Analysis of Salary earned by a Data Science employee based on Experience, Employment Type and Job Title.

1st Rishitha Garnepudi
dept of Computer Science
University of North Texas
Denton, Texas
RishithaGarnepudi@my.unt.edu
11716279

2nd Reethu Karnati
dept of Computer Science
University of North Texas
Denton, Texas
ReethuKarnati@my.unt.edu
11719982

3rdAkshaya Yalla
dept of Computer Science
University of North Texas
Denton, Texas
AkshayaYalla@my.unt.edu
11725870

4thBharath Reddy Gurram
dept of Computer Science
University of North Texas
Denton, Texas
BharathReddyGurram@my.unt.edu
11708657

***Abstract*—Many data science aspirants and current employees want to know about the current market availability to better their future. For this reason, they need thorough information about the statistics of how, when and where they can secure a position in the Data Science field for their education. The project we are working on provides them with clear visual statistics for their selection of their job choice. This project investigates the relationship between salary, experience level, employment type, and job title for data science professionals. The analysis utilizes a dataset of data science designations and their corresponding salary information. Job titles are categorized based on an employee's role, and salary ranges are calculated based on experience levels. The project also assesses the accuracy of job title categorization and salary calculations to ensure the reliability of the analysis.**

***Index Terms*—Data Science, Salary, Employment Type, Experience, Job Title, Exploratory Data Analysis, Accuracy, Statistics.**

I. Introduction

'Analysis of Salary earned by a Data Science employee based on Experience, Employment Type and Job Title.' This project helps many data science aspirants who are confused with what they can do after the completion of their education. Students who are about to start their career and are in a dilemma want clear information about what kinds of jobs are present and what kind of jobs they can do. This will help them choose a path they want. For this project we are using a data set with attributes: Work year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in Dollars, Employee Residence, Remote Ratio, Company Location and Company Size. Among these attributes we are using the main three which are Job Title, Employment Type and Experience Level. These three elements, when correlated with salary, provide data science aspirants with a detailed and visual analysis of what designation will be the best for them and which role is currently best in the market.

II. Motivation

The motive of this project is to determine how much salary a Data Science employee earns based on their experience level(how many years they have worked), type of employee they are(part time or full time) and their designated job title. Exploratory data analysis is implemented to understand the data and draw patterns, anomalies and conclusions from the dataset. It is also used in this project to visually display the data. Useful insights would be drawn from the data which would be helpful to solve business problems.

III. Significance

It gives a slight idea to future Data Science aspirants about the jobs available in the market, the salary of the job, the designations in that field, how impactful the company size is on the salary or the designation. It also aids the current Data Science employees to evaluate their current income and make informative decisions to improve their careers.

IV. Objectives

- Identify the key factors that influence salary for data science professionals, specifically considering experience, employment type, and job title.
- Determine the average salary range for data science professionals at different years of work experience they have.
- Analyze salary differences between full-time and part-time workers for the data science profession.
- Compare salaries for different data science job titles to understand the impact of job specialization.
- Identify emerging trends in data science salaries over time and to provide insights for data science professionals, employers, and policymakers based on the findings of the analysis.

V. Features

The features include data collection, data exploration, data pre-processing, model training, model evaluation, feature selection, data visualization, analysis, deployment.

A. Data Collection

Data collection is the first step in the process. Here raw data is collected to be used for the project according to its requirements. Raw data is, in most cases, messy, incomplete, and may have errors. In this step, data can either be quantitative or qualitative. Data is collected through various processes like research, surveys, interviews and questions.

B. Data Pre-processing

Data preprocessing is the second step in this process. Here raw data which has been collected in the previous step is cleaned, meaning all the errors are removed and it will not be messy or incomplete as it will be thoroughly analyzed and cleaned. Here this data can be in the form of video, audio or messages.

C. Model Training

This is the third step in the process. Here, machine learning algorithms are used on the training part of the data to help identify and learn good values for all attributes involved. This step results in a working model which is later tested and deployed to the customer or end user.

D. Model Evaluation

The next step is model evaluation. Here different evaluation metrics are used to understand machine learning algorithms performance, weaknesses and strengths. The metrics include accuracy, precision, confusion matrix and AUC curve.

