Subjective Questions And Answer:

Assignment-based Subjective Questions:

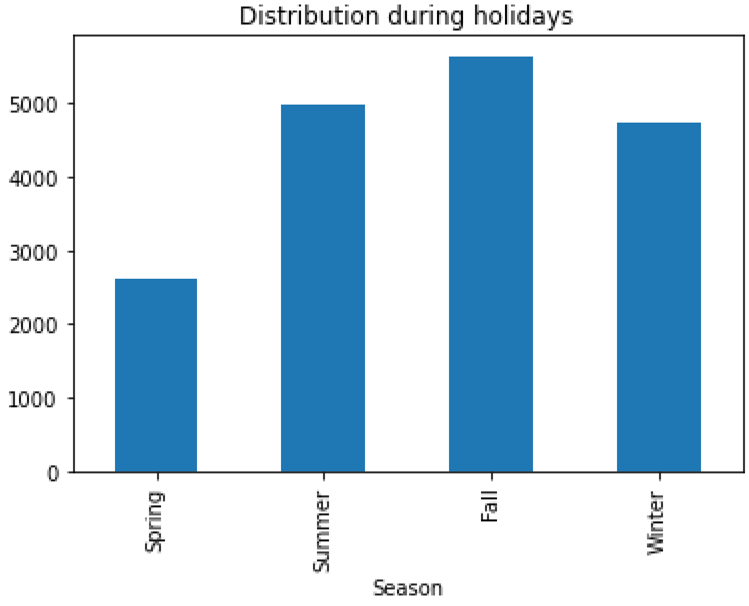
Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Based on year Variable, 2019 has good sales in bikes than 2018,May this happen due to increase of popularity

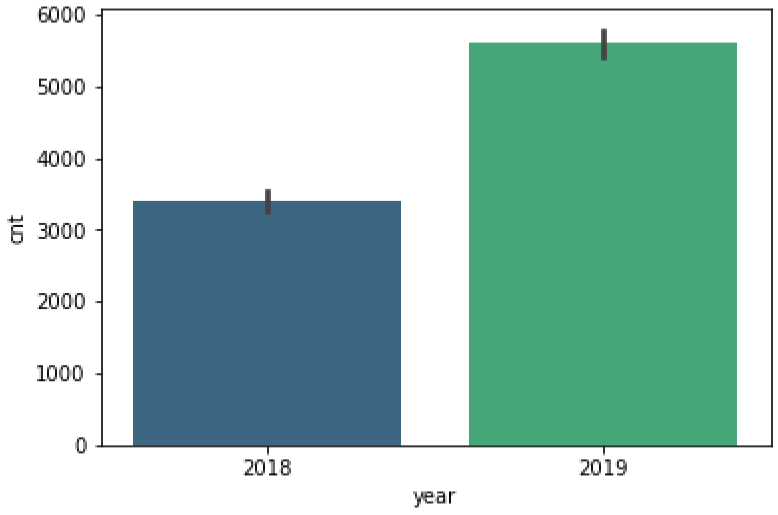
If we take the season The Spring has very low sales than any other Season

