Homework 1

Amin Baabol

Instructions

Answer all questions stated in each problem. Discuss how your results address each question.

Submit your answers as a pdf, typeset (knitted) from an Rmd file. Include the Rmd file in your submission. You can typeset directly to PDF or typeset to Word then save to PDF In either case, both Rmd and PDF are required. If you are having trouble with .rmd, let us know and we will help you.

This file can be used as a template for your submission. Please follow the instructions found under "Content/Begin Here" titled . No code should be included in your PDF submission unless explicitly requested. Use the `echo = F` flag to exclude code from the typeset document.

For any question requiring a plot or graph, answer the question first using standard R graphics (See ?graphics). Then provide a equivalent answer using `library(ggplot2)` functions and syntax. You are not required to produce duplicate plots in answers to questions that do not explicitly require graphs, but it is encouraged.

You can remove the `Instructions` section from your submission.

Please answer the following questions from **Handbook of Statistical Analyses in R** (HSAUR) and the written questions. Refer to **R Graphics Cookbook or Modern Data Science with R** for any ggplots.

1. Question 1.1, pg. 23 in **HSAUR**. *You will need to make some assumptions to answer this question. State how you interpret the question and list your assumptions*. Problem 1.1:Calculate the median profit for the companies in the United States and the median profit for the companies in the UK, France and Germany.

Assumming that the data stored in the package "HSAUR" is cleaned, meaning it has a complete data available on the countries we're performing the query on and that any missing data/values are taken out.

#Interpretation As indicated by the results the median profit for the companies in the United States has the highest median, followed by Germany. This isn't surprising knowing that the United States is the largest economy in the world while Germany is the biggest economy in Europe.

```
##          country median
## 1         France  0.190
## 2        Germany  0.230
```

```
## 3 United Kingdom  0.205
## 4  United States  0.240
```

2. Question 1.2, pg. 23 in **HSAUR** Problem 1.2:Find all German companies with negative profit.

#Interpretation

There are only thirteen companies on this list, I would argue this is a sign of a healthy economy that only thirteen companies reported a negative profit. I'm not sure of this is their quarterly reporting or fiscal/calender year.

```
##  [1] "Allianz Worldwide"       "Deutsche Telekom"
##  [3] "E.ON"                    "HVB-HypoVereinsbank"
##  [5] "Commerzbank"             "Infineon Technologies"
##  [7] "BHW Holding"             "Bankgesellschaft Berlin"
##  [9] "W&W-Wustenrot"           "mg technologies"
## [11] "Nurnberger Beteiligungs" "SPAR Handels"
## [13] "Mobilcom"
```

3. Question 1.3, pg. 23 in **HSAUR** Problem 1.3: Which business category are most of the companies situated at the Bermuda island working in?

