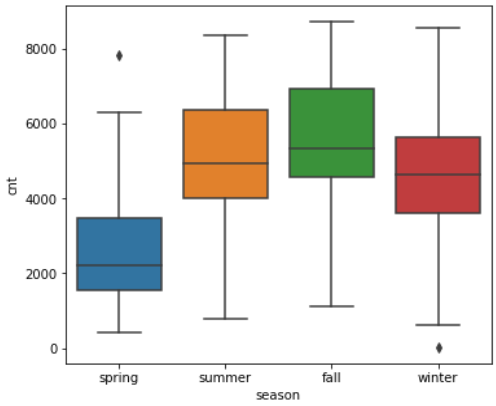
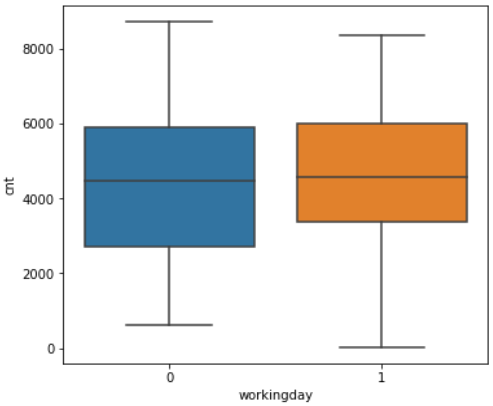
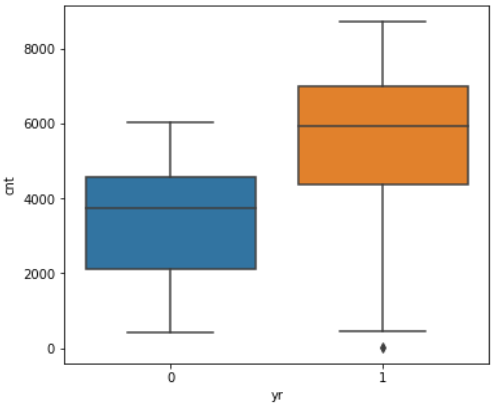
Subjective Questions

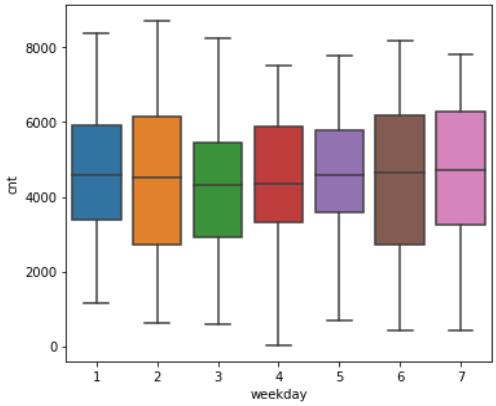
Assignment-based Subjective Questions:

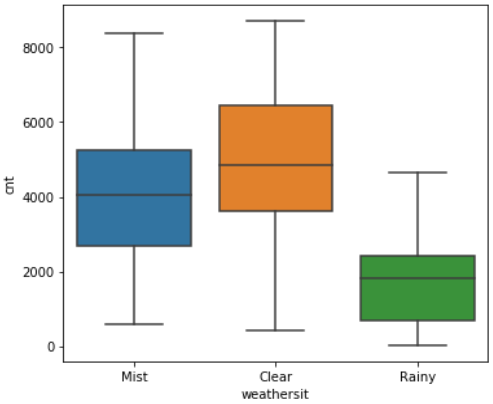
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

