**IS 733 LAB 2:**

**Report on Task 2:**

* **Number of attributes:** 11
* **Numeric attributes:** 4 ; **Discrete attributes:** 7
* **Report to manager:** Firstly, there are many discrete attributes so while profiling the data it produced many NaN values. So, while profiling doesn’t provide a desired report to draw conclusions. Secondly, to understand the data, an instance like the mean salary for every unique experience level is plotted with the help of a bar graph between the ‘experience\_ level’ and’ salary’ in USD. Although we can retrieve data by performing equations between different attributes. Sometimes it is difficult to understand (in the case of ‘job\_title’ vs ‘salary’) as they are producing huge data and it is a bit difficult to analyze, salary in different currencies would be another major issue to deal with and it would make the situation more complex. So, data cleaning or pre-processing is required.
* **Pre-processing:** There are no missing values found and dropping attributes like ‘salary’ and ‘salary\_currency’ because rather than having salary in multiple different currencies we can use the attribute ‘salary\_in\_usd’ so that data can be reduced. While plotting the boxplot, salary has the highest number of outliers so it is justified that ‘salary’ can be removed as well as the dependent column(as the ‘salary\_currecy’ is used to refer to the ‘salary’) ‘salary\_currency’ can also be dropped. Secondly, checking for duplicate values and handling of duplicate values are done by dropping duplicates. However, we can perform discretization and sampling(which is mostly useful for algorithm training) too.