If we take holiday,in my observation working day has better sales than holiday

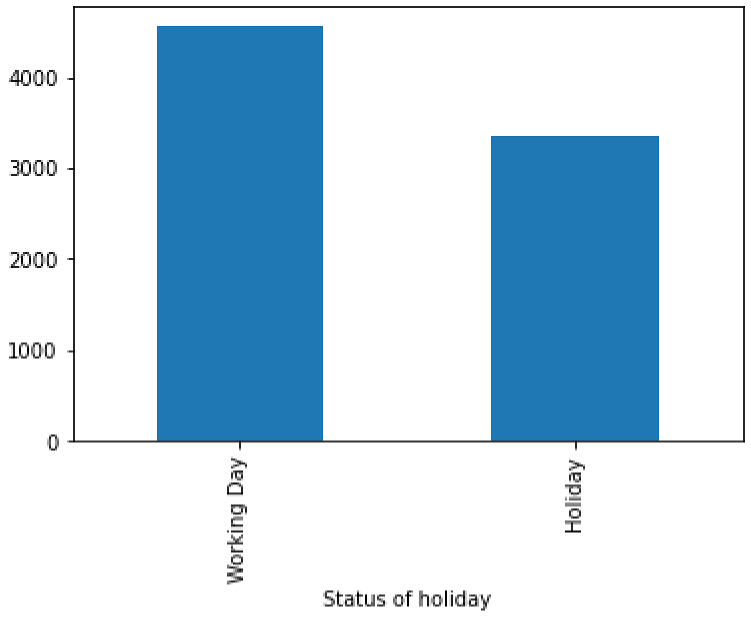
These are my observation in categorical



Seasons



Year



Holiday

Q2: Why is it important to use drop_first=True during dummy variable creation?

ANS: This thing is Dummy Variable trap, which means when we create the new Variable with One-Not Coding method,which will make high multicolinearity

If we use drop_first=True while create Dummy variable,it delete first variable which we got

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

ANS: By looking the pairplot I found that **temp** and **atemp** column has highly co-relate with target variable "cnt"

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

ANS: The Assumptions are

1)Linearity,

2).Errors are Normally distributed

3).No Multicollinearity

Q5: 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Based on my obsevation the significate variable for bike demands are:

1).Tempreature

2).Holiday

3).Weather

General Subjective Questions

1). Explain the linear regression algorithm in detail.

The Linear regression algorithm is simple and most useful algorithm in ML world.This one is fall under Supervised Algorithm category,it can be use for to find the relationship between independent and dependent variables.

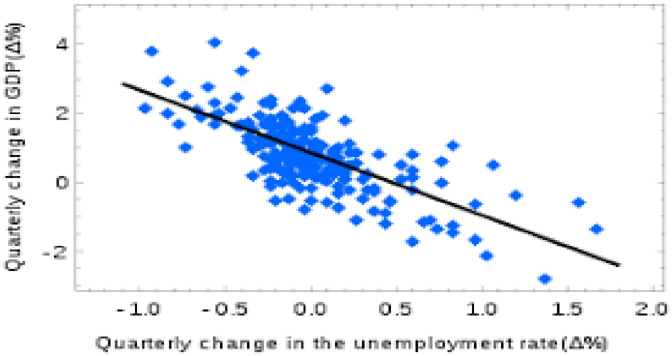
It has Two category:

a).Simple Linear Regression

b).Multiple Linear Regression

A).Single linear Regression:

In industry we not use this algoritm as much because it is work in a principle of single independent and dependent variable.

