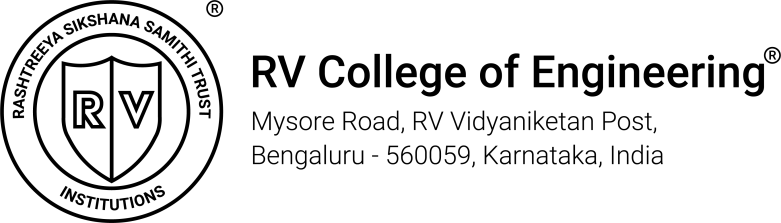
A close-up of a white background

Description automatically generated



**RV COLLEGE OF ENGINEERING®**

**(Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi)**

**Linear Regression on**

**Automobile Dataset**

**ARTIFICIAL INTELLIGENCE & MACHINE LEARNING**

**22MCE1A1T**

**PROJECT REPORT**

Submitted by

**Name (USN)**

Under the guidance of

**Prof. Name**

Assistant Professor,

Department of Computer Science and Engineering

RV College of Engineering

Bengaluru - 560059

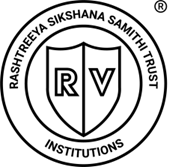
***Submitted in partial fulfillment of the requirements for the award of degree of***

**MASTER OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

[**Bengaluru**](http://www.bengaluruairport.com/)**– 560059**



**CERTIFICATE**

Certified that the project work titled **“Linear Regression on Automobile Dataset”** carried out by **Name (USN)** bona fide student of **RV College of Engineering, Bengaluru** submitted in partial fulfillment for the award of **Master of Technology** in **Computer Science Engineering** of **RV College of Engineering, Bengaluru affiliated to Visvesvaraya Technological University, Belagavi** during the year **2023-24**. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirement in respect of the project work prescribed for the said degree.

**Prof. Name**

Assistant Professor, Department of Computer Science Engineering, RVCE, Bengaluru-560059

Table of Contents

[CHAPTER 1 4](#_TOC_250040)

[INTRODUCTION 4](#_TOC_250039)

* 1. [METHODOLOGY 4](#_TOC_250038)

[CHAPTER 2 6](#_TOC_250034)

[LINEAR REGRESSION 6](#_TOC_250033)

* 1. TYPES OF LINEAR REGRESSION 6
  2. ASSUMPTIONS OF LINEAR REGRESSION 7
  3. MODEL EVALUATION 8

[CHAPTER 3 10](#_TOC_250020)

[LINEAR REGRESSION ON AUTOMOBILE DATASET 10](#_TOC_250019)

* 1. [SETTING UP THE ENVIRONMENT 10](#_TOC_250018)
  2. DATA PREPROCESSING 11
  3. ASSUMPTIONS OF LINEARITY 12
  4. APPLYING LINEAR REGRESSION 13

[CHAPTER 4 25](#_TOC_250016)

[CONCLUSION 25](#_TOC_250015)

* 1. [CONCLUSION 25](#_TOC_250014)

**Chapter 1: INTRODUCTION**

Linear regression is a foundational statistical technique widely employed in data science for modeling relationships between variables. In this project, we delve into the practical application of linear regression through hands-on analysis of a real-world dataset. By leveraging the power of Python programming and essential libraries such as NumPy, Pandas, Statsmodels, Matplotlib, Seaborn, and Scikit-learn, we aim to gain valuable insights into the dynamics of the data and derive meaningful predictions through regression modeling.

**Methodology:**

By the following methodology, we aim to demonstrate the practical application of linear regression in data science and derive actionable insights from the analysis of the dataset.

*Setting Up the Environment:* The initial step involves setting up the project environment with essential Python libraries. We rely on a comprehensive mix of libraries, each serving a specific purpose, such as data manipulation, visualization, and statistical modeling. This ensures that we have the necessary tools to handle, process, and analyze the dataset effectively.

*Data Exploration:* We begin by loading and exploring the dataset to gain insights into its structure and key characteristics. Through preliminary exploration, we identify potential areas of interest and formulate hypotheses for further analysis.

*Preprocessing and Analysis:* Data preprocessing is essential to ensure the quality and suitability of the data for regression analysis. We clean the data, handle missing values, and transform variables as needed to prepare them for modeling. Subsequently, we build our linear regression model, selecting relevant variables, checking for multicollinearity, and ensuring that the data meets the assumptions of linear regression.

