<u>Salary Prediction of Data Scientist</u>

M.Sc Data Science(sem-1)

Student Id – 202318026, 202318059

Course Id – IT606

INTRODUCTION

1.1 INTRODUCTION

- Now days, Major reason an employee switches the company is the salary of the employee. Employees keep switching the company to get the expected salary. And it leads to loss of the company and to overcome this loss we came with an idea what if the employee gets the desired/expected salary from the Company or Organization. In this Competitive world everyone has a higher expectation and goals. But we cannot randomly provide everyone their expected salary there should be a system which should measure the ability of the Employee for the Expected salary. We cannot decide the exact salary but we can predict it by using certain data sets. A prediction is an assumption about a future event.

- In this project the main aim is predicting salary and making a suitable user-friendly graph. So that an Employee can get the desired salary on the basis of his qualification and hard work. For developing this system, we are using a Linear regression algorithm of supervised learning in machine learning.

- Supervised learning is basically a learning task of a learning function that maps an input to an output of given example. In supervised learning each example is pair having input parameter and the desired output value.

- Linear regression algorithm in machine learning is a supervised learning technique to approximate the mapping function to get the best predictions. The main goal of regression is the construction of an efficient model to predict the dependent attribute from a bunch of attribute variables. A regression problem is when the output value is real or a continuous value like salary.

- With the expansion of businesses, so does the number of employees. Firms essentially raise their employees' compensation in order to keep their talents and raise them. While there are no major obstacles in the way of salary increases in small businesses, this process should be carried out carefully in large firms in accordance with a number of guidelines to avoid negative consequences that could disrupt workforce mobilization. For organizations with a significant number of employees, developing a model in which market conditions are fully defined and all economic elements are taken into account necessitates a months-long procedure. In this regard, research developed a technique for estimating wage increases based on machine learning.Certain characteristics were determined and a specific scale was developed to determine the functional outcome of this study.

1.2 PROBLEM STATEMENT

- In order to make the final gift on the job, employers must consider a variety of elements, including demographics and others. Although this work is performed by experienced human resource managers in cooperation with the appropriate departmental level management, it is always a difficult decision. For these decision-makers, any form of automated decision-making system can be highly useful in generating acceptable wage suggestions. Companies typically have their own compensation forecasting system, which uses internal data to forecast new hire earnings. In this regard, research developed a machine learning-based approach for estimating wage increases.
- Overspending on new staff is a common concern for businesses. And getting a head start on the new employee's salary expectations will benefit both the person and the organization. Overspending to hire new employees, as we all know, will have a severe impact on the company's financial health.

1.3 OBJECTIVE

- Salary is a continuous variable that may be calculated using regression analysis. A model for raw characteristics with two goals: straightforward wage prediction and illustrating the influence of embedded feature selection. Using job pay datasets, the purpose of this research is to create a strong machine learning model that can predict future employee wages based on current employee salary data.

1.4 METHODOLOGY

In order to obtain useful information on online recruitment of IT professionals, we compare and contrast different strategies with machine learning models. The procedure follows the best practices in literature and industry, which includes different categories:

I. Data Collection: A Python web browser is designed to analyze and collect information needed for a website.

II. Correcting Inconsistencies: To remedy the inaccuracies and inconsistencies found in the survey, subject matter knowledge was required. We cleaned up the data by first converting all of the questions into variables, then translating all of the details into English. We put together several different words in the replies that refer to the same term. Corrected errors and misspellings. Finally, commas, apostrophes, quotation marks, question marks, and other punctuation symbols have been eliminated.

III. Manual feature engineering: Features which are not important are discarded and some are suspended (e.g. converted into numerical features) by misusing domain information.

IV. Data set definition: To give original and integrated data definition, mathematical techniques and basic models were applied.

V. Deploying the model: The final project was deployed on localhost.

Machine learning is a branch of artificial intelligence that uses mathematical approaches to make machines more environmentally friendly. permits the active gadget to decide on hosting responsibilities. These programmes, or algorithms, may be used to test and expand over time by examining fresh records. The reason why computers can deduce meaning from records. Statistics are therefore the key to unlocking Machine Learning. A collection of ML rules may be highly useful as the most significant ML data you should have. to learn about a building's construction It's an intellectual property door founded on the premise that computers can search through records, discover styles, and make decisions with little human interaction.

Researchers using artificial intelligence information wish to examine if computers should study information based on sample recognition and the notion that computer programmes can test without being built to conduct good activities. When models are given fresh statistics, they may be able to adapt autonomously, which is a typical property of device knowledge. They study math in order to make consistent, repeatable judgments and outcomes. Not only is there new technological understanding now, but there is also fresh energy.

There are several perspectives on salary, including the use of despair and effective retreat, and the use of statistical sets to play statistics and enquiries. ¾ size of information is used to train the calculations, and ¼ size of information is designated as the Test set.

Highlighting, system efficiency and ingenuity have dramatically changed the world view of cost estimates with greater accuracy and forecasting. In addition, over the next few years significant improvements can be made to the use of these enhancements in anticipation of price payments. gadget learning has the following algorithms:

● **Supervised learning**

● **Unsupervised learning**

1.4.1 SUPERVISED LEARNING

This is a list of predictions. These predictions are unrelated variables. This learning algorithm's goal is to make predictions based on this set of independent variables. Variability in result predictability. This is a conditional variable. A function is formed from a group of independent variations that aids in the delivery of our intended output inputs. The machine is constantly trained in order to attain a particular level of accuracy in our training data. Linear regression, hindsight, KNN decision tree, random forest, and other guided readings are examples.

1.4.2 UNSUPERVISED LEARNING

No specific aim or outcome can be estimated or predicted with this approach. It is used to join many groups, which are then divided into various groups in order to interfere. K-Means are another example of unregulated learning.

1.5 ORGANIZATION

- Overall performance of controllable separators practical training is considered with limited memory accessible on web scraping. Accumulating training records and is often used directly in the steps of the section, reducing weight for customers. The report is composed of: section I describes the features used in determining the salary. Section II explains the proposed method and section IV defines test results and finally section V describes the conclusion.

