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?

→Month 8, 9, 10 shows high in rent
Summer and Fall shows high in rent
2019 year is showing high in rent
Working day prefer over weekend

==
2.Why is it important to use drop_first=True during dummy variable creation?
➔ categorical data needs to scale and can be shown as less one column in binary form eg: quarter month can be shown as

0 0 -> 1st quartor
0 1-> 2nd quartor
1 1-> 3rd quartor

3.Looking at the pair-plot among the numerical variables, which onehasthe highest correlation with the target variable?
-> Temp

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
->R squared should be high between target and predictors and p value should be below 0.005 for every predictor.

VIF should be less than 5 between predictors

==

5.Based on the final model, which are the top 3 featurescontributing significantly towards explaining the demand of the shared bikes?
Month 9, 8 and 6

General Subjective Questions

1.Explain the linear regression algorithm in detail.

=>Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

2.Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed

3.What is Pearson’s R?

→The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short). When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (ρ). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables. The scatterplot shown on this page depicts such a relationship. It is a positive relationship because high scores on the X-axis are associated with high scores on the Y-axis.

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

→Scaling performed to range columns in equal fashioned

The terms *normalization* and *standardization* are sometimes used interchangeably, but they usually refer to different things. *Normalization* usually means to scale a variable to have a values between 0 and 1, while *standardization* transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a **z-score**, and data points can be standardized with the following formula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A z-score standardizes variables.

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

=>**infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

=>The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.