Aradhya Mathur

Data Mining Assignment 1

REPORT

1. **Part I: Textbook problems**

   - 1.1, 1.2, 1.4, 1.5, 1.7, 1.9 (watch out for plagiarism)

1.1 What is *data mining*? In your answer, address the following

Data Mining is the process involving extraction (mining) of interesting pattern or knowledge from huge amount of data.

a) Is it another hype?

Availability of gigantic quantity of data and the urgency of turning them into useful knowledge and information is the reason behind the need for data mining. Data mining can be seen as the outcome of natural development of IT. Data mining field is dynamic and has bright future. To sum it up, it is not another hype, interdisciplinary nature of data mining contributes to vast real-life applications and is one of the most important methods of data extraction and organization.

(b) Is it a simple transformation or application of technology developed from *databases*, *statistics*, *machine learning*, and *pattern recognition*?

It isn't a simple transformation. Data mining is application of technology which has incorporated techniques from fields such as databases, machine learning, statistics and visualizations.

(c) We have presented a view that data mining is the result of the evolution of *database technology*. Do you think that data mining is also the result of the evolution of *machine learning research*? Can you present such views based on the historical progress of this discipline? Address the same for the fields of *statistics* and *pattern recognition*.

Data mining is the evolution of database technology and yes, it is also the result of evolution of machine learning research. With abundance of data and need to extract important patterns and knowledge from it, algorithms are required. In the past obtaining knowledge from huge data was slow, tedious job and with advancement in machine learning research (for example KNN algorithms, Boosting Algorithms), new algorithms with high efficiency and effectiveness have made the process of data mining smoother, simpler and fast. Same goes for the field of statistics, statistics provides the fundamental concepts on which patterns can be extracted which ultimately helps in obtaining knowledge from large data. New statistical methods (for example Naive Bayes) provide a new method to group, normalize data, eventually evolving data mining process. Pattern recognition has similar effect on data mining field. Evolution of pattern recognition algorithms (For example Fuzzy models) have led to highly efficient and accurate methods of finding and extracting patterns. This has a positive effect on evolution and development of Data mining techniques.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Steps involved are:

Data Cleaning: It is a process in which inconsistent data and noise are removed.

Data Integration: Process in which data from diverse data sources are combined.

Data Selection- Process in which relevant data associated with analytic task is retrieved from database.

Data Transformation – Process of transforming and consolidating data into forms apt for mining by applying operations.

Data Mining – Necessary process of extracting data patterns by applying intelligent and efficient methods on the data.

Pattern Evaluation - Process that identifies really interesting patterns illustrating knowledge built on some interestingness measures.

Knowledge Presentation- Process in which mined knowledge is presented to user using visualization and knowledge representation techniques.

1.2) How is a *data warehouse* different from a *database*? How are they similar?

Differences between a database and a data warehouse:

Sr. No.	Data Warehouse	Database
1.	Store of data collected from range of sources, stored in a unified schema	Interrelated data collection
2.	Quickly analyse huge amount of data and provide diverse viewpoints for analysts	Inserts, replaces, updates and deletes huge numbers of short online transactions rapidly.
3.	Historical data	Current, real-time data
4.	Complex queries for detailed analysis	Normal transactional queries

Similarities:

1. Used for storing data.
2. Multi-user access supported.
3. Queries are required for accessing the data.

1.4) Present an example where data mining is crucial to the success of a business. What *data mining functionalities* does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

Walmart operates chain of hypermarkets and department stores across USA (Example is valid for every department store). Walmart requires cross market analysis and customer profiling and using data mining functionalities like association, correlation can be crucial. Using data mining functionalities, Walmart can perform Association/Correlations between product sales and can determine which products are preferred by what types of customers. Based on findings, Walmart can develop and work on effective and efficient marketing and sales strategies.

The patterns can be generated alternatively but realistically query processing doesn't have a way for finding association rules and similarly simple statistics analysis can't handle huge amount of data, like in case of Walmart. Apart from that it would require a load of manual work by experts.

1.5) Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression.

Discrimination and Classification

Difference

Discrimination is comparison of the overall features of objects from target class against the overall features of objects from different classes. While class objects with unknown label can be predicted by classification.