SYSTEM DEVELOPMENT

2.1 FLOW CHART
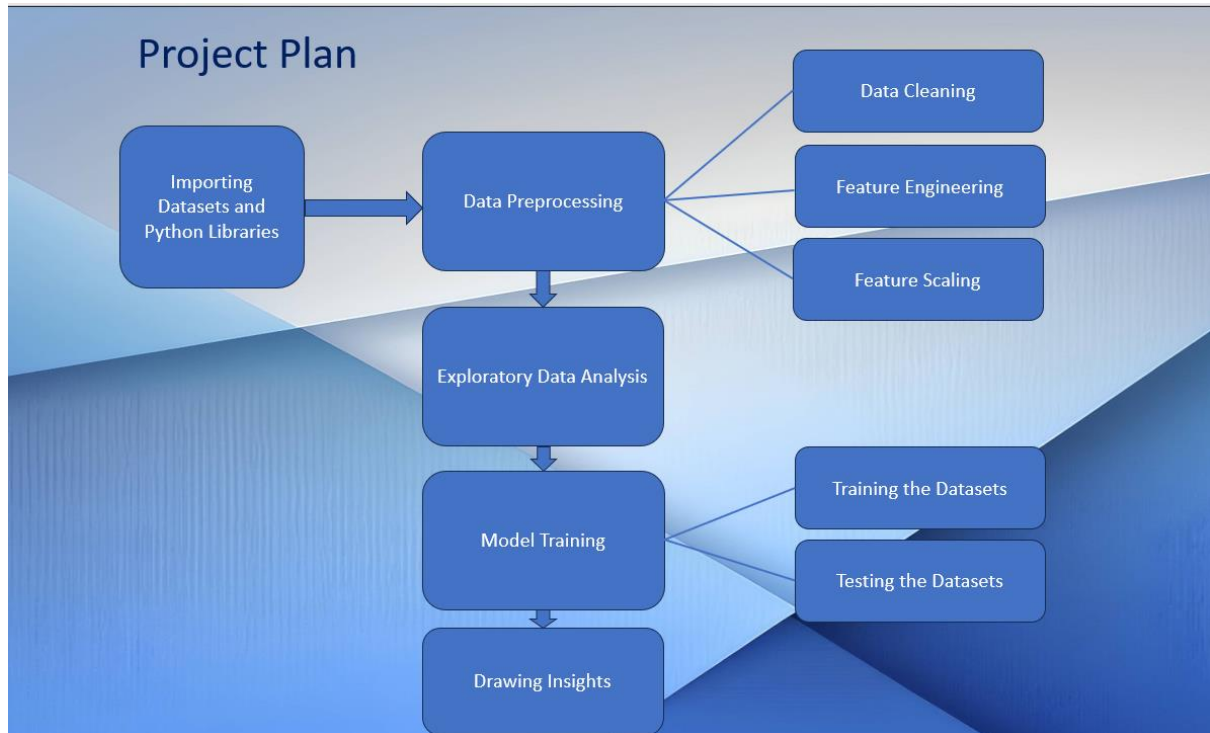
PYTHON

Python is a language-based dynamic semanticization system that is state-of-the-art. Its high level of integrated data structures, as well as dynamic typing and dynamic linking, are widely utilized in rapid application creation and as a scripting and pasting language to connect existing components. Python's straightforward, easy-to-understand syntax promotes readability while lowering programme maintenance costs.

Python encourages system flexibility and code reuse through modules and packages. Python interpreter and standard library that can be freely released in source or binary form across any big platform.Performance and consistency are two advantages of Python that make it ideal for AI and AI organizations. To perform the proper thing, there are links to specific archives, AI frameworks, and AI (ML) Private Fields.

NUMPY

NumPy is a Python library that provides aid for massive, multi-dimensional collections and matrices, as well as a massive number of mathematical functions for interacting with them Travis Oliphant invented NumPy in 2005 by integrating the competing capabilities of

Numarray in Numbers, as well as a broad variety of other features. NumPy is free software with a large number of vendors. NumPy is a project supported by NumFOCUS.

NumPy in Python is similar to MATLAB in that both translate and allow customers to construct projects more quickly as long as many jobs are centered on clusters or networks rather than scales. There are various alternatives to these crucial ones.

SCIKIT-LEARN

The greatest Scikit-learn machine library for Python. Many useful machine learning techniques and mathematical modeling approaches, such as division, deceleration, integration, and size reduction, are available in the sklearn library. Sklearn uses machine learning models. Scikit-Learn is expensive and should not be used to read, manipulate, or summarize data. Some are there to assist you in translating the spread.

Scikit-learn comes with many features. Some of them are here to help us explain the spread:

• Supervised learning algorithms: Consider any professional reading algorithms you've examined that may be considered scientific. The science toolset includes anything from standard line models to SVM and decision trees. The expansion of machine learning algorithms is one of the key reasons for scientists' increasing usage. I started using scikit, and I would encourage young people to do so as well. I'll work on supervised learning issues.

• Unchecked learning algorithms: Compilation, feature analysis, key component analysis, and unsupervised neural networks are only some of the machine learning methods available.

• Contrary verification: a variety of methods are used by sklearn to ensure the accuracy of the models followed with discrete details.

• Feature removal: Scientific learning to remove images and text elements.

• Datasets for different toys: This was useful when studying science. I have studied SAS for different educational data sets. It helped them a lot to support when they read the new library.

PANDAS

Pandas is an open-source library that makes operating with relational or labelled statistics easy and intuitive. It includes a ramification of facts formats and strategies for working with numerical facts and time collection. The NumPy Python library gives the muse for this library. Pandas is quick and provides users with excellent performance and productivity. Pandas is a game-changer when it comes to analyzing data using Python, and it is one of the most popular and commonly used data munging/wrangling tools, if not THE most popular.

