****

**COURSE CODE: CSE303(1)**

LAB GROUP ASSIGNMENT - #2

***Module Development on Statistical Functions***

**Group: 08**

# In this assignment we had to develop a module that contains the definition of some given statistical functions and then compare the output of our own module with the output of Pandas statistical function.

**SPRING 2021**

# Problem statement:

# In this assignment we had to develop a module that contains the definition of some given statistical functions and then compare the output of our own module with the output of Pandas statistical function.

# Abstract:

One of the initial courses for Intelligence & data science students is Statistics for Data Science. It can provide students with an introduction to analyze & work with different data frame, plots & measure the statistical measures with real life experiences. So, in this group assignment we made those statistical functions as well as made some extra functions.

To make this assignment we used our previous assignment’s dataset which was “Australian Credit Approval”.

# Reading the file:

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Dataframe Info:

Graphical user interface, text

Description automatically generated

Figure: Output of df.info()

# **Statistical Functions**

# Count ()

Text

Description automatically generated

1. Describe()

# Text Description automatically generated

1. Min, max

Graphical user interface, text

Description automatically generated

1. Argmin, argmax
2. Idxmin (), Idxmax () or index min & index max function:

This code is for compute the index label where minimum value & maximum value obtained, respectively

Text

Description automatically generated

Figure: Screenshot for executing idxmin() &idxmax()

Text

Description automatically generated

Figure: Screenshot of idxmin() &idxmax() function output

1. Quantile:

A quantile is a word that describes a part of a data set which specifies how many values in a distribution are above or below a specified limit. Here I computed the first quartile (1/4), the second quintile (1/2), and the third quantile (3/4) using my own module function &Pandas function.

Text

Description automatically generated

Figure: Screenshot for executing the ‘Quantile’ cell

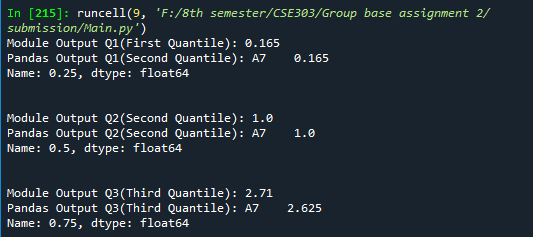


Figure: Screenshot of the output for quantile

1. Sum

Text

Description automatically generated

1. Mean()

Text

Description automatically generated

1. Median()
2. Mean Absolute Deviation(MAD) from the mean value:

The average distance between each data value and the mean is the mean absolute deviation of a data collection. This function is a term for how a data set's variance is described.

Text

Description automatically generated

Figure: Screenshot for executing Mean Absolute Deviation cell

Text

Description automatically generated

Figure: Screenshot of the output for Mean Absolute Deviation

1. Product ()

Text

Description automatically generated

1. Var():

Variance is a calculation of how much each data point differs from the mean (the average of all data points). So, the function is for determine the variance of attribute from dataset.

Text

Description automatically generated

Figure: Screenshot for executing the ‘Var’ cell

Text

Description automatically generated

Figure: Screenshot of the output for Variance

1. SD(): Standard deviation- The square root of the variance is used to calculate standard deviation, which measures how far a collection of data is from the mean.

Text

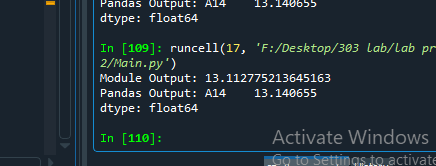
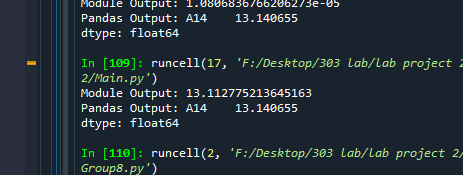
Description automatically generated

