Assignment 1

Garcia, Jorge A.
Department of Physics
New Mexico State University

Sanderson, Kelly
Department of Astronomy
New Mexico State University

(Dated: September 20, 2019)

Course: C S 508

Instructor: Dr. Tuan Le

I. DATA EXPLORATION

In the following section, we rely on the *pandas* library in Python to manipulate the data and calculate the necessary parameters.

A. Average, Standard Deviation, Minimum, and Maximum

The summary statistics for the Breast Cancer Coimbra data set are shown in Table I.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
mean	57.302	27.582	97.793	10.012	2.695	26.615	10.181	14.726	534.647
std	16.113	5.020	22.525	10.068	3.642	19.183	6.843	12.391	345.913
min	24.000	18.370	60.000	2.432	0.467	4.311	1.656	3.210	45.843
max	89.000	38.579	201.000	58.460	25.050	90.280	38.040	82.100	1698.440

TABLE I. Summary Statistics for each attribute in the data set.

B. Covariance and Correlation

The covariance for all attribute combinations is shown in Table II.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Age	259.621	0.689	83.515	5.271	7.455	31.721	-24.238	0.548	75.030
BMI	0.689	25.202	15.700	7.344	2.093	54.854	-10.400	12.152	389.060
Glucose	83.515	15.700	507.383	114.444	57.115	131.826	-18.824	81.309	2063.870
Insulin	5.271	7.344	114.444	101.360	34.181	58.222	-2.156	18.304	607.206
HOMA	7.455	2.093	57.115	34.181	13.264	22.860	-1.404	10.429	326.971
Leptin	31.721	54.854	131.826	58.222	22.860	367.998	-12.522	60.905	92.947
Adiponectin	-24.238	-10.400	-18.824	-2.156	-1.404	-12.522	46.831	-21.399	-475.084
Resistin	0.548	12.152	81.309	18.304	10.429	60.905	-21.399	153.528	1570.726
MCP.1	75.030	389.060	2063.870	607.206	326.971	92.947	-475.084	1570.726	119655.571

TABLE II. Covariance

The correlation between all attribute pairs can be seen in Table III.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Age	1.000	0.009	0.230	0.032	0.127	0.103	-0.220	0.003	0.013
BMI	0.009	1.000	0.139	0.145	0.114	0.570	-0.303	0.195	0.224
Glucose	0.230	0.139	1.000	0.505	0.696	0.305	-0.122	0.291	0.265
Insulin	0.032	0.145	0.505	1.000	0.932	0.301	-0.031	0.147	0.174
HOMA	0.127	0.114	0.696	0.932	1.000	0.327	-0.056	0.231	0.260
Leptin	0.103	0.570	0.305	0.301	0.327	1.000	-0.095	0.256	0.014
Adiponectin	-0.220	-0.303	-0.122	-0.031	-0.056	-0.095	1.000	-0.252	-0.201
Resistin	0.003	0.195	0.291	0.147	0.231	0.256	-0.252	1.000	0.366
MCP.1	0.013	0.224	0.265	0.174	0.260	0.014	-0.201	0.366	1.000

TABLE III. Correlation

C. Histogram

The histograms for each quantitative attribute are shown in Figure 1.

