Data Analysis using Graphs from HSAUR

Yamuna Dhungana

The following questions are from **Handbook of Statistical Analyses in R** (HSAUR) and the written questions. Refer to **R Graphics Cookbook or Modern Data Science with R**

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.*

Here, let us assume the data we have the data given and we will remove all the NAs from the data.

```
##          country median
## 1         France  0.190
## 2        Germany  0.230
## 3 United Kingdom  0.205
## 4  United States  0.240
```

Here, we got the median of the profit for the four countries.

2. Question 1.2, pg. 23 in **HSAUR**

```
##  [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"
```

The basic concept of mathematics says that zero is the point of neutralization or the neutral point.Neither profit nor loss is seen at a point. Values less than zero are loss whereas, values greater than zero are profit. I am using the same concept for coding in this question.

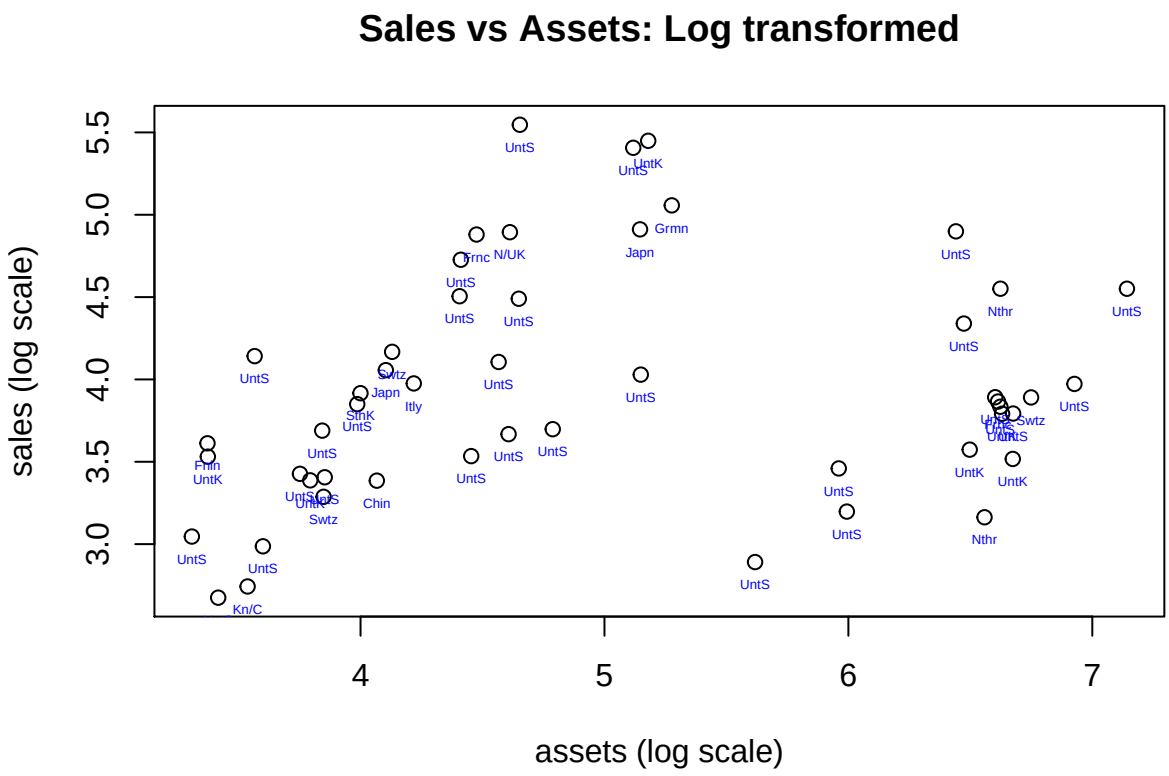
3. Question 1.3, pg. 23 in **HSAUR**

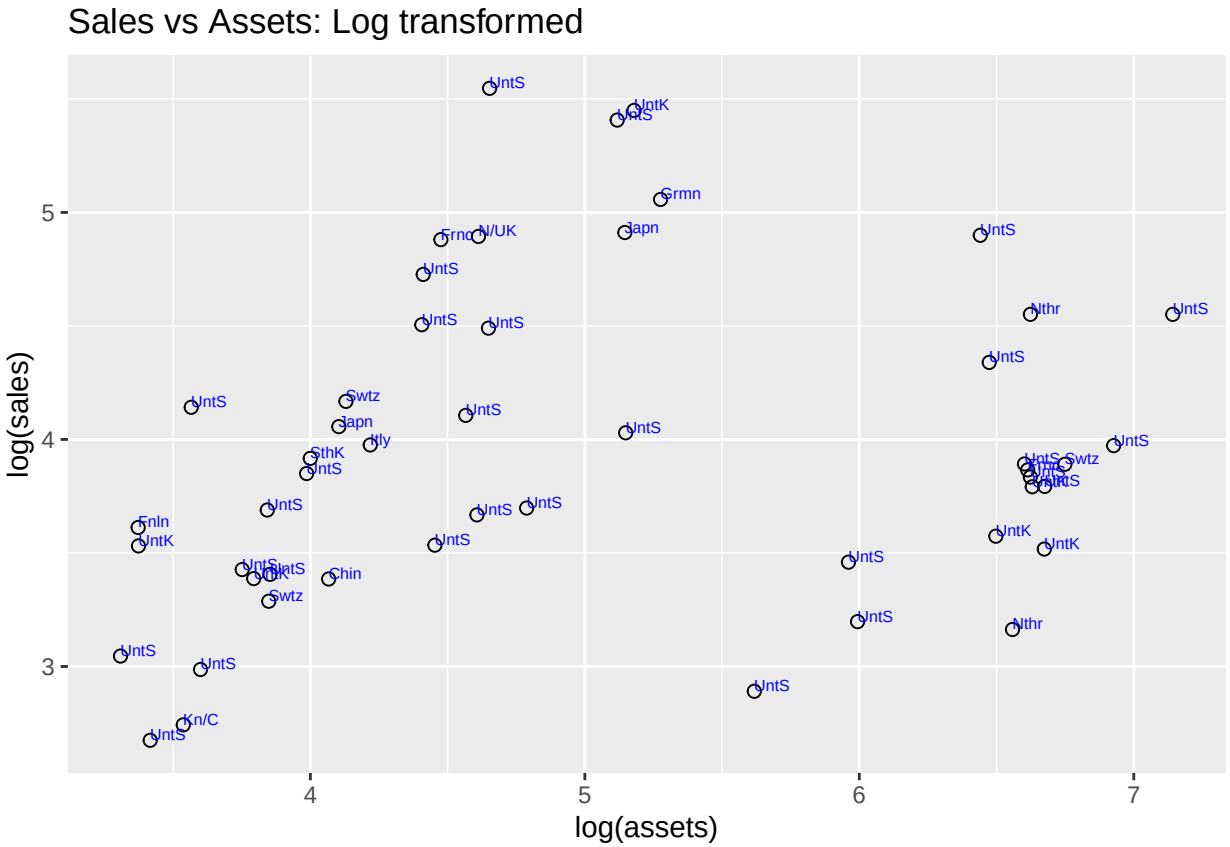
```
##
##                         Insurance                   Conglomerates
##                                10                               2
##             Oil & gas operations                         Banking
##                                2                               1
##                    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
```

The maximum no of business that the Burmuda island's company involved was insurance. Likewise, there were also involved in Conglomerates, Oil & gas operations, and six more other categories.

