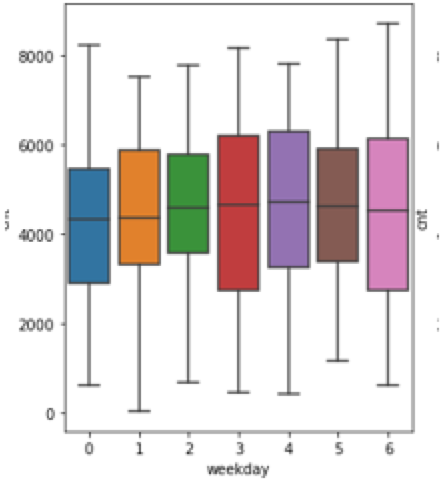
Assingment-based Subjective Questions:

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

-> Out of all categorical variables weekday, mnth and rain are final considered for building the model which have required model with r2_sore of 1.

Out of weekdays, Thursday has the highest correlation with dependent variable. It can also be seen for the box plot that maximum cnt are on Thursday.



Out of all months, September has positive correlation and February has negative which features considered in the final model.

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

-> I case of categorical variable once all variables are converted to Boolean the variable can be completely defined by p-1 categories where p is total number of categories in the variable. drop_first = True drops the first categorical variable and keeps all other variable in data frame. This is useful to reduce total number of features and reach the convergent solution faster.

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

-> Among the numerical varaibles, ‘registered’ has the highest correlation with the target variable.

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

-> Following assumptions are validated after building the model on the training set:
Linear relation between X and Y: Linearity of the variable can be seen from scatter plot of the independent variable with target variable.
Normal distribution of error term: This ca be validated by plotting histogram of error term. If all error terms are normally distributed with mean at zero the assumption holds valid.
Constant Variance: Scatter plot of error term shall be evenly distributed and no identifiable patter shall be visible.
Independent of each other: Correlation matrix and VIF are used for finding dependencies.

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

-> registered, casual and mnth(September) are top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

-> 1. Linear regression algorithm studies linear relation between target variable and independent variables and model relation between them as a straight-line equation which it uses to predict target variable from the test data set.
If there are *n* independent variables, then the relation between target and independent variable will look like.

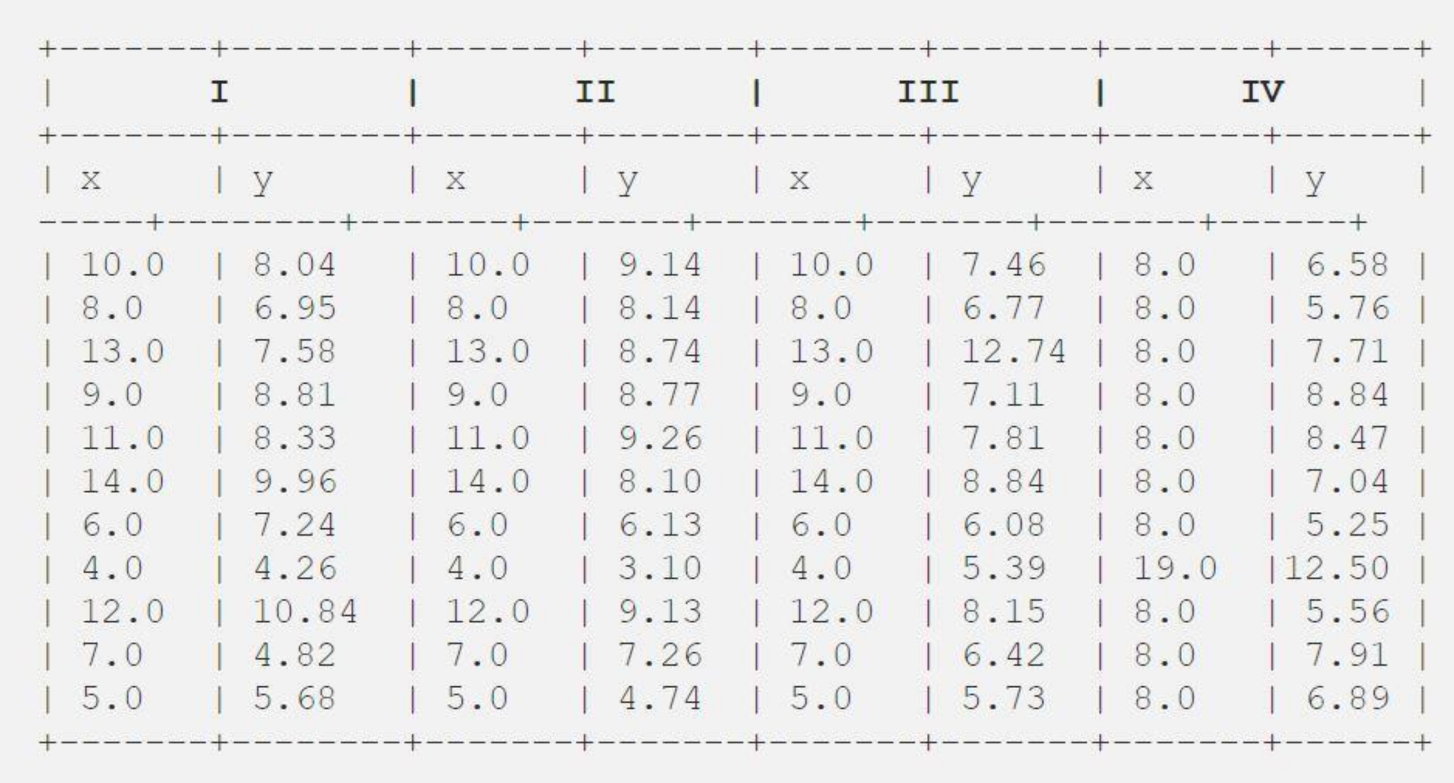
$$y = \beta_0 + \beta_1 * X1 + \beta_2 * X2 \ldots\ldots\ldots\ldots + \beta_n * Xn$$