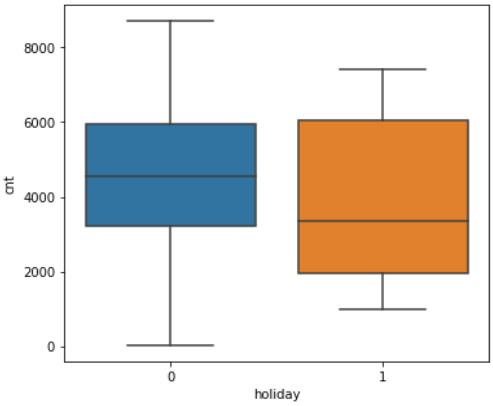


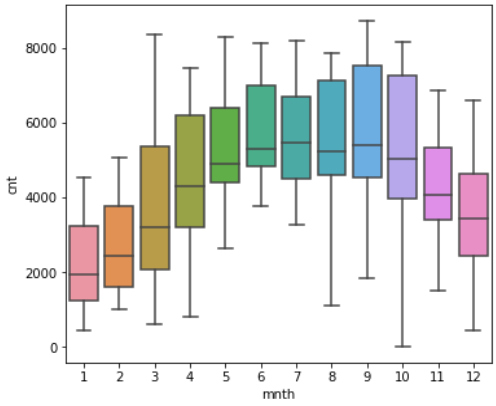












1. We can see that during **spring** there are **very less** number of riders and more during summer and fall.
2. We see there are a **lot** of riders during **2019** compared to 2018
3. If the weather situation is **rainy** we see a lot of less usage of the bikes.
4. We see **high** usage on a **non-holiday** than a holiday. We see no much difference of workingday and weekday.
5. When looked month wise we see a kind of **increasing** trend **till July** and then **decreases towards the year end.** This might be due to high work season in the mid of the year.
6. All the above interpretations are using median data.

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

It is very important to use drop_first=True as it will help us create n-1 dummy variables to a particular categorical variable. If we don't use it, then if a variable has n categories then, it will create n dummy variables. But why do we need *N-1* dummy variables? It's because let's say we have a variable called gender.

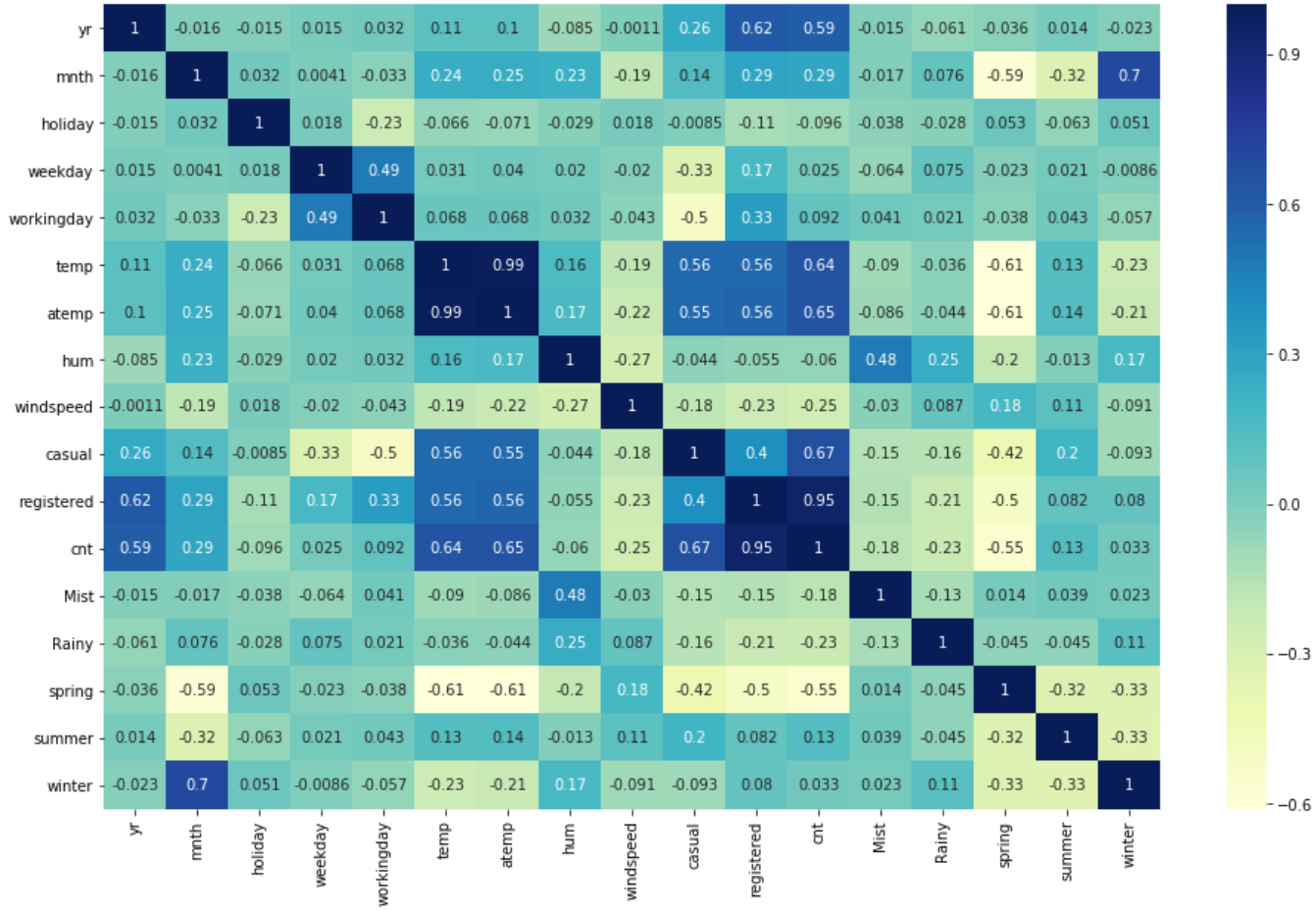
If this variable has two values Male and Female, then if we are making it as a numerical variable, then we can represent Male = 0 and Female = 1, we don't need a dummy column here as the existing column itself tells that 0 is male and 1 is female. Let's say we have another variable called Season having values Rainy, Sunny, Cloudy. In this case, if we create suppose three dummy variables, it looks something like below.

