Preprocesamiento de Datos

En el preprocesamiento de datos se realizaron los siguientes pasos:

Análisis General de los datos

Dataframe Winemag

Se obtienen del dataframe 'winemag-data-130k-v2.csv' desde Kaggle, el cual contiene:

- 129971 entradas y 14 columnas
- Las columnas son tipo objeto, categóricas a excepción del índice, la calificación del vino (80 a 100, tipo entero) y el precio.
- La mayoría de las columnas, excepto el índice, contienen entradas vacías, por lo cual requiere realizar una limpieza del Dataframe

Country: A nivel de país se encuentran 44 entradas de países. No se encuentran valores nulos

Description: Es una variable descriptiva del vino, que no será usada para este análisis

Designation: Es una descripción simple del vino, que no será usada para este análisis

Points: Es el puntaje (calificación) del vino, con valores enteros de 80 a 100, inclusive, de tipo entero. Esta es una de las variables de salida a analizar. Los valores nulos serán omitidos del análisis.

Price: Es el precio del vino en dólares, de tipo flotante. Esta es otra de las variables de salida a analizar. Los valores nulos serán omitidos.

Province, region_1, region_2: Es información donde se produce el vino, desde una provincia o región general a una más específica. Esta información será procesada para obtener la más detallada, y de no tenerla, se usará la general. También se complementa con la información de país, ya que existen varios sitios con el mismo nombre de "región" y se debe especificar para no tener valores erróneos. Si no se tienen datos de la provincia o región, los datos serán omitidos.

Taster_name: Es el nombre del catador o quien da la calificación del vino. Los valores nulos serán reemplazados por "Anónimo", ya que la información es importante. Debido a la cantidad de catadores (20), se realizará un one-hot enconding para el análisis.

Taster_twitter_handle: Este valor es analizado para ver si se puede encontrar información adicional de algún catador con datos nulos, pero en la revisión se encuentra que solo algunos de los catadores tienen la información de twitter o de resto no hay información. Como esta variable no proporciona ningún dato adicional y tiene una correlación directa con el nombre, no será usada para este análisis.

Title: Es el nombre o título del vino. Es usado para obtener el año de producción del vino. El resto de los valores serán omitidos para este análisis

Variety: Es el varietal o tipo de uva usada para la fabricación del vino. Debido a la gran cantidad de varietales y tipos (708 diferentes entradas, como tipo *blend*, o mezcla de diferentes cepas), esta variable no será usada para este análisis

Winery: Es el nombre de la casa que produce el vino. Debido a la gran cantidad de casas productoras (16757), la cual no aporta información del proceso, esta variable no será usada para este análisis.

Dataframe Data_Clima_Final

Desde "Global Climate Monitor", se obtienen los valores mensuales, desde el año 2000 hasta el año 2013 (es la información que está completa) de los siguientes parámetros:

- Temperatura Anual (Promedio)
- Temperatura Máxima Anual
- Temperatura Mínima Anual
- Precipitación Anual (Promedio)
- Evotranspiración Anual (Promedio)

Esta información se obtiene en dataframes separados que se consolidan, donde estos tienen en común: Año (agno) y coordenadas (formato WGS84), lo que permite hacer un "merge" de toda la información para así obtener un Dataframe final.

Las medidas de las variables obtenidas se tienen: temperatura: grados centígrados, precipitación anual: milímetros, evotranspiración: milímetros.

La información de localización de datos de GCM, se realiza de forma distribuida en las regiones, por lo que obtienen puntos equidistantes en los planos.

Comentarios Menores

- Se realiza una conversión de datos, de coordenadas WGS84 a dos columnas: Latitud y Longitud
- Se reemplaza, en 'country', el valor US a USA, para que sea correctamente analizado por la API de georreferenciación.
- Se emite el análisis del país Egipto porque no contiene información de precios.
- El análisis de años de producción del vino se ve reducido a los años 2000 a 2012 debido a la limitante en los datos de clima
- Se realiza *one-hot encoding* a las variables 'country' (44) y 'taster_name' (20)
- Se realiza una unión de parámetros: "region" + "," + "country", para complementar el dato de localización
- Se usa geocoders y GoogleV3 API como la aplicación de georreferenciación. Inicialmente se usó Nominatim, pero tenía fallas de búsqueda con algunas localizaciones.

- Se realiza la conversión y reemplazo de “region” por dos columnas numéricas: “Latitude” y “Longitude”
- Se procede a unir los dos dataframes finales, relacionando el año, y luego realizando un análisis para encontrar la ubicación de “región” más cercana a los datos provistos por GCM con sus respectivas variables.

Conclusiones

El dataframe obtenido, tiene dimensiones de 70’030 entradas validas con 70 columnas numéricas, sin valores nulos, con información completa de las variables de clima, latitud y longitud, calidad, precio, catador, país y año.

Análisis Exploratorio de Datos

Se despliega el dataset final para el Análisis Exploratorio de Datos (ó EDA, por sus siglas en inglés).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 70030 entries, 0 to 70029
Data columns (total 70 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   points                          70030 non-null  int64  
 1   price                           70030 non-null  float64
 2   Year                            70030 non-null  int64  
 3   Lat_x                           70030 non-null  float64
 4   Long_x                          70030 non-null  float64
 5   temp_anual                      70030 non-null  float64
 6   temp_max_anual                  70030 non-null  float64
 7   temp_min_anual                  70030 non-null  float64
 8   pre_anual                       70030 non-null  float64
 9   etp_anual                       70030 non-null  int64  
 10  country_Argentina               70030 non-null  uint8  
 11  country_Australia               70030 non-null  uint8  
 12  country_Austria                 70030 non-null  uint8  
 13  country_Bosnia and Herzegovina  70030 non-null  uint8  
 14  country_Brazil                  70030 non-null  uint8  
 15  country_Bulgaria                70030 non-null  uint8  
 16  country_Canada                  70030 non-null  uint8  
 17  country_Chile                   70030 non-null  uint8  
 18  country_China                   70030 non-null  uint8  
 19  country_Croatia                 70030 non-null  uint8  
 20  country_Cyprus                  70030 non-null  uint8  
 21  country_Czech Republic          70030 non-null  uint8  
 22  country_England                 70030 non-null  uint8  
 23  country_France                  70030 non-null  uint8  
 24  country_Georgia                 70030 non-null  uint8  
 25  country_Germany                 70030 non-null  uint8  
 26  country_Greece                  70030 non-null  uint8  
 27  country_Hungary                 70030 non-null  uint8  
 28  country_India                   70030 non-null  uint8  
 29  country_Israel                  70030 non-null  uint8  
 30  country_Italy                   70030 non-null  uint8  
 31  country_Lebanon                 70030 non-null  uint8  
 32  country_Macedonia               70030 non-null  uint8  
 33  country_Mexico                  70030 non-null  uint8  
 34  country_Moldova                 70030 non-null  uint8  
 35  country_Morocco                 70030 non-null  uint8  
 36  country_New Zealand             70030 non-null  uint8  
 37  country_Peru                    70030 non-null  uint8  
 38  country_Portugal                70030 non-null  uint8  
 39  country_Romania                 70030 non-null  uint8  
 40  country_Serbia                  70030 non-null  uint8  
 41  country_Slovakia                70030 non-null  uint8  
 42  country_Slovenia                70030 non-null  uint8  
 43  country_South Africa            70030 non-null  uint8  
 44  country_Spain                   70030 non-null  uint8  
 45  country_Switzerland             70030 non-null  uint8  
 46  country_Turkey                  70030 non-null  uint8  
 47  country_USA                     70030 non-null  uint8
```

```
48  country_Ukraine                70030 non-null  uint8
49  country_Uruguay                70030 non-null  uint8
50  taster_name_Alexander Peartree 70030 non-null  uint8
51  taster_name_Anna Lee C. Iijima 70030 non-null  uint8
52  taster_name_Anne Krebiehl MW   70030 non-null  uint8
53  taster_name_Anonimo            70030 non-null  uint8
54  taster_name_Carrie Dykes       70030 non-null  uint8
55  taster_name_Christina Pickard  70030 non-null  uint8
56  taster_name_Fiona Adams        70030 non-null  uint8
57  taster_name_Jeff Jenssen       70030 non-null  uint8
58  taster_name_Jim Gordon         70030 non-null  uint8
59  taster_name_Joe Czerwinski     70030 non-null  uint8
60  taster_name_Kerin O'Keefe      70030 non-null  uint8
61  taster_name_Lauren Buzzeo      70030 non-null  uint8
62  taster_name_Matt Kettmann      70030 non-null  uint8
63  taster_name_Michael Schachner  70030 non-null  uint8
64  taster_name_Mike DeSimone      70030 non-null  uint8
65  taster_name_Paul Gregutt       70030 non-null  uint8
66  taster_name_Roger Voss         70030 non-null  uint8
67  taster_name_Sean P. Sullivan   70030 non-null  uint8
68  taster_name_Susan Kostrzewa    70030 non-null  uint8
69  taster_name_Virginie Boone     70030 non-null  uint8
```

	points	price	Year	Lat_x	Long_x	temp_anual	temp_max_anual	temp_min_anual	pre_anual	etp_anual	country_Arg
0	87	15.0	2011	41.75	-5.75	13.01	19.44	6.61	388.5	1200	0
1	87	15.0	2011	41.75	-5.75	13.01	19.44	6.61	388.5	1200	0
2	87	17.0	2011	41.75	-5.75	13.01	19.44	6.61	388.5	1200	0
3	91	12.0	2011	41.75	-5.75	13.01	19.44	6.61	388.5	1200	0
4	87	8.0	2011	41.75	-5.75	13.01	19.44	6.61	388.5	1200	0
...	...	...	...	...	...	...	...	...	...	...	...
70025	84	25.0	2012	34.75	-118.25	13.96	20.23	7.75	325.0	1176	0
70026	90	21.0	2012	46.75	6.75	9.15	13.46	4.87	1382.5	696	0
70027	89	14.0	2012	45.25	6.25	5.61	9.20	2.05	1386.2	714	0
70028	89	18.0	2012	45.75	5.75	10.71	15.25	6.22	976.2	807	0
70029	87	25.0	2012	40.25	15.25	15.09	18.80	11.40	269.6	1077	0

70030 rows × 70 columns

EDA ANALYSIS

Statistical Theory

Count

Is the total count of values in a variable or column. This value is usefull to check it against the unique values to understand the ratio for the total entries. Using One-Hot encoding will use the same result with the sum of the values.

- df['column'].count()
- df['column'].nunique()
- df.nunique()

Maximum

Is the maximum value of a variable. This allows to understand if there are outliers and if this case, it is suggested to do a graphic checking this values. This does not apply for binary variables (one-hot encoding)

- df['column'].max()
- df.max()

Minimum

Is the minimum value of a variable. This allows to understand if there are outliers and if this case, it is suggested to do a graphic checking this values. This does not apply for binary variables (one-hot encoding)

- df['column'].min()
- df.min()

Mean (simple average)

Is the statistical measurment of the mean average value. This value is calculated as the sum of all the values, divided by the total number (count) of the values. This does not apply for binary variables (one-hot encoding)

- df['column'].mean()
- df.mean()

Median

Is the statistical measurement of the median average value. This value is calculated as the middle value of all the items placed in an array, if it is odd, or the sum divided by two of the two middle values, it it is odd. This measurement is not affected by outliers. This does not apply for binary variables (one-hot encoding)

- df['column'].median()
- df.median()

Mode

Is the statistical measurment of the most frequent or often value. This value is calculated as the value that is most repeated in an array. This measurement is not affected by outliers. This does not apply for binary variables (one-hot encoding)

- df['column'].mode()
- df.mode()

Skewness/Symmetry or Balanced Information

This allows to see if the information is skewed or not balanced, for this:

- if mean > median -> Skewed to the right: Outliers to the right
- mean ~ median -> Symmetrical (no significant or balanced outliers)
- mean < median -> Skewed to the left: Outliers to the left
- Analize mean vs median values

Variance (sigma^2 or s^2)

Is the statistical measurement of the data dispersion around the mean values. It is calculated as sigma ^2, that is the delta between each value and the median elevated to the square and divided by the population count (if it is a sample, then it is divided by n-1). It is elevated to the square so all deltas are positive, thus, it gives us a number on how "dispersed" are the values. This is affected by outliers, specially high number outliers.

- df['column'].var()

Standard Deviation (sigma or s)

Is more meaningful than the variance. It is root square of the variance, thus, it has the same unit of measurement than the sample. This allows to give meaning to the value

- df['column'].std()

Coeficient of Variation (CV) or Relative Standard Deviation

It is the standard deviation divided by the mean of the values. This allows to compare the variation of values between different units (has no unit of measurement)

Measures of Relationship between variables

Covariance (Cov(x,y))

Can be positive, zero or negative. It is the measurement on how two variables are related , meaning that while one variable grows, the second one grows, shrinks or has no change.

> _> 0 -> The two variables move together (direct relationship)
>
> _< 0 -> The two variables move in oppositve direction (inverse relationship)
>
> _~ 0 -> The variables are independent

It is calculated by the multiplication of the delta of each X and Y value with its mean value, and then divided by the sample size - 1 (or population size).

(Xi-Xmean)x(Yi-Ymean)/(n-1)

The problem of the Covariance is that the value can be anything and has no relative meaning

- df['column'].cov()

Correlation

It adjust the covariance, by regularizing it to a value between [-1 to 1], following the similar analysis of covariance, but between a certain range.

It is calculated as the covariance of two variables, divided by the product of the standard deviation of both variables. This will always be a number between -1 (imperfect or inverse correlation), 0 (no relationship) and 1 (perfect or direct relationship).

Cov(x,y) / (s(x) x s(y))

- df['column'].corr()

Quantiles (25% 50% 75% 90% 99%)

It is the analysis of the values in the amount of the percentage of the count of the values, ordered by size. This allows to understand where are distributed the values along the total amount of entries

- df['column'].quantile()

Simpson's Paradox

One or two conclusions cannot bias a third one. For example, there are more fatalities when an helicopter is used rather when an ambulance is used. This does not mean that the helicopter is more dangerous, but is because the cases where an helicopter is used is in very complicated cases or risk sites.

Sesgado

This is the error due to: Recolection, Analysis, Interpretation or revision of the data. The first conclusions of this analysis was "sesgado"

Correlación

Is a Linear Relationship or proportion between two variables

In [1]:
```
#Importamos las librerías y seteamos opciones generales
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

pd.set_option("display.max_rows", 100)
pd.set_option("display.max_rows", None)
plt.style.use('seaborn-whitegrid')
```

In [2]:
```
#Se lee el dataframe final
df = pd.read_csv("Dataframe_final")
dfr = pd.read_csv('Wine_reviews_climate_prediction.csv')
```

In [3]:
```
#Se obtiene la información del dataframe
df.shape
```

Out[3]:
```
(70030, 71)
```

In [4]:
```
#Se procede a ver los typos de datos de cada columna del dataframe, verificand
o que no existan
#valores nulos. Se encuentran solo variables numéricas
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70030 entries, 0 to 70029
Data columns (total 71 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Unnamed: 0                       70030 non-null  int64  
 1   points                           70030 non-null  int64  
 2   price                            70030 non-null  float64
 3   Year                             70030 non-null  int64  
 4   Lat_x                            70030 non-null  float64
 5   Long_x                           70030 non-null  float64
 6   temp_anual                       70030 non-null  float64
 7   temp_max_anual                   70030 non-null  float64
 8   temp_min_anual                   70030 non-null  float64
 9   pre_anual                        70030 non-null  float64
 10  etp_anual                        70030 non-null  int64  
 11  country_Argentina                70030 non-null  int64  
 12  country_Australia                70030 non-null  int64  
 13  country_Austria                  70030 non-null  int64  
 14  country_Bosnia and Herzegovina   70030 non-null  int64  
 15  country_Brazil                   70030 non-null  int64  
 16  country_Bulgaria                 70030 non-null  int64  
 17  country_Canada                   70030 non-null  int64  
 18  country_Chile                    70030 non-null  int64  
 19  country_China                    70030 non-null  int64  
 20  country_Croatia                  70030 non-null  int64  
 21  country_Cyprus                   70030 non-null  int64  
 22  country_Czech Republic           70030 non-null  int64  
 23  country_England                  70030 non-null  int64  
 24  country_France                   70030 non-null  int64  
 25  country_Georgia                  70030 non-null  int64  
 26  country_Germany                  70030 non-null  int64  
 27  country_Greece                   70030 non-null  int64  
 28  country_Hungary                  70030 non-null  int64  
 29  country_India                    70030 non-null  int64  
 30  country_Israel                   70030 non-null  int64  
 31  country_Italy                    70030 non-null  int64  
 32  country_Lebanon                  70030 non-null  int64  
 33  country_Macedonia                70030 non-null  int64  
 34  country_Mexico                   70030 non-null  int64  
 35  country_Moldova                  70030 non-null  int64  
 36  country_Morocco                  70030 non-null  int64  
 37  country_New Zealand              70030 non-null  int64  
 38  country_Peru                     70030 non-null  int64  
 39  country_Portugal                 70030 non-null  int64  
 40  country_Romania                  70030 non-null  int64  
 41  country_Serbia                   70030 non-null  int64  
 42  country_Slovakia                 70030 non-null  int64  
 43  country_Slovenia                 70030 non-null  int64  
 44  country_South Africa             70030 non-null  int64  
 45  country_Spain                    70030 non-null  int64  
 46  country_Switzerland              70030 non-null  int64  
 47  country_Turkey                   70030 non-null  int64  
 48  country_USA                      70030 non-null  int64  
 49  country_Ukraine                  70030 non-null  int64  
 50  country_Uruguay                  70030 non-null  int64  
 51  taster_name_Alexander Peartree   70030 non-null  int64
```

```
 52  taster_name_Anna Lee C. Iijima  70030 non-null  int64
 53  taster_name_Anne Krebiehl MW    70030 non-null  int64
 54  taster_name_Anonimo             70030 non-null  int64
 55  taster_name_Carrie Dykes        70030 non-null  int64
 56  taster_name_Christina Pickard   70030 non-null  int64
 57  taster_name_Fiona Adams         70030 non-null  int64
 58  taster_name_Jeff Jenssen        70030 non-null  int64
 59  taster_name_Jim Gordon          70030 non-null  int64
 60  taster_name_Joe Czerwinski      70030 non-null  int64
 61  taster_name_Kerin O’Keefe       70030 non-null  int64
 62  taster_name_Lauren Buzzeo       70030 non-null  int64
 63  taster_name_Matt Kettmann       70030 non-null  int64
 64  taster_name_Michael Schachner   70030 non-null  int64
 65  taster_name_Mike DeSimone       70030 non-null  int64
 66  taster_name_Paul Gregutt        70030 non-null  int64
 67  taster_name_Roger Voss          70030 non-null  int64
 68  taster_name_Sean P. Sullivan    70030 non-null  int64
 69  taster_name_Susan Kostrzewa     70030 non-null  int64
 70  taster_name_Virginie Boone      70030 non-null  int64
dtypes: float64(7), int64(64)
memory usage: 37.9 MB
```

In [5]:
```
#verificamos los valores únicos en el dataframe
df.nunique()
```

Out[5]:

Unnamed: 0	70030
points	21
price	344
Year	13
Lat_x	87
Long_x	223
temp_anual	1323
temp_max_anual	1392
temp_min_anual	1334
pre_anual	3371
etp_anual	435
country_Argentina	2
country_Australia	2
country_Austria	2
country_Bosnia and Herzegovina	2
country_Brazil	2
country_Bulgaria	2
country_Canada	2
country_Chile	2
country_China	2
country_Croatia	2
country_Cyprus	2
country_Czech Republic	2
country_England	2
country_France	2
country_Georgia	2
country_Germany	2
country_Greece	2
country_Hungary	2
country_India	2
country_Israel	2
country_Italy	2
country_Lebanon	2
country_Macedonia	2
country_Mexico	2
country_Moldova	2
country_Morocco	2
country_New Zealand	2
country_Peru	2
country_Portugal	2
country_Romania	2
country_Serbia	2
country_Slovakia	2
country_Slovenia	2
country_South Africa	2
country_Spain	2
country_Switzerland	2
country_Turkey	2
country_USA	2
country_Ukraine	2
country_Uruguay	2
taster_name_Alexander Peartree	2
taster_name_Anna Lee C. Iijima	2
taster_name_Anne Krebiehl MW	2
taster_name_Anonimo	2
taster_name_Carrie Dykes	2
taster_name_Christina Pickard	2