4. Question 1.4, pg. 23 in **HSAUR**





5. Question 1.5, pg. 23 in **HSAUR**

```
##                     country mean_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
```

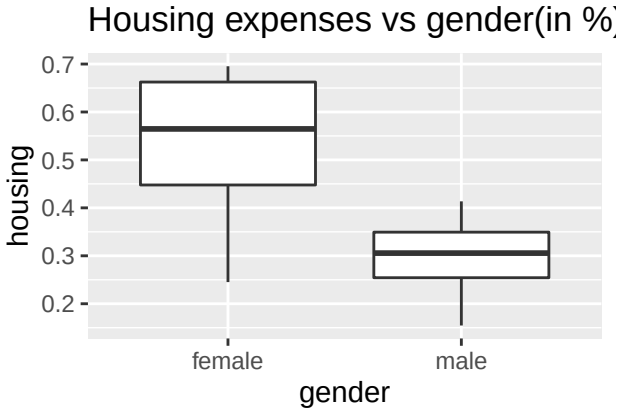
```
##                       country  n
## 1                       China  1
## 2                      France  1
## 3                     Germany  1
## 4                       Japan  1
## 5 Netherlands/ United Kingdom  1
## 6                 South Korea  1
## 7                 Switzerland  3
## 8              United Kingdom  3
## 9               United States 20
```

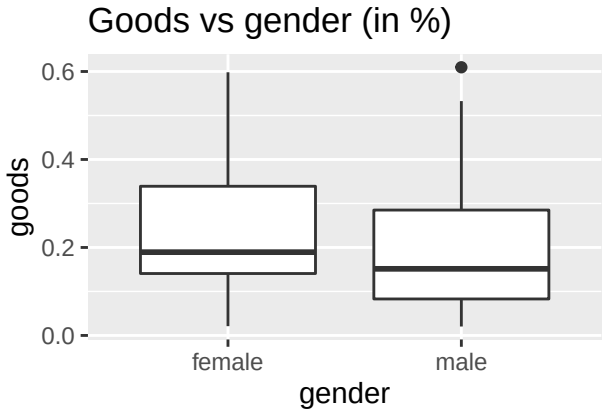
The first data denotes the mean sales of the company of the countries, the Second data denotes the number of companies that have profited more than 5 billion dollars.Here United states have the maximum no of companies that has the companies whose profit is more than 5 billion dollars. There were 20 such countries in the United States. Likewise, there were 3-3 companies in the United Kingdom and Switzerland.

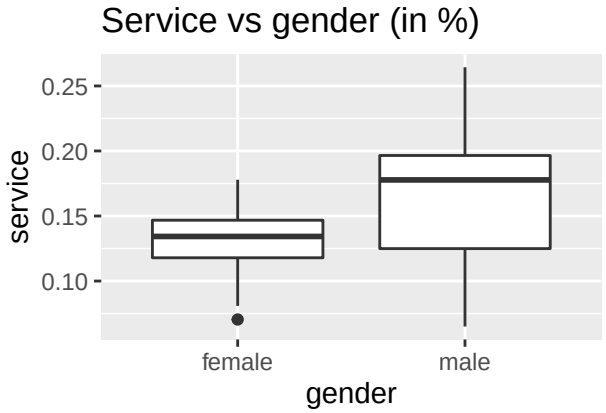
6. Question 2.1, pg. 41 in **HSAUR**

```
##   housing food goods service gender
```

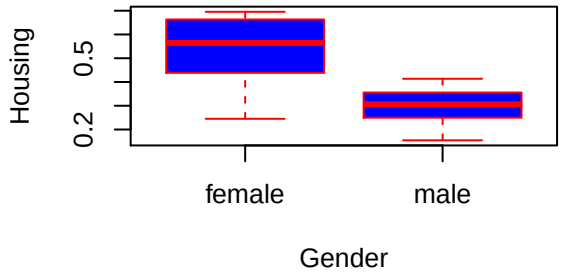
```
## 1     820  114   183     154 female
## 2     184   74     6      20 female
## 3     921   66  1686     455 female
## 4     488   80   103     115 female
## 5     721   83   176     104 female
## 6     614   55   441     193 female
```