MATPLOTLIB

For 2D array charts, Matplotlib is an excellent Python visualization library. One of the most important benefits of visualization is that it allows us to see large amounts of data in easily understood images. Line, bar, scatter, histogram, and more graphs are available in Matplotlib. Matplotlib is free to use because it is open sourceMatplotlib is primarily written in Python for platform portability.

2.3 DATA SET USED

While selecting or importing information sets, the type of method used, whether or not supervised, unsupervised, or semi-supervised, as well as the number of legitimate records and attributes, need to all be taken into consideration essential selection criteria. In this work, a Kaggle dataset with 50000 facts and 15 attributes is employed, in addition to the supervised technique. The range of statistics and attributes in this dataset have been chosen because they may be enormously sufficient to permit for the improvement of a green model and a huge range of options inside the prediction. The document layout is CSV. The dataset's houses are logical and competent, taking into consideration an accurate and green wage projection.

● The data is scraped from the website of Glass door using the scraper.py code.

● Job Description is broken into languages required for the job.[Python, R, Java, HTML, Excel,Spark, AWS]

● Salary is broken into hourly salary or employer provided.

● Company name simplified.

● Age of the company is simplified. If not available replace with the age calculated by DOB

● New CSV file created.

```
df = pd.read_csv("Salary_data.csv")
df.head()
```

	Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
0	0	Data Scientist	$53K-$91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense	$50 to $100 million (USD)	-1
1	1	Healthcare Data Scientist	$63K-$112K (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care	$2 to $5 billion (USD)	-1
2	2	Data Scientist	$80K-$90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services	$100 to $500 million (USD)	-1
3	3	Data Scientist	$56K-$97K (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	Government	Energy	Oil, Gas, Energy & Utilities	$500 million to $1 billion (USD)	Oak Ridge National Laboratory, National Renewa...
4	4	Data Scientist	$86K-$143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	Commerce Signals, Cardlytics, Yodlee

2.4 DATA PROCESSING

The data cleaning approach entails developing visualist and analytic methods that can handle multivariate, multidimensional data sets and are valuable in a variety of domains such as scientific visualization, business intelligence, machine learning and statistics. This method involves detecting the data, extracting it, cleaning it, and integrating it into a dataset that can be studied as needed. Because obtaining a perfect dataset that is noiseless is impossible, it is the programmers' obligation to remove as much noise, incompleteness, and other restrictions as possible to reduce errors. A decent enough model must exist.

It's a difficult task to improve the general accuracy of the ML model by removing all unneeded and useless features that don't contribute to the target variable. This study is based on data cleaning, which removes any null, missing, duplicate, and inconsequential values from the dataset. It not only cleans but also reduces the number of records and characteristics in the dataset. In ML, choosing the right feature is critical for improving the model's performance. The accuracy of the model can be improved by using the dataset attributes during training. Variables that are no longer necessary or are inconsequential must be eliminated. No data cleansing poses a risk of negatively affecting the model's forecast.There were attributes deleted from this model, and the remaining records were approximately which were efficient and sufficient enough to produce an accurate machine learning model capable of accurately predicting pay.

When data is collected and transformed into usable information, it is called data processing. Data processing is usually done by a data scientist or a team of data scientists, and it is critical that it is done correctly so that the final output, or data extraction, is not harmed.

Data processing begins with data in its immature form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context needed for computer translation and staffing throughout the organization.

FEATURES IN DATASET

```
[ ] df.columns

    Index(['Unnamed: 0', 'Job Title', 'Salary Estimate', 'Job Description',
           'Rating', 'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
           'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors'],
          dtype='object')
```

2.5 DATA VISUALIZATION

In the previous 45 years, scientific data visualization has changed dramatically. Data visualization is the process of presenting information in the form of graphs and charts. It serves as a conduit for data and images. This is especially important as data visualization makes patterns and trends more visible. Regular analyses, such as predictive analytics, are

aided by machine learning, which is a useful tool for displaying visualization. Data visualization is a field-independent technology that helps all tasks in some way. It is a useful tool for both demonstrating the relevance of huge data and assisting in its analysis.

Because the value of data is determined by its meaning, and visualization improves the meaning of data, this tool is essential. The ideal way to depict the age parameter is with a histogram, which represents the flow of the frequency, whereas parameters like educational level are best represented with a bar chart. We can explore three components at a time in a pretty simplistic method using interactive graphs using the Plotly package.

2.6 DATA CLEANING

I. Only the numeric data from Salary was parsed.

II. Separate columns were created for the employer's pay and hourly wages.

III. Because the Salary column had a few empty entries, the rows lacking Salary were eliminated.

IV. Made a separate column for the Company State by parsing the rating out of the company text.

V. A new column was added to check if the job is at the company's headquarters. A new column was added to check the company's age by using the founding date.

VI. Columns were added to check if the various talents were stated in the job description.

VII. A new column for simplified job titles and seniority has been added.

2.7 MODEL FITTING

Without the useful information collected from the data, the data has no meaning. Predictive analytics is now concerned with data analysis in order to obtain useful information. All of this is only feasible thanks to techniques that allow the ML model to perform various jobs. This paper is driven using three machine learning algorithms i.e Multiple Linear Regression and Random forest.

MULTIPLE LINEAR REGRESSION

The general form of the equation for linear regression is:

$y = A * x + B$

where,y is the dependent variable, x is the independent variable, and A and B are the equation's coefficients. The distinction between linear regression and multiple regression is that multiple regression requires the ability to handle many inputs, whereas linear

regression only requires one. To account for this modification, the multiple regression equation takes the following form:

y = A1 * x1 + A2 * x2 + … + An * xn + B

RANDOM FOREST

Random Forest is a well-known supervised learning technique-based machine learning algorithm. It can be used for both classification and regression problems in machine learning. It is based on ensemble learning, which is a way of combining multiple classifiers to solve a complicated problem and improve the performance of the model.

