HW8_Kim_Jaiyool

Due Wednesday Nov 13, 2019

2019-11-11

Problem 1

Work through the Swirl "Exploratory_Data_Analysis" lesson parts 1 - 10.

My answer : I did the above things !!

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.

My answer : I named this R-markdown file as HW8_Kim.

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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
# Data Handling to make the tidy data set.
Raw.data <- read.csv("EdStatsData.csv",skip = 1, header = F)

# Rename the names of columns
```

```r
colnames(Raw.data) <- c("Country_Name", "Country Code","Indicator_Name",
                        "Indicator Code","1970","1971","1972","1973","1974","1975","1976","1977",
                        "1978","1979","1980","1981","1982","1983","1984","1985","1986","1987",
                        "1988","1989","1990","1991","1992","1993","1994","1995","1996","1997",
                        "1998","1999","2000","2001","2002","2003","2004","2005","2006","2007",
                        "2008","2009","2010","2011","2012","2013","2014","2015","2016","2017",
                        "2020","2025","2030","2035","2040","2045","2050","2055","2060","2065",
                        "2070","2075","2080","2085","2090","2095","2100")


## Load the South Korea's data.
Data.of.South.KOR <- Raw.data[grepl("Korea, Rep",Raw.data[["Country_Name"]]),]

## Load the Arab World's data.
Data.of.Arab.World <- Raw.data[grepl("Arab World",Raw.data[["Country_Name"]]),]


## Remove the 'Indicator code' column (4th), and redundant 'NA' Column
Data.of.South.KOR1 <- Data.of.South.KOR[,-c(4,70)]

## Remove the 'Indicator code' column (4th), and redundant 'NA' Column
Data.of.Arab.World1 <- Data.of.Arab.World[,-c(4,70)]


## Subset data for South Korea

## We will use these below 3 column variables.

# 1) Adjusted net enrollment rate, lower secondary, both sexes (%)

# 2) Adjusted net enrollment rate, lower secondary, female (%)

# 3) Adjusted net enrollment rate, lower secondary, gender parity index (GPI)

## to see the Densigram (Composite word for Density + Histogram) in 1 plot.

Subset.Data.for.South.KOR <- Data.of.South.KOR1[1:3,]


# Make more tidy data set using the 'gather' function. (to make the 'Year' and corresponding 'Value' va
Subset.Data.for.South.KOR.1 <- gather(Subset.Data.for.South.KOR , "Year", "Value", 4:68)

# Separate the Indicator_Name Column using the 'spread' function
Subset.Data.for.South.KOR.2 <- spread(Subset.Data.for.South.KOR.1 , Indicator_Name, Value)


## Now, define the tidy data set as Tidy.Data.Set.for.South.Korea
Tidy.Data.Set.for.South.Korea <- Subset.Data.for.South.KOR.2



## Subset data for Arab World
```

Table 1: Summary Table for 3 Indexes of South Korea

	Country_Name	Country Code	Year	Adjusted net enrolment rate, lower secondary, both sexes (%)	Adjusted net enrolment rate, lower secondary, female (%)	Adjusted net enrolment rate, lower secondary, gender parity index (GPI)
	Korea, Rep. :65	KOR :65	Length:65	Min. :36.38	Min. :29.57	Min. :0.6918
	Afghanistan : 0	ABW : 0	Class :character	1st Qu.:70.09	1st Qu.:65.67	1st Qu.:0.8843
	Albania : 0	AFG : 0	Mode :character	Median :91.48	Median :91.64	Median :0.9904
	Algeria : 0	AGO : 0	NA	Mean :81.74	Mean :79.77	Mean :0.9369
	American Samoa: 0	ALB : 0	NA	3rd Qu.:95.64	3rd Qu.:95.17	3rd Qu.:1.0027
	Andorra : 0	AND : 0	NA	Max. :96.88	Max. :97.18	Max. :1.0190
	(Other) : 0	(Other): 0	NA	NA's :30	NA's :30	NA's :30

Table 2: Summary Table for 2 Indexes of Arab World

	Country_Name	Country Code	Year	Adjusted net enrolment rate, primary, both sexes (%)	Adjusted net enrolment rate, primary, gender parity index (GPI)
	Arab World :65	ARB :65	Length:65	Min. :54.82	Min. :0.6564
	Afghanistan : 0	ABW : 0	Class :character	1st Qu.:66.09	1st Qu.:0.7692
	Albania : 0	AFG : 0	Mode :character	Median :71.81	Median :0.8447
	Algeria : 0	AGO : 0	NA	Mean :72.22	Mean :0.8346
	American Samoa: 0	ALB : 0	NA	3rd Qu.:80.81	3rd Qu.:0.9263
	Andorra : 0	AND : 0	NA	Max. :86.10	Max. :0.9662
	(Other) : 0	(Other): 0	NA	NA's :20	NA's :20