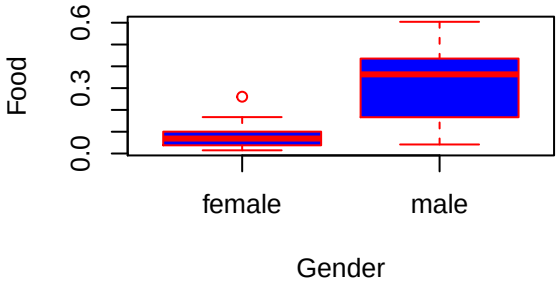




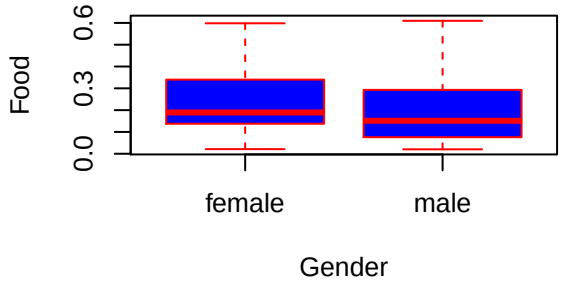
x plot plotted with housing against gende



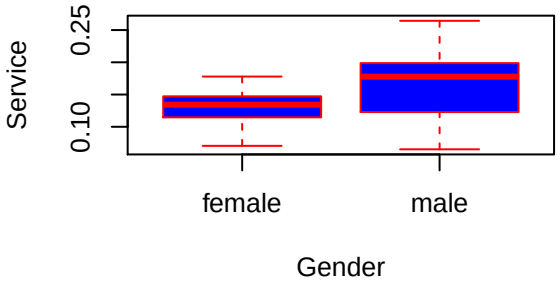
ox plot plotted with food against gender (

