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?

Ans1: After cleaning and obtaining derived metrics, I had 5 categorical columns i.e. year, month, season, weather and working/non-working. Out of those when I ran a correlation matrix, I saw that for seasons, winter had the highest co-relation with variable cnt, for weather, misty and light snow weather negatively impacts the demand of bikes, while working/non-working does not seem to be impacting the demand much.

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

Ans 2: When we use the get_dummy() method, pandas library will create a column for all the values present in the categorical column. But since we actually need just n-1 columns (n being the number of unique values in the categorical column) when we use drop_first=true, pandas library will drop the first column and we will have n-1 columns in our dataset. This way one extra column will not add complexity to the model.

For e.g. suppose a housing dataset has a column 'furnishing status' with 3 values 'furnished', 'semi-furnished' and 'unfurnished'. Then while using dummy variables, if we just have 2 columns 'furnished' and 'semifurnished', then the values '1, 0' would imply furnished, '0,1' would imply semi-furnished and '0,0' would automatically imply unfurnished.

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

Ans 3: As per the pair plot, Temperature (aTemp) has the highest co-relation with the target variable.

When looking at the exact corelation co-efficient using a heatmap, it shows a co-relation co-efficient of 0.63 with variable 'cnt'.

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

Ans 4: As per Linear Regression theory we have the following assumptions:

Assumption 1: Linear Regression assumes that there is a Linear relationship between the independent variables. - For this I plotted the graphs of all the input variables against the target variable. They seemed to be having a linear relationship.

Assumption 2: There is little or no collinearity among input variables. - This was validated by analysing the VIFs of all my input variables.

Assumption 3: Error is normally distributed and mean is 0 and std deviation 1. - For this I did a distribution plot of all the error terms in my model. They formed a normal distribution curve with mean 0 and standard deviation of 1.

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

Ans 5: In my final model, I was left with the below variables finally contributing towards demand of bikes. Their co-efficients are as below:

yr : 2075.3028

atemp : 5295.4750

windspeed : -1490.9861

Misty Cloudy : -540.6742

summer : 609.2490

winter : 909.5742

Hence Temperature, Windspeed and Year are the top 3 features contributing significantly towards explaining the demand of shared bikes, since they have the highest co-efficients in the model results.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans 1: Linear regression is an algorithm that explains the relationship between a dependent variable and one or more independent variables through a linear relationship. This algorithm is used to predict the outcome of future events using past data.

It is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

There are certain key assumptions in Linear Regression algorithm:

a. First important assumption of linear regression is that the dependent and independent variables should be linearly related.
b. The second assumption relates to the normal distribution of residuals or error terms, i.e., if residuals are non-normally distributed, the model-based estimation may become too wide or narrow.
c. The third assumption relates to multicollinearity, where several independent variables in a model are highly correlated. More correlated variables make it difficult to determine which variable contributes to predicting the target variable.
d. Another assumption of linear regression analysis is referred to as homoscedasticity. Homoscedasticity relates to cases where the residuals (error terms) between the independent and dependent variables remain the same for all independent variable values.

2. Explain the Anscombe's quartet in detail.

Ans 2: Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

3. What is Pearson's R?

Ans 3: Pearson Correlation is a statistical method that measures the similarity or correlation between two data objects by comparing their attributes and calculating a score ranging from -1 to +1. A high score indicates high similarity, while a score near zero indicates no correlation. This method is parametric and relies on the mean parameter of the objects, making it more valid for normally distributed data. It is calculated by dividing the covariance of the variables by their standard deviations.

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

Ans 4: Scaling is the process of converting data points within a range to suit the range of all other variable. E.g. if salary is provided in lakhs and age is provided in years, these data points are on a very different scale. When we build a Linear model on such differing scales, the coefficients are not easy to interpret. Hence through scaling we can bring them all in one scale, such as 0-1.

Normalized scaling is also called min-max scaling. It brings all the data in the range 0 to 1.

Standardized scaling 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).

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

Ans 5: VIF represents the correlation between the independent variables. It is calculated as below:

$$VIF = 1/1-R^2$$

This means when R^2 is 1 i.e. 100%, VIF will be infinite. That means the model is working with no error. This implies overfitting.

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

Ans 6: A Q-Q plot is also called quantile-quantile plot. It is created by plotting a scatter plot of two quantiles against each other. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

It helps in assessing the similarity between theoretical and sample quantiles. E.g. if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.