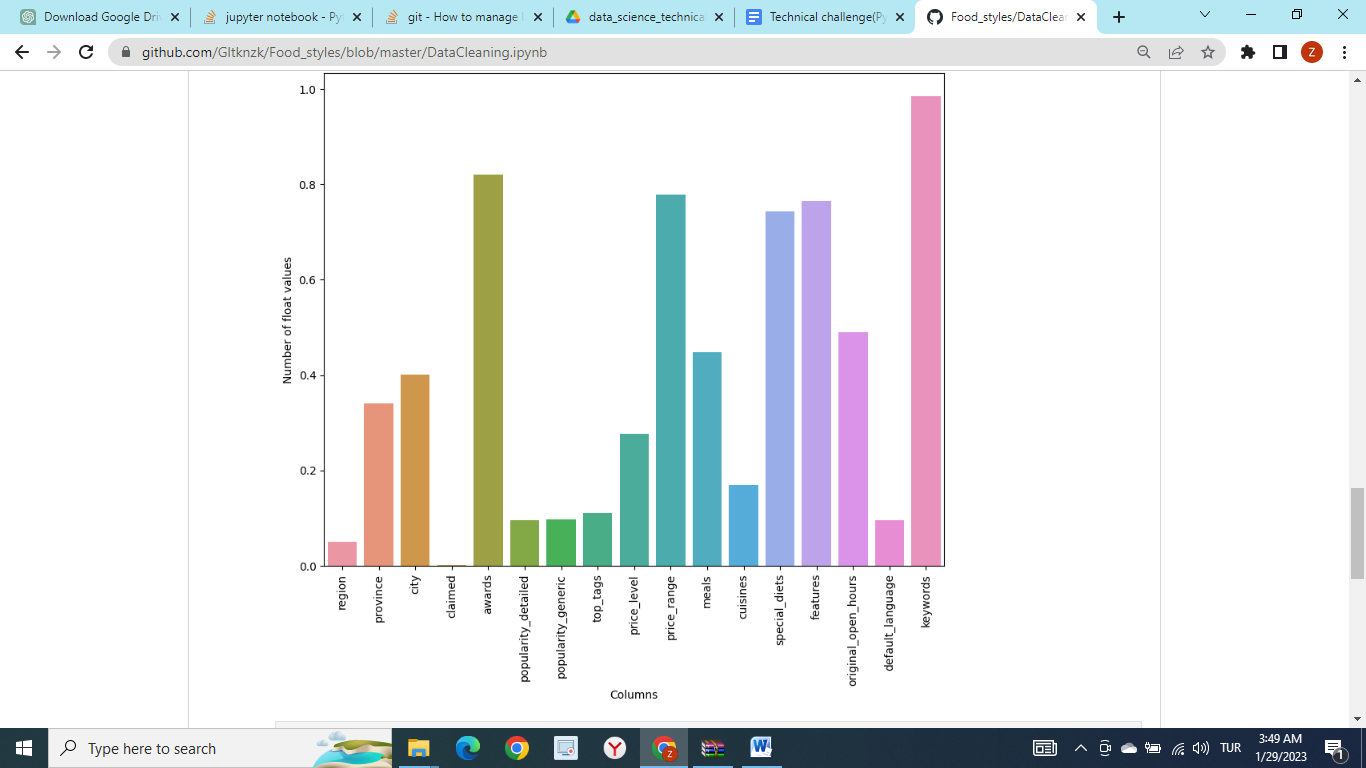
**Technical challenge**

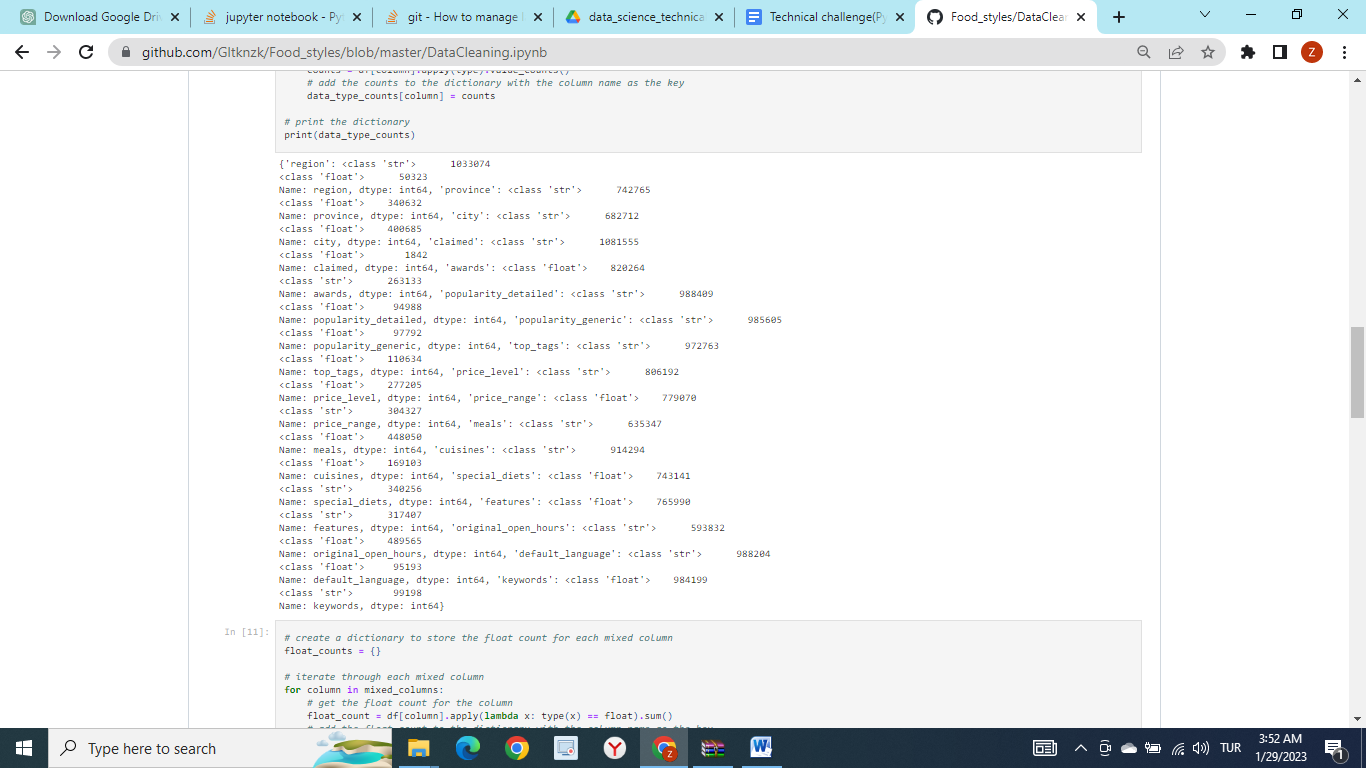
**Data Cleaning**

1. Identify the columns with mixed data types.

As seen on the the graph, 17 columns have mixed datatypesb (str, int and float). Float data type represents NaN values.



1. For each column, count the number of rows per data type.



1. Would removing missing values solve the mixed data type problem?

The best practice for handling missing values in a dataset depends on the specific situation and the goals of the analysis. Generally, the following strategies are commonly used:

Remove rows or columns with missing values: This approach is appropriate if the proportion of missing values is small and if the missing values are not informative. However, this approach can lead to a loss of important information and could bias the analysis.

Impute missing values: This approach involves replacing missing values with estimates based on the other values in the dataset. Common methods include mean imputation, median imputation, and multiple imputation. This approach can be useful if the proportion of missing values is moderate, but it can be problematic if the data is missing not at random.

Keep the missing values: This approach is appropriate if the data is missing at random and if the missing values are informative. For example, if a variable is missing because it was not applicable or not asked, then keeping the missing values can be useful.