```
## We will use these below 2 column variables.

# 1) Adjusted net enrollment rate, primary, both sexes (%)

# 2) Adjusted net enrollment rate, primary, gender parity index (GPI)

## to see the Densigram (Composite word for Density + Histogram) in 1 plot.

Subset.Data.for.Arab.World <- Data.of.Arab.World1[c(5,7),]


# Make more tidy data set using the 'gather' function. (to make the 'Year' and corresponding 'Value' va
Subset.Data.for.Arab.World.1 <- gather(Subset.Data.for.Arab.World, "Year", "Value", 4:68)

# Separate the Indicator_Name Column using the 'spread' function
Subset.Data.for.Arab.World.2 <- spread(Subset.Data.for.Arab.World.1 , Indicator_Name, Value)


## Now, define the tidy data set as Tidy.Data.Set.for.Arab.World
Tidy.Data.Set.for.Arab.World <- Subset.Data.for.Arab.World.2


## Make Table using 'kable' option.

#kable(Tidy.Data.Set.for.South.Korea,caption="Tidy Data Set for South Korea's 3 Indexes") %>%
 # kable_styling(latex_options="scale_down")


kable(summary(Tidy.Data.Set.for.South.Korea),caption = "Summary Table for 3 Indexes of South Korea") %>%
  kable_styling(latex_options="scale_down")
```

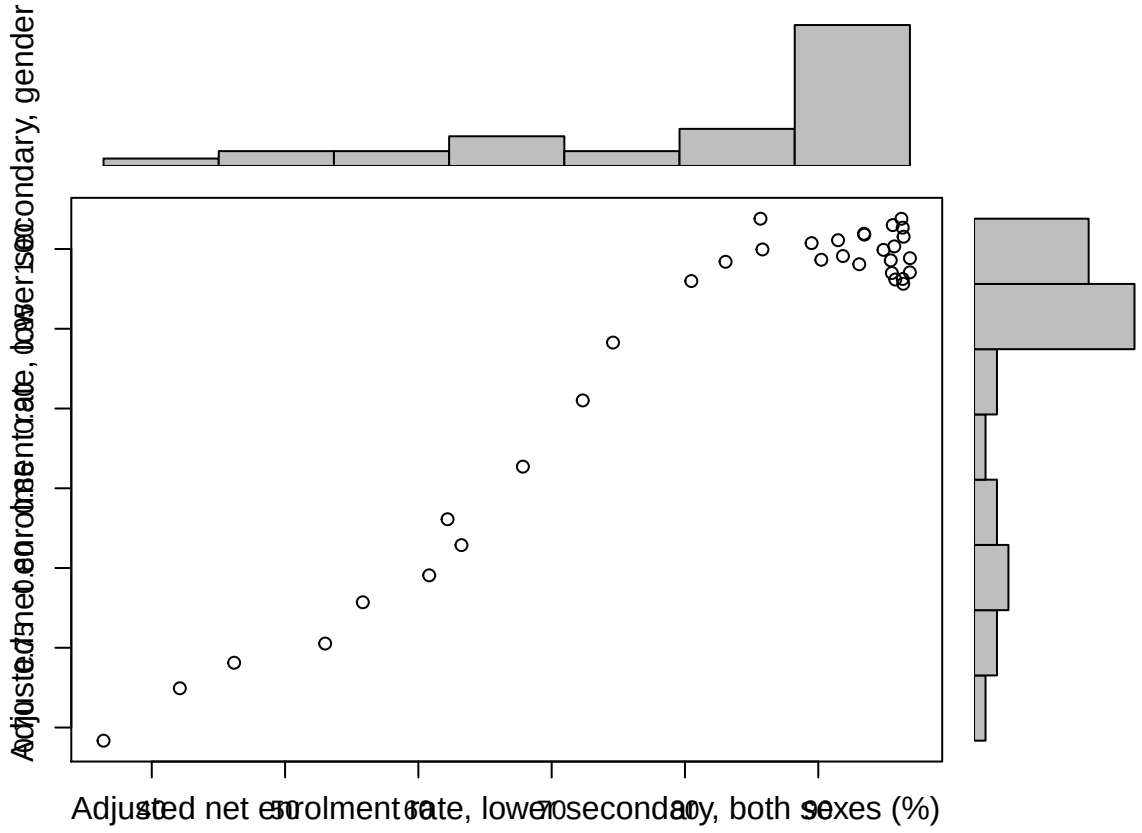
```
## Make Table using 'kable' option.

