Processor analysis

Łukasz Berwid

2/2/2020

1. Download and load data

```
machine_data <- read.csv("http://mlr.cs.umass.edu/ml/machine-learning-databases/cpu-performan
ce/machine.data", header=FALSE, colClasses = c("factor", "character", "integer", "integer",
"integer", "integer", "integer", "integer", "integer"), col.names = c('vendor name' ,'Model N
ame','MYCT','MMIN','MMAX','CACH','CHMIN','CHMAX','PRP','ERP'))
kable(head(machine_data))
```

vendor.name	Model.Name	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
adviser	32/60	125	256	6000	256	16	128	198	199
amdahl	470v/7	29	8000	32000	32	8	32	269	253
amdahl	470v/7a	29	8000	32000	32	8	32	220	253
amdahl	470v/7b	29	8000	32000	32	8	32	172	253
amdahl	470v/7c	29	8000	16000	32	8	16	132	132
amdahl	470v/b	26	8000	32000	64	8	32	318	290

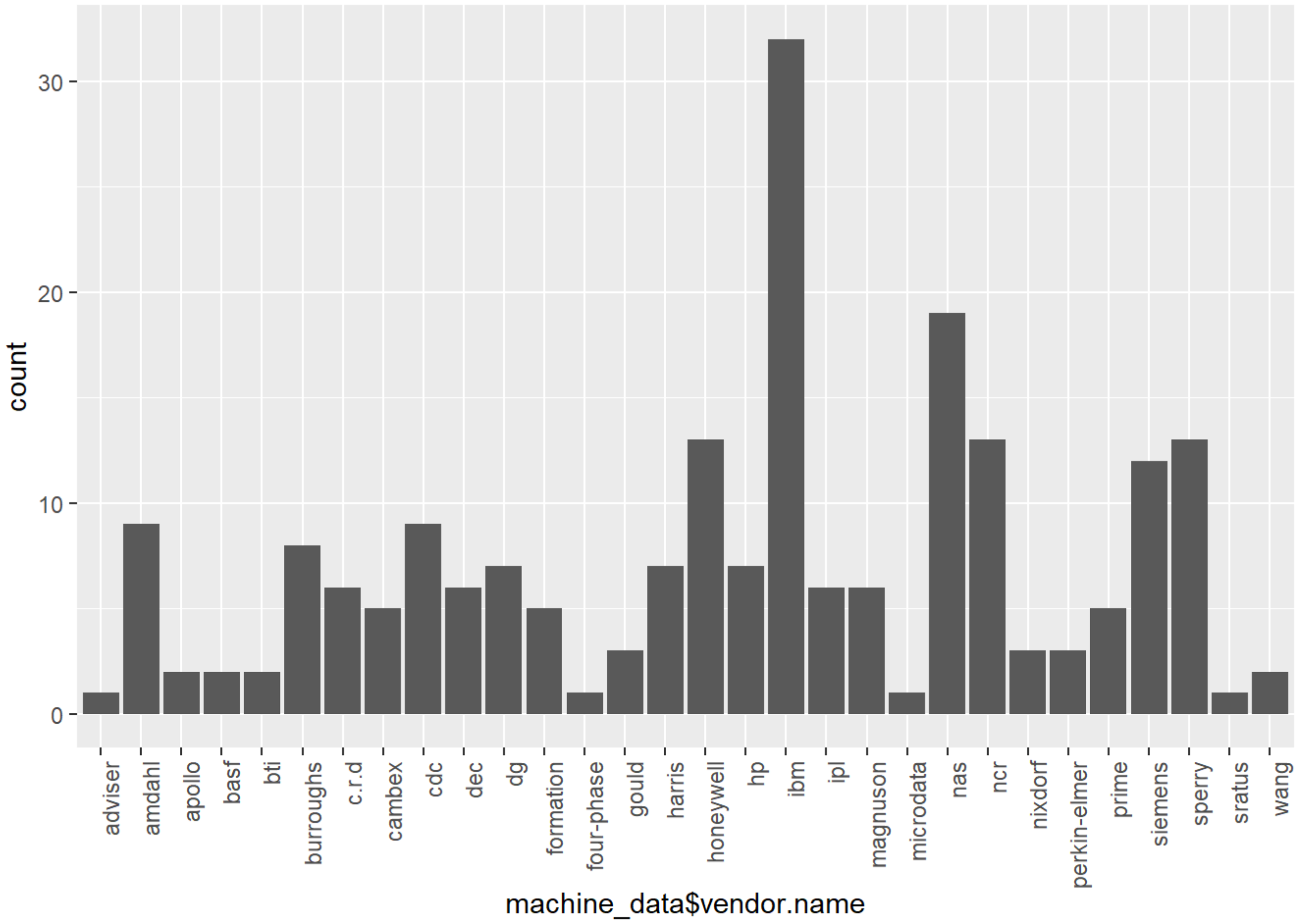
2. Missing values

```
cols_with_missing_names <- colnames(machine_data)[apply(machine_data, MARGIN = 2, function(a)
any(is.na(a)))]
NameList <- cols_with_missing_names
idx <- match(NameList, names(machine_data))
kable(colSums(is.na(machine_data[,c(idx)])), row.names = NA, col.names = 'missing count')
```

