Stat 537: Homework 3

Brandon Fenton and Kenny Flagg

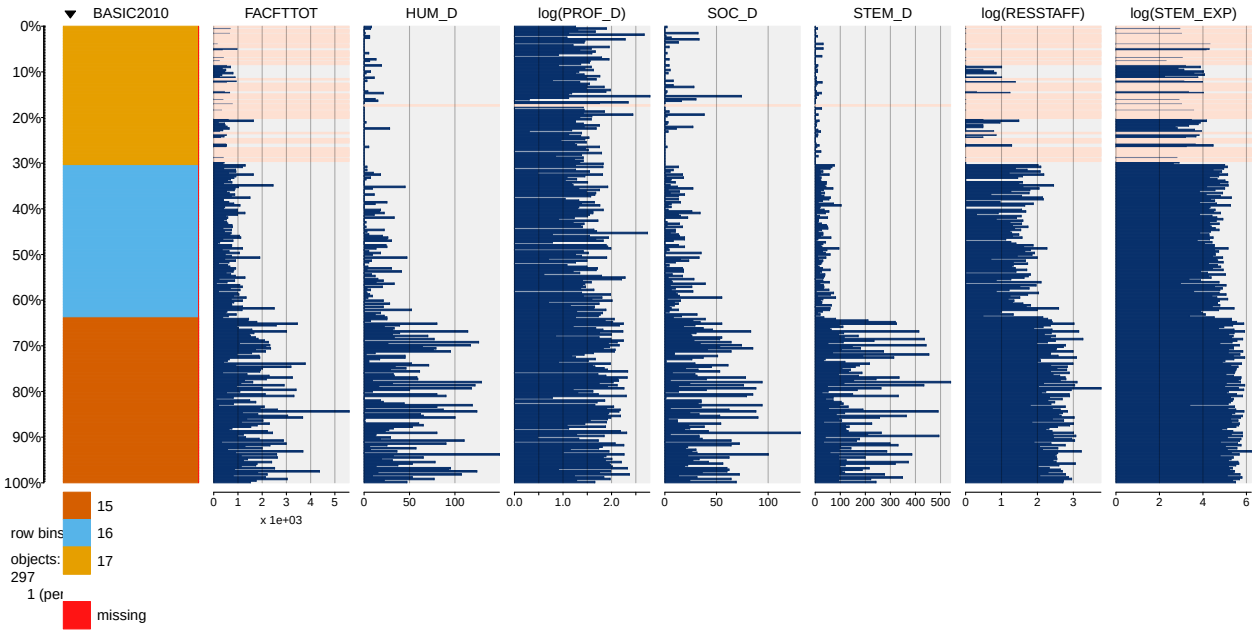
Due Friday, Feb 5 at end of day

1. *Summarize the missing data patterns. How many of each kind are present? Ignoring the name of the institutions, how many totals cells are missing out of the N=297 by Q=8 matrix of information?*

	NAME	BASIC2010	HUM_D	PROF_D	SOC_D	STEM_D	FACFTTOT	RESSTAFF	STEM_EXP	
248	1	1	1	1	1	1	1	1	1	0
48	1	1	1	1	1	1	0	0	0	3
1	1	1	0	0	0	0	0	0	0	7
	0	0	1	1	1	1	49	49	49	151

