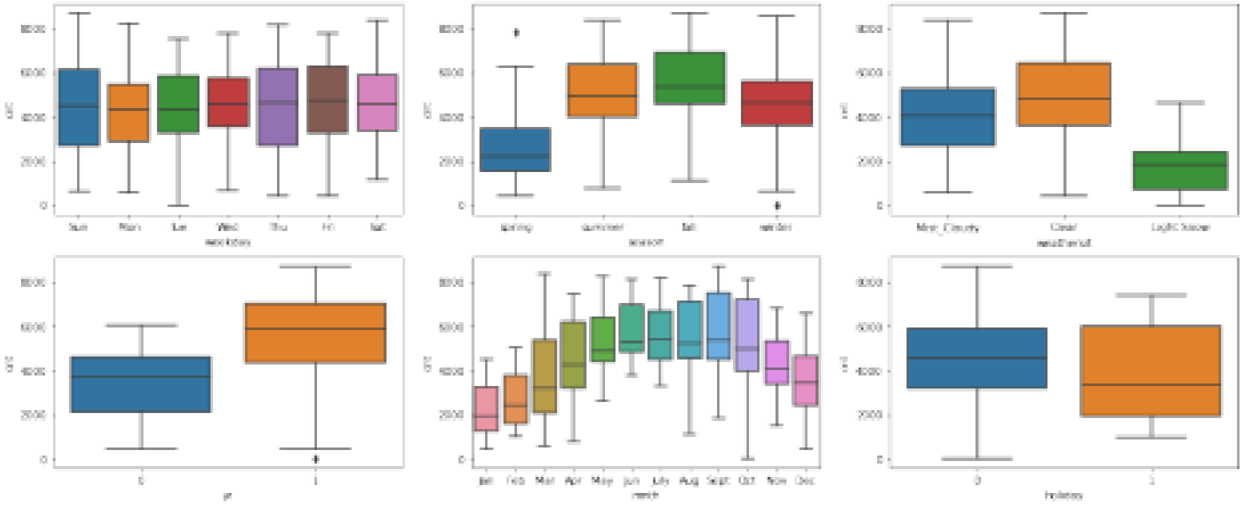
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Categorical variables mnth, yr, weekday, holiday, season and weathersit were observed. From the boxplots plotted, we can conclude that;

- **mnth**: demand increases from January to September and then gradually decreases towards end of the year (winter months)
- **yr**: demand is increasing year-on-year
- **season**: demand is highest in fall and lowest in spring □ **holiday**: demand is more on non-holidays as compared to holidays
- **weathersit**: demand is highest when weather is clear and lowest during light snow.

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

Helps to reduce extra column/s created during dummy variable creation to reduce the correlations created among dummy variables.

Ex: Drop initial variables ('season','mnth','weekday','weathersit') and keep Dummy variables ('spring', 'summer', 'winter', 'August', 'December', 'February', 'January', 'July', 'June', 'March', 'May', 'November', 'October', 'September', 'Monday', 'Saturday', 'Sunday', 'Thursday', 'Tuesday', 'Wednesday', 'Light Snow', 'Mist')

Helps to reduce correlations in the given variables.

Ex: Drop 'atemp' variable and keep 'temp' variable

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

From pair-plot ‘temp' variable has the highest correlation with Target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- **Linear relationship:** Strong linear relationship observed in the pair plots between ‘temp’ and ‘cnt’. The same was confirmed in the final model summary. The coefficient of ‘temp’ was **>0.5**
- **No or little multicollinearity:** Variables with **p-values>0 and VIF> 0.05** were dropped from the final model
- **No auto-correlation:** Variables ‘temp’ and ‘atemp’ had strong correlation. The later was

dropped from the final model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Yr, windspeed, Light Snow

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a method of finding the best straight-line fitting to the given data, - finding the best linear relationship between the independent and dependent variables.
This model is to find a linear relationship between the input variable(s) X and the single output variable y.

- Simple linear regression: When there is only single independent/feature variable X then it is called as simple linear regression.
- Multiple linear regression: When there are multiple independent/feature variables Xi then it is called Multiple linear regression.