*Dealing with Outliers:* Outliers can significantly affect the performance of a regression model, leading to biased estimates and reduced predictive accuracy. We employ strategies to detect and mitigate outliers, such as removing extreme data points or transforming variables to achieve a more balanced distribution.

*Regression Modeling:* With the preprocessed data, we proceed to build our linear regression model using Scikit-learn. We partition the data into training and testing sets, train the model on the training data, and evaluate its performance on unseen test data. Through regression analysis, we aim to predict outcomes based on the input variables and assess the model's accuracy and reliability.

*Model Evaluation*: We evaluate the performance of the regression model using various metrics, such as the coefficient of determination (𝑅2), mean squared error (MSE), and residual analysis. This allows us to gauge how well the model fits the data and make informed decisions about its effectiveness in predicting outcomes.

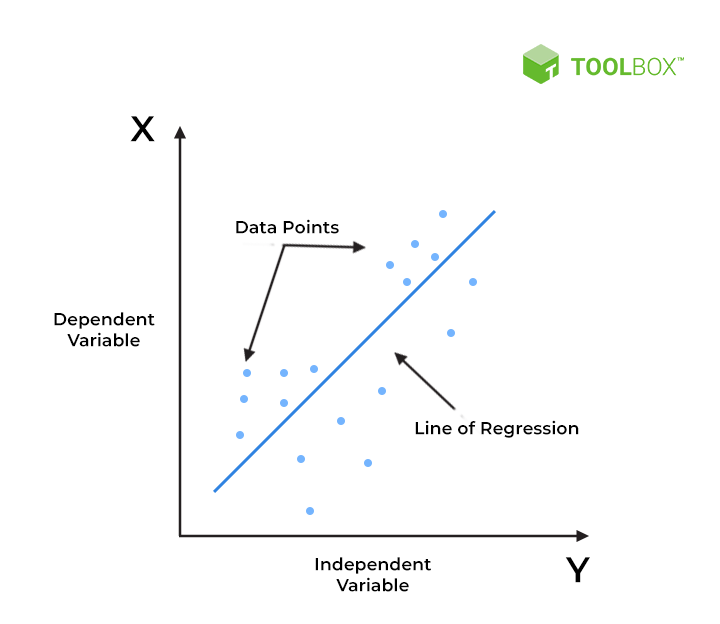
*Interpretation and Conclusion:* Finally, we interpret the results of the regression analysis and draw conclusions based on the insights gained from the model. We discuss the implications of our findings, potential areas for further exploration, and the practical applications of linear regression in real-world scenarios.

## Chapter 2: LINEAR REGRESSION

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the independent variables and the dependent variable, where the dependent variable can be predicted as a linear combination of the independent variables, plus an error term.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output.



**TYPES OF LINEAR REGRESSION:**

***Simple Linear Regression***

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

***Multiple Linear Regression:***

Multiple Linear Regression extends the concept of simple linear regression by allowing for more than one independent variable to predict the dependent variable. In MLR, the relationship between the dependent variable 𝑦 and multiple independent variables x, y, z is modelled using y = a + bx + cy + dz.

The MLR model estimates the coefficients​ that best fit the observed data. These coefficients provide insights into the strength and direction of the relationships between the independent variables and the dependent variable. For example, a positive coefficient suggests that an increase in the corresponding independent variable is associated with an increase in the dependent variable while a negative coefficient suggests the opposite.

MLR assumes that the relationship between the dependent variable and the independent variables is linear, meaning that the change in the dependent variable is proportional to the change in each independent variable. Additionally, MLR assumes that the observations are independent of each other, the variance of the error terms is constant across all levels of the independent variables (homoscedasticity), and the error terms are normally distributed.

**ASSUMPTIONS MADE IN LINEAR REGRESSION:**

***Linearity***: The assumption of linearity in linear regression signifies that the relationship between the dependent variable and the independent variables can be adequately represented by a straight line. This implies that changes in the independent variables result in proportional changes in the dependent variable. When examining scatter plots of the variables, linearity can be visually assessed. If the relationship appears to be curved, it may indicate violations of the linearity assumption, necessitating transformations or consideration of alternative modeling techniques.

***Independence***: Independence assumes that the observations used in the regression analysis are not influenced by each other and are, therefore, independent. The value of one observation should not be dependent on or influenced by the value of another observation. This assumption is crucial to avoid bias and ensure that the regression coefficients are estimated accurately. Violations of independence may occur in longitudinal or time-series data, where observations over time may exhibit autocorrelation, meaning that the value of one observation is correlated with the value of a previous observation.

