Subjective Questions And Answer:

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?

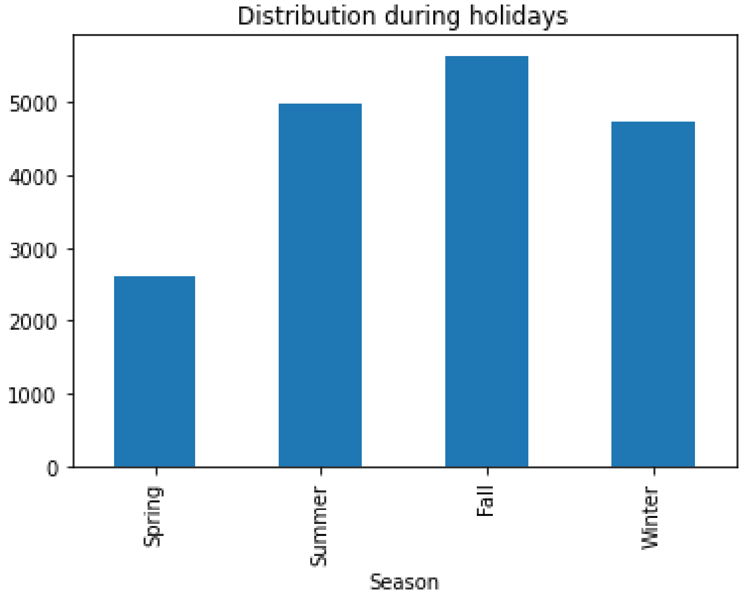
Ans: My analysis of the categorical variables from the dataset which are Based on

1.year Variable, 2019 has good sales in bikes than2018,May this happen due to increase of the popularity

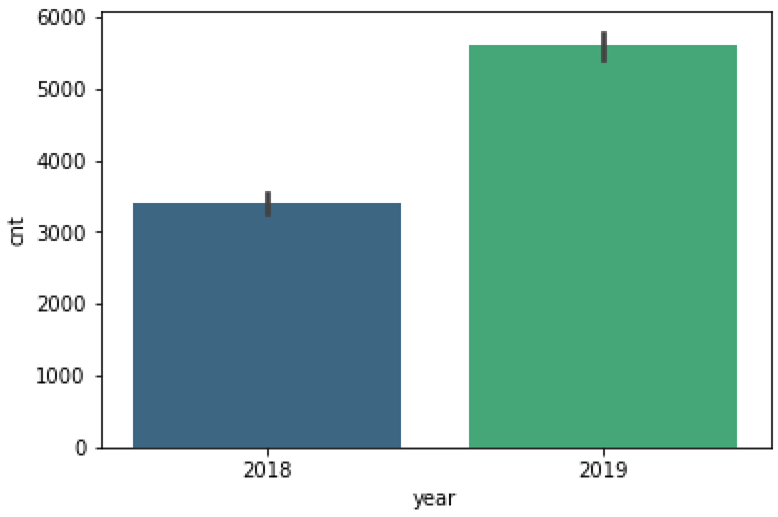
2.Next the season ,The Spring has very low sales than any other Season from model.

3.Next holiday, in my observation working day has better sales than the holiday.

