Homework 8

Due Wednesday Nov 13, 2019

2019-11-11

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.

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?

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

Answer

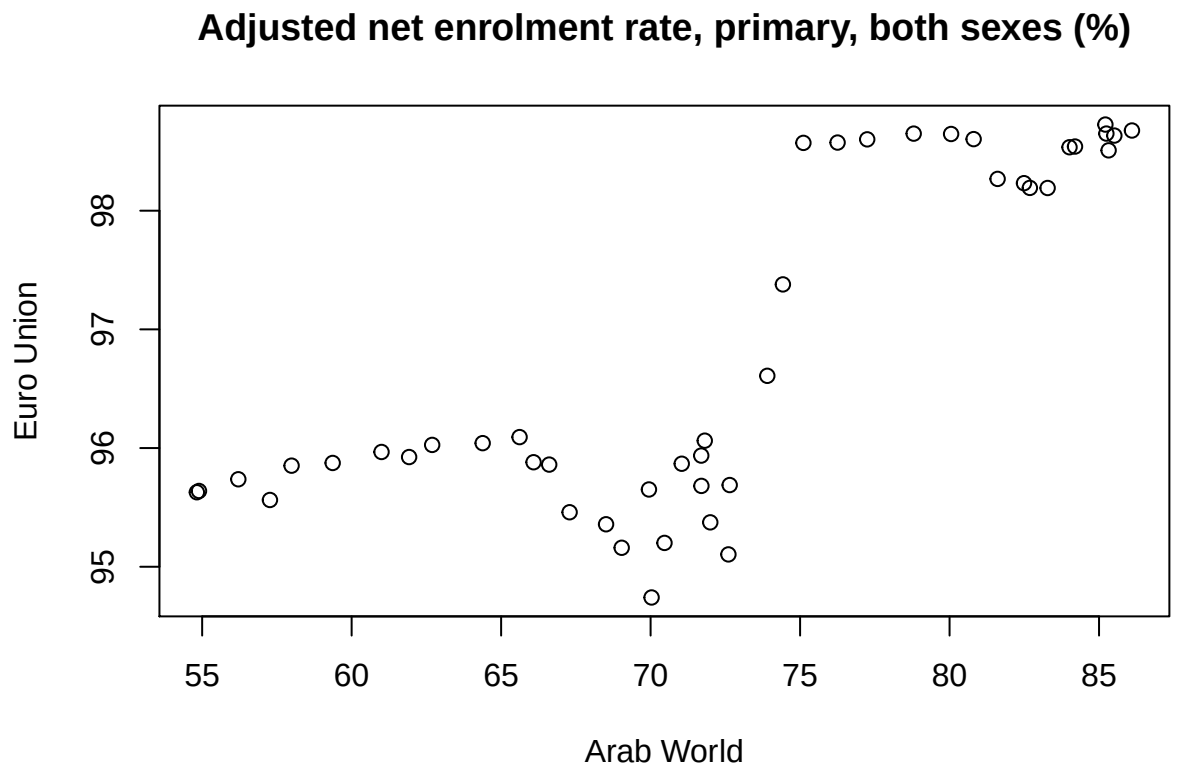
```
EdStatsCountry.Series <- read.csv("Edstats_csv/EdStatsCountry-Series.csv")
EdStatsCountry <- read.csv("Edstats_csv/EdStatsCountry.csv")
EdStatsData <- read.csv("Edstats_csv/EdStatsData.csv")
EdStatsFootNote <- read.csv("Edstats_csv/EdStatsFootNote.csv")
EdStatsSeries <- read.csv("Edstats_csv/EdStatsSeries.csv")
```

```
# merge EdStatsCountry.Series and EdStatsCountry by "Country.Code"
# name as df1
colnames(EdStatsCountry.Series)[1] <- "Country.Code"
df1 = merge.data.frame(EdStatsCountry.Series, EdStatsCountry, by = "Country.Code")

# merge df1 and EdStatsData by "Country.Code"
# name as df2
df2 = merge.data.frame(df1, EdStatsData, by = "Country.Code")
colnames(df2)[2] = "Series.Code"
```

```r
# merge df3 and EdStatsFootNote
colnames(EdStatsFootNote)[c(1,2)] <- c("Country.Code", "Series.Code")
EdStatsFootNote = EdStatsFootNote[,-5]
df3 = merge.data.frame(df2, EdStatsFootNote, by = "Seiries.Code")

# levels(EdStatsData$Country.Name)
# levels(EdStatsData$Indicator.Name)[1:30]
# View(EdStatsData[EdStatsData$Indicator.Name == "", ])
```

```r
ed_arb = EdStatsData[EdStatsData$Country.Name == "Arab World" & EdStatsData$Indicator.Code == "SE.PRM.T
ed_eu = EdStatsData[EdStatsData$Country.Name == "European Union" & EdStatsData$Indicator.Code == "SE.PR
ed_arb_data = as.numeric(ed_arb[5:49])
ed_eu_data = as.numeric(ed_eu[5:49])
```

```r
plot(ed_arb_data, ed_eu_data,
     xlab = "Arab World",
     ylab = "Euro Union",
     main = "Adjusted net enrolment rate, primary, both sexes (%)")
```

