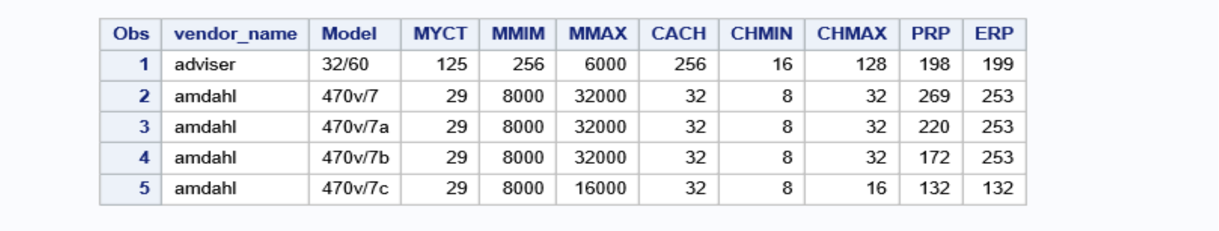
Part 1: Dataset description (Using SAS)

What is the data set?

The name of the dataset is Computer Hardware Dataset. The dataset is about Relative Performance of CPU.



Obs	vendor_name	Model	MYCT	MMIM	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253
5	amdahl	470v/7c	29	8000	16000	32	8	16	132	132

The following code shows that that dataset is read and stored in data frame named "Data".

Where did you get the data?

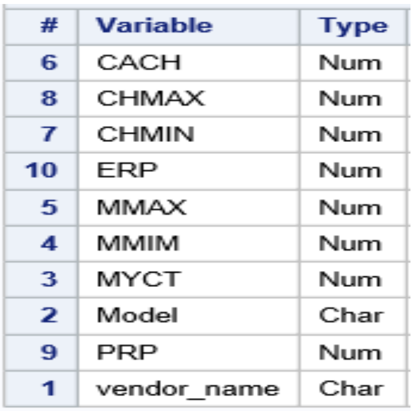
The data is taken from UCI Machine Learning Repository. (http://archive.ics.uci.edu/ml/datasets/Computer+Hardware)

What are the variables of interest?

These are the attributes of the dataset taken. We have focused on

- Model,
- PRP
- Vendor_name variabes.

(Appendix 1)



#	Variable	Type
6	CACH	Num
8	CHMAX	Num
7	CHMIN	Num
10	ERP	Num
5	MMAX	Num
4	MMIM	Num
3	MYCT	Num
2	Model	Char
9	PRP	Num
1	vendor_name	Char

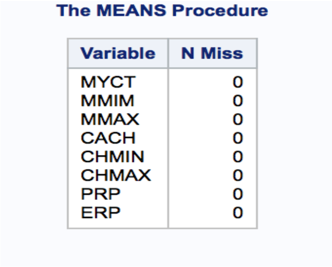
Are there any problems with the data?

No , it is clean dataset. There are no missing values.

To find this the following can be done:

```
PROC MEANS data=Data NMISS;
RUN;
```

The output for the above code will be as follows:

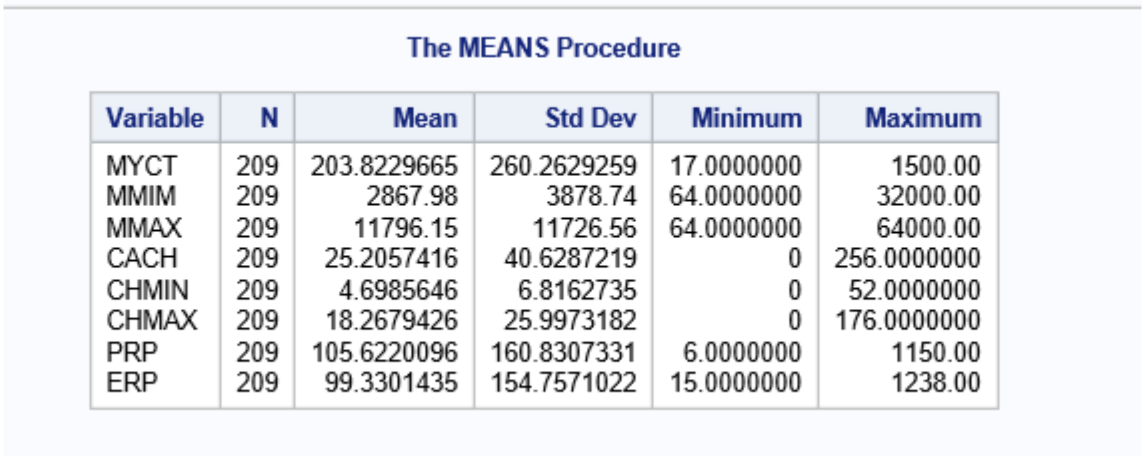


The MEANS Procedure

Variable	N Miss
MYCT	0
MMIM	0
MMAX	0
CACH	0
CHMIN	0
CHMAX	0
PRP	0
ERP	0

Part 2: Summary

Summary of the dataset (Using SAS): there are 209 observations,



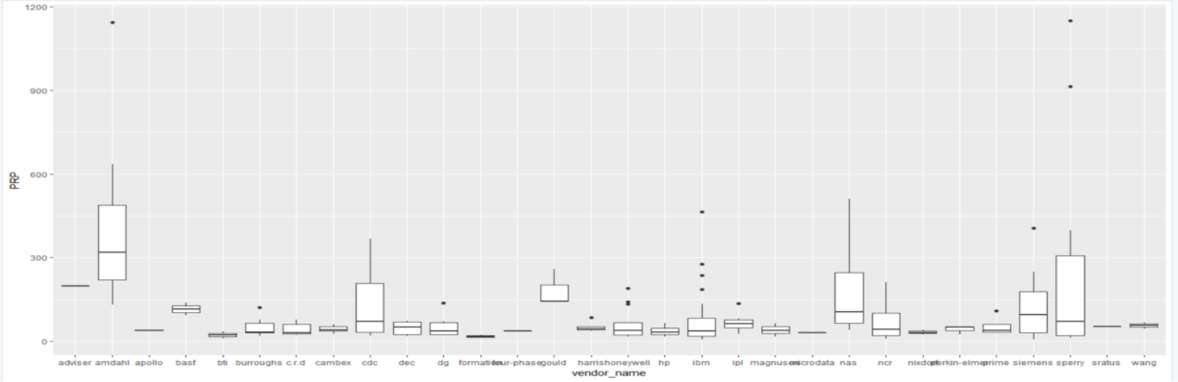
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
MYCT	209	203.8229665	260.2629259	17.0000000	1500.00
MMIM	209	2867.98	3878.74	64.0000000	32000.00
MMAX	209	11796.15	11726.56	64.0000000	64000.00
CACH	209	25.2057416	40.6287219	0	256.0000000
CHMIN	209	4.6985646	6.8162735	0	52.0000000
CHMAX	209	18.2679426	25.9973182	0	176.0000000
PRP	209	105.6220096	160.8307331	6.0000000	1150.00
ERP	209	99.3301435	154.7571022	15.0000000	1238.00

Minimum value of Maximum memory capacity and minimum memory capacity are same=64.

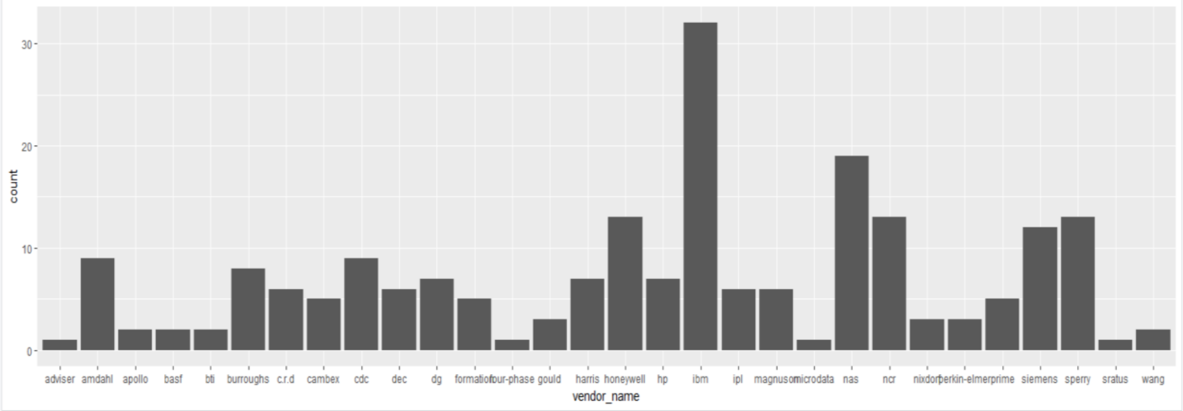
Observations: (Using R GGPlot):

Vendor Vs PRP (Box and Whiskers plot):



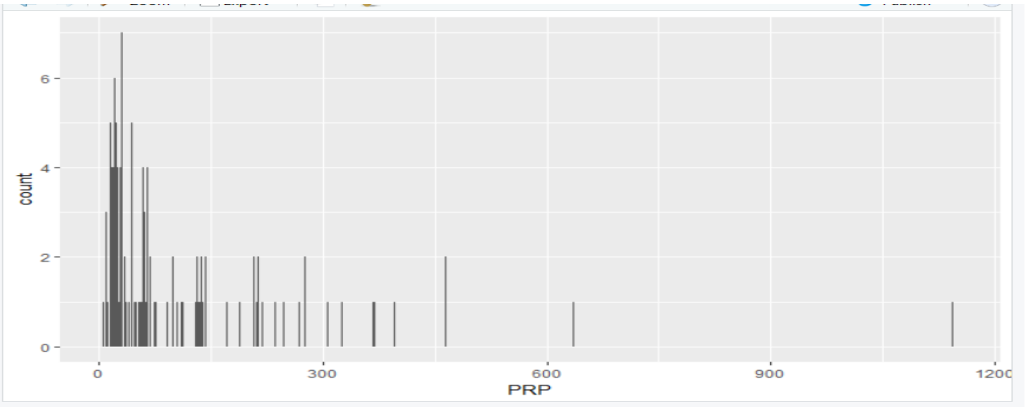
The vendors “amdahl” and “nas” produces computer hardware CPU with more performance when excluding outliers.