A random forest is made up of a large number of individual decision trees that work together as an ensemble, as the name suggests. The random forest creates a class prediction for each tree, and the class with the highest votes becomes our model's forecast.

IMPORTING DATASETS

```
df = pd.read_csv("Salary_data.csv")
df.head()
```

	Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
0	0	Data Scientist	$53K-$91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense	$50 to $100 million (USD)	-1
1	1	Healthcare Data Scientist	$63K-$112K (Glassdoor est.)	What You Will Do:\n\nl. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care	$2 to $5 billion (USD)	-1
2	2	Data Scientist	$80K-$90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services	$100 to $500 million (USD)	-1
3	3	Data Scientist	$56K-$97K (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	Government	Energy	Oil, Gas, Energy & Utilities	$500 million to $1 billion (USD)	Oak Ridge National Laboratory, National Renewa...
4	4	Data Scientist	$86K-$143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	Commerce Signals, Cardlytics, Yodlee

We can see from the data description that there are no missing values. But if we check the dataset the missing values are replaced with invalid values like '?'. Let's replace such values with 'nan' and check for missing values again.

˅ Exploring the dataset

```
[4] df.shape
```

```
(956, 15)
```

```
[5] df.columns
```

```
Index(['Unnamed: 0', 'Job Title', 'Salary Estimate', 'Job Description',
       'Rating', 'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
       'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors'],
      dtype='object')
```

```
df.dtypes
```

```
Unnamed: 0             int64
Job Title            object
Salary Estimate      object
Job Description      object
Rating              float64
Company Name         object
Location             object
Headquarters         object
Size                 object
Founded               int64
Type of ownership    object
Industry             object
Sector               object
Revenue              object
Competitors          object
```

DIFFERENT PLOTS

● Histograms for various fields that might contribute to the salary predictions.

● Correlation between these fields plotted using diverging palette

● Bar plots of different location and company textas and Headquarter.
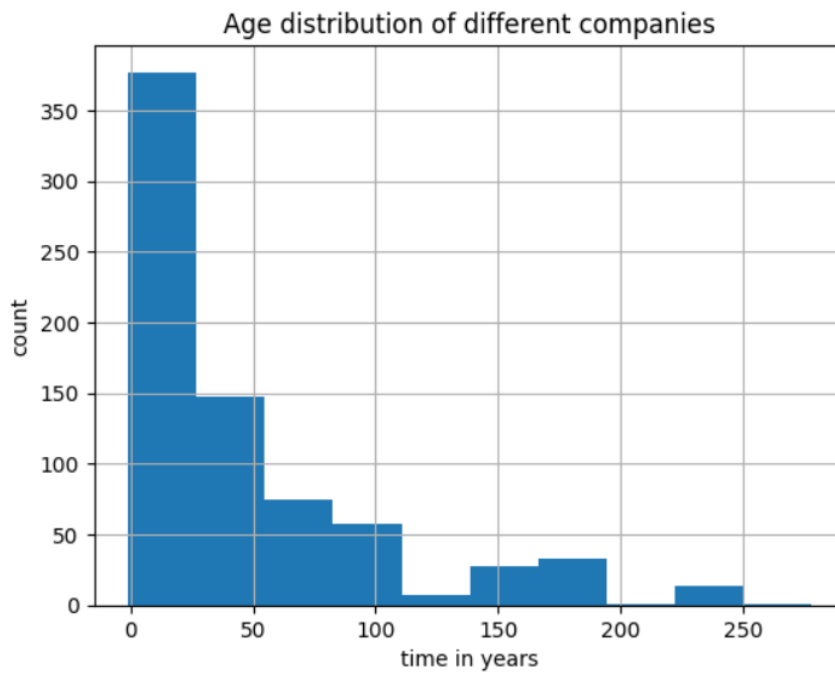
```
df['Rating'].hist()
plt.xlabel('Ratings')
plt.ylabel('count')
plt.title('Ratings of different companies')
```

Text(0.5, 1.0, 'Ratings of different companies')
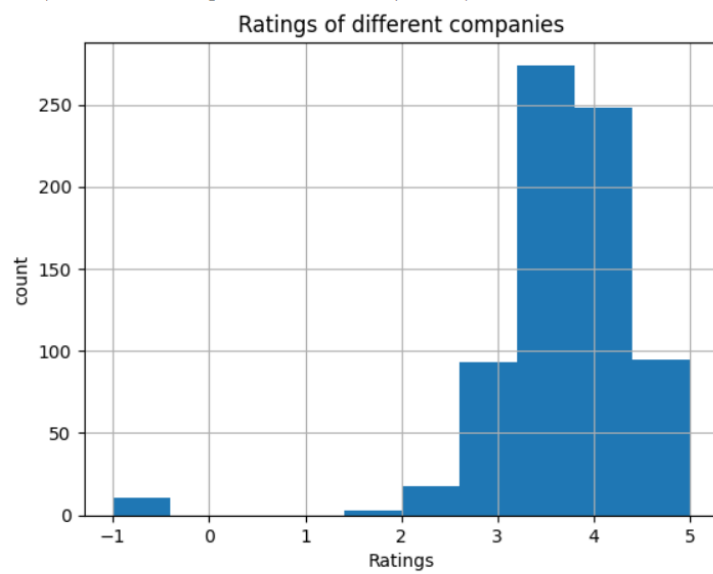
```
df['company_age'].hist()
plt.xlabel('time in years')
plt.ylabel('count')
plt.title('Age distribution of different companies')
```

Text(0.5, 1.0, 'Age distribution of different companies')


```
df['average-salary'].hist()
plt.xlabel('Annual Salary in $')
plt.ylabel('count')
plt.title('Histogram of Average Salary Histogram')
```

