General Subjective Questions

1) **Explain the linear regression algorithm in detail.**

   -Linear Regression Algorithm

   The linear regression algorithm is based on the supervise learning method.As the name itself suggests it performs a regression task where a model is trained to predict the value based on the other variable.

   Mathematically linear regression can be represented

   y = mx +c , where

   y = dependent variable

   x = independent variable

   m is the slope of the line

   c = y intercept of the line

   When training the model, the model gets to obtain the best regression line by finding the best m and c values.

   After finding the best m and c values, the model aims to predict the y value such that there is a very minimal difference between predicted value and the true value. This is called the cost function.

   The cost function is the **Root mean squared Error** between the predicted y and the true y.

   We use the gradient descent in order to minimize the RMSE and achieve best fit line we start with random m and c values , and will keep on updating them iteratively till we have reached a minimal value point.

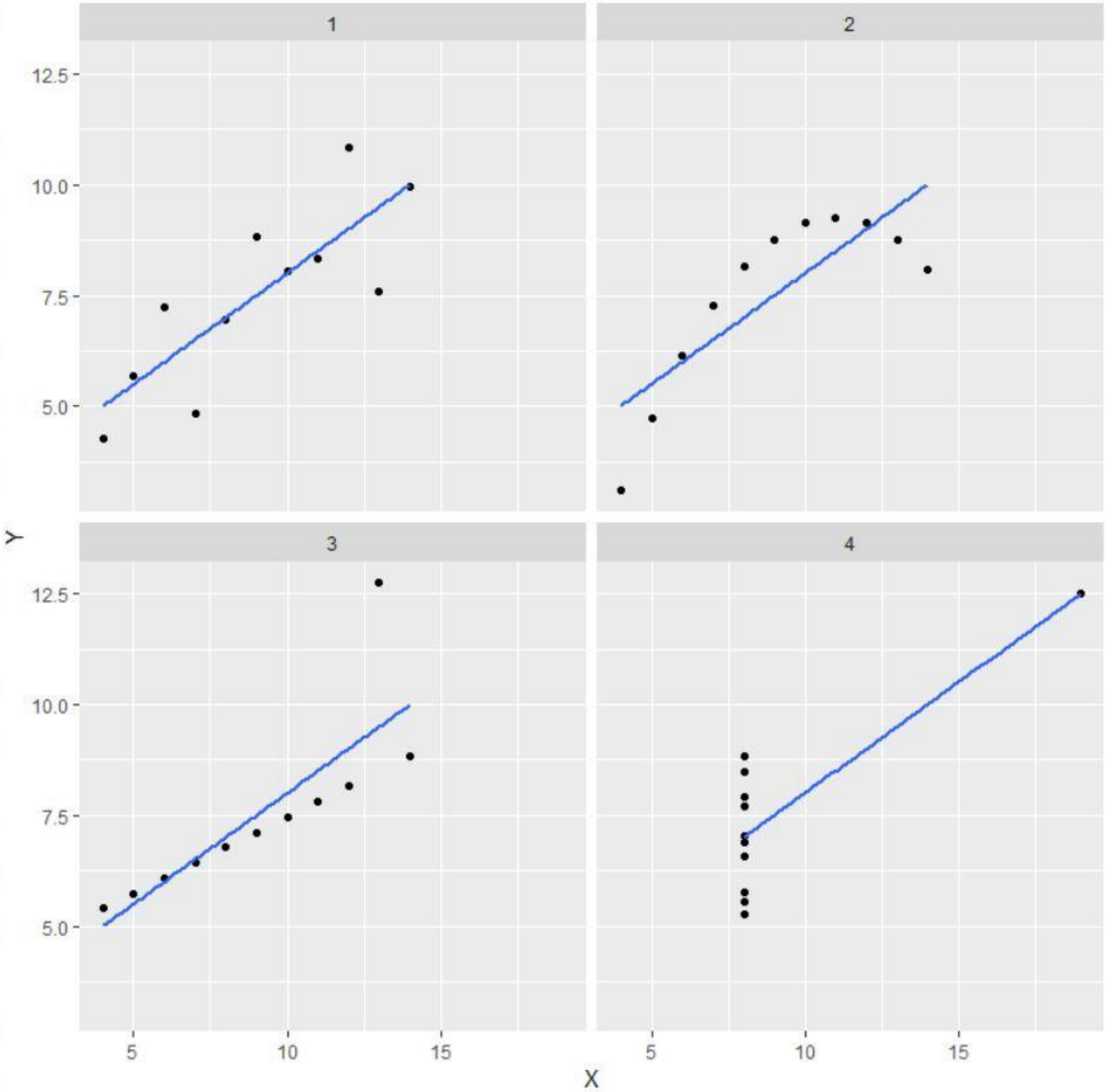
2) **Explain the Anscombe's quartet in detail.**

   -The quartet was designed by the statistician Francis Anscombe to explain the importance of plotting of graphs before analyzing the data and also the effect of outliers on statistical properties.