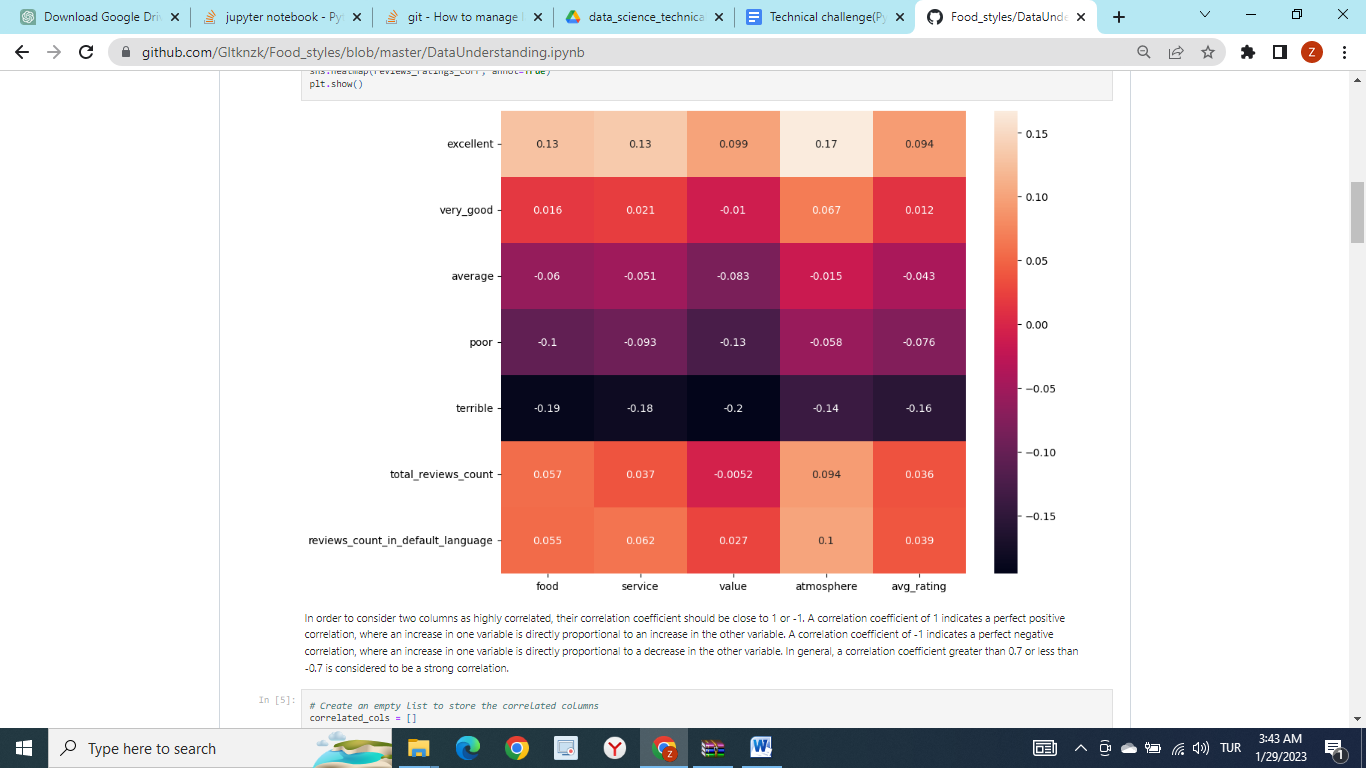
Ultimately, the best approach will depend on the specific dataset and the goals of the analysis. It is always a good idea to explore the missingness patterns and the distribution of the data before making a decision on how to handle missing values.

