Homework 8

Due Wednesday Nov 13, 2019

2019-11-09

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.

```
swirl::install_course("Exploratory Data Analysis")
swirl() # Did up to Part 5
```

Problem 2

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

The filename should be: HW8_lastname, i.e. for me it would be HW8_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?

```
setwd("/Users/samsung/Desktop/STAT_5014")

edstats <- read.csv("EdStatsData.csv", header = T)
names(edstats)[1] <- "Country.Name"
names(edstats)[-(1:4)] <- substr(names(edstats)[-(1:4)], 2, 5)
# There were a total of 886,930 x 66 = 58,537,380 observations

edstats.data <- as.vector(t(edstats[,-(1:4)]))
na.values <- which(is.na(edstats.data))
edstats.data <- edstats.data[-na.values]
```

```
edstats.names <- edstats[,(1:4)]
rows <- rep(1:nrow(edstats.names), each = length(names(edstats)[-(1:4)]))
edstats.names <- edstats.names[rows,]
edstats.names <- edstats.names[-na.values, ]
```

```
edstats.years <- names(edstats[,-(1:4)])
edstats.years <- rep(edstats.years, nrow(edstats))[-na.values]
```

```
edstats.df <- cbind.data.frame(edstats.names, Years = edstats.years, Data = edstats.data)
dim(edstats.df)
```

```
## [1] 5082201       6
```

```
# There were a total of 5,082,201 data points.
```

There were a total of 58,537,380 data points including NAs in the original data set. This was reduced to 5,082,201 data points after NAs were removed.

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

```
# Tibble
edstats.df <- as_tibble(edstats.df)

edstats.df.slv <- edstats.df %>%
  filter(Country.Name == "El Salvador")

edstats.df.kor <- edstats.df %>%
  filter(Country.Name == "Korea, Rep.")

edstats.df.slv <- edstats.df.slv[, -c(2, 4)]
edstats.df.kor <- edstats.df.kor[, -c(2, 4)]

edstats.df.both <- left_join(edstats.df.slv, edstats.df.kor,
                             by = c("Indicator.Name", "Years"))
edstats.df.both
```

```
## # A tibble: 31,426 x 6
##    Country.Name.x Indicator.Name         Years Data.x Country.Name.y Data.y
##    <fct>          <fct>                  <fct>  <dbl> <fct>           <dbl>
##  1 El Salvador    Adjusted net enrolmen~ 2000    48.7 Korea, Rep.      96.2
##  2 El Salvador    Adjusted net enrolmen~ 2001    50.5 Korea, Rep.      95.6
##  3 El Salvador    Adjusted net enrolmen~ 2002    56.9 Korea, Rep.      91.8
##  4 El Salvador    Adjusted net enrolmen~ 2005    62.3 Korea, Rep.      96.3
##  5 El Salvador    Adjusted net enrolmen~ 2006    57.1 Korea, Rep.      96.9
##  6 El Salvador    Adjusted net enrolmen~ 2007    58.6 Korea, Rep.      96.4
##  7 El Salvador    Adjusted net enrolmen~ 2008    61.1 Korea, Rep.      95.8
##  8 El Salvador    Adjusted net enrolmen~ 2009    64.3 Korea, Rep.      95.5
##  9 El Salvador    Adjusted net enrolmen~ 2010    66.9 Korea, Rep.      96.9
## 10 El Salvador    Adjusted net enrolmen~ 2011    68.3 Korea, Rep.      94.9
## # ... with 31,416 more rows
```

I chose El Salvador and Korea, Rep.

There were many indicators. A small subset is seen above.

Problem 3

Using base plotting functions, recreate the scatter plot shown in class with histograms in the margins. You do not have to make the plot the same, just have a scatter plot with marginal histograms. Demonstrate the plot using suitable data from problem 2.

