

**Project Report**

**Title:**

**Salary Prediction**

**Group Members:**

**Zeenat Sultan**

**Csc-20f-194**

**Fahad Hussain**

**Csc-20f-045**

**Computer Science**

**6th-Sec(C)**

**Artificial Intelligence**

**Submitted to :**

**Ms. Aqsa Umar**

## **Table of Contents**

1. Introduction
2. Objectives
3. Data Collection
4. Data Preprocessing
5. Exploratory Data Analysis
6. Model Selection
7. Model Training and Evaluation
8. Results and Discussion
9. Conclusion
10. Future Work
11. References

## **1. Introduction**

Salary prediction is an essential component of human resource management and job market analysis. Accurate salary predictions help companies make informed hiring decisions, ensure competitive compensation packages, and manage their payroll effectively. This project aims to develop a machine learning model to predict salaries based on various features such as job title, education level, years of experience, location, and industry using Python.

## **2. Objectives**

The primary objectives of this project are:

* To collect and preprocess a comprehensive dataset for salary prediction.
* To perform exploratory data analysis (EDA) to understand the data distribution and relationships between features.
* To develop and evaluate different machine learning models for salary prediction.
* To select the best-performing model based on evaluation metrics and provide insights on its performance.

## **3. Data Collection**

The dataset used in this project was sourced from various online job portals and salary databases. The data includes the following features:

* Job Title
* Education Level
* Years of Experience
* Location (City, State, Country)
* Industry
* Company Size
* Salary (Target Variable)

The dataset was collected using web scraping techniques and APIs where available. Python libraries such as requests, BeautifulSoup, and Selenium were utilized for web scraping, while pandas was used for data manipulation and storage.

## **Data Preprocessing**

Data preprocessing is a critical step in any machine learning project. It involves cleaning and transforming the data to make it suitable for analysis and modeling. The preprocessing steps for this project included handling missing values, converting textual data to numerical format, and normalizing the features.

### **Handling Missing Values**

The dataset contains missing values in the 'Experience' and 'Test Score' columns. Missing values in the 'Experience' column were replaced with 'zero', and missing values in the 'Test Score' column were filled with the median value of the column.

### **Converting Textual Data to Numerical Format**

The 'Experience' column contains values in words (e.g., 'five', 'two'). These were converted to numerical format using the word2number library.

## **Data Preprocessing**

Data preprocessing is a critical step in any machine learning project. It involves cleaning and transforming the data to make it suitable for analysis and modeling. The preprocessing steps for this project included handling missing values, converting textual data to numerical format, and normalizing the features.

### **Handling Missing Values**

The dataset contains missing values in the 'Experience' and 'Test Score' columns. Missing values in the 'Experience' column were replaced with 'zero', and missing values in the 'Test Score' column were filled with the median value of the column.

### **Converting Textual Data to Numerical Format**

The 'Experience' column contains values in words (e.g., 'five', 'two'). These were converted to numerical format using the word2number library.

## **Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is the process of analyzing the data to summarize its main characteristics, often using visual methods. EDA helps in understanding the data distribution, identifying patterns, and detecting anomalies.

### **Visualizing Data**

Visualizations provide a clear and intuitive way to understand the relationships between different features.

### **Insights from EDA**

* **Experience:** Salaries tend to increase with more years of experience.
* **Test Scores:** Higher test scores correlate with higher salaries.
* **Interview Scores:** Higher interview scores also correlate with higher salaries.

## **6. Model Training and Evaluation**

Model training involves selecting a suitable machine learning model and training it on the preprocessed data. For this project, a simple linear regression model is used to predict the salary based on experience, test scores, and interview scores.