```
taster_name_Fiona Adams              2
taster_name_Jeff Jenssen             2
taster_name_Jim Gordon               2
taster_name_Joe Czerwinski           2
taster_name_Kerin O'Keefe            2
taster_name_Lauren Buzzeo            2
taster_name_Matt Kettmann            2
taster_name_Michael Schachner        2
taster_name_Mike DeSimone            2
taster_name_Paul Gregutt             2
taster_name_Roger Voss               2
taster_name_Sean P. Sullivan         2
taster_name_Susan Kostrzewa          2
taster_name_Virginie Boone           2
dtype: int64
```

Con esto verificamos que existen:

- 21 valores de puntaje (80 a 100 inclusive)
- 13 valores de años (2000 a 2012 inclusive)
- Valores Binarios para los datos de one-hot encoding

In [6]:
```
#los valores máximos que pueden ser outliers o erroneos
df.iloc[:,:11].max()
```

Out[6]:
```
Unnamed: 0        70029.00
points              100.00
price              2500.00
Year               2012.00
Lat_x                52.25
Long_x              176.75
temp_anual           26.91
temp_max_anual       33.09
temp_min_anual       21.70
pre_anual          2791.70
etp_anual          2106.00
dtype: float64
```

Se encuentran valores aceptables en todas las variables. Para el caso de precio es posible tener un vino de alto precio (USD $2200), valores de temperatura correctos y de precipitación anual

In [7]:
```
#Los valores mínimos que pueden ser outliers o erroneos
df.iloc[:,:11].min()
```

Out[7]:
```
Unnamed: 0          0.00
points             80.00
price               4.00
Year             2000.00
Lat_x             -45.25
Long_x           -123.75
temp_anual          -0.88
temp_max_anual       2.75
temp_min_anual      -9.80
pre_anual            9.00
etp_anual          378.00
dtype: float64
```

No se encuentran valores mínimos erroneos o atípicos. Se encuentran valores de al menos USD $4 como precio mínimo lo que es correcto, valores de temperatura correctos y de precipitación anual

In [8]:
```
#Los valores promedio (mean)
df.iloc[:,:11].mean()
```

Out[8]:
```
Unnamed: 0        35014.500000
points               88.313965
price                37.902842
Year               2009.066700
Lat_x                32.218406
Long_x              -57.084014
temp_anual           12.988201
temp_max_anual       18.630808
temp_min_anual        7.390167
pre_anual           779.388079
etp_anual          1046.453506
dtype: float64
```

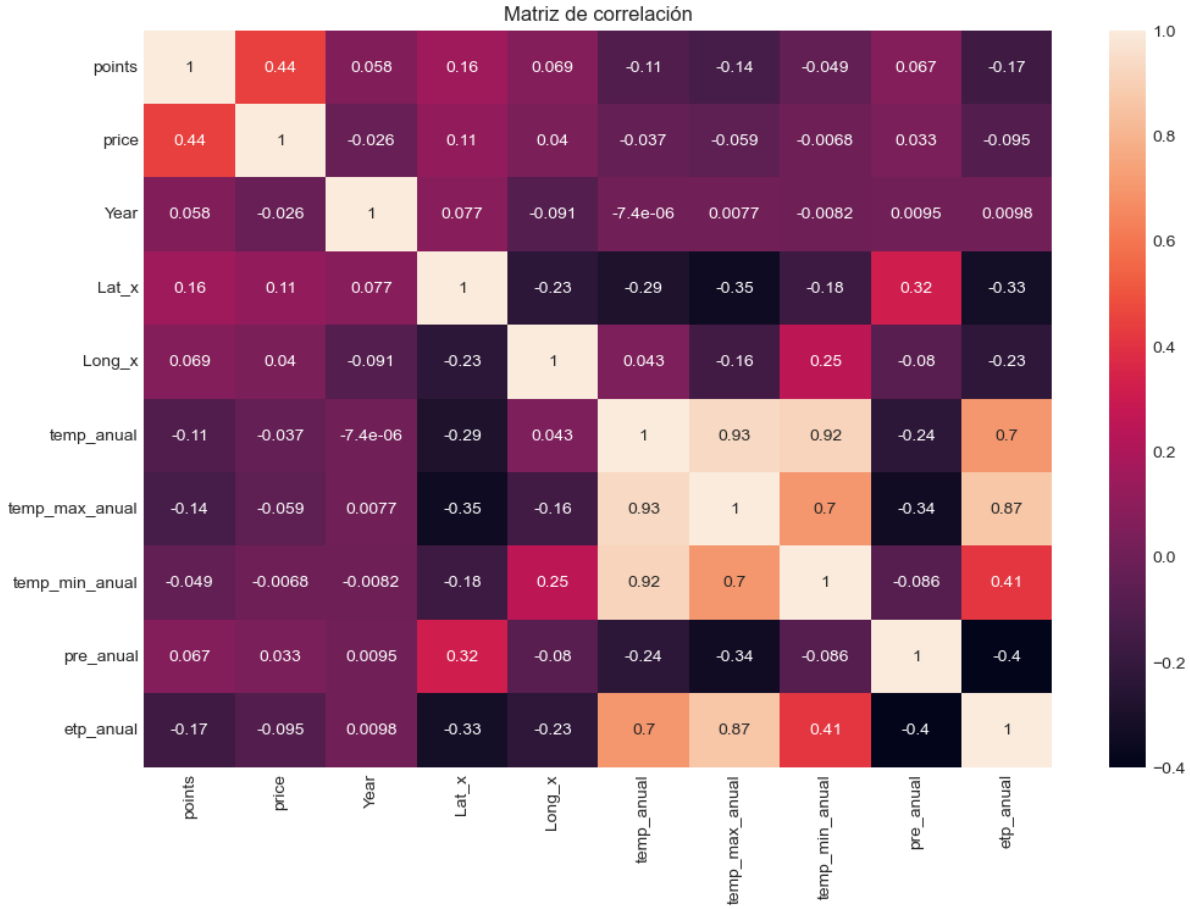
- En este caso el promedio de los puntos está en la parte baja pero no muy alejada de la medida central "90"
- El promedio del precio se sitúa en la parte baja de los valores extremos, por lo que se puede inferir que la mayoría de vinos calificados son de precio relativamente bajo
- El promedio del año es de 2009, lo que quiere decir que la mayoría de vinos calificados están con una vida de al menos 8 años (winemag es un dataframe del año 2017)
- El promedio de la temperatura promedio se situa en 13 grados, que es una temperatura relativamente baja
- El promedio de la temperatura máxima anual es de 18 grados, que es una temperatura relativamente baja
- El promedio de la temperatura mínima anual es de 7 grados, que es una temperatura relativamente baja
- El promedio de la precipitación anual es de 7,8 cm, lo que es una precipitación relativamente baja

In [9]:
```
#La mediana de los valores (median)
df.iloc[:,:11].median()
```

Out[9]:
```
Unnamed: 0        35014.50
points               88.00
price                28.00
Year               2010.00
Lat_x                39.75
Long_x              -74.25
temp_anual           13.61
temp_max_anual       18.87
temp_min_anual        7.52
pre_anual           737.40
etp_anual          1047.00
dtype: float64
```

- La mediana del puntaje es muy cercano a su promedio, por lo cual es una variable balanceada
- La mediana de precio es de USD $28, mucho menor que su promedio, por lo cual inferimos que tenemos valores de precios bastante altos que afectaron la mediana y tiene cierto desbalanceo
- La mediana de años es de 2010, lo que posiblemente indique que se tiene un pqueño desbalanceo hacia los valores bajos (es decir, se tienen más vinos que son más añejos que recientes)
- La mediana de la temperatura promedio anual es muy cercana al promedio de la temperatura promedio, por lo cual inferimos que se tienen valores balanceados
- La mediana de la temperatura máxima anual es muy cercana al promedio de la temperatura máxima, por lo cual inferimos que se tienen valores balanceados
- La mediana de la temperatura mínima anual es muy cercana al promedio de la temperatura mínima, por lo cual inferimos que se tienen valores balanceados
- La mediana de la precipitación anual es relativamente cercana al promedio de la precipitación promedio, por lo cual inferimos que se tienen valores balanceados
- La mediana de la evotranspiración anual es muy cercana al promedio de la evotranspiración promedio, por lo cual inferimos que se tienen valores balanceados

In [10]:
```
df_simple = df.iloc[:,1:11]
corrMatrix = df_simple.corr()
plt.figure(figsize=(12,8), dpi= 100)
sns.heatmap(corrMatrix, annot=True)
plt.title("Matriz de correlación")
plt.show()
```



En la matriz de correlación se analiza la correlación lineal de las variables.

Conclusiones Entre las vairables más interesantes en la matriz, se encuentran:

Puntos Cierta correlación con el precio, una baja correlación con la Latitud y una correlación baja inversa con la temperatura anual máxima y evotranspiración anual.

Precio Cierta corrleación con la Latitud

Año No se encuentra correlación alguna

Latitud Alguna correlación inversa con la temperatura promedio y máxima, y una correlación con la precipitación anual

Longitud Alguna correlación con la temperatura mínima anual y parte de correlación inversa con la precipitación anual

temp anual promedio Correlación con la temperatura máxima anual y mínima anual, así como con la evotranspiración, y alguna correlación inversa con la Latitud

temp max anual Correlación con la temperatura mínima anual y la evotranspiración. Correlación inversa con la Latitud

temp min anual Alguna Correlación con la evotranspiración y Longitud

Precipitación anual Correlaciones con las Latitudes, y correlación inversa con la temperatura máxima

evotranspiración anual Correlacion inversa con la Latitud, y correlaciones con la temperatura anual promedio, temperaturas máximas y mínimas

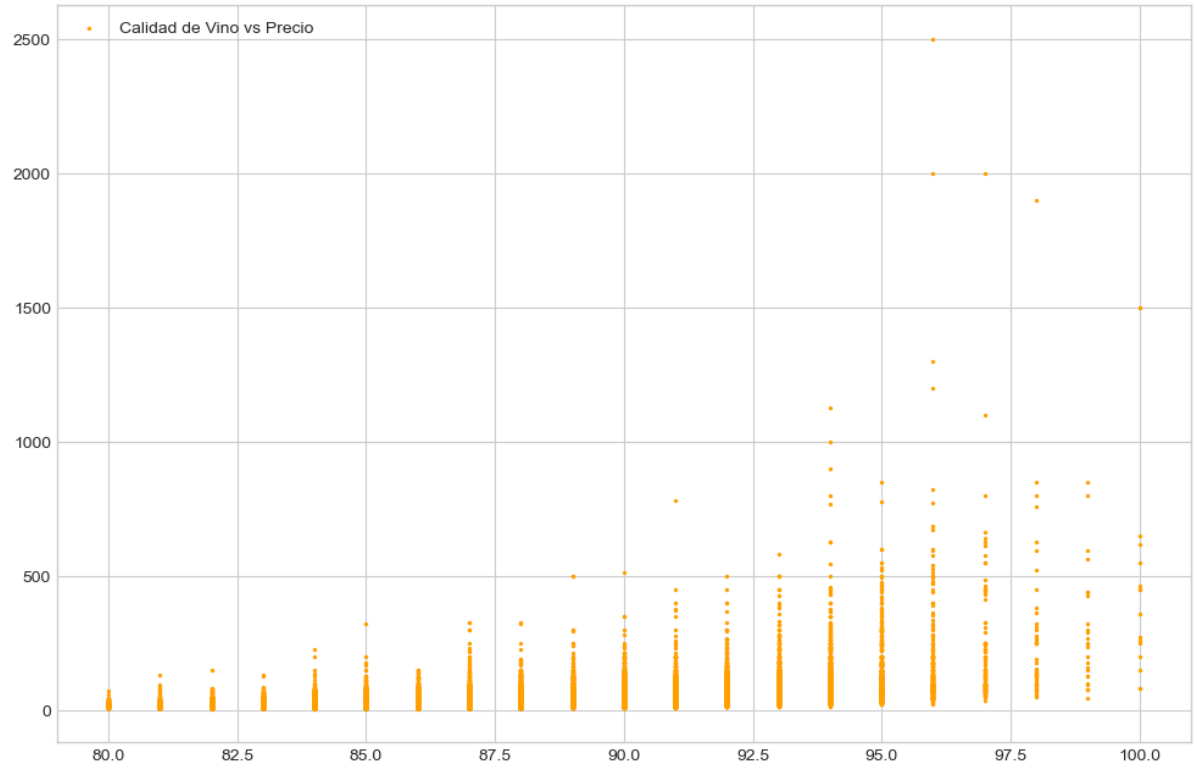
En las siguietnes gráficas, se verifican las gráficas de correlaciones entre variables, para comprender más su interrelación

In [11]:
```
#pd.plotting.scatter_matrix(df, alpha = 0.3, figsize = (16,12), diagonal = 'kd
e');
```

Relaciones Entre Variables

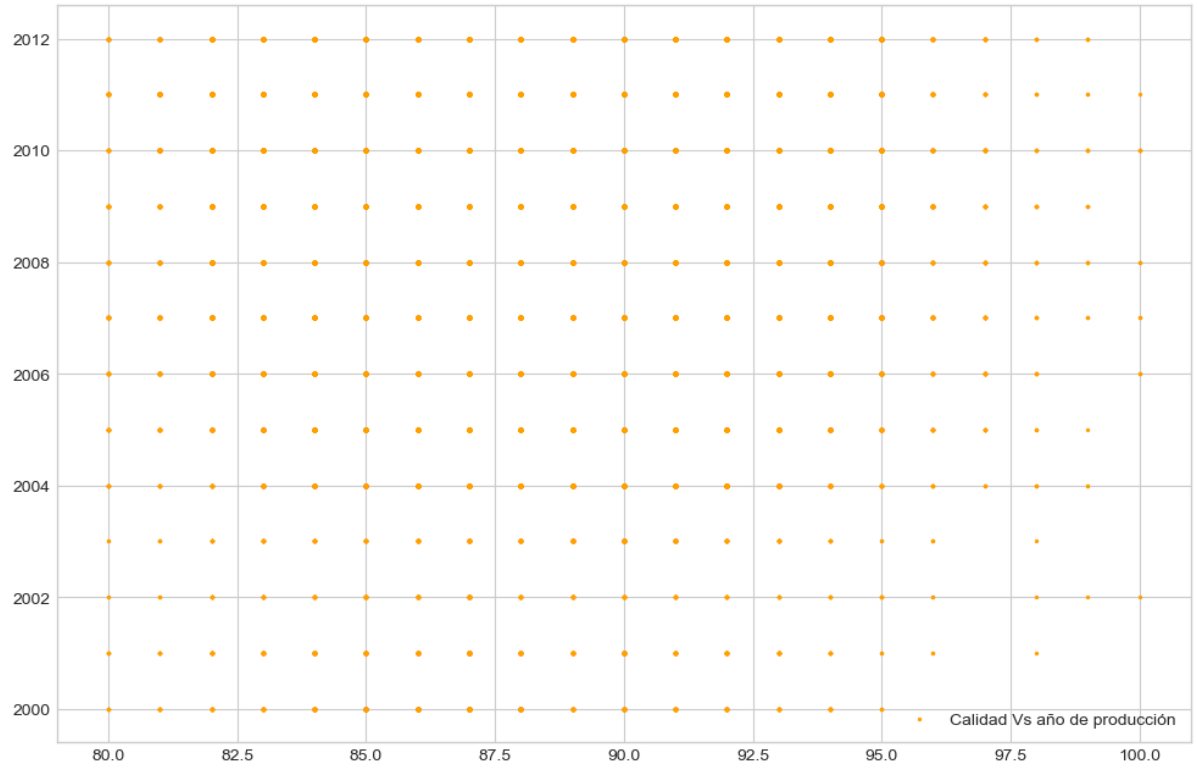
En este capítulo se realiza la comparación de cada una de las variables, para encontrar si existe una relación no lineal.

In [12]:
```
#Calidad vs. Precio
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['price'], s=2, label="Calidad de Vino vs Precio",
color="#FFA000")
plt.legend();
plt.show()
```



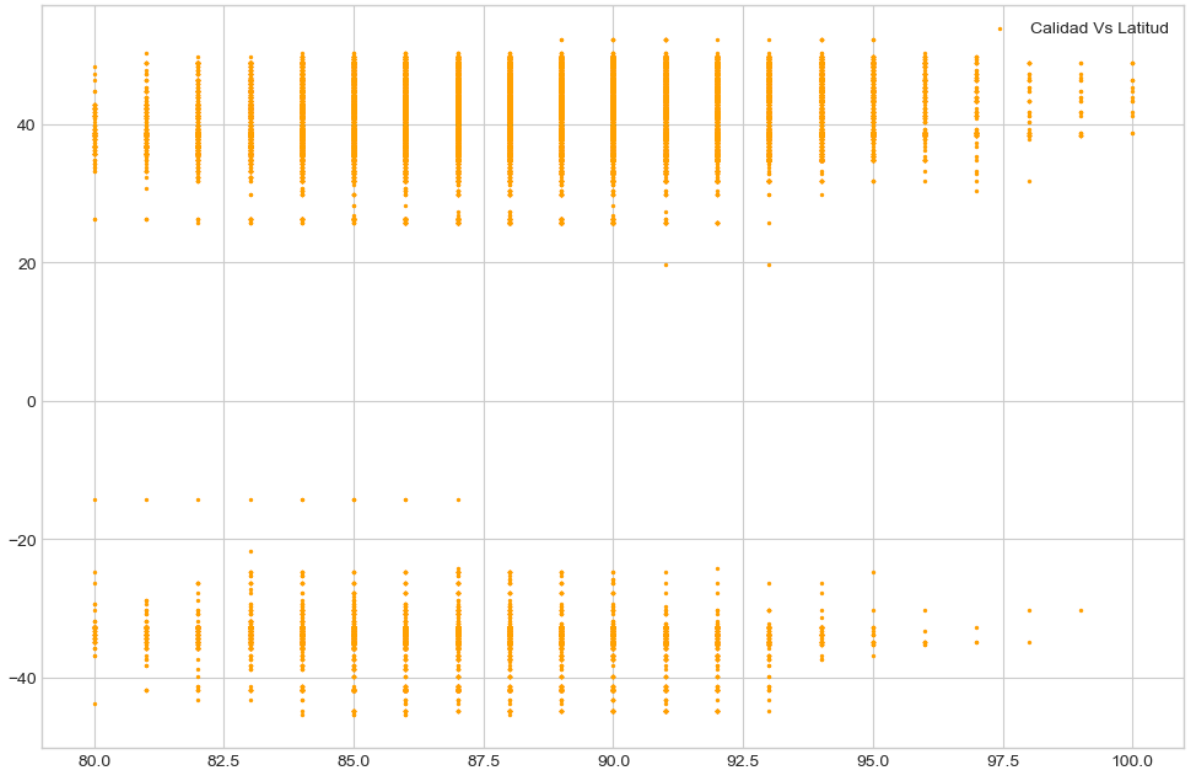
Como se aprecia, se pueden ver una correlación leve como lo mostraba la matrix de correlación. También se encuentran alguno valores atìpicos que no inciden en este análisis, pero si se encuentra que a mayor calificación, los precios de los vinos tienden a aumentar.

In [13]:
```
#Calidad vs. Año
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['Year'], s=2, label="Calidad Vs año de producció
n", color="#FFA000")
plt.legend();
plt.show()
```



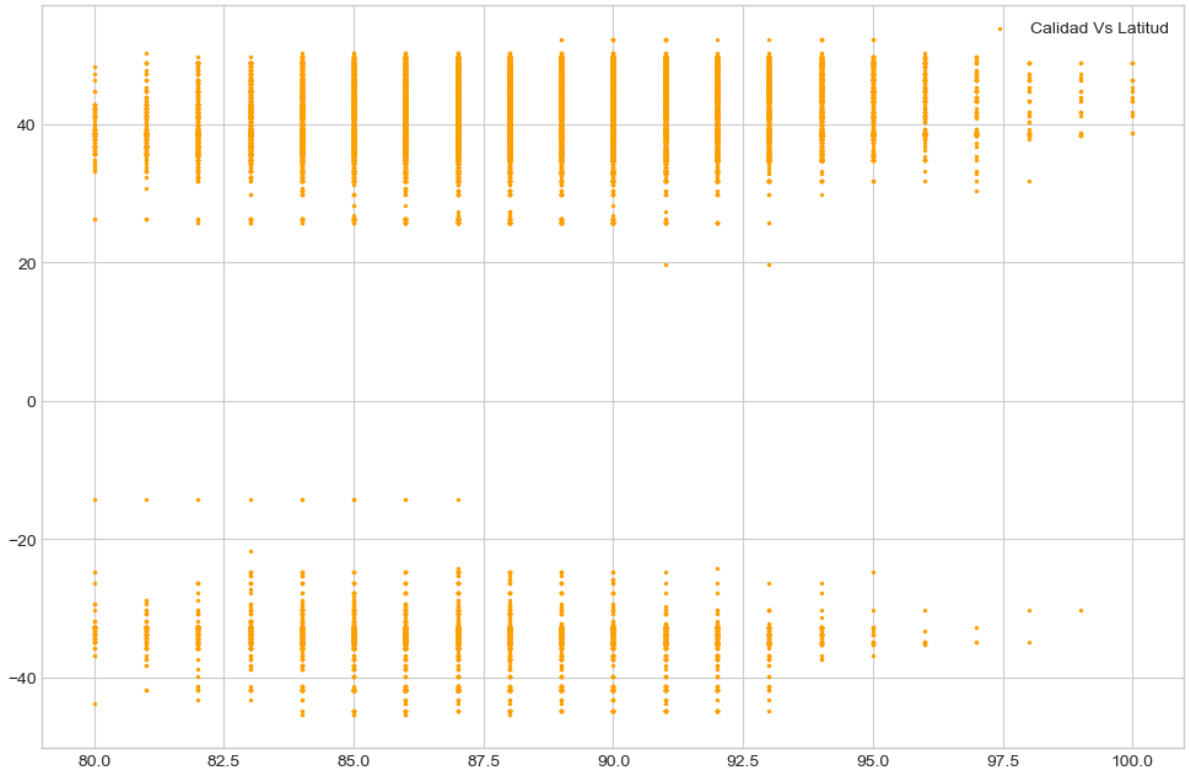
En el caso de la calidad vs años, no hay una clara tendencia, se puede alcanzar a apreciar que para los años 2006 y 2008 se tienen varios casos de muy buenos vinos.

In [14]:
```
#Calidad vs. Latitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['Lat_x'], s=2, label="Calidad Vs Latitud", color
="#FFA000")
plt.legend();
plt.show()
```

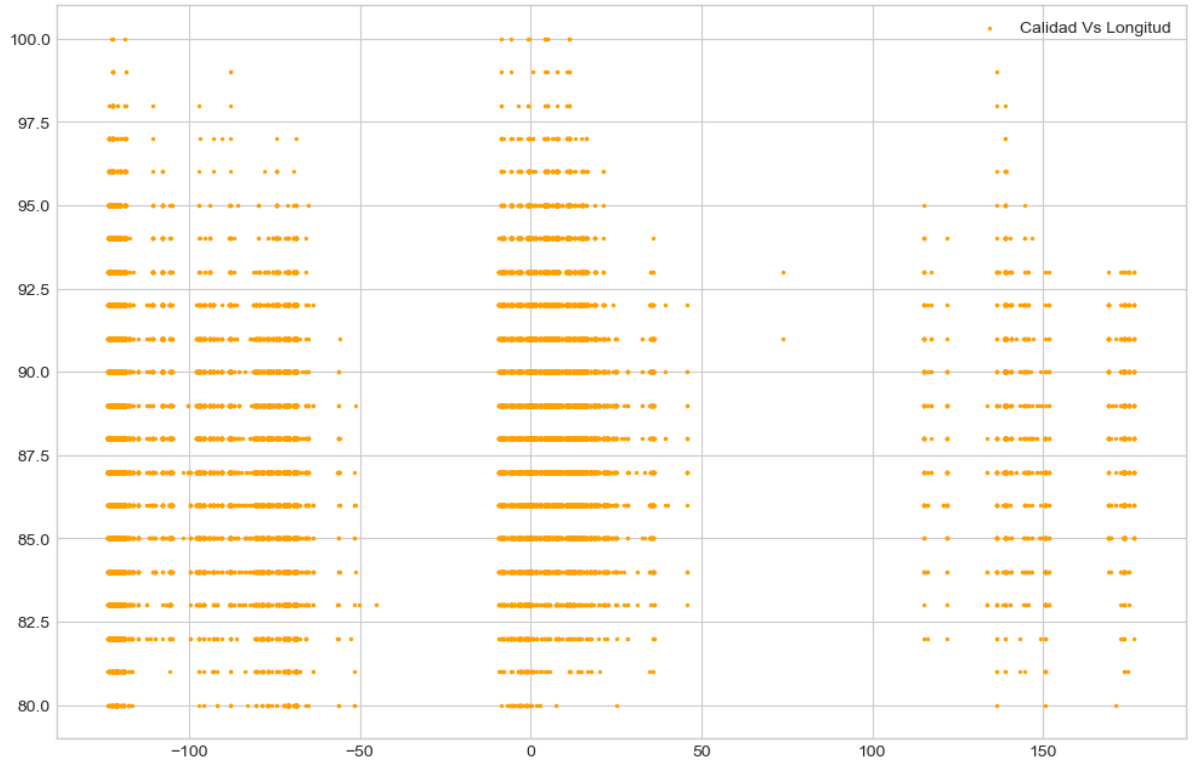


Se aprecia que los valores se agrupan en diferentes grandes subgrupos; por lo que el análisis se debe realizar en estos dos sub-grupos, parte positiva y parte negativa. Más abajo se realiza este análisis a detalle.

In [15]:
```
#Calidad vs. Latitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['Lat_x'], s=2, label="Calidad Vs Latitud", color
="#FFA000")
plt.legend();
plt.show()
```

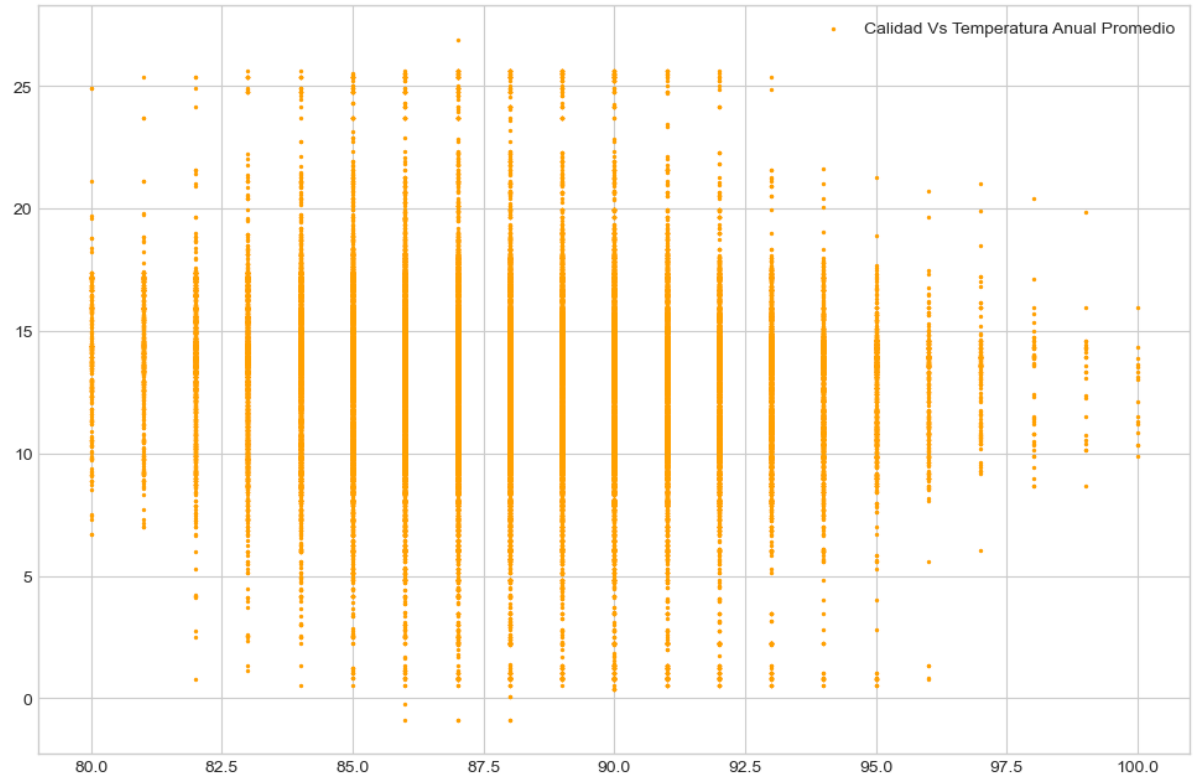


In [16]:
```
#Calidad vs. Longitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['points'], s=2, label="Calidad Vs Longitud", colo
r="#FFA000")
plt.legend();
plt.show()
```



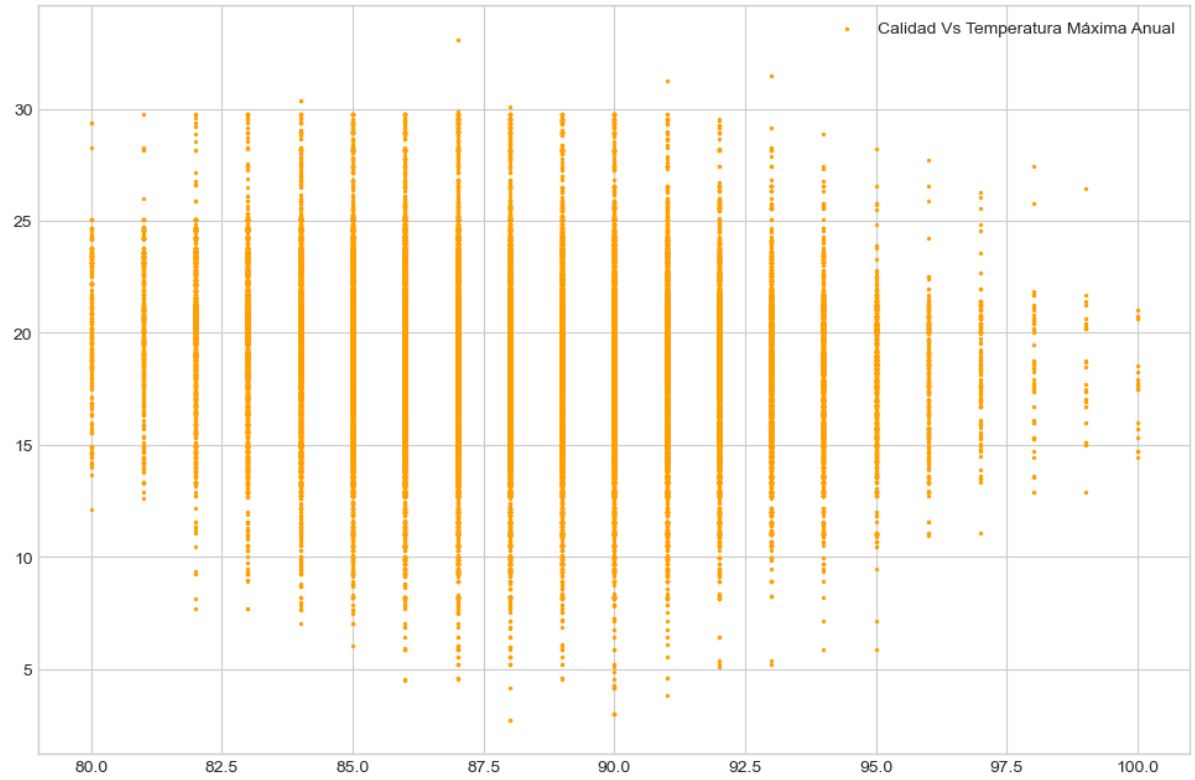
Para la Calidad contra la longitud, se representa mejor la longitud en el eje X, se tiene un tema similar que el anterior, ya que se debe analizar es por zonas y no en general, aunque se encuentra cierta clusterización en las zonas. ¿Sería recomendable separar estas zonas en diferentres nuevas columnas y realizar nuevamente un análisis de correlación?

In [17]:
```
#Calidad vs. Temperatura anual Promedio
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_anual'], s=2, label="Calidad Vs Temperatura
Anual Promedio", color="#FFA000")
plt.legend();
plt.show()
```



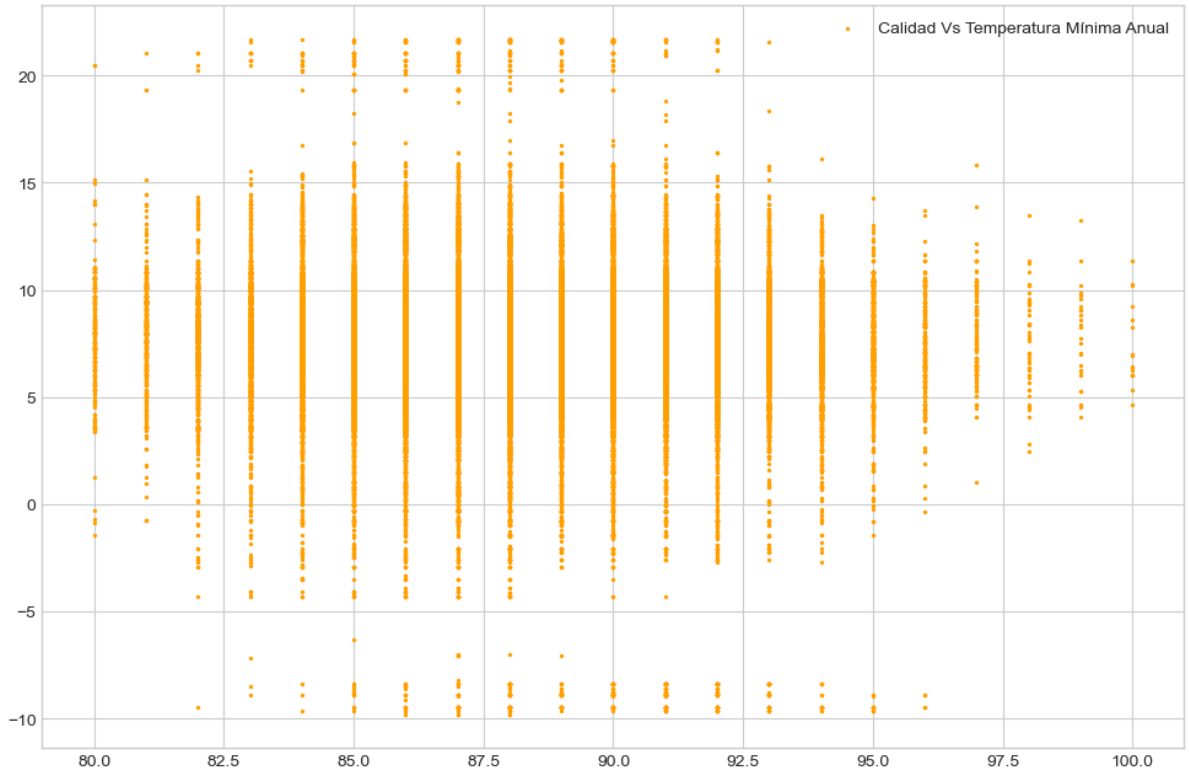
En este caso, se ve una leve relación de la calidad con la temperatura. Este análisis se puede ver mejor en un histograma, para así poder ver en donde existe mayor densidad y así intentar determinar la mejor temperatura promedio, sobre todo en puntajes muy altos. Si nos centramos en la calidad de 98 a 100, las temperaturas ideales parecen estar entre los 9 y 16 grados de promedio.

In [18]:
```
#Calidad vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_max_anual'], s=2, label="Calidad Vs Tempera
tura Máxima Anual", color="#FFA000")
plt.legend();
plt.show()
```



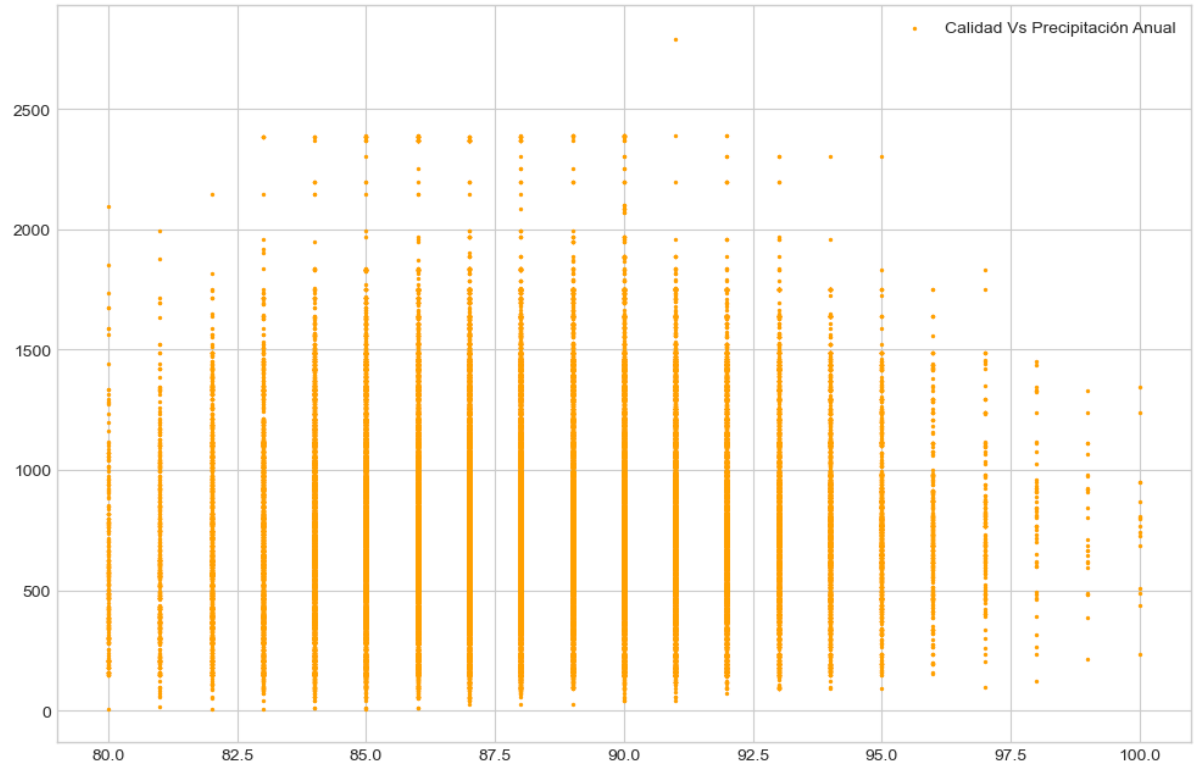
Similar como se aprecia en la temperatura promedio, se encuentra una leve correlación, principalmente en los vinos de alta calidad, donde los valores de temperatura máxima ideal se encuentran alrededor de los 14 y 22 grados centigrados en promedio

In [19]:
```
#Calidad vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_min_anual'], s=2, label="Calidad Vs Tempera
tura Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```



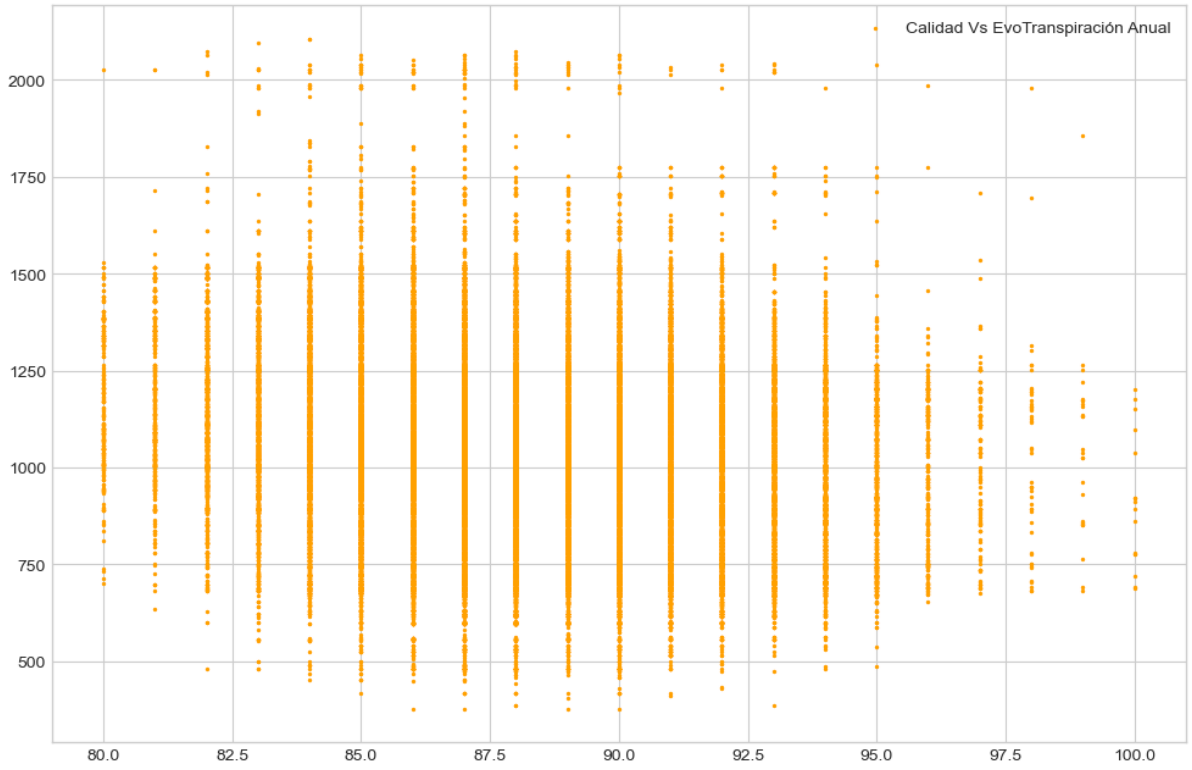
Nuevamente se encuentra cierta correlación con la temperatura mínima anual. En este caso, los valores ideales se encuentran entre los 5 y 11 grados centigrados. Es imporante remarcar que es posible sembrar vinos con diversas temperaturas extremas, el objetivo de este análisis es encontrar las variables que más favorecen a tener un vino de excelente calidad.

In [20]:
```
#Calidad vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['pre_anual'], s=2, label="Calidad Vs Precipitació
n Anual", color="#FFA000")
plt.legend();
plt.show()
```



En el caso de la precipitación anual, se encuentra también una clusterización importante, teniendo valores entre 250mm y 1400mm de precipitación anual. Esto permite inferir que es importante la precipitación, sin embargo lugares con muy poca precipitación (desérticos) o con mucha lluvia no son aptos para tener buenas cepas de calidad de vino (aunque esto puede ser también indicio de la paradoja de Simpson, donde nuestras muestras esten en ciertos sectores especificos y lleven a una conclusión incorrecta)

In [21]:
```
#Calidad vs. EvoTranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['etp_anual'], s=2, label="Calidad Vs EvoTranspira
ción Anual", color="#FFA000")
plt.legend();
plt.show()
```



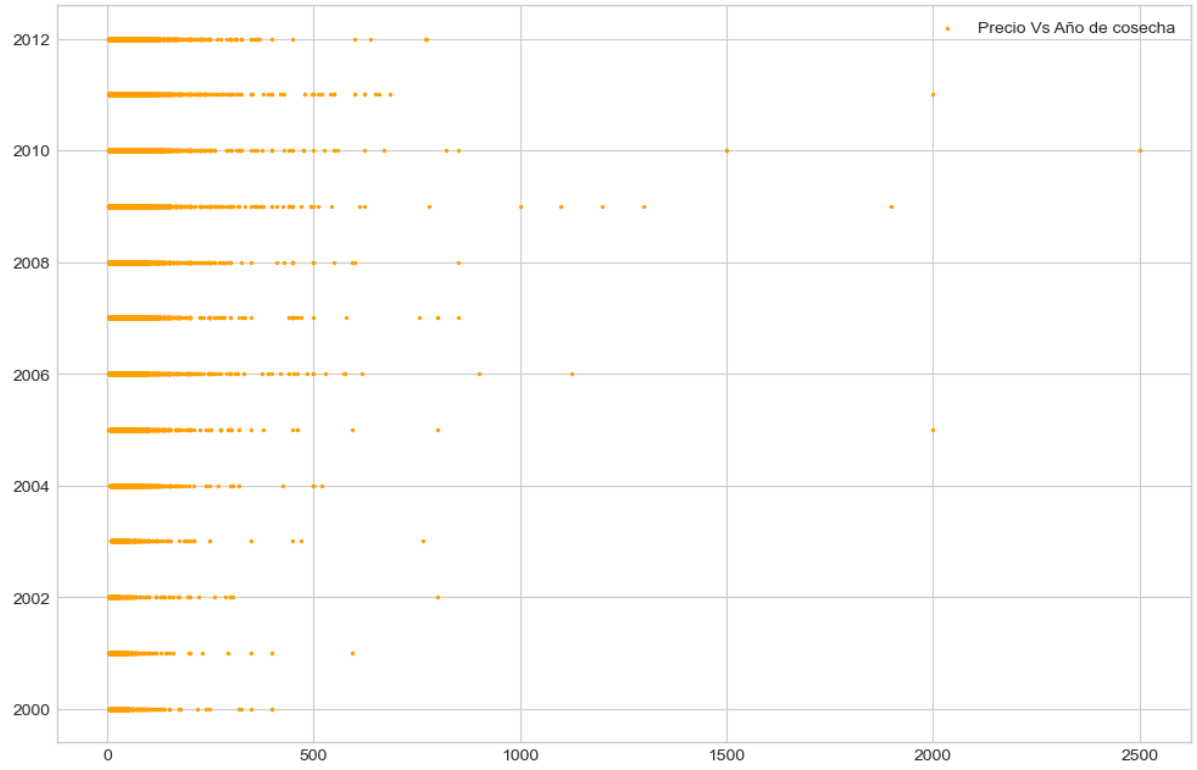
Se encuentra una clusterización interesante en la EvoTranspiración Anual. Es importante redoordar el concepto de evotranspiración: "La evaporación es el fenómeno físico en el que el agua pasa de líquido a vapor, producido desde:

- La superficie del suelo y vegetación de forma inmediata luego de la precipitación
- Desde la superficie del agua (en este caso no aplica)
- Desde el suelo, agua filtrada que se evapora desde la parte superficial del suelo La transpiración es el fenómeno biológico por el que las plantas pierden agua a la atmósfera. Toman parte del agua con sus raíces, el cual una parte se usa para su crecimiento y el resto la transpiran.

Como ambas mediciones son dificiles de realizar por separado, y en la mayoría de los casos lo más importante es la cantidad total de agua que se pierde a la atmosfera, se considera conjuntamente bajo el concepto de EvoTranspiración. Basicamente se resume en centrar la cuantificación de recursos hidrológicos de un área; lo que llueve, menos lo que se *evotranspira*, es el volumen de agua disponible"

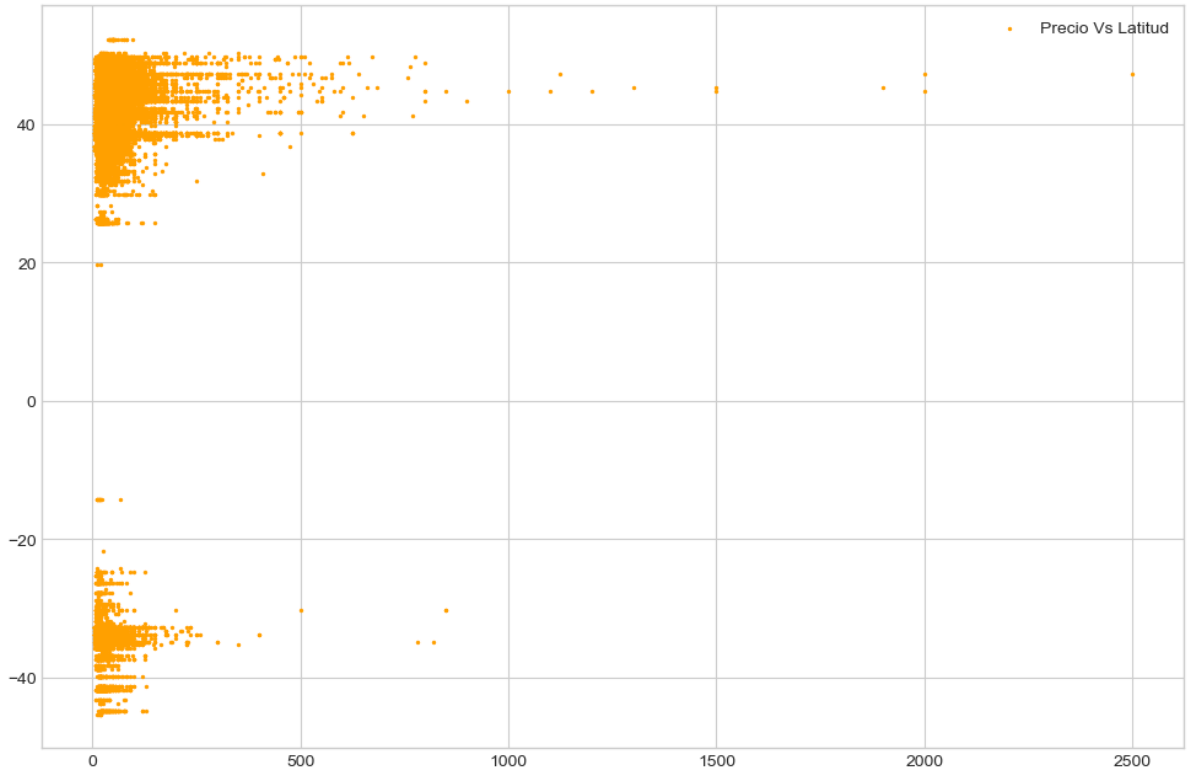
Con este valor, lo que obtenemos es la cantidad de agua que se requiere para un correcto cultivo y desarrollo de los viñedos; es aun más interesante encontrar que los valores de evotranspiración para los vinos de alta calidad oscilan entre los 650mm y 1300mm. También se encuentran algunos valores "atípicos" alrededos de los 1400mm a 1900mm, para esto se propone enfocar un nuevo análisis solo en estos valores vs la cepa del vino, para así entender si esto se puede dar dependiendo del tipo de uva.

In [22]:
```
#Precio vs. Año de Cosecha
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['Year'], s=2, label="Precio Vs Año de cosecha", co
lor="#FFA000")
plt.legend();
plt.show()
```



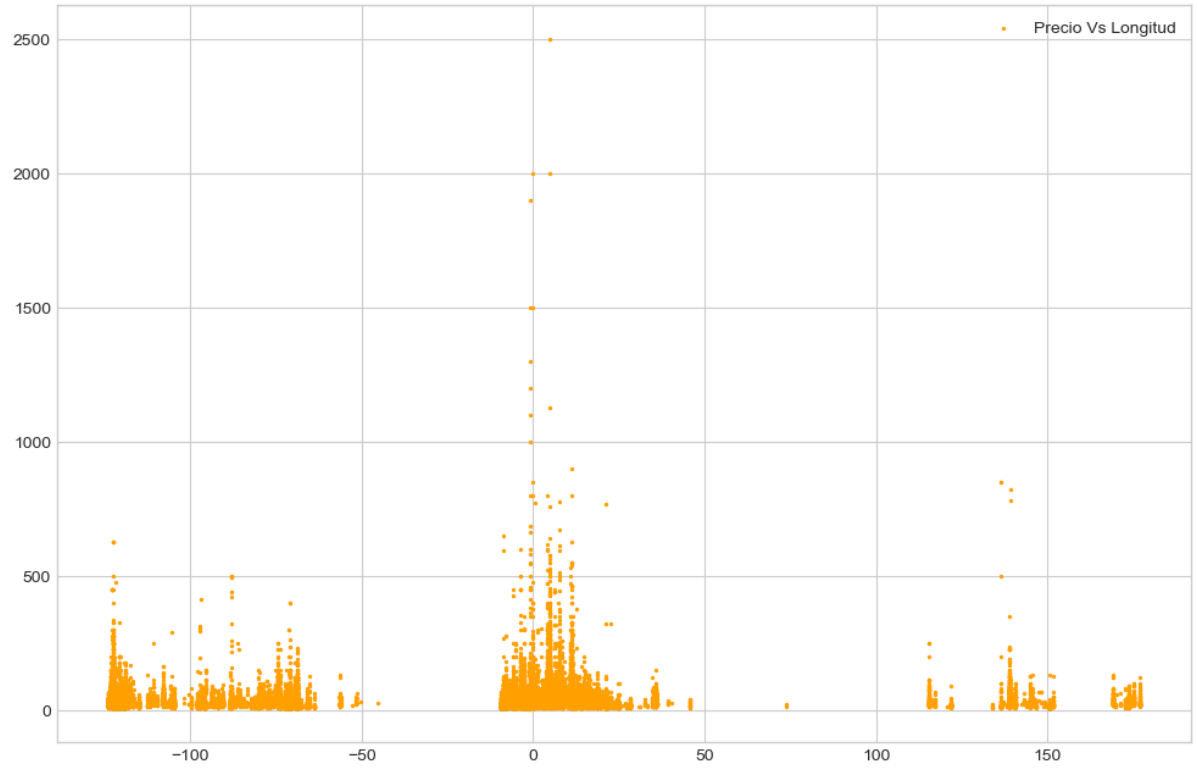
Relacionando el precio contra el año de siembra, no se encuentra una correlación de los precios del vino. Si se encuentran algunos valores atípicos en el año 2009, lo cual puede ser resultado de la crisis mundial de 2008, y donde posiblemente se realizó una menor producción de vino, a menos producción, mayor precio; sin embargo también puede ser dado por un sesgo en la información original ya que winereviews se basa en calificar vinos en general y no en vinos de alto precio.

