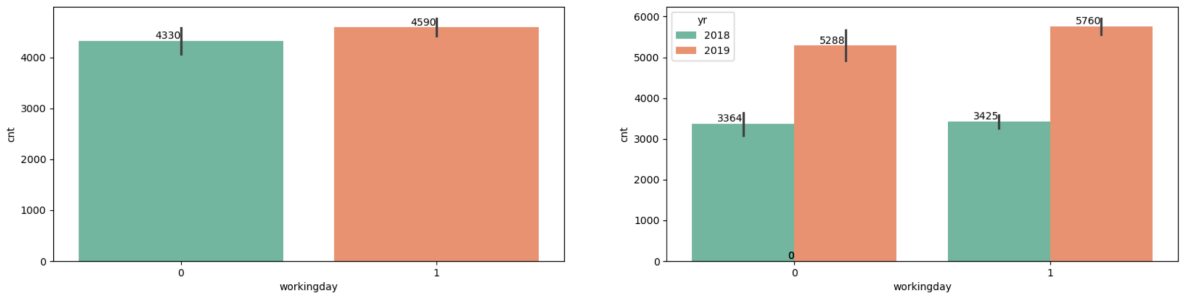
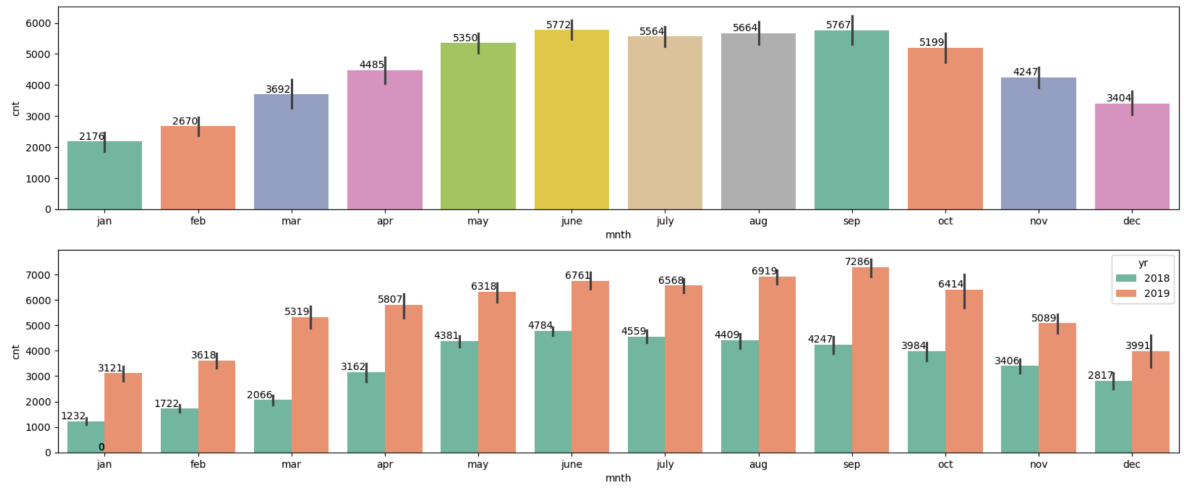
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

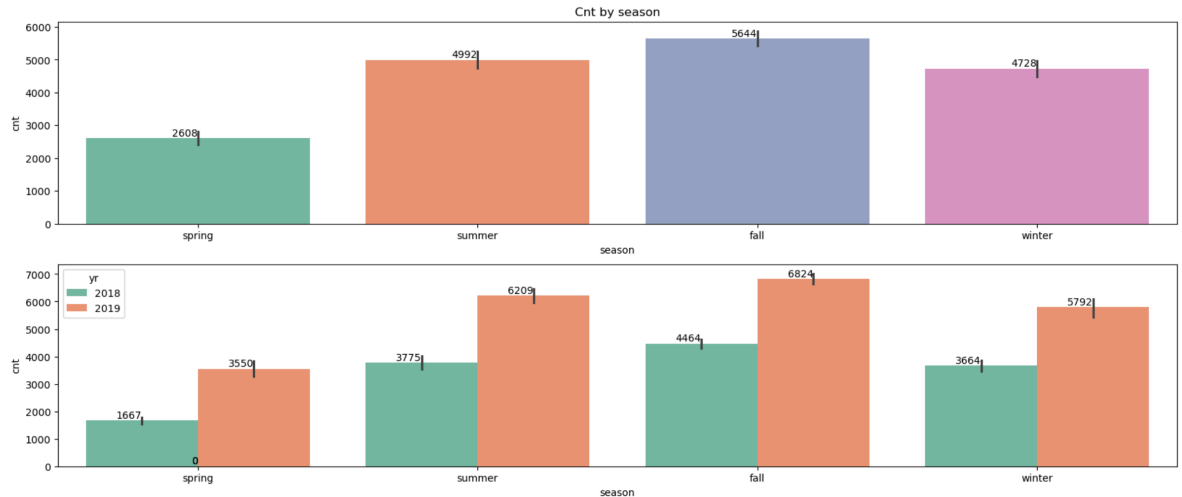
- The count of booking is more on working days compared to non-working days. Also, number of bike booking has increased in 2019 for both working and non-working days.



