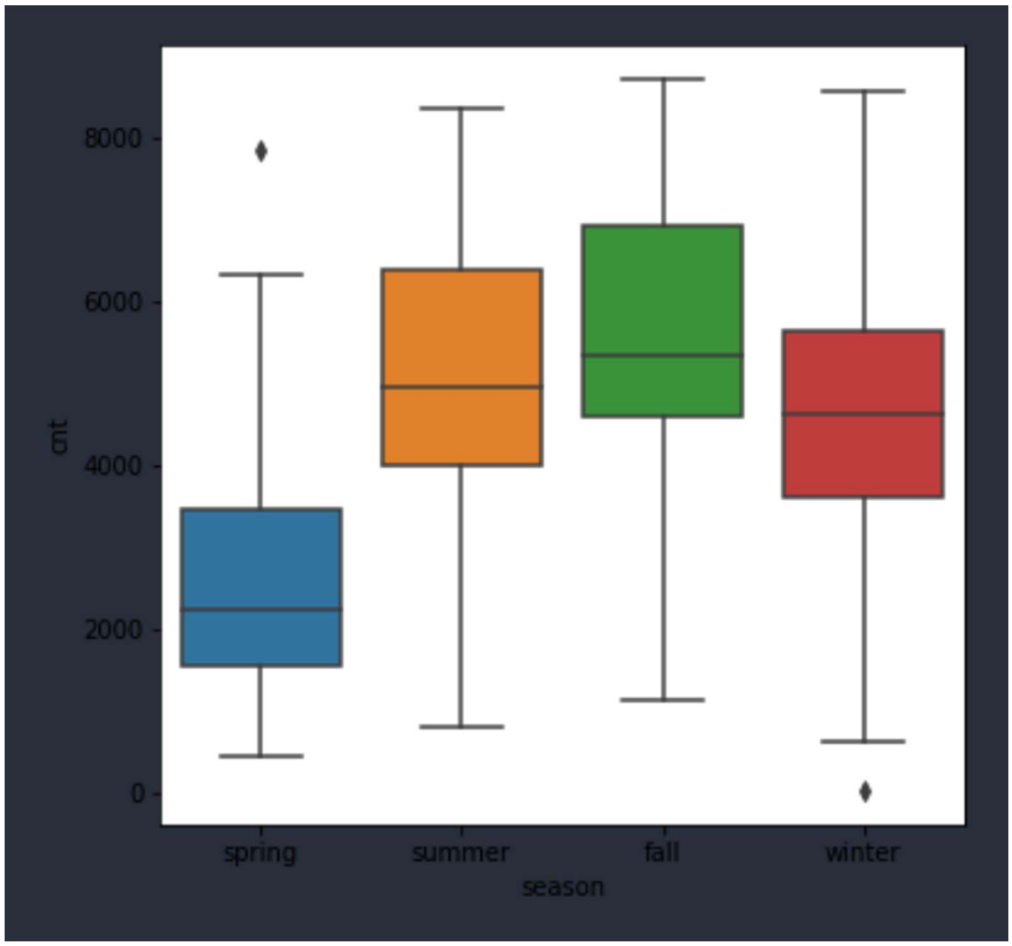
Assignment-based Subjective Questions – Answers

