Salary Prediction

Prepared By: Erum Akmal

1. **Problem Statement for Project:** In this project, I have predicted the salary of individuals based on their experience. I have used a linear regression model in this project which can benefit companies in improving their salary ranges based on candidate's experience.

2. **Data Sources and Description:** For data sources I searched on Kaggle. For this project, I worked on salary detection project with linear regression. My data set includes columns experience (in months) and salary (in thousands).

3. **Data Mining Operations:**
   -Data wrangling: Data cleaning was not needed for this project since all data was numeric and there were no null values.

   After calculation of correlation coefficient, data was normalized by using a minmax scaler. In the next step, splitting of the data was done into 80% training and 20% testing. After that, linear regression and model evaluation was performed, followed by the generation of graphs.

   -Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

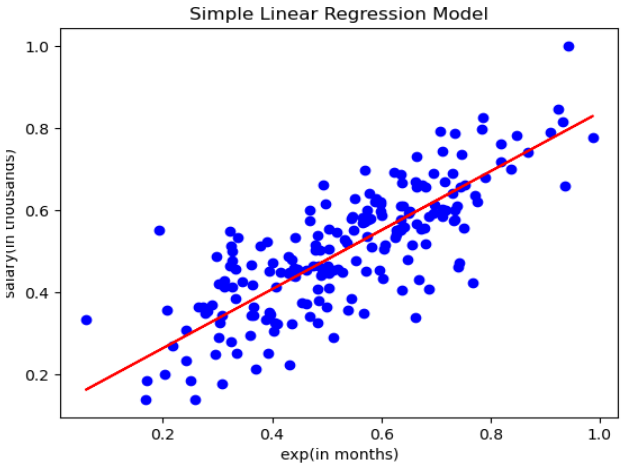
   -Data Modeling: I have performed the analysis below in my project:

   - Correlation metrics
   - No cleaning required in dataset
   - Min Max Scaler
   - Splitting dataset into 80-20
   - Modelling with Linear regression
   - Evaluation: MSE, MAE, r2_score
   - Scatter plot graph
   - Correlation heatmap
   - Histogram of diff or target and prediction
   - Box plot for dataset

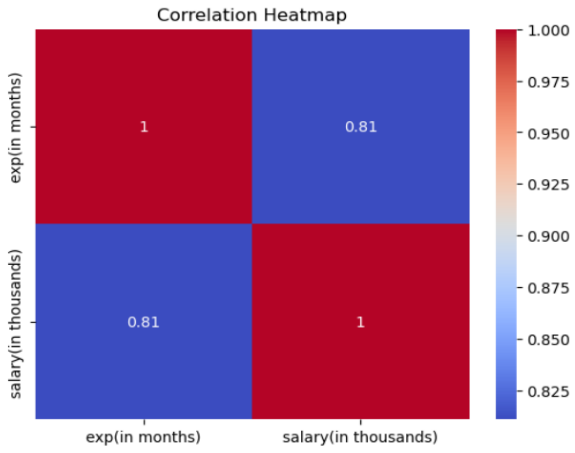
4. **Model Outputs:**
   Below is my evaluation of the model performance
   - Mean squared error: 0.00863
   - Mean absolute error: 0.0735
   - r2score: 0.6208

After the analysis, I can observe a positive correlation between experience and salary. A 62% variance in salary is explained here.

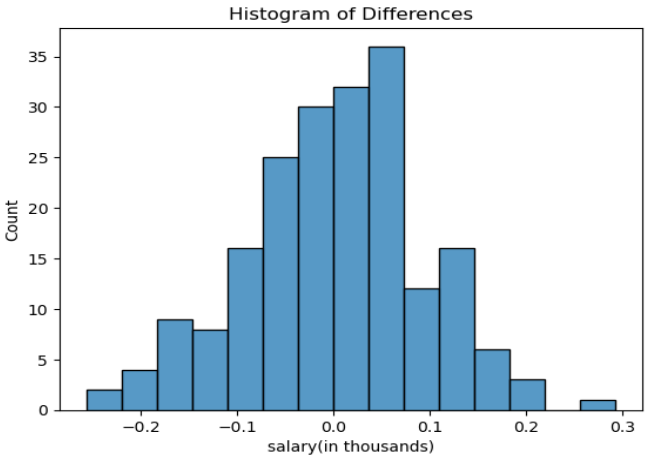
Here is my first graph for simple linear regression model where experience (in months) is shown on x-axis and salary (in thousands) is shown on y-axis.



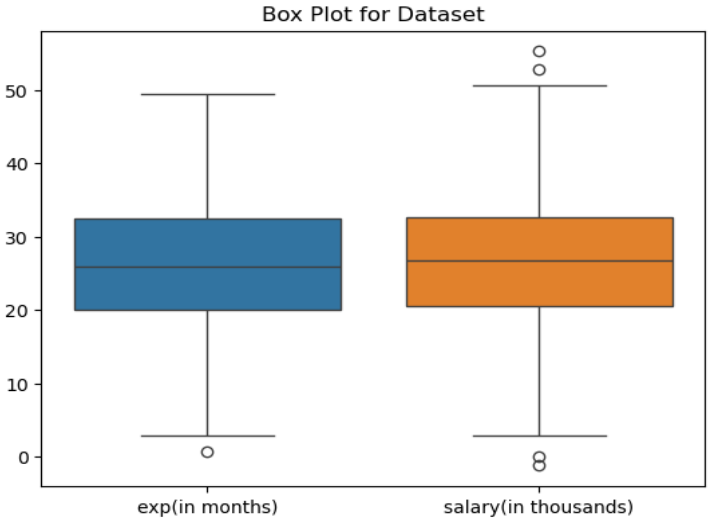
My second graph is a correlation heatmap for experience (in months) and salary (in thousands).



My third graph in this project is for Histogram of differences for experience (in months) and salary (in thousands).



My fourth and last graph for this project is a box plot for the data set.



5. Limitations and Challenges:

Some of the limitations of this study are that I have only used one variable experience. There could be multiple other factors that can potentially affect salary.

After looking at r2 score, it can be observed that there is 62% variance in salary.

6. Conclusion:

I was able to solve the problem partially. By adding more data and variables in this study, a better model could have been predicted for salary.

Linear regression on a simple data set has been performed in this study. A linear positive correlation is shown in graphs above which means that with increase in experience in months; salary in thousands is likely to be increased.

7. References

Reference for Dataset
https://www.kaggle.com/datasets/saquib7hussain/experience-salary-dataset