Similarity

Both classification and discrimination measure nominal data type and analyse object.

Characterization and Clustering

Difference:

Characterization refers to summary of overall characteristics of the target class. While clustering is the process of grouping of data points into clusters so that the objects lie in the same group

Similarity:

Characterization and clustering both group objects or related data to compare against data set values.

Classification and Regression

Difference:

Classification involves predicting discrete objects with unknown label. While regression is used to predict continuous variables.

Similarity:

Classification and regression both are used for prediction analysis.

1.7) *Outliers* are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable

Detecting fraudulent use of credit cards is immensely important process and finding exceptions (outliers) in credit card transactions can really help. Two methods proposed below can be easily used to detect outliers.

1) Clustering method: Regular data objects lie in large clusters while outliers either lie in small clusters or don't lie in any. Using this assumption, we can detect outliers.
2) Statistical method: Regular data follow stochastic model and the data that don't follow are outliers. Using this assumption, we can detect outliers.

Clustering method is more reliable as the statistical method won't be able to handle gigantic amount of credit card transaction data and the assumption might not be be correct for given transactional data. But even for clustering method, reliability depends on the used clustering algorithm.

1.9) What are the major challenges of mining a huge amount of data (e.g., billions of tuples) in comparison with mining a small amount of data (e.g., data set of a few hundred tuple)?

Challenges are:

1) Efficiency and Scalability: Data Mining algorithm needs to be efficient and scalable in order to extract information from gigantic amount of data in the data set in acceptable run time.
2) Complexity and Cost: Huge datasets means complex data and in order to reduce computational complexity of applying data mining methods to large size of databases, parallel and distributed data mining algorithms are introduced. Database updates can be integrated without the need to mine entire data from start using incremental data mining algorithms, hence reducing the high cost.

Part II: Playing with data

- Decide on your programming language (why wait?)

Python

- Select a small dataset (e.g., Iris from the UCI repository)

I have selected https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

- Compute some statistics or plot the data in some way, with your own code
- Interpret the statistics or plots (what do they tell you about the data?)
- Bonus points: up to 1 pt (for a total of 6 pts) for all assignments
- Above and beyond what's asked, and relevant (!)
- report + code Due 9/13 (in one week)
- Upload two separate files:
  - one for the code (or Jupyter Notebook)
  - one for the stand-alone report (everything else, not Jupyter Notebook)

Dataset Description

There are 10 attributes, all quantifiable, and a binary dependent variable, representing the cancer patients and healthy people. The attributes are anthropometric data and parameters can be gathered in a normal blood test.

Initial process

Step 1) Read and show data

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1

Step 2) Summarize data

Step 3) Convert into dataframe

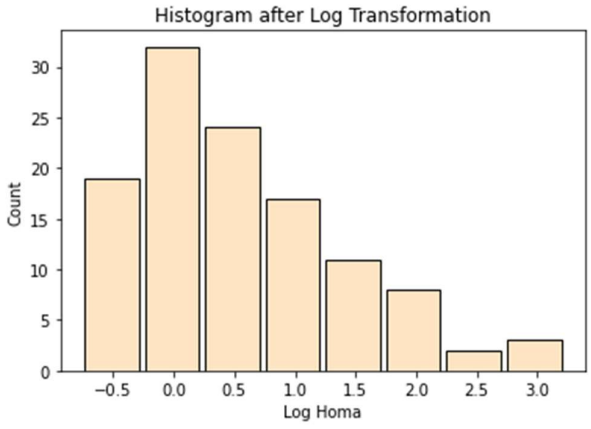
Statistics

1) Skewness of each attribute: Except classification all the attributes are positively skewed. Skewness of Age attribute is lowest, age attribute is almost symmetrical.
2) Summarized Age attribute
3) Determined mean, median, mode, standard deviation, variance, 75th percentile, 25th percentile, Interquartile range, minimum, maximum, geometric mean, fmean, low median, high median, quantile, median at interval=2
4) Reshaped Age into 1D array to find multimode, population sd, population variance, rank data, Z score, 1 sample t-test.
5) Determined summary of other 9 attributes.
6) Frequency table of classification to determine number of healthy people and patients respectively