Rainy	Sunny	Cloudy
0	0	1
0	1	0
1	0	0

We can still eliminate one of the three columns, for example, Cloudy is eliminated. Then still the value 0 0 in rainy and sunny tell that it's a cloudy day. This way we can avoid having more features.

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

Amongst all the numerical variables, ***atemp*** has the highest correlation of 0.65 with ***cnt*** . And also ***temp*** has 0.64.



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

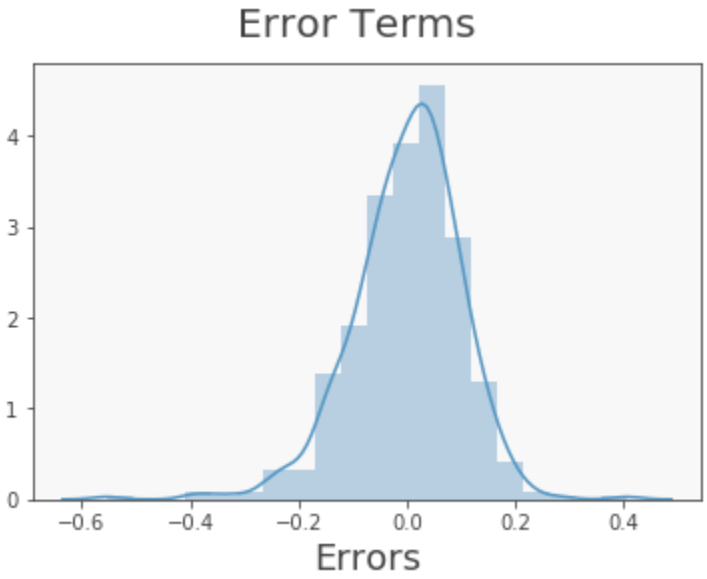
Answer:

Assumption on linearity: We finally got an equation

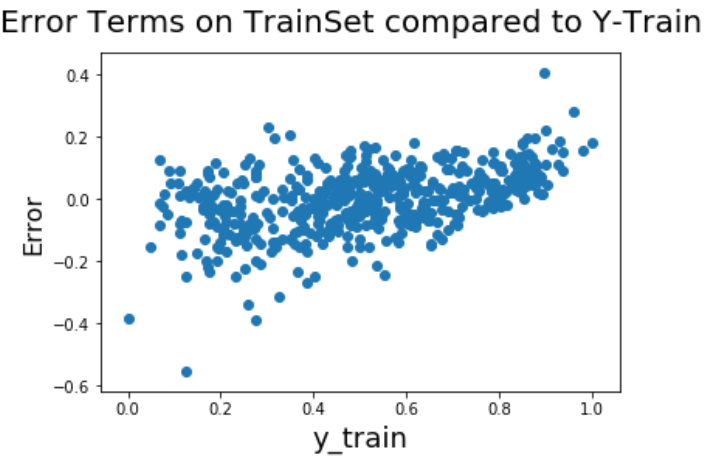
$$cnt=0.245744\times yr+0.057024\times workingday-0.192617\times windspeed-0.090099\times Mist-0.320724\times Rainy-0.238120\times spring-0.040292\times summer-0.118632\times December-0.123145\times January-0.112736\times November+0.055846\times September+0.066525\times Monday+0.535951$$

Residual Assumptions:

- ***Normal Distribution:*** We received a normal Distribution of the error terms.



- ***Zero Mean***: In the above figure you can see that the errors are having a mean of zero as well.
- ***Constant Variance/homoscedasticity and Independence***: In the figure below, we can see that there is a similar pattern across the whole dataset which implies there is a constant variance as well, and no specific patterns are observed in different regions which also implies that error terms are independent of each other



Assumptions on estimators:

1. Independent variables are measured without error.
2. Independent variables are linearly independent of each other, no multicollinearity. The model evaluated has been made sure that there are no VIF values >= 5

	Features	VIF
2	windspeed	3.92
1	workingday	3.20
5	spring	2.38
0	yr	1.87
6	summer	1.77
8	January	1.63
3	Mist	1.54
11	Monday	1.54
9	November	1.22
10	September	1.20
7	December	1.16
4	Rainy	1.07

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

Answer.

- The Top Three predictors are in terms of absolute values of Beta values.
    1. Rainy - 0.32
    2. yr +0.24
    3. Spring -0.23

- Top Three Predictors affecting negatively are:
    1. Rainy - 0.32
    2. Spring -0.23
    3. WindSpeed -0.19

- Top Three Predictors affecting positively are:
    1. Yr +0.24
    2. Monday +0.066
    3. Workingday +0.057

	0
Rainy	-0.320724
spring	-0.238120
windspeed	-0.192617
January	-0.123145
December	-0.118632
November	-0.112736
Mist	-0.090099
summer	-0.040292
September	0.055846
workingday	0.057024
Monday	0.066525
yr	0.245744
const	0.535951

General Subjective Questions:

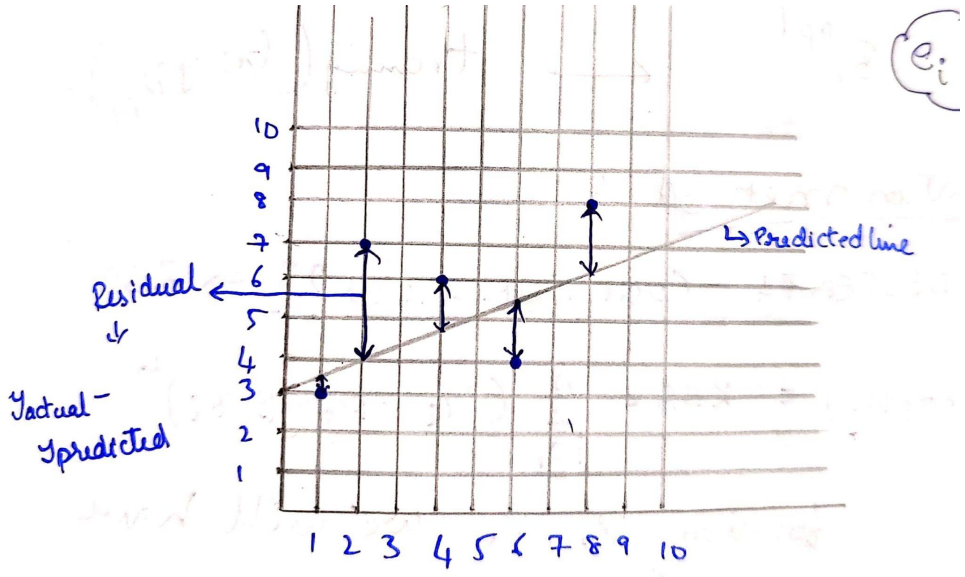
1. Explain the linear regression algorithm in detail.