**Data understanding**

1. Are the review columns correlated with the rating columns?

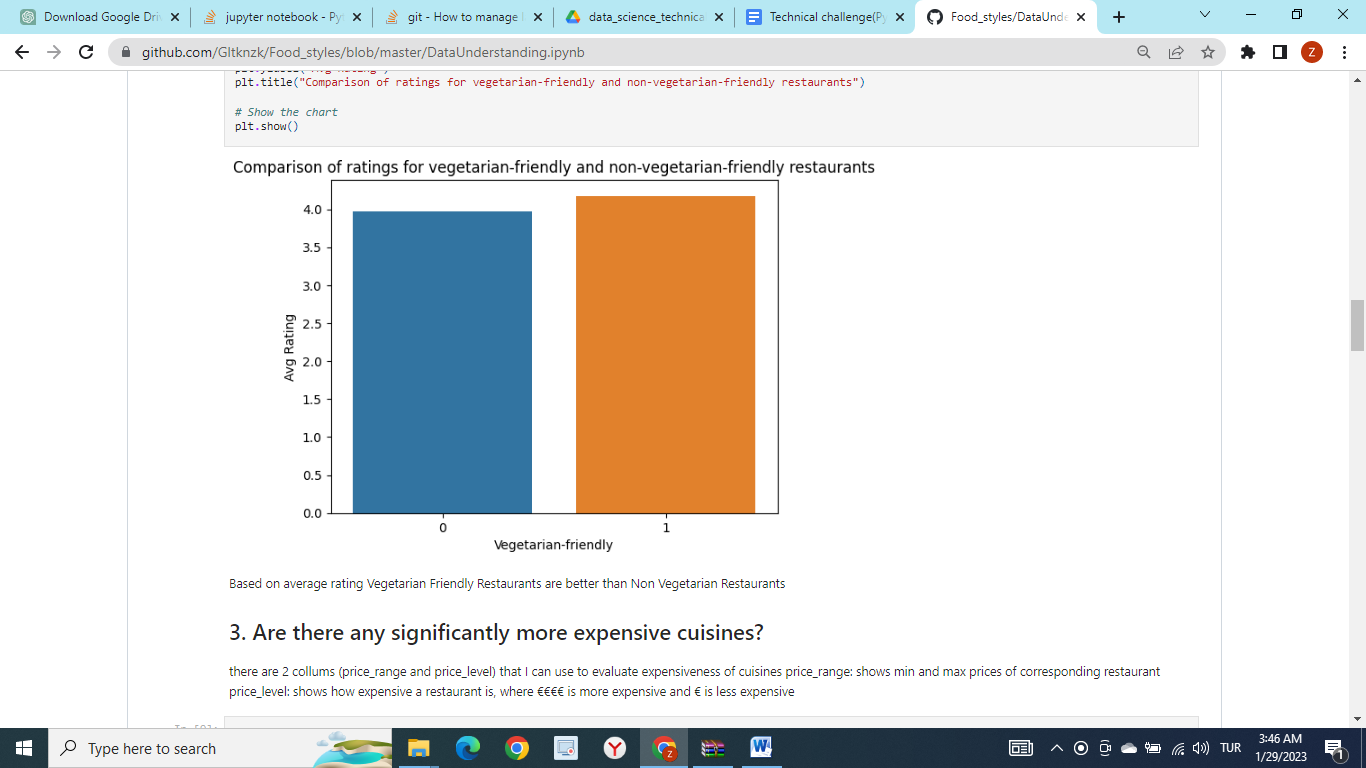
* Review columns: ["excellent", "very\_good", "average", "poor", "terrible", "total\_reviews\_count", "reviews\_count\_in\_default\_language"]
* Rating columns: ["food", "service", "value", "atmosphere", "avg\_rating"]

There is no strong correlation; correlation between any columns is not higher than 70 percent.



In order to consider two columns as highly correlated, their correlation coefficient should be close to 1 or -1. A correlation coefficient of 1 indicates a perfect positive correlation, where an increase in one variable is directly proportional to an increase in the other variable. A correlation coefficient of -1 indicates a perfect negative correlation, where an increase in one variable is directly proportional to a decrease in the other variable. In general, a correlation coefficient greater than 0.7 or less than -0.7 is considered to be a strong correlation.