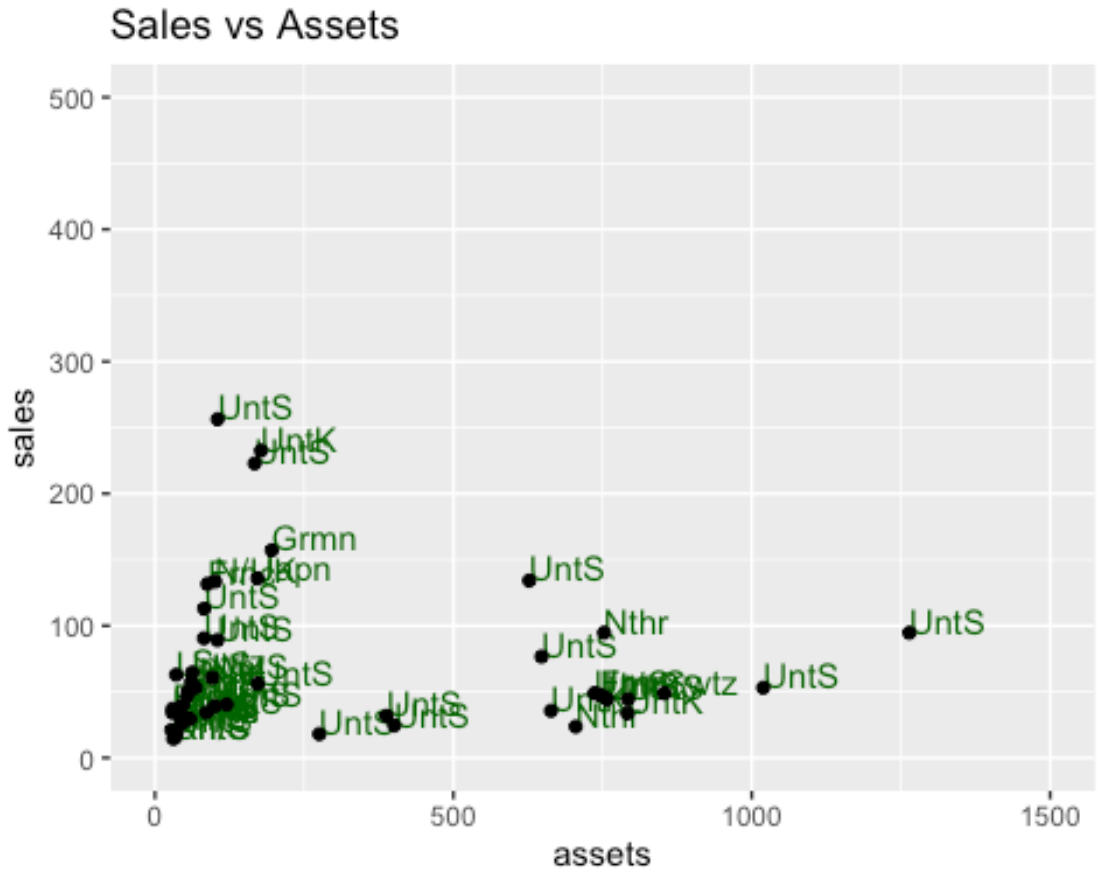
#Interpretation Insurance is the single largest market in the Bermuda Island, followed by Conglomerates and oil & gas operations. This makes sense since it's a popular travel destination.

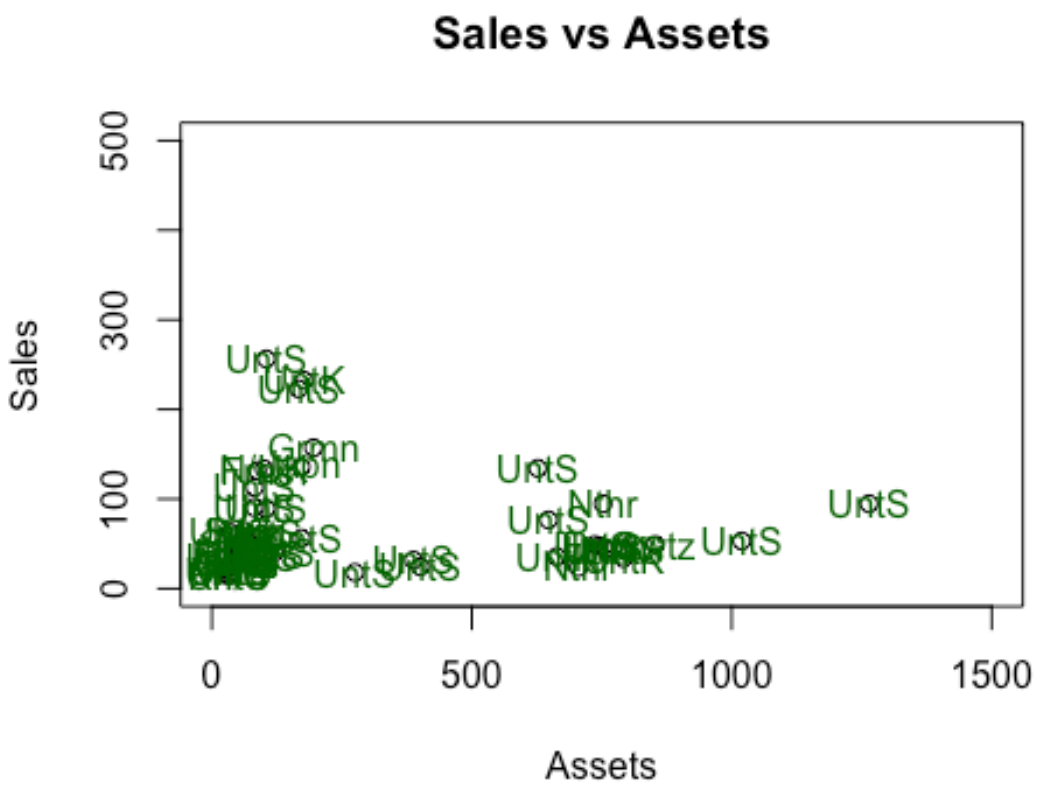
```
##
##                         Insurance                     Conglomerates
##                                10                                 2
##             Oil & gas operations                           Banking
##                                 2                                 1
##                    Capital goods              Food drink & tobacco
##                                 1                                 1
##                     Food markets                             Media
##                                 1                                 1
##              Software & services               Aerospace & defense
##                                 1                                 0
##     Business services & supplies                         Chemicals
##                                 0                                 0
##                     Construction                 Consumer durables
##                                 0                                 0
##           Diversified financials             Drugs & biotechnology
##                                 0                                 0
## Health care equipment & services     Hotels restaurants & leisure
##                                 0                                 0
##    Household & personal products                         Materials
##                                 0                                 0
##                          Retailing                   Semiconductors
##                                 0                                 0
##  Technology hardware & equipment      Telecommunications services
```

```
##                                0                                0
##               Trading companies                    Transportation
##                                0                                0
##                        Utilities
##                                0
```