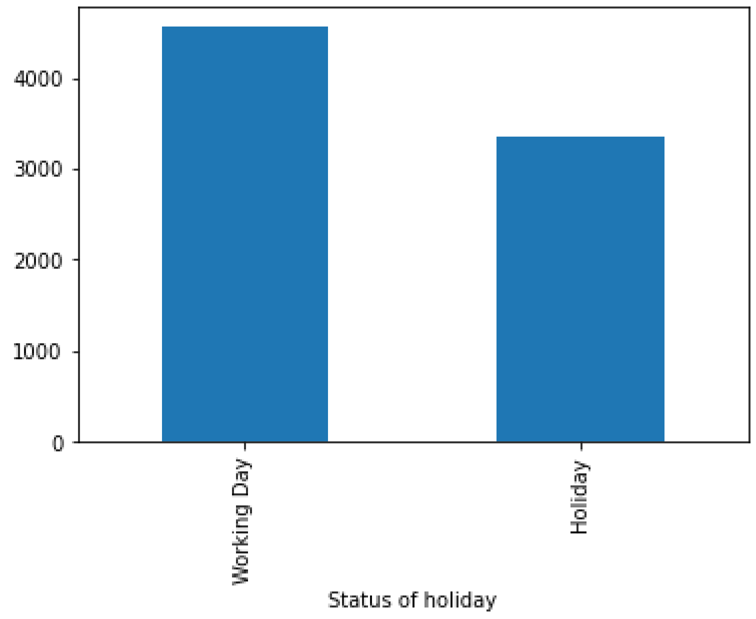
These are my observation in categorical



Seasons



Year



Holiday

2) Why is it important to use drop_ first=True during dummy variable creation?

ANS: 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 will get

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

ANS: Observing the pairplot I analysis that **temp** and **atemp** columns have highly co-relate with target variable "cnt"

4) 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

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the sharedbikes?

Ans: from my obsevation based on that the significate variable for bike

demands are:

1).Temperature

2).Holiday3).Weather

General Subjective Questions

1)Explain the linear regression algorithm in detail.

Ans: A Linear regression algorithm(Model) attempts to explain the relationship between a dependent variable(y)and independent variable(x) using by straight line y=a+bx.

Here , the independent variable(x) is known as the predictor variable and the dependent variable(y) is also known as output variable.

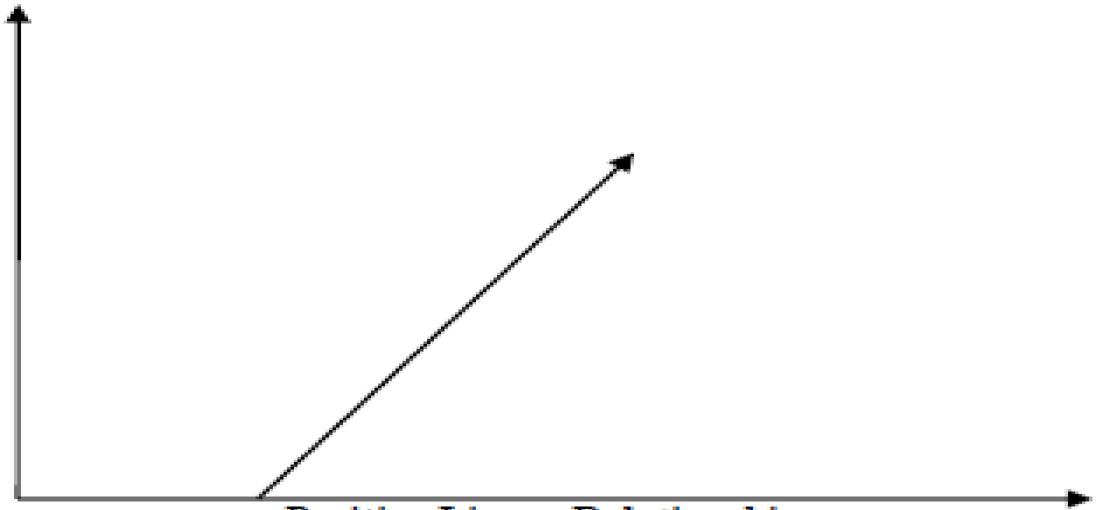
In linear regression we fit the best line ,the line which fits the given scatter-plot in the best way .

