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?

After performing EDA on the dataset below are the observations found:

- From the below plots we see that the bike sharing demand is the highest -

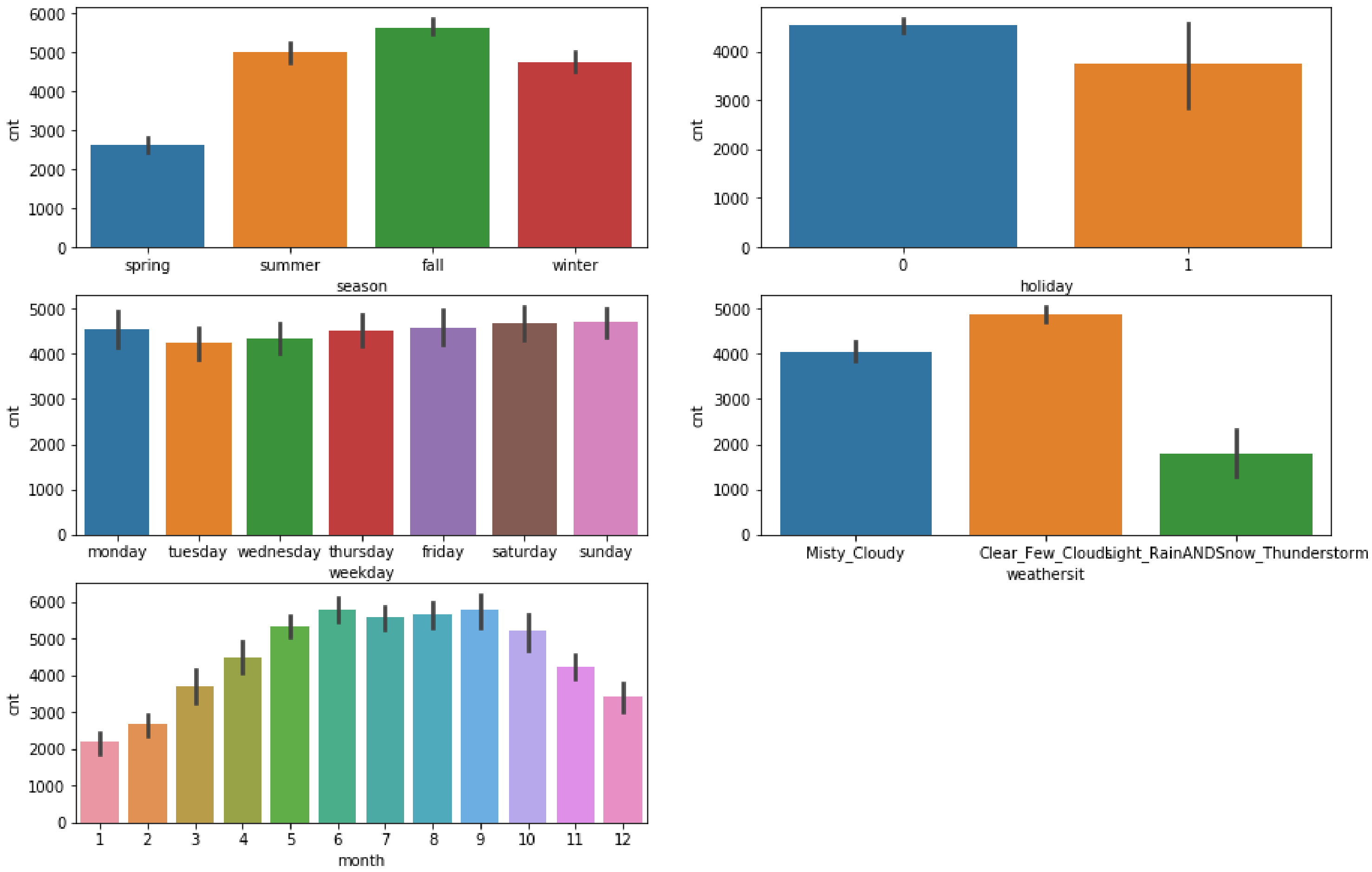
1. During summer and fall

2. When it’s a holiday

3. When the weather situation is Clear Few Clouds and also when its misty cloudy

4. During the months June and September

(Note-The above inference is across both years(2018 and 2019)