Figure: Screenshot for executing the Standard deviationcell

Text

Description automatically generated

Figure: Screenshot of the output for Standard deviation

1. Skewness()
2. 
3. Kurtosis()
4. 
5. Cumulative\_sum()

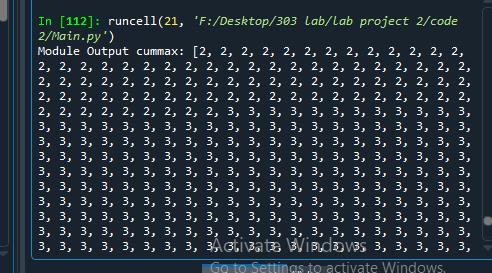
A picture containing background pattern

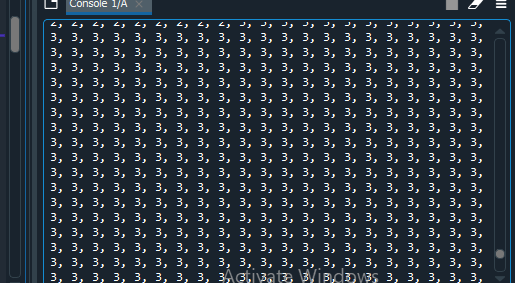
Description automatically generated

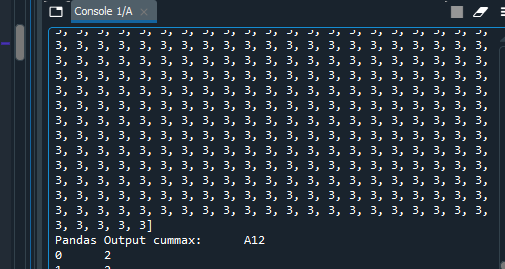
Text

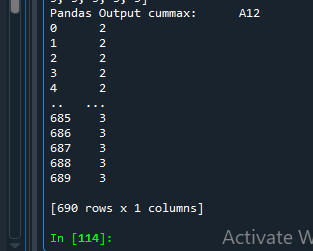
Description automatically generated with medium confidence

1. Cumin() &Cumax()









1. Cumprod()

Graphical user interface, text

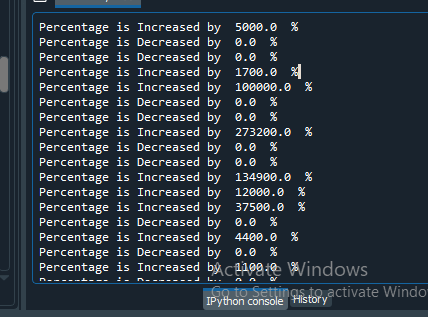
Description automatically generated

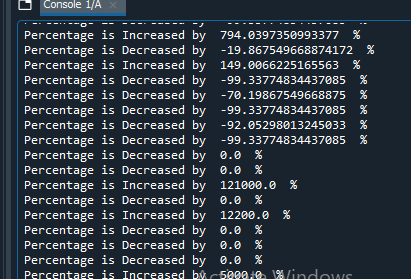
1. Diff()

Graphical user interface, text

Description automatically generated

1. Pct\_change()





**Extra Implementations:**

1. IQR:

This is the interquartile range function Q3(third quartile)- Q1(first quartile). This function is used for various statistical measure like box plot.

Graphical user interface

Description automatically generated with medium confidence

Figure: Screenshot for executing the IQR function cell

Text

Description automatically generated

Figure: Screenshot for output of the IQR function

1. Median Absolute deviation ():

In this function I made median absolute deviation to defines how spread out the set of data is & compare it with the pandas mean absolute deviation.

Text

Description automatically generated

Figure: Screenshot for executing the median absolute deviation function cell

Text

Description automatically generated

Figure: Screenshot ofmedian absolute deviation function cell output

1. Absolute ():Absolute means the magnitude of a real number.Here, I created the absolute function & compared the value with Pandas abs () function which is giving the same result. I made this module to use these absolute values in other functions like mean absolute deviation & median absolute deviation.

Graphical user interface, text

Description automatically generated

Figure: Screenshot for executing the absolute function cell

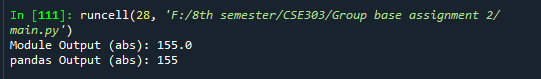


Figure: Screenshot for output of the absolute function

1. Sorting: I made this function to sort the values in ascending order. That’s how by using this module sort function we created few more function easily by calling sort().

