###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?

Answer: Season, year, month, holiday, weathersit, and workingday are categorical variables in this dataset. Seasons have a huge impact on our model both throughout and outside of these years. Out of them, the variables listed below have a substantial impact on our model. $count=0.4914\times temp+0.0916\times September+0.0645\times Saturday +0.0527\times summer+0.0970\times winter+0.2334\times Year+0.0566\times workingday-0.03041\times lightsnow-0.0786\times mistcloudy-0.065\times spring.$

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

Answer: When we use the get dummies method to construct dummy variables, the columns are created based on the columns' unique values. However, the first column displays the same result as the other columns. As a result, we eliminate the first column to reduce redundancy.

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

Answer: Among all numeric variables, temp 0.627044 has the highest value.

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

Answer: We calculate the R square of the model once we've built it. The model's confidence is represented by R square. We also have the p value, which is nothing more than the correlation between the target variable and the p value, so the lower the p value, the better the model. We validate the LR model using residual analysis of the train data and regression plots.

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

Answer: 1.temp = 0.4914 2. Year = 0.2334 3. $winter$ = 0.0970 4.$September$ = 0.0916

###General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a supervised learning machine learning algorithm. It carries out a regression task. Based on independent variables, regression models a goal prediction value. It is mostly utilised in forecasting and determining the link between variables. Different regression models differ in terms of the type of relationship they evaluate between dependent and independent variables and the amount of independent variables they employ. Linear regression is used to predict the value of a dependent variable (y) based on the value of an independent variable (x). As a result of this regression technique, a linear relationship between x (input) and y (output) is discovered (output). As a result, the term Linear Regression was coined. In the diagram above, X (input) represents job experience and Y (output) represents a person's wage. For our model, the regression line is the best fit line.

2. Explain the Anscombe's quartet in detail.

Answe : Anscombe's Quartet is a collection of four data sets that are essentially equal in terms of simple descriptive statistics, but have some quirks in the dataset that trick the regression model if formed. When plotted on scatter plots, they have extremely different distributions and appear differently. Francis Anscombe, a statistician, built it in 1973 to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of other data on statistical features. There are four data set plots with virtually identical statistical observations and statistical information, including variance and mean of all x,y points in all four datasets. This emphasises the importance of visualising data before applying various algorithms to create models from it, implying that data features must be plotted in order to see the distribution of samples, which can aid in identifying various anomalies in the data such as outliers, data diversity, linear separability, and so on. Furthermore, Linear Regression can only be used to fit data with linear connections and is incapable of handling any other datasets.

3. What is Pearson's R?

Answer: The Pearson product-moment correlation coefficient (abbreviated as r) is a measure of the strength of a linear relationship between two variables. A Pearson product-moment correlation seeks to create a line of best fit across the data of two variables, and the Pearson correlation coefficient, r, indicates how far these data points are from this line of best fit (i.e., how well the data points match this new model/line of best fit).

The Pearson correlation coefficient, r, can be anything between +1 and -1. A value of 0 implies that the two variables have no relationship. A positive relationship is shown by a value greater than 0; that is, when the value of one variable rises, so does the value of the other variable.

A negative relationship is indicated by a value less than 0; that is, as the value of one variable rises, the value of the other variable falls. The diagram below depicts this: Different Values of the Pearson Coefficient How can we use the Pearson correlation coefficient to determine the strength of an association? The Pearson correlation coefficient, r, will be closer to +1 or -1 depending on whether the relationship is positive or negative, depending on the strength of the association between the two variables. A score of +1 or -1 indicates that all of your data points are contained on the line of best fit, and no data points exhibit any deviation away from it. Variation around the line of best fit is indicated by r values between +1 and -1 (for example, r = 0.8 or -0.4). The closer r is to 1, the better.Variation around the line of best fit is indicated by r values between +1 and -1 (for example, r = 0.8 or -0.4). The larger the variation around the line of best fit, the closer r is to 0. The graphic below depicts many relationships and their correlation coefficients: Different Pearson Correlation Coefficient values

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

Answer : The majority of the time, the acquired data set contains features with a wide range of magnitudes, units, and ranges. If scaling is not done, the algorithm will only consider magnitude rather than units, resulting in erroneous modelling. To solve this problem, we must scale all of the variables to the same magnitude level. Scaling only changes the coefficients and not the other parameters such as the t-statistic, F-statistic, p-values, R-squared, and so on. Normalization/Min-Max Scaling: It gathers together all of the data in the 0 to 1 range. sklearn.preprocessing. MinMaxScaler is a Python module that aids in the implementation of normalisation.

Standardization Scaling: When values are replaced by their Z scores, this is referred to as standardisation. It converts all of the information into a conventional normal distribution with a mean of zero and a standard deviation of one.

sklearn.preprocessing.scale is a Python module that aids in the implementation of standardisation. Normalization has the disadvantage of losing some data information, particularly about outliers, when compared to standardisation.

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

Answer: VIF = infinity if there is perfect correlation. This demonstrates that two independent variables have a perfect correlation. We get R2 = 1 in the event of perfect correlation, which leads to 1/(1-R2) infinite. To overcome this issue, we must remove one of the factors that is producing the perfect multicollinearity from the dataset.

An infinite VIF value suggests that a linear combination of other variables may exactly express the related variable (which show an infinite VIF as well).

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

Answer: Q-Q charts (Quantile-Quantile plo ll below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it.

Q Q plots are used to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45 degree angle is drawn; if the two data sets are from the same distribution, the points will fall on that reference line.

The 45 degree reference line on a Q Q plot: plot qq

The points in the Q–Q plot will roughly lie on the line y = x if the two distributions being compared are similar. The points in the Q–Q plot will almost always, but not always, sit on a line if the distributions are linearly connected.

In [ ]:

```

```

In [ ]:

```

```