LAB 2 REPORT

This report summarizes the analysis conducted on various benchmark data extracted from the "all-data.csv" file.

1. Statistical Computation
   For the Int1992 dataset, I computed the following statistics:

   Mean, Variance, Minimum and Maximum

```
> mean(int92.dat$perf)
[1] 124.2859
> sd(int92.dat$perf)
[1] 78.0974
> min(int92.dat$perf)
[1] 36.7
> max(int92.dat$perf)
[1] 366.857
```

2. Sort the column to look for outliers or unusual patterns.
   Sorted the transistors values from int92.dat.

```
> sort(int92.dat$transistors)
 [1]  0.58  0.85  0.85  0.85  0.85
 [6]  0.85  1.50  1.50  1.50  1.60
[11]  1.60  1.68  1.68  1.68  1.68
[16]  1.68  1.68  1.80  1.80  1.80
[21]  2.30  2.30  2.30  2.30  2.30
[26]  2.30  2.30  2.30  2.30  2.60
[31]  2.80  2.80  2.85  2.85  2.85
[36]  2.85  3.10  3.10  3.10  3.20
[41]  3.20  3.30  3.60  3.60  3.60
[46]  9.30  9.30  9.30  9.30  9.30
[51]  9.30 15.00
>
```

   From this we can see the distribution of transistor counts. The highest value is 15.00, which stands out as significantly higher than the rest. We can also see that after values around 3.0 to 4.0 it goes up drastically to 9.30.

3. Compute the fraction of NA values to the total number of values in the column to see if there are enough values for the column to be useful.

```
> nrow(int00.dat)-sum(is.na(int00.dat$voltage))
[1] 125
> sum(is.na(int00.dat$voltage))
[1] 131
> nrow(int00.dat)
[1] 256
>
```

Computing nrow() we find that the dataframe has 256 rows. Doing sum(is.na()) for a column we can find out the number of NA's in it. While computing sum(is.na(int00.dat$voltage)) we find there are 131 NA's in the voltage column. Subtracting the number of rows and the number of NA's in the voltage column we get 125 non NA entries.

The fraction of NA values in the voltage column of int00.dat is computed as:

```
> sum(is.na(int00.dat$voltage))/nrow(int00.dat)
[1] 0.5117188
```

This shows that 51.17% of the values in the voltage column are missing. Since more than half of the data is missing, this column may not be very reliable or useful for analysis.

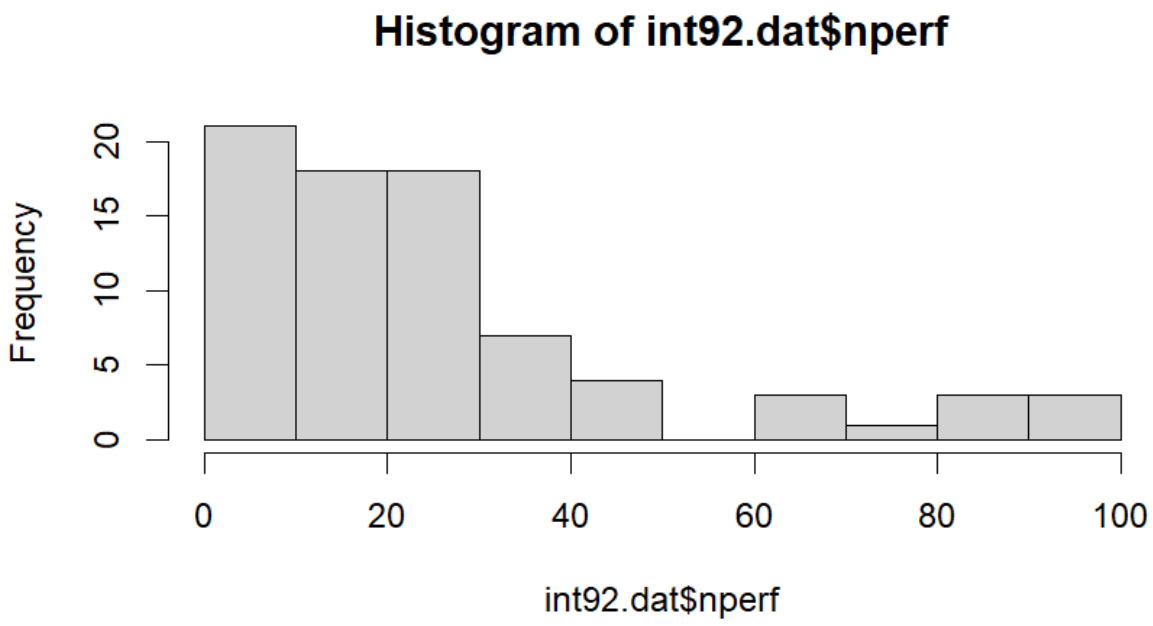
4. Use the table() function to determine if the distribution of values appears to have any anomalies.

```
> table(int92.dat$channel)

0.25 0.29  0.3 0.35  0.4 0.45  0.5 0.55  0.6 0.65  0.7 0.75  0.8    1
   4    1    1    4    4    2   16    2   11    3    1   12   13    1
```

This creates a frequency table showing the counts of each unique value in the channel column. From this we see that 0.29, 0.3, 0.7 and 1 are values with unusually low counts compared to others and the value 0.5 has the highest frequency (16).

5. Additional Verification
We can plot a histogram to visualize column nperf and get an idea of how the values vary.



The peak or cluster of data is located between 0-20 on the horizontal axis. This interval has the highest bar, indicating the highest frequency of data points. The lowest cluster is between 50 - 60 on the horizontal axis.

6. Removal of Rows/Columns: If a column has more than 50% missing data, as with voltage, it's often better to drop the column unless it holds critical information. Similarly, if only a few rows in the dataset are missing values, we can remove those without losing much important information.

   Imputation: For columns with fewer missing values, we can use imputation techniques. For example: Replace missing values with the mean or median of the existing data.