E. Data Exploration

In data exploration, the data is explored thoroughly to understand the attributes characteristics and the correlations among the attributes.

F. Feature Selection

In this step, the important or most impactful attributes are selected for the functioning without any changes or modification on the attributes.

G. Data Visualization

The correlation between independent and dependent attributes are visually displayed using methods like bar chars, pie plots, histograms, scatterplots, heatmaps, line chart and box plot.

H. Analysis

Here, the analysis of the data takes place based on the graphical displays of the correlation between the independent variables and dependent variables.

I. Deplotment

This is the final step. Here the end model or product is deployed to the customer or end user.

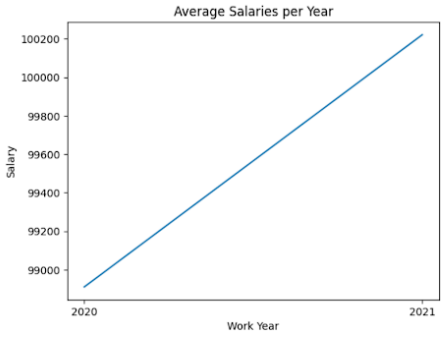
VI. Data Description

We are using the Data Science Jobs Salaries dataset from Kaggle. The attributes present in the data are work year, experience level, employment type, job title, salary, salary currency, salary in US dollars, employee residence, remote ratio, company size and company location.This dataset has been chosen by the whole team because it provides a lot of attributes which gives a broad scope for analysis and understanding the situation of current Data Science employees or future aspirants. The data being used in this project gives a visual representation of how the salary of a Data Science employee is dependent on the main three aspects which are Job Title, Employment Type and Experience level.

VII. Detailed Information of Dataset

A. Work Year includes

The year during which the salary was paid. There are two types of work year values:
2020: Year with a definitive amount from the past
2021e: Year with an estimated amount (e.g. current year)



B. Experience Level includes

- EN: Entry-level / Junior
- MI: Mid-level / Intermediate
- SE: Senior-level / Expert
- EX: Executive-level / Director

C. Employment Type includes

- PT: Part-time
- FT: Full-time
- CT: Contract
- FL: Freelance

D. Job Titles include

The role worked in during the year.

E. Salary includes

The total gross salary amount paid.

F. Salary Currency

The currency of the salary paid.

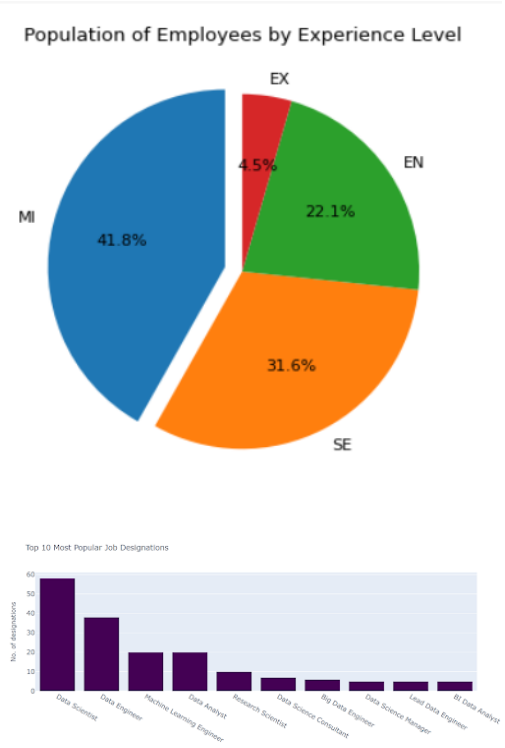
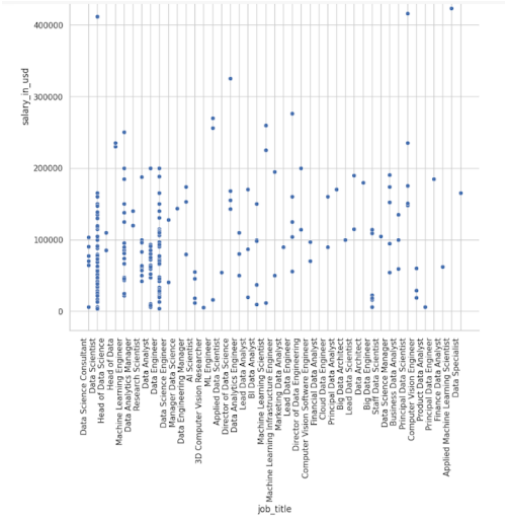
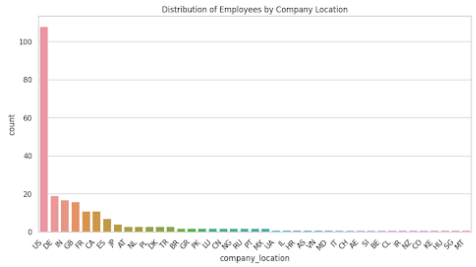