4. Question 1.4, pg. 23 in **HSAUR** Problem 1.4:For the 50 companies in the Forbes data set with the highest profits, plot sales against assets (or some suitable transformation of each variable), labelling each point with the appropriate country name which may need to be abbreviated (using abbreviate) to avoid making the plot look too 'messy'.c

#Interpretation Both y-axis snd x-axis is measured in $billions. The plot indicates that most companies sales are under 100billion and have equally assets worth less than 100 billionfor most countries. There are outliers as indicated in the plot but because most data falls between 0 and 100-250 billion the probability of the mean falling somewhere there is high under normal gaussian distribution.





5.

Question 1.5, pg. 23 in **HSAUR** Problem 1.5:Find the average value of sales for the companies in each country in the Forbes data set, and find the number of companies in each country with profits above 5 billion US dollars

Assummuing that the NA values are removed from the calculations which will increase the bias of the data, also the average sales value is calculated after filtering out the companies with less than $5bil profit which is starkly different from when it is not filtered.

#Interpretation The average sales seems to be all over the place, meaning there are quiet a lot of countries making sales on average over 5 billion USD, however, when I filtered the forbes data using the 5 billion USD profits as a constraint not many countries made this selection. The first output shows all companies including their respective countries and their average sale, the second output shows the companies that made at least 5 billion USD, the average sales of the country in which they are based. Since profit is the difference between revenue and operation cost, I would expect the companies with at least 5 billion profit to be from the developed countries.

```
##                          Country Average Sales
## 1                         Africa      6.820000
## 2                      Australia      5.244595
## 3     Australia/ United Kingdom     11.595000
## 4                        Austria      4.142500
## 5                        Bahamas      1.350000
```

```
## 6                       Belgium     10.114444
## 7                       Bermuda      6.840500
## 8                        Brazil      6.338667
## 9                        Canada      6.429643
## 10               Cayman Islands      1.660000
## 11                        Chile      1.602500
## 12                        China      5.099600
## 13               Czech Republic      1.805000
## 14                      Denmark      6.349000
## 15                      Finland     10.291818
## 16                       France     20.102063
## 17       France/ United Kingdom      1.010000
## 18                      Germany     20.781385
## 19                       Greece      2.528333
## 20              Hong Kong/China      2.044000
## 21                      Hungary      3.370000
## 22                        India      3.868148
## 23                    Indonesia      2.450000
## 24                      Ireland      4.765000
## 25                      Islands      6.670000
## 26                       Israel      2.060000
## 27                        Italy     10.213902
## 28                        Japan     10.190633
## 29                       Jordan      1.330000
## 30                    Kong/China      5.717500
## 31                         Korea     15.005000
## 32                      Liberia      3.780000
## 33                    Luxembourg     14.185000
## 34                      Malaysia      1.716250
## 35                        Mexico      3.937647
## 36                   Netherlands     17.020714
## 37  Netherlands/ United Kingdom     92.100000
## 38                   New Zealand      2.640000
## 39                        Norway     10.780000
## 40                      Pakistan      1.230000
## 41       Panama/ United Kingdom      5.930000
## 42                          Peru      0.170000
## 43                   Philippines      1.565000
## 44                        Poland      4.410000
## 45                      Portugal      3.884286
## 46                        Russia      7.672500
## 47                     Singapore      3.685000
## 48                  South Africa      4.124000
## 49                   South Korea      7.969333
## 50                         Spain      7.843448
## 51                        Sweden      7.665769
## 52                   Switzerland     12.456765
## 53                        Taiwan      2.751429
## 54                      Thailand      2.513333
## 55                        Turkey      4.713333
```

