Jacob Vargo

CS 425 project 1

1) Data exploration

The auto-mpg.data file is in a csv format. The auto-mpg.names file describes the meaing and types of the fields in the data file. The values referred to as continuous are doubles, and the multi-valued discrete numbers refer to integers. The string values are character strings. There are 6 missing data points in the horsepower field of the auto-mpg.data file. These missing data point are also noted in the auto-mpg.names file. The missing entries are represented as a '?', and all string fields are present and do not appear to be anomalous.

2) Data preparation

Mpg:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 23.51 | 1.56 | 9 | 46.6 | 17.5, 23, 29 | 398 |

Cylinders:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 5.45 | 0.20 | 3 | 8 | 4, 4, 8 | 398 |

Displacement:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 193.43 | 5.97 | 68 | 455 | 104, 151, 262 | 398 |

Horsepower:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 104.47 | 4.15 | 46 | 230 | 75, 94, 129 | 392 |

Weight:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 2970.42 | 136.50 | 1613 | 5140 | 2223, 2807, 3609 | 398 |

Acceleration:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 15.57 | 0.97 | 8 | 24.8 | 13.8, 15.5, 17.2 | 398 |

Model year:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 76.01 | 4.12 | 70 | 82 | 73, 76, 79 | 398 |

Origin:

| Mean | SD | Min | Max | Quartiles | Number of Values |
|---|---|---|---|---|---|
| 1.57 | 0.05 | 1 | 3 | 1, 1, 2 | 398 |

To deal with missing and anomalous data I decided to simply remove the affected data entries. I removed the data entirely because I do not yet have a suitable method of estimating the data entry that would not introduce unwanted bias.

3) Dimension reduction

To start, I decided to not reduce the dimensionality of the data and wait to see if later steps will require the reduction.