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.

Seasonal:

- Fall season experienced a surge in bookings, with overall bookings significantly increasing across all seasons from 2018 to 2019.

Months:

- Peak months were May-October, with bookings steadily rising towards mid-year before a gradual decline later. Bookings grew compared to 2018 in each month.

Weather:

- Sunny weather correlated with higher bookings, as expected. Bookings during all weather types increased in 2019 compared to 2018.

Day of Week:

- Thursdays, Fridays, Saturdays, and Sundays saw higher booking volumes compared to weekdays.

Holidays:

- Bookings were lower on non-holidays, suggesting people prioritized spending time with family.

Working vs. Non-Working Days:

- Booking counts were relatively even between working and non-working days, with a positive increase from 2018 to 2019.

Overall:

- 2019 saw a substantial rise in bookings compared to 2018, indicating positive business growth.

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

Ans. In regression analysis, setting "drop_first=True" when creating dummy variables helps prevent multicollinearity, where variables are highly correlated and redundant. This ensures one category serves as a reference point, allowing for clear interpretation of the impact of other

categories on the outcome variable. Skipping this step can lead to perfect multicollinearity, where one dummy variable is perfectly predictable from the others. This can make it difficult to determine the independent effect of each category and increase the variance of coefficient estimates, potentially rendering the model unreliable.

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

Ans.

Among the numerical variables, temp has the highest correlation with the target variable.

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

Ans.

Normality of error terms:

- Error terms should follow a normal distribution.

Multicollinearity check:

- There should be no significant multicollinearity among variables.

Linear relationship validation:

- Variables should exhibit linearity.

Homoscedasticity:

- There should be no discernible pattern in residual values.

Independence of residuals:

- Residuals should not show autocorrelation.

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

Ans.

- Temperature (temp)
- light precipitation
- Year (yr)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans.

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables. Its objective is to find the best-fitting linear equation that describes how the independent variables affect the dependent variable.

In linear regression, the relationship between the independent variables (x_1, x_2, ..., x_n) and the dependent variable (y) is represented by the equation:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

Here, y is the dependent variable, x_1, x_2, ..., x_n are the independent variables, β_0, β_1, β_2, ..., β_n are the coefficients (also known as weights or parameters), and ε is the error term.

The goal of linear regression is to estimate the coefficients (β_0, β_1, β_2, ..., β_n) that minimize the difference between the predicted values and the actual values of the dependent variable. This is typically achieved using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values.

The linear regression algorithm involves several steps:

1. Data preprocessing: This includes cleaning the data, handling missing values, and scaling or normalizing the features as required.

2. Model training: The algorithm fits the linear regression model to the training data by estimating the coefficients using optimization techniques like gradient descent or analytical solutions such as the normal equation.

3. Model evaluation: The performance of the model is assessed using evaluation metrics like mean squared error (MSE), root mean squared error (RMSE), or R^2 score to determine how well the model fits the data.

4. Prediction: Once the model is trained and evaluated, it can be used to make predictions on new or unseen data by inputting the values of the independent variables into the linear equation.

Linear regression is widely applied in various domains such as economics, finance, healthcare, and social sciences for tasks like predicting house prices, estimating sales, analyzing the impact of variables on outcomes, and more. Its simplicity, interpretability, and ease of implementation make it a valuable tool for data analysis and decision-making.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of summary statistics alone in understanding the underlying relationships between variables.

The quartet consists of four datasets, each containing 11 data points:

1. Dataset I: A simple linear relationship between two variables.

2. Dataset II: A non-linear relationship between two variables.

3. Dataset III: A strong linear relationship between two variables with an outlier.

4. Dataset IV: No apparent relationship between two variables, but with an influential outlier that greatly affects the correlation coefficient.

Despite their different characteristics, all four datasets have the same mean, variance, correlation coefficient, and linear regression line parameters. This highlights the danger of relying solely on summary statistics like mean and correlation without visualizing the data.

Anscombe's quartet underscores the importance of data visualization in exploratory data analysis and the interpretation of statistical models. It emphasizes that visual inspection of data plots is essential for uncovering patterns, relationships, and outliers that may not be apparent from summary statistics alone. Additionally, it serves as a cautionary example against drawing conclusions based solely on numerical summaries without considering the context and distribution of the data.

3. What is Pearson's R?

Ans.

