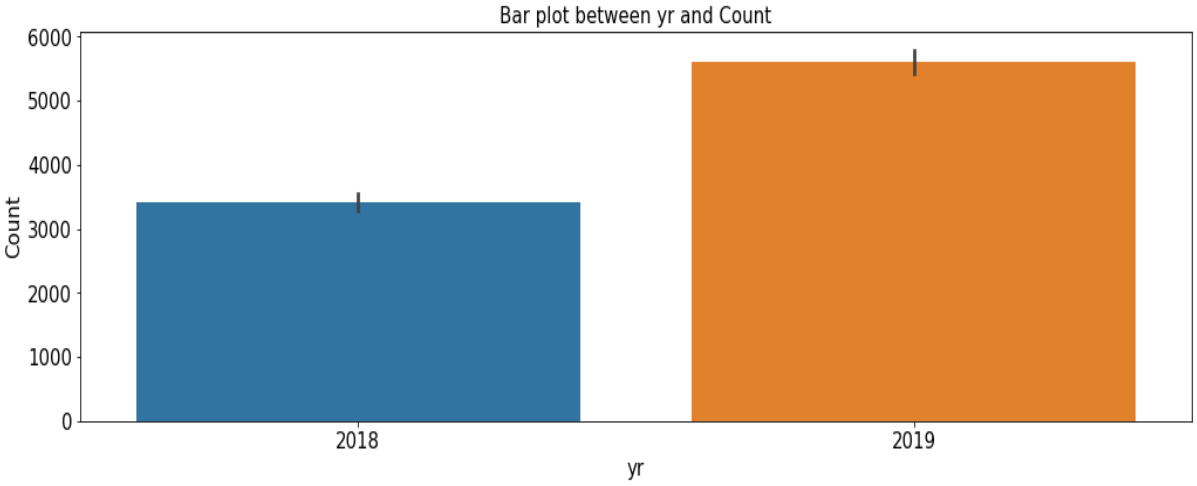
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

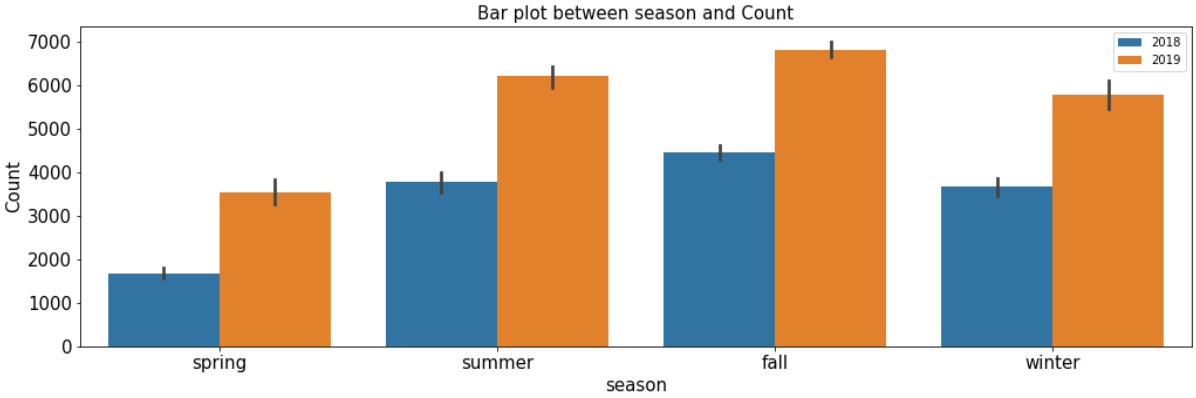
Ans: I have included analysis which showed one trend or another and the rest has been discarded, the analysis are as follows and I will explain it with the help of graphs to support my inference

Year vs Count:



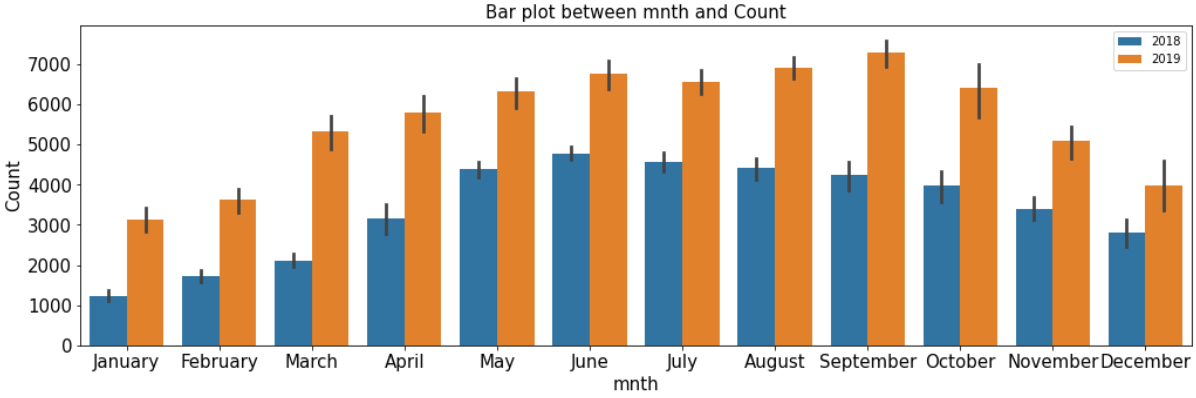
From the plot it is obvious that there is an increase in the number of bike counts, which shows us positive growth of the company, it could mean as the business gets older and older the number of customer increases (excluding the covid years), this although is constrained by the yearly performance, so far we have only seen the data for 2 years

Season vs Count with hue as year:



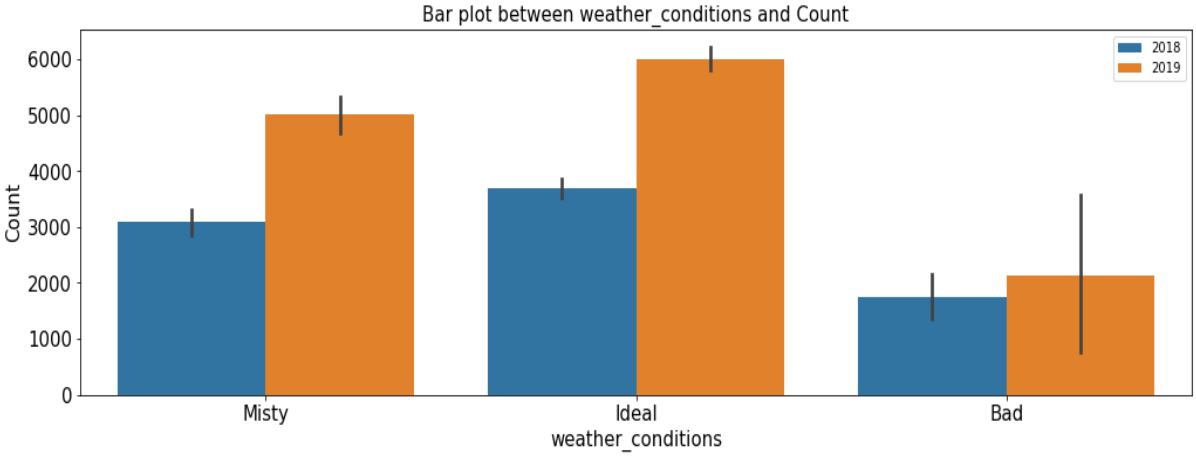
The same trend can be seen where there is growth in 2019, the depended variable that is Count, is the highest in summer and fall, This can be attributed to the fact that due to summers, most students have their holidays and people who work tend to take holidays during this season there is a rise in bikes being rented

Month vs Count with hue as year:



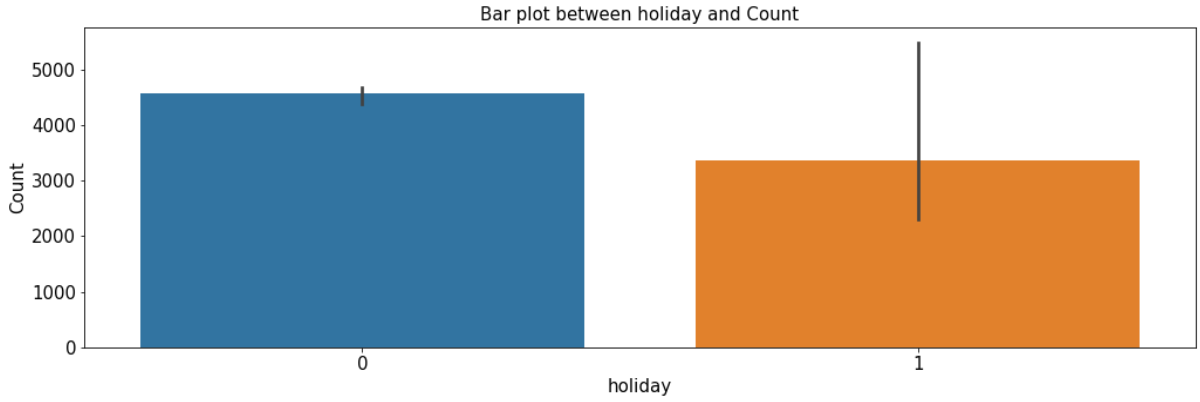
The months belonging to summer are from June to August and months belonging to fall are from September to November and our plots confirms our previous inference as these months attract the most bookings.

Weather Conditions vs Count with hue as year:



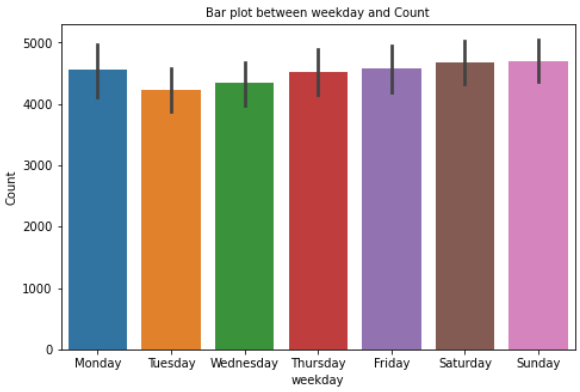
People usually tend to rent bikes when the weather is Ideal or Misty(with some visibility) we observe that the count goes up both for 2018 as well for 2019 for both weather conditions

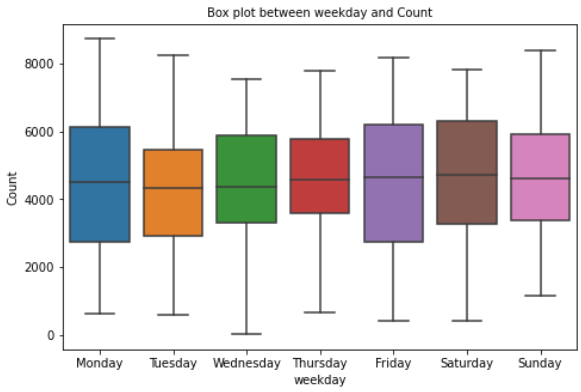
Holiday vs Count:



People who tend to work or study tend to book more bikes

Weekday vs Count:





The bike count are similary throughout the week indicating a good sign that people are renting bikes daily from this company

2)Why is it important to use drop_first=True during dummy variable creation? (2 marks)

ANS: Let's analyse this step by step with first understanding how we create dummies. We use pandas' **get_dummies()** function to create dummies which creates "K" dummies and the reason we use **drop_first = True** is to drop one level of category which is beneficial for our model, so it can converge or rather learn at a faster rate and to reduce redudancy; there is no harm in keeping that extra category it's just so for the convenience of our model, we will explain this better with the help of an example,

Example: Let's assume there exists a dataframe which has the column; Levels and it has 3 unique levels; Level_1, Level_2, Level_3, when

drop_first = False then all three levels will be included, if it is set to **True** only two levels will be included. It goes by the logic that if one variable is not Level_1 or Level_2 then it is obvious that it is Level_3

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

Ans: temp and **atemp** has the highest correlation with the dependent variable but we dropped atemp later on due to its high collinearity with temp

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

Ans: There are four assumptions of Linear Regressions and they are as follows, I will explain each one them first and then proceed with how I validated them

1. **There should exist some linear relationship with X and Y;** which we already saw between temp vs Count, atemp vs Count and a moderate linear relationship with windspeed vs Count where Count was our targeted variable

2. **Error terms should be normally distributed –** We used our training set to plot the residuals which is nothing but the error terms (y_train – y_train_pred) y_train_pred contains the values predicted by our model using the training set; we plotted the residuals using histogram which showed the error terms to be normally distributed around zero which held our second assumption

3. **Error terms should be independent –** For this we plotted our predicted values on the X axis and the residuals on the Y-axis and observed that there was no relationship between them, no shape or pattern was observed and the points were scattered, this was enough to validate our 3rd assumption

4. **Error terms should have constant variance(Homoscedasticity) –** We plotted the residuals with the predicted values and observed no pattern indicating that error terms were scattered around.

