Assignments for "What is Exploratory Data Analysis?"

To complete this assignment, send the Github link of the Jupyter notebook file containing solutions to the following questions:

In [ ]:

1. What is the purpose the data exploration?

- Geting information and wide aspect on our data.
- To increase the success on result;
  - Find extreem values (outliners) and if possible and required overcome them.
  - Find data errors (dublicates, missing values etc)
- To fit out data on a model determine the how out data values distributed
- Find usable features and relations for using in our analysis.

1. Suppose you are working on a dataset containing customer reviews of an e-commerce company's products. Customer reviews collected through the company's website are rated between 1 and 5 depending on whether the content is positive or negative.
   - What kind of problems do you expect to encounter in this raw data?
   - What kind of problems do you expect to encounter in this raw data?
   - If your task is to identify the characteristics that reveal that the customers' comments are positive or negative, how would you do that and what methods would you use for this?
   - What are the useful features that can be extracted from raw data? How can you access this data and understand whether it is useful?

- Some values may be null or out of the normal (1-5) range, some reviews may be more than one time occured by the same customers.
- We can group data by some categorical or customer specific values and analyz our results on these groups to identify the characteristics those can be effective on positive or negative review.
- Age, sex, country,... these features may be extracted from raw data.
- We use pandas dataframe to access this data and filter by these feautures and check if our target is related with these features using statistics and visualization methods.

1. Why do you think missing values should be taking care of?

- Some statistical python methods produces exceptions on missing values.
- If density of missing values are large this can affect our analysis

1. Do you think that outliers have an impact on a dataset? If so, how would you explain this impact?

Outliers, if exists in large number or have very extreme values, may change our data characteristics. If our data values normally distributed outliers may break this distribution and make skew our distribution graph right or left.

1. Please briefly summarize you first actions when you start analyzing the data?

- Check if missing values exists
- Use describe to view general statistics of each features.
- Draw distribution graph to detect the distribution of data values for interested features.