In [23]:
```
#Precio vs. Latitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['Lat_x'], s=2, label="Precio Vs Latitud", color="#FFA000")
plt.legend();
plt.show()
```



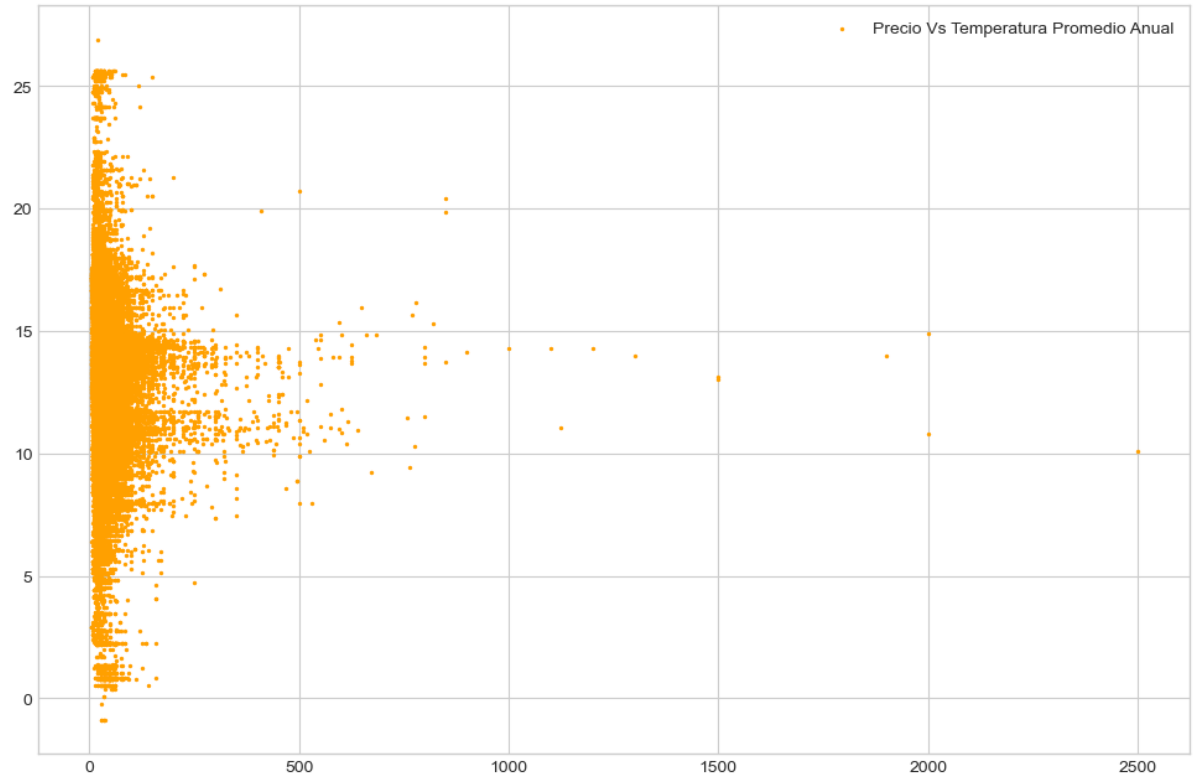
Es interesante encontrar que los precios atípicos se encuentran alrededor de la Latitud 45°, ya que justo pertenece a la zona de California (Napa Valley) y la misma concuerda con las regiones de España, Italia y Francia. Se podría tambien analizar la zona de -30°, donde se encuentran varios vinos de mejor precio. Se puede analizar, en esta zona, si se pudieran encontrar aun más zonas con las temperaturas ideales

In [24]:
```
#Precio vs. Longitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['price'], s=2, label="Precio Vs Longitud", color
="#FFA000")
plt.legend();
plt.show()
```



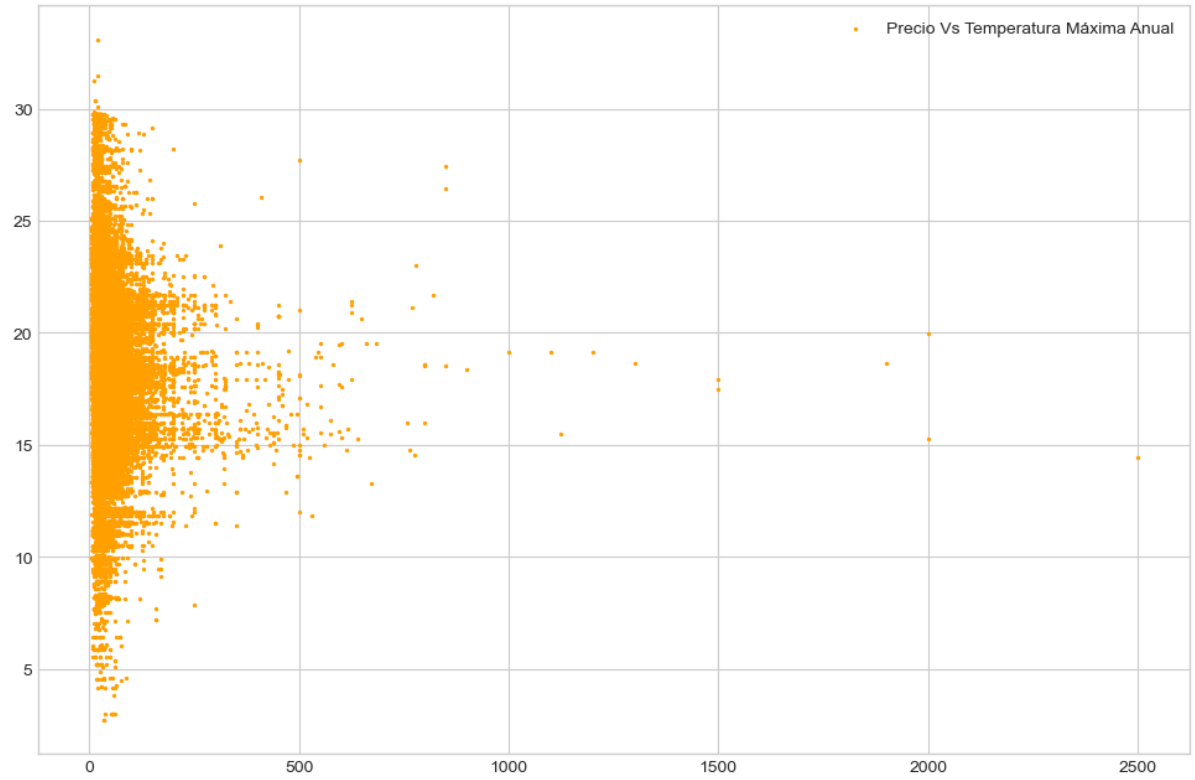
En este caso vemos que los precios atípicos se centran alrededor de los 0°; lo cual quiere decir que estos vinos tienden a ser los de la zona cercana a España, Francia e Italia. Algunos otros valores elevados son los de la zona de Australia. Es interesante que la zona de la Costa Oeste de Estados Unidos tienen vinos que oscilan alrededor de los 250 dólares, precios bastante acsequibles.

In [25]:
```
#Precio vs. Temperatura Promedio Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['temp_anual'], s=2, label="Precio Vs Temperatura P
romedio Anual", color="#FFA000")
plt.legend();
plt.show()
```



En este caso, los vinos más caros oscilan en las temperaturas promedio anuales de 10 a 15°. Esto también permite identificar que una buena zona para sembrar vinos, que pueden ser altamente rentables son en la misma zona de vinos de buena calidad.

In [26]:
```
#Precio vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['temp_max_anual'], s=2, label="Precio Vs Temperatu
ra Máxima Anual", color="#FFA000")
plt.legend();
plt.show()
```

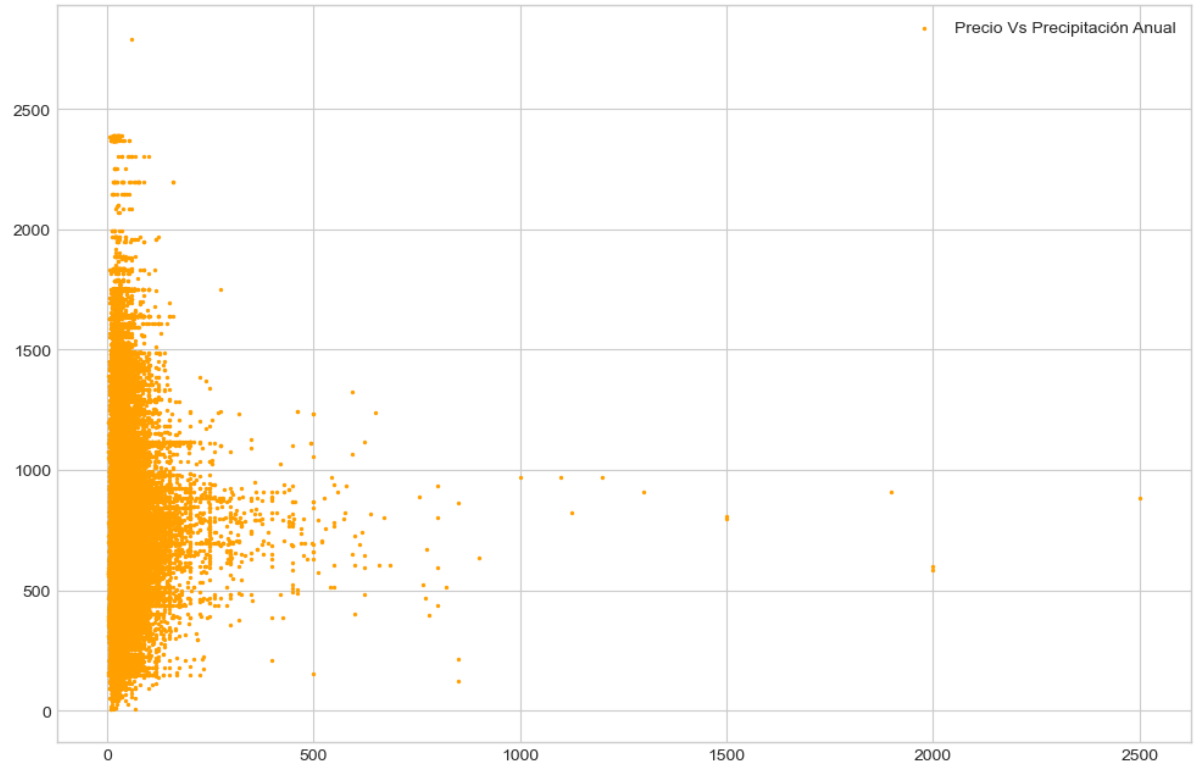


Nuevamente, los vinos con mayor precio se ubican dentro del rango de los 14 y 22 grados, validando la hipótesis de que la temperatura máxima anual varía entre este rango.

In [27]:
```
#Precio vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['temp_min_anual'], s=2, label="Precio Vs Temperatu
ra Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```

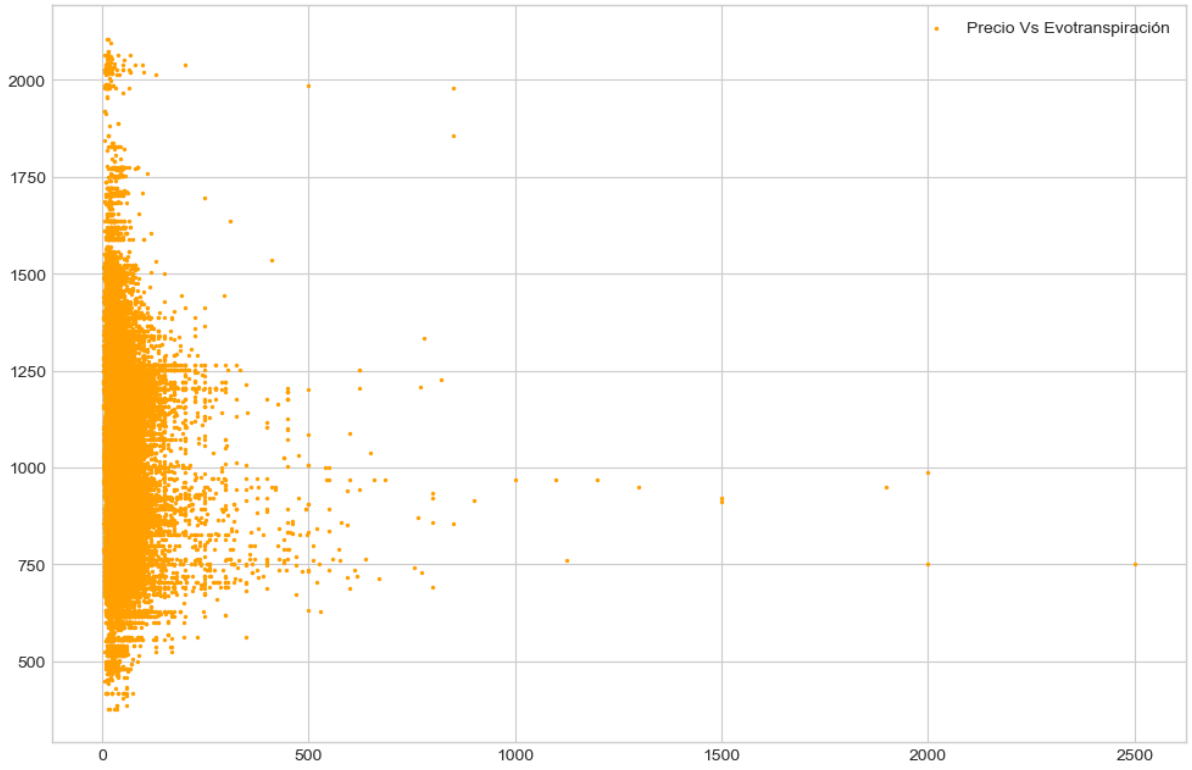
Así como con los otros datos de temperatura, los vinos de mayor precio varían su temperatura mínima entre los 4 y 13 grados centigrados. Esto permite seguir validando las interpretaciones anteriores, ya que un vino de alto precio no seria comercialmente rentable si no es vendido de manera racional por el público.

In [28]:
```
#Precio vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['pre_anual'], s=2, label="Precio Vs Precipitación
Anual", color="#FFA000")
plt.legend();
plt.show()
```



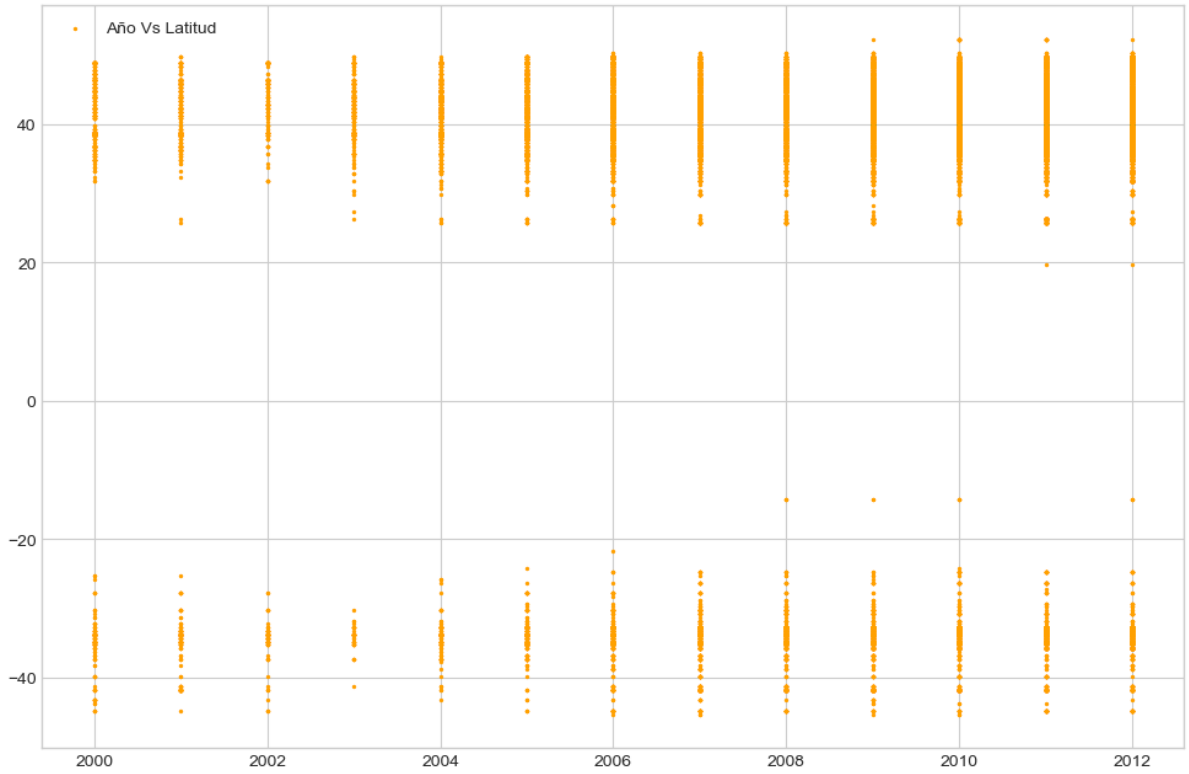
Para la precipitación anual encontramos que los vinos con mayor precio se ubican, con respecto a la precipitación, con valores entre los 200mm y los 1400mm al igual que el análisis de calidad

In [29]:
```
#Precio vs. Evotranspiración
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['price'], df['etp_anual'], s=2, label="Precio Vs Evotranspiración", color="#FFA000")
plt.legend();
plt.show()
```



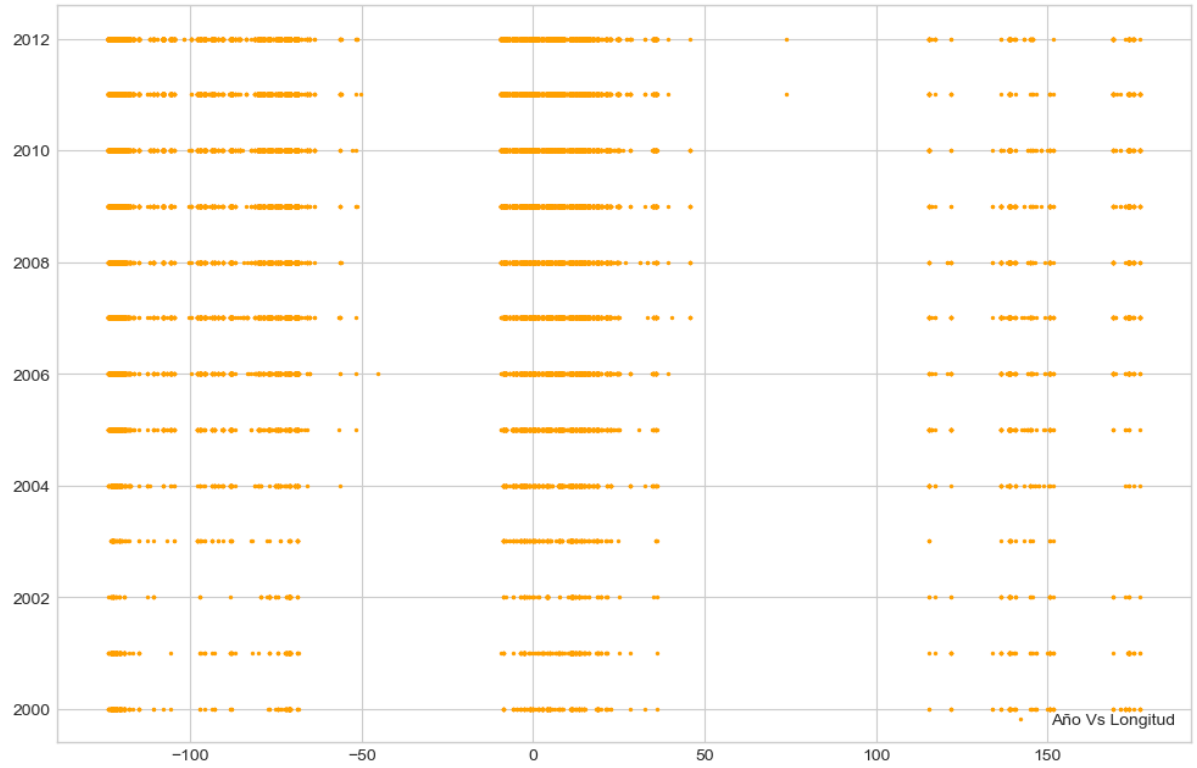
En este caso encontramos que los valores para los vinos de mayor precio se encuentran entre los 600mm y los 1300mm. También se encuentran algunos valores atípicos alrededor de los 180mm a los 1900mm, por lo cual resulta interesante analizar estos vinos para encontrar si correspoden a una cepa en partícular que requiera mayor evotranspiración (agua requerida para la siembra)

In [30]:
```
#Año vs. Latitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['Lat_x'], s=2, label="Año Vs Latitud", color="#FFA0
00")
plt.legend();
plt.show()
```



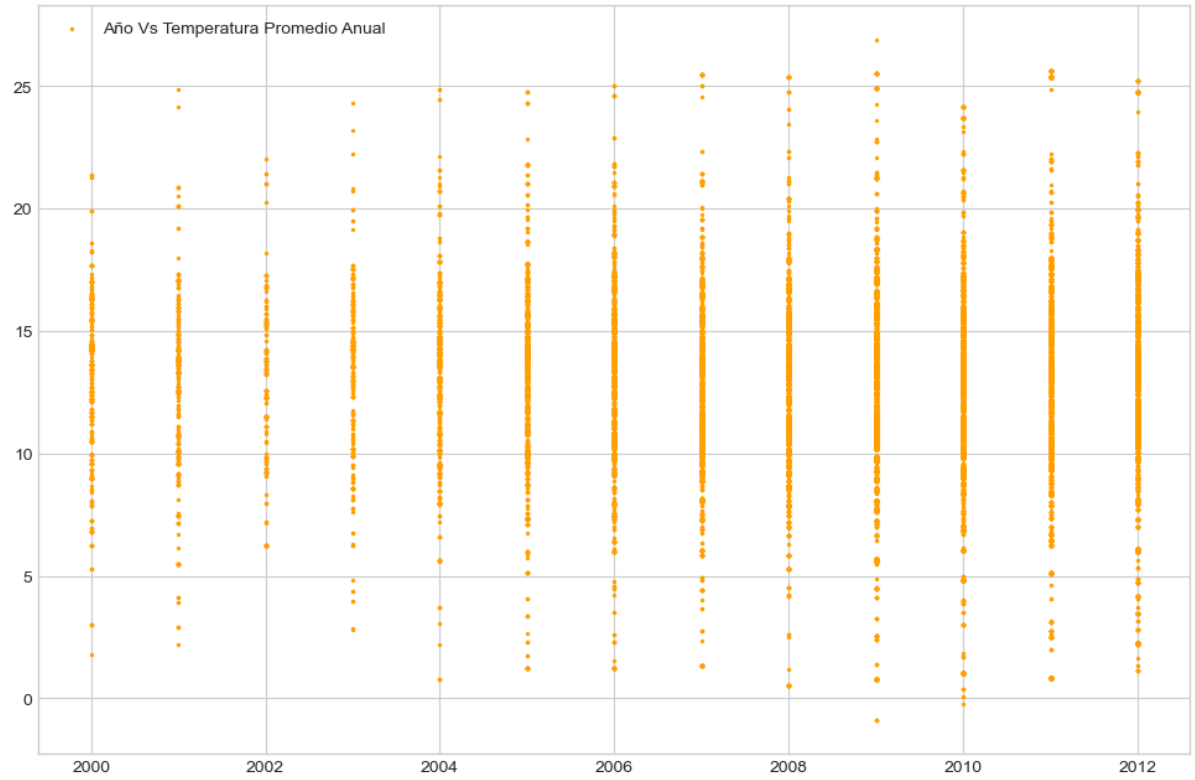
Este análisis nos permite ver la relación de las muestras por año, vs la Latitud de la cosecha. En este caso encontramos que hay una concentración de muestras mucho mayor para la latitud positiva, así podemos determinar que nuestro dataframe se encuentra un poco desbalanceado en esta variable. También se encuentra cierto desbalanceo para las variables de los años 2000 a 2006, posiblemente por que hay mayor producción de vino en los ultimos años (el vino es uno de los productos con un crecimiento YoY alto; anexar referencias de este análisis)

In [31]:
```
#Año vs. Longitud
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['Year'], s=2, label="Año Vs Longitud", color="#FF
A000")
plt.legend();
plt.show()
```



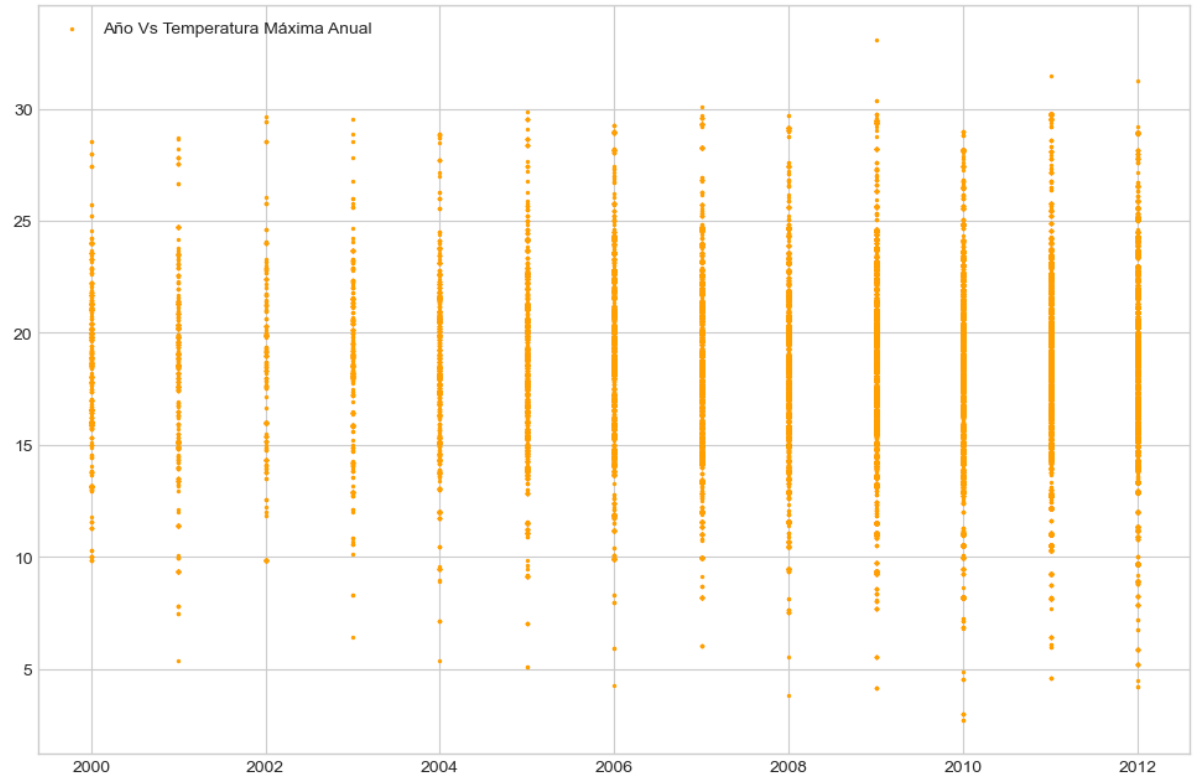
Para la Longitud encontramos la misma conclusión que para la Latitud, así mismo, tambien vemos un dataframe desbalanceado, teniendo gran cantidad de vinos de la parte Oeste de Europa, y menos cantidad de datos para la parte Americana y Asiática-Oceánica

In [32]:
```
#Año vs. Temperatura Promedio Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['temp_anual'], s=2, label="Año Vs Temperatura Prome
dio Anual", color="#FFA000")
plt.legend();
plt.show()
```



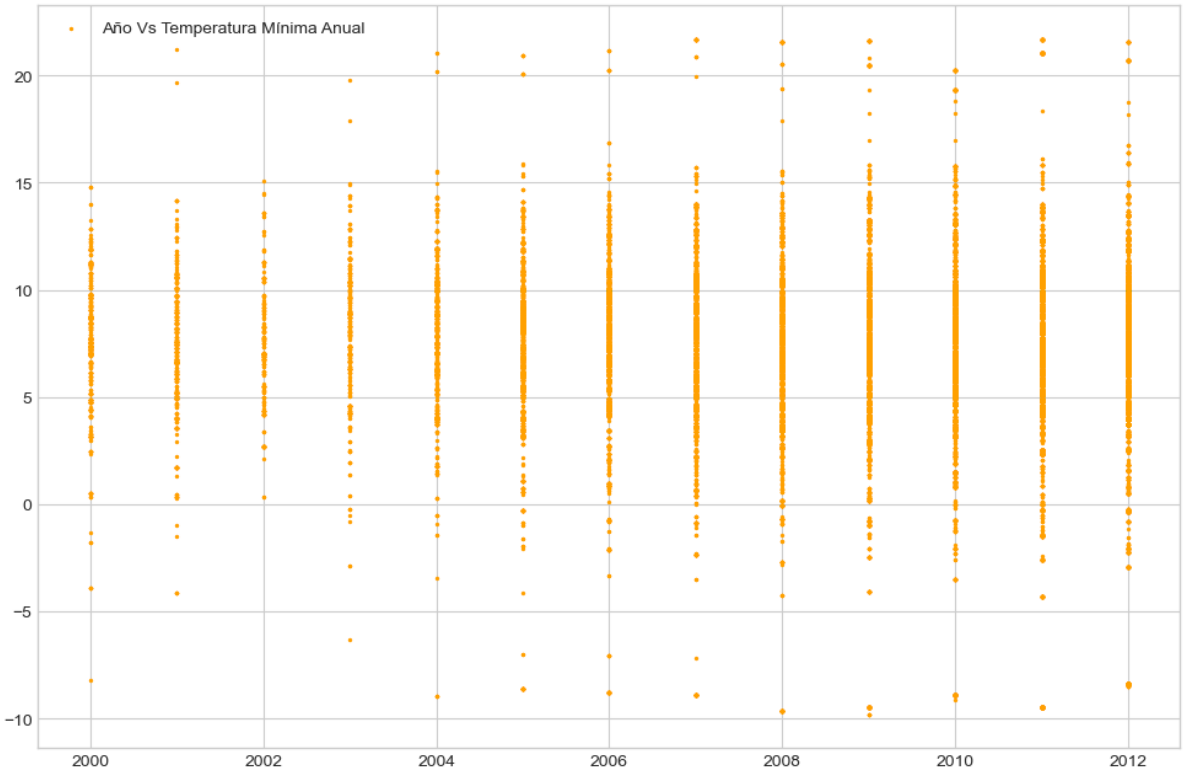
Se puede apreciar que a medida que aumentan los años, se ve un crecimiento de los valores extremos de las tempreaturas promedio, esto puede deberse principalmente al calentamiento global o también al desbalanceo del dataframe y el cambio climático, y el cambio en las zonas donde se tienen máyores entradas de datos

In [33]:
```
#Año vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['temp_max_anual'], s=2, label="Año Vs Temperatura M
áxima Anual", color="#FFA000")
plt.legend();
plt.show()
```



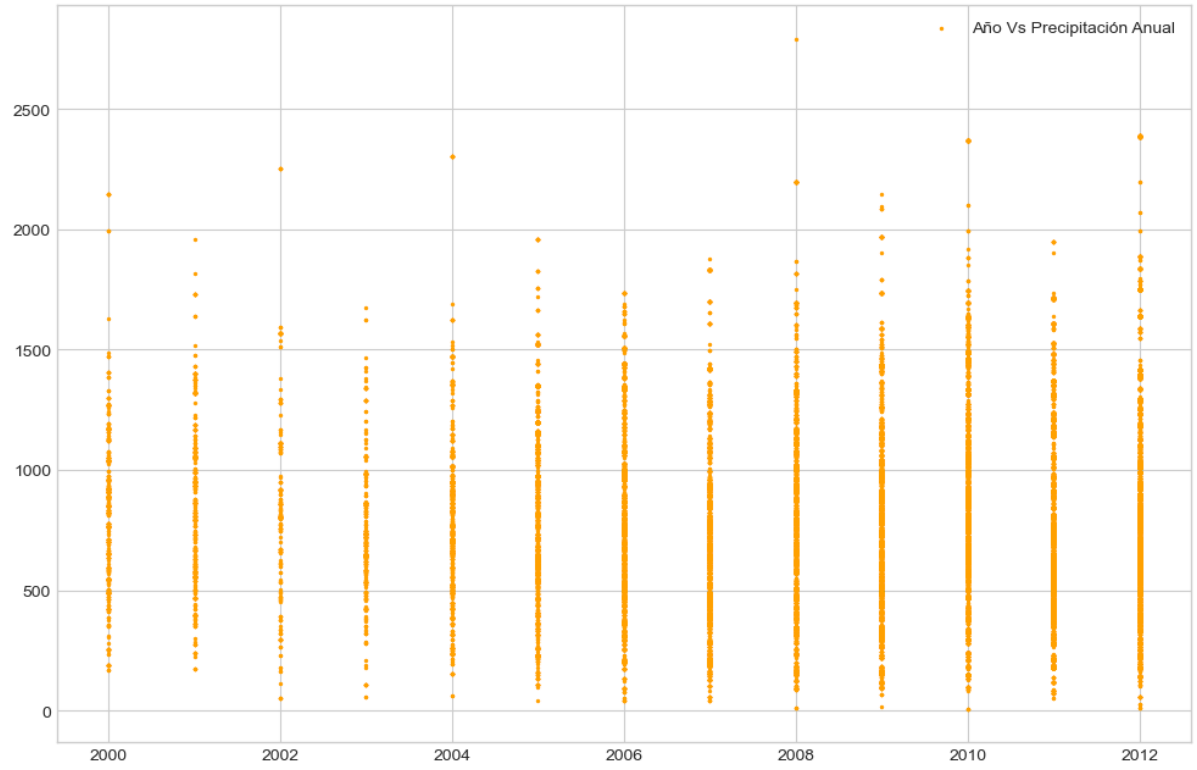
Nuevamente se aprecia un incremento del rango de temperatura máxima anual en los últimos años.

In [34]:
```
#Año vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['temp_min_anual'], s=2, label="Año Vs Temperatura M
ínima Anual", color="#FFA000")
plt.legend();
plt.show()
```



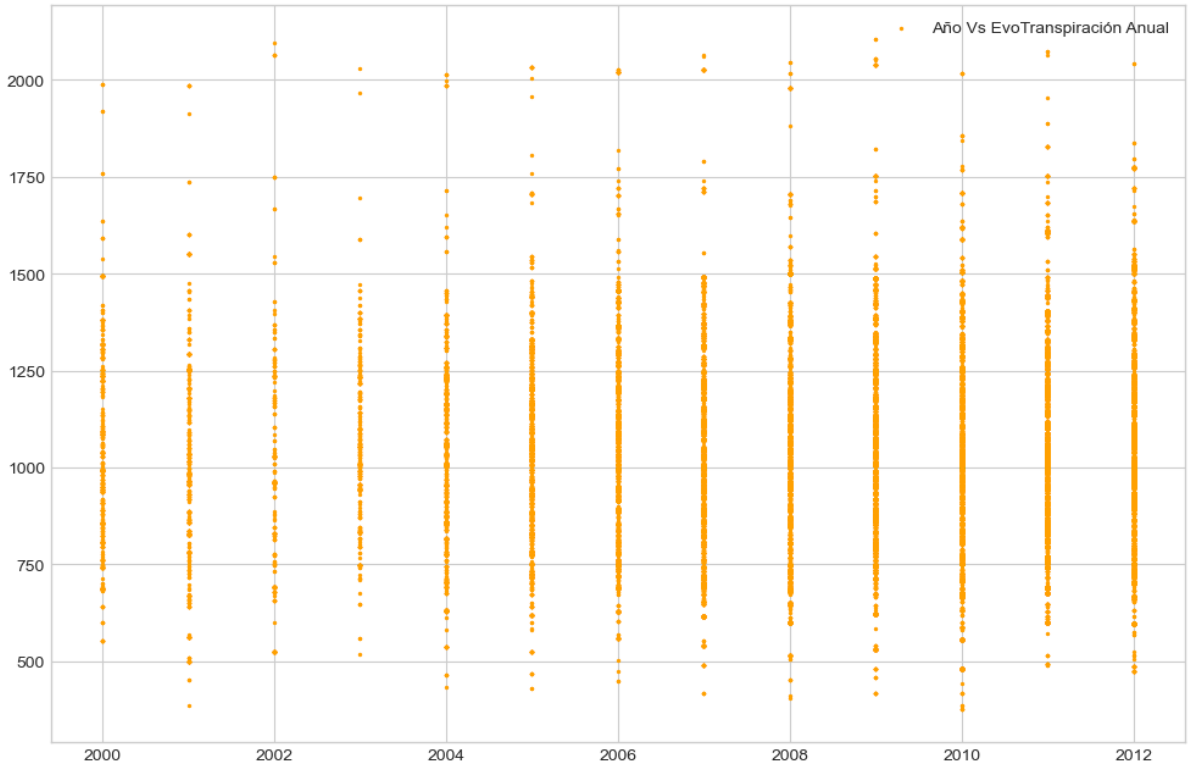
Encontramos nuevamente la misma conclusión para las temperaturas mínimas anuales.

In [35]:
```
#Año vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['pre_anual'], s=2, label="Año Vs Precipitación Anua
l", color="#FFA000")
plt.legend();
plt.show()
```



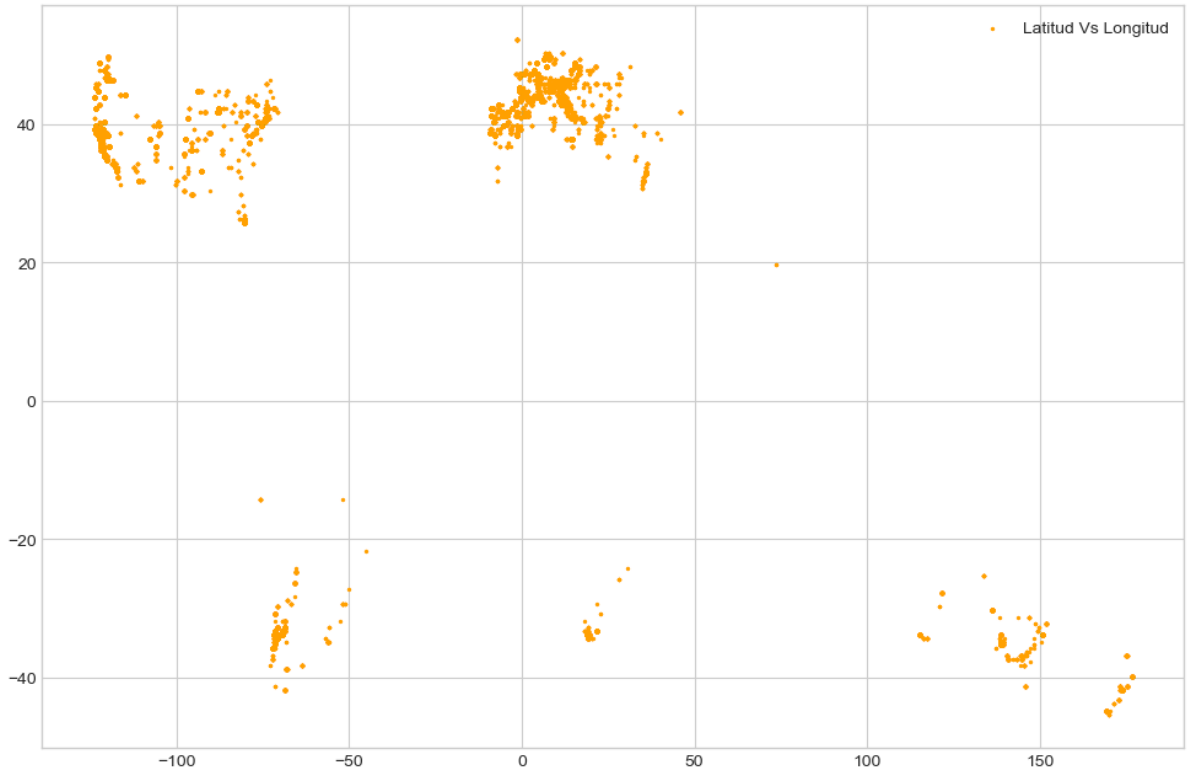
Con respecto a la precipitación, se pueden tambien ver casos de mayor precipitación hacia los años recientes, pero en terminos generales el rango principal se mantiene en los años

In [36]:
```
#Año vs. EvoTranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['etp_anual'], s=2, label="Año Vs EvoTranspiración Anual", color="#FFA000")
plt.legend();
plt.show()
```



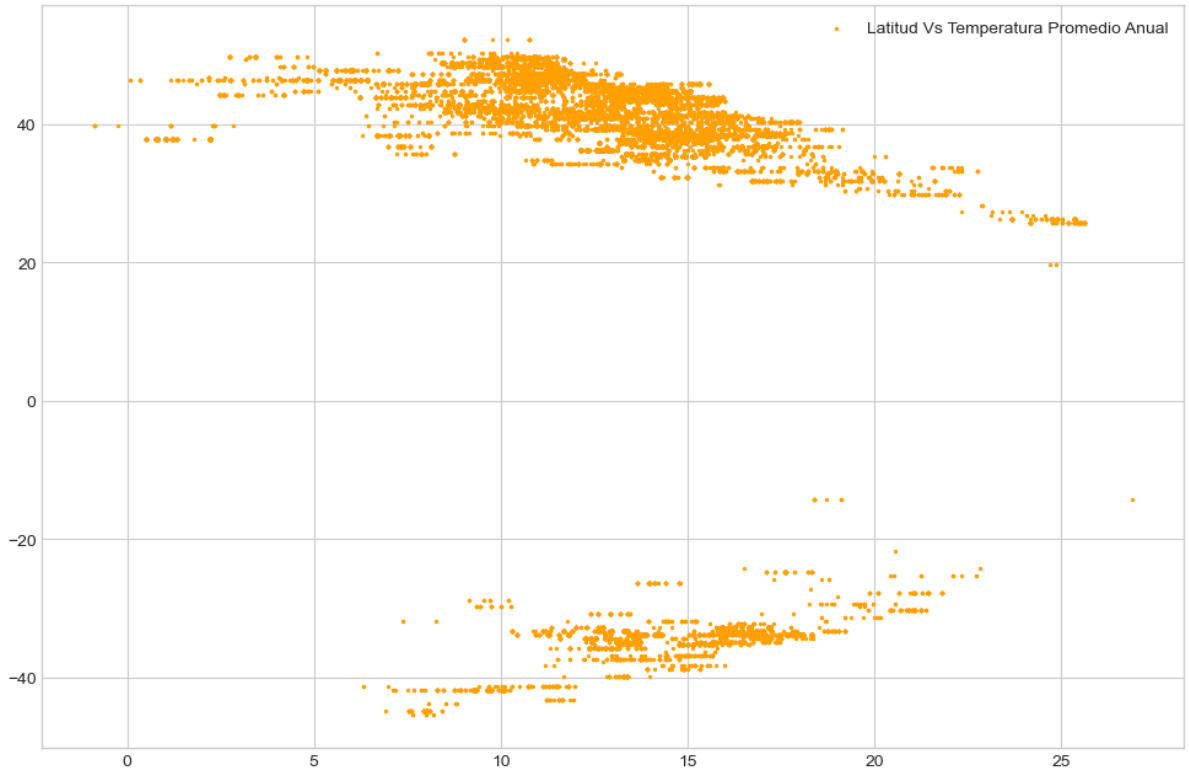
Con respecto a la evotranspiración, se encuentra que el rango general es el mismo para todos los años con un desbalanceo de datos en los años más recientes. Ene ste caso, se esperaba que la evotranspiración se mantuviera en valores estables

In [37]:
```
#Latitud vs. Longitud (Invertido para poder ver el mapa mundial y los vinos de la muestra)
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['Lat_x'], s=2, label="Latitud Vs Longitud", color="#FFA000")
plt.legend();
plt.show()
```



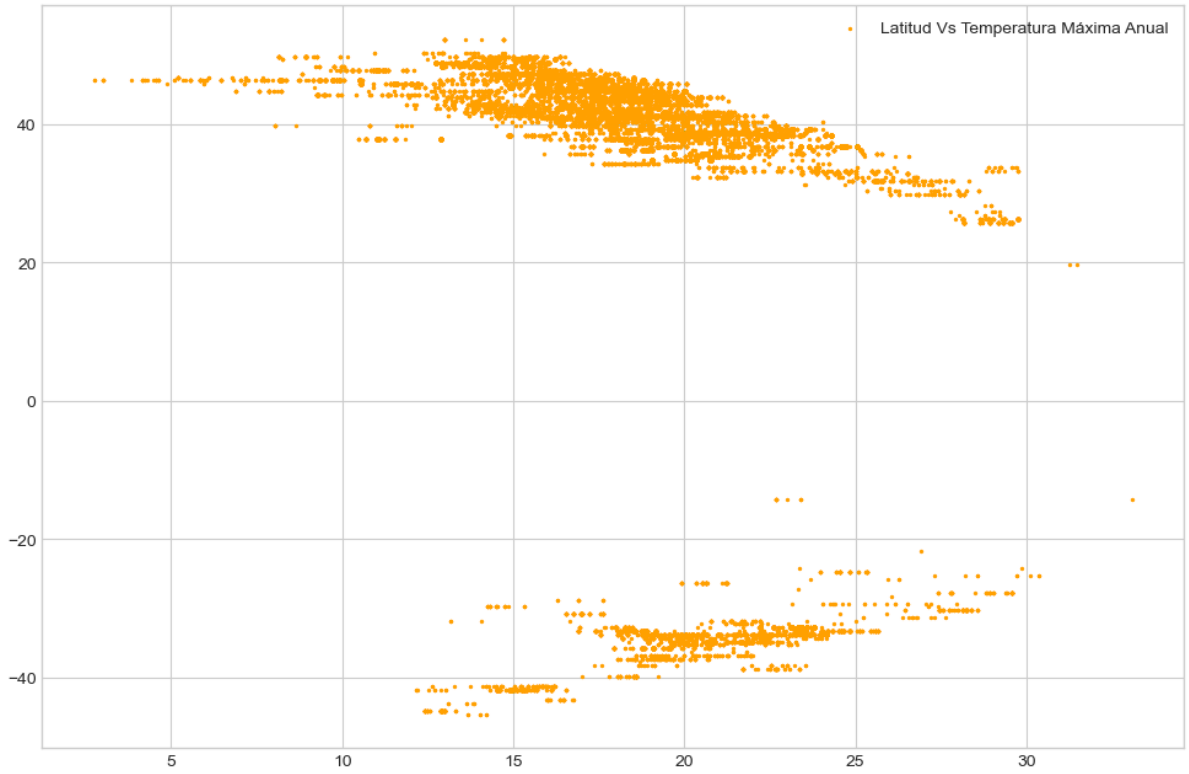
En este caso se aprecia el muestreo en un mapa imaginario, donde vemos las muestras de Estados Unidos y Canadá, luego en la parte inferior cosechas en chile y argentina, con alguos valores en perú, Brazil y cosechas en la costa este de Argentina. Luego se puede ver claramente Europa con la mayor clusterización de muestras, parte del muestreo en la parte sur de África, y por último valores en la India, China y Australia. Lamentablemente los datos obtenidos tienen un pocos datos de zonas que no son habiltualmente usadas para la siemrba de vinos, lo que perimitiría encontrar información más valiosa para las conclusiones finales.

In [38]:
```
#Latitud vs. Temperatura Promedio Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_anual'], df['Lat_x'], s=2, label="Latitud Vs Temperatura
Promedio Anual", color="#FFA000")
plt.legend();
plt.show()
```



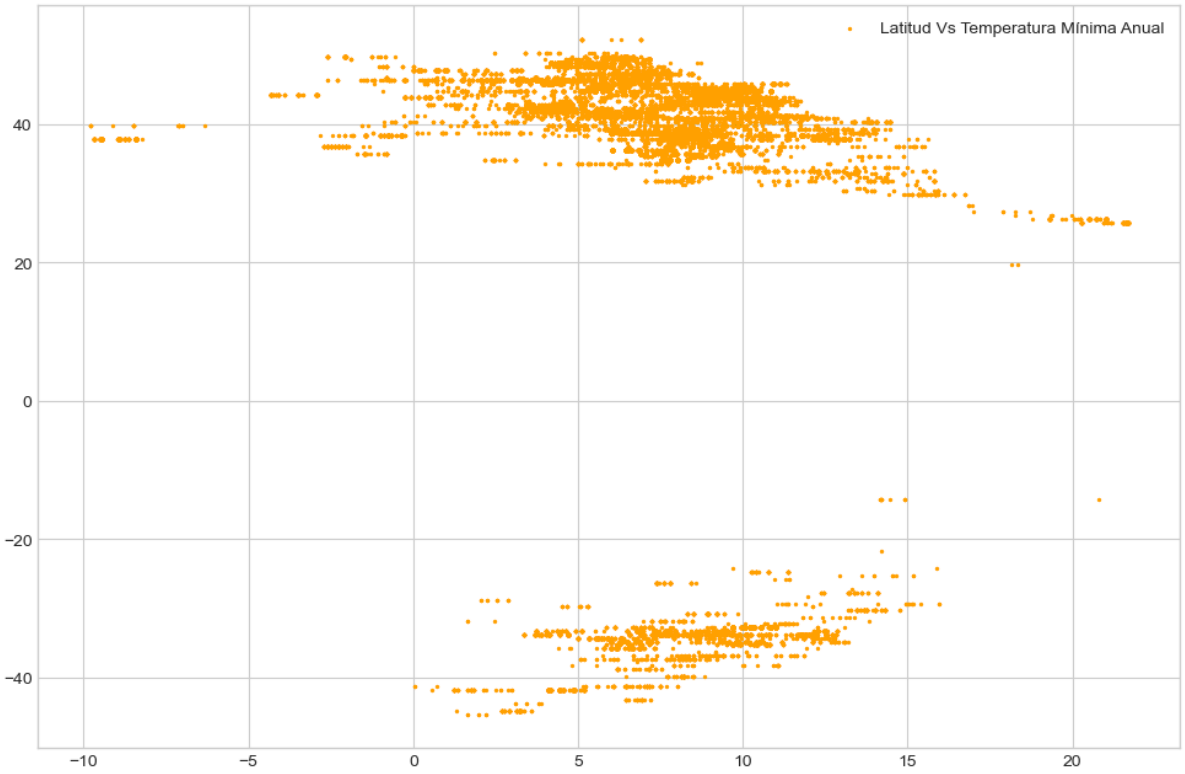
En este caso encontramos que los valores de tempreatura promedio extremos se encuentran en la parte positiva de la Latitud, posiblemente por la cantidad de muestreos; pero esto también permite inferir que entre más cerca del Ecuador, las temperaturas promedio anuales comienzan a aumentar. También encontramos que en la misma Latitud se pueden encontrar amplios rangos de temperaturas, que indica como conclusión que no solo seleccionar una Latitud con buena calidad es suficiente, pero también hay que evaluar los valores de temperatura. Otro punto interesante es que es posible buscar algunas zonas al Sur del trópico de Capricornio que tengan los rangos de temperatura ideales para el proyecto de cultivo.

In [39]:
```
#Latitud vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_max_anual'], df['Lat_x'], s=2, label="Latitud Vs Temperat
ura Máxima Anual", color="#FFA000")
plt.legend();
plt.show()
```



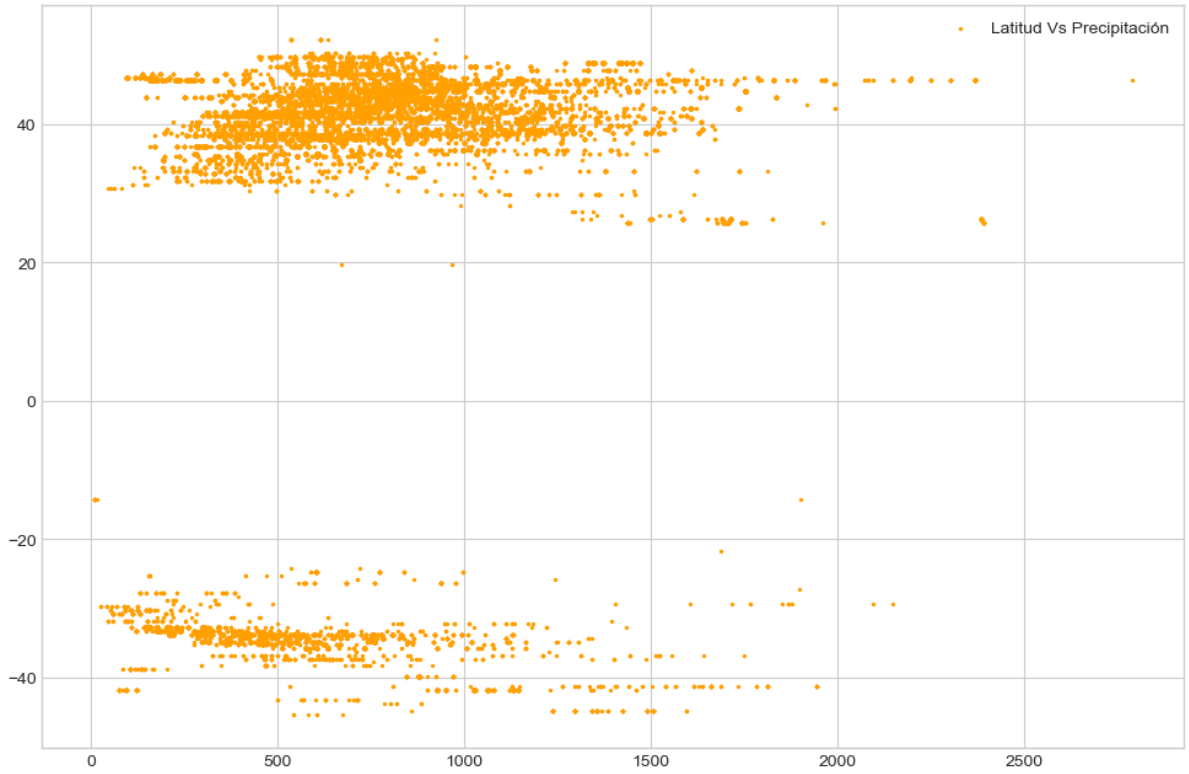
Se obteiene una conclusión similar a la anterior, con variaciones de temperatura cercana al Ecuador, y con algunas zonas en el sur que podrían evaluarse con respecto a las mejores temperaturas.

In [40]:
```
#Latitud vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_min_anual'], df['Lat_x'], s=2, label="Latitud Vs Temperat
ura Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```



