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 : Analysis has been done on the categorical variables. The following are some of the inference that I have obtained.

- The Number of bike booked is the highest in Fall season.
- The month of September has the hightest number of bookings followed by October and August
- An Increasing trend can be seen from March to October on number of bookings
- A Clear Weather has attracted more bookings while rainy weather has the lowest number of bookings
- There is a good increase in number of bookings in 2019 than previous year which shows the business is doing good.

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

Ans: When creating dummy variables from categorical variables in the context of linear regression or other modeling techniques, using drop_first=True is a common practice. The primary reason for using drop_first=True is to avoid multicollinearity, a situation where one predictor variable in a multiple regression model can be predicted from the others with a high degree of accuracy.

We also use it to decrease the number of Variables created while creating dummy variables. Let's say we have three types of values in the categorical column and we want to create dummy variables for that column. If one variable is not A or B, then it's obviously C. So we do not need the 3rd variable to identify the C.

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

Ans: Temp has the highest correlation with the target variable(0.63)

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

Ans :

I. Error terms should be normally distributed.
II. There should be insignificant multicollinearity among variables
III. There should be no visible pattern in residual values.

We validate the above rules by doing the following.

Residual Analysis:

We start by examining the residuals (the differences between the observed and predicted values) by plotting a histogram or a Q-Q plot. We do this to check for normality of residuals. A histogram or Q-Q plot that roughly follows a normal distribution indicates that the assumption of normality is met.

Scatterplots:

We plot the residuals against the predicted values or each predictor variable and then look for patterns or trends in the residuals. A random scatter of points around zero indicates homoscedasticity (constant variance of residuals). Patterns may suggest heteroscedasticity.

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

Ans: Based on my final model, the top three features contributing significantly towards explaining the demand of the shared bikes are

1. Temp
2. Year
3. Light rain

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a fundamental algorithm in machine learning, used for **predicting numerical values based on input features**. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).

a) **Model Definition:** The linear regression model is represented by a linear equation:

$y = mx + c$

where: y: dependent variable x: independent variable m: slope of the line c: y-intercept

b) Least Squares Method:

The goal of linear regression is to find the values of m and c that minimize the **sum of squared errors (SSE)**:

$SSE = \Sigma(yi - \hat{y}i)^2$

where:

- yi: actual value of the dependent variable for the i-th observation
- $\hat{y}i$: predicted value of the dependent variable for the i-th observation

There are various methods to find the optimal values of m and b, with **the least squares method being the most common**. This method involves solving a system of linear equations derived from the partial derivatives of SSE with respect to m and b.

c) Algorithm Steps:

- **Data Preparation**:
  - Collect data for independent and dependent variables.
  - Check for missing values and handle them appropriately.
  - Normalize or standardize the data if necessary.
- **Model Building**:
  - Choose the appropriate model (simple or multiple linear regression).
  - Initialize the values of m and c(e.g., randomly).
- **Model Training**:
  - Calculate the predicted values for each data point.
  - Compute the sum of squared errors.
  - Update the values of m and b using an optimization algorithm (e.g., gradient descent) to minimize the SSE.
- **Model Evaluation**:
  - Evaluate the performance of the model using metrics like R-squared, mean squared error, etc.
  - Analyze the model coefficients and interpret their meaning.
- **Model Tuning**:
  - If necessary, fine-tune the model to improve performance.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four data sets designed to highlight the limitations of relying solely on summary statistics to analyze data. Each data set has the same mean, variance, correlation coefficient, and linear regression line, yet they appear very different when plotted visually.

Purpose:

- Demonstrates the importance of data visualization in statistical analysis.
- Shows that summary statistics can be misleading if the underlying data distribution is not considered.
- Highlights the potential for outliers and other influential observations to distort statistical analysis.