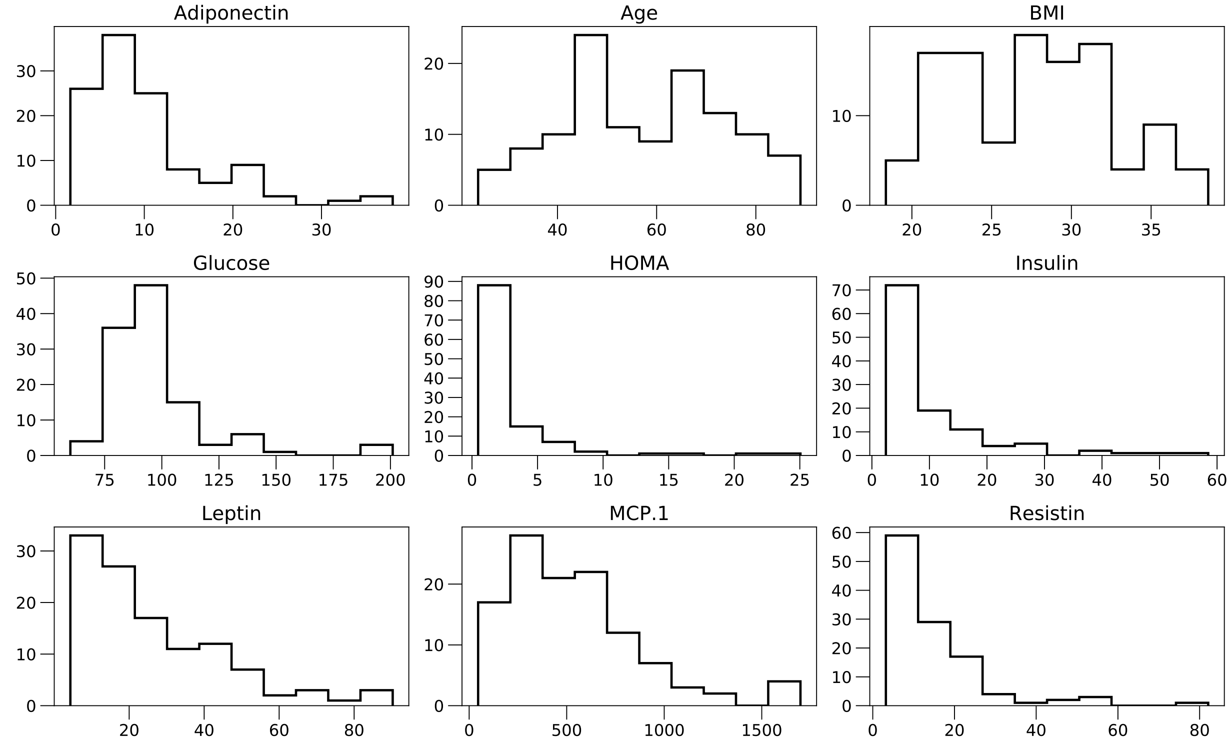

FIG. 1. Histogram of values for each attribute. Histogram bins sizes are 10 units for each attribute.

D. Box Plots

Box plots are shown for each qualitative attribute in Figure 2. These plots reveal that all of the attributes have at least one outlier aside from BMI.

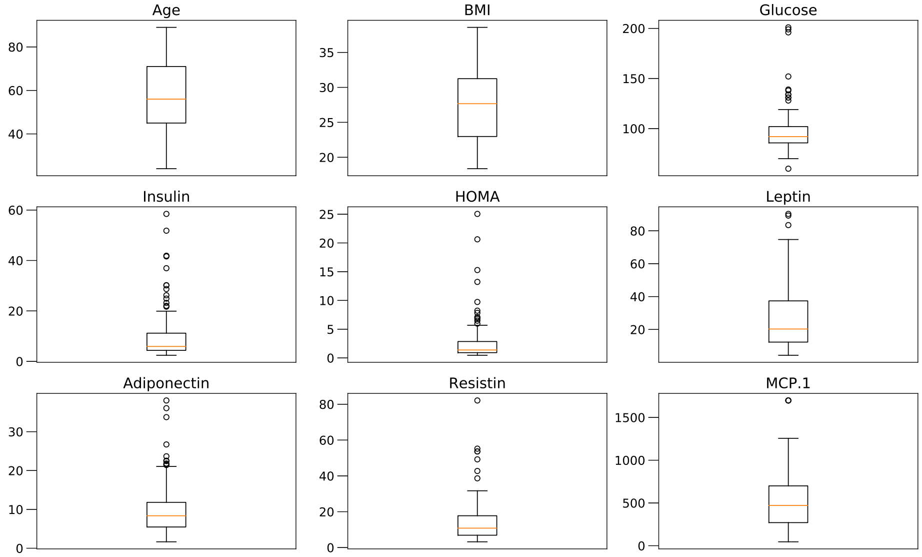

FIG. 2.

E. Scatter Plots & Observable Correlations

Scatter plots between Age, BMI, Glucose, and Insulin can be observed in Figure 3.