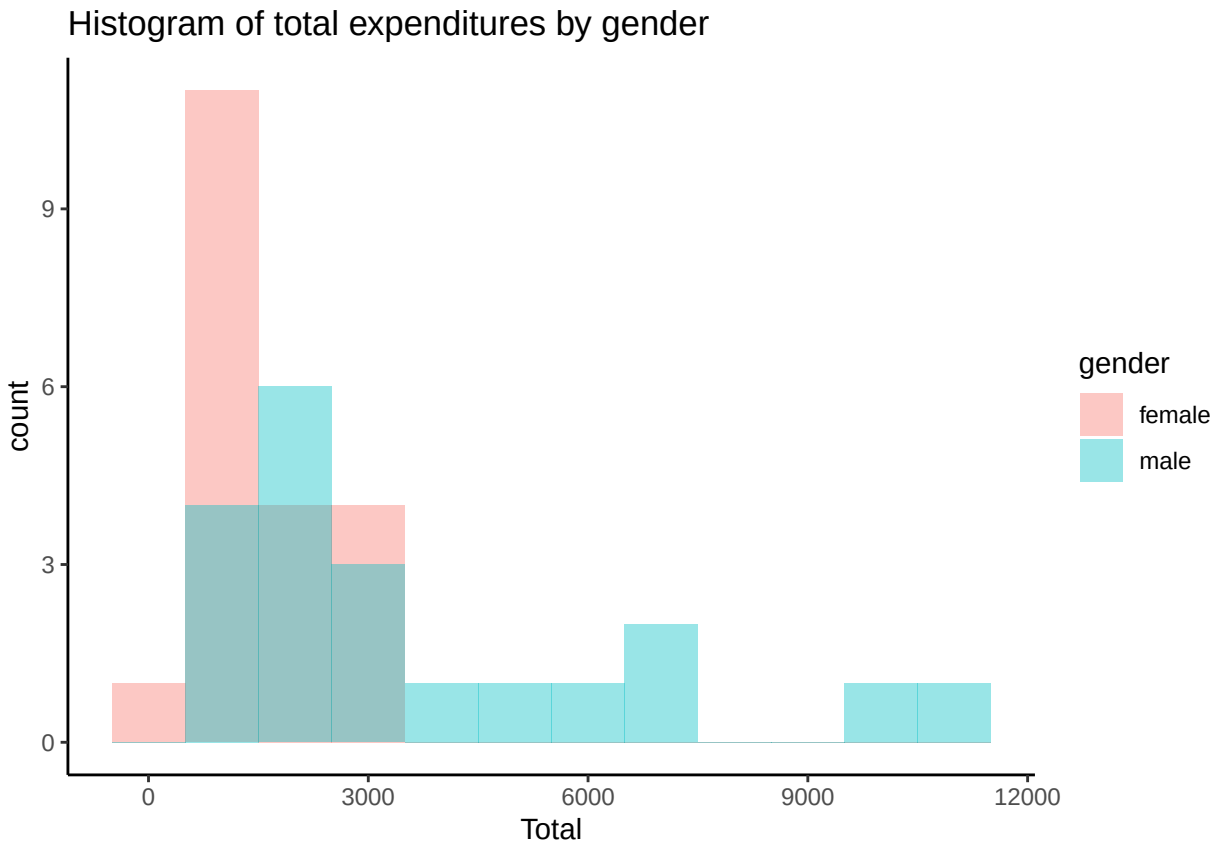


ox plot plotted with good against gender



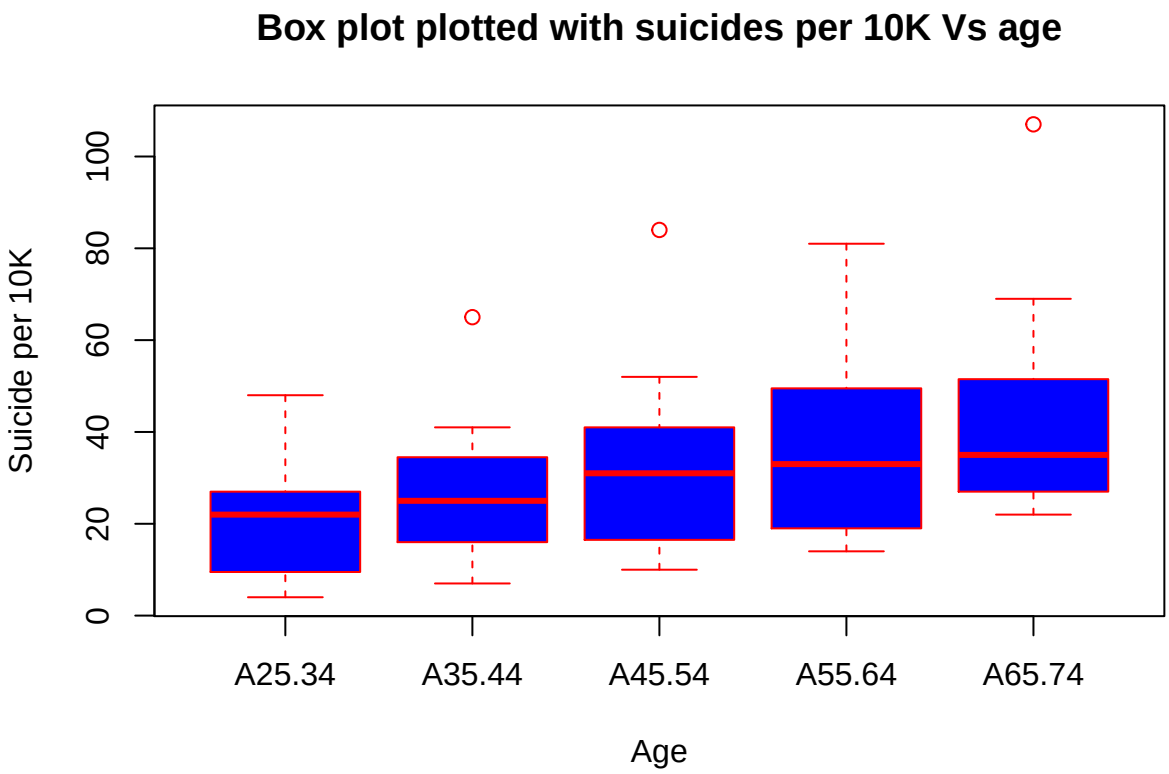
x plot plotted with service against gende