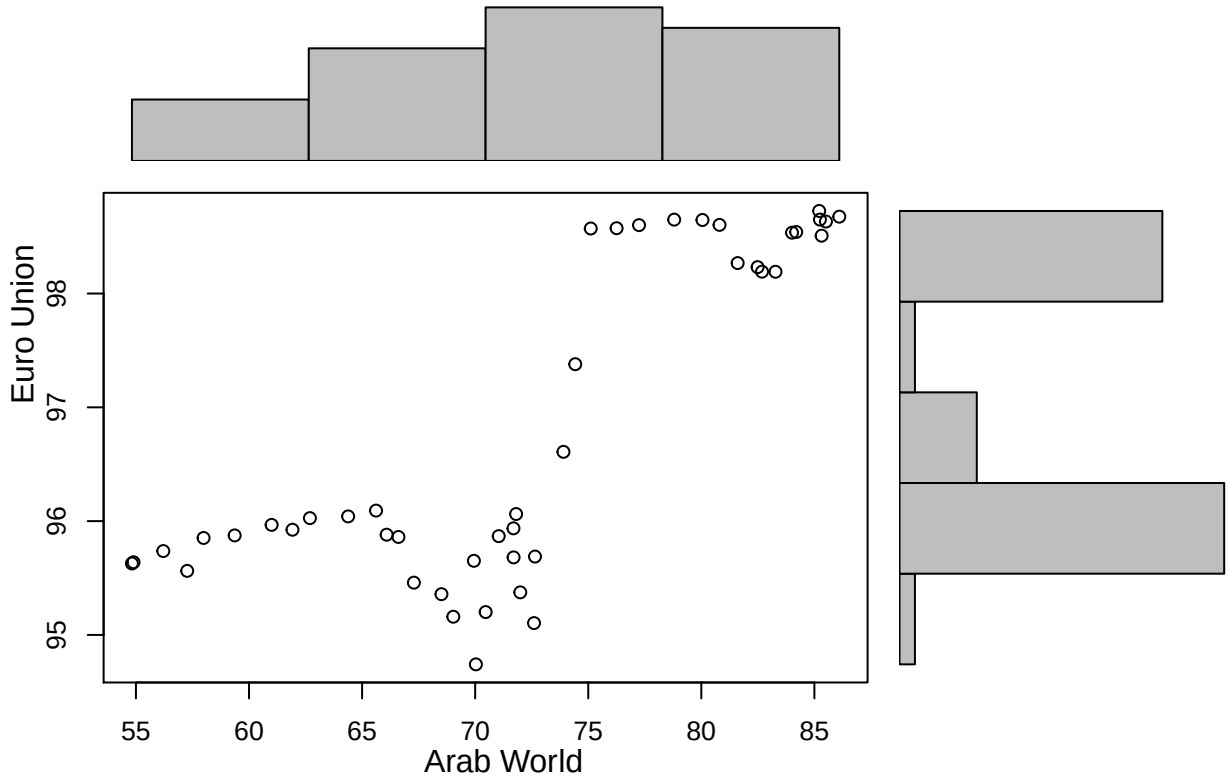


Problem 3

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.

