Homework 1

Snigdha Peddi

9/5/2020

Question 1: Calculate the median profit for the companies in the US and the median profit for the companies in the UK,France and Germany.

Answer: To answer this question I have used *dplyr* package to aggregate and summarize data. I used group by function to group the data by country and then used summarize function to calculate the median profit making sure to remove any NA values(na.rm=True) while summarizing. Then I filtered the countries, United States,United Kingdom, France and Germany.

```
## # A tibble: 4 x 2
##   country        Median_Profit
##   <fct>                  <dbl>
## 1 France                 0.19
## 2 Germany                0.23
## 3 United Kingdom         0.205
## 4 United States          0.24
```

Question 2: Find all German companies with negative profit

Answer: I have used *dplyr* package to select the columns name, country and profits which are more relevant for this problem. Then grouped by country and filtered data where country is Germany and profits are less than zero.

```
## # A tibble: 6 x 3
## # Groups:   country [1]
##   name                   country profits
##   <chr>                  <fct>     <dbl>
## 1 Allianz Worldwide      Germany   -1.23
## 2 Deutsche Telekom       Germany  -25.8
## 3 E.ON                   Germany   -0.73
## 4 HVB-HypoVereinsbank    Germany   -0.87
## 5 Commerzbank            Germany   -0.31
## 6 Infineon Technologies Germany   -0.51
```

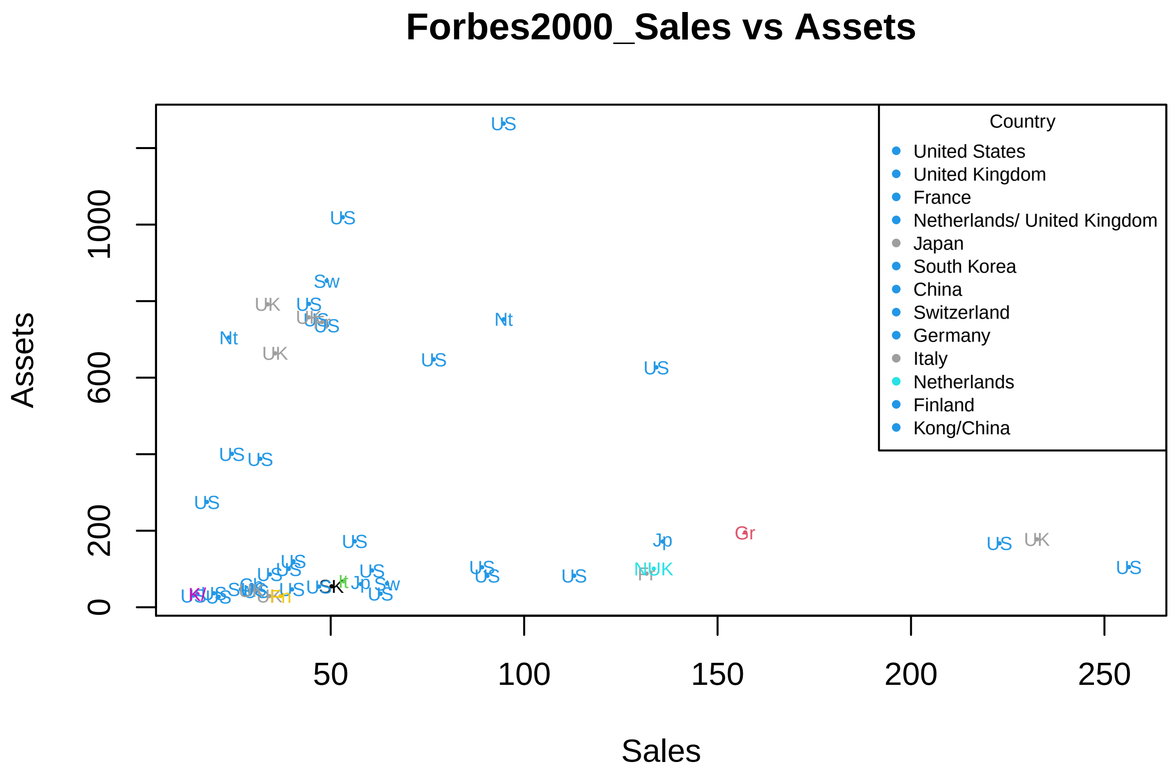
Question 3: To which business category do most of the Bermuda island companies belong?

