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. Following can be inferred from the Categorical Variables analysis:

- Most bikes were sold in the fall season.
- In 2019, more bikes were sold as compared to 2018.
- When weather was Clear and Partly Cloudy more bikes were taken by the customers
- Most bikes were taken by customer from months of Aug-Oct

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

Ans. When drop_first is true it removes the first column which is created for the first unique value of a column. If we don't set it as true, it'll keep the original column.

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

Ans.

- Temp & atemp has the highest correlation.
- Registered & cnt has the second highest correlation
- Casual & cnt also has very high correlation

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

Ans.

- Multivariate normality - We can validate the relationship by plotting the correlation matrix
- Linear relationship - We can plot pairplot to validate the Linear Relationship

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

Ans. Top 3 features contributing significantly towards the demands of share bikes are:

1. Yr
2. Windspeed
3. atemp

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans.
Machine Learning: It is the field of study that gives computers the ability to learn without being explicitly programmed. It is one of the most advancing technologies. Machine Learning can be further classified into 3 types:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning:

It is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labelled for a particular output. The model is trained until it can detect the underlying patterns and relationships between the input data and the output labels, enabling it to yield accurate labelling results when presented with new data.

There are many learning algorithms such as Linear Regression, Logistic Regression, etc.

Linear Regression:
It is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

The strength of the linear regression model can be assessed using 2 metrics:
1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

The assessment of model can be assessed with the help of the following:
1. t statistic: It is used to determine the p-value and hence, helps in determining whether the coefficient is significant or not.
2. F statistic: It is used to assess whether the overall model fit is significant or not. Generally, the higher the value of the F statistic, the more significant a model.
3. R-squared: The R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics. There are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It is used to show both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R?

Ans.

It is the covariance of the two variables divided by the product of their standard deviations.

The Pearson's correlation coefficient varies between -1 and +1:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

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

Ans. It is a step during 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**.**

The dataset which we use in the model sometimes contains very high values that need to be normalised. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modelling.

Difference between normalized scaling and standardized scaling

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Ans. If there is perfect correlation i.e. Correlation value =1 for a parameter, then VIF = infinity as VIF is proportional to inverse of correlation. 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.

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

Ans. 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, if 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.