Text(0.5, 1.0, 'Histogram of Average Salary Histogram')
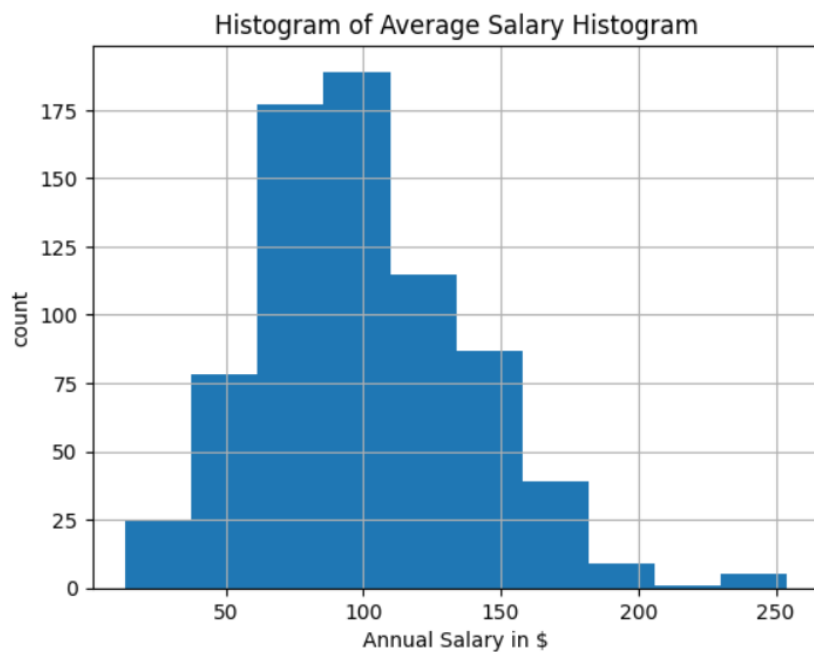
CORRELATION MAP

```python
correlation_matrix = df[['company_age', 'average-salary', 'Rating']].corr()
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```


Correlation Heatmap
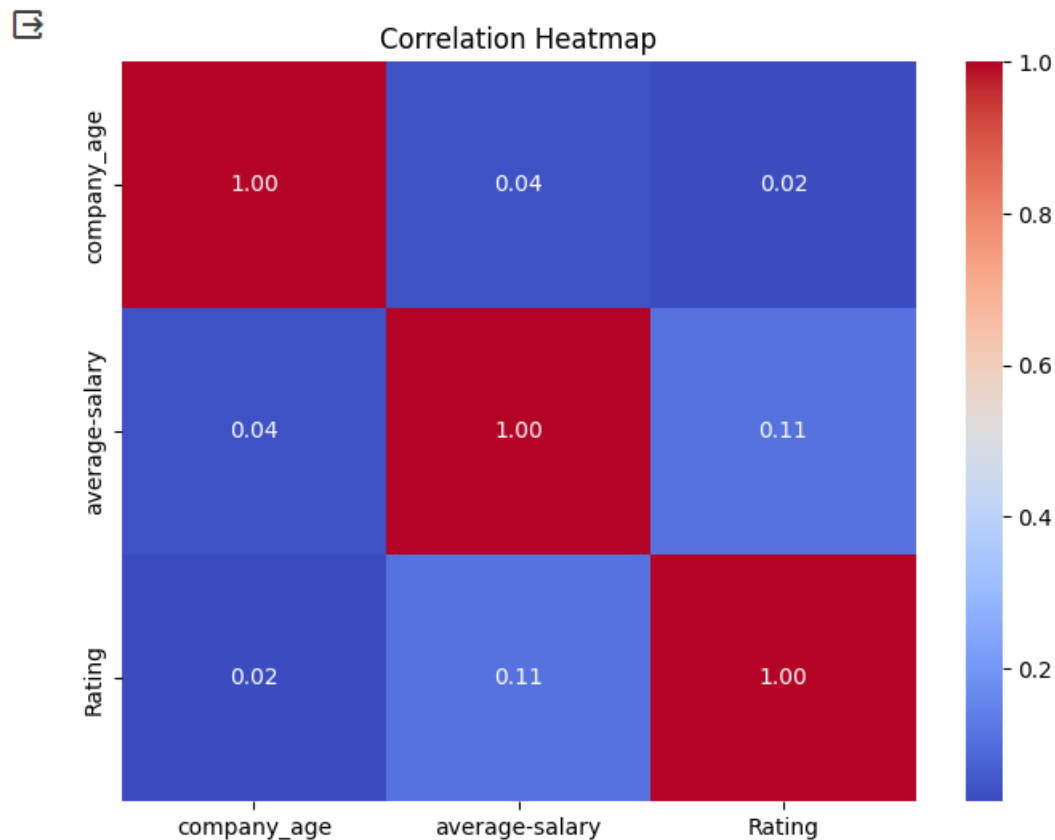
We must convert the categorical data to numerical before applying any imputation algorithms.We could use get dummies, but since most of the columns only have two categories, we'll use mapping instead. Why? Because there are just two categories, the two columns created after get_dummies will have a high degree of correlation because they both explain the same thing. In any event, one of the columns will have to be removed. As a result, we'll employ mapping for these columns. We'll use get dummies for columns with more than two categories.

```python
def ownership_simplifier(text):
    if 'private' in text.lower():
        return 'Private'
    elif 'public' in text.lower():
        return 'Public'
    elif ('-1' in text.lower()) or ('unknown' in text.lower()):
        return 'Other Organization'
    else:
        return text

df['Type of ownership'] = df['Type of ownership'].apply(ownership_simplifier)
```

```
[50] df['Type of ownership'].value_counts()
```

```
Private                          410
Public                           193
Nonprofit Organization            55
Subsidiary or Business Segment    34
Government                        15
Hospital                          15
College / University              13
Other Organization                 5
School / School District           2
Name: Type of ownership, dtype: int64
```

```
[51] df['Revenue'].value_counts()
```

```
Unknown / Non-Applicable             203
$10+ billion (USD)                   124
$100 to $500 million (USD)            91
$1 to $2 billion (USD)                60
$500 million to $1 billion (USD)      57
$50 to $100 million (USD)             46
$25 to $50 million (USD)              40
$2 to $5 billion (USD)                39
```

Salary is a continuous variable for which regression analysis can be used. By avoiding overfitting and selecting the most relevant features using regularisation algorithms, it is possible to do embedded feature selection and maybe improve the models.