Here best fit line is obtained by minimizing a quantity called residual sum of squares(rss)

This algorithm is simple and most useful in Machine Learning world.we are using this to find the relationship between independent and dependent variables.
the linear relationship can be positive or negative in nature as explained below−

1.Positive Linear Relationship:

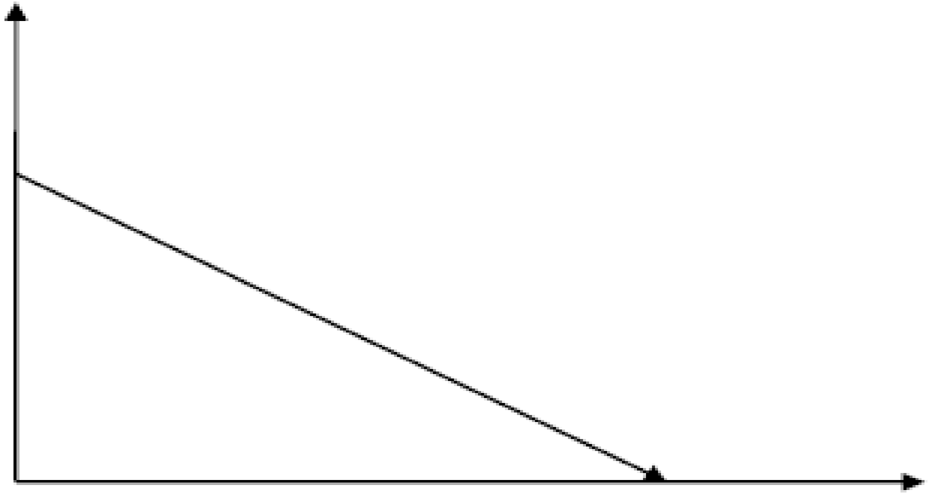
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

2.Negative Linear relationship:

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph



Negative Linear Relationship

It has Two types

1).Simple Linear Regression

2).Multiple Linear Regression

1).Single linear Regression:

industry we donot use this algoritm as much because it is working on principle of single independent variable(x) and dependent variable(y).

Ex.. **y=a+bx**

2). 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.

Ex: **Y=b0+b1x1+b2x2+b3x3....**

We will assume the Best fit line Algorithm:

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

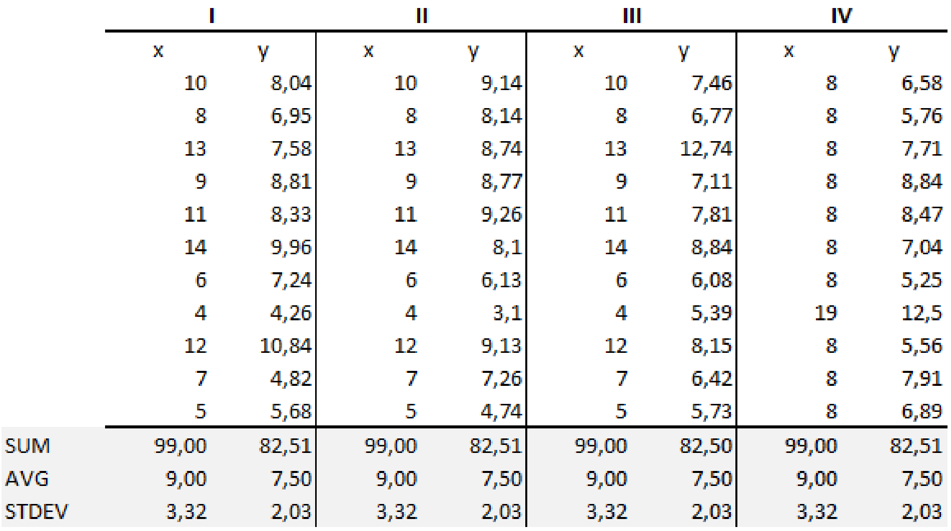
Normaly distributed: The Target variable and predictor variable should be normally distributed

Homasedent casity:The variance of the residuals are constant.

2)Explain the Anscombe's quartet in detail.