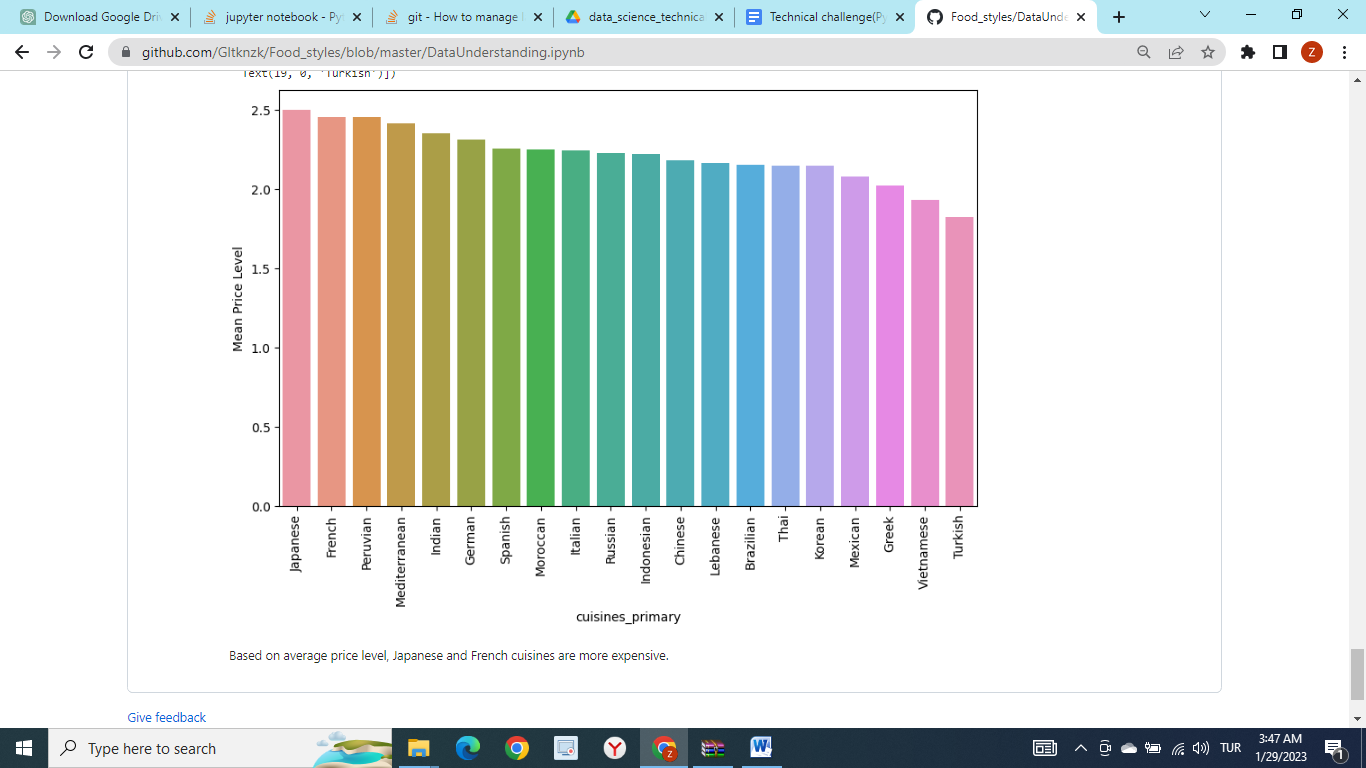
1. Are vegetarian-friendly restaurants *better* than non-vegetarian ones?



Based on average rating Vegetarian Friendly Restaurants are better than Non Vegetarian Restaurants