5)Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANS: The top three features contributing significantly are:

1.temp

2.snow_rain

3.yr

General Subjective Questions

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

ANS: Linear regression is a type of Machine Learning algorithm which is based on supervised learning which means we already have the outputs we can compare our predicted values with and simultaneously improve our model. This Regression model predicts a value(dependent value) based on other values (independent variable). In mathematical terms it tries to fit a best fit line through a given set of data points, we can say that the line is the predicted value and the other set of points are the actual values

Linear Regression is of two types:

1) **Simple Linear regression (Y = mX + C):** It has one dependent value and one independent value. Over here we are trying to fit a simple line

2) **Multiple Linear regression (Y = C + m1X1 + m2X2 +……+ mnXn) :** It has one dependent variable and multiple independent variables. Over here instead of trying to fit a line we try to fit a hyper plane

Y – The dependent variable
X or X1, X2……Xn – Independent variables
m or m1,m2…mn – Slope of the line
C – The constant or the Y intercept

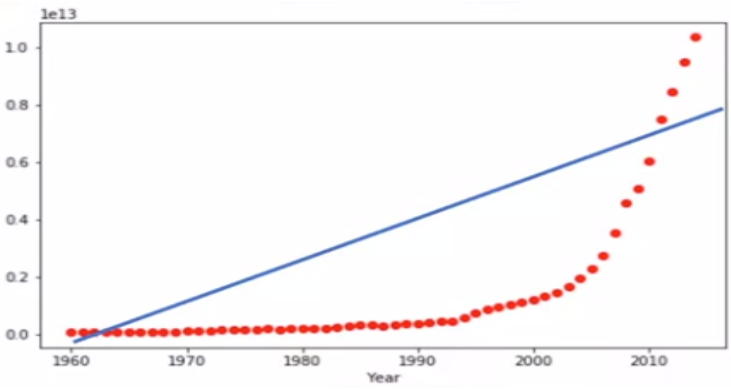
The Linear regression algorithm works on **m** and **C** and updates it so that the model finds a line best suited for those data points; in layman terms it tries to find a line best suited for a particular problem. It tries to reduce the residual sum of squares so that it can optimize the line w.r.t the expected value
It uses the Residual sum of squares and Total sum of squares to give us an idea of the accuracy of the model with "N" features; Its upto us which features we need to include or exclude based on our domain knowledge, in other words it relies on minimizing the error between the actual value and the predicted values.

This leads us to the problem of **over fitting** and **under fitting**, **overfitting** is where the model memorizes the data instead of learning it; this is the case where there are too many variables and making the model so complex and it ends up failing to generalize, underfitting is where the model is just too simple and lack features which leads to the model failing to learn from the training set

We also need to **validate some assumptions of Linear regression** and violating any one of them can lead to serious errors in our model with our model failing to learn or generalizing on the dataset

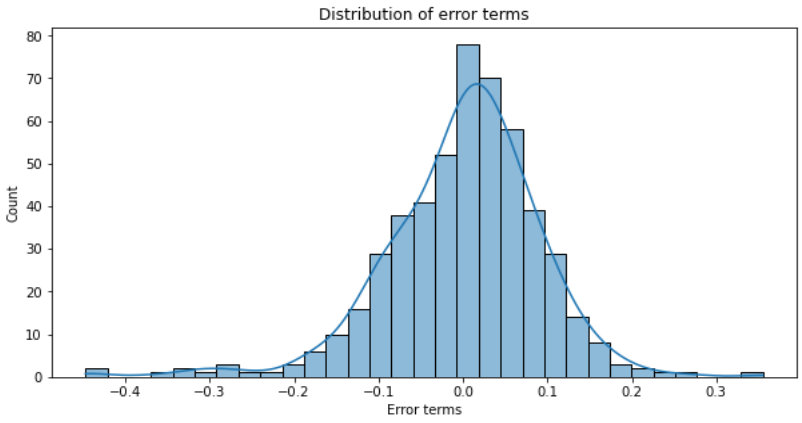
There are some assumptions to validate in linear regression on the training set and violating any of these can introduce some errors in our model

