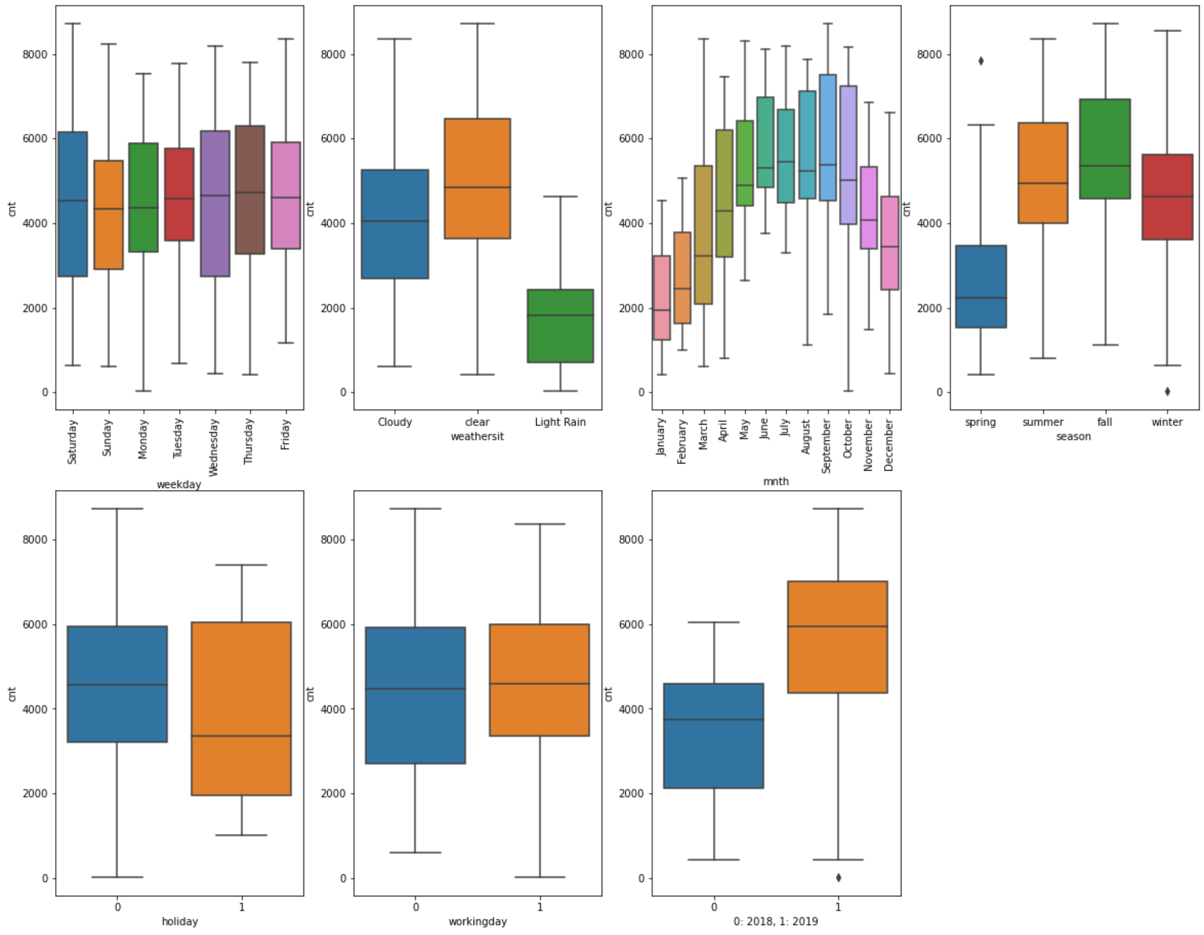
Ritesh Dubey

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)
Answers: By looking at the categorical variable vs target variable cnt below are the things which we can inferred :-



- Weekdays does not give any clear picture about the demand
- Clear weathersit has very positive correlation with demand
- Demands get increases by months from feb to september but we can observer little drop in july due to heavy rain
- Demands is high in fall season and low in spring season

- Holiday is showing very weird relation as when there is holiday demand is decreased
- Working day has no effect with median of demand only 25 percentile has increased
- Demands grows with Year

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

- drop_first=True drops the categorical variable from which dummy variables were created.
- If we dont drop that then extra categorical variables which create multicollinearity issues.
- And it will also affect the overall adjusted R square.