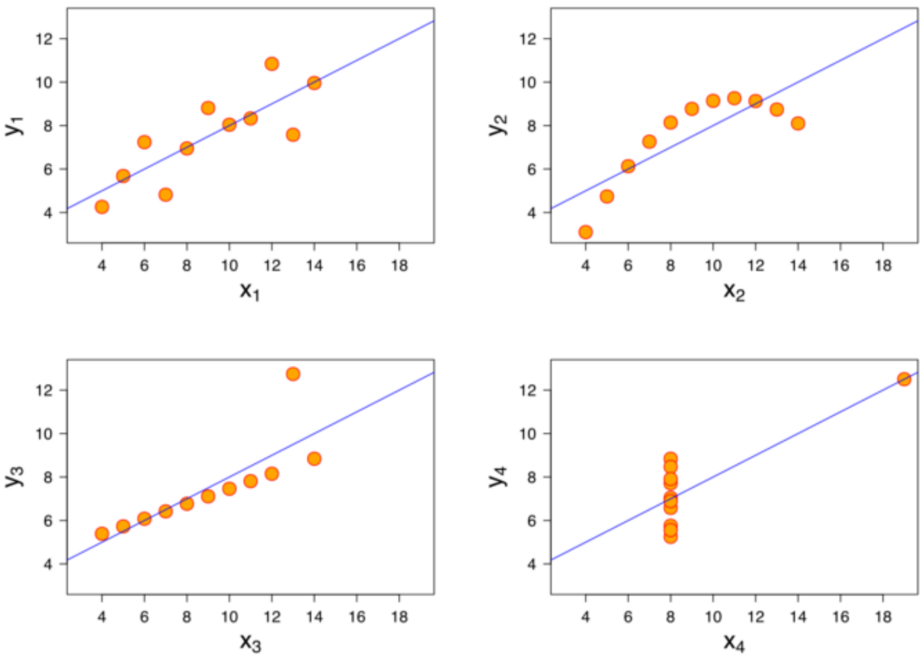
Ans: Anscombe's quartet is a method which compress the four dataset in eleven(x, y) pairs, these dataset is also give the same descriptive statistics, but when we see the same statistics visually that things will completely change

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

Generally we observe The Anscombe’s quartet is nothing it only explain the importance of visualization.

3)What is Pearson’s R?

Ans: The Pearson’s R is also called pearson coefficient,which is used to calculate the linear relationship between the two variables(x,y),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, the formulae which is required to calculate 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 the two variables

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

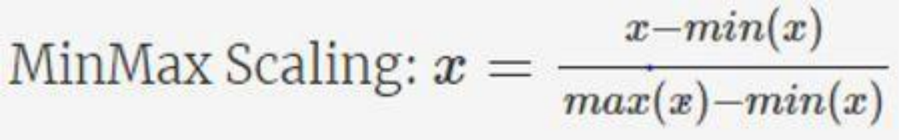
Ans: Scaling(feature scaling) is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes orvalues or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus,tackle this issue.

Normalization:

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

Formulae:

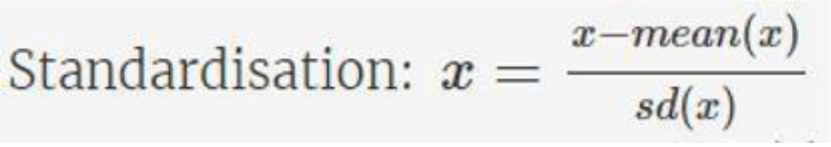


$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standarization:

standarization replace the values with Z-Score

Formule:



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

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used forscaling.
2.	It is used when features are of differentscales.	It is used when we want to ensure zeromean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans: when we have the perfect correlation, then only VIF = infinity. That means a large value of VIF indicates that there is acorrelation between the variables. If the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R^2) =1, which lead to 1/ (1-R^2) infinity. To solve this we need to drop one of the variables from the dataset which is causing thisperfect multicollinearity.

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 Highcolinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets comefrom populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence

The Q-Q Plot which stands for quantile-quantile plot,it is usedto 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 be have to check from population with a common distribution

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

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