Exploratory Data Analysis (EDA)

STAT/BIOS 823

Homework 6

Directions

Using `RMarkdown` in `RStudio`, complete the following questions. Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a `.Rmd` file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit your homework on the Blackboard.

If you have questions, please post them on the lesson discussion board. **All** questions are mandatory and the `R-code` and output must be clearly shown.

Homework submitted after the due date will attract a penalty of 10 points per day after the due date.

---

1. Use the built-in dataset `cars`.

(a) Reproduce Figure 1 by creating a scatterplot of `speed` versus {distance} starting with the code `plot(dist ~ speed, data = cars`. Add the following details into the plot: `main title` "Scatterplot of Speed versus Distance", `sub title` "Using plot() in R", `x axis title` "Speed (miles per hour)", `y axis title` "Stopping Distance (feet)". Make the main title to be `red`, axis colors to be `magenta`. Use filled circles for symbol type and make the symbol color to be blue and axis labels to be dark green. Make the fonts of the titles, label axes and symbol sizes to 1.5.

(b) From Figure 1, create Figure 2. This can be done by turning off the axes using `plot(...,axes=FALSE)`. Add a line of best fit using `abline(lm(dist ~ speed, data=cars))`. Add a grid using `grid()`. Add a box around the plotting area using `box(col="red", lwd=3, lty=3)`. Add a legend to the plot using `legend("topleft", inset = 0.01, title = "Distance vs. Speed", legend = c("Observation"), col=c("blue"), pch=19,horiz=TRUE)`. You can add these commands to get back the axes labels: `axis(1)` and `axis(2)`.