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

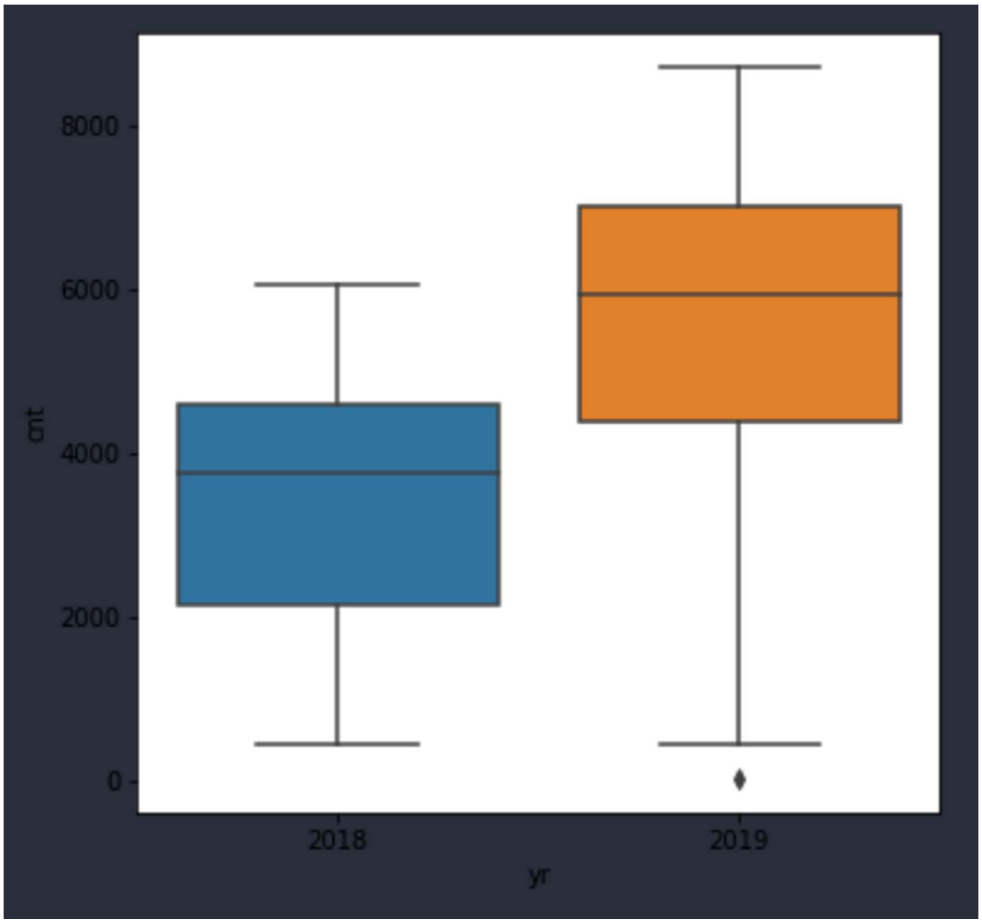
Answers – We have following categorical variables –

'season','yr','mnth', 'holiday','weekday','workingday', 'weathersit'

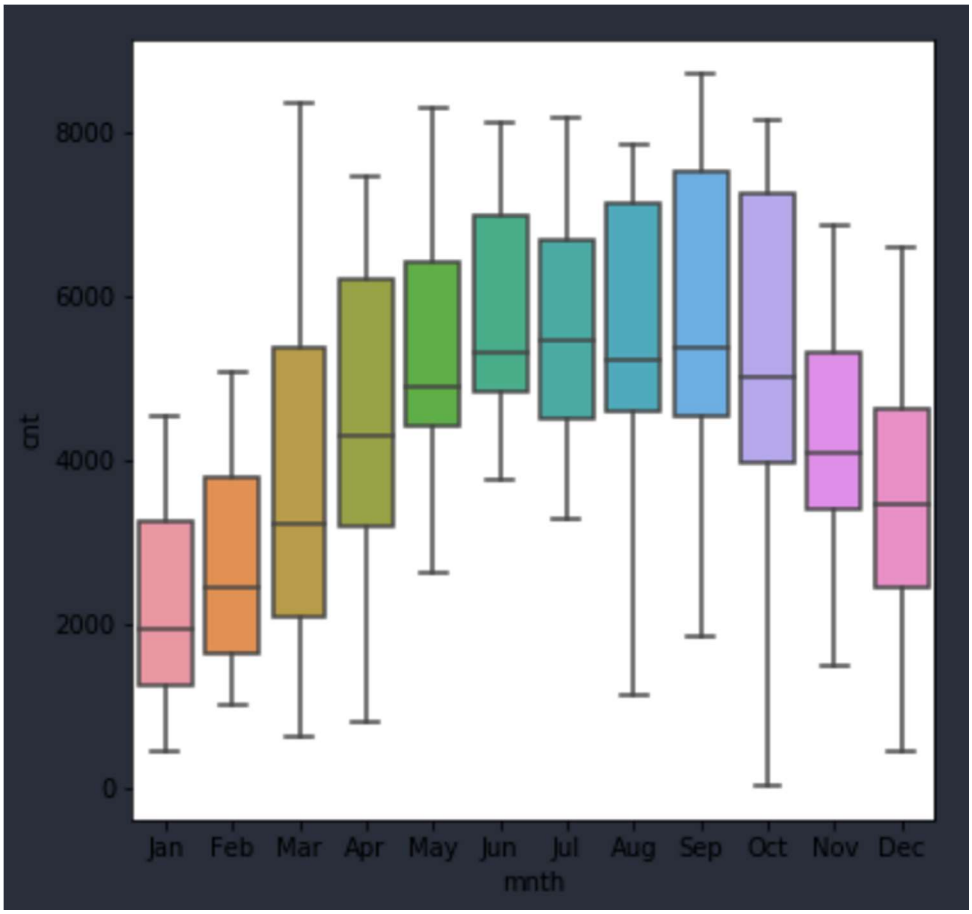
- **Season** – Fall and summer have comparatively high count and spring have some outliers in the data.



