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? (3 marks)

Ans:

a) yr: Bike sharing was more popular in 2019 than in 2018
b) Holiday: Fewer people hired bikes on a holiday
c) Weathersit_Fair_WX and weathersit_Manageable_WX increase the number of bikes being hired.
d) Winter Months such as mnth_Dec, mnth_Jan, and mnth_Nov decrease the number of Bikes being hired
e) mnth_Sep: People hired more bikes in the month of September
f) season_Summer and season_Spring: People hired fewer bikes in the Summer and Spring Season.

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

Ans: The reason for the following are:

a) To Reduce Multicollinearity among the Dummy Variables.
b) To make our data clean by getting rid of redundant variables.
c) To avoid Dummy Variable Trap and make the model easier to interpret.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: Going by the pair_plot, the target variable (cnt) has the highest correlation with Temp and atemp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: I validated the assumptions of Linear Regression By:

a) Checking if the Residual Errors are Normally Distributed
b) Checking if there is a linear relationship between the y-Test and the y- Predicted Variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

a) Weathersit_Fair_WX: When the weather is fair, it increases the bike being hired by 0.3175. Good weather encourages people to hire more Bikes.
b) Season_Spring: When it is spring, the bike hired decreases by 0.2527. Spring is still a cold season in certain areas, especially when the snow is melting.
c) yr: When the year is 2019, the bike hired increases by 0.2424. This could be cause hired bikes became more popular and/or people were motivated to exercise to keep fit, hence cycling became trendy.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: One of the popular machine learning algorithms is Linear Regression. Linear Regression is used for supervised learning. The target variables have to be given and the target variables should be numerical in nature. In regression models, the target variable (y) is predicted based on the independent variables(x variables). Linear Regression is highly useful in showing the relations between variables and forecasting.

Linear Regression can be of two types:

a) Simple Linear Regression: It is a linear regression that has just one independent variable which is used to predict the value of the dependent variable.
b) Multiple Linear Regression: It is a linear regression that has more than one independent variable which is used to predict the value of the dependent variable

The process to perform a Linear regression is:

- First, we understand the data and clean the data till it's in the correct format (EDA Process).
- Then, the data is divided into training set and test set.
- Next, we create dummy variables if we have categorical variables and rescale the data.
- After this,we use the training data set and conduct an RFE (Automated Approach) to select n number of variables if the number of variables is large.
- Then we will keep on creating models by deleting some variables (one by one) and/or adding some variables (again one by one) till we get a good r2 and adjusted r2, all the coefficients are significant and the VIFs of all the coefficients are not greater than 5.
- Finally, we evaluate our final model on the test data set and check if the residual errors are normally distributed and if the r2 and the adjusted r2 of the model on the test data set have a high value and are within 3% of each other.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans: Anscombe’s quartets consist of 4 datasets having almost the same simple statistical properties but shown to be wildly different when they are graphed. Statistician Francis Anscombe constructed Anscombe’s quartet in 1973 to emphasise why it is important to graph the data before analysing it and also to show how outliers can manipulate the statistical properties of the data

3. What is Pearson’s R? (3 marks)

Pearson's R also called Pearson's correlation coefficient, is used to measure the correlation between two variables. This method is frequently used in linear regression. The range of Pearson's R is between +1 and -1. A Pearson's R of -1 means that there is a perfectly negative and linear correlation between the variables while a Pearson's R of +1 states that there is a perfect negative linear correlation between the variables. Lastly, a Pearson's R of 0 means there is no correlation between the variables. The values between 1 and 0 indicate relative collinearity between the two variables.

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

Ans: Scaling is a method of compressing the data within a certain range. Scaling aid the model in learning the data by smoothing the flow of gradient descent and allows the algorithms to attain the minima of the cost function quickly.

Some of the differences between normalized scaling and standardized scaling:

a) Normalized scaling: Scaling uses maximum and minimum values of the variable
Standardized scaling: Scaling uses mean and Standard deviation

b) Normalized scaling: Useful when the variables are of various scale
Standardized scaling: Useful when variables of the mean are required to be zero and have only one standard deviation.

c) Normalized scaling: The scale has a range of between 1 and 0
Standardized scaling: No such fixed range.

d) Normalized scaling: Affected by outliers
Standardized scaling: Not affected by outliers

e) Normalized scaling: Useful when the distribution of the variables is unknown
Standardized scaling: Useful when the distribution of the variables are Normal or Gaussian

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

Ans: When the VIF is infinite, it means that there is a perfect correlation between the independent variables and dependent variables in the model. For this, the r squared of the model will be exactly 1. The formula for VIF is 1 / 1- r squared which will be 1/0 which is infinity. High VIF tells us that there is a strong presence of multicollinearity among the variables in the model.

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

Ans: A Q-Q plot is a scatterplot, which is formed by plotting two sets of quantiles. It is used to determine if the two samples of data came from the same population or not. It can also help us to know if the residuals are normally distributed or not. It can also inform us about the skewness of the distribution. These abilities of Q-Q plot help with linear regression especially as we have to treat the test dataset and test dataset as two different datasets and also confirm certain assumptions that the linear regression has to follow in order for the model to be valid.