#kable(Tidy.Data.Set.for.Arab.World,caption="Tidy Data Set for Arab World's 2 Indexes") %>%
#  kable_styling(latex_options="scale_down")


kable(summary(Tidy.Data.Set.for.Arab.World),caption = "Summary Table for 2 Indexes of Arab World") %>%
  kable_styling(latex_options="scale_down")
```

My ans : Based on the above my R code, in the 'Tidy.Data.Set.for.South.Korea' data set, there are 65 obs. of 6 variables (mentioned in the above R code), however, what I am interested in this tidy data set is the number of data point for three column variables I have mentioned in the above R code. Thus, $65 \cdot 3 = 195$ data points (including 'NA' value) exist in the 'Tidy.Data.Set.for.South.Korea' data set needed to draw the so-called 'Densigram' (which is a composite word (Density + Histogram)) in 1 plot simultaneously. In a similar way, in the 'Tidy.Data.Set.for.Arab.World' data set, there are $65 \cdot 2 = 130$ data points to draw the same thing with the above. The Tidy data set for selected indexes for each nation (South Korea, and Arab World) and summary tables are summarized in the below tables.

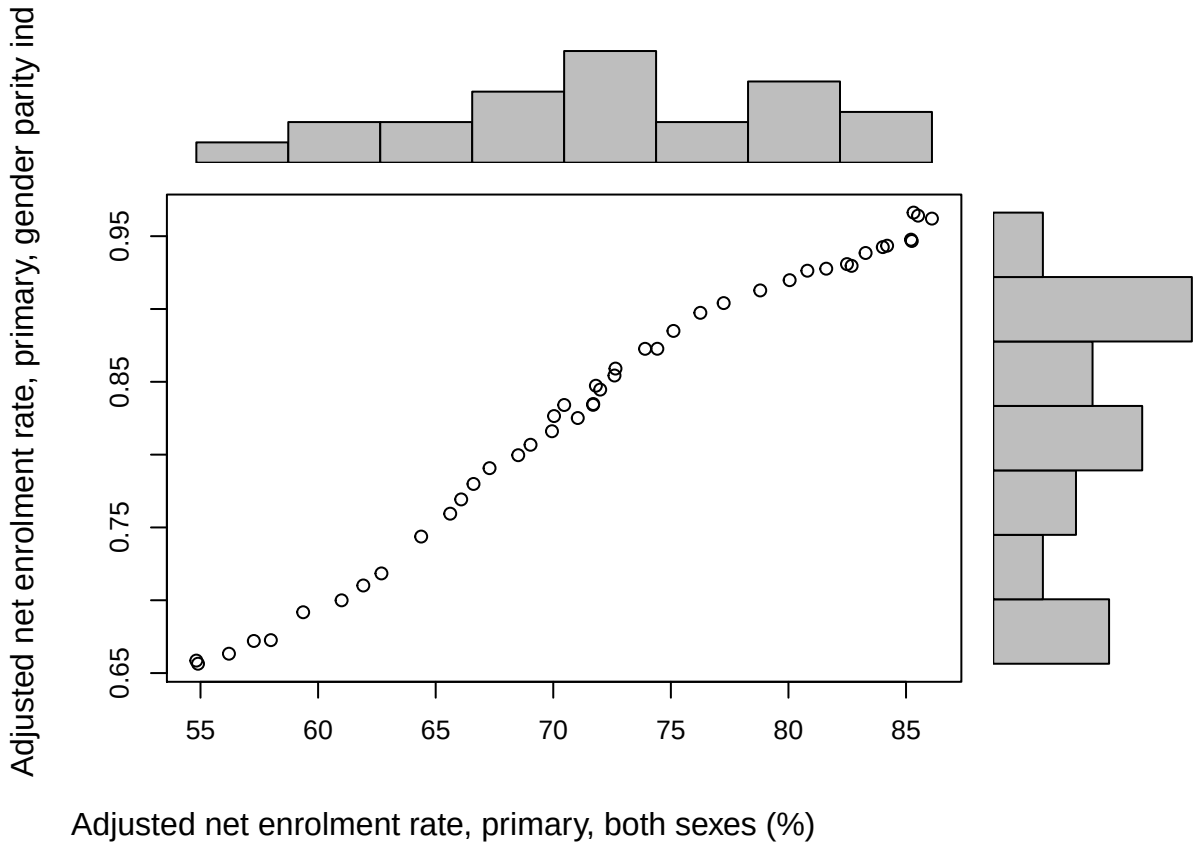
Problem 4

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.