The independent variable is also known as the predictor variable. The dependent variables are also known as the output variables.

$Y = \beta_0 + \beta_1 X$ (SLR)

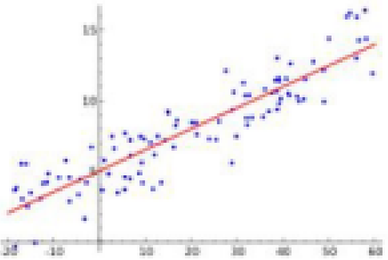
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$ (MLR)

Where:

Y = how far up ↑ and X = how far along →

$\beta_1, \beta_2 .. \beta_p$ = Slope or Gradient (how steep the line is)

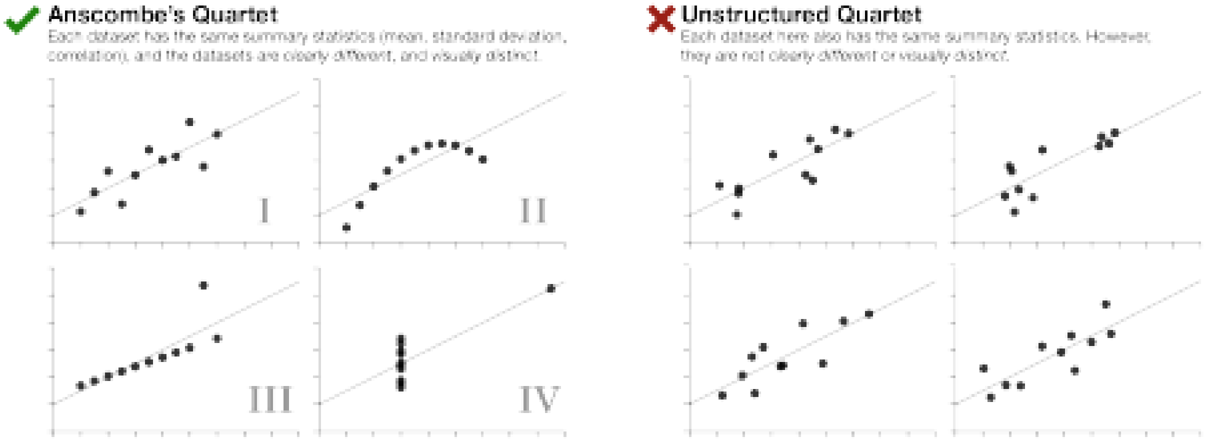
β_0 = value of Y when X=0 (Y-intercept)



As part of linear regression, there can be multiple lines which can be drawn from the data points as part of a scatter plot but a regression model can help to identify the model that is the best fit line from the data points.

2. Explain the Anscombe's quartet in detail. (3 marks)
Anscombe's Quartet demonstrates the importance of data visualization. An effective (and often used) tool used to demonstrate that visualizing your data is in fact important is Anscombe's Quartet. Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar. However, after visualizing (plotting) the data, it becomes clear that the datasets are markedly different. The effectiveness of Anscombe's Quartet is not due to simply having four different datasets which generate the same statistical properties, it is that four clearly different and visually distinct datasets are producing the same statistical properties. In contrast the "Unstructured Quartet" on the right in below Figure also shares the same statistical

properties as Anscombe's Quartet, however without any obvious underlying structure to the individual datasets, this quartet is not nearly as effective at demonstrating the importance of visualizing your data.

