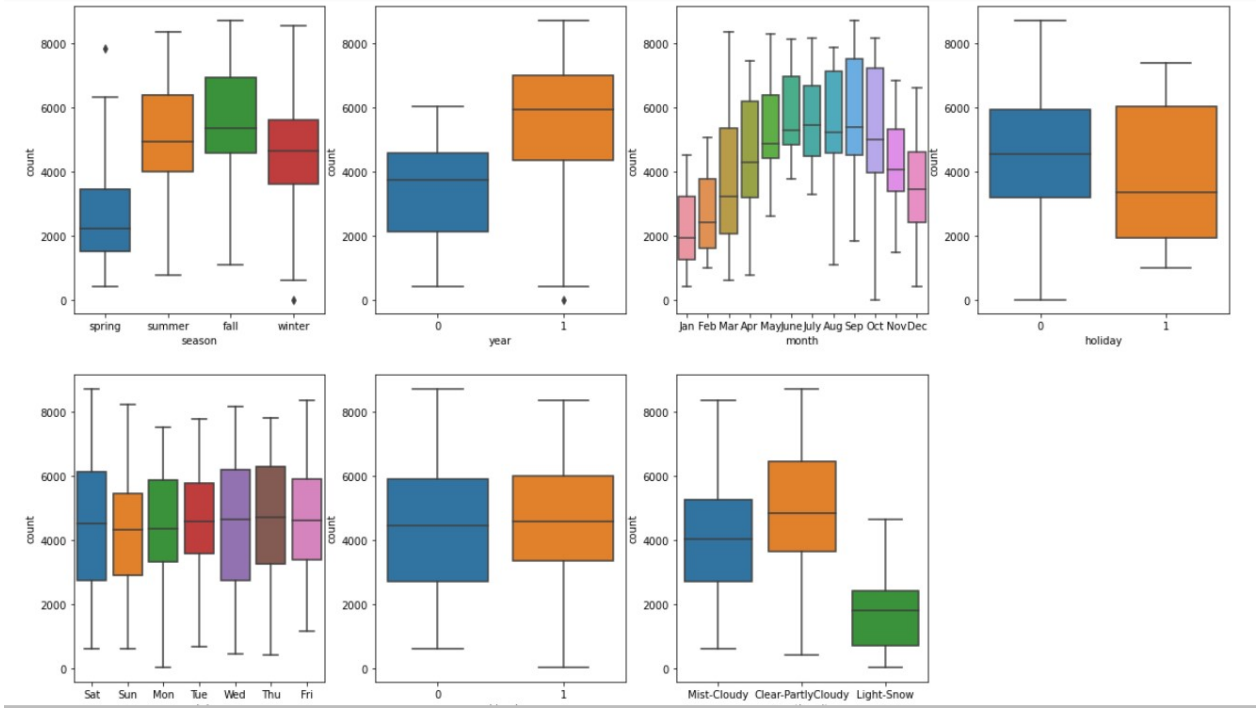
Assignment-based Subjective Questions

1. *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*
   Answer:
   #Bike Rentals are more in Fall and summer. It is highest in Fall. It is high in winter too.
   #Bike Rentals are increased in year(0:2018) to year(0:2019).
   #Bike Rentals are more during April to October and highest at September.
   #Bike Rentals are more on Saturday, Wednesday and Thursday.
   #Bike Rentals are more in partly cloudy weather



2. *Why is it important to use drop_first=True during dummy variable creation?*
   Answer:
   To Describe n levels of categorical variable only n-1 dummy variables are required. If we do not pass the drop_first=True then n dummy variables will be created. So It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?*
   Answer:
   'temp' and 'atemp'. So in feature selection 'temp' was selected.

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

   Answer:

   - Predicted values have linear relationship with actual values.
     - Validated with a scatter plot between actual & predicted values in train & test data set and best fit line were drawn.
   - Error terms are normally distributed
     - Normal distribution was checked using a histogram plot.
   - No Overfit & Underfit situation
     - R2 and Adjusted R2 was calculated for train and test data set and found very close to each other. Which ensured that there is no overfit & underfit.
     - Training Set: R2 = 82.6 % , Test Set: R2 = 80.8 %
     - Traning Set: Adjusted R2 = 82.6 % , Test Set: Adjusted R2 = 78.4 %
   - No Multicollinearity
     - Heatmap was created to check the co linearity among the variables.
     - VIF is calculated and considered the variables VIF > 5.
   - Categorical variable are converted to numeric dummy variables.

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

   Answer:

   Final Liner Regression Model Equation:

   'count' = 0.0902 + 0.2334x'year' + 0.0566x'workingday' + 0.4914x'temp' + 0.0916x'Sep' + 0.0645x'Sat' - 0.3041x'Light Snow' - 0.0786x'Mist + Cloudy' - 0.0650x'spring' + 0.0527x'summer' + 0.0970x'winter'

   - Top 3 Features:
     - Temperature: ‘temp’
     - Year: ‘yr’
     - Mist+Cloudy in weathersit

General Subjective Questions

1. *Explain the linear regression algorithm in detail*

   Answer:

   Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

   It is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as we can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, we can write a linear regression equation as:

   y = a + bx

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

Real Life Examples:

- Businesses often use linear regression to understand the relationship between advertising spending and revenue.
- Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.
- Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields.
- Data scientists for professional sports teams often use linear regression to measure the effect that different training regimens have on player performance.