```
### my function than enable us to draw the scatter plot and histogram simultaneously
## in one plot using basic plot and hist function.

scatterhist = function(x, y, xlab="", ylab=""){

  zones=matrix(c(2,0,1,3), ncol=2, byrow=TRUE)

  layout(zones, widths=c(4/5,1/5), heights=c(1/5,4/5))

  xhist = hist(x, plot=FALSE)

  yhist = hist(y, plot=FALSE)

  top = max(c(xhist$counts, yhist$counts))

  par(mar=c(3,3,1,1))

  plot(x,y)

  par(mar=c(0,3,1,1))

  barplot(xhist$counts, axes=FALSE, ylim=c(0, top), space=0)

  par(mar=c(3,0,1,1))

  barplot(yhist$counts, axes=FALSE, xlim=c(0, top), space=0, horiz=TRUE)

  par(oma=c(3,3,0,0))

  mtext(xlab, side=1, line=1, outer=TRUE, adj=0,
        at=.8 * (mean(x) - min(x))/(max(x)-min(x)))

  mtext(ylab, side=2, line=1, outer=TRUE, adj=0,
        at=(.8 * (mean(y) - min(y))/(max(y) - min(y))))
}


## Draw the Densigram using 'with' function.

with(Tidy.Data.Set.for.South.Korea, scatterhist(Tidy.Data.Set.for.South.Korea$`Adjusted net enrolment ra
```




```
## Draw the Densigram using 'with' function.