Fig. 1. Top 10 most popular job designations.





G. Salary in US dollars

The salary in USD.

H. Employee Residence

Employee's primary country of residence during the work year.

I. Remote Ratio

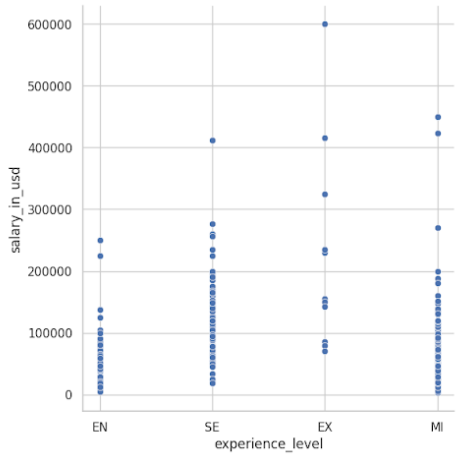
- 0: No remote work (less than 20
- 50: Partially remote
- 100: Fully remote (more than 80

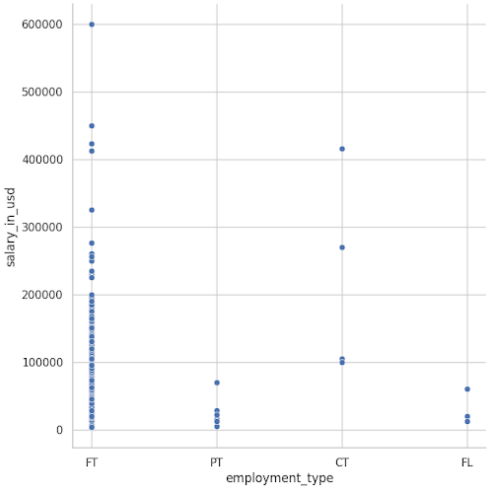
J. Company location

The country of the employer's main office or contracting branch.

K. Company Size

- S: less than 50 employees (small)
- M: 50 to 250 employees (medium)
- L: more than 250 employees (large)





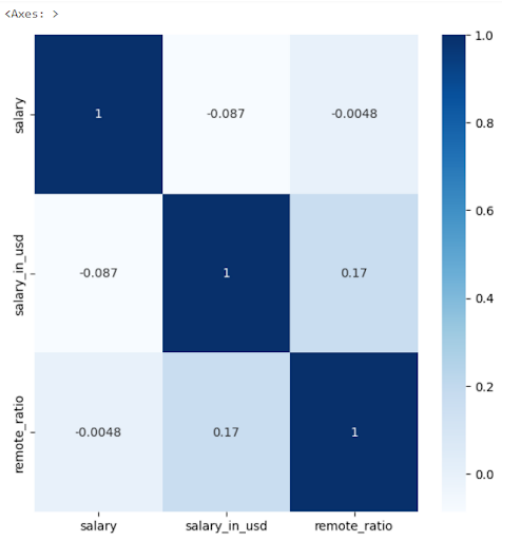


Fig. 2. Salary in USD

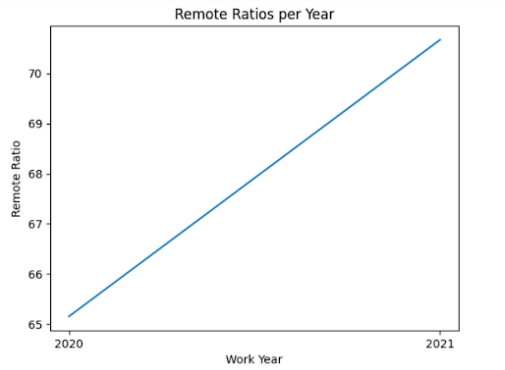


Fig. 3. Example of a figure caption.