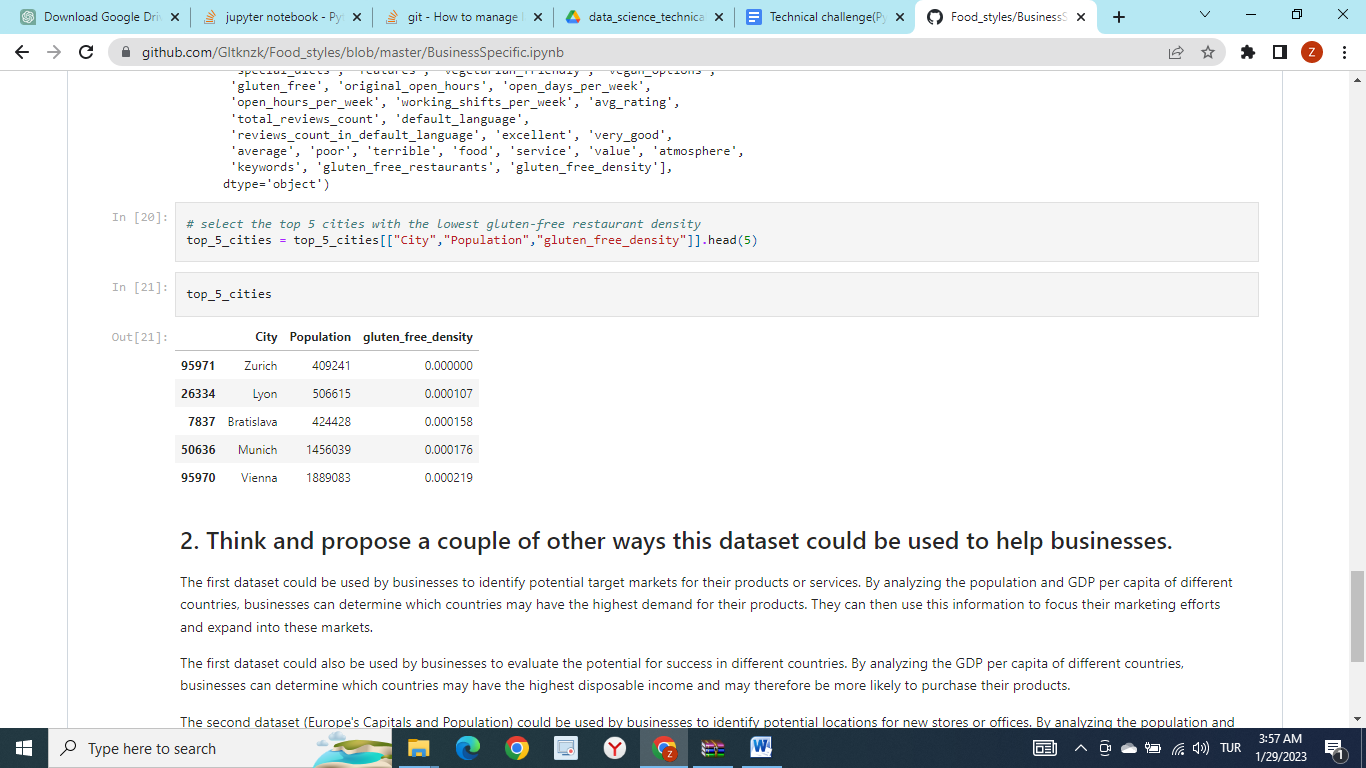
1. Are there any significantly more expensive cuisines?

Based on average price level, Japanese and French cuisines are more expensive.



**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?



Zurich, Lyon, Bratislava, Munich and Vienna seem the cities considering gluten free restaurant density.

1. Think and propose a couple of other ways this dataset could be used to help businesses.

The first dataset (restaurants) could be used by businesses to identify potential target markets for their products or services. By analyzing the population and GDP per capita of different countries, businesses can determine which countries may have the highest demand for their products. They can then use this information to focus their marketing efforts and expand into these markets.

The first dataset could also be used by businesses to evaluate the potential for success in different countries. By analyzing the GDP per capita of different countries, businesses can determine which countries may have the highest disposable income and may therefore be more likely to purchase their products.

The second dataset (Europe's Capitals and Population) could be used by businesses to identify potential locations for new stores or offices. By analyzing the population and land area of different cities within a country, businesses can determine which cities have the highest population density and may therefore have the highest demand for their products or services. They can then use this information to choose the best location for their new store or office.

The second dataset could be used by businesses to compare the cost of operating in different cities. By analyzing the land area of different cities, businesses can determine which cities have the highest cost of land and therefore the highest costs of operating. They can then use this information to choose the most cost-effective location for their business operations.

**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)?

Density of Italian restaurants on Paris map is created by using Folium. Location for a new restaurant can be decided based on the map below. This map can be found on repo.