missing count

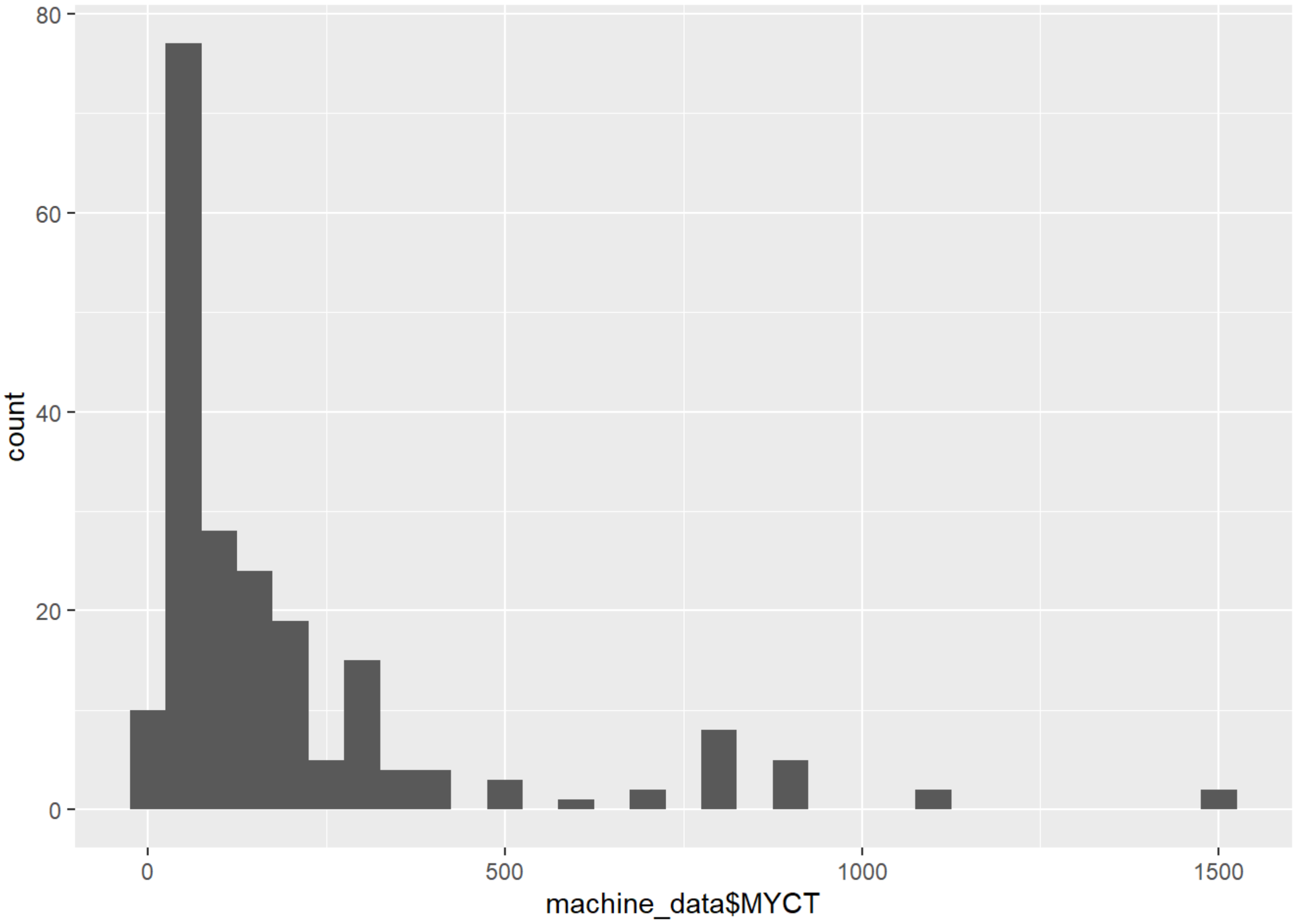
3. Vendors histogram

```
ggplot(data.frame(machine_data$vendor.name), aes(x=machine_data$vendor.name)) +
  geom_bar() + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



4. MYCT chart

