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:

There are 6 categorical variables present in the data set:

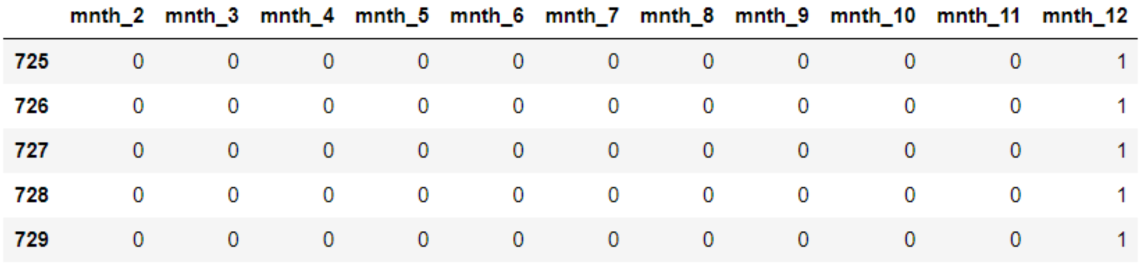
- **Season:** we can see from the boxplot that season 3(fall) has more demand in bike hirings.
- **Month:** we can see from the boxplot that mnth_9(September) has more demand in bike hirings.
- **Holiday:** Bike hirings count is less in holidays.
- **Weather Situation:** If weather situation is clear then bike hirings are more.
- **Weekday:** There is not much difference in weekday but we can see on Sunday there is slight demand.
- **Working day:** There is not much difference in working day in demand of bike hires.

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

Ans:

We have to drop first column in case of creating dummies in order to reduce multi-collinearity. If we have 'k' levels in a variable then we need to create k-1 dummy variables in order to reduce the redundancy.

For example, we can consider Month which has 12 months but we can show using 11 dummies.



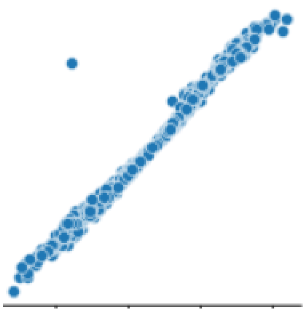
	mnth_2	mnth_3	mnth_4	mnth_5	mnth_6	mnth_7	mnth_8	mnth_9	mnth_10	mnth_11	mnth_12
725	0	0	0	0	0	0	0	0	0	0	1
726	0	0	0	0	0	0	0	0	0	0	1
727	0	0	0	0	0	0	0	0	0	0	1
728	0	0	0	0	0	0	0	0	0	0	1
729	0	0	0	0	0	0	0	0	0	0	1

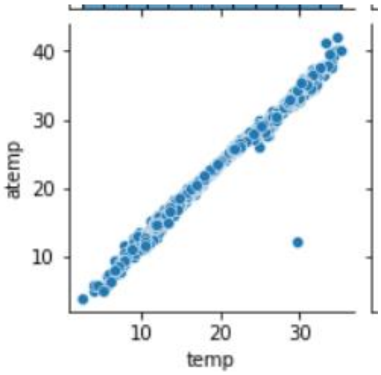
The above data shows that the month is December.

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

Ans:

We can see that atemp and temp has high correlation among numerical variables.





temp: 0.99(corr)

atemp: 0.99(corr)

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

Ans:

We can validate assumptions by below factors:

- All the residual (error) terms are linearly distributed
- All the coefficient values are greater than 0 and the p values of them are significant (p<0.05)
- There is a linear relation between y(cnt) and x (all predictor variables)
- The training set and testing set r2 scores difference is below 5.
- The error terms have constant variance.
- F-statistic value is greater than 1.

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

Ans:

Top 3 features based on model are:

- **Temperature:**

  A unit increase in temperature then bike hire number increases by **0.5687** units.

- **Weather Situation (`Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds`):**

  A unit increase in this variable then bike hire number decreases by **0.2541** units.

- **Year:**

  A unit increase in year then bike hire number increases by **0.2337** units.

General Subjective Questions:

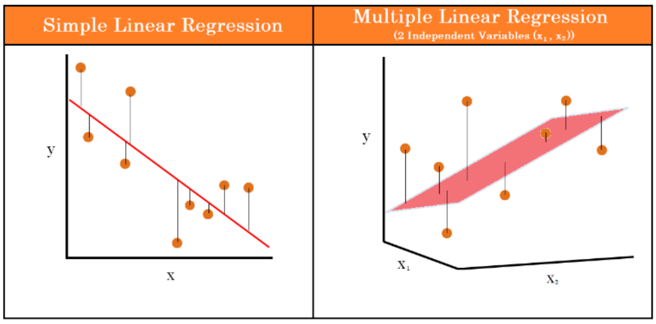
1. Explain the linear regression algorithm in detail?

