# HW5

Tanner Glass

10/27/2020

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Exploratory Data Analysis and plotting. To begin the homework, we will as usual, start by loading, munging and creating tidy data sets. In this homework, our goal is to create informative (and perhaps pretty) plots showing features or perhaps deficiencies in the data.

## Problem 1

Work through the Swirl "Exploratory_Data_Analysis" lesson parts 1 - 10. If you need some review of ggplot, see the tutorial on Rstudio.cloud.

Nothing to submit.

## Problem 2

Create a new R Markdown file within your local GitHub repo folder (file–>new–>R Markdown–>save as).

The filename should be: HW5_lastname, i.e. for me it would be HW5_Settlage

You will use this new R Markdown file to solve the following problems.

## Problem 3

Using tidy concepts, get and clean the following data on education from the World Bank.

http://databank.worldbank.org/data/download/Edstats_csv.zip

How many data points were there in the complete dataset? In your cleaned dataset?

Choosing 2 countries, create a summary table of indicators for comparison.

```
rm(list = ls())
#library(readr)
#EdStatsData <- read_csv("EdStatsData.csv")
#EdStatsData<-saveRDS(EdStatsData,"EdStatsData.RDS")
EdStatsData<-readRDS("EdStatsData.RDS")
#View(EdStatsData)
```

We have successfully read the data set. We observe that there are many missing data points in the set as not every country for every given year is going to be able to produce the result suggested by the indicator value. In total, there are 886,930 entries that include NA's and with all the indicator and country values. In our cleaning. Our objective will be to create a data set that is clean that has all the observation that are able to be used. We will make each indicator a variable and the subsequent years will be a result of an

observation. observation 1 will be the year 1970 and the last observation for a given indicator will be 2020, the last supported year in the data set. Country will remain a factor.

```
educationaldf<-data.frame(c(rep(0,11858)))
for (i in 1:3665){
  allyears<-c()
  for (j in 1:242){
    allyears<-c(allyears,as.double(EdStatsData[(j-1)*3665+i,5:53]))
  }
  assign(EdStatsData$`Indicator Name`[i],allyears)
  educationaldf<-cbind(educationaldf,eval(as.name(EdStatsData$`Indicator Name`[i])))

}

educationaldf<-saveRDS(educationaldf,"educationaldf.RDS")
```

```
educationaldf<-readRDS("educationaldf.RDS")
educationaldf<-educationaldf[,-1]
names(educationaldf)<-EdStatsData$`Indicator Name`[1:3665]

educationaldf<-saveRDS(educationaldf,"educationaldf.RDS")
```

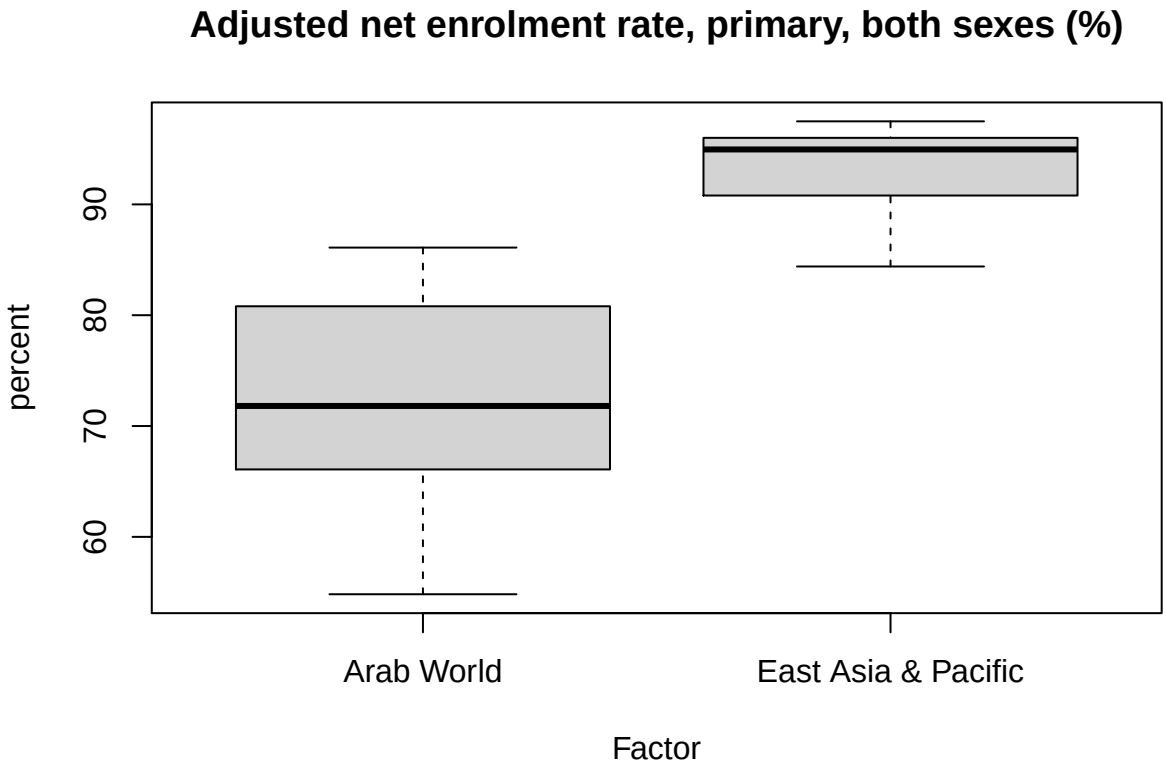
```
educationaldf<-readRDS("educationaldf.RDS")

country<-c()
for (i in 1:242){
country<-c(country,rep(EdStatsData$`Country Name`[(i-1)*3665+1],49))
}
educationaldf<-cbind(country,educationaldf)
educationaldf<-saveRDS(educationaldf,"educationaldf.RDS")
```

