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?

I conducted an analysis on the categorical columns using boxplots and bar graphs, and based on the visualizations, the following observations can be made:

1. Bookings were highest during the fall season, with a significant increase in the number of bookings from 2018 to 2019 in each season.
2. The majority of bookings were made between May and October, with a trend of increasing bookings from the beginning of the year until mid-year, and then decreasing towards the end of the year.
3. Clear weather was a significant factor in attracting more bookings.
4. Thursdays, Fridays, Saturdays, and Sundays had higher booking numbers than the start of the week.
5. Bookings were fewer on non-holidays, which is reasonable as people may prefer to spend time at home with family during holidays.
6. Booking numbers were almost the same for working and non-working days.
7. In 2019, there were more bookings than the previous year, indicating good progress in terms of business.

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

It is important to use drop_first=True during dummy variable creation to avoid the "dummy variable trap". The dummy variable trap occurs when we include all the categorical variables as dummy variables in the regression model, without dropping one of the categories. This leads to a situation called multicollinearity, where the independent variables become highly correlated with each other. This, in turn, can affect the performance of the regression model and cause issues such as overfitting.

By setting drop_first=True, we drop the first category of each categorical variable, which then becomes the reference category. This means that the remaining categories of the variable are represented by their respective dummy variables, which helps to avoid the dummy variable trap

and reduce multicollinearity in the regression model. It also reduces the number of variables in the model, making it less complex and easier to interpret.

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

'temp' variable 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?

I have verified the assumptions of the Linear Regression Model, based on the following five criteria:

- Normality of error terms: The error terms must follow a normal distribution.
- Multicollinearity check: There should be no significant multicollinearity among the variables.
- Linear relationship validation: A linear relationship should be evident among the variables.
- Homoscedasticity: There should be no discernible pattern in the residual values.
- Independence of residuals: There should be no autocorrelation present in the residuals.

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. const
2. year
3. sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method for modeling the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as the predictor or explanatory variables). The goal of linear regression is to find the best-fit line that describes the relationship between the variables.

In simple linear regression, there is only one independent variable, and the equation for the best-fit line is of the form y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the intercept. The goal is to find the values of m and b that minimize the sum of the squared differences between the actual values of y and the predicted values of y (i.e., the errors).

Multiple linear regression is a more complex form of linear regression that involves multiple independent variables. The equation for the best-fit line is of the form y = b0 + b1x1 + b2x2 + ... + bnxn, where b0 is the intercept, b1, b2, ..., bn are the slopes, and x1, x2, ..., xn are the independent variables. The algorithm for multiple linear regression is similar to that for simple linear regression, except that it involves finding the values of b0, b1, b2, ..., bn that minimize the sum of squared errors for the multiple independent variables.

In summary, linear regression is a widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables. The algorithm involves data collection, preprocessing, model training, model evaluation, and model deployment, and can be used for both simple and multiple linear regression.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have identical statistical properties, despite having vastly different appearances when plotted. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it.

Each dataset consists of 11 x,y pairs of data. When plotted, each dataset appears to have a linear relationship between x and y, and they all have a correlation coefficient of approximately 0.82. However, the data points are distributed very differently in each dataset, resulting in vastly different plots.

The first dataset is a simple, well-behaved linear relationship. The second dataset is also a linear relationship, but with one outlier that drastically changes the regression line. The third dataset has a non-linear relationship that is masked by one outlier. The fourth dataset has a strong linear relationship, but it is driven entirely by one outlier, and the other 10 points have no relationship at all.

The key takeaway from Anscombe's quartet is that a summary statistic, like a correlation coefficient or regression line, can be misleading if we don't take the time to visualize the data. By creating these datasets with identical statistics but vastly different appearances, Anscombe showed that we need to be cautious in relying solely on summary statistics to understand data.

In summary, Anscombe's quartet is a powerful demonstration of the importance of visualizing data before analysis. It reminds us that a good analysis requires both numerical and visual exploration of the data.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It is a statistic that ranges from -1 to +1, where -1 represents a perfect negative linear relationship, 0 represents no linear relationship, and +1 represents a perfect positive linear relationship.

The Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. This measures the degree to which the two variables vary together, while controlling for their individual scales. The resulting value indicates the strength and direction of the linear relationship between the two variables.

Pearson's R is commonly used in fields such as psychology, economics, and social sciences to measure the relationship between two continuous variables. It is a widely used statistic because it is easy to interpret and understand, and it is sensitive to both positive and negative relationships.

However, it is important to note that Pearson's R only measures linear relationships and may not capture other types of relationships, such as non-linear or categorical relationships. Additionally, Pearson's R does not imply causation and should be interpreted with caution.

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

Scaling is a preprocessing technique in which the range of values of a feature is transformed to a specific scale. The goal of scaling is to bring all features to a similar range so that they can be

compared on equal footing and the model can converge more quickly and produce more accurate results.

Scaling is performed to ensure that features with large scales do not dominate features with small scales during modeling, as some machine learning algorithms are sensitive to the scale of the input data.

Normalized scaling and standardized scaling are two common types of scaling techniques. Normalized scaling involves scaling the values of a feature to be between 0 and 1. This can be achieved by subtracting the minimum value of the feature and then dividing by the range of the feature. Normalization preserves the shape of the distribution of the data and is useful when the distribution of the data is skewed.

Standardized scaling involves scaling the values of a feature to have a mean of 0 and a standard deviation of 1. This can be achieved by subtracting the mean of the feature and then dividing by the standard deviation of the feature. Standardization transforms the data to have a normal distribution and is useful when the data is normally distributed or when the model assumes normally distributed data.

In summary, scaling is a preprocessing technique used to transform the range of values of a feature to a specific scale. Normalized scaling and standardized scaling are two common types of scaling techniques used to bring features to a similar range, with different benefits depending on the distribution of the data.

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

The Variance Inflation Factor (VIF) is a measure of the degree of multicollinearity among the independent variables in a linear regression model. A high VIF value indicates that a variable is highly correlated with other variables, which can lead to unreliable and unstable coefficient estimates in the model.

In some cases, the VIF value can be infinite, indicating perfect multicollinearity among the variables. This happens when one or more variables in the model can be expressed as a linear combination of the other variables. In other words, one or more variables can be perfectly predicted from the other variables in the model.

This perfect multicollinearity can arise in situations such as when there are duplicate variables or when one variable is calculated as a linear combination of other variables. For example, if a model includes both height in inches and height in feet as independent variables, the VIF for one of these variables will be infinite because it can be perfectly predicted from the other variable.

In such cases, it is important to identify and remove the redundant variables to ensure the stability and reliability of the model. This can be done through techniques such as correlation analysis or dimensionality reduction methods like principal component analysis.

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

A Q-Q (quantile-quantile) plot is a graphical technique used to compare the distribution of a sample of data to a theoretical distribution, such as a normal distribution. The sample data is plotted against the expected values from the theoretical distribution, with deviations indicating differences between the two distributions.

In linear regression, Q-Q plots are useful for assessing the normality assumption of the residuals, which is a critical assumption of the linear regression model. A Q-Q plot can help determine whether the residuals follow a normal distribution, which is necessary for accurate predictions and reliable inference. If the residuals do not follow a normal distribution, the model may need to be modified or other statistical techniques may need to be used.