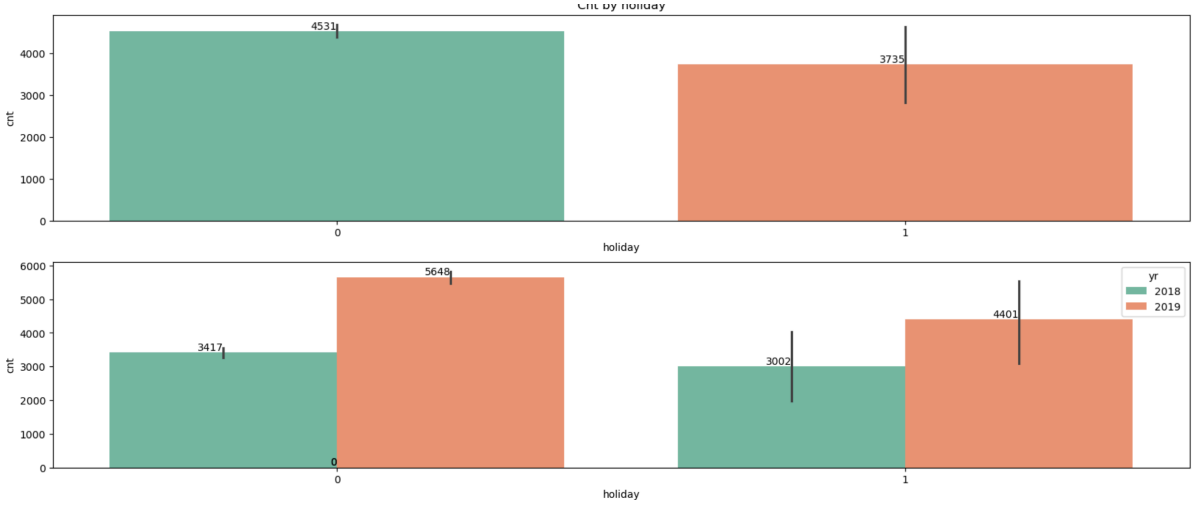
- Sep 2019 had the highest booking for the bike. The count of booking in year 2019 is more than 2018 for all the months.



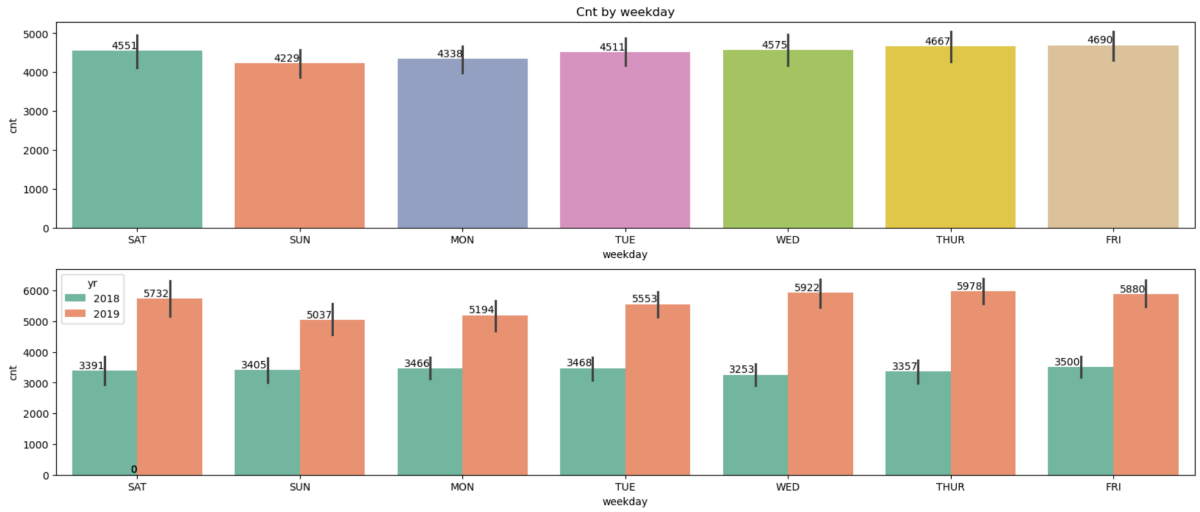
- Maximum booking was done in fall season for both the years. The booking was more in 2019 for almost all the season compared to 2018.