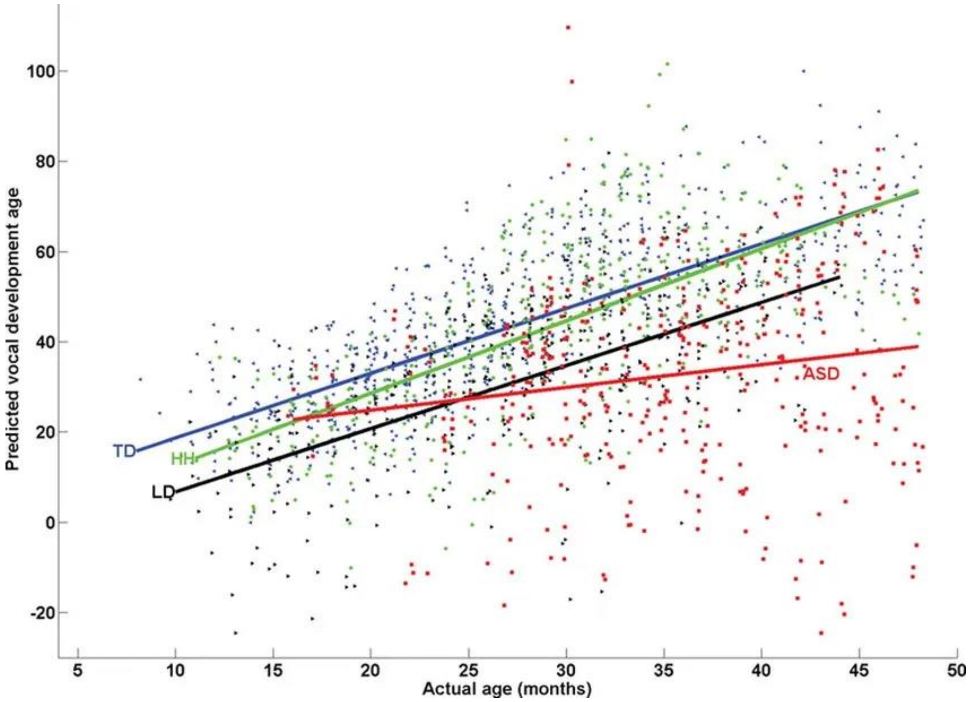


Formulae:

Y=b0+b1x

B). Multiple Linear Regression:

It is mostly used regression model in linear regression,it works on a principle of Multiple Independent variable which can be use to find single prdictor variable.



Formulae:

Y=b0+b1x1+b2x2+b3x3....

In linear regression model we use the line to fit and predict perfect value,that is the main purpose of linear model.It work on residual or error term,we can find the best fit line with calculating the Residual

We can Assume the Best ALgorithm:

Linearity: There should be linear relationship between independent variable and dependent variable.

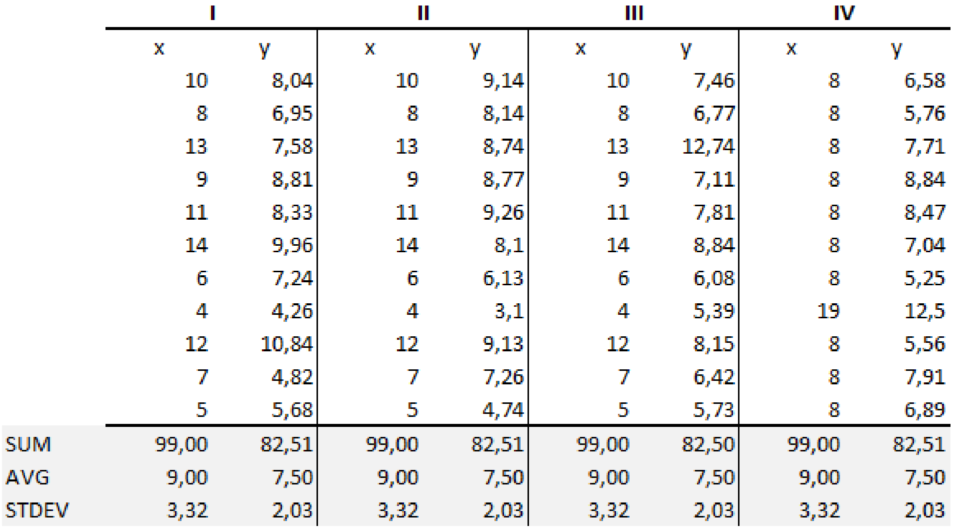
Normaly distributed: The Target variable and predictor variable should be normally distributed