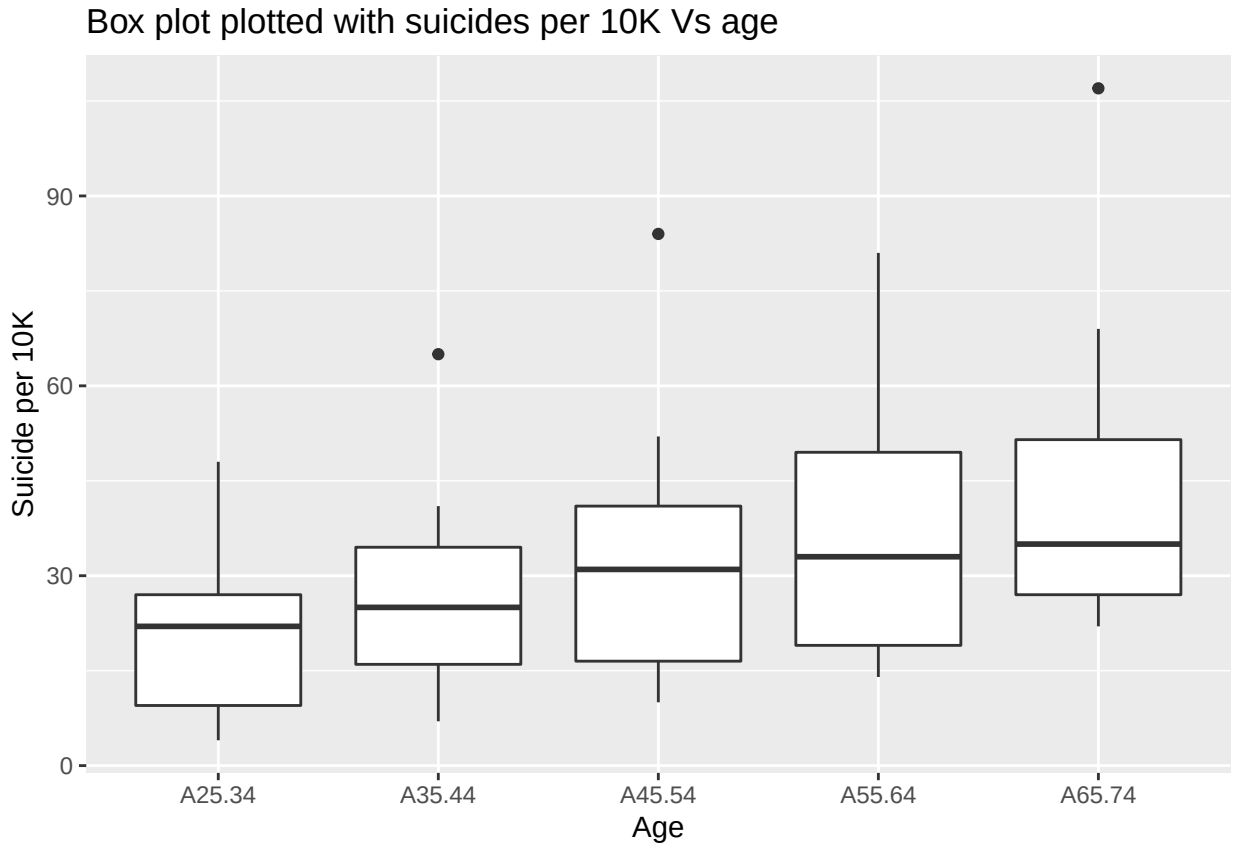




From the graph named, "Box plot plotted with total against gender", "Total expenses per gender" it is visible that the male spends more money on food, goods, service, and housing than the female.The graph-3 and graphs-4 are plotted with the housing, food, goods, and service against gender. These two graphs are similar only plotted with base R and ggplot. From the first graphs of graphs-3 and graphs-4, the expenditure on housing is much larger for female than a male. The second graph(right) shows the expanses in food. The female spends much less money on food than a male. The difference in the expenditure seems much larger.From the graphs, the lower-left denotes the graphs for goods. According to the graphs, the male and female expenditure is equal for good. Likewise, in the last graphs (last right) which is plotted for the service. This graph shows the expenditure on service, which is much less for females than males. The male spent more money on service.Hence, we can understand that males spend more money on food and service. Females spent more money on housing and both genders spent nearly equal money on goods. The last graph denotes the histograms for both females and males. The female spends less than 3000 whereas male spent less than 10,000.