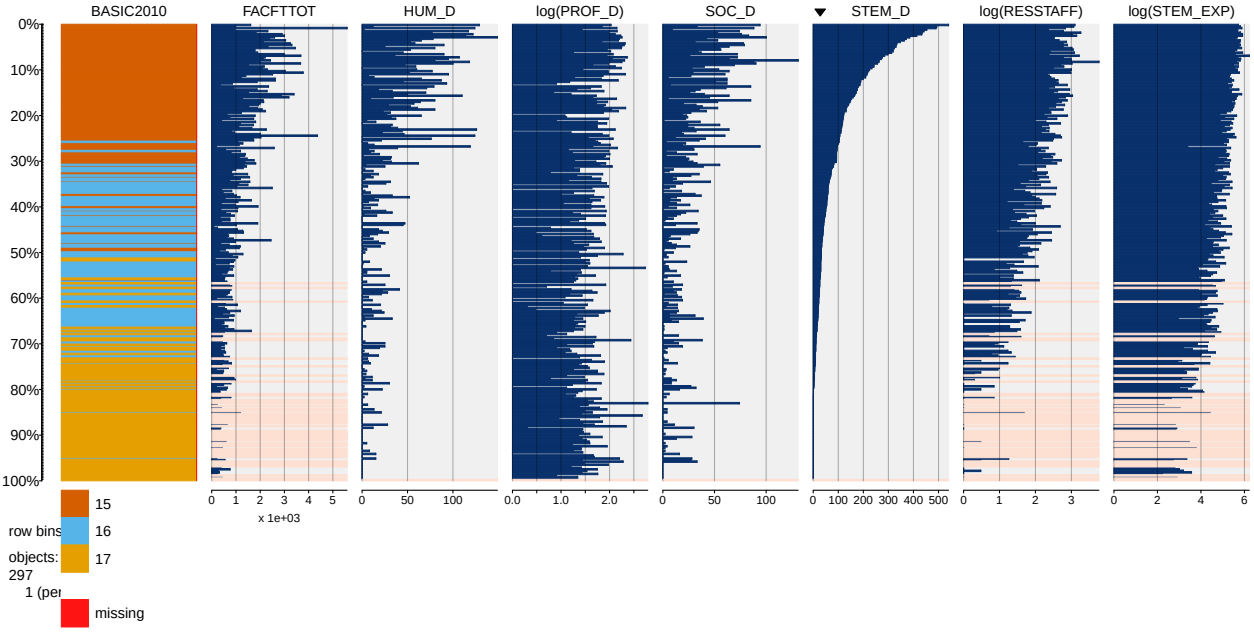
There are 248 institutions with no missing cells. 48 institutions are missing the number of full-time faculty, number of non-faculty research staff, and amount of STEM research expense. One institution is missing all variables except for the 2010 Basic Classification. In total, 151 of the 2376 cells are missing.

2. *Make a tableplot of the data set cc2010Ps except the university names. It should be sorted based on the classification codes that Carnegie created. Describe the differences in the responses based on the tableplot results.*



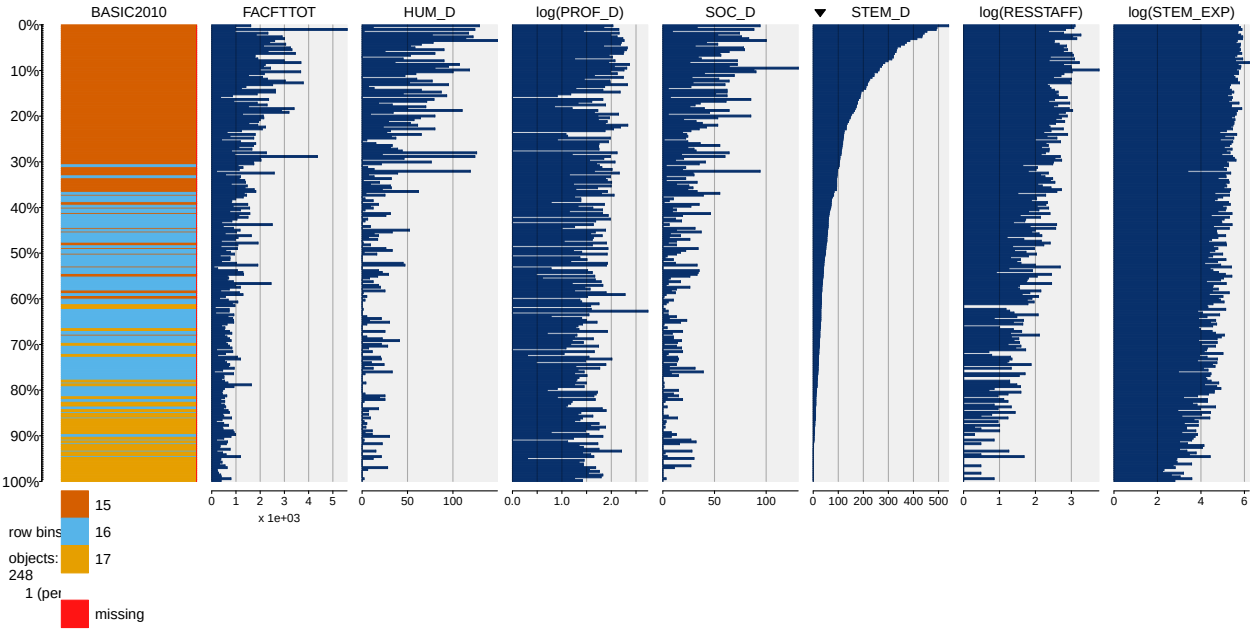
Institutions with classification 15 tend to have higher numbers of faculty members and research staff, have more research doctorates, and spend more on STEM research than institutions with classifications 16 and 17. All of the institutions with missing responses have classification 17.