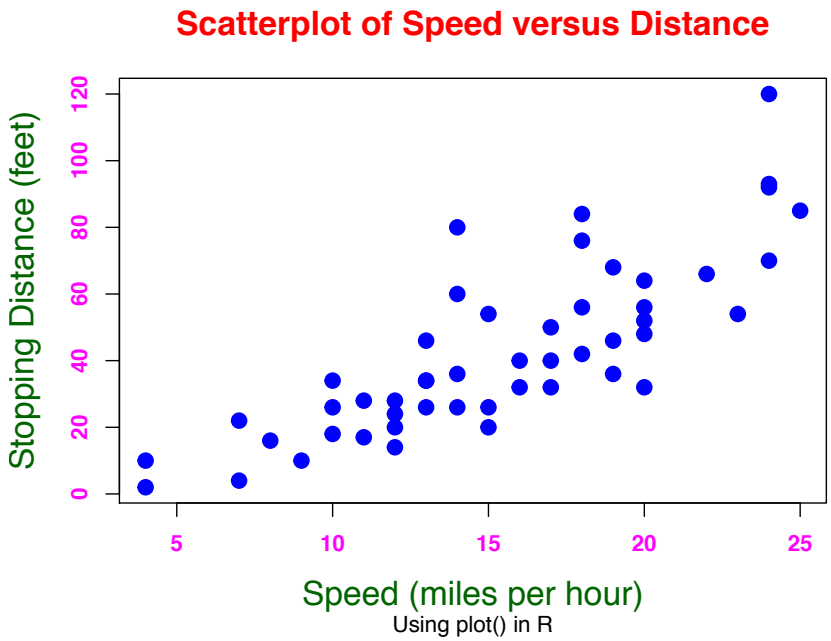

Figure 1: Scatterplot of Speed versus Distance

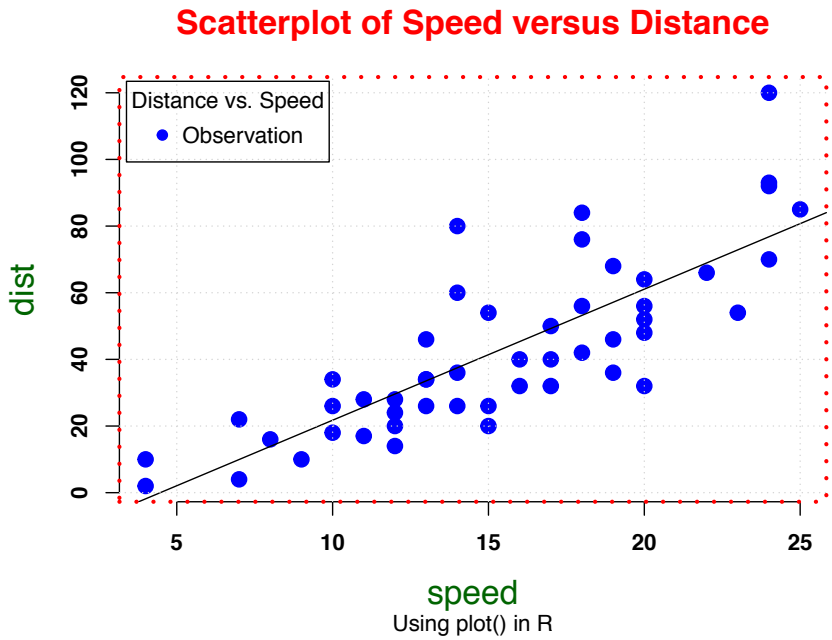

Figure 2: Scatterplot of Speed versus Distance

2. Use the `economics` built-in dataset and library `ggplot2`. Plot the time series of unemployment (Figure 3). Starting code is `ggplot(economics, aes(date, unemploy)) + geom_line()`

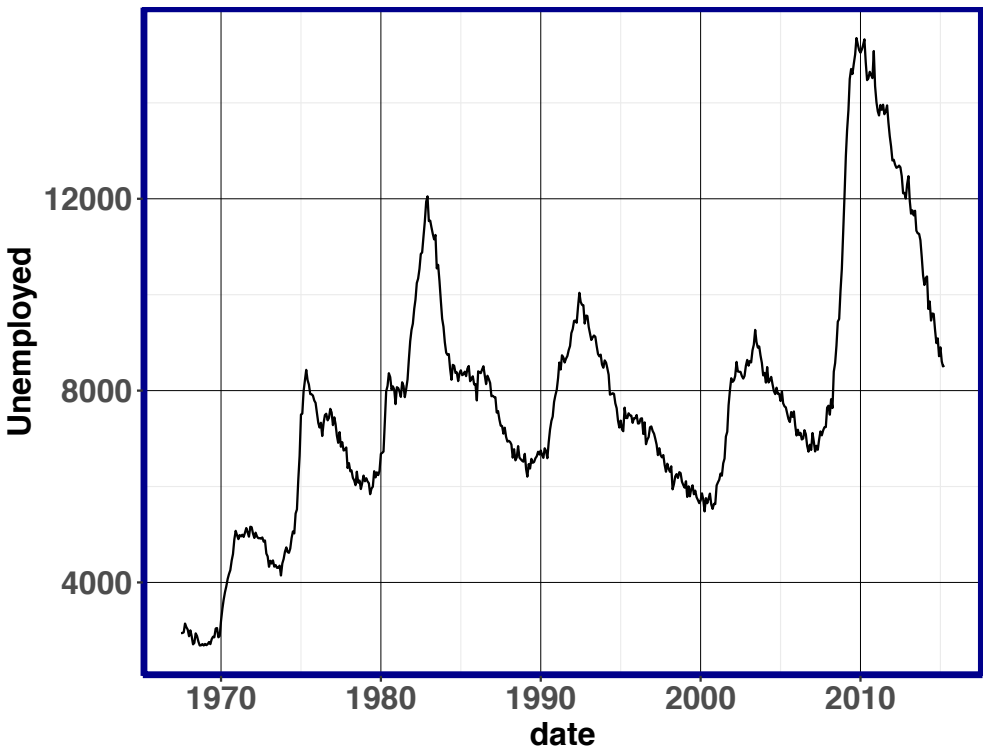

Figure 3: Time Series Graph

3. Use the built-in dataset `survey` that contains the results of a survey given to 237 students at the University of Adelaide. Download and install the `MASS` package and use the `R` help documentation to examine the contents of survey. Install `Hmisc` package by typing `install.packages("Hmisc",dep=TRUE)`.