Linear Regression is Supervised learning method that falls into the regression technique. In this, using values of training data with N-Independent variables, we predict the target dependent variable.

There are two types of Regression:

1. Simple Linear regression : Where we have 1 independent and 1 dependent variable
2. Multiple Linear regression: Where we have N independent and 1 dependent variable.

The idea behind Linear regression is, we assume there is a linear relation between any independent variable and the target variable. We try to fit a line/hyperplane which should be the best fit in terms of prediction.



Example: given the area of a house, predicting the rental price of the house in Hyderabad.

The equation of a linear regression best fit line is of the form:

1. y = B0 + B1.x for a single independent variable.
2. And y= B0 + B1x1 + B2x2 + B3x3 + …. + Bnxn for n independent variables. The parameters to be learnt are the B coefficients.

In order to find the best parameters, we look at a term called Cost-Function which needs to minimized. We look at something called **Residual Sum of Squares(RSS)** which is the sum of squares of error terms.

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2 + \ldots + e_n^2 \text{ where } e_i = y_i - ypred_i$$

- We can use techniques like **Gradient Descent Algorithm** and find the best values of the coefficients after some iterations using a suitable learning rate.
- R2 score is a good way of understanding how good a linear model is.

There are few assumptions of linear regression:

1. No Multicollinearity amongst the independent variables(Analysing p-values and VIF scores).
2. Errors terms are independent of each other
3. .Error distribution has a mean of 0.
4. Homoscedasticity i.e., const variance in the error distribution.
5. There is a linear relationship using independent variables v/s predicted variable.

This way linear regression has its own assumptions and considerations with many perks to build simple models.

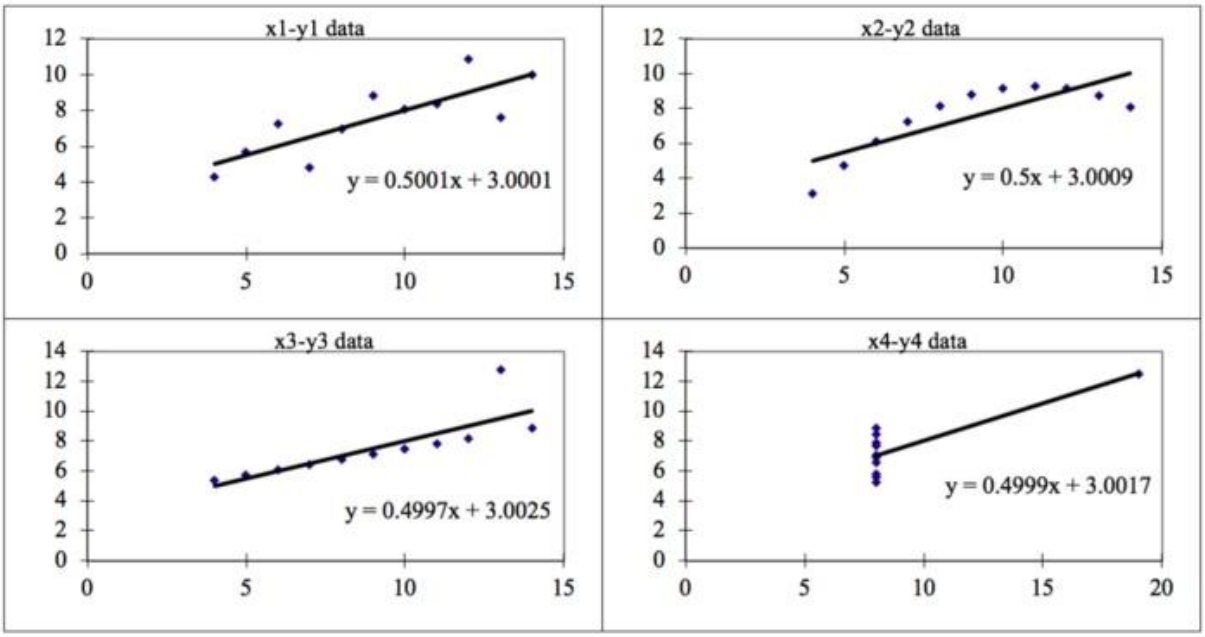
2. **Explain the Anscombe's quartet in detail**