1. **There should exist some linear relationship between X and Y**: There is no point in utilizing this algorithm if there doesn't exist any linearity between X and Y



The Blue line indicates a linear relationship and the red one a non-linear

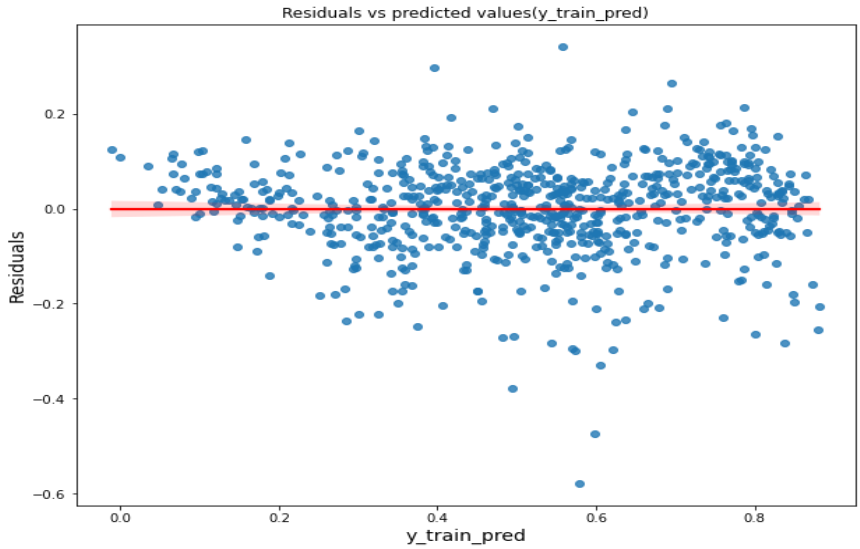
2. **Error term should be normally distributed around zero**: The mean is also very close to Zero



3. **Error terms should be independent of each other &**

4. **Errors have constant variance :**

   From the graph we can make out that the points are scattered around thereby validating our third and fourth assumption



In the end we save our model if we are satisfied with its accuracy and can use it for further predictions

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

ANS: According to Wikipedia **Anscombe's quartet** is a set of four data sets each of them having extremely similar descriptive statistics (Mean, standard deviation etc). It was devised by a statistician who goes by the name of **France Anscombe** ; he wanted to

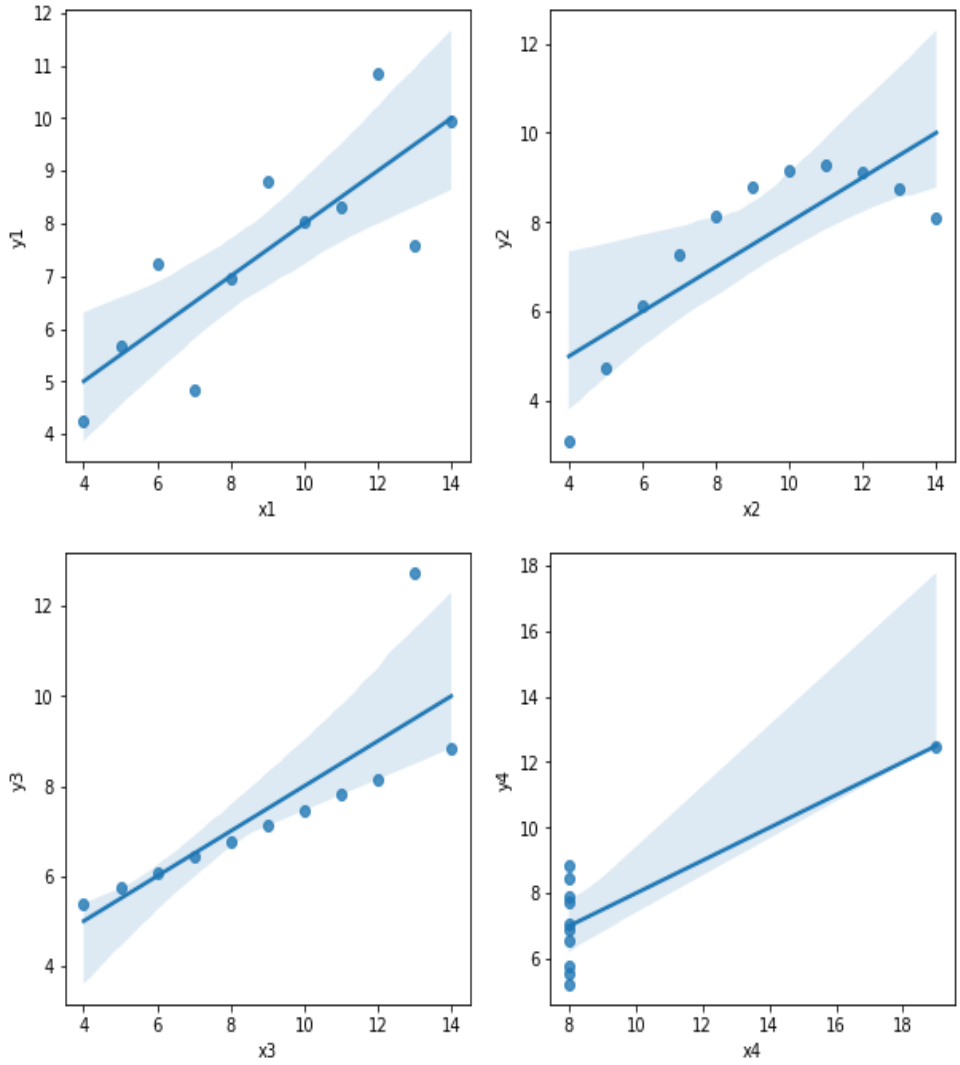
show the need of graphing the data and not relying only on descriptive statistics to conclude. Following is a dataframe consisting of Anscombe's quartet and its descriptive statistics;