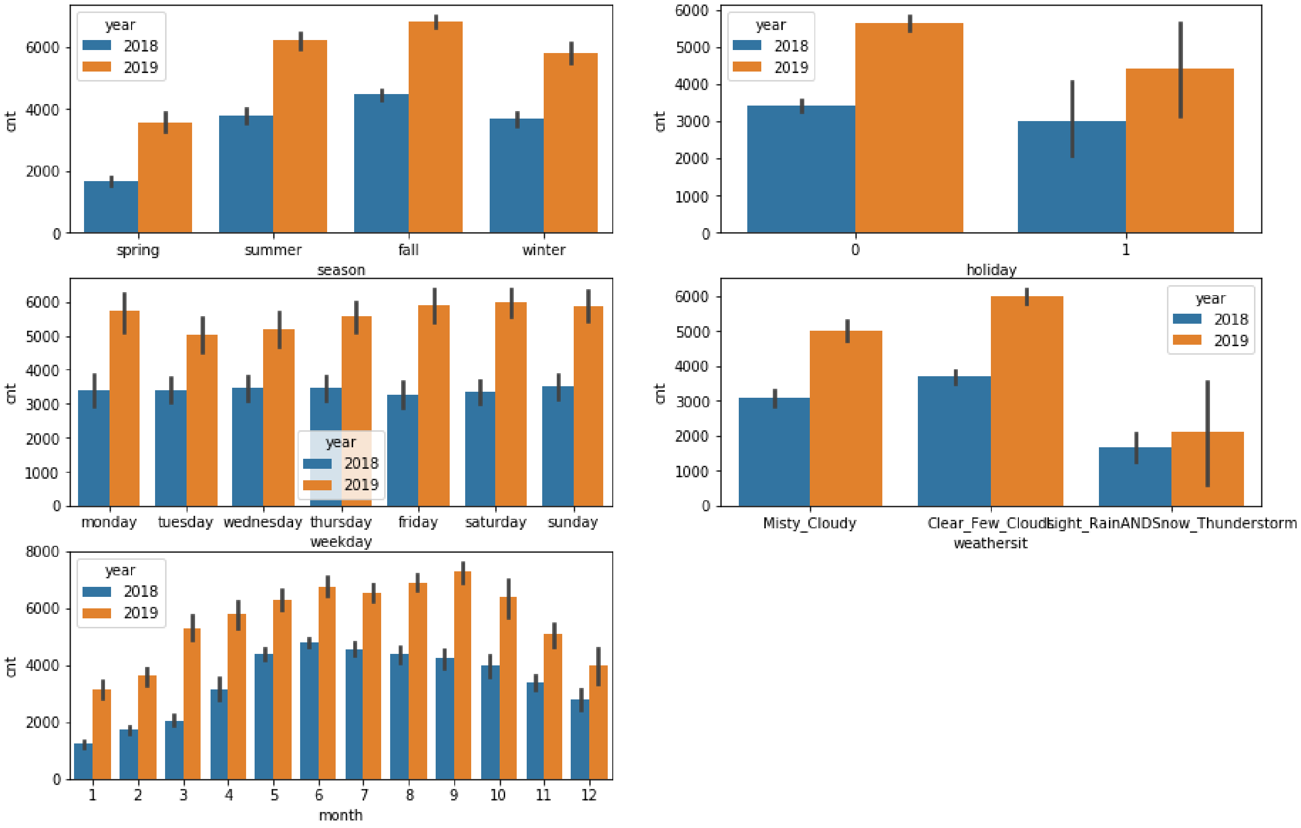


- From the below plots :

1. From the below plots its clear that the bike sharing demand was significantly higher in 2019 when compared to 2018.

2. From the plot 5 we can see that bike sharing count has gradually increased from Jan to September and then it has dropped gradually till december

3. We dont see much difference in the counts accross weekdays, Monday, Friday and Saturday having slightly higher counts.