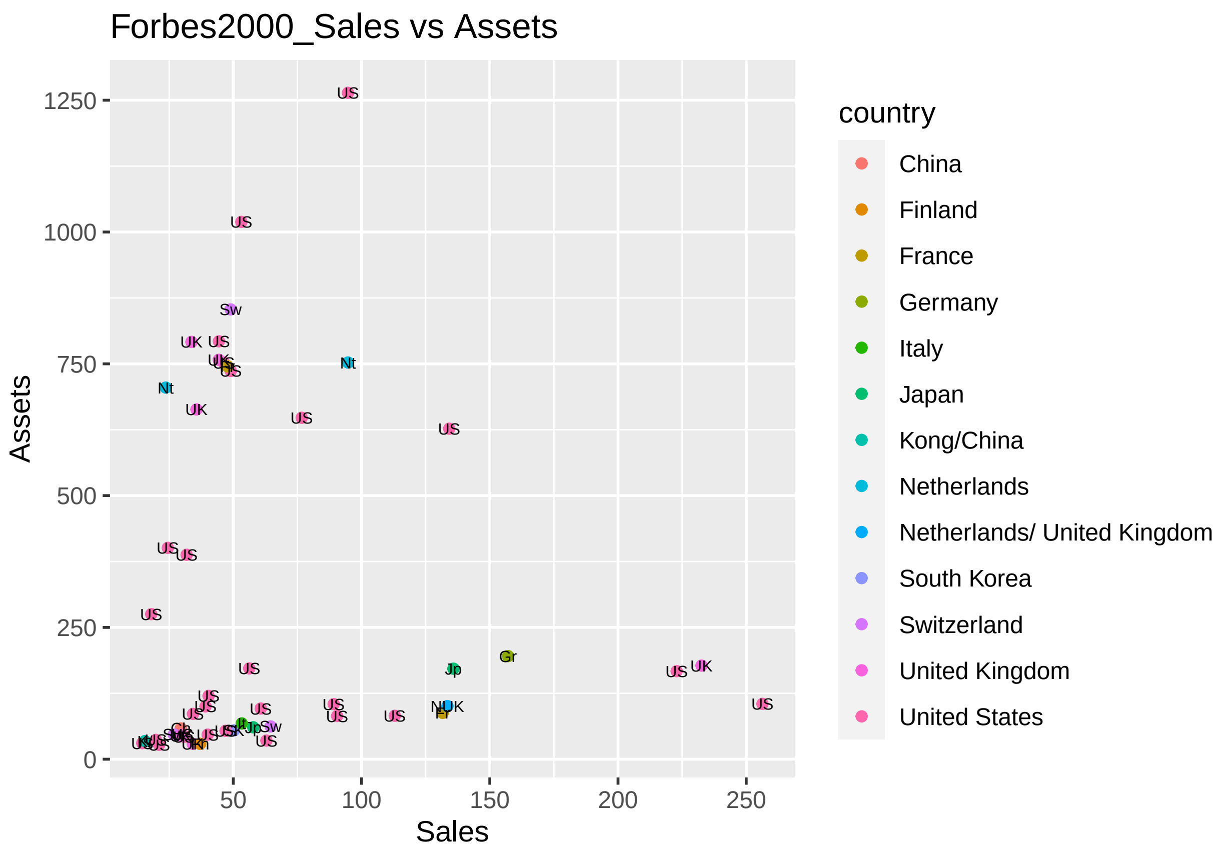
Answer: I have selected the relevant variables from the Forbes2000 data set. Filtered the data set by country and selected all the data that belongs to Bermuda.Then used count function from *dplyr* package to count the number of unique categories the companies of the country belong to. I have sorted that observations(count) to have highest at the top and then sliced the data set to display the category to which most of the Bermuda companies belong to.

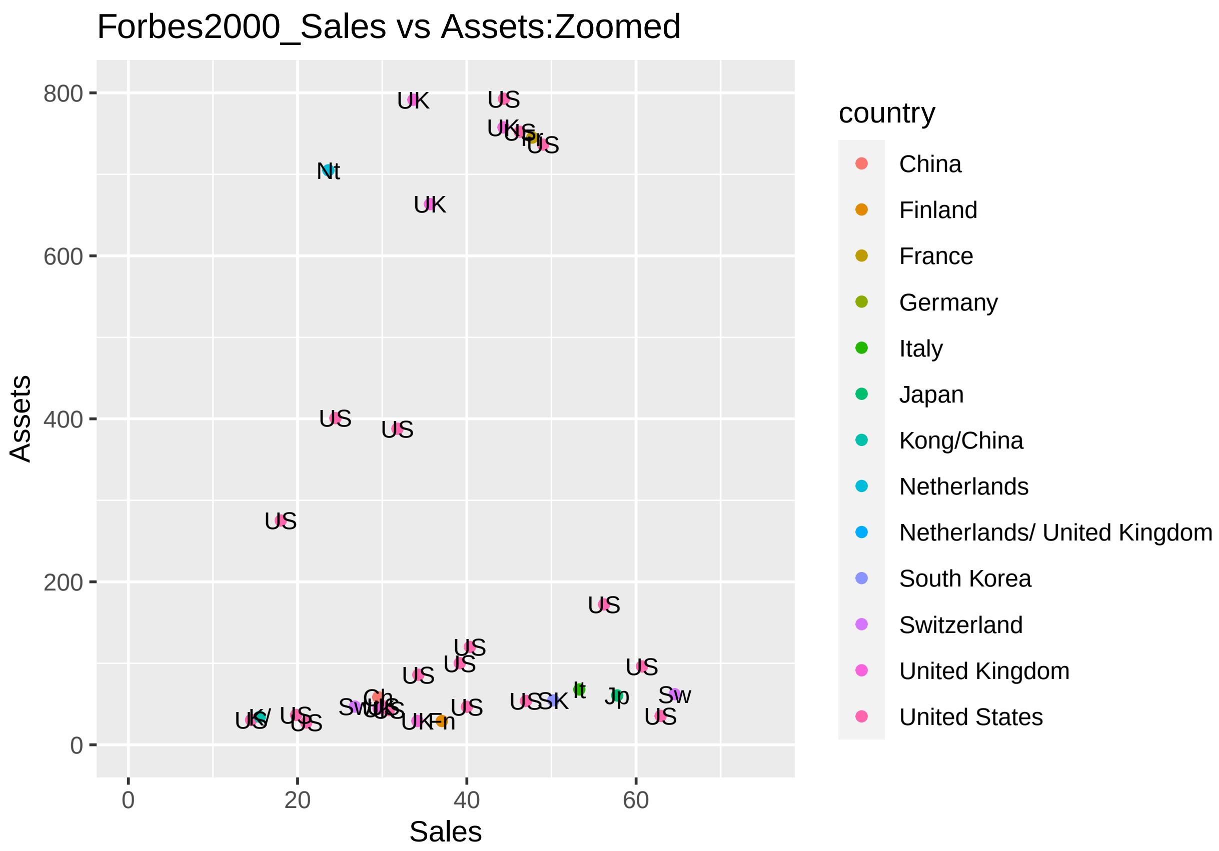
```
##   country  category  n
## 1 Bermuda Insurance 10
```