col_0	People
Classification	
1	52
2	64

7) Performed groupby().mean() operation on classification. **Inference: Higher glucose, insulin, HOMA, Resistin and MCP.1 levels are seen in Patients.**
8) Performed splitting operation to split healthy people and patients
9) Reshaped Leptin attribute to 1D array followed by Normalizing leptin data.
10) Reshaped Resistin attribute to 1D array
11) Used reshaped leptin and resistin data to find correlation coef, covariance, pearson correlation coef, P value, Spearman correlation coef, kendall correlation coef, linear regression and 2 Sample t-test.
12) Applied transformation to HOMA attribute to reduce skewness

12.1) Log transformation: Reshaped HOMA attribute to 1D array
Reduced skewness of HOMA from 3.812 to 0.914 by applying log transformation.
Histogram: Unimodal, right skewed

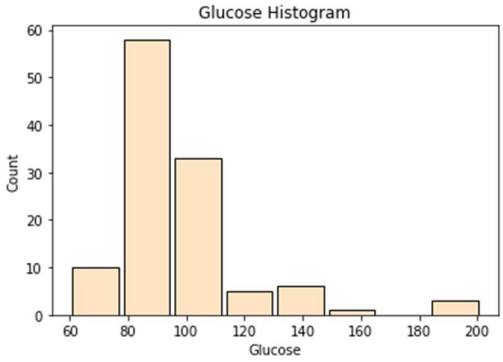


12.2) SQRT transformation: R educed skewness of HOMA from 3.812 to 2.176 by applying sqrt transformation.
Histogram: Unimodal, right skewed
12.3) CBRT transformation: Reduced skewness of HOMA from 3.812 to 1.699 by applying cbrt transformation.
Histogram: Unimodal, right skewed
Inference: Comparing log, sqrt and cbrt transformation, we can infer that log transform works best followed by cbrt and sqrt respectively.

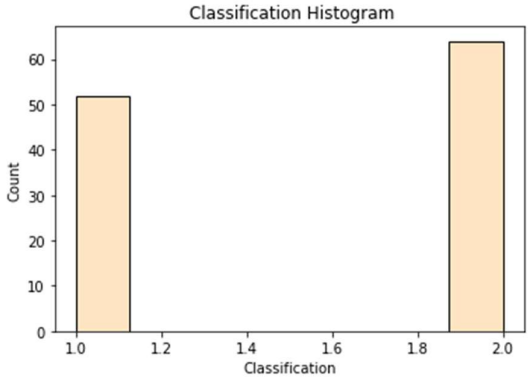
Plotting

a) Histograms

1) Histogram of Age attribute: Symmetrical, Large spread uniform
2) Histogram of BMI attribute: Almost symmetrical and large spread uniform
3) Glucose Histogram: Unimodal, right skewed

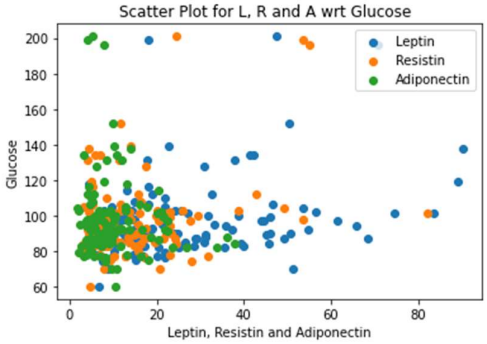


4) Insulin Histogram: Unimodal, right skewed
5) Homa Histogram: Unimodal, right skewed
6) Leptin Histogram: Unimodal, right skewed
7) Normalized Leptin Histogram: Unimodal, right skewed
8) Adiponectin Histogram: Unimodal, right skewed
9) Resistin Histogram: Unimodal, right skewed
10) MCI Histogram: Unimodal, right skewed
11) Classification Histogram: Left skewed

