STAT 40001/MA59800 **Statistical Computing** **Fall 2017**

Homework 2-Solution

Name:

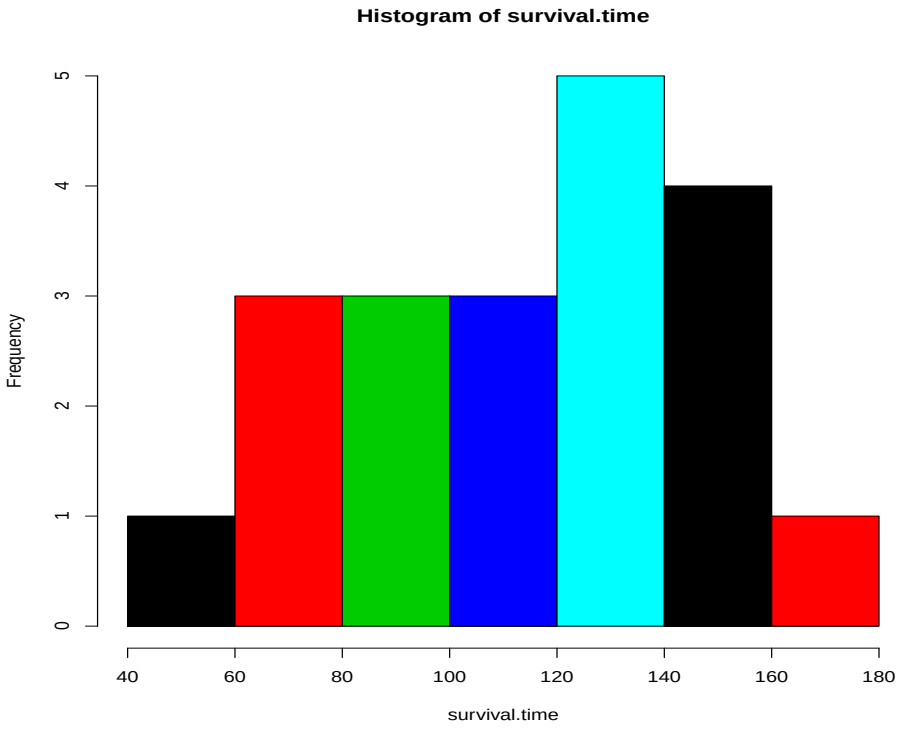
Due : September 21, 2017 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output with ERRORS:). Some of the questions below can be answered with very little or no programming. However, write R code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) The data frame ***Rat*** from the ***PASWR*** package has the survival time in weeks of 20 male rats exposed to high levels of radiation. Draw a histogram of the survival times of the rats.

Solution: We access the Rat dataframe using the code below

```
>library(PASWR)
>data(Rat, package="PASWR")
>attach(Rat)
> Rat
   survival.time
1            152
2            152
3            115
4            109
5            137
6             88
7             94
8             77
9            160
10           165
11           125
12            40
13           128
14           123
15           136
16           101
17            62
18           153
19            83
20            69
> hist(survival.time,col=c(1,2,3,4,5))
```



Q.N. 2) A data set 'Gapminder' is available in the attached file with this assignment. Gapminder contains data on life expectancy, population and GDP for 142 countries from 1952 to 2007.Below are the variables:
country = Name of the country
continent= Name of five contienent
year=ranges from 1952 to 2007 in increments of 5 years
lifeExp= life expectancy at birth(in years)
pop= population
gdpPercap=GDP per capita

a) Import the data in R and print first 5 rows(observations)
b) Install and load the library ggplot2.
c) How many unique countries are represented per continent?
d) Map 'gdpPercap' to the x-axis and 'lifeExp' to the y-axis.
e) Add points to the plot, make the points size 3 and map continent onto the aesthetics of the point and print the graph. Change the scale of x-axis.
Hint: See section 6.1 and 6.2 in the link below and follow the instructions.
http://nagraj.net/bims8382-textbook/data-visualization-with-ggplot2.html