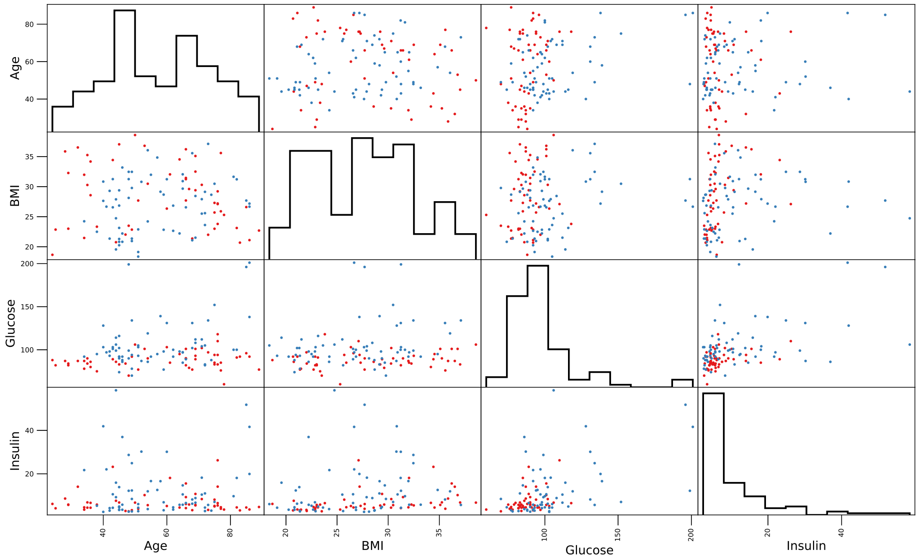

FIG. 3. Joint distribution for each pair of attributes selected from Age, BMI, Glucose, and Insulin.

F. Parallel Coordinates

A parallel coordinates plot of the attributes can be seen in Figure 4.

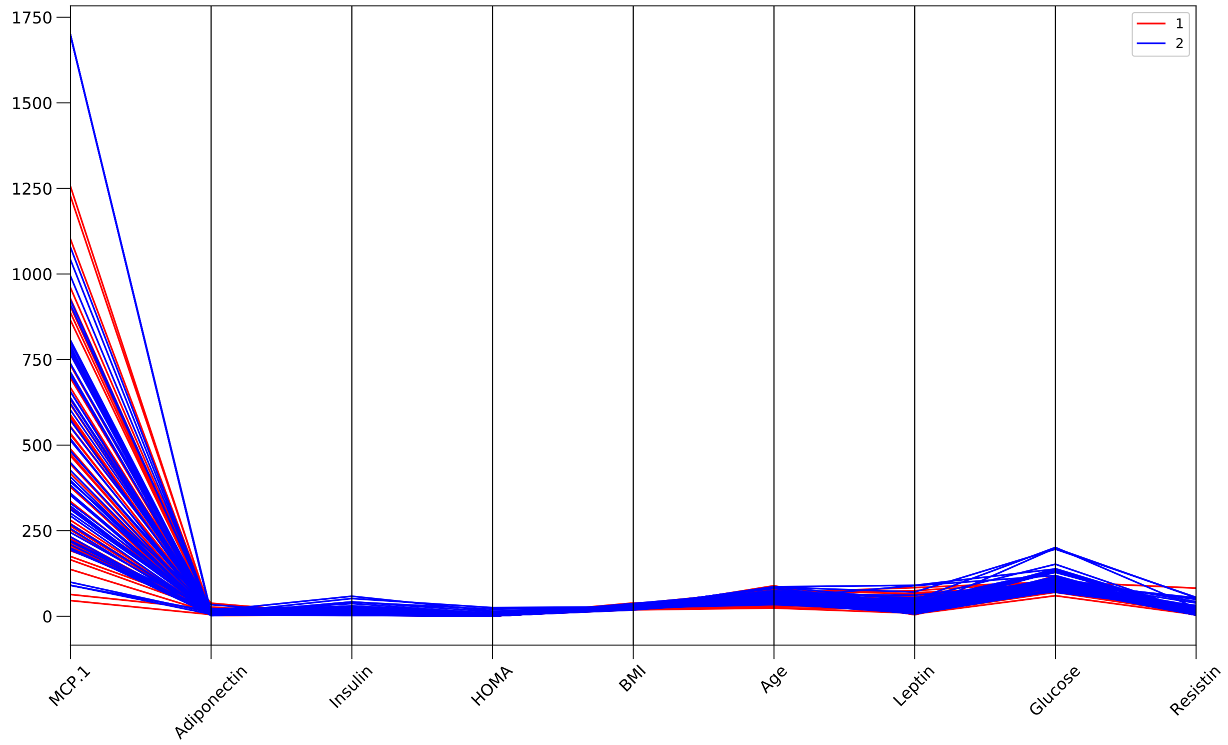

FIG. 4. Parallel coordinates plot of the attribtues in the data.

