HW8_Brown

Nick Brown

10/30/2019

##Problem 1:

Completed Swirl Exploratory Data Analysis lessons 1 through 10.

##Problem 2:

HW8 created as necessary.

##Problem 3:

Clean data from the World Bank

```
zip_edu_url <- "https://databank.worldbank.org/data/download/Edstats_csv.zip"

############## Read all files from a .zip file ##############################
#(citation: https://hydroecology.net/downloading-extracting-and-reading-files-in-r/)

# create a temporary directory
td = tempdir()
# create the placeholder file
tf = tempfile(tmpdir=td, fileext=".zip")
# download into the placeholder file
download.file(zip_edu_url, tf)

# get the name of the first file in the zip archive
fname = unzip(tf, list=TRUE)$Name[1]
# unzip the file to the temporary directory
unzip(tf, files=fname, exdir=td, overwrite=TRUE)
# fpath is the full path to the extracted file
fpath = file.path(td, fname)

# stringsAsFactors=TRUE will screw up conversion to numeric!
edu = read.csv(fpath, header=TRUE, row.names=NULL,
               stringsAsFactors=FALSE)
#############################################################################################
```

How many data points were there in the complete dataset?

There are 886,930 observations of 70 variables in the complete dataset

```
###### Clean Dataset ######################

clean_edu <- edu[!is.na(edu$X1970),1:70]
```

How many data points were there in the cleaned dataset?

There are 72,288 observations of 70 variables in the cleaned dataset

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

```
###### Choose two countries ######################

Korea <- subset(clean_edu, Country.Code == "KOR" )
Turkey <- subset(clean_edu, Country.Code == "TUR" )
korea_indicator <- summary(Korea$Indicator.Name)
turkey_indicator <- summary(Turkey$Indicator.Name)
kable(cbind(korea_indicator,turkey_indicator), caption="Summary table of indicators for comparison")
```

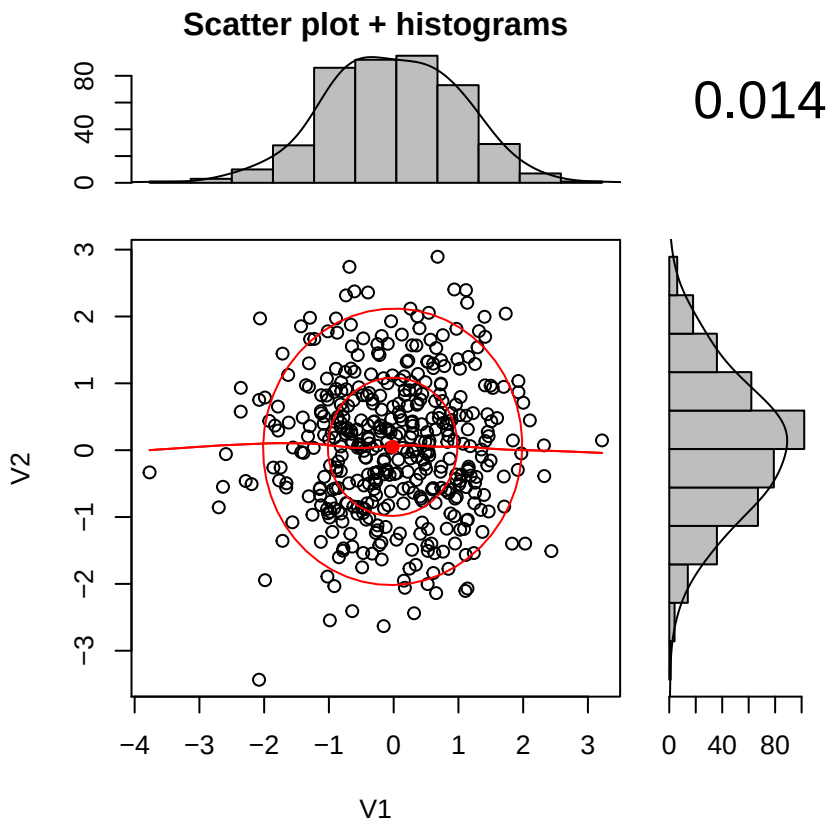
Table 1: Summary table of indicators for comparison

	korea_indicator	turkey_indicator
Length	435	416
Class	character	character
Mode	character	character

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.

```
###### Scatterplot with marginal histograms ######################

Korea <- subset(clean_edu, Country.Code == "KOR" )
Turkey <- subset(clean_edu, Country.Code == "TUR" )

x=na.omit(Korea$X1970)
y=na.omit(Korea$X2000)

x=x[-(length(y)+1):-length(x)]

x=rnorm(x)
y=rnorm(y)

scatter.hist(x,y)
```



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.

```
###### Scatterplot with marginal histograms ######################

Korea <- subset(clean_edu, Country.Code == "KOR" )
Turkey <- subset(clean_edu, Country.Code == "TUR" )

x=na.omit(Korea$X1970)
y=na.omit(Korea$X2000)

x=x[-(length(y)+1):-length(x)]

x=rnorm(x)
y=rnorm(y)

hist_top <- ggplot()+geom_histogram(aes(x))
empty <- ggplot()+geom_point(aes(1,1), colour="white")+
         theme(axis.ticks=element_blank(),
               panel.background=element_blank(),
               axis.text.x=element_blank(), axis.text.y=element_blank(),
               axis.title.x=element_blank(), axis.title.y=element_blank())
```

```
scatter <- ggplot()+geom_point(aes(x,y))
hist_right <- ggplot()+geom_histogram(aes(y))+coord_flip()

grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```