Solution: We will save the data in local drive and import data in R using R code bellow

```
> data=read.csv("C://Aryal//Purdue-Northwest//STAT 40001//Assignments//Gapminder.csv")
> head(data,5)
      country continent year lifeExp      pop gdpPercap
1 Afghanistan      Asia 1952  28.801  8425333  779.4453
2 Afghanistan      Asia 1957  30.332  9240934  820.8530
3 Afghanistan      Asia 1962  31.997 10267083  853.1007
4 Afghanistan      Asia 1967  34.020 11537966  836.1971
5 Afghanistan      Asia 1972  36.088 13079460  739.9811
```

b) We use R code below to install ggplot2

```
>install.packages("ggplot2")
>library(ggplot2)
```

*c) We could use subset function and choose unique countries for each continent but the code below using count function in*plyr *package*

```
> library(plyr)
> count(country, "continent")
  continent freq
1    Africa   52
2  Americas   25
3      Asia   33
4    Europe   30
5   Oceania    2
```

d) We use R code below in ggplot2 to display the information

```
library(ggplot2)
ggplot(data, aes(x = gdpPercap, y = lifeExp))
ggplot(data, aes(x = gdpPercap, y = lifeExp)) + geom_point()
ggplot(data, aes(x = gdpPercap, y = lifeExp)) + geom_point()
+geom_point(aes(color=continent))
```

See figure 1.
We can change the scaling using the R code below

```
ggplot(data, aes(x = gdpPercap, y = lifeExp)) + geom_point()
+geom_point(aes(color=continent), size=3)+scale_x_log10()
```