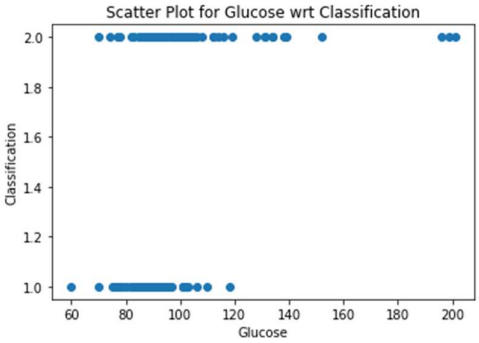


b) Scatter Plots

1) Leptin, Resistin and Adiponectin with respect to Age: Adiponectin and resistin has positive correlation while leptin has almost 0 correlation with respect to Age.
2) Leptin, Resistin and Adiponectin with respect to Glucose: Adiponectin, leptin and resistin have positive correlation with respect to glucose and outliers are noticeable.

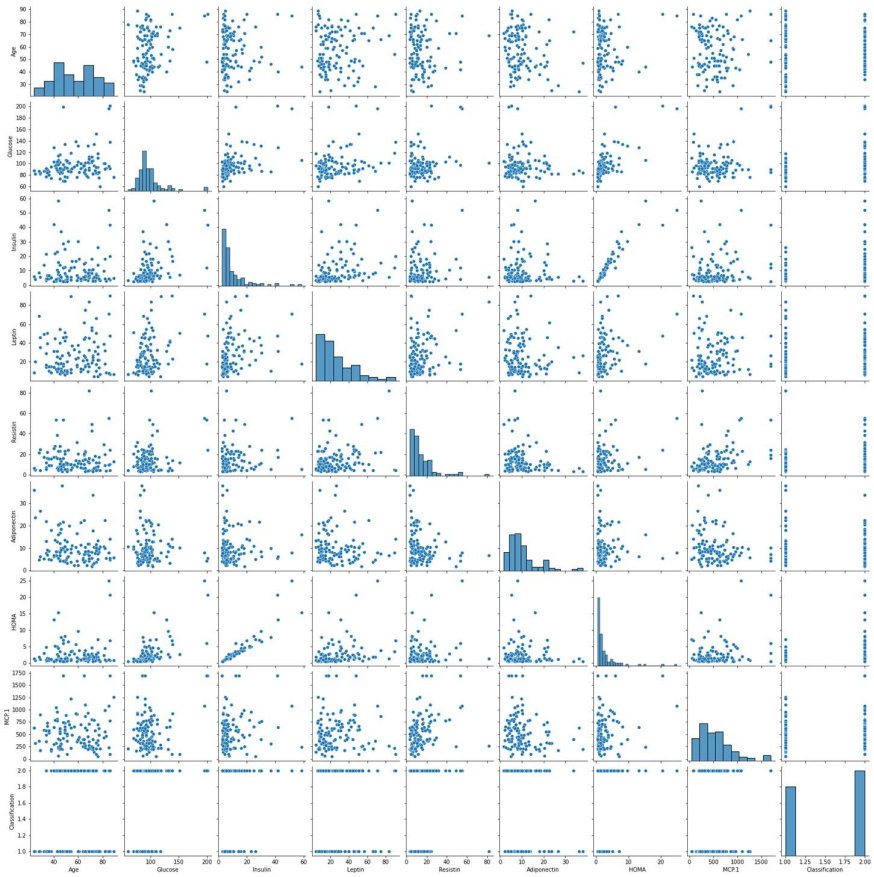


3) Leptin, Resistin and Adiponectin and Insulin with respect to HOMA: Adiponectin, leptin and resistin have positive correlation with respect to HOMA. Insulin has strong positive correlation. Some outliers are noticeable.
4) Age with respect to BMI: No correlation noticeable, highly scattered plot.
5) Resistin with respect to classification: Patients tend to have higher resistin, outliers are noticeable.
6) Glucose with respect to classification: Patients tend to have higher glucose level.



