Code.Hub

Project 2:
Data exploration and preprocessing

This assignment aims to get you acquainted with Exploratory Data Analysis (EDA), an essential part of Data Science. You will attempt to explore a given dataset and extract information about its structure, quality and main characteristics.

Deliverable

The solution should be submitted in a jupyter notebook. This notebook should:

- be reproducible
- be well formatted and documented
- clearly indicate which question you are answering at each point

Dataset

The dataset you have to investigate contains AirBnB listings in Athens, Greece. It comprises of two parts:

- **listings_cleaned.csv**: information on listings and hosts
- **calendar.csv**: bookings calendar for the next few months

You have to explore both files and answer the following questions.

Questions

1. How many samples and features does each file have?
2. What are the types of your features?
3. Are there any missing values? If yes, how many and how many rows are affected?

4. How many listings per neighborhood are there?
5. How many listings per room type are there?
6. How many listings per room number are there?
7. What is the distribution of listings per host? What are the most listings that a single host has?
8. When was the first host registered?
9. What year had the most hosts registered?
10. What is the range of the calendar, i.e. when does it start and when does it end?
11. What is the distribution of bookings per month? Can you identify the time periods with the biggest percentage of bookings over listings?
12. Which time periods are the prices higher?
13. How many identified hosts are there? What is their percentage over all hosts?
14. What are the top-10 most common amenities provided by the hosts?
15. Can you identify the top-10 rated listings? Are they by the same host?
16. Can you identify the top-5 rated locations/neighborhoods?
17. Can you identify the time periods when most reviews are submitted?
18. What is the distribution of score ratings? Are there lots of reviews scoring < 50?
19. What is the distribution of price for each room type?
20. Can you identify which days of the week have the highest mean prices?
21. The feature “amenities” in **listings_cleaned.csv** does not seem to be in an appropriate form for analysis. Can you find a way to represent it in numerical form?
22. What is the scale of your numeric features? Are they similar?

Further analysis

Perform an Exploratory Data Analysis on the listings dataset. Treat 'price' as the target variable. These steps aren’t necessary, they might help you get a better understanding of your data. Some ideas you might want to explore:

- Check the distribution of all continuous features. Are there any outliers?

- Check the distribution of all discrete features. Are they evenly distributed?
- Check the distribution of the target variable. What can you tell from this?
- Can you find any features that are related to the target variable?
- Are there any wrong values in the dataset?
- Can you make any assumptions on why the values are missing in the dataset?

Preprocessing

In this step you must bring the dataset in a format understandable by most machine learning algorithms. For this you must:

- Handle any missing values in the dataset.
- Encode all categorical features.
- Scale all features.

General Notes

- Some questions might require studying the material provided to you, which goes in some regards beyond what was covered in class. Don't forget that you can always search for new resources on your own.
- The library documentations are your friends! Remember that you can easily access them through the Help tab of your jupyter notebook. Or just Google it!
- If you cannot answer one question, move on to the next.

Good Luck