Práctica 2: Introducción a pandas

Objetivos generales

- Leer un dataset utilizando pandas, explorarlo y preparar los datos para un posterior análisis.
- Extraer conclusiones basadas en datos, utilizando dos fuentes distintas.

Descripción de la práctica

El objetivo de una tarea de análisis de datos es extraer conclusiones sobre los mismos. Sin embargo, para poder llegar a ese punto es necesario llevar a cabo un exhaustivo proceso de entendimiento, limpieza y preparación.

El objetivo de esta práctica es **manipular y preparar un dataset** que contiene datos de la empresa Airbnb. Estos datos contienen información sobre los alquileres de la empresa en la ciudad de Nueva York. También se usará un fichero con datos poblacionales de la ciudad de Nueva York. Se proporcionan junto a este enunciado dos archivos: airbnb.csv y population.csv.

Se pide entregar **un fichero llamado airbnb_main.py** que contenga las siguientes **funciones**:

- Una función **airbnb_read** que tome como argumento el path al archivo airbnb.csv, lea el contenido y devuelva el mismo en formato DataFrame.
- Una función **airbnb_size** que tome como argumento un DataFrame y devuelva sus dimensiones en formato (n_filas, n_columnas).
- Una función **airbnb_columns** que tome como argumento un DataFrame y devuelva una **lista** con los nombres de las columnas, **en el orden original**.
- Una función **airbnb_countries** que tome como argumento el DataFrame inicial y devuelva un objeto **Series**. El índice deberá ser el nombre de cada país en el dataset y el valor corresponderá al número de alojamientos que hay en ese país.
- Una función **airbnb_boroughs** que tome como argumento el DataFrame inicial y devuelva un objeto **Series**. El índice deberá ser el nombre de cada "neighbourhood group" en el dataset y el valor corresponderá al número de alojamientos que hay en ese "neighbourhood group".
  Pista: presta especial atención a los nombres de los distintos grupos y renombra donde sea necesario.
- Una función **airbnb_price**. Esta función debe tomar como argumento el DataFrame inicial, y devolver **un nuevo DataFrame.** Este DataFrame será idéntico al original salvo porque la columna “price” de este DataFrame deberá tener formato numérico, y todas las filas cuyo ”price” sea NaN deben ser eliminadas.

Pista: Puede ser útil utilizar el modulo "re", lambda functions y el método apply de pandas.

- Una función **airbnb_aggregate**. Esta función debe tomar como argumento un DataFrame y devolver un nuevo DataFrame. Este DataFrame representará el precio medio, mediano, mínimo y máximo por cada uno de los "neighbourhood group".
- Una función **airbnb_totals**. Esta función debe tomar como argumento el DataFrame original y devolver un nuevo DataFrame con el número total de airbnbs, agrupado por "neighbourhood group".
- Una función **population_totals**. Esta función debe tomar como argumento el DataFrame correspondiente al dataset de population.csv. La función deberá devolver un nuevo DataFrame correspondiente a la población total, agrupada por “Borough”
- Una función **airbnb_population. Esta función deberá tomar dos argumentos**, el DataFrame correspondiente al resultado de airbnb_totals, y el DataFrame correspondiente al resultado de population_totals. La función deberá devolver un nuevo DataFrame, con dos columnas: La columna “Borough” y una columna llamada “airbnbs_per_thousands” que represente el número total de airbnbs por cada 1000 habitantes en ese “Borough”.

Ejemplo de funciones

```
def airbnb_read(path_to_file):
    """
    This function reads a pandas DataFrame from a path to a csv file

    Args:
        path_to_file(str): The path to the csv file.
    """

def airbnb_size(df):
    """
    This function takes a pandas DataFrame and returns its dimensions (n_rows, n_cols)

    Args:
        df(pd.DataFrame): A Pandas Datarame
    """

def airbnb_population(df_airbnb, df_population):
    """
    This function takes two DataFrames, and returns a new                    DataFrame
containing number of airbnbs per 1000 inhabitants.

    Args:
        df_airbnb(pd.DataFrame): The DataFrame produced by airbnb_totals.
        df_population(pd.DataFrame): The  DataFrame produced by population_totals.
    """
```

Evaluación:

- 10% por cada función bien implementada

- La calidad del código (fácil lectura, documentación, estilo…) es importante y afectará a la nota final.