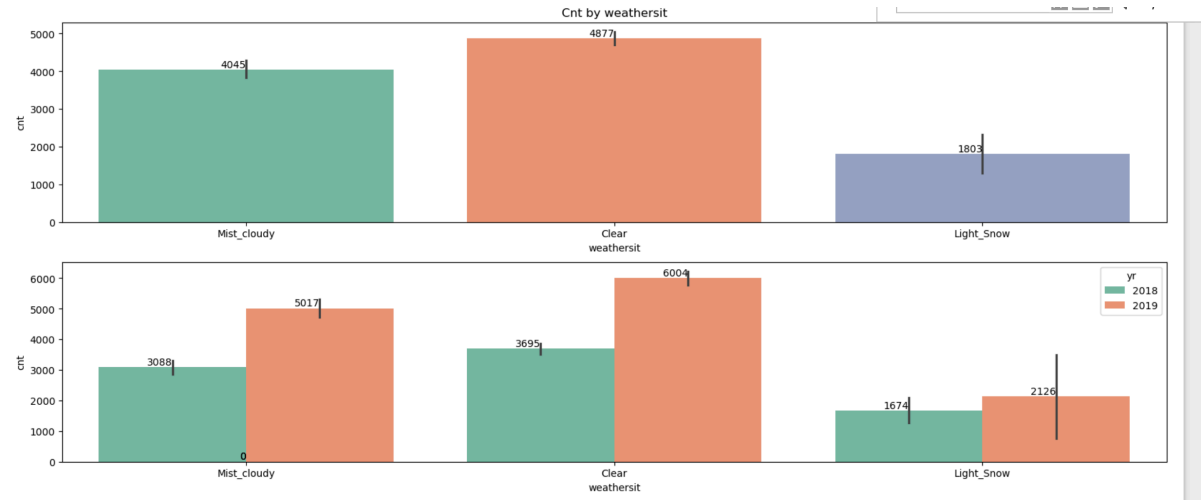
- The number of bookings increased when it wasn't holiday for both the year 2018 and 2019



- The count of booking was more on weekdays - Thur and second highest on Fri for 2019