with(Tidy.Data.Set.for.Arab.World, scatterhist(Tidy.Data.Set.for.Arab.World$`Adjusted net enrolment rat
```



My ans : Actually, using the function 'scatterhist' function, I made the above two Densigrams for 2 column variables in each national tidy data set seen above.

Problem 5

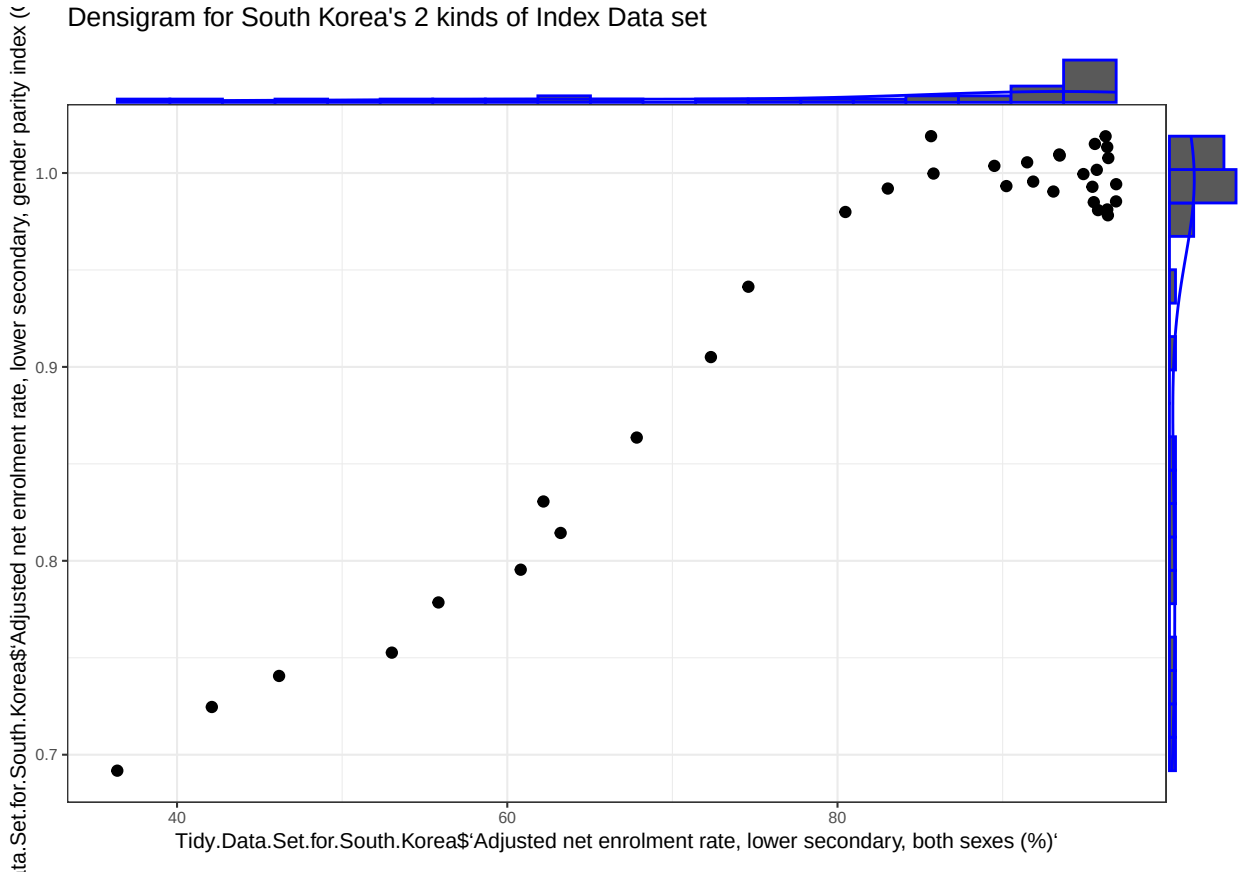
Recreate the plot in problem 3 using ggplot2 functions. Note: there are many extension libraries for ggplot, you will probably find an extension to the ggplot2 functionality will do exactly what you want.

For South Korea's Tidy Data set (Defined in the above) :

```
library(ggplot2)
library(ggExtra)

p <- ggplot(Tidy.Data.Set.for.South.Korea,
            aes(Tidy.Data.Set.for.South.Korea$`Adjusted net enrolment rate, lower secondary, both sexes
                Tidy.Data.Set.for.South.Korea$`Adjusted net enrolment rate, lower secondary, gender pari


p2 <- p + ggtitle("Densigram for South Korea's 2 kinds of Index Data set") + theme_bw(8)