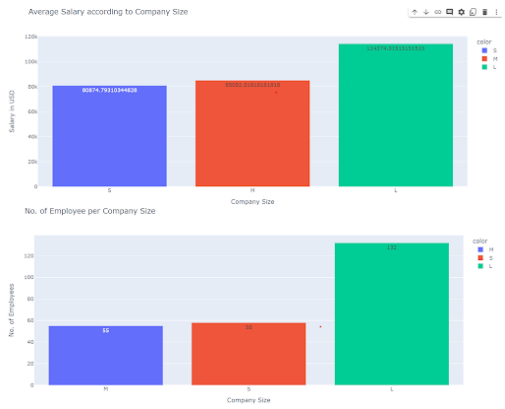


Fig. 4. Example of a figure caption.

VIII. Detail Design of Features

The features of 'Analysis of Salary earned by a Data Science employee based on Experience, Employment Type and Job Title' include data collection, data exploration, data pre-processing, feature selection, data visualization and Analysis.

A. Data Collection

We have taken the dataset from kaggle. This dataset includes attributes such as Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in Dollars, Employee Residence, Remote Ratio, Company Location, and Company Size. In this we use Experience Level, Employment Type and Job Title as the main attributes.

B. Data Pre-processing

For the data preprocessing we have first printed the number of duplicate rows present in the dataset and dropped them. We have also looked for null or missing values to remove them from the dataset, but they aren't present in the dataset.

C. Model training

In this project, we have used the Machine learning algorithm Linear Regression to train and build the model.

D. Model Evaluation

After using ML algorithm Linear Regression, the results we got are as Mean Squared Error is 9653537230.921192, Root Mean Squared Error is 98252.41590373842 and R-square is -0.4862211305910329.

E. Data Exploration

In data exploration, the data is explored thoroughly to understand the attributes characteristics and the correlations among the attributes.

F. Feature Selection

We took Experience Level, Employment Type and Job Title attributes as independent variables because their correlation with the salary attribute is statistically and analytically more than other attributes.

G. Data Visualization

We have visually displayed the data using EDA to understand the relationship between three independent variables namely, Experience Level, Employment Type and Job Title and one dependent variable which is Salary.

H. Analysis

Analyzation of the data based on the graphical displays of the correlation between the independent variables Experience Level, Employment Type and Job Title and dependent variable Salary is accomplished.

I. Deployment

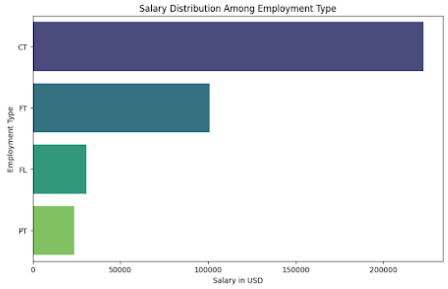
As the project is finished, it is delivered to the end users.


```
Statistics for Experience Level 'EN':
Mean Salary: $59753.46
Median Salary: $58800.50
Salary Range: $246000.00

Statistics for Experience Level 'SE':
Mean Salary: $128841.30
Median Salary: $120000.00
Salary Range: $392948.00

Statistics for Experience Level 'EX':
Mean Salary: $226288.00
Median Salary: $154963.00
Salary Range: $529671.00

Statistics for Experience Level 'MI':
Mean Salary: $85681.65
Median Salary: $72812.50
Salary Range: $447124.00
```




```
Statistics for Employment Type 'FT':
Mean Salary: $100986.63
Median Salary: $85000.00
Salary Range: $597124.00

Statistics for Employment Type 'PT':
Mean Salary: $23748.14
Median Salary: $15966.00
Salary Range: $64906.00

Statistics for Employment Type 'CT':
Mean Salary: $222750.00
Median Salary: $187500.00
Salary Range: $316000.00

Statistics for Employment Type 'FL':
Mean Salary: $30666.67
Median Salary: $20000.00
Salary Range: $48000.00
```