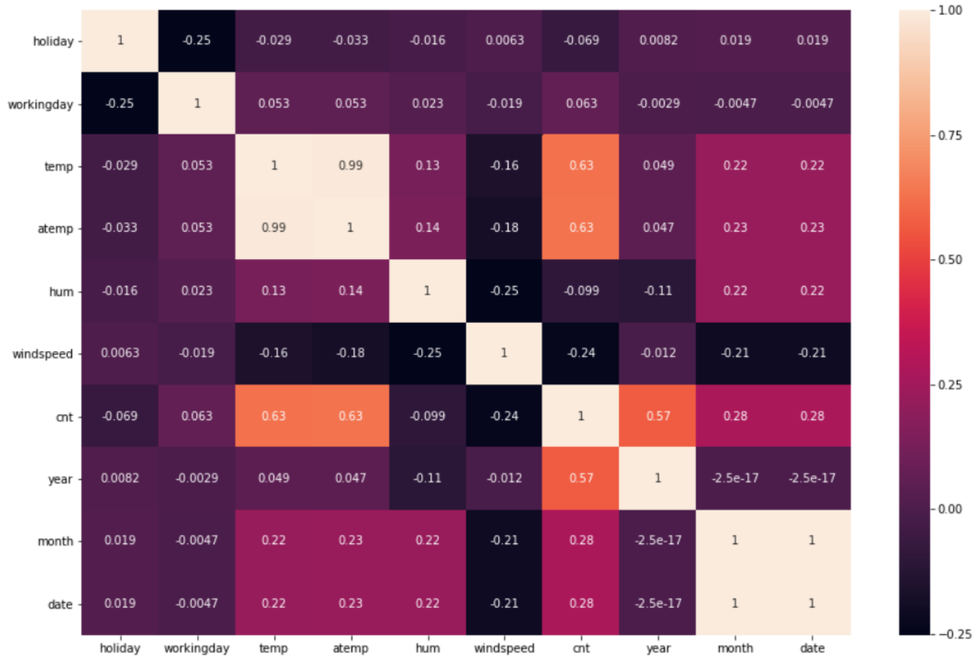
2. Why is it important to use drop_first=True during dummy variable creation?

Whenever we create dummy variables, if there are n levels then it creates n columns. But we might not need n columns, we'll only need (n-1) columns.

Eg. Lets say the column season has values 'Summer', 'Monsoon', 'Autumn', 'Spring'. There are 4 levels here. When we create dummies for this column, then if one variable is not 'Summer', not 'Monsoon', not 'Autumn' then it has to be 'Spring'.

Eg. Season=pd.get_dummies(data['season'],drop_first=True)

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

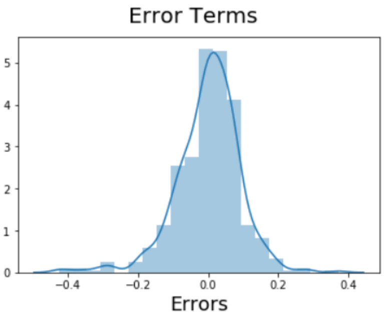


Looking at the heatmap we see that atemp and temp have the highest correlation with target variable. Since in my model I have dropped temp and kept only atemp, we can say that atemp has the highest correlation with the target variable.

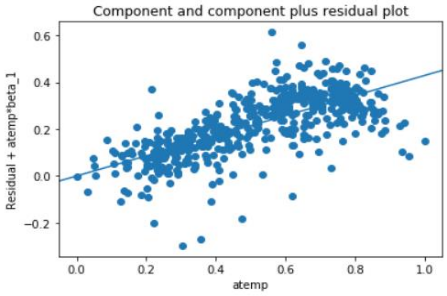
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

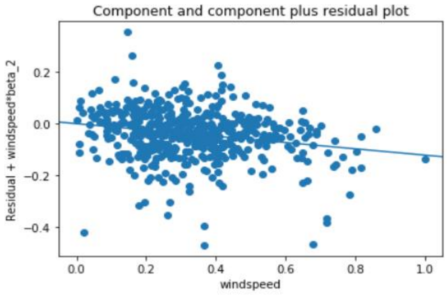
After building the linear regression model on the train set I have checked for the below assumptions:

