Assignment-based Subjective Questions

Question 1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer - Top variables that are seen effecting and benefitting the Bike Rental count are as follows:

- Spring season: -0.6842
- Temperature: 0.3999
- Mist: -0.3647
- Sun: 0.2749
- working_day: 0.2327

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

Answer - It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer - It is observed that atemp and temp are highly correlated and one can be dropped to avoid multicollinearity.

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

Answer - Concluded based upon the Linear Regression -

- The error terms are normally distributed.
- The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.
- The predicted values have linear relationship with the actual values.

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

Answer –

1. Bike Rentals are more during the Fall (Monsoon) season
2. Bikes seem to be rented more in Partly cloudy weather.
3. Bikes seem to be rented more on working days

General Subjective Questions

Question 1 - Explain the linear regression algorithm in detail.

Answer - Linear Regression 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. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

y= mx+c

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Question 2 - Explain the Anscombe's quartet in detail

Answer - Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but 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.

Illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Question 3 - What is Pearson's R?

Answer - The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer - 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.

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. 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

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.

Min Max Scaling: x = {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 zero and standard deviation one

Standardisation: x = {x-mean(x)}/sd(x)

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

Answer - If there is perfect correlation, then VIF = infinity.

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. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6 - What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer - Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages -

It can be used with sample sizes also

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.