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

Answer: For this Question, I have used functions, select and arrange from *dplyr* package. I have selected relevant variables and used arrange function to sort the data set from highest profits to lowest profits. Subset top 50 rows to have top 50 profit making companies. Used this Subset of data to plot Sales vs assets using base R function and also using ggplot.Used geom_text to add abbreviated country labels.As the plot is overcrowded, zoomed the plot to see the trend by resetting the scale of x axis to 0 and 75 and y axis to 0 and 800.




Forbes2000_Sales vs Assets
Assets
1250
1000
750
500
250
0
Sales
50
100
150
200
250
country
China
Finland
France
Germany
Italy
Japan
Kong/China
Netherlands
Netherlands/ United Kingdom
South Korea
Switzerland
United Kingdom
United States
US
Sw
UK
Nt
Jp
Gr
Fr
NtUK



Question 5: Find the average values 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

Answer: Grouped the data by country and then computed the mean sales using summarize function. Filtered the Forbes data to get the companies making profits over 5 billion, grouped by country,summarized the length of the companies and arranged by highest at the top.

```
## # A tibble: 6 x 2
##   country                     avg_sales
##   <fct>                           <dbl>
## 1 Africa                           6.82
## 2 Australia                        5.24
## 3 Australia/ United Kingdom       11.6 
## 4 Austria                          4.14
## 5 Bahamas                          1.35
## 6 Belgium                         10.1 
```

```
## # A tibble: 6 x 2
##   country             n
##   <fct>           <int>
## 1 United States      20
## 2 Switzerland         3
## 3 United Kingdom      3
## 4 China               1
## 5 France              1
## 6 Germany             1
```

Question 6: The data in the household data table 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,service 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.

Answer:

Plot 1 and Plot 2

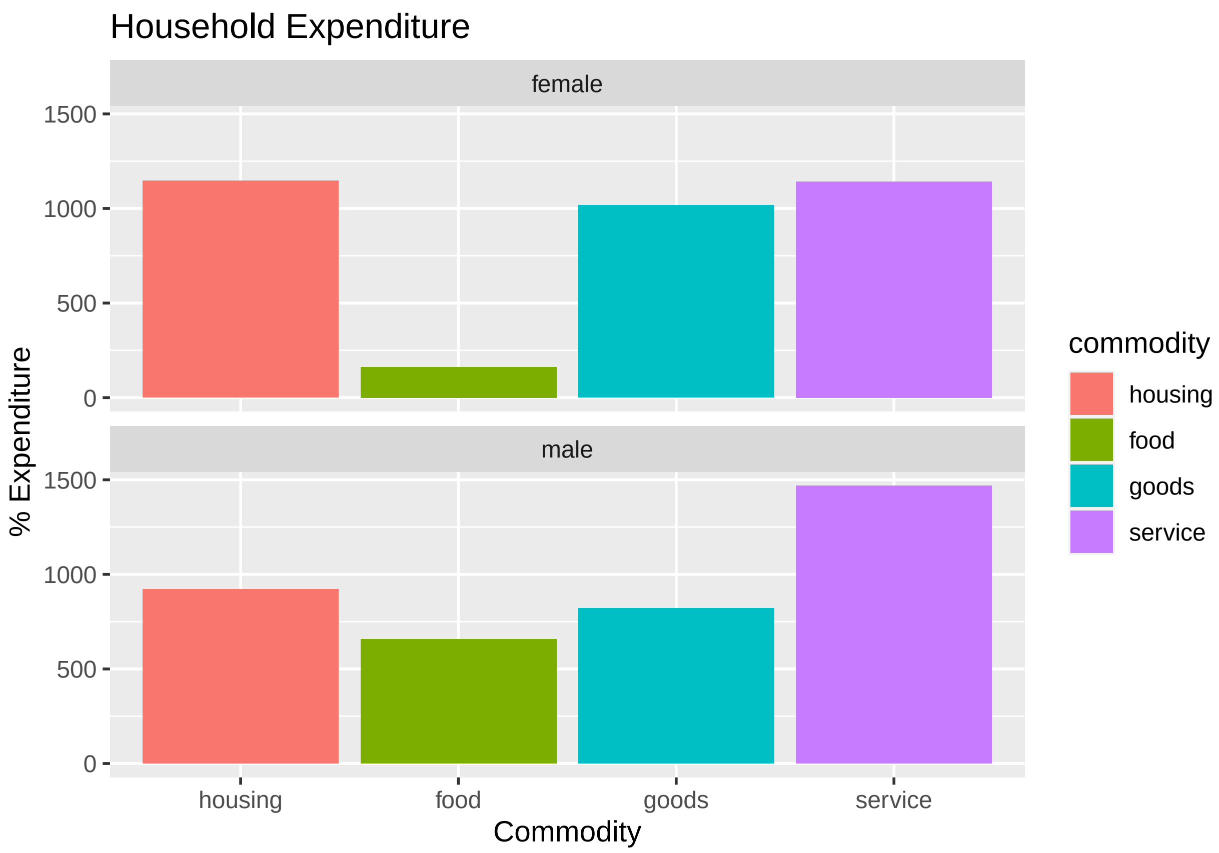
To answer this question, first I have used mutate function from *dplyr* package to compute the percent expenditure of four commodities against the total expenditure. Then used gather function from *tidyr* package to convert the data from wide form to long form with categorical variables to commodity and percentage values to Percent_Expenditure. Verified the data and Used ggplot to geom_bar to plot a barplot with Percent expenditure on Y axis and Commodities on X axis, and facet wrapped by gender. This plot clearly shows that the amount of money spent by Males and Females differs between each commodity. For example, males tend to spend more on services whereas females tend to spend more on housing.Then,plotted total_expenditure on Y axis and gender on X axis to investigate the spending behavior of males vs females and it was observed that on average males spend more than females.

