ASSIGNMENT SUBJECTIVE QUESTIONS:

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

Ans:

1)The demand of bike is less in the month of spring when compared with other seasons
2)The demand bike increased in the year 2019 when compared with year 2018.
3)Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
4)The demand of bike is almost similar throughout the weekdays.
5)There is no significant change in bike demand with working day and nonworking day.
6)The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall.
7)We do not have any date for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

Ans:

A variable with **'n'** values can be represented by n-1 variables. If we remove the first column then also, we can represent the data. If the value of the variable from 2 to n is 0, it means the value of 1st variable is 1.

Eg : If I take a bike with three scenarios such as ON, OFF, RUNNING and represent them in binary format then we have:

BIKE	**OFF**	**ON**	**RUNNING**
OFF	**1**	**0**	**0**
ON	**0**	**1**	**0**

RUNNING	**0**	**0**	**1**

Even if we remove the first column we can clearly understand what the data is trying to explain us in the tabular format.

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

Ans.

‘temp’ has the highest correlation coefficient of 0.68.

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

Ans.

By plotting the residuals distribution. It came out to be a normal distribution of value 0.

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

Ans.

The following are the top three features contributing significantly towards explaining he demands of the shared bikes:

1. temp : 0.518
2. year : 0.232
3. weathersit Light rain : -0.287

General Subjective Questions:

Q.1 Explain the linear regression algorithm in detail.

Ans:

A Linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable only to numerical variables.

While doing linear regression the following steps need to followed:

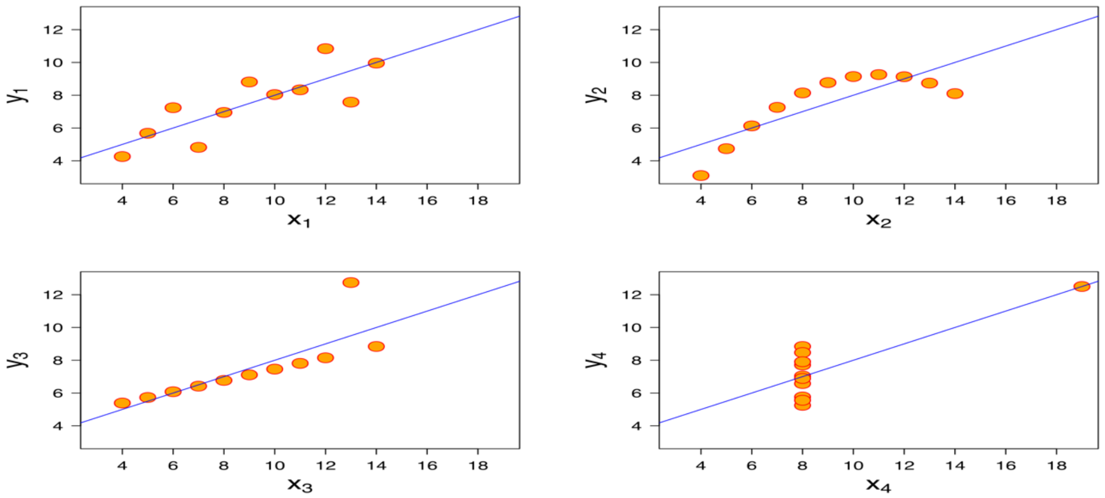
1. The data set is divided into training and test data.
2. The train data is divided into dependent and independent datasets.
3. A linear model is fitted using the training dataset.
4. In case there are multiple variables then there is hyperplane instead of a line.
5. The predicted variable is then compared with test data and assumptions are drawn accordingly.

Q.2 Explain the Anscombe's quartet in detail.

Ans.

Anscombe's quartet comprises of 4 data sets that have nearly identical simple descriptive statistics but have different distribution when visualised graphically.

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
2. The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.



3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Q.3 What is Pearson's R?

Ans.

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.

1. A value of 1 means a total positive linear correlation.
2. A value of 0 means no correlation.
3. A value of -1 means total negative correlation.

Q.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.

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…

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

Ans.

Formulae for $\mathbf{VIF_i: 1 / (1 - R_i^2)}$

If R^2 is 1 then VIF becomes infinite. It means that there is perfect correlation between the features.

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

Ans.

A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two sets of data came from the same distribution. It is visual check of data. If the data is from same source then the data will visualise as straight line.