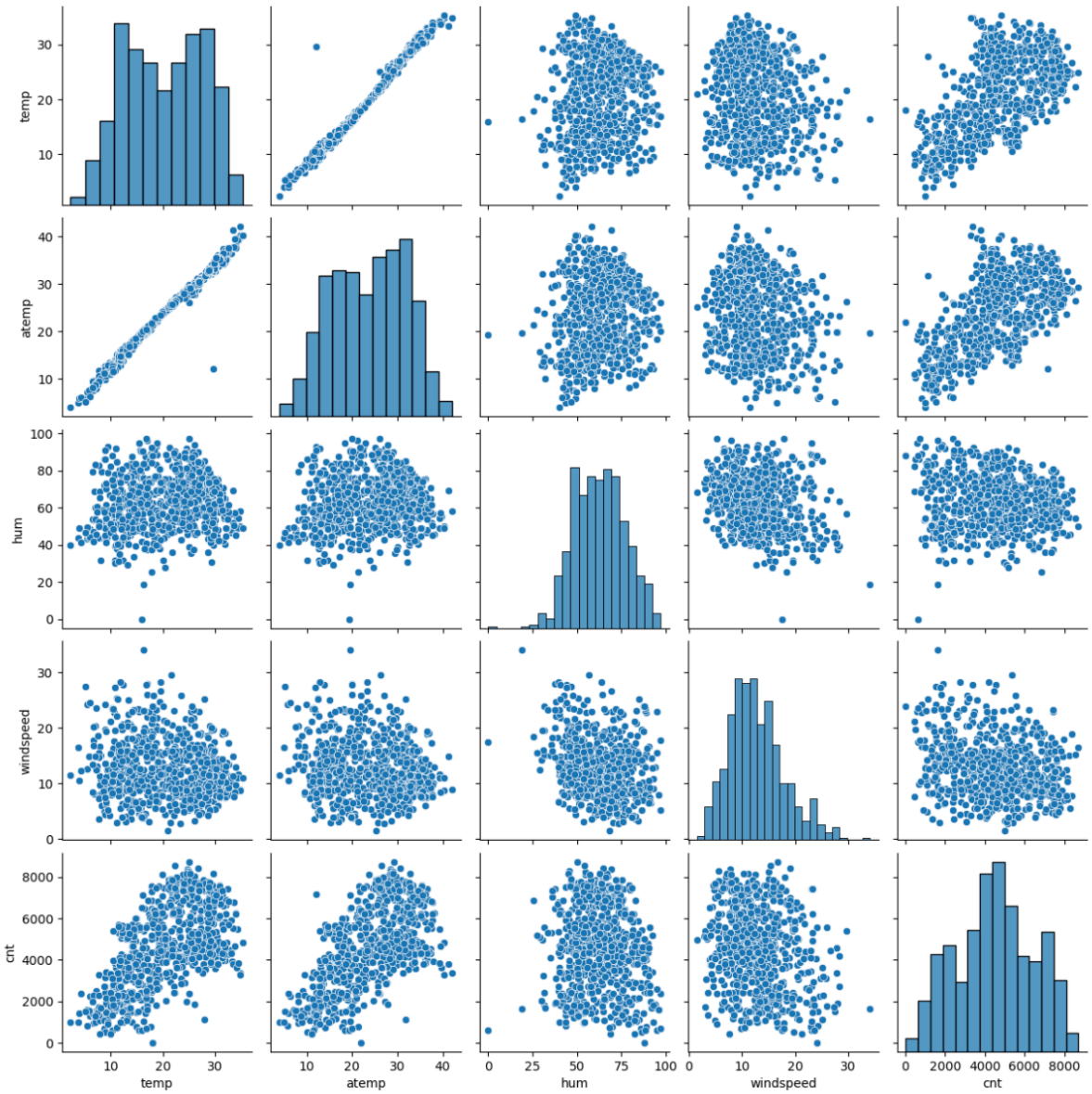
- The count of booking was more on weekdays - Thur and second highest on Fri for 2019



Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

It is important to use drop_first=True as this handles the multicollinearity. For m variable, it will create m-1 columns.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



Temp and temp variable has the highest correlation with target plot

---

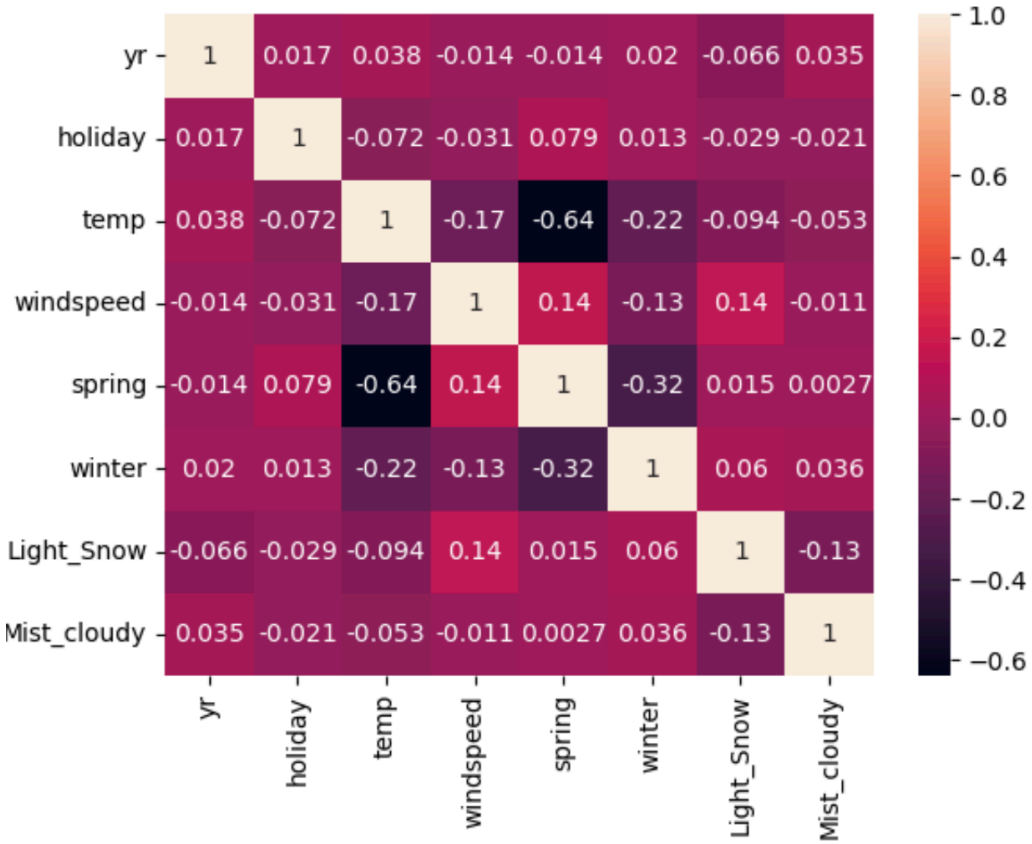
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