We have finally achieved our goal. Rstudio cloud does not like this large of a data set to work with. To reduce crashing of the workspace, we save as an object and now work down. The presented code in the document provides how i cleaned the data set. Upon viewing the data set, it is set to have data entries for each of the indicator and is factored by country. Now, for the plotting and comparison. For the two countries we chose, we chose to have Arab world and East Asia and Pacific. The instruction state to use all indicators. However, we due to limitations of rcloud, i will not be able to achieve for all indicators. I have shown below in code how i can summarize indicators with my newfound data set. To prevent further crashes we subset the data set into the necessary components and then process as such. We summarize between the indicator known as enrollment rate for primary schools in both sexes. We can simply pick any indicator and continue this process to study the data further.

```
educationaldf<-readRDS("educationaldf.RDS")
p3df<-cbind(educationaldf[1:97,1],educationaldf[1:97,6])
p3df<-saveRDS(p3df,"p3df.RDS")
```

```
p3df<-readRDS("p3df.RDS")
boxplot(as.double(p3df[,2]) ~ as.factor(p3df[,1]),data = p3df,na.rm=TRUE,main="Adjusted net enrolment ra
```




We can see from this graph that there is a higher enrollment rate in East Asia and Pacific. We also can note the variability is lower. It however has a skew of the left which is similar to that of the arab world plot.

The cleaned data set has a total of 3665 variables with 1 addional variable being the factor which is country. Each indicator has 11858 observations.

## Problem 4

Using *base* plotting functions, create a single figure that is composed of the first two rows of plots from SAS's simple linear regression diagnostics as shown here: https://support.sas.com/rnd/app/ODSGraphics/examples/reg.html. Demonstrate the plot using suitable data from problem 3.

We are tasked with creating a variety of plots including: residuals vs. predicted. studentized residuals vs. predicted, studentized residuals vs. leverage, residual vs. quantile, population vs quantile, and cooks distance. We find that the data we are given is time oriented where each entry per given country is chronologically ordered by years. We use a set of data from our previous data from. We will look at data from Arab world enrollment rate into primary schools through the years 1970-2020. Note that are not entries from 2015-2020 so we will throw those out.

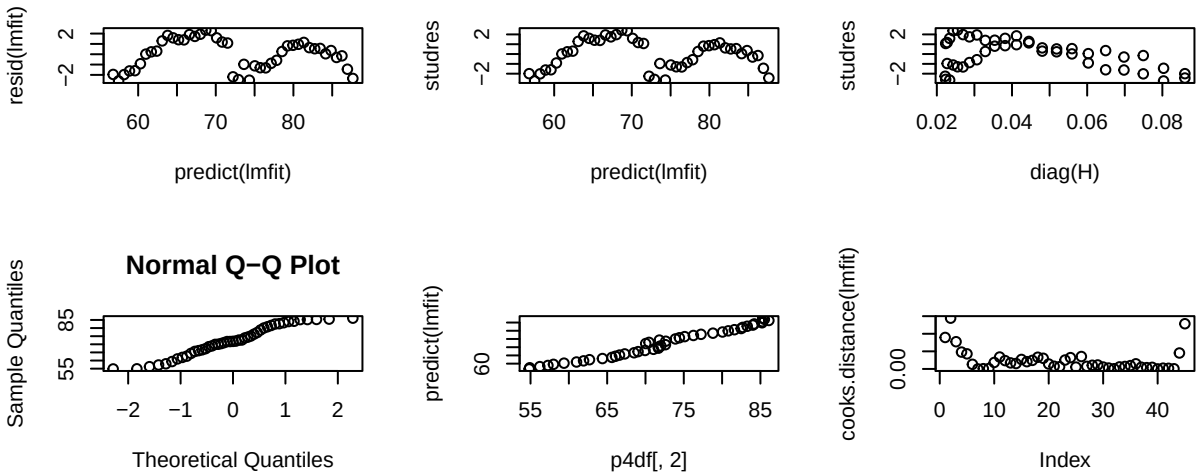
```
p4df<-cbind(c(1970:2014),p3df[1:45,2])

X<-cbind(rep(1,45),c(1970:2014))
H<-X%*%solve(t(X)%*%X)%*%t(X)


lmfit<-lm(as.numeric(p4df[,2])~as.numeric(p4df[,1]))
studres<-resid(lmfit)*sqrt(43/(45-resid(lmfit)^2))


par(mfrow=c(3,3))
plot(resid(lmfit)~predict(lmfit))
```

```
plot(studres~predict(lmfit))
plot(studres~diag(H))
qqnorm(as.numeric(p4df[,2]))
plot(p4df[,2],predict(lmfit))
plot(cooks.distance(lmfit))
```




With only using those functions provided in Base R. We have recreated the plots of the first two rows of the SAS functions.

