CMDA 3654: Assignment #6

Due on Friday, Nov 11, 2016

Jung Choi

"I have neither given nor received unauthorized assistance on this assignment."

November 10, 2016

1 Knitr

1. (20 points) Using knitr, format the answers, R code, and R output for the following problems into a pdf file. You can practice generating HTML and RHTML if you like too but that is for your own purposes. Please submit the R Markdown file (.Rmd or .Rnw) that you used to in your write-up.

```
install.packages(’knitr’, dependencies = TRUE)

library(knitr)
?knit
```

2 (20 points) Barplots using ggplot2

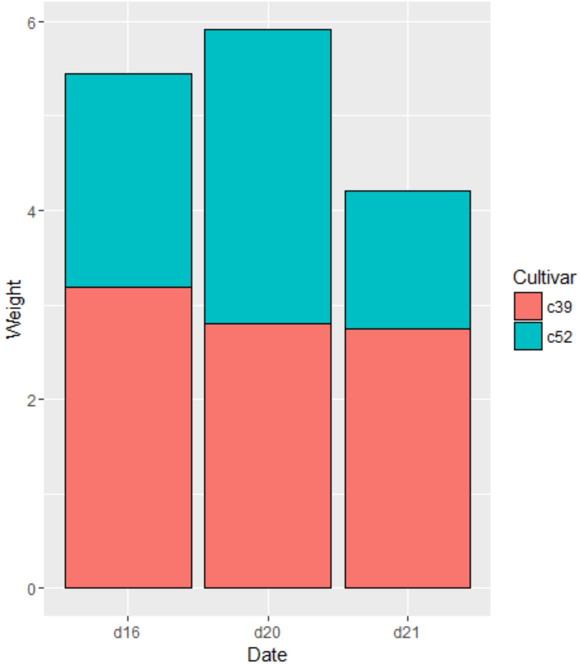
2. Barplots using ggplot2. You will need the R packages ggplot2 and gcookbook for problems 2-5. (a) Load the **cabbage_exp** dataset in R, and describe the dataset in your own words, in 2-3 lines. Note that is a summarized version of the **cabbages** dataset in the MASS package, and you can get detailed information from the **MASS** package.

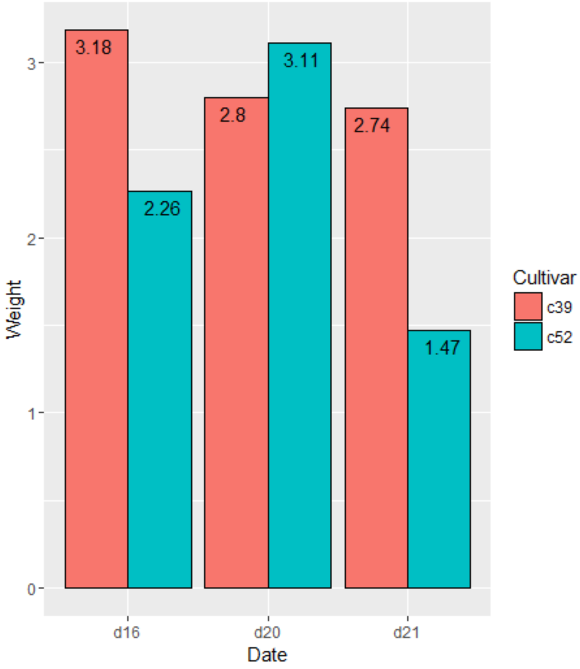
cabbage_exp is a summarized dataset of cabbages which it contains all the columns in the dataset cabbages except Vitc. The Vitc is the Ascorbic acid content labeled in integer values. The **cabbage_exp** contains mean, standard deviation, counts, and standard error of the mean for the cabbages values.This purpose of this summarized verision is to make it easier use for the graphs.

```
library(MASS)
library(ggplot2)
install.packages("gcookbook")
library(gcookbook)
cab = data.frame(cabbage_exp) # summarized version of the cabbage dataset in the
    MASS package
cab
cabbages
```