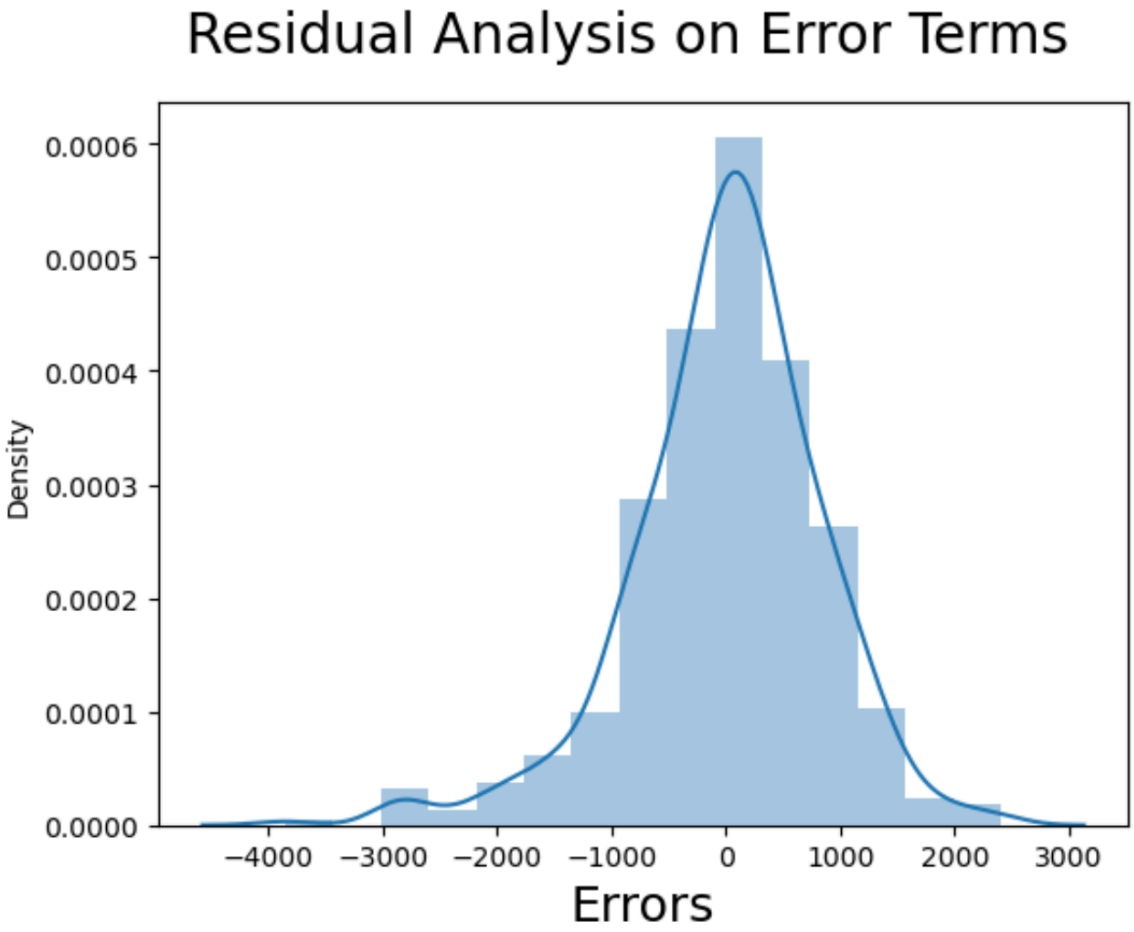
Linear regression assumptions is validated as mentioned below :