y – is the target variable
Xn – are independent variable
Beta – is the slope of straight line for each in dependent variable provided all variables are independent.

2. Linear regression algorithm finds the best fit line for each independent variable assuming that other variable do not have any effect on it by minimizing the expression of residual Sum of Squares (RSS).
3. It is not necessary that having least RSS will give the best fit and correct prediction. To find relevance of each of the independent variable, linear algorithm evaluates Null Hypothesis for each of the variable.
4. Once the correlate variables (features) are selected and relevance is evaluated, prediction on test data can be made and model can be evaluated on test data by calculating RSS for test data.

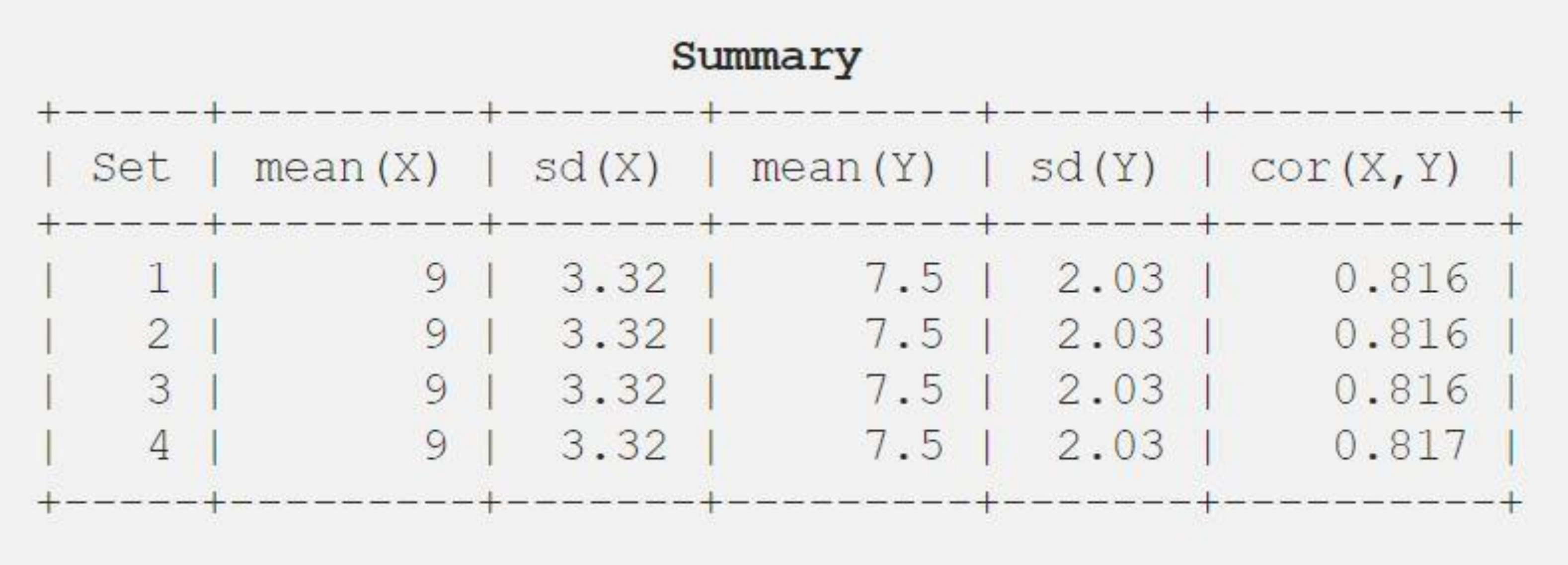
2. **Explain the Anscombe’s quartet in detail.**

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.