(b) Construct a side-by-side bar plot with Date in the x-axis, Weight in the y-axis, and different colors for different cultivar. Add labels to each bar with the corresponding Weight.

```
p=ggplot(cab, aes(x=Date, y=Weight, fill=Cultivar)) +  geom_bar(stat="identity",
    position = "dodge",colour="black")
# show what happens without stat="identity" (by default it tries to plot counts)
ggplot(cab, aes(x=Date, y=Weight, fill=Cultivar)) +  geom_bar(stat="identity",
    colour="black") #stacked
# add text
p+geom_text(aes(label=Weight), vjust=1.5, colour="black",position=position_dodge
    (1))
```


6
4
2
0
Weight
d16
d20
d21
Date
Cultivar
c39
c52


3.18
2.26
2.8
3.11
2.74
1.47
3
2
1
0
Weight
d16
d20
d21
Date
Cultivar
c39
c52

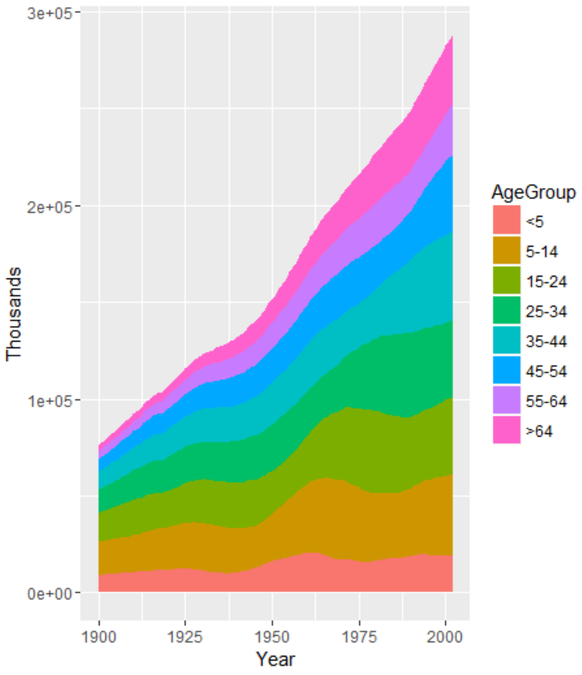
3 (20 Points) Area graphs using ggplot2

3. Load the **uspopage** dataset in R, and describe the dataset in your own words, in 2-3 lines.

This is the dataset from U.S. Census data which it contains age distribution of population in United States from 1900 to 2002. According to it's description, these varibles are estimated values. There are three variables that is used in this dataset which are Year, AgeGroup, Numb. of people in thousands

(b) Construct a stacked area graph with Year in the x-axis, population (in thousands) in the y-axis, and different age groups in different layers.

```
ggplot(uspopage, aes(x=Year, y=Thousands, fill=AgeGroup)) + geom_area()
```

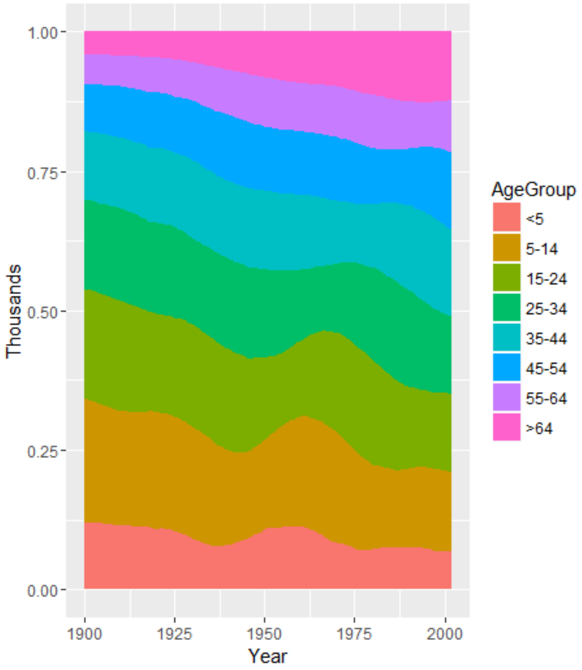


(c) Next, construct a proportional stacked area graph, i.e., for each Year, compute the contribution from each age group to the total population as a fraction of the total population. Construct a stacked area graph as before, with the proportions.

```
ggplot(uspopage, aes(x=Year, y=Thousands, fill=AgeGroup)) + geom_area(position="
    fill")
```

4 (20 Points) Scatterplots using ggplot2

(a) Load the heightweight dataset in R (in package gcookbook), and describe the dataset in your own words, in 2-3 lines.