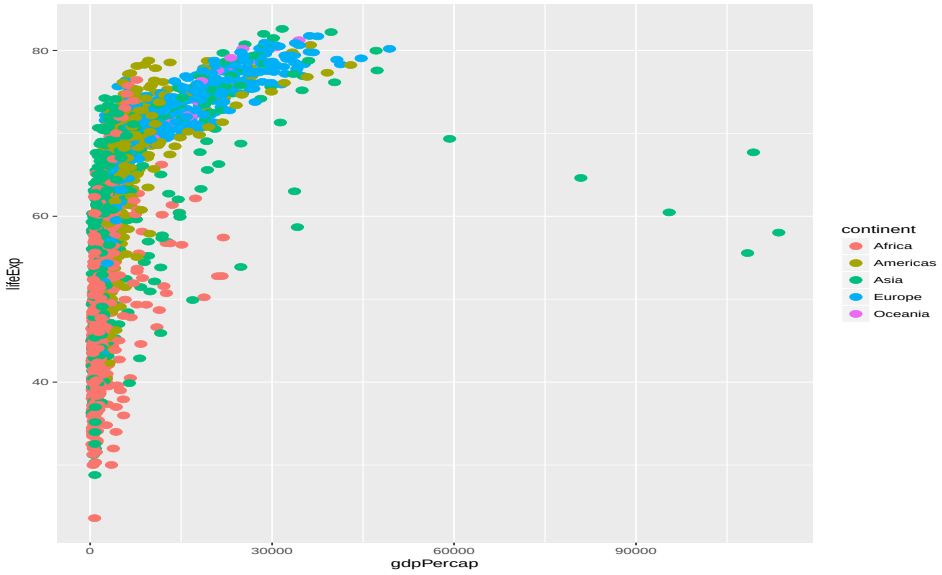


Figure 1: Gapmider data display without changing the scale

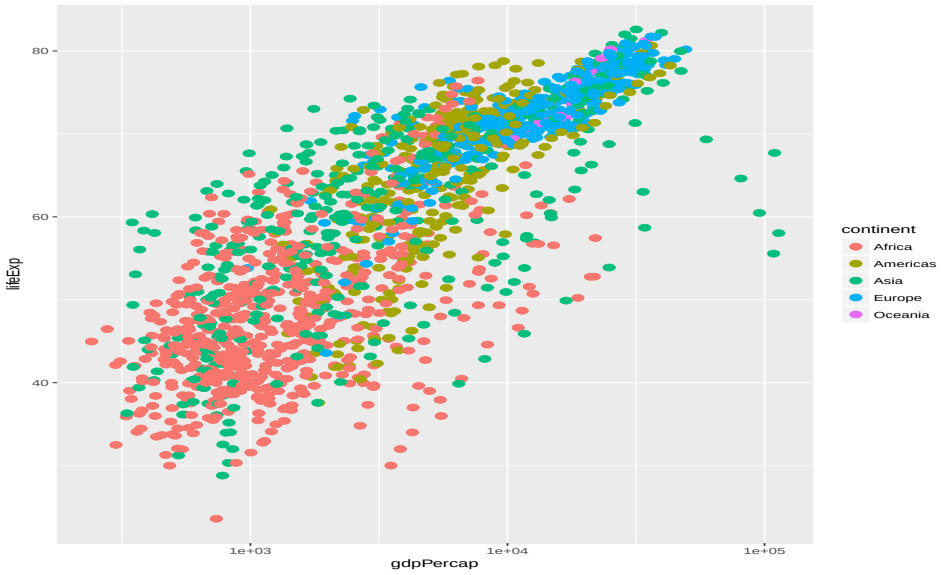
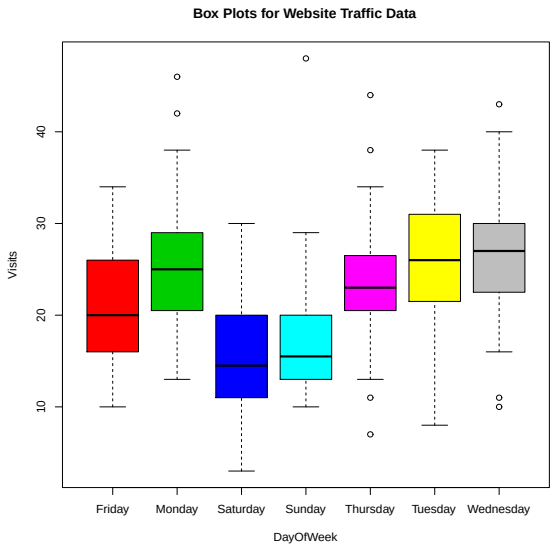


Figure 2: Gapmider data display after changing the scale

Q.N. 3) The number of visits to a website on each day by visitors is recorded. If a user accesses the site after 30 minutes of inactivity, that will be logged as a new visit. The data is available in the Blackboard as "website traffic".
a) Create a chart (side-by-side box plot) that shows the variability in website traffic for each day of the week.
b) Recreate the graph to display the box plot in the order of the days of a week.
b) Calculate the numerical summary of the website traffic data for each day of the week.

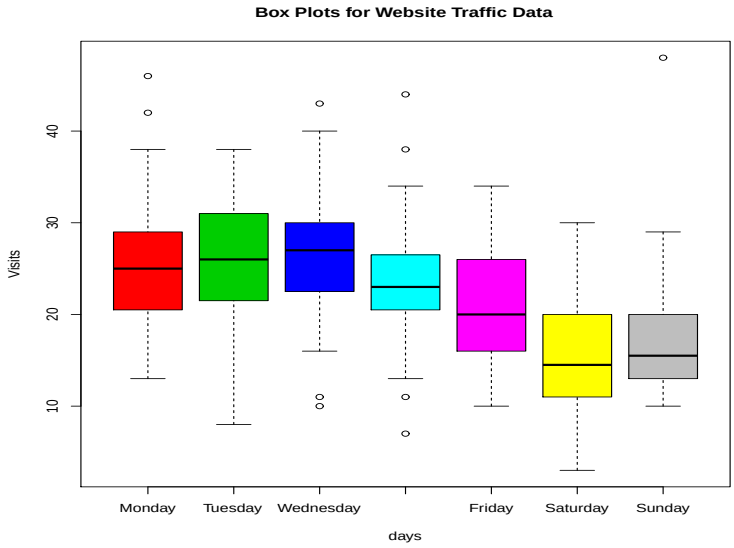
Solution:
a) We can use R code below to import the data and draw side-by-side box plot to display the information

```
data=read.csv("C://STAT 40001//Assignments//website traffic.csv")
attach(data)
names(data)
plot(Visits~DayOfWeek, col=c(2,3,4,5,6,7,8), main="Box Plots for Website Traffic Data")
```