1. Normality of error terms

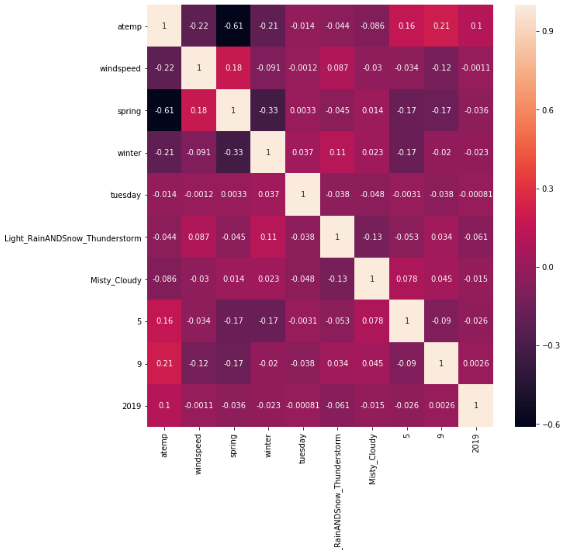


2. Linearity : The features should show linearity

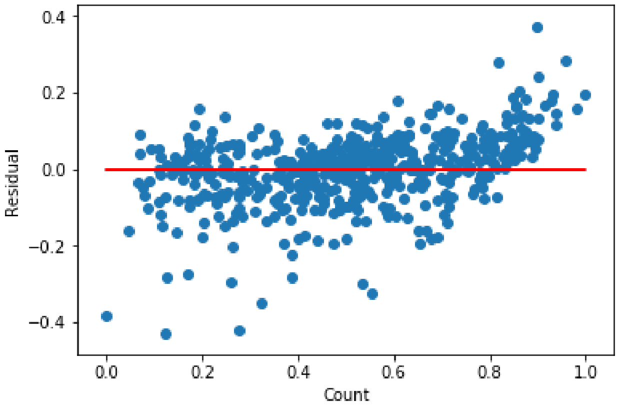




3. Multicollinearity: There should not be inter-associations among predictor variable.



4. Homoscedasticity: The probability distribution of errors has constant variance.



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

   The top 3 featues based on the final model are :

   - atemp
   - Year 2019
   - Month September

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is statistical model which attempts to explain the relationship between a dependent variable and an independent variable suing a straight line. The independent variables is also known as **predictor**, the dependent variable is known as **output/target.**

Best Fit Line : There is a notation for the best fit line, a line that fits a given scatter plot in the best way.

Standard Notation of Linear Regression: $y = B_0 + B_1X$

B_0 → Intercept

B_1 → Slope

There are 2 types of linear regression – Simple Linear Regression and Multiple Linear Regression

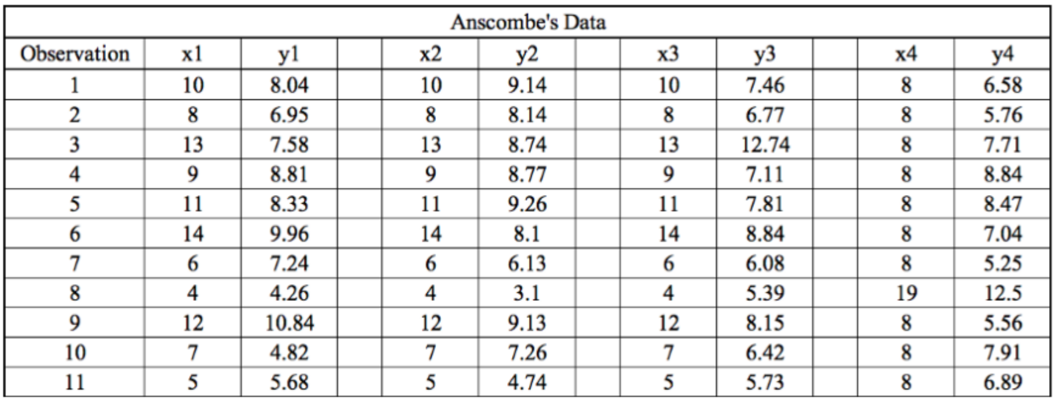
Assumptions of linear regression:

1. While building a linear regression model we assume that the target and input variables are linearly related. X and y are in linear relationship.
2. Error terms are distributed normally(with mean 0).
3. Error terms are independent of each other.
4. Homoscedasticity - Error terms have constant variance.
5. Multicollinearity – LR model assumes that there’s no inter-associations among predictor variables.

