# 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