Text

Description automatically generated

Figure: Screenshot for executing the cell

Output:

A picture containing calendar

Description automatically generated

Background pattern

Description automatically generated

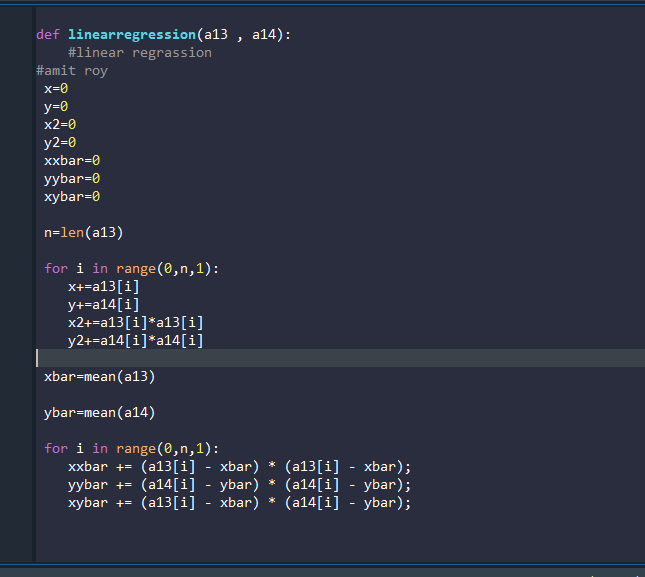
Text

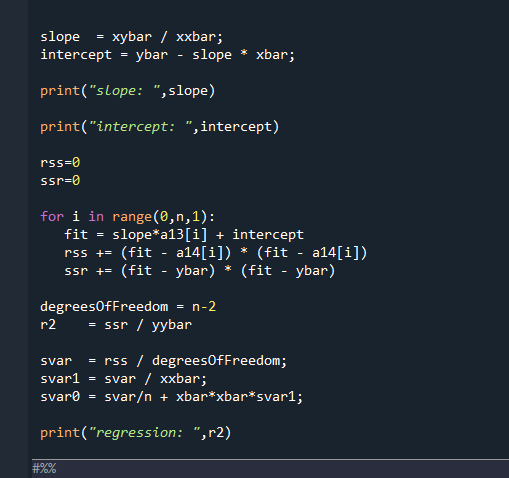
Description automatically generated with medium confidence

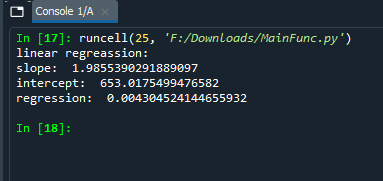
Figure: Screenshot for output of the sort() function

**Amit roy’s part**

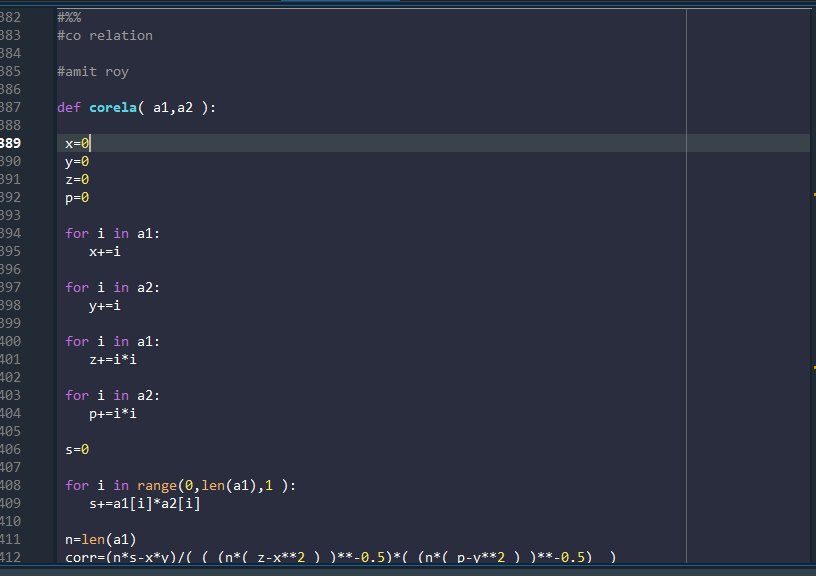
Linear regression:

