Instrucciones Miniproyecto 1

1. Introducción

Para este miniproyecto se les pedirá que trabajen con el dataset Red Wine Quality, el cual consiste en una muestra de distintas variantes del vino tinto de origen portugués: Vinho Verde. El dataset es un archivo csv que se les entregará y podrán leerlo mediante la función de pandas `read_csv`.

El dataset tiene las siguientes columnas:

- **fixed acidity ($\mathbb{R}_+$):** la cantidad de la mayoría de los ácidos involucrados en el vino y que no son volátiles (g/dm^3).
- **volatile acidity ($\mathbb{R}_+$):** cantidad de ácido acético en el vino (g/dm^3).
- **citric acid ($\mathbb{R}_+$):** cantidad de ácido cítrico en el vino (g/dm^3).
- **residual sugar ($\mathbb{R}_+$):** la cantidad de azúcar que queda en el vino luego de que la fermentación se detiene (g/dm^3).
- **chlorides ($\mathbb{R}_+$):** la cantidad de sal en el vino (g/dm^3).
- **free sulfur dioxide ($\mathbb{R}_+$):** la cantidad SO2 que no ha reaccionado en el vino (g/dm^3).
- **total sulfur dioxide ($\mathbb{R}_+$):** la cantidad total (que ha reaccionado y que no ha reaccionado) de SO2 en el vino (g/dm^3).
- **density ($\mathbb{R}_+$):** densidad del vino (g/cm^3).
- **pH ([0, 14]):** nivel de pH del vino (asociado a su acidez).
- **sulphates ($\mathbb{R}_+$):** cantidad de sulfato de potasio en el vino (g/cm^3).
- **alcohol ([0, 100]):** porcentaje de alcohol por volumen de vino (% vol.).
- **quality ([0, 10]):** entero que representa la calificación del vino basada en datos sensoriales.

El miniproyecto está pensado para ser desarrollado con los contenidos vistos durante el curso, junto con los módulos `NumPy`, `Pandas`, `thinkstats2` y `thinkplot`, no es necesario ningún otro modulo para poder realizar las actividades que se presentan a continuación.

2. Código inicial

Puede usar las siguientes líneas de código para iniciar el miniproyecto. Estas líneas importan todos los paquetes necesarios, lee los datos del archivo y los deja en el dataframe.

Importante: para que esto funcione el archivo en que se ejecuta el código debe estar en el mismo directorio que el archivo con los datos y los paquetes `thinkplot` y `thinkstats2`.

```
import nsfg
import numpy as np
import pandas as pd
import thinkplot
import thinkstats2
```

```
df = pd.read_csv("winequality-red.csv")
df.head()
```

3. Por desarrollar

- **Parte 1 (2 pts):** Utilice los métodos `isnull` y `sum` para ver cuántos datos faltantes hay en el dataframe. Luego, corrobore que los tipos de datos de cada columna son correctos (todas las columnas deben ser floats, a excepción de "quality" que es un entero). En caso de ser necesario, realice una limpieza correspondiente.

- **Parte 2 (2 pts):** Cree una nueva columna llamada "good", que tome el valor de 1 en caso de que la calidad del vino sea mayor o igual a 7 y que sea cero en caso contrario (Hint: para esto le puede ser útil el método `apply` de `pandas.DataFrame` o de `pandas.Series`).

- **Parte 3 (3 pts):** Responda las siguientes preguntas dejando evidencia del código utilizado para llegar a los resultados: ¿Cuántos vinos tienen una calificación de 10? Y, ¿una calificación de 3? ¿Cuántos vinos son considerados buenos según lo definido en la parte 2?

- **Parte 4 (7 pts):** Grafique un histograma de la variable “good" y otro histograma de la variable “quality”. Comente sobre los valores que toman ambas variables y como estos distribuyen en la muestra.

- **Parte 5 (10 pts):** Calcule el promedio y desviación estándar de cada columna del dataset, exceptuando la columna “good”. Comente por qué hay diferencias tan grandes entre algunas desviaciones estándar.
- **Parte 6 (10 pts):** Grafique el histograma de la columna residual “sugar”. Luego muestre los 5 valores más grandes de dicha columna con sus frecuencias respectivas. Después, basándose en el promedio, desviación estándar e histograma de esta columna comente sobre si los 5 valores más grandes de esa columna deberían ser considerados *outliers* o no. Finalmente, muestre los 5 valores más pequeños de la columna con sus respectivas frecuencias y comente si debiesen ser considerados *outliers*.
- **Parte 7 (10 pts):** Ahora separe los datos entre los vinos buenos y los vinos malos según lo descrito en la parte 2. Luego, en un mismo grafico muestre los histogramas de la variable alcohol en ambos grupos. Después, en un nuevo grafico de barras muestre las funciones de probabilidad (PMFs) de la variable alcohol en ambos grupos. Responda: ¿Qué puede decir sobre la distribución de la variable alcohol en los dos grupos? ¿Qué grafico es preferible utilizar para comparar las distribuciones? ¿Por qué?
- **Parte 8 (6 pts):** Calcule el efecto del tamaño de Cohen de la variable alcohol de los vinos buenos vs los vinos malos. ¿Hace sentido con lo que detectó en la parte 7?
- **Parte 9 (10 pts):** Ahora suponga que le interesa el experimento de elegir 10.000 vinos al azar y ver cuántos vinos de los elegidos son buenos (según lo definido en la parte 2). Suponga también que la cantidad de vinos elegidos que son buenos es una variable aleatoria y su distribución es binomial, donde el parámetro p es la probabilidad de elegir un vino bueno en el dataset. Usando `random.binomial` de `NumPy` genere una muestra de tamaño 100.000 de la variable aleatoria descrita anteriormente y grafique un histograma de la muestra obtenida. Finalmente calcule el promedio de la muestra obtenida y responda: ¿Por qué hace sentido que ese sea el promedio de la muestra?

4. Referencias

Red Wine Quality: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009