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

- More bookings are observed in year 2019 than in 2018.
- During an year, the bookings are observed to increase from Jan to Jun and slowly reduce from Jul to Dec.
- During an year, less number of bookings are observed during **spring** season.
- More bookings are oberved during **summer** and **fall** seasons.
- More bookings are observed during better weather conditions and vice versa.
- Also during an holiday the average number of bookings appear to be lower than on a non holiday. This suggests many people may be using their bikes for commuting to work space.
- No difference found in distribution of bookings across all weekdays.
- Also no difference found in median bookings between a working day and a non workin day.
- Hence the columns **weekday**, **workingday** can be removed
- Above observations indicate the categorical variables: **season**, **yr**, **holiday**, **weathersit**, **mnth** have an effect on target variable **cnt**.

Q) Why is it important to use drop_first=True during dummy variable creation?

- Dummy variables are created, in general, to detect the presence of a category.
- When creating dummy variables, one of the variable can be derived from the rest of all other variables represent their non presence by holding value of zero(0).
- It also reduces correlations created within dummy variables.

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

- Column **registered** has highest correlation with target variable.

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

- By visualizing the residuals (difference in model predictions and actual values) of train set have followed a normal distribution or not.
- We can also use statistical tests like Shapiro wilk test for verifying the normality of residuals

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

- The top three features contributing towards the explaining the demand of shared bikes are: casual, year and weather situation.

Q) Explain the linear regression algorithm in detail?

- Linear regression algorithm is used to predict the value of a variable based on the value of one or more other variables. The variable which is predicted is called the dependent variable. The variables which are used to predict the other variable's value are called the independent variables.
- A linear equation is generally represented as

$$Y = \beta 0 + \beta 1 * x$$

Where

1. Y is the response or the target variable
2. x is the independent feature
3. β1 is the coefficient of x
4. β0 is the intercept

- The aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) is as small as possible. Error is the distance between the points to the regression line.

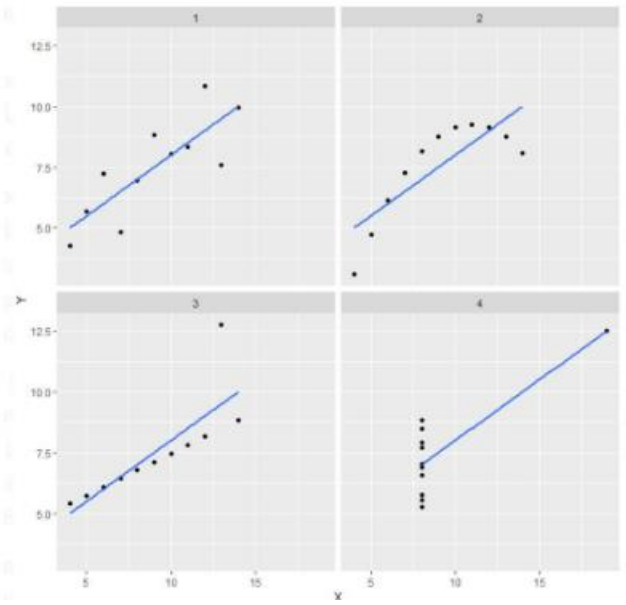
Q) Explain the Anscombe's quartet in detail?

- 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.
- The summary of descriptive statistics of these standard data sets is shown below

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

- When linear regression is performed on these data sets, it appears as shown below



- The above plots show that they are different even when their descriptive measures are same. So it is always important to visualize the date.

Q) What is Pearson's R?

- Pearson's R is simply known as correlation coefficient.
- It is the ratio between covariance of two variables and the product of their standard deviations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

- It shows the relation between two variables.
- It's value ranges between -1 and 1.
- -1 indicates a high negative correlation and +1 indicates a very high positive correlation between two variables.

Q) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is a data pre processing step performed to reduce high magnitudes of different variables into a specific range of values.
- Scaling helps in preforming algorithm calculations at a faster pace.
- It also transforms variables captured in different units into same scale.
- Normalized scaling refers to reducing the feature values to a specific range such as 0 to 1.

- Whereas Standarized scaling refers to transformation of features by subtracting from mean and dividing by standard deviation. This results in normal distribution

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

- Variance inflation factor is defined as $1 / (1-R^2)$.R^2 represent
- Infinite value of VIF indicates there is a perfect correlation between two independent variables. In case of perfect correlation, R^2 value is 1.

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

- 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.
- In linear regression it helps 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.