- There were no significant correlation between the final input variables. I checked that VIF value is less than 5 and p-value is less 0.05

	Features	VIF
3	windspeed	4.79
2	temp	4.05
0	yr	2.08
4	spring	1.68
7	Mist_cloudy	1.52
5	winter	1.36
6	Light_Snow	1.08
1	holiday	1.04



- In residual analysis, it was found that error terms are normally distributed



- Predictors and target variable has linear correlations.
- There are no correlation with residuals

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

As per the final model, top 3 features contributing significantly are :
- Year
- Windspeed
- temp

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is method which is used to analyse the relationship between the target variable or output variable and predicator variables. The goal is to find the best fit line between the predicated values and the actual values.

Assumption of linear relationship :

Y = mX+c

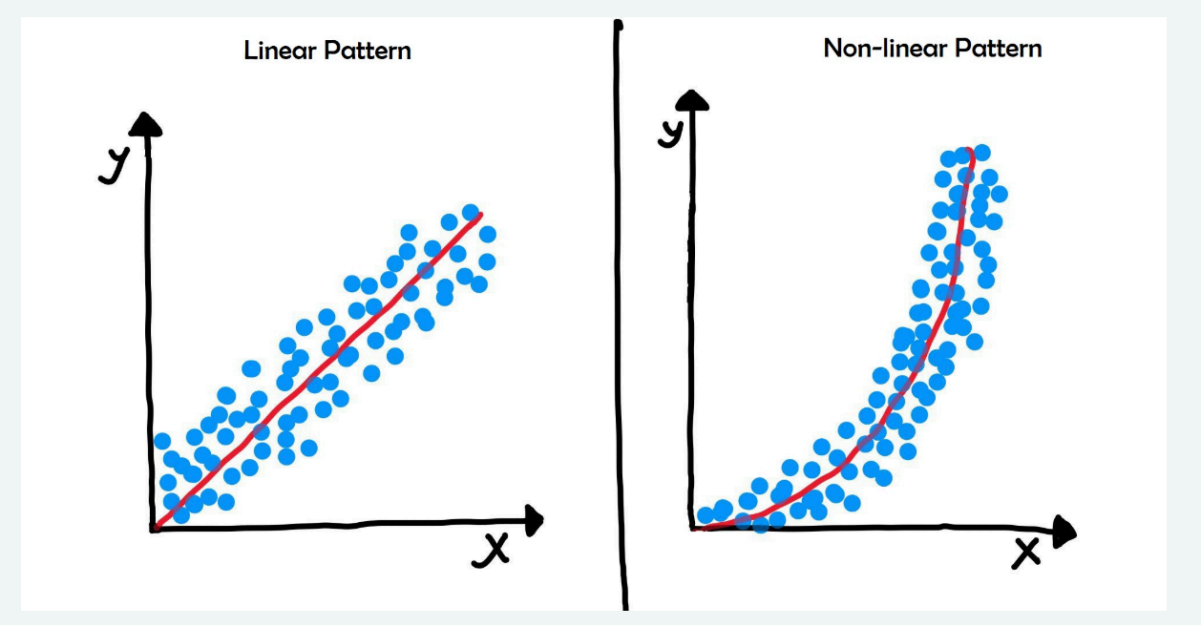
Where y -> dependent variable ; X -> Independent variable ; C-> constant also known Y-intercept as it is the value if y when x = 0; m -> slope

There are 2 types of linear regression :

- Positive Linear Regression - The target variable increases with the increase of input variables or decreases with decrease of input variables
- Negative Linear Regression - When input variable increases target variable decreases and vice versa.