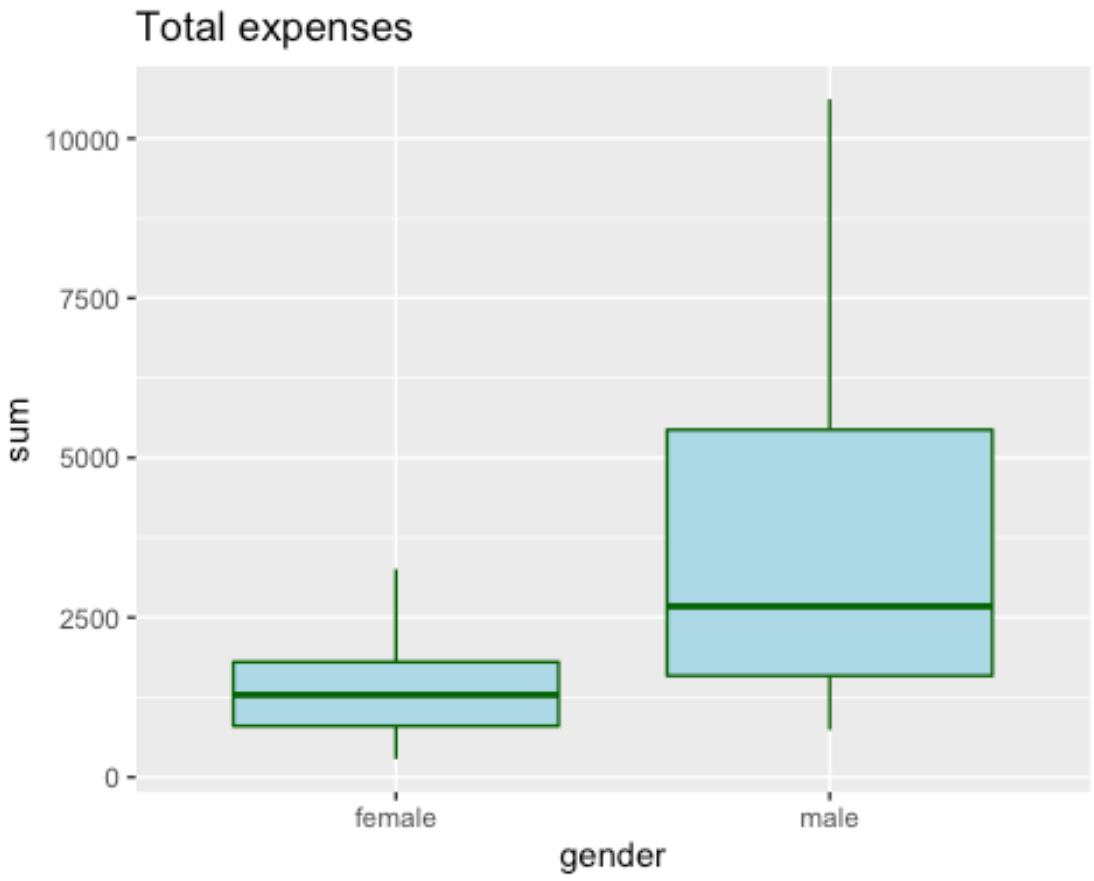
```
## 56                United Kingdom     10.445109
## 57    United Kingdom/ Australia     10.010000
## 58  United Kingdom/ Netherlands      7.540000
## 59 United Kingdom/ South Africa      2.060000
## 60                 United States     10.058256
## 61                     Venezuela      0.980000

##                         Country Average Sales Num. of Companies
## 1                         China       29.5300                 1
## 2                        France      131.6400                 1
## 3                       Germany      157.1300                 1
## 4                         Japan      135.8200                 1
## 5 Netherlands/ United Kingdom      133.5000                 1
## 6                   South Korea       50.2200                 1
## 7                   Switzerland       46.7600                 3
## 8                United Kingdom      103.6867                 3
## 9                 United States       77.2835                20
```