Ans:

Linear regression algorithm is a supervised learning model that shows the linear relationship between two variables which are dependent and predictor variables.

There are 2 types of linear regression algorithms

- Simple linear regression.
- Multiple linear regression.


Simple Linear Regression
y
x
Multiple Linear Regression
(2 Independent Variables (x_1, x_2))
y
x_1
x_2

In the above diagram x is predictor variable y is dependent variable

Simple Linear Regression:

In simple linear regression there will be one independent variable and one dependent variable. We need to find the one dependent variable which has linear relationship. The relationship will be a straight line in simple linear regression.

(Y=β0+β1X) is the equation of simple linear regression.

Where Y - Dependent variable

β0 - Intercept

β1 – Coefficient

X - Dependent Variable

Multiple Linear Regression:

In Multiple linear regression there will be more than one dependent variable for one independent variable. The relationship will be a hyper plane.

(Y=β0+β1X1+β2X2……..+βnXn) is the equation of multiple linear regression.

Linear regression is used to find the best fit linear line between two variables where residuals(errors) are minimized.

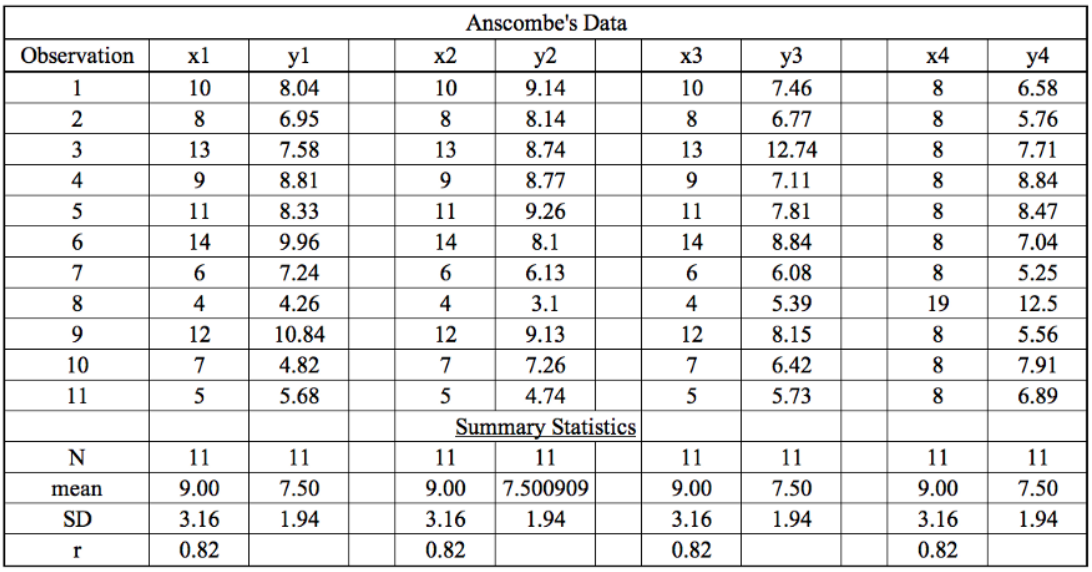
2. Explain the Anscombe's quartet in detail?

Ans:

It is defined as group of 4 datasets which are identical in simple descriptive statistics but appear differently when we observe the scatter plots.