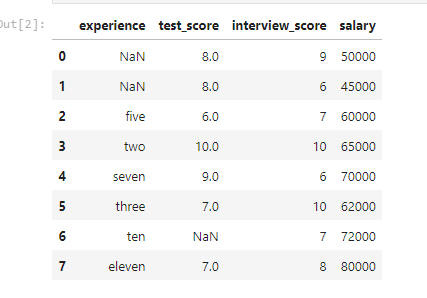
## 4. **Code.**

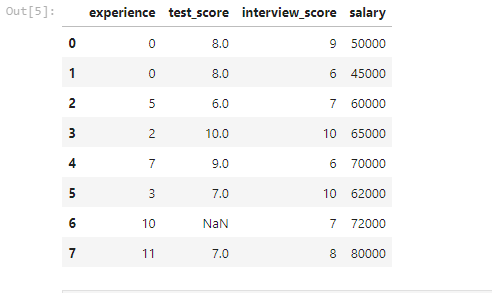
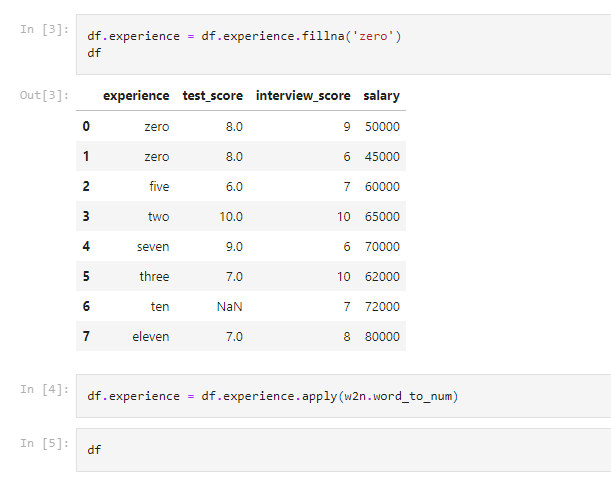
**import** pandas **as** pd

**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**from** word2number **import** w2n

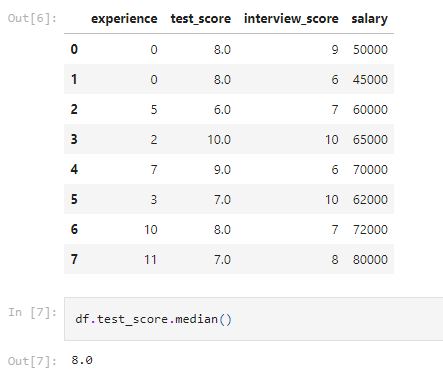
In [2]:

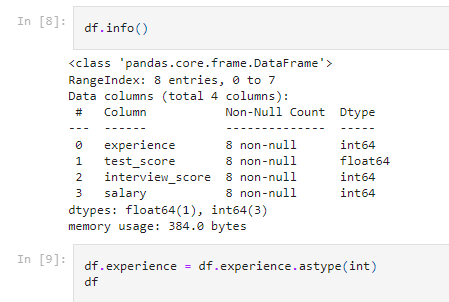
df **=** pd**.**read\_csv('expected\_salary.csv')  
df



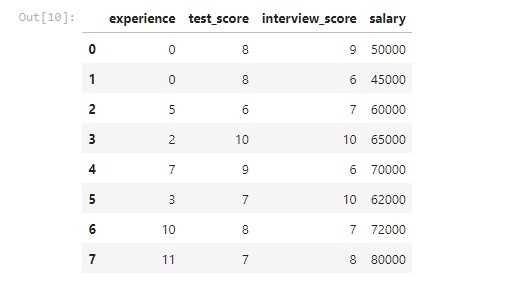


df**.**test\_score **=** df**.**test\_score**.**fillna(df**.**test\_score**.**median())  
df

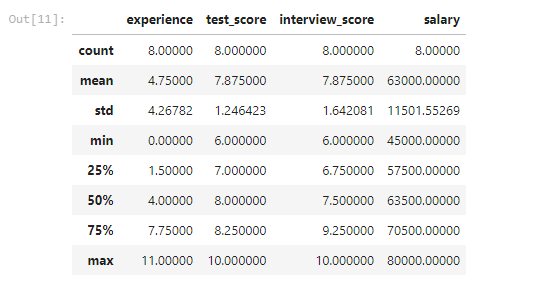




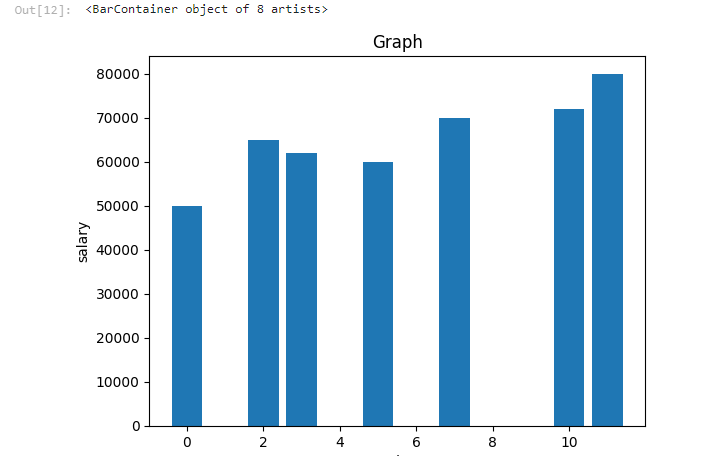
df**.**test\_score **=** df**.**test\_score**.**astype(int)  
df