# p <- ggplot(Dataset, aes('X' Axis Column, 'Y' Axis Column)) + geom_point()

# Using a "densigram" plot
# par(mfrow=c(2,1))
ggMarginal(p2, type = "densigram",colour="blue",bins=20,size=15)
```

```
## Warning: Removed 30 rows containing missing values (geom_point).
```

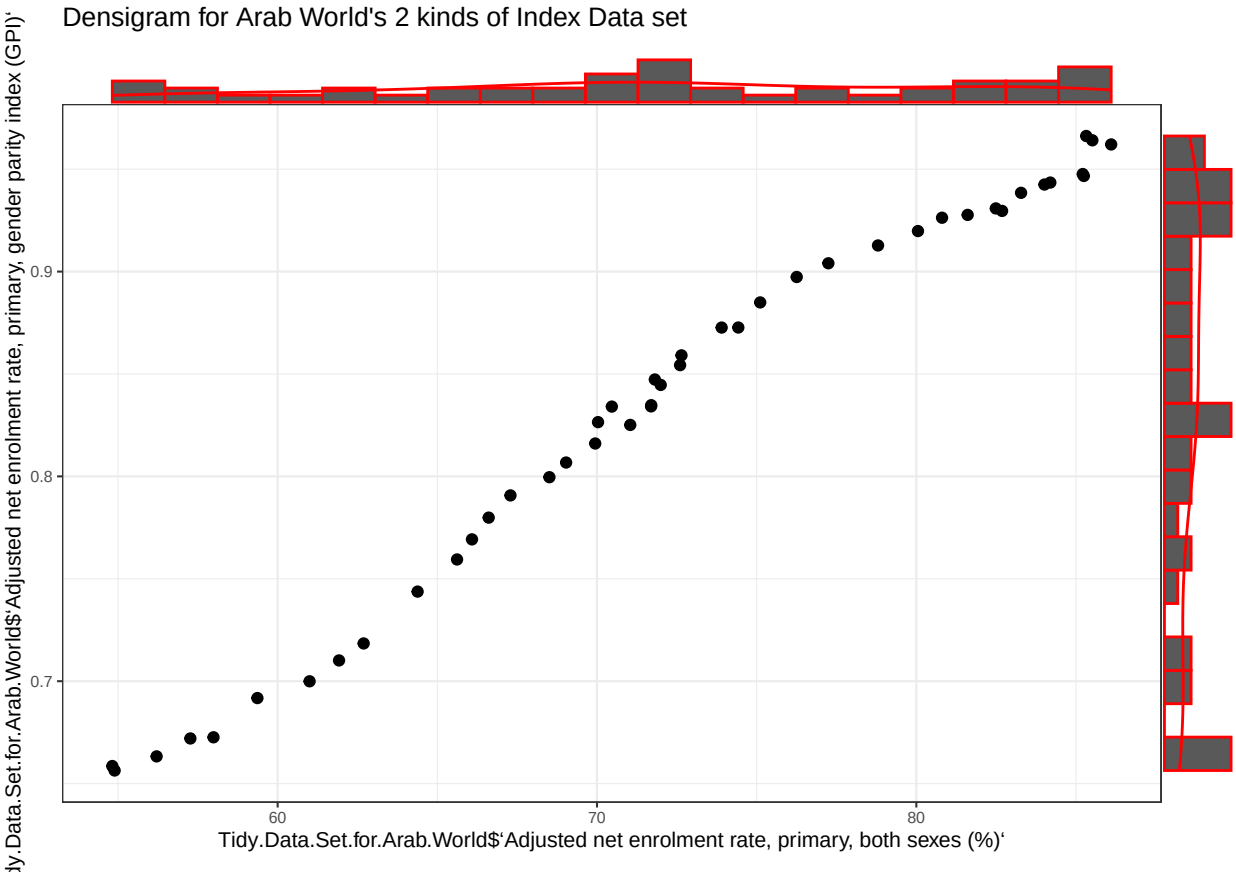
```
## Warning: Ignoring unknown parameters: bins

## Warning: Ignoring unknown parameters: bins

## Warning: Ignoring unknown parameters: bins

## Warning: Ignoring unknown parameters: bins

## Warning: Removed 30 rows containing missing values (geom_point).
```



For Arab World's Tidy Data set (Defined in the above) :

```
library(ggplot2)
library(ggExtra)

####

p3 <- ggplot(Tidy.Data.Set.for.Arab.World,
             aes(Tidy.Data.Set.for.Arab.World$`Adjusted net enrolment rate, primary, both sexes (%)`,
                 Tidy.Data.Set.for.Arab.World$`Adjusted net enrolment rate, primary, gender parity index

p4 <- p3 + ggtitle("Densigram for Arab World's 2 kinds of Index Data set") + theme_bw(8)



ggMarginal(p4, type = "densigram",colour="red",bins=20,size=15)
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
## Warning: Ignoring unknown parameters: bins
## Warning: Ignoring unknown parameters: bins
## Warning: Ignoring unknown parameters: bins
## Warning: Ignoring unknown parameters: bins
## Warning: Removed 20 rows containing missing values (geom_point).
```



My ans : Using the 'ggplot2' package (actually, there are plenty of flavors to capitalize), we can draw more fancy Densigrams compared to the prior Densigrams !!!

Problem 6

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

My ans : OK.