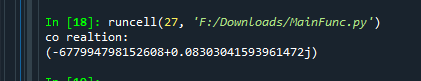




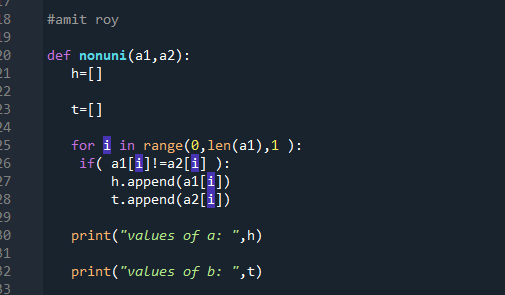


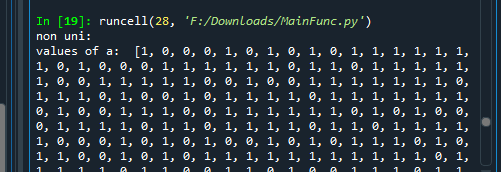
Co relation

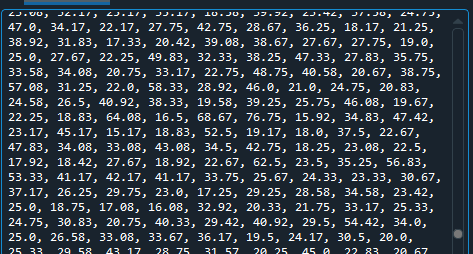




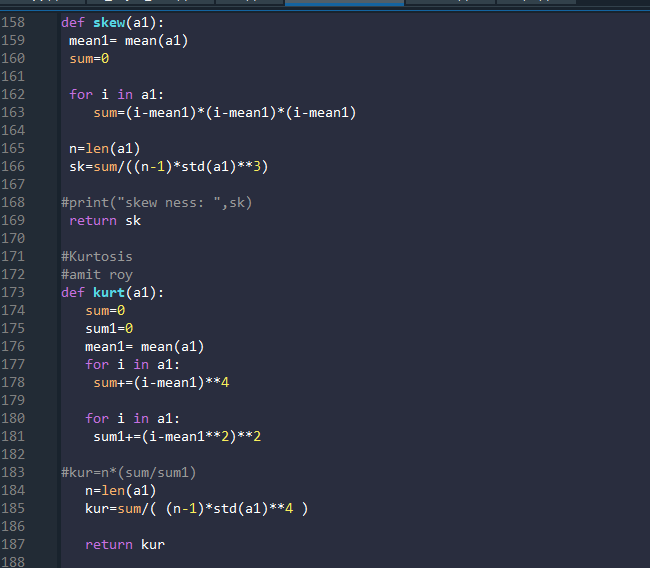
Non uni:

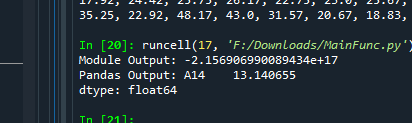


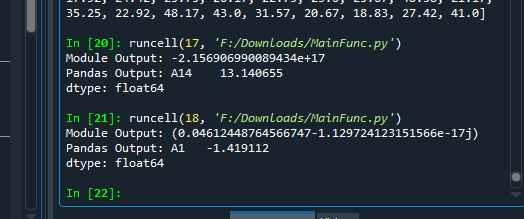




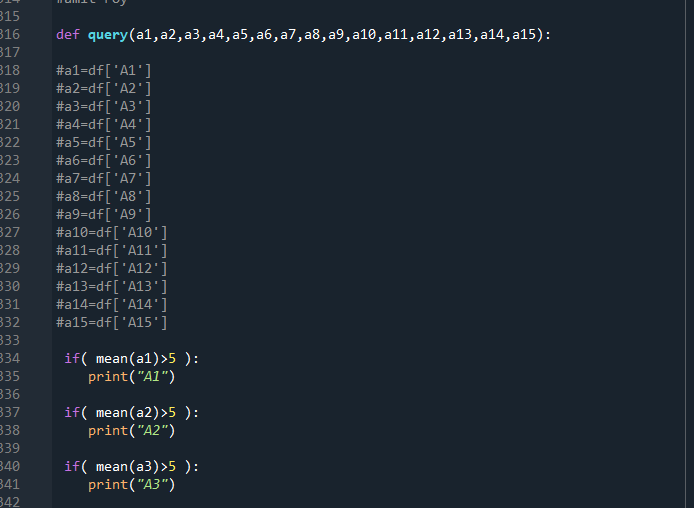
Skewness and kurtosis:

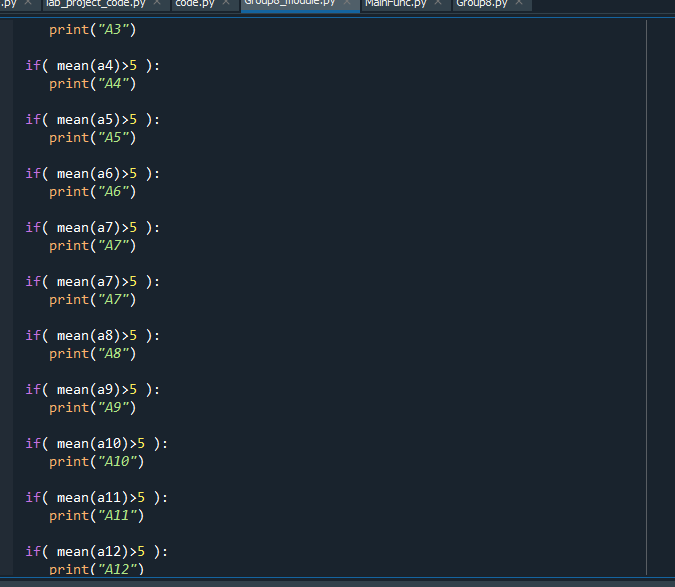


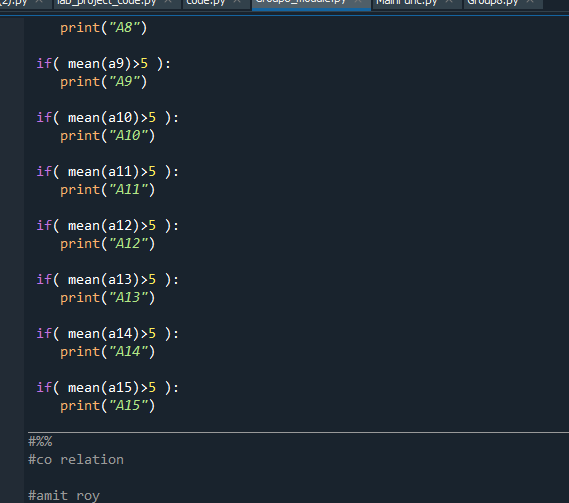


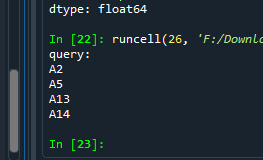


Query:

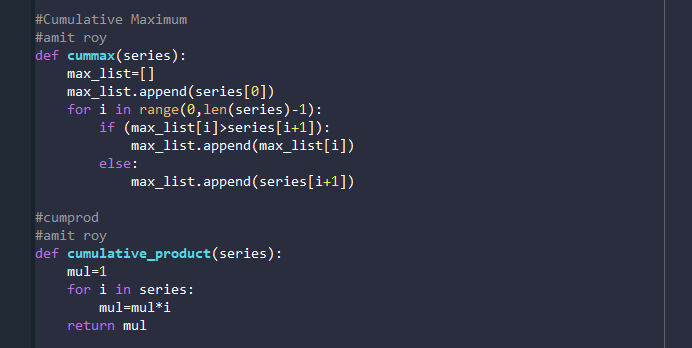




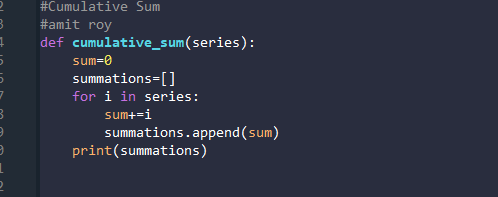




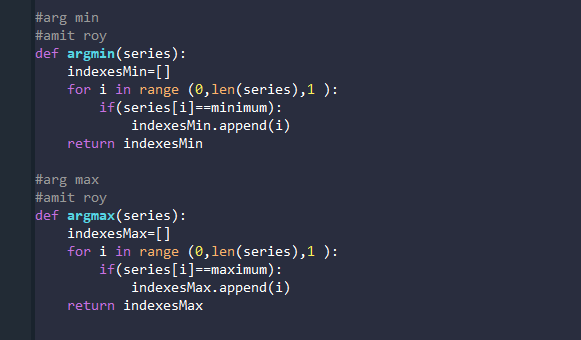
Cumax and cumin and cuprod sumuletive sum