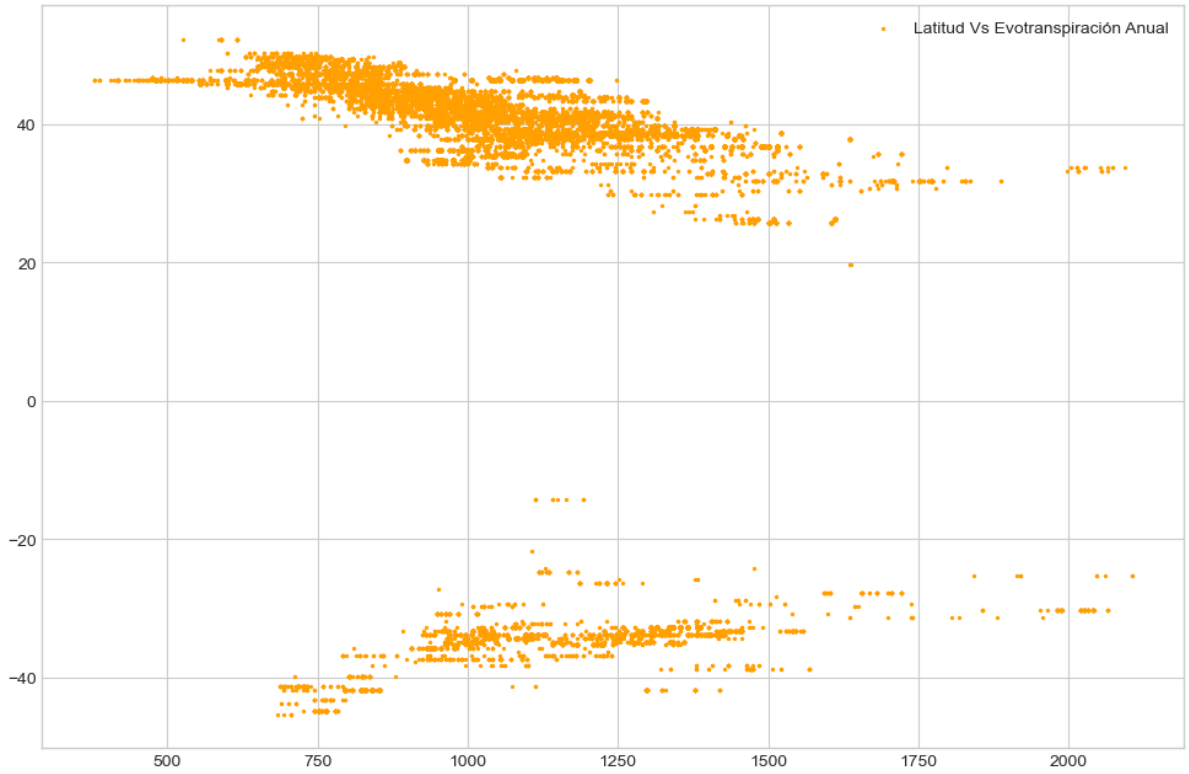
Para este caso, encontramos nuevamente la misma conclusión anterior.

In [41]:
```
#Latitud vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['pre_anual'], df['Lat_x'], s=2, label="Latitud Vs Precipitació
n", color="#FFA000")
plt.legend();
plt.show()
```



Para la precipitación, encontramos el desbalance del dataframe, teniendo muchos más datos para la parte positiva de la Latitud. Los rangos de mejor precipitación se mantienen en ciertos sitios pero se pueden analizar nuevas zonas en la parte sur para nuevos cultivos.

In [42]:
```
#Latitud vs. Evotranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['etp_anual'], df['Lat_x'], s=2, label="Latitud Vs Evotranspirac
ión Anual", color="#FFA000")
plt.legend();
plt.show()
```



Nuevamente, con respecto a la Evotranspiración, se encuentran los mismos resultados; sin embargo, a medida que se acerca a la zona del Ecuador se ve una tendencia que los viñedos requieren más agua para poder producirse.

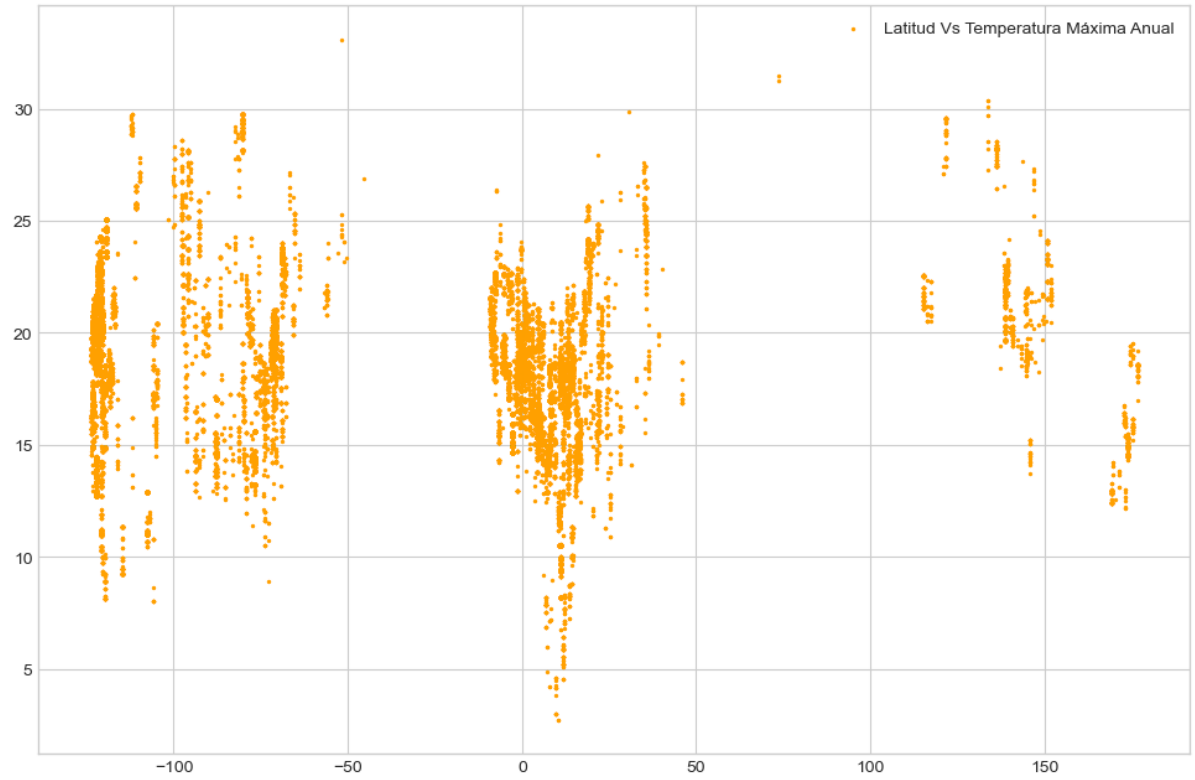
In [43]:
```
#Longitud vs. Temperatura Promedio Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['temp_anual'], s=2, label="Latitud Vs Temperatura
Promedio Anual", color="#FFA000")
plt.legend();
plt.show()
```

AMERICA Se encuentra cierta tendencia, que a medida que se mueve hacia la costa este, los valores de temperatura promedio comienzan a aumentar. Esto permite encontra mejores zonas en esta parte para las cosechas de vino.

EUROPA/AFRICA En este caso, aunque son zonas diferentes, podemos ver que la mayoría de vinos se sentra entre los 9 y 16 grados, temperatura óptima para la cosecha de vinos.

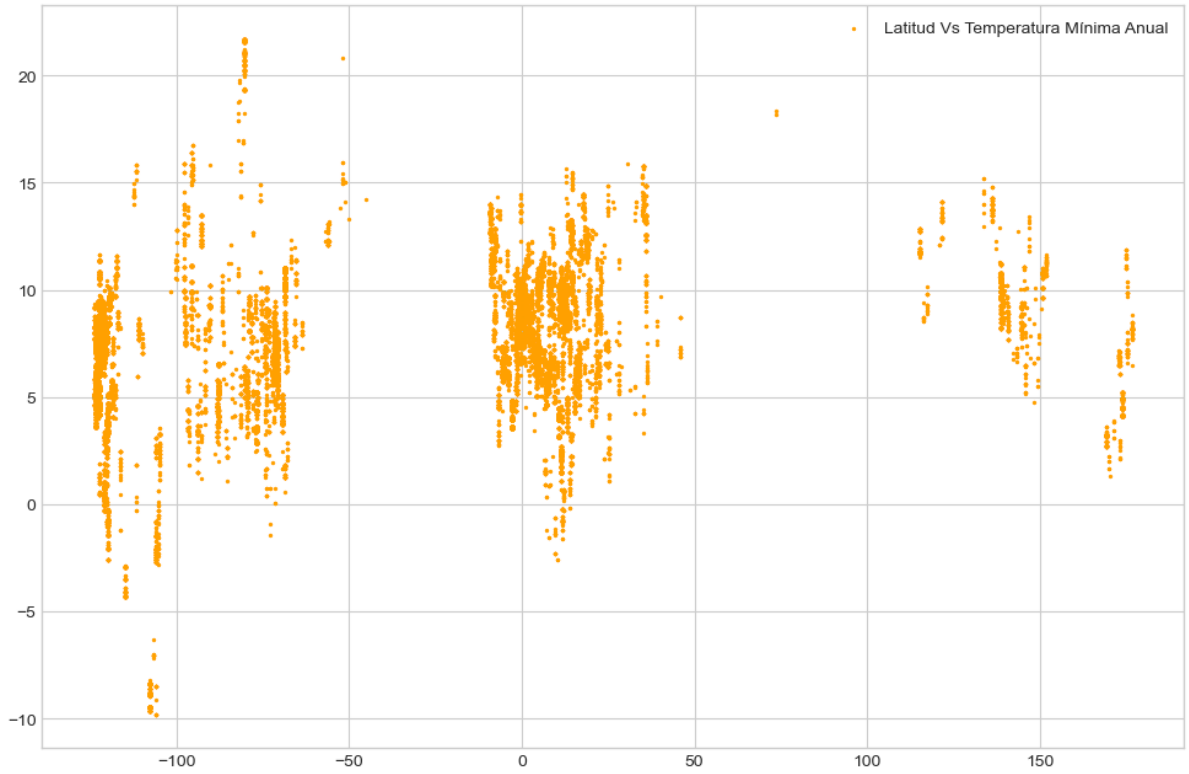
ASIA/OCEANÍA Acá se encuentran valores con temperaturas muy buenas para la siembra, con algunos casos de temperaturas un poco más altas. En este caso, se pueden tambien buscar zonas para nuevos cultivos de vino

In [44]:
```
#Longitud vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['temp_max_anual'], s=2, label="Latitud Vs Tempera
tura Máxima Anual", color="#FFA000")
plt.legend();
plt.show()
```



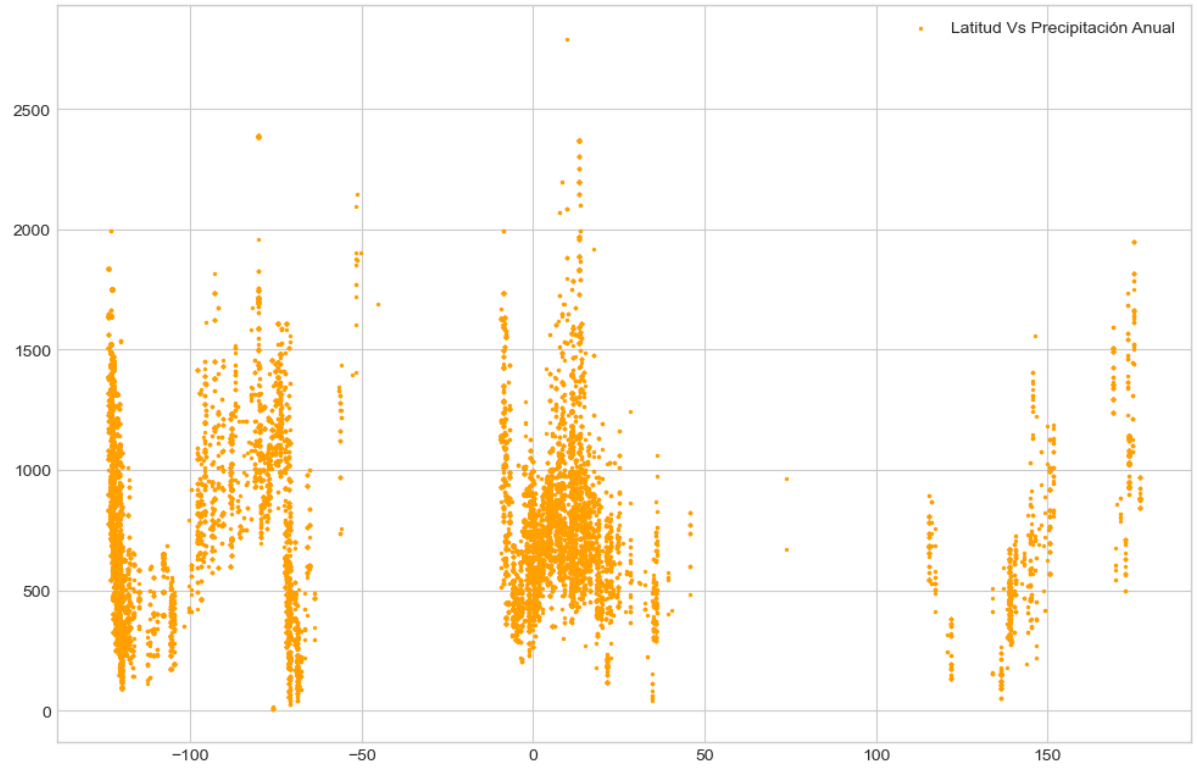
En esta comparativa vemos nuevamente algo similar a la comparativa anterior, sin embargo se ve en ASIA y OCEANIA que existe una mayor temperatura máxima, por lo cual las zonas a proponer para cultivos no deben tener temperaturas tan extremas. Acá una opción es buscar lugares cercanos al oceano para que las corrientes de viento refresquen dichos cultivos.