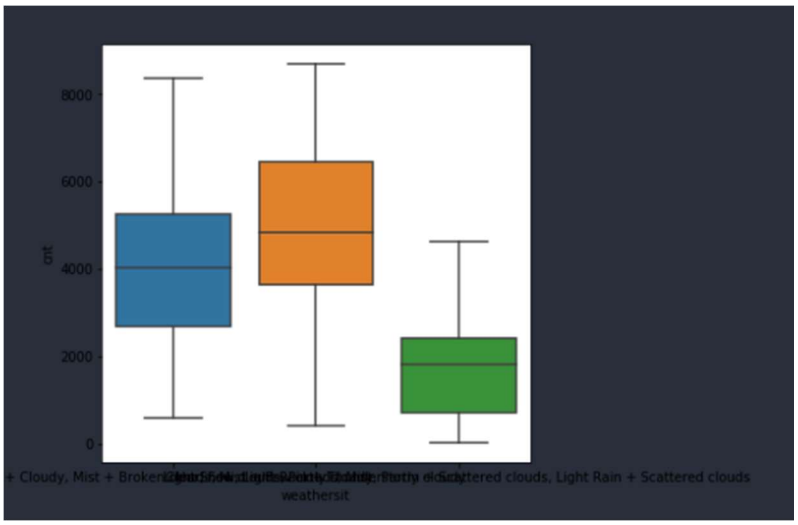
- **Yr** – Count is increasing over the year as 2019 have high count then 2018. And also 2019 has some outliers.



- **<u>Mnth</u>** – Summer and fall have more count during the year.



- Weekday, holiday and working day are mostly equally distributed with count.
- Weather sit - `*Clear, Few clouds, Partly cloudy, Partly cloudy*` have high count then other two weather sit.



2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
Answers – this is important to remove redundant column in our analysis.

For an example – I have used drop_first = True for ‘weekday’ variable. Now using `drop_first = Ture` for 'weekday', where we can see if 6 days flags are 0, that mean it’s the seventh remaining day, such as Friday.

Removing first column is more significant where we have general value (or less category in data), for example Gender (as M and F) or result (as pass or fail).