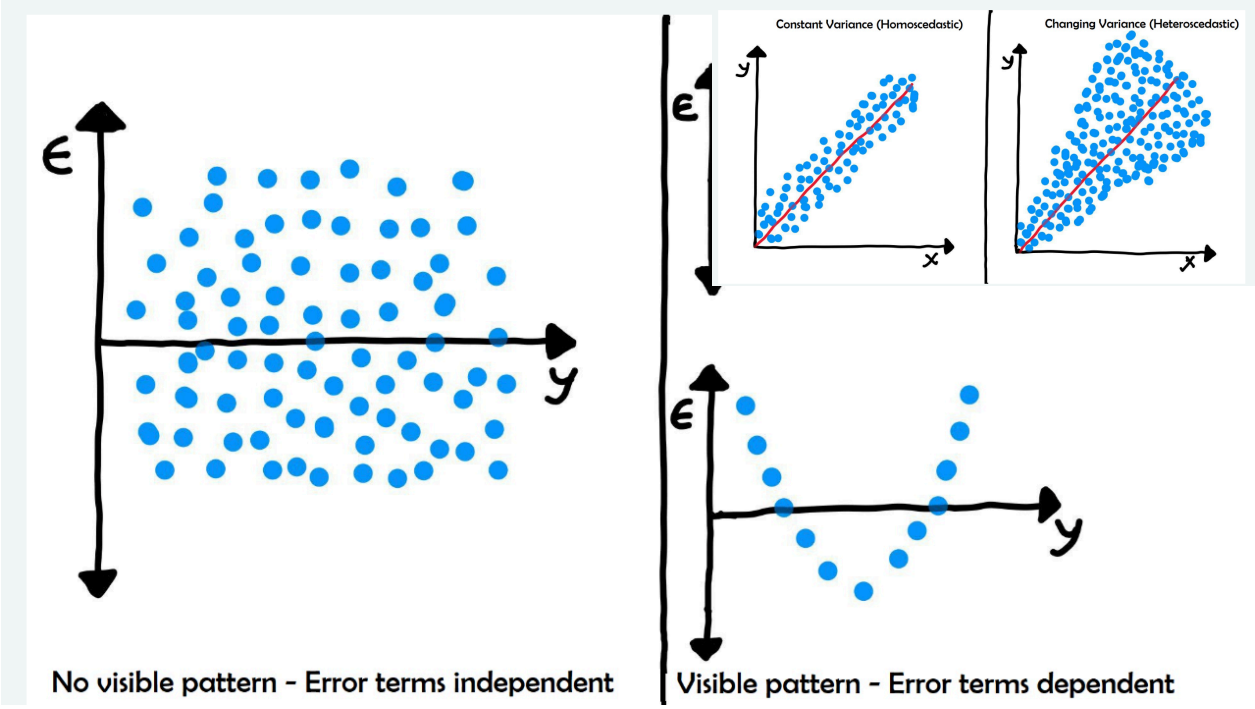
Assumptions made on linear regressions :

- Input and output variable should have some sort of linear relationships in order to fit the linear model between them



- Error terms are normally distributed. Error terms follow the normal distribution with mean equal to 0 in most of the cases

- Error terms should not be dependent on each other, there should be no visible pattern on the error terms



- Error term should have constant variance, they should not increase or decrease as the error values changes. Variance should be independent of any pattern.

Linear regression is of 2 types :

- Simple linear regression
- Multiple linear regression

Simple Linear regression - Model is built and predictions are done with 1 input and 1 output variable

Multiple Linear regression - In multiple linear regression, we use more than 1 input variables.

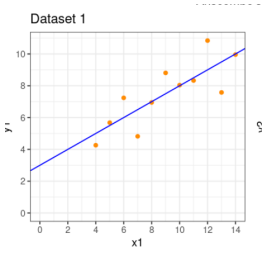
NOTE - R - Squared value increases or remain the same when we add more variables for model building. Addition of new variable never decreases R-squared value irrespective of its significance. Therefore, we use adjusted R-squared values.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

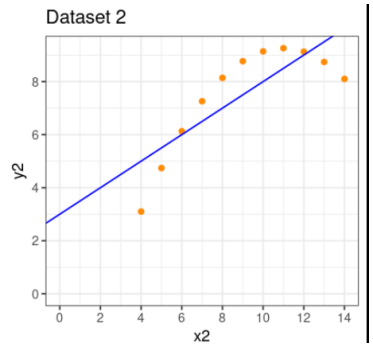
<Your answer for Question 7 goes here>

Anscombe's quartet explains the importance of data visualisation. It consists of 4 data-sets with similar statistical properties but entirely different graphs. It consists of 11 points on (X,Y). All the 4 data-sets have same mean and standard deviations but with different distributions.

