**1 to 2 – descriptive statistics**

**-mean,median,standard deviation,mean absolute deviation – this is achieved using data.describe()**

**graphs and plots used to visualize the univariant and bivariant data .**

**Univariant is just visualizing only one variable – this achieved using sns,matplotlib**

**Bivariant is visualizing 2 variables.**

**2to 2.45 – distribution-why distribution is important(distribution is used in inferential statistics to find the population value from sample value)**

**2.45-corelation which is used for bivariant analysis**

**Covariance is used in dimensionality reduction and principle component analysis**

**If co variance is positive then variables move together**

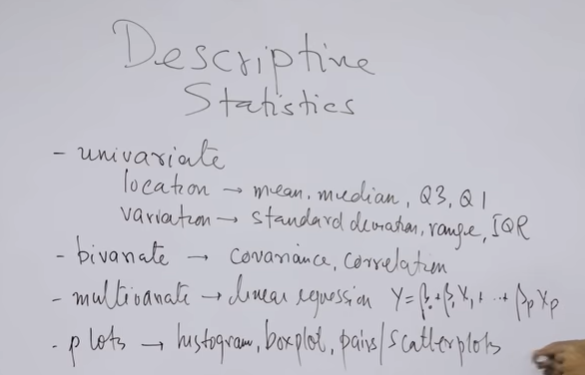
**If it is negative then moves opposite direction**

**If 0 then independent**

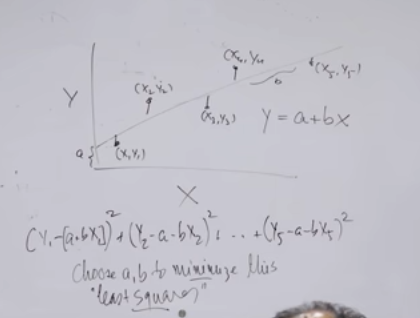
**2.58 – correlation coefficient –**units of measurement of co relation (linear relation) between -1 and 1

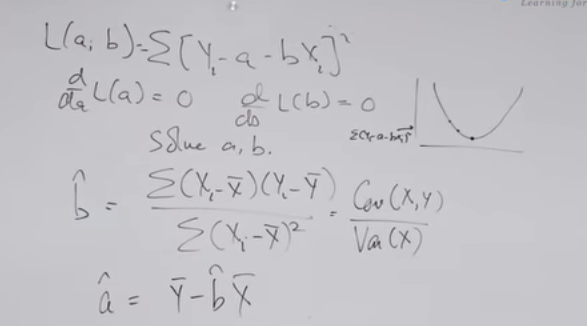
**This we can achieve using mydata.corr() and use heatmap**

**3.30 – over view of what learnt in descriptive statistics**



**3.45 – linear regression**



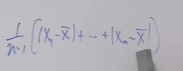


4.15 – linear regression –how to calculate intercept and co eff manually and compare with python prediction

Descriptive statistics:

Mean absolute deviation:

1. It is used to calculate the deviation between the data point and mean value.

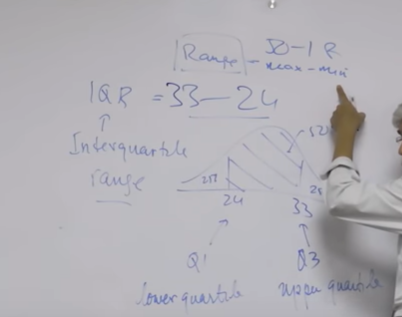


|x1-x-| > || represents absolute value of difference between data and mean value(x- (x bar))

1. Since it is not squaring the difference between the data and mean we can find the outliers easily, which is not possible in standard deviation. So it is mostly used in machine learning algorithms

Inter quartile range: Q3 –Q1 (75% minus 25%)

Range is max- min



With the 5point summary (mean, 25%, 50%, 75%.max) we can find the location which is median and two measures of dispersion (range and Inter quartile range).

Those two were twisted to give summary number, which is mean, and range number.

So the conclusion is half of my customer is between the age of 24 and 33 with the difference of 9years range which I am interested in which is 25%to 75% (interquartile range).

Therefore, when building the model we have to make sure that model performs for this 33 old and also for 24-year old.50 year old may or may not okay with the model, which is the maximum value since we are interested only with, inter quartile range.