   Anscombe's quartet consists of 4 data sets , each of which contain 11 (x,y0 points. They have almost identical simple statistical properties but appeared differently when observed through graph.

   The $ datasets have the same

-Mean(x) : 9	Mean(y):7.5	Corr(x,y)=0.816
Sd(x): 3.32	Sd(y):2.03	

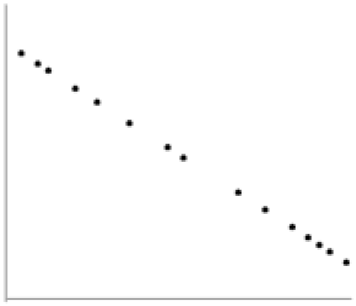


As you see all the 4 datasets have identical properties but from graph it seems different.

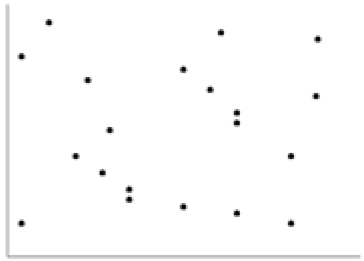
- Top left graph indicates x and y have linear relationship
- Top right graph indicates x and y have a non-linear relationship
- Bottom left in this one , the points seems to have a linear relationship except for the last point which is an outlier , which os far away from the line
- Bottom right it shows only one high leverage point is enough to produce a high correlation coefficient

3) **What is Pearson’s R?**

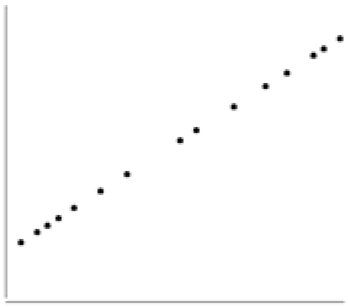
- Pearson’s correlation coefficient or simply Pearson’s R is the measure of the strength variables which explains the linear association between them.
  If both the variables are directly proportional i.e if both x and y tends to go up and down together then the Pearson’s R is positive else if the variables are inversely proportional i.e when one goes down other goes up.
- The Pearson’s value lies between –1 and 1 ,however we have 3 ideal cases.



The data lies on a perfect straight line, the slope is negative. in this case r=-1



The data is scattered all over, there is no linear association, here the r=0



The data lies on a perfect straight line, the slope is positive. in this case r=1

Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) * stdv(Y))

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

Scaling is a technique to normalize\standardize the range of independent variables, it is performed during data pre processing

When we are training a model in which a coefficient has to be calculated , all the variables used for the training must be in a comparable scale to increase interpretability mainly to which scaling is used.

Two important scaling techniques are

Min- Max Normalization	**Standardization**
In this technique the values are shifted and rescaled they end up between 0 and 1	In this technique the values are centered around the mean with a std deviation value is 1.
If outliers are present it will be affected by normalization	Even if outliers are present it , it will not be affected by standardization
Mostly helpful in cases where data have distribution other than Gaussian	Mostly helpful in cases where data follows Gaussian distribution
X ` = (X-Xmin)/(Xmax-Xmin)	X` = (X- μ)/ σ

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

Variation inflation factor (VIF) , it is a measure of multi collinearity of the variables in the mulit linear regression.

MultiCollinearity is a issue for the interpretation, detecting mulitcollinearity is same as detecting association in the predictors of the model.

If there is an exact collinearity among the predictors of the model then the VIF value becomes infinite.

The VIF values that can be considered

VIF > 10 , the variables should be eliminated

VIF > 5 , it is OK to be inspected

VIF < 5 , No need to eliminate the variable , consider for prediction

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

QQ is a graphical technique used to determine if the two data sets comes from population with a common distribution.