3. *Now consider re-sorting the observations based on other variables. You can use itableplot() or just re-sort manually using the sortCol= in the function call. For example, try sortCol=2 to sort based on total number of faculty. What does this show you about the missing observations? Does this support a MCAR assumption? Just use the plot to provide a short discussion.*

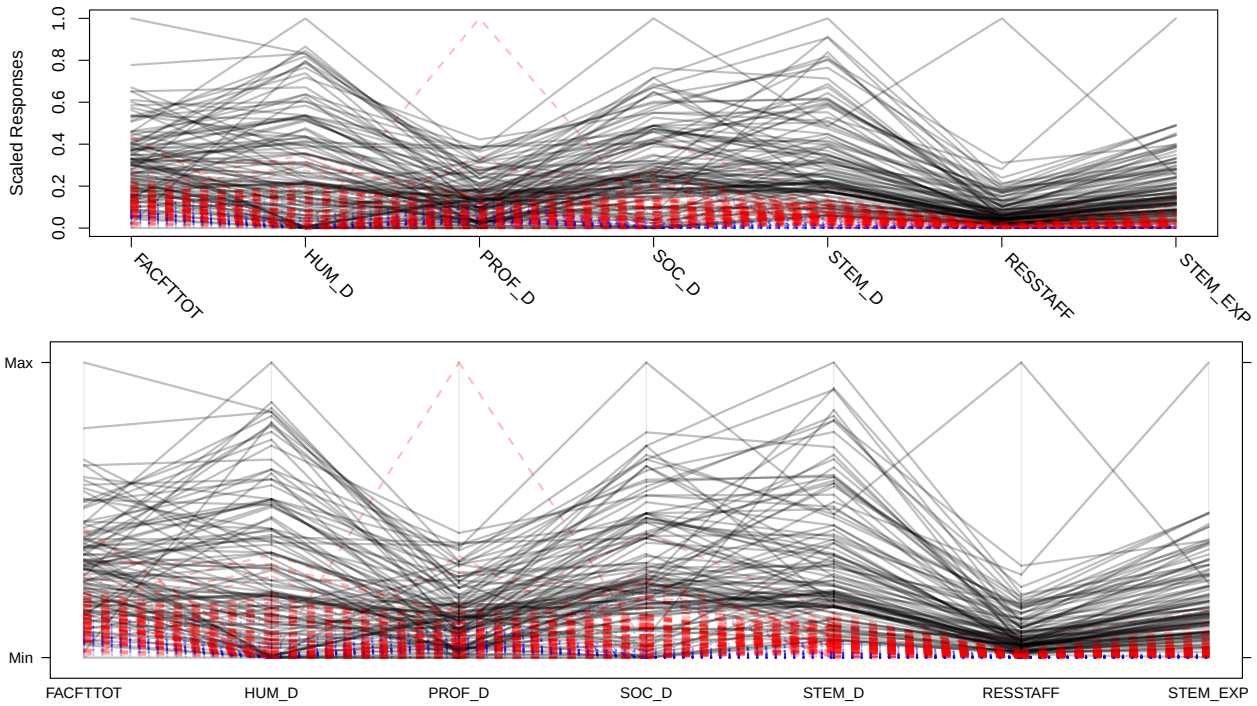


After sorting by number of STEM research doctorates, it can be seen that missingness is associated with low numbers values of that variable. This does not support a missing-completely-at-random asspumtion.