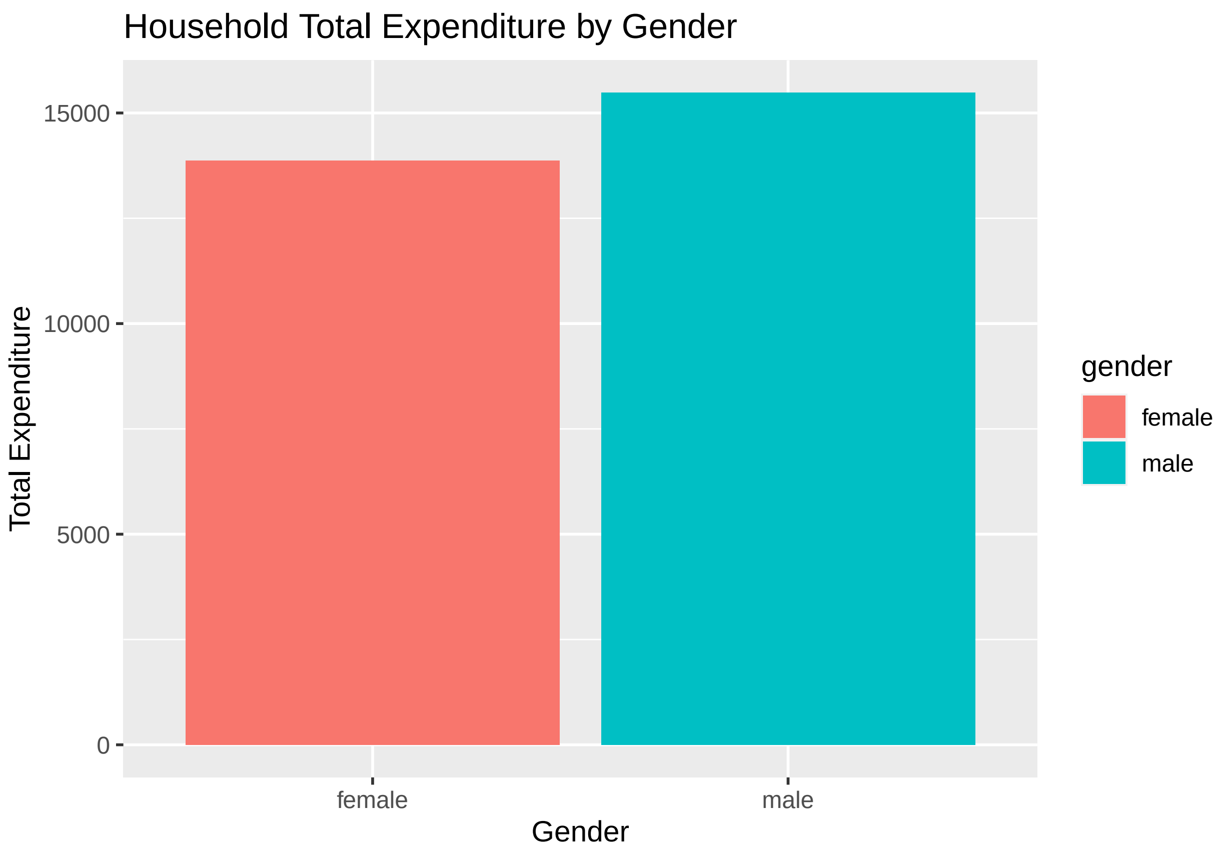
Plot 3

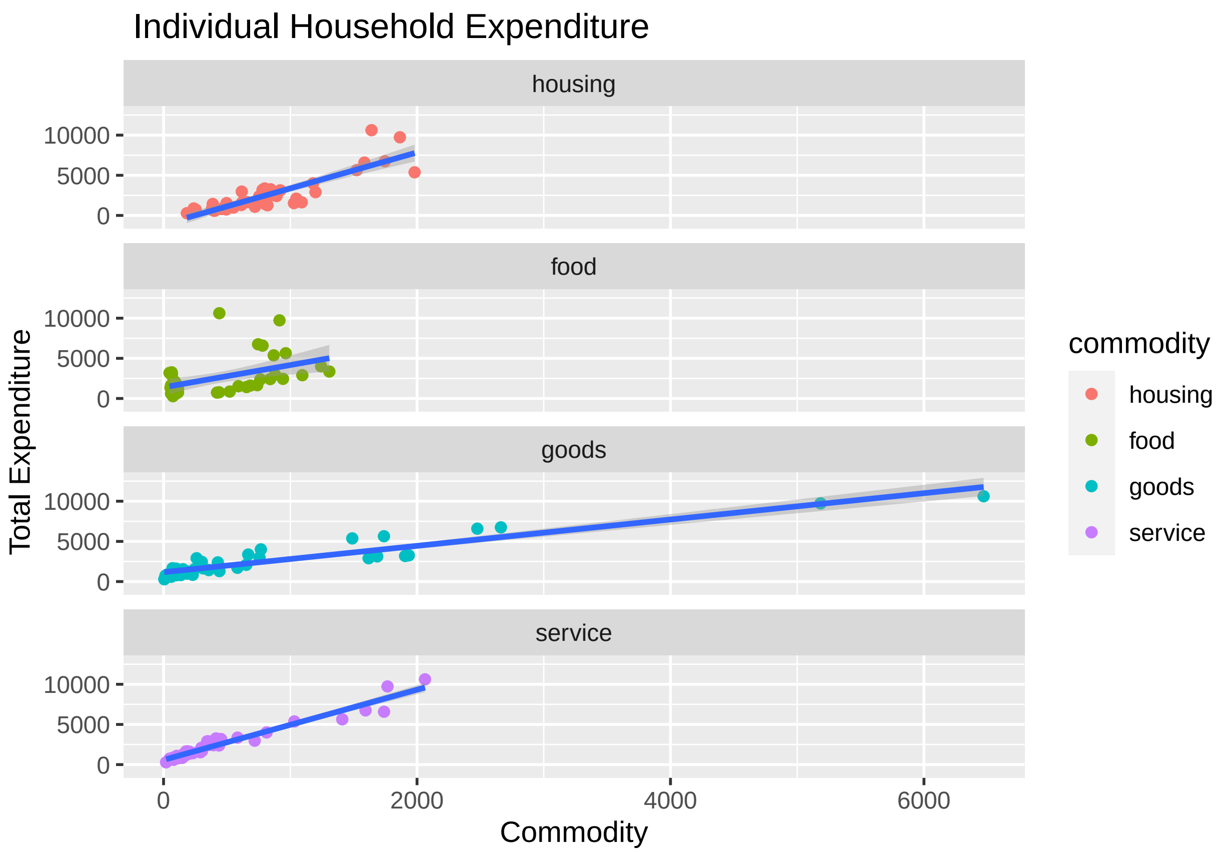
To show the relationship between the Total Expenditure and household expenditure I used a scatter plot between Total_Expenditure vs commodity. For this purpose, used mutate function to compute total expenditure and converted the data into long form using gather function. The final plot clearly shows that as the total expenditure increases the household service expenditure also increases (shown by very less smoothing around the linear plot). At the same time the relation between food expenditure and the total expenditure cannot be clearly determined as shown by the line that does not fit properly. Additionally the relationship between total expenditure and food expenditures cannot be determined because the data does not fit a line.The other two variables do increase with the total expenditure but not same as services.