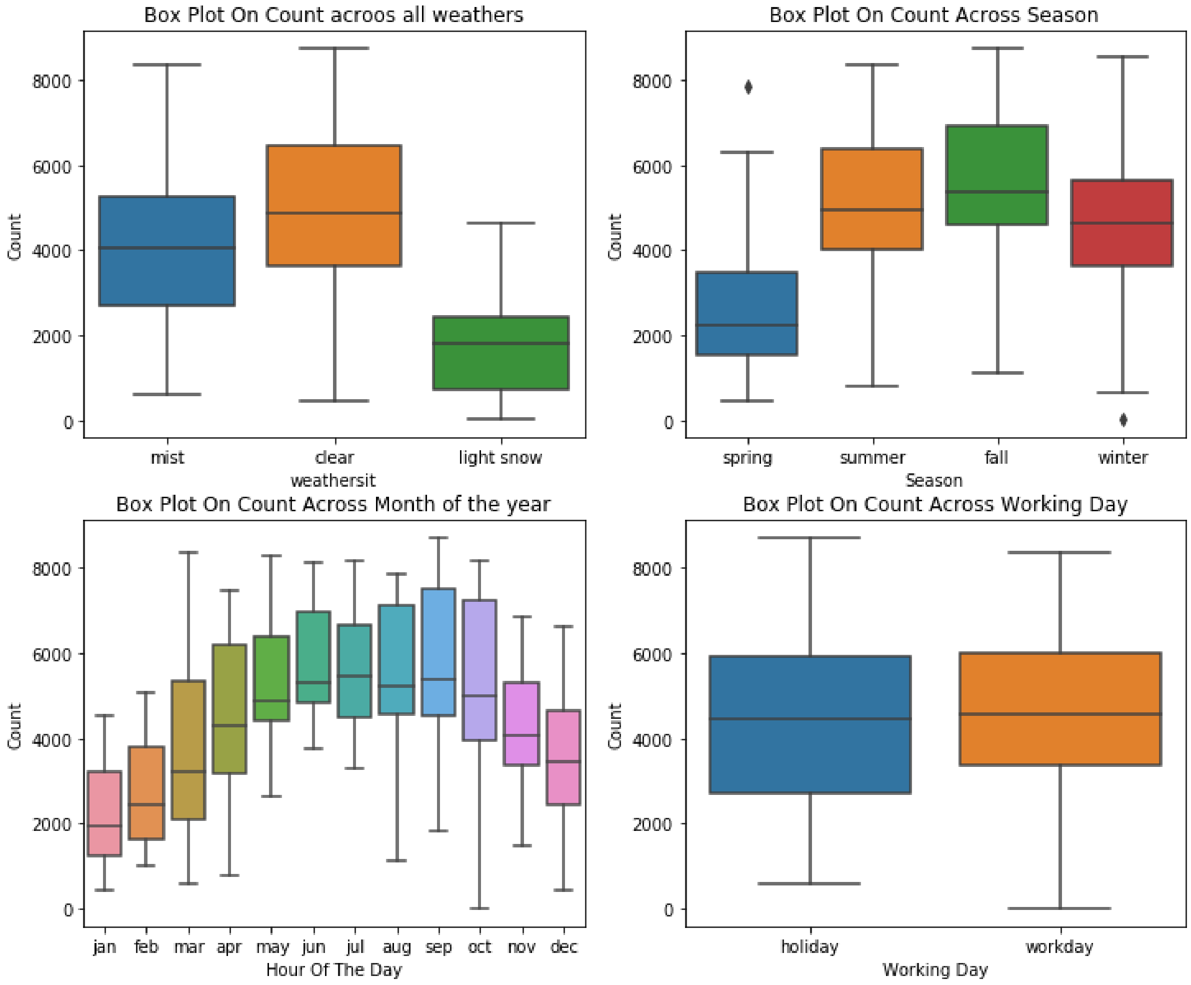
We use a QQ plot in linear regression when we receive a train and test dataset separately instead of us splitting, so that we can confirm both the data sets come from the population of the same distributions.

- It can be used even if the sample sizes are not equal
- It helps in checking if two data sets have come from population with same distribution , common scale , similar distributional shapes

As concluding QQ plot is a plot of quantiles of the train data set against the quantiles of the test data set

Assignment-based Subjective Questions

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


Box Plot On Count acroos all weathers

Box Plot On Count Across Season

Box Plot On Count Across Month of the year

Box Plot On Count Across Working Day

- The bike count is less during light snow and high during clear weather
- The bike count is more during fall and less during spring
- The demand for bike is more in the month of July and less in January
- The median of the bike count is almost same during working day and holiday

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

   -To avoid the redundancy features the drop_first which allows the reference variable to be dropped

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

   -Temperature 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?**
   - The error terms graph during residual analysis, the mean of the error is zero the errors are normally distributed.
   - The graph b/w the residual value and predicted value there is no particular pattern so the homoscedasticity assumption is confirmed
   - There are linear relationships between the dependent and the independent variables

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - light snow -0.2753
   - spring -0.2666
   - yr 0.2511
   - windspeed -0.1886
   - sun 0.0592
   - workingday 0.0538

   The above mentioned are the top predictors which influence the bike demand

Among them light snow , spring and windspeed shows –ve effect and year ,Sunday and workingday shows +ve effect