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

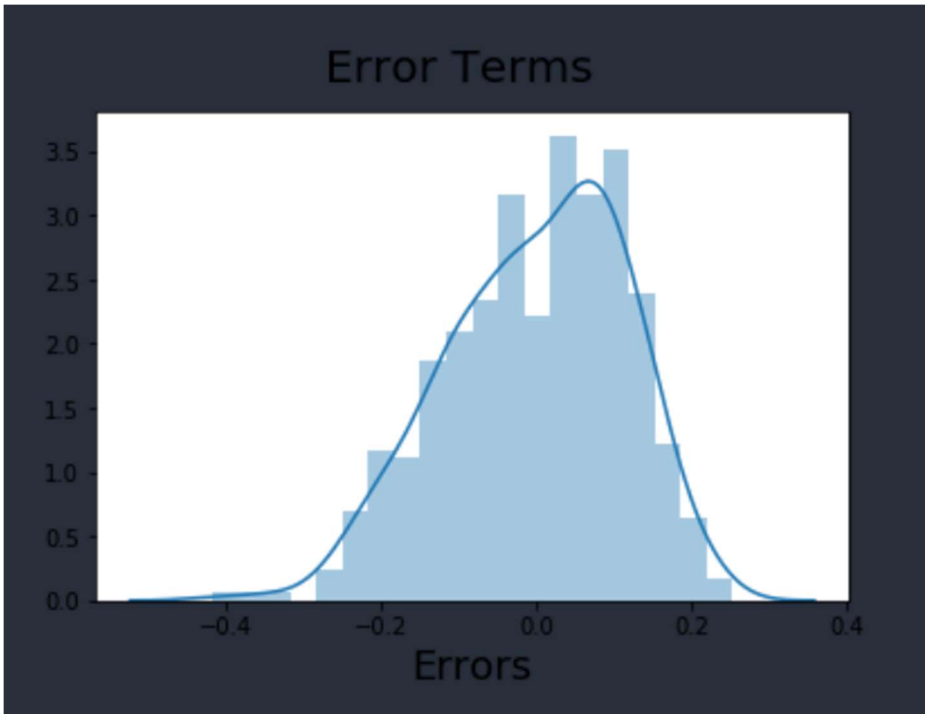
Answers – “registered” numerical variable have highest correlation with “cnt” target variable.

Following is the illustrations in pair plot.

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

Answers – I have performed Residual Analysis of the train data to see the error term distribution.

Error Term is almost centered to zero, which says our assumption of linear regression is good in shape.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answers – “casual”, “wind speed” and weekday are significantly toward the demand of shared bike.

General Subjective Questions

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

Answers – Linear regression is predictive analysis to predict target variable with available feature and figure out the best significant combination of available feature to predict target variable with minimum error.

Equation of linear regression: Y = MX+C

C is intercept on Y axis.

M is slope / gradient

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answers – Dataset with same statistics but different graph.

For example –

Dataset –

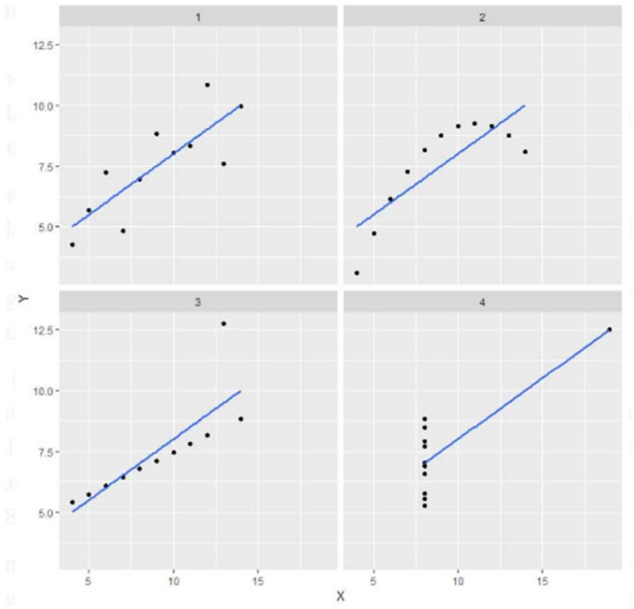
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Statistics –

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

Graph –



3. What is Pearson's R? (3 marks)

Answers – This is a correlation between variables. Correlation between data sets is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answers – This is method to transform variables in same range to have comprehensive visualize relationship. It's very common to have data in different units and magnitude.

Scaling is performed to bring data in same scale, which will help us analyze data in comprehensive visualization.

There are two method of scaling –

1 – Min Max scaler

2 – Standard scaler

Difference between min max scaler and standard scaling –

Min max scaling rescale the data set such that all feature are in the range of 0 to 1 and standard scaling remove the mean and scale the data to unit variance

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answers – VIF infinity happen when there is a perfect correlation between variables. In this case variable in insignificant and need to remove from model development.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answers – Q-Q plots are used to visualize dataset distribution that will help us in linear regression.