**Squad Project: Fines and Wines**

Línea horizontal

1. Introduction

Welcome to the final project of **Foundations II**. With this project you will enhance your data processing and transformation skills in Python (you have the flexibility to use another language if you prefer). Dive into two distinct Exploratory Data Analysis (EDA):

1. **Madrid Traffic Fines Analysis:** Embark on a journey through the open data universe, focusing on the allocation of traffic fines to drivers in Madrid. This exercise challenges you to navigate an open data portal to explore real-life infractions.
2. **Wine Popularity Study:** Savor the intricacies of the wine world as you analyze datasets to unravel the factors contributing to wine popularity.

In these exercises, you'll be handling multiple datasets, each bringing its own challenges and insights. There are **two optional tasks available**, one for each EDA. The Squad must **choose one** of these **to deepen the analysis** of one of the two datasets.

Let's delve deep into the world of data analysis with Python.

2. Main objectives

This project aims to solidify your foundation in data science by focusing on key areas:

* **Data Manipulation:** Load and manipulate different types of files using Pandas.
* **Data Structures:** Enhance your understanding and handling of Python’s essential data structures – lists, tuples, dictionaries.
* **Numerical Operations with Numpy:** Sharpen your skills in dealing with numeric vectors and matrices using the Numpy library.
* **Data Transformation:** Transform raw data into insightful information with Pandas.
* **Visualization:** Bring data to life through compelling visual narratives using matplotlib or seaborn.

3. General analysis

**1 - MADRID OPEN DATA**

This EDA challenges you to discover patterns and insights within a dataset measuring the dynamics of real urban life. Madrid's Open Data Portal:

[Portal de datos abiertos del Ayuntamiento de Madrid](https://datos.madrid.es/portal/site/egob)

To get you started, the materials provided include a .zip file containing data on traffic fines (“Multas de circulación: detalle“). This file consists of 12 .csv files detailing infractions committed by drivers in Madrid during 2020.[[1]](#footnote-1)

These files have been sourced from the aforementioned link. In fact, more data can be found in the “Catálogo de datos” section. Within the “Conjunto de datos”, you will discover complete catalogue of data from Madrid City Council. You can filter datasets to easily locate traffic fine data from the “Dirección General de Gestión y Vigilancia de la Circulación”. Apart from the provided files for 2020, there is a detailed description of the data. The available data has been published since 2014.

The data available for download and consultation are divided into:

* Year and month
  + **Detail**: Download file in .csv format
  + **Grouped-excluded**: Download file in .csv format

For our case study, we will only download the .csv files from the "Detail" category. **We will analyze data from both the year 2020 and the most recent full year available.**

**Feature Information:**

In the dataset, you will encounter a section titled “Documentación asociada,” which is essential for our analysis. Specifically, the “Estructura y formato de la información. Multas” file will be of great interest. To summarize, the data frame comprises 14 variables, each with a specific meaning:

* **CALIFICACIÓN:** Severity of the reported infraction.
* **LUGAR:** Location of the infraction.
* **MES +** **ANIO +** **HORA:** Date and time of the sanction.
* **IMP\_BOL:** Fine amount in euros, based on **CALIFICACIÓN**.
* **DESCUENTO:** Reduction in the infraction price (50% if paid within the period specified in the Road Safety Law), applicable for qualifying infractions.
* **PUNTOS:** Points deducted from the offender's driving license.
* **DENUNCIANTE:** Agent from the public body responsible for traffic regulation in the Community of Madrid who issues the complaint against the offender.
* **HECHO\_BOL:** Reason or cause for the driver’s sanction.
* **VEL\_LIMITE:** Maximum allowed speed where the offense occurred.
* **VEL\_CIRCULA:** Speed of the driver at the time of the offense.
* **COORDENADA-X:** Geographical X-coordinate identifying the offender in offenses related to traffic light violation, restricted lane disobedience, or unauthorized access to Madrid Central.
* **COORDENADA-Y:** Geographical Y-coordinate for the same contexts as **COORDENADA-X**.