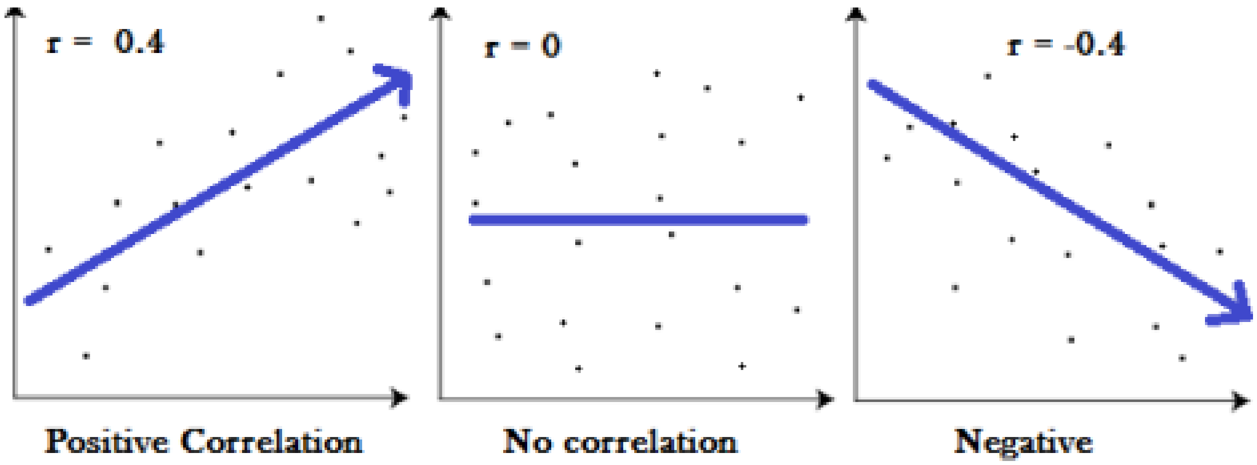


3. What is Pearson’s R? (3 marks)

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression. The formula below returns a value of ‘r’ between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$



Positive Correlation

No correlation

Negative

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

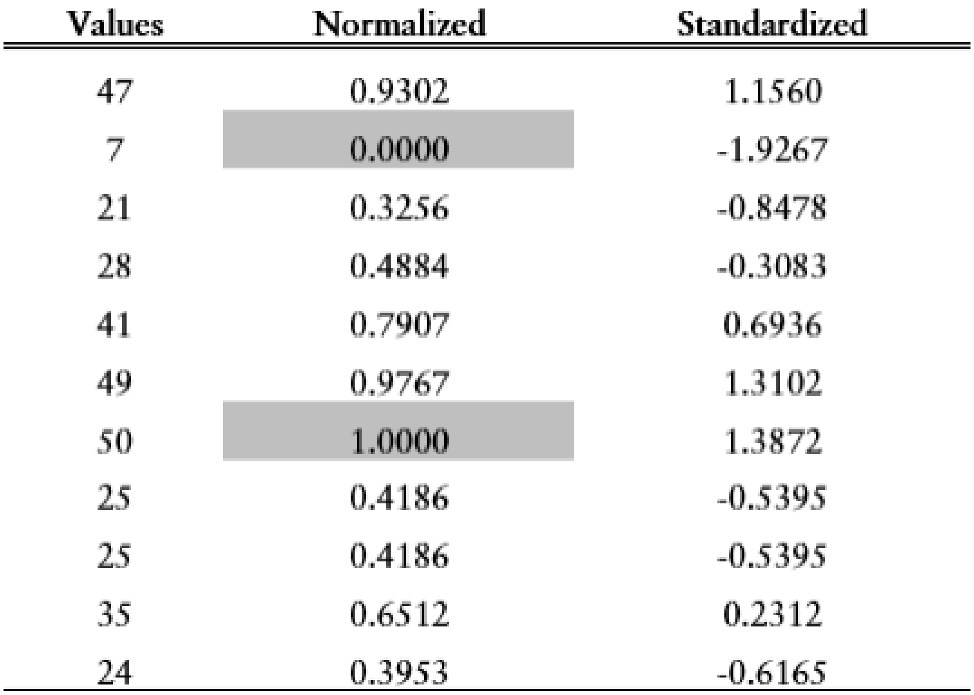
   Feature scaling or simply scaling means adjusting data that has different scales so as to avoid biases from big outliers. The most common techniques of feature scaling are Normalization and Standardization.

   Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   While Standardization transforms the data so as to have zero mean and a variance of 1. The table below compares raw data with its two transformations, the second column is processed through normalization and the third column is calculated using the standardization function:

$$x_{new} = \frac{x - \mu}{\sigma}$$



Values	Normalized	Standardized
47	0.9302	1.1560
7	0.0000	-1.9267
21	0.3256	-0.8478
28	0.4884	-0.3083
41	0.7907	0.6936
49	0.9767	1.3102
50	1.0000	1.3872
25	0.4186	-0.5395
25	0.4186	-0.5395
35	0.6512	0.2312
24	0.3953	-0.6165

Feature scaling does not alter how data is shown. If you plot the three columns above, you will get exactly the same figure. This step is vital for the success of any machine learning model with the exception of the Random Forest algorithm who can be run without the need to scale data although it's always best practice to do so. Scaling (precisely Normalization) is also helpful to detect historical highs and lows.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
Variation Inflation Factor (VIF) calculates how well one independent variable is explained by all the other independent variables combined.
The common heuristic we follow for the VIF values is:
   - **10**: Definitely high VIF value and the variable should be eliminated.
   - **5**: Can be okay, but it is worth inspecting.
   - **< 5**: Good VIF value. No need to eliminate this variable.

$$VIF_1 = \frac{1}{1 - R^2}$$

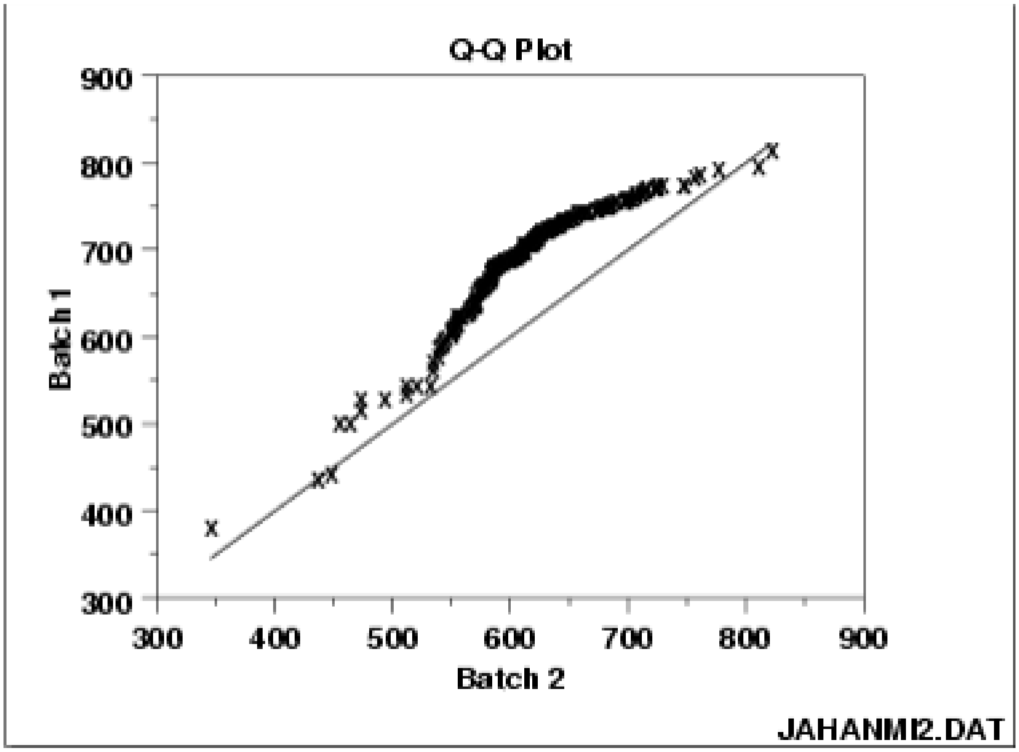
R-square in this formula is the coefficient of determination from the linear regression model which has:

- X1 as dependent variable
- X2 and X3 as independent variables

R-square =1 means perfect correlation. If this is the case, then VIF will be **infinity**.

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.



Q-Q plot 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 the same distributions.