# Report:

I used the table, max, min and mean command for all columns in the **fp06.dat** dataframe to verify which columns are usable.

Most of the columns yielded reasonable results. For instance, the threads column had the following distribution

```
> table(fp06.dat$threads)

  1   2
103  90
```

All entries were concentrated in 2 categories. Other columns that exhibited a close distribution of entries were voltage, dieSize, and transistor.

Some columns such as the transistor had NA values, as can be seen:

```
> mean(fp06.dat$transistors)
[1] NA
> mean(fp06.dat$transistors, na.rm = TRUE)
[1] 518.6663
```

The ratio of NA to total entries was 0.155, which I believe is acceptable.

The featureSize column had an abnormal entry of 0.13. This can be seen from the following analysis:

```
> mean(fp06.dat$featureSize, na.rm = TRUE)
[1] 0.05636788
> median(fp06.dat$featureSize)
[1] 0.045
> min(fp06.dat$featureSize)
[1] 0.032
> max(fp06.dat$featureSize)
[1] 0.13
> sd(fp06.dat$featureSize)
[1] 0.0164581
```

The mean and median average at about 0.05, with a minimum of 0.032 and a standard deviation of 0.016. However, the max value recorded is approximately 5 standard deviations above the mean and median.

The FO4delay column also exhibited an odd distribution.

```
> mean(fp06.dat$FO4delay)
[1] 11.98724
> median(fp06.dat$FO4delay)
[1] 12.6
```

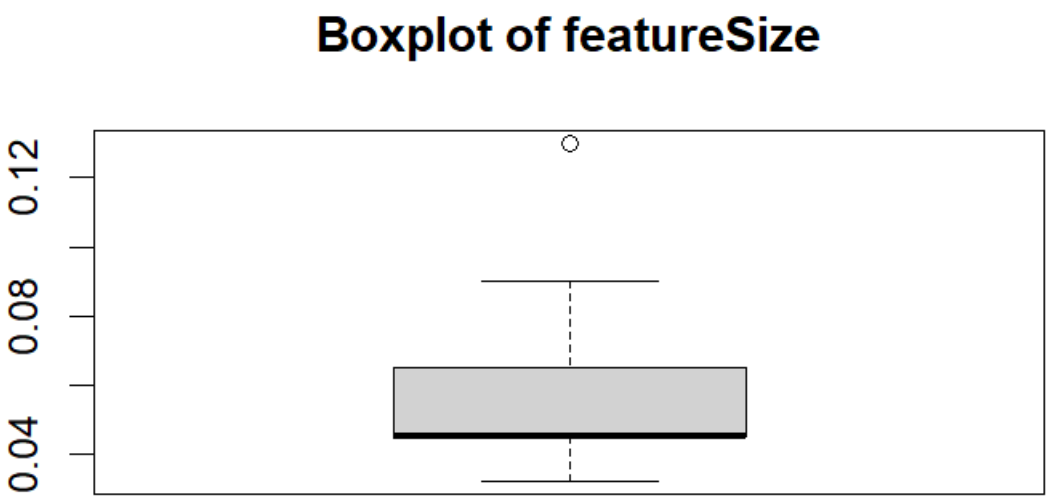
```
> sd(fp06.dat$FO4delay)
[1] 3.082101
> max(fp06.dat$FO4delay)
[1] 25.2
> min(fp06.dat$FO4delay)
[1] 6.48
```

Again, the mean and median center at around twelve, with a standard deviation of 3, however, the maximum recorded is 25.2 which is 4 standard deviations above the mean.

Lastly, I also checked the ratio of NA entries to total entries. I found that the L3cache column is unreliable as it has a 63% NA ratio:

```
> nrow(fp06.dat)
[1] 193
> sum(is.na(fp06.dat$L3cache))
[1] 124
> 124/193
[1] 0.642487
```

We could further enhance our data cleaning operation by looking at boxplots for each column. For example, the feature size boxplot very clearly highlights the outlier in the data:




In conclusion, most of the data in the **fp06.dat** dataframe is centered around the 1st and 3rd quartiles of its respective columns. Most columns do not have outliers. One one column (L3cache) has a great ratio of NA entries. 9 out of the 16 columns have no NA entries at all. Considering the above, I believe it is a good dataset for statistical analysis.