6. Question 2.1, pg. 41 in **HSAUR** Problem 2.1:2.1 The data in Table 2.3 are part of a data set collected from a survey of household expenditure and give the expenditure of 20 single men and 20 single women on four commodity groups. The units of expenditure are Hong Kong dollars, and the four commodity groups are housing: housing, including fuel and light,food: foodstuffs, including alcohol and tobacco, goods: other goods, including clothing, footwear and durable goods, services: services, including transport and vehicles.The aim of the survey was to investigate how the division of household expenditure between the four commodity groups depends on total expenditure and to find out whether this relationship differs for men and women.

Use appropriate graphical methods to answer these questions and state your conclusions.

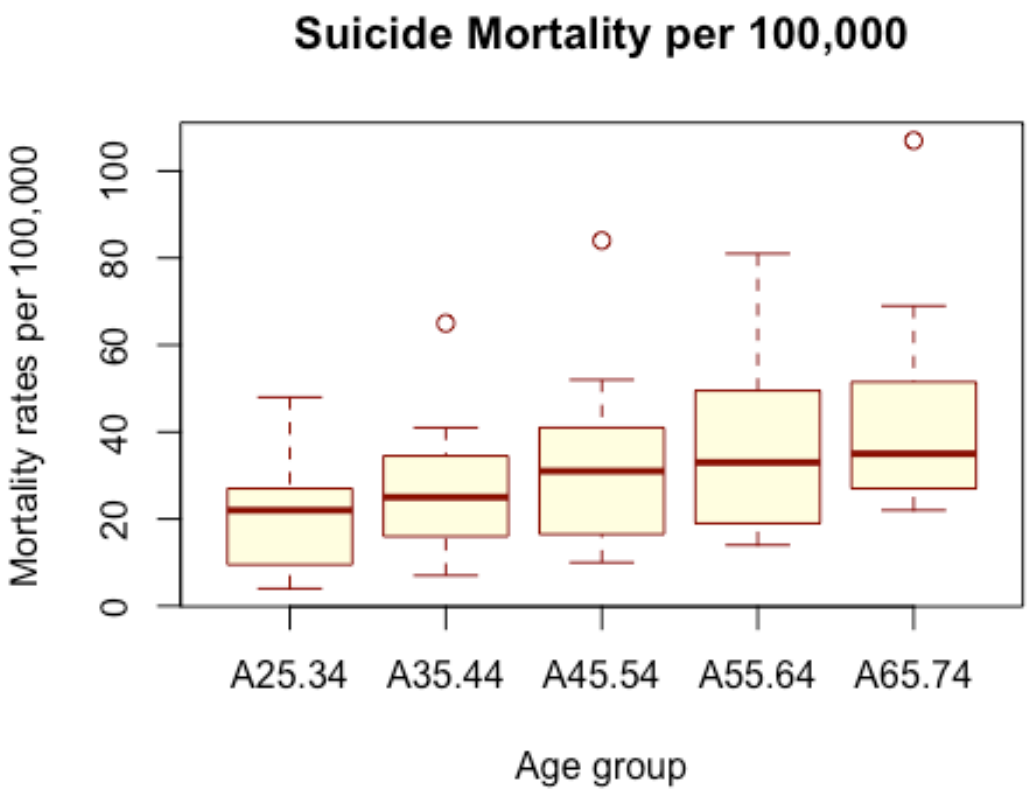
#Interpretion contrary to popular opinion, these plots indicate men out-spend women on almost every category

