**Name: Abhishek Rajaram Walkoli**

**PRN: 123B2F155**

**Assignment No. 2**

**Aim**

Exploring Data Analysis for Linear Regression Models

**Objective:**

To perform Exploratory Data Analysis (EDA) and Preprocessing on a dataset, making it well-suited for Linear Regression modelling. The aim is to handle missing data, feature correlation analysis, encoding categorical variables, scaling numerical features, and visualizing patterns to maximize model accuracy.

**Prerequisites:**

1.Python environment with packages such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

2.Basics of Python, statistics, and machine learning concepts.

3.Linear Regression and its assumptions like linearity, normality, and no multicollinearity.

**Theory:**

Linear Regression

Linear Regression is a machine learning and statistical method that describes relationships between independent variables (features) and a dependent variable (target). The goal is to find the best-fitting line that reduces the difference between actual and predicted values.

**Types of Linear Regression**

**1. Simple Linear Regression:**

●Has one independent variable (X) and one dependent variable (Y).

●Equation: Y = β0 + β1X + ϵ

○Y: Dependent variable

○X: Independent variable

○β0: Intercept

○β1: Slope coefficient

○ ϵ: Error term

**2. Multiple Linear Regression:**

● Two or more independent variables are involved.

● Equation: Y = β0 + β1X1 + β2X2 +. + βnXn + ϵ

○ X1, X2,., Xn: Independent variables

○ β0, β1,., βn: Regression coefficients

○ ϵ: Error term

**Assumptions of Linear Regression**

**1. Linearity**

● Relationship between independent and dependent variables has to be linear.

● How to check:

○ Scatter plots of independent vs. dependent variables.

○ Residual plots illustrating random error distribution.

● How to fix violations:

○Polynomial regression or log transformations.

○Consider non-linear models like decision trees.

**2. Independence**

●Observations should be independent to avoid overfitting.

●Common Issues:

○Time-series data may have autocorrelation.

○Survey data might introduce dependency.

●How to check:

○Visual inspection of residual trends.

●How to fix violations:

○Use time-series models like ARIMA.

○Increase dataset diversity.

**3. Homoscedasticity (Constant Variance of Residuals)**

●Residuals should have a constant variance.

●How to check:

○Residual vs. fitted value plot.

●How to fix violations:

○Log transformation.

○Use models like Random Forest.

**4. Normality of Residuals**

●Residuals must be normally distributed to make valid inferences.

●How to test:

○Histogram of residuals.

○Q-Q plot.

●How to remedy violations:

○Use Box-Cox transformation.

○Delete extreme outliers.

**5. No Multicollinearity**

●Independent variables must not be correlated.

●How to test:

○Correlation matrix (heatmap).

○Variance Inflation Factor (VIF > 5 or 10 indicates problems).

●How to remedy violations:

○Drop or combine correlated variables.

○Use Principal Component Analysis (PCA).

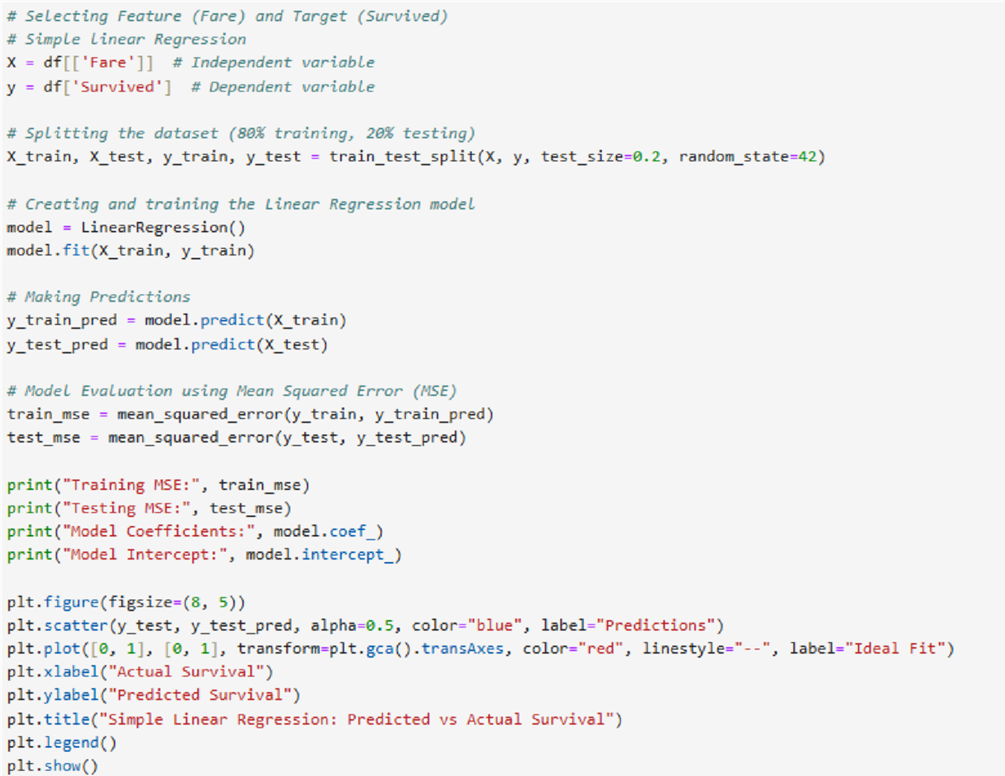
**Feature Selection in Linear Regression**