***Homoscedasticity***: Homoscedasticity refers to the uniformity of variance in the error terms across all levels of the independent variables. In simpler terms, it means that the spread or dispersion of the residuals (the differences between the observed and predicted values) remains constant across the range of the independent variables. Homoscedasticity is typically assessed by examining residual plots, where the residuals are plotted against the predicted values or independent variables. If the spread of the residuals appears to change systematically with the predicted values or independent variables, it suggests heteroscedasticity. To address this, transformations of the dependent variable or the use of robust standard errors may be considered.

***Normality***: The assumption of normality pertains to the distribution of the error terms in the regression model. It assumes that the error terms follow a normal distribution with a mean of zero. In practical terms, this means that the residuals should be symmetrically distributed around zero, resembling a bell-shaped curve when plotted. Departures from normality may manifest as skewness or kurtosis in the distribution of residuals. Techniques such as transformation of variables or the use of robust standard errors can help mitigate the effects of non-normality in the error terms.

**MODEL EVALUATION:**

Model evaluation is the process of assessing the performance and effectiveness of a predictive model in terms of its ability to accurately capture the underlying patterns in the data and make reliable predictions. It involves using various metrics and techniques to measure how well the model fits the observed data and how accurately it predicts outcomes for new, unseen data. The goal of model evaluation is to determine whether the model meets the requirements and expectations of the problem at hand and to identify areas for improvement if necessary. It helps to ensure the reliability, validity, and generalizability of the model's predictions and enables stakeholders to make informed decisions based on the model's outputs.

***Coefficient of Determination*** *(R^2)*: The coefficient of determination, often denoted as R^2, measures the proportion of the variance in the dependent variable that is explained by the independent variables in the regression model. It ranges from 0 to 1, where 0 indicates that the model does not explain any of the variability in the dependent variable, and 1 indicates that the model explains all of the variability. R^2 can be interpreted as the proportion of the total variation in the dependent variable that is accounted for by the regression model. A higher R^2 value indicates a better fit of the model to the data, although it does not necessarily imply a more predictive model.

***Mean Squared Error*** *(MSE)*: Mean squared error is a measure of the average squared difference between the observed values of the dependent variable and the values predicted by the regression model. It provides a quantitative assessment of the model's accuracy in predicting the dependent variable. The MSE is calculated by taking the average of the squared residuals (the differences between the observed and predicted values) across all observations. Lower values of MSE indicate better predictive performance, as they indicate that the model's predictions are closer to the observed values on average.

***Root Mean Squared Error*** *(RMSE)*: Root mean squared error is the square root of the mean squared error. It is a measure of the standard deviation of the residuals and provides a more interpretable metric of the model's predictive accuracy, as it is expressed in the same units as the dependent variable. Like MSE, lower values of RMSE indicate better predictive performance, as they signify smaller discrepancies between the observed and predicted values of the dependent variable.

### 

### Chapter 3: Linear Regression on Automobile Dataset

### SETTING UP THE ENVIRONMENT

Before delving into the data, it’s crucial to set up our environment with the right tools. We used a variety of Python libraries, each serving a specific purpose. This eclectic mix of libraries forms the backbone of our data analysis, allowing us to handle, process, and visualize data effectively:

NumPy: Essential for numerical operations.

Pandas: Perfect for data manipulation and analysis.

Statsmodels: Provides classes and functions for the estimation of statistical models.

Matplotlib and Seaborn: Our go-to libraries for data visualization.

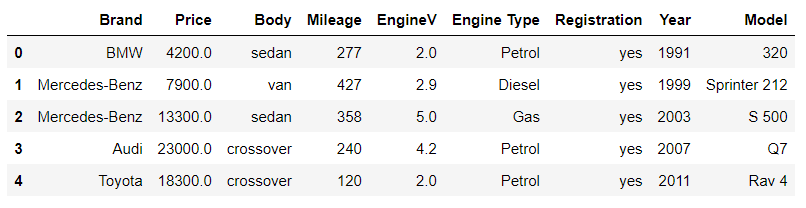
Scikit-learn: A comprehensive library for machine learning, including linear regression models.

A screenshot of a computer code

Description automatically generated

**DATA EXPLORATION**

The first step in our analysis was to load and explore the dataset. We are using the automobile dataset that contains the description as shown below. The following snippet shows the sample rows of the dataset:



The following snippet shows the description of the columns data:

A screenshot of a computer screen

Description automatically generated

**DATA PREPROCESSING**

Data preprocessing is a vital step in any data science project. We clean the data, handled missing values, and transformed variables to prepare them for analysis before we embark on building our linear regression model.

