ASSIGNMENT QUESTION

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

Ans- The dataset have 7 categorical variables they are as follows

1. Season 2. Year 3. Month 4. Holiday 5.Weekday 6.Working day 7.Weathersit

- **Season**

The Season variable show a positive correlation with the dependent variable i.e CNT ,about 0.4 is the correlation matrix of the season with cnt column .The Fall season and Summer season has good number of people taking the shared bikes.

- **Year**

The Year variable show a positive growth with the cnt column , 0.57 is the correlation matrix between year and cnt column and also year on year growth will be about 0.2288 which mean if also one bike increased per year the growth of the company will increased by 0.2288 times.

- **Month**

The month variable show a relationship with cnt column but not as strong. Only month 6,7,8,9 (June, July, August, September) have high number of people take service than rest of the month

- **Holiday**

The Holiday variable show some negative relationship with cnt column, -0.069 is the correlation matrix between holiday and cnt column. The people on holiday wants to takes rest of go some where for outing with there family due to which the no of shared bike decreased.

- **Weekday**

The Weekday has some positive relationship with cnt column. About 0.036 is the correlation matrix between weekday and cnt column .The shared bike no increased mostly on Sunday.

- **Working day**

The Working day has negative correlation matrix about -0.028 with cnt column.

- **Weather sit**

The weather has a good negative relationship with cnt , its correlation matrix is -0.3 this is due to mostly due Rainy weather where people often choose cab or prefer work from home in such days due to which bike sharing number is decreased on these days.

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

Ans-

- A dummy variable is a binary variable that takes a value of 0 or 1 and are use in regression model to represent the factor.
- A dummy variable is basically create another column equal to number of unique value in that column in 0 or 1
- For example- City column has 3 values name DELHI, MUMBAI, KOLKATA so the dummy variable create 3 column of each and give them a binary coding
  DELHI- 1 0 0 , MUMBAI – 0 1 0 , KOLKATA – 0 0 1
- So if we remove the first column we still be able to identify the first column with binary code of rest of the two column .
- So in regression model we have keep the no of variable which have more information and less in number . also if we keep the first variable our model complexity increases
- Due to this we have to use drop_first=True during dummy variable creation.

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

Ans:- The Temperature has highest correlation with cnt (target) variable.

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

Ans:- The Assumptions of Linear Regression are as follows:

- **Linearity**: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.
- **Normality**: The X and Y variables should be normally distributed.
- **Homoscedasticity**: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.
  Error Term : y act – y pred
- **No Multicollinearity**: The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.
- **No Autocorrelation**: The error terms (yact – ypred) should be independent of each other. Autocorrelation can be tested using the DURBIN WATSON TEST. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.

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

Ans:- The Top 3 variable that significantly contribute towards the demand of shared bikes.

- **Temperature** = Its coefficient is 0.5316 which mean any unit change in temp will increase bike sharing by 0.5316 value.
- **Year** = Its coefficient is 0.2288 which mean any unit change in year will increase bike sharing by 0.2288 value. (This also mean that the YoY increase in bike sharing is 0.2288)
- **Weathersit_3** =Its coefficient is -0.2384 which mean any unit change in weathersit_3 will decrease bike sharing by 0.2384 value.

GENERAL SUBJECTIVE QUESTIONS

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

Ans:-

Linear regression is one of the very basic forms of Supervised machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression the linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, A linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(inter cept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

Steps in linear regression algorithm

Step1- Importing and loading the data set and required library.

Step2- Data Cleaning and basis sanity check.

It include checking data shape ,info, null values, any duplicate value , dropping redundant column etc

Step3- Visualise the data

It include plotting the pair plot , heat map, box or bar plot between variable to check the linear relationship between variable.

Step4- Training the Model.

a). Creating the Dummy variable (wherever required)

b). Rescaling the data.

c). Splitting the data into test and train data set.

d). Creating the X and y i.e. deciding the independent variable and dependent variable and creating X and y for train data set.

e). Crete the OLS model and training the model by fitting X and y value in model

f). Check the parameter and check the summary of the created model

g). Checking the multicollinearity between independent variable with the help of VIF (Variance influencing factor)

Step5- Residual analysis and validating the assumption

a). First calculate the y prediction value from the model which we have created

b) Second calculate the residual which is difference between y_true value and y predicted value

c) Plotting the histogram for residual and checking that the residual should form a normal distribution curve.

d) Check for No Multicollinearity wit help of VIF value if VIF value is more than 5 then high multicollinearity is present.

