Introduction

In this project you will use all the skills you acquired in the course to perform a data analysis task. You are given several datasets and will have to analyze each one of them separately and all together, come up with insights, and report your findings. As an aspiring data scientist, you should explain the reasoning behind your analysis and walk the reader through your motivation and logic. Any result presented must be described and discussed. **You should submit a single Jupyter notebook that includes all the logic you wrote. Text should be included in Markdown cells.**

Customer Dataset

The file 'customers.csv' contains a dataset of customer data. The information was gathered directly from the customers. Some information may be missing not reliable. You are requested to prepare the data for analysis, organize it, and perform some simple analyses.

The dataset contains 9189 samples, each one has 9 variables describing demographic and personal information as follows:

- 'name' – The first and last name of the customer, separated by "__".
- 'age' - The age of the customer.
- 'gender' - The gender of the customer.
- 'email' - The domain of the email only (the string after the '@' sign).
- 'business_nature' - What is the industry where the customer works.
- 'Company' - The name of the company, where the customer works.
- 'position' - The job position of the customer.
- 'payment_method' - The way the customer pays to the client.
- 'target' - Some variable

Data cleaning and exploration:

1. Warmup: split the 'name' feature into two features: first and last name.
2. Split the 'target list' feature to extract the values of all the 'target' values.
   Create three aggregated features from the list of target values. (Mean, SD, Count)
3. Summarize the data (report the distribution of the features).
   Use meaningful visualizations that will help understand the data. Remember to describe and discuss the results.
4. Handle outliers, missing values and points that might skew the data.
   You should write this logic as a "Data_Cleaner" Class inside the Jupyter notebook.
   The Constructor for the class should be initialized with a dataframe. Each cleaning task should be implemented as a separate method in the class. You can implement as many helper functions as you need. Make sure to explain your logic of each task you implement and include a docstring for every method.
5. What can you infer from the data?
   Describe the customers based on the data.

Extending the data:

6. If you could take data from the web, what features would you add to this dataset to improve it? How would you relate the external dataset to samples in our dataset? What possible caveats you might face?

Company Dataset

The file 'companies.xlsx' contains a dataset with data about companies. The information was gathered automatically. Some information may be missing not reliable. You are requested to prepare the data for analysis using a script, organize it, and perform some simple analyses.

The dataset contains 11 samples, each one has 7 variables:

- 'company_id' – A unique identifier of the company.
- 'name' – The name of the company.
- 'domain' – The domain name of the company.
- 'business_nature' – The industry to which the company belongs.
- 'employee_number' – Number of employees in company.
- 'type' – Indication whether this is a private or public company.
- 'market_cap' – The company valuation in $.

Data cleaning and exploration:

7. Summarize the data (report the distribution of interesting features).
   Use meaningful visualizations that will help understand the data. Remember to describe and discuss the results.
8. Handle outliers, missing values and points that might skew the data.
   Try to reuse the code you wrote for the first dataset whenever this makes sense.

Aggregate Customer Data into Companies Data

In this part you will combine the data from both datasets to further analyze the company dataset.

Customer dataset data aggregation:

9. Extract information about companies from the customer dataset by performing aggregations (count, sum, average, min, max, etc.).
10. Summarize the aggregated data on the company level.
    Use meaningful visualizations that will help understand the data. Remember to describe and discuss the results.

Dataset merging

11. Combine the original company dataset with the aggregated company data you created.
12. Analyze the combined dataset.
    What can you learn on the companies given the aggregated customer data?

Good Luck!