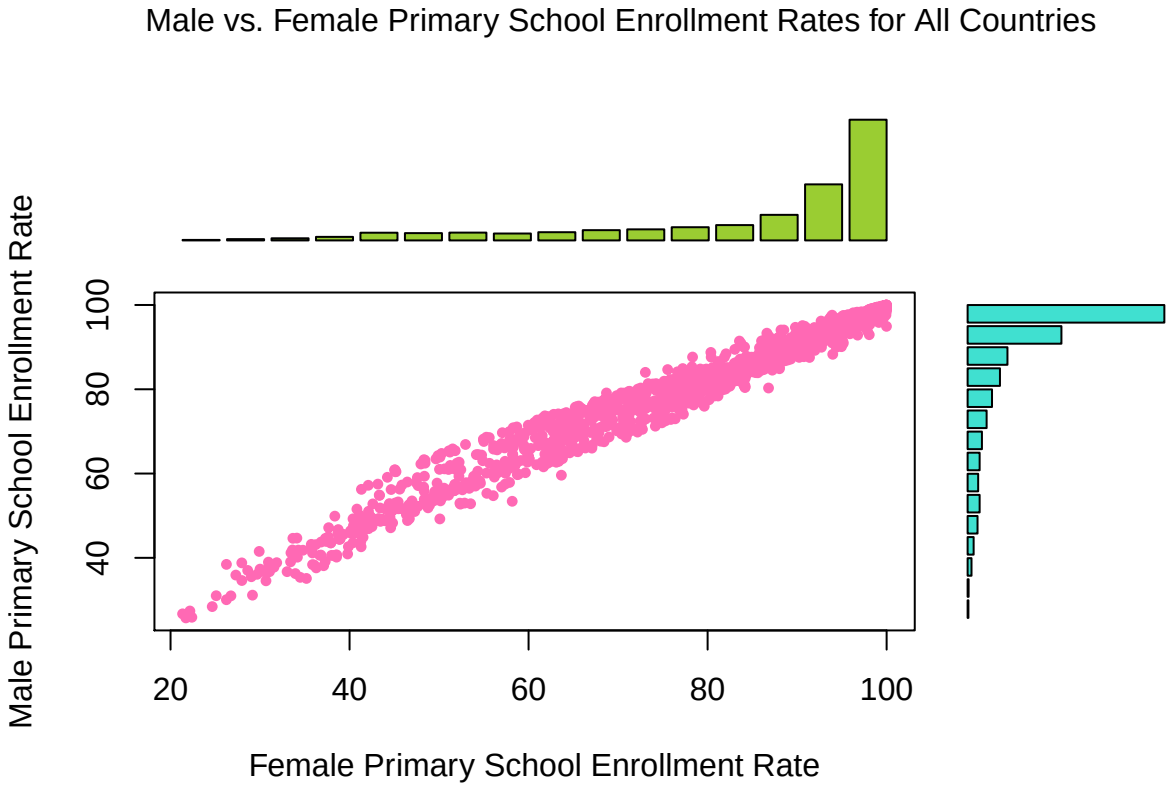
```
edstats.df.se.prm.tenr <- edstats.df %>%
  filter(Indicator.Code == "SE.PRM.TENR")
edstats.df.se.prm.tenr.fe <- edstats.df %>%
  filter(Indicator.Code == "SE.PRM.TENR.FE")
edstats.df.se.prm.tenr.both <- inner_join(edstats.df.se.prm.tenr.fe, edstats.df.se.prm.tenr,
```

```
                                        by = c("Country.Name", "Country.Code", "Years"))
names(edstats.df.se.prm.tenr.both)[6] <- "Female"
names(edstats.df.se.prm.tenr.both)[9] <- "Male"
```

```
# Add histograms to a scatterplot
par(fig = c(0, 0.8, 0, 0.8), new=TRUE)
```

```
## Warning in par(fig = c(0, 0.8, 0, 0.8), new = TRUE): calling par(new=TRUE)
## with no plot
```

```
with(edstats.df.se.prm.tenr.both, plot(Female, Male, pch = 16, cex = 0.75,
                                       xlab = "Female Primary School Enrollment Rate",
                                       ylab = "Male Primary School Enrollment Rate",
                                       col = "hotpink"))
par(fig = c(0, 0.8, 0.45, 1), new=TRUE)
barplot(hist.female$density, axes = FALSE, col = "yellowgreen")
lines(hist.female$mids, hist.female$density, lwd = 2, col = "navyblue")
par(fig = c(0.65, 1, 0, 0.8), new=TRUE)
barplot(hist.male$density, axes = FALSE, horiz = TRUE, col = "turquoise")
lines(hist.male$mids, hist.male$density, lwd = 2, col = "navyblue")
mtext("Male vs. Female Primary School Enrollment Rates for All Countries",
      side = 3, outer=TRUE, line = -2)
```