II. DATA PROCESSING

A. Random Sampling

A random sample of n-dimensional data points is shown in Table IV.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
12	25	22.860	82	4.090	0.827	20.450	23.670	5.140	313.730	1
91	82	31.217	100	18.077	4.459	31.645	9.924	19.947	994.316	2
103	72	29.136	83	10.949	2.242	26.808	2.785	14.770	232.018	2
97	40	27.636	103	2.432	0.618	14.322	6.784	26.014	293.123	2
53	45	20.830	74	4.560	0.832	7.753	8.237	28.032	382.955	2
69	44	19.560	114	15.890	4.468	13.080	20.370	4.620	220.660	2
31	53	36.790	101	10.175	2.535	27.184	20.030	10.263	695.754	1
99	69	28.444	108	8.808	2.346	14.748	5.288	16.485	353.568	2
46	75	25.700	94	8.079	1.873	65.926	3.741	4.497	206.802	1
9	75	23.000	83	4.952	1.014	17.127	11.579	7.091	318.302	1

TABLE IV. Random Sample

B. Principal Component Analysis (Bonus Question)

Using the *sklearn* library, PCA was performed on the data and reduced to a 2-dimensional vector. The new attributes are plotted in Figure 5

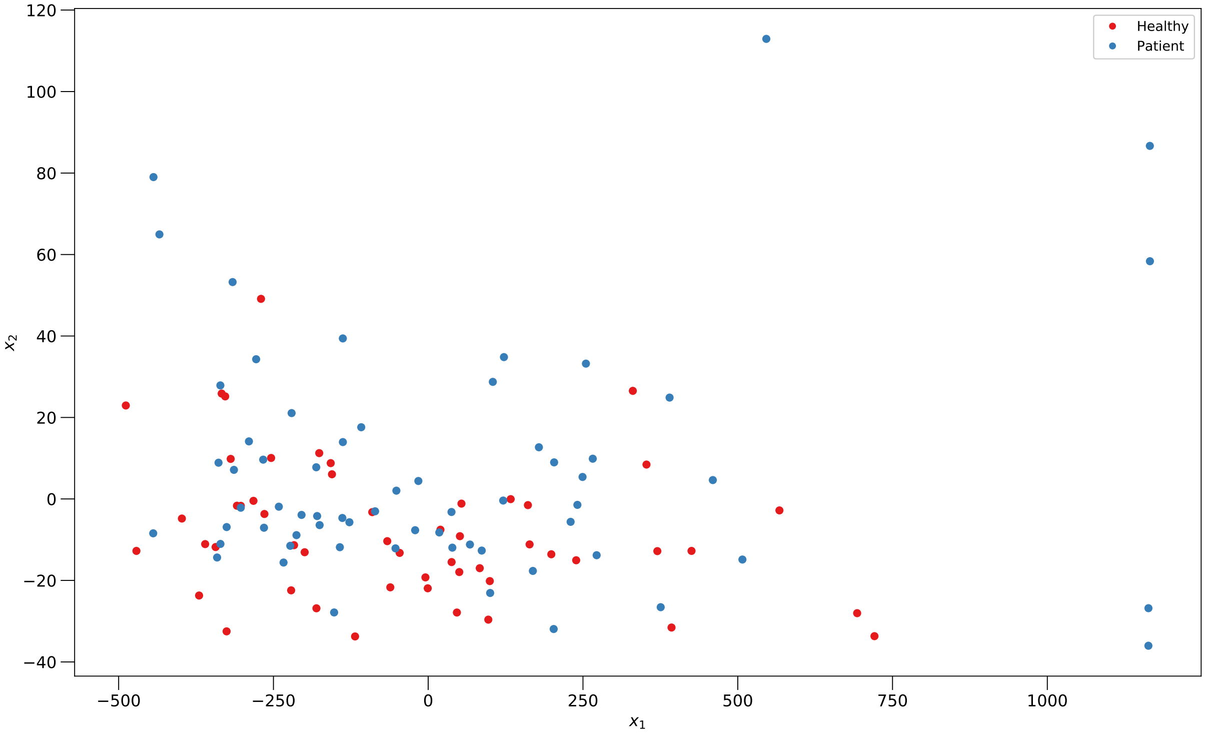

FIG. 5. Scatter plot of the resultant attributes from PCA.

[1] Melissinos, A. C. and Napolitano, J., *Experiments in Modern Physics*, 2nd ed. (Academic Press, 2003).
[2] Papavassiliou, V., New Mexico State University (2018).