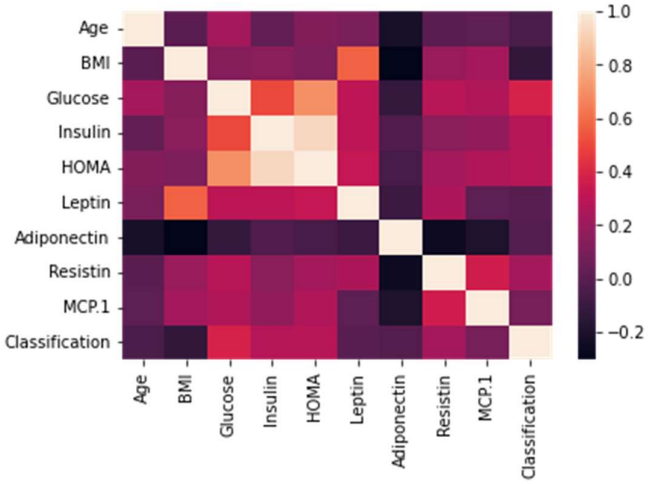
7) HOMA with respect to classification: Patients tend to have higher HOMA, outliers are noticeable.
8) Glucose, Resistin, Age and BMI with respect to Classification: Patients tend to have higher glucose and resistin level. Age and BMI of patients and healthy people are almost similar.

c) Scatter Matrix: Gives scatter plots and histograms of attributes with respect to each other



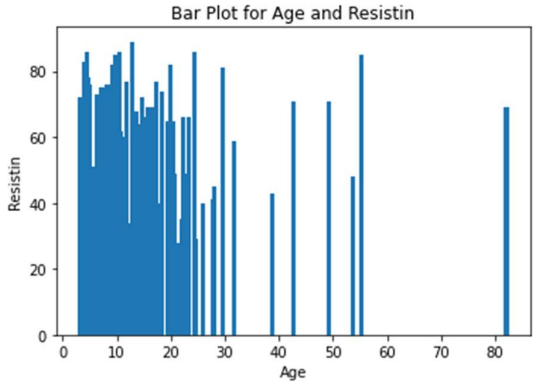
MCP.1 level and Glucose are positively correlated. Same for HOMA and Glucose. All the attributes with respect to Age are not correlated.

d) ***Heatmap***: HOMA and Insulin are highly correlated



e) Bar Plot

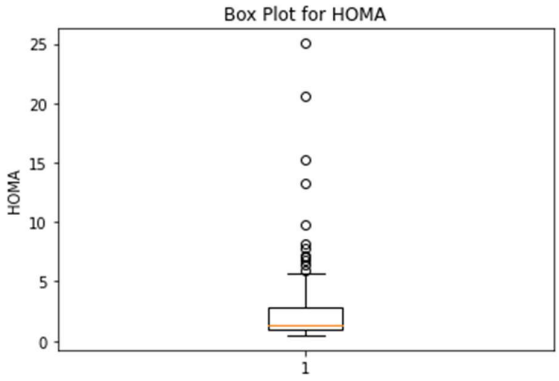
1. Leptin vs Resistin: Majority of people have Leptin and Resistin in the range 0-50
2. Reshaped Age vs Reshaped Resistin: Most people in the age group 0-30 have resistin level6 0-80



3. Age vs Leptin: Most people have leptin in the range 0-55
4. Homa vs Age: Majority of people have HOMA level in the range 0-8

f) Box Plot

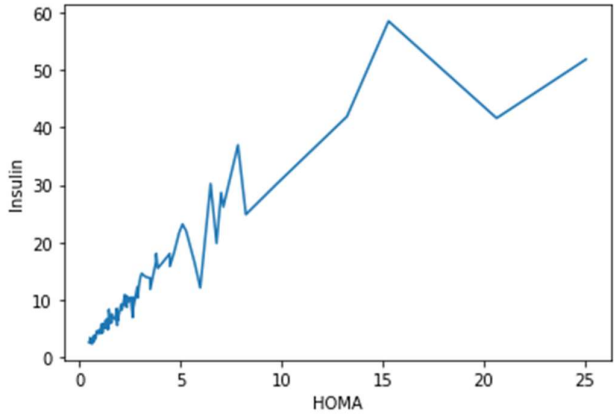
1. Box plot for Age: No outlier detected
2. Box plot for Leptin: Outliers detected
3. Box plot for HOMA: Outliers detected



