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

The categorical variable in the dataset were season,weathersit,holiday,mnth,yr and weekday.

1. **Season -** The spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
2. **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable.Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
3. **Holiday -** rentals reduced during holiday.
4. **Mnth -** September saw highest no of rentals while December saw least.This observation is on par with the observation made in weathersit.The weather situation in december is usually heavy snow.
5. **Yr** - The number of rentals in 2019 was more than 2018

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

If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller.

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

temp" and "atemp

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

Residuals distribution should follow normal distribution and centred around 0

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

1.temp - coefficient : 0.491508

2.yr - coefficient : 0.233482

3.weathersit_Light Snow & Rain - coefficient -0.285155

1. Explain the linear regression algorithm in detail.

Linear regressioin is the algorithm for supervising machine learning used for prediction of numeric values. Most basing equation of linear regression of one variable is y=mx+c

This type of equation is linearly related with the depandant variable(x).

Regression is performed when the dependent variable is of continuous data type and Predictors

or independent variables could be of any data type like continuous, nominal/categorical etc.

Regression method tries to find the best fit line which shows the relationship between the

dependent variable and predictors with least error

Categories:

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using

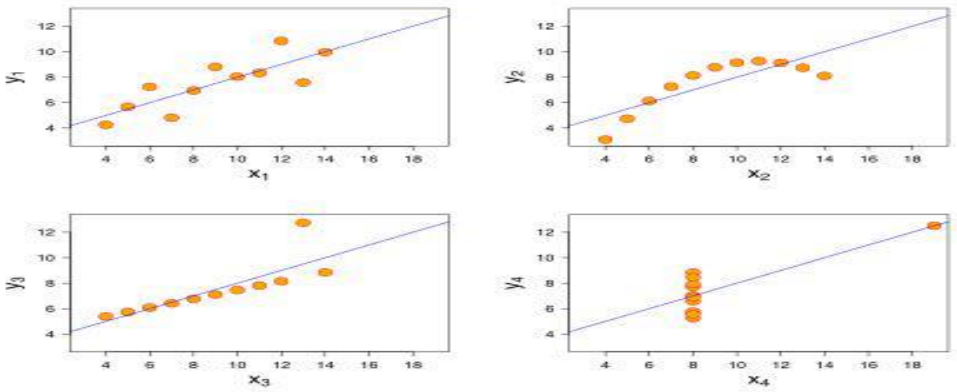
only one independent variable.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted

using multiple independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look
totally different when plotted on a graph.It was developed to emphasize both the importance of
graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution 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.
- 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.

3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we *draw a line graph to represent the data?*
r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association

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

Feature **scaling** is a method used to normalize or standardize the range of independent variables
or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

VIF - the variance inflation factor - The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will
have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions.A Q-Q plot is a scatterplot created by plotting
two sets of quantiles against one another. If
both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?