Technical challenge

Data Scientist

The challenge

You are given a dataset (**assets/dataset.zip**) containing information about restaurants all over Europe. The goal of the challenge is to make sense of the dataset and understand it in such a way that informed, data-based business decisions can be made. To make things easier, the challenge is split into three conceptually independent parts, with action points at each part.

Data cleaning

1. Identify the columns with mixed data types.
2. For each column, count the number of rows per data type.
3. Would removing missing values solve the mixed data type problem?

Data understanding

1. Are the review columns correlated with the rating columns?
- Review columns: `["excellent", "very_good", "average", "poor", "terrible", "total_reviews_count", "reviews_count_in_default_language"]`
- Rating columns: `["food", "service", "value", "atmosphere", "avg_rating"]`
2. Are vegetarian-friendly restaurants *better* than non-vegetarian ones?
3. Are there any significantly more expensive cuisines?

Business-specific

1. In the **assets** directory, you will see a very small dataset called **europe_capitals_population_and_area.csv**. A gluten-free restaurant wants to open a new restaurant in a European capital where gluten-free restaurants are underrepresented. Assuming there are no other factors, except population and gluten-free restaurant density, what would be the top 5 capitals to open that restaurant?
2. Think and propose a couple of other ways this dataset could be used to help businesses.

Bonus

1. In the **assets** directory, you will see a file called **paris_bounding_polygon.json**. This contains a list of latitude and longitude coordinates that define a polygon that is considered to represent the Paris city area. For simplicity, we assume the population distribution is uniform in the Paris city area. An Italian restaurant wants to open a restaurant in Paris in a zone where there are the fewest Italian restaurants. What is the

best location to open the restaurant (the answer can be a single point or a bounding box/polygonal region depending on the implementation)?

Delivering the solution

- Create a GitHub repository for the solution;
- For each action point, create a separate commit (or multiple commits if the action point requires);
- Keep in mind that notebook cells should contain relevant output;
- After finishing the challenge, make sure to add us as collaborators to the repository.
    - https://github.com/marianstefi20

Timeline

- The challenge should be completed between 4-6 hours.
- It is a bonus if you can finish it in a single day.
- If you can not finish in a single day, you can split the work into two days, but please have continuity (if you do not finish the first day, please continue the second morning).

Things to consider

- It's great if you manage to cover all the action points in the given time period, but we'll also pay attention to the following:
    - how was the data analyzed;
    - how were hypotheses defined;
    - what techniques were used in order to validate/invalidate hypotheses;
    - how were the corresponding conclusions/findings presented (they should be clear, interpretable, and relevant)
- Code quality, correctness, performance

Good luck! 👩‍💻