Anscombe's quartet is a group of four datasets which have similar descriptive statistics like mean , R2 , standard deviation etc.. but when plotted as a scatter plot, they all show a different distribution.

This was observed by a statistician **Francis Anscombe** in 1973 where he showed a special dataset as below split into 4 sets.

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	

As we can see they have similar statistics, but when plotted as scatter plots, it looked like



Let's explain each of the quartets:

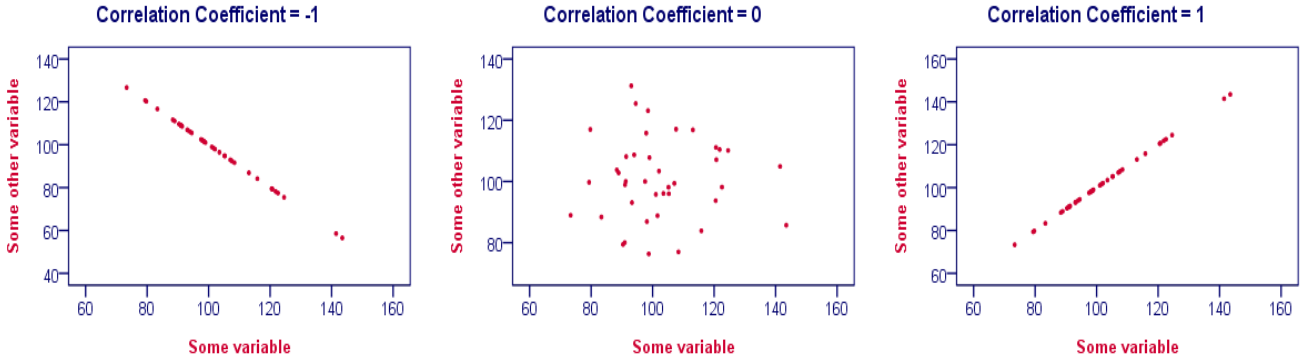
1. The first data looks like it's a good fit for linear regression.
2. The second plot looks like its not a good idea to use linear regression as the data looks to be non-linear.
3. The third plot if we see, it was almost a very good linear fit, but due to outliers there a little realignment of the line.
4. The fourth shows outliers in the dataset which the model couldn't handle though there is regression line is being fit.

With these above descriptions, we were able to see how the linear regression model was fooled with the data being selected. This way anscombe's quartet gives a very good analysis on the data by which we can decide upon choosing linear regression as a good model or not!