Visualization:

Histogram(frequency or count distribution)

Scatterplot is used to find the numerical variable correlation

These plots are used to visualize the distribution of a data set (numerical values) :

distplot- single value distribution - sns.distplot(tips['total\_bill'],kde=False,bins=30)

scatterplot - 2 numerical variables distribution - sns.scatterplot(x='total\_bill',y='tip',data=tips,hue='sex')

pairplot- entire dataframe - sns.pairplot(tips,hue='sex',palette='coolwarm')

Categorical Data Plots:

Countplot – used to take count of categorical variable- **sns.countplot(x='sex',data=tips)**

boxplots and violinplots - the distribution of categorical data

**sns.boxplot(x="day", y="total\_bill", data=tips,palette='rainbow')**

**sns.violinplot(x="day", y="total\_bill", data=tips,palette='rainbow')**

barplot is a general plot that allows you to aggregate the categorical data based off some function, by default the mean:

**sns.barplot(x='sex',y='total\_bill',data=tips)**

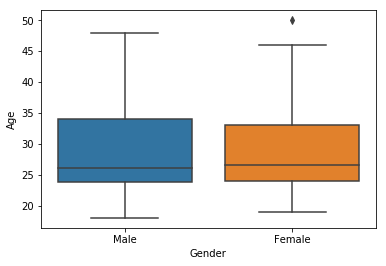
Questions to understand the data:

1. How many products purchased based on the product model?
2. How the user’s age is distributed?
3. Does age affect the product purchase?
4. Does gender affect the product purchase? .What is the count of the product based on gender?
5. Does marital status affect the product purchase? .What is the count of the product based on marital status?
6. How income part of product purchase?
7. What is the min and maximum age of the male and female on the data?
8. In which age the maximum number of usage is marked?
9. Does marital status affects the usage?
10. Who has highest income on the gender?
11. Which gender uses the product more?
12. Which gender has the maximum miles?

Numerical plot as y-axis:

Histogram used for frequency or count.

Box plot used to identify the outliers which is plotted between categorical column (gender) in the x axis and numerical (age ) in y axis means how gender male and female got distributed for the age.



In the whisker plot, lower and upper line is min and max.

lower line is marked at min value from describing the data.

upper line is marked with the length l=1.5 \* IQR (inter quartile range) from 75% line of the box

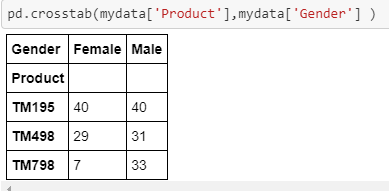
Box is plotted between IQR 25% Q1 to 75% Q3.Middle line in the box is median (50%).

The records which are higher than upper line of whisker plot is called as outliers.

Categorical variables:

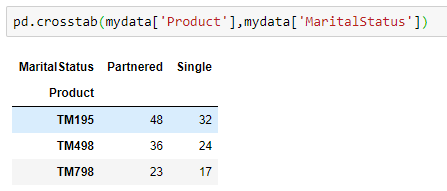
Does male and female have the same preference while getting the treadmill? .

The answer is below.



Is there any preference while purchasing the tread mill based in the marital status?

The answer is below.

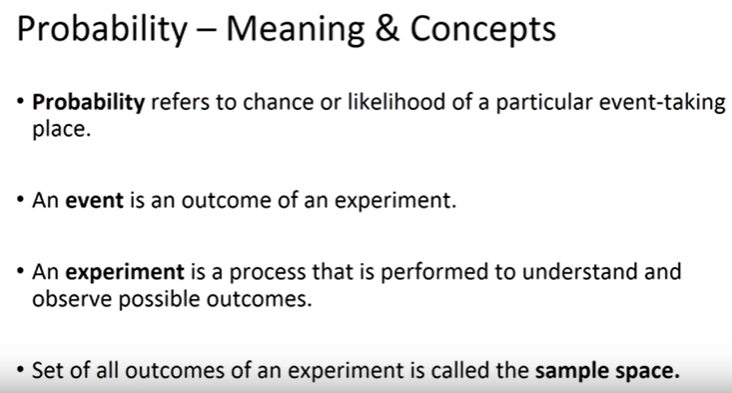


Distribution:

What is the common for observed set of data and future data . That commonality we can say as distribution.

4.26 :

Probablity :



<https://towardsdatascience.com/what-is-bayes-rule-bb6598d8a2fd>