IX. Data set analysis

A. Salary Analysis by Experience

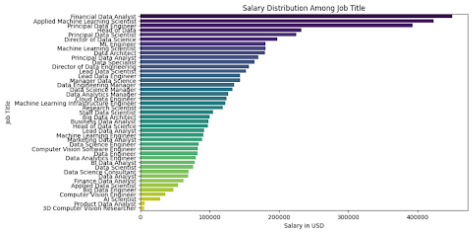
The above graph displays the distribution of salaries among experience levels. The experience levels include Entry Level, Mid Level, Senior Level and Executive Level. The range of Salary in US dollars ranges from 50,000 USD to 200,000 USD. The salary distribution of executive level is 200,00 US dollars being the highest among all. The least distribution of salaries among experience level is Entry Level with value of almost 60,000. The second least is mid level with a value of 85,000. The second highest is Senior Level with a value of 130,000. Below the graph is the statistics for the four experience levels. The statistics include mean, median and range. For the Entrance Level the mean salary is 226288.00 USD, and the median salary is 154963.00 USD and the salary range is 529671.00 USD. For the Senior Level the mean salary is 128841.30 USD, the median salary is 120000.00 USD and the salary range is 392948.00 USD. For the Mid Level the mean salary is 85681.65 USD, the median salary is 72812.50 USD and the salary range is 447124.00 USD. For the Entrance Level the mean salary is 59753.46 USD the median salary is 58800.50 USD, and the salary range is 246000.00 USD.

B. Salary Analysis by Employment Type

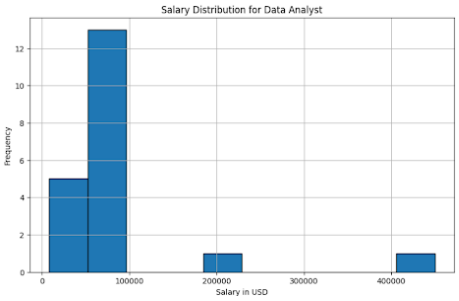
The above graph displays the distribution of salaries among employment types. The employment types include Full-time, Part-time, Contract and Freelance . The range of Salary in US dollars ranges from 50,000 to 200,000 in US dollars. The salary distribution for Contract type employment is above 200,000 USD, being the highest. The second highest is Full-time with 100,000 USD. The third highest is Freelance with 40,000 USD and the least type of employment is Part-time with 30,000 USD. Below the graph is the statistics for the four employment types. The statistics include mean, median and range. For contract type employment the mean salary is 222750.00 USD, and the median salary is 187500.00 USD and the salary range is 316000.00 USD. For full-time employment type, the mean salary is 100986.63 USD, the median salary is 85000.00 USD and the salary range is 597124.00 USD. For freelance employment type the mean salary is 30666.67 USD, the median salary is 20000.00 USD and the salary range is 48000.00 USD. For the part-time employment type the mean salary is 23748.14 USD, the median salary is 15966.00 USD, and the salary range is 64906.00 USD.

C. Salary Analysis by Job Title

The first graph displays the salary distribution among each graph title. There are a total of 43 job titles present in the graph. The highest earning job title is ‘Financial Data Analyst’ with a salary over 400,00 USD. The leasts earning job title is ‘3D Computer Vision Researcher’ with a salary just above 0. Below the graph, we have a sample graph of ‘Salary Distribution for AI Scientists’ and its statistics. The Mean Salary is 28599.60 USD, the Median Salary is 18102.00 USD and the Salary Range is 43000.00 USD. Below the




```
Statistics for Job Title 'AI Scientist':
Mean Salary: $28599.60
Median Salary: $18102.00
Salary Range: $43000.00
```




```
Statistics for Job Title 'Data Analyst':
Mean Salary: $69329.15
Median Salary: $71984.00
Salary Range: $193928.00
```

statistics for AI SCientist, there is another sample graph for 'Salary Distribution for Data analyst'. Below the graph are its statistics. The Mean Salary is 69329.15 USD, the Median Salary is 71984.00 USD, and the Salary Range is 193928.00 USD.