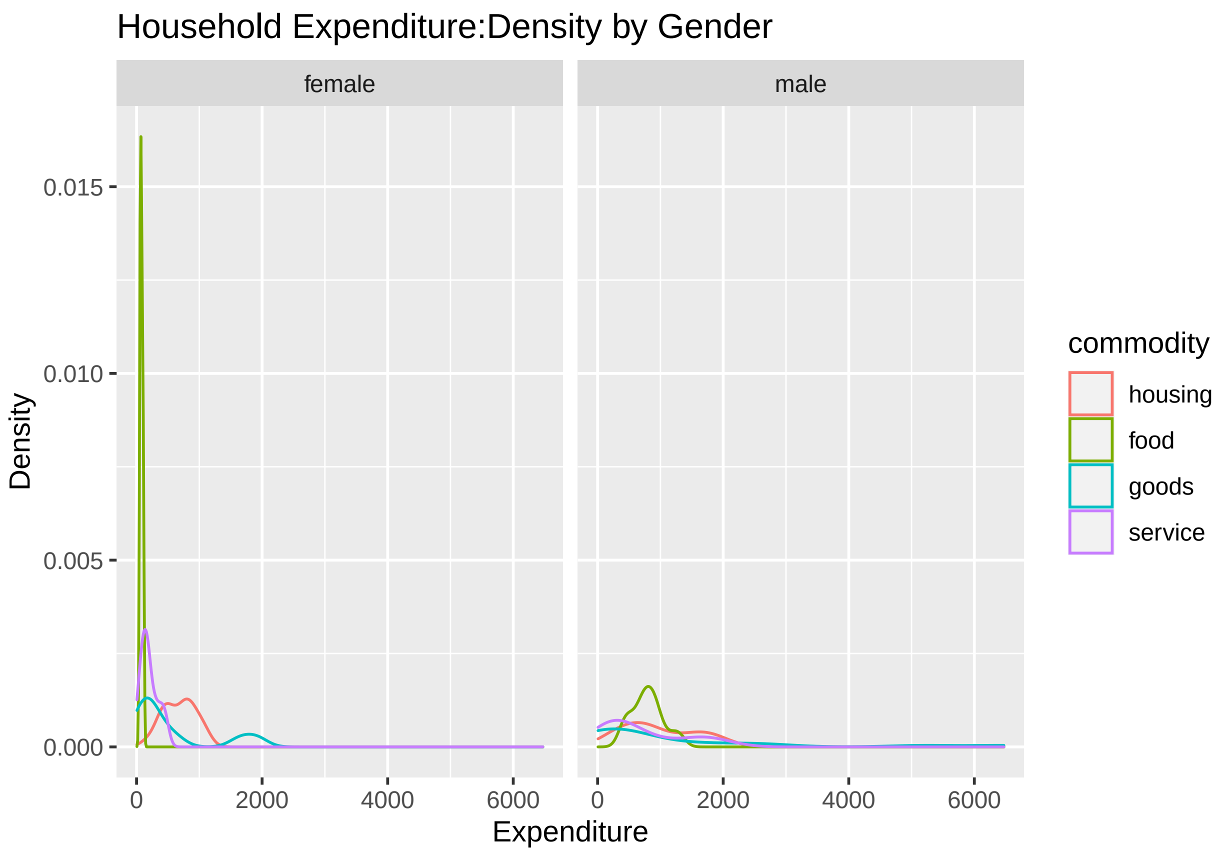
Plot 4 and Plot 5

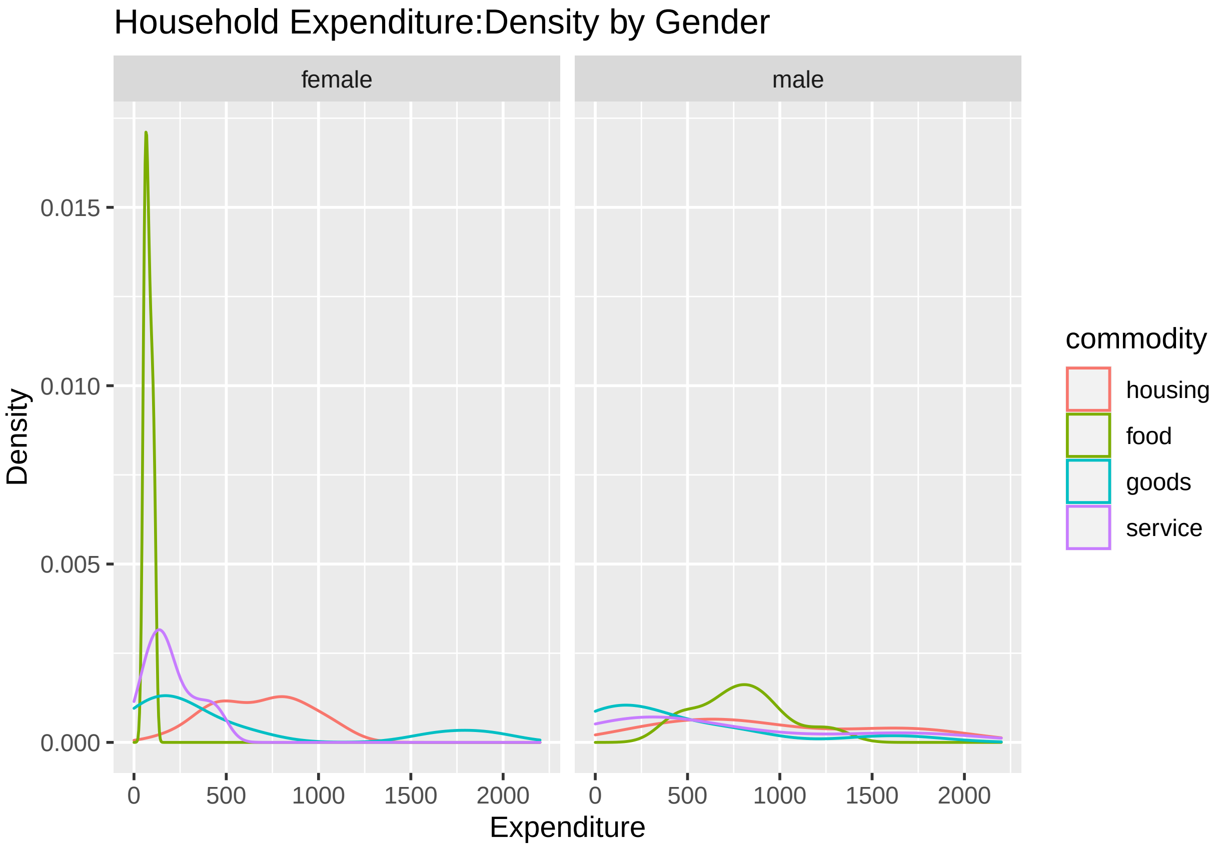
To show the relationship between the expenditure behavior based on gender I have plotted density plot between Expenditure and wrapped by gender. It clearly shows that most of the females spend less on food and more on housing. Whereas, males spend more on food and less on goods.In the last plot, we will examine the density curves of each expenditure by gender.The last plot was zoomed in to show the expenditure patter clearly.