**Objectives:**

Given the previously described datasets and their variables, the following tasks are requested:

1. Merge the twelve dataframes, spanning from January to December for both specified years, into a single dataframe. Then, display the total dimensions (shape) of the resulting combined dataframe.
2. As we will not be working with geographic data, delete the **COORDENADA-X** and **COORDENADA-Y** columns.
3. Examine the unique values of the variables. Some may have unique values, making them constants. Evaluate their relevance in the dataframe.
4. The **VEL\_LIMITE** and **VEL\_CIRCULA** columns appear to have empty values. This occurs when the infraction is not related to a speed limit. Convert all anomalies to null values. *Hint: Explore the use of regex (regular expressions). Investigate the use of a pattern like r'^\s\*$' within the replace function to handle specific data anomalies.*
5. The columns **VEL\_LIMITE** and **VEL\_CIRCULA**, though numerical, are recognized as "object" type. After checking the frequency of values, you'll find they are counted as text strings. Therefore:
   1. Convert nulls in these two variables to 0.
   2. Change the variable type to numeric.
   3. Identify the most common speed limit, excluding zero.
6. Create a new column called **DIFFERENCE\_KMH**, calculated by subtracting the speed limit from the driver's speed at the time of the infraction. Use this to identify the top 10 drivers who exceeded the speed limits by the highest margins.
7. Filter the dataframe for complaints that resulted in point deductions (other than zero points). Group this data by the public agent issuing the complaint. Which agent has the highest average points deduction?
8. Remove the decimal part of the hours column and graphically represent the number of infractions for each hour. Identify the peak hours with the most infractions.
9. Graphically display the fines issued during the months of the most recent full year available. Can any analysis be derived from this data? Compare it with the 2020 analysis to potentially predict the impact of Spain's COVID-19 confinement on traffic infractions.
10. Display the number of infractions (without accumulated frequency) by each public agent, categorized by the infraction classification.
11. Imagine landing a Junior Data Scientist role at a firm specializing in the comparative analysis of fines. Your boss, setting high standards, assigns you to conduct a comparative study between data from the most recent full year and the available 2020 records. You have the discretion to include additional data from other years if it enhances the analysis. Use relevant graphics to augment your analysis and weave an engaging narrative!
12. **Option 1:** Imagine landing a Junior Data Scientist position at a firm specializing in comparative analysis of fines. Your boss, renowned for high expectations, assigns you the task of conducting a comparative study between the data from the most recent full year and the provided 2020 records. You have the liberty to include data from other years if it significantly enhances your analysis. Utilize relevant graphics to enrich your data exploration and create a compelling narrative!

**Exercise 2 - Wines**

Dive into the intriguing world of wines as we extract fascinating insights from data on a variety of global wines. This information, sourced from the website [Wine Enthusiast](https://www.winemag.com/?s=&drink_type=wine), encompasses details about the wine's origin, score, price, tasters' notes, and other general information about each wine.

For more comprehensive insights into these datasets, refer to the original Kaggle data available at <https://www.kaggle.com/zynicide/wine-reviews>. In this exercise, our focus will be on the following specific data sources:

* **winemag-data\_first150k.csv**
* **winemag-data-130k-v2.json**

In both datasets, we find very similar information, whether it's the columns in the .csv file or the key fields in the .json file. The data includes the following features:

* **country:** Origin country of the wine.
* **description:** Taster's notes on wine properties.
* **designation:** Wine category.
* **points:** Score from winemag users.
* **price:** Wine price.
* **province:** Province in the wine's country of origin.
* **region\_1:** Primary wine region.
* **region\_2:** Secondary wine region.
* **variety:** Grape type.
* **winery:** Winery responsible for production and maturation.

With the provided information, complete the following tasks:

1. Load both files: the .csv as a pandas dataframe and the .json as json.
2. Get a single dataframe that is the union of both files that only has the following columns:
   * **country**
   * **designation**
   * **points**
   * **price**
   * **state/province**
   * **winery**
3. Check for and remove any duplicate rows in the dataframe.
4. Analyze null values: Assign zero to nulls in the price column, then remove remaining null values. What is the final row count?
5. Extract a subset of wines from Spain with 'reserva' in their designation. *Hint: To search for a specific text string within a column, utilize the* ***str.contains*** *method, which is highly effective for text-based queries in pandas:* <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.contains.html>
6. Compare the average score of Spanish 'reserva' wines with the overall average score of all wines. Determine whether the Spanish 'reserva' wines have a higher or lower average score compared to the global average.
7. Re-examine the subset of Spanish 'reserva' wines. Considering the fame of Spain's Rioja wines (originating from the Rioja province in northern Spain; labeled as North Spain in the dataset), it is presumed that this region predominates in the subset. Employ graphical analysis to validate the accuracy of this assumption.
8. Display the top 10 highest-rated wines.
9. Calculate the average wine price per country. Identify the country with the highest average price. Does this country appear in the previously calculated top 10 highest-rated wines?
10. Graphically represent the price versus score for the Spanish 'reserva' wines by province. Which province has the most expensive wine? Does North Spain have the highest-rated 'reserva' wine?
11. **Option 2:** Imagine landing a Junior Data Scientist role at a firm specializing in the comparative analysis of wines. Your boss, renowned for high standards, tasks you with conducting a comparative study of wines globally. You have the freedom to incorporate additional data if it significantly enhances your analysis. Employ relevant graphics to augment your data exploration and craft a compelling narrative!

4. Requirements

It is important to document your own work on every project you do. This way, after a while, you can easily remember what your own code does, and what you did in each project.

* Please submit each of the two EDAs as a **separate repository**.
* You will be required to use GitHub. Two team members will **create the repositories** for each EDA.
* Other team members must **clone** these repositories and work on them from their local machines.
* **Contribute** to these repositories **with commits** as the project progresses.
* We strongly recommend that all team members actively contribute with commits. Remember that GitHub tracks each individual's contributions, and all commits made in repositories are recorded in GitHub's statistics.
* We highly encourage you to create a **roadmap** and develop both projects **using agile methodologies**. Remember, working in a group isn't about doing the same task simultaneously, but rather organizing yourselves to work optimally. Do you have a clear understanding of your most effective way of working? For this purpose, you can use either **ClickUp**, as you have learned in the previous module, **or** you can set up a project with a board on **GitHub!** Here's a **hint**: [Board · Roadmap (github.com)](https://github.com/orgs/remix-run/projects/5/views/1) → Remember to **make it** **Public!**
* It is essential that you add comments on the most important elements during the development of the project. **Markdown** cells, **comments** in code cells, **docstrings**, type hints **and** a complete **Readme.md**. If a programmer opens your code, he/she must understand how the program flows by reading your comments.
* You can do the exercises in whatever software that supports code, like VS code, jupyter notebook, jupyter lab, Databricks, google Colab, etc.
* All code, including comments, must be written in English.
* You can use the main Python libraries for Data Science as numpy, pandas, seaborn and matplotlib, anyway you can use related or similar libraries as your convenience.
* All files that were used in the creation of the project as well as its solution must be attached.
* Remove files that are not used or are not necessary to evaluate the project.

5. Resources

The following webpages contain useful resources for the development of both projects. You can find ideas for your graphs or to complement your analysis:

* Seaborn documentation: <http://seaborn.pydata.org/>
* Seaborn Tutorial: <https://machinelearningmastery.com/seaborn-data-visualization-for-machine-learning/>
* Seaborn examples: <https://www.python-graph-gallery.com/seaborn/>
* Pandas tutorial: <https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html>
* Numpy tutorial: <https://numpy.org/doc/stable/user/quickstart.html>
* Matplotlib tutorial: <https://matplotlib.org/stable/tutorials/index.html>
* Matplotlib documentation: <https://matplotlib.org/>
* Numpy documentation: <https://numpy.org/doc/>
* Pandas documentation: <https://pandas.pydata.org/docs/>

6. Presentation

To evaluate the project you will need to:

* Submit the links to **both GitHub repositories**, which should include all documents and additional files, for review by the tribunal prior to your presentation. Adhere to the deadlines set **in Google Classroom**; all materials must be submitted **a day before the defense of your work**. The tribunal will review the last commit registered on GitHub before the deadline. While you may update your project with future versions, only the last version submitted before the deadline will be considered for evaluation.
* Present a **20/25-minute talk** with your Squad, highlighting the project's key aspects. The presentation time must be equally distributed among all Squad members. Choose to showcase the notebook directly or use slides with essential code snippets. **Each Squad must select 1 of the optional exercises (avoid choosing the same one) and weave an engaging narrative.**

1. Due to data protection reasons, these files only report the infraction type. Personal re-identification of the offenders’ information is strictly prohibited. [↑](#footnote-ref-1)