Pearson's correlation coefficient, denoted as r, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where:

r=+1r=+1: Represents a perfect positive linear relationship.

r=−1r=−1: Indicates a perfect negative linear relationship.

r=0r=0: Suggests no linear relationship between the variables.

To calculate r, we divide the covariance of the two variables by the product of their standard deviations. While r is sensitive to the variables' measurement scales, it assumes a linear relationship and normal distribution of both variables.

Pearson's r is extensively used in statistics, research, and data analysis to understand relationships between variables. It helps researchers and analysts comprehend how changes in one variable correspond to changes in another, offering valuable insights into data patterns and dependencies.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans.

Scaling is the process of adjusting numerical features within a dataset to a comparable scale or range. It's a crucial step to ensure that all features contribute equally to the analysis and avoid biases in the model based on the magnitude of values.

Scaling becomes essential because many machine learning algorithms are sensitive to the scale of input features. For instance, algorithms like k-nearest neighbors (KNN) and support vector machines (SVM) compute distances between data points, and features with larger scales could dominate these calculations, potentially leading to biased outcomes.

Two common techniques for scaling are normalized scaling and standardized scaling:

1. Normalized scaling, also known as Min-Max scaling, transforms features to a predetermined range, often between 0 and 1. While it retains the original data distribution, it may be influenced by outliers.

2. Standardized scaling, or z-score scaling, standardizes features to have a mean of 0 and a standard deviation of 1. This centers the data around 0 and adjusts it based on the standard deviation, making it less susceptible to outliers. Standardized scaling is often preferred when dealing with outliers or when algorithms assume normally distributed features.

In essence, scaling is performed to ensure that all features are brought to a consistent scale, thereby enhancing the performance and reliability of machine learning algorithms. Both normalized scaling and standardized scaling offer distinct advantages and considerations, catering to different data scenarios and algorithm requirements.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans.

Sometimes, the value of the Variance Inflation Factor (VIF) can become infinite. This occurs when one or more independent variables in a regression model are perfectly linearly related to each other. When two or more variables are perfectly collinear, it means that one of the variables can be expressed as a linear combination of the others. As a result, the matrix inversion used in calculating the VIF becomes impossible, leading to an infinite value for the VIF.

Perfect multicollinearity can arise due to various reasons, such as data duplication, coding errors, or including derived variables that are linearly dependent on each other. Regardless of the cause, infinite VIF values indicate a serious issue in the regression model, as it undermines the model's stability and interpretability.

To address this problem, it's essential to identify and resolve the multicollinearity issue by removing one of the correlated variables or using dimensionality reduction techniques like Principal Component Analysis (PCA). By resolving multicollinearity, we can ensure the reliability and accuracy of the regression analysis results.

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

Ans.

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a given data distribution is similar to a theoretical distribution, such as 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 data distribution closely follows the theoretical distribution, the points on the Q-Q plot will fall approximately along a straight line.

The use of a Q-Q plot in linear regression is essential for several reasons:

1. Assumption Checking: Q-Q plots are commonly used to check the assumption of normality in linear regression residuals. Linear regression models assume that the residuals (i.e., the differences between observed and predicted values) are normally distributed. By visually inspecting the Q-Q plot of the residuals, we can determine whether they deviate significantly from a normal distribution. If the points on the Q-Q plot deviate from the straight line, it suggests that the normality assumption may not hold.

2. Detecting Outliers: Q-Q plots can also help in identifying outliers in the data. Outliers are data points that significantly deviate from the overall pattern of the data distribution. In a Q-Q plot, outliers appear as points that fall far away from the diagonal line, indicating that they may not be consistent with the assumed distribution.

3. Model Performance Evaluation: By examining the shape of the Q-Q plot, we can gain insights into the performance of the linear regression model. A Q-Q plot with points closely aligned along the diagonal line suggests that the model's residuals are normally distributed, indicating a good fit between the model and the data. On the other hand, deviations from the diagonal line may indicate model misspecification or violation of assumptions, requiring further investigation and potential model refinement.

In summary, Q-Q plots play a crucial role in linear regression analysis by helping to assess the normality of residuals, detect outliers, and evaluate the overall performance of the regression model. They provide valuable visual insights that aid in the interpretation and validation of regression results, ultimately enhancing the reliability and accuracy of statistical analyses.