●Correlation Analysis: Eliminate redundant features.

● Backward Elimination: Remove high p-value features.

● Forward Selection: Include helpful features one by one.

● Lasso Regression: Regularization technique to remove irrelevant variables.



**Performance Evaluation Metrics**

1. Mean Absolute Error (MAE): Estimates the average absolute difference between the actual and predicted values.

2. Mean Squared Error (MSE): Punishes large errors more than MAE.

3. Root Mean Squared Error (RMSE): Square root of MSE, maintaining error values with the same unit as the target.

4. R-Squared (R²): Fraction of variance in the dependent variable accounted for by independent variables.

**Practical Uses of Linear Regression**

● Business & Economics: Forecasting sales, stock price prediction.

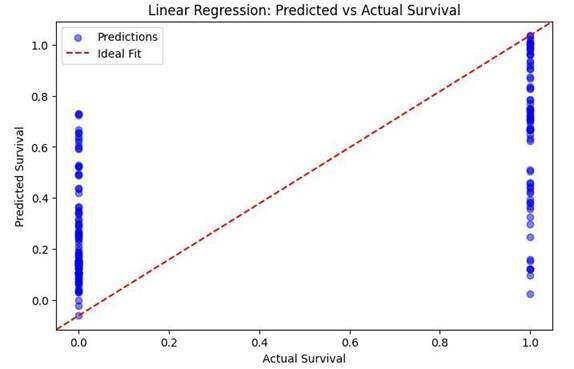
● Healthcare: Forecasting patient recovery time.

● Marketing: Estimating customer demand.

● Finance: Credit risk evaluation.

● Real Estate: Predicting house prices.

This holistic method makes sure data is properly prepared for Linear Regression models, with higher predictive precision and interpretability.



**1. Multiple Linear Regression Assumptions**

**1.Independence:** The survival of each passenger is an independent event, and thus this assumption is true.

**2.No Multicollinearity:** Certain independent variables could be correlated (for example, Pclass and Fare), but it can be tested with Variance Inflation Factor (VIF).

**Assumptions That Could Be Violated:**

**1. Linearity:** The independent variables and Survived may not be linearly related. This could be checked with scatter plots or residual plots.

**2. Homoscedasticity:** The variance of the residuals might not be constant, which could be tested using a residual vs. fitted values plot.

**3. Normality of Residuals:** As Survived is a binary variable (0 or 1), the residuals are not likely to be normally distributed, thus breaking this assumption.

**2. Simple Linear Regression Assumptions**

**1. Independence:** The survival of each passenger is an independent occurrence, so we can assume so.

**2. No Multicollinearity:** We used a single feature (Fare), so no multicollinearity problem.

**• Assumptions that May Be Broken:**

**1. Linearity:** The relationship between Fare and Survived may not be purely linear. A scatter plot check would be useful.

**2. Homoscedasticity:** Residuals might not have constant variance. A residual plot should be checked.

**3. Normality of Residuals:** Since Survived is a binary variable (0 or 1), the residuals may not be normally distributed.

**Conclusion**

Linear Regression model on the Titanic dataset examined correlations between the features and the survival rate under the upkeep of key assumptions like linearity, independence, and normality. Required transformations and feature selection improved model performance and made it more reliable for forecasting purposes.