Answer

```r
# experiment data
# x <- rnorm(1000, mean = 5, sd = 0.5)
# y <- rnorm(1000, mean = 5, sd = 0.5)

# data from problmem 2
x = ed_arb_data
y = ed_eu_data

  xlab = "Arab World"
  ylab = "Euro Union"
  zones=matrix(c(2,0,1,3), ncol=2, byrow=TRUE)
  layout(zones, widths=c(5/7,2/7), heights=c(2/7,5/7))
  xhist = hist(x, breaks = "FD", plot=FALSE)
  yhist = hist(y, breaks = "FD", 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=2, outer=TRUE, adj=0,
    at=.5 * (mean(x) - min(x))/(max(x)-min(x)))
  mtext(ylab, side=2, line=2, outer=TRUE, adj=0,
    at=(.7 * (mean(y) - min(y))/(max(y) - min(y))))
```

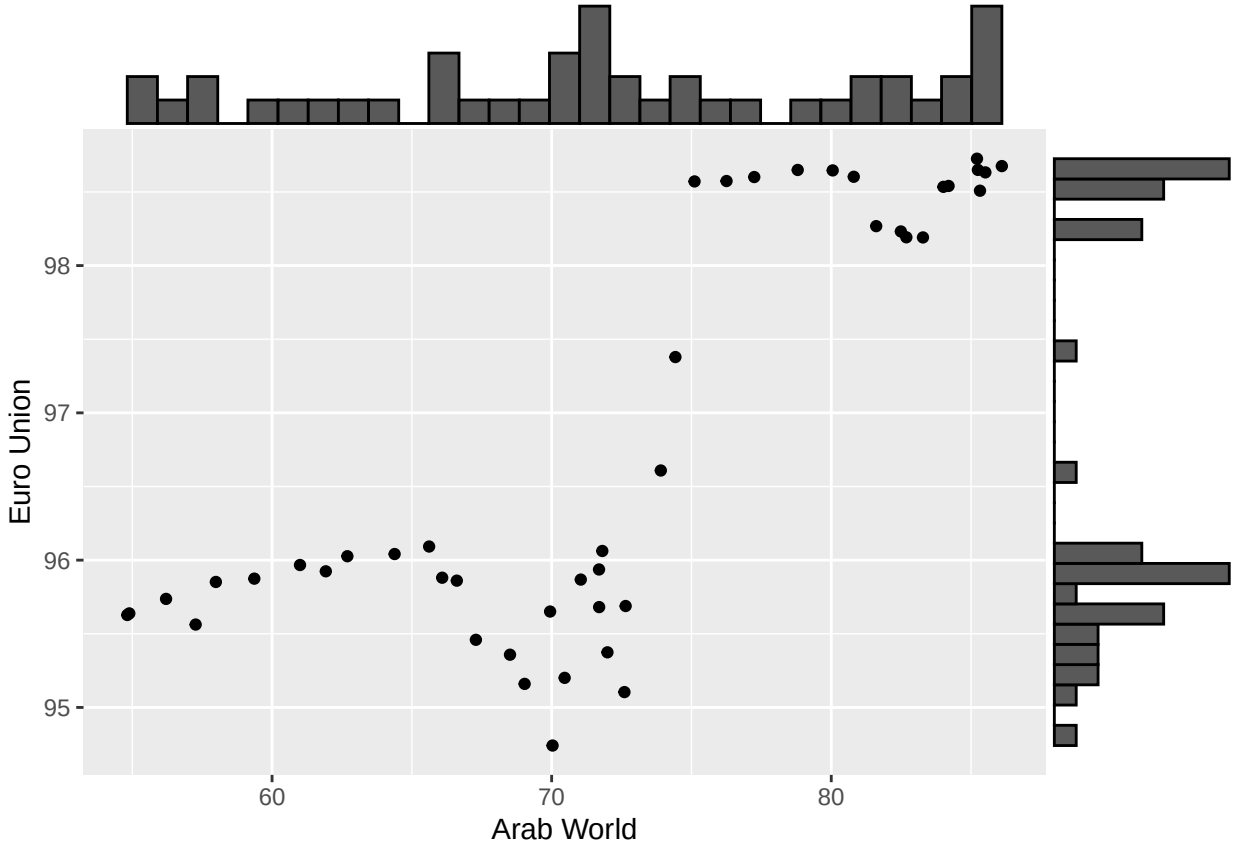


Problem 4

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.

Answer

```
library(ggplot2)
library(ggExtra)
df = data.frame(x,y)

plot_center = ggplot(df, aes(x=x, y=y)) +
  geom_point() +
  xlab(xlab) +
  ylab(ylab)
#  geom_smooth(method="lm")

# default: type="density"
ggMarginal(plot_center, type="histogram")
```



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