b) Note that the previous codes produce the box-plots in the alphabetical order of the days. In order to place the days of the week in order we can use the R code below

```
data=read.csv("C://STAT 40001//Assignments//website traffic.csv")
attach(data)
days=factor(DayOfWeek, c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Su
plot(Visits~days, col=c(2,3,4,5,6,7,8), main="Box Plots for Website Traffic Data")
```



c) We can use R code below to calculate the numerical summary broken down by the day of the week

```
> tapply(Visits, days, summary)
$Monday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   20.50   25.00   25.32   29.00   46.00

$Tuesday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   21.50   26.00   25.77   31.00   38.00

$Wednesday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   22.50   27.00   26.74   30.00   43.00

$Thursday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   20.50   23.00   23.71   26.50   44.00

$Friday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   16.25   20.00   20.77   26.00   34.00

$Saturday
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   11.25   14.50   15.27   19.75   30.00

$Sunday
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   13.00   15.50   17.63   20.00   48.00
```

Q.N. 4) Table 1 and Table 2 below are the test scores of 10 students in Test 1 and Test 2

Name	Test 1
Ana	56
Brian	78
Cathy	87
Dough	89
John	95
Lucas	98
Marcus	59
Nabin	78
William	87
Zoe	98

Table 1: Test 1 Scores

Name	Test2
Ana	86
Brian	67
Cathy	78
Dough	89
John	87
Lucas	67
Marcus	94
Nabin	78
William	81
Zoe	83

Table 2: test 2 scores

a) Use ***merge(.,.)*** to create a single table containing the student's test 1 and test 2 scores.
b) How many students did better in the second test?
c) How many students did better in the first test?
d) How many students have the same score in both tests?
e) Calculate the average and standard deviation of both tests.

Solution: We can save both tables and import them individually and merge them. Alternatively, we can use simply use the following codes to merge them

```
Names=c("Ana","Brian","Cathy","Dough","John","Lucas","Marcus","Nabin","William","Zoe")
Test1=c(56,78,87,89,95,98,59,78,87,98)
Test2=c(86,67,78,89,87,67,94,78,81,83)
Table1=data.frame(Names,Test1)
```

```
Table2=data.frame(Names,Test2)

> merge(Table1, Table2)
     Names Test1 Test2
1      Ana    56    86
2    Brian    78    67
3    Cathy    87    78
4    Dough    89    89
5     John    95    87
6    Lucas    98    67
7   Marcus    59    94
8    Nabin    78    78
9  William    87    81
10     Zoe    98    83
```

b) How many students did better in the second test?
Solution:

```
> sum(Test2>Test1)
[1] 2
```

There are 2 students that did better on the second test.
c) How many students did better in the first test?
Solution:

```
> sum(Test1>Test2)
[1] 6
```

There are 6 students that did better on the first test.
d) How many students have the same score in both tests?
Solution:

```
> sum(Test1==Test2)
[1] 2
```

There are 2 students who have the same score in both tests.
e) Calculate the average and standard deviation of both tests.
Solution: below are the mean and standard deviation of test 1 and test 2

```
>mean(Test1)
 [1] 82.5
> mean(Test2)
  [1] 81
> sd(Test1)
  [1] 14.96106
> sd(Test2)
 [1] 8.869423
```

Q.N. 5) The data set provided below concerns the Auto-MPG.
https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data

The list of the variables are provided below

1. mpg:	2. cylinders	3. displacement
4. horsepower	5. weight	6. acceleration
7. model year	8. origin	9. car name

a) Import the data in R
b) Please replace the variables $\boldsymbol{V_1, V_2, \cdots, V_9}$ by the names names provided in the table above
c) There are some missing values marked as "?". Please remove these missing value and identify the dimension of complete data.
d) Create a parallel box-plot for (complete data) to display the mpg by number of cylinders.