7. Question 2.3, pg. 44 in **HSAUR**


Box plot plotted with suicides per 10K Vs age
Suicide per 10K
0
20
40
60
80
100
A25.34
A35.44
A45.54
A55.64
A65.74
Age



There are two graphs named, "Box plot plotted with suicide per 10k vs age". This is plotted with the same data. One of them is with base and the other with a plot. From the graphs, we can see the suicide among the different age groups. The age group between 35 to 54 has outliers. The plot of age 55 to 64 has a slightly large suicide rate than other age groups. The age group between 25 to 34 and 35 to 44 is similar. These two groups are slightly smaller than others.The other thing that we can see from the graphs is that the median of all the age groups is somewhat equal.

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 **chickwts**. Use the R function `mad()` to verify your answer.

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

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

Both the methods exihibit the same result.

9. Using the data matrix **state.x77**, find the state with the minimum per capita income in the New England region as defined by the factor *state.division*. Use the vector *state.name* to get the state name.

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

The state with the minimum per capiya income in the New England region is Maine with the income 3694.

10. Use subsetting operations on the dataset **Cars93** to find the vehicles with highway mileage of less than 25 miles per gallon (variable *MPG.highway*) and weight (variable *Weight*) 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 **mycars** from the variables *Min.Price, Max.Price, MPG.city, MPG.highway, EngineSize, Length, Weight* from the **Cars93** dataframe from the **MASS** package. Use it to create a list object named *cars.stats* containing named components as follows:

a) A vector of means, named *Cars.Means*

```
##   Min.Price   Max.Price    MPG.city MPG.highway  EngineSize      Length
##   17.125806   21.898925   22.365591   29.086022    2.667742  183.204301
##      Weight
## 3072.903226
```

b) A vector of standard errors of the means, named *Cars.Std.Errors*

```
## $Min.Price
## [1] 0.906921
##
## $Max.Price
## [1] 1.143805
##
## $MPG.city
## [1] 0.5827473
##
## $MPG.highway
## [1] 0.5528742
##
## $EngineSize
## [1] 0.1075695
##
## $Length
## [1] 1.514196
##
## $Weight
## [1] 61.16942
```

12. Use the `apply()` function on the three-dimensional array **iris3** to compute:

a) Sample means of the variables *Sepal Length, Sepal Width, Petal Length, Petal Width*, for each of the three species *Setosa, Versicolor, Virginica*

```
##          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 *Sepal Length, Sepal Width, Petal Width* 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 **state.x77** and the `tapply()` function to obtain:

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

```
##     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 *state.division*

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

c) The number of states in each region

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

14. Using the dataframe **mtcars**, produce a scatter plot matrix of the variables *mpg, disp, hp, drat, qsec.* Use different colors to identify cars belonging to each of the categories defined by the *carsize* variable in different colors.

15. Use the function `aov()` to perform a one-way analysis of variance on the **chickwts** data with *feed* as the treatment factor. Assign the result to an object named *chick.aov* and use it to print an ANOVA table.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## 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
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ feed, data = 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 `ttest()` 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 *weight* variable (in the **chickwts** dataset) is equal to 240 against the two-sided alternative. *For this problem, please show the code of function you created as well as show the output. You can do this by adding* `echo = T` *to the code chunk header.*

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

```
##
##  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
```

```
## [1] "T value and two sided P values returned by the funtion: "
```

```
##            P        T
## 1 0.02439824 2.299879
```

```
## Hypothesis Result:
```

```
## [1] "Rejected! The true mean is NOT 240 !!"
```