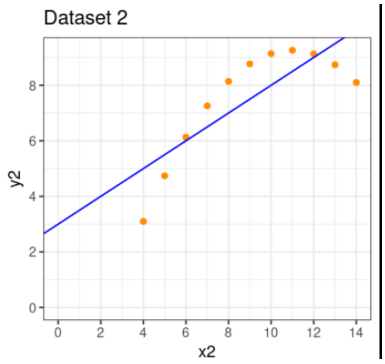
- Data-set 1 graph plot shows - X and Y have linear relationship



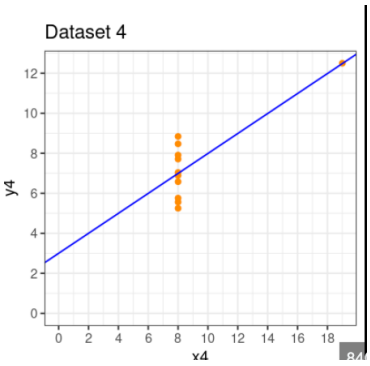
- The data-set 2 graph plot shows - X and Y have non-linear relationship



- The data-set3 graph plot shows - X and Y have a linear relationship with one outlier



- The data-set4 graph plot shows - X and Y have no linear relation with one outlier



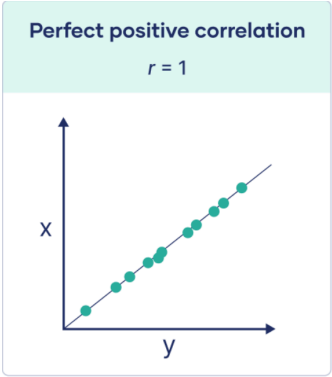
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
The Pearon's R tells us strength of linear relationships between the two variables. Its value varies

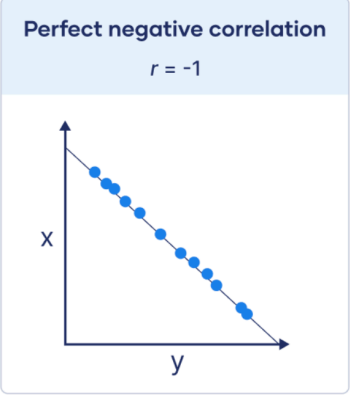
from -1 to +1.

There are three different types of correlations -

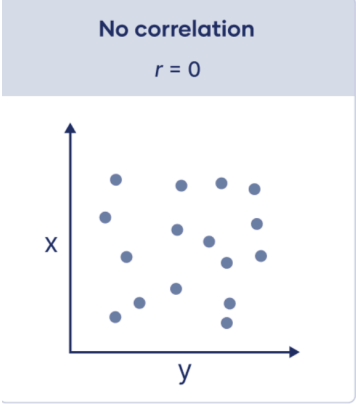
- Positive correlation : Change in X(Input variable) changes the Y in the same direction. The value of correlation is from 0 to 1



- Negative correlation : Change in X(Input variable) changes the Y in opposite direction. The value of correlation is from -1 to 0



- No Correlation : There are no relationship between input and output variables. The value of correlation is 0



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

In scaling, we transform the data to fit in scales. We need scaling for the use of interpretation and faster convergence for gradient descent method. Scaling doesn't affect the model accuracy. It only affects the coefficients.

- Standardized scaling brings all the data into a standard normal distribution with mean 0 and standard deviation = 1. However, Normalised scaling brings all the data in the range of 0 and 1.

- Standardised scaling are not affected by the outliers. However, Normalised scaling are affected by the outliers.

- Standardized saling is used when data is normally distributed. However, Normalised scaling is used for different scale.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF is infinity when there is exact linear relationships between the input variables. It also shows multicollinearity is present as VIF value is extremely high.

This can happen because of following reasons :

- There might be duplicate variables present on the data-sets, so we should be removed.
- While creating dummy variables we missed dropping one category from categorical variables.
- Input variables are highly correlated to each other. We should consider dropping these columns

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

The Q-Q(Quantile-Quantile) plots the graphs of two distribution against each other and helps us in our analysis and comparison.

If the two distributions we are comparing are exactly the same, all the points in the Q-Q plot will fall perfectly on a 45 degree line (y = x). If points are closer to the line, it shows data is normally distributed. However, if points are far away from the line, it shows data might have outliers.

In Linear regression, we can create Q-Q plot between train and test data set to conclude if both the dataset have common or different distribution. We can also compare the tail behaviour along with the type of distributions.