In [45]:
```
#Longitud vs. Temperatura Mínima  Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['temp_min_anual'], s=2, label="Latitud Vs Tempera
tura Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```



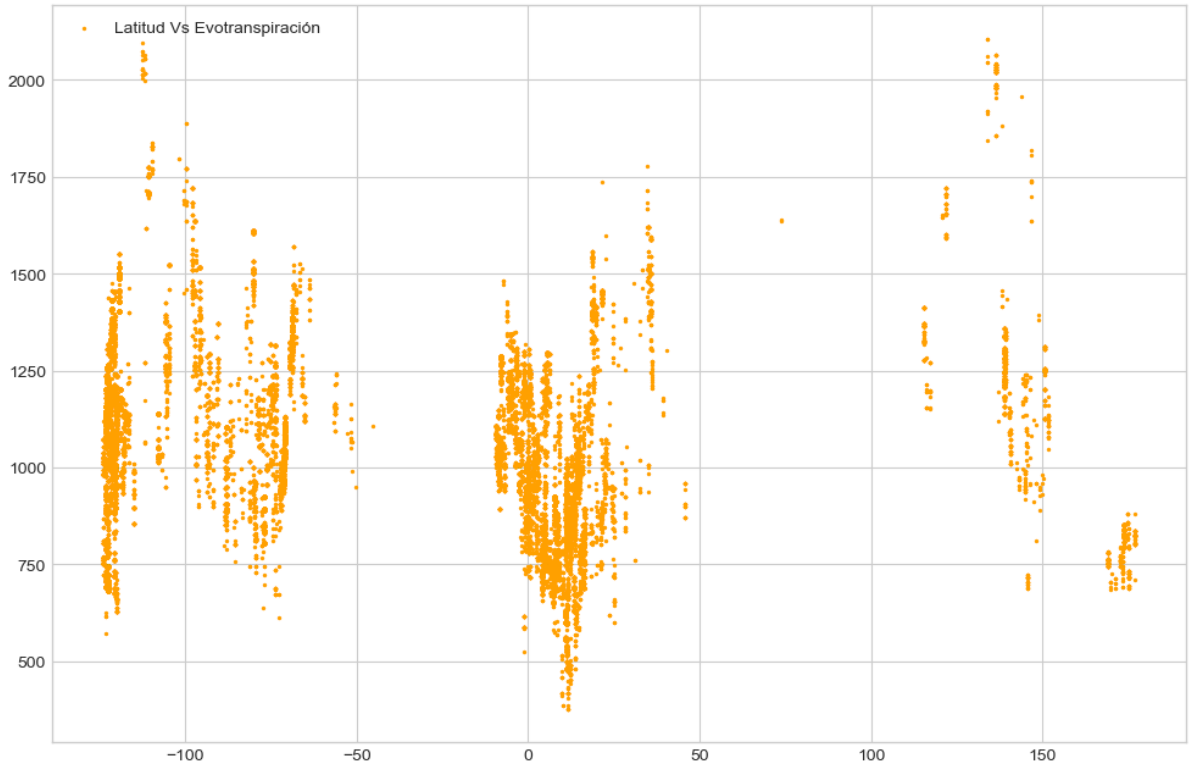
Inversamente, con la comparativa anterior, la zona Americana tiene temperaturas mínimas más bajas, por lo que las zonas del sur que se recomienden no pueden bajar tanto de temperatura para estar sobre el rango ideal. En este caso, Asia y Oceanía presenta para temperaturas mínimas ideales

In [46]:
```
#Longitud vs. Precipitación
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['pre_anual'], s=2, label="Latitud Vs Precipitación Anual", color="#FFA000")
plt.legend();
plt.show()
```



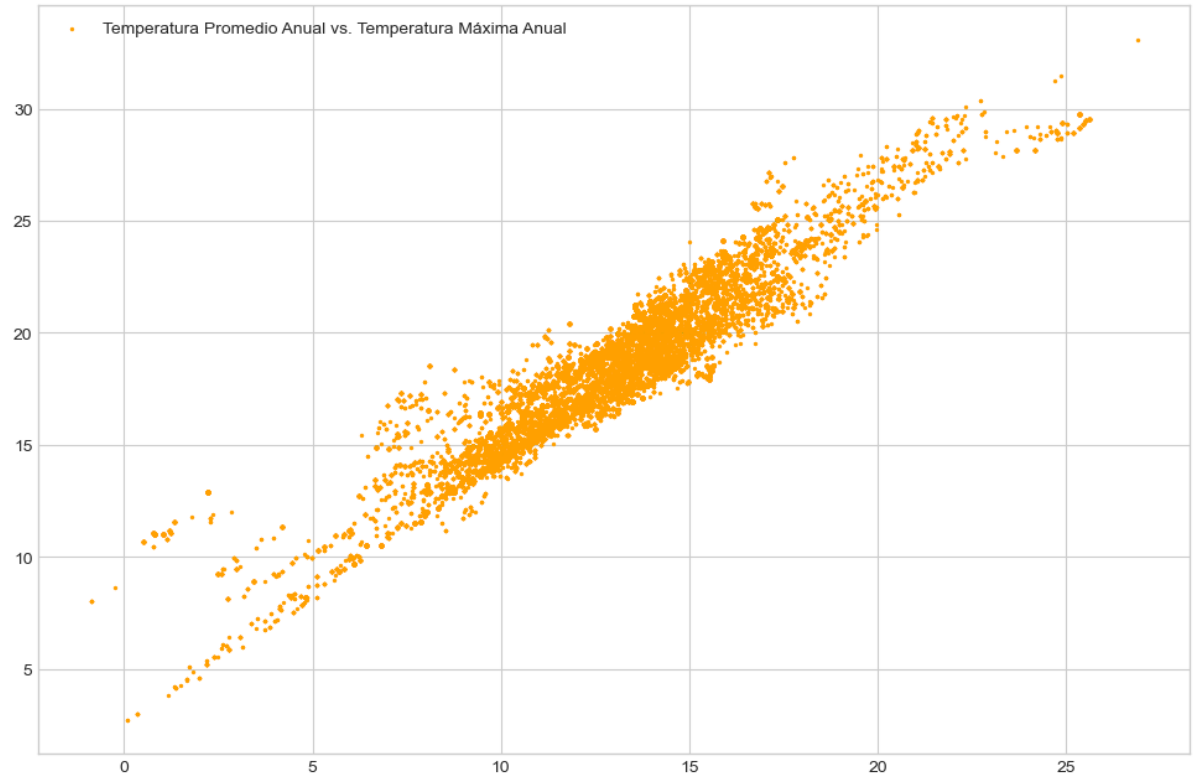
En el caso de la precipitación, la mayoría de regiones se encuentran en zonas con precipitación recomendable para el cultivo. Es interesante ver que en la costa Oeste de Estados unidos tiene un rango ideal, pero a medida que se avanza hacia el este los valores de precipitación son menores, debido a que la zona Oeste esta pegada al oceano atlantico y sus correintes de aire. Caso contrario sucede entre chile y argentina donde se puede apreciar que hay un decenso en la precipitación anual, y donde vuelve a aumentar a medida que se acerca al Oceano Pacífico.

In [47]:
```
#Longitud vs. Evotranspiración
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Long_x'], df['etp_anual'], s=2, label="Latitud Vs Evotranspira
ción ", color="#FFA000")
plt.legend();
plt.show()
```



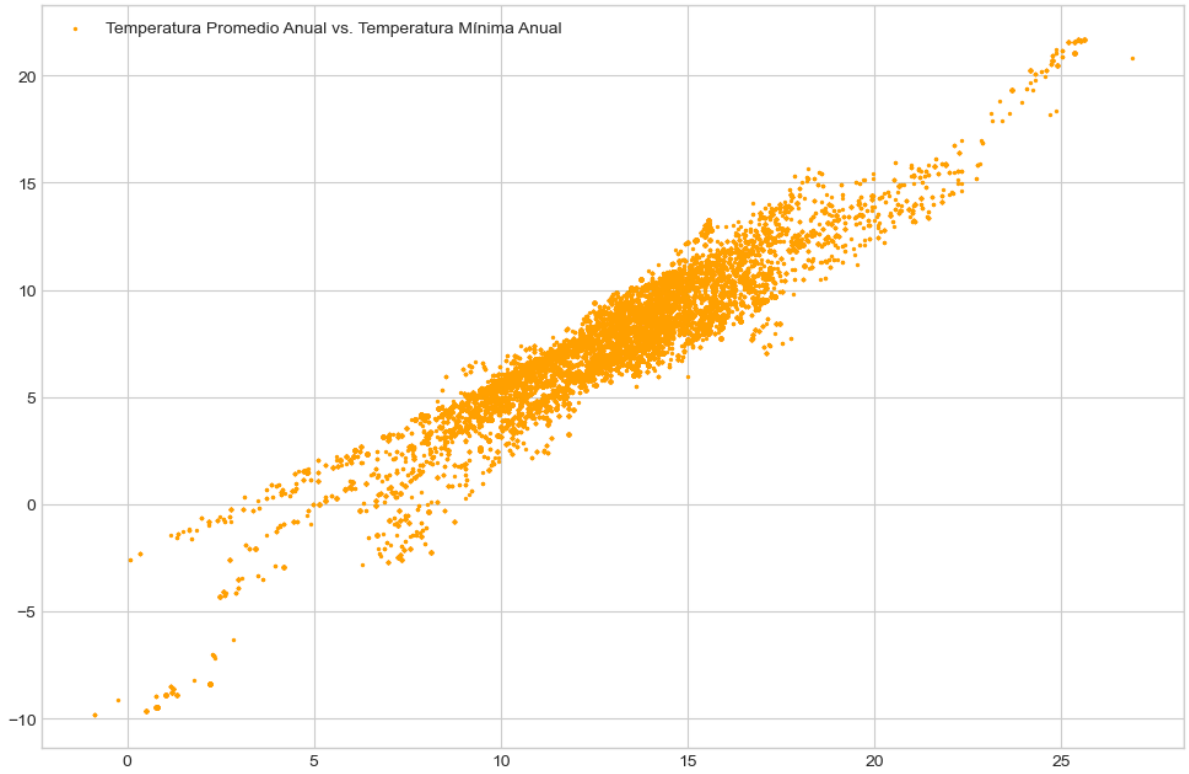
En el caso de la Evotranspiración se ve algo similar invertido que con la precipitación (vease la forma de N en el caso Americano), lo que permite deducir que a menor precipitación, los viñedos requieren mayor agua para sus cultivos. Esto permite plantear el riego del viñedo en los casos que la precipitación no sea la ideal

In [48]:
```
#Temperatura Promedio Anual vs. Temperatura Máxima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_anual'], df['temp_max_anual'], s=2, label="Temperatura Pr
omedio Anual vs. Temperatura Máxima Anual", color="#FFA000")
plt.legend();
plt.show()
```



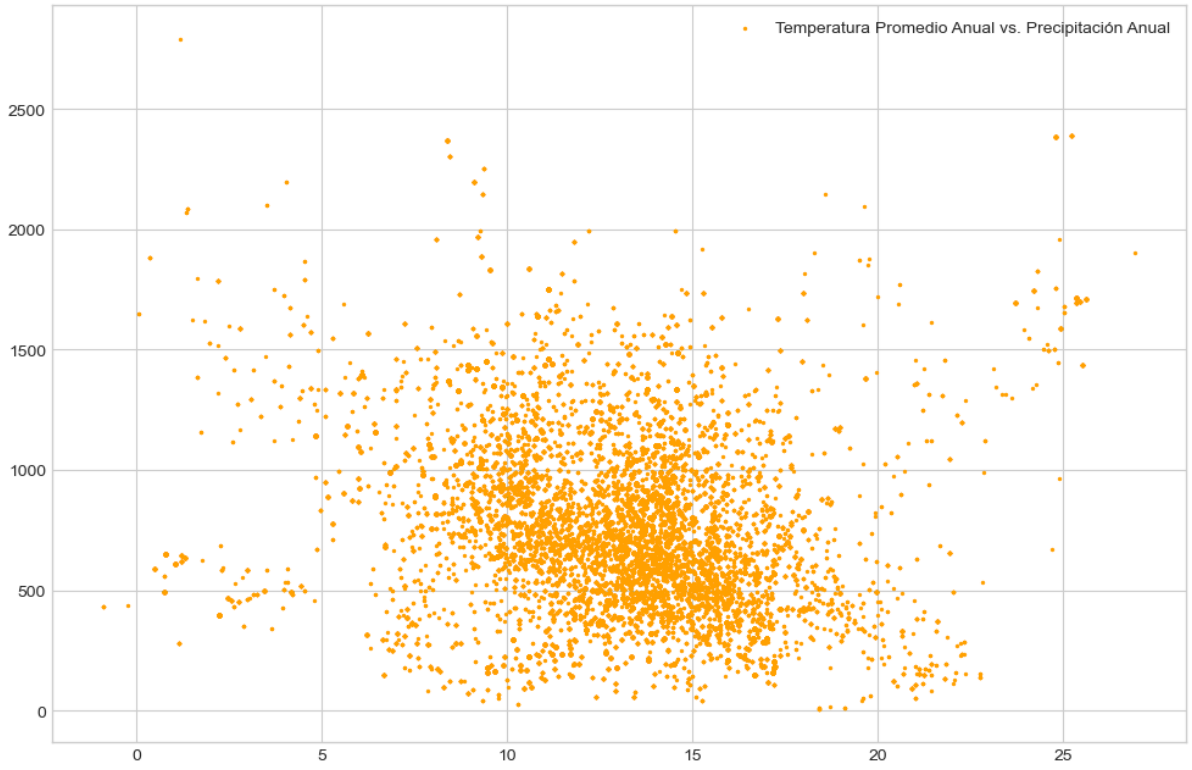
Tal como se esperaba, hay una correlación entre la temperatura anual promedio y la máxima. Por supuesto, existen algunos casos atípicos donde podemos ver que esta correlación varía, pero es posible que sea por zonas que tengan ciertos temas climatológicos especiales, por ejemplo zonas cerca a glaciares o rodeados por oceanos.

In [49]:
```
#Temperatura Promedio Anual vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_anual'], df['temp_min_anual'], s=2, label="Temperatura Pr
omedio Anual vs. Temperatura Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```



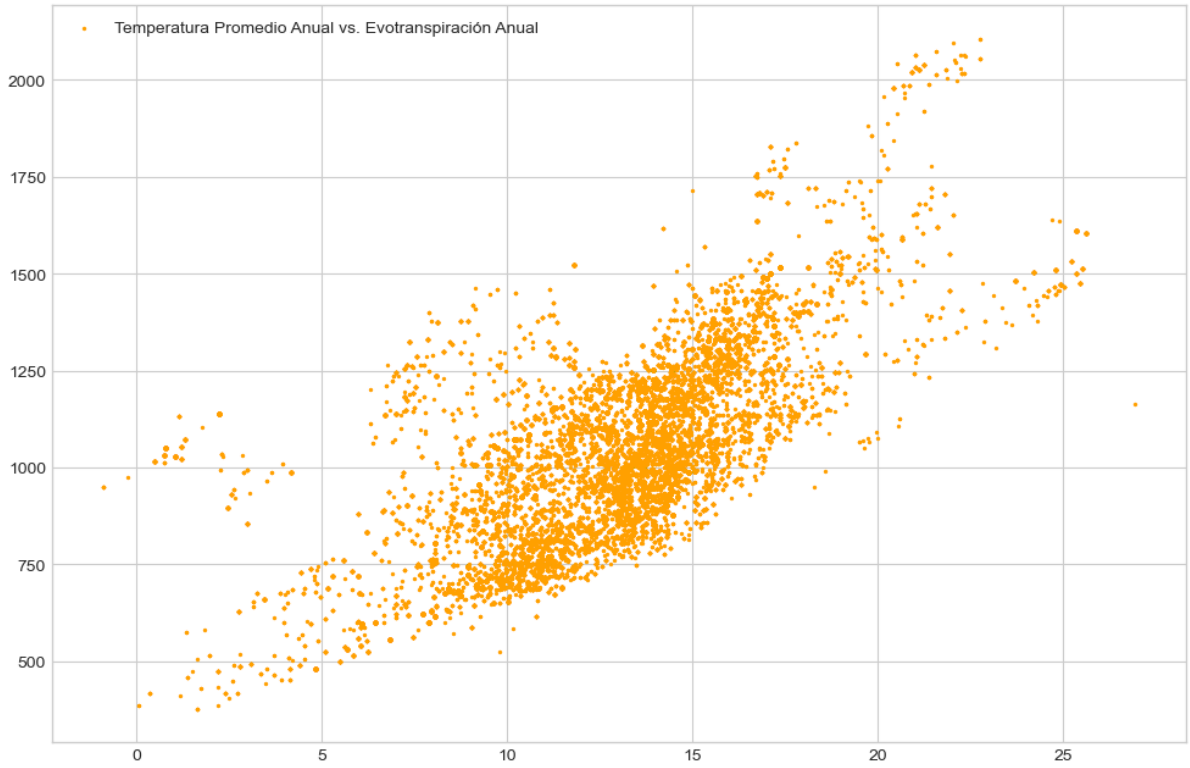
Tal como el análisis anterior, se encuentra una correlación con algunos valores atípicos.

In [50]:
```
#Temperatura Promedio Anual vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_anual'], df['pre_anual'], s=2, label="Temperatura Promedi
o Anual vs. Precipitación Anual", color="#FFA000")
plt.legend();
plt.show()
```



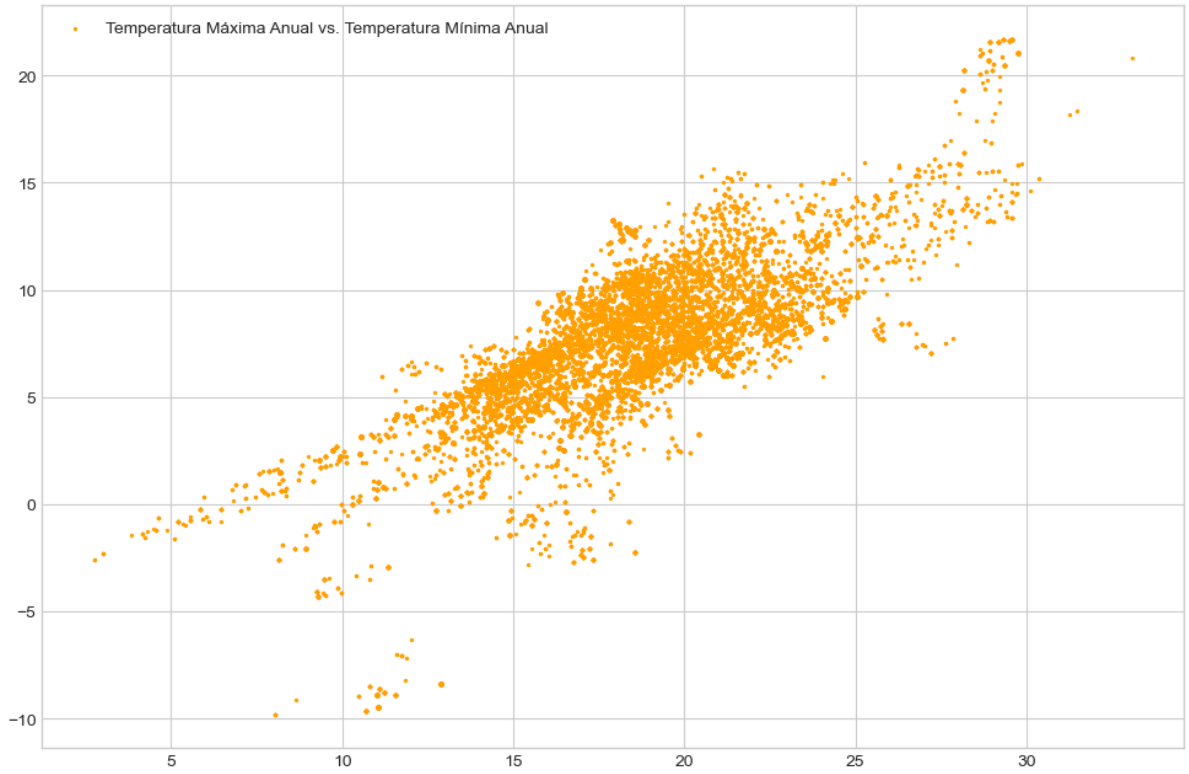
En este caso no se encuentra relación alguna entre la temperatura promedio anual y la precipitación anual, solo se percibe una clusterización de los valores ideales.

In [51]:
```
#Temperatura Promedio Anual vs. Evotranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_anual'], df['etp_anual'], s=2, label="Temperatura Promedi
o Anual vs. Evotranspiración Anual", color="#FFA000")
plt.legend();
plt.show()
```



En este caso si es interesante ver cierta correlación entre los valores de temperatura promedio y evotranspiración; esto debido a que el agua que requiere el viñedo está relacionada con la evaporación y esta por la temperatura de la zona.

In [52]:
```
#Temperatura Máxima Anual vs. Temperatura Mínima Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_max_anual'], df['temp_min_anual'], s=2, label="Temperatur
a Máxima Anual vs. Temperatura Mínima Anual", color="#FFA000")
plt.legend();
plt.show()
```

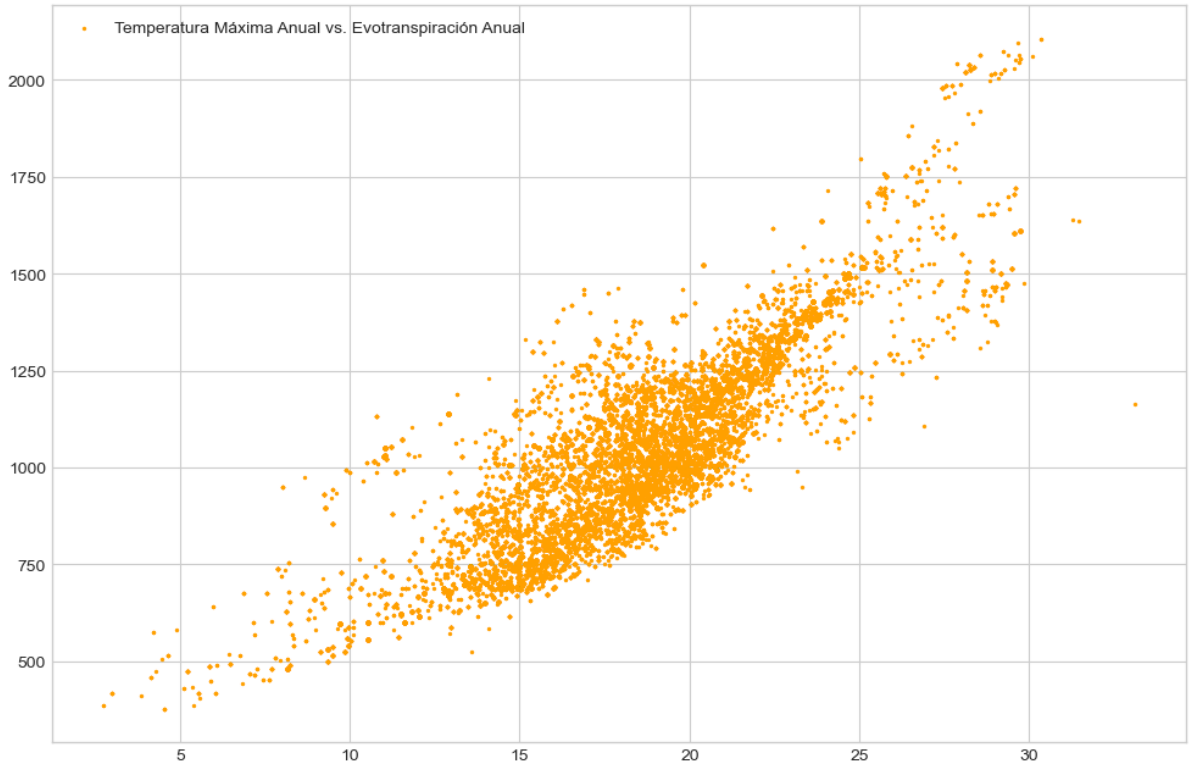


En este caso, se esperaba cierta correlación entre la temperatura máxima anual y mínima anual, debido a las zonas del muestreo y las estaciones, sin embargo la mayor parte de las temperaturas se encuentran en la clusterización de datos en el centro

In [53]:
```
#Temperatura Máxima Anual vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_max_anual'], df['pre_anual'], s=2, label="Temperatura Máx
ima Anual vs. Precipitación Anual", color="#FFA000")
plt.legend();
plt.show()
```

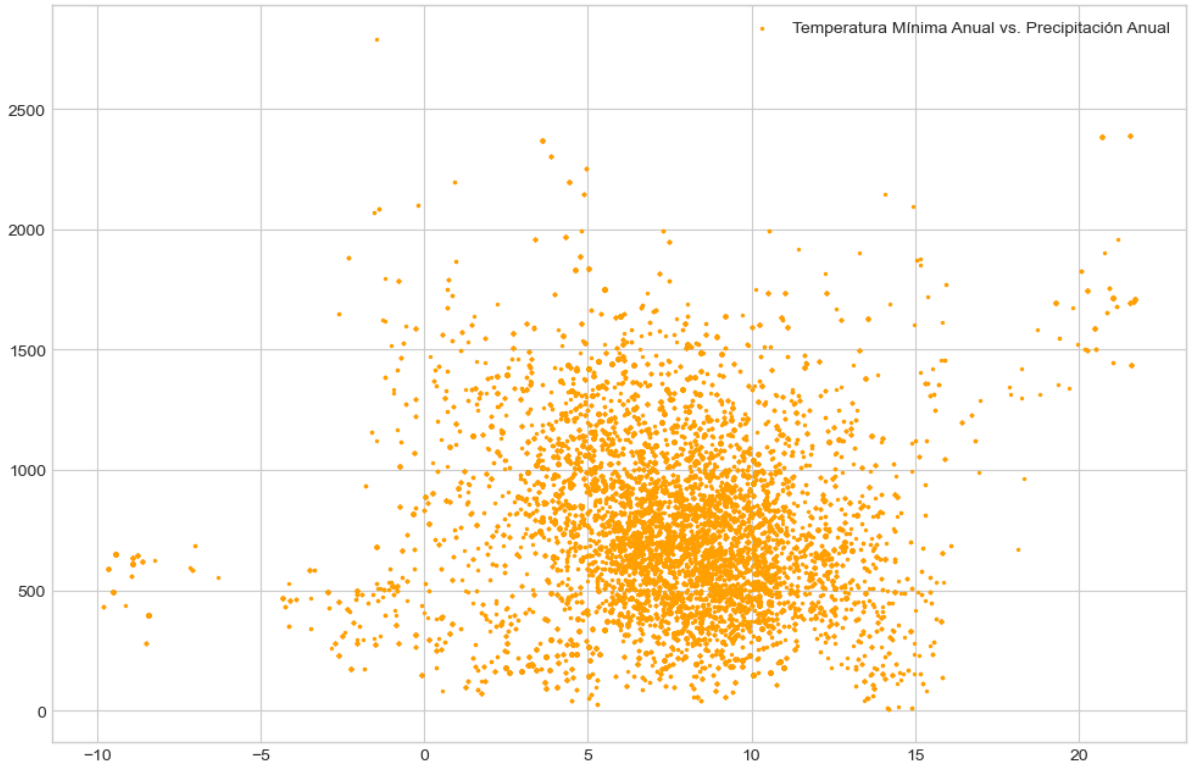
Se encuentra una clusterización entre la Temperatura Máxima Anual y la precipitación en el centro

In [54]:
```
#Temperatura Máxima Anual vs. Evotranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_max_anual'], df['etp_anual'], s=2, label="Temperatura Máx
ima Anual vs. Evotranspiración Anual", color="#FFA000")
plt.legend();
plt.show()
```



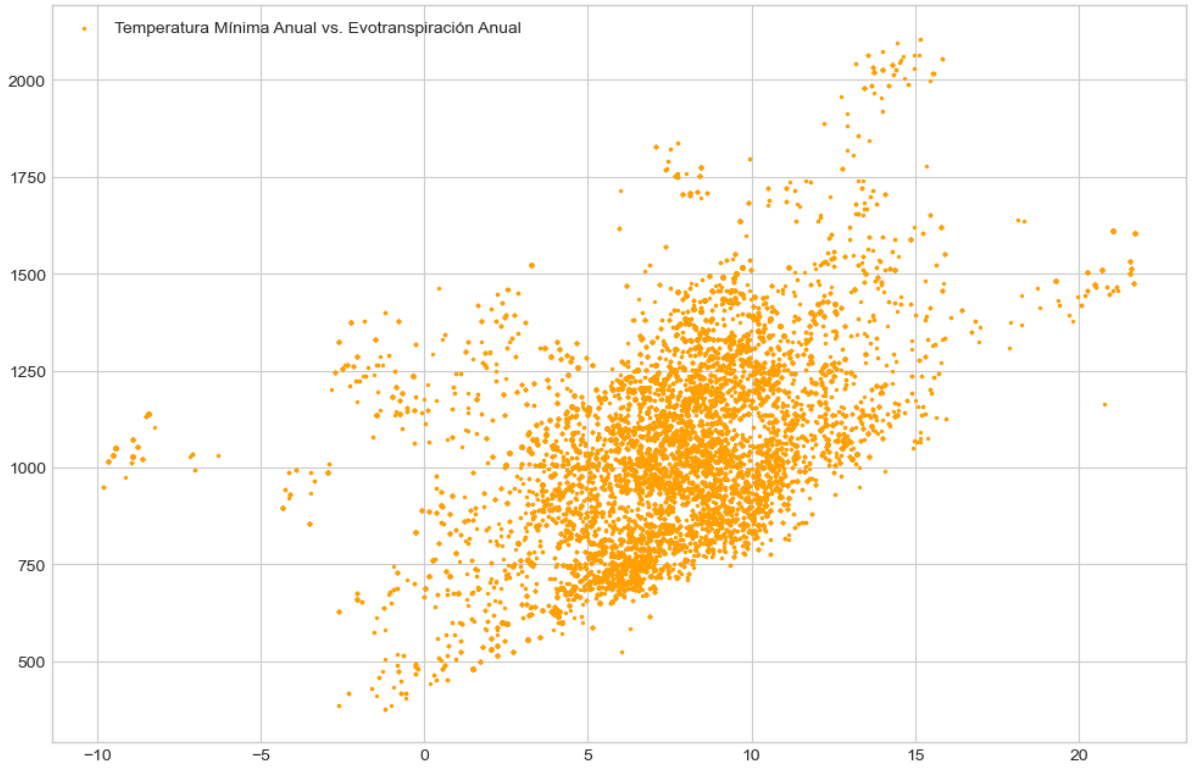
Al igual que con la temperatura promedio, se encuentra una correlación entre la temperatura máxima y la evotranspiración, a mayor temperatura, las plantas requieren mayor agua.

In [55]:
```
#Temperatura Mínima Anual vs. Precipitación Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_min_anual'], df['pre_anual'], s=2, label="Temperatura Mín
ima Anual vs. Precipitación Anual", color="#FFA000")
plt.legend();
plt.show()
```



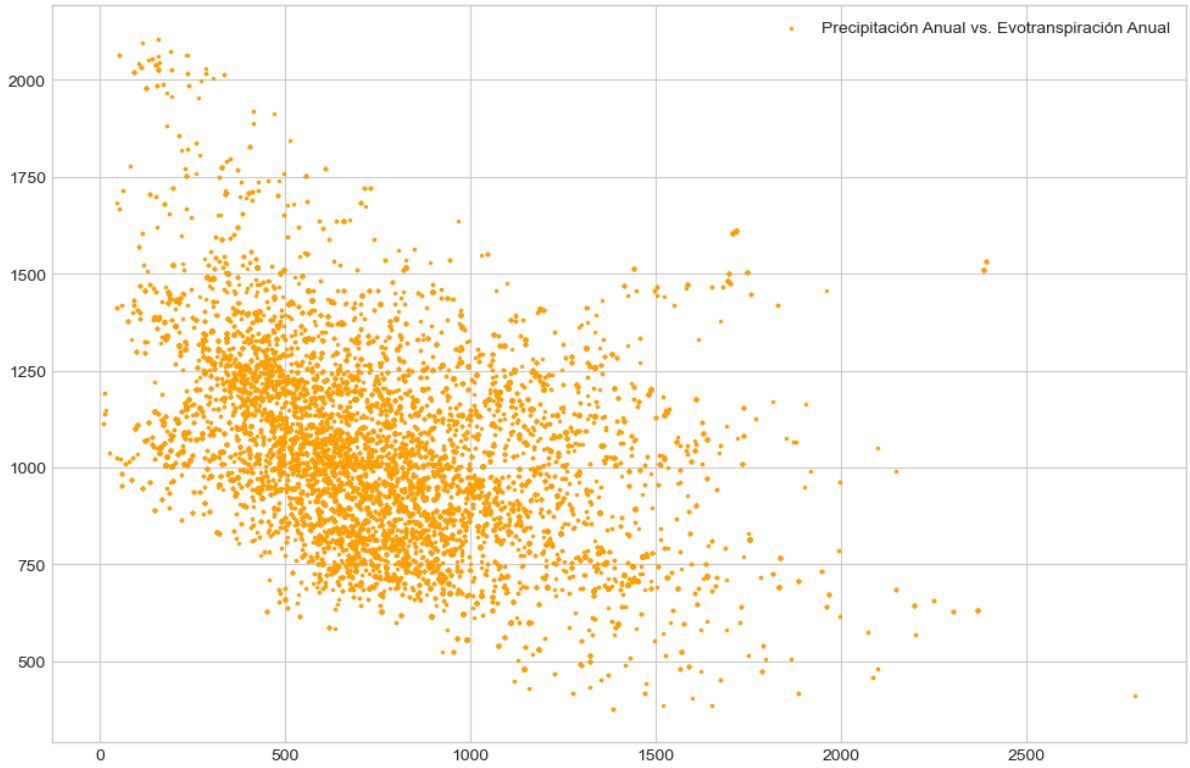
Nuevamente no se encuentra correlación entre la precipitación y la temperatura mínima anual, sin embargo es interesante ver la clusterización de datos

In [56]:
```
#Temperatura Máxima Anual vs. Evotranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['temp_min_anual'], df['etp_anual'], s=2, label="Temperatura Mín
ima Anual vs. Evotranspiración Anual", color="#FFA000")
plt.legend();
plt.show()
```



Aunque en este caso no es tan claro, se encuentra cierta clusterización entre la temperatura mínima anual y la evotranspiración. La parte de temperaturas bajo cero con mayor evotranspiración se puede deber a que el agua congelada no puede ser apropiadamente por las plantas y en este caso van a requerir más agua

In [57]:
```
#Precipitación Anual vs. Evotranspiración Anual
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['pre_anual'], df['etp_anual'], s=2, label="Precipitación Anual
vs. Evotranspiración Anual", color="#FFA000")
plt.legend();
plt.show()
```



Por último, se encuentra cierta relación entre la precipitación y la evotranspiración, ya que la evotranspiración depende de la precipitación. A mayor cantidad de lluvias, las plantas requieren menor agua, posiblemente por la humedad. Es posible también entender ciertos niveles atípicos por el análisis anteriormente mencionado, donde algunas zonas pueden tener condiciones de congelamiento de agua que puede no ser apropiado para la absorción del agua

In [58]:
```
#from scipy.stats import f_oneway

#grps = [d['points'] for _, d in dfr.groupby('country')]

#F, p = f_oneway(*grps)
#print(F, p)
```

Visualización Con Países de vinos de Alta Calidad

Para afirmar las conclusiones previas, se realiza un análisis de los vinos con mejor calidad.

Para esta vusalización de utiliza el Dataframe 2 (dfr), el cual contiene la información de países para así hacer un analisis por zonas. En este caso se analizarán los dataframes con puntaje alto (>96) y precio alto (100)

In [59]:
```
dfr.head()
```

Out[59]:

	Unnamed: 0	country	description	points	price	taster_name	variety	winery	Year
0	0	Portugal	This is ripe and fruity, a wine that is smooth...	87	15.0	Roger Voss	Portuguese Red	Quinta dos Avidagos	2011
1	1	Portugal	This is a solid mineral and tannin dominated w...	87	15.0	Roger Voss	Portuguese Red	Quinta Nova de Nossa Senhora do Carmo	2011
2	2	Portugal	This is a hard, edgy and tannic wine that has ...	87	17.0	Roger Voss	Portuguese Red	Quinta dos Aciprestes	2011
3	3	Portugal	The Reserva version of Mural makes a powerful ...	91	12.0	Roger Voss	Portuguese Red	Quinta do Portal	2011
4	4	Portugal	There is a strong barnyard aroma here, althoug...	87	8.0	Roger Voss	Portuguese Red	Borges	2011

In [60]:
```
dfpoints = dfr[dfr.points > 96]
dfprice = dfr[dfr.price > 199]
```

In [61]:
```
dfpoints.info()
```

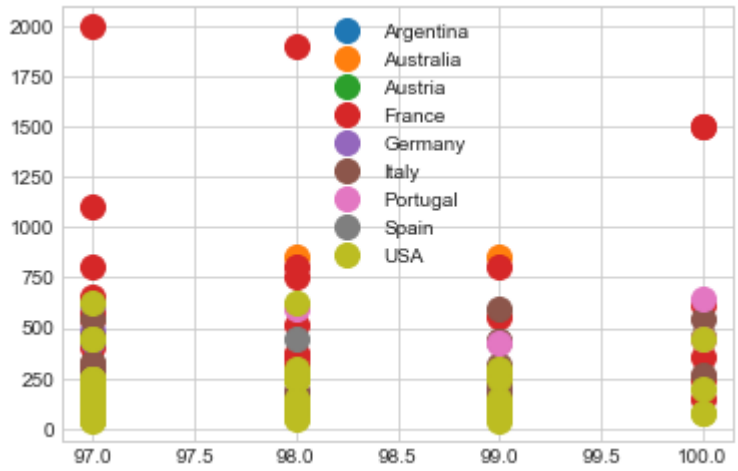
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 235 entries, 229 to 69972
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Unnamed: 0      235 non-null    int64  
 1   country         235 non-null    object 
 2   description     235 non-null    object 
 3   points          235 non-null    int64  
 4   price           235 non-null    float64
 5   taster_name     235 non-null    object 
 6   variety         235 non-null    object 
 7   winery          235 non-null    object 
 8   Year            235 non-null    int64  
 9   region          235 non-null    object 
 10  Latitude        235 non-null    float64
 11  Longitude       235 non-null    float64
 12  Lat_x           235 non-null    float64
 13  Long_x          235 non-null    float64
 14  temp_anual      235 non-null    float64
 15  temp_max_anual  235 non-null    float64
 16  temp_min_anual  235 non-null    float64
 17  pre_anual       235 non-null    float64
 18  etp_anual       235 non-null    int64  
dtypes: float64(9), int64(4), object(6)
memory usage: 36.7+ KB
```

In [62]:
```
dfprice.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 532 entries, 272 to 69972
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Unnamed: 0      532 non-null    int64  
 1   country         532 non-null    object 
 2   description     532 non-null    object 
 3   points          532 non-null    int64  
 4   price           532 non-null    float64
 5   taster_name     532 non-null    object 
 6   variety         532 non-null    object 
 7   winery          532 non-null    object 
 8   Year            532 non-null    int64  
 9   region          532 non-null    object 
 10  Latitude        532 non-null    float64
 11  Longitude       532 non-null    float64
 12  Lat_x           532 non-null    float64
 13  Long_x          532 non-null    float64
 14  temp_anual      532 non-null    float64
 15  temp_max_anual  532 non-null    float64
 16  temp_min_anual  532 non-null    float64
 17  pre_anual       532 non-null    float64
 18  etp_anual       532 non-null    int64  
dtypes: float64(9), int64(4), object(6)
memory usage: 83.1+ KB
```

In [63]:
```
np.random.seed(1984)

groups = dfpoints.groupby('country')
plt.figure(figsize=(12,8), dpi= 100)
fig, ax = plt.subplots()
#ax.margins(0.05)
for name, group, in groups:
    ax.plot(group.points, group.price, marker='o', linestyle='', ms=12, label=
name)
ax.legend()

plt.show
```

Out[63]:
```
<function matplotlib.pyplot.show(close=None, block=None)>