Linear regression is used in a wide variety of real-life situations across many different types of industries. Fortunately, statistical software makes it easy to perform linear regression.

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

Answer:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe, is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once we plot each data set. As we can see, the data sets have very different distributions so they look completely different from one another when we visualize the data on scatter plots.

It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

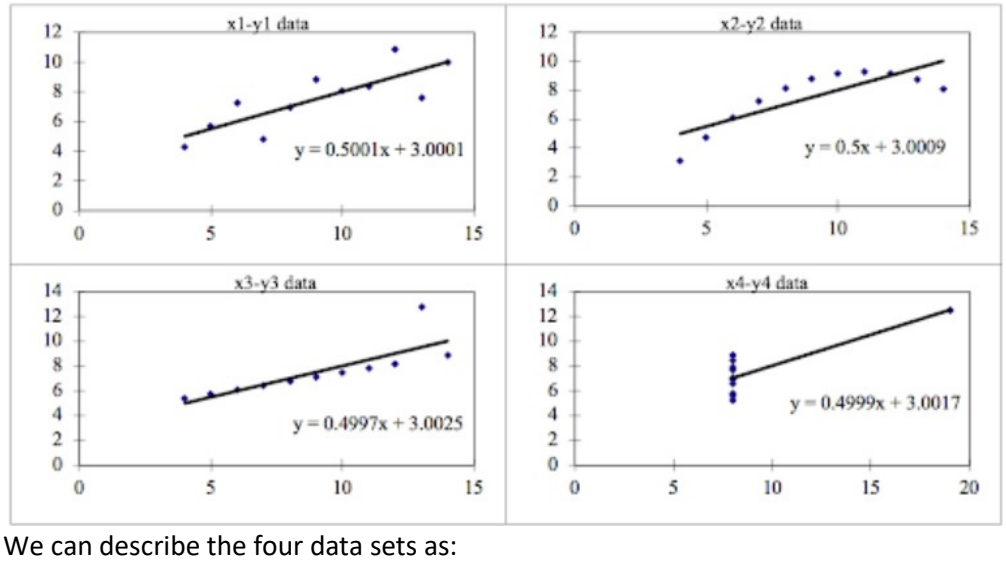
We can define these four plots as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as we can see below:



We can describe the four data sets as:

- Data Set 1: fits the linear regression model pretty well

- Data Set 2: cannot fit the linear regression model because the data is non-linear
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

  As we can observe, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

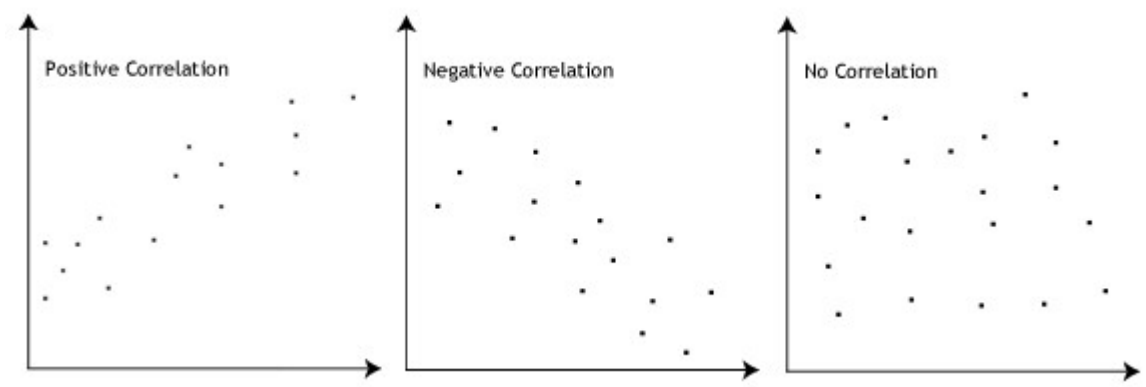
3. *What is Pearson's R?*

   Answer:

   In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



pearson r formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

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

Answer:

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.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.
Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

   Answer:
   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

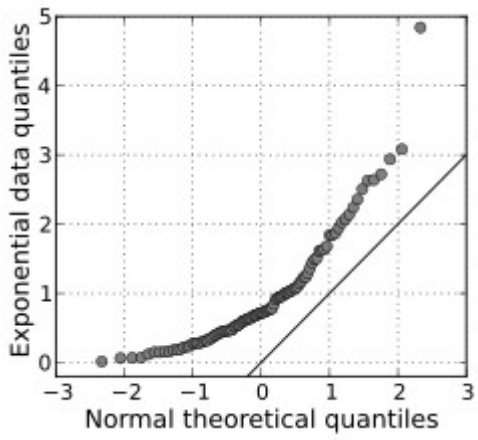
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.