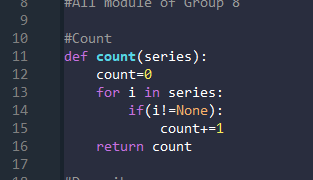
# 



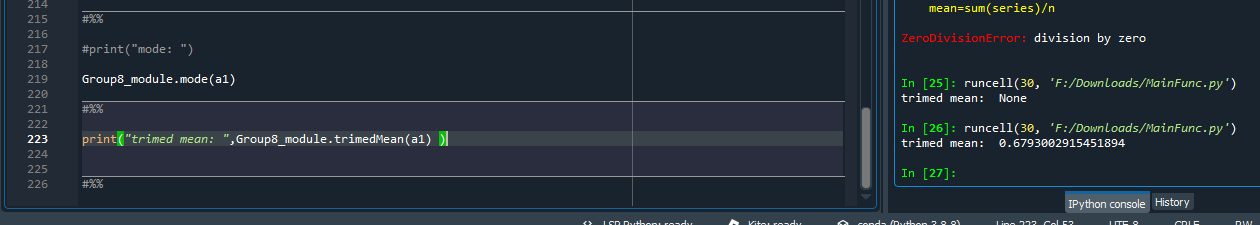
Argmax and argmin:

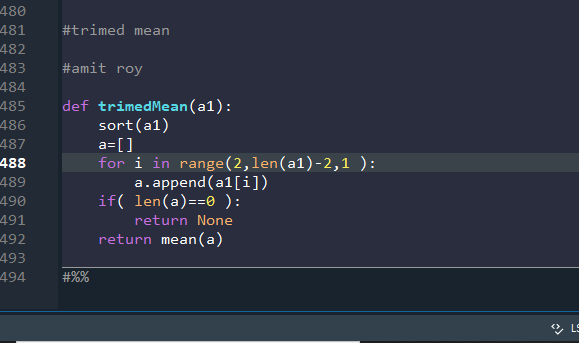


Counting non null:



Trimed mean





# **Discussion**

In this assignment, we had been asked to create our own statistical module & compare those with Pandas module. By doing this assignment we had face some difficulties as we can’t use any built-in panda’s function. So, we had to create own module like sort, abs, sum, mean & then we used those modules for creating other functions to avoid repentance of same functions.

Again, while comparing our module, we find some differences in kurtosis value with panda’s module because of excess kurtosis. But by doing it we have learnt to implement statistical functions by our own so that we understand every behind that. As well as by comparing these modules we are now pretty confidence with ourselves. So, we enjoyed again to implement statistical functions & so made some extra functions.

Another thing we see that the values of the pandas and the naturally calculated value is different the cause is given below:

In probability theory and statistics "kurtosis (from Greek: κυρτός, kyrtos or kurtos, meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable". In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.

Depending on the particular measure of kurtosis that is used, there are various interpretations of kurtosis, and of how particular measures should be interpreted.

The kurtosis is the fourth standardized moment, defined as (Check Diagram 1). Where μ4 is the fourth central moment and σ is the standard deviation. Several letters are used in the literature to denote the kurtosis. A very common choice is κ, which is fine as long as it is clear that it does not refer to a cumulant. Other choices include γ2, to be similar to the notation for skewness, although sometimes this is instead reserved for the excess kurtosis.

The kurtosis is bounded below by the squared skewness plus 1: (Check Diagram 2). Where μ3 is the third central moment. The lower bound is realized by the Bernoulli distribution. There is no upper limit to the excess kurtosis of a general probability distribution, and it may be infinite. A reason why some authors favor the excess kurtosis is that cumulants are extensive. Formulas related to the extensive property are more naturally expressed in terms of the excess kurtosis. For example, let X1, ..., Xn be independent random variables for which the fourth moment exists, and let Y be the random variable defined by the sum of the Xi.

What is Kurtosis? For symmetric, unimodal distributions:

•A positive value tells you that you have heavy-tails (i.e. a lot of data in your tails).

