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?
   Based on my analysis on the categorical columns using bar chart and box plots, following are the inference:
   [Answer]
   a. The average bike rental is higher in Fall and Summer.
   b. Booking Rentals in the year 2019 are higher than 2018 in all quarters.
   c. In 2019 the rentals are highest in September than in any other month, whereas in 2018, June seems to be having good business than any other month.
   d. The effect of days is not having much difference in the rentals as its pretty much the same on average among all the days.
   e. Where there is Clear Weather, the rental business has good booking than any other weather condition.
   f. Across working day or non-working day is not making much of a difference.

2. Why is it important to use drop_first=True during dummy variable creation?
   [Answer]
   When we create dummy variables for a categorical feature with 'n' categories, we typically end up with 'n' dummy variables. Since, these variables are not entirely independent. If we know the values of n-1 dummy variables automatically determines the value of the remaining one (since they all sum to 1). This is to avoid Multicollinearity which occurs when there's a high degree of linear correlation relationships between independent variables and all this can be avoided since we can infer the nth variable based on n-1 variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   [Answer]
   The 'temp' independent 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?
   [Answer]
   I validated the assumptions of linear regression model on below 4 assumptions:
   a. Normality: The error terms are normally distributed with mean 0.
   b. Linearity: Using a scatter plot, I validated the linearity is visible among variables
   c. Multicollinearity check: Using a correlation heatmap plot, Among the independent variables, I could not find any relationship. Also, the VIF was <= 5 in the model
   d. Homoscedasticity: I plotted the residuals against the fitted values, I found a random scatter with no clear pattern.

e. Independence of Residuals: I plotted the residuals (errors) against the fitted values and were randomly scattered around zero with no apparent patterns. This suggests no dependence between the errors.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   [Answer]
   1. Temp (positive correlation, which means when temperature rises the booking increases)
   2. Season_Spring (negative correlation, as the seasons move towards spring season the bookings decrease)
   3. Windspeed (negative correlation, when its more windy there seems to be reduced booking)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   [Answer]
   Linear regression is a fundamental statistical technique used for modelling the relationship between a dependent variable (what you want to predict) and one or more independent variables (what you think might influence the dependent variable). It assumes this relationship can be represented by a straight line in the form of

   $y = mx + c$

Dependent Variable (y): This is the variable that we are trying to predict or explain. It's the numerical outcome that we are interested in.
Independent Variable(s) (x): These are the features or factors that might influence the outcome of dependent variable. They can be continuous (like height) or categorical (like hair color). In ***simple linear regression***, there's one independent variable, but ***multiple linear regression*** can handle multiple X variables.

The Assumption is that there is a linear relationship between the independent and dependent variables. This means changes in the independent variable result in proportional changes in the dependent variable. Imagine a straight line:

Positive Linear Relationship: As the X value increases, the Y value increases (positive slope).
Negative Linear Relationship: As the X value increases, the Y value decreases (negative slope).
No change in Y with changes in X (zero slope).

The algorithm aims to find the equation of a straight line that best fits the data points in a scatter plot (X vs. Y or X's vs Y in a multidimension). This "best fit" is achieved by minimizing the sum of the squared errors between the actual Y values and the Y values predicted by the line for each data point.

2. Explain the Anscombe’s quartet in detail.

[Answer]

Anscombe's quartet is a set of four seemingly unrelated data visualizations created by statistician Francis Anscombe in 1973. The purpose of this quartet is to highlight the importance of data visualization and the limitations of relying solely on summary statistics to understand data.

It consists of four sets of data, each containing 11 data points representing the relationship between an independent variable (X) and a dependent variable (Y). Despite having nearly identical summary statistics (mean, variance, correlation coefficient), the four datasets look very different when plotted visually. Visualization techniques like scatter plots can reveal hidden patterns, outliers, and non-linear relationships that might be missed by solely relying on numbers.

Anscombe's quartet is a valuable tool for understanding the importance of data visualization in statistical analysis. By incorporating visualization techniques, we can gain a deeper understanding of the dataset and could avoid misleading conclusions.

3. What is Pearson’s R?

[Answer]

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure used to quantify the linear relationship between two continuous variables. It represents the strength and direction of that association.

Pearson's R ranges from -1 to +1.

- +1: Indicates a perfect positive correlation. As one variable increases, the other variable increases proportionally.
- 0: Indicates no linear correlation. There's no predictable relationship between the two variables.
- -1: Indicates a perfect negative correlation. As one variable increases, the other variable decreases proportionally.
- Values closer to +1 or -1 represent stronger linear relationships, while values closer to 0 indicate weaker or no linear relationships.

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

   [Answer]

During the pre-processing step while building a machine learning model, we do scaling which is the process of transforming features in the data to a common range. This is an important pre-processing step.

Scaling helps put all features on a similar scale. If features have vastly different scales (e.g., one feature in thousands, another in tenths), algorithms can become sensitive to the units used and prioritize features with larger scales, so Scaling is used to prevent these issues and potentially improving the algorithm's performance. Also, many optimization algorithms used in machine learning, like gradient descent, rely on

calculating the magnitude (size) of gradients to update model parameters. Features with large scales can have a disproportionate effect on these calculations, hindering the optimization process. Scaling helps ensure all features contribute equally to the optimization process.

Normalization preserves the original data distribution (relative distances between values are maintained). It's useful when the absolute values of features might be significant, or when we want to bound the data within a specific range. While Standardization results in a standard normal distribution with a mean of 0 and a standard deviation of 1. It's useful when the distribution of features is important, and we want to focus on the number of standard deviations a data point lies from the mean.

We can use normalization if the presence of outliers or the absolute values of features is very important in the analysis. Similarly, we can use standardization if the algorithms being used are sensitive to feature means and variances.

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

An infinite VIF value signifies perfect collinearity. This means one variable can be expressed as an exact linear combination of other variables in the model. In simpler terms, the information contained in that variable is completely redundant with the information provided by other variables.

The formula for VIF involves dividing 1 by (1 - R^2), where R^2 is the coefficient of determination. In the case of perfect collinearity, R^2 will be equal to 1. Dividing by 1 minus 1 (0) resulting in infinity.

A model with variables exhibiting perfect collinearity can be unreliable. The coefficients of these variables will have inflated variances, making it difficult to assess their true significance and interpret their individual effects. By using correlation analysis, One of the solution could be the variables could be analysed for multi-collinearity and one of the variable could be dropped. In some cases, we can also be able to combine collinear variables into a single new variable that captures the shared information. But this requires good understanding of the domain.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression(3 marks)

[Answer]
A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to compare the quantiles (distribution) of two datasets. In linear regression, it helps to assess how well the errors (residuals) of the model follow a specific theoretical distribution, typically a normal distribution. It is it's a valuable technique for assessing one crucial assumption in linear regression: normality of residuals.

Linear regression relies on several assumptions, and normality of residuals is one of them. Here's why it's important:

Reliable Hypothesis Tests: Many statistical tests used in linear regression, like p-values for coefficient significance, assume normality of residuals. Deviations from normality can make these tests unreliable.
Confidence Intervals: Confidence intervals for the coefficients also depend on the normality assumption. Non-normality can lead to inaccurate confidence intervals.

After calculating the quantiles and plotting them as a pair whereby x-axis represents the quantiles of the theoretical normal distribution and the y-axis represents the corresponding quantiles of the residuals., If the residuals are normally distributed, the points on the Q-Q plot should roughly fall along a straight diagonal line. This indicates that the observed quantiles of the residuals closely match the expected quantiles of a normal distribution.