Vendor_name distribution:



IBM produces more number of Computer Hardware. This is a comb distribution (the bars are alternately tall and short).

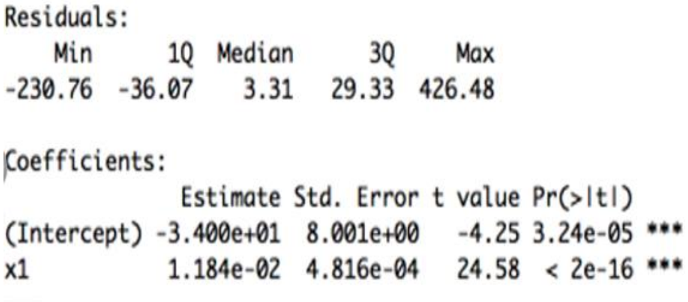
Distribution of PRP:
the factor value (PRP) is continuously valued**.**
The published relative performance is concentrated between 21 -100 and <4 above 500. The distribution is Right-skewed. The distribution is shown in appendix (2)



Part 3: Analysis (Using R)

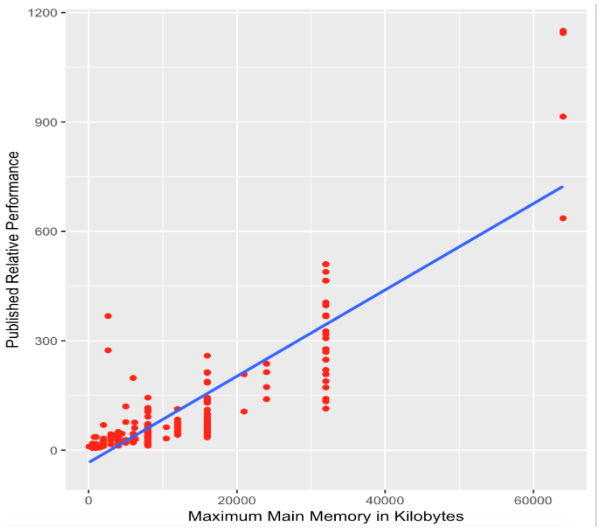
Covariate is MMAX and Response variable is PRP.

In the summary, we can clearly see that the intercept is negative i.e., -3.400e+01. The slope is 1.184e-02.
Here, since the p value < 0.05, we reject the null hypothesis.
Also, since |t|>2, we reject the null hypothesis.



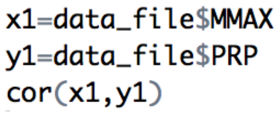
```
Residuals:
    Min      1Q  Median      3Q     Max 
-230.76  -36.07    3.31   29.33  426.48 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.400e+01  8.001e+00   -4.25 3.24e-05 ***
x1           1.184e-02  4.816e-04   24.58  < 2e-16 ***
```



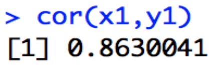
To find the correlation we can use the following code:

Code:



```
x1=data_file$MMAX
y1=data_file$PRP
cor(x1,y1)
```

The output will be as follows:



```
> cor(x1,y1)
[1] 0.8630041
```

Observations:

The correlation coefficient between the maximum main memory in kilobytes and published relative performance is 0.86 which is close to 1, which means the variables are positively linearly related.

From the above model, we can see that as the maximum main memory in kilobytes increases the relative performance also increases. We can see that, when the maximum memory is 12000 kb the relative performance is 113 and when maximum memory is 24000 kb the relative performance is 237.

Appendix:

1. Attribute Information:

1.vendor name:
2. Model Name: many unique symbols
3. MYCT: machine cycle time in nanoseconds (integer)
4. MMIN: minimum main memory in kilobytes (integer)
5. MMAX: maximum main memory in kilobytes (integer)
6. CACH: cache memory in kilobytes (integer)
7. CHMIN: minimum channels in units (integer)
8. CHMAX: maximum channels in units (integer)
9. PRP: published relative performance (integer)
10. ERP: estimated relative performance from the original article (integer)

2. Class Distribution of published relative performance(PRP)

PRP Value Range	Number of Instances in Range
1.0-20	31
2. 21-100	121
3. 201-300	13
4. 301-400	7
5. 401-500	4
6. 501-600	2
7. >600	4