2. Explain the Anscombe’s quartet in detail.

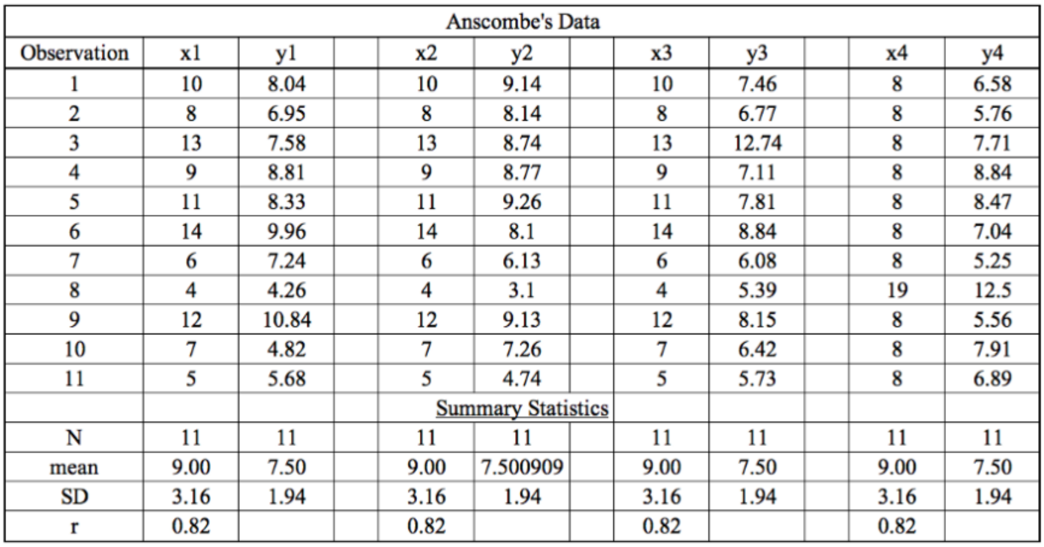
Anscombe’s Quartet is a group of four data sets which are nearly identical in simple descriptive statistics, but they appear differently when plotted on scatter plots.

It was developed by Francis Anscombe in 1973, to show the importance of plotting charts before building the model.



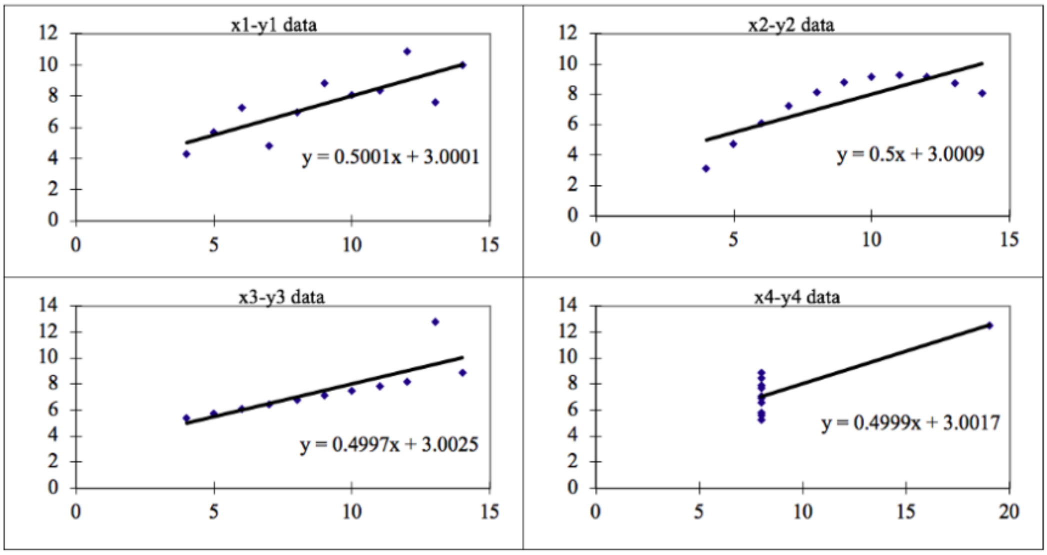
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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 observations for the above data sets are similar