<Figure size 1200x800 with 0 Axes>
```



Se puede apreciar que los vinos de alta calidad provienen de los países:

- Argentina
- Australia
- Austria
- Francia
- Alemania
- Italia
- Portugal
- España
- Estados Unidos

In [64]:
```
#Histograma de Calidad Vs Países
dfcountry = dfpoints['country'].value_counts()
dfcountry
```

Out[64]:
```
USA          112
Italy         48
France        40
Portugal      15
Germany        6
Spain          5
Australia      5
Austria        3
Argentina      1
Name: country, dtype: int64
```

Se puede ver que la información obtenida no está balanceada, ya que la mayoría de vinos de buena calidad están en Estados Unidos, Italia y Francia (85% de la muestra), lo cual hara que este sesgo no permita tener conclusiones definitivas sobre algunos parámetros.

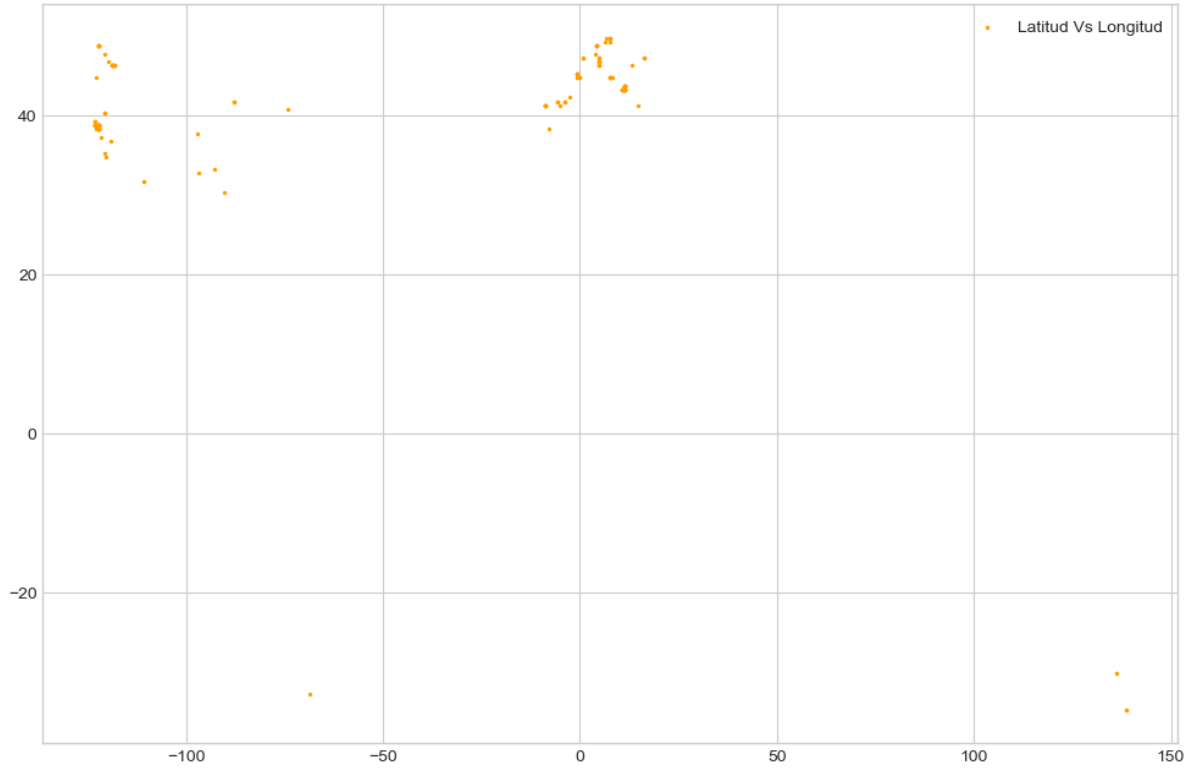
In [65]:
```
dfcountry = dfpoints['region'].value_counts()
dfcountry
```

Out[65]:

Napa Valley, USA	24
Brunello di Montalcino, Italy	14
Toscana, Italy	11
Russian River Valley, USA	11
Columbia Valley (WA), USA	10
Port, Portugal	10
Champagne, France	9
Sonoma Coast, USA	8
Walla Walla Valley (WA), USA	7
Barolo, Italy	7
Walla Walla Valley (OR), USA	6
Oakville, USA	6
Bolgheri Superiore, Italy	5
Saint-Julien, France	4
Langhe, Italy	4
Anderson Valley, USA	4
Bolgheri Sassicaia, Italy	4
Douro, Portugal	4
Sonoma County, USA	4
Bâtard-Montrachet, France	3
Atlas Peak, USA	3
Vouvray, France	3
Rheingau, Germany	3
Ribera del Duero, Spain	3
St. Helena, USA	3
Burgenland, Austria	3
Chevalier-Montrachet, France	2
Oak Knoll District, USA	2
Green Valley, USA	2
Margaux, France	2
Stags Leap District, USA	2
Pauillac, France	2
Pessac-Léognan, France	2
Barossa Valley, Australia	2
Diamond Mountain District, USA	2
Saint-Émilion, France	2
South Australia, Australia	2
El Dorado, USA	1
Wahluke Slope, USA	1
Corton-Pougets, France	1
Mosel-Saar-Ruwer, Germany	1
Santa Cruz Mountains, USA	1
Barbaresco, Italy	1
Toro, Spain	1
Clos de Tart, France	1
Clos de la Roche, France	1
Howell Mountain, USA	1
Alexander Valley, USA	1
Saint-Estèphe, France	1
Washington, USA	1
Pfalz, Germany	1
Knights Valley, USA	1
Napa-Mendocino-Sonoma-Marin, USA	1
Charmes-Chambertin, France	1
Mosel, Germany	1
North Coast, USA	1
Carneros, USA	1

```
Pomerol, France                          1
Sauternes, France                        1
Santa Maria Valley, USA                  1
Mount Veeder, USA                        1
Chablis, France                          1
Sonoma Valley, USA                       1
Taurasi, Italy                           1
Alentejo, Portugal                       1
Fort Ross-Seaview, USA                   1
Barossa, Australia                       1
Colli Orientali del Friuli, Italy        1
Willamette Valley, USA                   1
Arroyo Grande Valley, USA                1
Grands-Echezeaux, France                 1
Rutherford, USA                          1
California, USA                          1
Rioja, Spain                             1
Montrachet, France                       1
Mendoza, Argentina                       1
Bienvenues Bâtard-Montrachet, France     1
Name: region, dtype: int64
```

In [66]:
```
#Latitud vs. Longitud (Invertido para poder ver el mapa mundial y los vinos de la muestra)
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(dfpoints['Long_x'], dfpoints['Lat_x'], s=2, label="Latitud Vs Longitud", color="#FFA000")
plt.legend();
plt.show()
```



Como se esperaba, la representación gráfica de los vinos está centrada principalmente en la zona de la Costa Oeste de Estados Unidos, y Europa Oeste

In [67]:
```
pd.options.display.float_format = '{:.5f}'.format
dfpoints.var()
```

```
C:\Users\Public\Documents\Wondershare\CreatorTemp/ipykernel_16796/675791168.p
y:2: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with
'numeric_only=None') is deprecated; in a future version this will raise TypeEr
ror.  Select only valid columns before calling the reduction.
  dfpoints.var()
```

Out[67]:
```
Unnamed: 0       379114843.82135
points                   0.86550
price                74074.80480
Year                     5.79353
Latitude               157.54530
Longitude             4639.52248
Lat_x                  158.05576
Long_x                4632.36186
temp_anual               4.53836
temp_max_anual           7.42201
temp_min_anual           4.74840
pre_anual            91507.14935
etp_anual            46346.31849
dtype: float64
```

Realizando el cálculo de Varianza, se encuentra: El Puntaje no tiene una varianza ya que el análisis está hecho con datos de los mejores vinos El precio tiene una varianza alta, por lo que se pueden encontrar vinos de diferentes precios La varianza de año es acorde con los valores de años usados La Latitud no tiene una varianza tan alta como la lingitud ya que la mayoría de valores se encuentran agrupados en la zona del trópico de capricornio La temperatura tiene una varianza baja. Esto es por que posiblemente la mayoría de valores se encuentran agrupados, así como la temperatura máxima y mínima Los valores de Precipitación y Evotranspiración si tienen una alta varianza. Esto si es más interesante de analizar, ya que se esperaba una varianza muy baja

In [68]:
```
dfpoints.std()
```

```
C:\Users\Public\Documents\Wondershare\CreatorTemp/ipykernel_16796/1982406448.p
y:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with
'numeric_only=None') is deprecated; in a future version this will raise TypeEr
ror.  Select only valid columns before calling the reduction.
  dfpoints.std()
```

Out[68]:
```
Unnamed: 0       19470.87168
points               0.93032
price              272.16687
Year                 2.40697
Latitude            12.55170
Longitude           68.11404
Lat_x               12.57202
Long_x              68.06146
temp_anual           2.13034
temp_max_anual       2.72434
temp_min_anual       2.17908
pre_anual          302.50149
etp_anual          215.28195
dtype: float64
```

Análisis ANOVA

En este apartado se analiza el ANOVA de algunos valores con respecto a las variables de país y región, para entender si existe una varianza muy alta al respecto

In [ ]:
```
#We import the library to do ANOVA one_way
from scipy.stats import f_oneway
```

In [120]:
```
#Se realiza el ANOVA de puntos por país en el Dataset dfpoints
grps = [d['points'] for _, d in dfpoints.groupby('country')]
F, p = f_oneway(*grps)
print(F, p)
```

```
1.5109784235297654 0.15432219907311287
```

En este caso, F no es un valor alto, y p es mayor a .05, lo cual quiere decir que no existe una diferencia estadísitcamente significativa entre las medias de los grupos, es decir, que la puntuación de los vinos no está directamente ligada a donde se ha cosechado el vino (país)

In [124]:
```
grps = [d['points'] for _, d in dfpoints.groupby('region')]
F, p = f_oneway(*grps)
print(F, p)
```

```
0.6567595685159985 0.9795574204365326
```

En este caso, F es menor a 1, lo cual indica que la variancia entre los grupos no es mayor que la variancia total, y como el valor de p es mayor a .05, indica que no existe una diferencia estadísticamente significativa entre las medias, es decir, que la puntuación de los vinos no está directamente ligada a donde se ha cosechado el vino (region)

In [125]:
```
grps = [d['points'] for _, d in dfpoints.groupby('variety')]
F, p = f_oneway(*grps)
print(F, p)
```

```
0.7483133935941751 0.8173788888234723
```

En este caso, F es menor a 1, lo cual indica que la variancia entre los grupos no es mayor que la variancia total, y como el valor de p es mayor a .05, indica que no existe una diferencia estadísticamente significativa entre las medias de los grupos, es decir, que la puntuación no está directamente ligada al varietal del vino

In [113]:
```
grps = [d['points'] for _, d in dfpoints.groupby('Year')]
F, p = f_oneway(*grps)
print(F, p)
```

```
1.0371161847970787 0.4145405075681076
```

En este caso, F es cercano a 1, lo cual indica que la variancia entre los grupos no es mayor que la variancia total, y como el valor de p es mayor a .05, indica que no existe una diferencia estadísticamente significativa entre las medias de los grupos, lo que quiere decir que la puntuación no está directamente ligada al año de cosecha del vino.

In [115]:
```
grps = [d['price'] for _, d in dfprice.groupby('country')]
F, p = f_oneway(*grps)
print(F, p)
```

```
3.198467398291683 0.0005277965011465536
```

En este caso, F es mayor a 1, lo cual indica que la variancia entre los grupos es mayor que la variancia total, pero el valor de p es menor a .05, indica que si existe una diferencia estadísticamente significativa entre las medias de los grupos, por lo cual podemos concluir que el costo del vino tiene relación con el país donde se produce.

In [126]:
```
grps = [d['price'] for _, d in dfprice.groupby('region')]
F, p = f_oneway(*grps)
print(F, p)
```

```
4.17194959399146 1.418515376111874e-25
```

En este caso, F es mayor a 1, lo cual indica que la variancia entre los grupos es mayor que la variancia total, pero el valor de p es menor a .05, indica que si existe una diferencia estadísticamente significativa entre las medias de los grupos, por lo cual podemos concluir que el costo del vino tiene relación con la región donde se cosecha el vino.

In [118]:
```
grps = [d['price'] for _, d in dfprice.groupby('variety')]
F, p = f_oneway(*grps)
print(F, p)
```

```
2.362509348890116 3.0420674255934102e-05
```

En este caso, F es mayor a 1, lo cual indica que la variancia entre los grupos es mayor que la variancia total, pero el valor de p es menor a .05, indica que si existe una diferencia estadísticamente significativa entre las medias de los grupos, por lo cual podemos concluir que el costo del vino varía dependiendo del varietal.

In [119]:
```
grps = [d['price'] for _, d in dfprice.groupby('Year')]
F, p = f_oneway(*grps)
print(F, p)
```

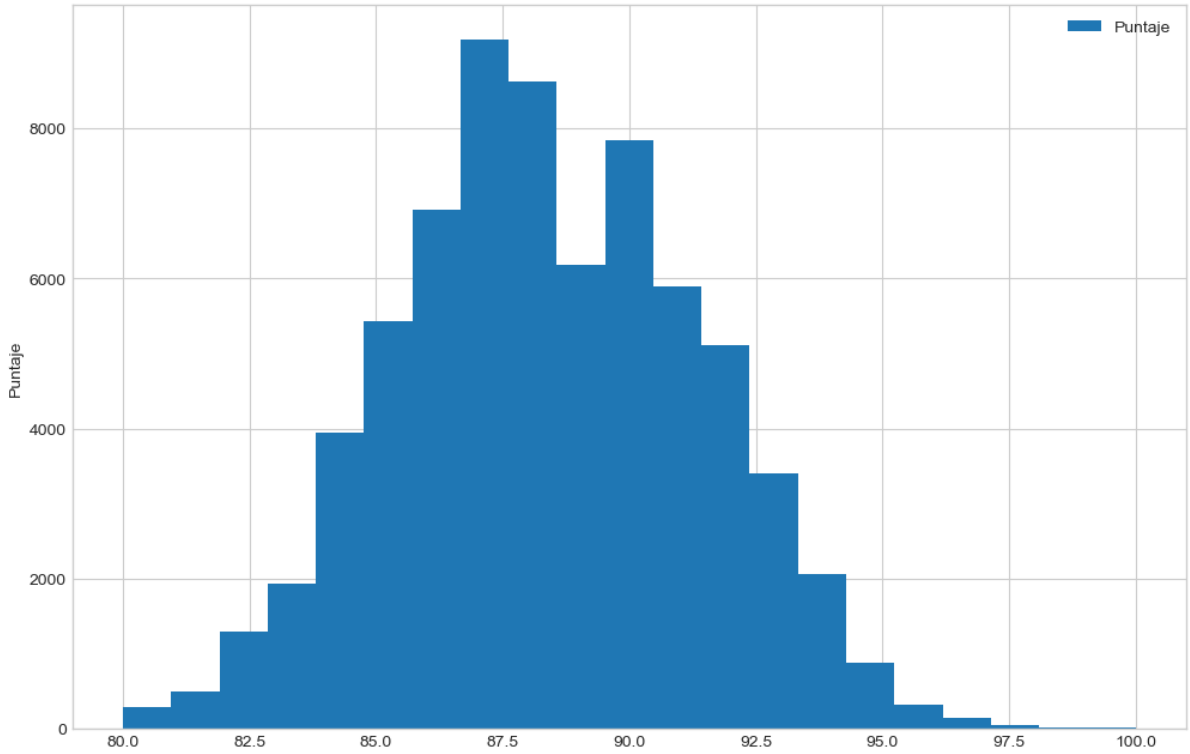
```
1.0683933538878336 0.3847538841986532
```

En este caso, F es cercano a 1, lo cual indica que la variancia entre los grupos no es mayor que la variancia total, pero el valor de p es mayor a .05, indica que si no existe una deferencia significativa en las medias, es decir, que no se encuentra una relación clara del costo del vino dependiendo de su año de cosecha

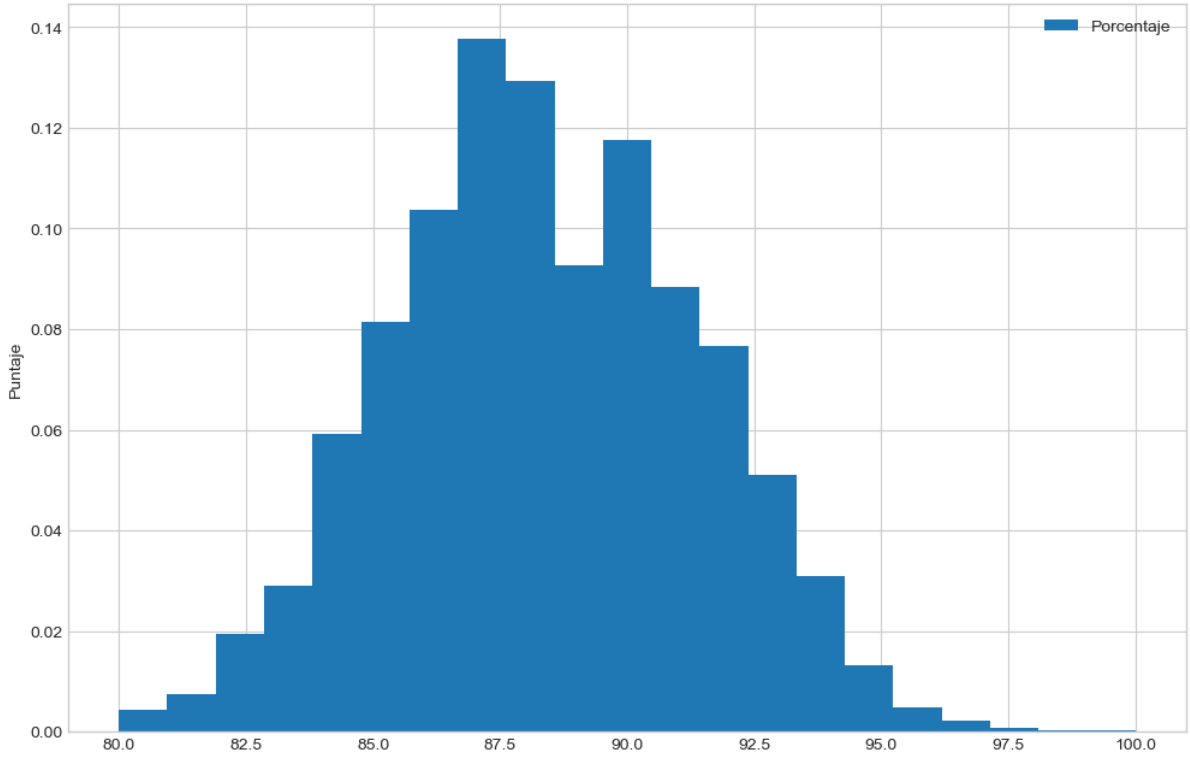
Visualizacion

Histogramas

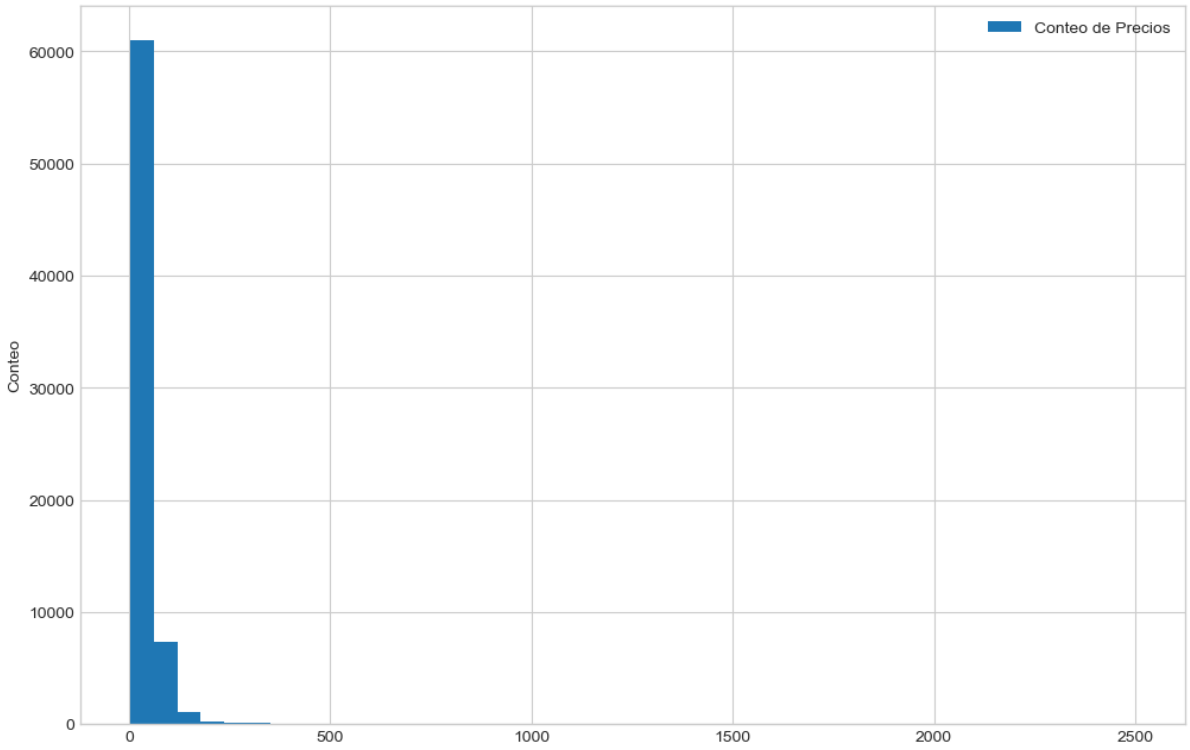
In [69]:
```
#Histograma de Puntos
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['points'], bins=21, label="Puntaje")
plt.legend()
plt.ylabel("Puntaje");
```



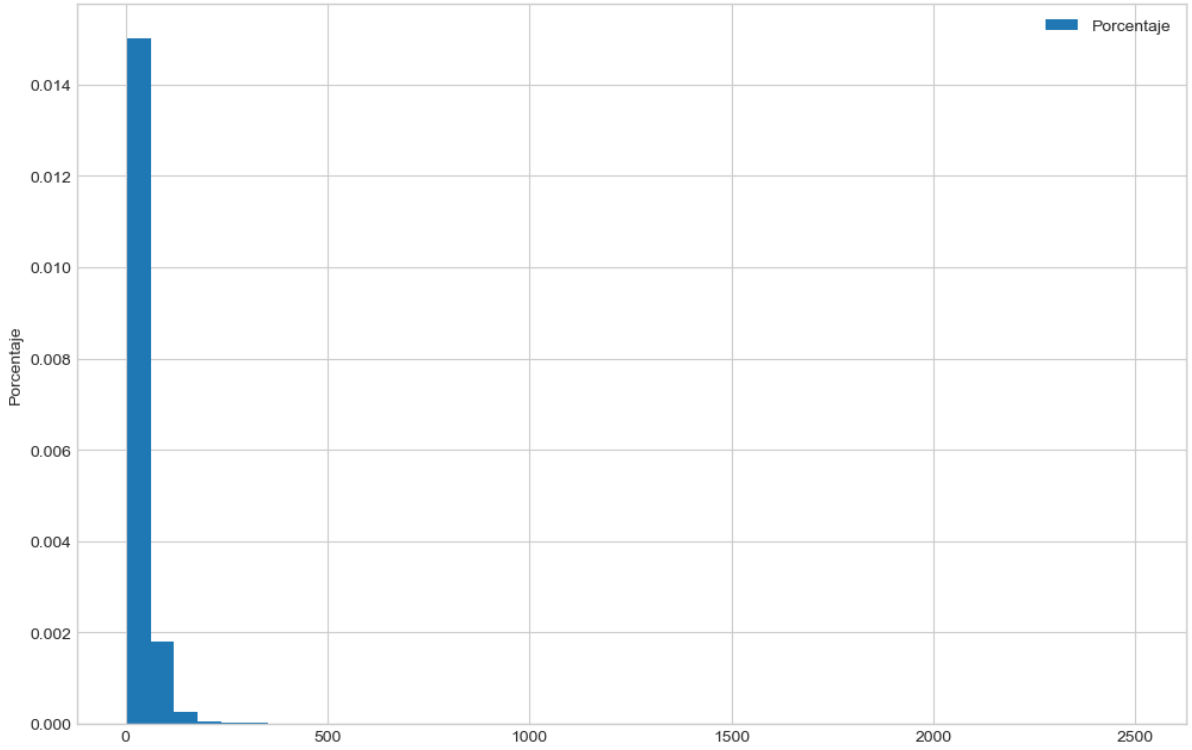
In [70]:
```
#Puntaje en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['points'], bins=21, label="Porcentaje", density=True)
plt.legend()
plt.ylabel("Puntaje");
```



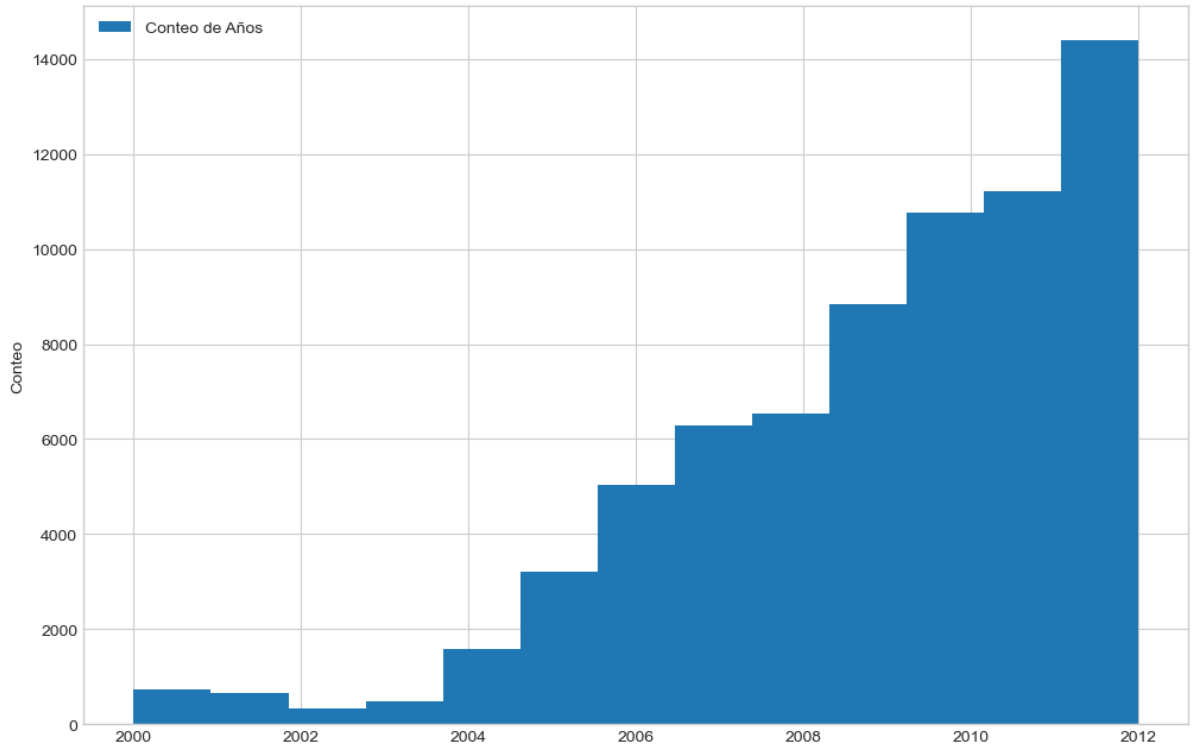
In [71]:
```
#Histograma de Precios
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['price'], bins=43, label="Conteo de Precios")
plt.legend()
plt.ylabel("Conteo");
```



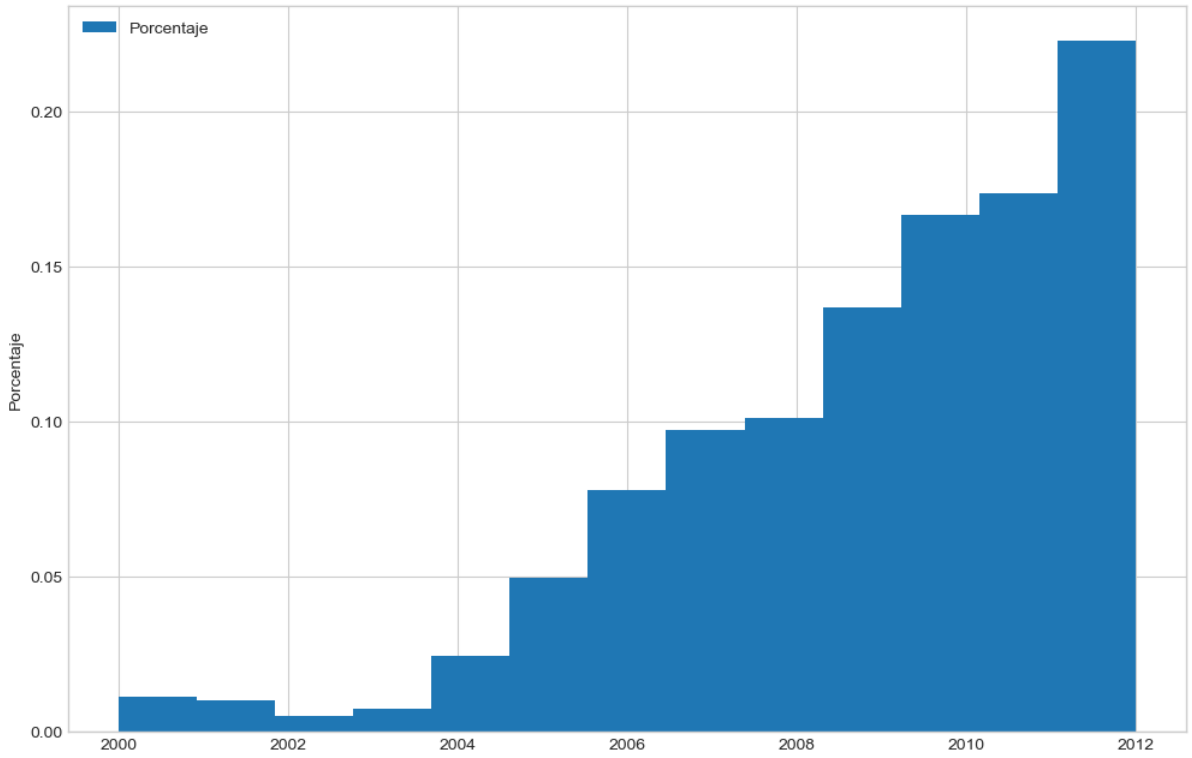
In [72]:
```
#Precio en porcentajes, por que no muestra los porcentajes correctos?
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['price'], bins=43, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



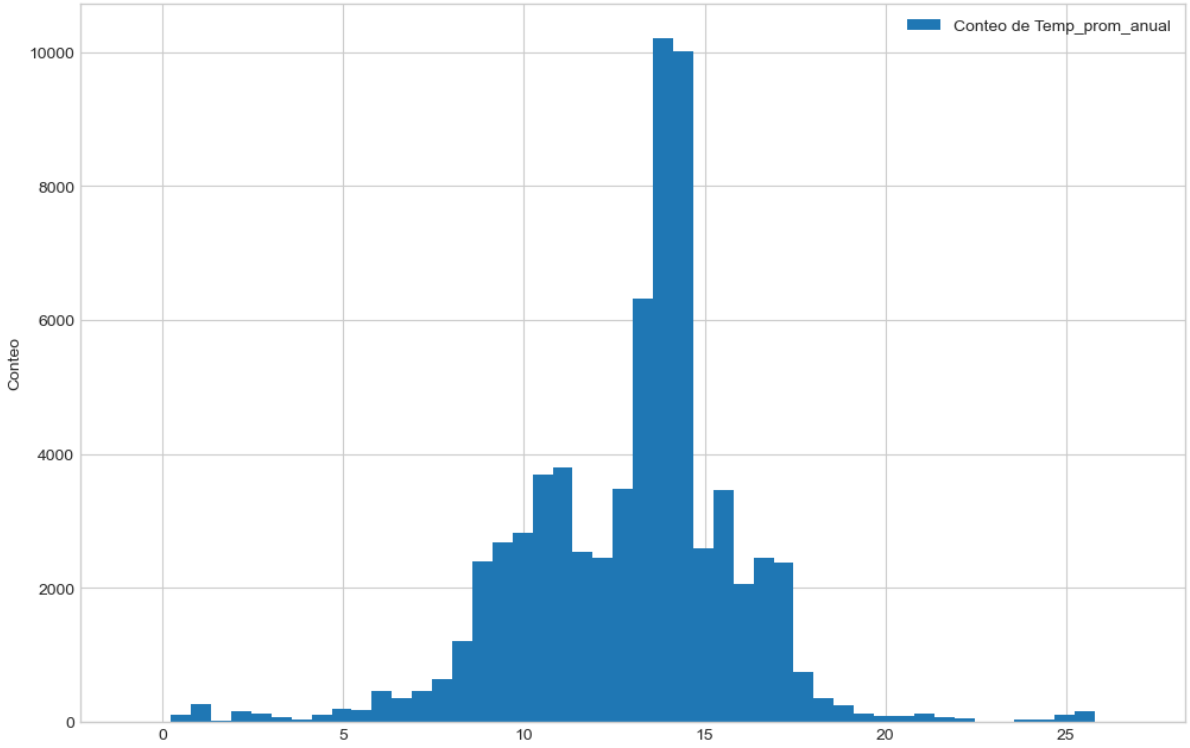
In [73]:
```
#Histograma de Años
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['Year'], bins=13, label="Conteo de Años")
plt.legend()
plt.ylabel("Conteo");
```



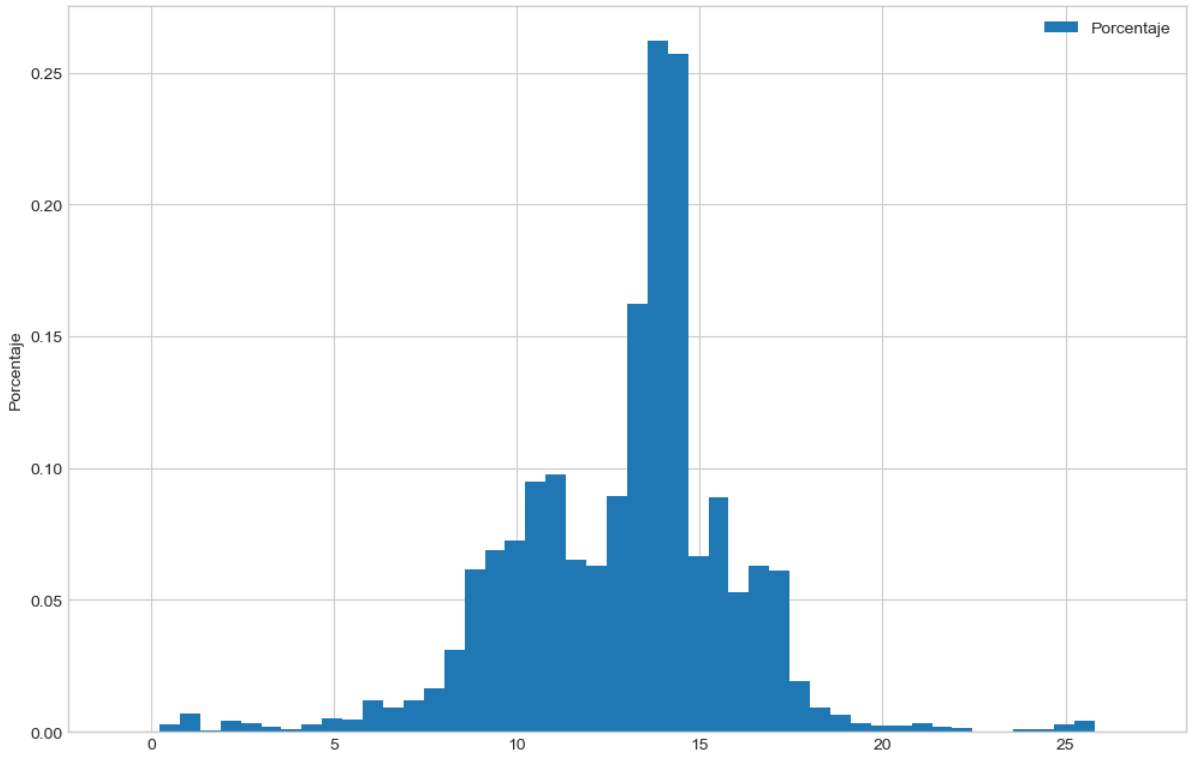
In [74]:
```
#Años en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['Year'], bins=13, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



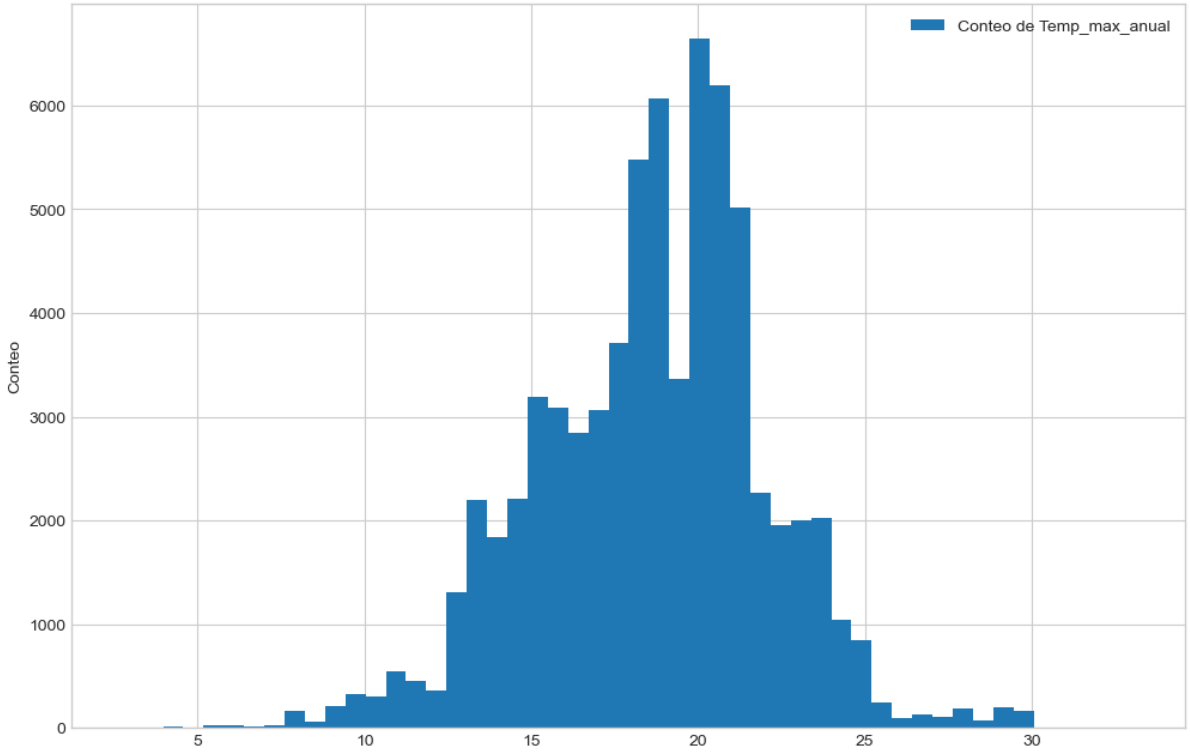
In [75]:
```
#Histograma de Temperatura promedio anual
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_anual'], bins=50, label="Conteo de Temp_prom_anual")
plt.legend()
plt.ylabel("Conteo");
```



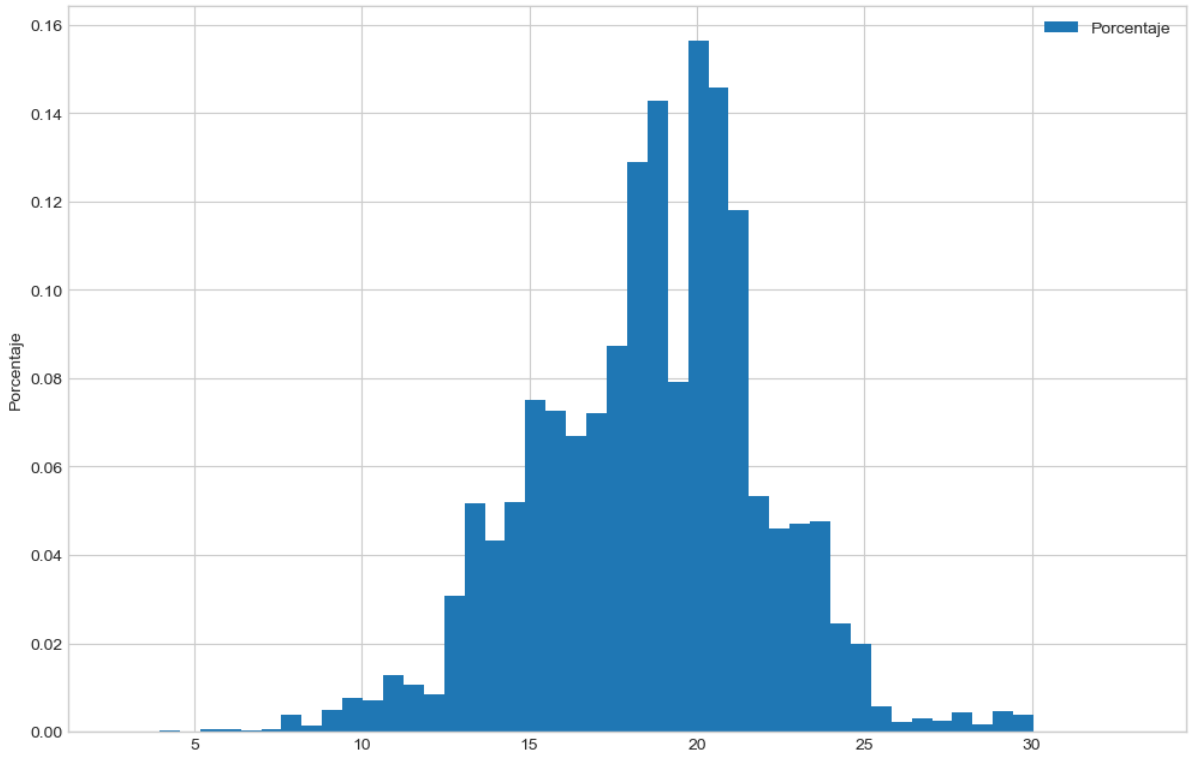
In [76]:
```
#Temp_anual en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_anual'], bins=50, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