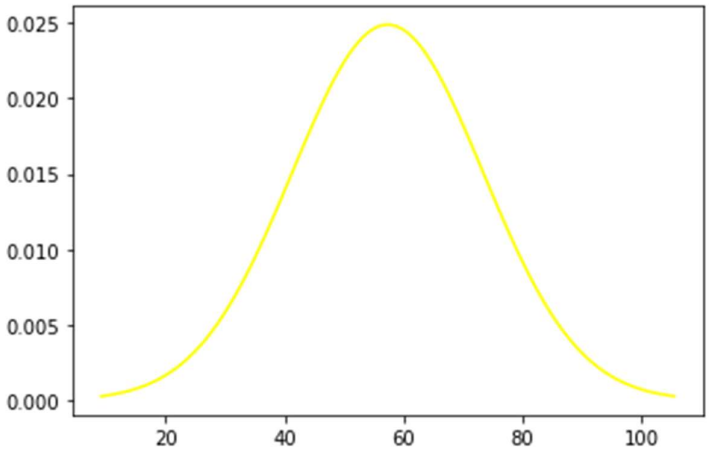
4. Finding outliers: 12 outliers detected, max = 25.050 and minimum =5.969

g) Line plot

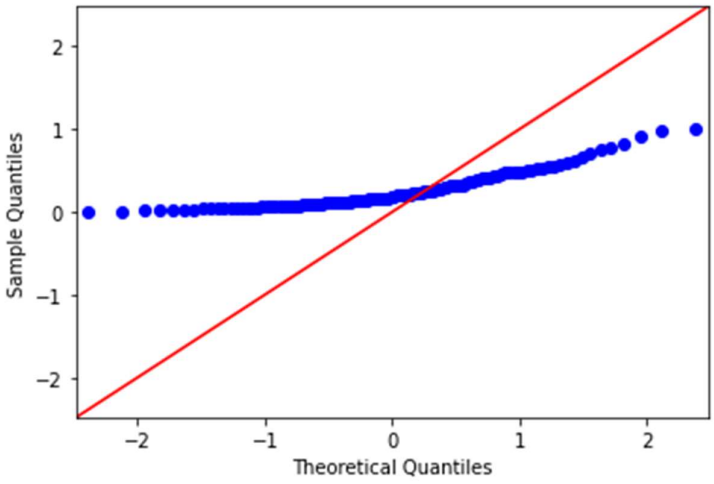
1) BMI vs MCP.1: Highly variable, can be inferred that BMI and MCP.1 levels are not related.
2) Classification vs HOMA: Patients tend to have higher HOMA levels
3) Insulin vs HOMA: It can be inferred higher the HOMA level, higher will be insulin level.



h) **Normal Distribution Curve:** Majority of data lies between 40-80. Mean is around 55. This is for Age attribute.

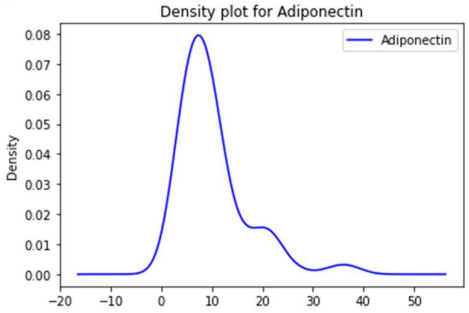


i) **QQ Plot**: Right skewed



j) Density Plot

1) Density plot for MCP.1: Majority of the data points reside between 0-1000
2) Density plot for Adiponectin: Majority of data points reside between 0-20



To conclude, Patients have significantly higher levels in glucose, insulin, HOMA, Resistin and MCP.1 compared to healthy people. HOMA levels are strongly related to Insulin level. Leptin and BMI are positively correlated along with Insulin and Glucose. Breast cancer model can be developed based on these attributes. Log transformation is superior than SQRT and CBRT transformations.

References:

1) Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer by Joana Crisóstomo, Paulo Matafome, Daniela Santos-Silva, Ana L. Gomes, Manuel Gomes, Miguel Patrício, Liliana Letra, Ana B. Sarmento-Ribeiro, Lelita Santos & Raquel Seiça. Dated: 18 February 2016.
2) Numpy.org: https://numpy.org/doc/stable/reference/routines.statistics.html
3) Matlplotlib.org: https://matplotlib.org/stable/plot_types/index