The heightweight is a dataset that contains height and weight of schoolchildren. There are five variables ,such as sex, age in years, age in months, heigh in inches, and weight in pounds.

(b) Construct a scatterplot with age in the x-axis, height in the y-axis, and different colors for male and female students. Add linear regression line through the scatterplot using stat_smooth().

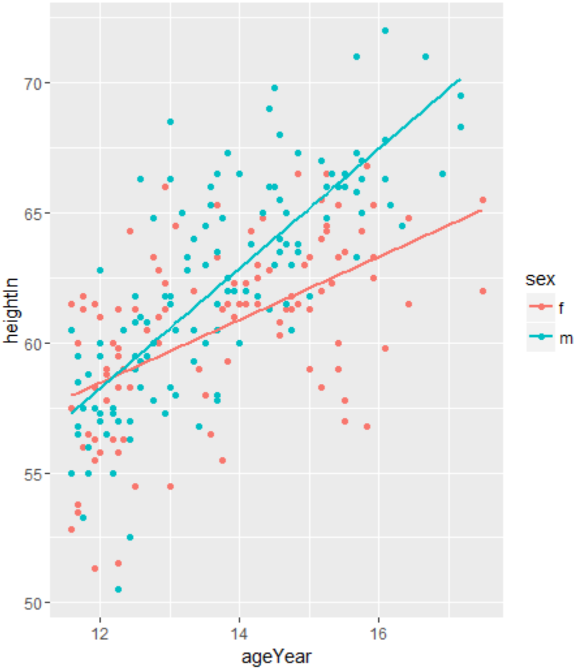
```
# ggplot(heightweight, aes(x=ageYear, y=heightIn)) + geom_point()
p = ggplot(heightweight, aes(x=ageYear, y=heightIn, colour=sex)) + geom_point()
# ggplot(heightweight, aes(x=ageYear, y=heightIn, shape=sex)) + geom_point()

# p + geom_point() + stat_smooth(method=lm) #95% confidence region by default
# p + geom_point() + stat_smooth(method=lm, level=0.99) #99% confidence region
p + geom_point() + stat_smooth(method=lm, se=FALSE) #no confidence region
```

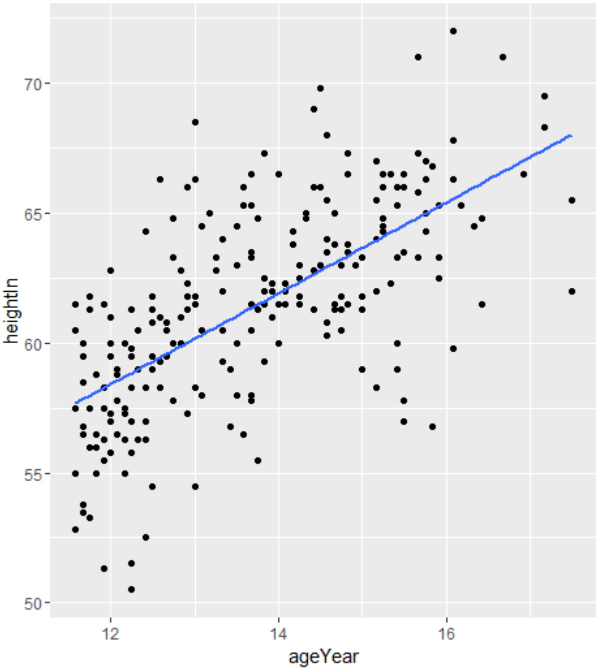
(c) Note that when using stat_smooth() in the previous problem, two regression lines are generated, one for male students and one for female students. Construct the same scatterplot as before, but with a single regression line for all students.

```
p2 = ggplot(heightweight, aes(x=ageYear, y=heightIn)) + geom_point()
p2 + geom_point() + stat_smooth(method=lm, se=FALSE) #no confidence region
```

5 (20 POINTS) GRAPHICAL PARAMETERS USING GGPLOT2.