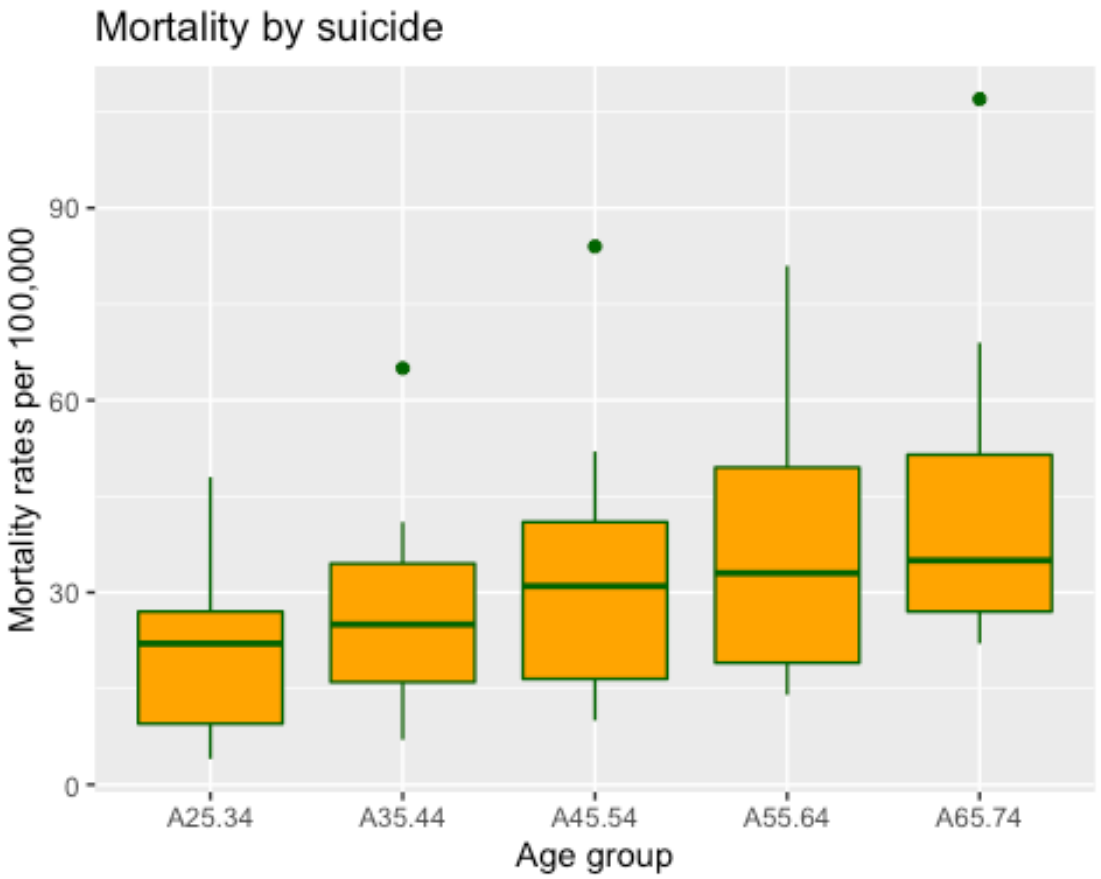

Total Expenditure Gender Comparison
Expenditure
0
4000
8000
female
male
Gender

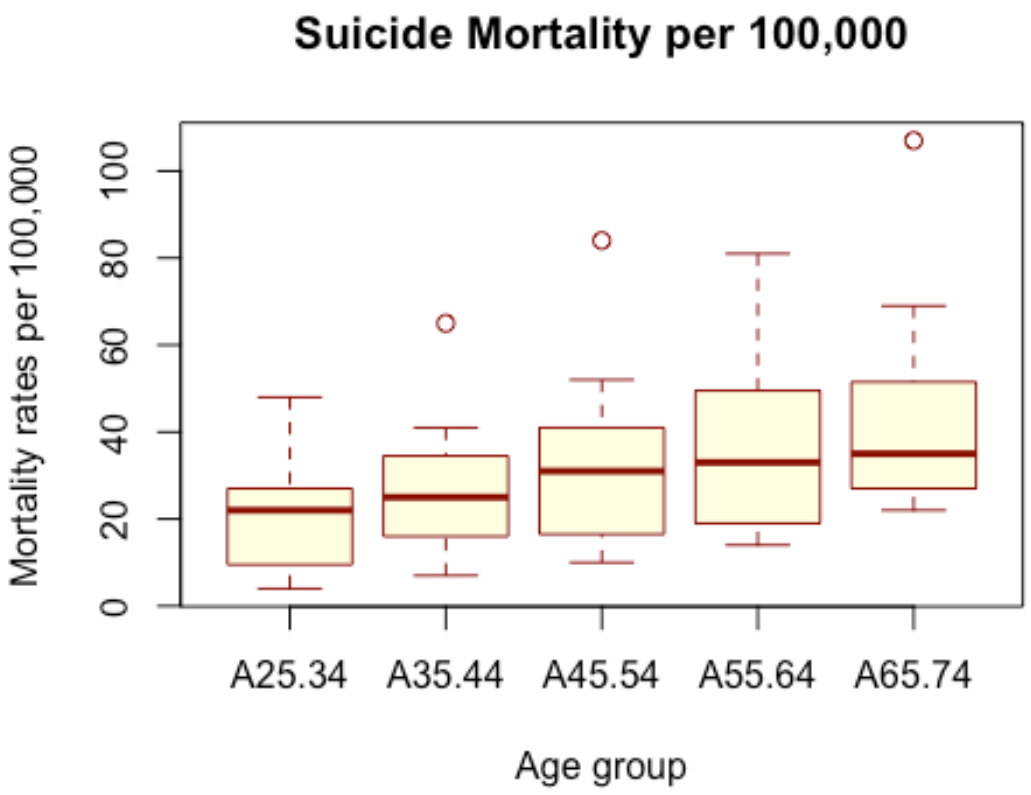


7. Question 2.3, pg. 44 in **HSAUR** Problem 2.3:Mortality rates per 100, 000 from male suicides for a number of age groups and a number of countries are given in Table 2.4. Construct sideby- side box plots for the data from different age groups, and comment on what the graphic tells us about the data.

#Interpretion Unfortunately, this plot indicates that there is an increasing suicide mortality rate for males as age goes up, with male ages 55 and older having the highest suicide rates


Suicide Mortality per 100,000
Mortality rates per 100,000
0
20
40
60
80
100
A25.34
A35.44
A45.54
A55.64
A65.74
Age group


Mortality by suicide
Mortality rates per 100,000
90
60
30
0
A25.34
A35.44
A45.54
A55.64
A65.74
Age group



8. Using a single R statement, calculate the median absolute deviation, $1.4826 \cdot median|x - \hat{\mu}|$, where $\hat{\mu}$ is the sample median. Use the dataset . Use the R function `mad()` to verify your answer.

```
## [1] 91.9212
```

```
## [1] 91.9212
```

9. Using the data matrix , find the state with the minimum per capita income in the New England region as defined by the factor . Use the vector to get the state name.

```
##       income  name    division
## Maine   3694 Maine New England
```

10. Use subsetting operations on the dataset to find the vehicles with highway mileage of less than 25 miles per gallon (variable) and weight (variable) over 3500lbs. Print the model name, the price range (low, high), highway mileage, and the weight of the cars that satisfy these conditions.

```
##          Model Price MPG.highway Weight
## 16 Lumina_APV  16.3          23   3715
## 17       Astro  16.6          20   4025
## 26    Caravan  19.0          21   3705
## 56        MPV  19.1          24   3735
## 66      Quest  19.1          23   4100
```

```