D. Code Analysis

For the code, we first imported required python libraries like pandas, numpy, matplotlib,etc. Then the dataset is loaded and read. Then we displayed the first five rows and the data type for each attribute. Then we print the shape of the data set and number of duplicate rows present in the dataset. Then the count of each attribute is displayed and duplicate rows are dropped. Then we once again print the count of each attribute after deleting the duplicate rows.

Next we focused on experience level. We grouped the data by experience level and calculated the salary for each level. Then we plotted a bar graph which displays the correlation between experience level and salaries. Then a pie chart which displayed the distribution of employees by their experience level. Next we used a scatter plot to display the correlation between salary in US dollars and experience level. At last we calculated the mean salary, median salary and the salary range for all four experience levels.

The next attribute we focused on was employment type. Here, similar to the above process we grouped the data by employment type and calculated its salary. Then we plotted a few graphs. The first one among them is a bar plot and the next one is a scatter plot. These both graphs show the correlation between employment type and the salary in US dollars. Finally we calculate the mean salary, median salary and salary range for all four employment types.

The third attribute we choose is the Job title. First we printed out all the names of the job titles in the dataset and its count which is 43. We then grouped the data by job titles and calculated the salary for each title and plotted different graphs for each. The first one is a bar graph which shows the correlation between the job titles and salary in US dollars. The next bar graph shows the top 10 job designations and the last graph is a scatter plot which displays the correlation between job title and salary in US dollars. We also plotted the mean, maximum and minimum salaries for each job title along with calculating the mean salary, median salary and salary range for each job title.

Then we plotted more graphs to show the relation between different attributes. Some of them include a bar graph which shows the correlation between company size and salary in US dollars and another bar graph which shows the correlation between number of employees and company size. Then we plotted another bar graph which displayed the correlation of the distribution of employees by company location. Two line graphs were also displayed which shows the average salaries per year and the remote ratios per year. The final graph is a histogram which shows the relation between remote ratio, salary in US dollars and salary.

Then we performed the machine learning algorithm linear regression and calculated the mean squared error, root mean squared error and r-squared. For this we first loaded the data and encoded categorical variables and then created dummy variables for those categorical columns. Then we split the dataset into training and testing sets. Then we evaluated the linear regression model and we got the results as Mean Squared Error is 9653537230.921192, Root Mean Squared Error is 98252.41590373842 and R-squared is -0.4862211305910329

X. Implementation

Our extensive information was collected from a variety of sources, such as business HR databases, employment websites, and industry surveys. To prepare the dataset for analysis, data cleaning was done, including addressing missing values, encoding category categories, and scaling numerical characteristics. To summarize the dataset, detailed descriptive statistics were performed. made use of a variety of visualizations, such as box plots, bar charts, and histograms, to comprehend the income distributions with respect to job title, employment type,

and experience.Several linear regression models were put into practise and refined in order to forecast earnings depending on employment type, job title, and experience.
Feature selection strategies were used to improve the interpretability and performance of the model.assessed the effectiveness of linear regression models using R-squared, MAE, and RMSE measures. To guarantee robustness, cross-validation was carried out and the models were tested on a holdout dataset.analysed model coefficients and predictor importance to obtain insightful knowledge about the factors influencing data science wages. determined the main variables affecting pay differences.The project's use of linear regression to analyze Data Science staff wages has advanced significantly. The report is a useful tool for learning about the factors that influence income in the field of data science since it summarizes the analysis's conclusions, suggestions, and consequences.

XI. Preliminary Results

A. Experience Level

Salary is significantly more for executive-level employees whose average salary is equal to 226,228 USD per year. Entry-level data scientists typically earn around 60,000 USD, while mid-level data scientists can earn around 85,000 USD in a year. Senior-level data scientists with more than five years of experience can earn upwards of 128,841 USD per year.