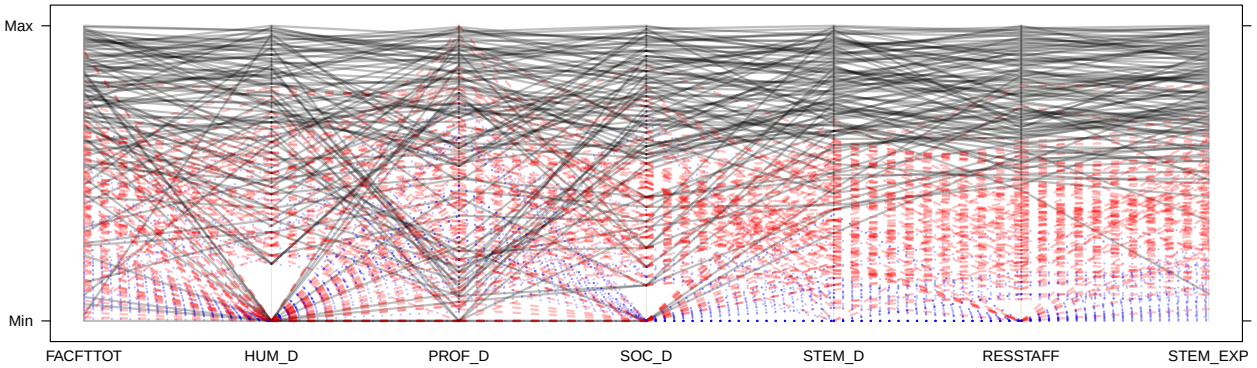
4. *Before proceeding further, clean out any universities that have any missing values in this data set. Remake a tableplot that is sorted based on one of the quantitative variables.*



5. *Make two parallel coordinate plots, one using matplot and one using parallelplot from lattice. In each, color the lines based on the BASIC2010 variable. I like to leave the coloring categorical variable in PCP plots which often alleviates the need to add a legend especially with an ordered categorical variable (if you know what direction it went). Make sure in each that I can read all the variable names in the versions you submit.*