3.**What is Pearson's R?**

Pearson's Correlation Coefficient **R** is a value ranging from -1 to +1 explaining how two numeric variables are related to each other

1. **R value cannot be below 1:** If the R value is as close to -1 then it means that if one variable value increases then the other variable value strictly decreases.
2. **R Value being 0 :** This indicates that both the variables are independent of each other
3. **R value cannot be above 1:** If the value of R is close to +1 then it means that if one variable increases then the other variable value also increases



The increase/decrease of one variable with other depends on how close the R value is to -1 and +1. R value is very sensitive to outliers as well. The Person R formula is given by:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

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

Scaling is a concept of bringing all the data values in a similar range usually 0-1. Scaling is performed for various reasons,

1. It will help us understand the Beta values comparatively much better.
2. It will help the gradient descent algorithm to converge quickly towards minima. This is because we can avoid different step sizes for different variable ranges.

If values are in different ranges and units like for example , 5000 grams and 5Kgs are one and the same, but if not scaled properly, variation in the y-ypred value will be more for grams than kgs which will end up showing a greater Beta value for this variable, which might mislead if other variables are on a shorter scale like KGs.

Scaling Types:

Normalized scaling: This will make sure that the values after scaling when plotted, show a normal distribution. One popular normalized scaling technique is **Min-Max Scaling** where the formula looks like,

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

Standardized scaling: This will make sure that the distribution of the standardized values has a mean of 0 and a standard deviation of 1.

$$X' = \frac{X - \mu}{\sigma}$$

Normalized scaling is affected with outliers as it is dependent on max(x), whereas standardized values arent affected. We used normalization when we want to get a normal bell curve in the data, and if already have a bell curve kind of data we proceed to standardize if possible to make the data more focused.

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

Answer: Since VIF= 1/(1-R2), it will be infinite when R2 is 1.

- That is when there is a perfect correlation of a variable with some other variable, we need to drop such variable which is showing a perfect multicollinearity
- Infinite VIF of a variable implies that it can be expressed exactly as a linear combination of all the other independent variables which have infinite VIF as well
- R2 of 1 implies that the other variables are explaining 100% variation of that variable which will show VIF as infinity.

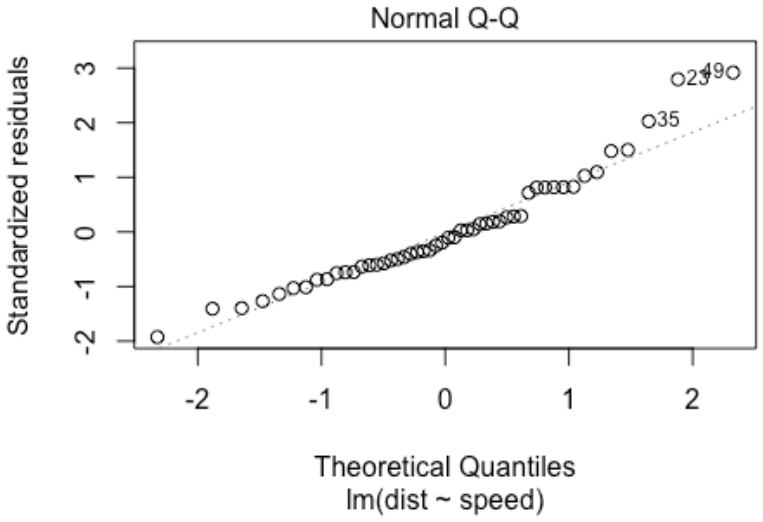
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**.

Q-Q plot will help us understand if two data sets have come from a similar kind of distribution. When we receive Train and test data sets separately, these plots will help us understand if we obtained data from a similar kind of distribution.

How is Q-Q plot formed?

If we have 9 data points. We take a normalized data and starting from the mean, we create half no. of quantiles to the left and other half to the right of it. Totally creating 10 theoretical quantiles with equal width.

Further, we calculate the z-scores of all these data points and again create 10 actual quantiles at these z-scores. Now plotting these values on a graph with theoretical quantiles on the x-axis and actual quantiles on the y-axis will give us the Q-Q plot. An example QQplot is shown below.



Using the plot we can observe if two datasets,

- Come from a common population with common distribution
- Have a common location and scale
- Have similar distribution shapes
- Similar tail behavior