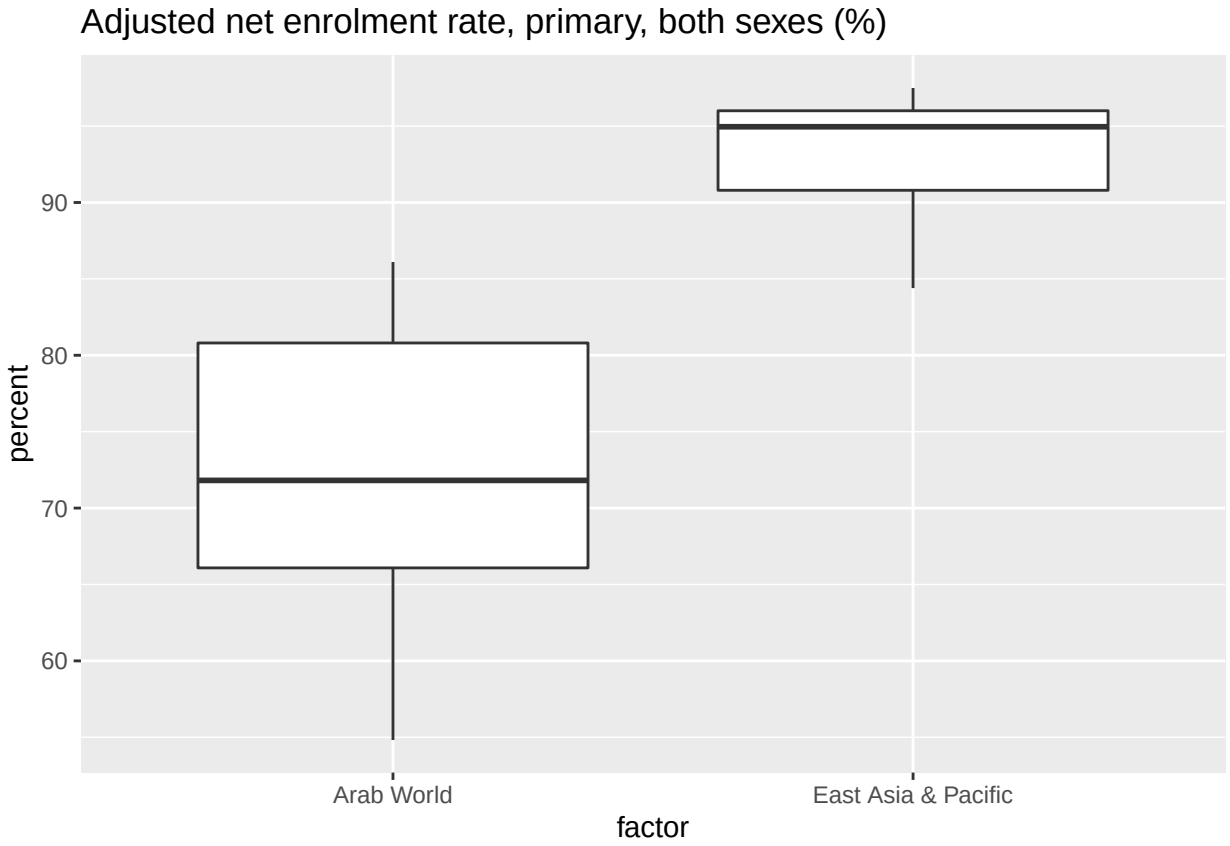
## Problem 5

Recreate the plot in problem 3 using ggplot2 functions. Note: there are many extension libraries for ggplot, you will probably find an extension to the ggplot2 functionality will do exactly what you want.

We aim to recreate the box plot made in problem 3. Note, that we can continue the same code to mimic ALL of the indicators given that Rstudio would cooperate.

```
#install.packages("ggplot2")
library("ggplot2")
p3df<-as.data.frame(p3df)
colnames(p3df)<-c("x","y")
ggplot(p3df, aes(x=as.factor(x) , y=as.numeric(y))) +
    geom_boxplot() +
  labs(title = "Adjusted net enrolment rate, primary, both sexes (%)",x="factor",y="percent")
```

```
## Warning: Removed 13 rows containing non-finite values (stat_boxplot).
```




The plot has been recreated using ggplot2. We successfully answered the problem.

## Problem 6

Finish this homework by pushing your changes to your repo.

**Only submit the .Rmd and .pdf solution files. Names should be formatted HW5_lastname_firstname.Rmd and HW5_lastname_firstname.pdf**