## 70 Silhouette  19.5          23   3715
## 89    Eurovan  19.7          21   3960
## 36   Aerostar  19.9          20   3735
## 87     Previa  22.7          22   3785
## 28    Stealth  25.8          24   3805
## 63   Diamante  26.1          24   3730
## 49      ES300  28.0          24   3510
## 50      SC300  35.2          23   3515
## 48        Q45  47.9          22   4000
```

11. Form a matrix object named from the variables from the dataframe from the package. Use it to create a list object named containing named components as follows:

a) A vector of means, named

b) A vector of standard errors of the means, named

12. Use the function on the three-dimensional array to compute:

a) Sample means of the variables , for each of the three species

```
##          Setosa Versicolor Virginica
## Sepal L.  5.006      5.936     6.588
## Sepal W.  3.428      2.770     2.974
## Petal L.  1.462      4.260     5.552
## Petal W.  0.246      1.326     2.026
```

b) Sample means of the variables for the entire data set.

```
## Sepal L. Sepal W. Petal L. Petal W.
## 5.843333 3.057333 3.758000 1.199333
```

13. Use the data matrix and the function to obtain:

a) The mean per capita income of the states in each of the four regions defined by the factor

```
##     Northeast         South North Central          West
##      4570.222      4011.938      4611.083      4702.615
```

b) The maximum illiteracy rates for states in each of the nine divisions defined by the factor

```
##        New England    Middle Atlantic     South Atlantic
##                1.3                1.4                2.3
## East South Central West South Central East North Central
##                2.4                2.8                0.9
## West North Central           Mountain            Pacific
##                0.8                2.2                1.9
```

c) The number of states in each region

```
##     Northeast         South North Central          West
##             9            16            12            13
```