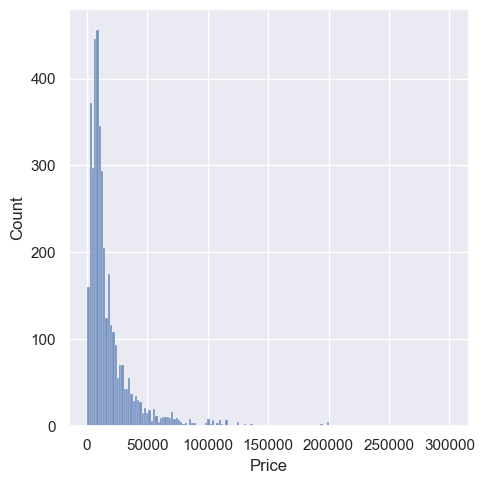
***Dealing with missing values:***

When examining the count descriptions for all columns, variations in the number of entries become apparent, indicating the presence of missing values in certain rows. We observe that the variables such as ‘Price’ and ‘EngineV’ contain a small proportion of missing values. Given that these missing entries constitute less than 5% of the entire dataset, a practical approach is to remove them.

***Dealing with outliers:***

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In the provided data, numerical information is present in the ‘Price,’ ‘Mileage,’ ‘EngineV,’ and ‘Year’ columns. Since ‘Year’ is a datetime, there is no outlier in it. So we skip it and examine the probability density functions (PDFs) for ‘Price,’ ‘Mileage,’ and ‘EngineV’ to gain insights into their distributions and identify any outliers that may impact our analysis.

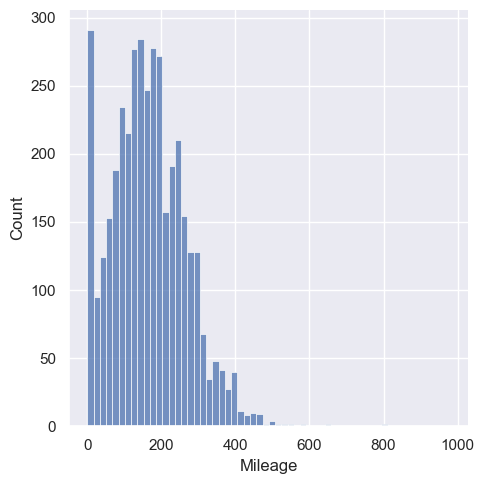
*For ‘Price’:*



Observing a pronounced right skew in the distribution of the ‘Price’ variable and the statistic description, we note the presence of outliers, particularly at the higher price range. Given their potential influence on regression analysis and central tendency measures, a prudent approach is to mitigate their impact.

To achieve a more normal distribution, a strategic step involves removing the top 2% of the highest-priced data points. This targeted adjustment not only aligns the data with a more symmetrical distribution but also fosters a robust foundation for regression analysis, ensuring more accurate and reliable modeling results.

*For ‘Mileage’:*



Just like the price, we can see outliers in Mileage at the higher values. Lets remove the top 1% of the values which are present in the right side.

*For ‘EngineV:*

A graph with numbers and symbols

Description automatically generated

The recorded sources indicate that the maximum engine displacement for cars typically reaches 6.5 liters. Consequently, a prudent course of action involves excluding data points where ‘EngineV’ exceeds this realistic threshold. By filtering out instances where engine displacement is mentioned as more than 6.5 liters, we ensure a more accurate and credible dataset for analysis, aligning with industry standards and enhancing the reliability of our findings.

**ASSUMPTIONS OF LINEARITY**

In order to evaluate the model, we are utilizing Ordinary Least Squares. OLS method is widely used to estimate the parameter of a linear regression model. OLS estimators minimize the sum of the squared errors, a difference between observed values and predicted values. While OLS is computationally feasible and can be easily used while doing any econometrics test, it is important to know the underlying assumptions of OLS regression.

