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.

- ✓ Bike demand in the fall is the highest.
- ✓ Bike demand takes a dip in spring.
- ✓ Bike demand in year 2019 is higher as compared to 2018.
- ✓ Bike demand is high in the months from May to October.
- ✓ Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
- ✓ The demand of bike is almost similar throughout the weekdays.
- ✓ Bike demand doesn't change whether day is working day or not.

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

Ans.

- ✓ It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
- ✓ For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it
- ✓ It is also used to reduce the collinearity between dummy variables.

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

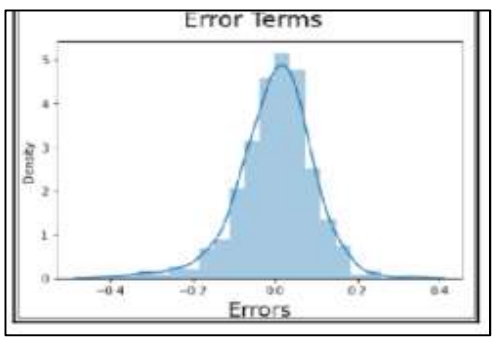
Ans.

- ✓ atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

Ans.

- ✓ We validate the assumptions of the linear Regression by plotting a distplot of the residuals and analyzing it to see if it is a normal distribution or not and if it has a mean=0. The diagram below shows that it is normally distributed with mean = 0



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 final model top three features contributing significantly towards explaining the demand are:

- ✓ Temperature (0.552)
- ✓ weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
- ✓ year (0.256)

General Subjective Questions

Explain the linear regression algorithm in detail?

Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable)

Explain the Anscombe's quartet in detail.?

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

What is Pearson's R?

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 scaling and standardized scaling?

Ans:

What?

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

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Difference:

In statistics, Standardization is the subtraction of the mean and then dividing by its standard deviation. In Algebra, Normalization is the process of dividing of a vector by its length and it transforms your data into a range between 0 and 1.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

Answer: The value of VIF is calculated by the below formula:

1 / (1 – R Square)

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.