Project 1 - Chewbacca Team

Data analysis and data scientist job positions across New York, Seattle, and San Francisco.

In this project we aim to analyze all the Data Analyst and Data scientist job openings that companies are offering in three different cities: New York, San Francisco and Seattle.

Our assumption is that New York and San Francisco offer more job positions than Seattle since the first two cities are famous for being the location of the majority of tech companies while Seattle's economy is based more on manufacturing companies.

Data cleaning optimization

We decided to analyze the data set “All_Data.csv” that we extracted from the “Data Scientist Job Market in the U.S. in 2018 - An outlook at data science job market in the U.S. in 2018 August.

We did pull out all the data related to the job positions we wanted to analyze(Data Analyst and Data scientist) and the cities. We created a different data frame analyzing each job position related to each city. (For example Data Analyst in Seattle & Data Scientist in Seattle).

In order to get a cleaner dataset we created a new column called “jobname” that includes all the job positions related to “Analyst” in one data frame and “Scientist” in another dataframe. This will make it easier to make the comparison among the two job positions across cities. We have also decided to create a new column called “city” which will group all the different strings for cities, for example all the values that contain the parameter “NY” within the new value “New York” in the new created column.

As a next step, it was the creation of two different datasets for each job position (Subset Scientist and Subset Analyst) for each city, for example Subset Scientist in Seattle.
All the new subset filtered has been merged with the concat function into one dataframe called “final”, a dataframe that we use to visualize the data, using the plot included in the Panda library. Library which has been uploaded as the first code in our jupyter notebook file.

Visualization and conclusion

As mentioned above, we have used the plot kind “bar” with the goal of visually checking our initial assumption. At first, the plot showing the distribution of data analyst and data scientist jobs in the three cities, secondly the singular distribution of the two jobs, and last the comparison between the two kind of data science jobs (# of data analyst VS # of data scientist)

The end of the analysis confirmed our initial hypothesis, the majority of the data science job openings are distributed among NY with 561 openings and San Francisco with 413 openings. Seattle data science job offers are only 243.
Furthermore, we can assume that data scientist roles in NY, San Francisco, and Seattle are widely more requested than data analyst roles.

Peter George, Anja, Federico