Each data set consists of 11 points:

- **Data set 1:** Linear relationship with no outliers.
- **Data set 2:** Non-linear relationship with an outlier.
- **Data set 3:** Linear relationship with a point far from the regression line.
- **Data set 4:** Non-linear relationship with several outliers

3. What is Pearson’s R?

Ans: Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

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

Pearson's correlation is calculated using the following formula:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

where:

- x_i and y_i are the individual data points,
- $\bar{x}$ and $\bar{y}$ are the means of the variables x and y.

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

Ans: Scaling is a preprocessing technique used in data analysis and machine learning to transform numerical features of different scales into a common scale. The goal of scaling is to standardize the range of the independent variables or features of the dataset. This is important because many machine learning algorithms are sensitive to the scale of the input features. Features with different scales can lead to biased or incorrect predictions, as some features may dominate others simply due to their scale.

Why Scaling is Performed:

Algorithm Sensitivity: Some machine learning algorithms, like k-nearest neighbors or support vector machines, are distance-based and sensitive to the scale of the features. Scaling helps these algorithms perform better.

Convergence in Gradient Descent: For optimization algorithms like gradient descent, scaling can help in faster convergence by ensuring that the steps taken during optimization are more consistent across features.

Regularization: Regularization techniques, such as L1 or L2 regularization, can be sensitive to the scale of features. Scaling helps in ensuring that regularization penalties are applied uniformly.

Normalized Scaling:

- Also known as min-max scaling.
- Transforms data to a range of 0 to 1 or -1 to 1 by subtracting the minimum value and dividing by the range (maximum - minimum).
- Preserves the relative order of the original data points.
- Sensitive to outliers, which can significantly affect the scaling range.
- Useful when the data is not normally distributed.

Standardized Scaling:

- Also known as Z-score normalization.
- Transforms data to a mean of 0 and a standard deviation of 1.
- Subtracts the mean and divides by the standard deviation.
- Does not preserve the relative order of the original data points.
- Less sensitive to outliers than normalized scaling.
- Useful when the data is normally distributed or needs to be normalized for statistical analysis.

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

Ans: The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a regression analysis. High VIF values indicate that there may be a problematic degree of correlation between predictor variables, which can lead to issues such as unstable coefficient estimates and inflated standard errors.

a) Perfect Multicollinearity:

This occurs when one variable is perfectly linear combination of other variables in the model. This means one variable can be predicted exactly from the other variables, leading to an undefined variance and an infinite VIF.

b) Near Perfect Multicollinearity:

Even when perfect multicollinearity doesn't exist, high correlations between independent variables can significantly inflate the VIF value. If two or more variables are highly correlated, they share a lot of information, leading to inflated variances and potentially infinite VIF values.

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

Ans: A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the observed data against the quantiles expected from the theoretical distribution. In the context of linear regression, Q-Q plots are particularly useful for checking the normality assumption of residuals.

Use and Importance of Q-Q Plot in Linear Regression:

a) Normality Assumption:

- Purpose: One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) are normally distributed. Deviations from normality can impact the validity of statistical inferences.
- Use of Q-Q Plot: By comparing the observed residuals against the quantiles of a standard normal distribution in a Q-Q plot, you can visually assess whether the residuals follow a normal distribution. If the points on the Q-Q plot roughly align with a straight line, it suggests that the residuals are approximately normally distributed.

b) Identification of Outliers:

- Purpose: Q-Q plots can help identify outliers or heavy-tailed distributions in the residuals.
- Use of Q-Q Plot: If the tails of the Q-Q plot deviate from a straight line, it may indicate the presence of outliers or non-normality in the residuals. This can guide further investigation and potentially lead to model refinement.

c) Model Adequacy Checking:

- Purpose: Q-Q plots are part of a set of diagnostic tools used to check the overall adequacy of the regression model.
- Use of Q-Q Plot: A well-behaved Q-Q plot supports the assumption that the model is appropriate for the data. Deviations from the expected pattern may suggest issues with the model or violations of underlying assumptions.