**Data Wrangling of Breast Cancer Capstone 1 Project**

1. **Raw Data Example-** Below is a 5 row sample of the original, unmodified data from the original source. A 5 row sample will be used for every sample in this document.

MCP.1 Classification

417.114 1

468.786 1

554.697 1

928.220 1

773.920 1

Age BMI Glucose Insulin HOMA Leptin Adiponectin Resistin

0 48 23.500000 70 2.707 0.467409 8.8071 9.702400 7.99585

1 83 20.690495 92 3.115 0.706897 8.8438 5.429285 4.06405

2 82 23.124670 91 4.498 1.009651 17.9393 22.432040 9.27715

3 68 21.367521 77 3.226 0.612725 9.8827 7.169560 12.76600

4 86 21.111111 92 3.549 0.805386 6.6994 4.819240 10.57635

1. **Data Indexed by Classification and Index renamed to Control/Patient**

Age BMI Glucose Insulin HOMA Leptin Adiponectin Resistin MCP.1

Control Or Patient

Control 48 23.500000 70 2.707 0.467409 8.8071 9.702400 7.99585 417.114

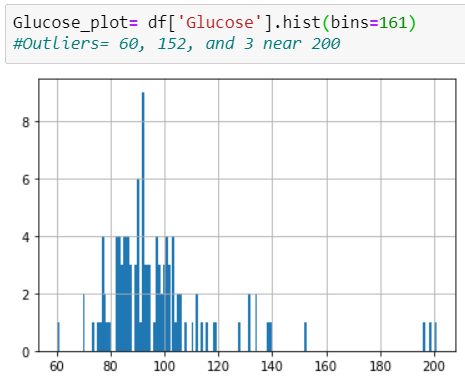
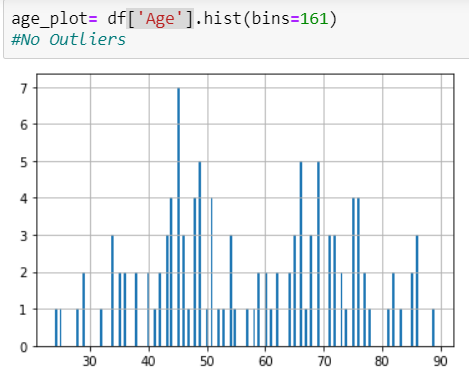
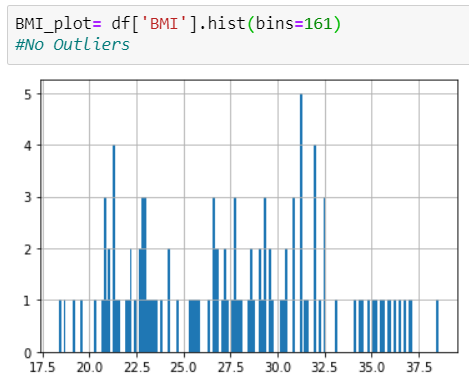
Control 83 20.690495 92 3.115 0.706897 8.8438 5.429285 4.06405 468.786

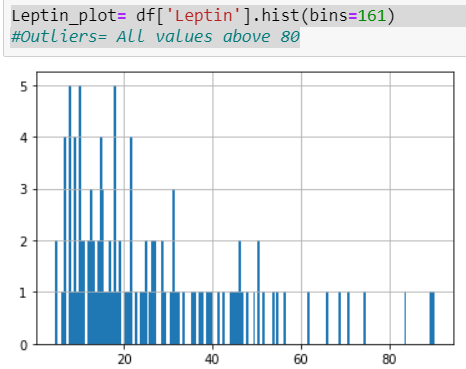
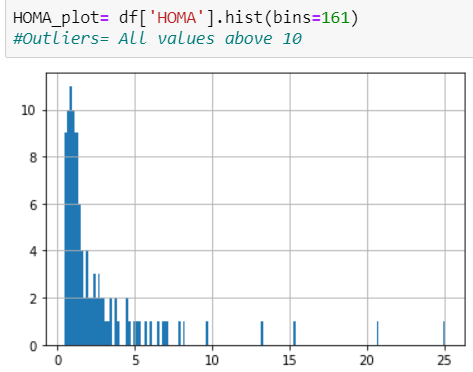
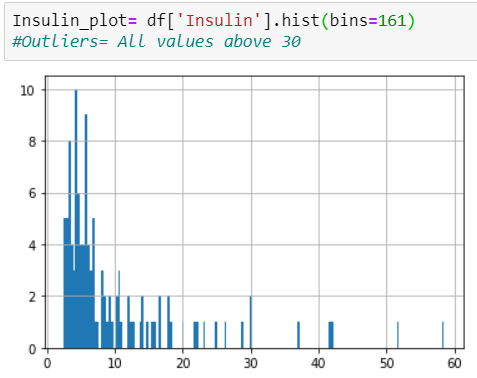
Control 82 23.124670 91 4.498 1.009651 17.9393 22.432040 9.27715 554.697

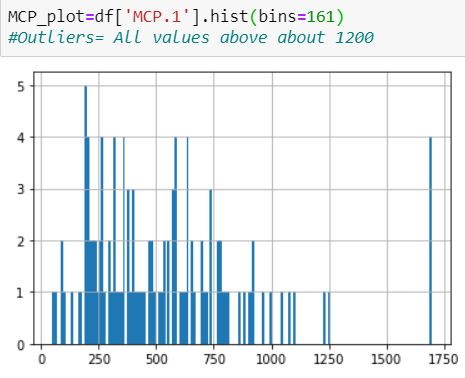
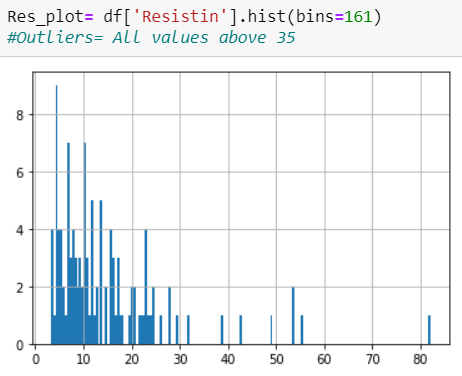
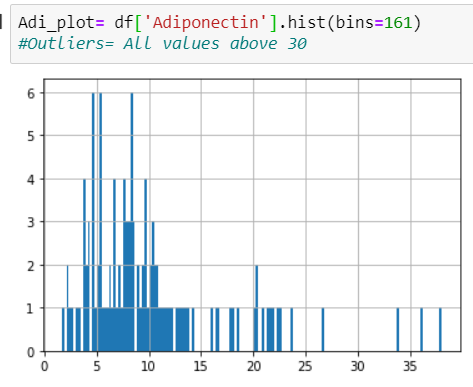
Control 68 21.367521 77 3.226 0.612725 9.8827 7.169560 12.76600 928.220

Control 86 21.111111 92 3.549 0.805386 6.6994 4.819240 10.57635 773.920

1. **Outliers were determined by plotting histograms for each variable and are shown below:**







1. **Means of Variables Grouped by Control/Patient-** This will give me a rough estimate of whether a relationship between variables is likely

Age BMI Glucose Insulin HOMA

Control Or Patient

Control 58.076923 28.317336 88.230769 6.933769 1.552398

Patient 56.671875 26.984740 105.562500 12.513219 3.623342

Leptin Adiponectin Resistin MCP.1

Control 26.637933 10.328205 11.614813 499.730692

Patient 26.596512 10.061167 17.253777 563.016500

<http://localhost:8889/notebooks/Data%20Science%20Capstone%201-%20Breast%20Cancer.ipynb>