```
ggplot(data.frame(machine_data$MYCT), aes(x=machine_data$MYCT)) +
  geom_histogram(binwidth = 50)
```



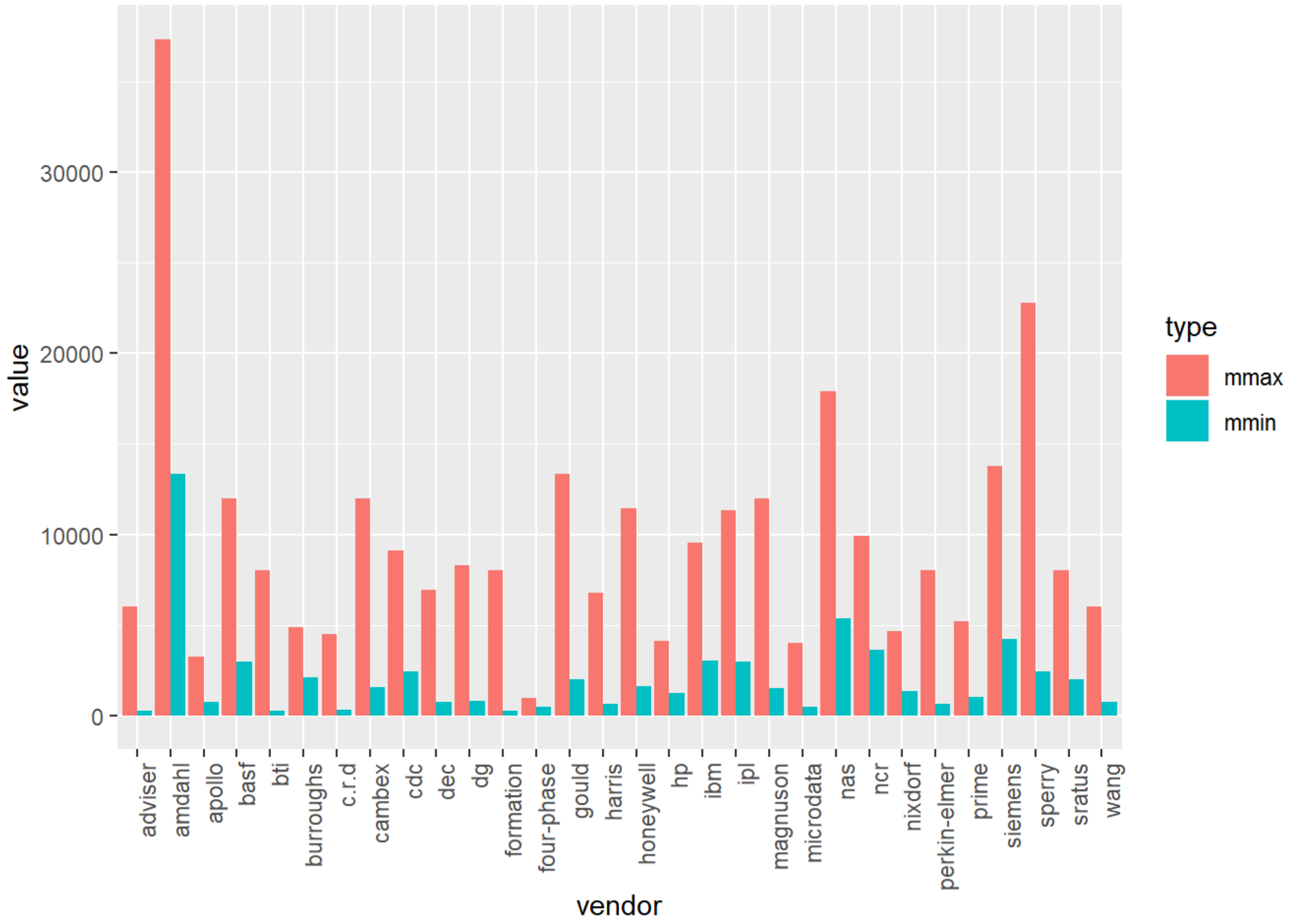
5. MMIN MMAX histogram

```
chart_data_max <- aggregate(machine_data$MMAX, list(machine_data$vendor.name), mean)
colnames(chart_data_max) <- c('vendor', 'value')
chart_data_max$type <- 'mmax'

chart_data_min <- aggregate(machine_data$MMIN, list(machine_data$vendor.name), mean)
colnames(chart_data_min) <- c('vendor', 'value')
chart_data_min$type <- 'mmin'

chart_data <- rbind(chart_data_max, chart_data_min)

ggplot(chart_data, aes(fill=type, y=value, x=vendor)) +  geom_bar(position="dodge", stat="ide
ntity") + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