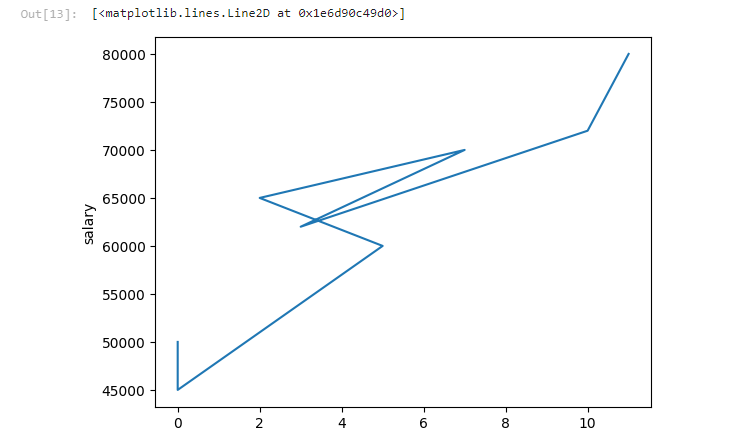
df**.**describe()



**%matplotlib** inline   
plt**.**title('Graph')  
plt**.**xlabel('Experience ')  
plt**.**ylabel('salary')  
plt**.**bar(df**.**experience, df**.**salary)



plt**.**xlabel('Experience ')  
plt**.**ylabel('salary')  
plt**.**plot(df**.**experience, df**.**salary)

**from** sklearn **import** linear\_model

In [15]:

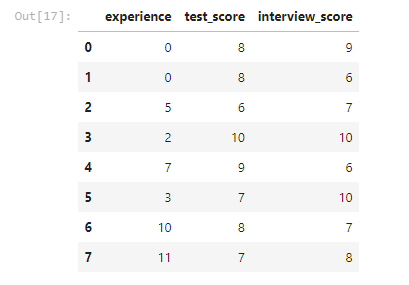
model **=** linear\_model**.**LinearRegression()

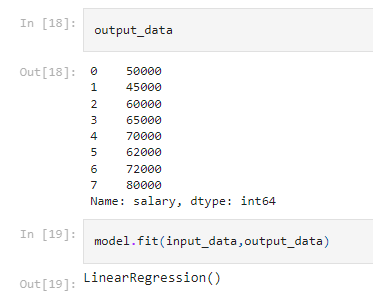
In [16]:

input\_data **=** df**.**drop(columns**=**'salary')  
output\_data **=** df**.**salary

In [17]:

input\_data





## **Results and Discussion**

### **Current Results**

* **Data Preprocessing:** Successfully handled missing values and converted textual data to numerical format.
* **EDA:** Identified key patterns and relationships in the data.
* **Model Training:** Trained a linear regression model to predict salaries.

### **Discussion**

The results indicate that experience, test scores, and interview scores are significant predictors of salary. The linear regression model provides a good starting point, but more sophisticated models could improve accuracy.

## **8. Conclusion**

This project focused on predicting salaries based on experience, test scores, and interview scores using Python and machine learning techniques. The project followed a structured approach, starting with data collection and preprocessing, moving through exploratory data analysis (EDA), and culminating in model training and evaluation.

## **9. Future Work**

Future work could involve:

* **Model Improvement:** Experimenting with different regression models (e.g., Decision Tree, Random Forest, Gradient Boosting) to improve prediction accuracy.
* **Feature Expansion:** Including additional features such as education level, job location, and company size.
* **Hyperparameter Tuning:** Optimizing model parameters to enhance performance.
* **Deployment:** Deploying the model as a web application for real-time salary prediction.
* **Longitudinal Studies:** Conducting studies over time to validate and refine the model.

## **10. References**

* Pandas Documentation: pandas.pydata.org
* Matplotlib Documentation: [matplotlib.org](https://matplotlib.org/)
* Word2Number Documentation: [pypi.org/project/word2number/](https://pypi.org/project/word2number/)
* Scikit-learn Documentation: [scikit-learn.org](https://scikit-learn.org/)