Assignment based subjective question's Answers:

1. It is observed that,

- spring season have less cnt of bike demand compared to summer & fall

- 2019 have more bike demand then 2018

- Bike demand have also relation on month where 1 & 12 month have less demand and June & July have more demand

- As per wheathersit box graph data, Clear wheathersit have more demand and Light Rain have less demand among the category

2. As per dummy variable we can always can work with N -1 count of columns and

dropped column can be considered when all other columns value is 0.

drop_first=True will drop first column and apply the N - 1 logic during splitting the column.

3. As per pairplot graph, after dropping registered and casual,

we can see 'temp' has highest correlation with target variable.

4. After building the linear regression model. below points are checked.

- Checked p-value of each variable

- Whose p-value is > 0.05 is removed from model

  as they are denoting that it is less significant in model prediction.

  and proves that Hypothesis of straight line is near to B1 (Beta 1) = 0

- Calculated VIF and removed the variable one by one whose VIF is > 5

  As VIF calculation is used to understand how well on independent variable is

  explained by all other independent variable combined.

  If we have multicollinearity in the variables than variable can be explained by

  other variable and hence variable is not require anymore in model building.

- Although multicollinearity will not impact on model, we should understand this from domain

  knowledge to understand more in detail.

5. As per my model "temp" , "yr" , "summer" are most significant towards explaining shared bike.

- As "temp" is highly correlated with target variable and the relation is also visually seen in pairplot
- yr is highly correlated with target variable and during manual model building adding "yr" variable shown significant improvement in R-squared value as well
- summer is highly correlated and also shows low p-value to validate this.

General Subjective Objective Answers:

1. Linear regression algorithm is one of the supervised algorithm to predict target value based on linear pattern independent variable.

We can apply algorithm based on below assumption

- target variable should be continous variable
- input can be catagorical or continous
- The relationship between dependent and independent variable should be linear

It is also called line fitting algorithm.

where straight line can be explained, $y = mx + c$

where, m is slop and c is intercept

- We can calculate and find best line via two method
  -- Cost function Method - By differenciating
  -- Gradiant Decent Method - By recurring with different slop

2. Anscombe's quartet comprises four dataset.

Some times statistically the data might look simillar but quite different in visualizing.

This theory explains important of graphing data visually.

This can be done using scatter plot diagram.

3. It is the way to measure linear correlation.

We use this person-r value to evaluate correlation of variable on target variable.

If value > 0.5, the String correlation is.

4. Scalling is performed to process the numerical variable into limited range.

- If we have numeric data with different variations in value, the process on data become difficult.
- The visual represention also difficult to do for different varible.
- By scalling we move the numeric value in the range which can be easily handled
- Scalling also helps in optimizing the processing happenning by python in backend.

normalized scalling will move numeric data around 0 - 1 value range.

Standardize scalling will move nueric data near to 0 and variable mean value will become 0.

5. When the relationship of any one independant variable can be explain almost 100%

using other independant variable then the value of VIF value will be very high

and represented by infinite value.

6. Like standard Normal destribution,

Q-Q plot is graphical method for determining if two sample of data come from the same population or not.

- If the two distributions which we are comparing are exactly equal then the points
  on the Q-Q plot will perfectly lie on a straight line
- Q-Q plots are used to find the type of distribution for a random variable
  whether it be a Gaussian Distribution, Uniform Distribution, Exponential
  Distribution or even Pareto Distribution, etc.