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	

When the data sets are plotted on scatter plot it looks like below



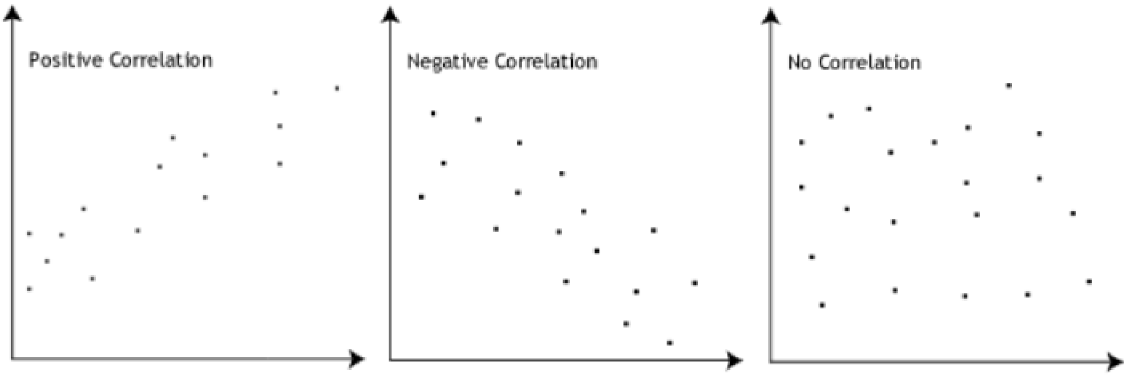
Dataset 1 fits the linear regression model pretty well. Dataset 2 did not fit linear regression model on the data well as the data is non-linear. Dataset 3 shows the outliers in the dataset which could not be handled by linear regression model. Dataset 4 shows the outliers in the dataset which could not be handled by linear regression model.

3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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



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

Feature scaling is the process of normalising the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
Normalisation – Its also known as min-max scaling, is a scaling method wherein the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

Standardisation - Standardisation or Z-score normalisation is another scaling technique whereby the values in a column are rescaled so that they demonstrate the properties of a standard Gaussian distribution, that is mean = 0 and variance = 1.

Normalisation vs standardisation - Standardisation is generally preferred over normalisation in most machine learning context as it is especially important when comparing the similarities between features based on certain distance measures. This is most prominent in Principal Component Analysis (PCA), a dimensionality reduction algorithm, where we are interested in the components that maximise the variance in the data.

Normalisation, on the other hand, also offers many practical applications particularly in computer vision and image processing where pixel intensities have to be normalised in order to fit within the RGB colour range between 0 and 255. Moreover, neural network algorithms typically require data to be normalised to a 0 to 1 scale before model training.

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

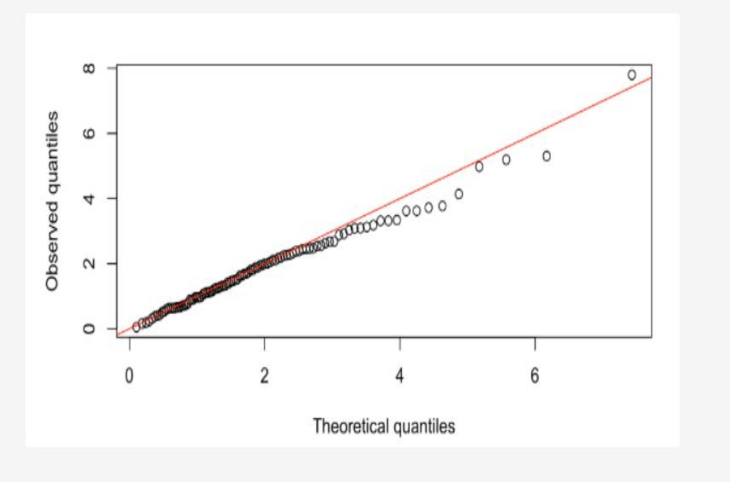
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

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

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.

Here is an example, where we are generating data x from a Gamma distribution with shape = 2 and rate = 1 parameter.

```
# Set seed for reproducibility
set.seed(2017);
# Generate some Gamma distributed data
x <- rgamma(100, shape = 2, rate = 1);
# Sort x values
x <- sort(x);
# Theoretical distribution
x0 <- qgamma(ppoints(length(x)), shape = 2, rate = 1);
plot(x = x0, y = x, xlab = "Theoretical quantiles", ylab = "Observed quantiles");
abline(a = 0, b = 1, col = "red");
```


Observed quantiles
Theoretical quantiles