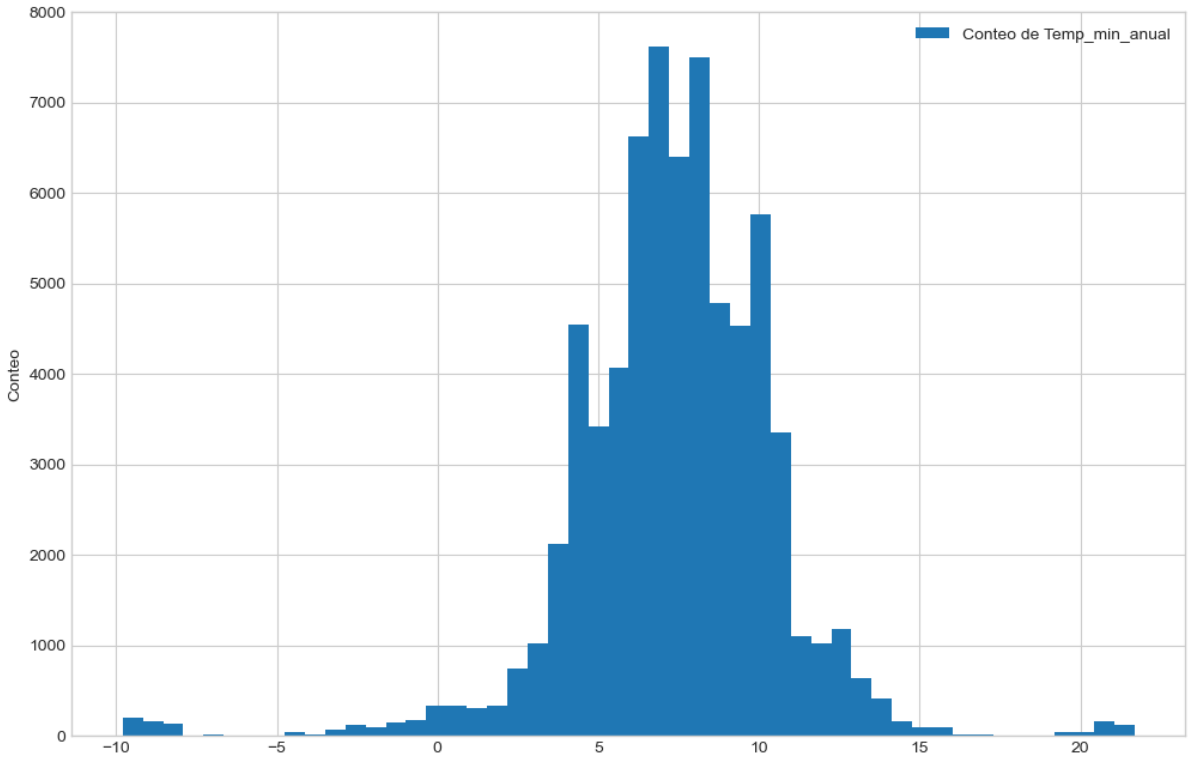
In [77]:
```
#Histograma de Temperatura maxima anual
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_max_anual'], bins=50, label="Conteo de Temp_max_anual")
plt.legend()
plt.ylabel("Conteo");
```



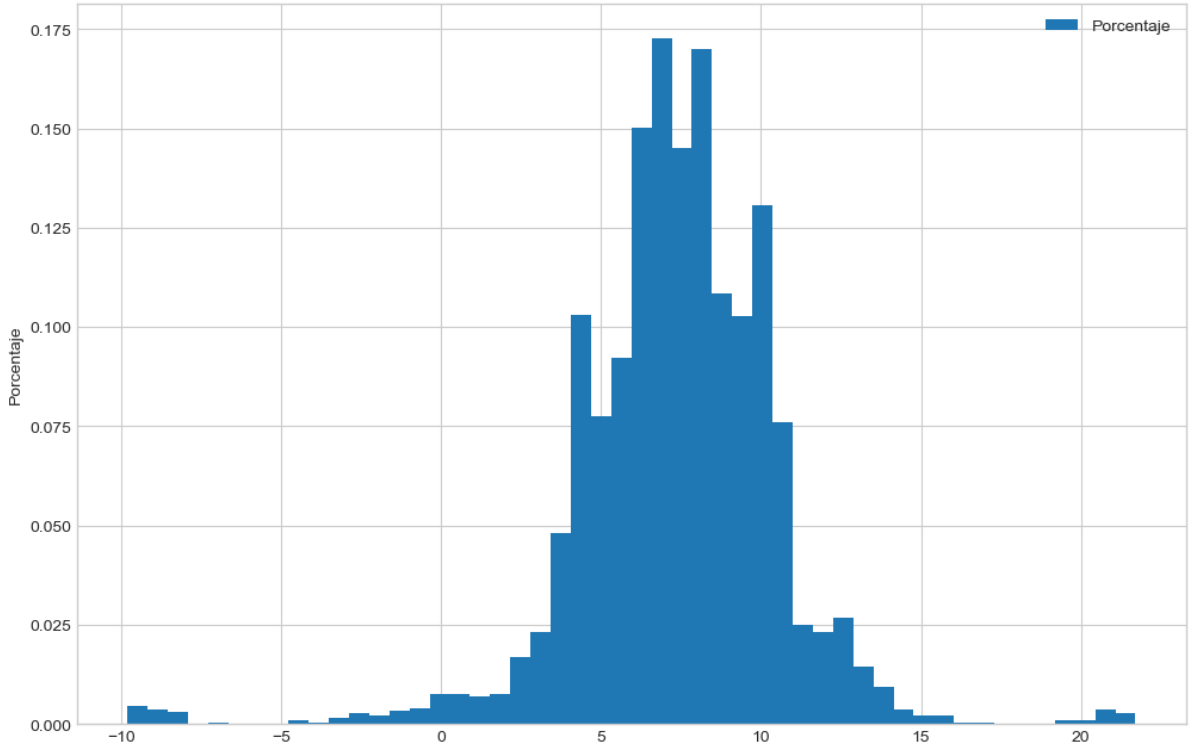
In [78]:
```
#Temp_max_anual en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_max_anual'], bins=50, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



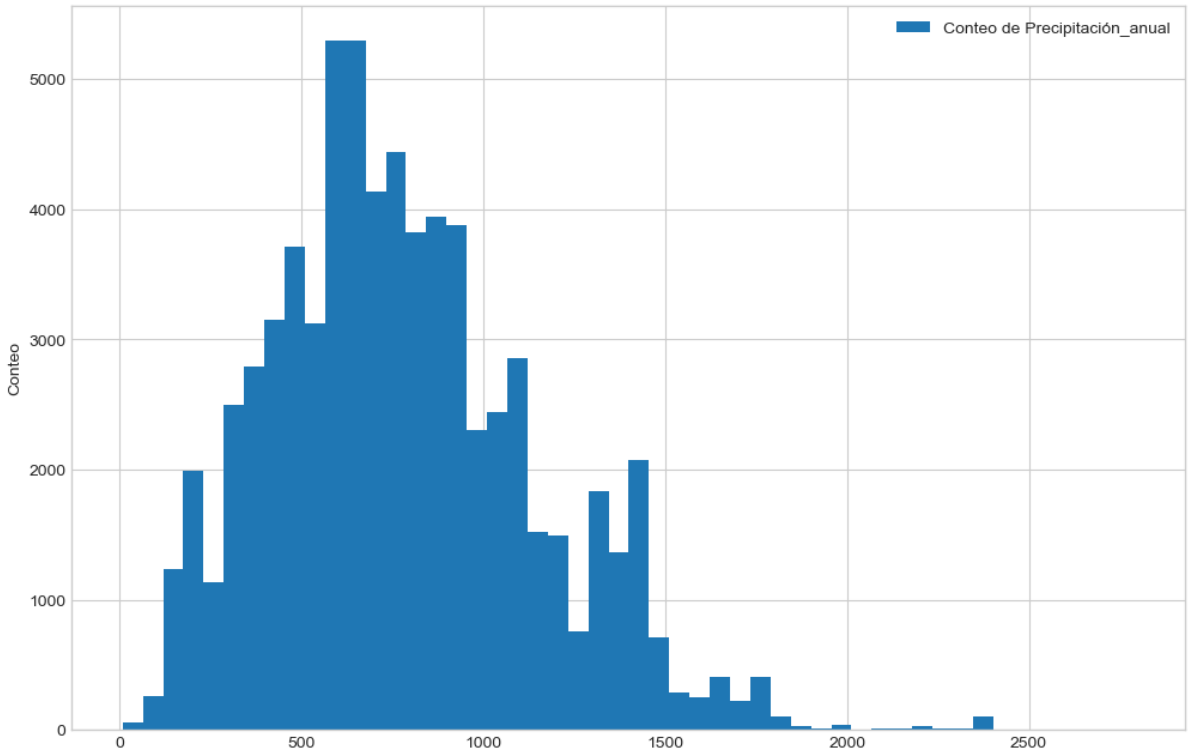
In [79]:
```
#Histograma de Temperatura mínima anual
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_min_anual'], bins=50, label="Conteo de Temp_min_anual")
plt.legend()
plt.ylabel("Conteo");
```



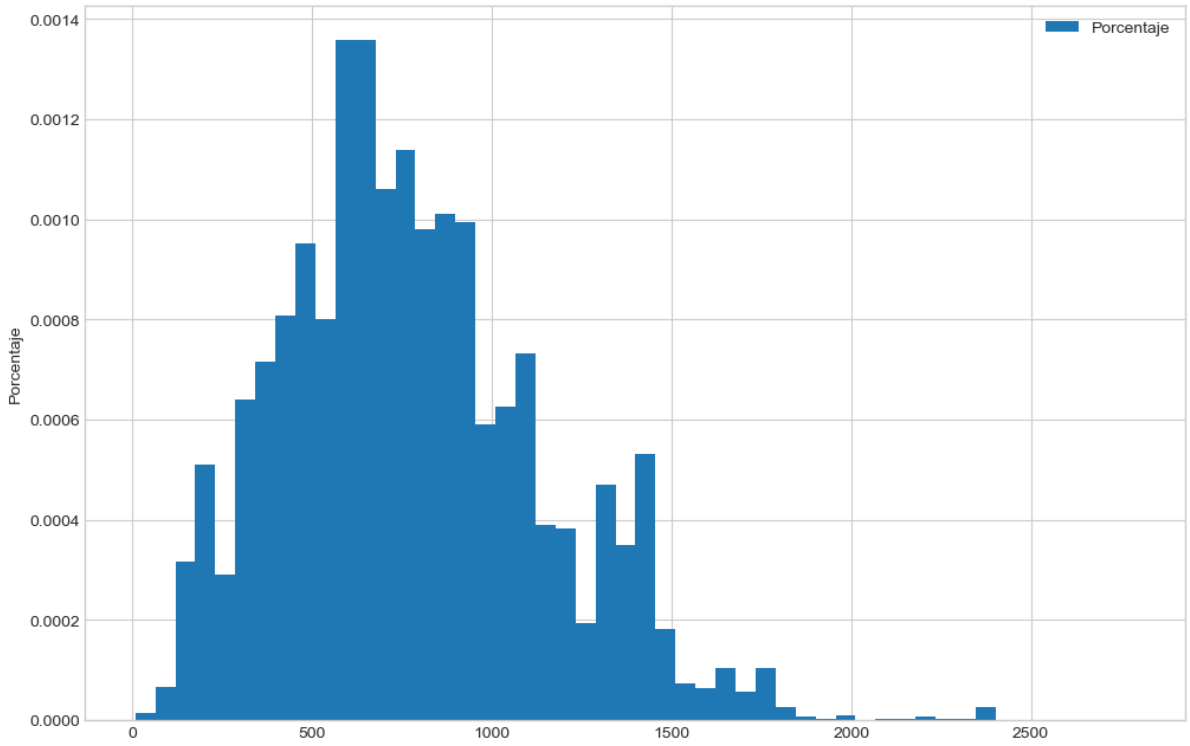
In [80]:
```
#Temp_min_anual en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['temp_min_anual'], bins=50, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



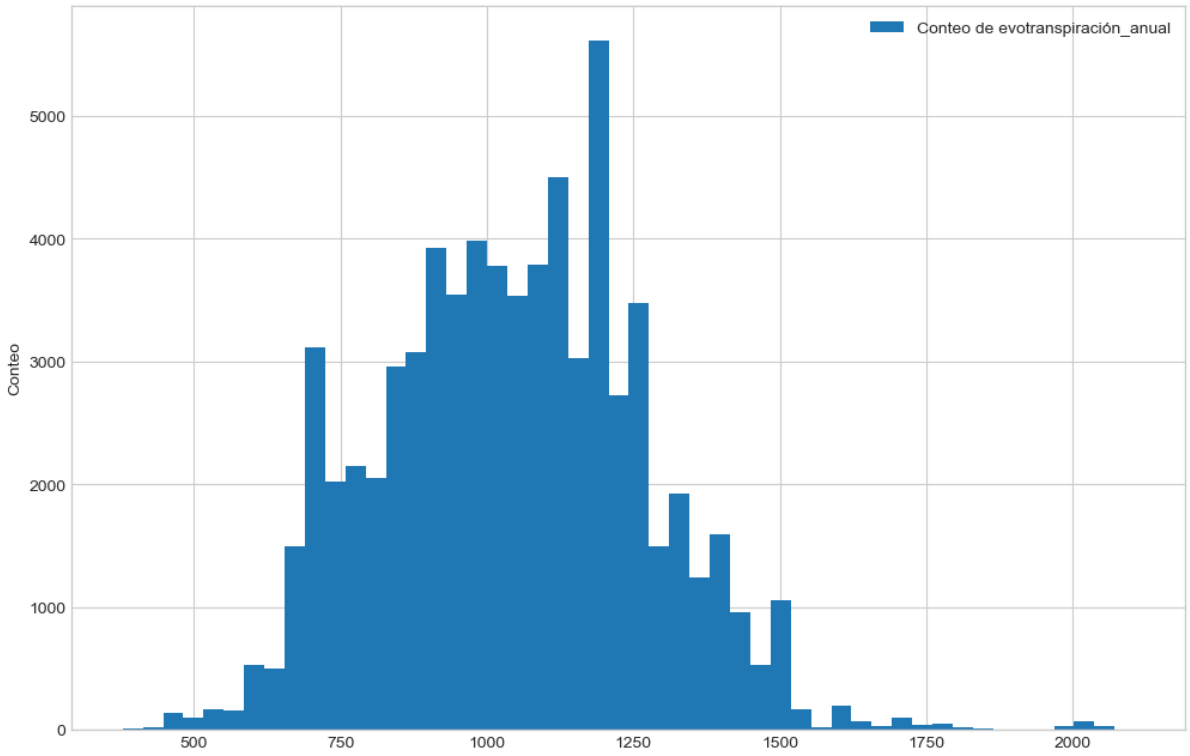
In [81]:
```
#Histograma de precipitación anual
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['pre_anual'], bins=50, label="Conteo de Precipitación_anual")
plt.legend()
plt.ylabel("Conteo");
```