(a) Use the `str()` function to examine the structure of the dataset. Use the `describe()` function in the `Hmisc` package. Use `des()`, `summ()` and `codebook()` functions from the `epiDisplay` package and `summary()` function to visualize summaries of the 12 variables in the dataset.

(b) Load `vcd` and `epiDisplay` packages. Use the `table()` and `tab1()` functions to *generate frequency tables* describing the distribution of each of the following categorical variables: `Sex, Exer` and `Smoke`.

(c) Produce contingency tables to explore the relationships between Sex and Exercise, Smoke and Exercise and Smoke and Sex. *Calculate the Pearson's Chi-squared test or Fisher's Exact Test if appropriate (if the expectation of at least one of the cell value is ≤ 5). From the test of independence of these categorical variables, what would be your conclusion?*

(d) Using the following code, write down the least squares regression equation describing the linear relationship between `hand span` and `height` and *calculate the Pearson's correlation coefficient. What do you notice about* r^2 *from the linear regression output and the correlation coefficient,* r*? Based on the Pearson's correlation matrix Figure 4, which continuous variables are highly correlated?*

```
data("survey")
ff <- lm(Height ~ Wr.Hnd, data = survey)
summary(ff)
# calculation of Pearson's correlation coefficient.
cor(survey$Wr.Hnd, survey$Height, use = "complete")
```

```
# This code was used to produce the correlation
# matrix
library(psych)
dat0 <- survey[, c("Pulse", "Age", "Height", "NW.Hnd",
    "Wr.Hnd")]
pairs.panels(dat0)
```

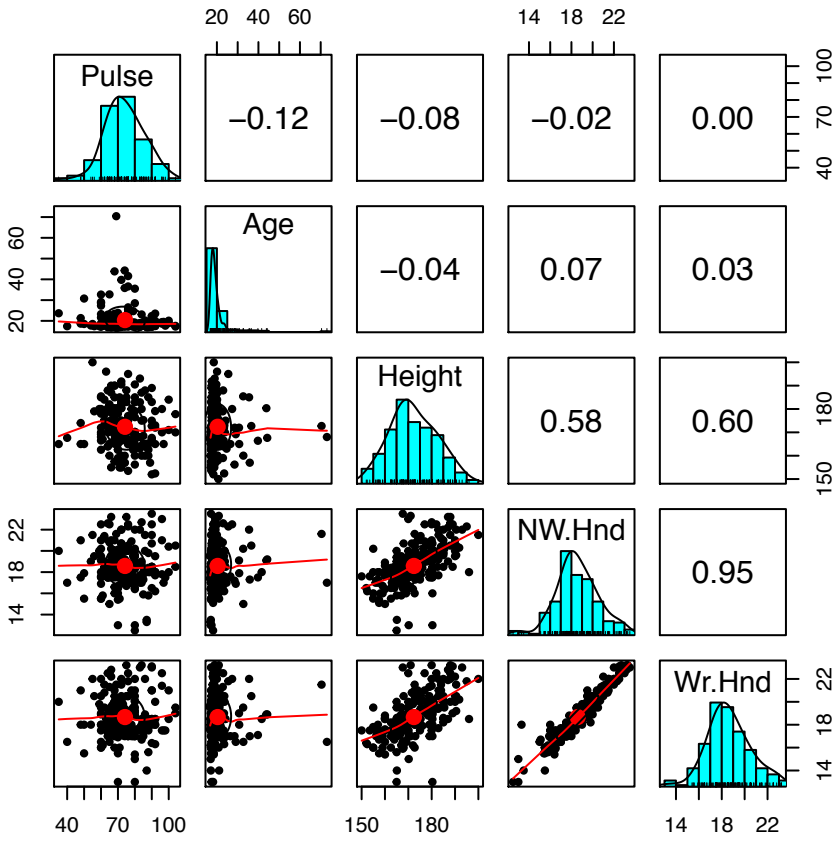

Figure 4: Pearson's Pairwise Correlation Coefficients