Solution:
a) We can use R code below to import the data

```
data=read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-
head(data,5)
```

b) We can use R code below to change the name of the variable

```
>  names(data)=c("mpg","cylinders","displacement","horsepower","weight","acceleration",
              "model year","origin", "car name")
>  head(data,5)
  mpg cylinders displacement horsepower weight acceleration model year origin                 car name
1  18         8          307      130.0   3504         12.0         70      1 chevrolet chevelle malibu
2  15         8          350      165.0   3693         11.5         70      1         buick skylark 320
3  18         8          318      150.0   3436         11.0         70      1        plymouth satellite
4  16         8          304      150.0   3433         12.0         70      1             amc rebel sst
5  17         8          302      140.0   3449         10.5         70      1               ford torino
>  dim(data)
[1] 398   9
```

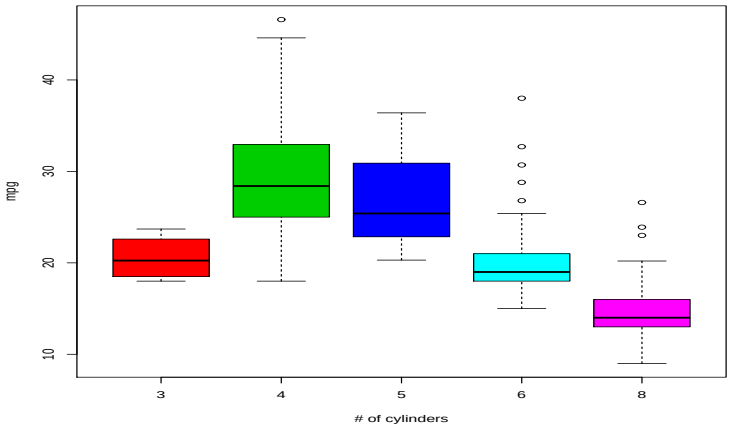
c) In order to remove missing values "?" first we need to convert them to "NA" and then remove

```
data1=read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
 na.strings="?")
head(data1,5)
> dim(data1)
[1] 398   9
> data2=na.omit(data1)
> dim(data2)
[1] 392   9
```

Note that 6 observations with missing values have been removed.

d) We can use R code below to create a box plot

```
plot(data2$V1~factor(data2$V2), xlab="# of cylinders", ylab="mpg", col=c(2,3,4,5,6))
```



Q.N. 6) Access the data from url ***http : //www.stat.berkeley.edu/users/statlabs/data/vote.data*** and store the information in an object named **vote** using the function *read.table()*. This includes the 1988 Stockton Primary Exit Poll Survey:
a) How many variables are included in the survey? Please print the variables.
b) One of the variable included is the voter's race. Note that following code are used.

$$\mathbf{0 = missing, 1 = White, 2 = Hispanic, 3 = Black, 4 = Asian, 5 = Other}$$

Display the distribution of the voter's race graphically.
Solution: We can then use the following R command to access the data

```
>vote<-read.table('http://www.stat.berkeley.edu/~statlabs/data/vote.data', header=T)
> names(vote)
[1] "precinct" "candidate" "race" "income"
> dim(vote)
[1] 1867    4
```

There are four variables included in the data set.

b) One of the variable included is the voter's race. Note that following code are used.

$$\mathbf{0 = missing, 1 = White, 2 = Hispanic, 3 = Black, 4 = Asian, 5 = Other}$$

Display the distribution of the voter's race graphically.

Solution: Since race is a categorical data we can display it graphically either using bargraph or piechart. R code below can be used to draw the pie chart as shown :

```
> vote<-read.table('http://www.stat.berkeley.edu/~statlabs/data/vote.data', header=T)
> table=table(vote$race)
> names(table)=c("missing", "White","Hispanic","Black","Asian", "Other")
> pie(table,col=c(1,2,3,4,5,6), main="Race distribution of 1988 Stockton Primary Exit Poll Survey")
```

Race distribution of 1988 Stockton Primary Exit Poll Survey