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

- “Temp” has highest correlation with target variable “cnt”

4. How did you validate the assumptions of Linear Regression after building the model on the
training set? (3 marks)
Answer: I have validated on using following assumption:-

- Distribution of error term by mean=0
- No pattern observed in error term vs y predicted value
- Multicollinearity using VIF
- Adjusted R squared value 83.2 %
- Demands and Index of actual vs predicted value completely matched

5. Based on the final model, which are the top 3 features contributing significantly towards
explaining the demand of the shared bikes?
Answer: Top Three Variable to focus on :

- temp, weathersit_clear, season_winter
- unit increment in temp raise the cnt by 0.49 units

- unit increment in weathersit_clear raise the cnt by 0.081 units
- unit increment in season_winter raise the cnt by 0.0831 units

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

- Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
- An example is let’s say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.
- Mathematically, we can write a linear regression equation as:

  $$y = a + bx$$

  Where a and b given by the formulas:
  Here, x and y are two variables on the regression line.

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

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

b = Slope of the line.

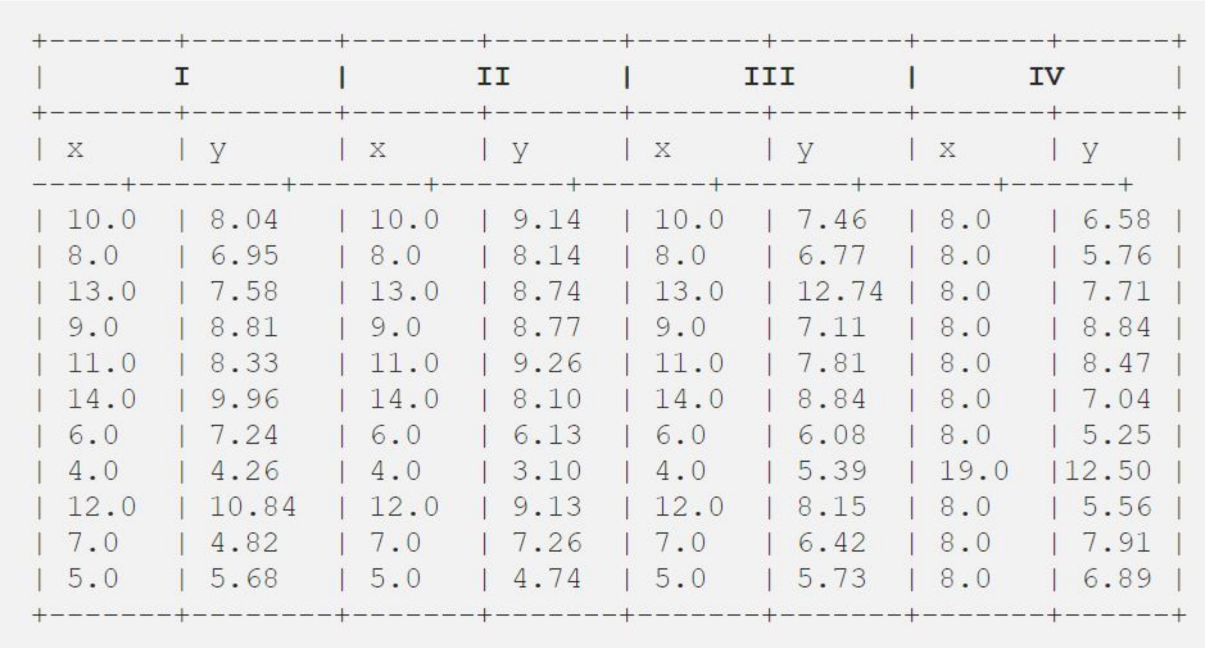
a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe’s quartet in detail.
Answer:
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.
Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11

data-points are given below.



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.

3. What is Pearson’s R?
Answer:

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗 _1=1/(1-R_1^2)

VIF infinit:
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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer:
The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the ‘y = x’ line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.