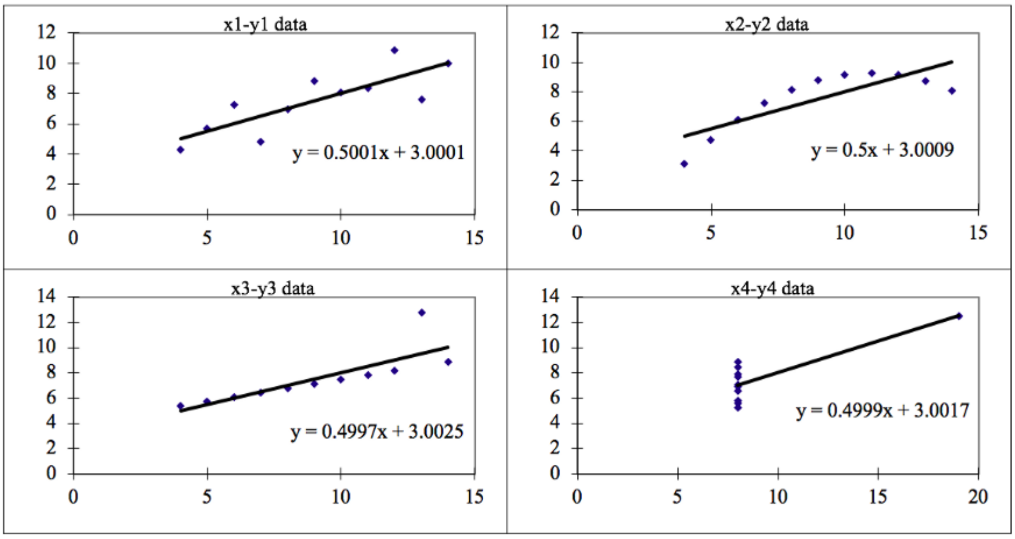
It was constructed by Franscis Anscombe in 1973 to illustrate the importance of plotting graphs before analysing the model.

To check this, take 4 data sets which are having similar statistics and also summary statistics as below:



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	

We can see the data is similar in all the data sets including mean and standard deviation.



But when we plot the above datasets, we can see clearly that all are different.

Hence all the data needs to be visualized before coming to an conclusion.

3. What is Pearson’s R?

Ans:

The Pearson’s correlation is the most used correlation method for numerical variables. It assigns the value from -1 to 1 where

0 – no correlation between A and B

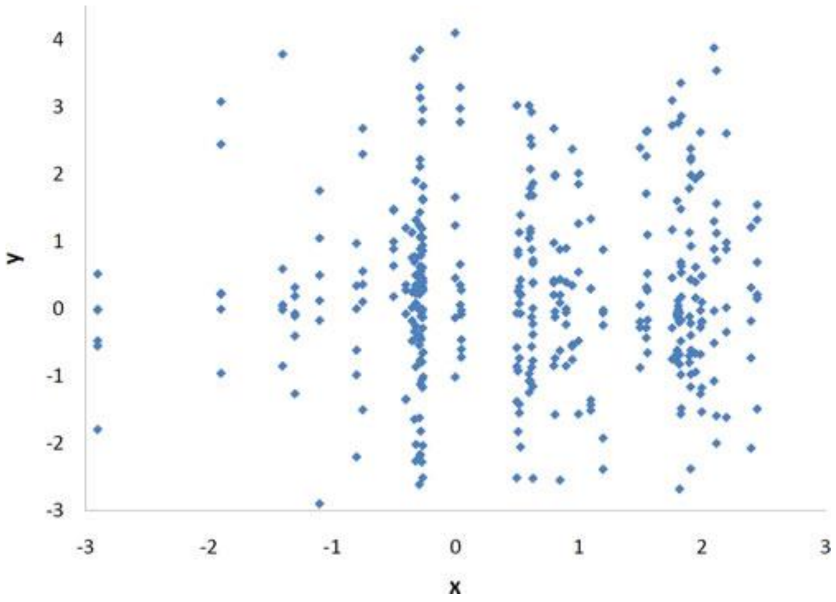


Fig: No correlation between x and y so r value = 0

1 – Strong Positive correlation between A and B such that if A goes up then B also goes up.

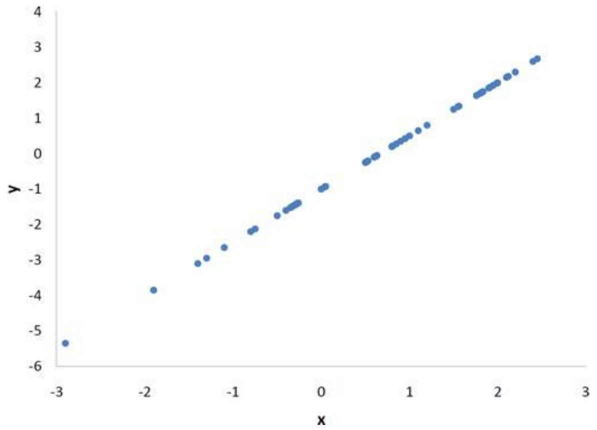


Fig: Positive correlation between x and y with r = 1.

-1 – Strong Negative correlation between A and B such that if A goes up then B goes down.