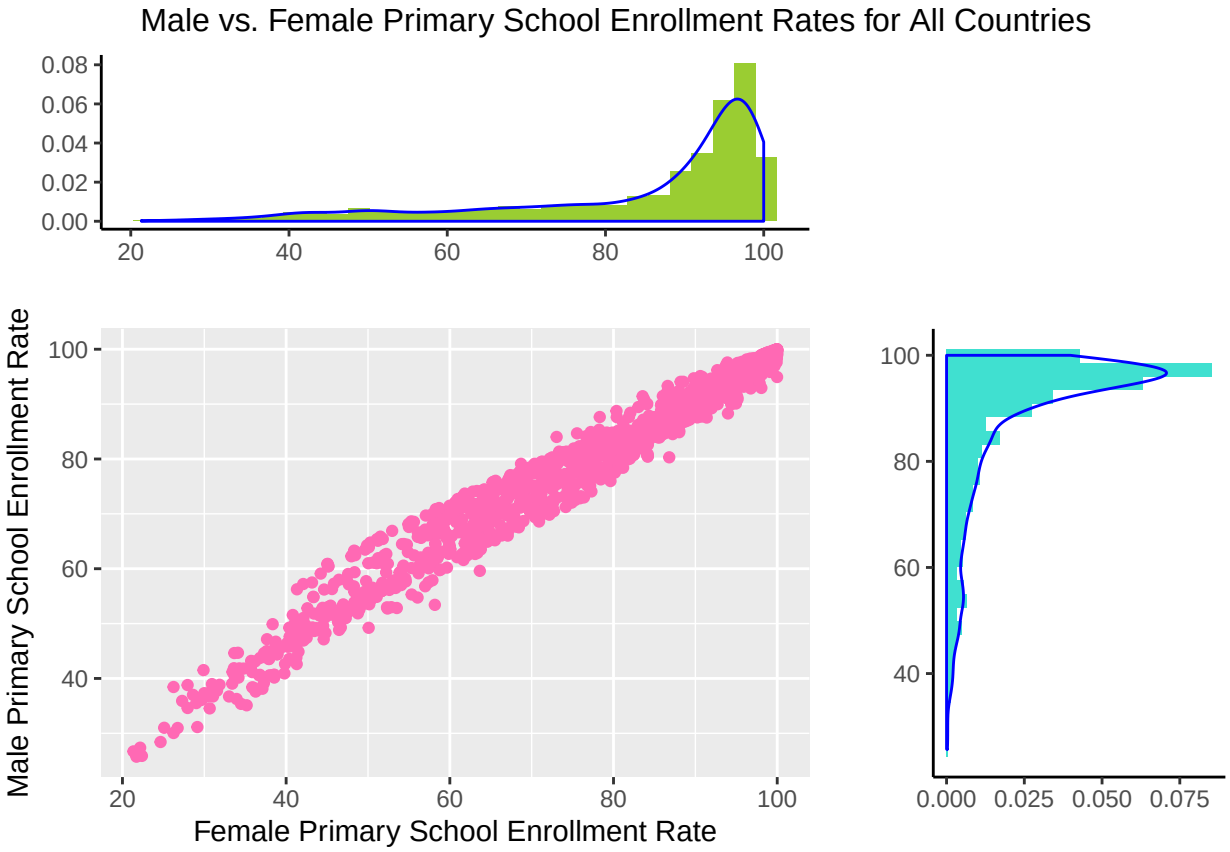


Problem 4

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.

```
edstats.df.se.prm.tenr.both.df <- as.data.frame(edstats.df.se.prm.tenr.both)

p1 <- ggplot(edstats.df.se.prm.tenr.both.df, aes(x = Female, y = Male)) +
  geom_point(color = "hotpink") +
  xlab("Female Primary School Enrollment Rate") +
  ylab("Male Primary School Enrollment Rate")
p2 <- ggplot(edstats.df.se.prm.tenr.both.df, aes(Female, ..density.., )) +
  geom_histogram(fill = "yellowgreen") +
  geom_density(col = "blue") + xlab("") + ylab("") +
  theme_classic()
p3 <- ggplot(edstats.df.se.prm.tenr.both.df, aes(Male, ..density..)) +
  geom_histogram(fill = "turquoise") +
  geom_density(col = "blue") + xlab("") + ylab("") +
  coord_flip() + theme_classic()
gg_figure <- ggarrange(p2, NULL, p1, p3,
                       ncol = 2, nrow = 2,  align = "hv",
                       widths = c(2, 1), heights = c(1, 2))
annotate_figure(gg_figure,
                top = "Male vs. Female Primary School Enrollment Rates for All Countries")
```



Problem 5

Finish this homework by pushing your changes to your repo.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW8_lastname_firstname.Rmd and HW4_lastname_firstname.pdf