	x1	y1	x2	y2	x3	y3	x4	y4
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47

```
round(df.describe(),2)
```

	x1	y1	x2	y2	x3	y3	x4	y4
count	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
min	4.00	4.26	4.00	3.10	4.00	5.39	8.00	5.25
25%	6.50	6.32	6.50	6.70	6.50	6.25	8.00	6.17
50%	9.00	7.58	9.00	8.14	9.00	7.11	8.00	7.04
75%	11.50	8.57	11.50	8.95	11.50	7.98	8.00	8.19
max	14.00	10.84	14.00	9.26	14.00	12.74	19.00	12.50

As one can see the descriptive stats are almost similar across the X and Y, so one would assume that the graphs or a scatterplot of these points will be the same as well; we will plot the graphs and then give our vinal verdict on it;



In the **top left** column we can see that there exists a linear relationship between X and Y

In the **top right** corner it is obvious that there is a non linear relationship between X and Y

In the **bottom left** corner there exits a perfect Linear relationship except for that one outlier maybe for this a different regression line would fit

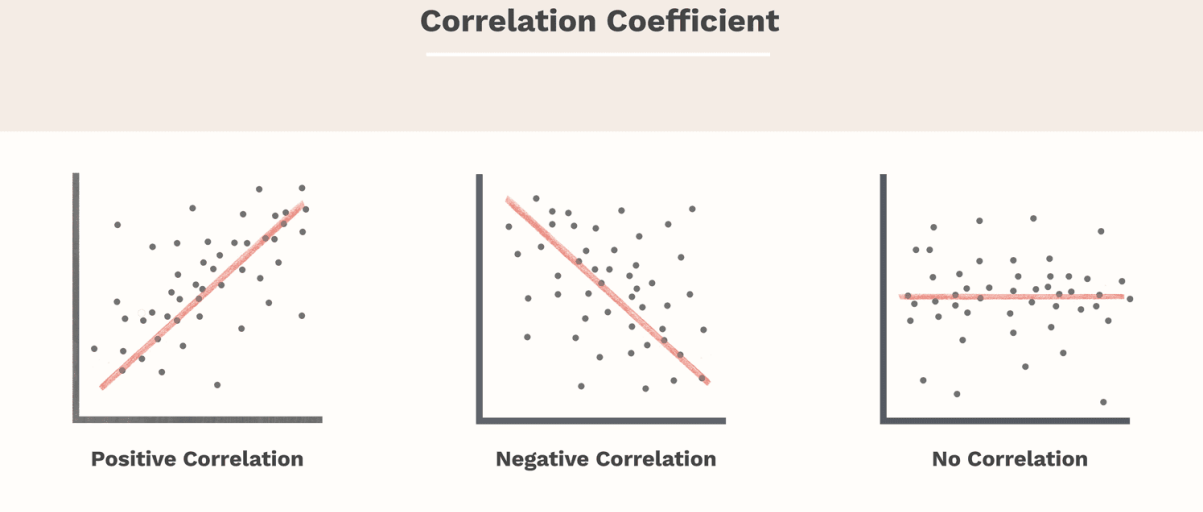
In the **last figure** we see that the points are sort of stacked as the value of X is constant with the exception of one point, this shows that one outlier is enough to promote high correlation coefficient.