***Assumption: The linear regression model is “linear in parameters”***

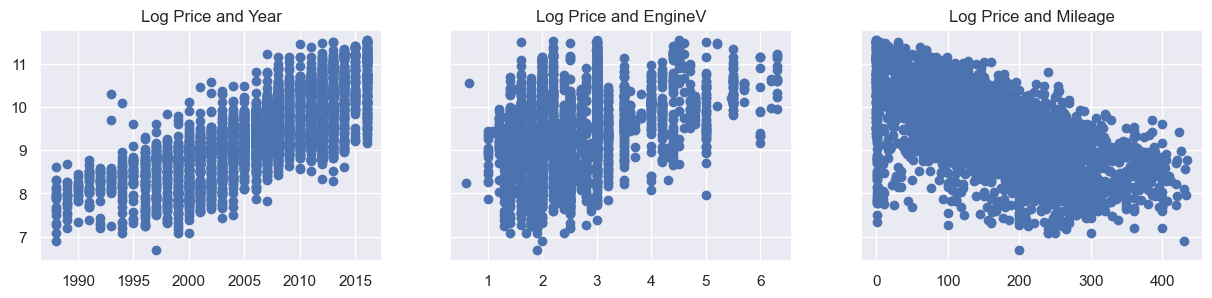
Let us check how the price is distrubuted with the other numberical columns “Year”, “EngineV” and “Mileage”

A graph with blue dots

Description automatically generated

The subplots and probability density function (PDF) analysis vividly indicate that the distribution of ‘Price’ follows an exponential pattern, deviating from the linear assumption of Ordinary Least Squares (OLS).

To address this, a practical strategy is to apply a log transformation to ‘Price.’ This transformation mitigates the exponential distribution, aligning the variable with the linearity assumption essential for OLS regression.



***Assumption: There is no multi-collinearity***

Multicollinearity in linear regression occurs when predictors are highly correlated, inflating standard errors and hindering coefficient interpretation. This phenomenon compromises the model’s stability and the ability to discern individual predictor impacts. Detection involves assessing correlation or Variance Inflation Factor. Strategies to mitigate multicollinearity include removing or combining correlated predictors, ensuring a reliable and interpretable regression model for accurate insights.

A screenshot of a computer code

Description automatically generated

A screenshot of a number

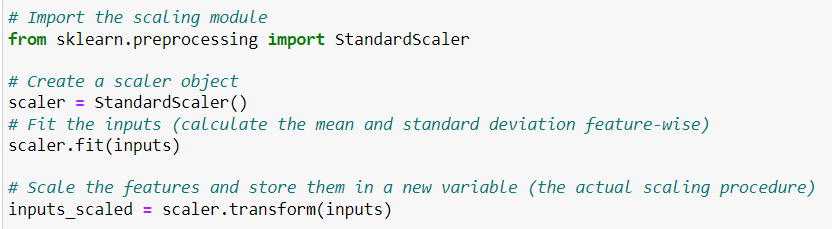
Description automatically generated

If the Variance Inflation Factor (VIF) surpasses 10, indicating significant multicollinearity with other variables, it is advisable to exclude the respective variable from the regression analysis. While retaining it may compromise the model’s stability, removing the high-VIF variable ensures a more robust and interpretable regression model. This strategic adjustment maintains the reliability of the analysis and enhances the accuracy of insights drawn from the remaining predictors, fostering a more effective and trustworthy modeling process.

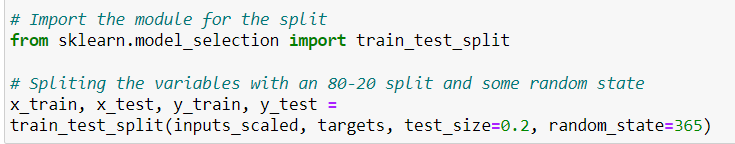
**APPLYING LINEAR REGRESSION**

Prior to conducting regression analysis, it’s essential to partition the data into target and input variables. Inputs are employed in developing the regression model, predicting the target variable. This segmentation is fundamental for the accurate execution and interpretation of linear regression. We choose our input variable as logarithmic price.

Importing essential libraries into the notebook and initializing a scaler object to transform the distribution of values into a normal distribution.



Partition the data into training and testing sets. The training set is utilized to train the machine learning model, while the test set is reserved for evaluating the model’s performance. Data division into training and testing sets is essential for machine learning. The training set instructs the model on data patterns, while the untouched testing set evaluates its predictive accuracy. This bifurcation ensures an unbiased assessment, gauging the model’s ability to generalize to new, unseen data. The training-test split is a fundamental step in model assessment, guaranteeing reliability and applicability to diverse scenarios.



Now the Regression is applied.

A white background with black text

Description automatically generated

That concludes the regression process for the provided data using `x\_train` as the input features and `y\_train` as the target variable. Predicting the outputs of the regression as follows.

