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:- We could see that In the Fall season the demand of bike increases. In the year 2019 bike demands significantly higher than in 2018. We could see that from June to October The bike demand is higher than the other months. On the day of Holidays the overall bike demand is less probably because of people like to stay at home on those days. From the weather data we can infer that as the weather goes bad the bike demand rapidly decreases.

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

Ans:- To create dummy variables of categorical variable which has n unique value we need n-1 dummy variable. Hence we drop the first column.

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

Ans:- variable ‘temp’ and ‘atemp’ have highest correlation with target variables

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

Ans:- We check R2 and adjusted R2 value should explain most of that predicted data. All the p-values should be nearly zero. From Residual analysis we can say Residuals is normally distributed. To check no perfect multicollinearity we ensure that the all the significant variables have low VIF value(vif<5).

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

Ans:- Winter season, Temperature and year 2019

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:- Linear regression is one of the most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. It is mostly used for finding out the relationship between variables and forecasting.

2. Explain the Anscombe's quartet in detail.

Ans:- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R?

Ans:- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Ans:- 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.

Normalized scaling - It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardized scaling - Standardization 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:- 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)

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

Ans:- 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.