```
[ ] df_model.head()
```

|	average-salary	company_rating	competitors	company_age	python_job	spark_job	aws_job	excel_job	job_in_headquarters	company_size_rank	...	job_title_simplified_mle	job_title_simplified_na	job_state_CA	job_state_IL	job_state_MA
0	72.0	3.8	0	49	1	0	0	1	0	4.0	...	0	0	0	0	0
1	87.5	3.4	0	38	1	0	0	0	0	NaN	...	0	0	0	0	0
2	85.0	4.8	0	12	1	1	0	1	1	4.0	...	0	0	0	0	0
3	76.5	3.8	3	57	1	0	0	0	1	5.0	...	0	0	0	0	0
4	114.5	2.9	3	24	1	0	0	1	1	2.0	...	0	0	0	0	0

5 rows × 48 columns
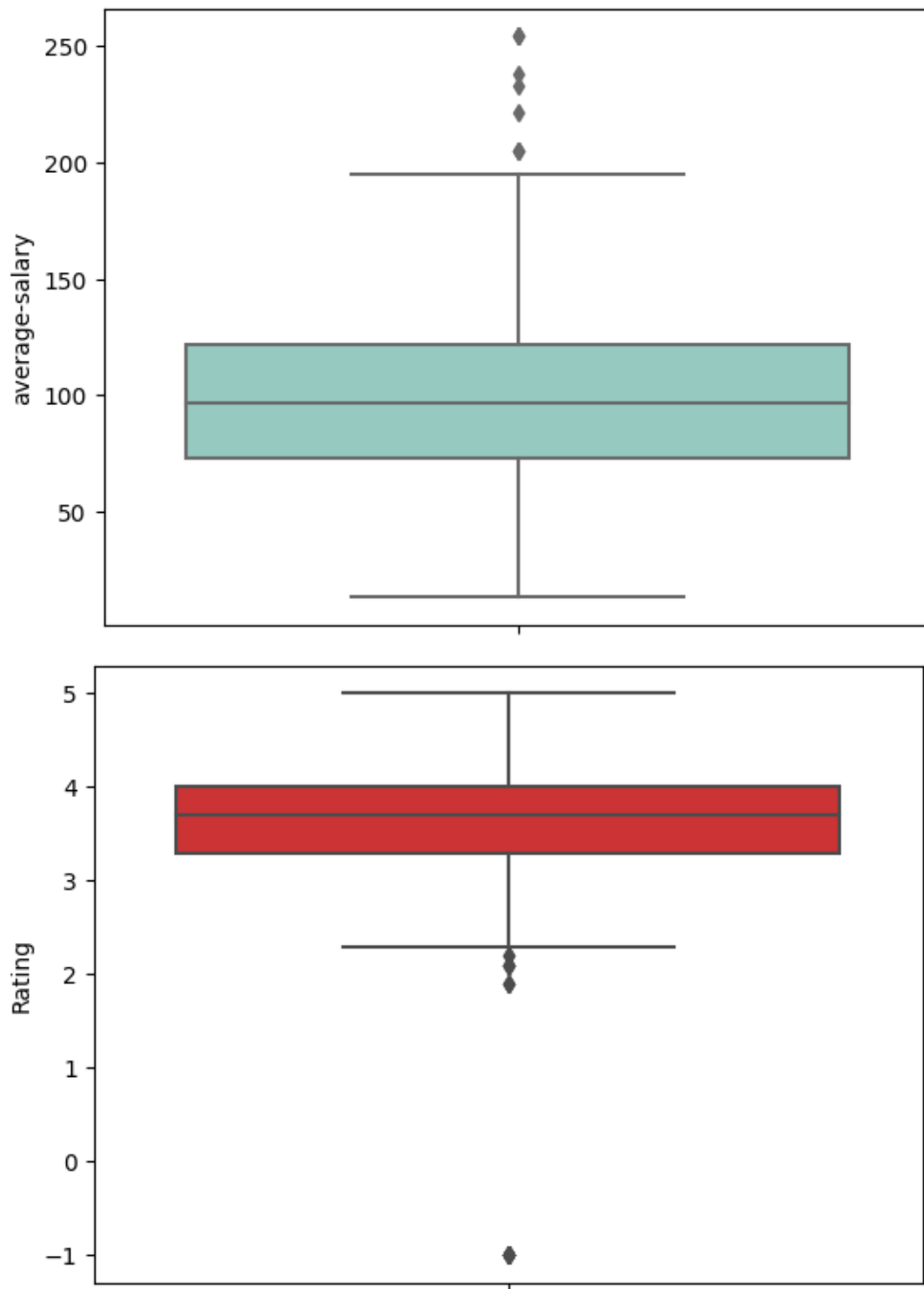
This graph is average salary and rating. We see this boxplot and identity some outliers so in average salary graph some value between 200 to 250 this is outliers and rating boxplot value is outliers is -1.
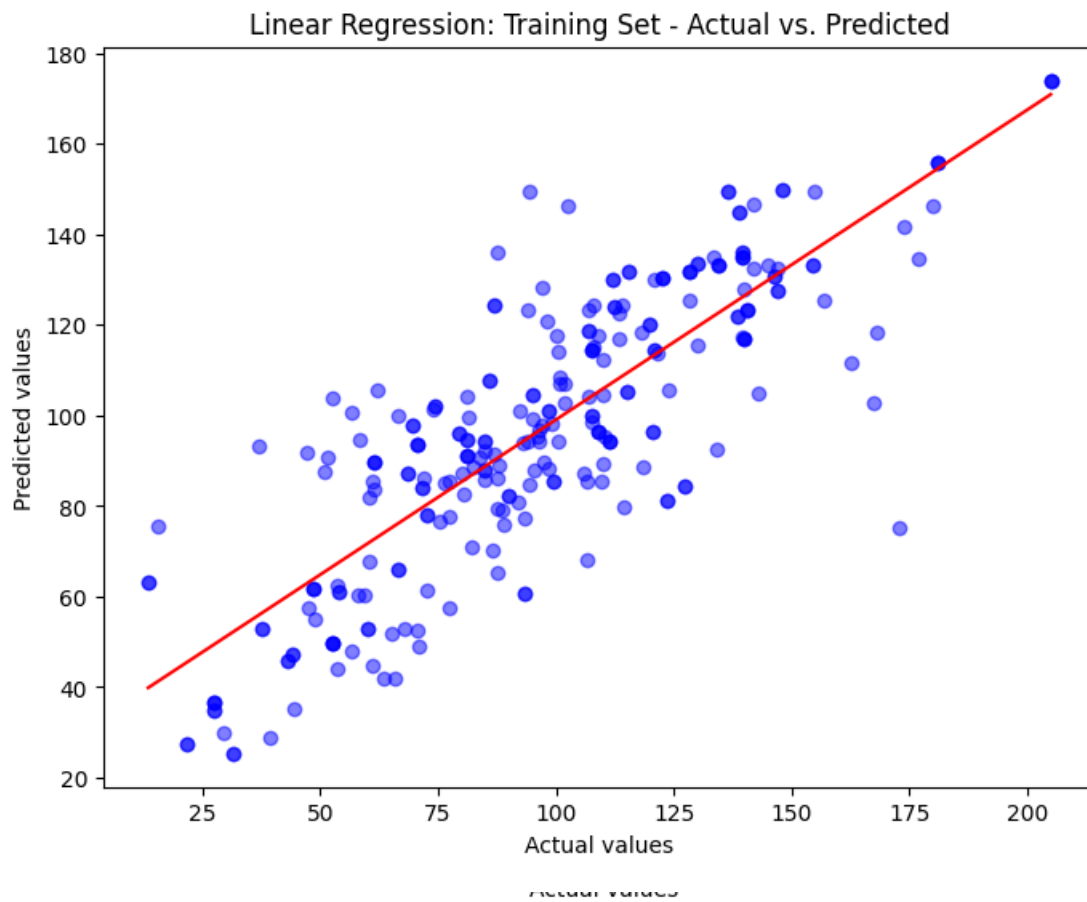
Top 15 States by Average Salary
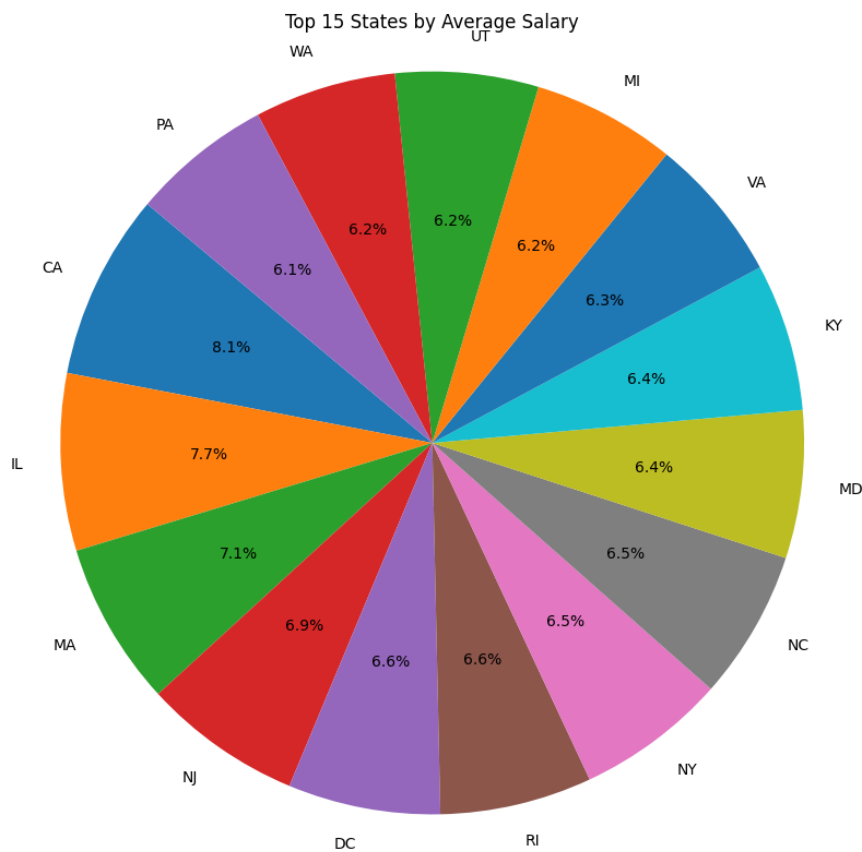
This diagram is top 15 states average salary of percentage.