•A negative value means that you have light-tails (i.e. little data in your tails). This heaviness or lightness in the tails usually means that your data looks flatter (or less flat) compared to the normal distribution. The standard normal distribution has a kurtosis of 3, so if your values are close to that then your graph is nearly normal. These nearly normal distributions are called mesokurtic. Kurtosis is the fourth moment in statistics.

Mesokurtic: Mesokurtic distributions are technically defined as having a kurtosis of zero, although the distribution doesn’t have to be exactly zero in order for it to be classified as mesokurtic. The most common mesokurtic distributions are:

•The normal distribution.

•Any distribution with a Gaussian (normal) shape and zero probability at other places on the real line.

•The binomial distribution is mesokurtic for some values (i.e. for p = 1/2±√(1/12).

Excess kurtosis?: Excess kurtosis is usually defined as kurt – 3 (see Important note about equations). It is a measure of how the distribution’s tails compare to the normal (Aldrich, E, 2014).

•Excess kurt for the normal distribution is 0 (i.e. 3 -3 = 0).

•Negative excess equals lighter tails than a normal distribution.

•Positive excess equals heavier tails than the normal.

Calculating Kurtosis: Important Note about formulas: There’s no real consensus for exactly what the correct equation is for calculating kurt. Which definition/equation you use is a matter of convention in your field, the particular software you’re working with, and sometimes the preference of the author. Therefore, it’s a good idea to check which formula you’re working with. This Cross Validated thread has an excellent rundown of the different equations and which software uses which equation.

For Minitab and SPSS, you can find the option in the “Descriptive Statistics” tab.

Kurtosis Excel 2013: KURT function:

Step 1: Type your data into columns in an Excel worksheet.

Step 2: Click a blank cell.

Step 3: Type “=KURT(A1:A99)” where A1:99 is the cell locations for your data.

Kurtosis Excel 2013: Data Analysis:

Step 1: Click the “Data” tab and then click “Data Analysis.”

Step 2: Click “Descriptive Statistics” and then click “OK.”

Step 3: Click the Input Range box and then type the location for your data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box.

Step 4: Click the radio button for Rows or Columns, depending on how your data is laid out.

Step 5: Click the “Labels in first row” box if your data has column headers. Step 6: Click the “Descriptive Statistics” check box.

Step 7: Select a location for your output. For example, click the “New Worksheet” radio button.

Step 8: Click “OK.”

I didnot mean to Teach How to do or Calculate Kurtosis by Excel or SPSS OR Minitab or "R". The point I am Making is as it says in Paragraph up "Depending on the particular measure of kurtosis that is used, there are various interpretations of kurtosis, and of how particular measures should be interpreted.". Might the Softwares you have used R and Excel their Interpretations are Different than each other or they used the Formula which have Different words and Symbols. That is why your Results are 1 Point different not a big Variation though. I hope i have made my point. Regards

# Appendix

## **Contribution Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **A. Coding Part** | | | |
| Sl. | Function name | Member Name | Remarks |
| 1. | Idxmin (), idxmax (),  Quantile (),  MAD(Mean Absolute Deviation).  Var (), SD(),  **Extra**:absolute(),mead(), Sort, IQR() | Adri Saha  (2019-1-60-024) | Write if you have any additional comments. |
| 2. | Skewness,curtosis  Cumax  Cumin  Counting non nul value  Median absolute difference  Percentage change  Cumulative product  Cumulative sum  Aurgmax,aurgmin  Extra:  Linear regression  Nonuni  Query  Co relation  Trimed mean | Amit Roy  (2019-1-60-006) |  |
|  |  |  |  |
|  |  |  |  |
| **B. Report Part** | | | |
| 1. | Problem statement, abstract,  Idxmin (),idxmax (),  Quantile (),  MAD (Mean Absolute Deviation).  Var (), SD(),  **Extra**:absolute(),mead(), Sort, IQR(), disscussion | Adri Saha  (2019-1-60-024) |  |
| 2 | Skewness,curtosis  Cumax  Cumin  Counting non nul value  Median absolute difference  Percentage change  Cumulative product  Cumulative sum  Aurgmax,aurgmin  Extra:  Linear regression  Nonuni  Query  Co relation  Trimed mean | Amit Roy |  |
| 3 | Linear Search  Reverse  Append function  Print function  Median | Moumita Das |  |
|  |  |  |  |