Household Expenditure
female
male
1500
1000
500
0
1500
1000
500
0
% Expenditure
housing
food
goods
service
Commodity
commodity
housing
food
goods
service


Household Total Expenditure by Gender
Total Expenditure
15000
10000
5000
0
female
male
Gender
gender
female
male

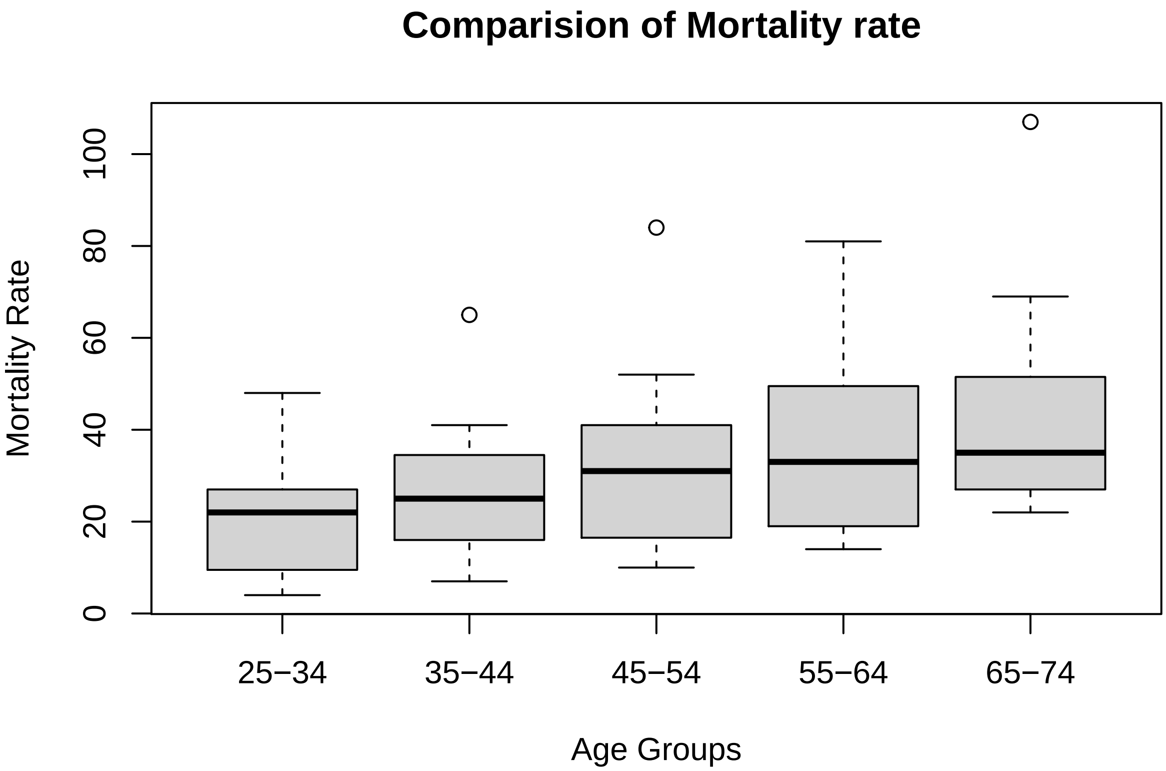

Individual Household Expenditure
housing
food
goods
service
Total Expenditure
10000
5000
0
Commodity
0
2000
4000
6000
commodity
housing
food
goods
service

