LINEAR REGRESSION ASSIGNMENT

SUBJECTIVE QUESTIONS

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   In examining the categorical variables within the dataset, a comprehensive analysis reveals captivating insights into their potential impact on the dependent variable. By analyzing the details of these categorical features we may understand how they may influence the outcome of the dataset. Here are some inferences to draw valuable insights into the relationships and patterns that emerge from this exploration:

   - The **year 2019** attracted more bookings overall as compared to the year 2018, which seems to be a positive indicator for the business.
   - **Fall season** has more bookings followed by **Summer season**, which seems to be fewer in winter and spring seasons.
   - **Clear weather**, of course, attracted more bookings as compared to misty and light snow & and rainy weather.
   - **Thur, Fri, Sat, Sun** have a greater number of bookings as compared to the start of the week.
   - Most of the bookings were done during the months of JUNE in the year 2018 and SEPT in the year 2019.
   - Booking seems to be increasing from the start of the year till mid and then drastically decreasing as the end of year approaches.

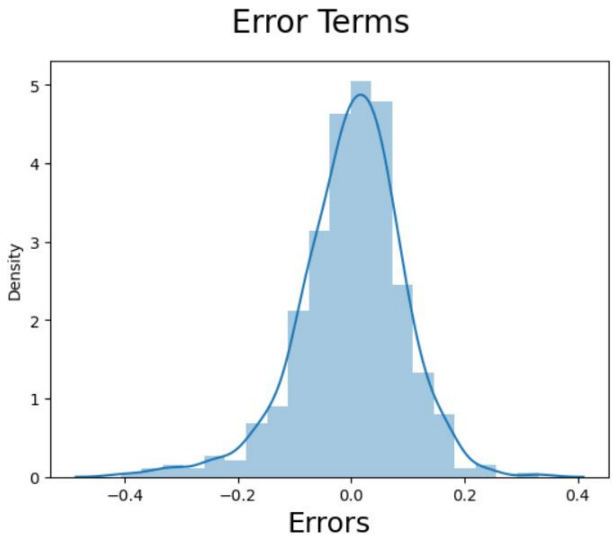
2. **Why is it important to use drop_first=True during dummy variable creation?**

- Dummy Variables are created to convert categorical data into numbers for statistical uses. While doing this, multicollinearity occurs which leads to unstable coefficient estimates and standard errors, making it difficult to interpret the individual impact of each dummy variable.
- Therefore, using drop_first = True is important as it helps in reducing the extra column created during Dummy Variable creation.
- By dropping it, we can ensure that the dummy variables are independent and the model can accurately estimate the effects without overlapping information.
- For example, if we want to create dummy variables for categorical columns having three categories, X, Y, and Z, we will create two dummy variables. If the variable/category is neither X nor Y, it implies that the variable must be Z. Therefore, we can conclude that information about one category can be easily predicted from others, hence, drop_first = True helps in removing that extra column.
- To make our statistical model work better and give reliable results, it's important to ensure that each dummy variable gives a unique answer without repeating too much.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Variable 'temp' has the highest correlation with the target variable, 'cnt', that is, 0.65.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Normality of error terms, that is, error terms are normally distributed.



- Multicollinearity check shows that there is insignificant multicollinearity among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**

   The top three features that significantly contribute to explaining the demand for shared bikes are:

- temp : A coefficient value of '0.5480' indicated that a unit increase in temp variable, increases the bike hiring numbers by 0.5480 units.
- Light_snowrain : A coefficient value of '-0.2829' indicated that a unit increase in the Light_snowrain variable, that is, light snow and rain, decreases the bike hiring numbers by 0.2829 units.
- winter : A coefficient value of '0.1293' indicated that a unit increase in the Winter variable increases the bike hiring numbers by 0.1293 units.

GENERAL SUBJECTIVE QUESTIONS

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a statistical model that helps in understanding the relationship between a dependent variable with given independent variable through a straight line. This means that when there is a change in the value of one or more independent variables, whether increase or decrease, the value of the dependent variable also changes accordingly.