Homasedecasity:The variance of the residuals are constan

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

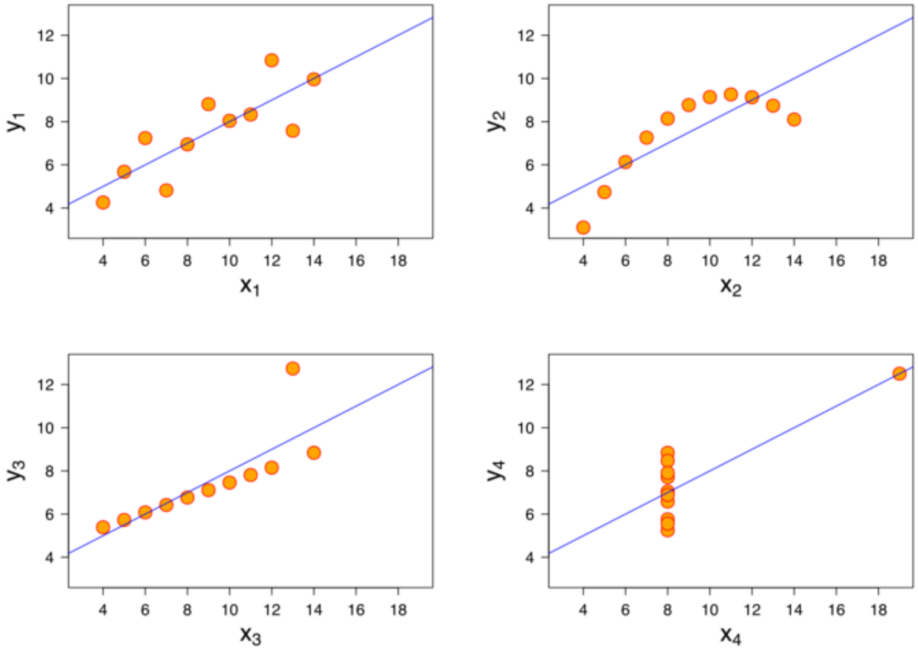
Ans: It is a method which compress the four dataset in eleven(x,y) pairs,these dataset is also give the same descriptive statistics,but when we see the same statistics visualy that things will completely change

Below image shown Anscombe's quartet:



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

It may seem like equal but thigs will change while see that visually



1).Image1: Looks everything perfect

2).Image2: Not distributed well

3).Image 3 and 4 has outliers

The Anscombe’s quartet is nothing it only explain the importance of visualization

3). What is Pearson’s R?

Ans:The Pearson’s R which also called pearson coefficient,which is used to calculate the linear relationship between the two variables,The range of the relation is lies between –1 to 1.It is common term to refer correlation

The drawbacks are:

It cannot capture the non-linear relationship between two variables

we can calculate a linear relationship between the two given variables,with the formulae of the pearson’s R:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

With the help of above formulae we can check the relationship between two variables

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

Ans: While we collect the data,the data come frrom own values which may be in in variation,which cannot be measure easily or may be in huge variation

Example: We take two universities A and B,In the University A university mesure the CGPA in 5 scale,and University B measure the score in CGPA 10,

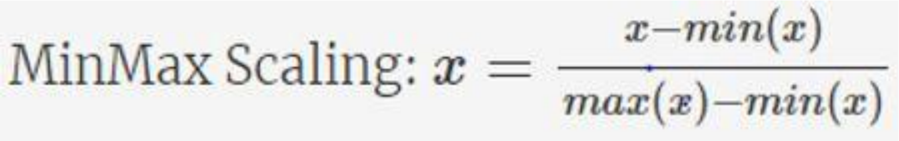
In this eample we can see the if one student score 4 in university B cannot equal to the Students with 4 CGPA in University A

If we want to handle this type of situation scaling is help us to calculate this one in relative term

Normalization:

It brings all the data into the scale value between 0 to 1

Formulae:

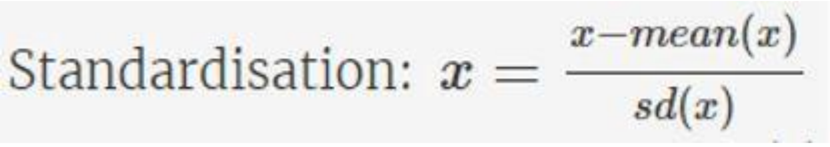


$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standarization:

Standarization replace the values with Z-Score

Formule:



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

If the values of VIF is high in nature which means ,there is high co-relation between those variables. Which one give the result as **Infinity,**

If we want to solve these type of highly co-relation factor we should drop that variable, which can help us to escape from High colinearity.

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

Ans:The Q-Q Plot which stands for quantile-quantile plot,it is used to check the data come from same common distribution.

While performing linear regression sometime the training dataset and test dataset come from out,at the time we need to check Q-Q plot,which one will explain wheather the data come from same distribution or different distribution

Importance:

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Use case:

1).It can check if the come from population with a common distribution

2). It can be cheak *have similar distributional shapes*

3).it can be use for to check the datasets are have similar distributional shapes

Explain:

Approach:

While plotting the 45 degree line on plot,if the the datasets are come from common distribtion ,the poins will fall on the reference line