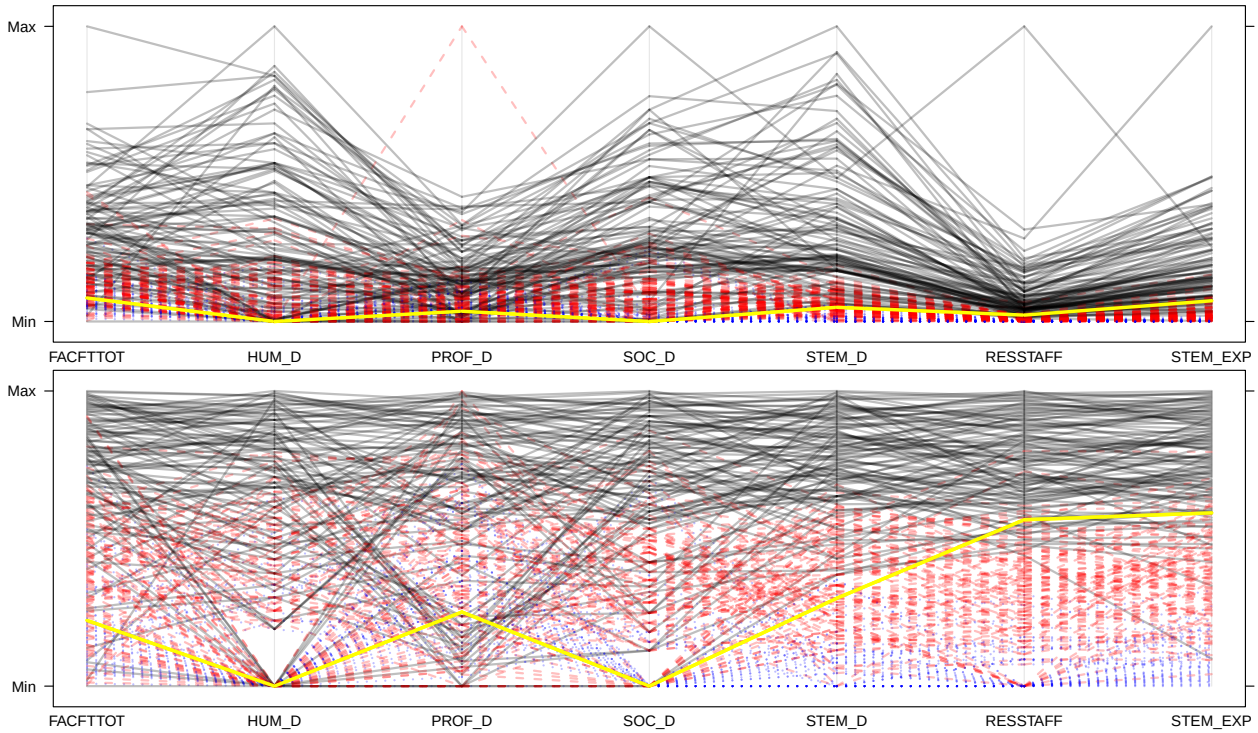
6. *The actual Carnegie rankings are based on initially ranking each quantitative variable to spread out the differences at small values and decrease the impacts of outliers. Use the rank function to generate a new data set that replaces the ranks on each of the quantitative variables. Remake your favorite PCP plot and discuss how different the classifications look in this plot.*



In this plot we are seeing the ordering of the variables without taking into account their magnitude. This method confers a number of advantages. As stated in the question this will increase small differences and decrease large differences, and this makes certain relationships between the observations in each category more apparent. The rank-based plot shows that there is relatively little overlap between the schools in categories 15 and 16 with respect to all variables except for **HUM_D** and **PROF_D**; it is also far more apparent in this plot that many schools have no research doctorates in the humanities or professional fields. The rank-based plot also makes more efficient use of the available vertical space, resulting in, for example, a much clearer picture of where the rankings for category 17 schools lie. However, there might be cases in which the magnitude of a particular value would be relevant, so either form of plot could be appropriate depending on the circumstances.

7. *Find Montana State in the data set and highlight its observation in one of your PCP plots, making plots of the ranked and the original version of the data set. I like to create color and line width id vectors to pass to col and lwd to enhance the plot with other information when I want to change the color or line*

width for just an observation or two but you are welcome to tackle this in other ways (matplot allows a second plot to be added using add=T in the second plot call). Keep the basic three color groups and try to make MSU identifiable. Characterize MSU relative to the other institutions on these variables.



In terms of the number of research staff and STEM research expenditure, Montana State lies in a region of overlap between category 15 institutions and category 16 institutions. However, MSU has a much lower number of faculty than most category 15 institutions and has no research doctorates in the humanities and social sciences. In the plot on the original scale, it is seen that several category 15 institutions also have 0 humanities research doctorates, but they have much larger numbers of faculty and professional research doctorates than does MSU.

R Code Appendix:

Problem 1:

```
require(pander)
require(mice)

#cc2010 <- read.csv("https://montana.box.com/shared/static/hlv178y436vggzmsxxzsdkoarfkopkvp.csv",
#                   header = TRUE)
#cc2010_PhD <- cc2010[(cc2010$BASIC2010>14 & cc2010$BASIC2010 < 18),]
#cc2010_PhD$BASIC2010 <- factor(cc2010_PhD$BASIC2010)
#cc2010Ps <- cc2010_PhD[,c("NAME", "BASIC2010", "FACFTTOT", "HUM_D", "PROF_D", "SOC_D",
#                          "STEM_D", "RESSTAFF", "STEM_EXP")]
#save(cc2010Ps, file = "cc2010Ps.RData")
load("cc2010Ps.RData")

cc2010.md <- md.pattern(cc2010Ps)
pander(cc2010.md, split.table = Inf)
```