- Dependent variable, in any given task, is the variable we aim to predict. In linear regression, it is denoted by 'y'.
- Independent variables are the variables that affect the dependent variable and are denoted as $X_1, X_2, ..., X_n$.
- The algorithm estimates coefficients ($b_1, b_2, ..., b_n$) for each independent variable and an intercept (b_0).

Therefore, the linear regression equation formed by these terminologies is:

$$y = b_0 + b_1X_1 + ... + b_nX_n$$

The primary objective in linear regression is to find the values of $b_0, b_1, b_2, ..., bn$ that minimize the sum of squared differences between the predicted and actual values, and this is known as Ordinary Least Square(OLS) regression.

Following are the steps for the model-

1. The dataset is split into Training and Test sets.
2. It is then divided into X and Y sets for model building.

3. Recursive Feature Elimination is done that removes features, fits a model, and then evaluates the performance until the desired number of features is reached.
4. Next, the Variance Inflation Factor is measured which quantifies how much a variable is contributing to the variance of a model. It is used to detect multicollinearity among predictor variables in a regression analysis.
5. Then the Linear Model is built by using statsmodel.
6. After the model is finalized, Residual Analysis is done of the Train data.
7. At last, prediction is done using the final model, followed by Model Evaluation.

There are two types of Linear Regression:

a) **Simple linear regression** – it explains the relationship between a dependent variable and one independent variable using a straight line. Equation – $y = b_0 + b_1.X$

b) **Multiple linear regression** – it is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
Equation - $y = b_0 + b_1X_1 + \ldots\ldots\ldots + b_nX_n$

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of plotting data before you analyze it and build your model.
- These four datasets have nearly the same statistical observations, which provide the same information for each x and y point in all four datasets.
- However, when you plot these datasets, they look very different from one another, that is, a set of four datasets that have nearly identical simple descriptive statistics but vary widely when graphed.
- It visualizes data before analyzing it and to highlight the limitations of relying solely on summary statistics.
- While sharing identical mean, variance, correlation, and linear regression line characteristics, the datasets within the quartet tell the significance of visually examining data and avoiding dependence solely on summary statistics.
- This quartet helps in understanding the importance of Visualization as by looking at the data structure graphically we can get a clear understanding of it.
- Therefore, we can say, it highlights that depending exclusively on numerical summaries can be misleading sometimes and we might fail to grasp the complexity within the data.

3. What is Pearson's R?

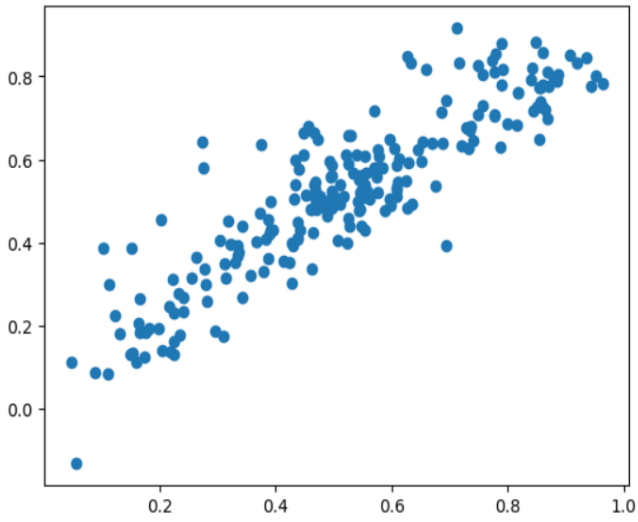
Pearson's R or Pearson's Correlation Coefficient is a statistical tool that measures the strength of linear relationship between the variables. It quantifies how well the relationship between variables can be described with the help of a straight line.

The Pearson's (r) or the value of (r) can range from -1 to 1:

- r = -1, indicates a perfect negative linear relationship
- r = 0, indicates no linear relationship
- r = 1, indicates perfect positive linear relation.

The sign of (r) indicates the direction of the relationship.

