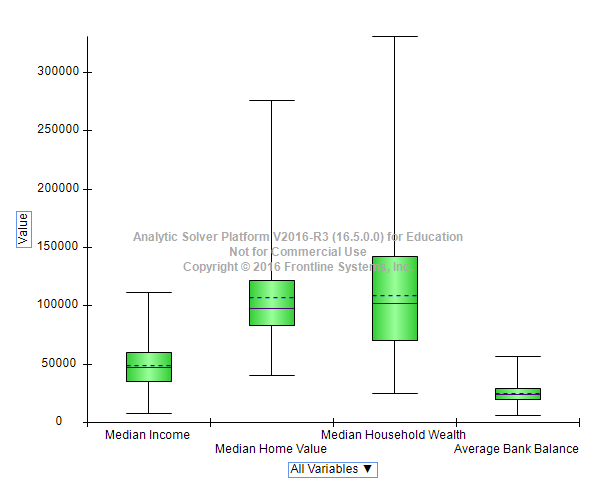
**ANLY500 Homework 8**

You can use R or any other appropriate software for this homework

10.2. Use the Excel file Banking Data.

a) Construct a boxplot for the Median Income, Median Home Value, Median Household Wealth, and Average Bank Balance.



b) What observations can you make about these data?

|  |
| --- |
| Wealth has most variability and positive skewed |

10.7. For the Excel file Colleges and Universities, normalize each column of the numerical data (i.e., compute a z-score for each of the values) and then compute the Euclidean distances between the following schools: Amherst, Cal Tech, and Duke.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Median SAT | | Acceptance Rate | | | Student | Top 10% HS | | | Graduation % | |
| Amherst | | 0.83 | | -1.2 | | | -0.22 | 0.6 | | | 1.31 | |
| Cal Tech | | 2.18 | | -0.53 | | | 4.67 | 1.76 | | | -1.11 | |
| Duke | | 0.75 | | -0.98 | | | 0.61 | 1.24 | | | 1.04 | |
|  |  | | Differences | | | | | | | | | Distance | |
| Amherst | Cal Tech | | 1.84 | | 0.45 | 23.92 | | | 0.92 | 5.84 | | 5.74 | |
| Cal Tech | Duke | | 2.06 | | 0.2 | 16.47 | | | 0.27 | 4.61 | | 4.86 | |
| Duke | Amherst | | 0.01 | | 0.05 | 0.69 | | | 0.2 | 0.07 | | 1.01 | |

10.11. Apply cluster analysis to the Excel file Sales Data, using the input variables Percent Gross Profit, Industry Code, and Competitive Rating. Create four clusters and draw conclusions about the groupings.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Avg of percent gross | Avg of competitive rate | Avg of industry |
| 1 | 0.51 | 2 | 1 |
| 2 | 0.2047 | 3.125 | 4.5 |
| 3 | 0.215 | 1 | 5 |
| 4 | 0.37 | 1 | 6 |
| **Grand Total** | **0.21193** | **3** | **4.48333** |

10.16. The Excel file Credit Risk Data provides a database of information about loan applications along with a classification of credit risk in column L. Convert the categorical data into numerical codes as appropriate. Sample 200 records from the data set. Then apply the k-NN algorithm to classify training and validation data sets and the additional data in the file. Summarize your findings.