B. Employement Type

Contract based data scientists typically earn more than full-time, part-time and freelance data scientists. The average salary of a contract based data scientist is 222,750 USD in each year. Next, the full-time data scientists earn more, this is likely due to the fact that full-time data scientists are typically more experienced and have more responsibility. The part-time data scientists earn very less when compared to other employment type data scientists whose average salary is around 23,748 USD per year. The freelance data scientists also earn less which is equal to 30,666 USD in a year.

C. Job Title

Data scientists with certain job titles, such as Financial Data Analyst, Applied Machine Learning Scientist and Principal Data Engineer, typically earn the highest salaries. These roles typically require more experience and expertise than other data science roles. The data scientists with roles such as Product Data Analyst and 3D Computer Vision Researcher earn less salary which is around 5,750 USD per year.

D. Linear Regression Results

We got the linear regression results values as Mean Squared Error is 9653537230.921192, Root Mean Squared Error is 98252.41590373842 and R-square is -0.4862211305910329

XII. Project Management

A. Implementation report of work completed

1) Description: We have precisely identified the issue and established the aims and objectives for this phase. In order to comprehend the characteristics, we gathered the dataset, carefully examined each feature, and created a few charts. We determined the independent and dependent variables as well as their correlations. We were able to summarize the correlations between the independent and dependent variables in this way. Next, we drew the flow diagram and established our project's process. We outlined every critical action needed to complete the job. Also, we have built a model for linear regression by considering salary as a dependent variable which depends on experience level, employment type and job title. We have found out the linear regression results as Mean Squared Error, Root Mean Squared Error and R-squared error values.

2) Responsibility:

- Rishitha Garnepudi: It was my responsibility to handle duplicate data and process the data for analysis. My responsibility was developing features that enhance the quality and usefulness of the data.I made sure that the data was compliant and worked with others to record the processes. In addition, I optimized processes and performed quality checks. Furthermore, I have performed analysis on data to build linear regression models and splitting of data into training and testing sets.
- Reethu Karnati: My responsibility is to find out the relation between the independent variable Experience Level to dependent level Salary. To find the average salary of each experience level and to perform statistics for attribute Experience Level with respect to Salary. Also, I have built a linear regression model by viewing the salary variable as a dependent variable that is influenced by the job title, employment type, and experience level.
- Akshaya Yalla: My responsibility is to find how the dependent variable salary depends on Employment Type. To find out which Employment Type has high salary and which has low salary and to perform analysis on them. Even more, I have evaluated logistic regression.
- Bharath Reddy Gurram: It is my responsibility to analyze the number of Designations or Job Titles available under Data Science employment. Also to check which Jobs has highest paying salaries and which has lowest paying salaries. Additionally, my goal is to determine the Mean Squared Error, and Root Mean Squared Error and R-squared error values of the linear regression results.

3) Contribution:

- Rishitha Garnepudi: 25
- Reethu Karnati: 25
- Akshaya Yalla: 25
- Bharath Reddy Gurram: 25

Acknowledgment

I am particularly thankful to Professor Sayed Khushal Shah for his guidance and insightful feedback throughout the

project. Their expertise greatly enriched the quality of our work.
Furthermore, I want to thank Ruthvik and Dinesh for their support, guidance and help. Their contributions were essential to the project's overall success.
Lastly, I am grateful to my family and friends for their unwavering support and understanding during the course of this project.
This project would not have been possible without the collective efforts of these individuals and organizations, and for that, I am truly thankful.

References

- Matbouli, Y.T.; Alghamdi, S.M. Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. Information 2022, 13, 495. https://doi.org/10.3390/info13100495
- Journal of Physics: Conference Series, Volume 1881, The 2nd International Conference on Computing and Data Science (CONF-CDS) 2021 28-30 January 2021, Stanford, United States Citation Biqi Li 2021 J. Phys.: Conf. Ser. 1881 032022
- Nan X, Yuming L, Jianfeng P and Haigang M From e-HRM to HRM: The Evaluation of Human Resource Data Management Technology of HRD 05 79-88

XIII. Github link

https://github.com/YallaAkshaya/Empirical-Analysis-Final-Project