14. Using the dataframe , produce a scatter plot matrix of the variables . Use different colors to identify cars belonging to each of the categories defined by the variable in different colors.

```
#install.packages("GGally")
#creating variables for the size of the car,m
carsize = cut(mtcars[,"wt"], breaks=c(0, 2.5, 3.5, 5.5),
labels = c("Compact","Midsize","Large"))

carsize = cut(mtcars[,"wt"], breaks=c(0, 2.5, 3.5,
                                      5.5), labels =
c("Compact","Midsize","Large"))

car.data <- data.frame(mtcars, sizes = carsize)

data("mtcars")

# Using base R
pairs(~mpg + disp + hp + drat + qsec, data=car.data,
      col = car.data$sizes,
      main="mtcars Scatterplot Matrix")

# Using ggplot
library(ggplot2)
ggpairs(car.data[,c("mpg", "disp", "hp", "drat", "qsec")],
        mapping = ggplot2::aes(colour = car.data$sizes))
```

15. Use the function to perform a one-way analysis of variance on the data with as the treatment factor. Assign the result to an object named and use it to print an ANOVA table.

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## chickwts$feed  5 231129   46226   15.37 5.94e-10 ***
## Residuals     65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: chickwts$weight
##               Df Sum Sq Mean Sq F value    Pr(>F)
## chickwts$feed  5 231129   46226  15.365 5.936e-10 ***
## Residuals     65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = chickwts$weight ~ chickwts$feed)
##
```

```
## $`chickwts$feed`
##                             diff         lwr       upr     p adj
## horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
## linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
## meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
## soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
## sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
## linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
## meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
## soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
## sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
## meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
## soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
## sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
## soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
## sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
## sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

16. Write an R function named for conducting a one-sample t-test. Return a list object containing the two components:

    - the t-statistic named T;
    - the two-sided p-value named P.

Use this function to test the hypothesis that the mean of the variable (in the dataset) is equal to 240 against the two-sided alternative. `echo = T`

To reduce the number of unknown parameters I assummed a 95% confidence interval which in return makes alpha 0.05. Furthermore, my theoretical mean,mu is 240 and my actual mean from the chickwts data is is computed by utilizing the mean function with the chickwts weight as input. My ttest function there has one required input parameter and that is the chickwts dataset.

#Interpretation The t-test function i wrote gave me a p-value of 0.024 and a t-statistic value of 2.30. After comparing those values againts the built-in t-test function, it turns out my values are in agreement with the the values of the t-test built-in function. Therefore, a p-value of 0.024 is less than the confidence level(alpha) of 0.05, hence the null hypothesis can be safely reject and the true mean is NOT 240!

```
#answering the first part of the question: function for conductiong a one-
sample t-test
#installing and importing the necessary libraries
library(HSAUR3)
library(stats)
data(chickwts)
#taking a glance at chickwts data
#head(chickwts)

  ttest=function(y,mu,alpha){
```

```
    avg.weight=mean(y$weight)
    p1=qt(alpha/2,(nrow(y)-1))
    p2=qt(1-alpha/2,(nrow(y)-1))
    std.dv=sqrt(var(y$weight))
    n=nrow(y)

    T=(avg.weight-mu)/(std.dv/sqrt(nrow(y)))
    P=2*(1-pt(T,n))
    return (c(P,T))}

t_test <- ttest(chickwts,240,0.05)
print(t_test)

## [1] 0.02439824 2.29987903

#comparing my t,p values against the the t-test built-in function
t.test(chickwts$weight,mu=240)

##
##  One Sample t-test
##
## data:  chickwts$weight
## t = 2.2999, df = 70, p-value = 0.02444
## alternative hypothesis: true mean is not equal to 240
## 95 percent confidence interval:
##  242.8301 279.7896
## sample estimates:
## mean of x
##  261.3099
```

#Citation “A Handbook of Statistical Analyses Using R, third Edition” by Everitt and Hothorn R Graphics Cookbook” by Winston Chang published through O’Reilly (Basic guide to Grammar of Graphics in R) www.google.com www.stackoverflow.com