Step6:- Prediction from final model

a) Now predicting the y value from the final model on base on X test value.
b) Creating a scatter plot of y test value and y predicted value and check for spread of data to evaluate the model .
The data should be form a linear relationship between y test data and y predicted data

Step 7:- Calculating the R-Square and adjusted R-square value

a). Calculating the R-Square and adjusted R-square value .

b). Comparing the R-Square and adjusted R-square value with the final model value and the value should be within 5% of value

This shows that the model is good

Step8:- Building the Best fit line equation form the final model.

2. Explain the Anscombe's quartet in detail.

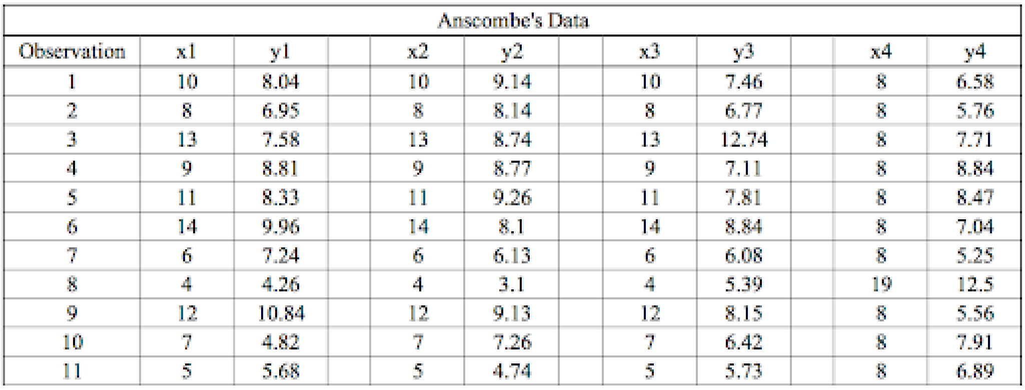
Ans. Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.

The data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

The linear regression Model can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

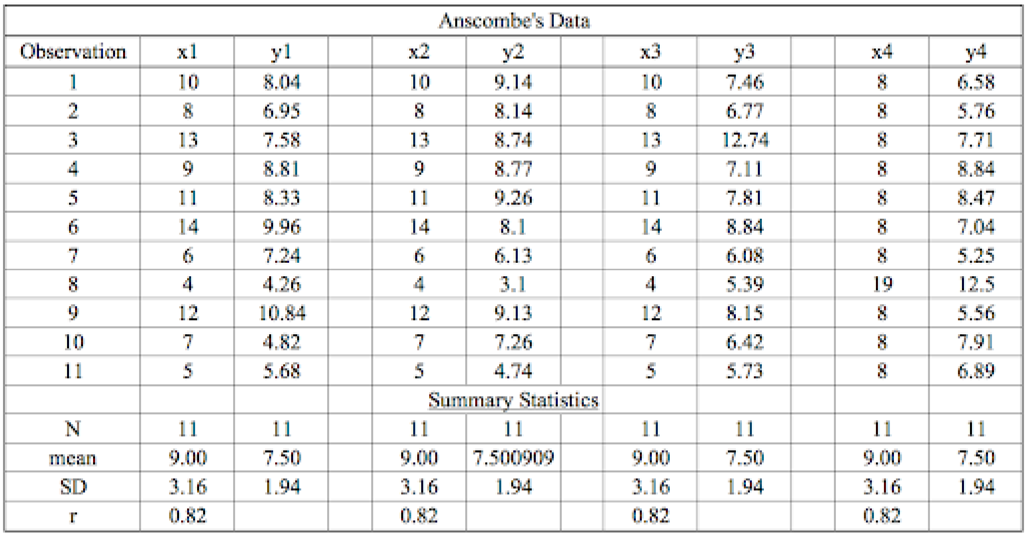
We can define these four plots as follows:



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

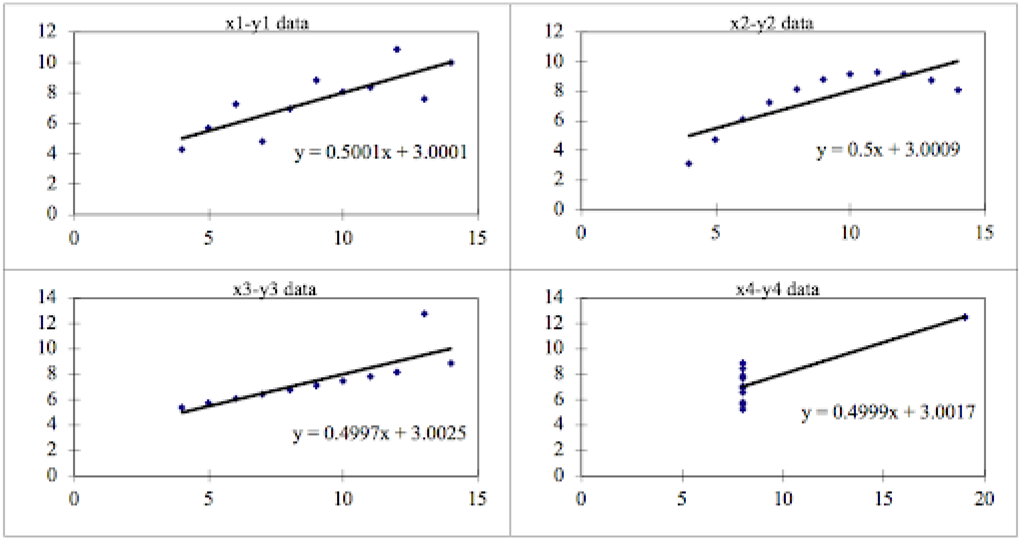
The statistical information for these four data sets are approximately similar. We can compute them as follows:



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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm.



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet helps to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

Ans:-

- The Pearson's R is also called as Pearson correlation coefficient.
- It is used to calculate the statistical calculation of the strength between two variable i.e. how two variable are dependent on one another.
- The value of Person r can varies from +1 to -1.
- If the value is zero than there is no correlation between two variables.
- If the value is greater than zero, there is a positive correlation between variable and if the value is less than zero than there is negative correlation between variable.

For Example :-

Lets take example of two common variable demand and supply.

A case of two variables affecting one another is demand and supply in an economy when the price of the product and the quantity demanded and supplied is known.

Pearson R shows that demand and supply have a positive correlation. As more consumers demand products, the supply amount will also increases as well.

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

Ans:-

Scaling :

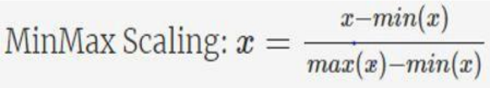
Its is one of the pre-processing step of linear regression algorithm .It is applied to independent variable to normalize the data within specific range. It also help in speeding up the calculation in algorithm.

The Scaling is performed due to in mostly data set the containing feature variable are in highly variance magnitude , unit, range. If scaling is not performed then the algorithm will only take magnitude in account and not units which will lead to incorrect modelling. So scaling is done to bring all the variables to the same level of magnitude.

Difference between Normalized and standardized scaling

a). Normalized scaling**.**

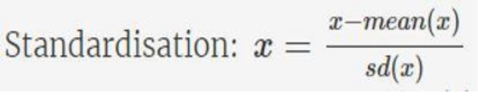
- It brings all of the data in the range of 0 and 1.
- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.



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

b). Standardization Scaling:

- Standardization replaces the values by their Z scores.
- It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



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

- **sklearn.preprocessing.scale** helps to implement standardization in python.

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

Ans:-

VIF(Variance Inflation Factor**)**

- The VIF is used to detect the multicollinearity between independent variable in multiple regression model.
- Generally, the VIF value more than 5 indicate a high multicollinearity.
- If VIF= infinite it basically indicate that there a perfect correlation between two independent variable.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables
- In the case of perfect correlation, the R2 =1, which lead to 1/(1-R2) infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?
Ans:-
Quantile-Quantile (Q-Q) plot, is a graphical tool for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution.

Use

- It helps in a scenario of linear regression when the training and test data set received separately and then we have to confirm using the Q-Q plot that both the data sets are from populations with the same distributions.
- It use to find out whether the two data come from common location and scale, or similar distributional shapes, or **similar ta**il behaviour.

Importance

Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.