- A Positive sign means that when one variable increases, the other also increases, whereas
- A Negative sign indicates that with the increase of one variable, the other will decrease.
- The only limitation is that Pearson's correlation measures only linear relationships and may not capture non-linear associations. Also, it is sensitive to outliers, meaning that extreme values can influence the results.



This scatter plot shows Positive Correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

In machine learning and statistics, scaling is the process of converting numerical features of a data set into a standard range (or distribution). The main purpose of scaling is to make sure that all features contribute the same amount to the analysis and the model training process. Scaling is the process of reducing the values of variables in a data set to a specific scale so that they are comparable and that variables with larger values do not dominate the analysis.

Scaling is performed for various reasons:

- Equal contribution: scaling all features ensures that they contribute the same amount to the model, otherwise, variables with larger scales will dominate and affect the model disproportionately.
- Gradient descent convergence: many machine learning algorithms (especially gradient descent optimization algorithms, such as linear regression and logistic regression, and neural networks) converge faster and work better when input features are similar in scale.
- Distance-based algorithms: algorithms that rely on distance between data points, such as K-nearest Neighbor and K-Means clustering, are sensitive to the size of features. Scalating prevents features with larger sizes from having a larger effect on distance calculations.

Difference between Normalized Scaling & Standardized Scaling

Normalized	Standardized
1. It uses Minimum and Maximum of features for scaling.	It uses Mean and Standard Deviation for scaling.
2. It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3. Scales values are between [0,1] or [-1,1].	It is not bound to a certain range.
4. It can be affected by outliers.	It is much less affected by outliers.
5. Scikit-Learn provides a transformer called *MinMaxScaler* for normalization.	Scikit-Learn provides a transformer called *StandardScaler* for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) serves as a tool for detecting multicollinearity in a regression model. Elevated VIF values suggest pronounced correlations among the independent variables, complicating the ability to discern the unique influence of each variable on the dependent variable. Although VIF is a valuable diagnostic tool, it can encounter issues and potentially reach infinity under specific circumstances.

Infinite VIF occurs when one or more independent variables are perfectly correlated, leading to a situation known as perfect multicollinearity. Perfect multicollinearity means that one or more variables can be exactly predicted using a linear combination of the

other variables in the model. To solve this we need to drop one of the variables from the dataset which is causing this.
If a variable in the model is a linear combination of other variables, it can result in a perfect multicollinearity scenario and, consequently, an infinite VIF.
When there is a linear dependence among the independent variables, the matrix inversion required in the VIF calculation becomes problematic, resulting in an infinite VIF.
Infinite VIF values can lead to unreliable coefficient estimates in regression models.
The standard errors of the coefficients become extremely large, making it difficult to draw valid statistical inferences.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess the normality of a dataset by comparing the quantiles of the observed data with the quantiles of a theoretical normal distribution. In a Q-Q plot, if the points closely follow a straight line, it indicates that the data is approximately normally distributed.

Use and importance of a Q-Q plot in linear regression:

- The Q-Q plot helps in visually inspecting how well the distribution of residuals matches a normal distribution. Deviations from a straight line in the Q-Q plot may indicate departures from normality.
- Non-normality in residuals can affect the accuracy and reliability of statistical tests associated with linear regression.

- Outliers in the data can impact the normality of residuals. Q-Q plots can reveal whether extreme values in the residuals are consistent with a normal distribution or if they deviate significantly.
- A linear regression model assumes that the residuals are normally distributed. A Q-Q plot helps evaluate whether this assumption holds true, ensuring the adequacy of the model.
- By examining the Q-Q plot, analysts can identify situations where the normality assumption might be violated. Addressing such violations might involve model adjustments, transformations, or considering alternative statistical methods.
- If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are normally distributed.
- Deviations from the straight line may indicate departures from normality, such as skewness or heavy tails in the distribution.
- A Q-Q plot is a valuable diagnostic tool in linear regression, providing a visual assessment of the normality of residuals. It helps ensure the validity of statistical inferences and aids in identifying areas where the model may need adjustments for better performance.