And from the graphs above one can clearly feel the need to visualize the data instead of only relying on basic descriptive statistics.

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

ANS: Pearson's R or Pearson's correlation coefficient gives us the linear correlation between two variables in other words it gives us the summary of strength of correlation between those two variable and it is given sum of products of X and Y subtracted by product of individual sum of X and Y divided by their standard deviation,

Pearson's r can take values ranging from -1 to 1 and can be categorized as Positive Correlation, Negative Correlation and No Correlation.



From the first graph we can see a **Positive correlation(greater than zero)** where one variable goes up the other variable goes up as well

From the second graph we can see a **Negative correlation(less than zero)** where one variable increases the other decreases

The third graph depicts **No correlation(equal to zero)** where the points are scattered everywhere and no visible pattern can be identified

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

ANS: Scaling or more commonly known as **Feature scaling** is done to standardize the data in Machine Learning in the pre processing phase ; i.e collapsing the data into a range of similar values this is done so that the larger values are not given more weightage for example The weight measured in pounds will always be greater than weight measured in Kgs which is true but logically it conveys the same thing; so to remove this mismatch we do Feature scaling but we don't do it in Linear regression as there is only one independent variable. This also helps the algorithm in increasing the accuracy of the model. The difference between **Normalized(MinMax)** and **standardised** scaling are as follows:

Normalized scaling	Standardized scaling
Transforms the data in the range of [0,1] or [1,-1]	Transforms the data by removing the mean and scaling the data about mean being 0 and SD 1

It used when we do not known the distribution of the features in our data	It is used when we are certain of the distribution of data(normal distribution)
It is sensitive to outliers	It is not as sensitive to outliers compared to normalized scaling
It is calculated by subtracting the data in individual column by the minimum value and diving it by the range of that column X – Xmin/Xmax - Xmin	It is calculated by subtracting the data in individual column by the mean of that column divided by the standard deviation of that column X – Mean/ SD

5)You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANS : The formula for VIF (Variance Inflation Factor) is given by **1/(1-R^2)**. When VIF is below 5 the Multi collinearity is relatively low; when it is in the range of 5-10 one should investigate the variable and anything beyond 10 should be eliminated at once.

Before we explain VIF let's understand the theory behind VIF, what VIF does is calculate the score of each dependent variable by using (1/(1 – R^2), it segregates dependent values and calculates how other

dependent variable can explain other dependent variables

For simplicity sakes you can assume that it sort of divides dependent variables into train and test.

The reason for VIF being “inf” is as the R^2(R square) approaches 1 i.e the dependent variables are able to explain the other dependent variables brilliantly; the value in the denominator becomes closer to zero as it indicates extremely high correlation between dependent variables and needs to be dropped after careful consideration of P-values as changing or dropping even one variable changes the VIF of other variables.

6)What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANS : Q-Q plot or Quantile-Quantile plot is a probability plot and is plotted using the quantiles of (x,y) where x represent the quantile of the first dataset and y represents the quantile of the second dataset. We compare two probability distributions (x,y)

It is used to determine if the two dataset come from a common distribution, we plot the points on the X,y axis and if the distributions are similar on both the datasets, then the points will (approximately) lie on the line with slope 1 i.e the identity line which has

an angle of 45 degrees from the X-axis (Normal distribution)

Let’s take an example from our Bike sharing assignment by taking the distribution of error terms or residuals, now according to an assumption in Linear regression residuals should be normally distributed; well approximately because the model which is built cannot be 100% linear; so a line is drawn at a 45 degree angle from the X-axis to fit normally distributed data points(residuals) and we see if the points are on the line or around the line, if we see such pattern then our residuals are normally distributed.