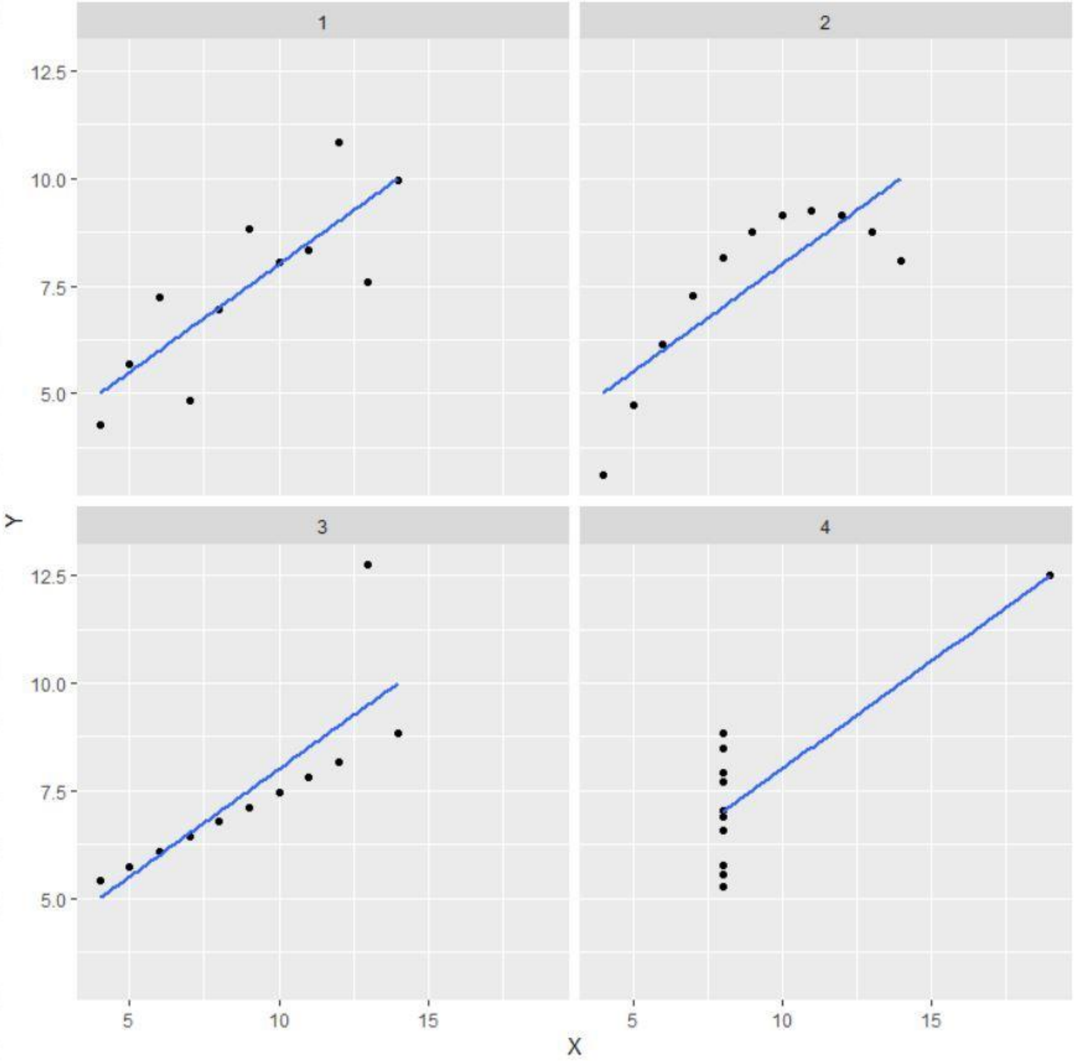
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y



```
                       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 |
+-----+---------+-------+---------+-------+----------+
```

It is mentioned in the definition that Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed.



Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

What is Pearson’s R?

-> Pearson’s R (Pearson is the correlation coefficient which is a measure of strength linear correlation between two data sets. Pearson’s R indicates how far are the data points from the best fit line. It is the ratio between the covariance of two variables and the product of their standard deviations.
Value of r will vary between -1 and 1. 0 means two data set do have any correlation.
Values near -1 or 1 means that data poitns are closer to the best fit line.
Signs indicate positive or negative correlation. 0 to -1 is negative correlation and 0 to1 is positive correlation.

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

-> Scaling is the method of normalizing the available data such that it can be represent within the required limited range. Scaling is performed because, the data on the scale is easy to compare that the data on different scales. Also, scaling helps regression model to converge faster as the model operation range reduces which in turn reduces the number of iterations to model convergence.
Normalized scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - x_{mean}}{sd(x)}$$

MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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

-> According VIF expression, $VIF = \frac{1}{1- Ri^2}$ VIF can be infinity only when R =1.

R =1 means that the variable into consideration has perfect linear correlation will all other variable.

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

Quantile-Quantile (Q-Q) plot, are plots of two quantiles against each other that help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
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.

In linear regression Q-Q plot helps to determine if training and test data received separately are from populaitions with same distributions.