Report

San Francisco Salaries Dataset

Name: Alaa Atef Abdelhaleem Ebrahim Bediwi

- I started by exploring the dataset , knowing its shapes, data types, size and checking the null values.
- Then , contained cleaning the data and dropping the un necessary column and also handling the missing values.
- After that ,I observed that the dataset contains many ‘Not Provided’ values so I replaced the those in each column with NAN value .
- Secondly, I applies to it a descriptive statistics to know about the average total salaries is paid for the employees which consists of (‘BasePay’,

‘OvertimePay’, ‘OtherPay’, ‘Benefits’) which totally affects the total averages of the emplyees.

- Applying some visualizations to the data for showing the distribution for each column.
- And then grouping it by JobTitle of employees to see the proportion of Eemployees in different departments and from it , I conclude that the highest percentage of employees is in [Transit Operator] job which their percentage represents 6.2%.
- Grouping the dataset by JobTitle and calculating some summary statistics for each group according to the ‘Year’ and ‘ TotalPay’ deducing that the highest average salaries is related to ‘Chief Investment Officer’ JobTitle.
- Finally, I applied simple correlation analysis to the dataset by first, calculating the correclation between several columns to each other and found that the highest correlations are between (BasePay , TotalPay) , (BaseBay, Benefits) , and

slightly small difference between (TotalPay , Benefits).

- Visualizing the results using scatter plots for fitting these data.