A white rectangular sign with blue text

Description automatically generated

Creating a plot comparing `y\_hat` (predicted values) against `y\_train` to visualize the behavior and alignment of the model’s predictions with the actual values.

A white background with black text

Description automatically generated



Observing the plot, it’s evident that Predictions and Targets align along the 45-degree line. This alignment signifies a close match between the regression predictions and the actual target values.

Distrubtion of Residuals: Once a regression model has been fit to a group of data, examination of the residuals, the deviations from the fitted line to the observed values, allows the modeler to investigate the validity of the assumption that a linear relationship exists. Plotting the residuals on the y-axis against the explanatory variable on the x-axis reveals any possible non-linear relationship among the variables or might alert the modeler to investigate lurking variables.

A white background with black text

Description automatically generated

A graph of a function

Description automatically generated

Ideally, this plot should exhibit a normal distribution. However, in our scenario, numerous negative residuals, significantly deviating from the mean, are observed. As per the definition of residuals (y\_train - y\_hat), negative values indicate that the predictions (y\_hat) surpass the actual targets (y\_train) by a considerable margin.

Results and Performance:

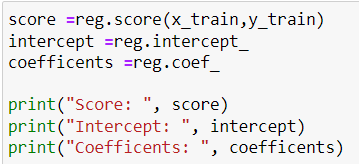
These values provide insights into the model's accuracy, the intercept term, and the influence of each feature on the target variable. The regression model’s performance score on the training data is captured by

score = reg.score(x\_train, y\_train)

The intercept and coefficients of the fitted model are determined by

intercept = reg.intercept\_

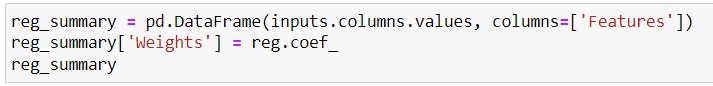
coefficients = reg.coef\_



A number with numbers on it

Description automatically generated with medium confidence

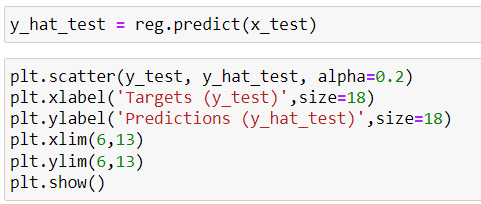
Generating a regression summary allows for a comparative analysis of features. The summary, stored in `reg\_summary`, comprises a DataFrame presenting each feature’s name in the ‘Features’ column and its corresponding weight in the ‘Weights’ column. This tabulation offers insights into the influence and contribution of each feature to the linear regression model.



A screenshot of a table

Description automatically generated

To evaluate the model’s performance, we can apply it to predict outcomes using the `x\_test` data. This testing phase assesses how well the model generalizes to new, unseen data. Creating a scatter plot with the test targets and the test predictions



A graph with a red line

Description automatically generated

The presence of a 45-degree alignment suggests that the regression is effectively predicting values close to the actual ones, indicating a reasonable performance in capturing the underlying patterns in the data.

To obtaining the actual prices, we take the exponential of the log\_price

A white background with black text

Description automatically generated

A computer code with text

Description automatically generated

A table with numbers and a few black text

Description automatically generated

In the above table we can see that how closely the model is trying predict the target price. we could understand the influence of each independent variable on our dependent variable. This interpretation is crucial as it goes beyond mere number-crunching, providing real-world context to our findings.

**Chapter 4: CONCLUSION**

The linear regression analysis undertaken in this project offers valuable insights into understanding the relationships between independent variables and a dependent variable, providing a robust framework for predictive modeling and inference. Through meticulous data exploration, preprocessing, and model development, several critical observations were made.

This project was more than just an academic exercise; it was a journey into the heart of data analysis. The insights gained from our linear regression model have the potential to inform decisions in a real-world context. The project highlighted the power of statistical methods in extracting meaningful information from data and the importance of a thorough understanding of the underlying assumptions of these methods.

While our project provided valuable insights, it also opened doors to further exploration. Different approaches to variable selection, using alternative regression models, or applying the model to different datasets could yield new findings. The field of data science is dynamic, and there’s always room for experimentation and learning.

In conclusion, the linear regression analysis conducted in this project offers a comprehensive framework for modeling and understanding the relationships between variables. By adhering to best practices in data analysis and model development, the project delivers actionable insights that can inform decision-making processes and drive informed business strategies. This conclusion encapsulates the key findings and implications of the linear regression analysis, highlighting the project's significance in deriving actionable insights from data.