(a) Load the faithful dataset in R, and describe the dataset in your own words, in 2-3 lines.


heightIn
70
65
60
55
50
12
14
16
ageYear
sex
f
m

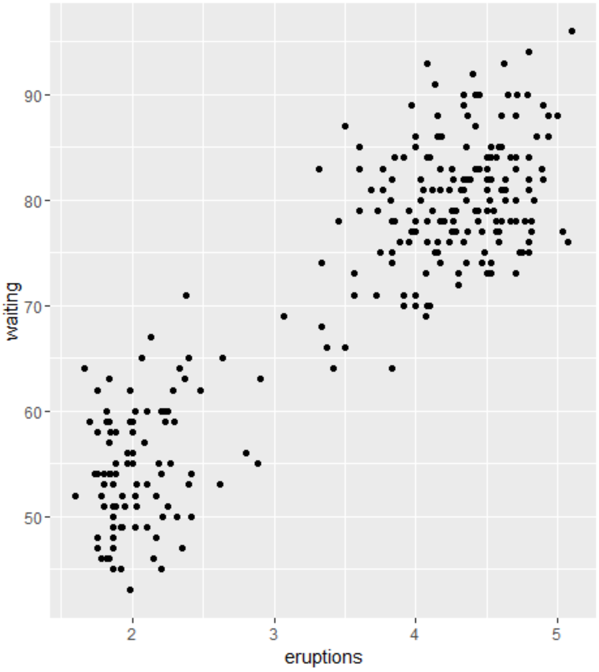

heightIn
70
65
60
55
50
12
14
16
ageYear

The faithful is a dataset collected from the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. It contains waiting time between eruptions and the duration of the eruption with 272 observations and 2 variables.

(b) Construct a scatterplot with eruption time in the x-axis, waiting time in the y-axis. What interesting features do you see in the plot?

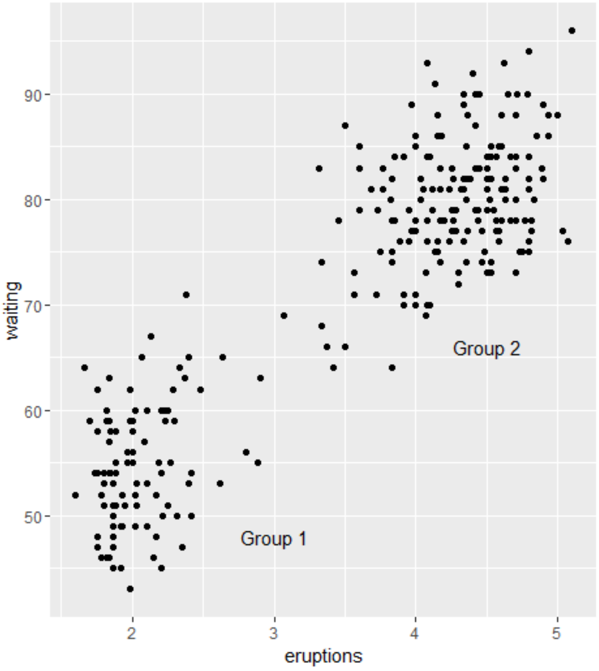
Shorter the eruption time will likely have shorter waiting time as well. It is opposite for the longer eruption time where it has longer waiting time.

```
ggplot(faithful, aes(x=eruptions, y=waiting)) + geom_point()
f=ggplot(faithful, aes(x=eruptions, y=waiting)) + geom_point()
```



(c) Add visual guides (e.g. text annotations, arrows, rectangles) to the scatterplot to emphasize the interesting features noted in the previous part.

```
f + annotate("text", x=3, y=48, label="Group␣1") +
  annotate("text", x=4.5, y=66, label="Group␣2")
# p + annotate("text", x=3, y=48, label="Group 1", family="serif",
#              fontface="italic", colour="darkred", size=3) +
#   annotate("text", x=4.5, y=66, label="Group 2", family="serif",
#            fontface="italic", colour="darkred", size=3)
# p + annotate("text", x=2, y=55, label="Group 1", family="serif",
#              fontface="italic", colour="darkred", size=10,alpha=0.5) +
#   annotate("text", x=4.5, y=80, label="Group 2", family="serif",
#            fontface="italic", colour="darkred", size=10,alpha=0.5)
```


waiting
90
80
70
60
50
2
3
4
5
eruptions
Group 1
Group 2