Linear Regression Assignment

Assignment-based Subjective Questions:

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

Ans 1. I conducted an analysis of the categorical columns using box plots and bar plots, which led to several interesting observations. Firstly, it appears that the fall season attracted a higher number of bookings, with a significant increase in booking counts across all seasons from 2018 to 2019. Most bookings were made during the months of May, June, July, August, September, and October, showing a clear upward trend from the beginning of the year until mid-year, followed by a decline as we approached the end of the year. Additionally, it is evident that clear weather contributed to increased bookings, which is quite intuitive. We also found that Thursdays, Fridays, Saturdays, and Sundays saw more bookings compared to earlier in the week. Interestingly, bookings were lower on non-holidays, likely because people prefer to spend time at home and with family during those times. Furthermore, the data indicates that bookings on working days and non-working days were almost equal. Overall, 2019 showed a notable increase in bookings compared to the previous year, reflecting positive growth in business.

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

Ans 2. Using the parameter “**drop_first=True**” during dummy variable creation is beneficial as it helps to reduce the extra column that would typically be generated. This approach effectively avoids redundancy in the dataset, streamlining the analysis and improving model efficiency.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans 3. The variable 'temp' exhibits the highest correlation with the target variable, indicating a significant relationship between them.

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

Ans 4. I validated the assumptions of the Linear Regression Model based on five key criteria. Firstly, I checked for the normality of error terms, ensuring that they are normally distributed. Next, I conducted a multicollinearity check to confirm that there is insignificant multicollinearity among the variables. I also validated the linear relationship, verifying that linearity is evident among the variables. Additionally, I assessed homoscedasticity, ensuring there are no visible patterns in the residual values. Finally, I examined the independence of residuals to confirm that there is no auto-correlation present.

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

Ans 5. The top three features that significantly contribute to explaining the demand for shared bikes are temperature (temp), the winter season, and the month of September.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans 1. Linear regression is a statistical approach used to understand the relationship between a dependent variable, which is the outcome we want to predict, and one or more independent variables, which are the predictors. Essentially, it seeks to find the best-fitting straight line that describes how changes in the independent variables affect the dependent variable. The goal is to minimise the differences between the actual data points and the predicted values generated by the model. To create this model, we start by estimating coefficients that indicate how much influence each independent variable has on the dependent variable. There are certain assumptions underlying linear regression, such as the belief that the relationship between the variables is linear, that the errors in predictions are independent, and that they have a constant variance and a normal distribution. After building the model, we evaluate its performance using various metrics to understand how well it predicts new data, ultimately enabling us to make informed predictions based on the relationships identified in the training data.

2. Explain the Anscombe's quartet in detail.

Ans 2. Anscombe's Quartet is a collection of four datasets that remarkably demonstrate how statistical summaries can be misleading without proper data visualisation. Each dataset contains pairs of values with similar statistical properties, such as the same mean and variance, yet their graphical representations reveal very different relationships. For instance, one dataset shows a straightforward linear relationship, while another exhibits a clear non-linear pattern. A third dataset features an outlier that dramatically skews the results, and the fourth dataset shows a vertical line indicating a perfect correlation in one direction. The main takeaway from Anscombe's Quartet is the critical importance of visualising data; relying solely on summary statistics can obscure significant insights and lead to incorrect conclusions. This highlights the necessity of exploring the underlying patterns in data before conducting any statistical analysis.

3. What is Pearson's R?

Ans 3. Pearson's R, or the Pearson correlation coefficient, is a measure that indicates the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where a value of 1 signifies a perfect positive correlation, meaning that as one variable increases, the other does as well. Conversely, a value of -1 indicates a perfect negative correlation, where one variable increases while the other decreases. A value of 0 suggests no linear correlation between the variables. To calculate Pearson's R, we assess how the two variables vary together compared to how much they vary individually. It's important to note that this measure assumes that both variables are continuous, normally distributed, and have a linear relationship, making it a valuable but specific tool for examining relationships in data.

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

Ans 4. Scaling is the process of adjusting the range and distribution of features in a dataset so that they are on a comparable scale. This practice is particularly crucial in machine learning, as many algorithms, especially those that rely on distance calculations or optimization techniques, perform better when the input features are standardized. Scaling helps to ensure that no single feature dominates others due to differences in magnitude, thus improving the model's performance and speed during training. There are two common methods of scaling: normalization and standardization. Normalization adjusts the features to a specific range, usually between 0 and 1, while standardization transforms the features so they have a mean of zero and a standard deviation of one. Each method serves different purposes depending on the nature of the data and the specific algorithm used.

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

Ans 5. The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. A VIF of 1 indicates no correlation among the kth predictor and the remaining predictors, while a VIF exceeding 5-10 indicates high multicollinearity.

- **Infinite VIF**: A VIF can become infinite when there is perfect multicollinearity, meaning one predictor is a perfect linear function of others. This occurs when:
  - A variable is a linear combination of other variables.
  - One variable is identical to another variable.

In such cases, the model cannot estimate the coefficients accurately because the relationship between predictors is not independent.

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

Ans 6. A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, most commonly the normal distribution. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the points lie approximately on a straight line, it indicates that the data follows that distribution. In linear regression, one of the assumptions is that the residuals (errors) are normally distributed. A Q-Q plot can be used to visually check this assumption.

Importance:

- The Q-Q plot helps in determining whether the residuals deviate from normality, which can impact the validity of hypothesis tests and confidence intervals derived from the model.
- It aids in diagnosing potential problems with the model, such as non-linearity or outliers, allowing analysts to take corrective measures.
- It provides a clear and visual means of assessing how well the data conforms to the normal distribution.