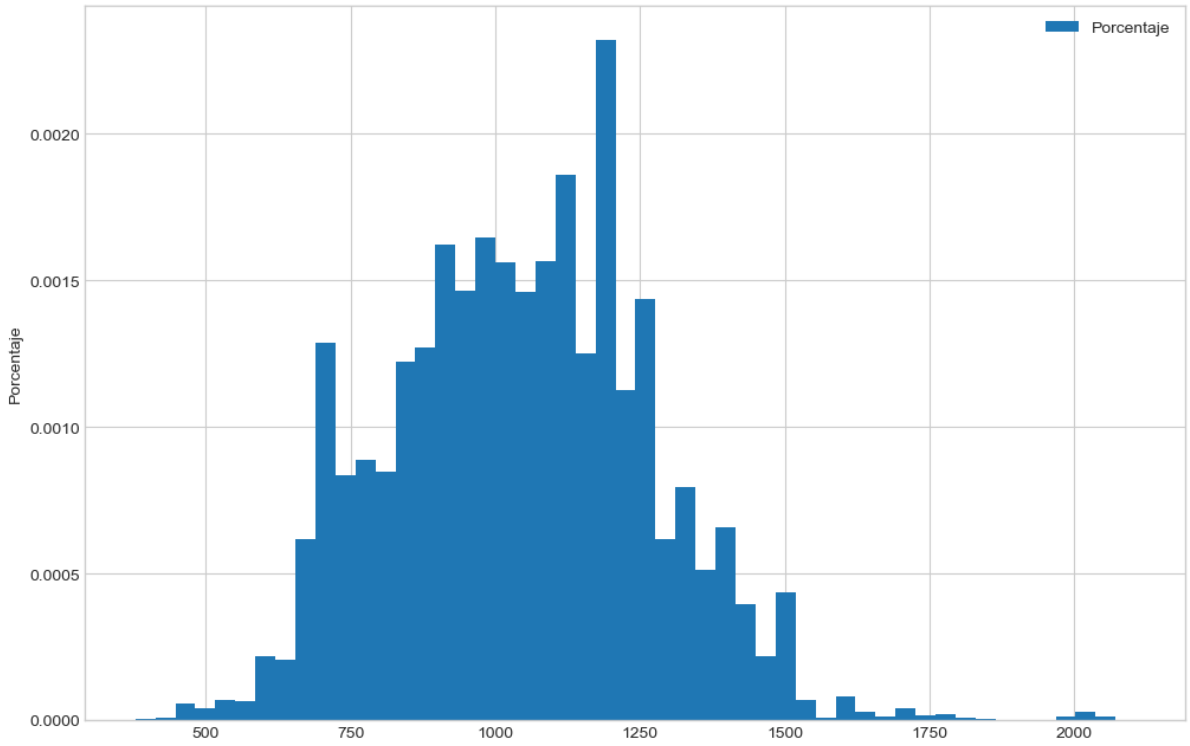
In [82]:
```
#Precipitación_anual en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['pre_anual'], bins=50, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



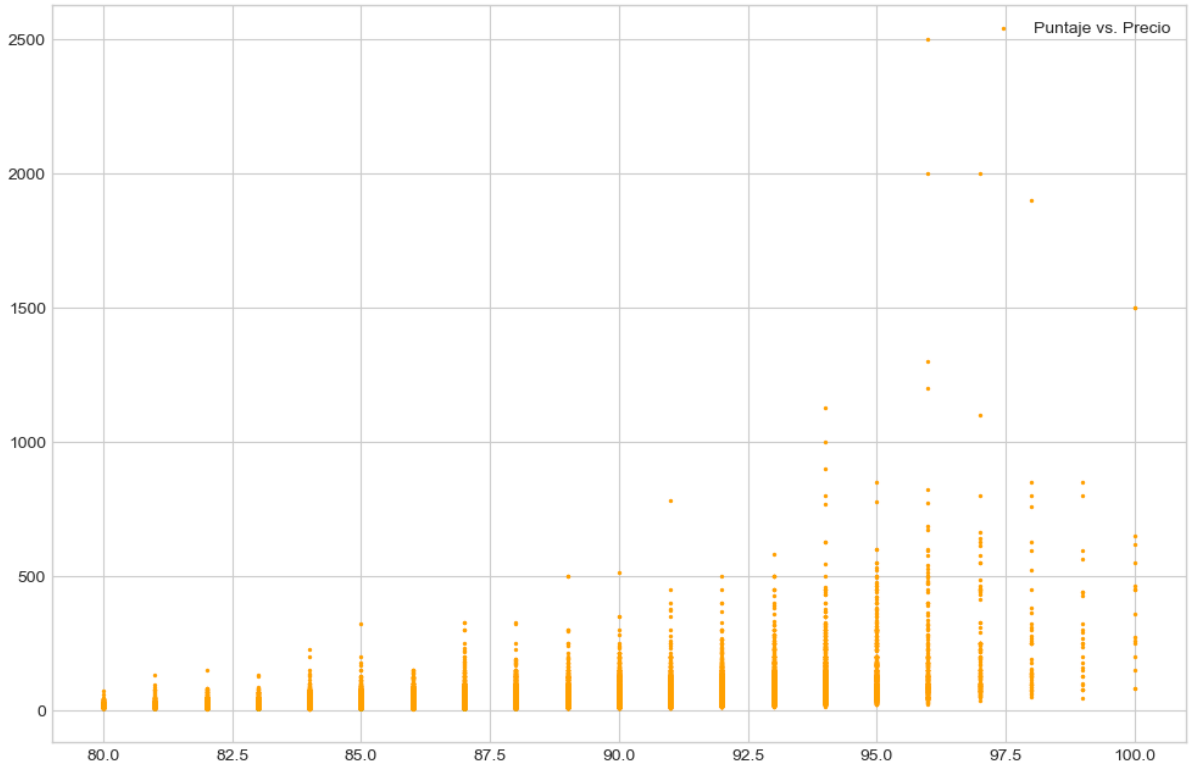
In [83]:
```
#Histograma de evotranspiración anual
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['etp_anual'], bins=50, label="Conteo de evotranspiración_anual")
plt.legend()
plt.ylabel("Conteo");
```



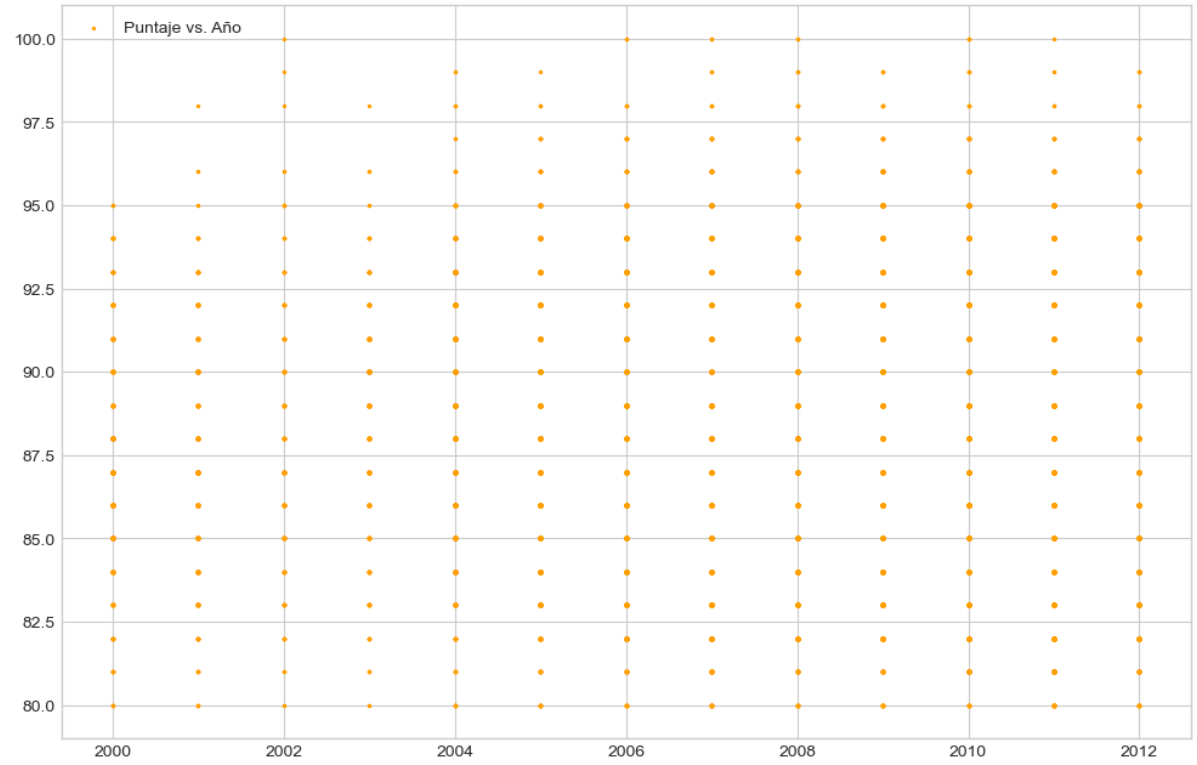
In [84]:
```
#evotranspiración_anual en porcentajes
plt.figure(figsize=(12,8), dpi= 100)
plt.hist(df['etp_anual'], bins=50, label="Porcentaje", density=True)
plt.legend();
plt.ylabel("Porcentaje");
```



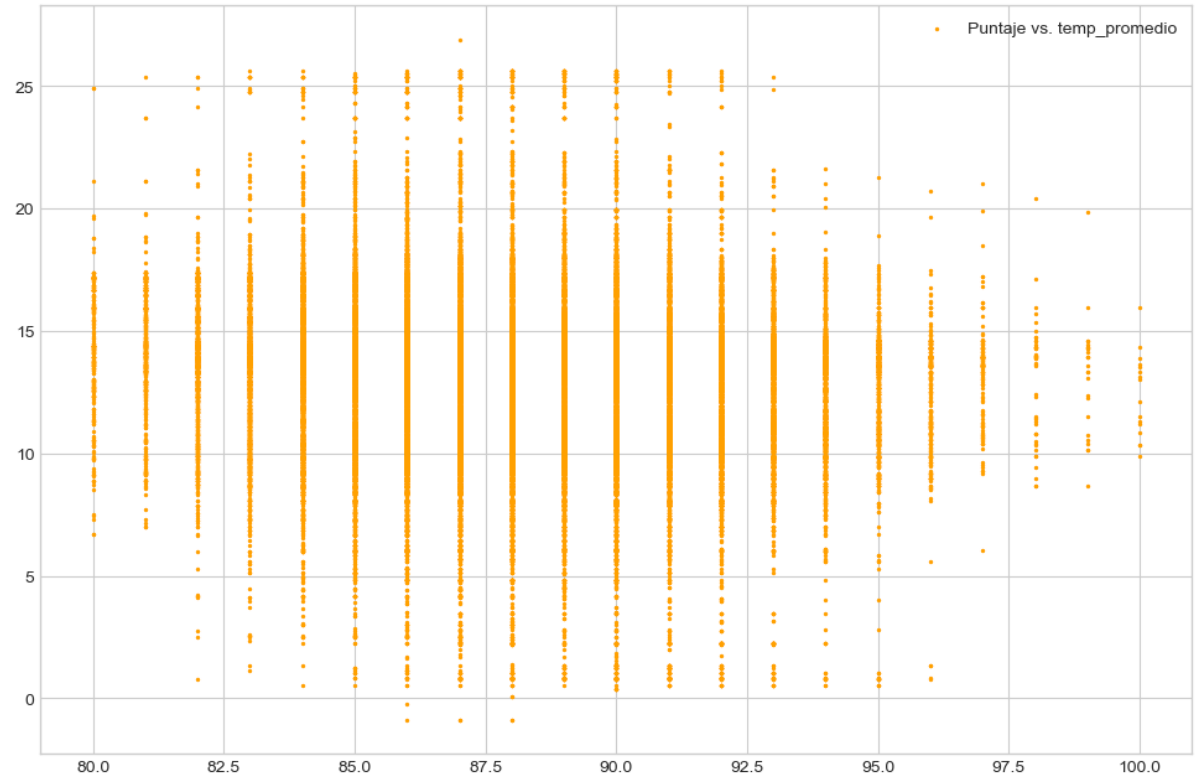
In [85]:
```
#Scatter entre puntaje y precios
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['price'], s=2, label="Puntaje vs. Precio", color
="#FFA000")
plt.legend();
plt.show()
```



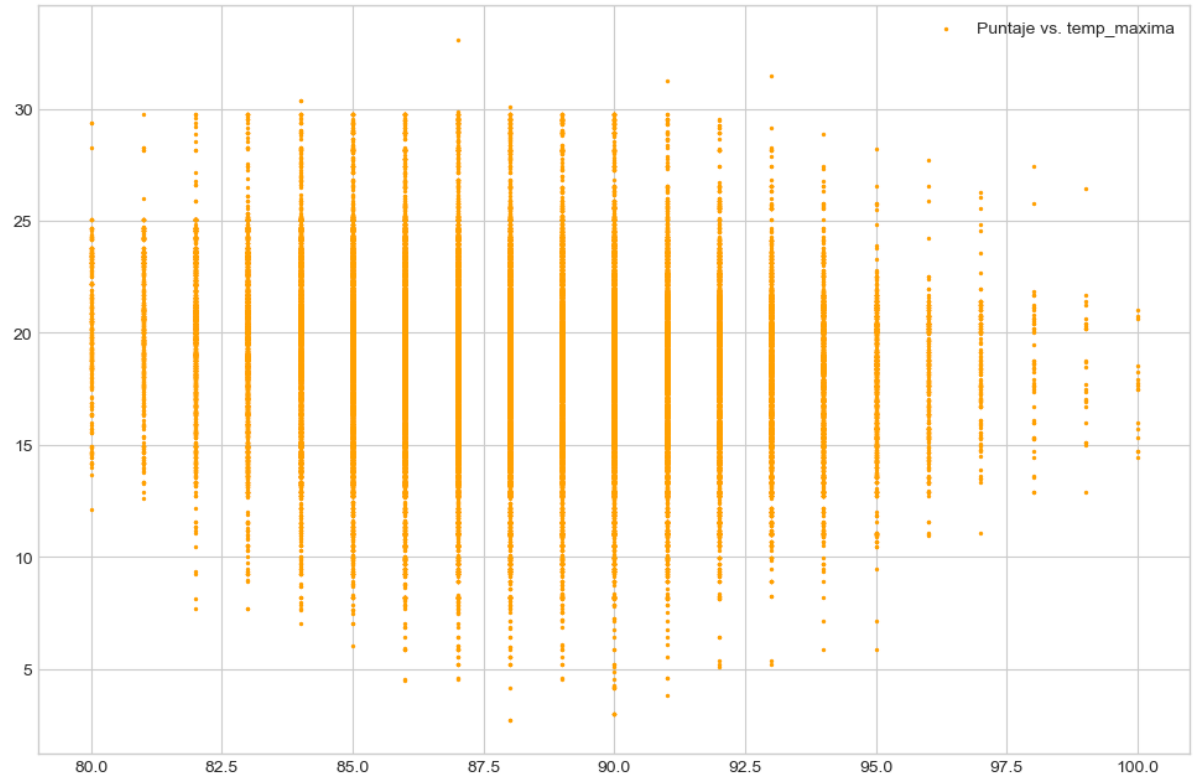
In [86]:
```
#Scatter entre puntaje y año
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['Year'], df['points'], s=2, label="Puntaje vs. Año", color="#FF
A000")
plt.legend();
plt.show()
```



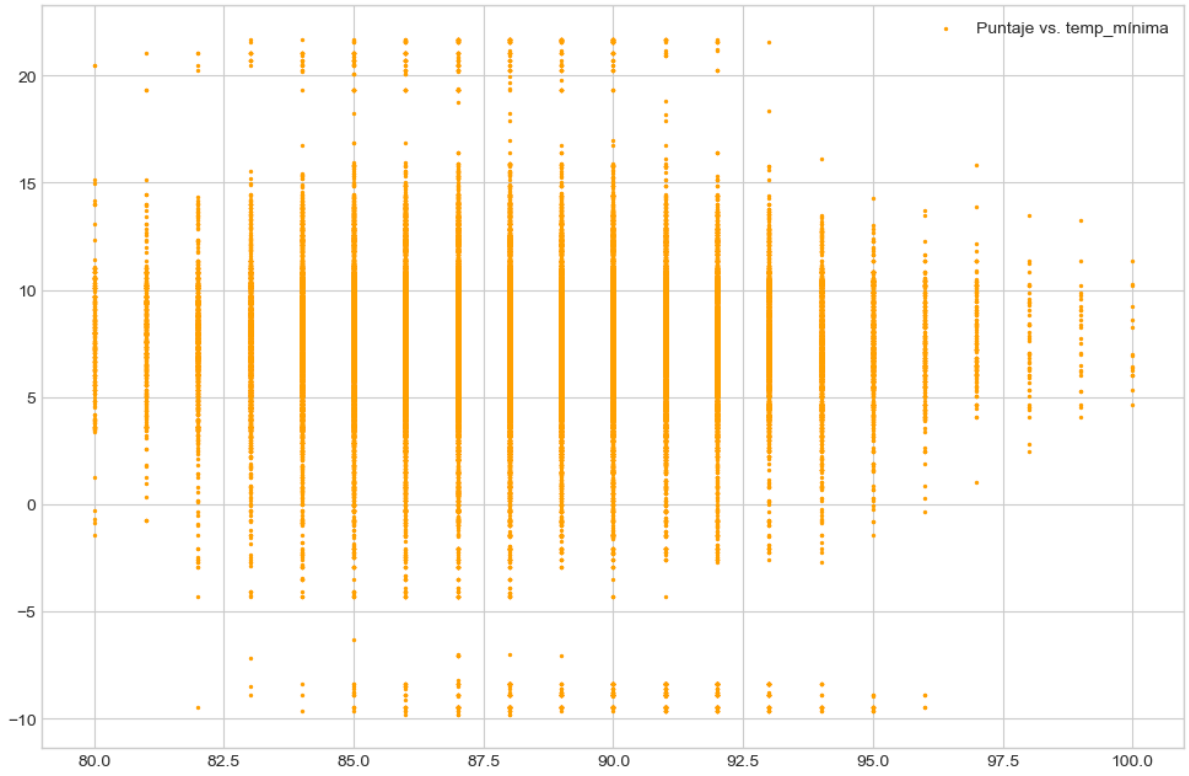
In [87]:
```
#Scatter entre puntaje y temperatura
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_anual'], s=2, label="Puntaje vs. temp_promedio", color="#FFA000")
plt.legend();
plt.show()
```



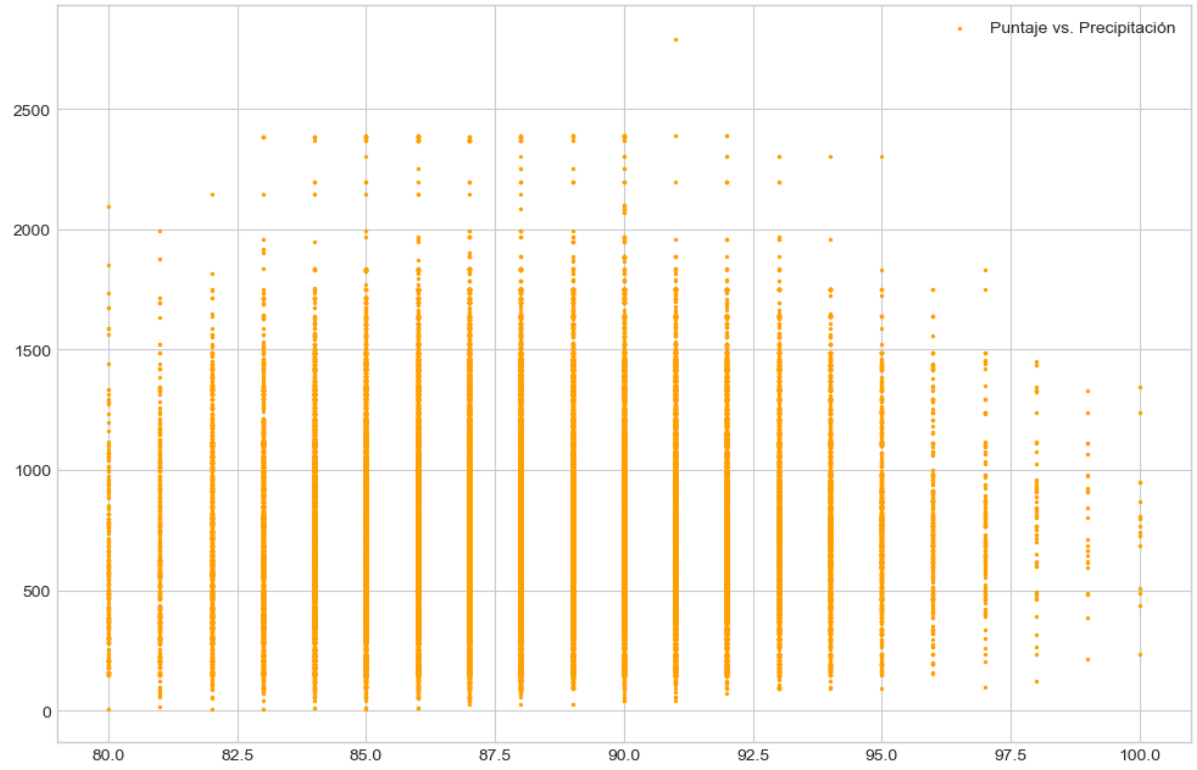
In [88]:
```
#Scatter entre puntaje y temperatura máxima
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_max_anual'], s=2, label="Puntaje vs. temp_maxima", color="#FFA000")
plt.legend();
plt.show()
```



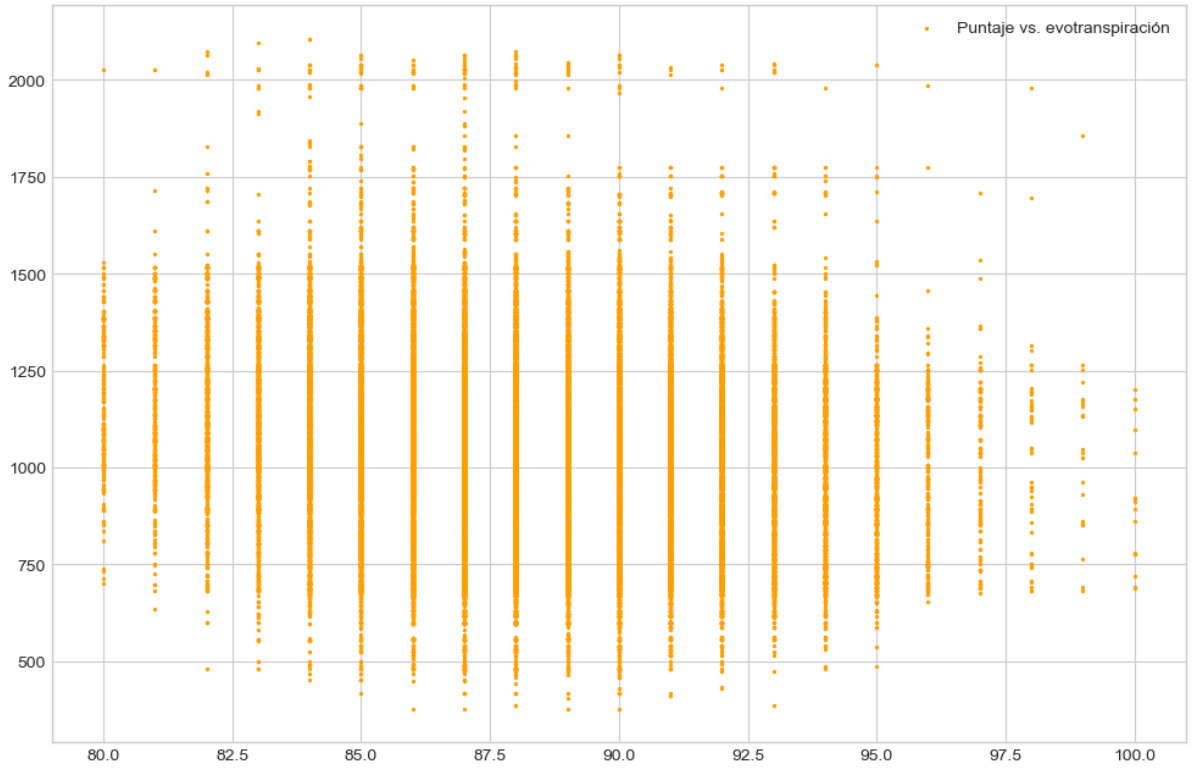
In [89]:
```
#Scatter entre puntaje y temperatura mínima
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['temp_min_anual'], s=2, label="Puntaje vs. temp_mínima", color="#FFA000")
plt.legend();
plt.show()
```



In [90]:
```
#Scatter entre puntaje y precipitación
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['pre_anual'], s=2, label="Puntaje vs. Precipitaci
ón", color="#FFA000")
plt.legend();
plt.show()
```



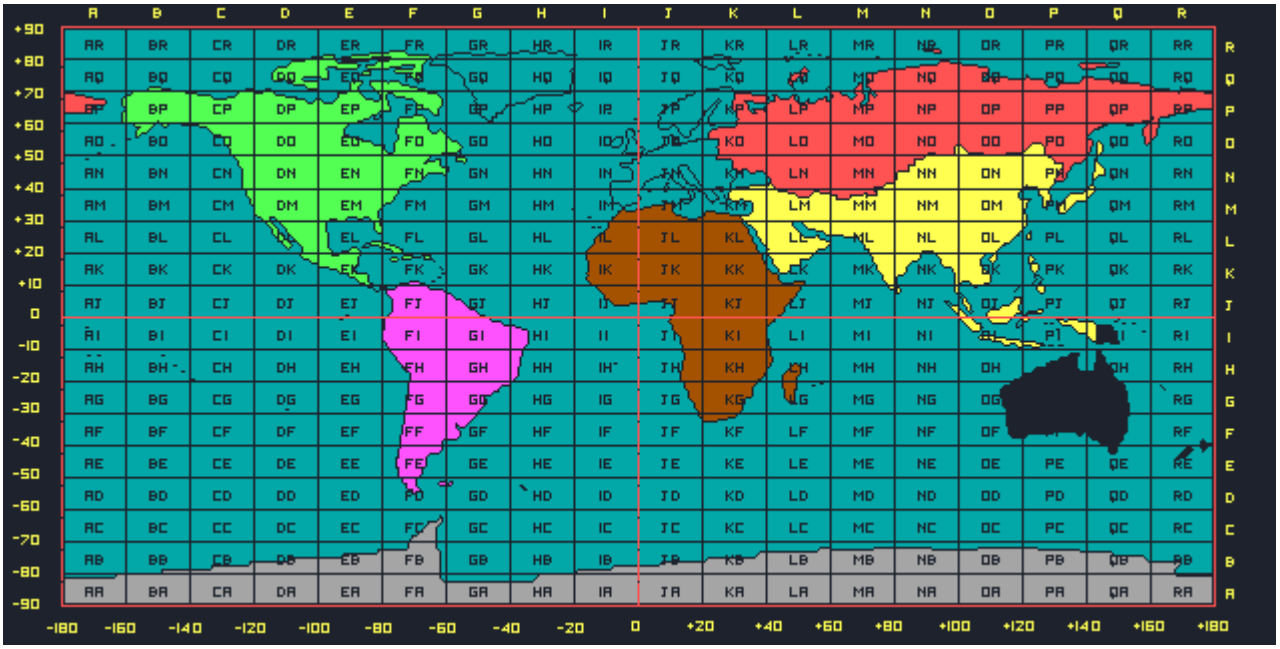
In [91]:
```
#Scatter entre puntaje y evotranspiración
plt.figure(figsize=(12,8), dpi= 100)
plt.scatter(df['points'], df['etp_anual'], s=2, label="Puntaje vs. evotranspir
ación", color="#FFA000")
plt.legend();
plt.show()
```



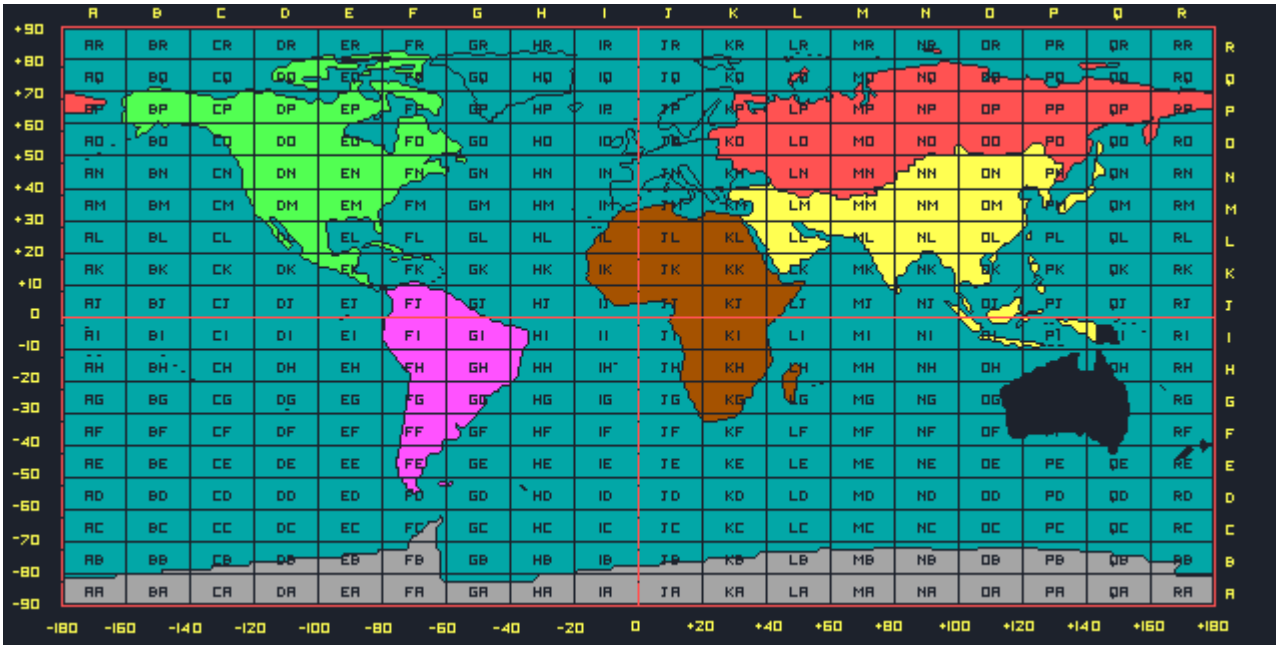
Histrogramas

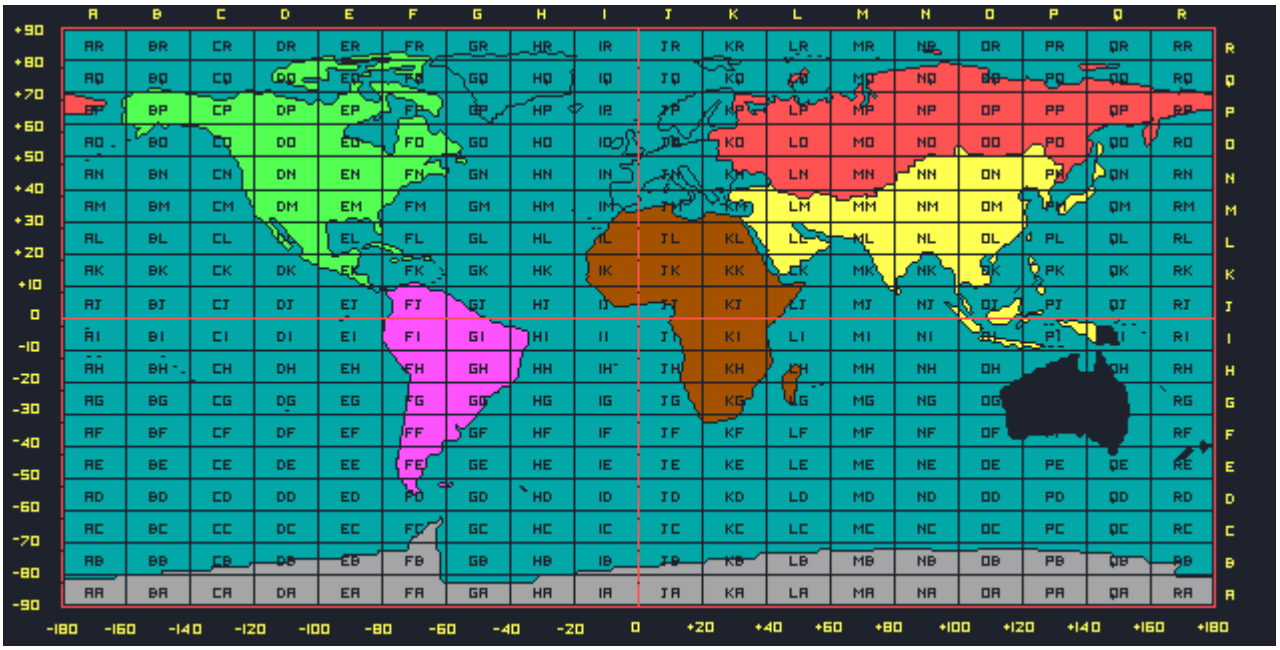
Latitudes Vs. Puntajes

Se utilizará,las calificaciones del 20% de los mejores vinos. Para esto, como los vinos son calificados de 80 a 100, el 20% es el puntaje de 96 a 100. Con esta información, separamos las diferentes zonas de Latitudes con respecto a los trópicos, los cuales determinan las zonas de calor.



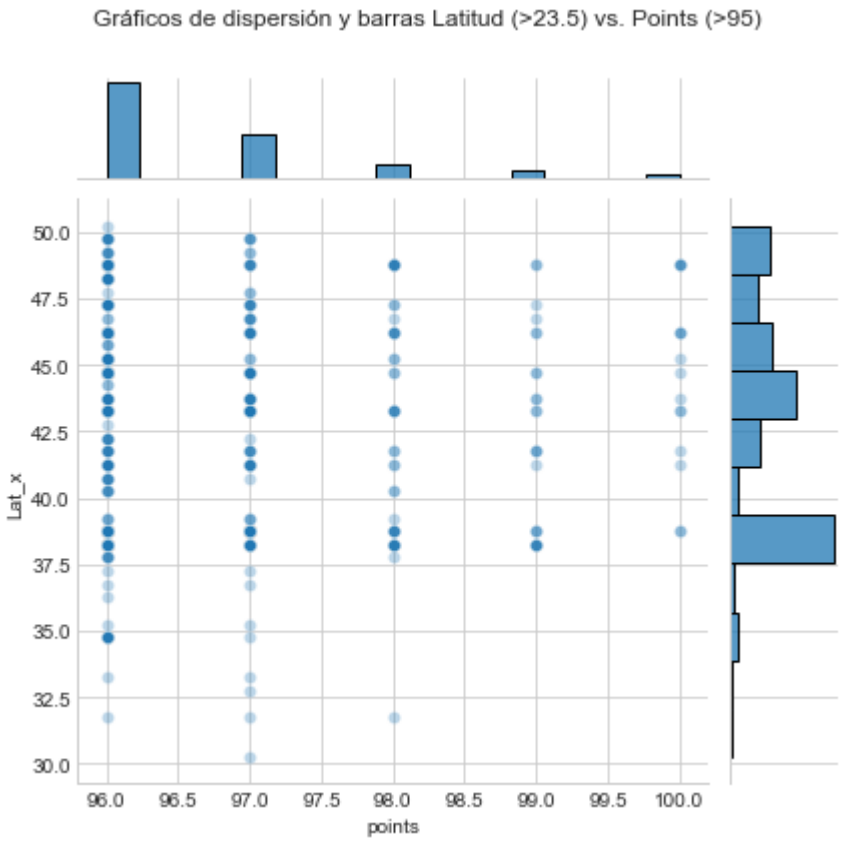
Con esto, realizamos el cruce gráfico de estas zonas, buscando ciertas conclusiones sobre las zonas geográficas contra la calidad de los vinos. No se graficarán datos sobre los círculos árticos y antárticos ya que no se encuentra información en nuestra base de datos de algún vino en estas zonas





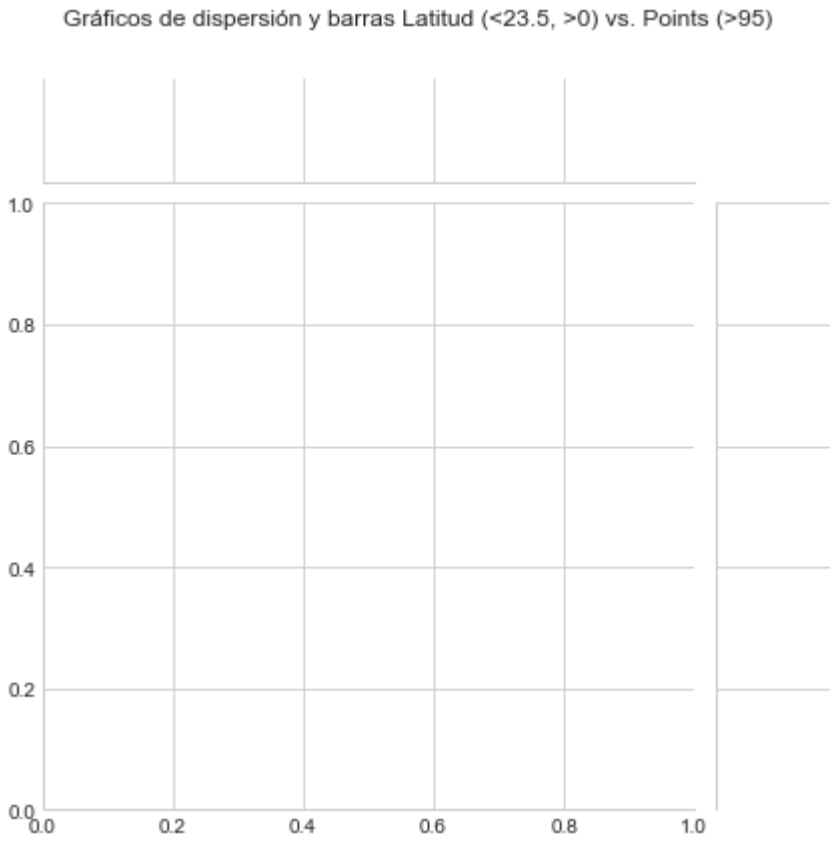
In [92]:
```
#Se carga la información en EDA
eda = df_simple
```

In [93]:
```
#Trópico de Cancer
df = eda[(eda.points>95) & (eda.Lat_x>23.5)]
p = sns.jointplot(x=df.points,y=df.Lat_x, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Latitud (>23.5) vs. Points (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```



Conclusion Se encuentra una mayor concentración de vinos de buena calidad en las latitudes cercanas a los 38°, 44°, 45° y 46°; esta característica tambien se puede percibir ligeramente con los vinos de puntaje 99 y 100, con buenos vinos en los 49°. Esto se da por que la inforamción del dataset no está balanceada.

In [94]:
```
#Ecuador Norte
df = eda[(eda.points>95) & (eda.Lat_x<23.5) & (eda.Lat_x>0)]
p = sns.jointplot(x=df.points,y=df.Lat_x, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Latitud (<23.5, >0) vs. Points
(>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```

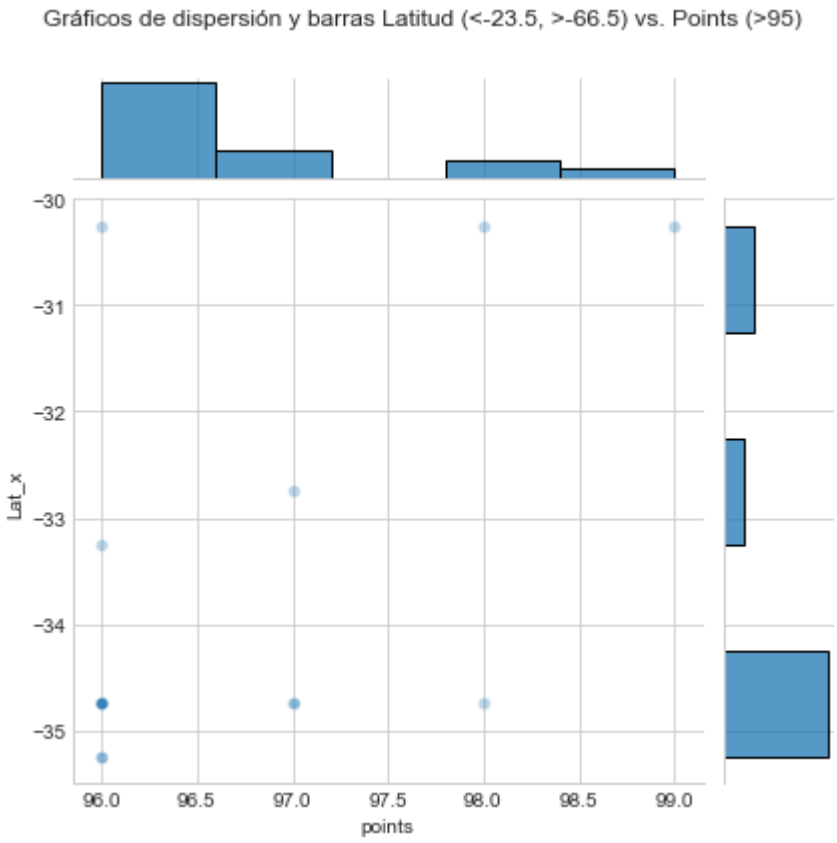


Conclusion No se enceuntra ningún vino en esta zona. El dataset no contiene suficiente información para poder deducir si se pueden tener vinos de buena calidad en esta zona.

In [95]:
```
#Ecuador Sur
df = eda[(eda.points>95) & (eda.Lat_x>-23.5) & (eda.Lat_x<0)]
p = sns.jointplot(x=df.points,y=df.Lat_x, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Latitud (>-23.5, <0) vs. Point
s (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```

Conclusion No se encuentra ningún vino en esta zona. El dataset no contiene suficiente información para poder deducir si se pueden tener vinos de buena calidad en esta zona.

In [96]:
```
#Trópico de Capricornio
df = eda[(eda.points>95) & (eda.Lat_x<-23.5) & (eda.Lat_x>-66.5)]
p = sns.jointplot(x=df.points,y=df.Lat_x, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Latitud (<-23.5, >-66.5) vs. P
oints (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```



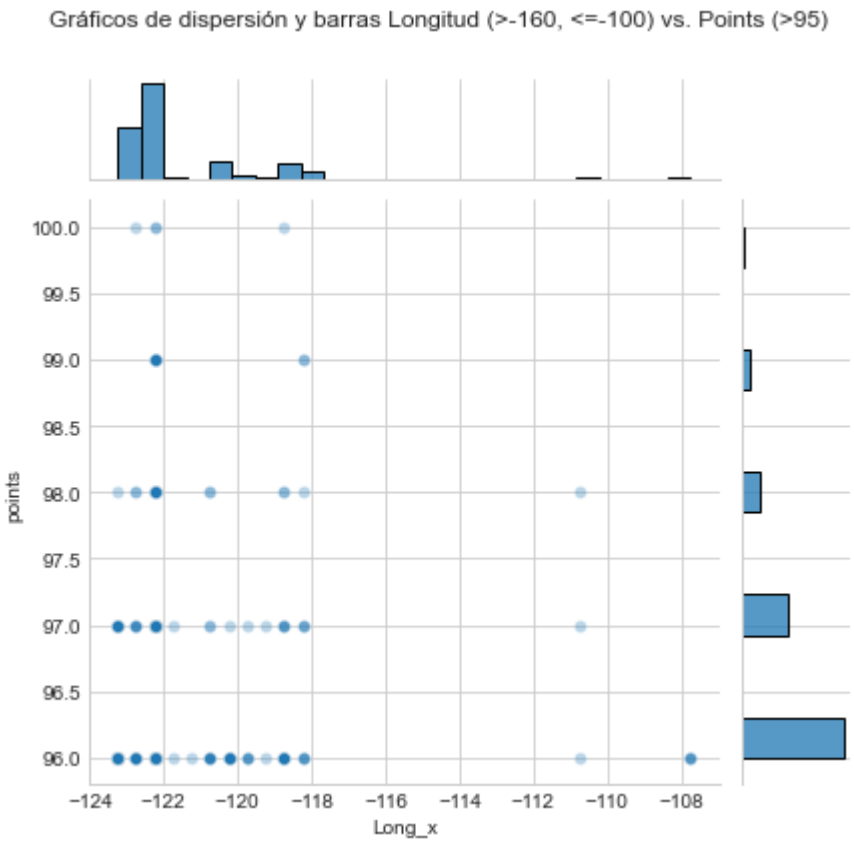
Conclusion En este gráfico se realiza con un muestreo de datos peuqueño de calificaciones en relación con la zona norte, debido a que la cantidad de muestras es mucho mayor en la zona norte a comparación de la zona sur. En este caso, se ve que los mejores vinos se encuentran alrededor de los -35°/-36° hacia el sur.

Longitudes Vs. Puntaje

Se realiza el análisis análogo al de Latitudes, esta vez no se tienen zonas para los meridianos, por lo que se separarán en las siguientes zonas:

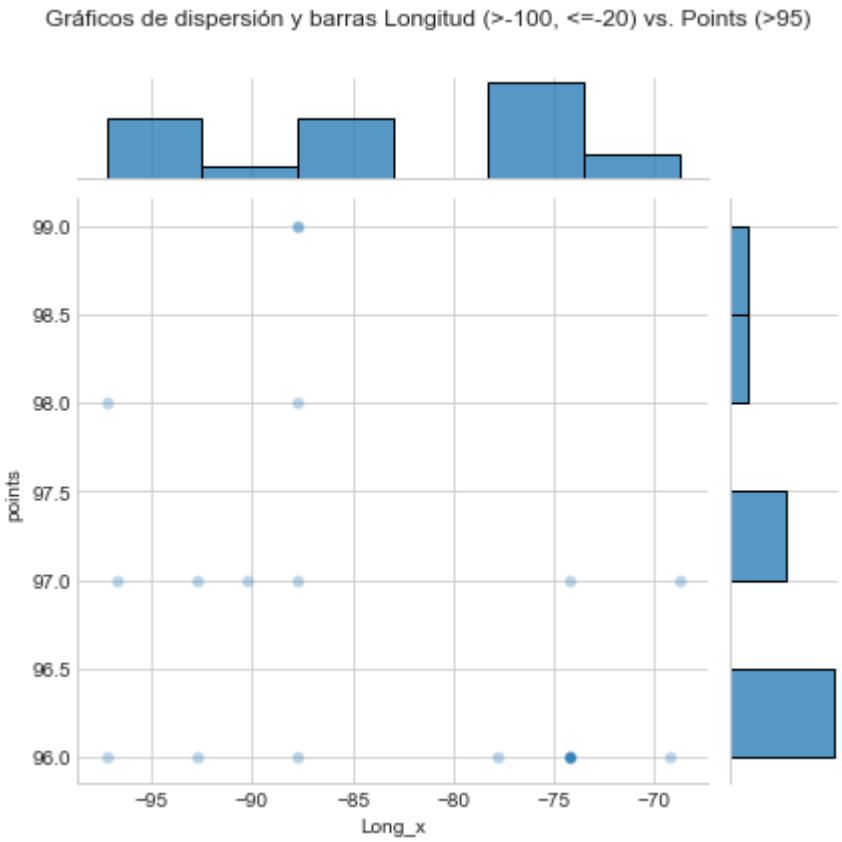
1. Este USA (EU): -160° a -100°
2. Oeste USA y LATAM (OUL): -100° a -20°
3. Este Europa y Africa (EEA): -20° a 30°
4. Oeste Europa y Emiratos Árabes (OEEA): 30° a 70°
5. Asia y Oceanía (AO): 70° a 180°

In [97]:
```
#Zona EU
df = eda[(eda.points>95) & (eda.Long_x>-160) & (eda.Long_x<=-100)]
p = sns.jointplot(x=df.Long_x,y=df.points, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Longitud (>-160, <=-100) vs. P
oints (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```



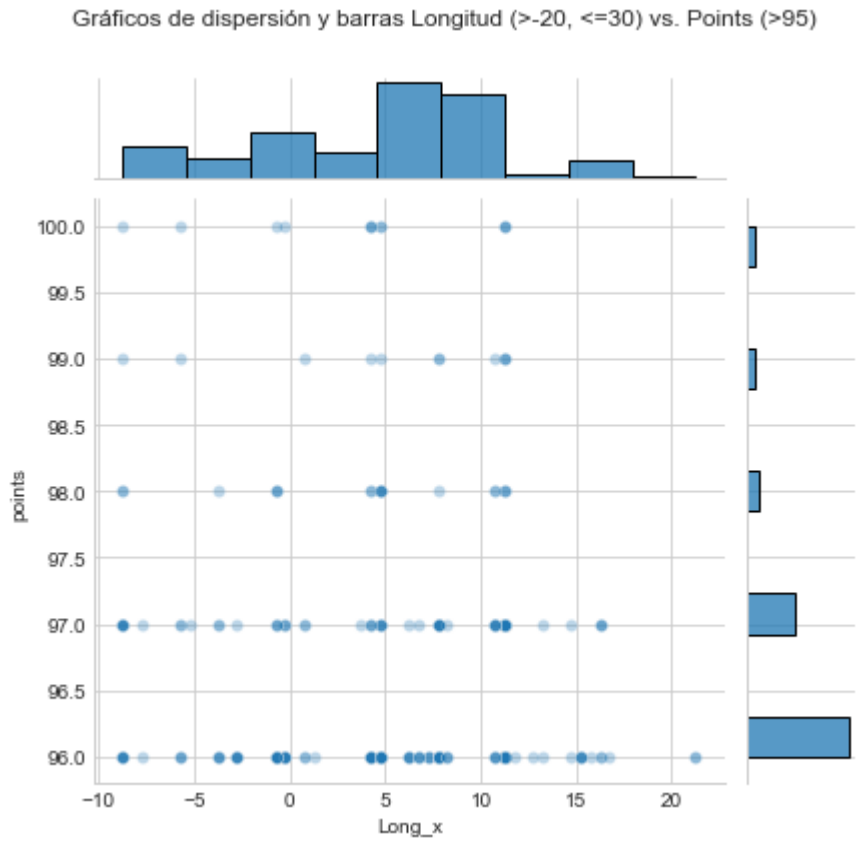
Conclusion Los mejores vinos se dan en las Longitudes alrededor de los -123.5° y -118°, que comprende la región de California. Aunque hay algunos vinos en la zona de -111° a -107°, es interesante analizar zonas con buenas latitudes para zonas de siembra y producción, sobre todo en lugares con acceso a buenas fuentes hídricas u oceanos.

In [98]:
```
#Zona OUL
df = eda[(eda.points>95) & (eda.Long_x>-100) & (eda.Long_x<=-20)]
p = sns.jointplot(x=df.Long_x,y=df.points, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Longitud (>-100, <=-20) vs. Points (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```



Conclusion Los mejores vinos se dan en múltiples longitudes, con concentración entre los -97° y -87°. Esto puede ser una conclusión basada en el agrupamiento de datos y no necesariamente como una conclusión general.

In [99]:
```
#Zona EEA
df = eda[(eda.points>95) & (eda.Long_x>-20) & (eda.Long_x<=30)]
p = sns.jointplot(x=df.Long_x,y=df.points, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Longitud (>-20, <=30) vs. Poin
ts (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```

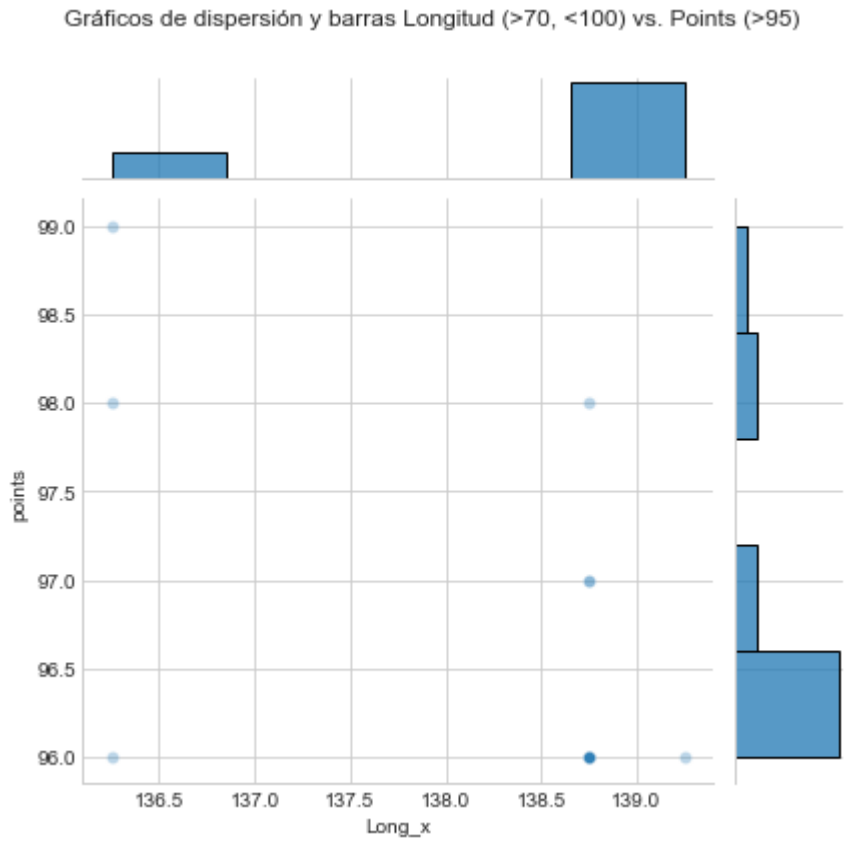


Conclusion Los mejores vinos se encuentran distribuidos, con una mayor concentración entre los -9° y 12°, la cual es la zona comprendida para España, Francia e Italia.

In [100]:
```
#Zona OEEA
df = eda[(eda.points>95) & (eda.Long_x>30) & (eda.Long_x<=70)]
p = sns.jointplot(x=df.Long_x,y=df.points, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Longitud (>30, <=70) vs. Point
s (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```

Conclusion No se encontraron datos de esta zona.

In [101]:
```
#Zona AO
df = eda[(eda.points>95) & (eda.Long_x>70) & (eda.Long_x<180)]
p = sns.jointplot(x=df.Long_x,y=df.points, alpha=0.3);
p.fig.suptitle("Gráficos de dispersión y barras Longitud (>70, <100) vs. Point
s (>95)")

p.fig.subplots_adjust(top=0.90)
plt.show()
```



Conclusion Los datos encontrados hacen referencia a Australia, con la mayoría de vinos entre los 138° y 140°.