Top 15 Industries by Average Salary
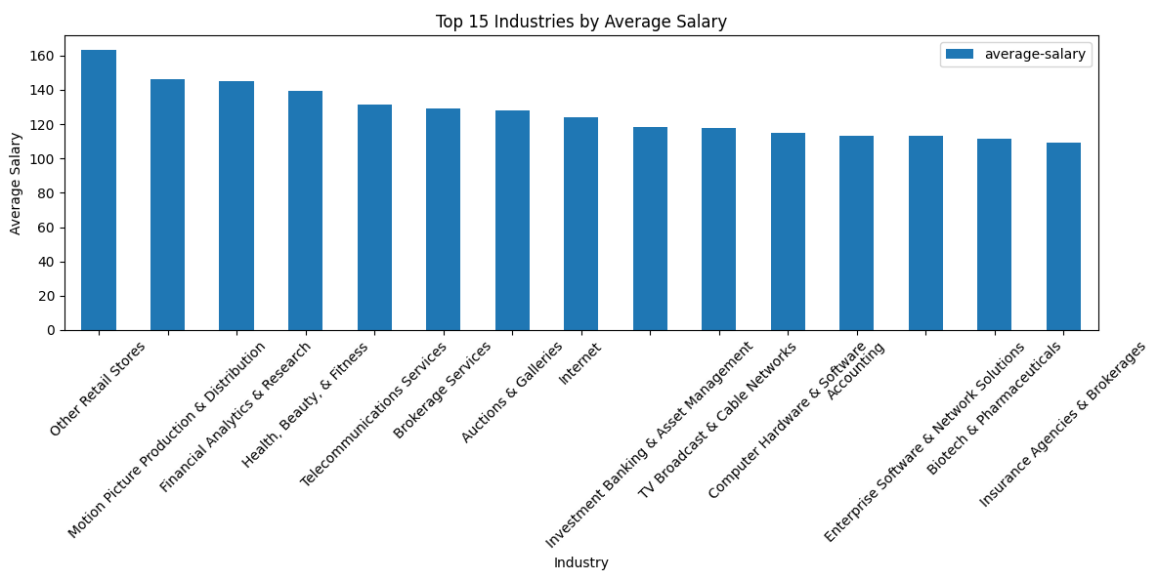
This graph is show top 15 industries by average salary.

Average Salary by Type of Ownership
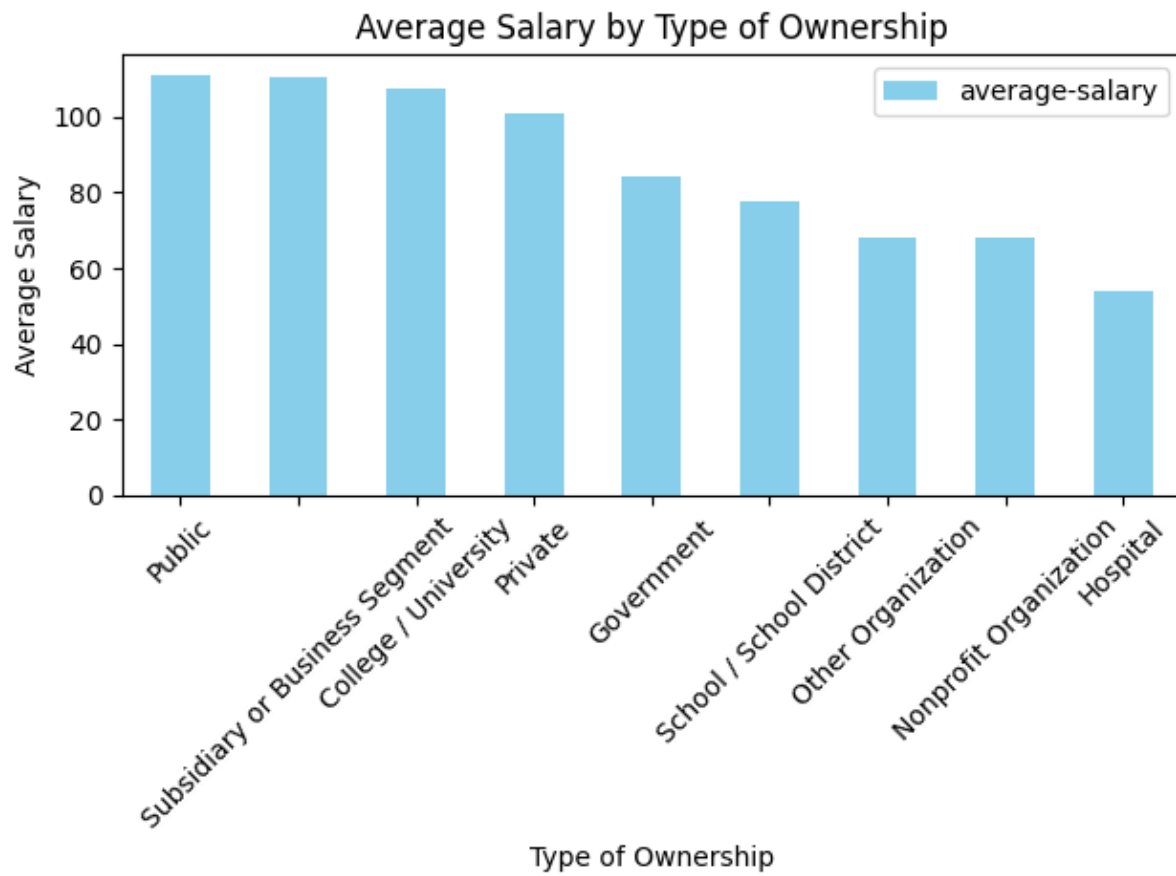
This graph is show average salary by type of ownership so type of ownership is public or private or government.

Linear Regression: Training Set - Actual vs. Predicted

Linear Regression: Test Set - Actual vs. Predicted

Building Model and Evaluation

For Regression Analysis :

After applying MSE on linear regression, Decision Tree Regression and Random Forest Regression, we found out Random Forest Regression is the best among the other the two.

```
Accuracy of the Linear Regression Model on Training set is: 85.76%
Accuracy of the Linear Regression Model on Test set is: 80.24%
```

```
Accuracy of the Decision Tree Regression Model on Training set is: 91.46%
Accuracy of the Decision Tree Regression Model on Test set is: 81.68%
```

```
Accuracy of the Random Forest Regression Model on Training set is: 95.39%
Accuracy of the Random Forest Regression Model on Test set is: 82.23%
```

CONCLUSION

The major goal of this project is to determine an Analysed them by graphs and tables, drawn insights and predicted salaries. To train and perform predictions using the ML model, Multiple linear regression and Random forest . The classification report is used as a comparison criteria to evaluate the overall efficiency of both algorithms once they have been imported. Random Forest outperforms the other THREE algorithms. The best prediction can then be picked.

References

https://github.com/goodluck08/practice_dataset/blob/main/glassdoor_jobs.csv

https://www.glassdoor.co.in/Salaries/data-scientist-salary-SRCH_KO0,14.htm