Problem 2:

```
require(tabplot)

# Name is first column, BASIC2010 is first of remaining columns
tableplot(cc2010Ps[-1], sortCol = 1, nBins = 297)
```

Problem 3:

```
tableplot(cc2010Ps[-1], sortCol = 6, nBins = 297)
```

Problem 4:

```
cc2010.clean <- na.omit(cc2010Ps)

tableplot(cc2010.clean[-1], sortCol = 6, nBins = 297)
```

Problem 5:

```
scale01 <- function(vec){
  vec1 <- vec[!is.na(vec)]
  (vec-min(vec1))/(max(vec1) - min(vec1))
}

cc.scaled <- sapply(cc2010.clean[,-c(1,2)], scale01)
colors <- ifelse(cc2010.clean$BASIC2010==15, "#00000040",
                 ifelse(cc2010.clean$BASIC2010==16, "#ff000040",
                        "#0000ff40"))
matplot(t(cc.scaled), ylab = "Scaled Responses", xaxt="n", type="l", col=colors,
        lty = as.numeric(cc2010.clean$BASIC2010), lwd=2)
tics <- 1:15
xlabs <- names(cc2010.clean[,-c(1,2)])
tck <- axis(1, at=tics, labels=FALSE)
text(tck, par("usr")[3], labels=xlabs, srt=315, xpd=TRUE, adj=c(-0.2,1.2, cex=0.9))
```

```
require(lattice)
parallelplot(cc2010.clean[,-c(1,2)], col = colors,
             lty = as.numeric(cc2010.clean$BASIC2010), horizontal=FALSE, lwd=2)
```

Problem 6:

```
cc2010.ranks <- cc2010.clean

# Rank each column except the first two (name and class)
cc2010.ranks[,-(1:2)] <- apply(cc2010.ranks[,-(1:2)], 2, rank)
parallelplot(cc2010.ranks[,-c(1,2)], col = colors,
             lty = as.numeric(cc2010.clean$BASIC2010), horizontal=FALSE, lwd=2)
```

Problem 7:

```
# Create new data frames with an extra MSU row at the end
msu <- rbind(cc2010.clean,
             cc2010.clean[cc2010.clean$NAME=="Montana State University",])
msu.ranks <- rbind(cc2010.ranks,
                   cc2010.ranks[cc2010.ranks$NAME=="Montana State University",])
colors[msu$NAME=="Montana State University"] <- "#ffff00ff"
parallelplot(msu[,-c(1,2)], col = colors,
             lwd = 2+as.numeric(msu$NAME=="Montana State University"),
             lty = as.numeric(msu$BASIC2010), horizontal=FALSE)
parallelplot(msu.ranks[,-c(1,2)], col = colors,
             lwd = 2+as.numeric(msu.ranks$NAME=="Montana State University"),
             lty = as.numeric(msu.ranks$BASIC2010), horizontal=FALSE)
```

About This Markdown File

- File creation date: 2016-02-02
- R version 3.2.3 (2015-12-10)
- R version (short form): 3.2.3
- Additional session information

```
## R version 3.2.3 (2015-12-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.3 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] lattice_0.20-33 tabplot_1.1     ffbase_0.12.1   ff_2.2-13
## [5] bit_1.1-12      mice_2.25       Rcpp_0.11.5     pander_0.6.0
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.8    MASS_7.3-44     grid_3.2.3      formatR_1.1
##  [5] evaluate_0.8    rpart_4.1-10    fastmatch_1.0-4 rmarkdown_0.9.2
##  [9] splines_3.2.3   tools_3.2.3     stringr_0.6.2   survival_2.38-3
## [13] yaml_2.1.13     htmltools_0.3   knitr_1.12.3    nnet_7.3-11
```