6. Table CHMAX gt 12 distibution

```
chmax_above_12 <- machine_data[which(machine_data$CHMAX>12),]
total <- NROW(chmax_above_12)
frquencies <- aggregate(chmax_above_12$vendor.name, list(chmax_above_12$vendor.name), FUN = N
ROW)
colnames(frquencies) <- c('vendor', 'count')
frquencies$frequency <- frquencies$count * 100 / total
kable(frquencies)
```

vendor	count	frequency
adviser	1	1.162791
amdahl	9	10.465116
bti	1	1.162791
burroughs	8	9.302326
cdc	6	6.976744
four-phase	1	1.162791
gould	3	3.488372
harris	7	8.139535
honeywell	8	9.302326
hp	5	5.813954
ibm	5	5.813954
magnuson	3	3.488372
microdata	1	1.162791
nas	8	9.302326
ncr	3	3.488372
prime	3	3.488372

vendor	count	frequency
siemens	6	6.976744
sperry	7	8.139535
sratus	1	1.162791

7. Companies with CHMIN lt 16

```
chim <- machine_data[which(machine_data$CHMIN<16),]
length(chim)
```

```
## [1] 10
```

```
kable(data.frame(with(chim, table(vendor.name))))
```

vendor.name	Freq
adviser	0
amdahl	5
apollo	2
basf	2
bti	2
burroughs	8
c.r.d	6
cambex	5
cdc	9
dec	6
dg	7
formation	5
four-phase	1
gould	3
harris	7
honeywell	13
hp	7
ibm	31
ipl	6
magnuson	6
microdata	1
nas	17
ncr	10
nixdorf	3
perkin-elmer	3
prime	5
siemens	11
sperry	10
sratus	1
wang	2

8. ERP distribution for top 4 vendors (by model count)

```
top_4 <- data.frame(sort(table(machine_data$vendor.name),decreasing=TRUE)[1:4])
kable(top_4)
```

Var1	Freq
ibm	32
nas	19
honeywell	13
ncr	13

```
companies <- top_4$Var1
subset <- subset(machine_data, machine_data$vendor.name %in% companies)
total <- NROW(subset)
subset <- aggregate(subset$vendor.name, list(subset$vendor.name, subset$ERP), FUN = NROW)
colnames(subset) <- c('vendor', 'ERP', 'Count')
subset$frequency <- subset$Count * 100 / total
kable(subset)
```

vendor	ERP	Count	frequency
ibm	101	1	1.298701
nas	107	1	1.298701
ibm	113	1	1.298701
ibm	116	1	1.298701
nas	117	1	1.298701
nas	119	1	1.298701
nas	120	1	1.298701
nas	126	1	1.298701
ncr	142	1	1.298701
ibm	15	1	1.298701
nas	151	1	1.298701
ibm	17	1	1.298701
ibm	171	1	1.298701
honeywell	175	1	1.298701
ibm	18	3	3.896104
honeywell	181	2	2.597403
ncr	19	1	1.298701
ncr	190	1	1.298701
honeywell	20	1	1.298701
ibm	20	4	5.194805
ibm	21	1	1.298701
ncr	21	1	1.298701
ibm	220	1	1.298701
honeywell	23	1	1.298701
honeywell	25	1	1.298701
ibm	26	2	2.597403
ncr	26	1	1.298701
nas	266	1	1.298701
nas	267	1	1.298701
nas	270	1	1.298701
honeywell	28	1	1.298701
ibm	28	3	3.896104
ncr	281	1	1.298701
honeywell	29	1	1.298701
nas	29	1	1.298701

vendor	ERP	Count	frequency
honeywell	30	1	1.298701
ibm	31	2	2.597403
honeywell	32	2	2.597403
ibm	35	1	1.298701
ncr	35	1	1.298701
ibm	350	1	1.298701
ibm	361	1	1.298701
nas	41	1	1.298701
ncr	41	1	1.298701
ibm	42	1	1.298701
nas	426	1	1.298701
ibm	45	1	1.298701
nas	46	1	1.298701
ncr	47	1	1.298701
nas	48	1	1.298701
nas	53	2	2.597403
honeywell	57	1	1.298701
ibm	59	1	1.298701
nas	603	1	1.298701
ncr	62	1	1.298701
ibm	65	1	1.298701
honeywell	73	1	1.298701
ibm	76	2	2.597403
ncr	78	1	1.298701
ncr	80	2	2.597403
ibm	82	1	1.298701
nas	86	1	1.298701
nas	95	1	1.298701