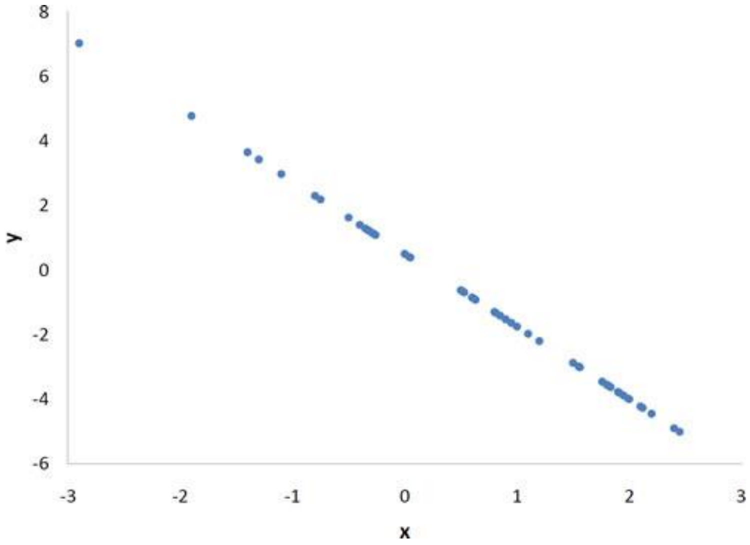


Fig: Negative correlation between x and y with r = -1.

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

Ans:

Scaling is a method which brings all features under same proportions to avoid the indifferences between them. Scaling is an important step in data processing before creating a model.

The most important scaling features are:

- Normalization and
- Standardization.

Differences between normalized and standardized scaling:

Normalized Scaling **(Min Max Scaling)**	**Standardized Scaling**
Scaled values are between [0,1]	Scaled values are not bounded by a range
We will use min and max values	We will use mean and standard deviation for scaling.
It Is called as scaling normalization.	It is called as z-score normalization.
Formula: **x = (x-min(x))/Max(x)-Min(x)**	**Formula**: **x = (x-mean(x))/sd(x)**

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

Ans:

The formula for VIF (Variance Inflation Factor) is

VIF = 1/1- R^2

Where R^2 is the correlation value.

If the correlation between A and B is 1 i.e., a strong positive correlation then the R^2 value will be 1 so the VIF value will be infinity (since 1/0 = infinity). So in case of multiple variables have VIF as infinity we will drop any one of them based on the probability or the business value such that other features correlation will decrease.

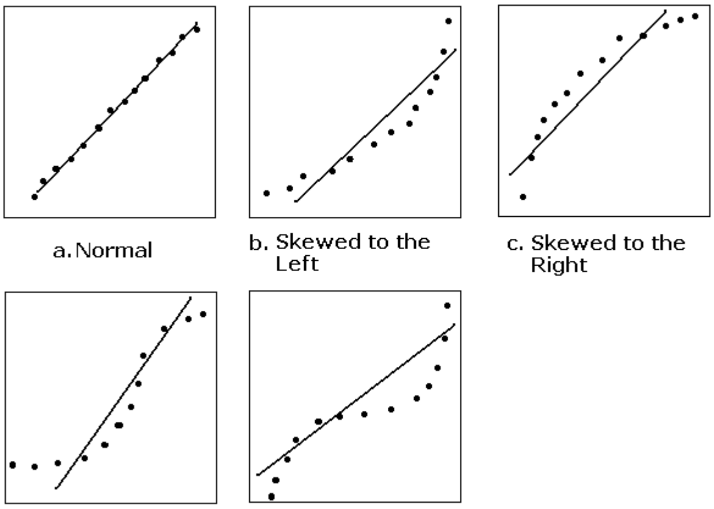
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Ans:

Q-Q (Quantile-Quantile) plot is used to find the distribution between the variables just by looking at the plot whether it is a normal distribution or not.

It also helps to determine the residuals whether they are in normal distribution or not.

It helps in determining that the two data sets belong to a common distribution.



a. Normal

b. Skewed to the Left

c. Skewed to the Right

d. Thick Tails

e. Thin Tails

We can see fig.(a) a normal Q-Q plot in which all the plotted points lie within the straight line.

Fig (b) and fig (c) are skewed Q-Q plots in which if the plotted points lie below the straight line called as left skewed plot and if it lies above the straight line called right skewed plot.

Fig (d) and fig (e) are tailed Q-Q plots in which if the plotted points centre follows the straight line but with less deviation from straight line then it is called thick tail and if the deviation is more from straight line called as thin tail Q-Q plot.