Assignment-based Subjective Questions

- From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?

  Answer : Observations from above boxplots for categorical variables:

  The year box plots indicates that more bikes are rent during 2019.
  The season box plots indicates that more bikes are rent during fall season.
  The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
  The month box plots indicates that more bikes are rent during september month.
  The weekday box plots indicates that more bikes are rent during saturday.
  The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

- Why is it important to use **drop_first=True** during dummy variable creation?

  Answer : we have number of categories, we will use dummy variables. Dropping one dummy variable to protect from the dummy variable trap.

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

  Answer : atemp and temp both have same correlation with target variable of 0.63 which isthe highest among all numerical variables.

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

  Answer :
  Comparisons of models theoretical calculations and results
  Comparisons of models coefficients and predictions with theory
  Gathering and incorporating new data to check model predictions
  Cross-validation/Data splitting

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

  Answer :
  weathersit(negative correlation).
  yr(Positive correlation).
  temp(Positive correlation).

General Subjective Questions

- Explain the linear regression algorithm in detail.

Answer : It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear

relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Types of Linear Regression -

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

- Explain the Anscombe's quartet in detail.

Answer : Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

- What is Pearson's R?

Answer : The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

- What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?

Answer : It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1.

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Answer : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer : Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.