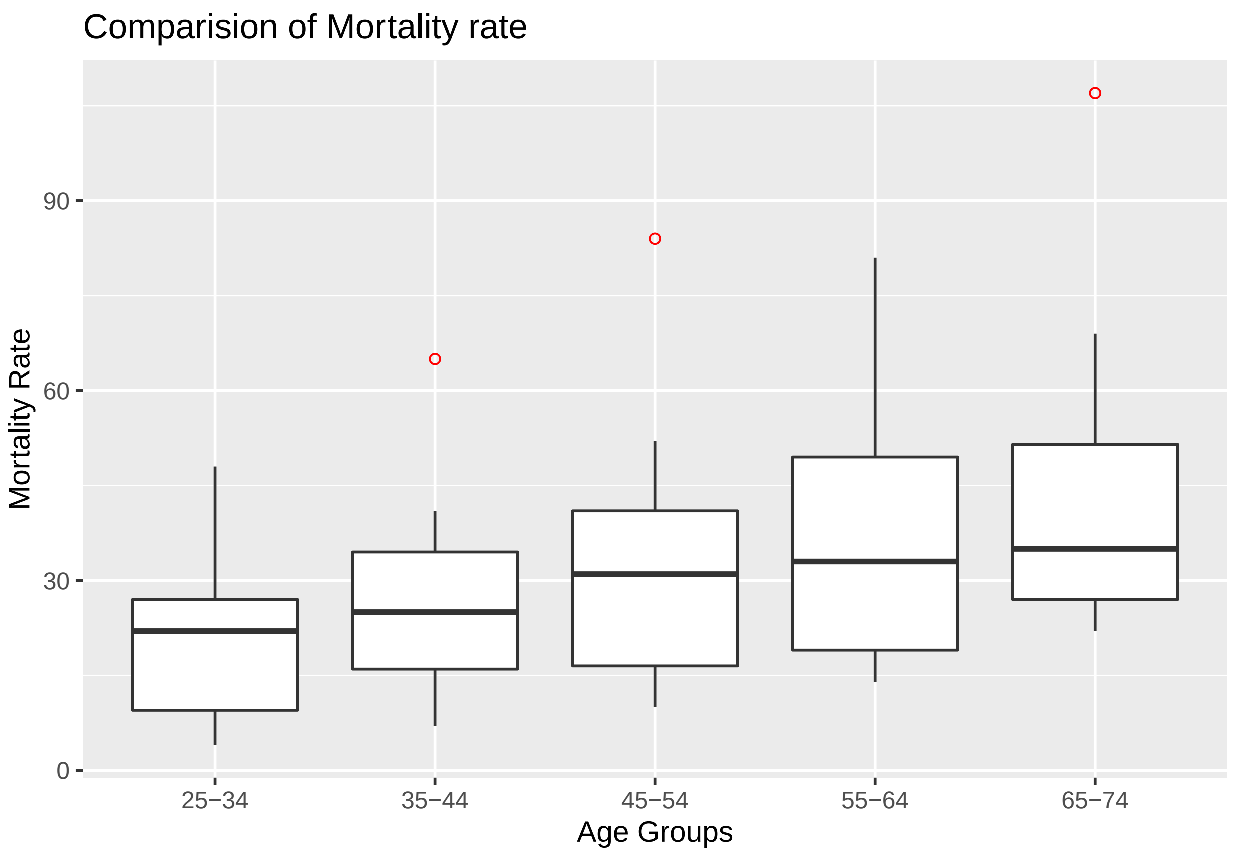

Household Expenditure:Density by Gender
female
male
Density
0.015
0.010
0.005
0.000
0
2000
4000
6000
Expenditure
commodity
housing
food
goods
service



Question 7: Mortality rates per 100,000 from male suicides for a number of age groups and a number of countries are given.Construct side by side box plots for the data from different age groups, and comment on what the graphics tells us about the data.

Answer: For this question, first the column names were changed and then the numerical columns were stacked one over the other using stack function. The resulting data set has numerical values as values and age group as ind.Used basic boxplot function to plot suicide rate vs age group. Below are few of my observations, * A lowest mortality rate of about 5 is observed in age groups 25-34 and highest is observed in age group 55-64. * Mortality rate of age groups 35-44 is between 8 and 42 (with one outlier with a rate about 65).With a lower overall mortality rate compared to other age groups. * Mortality rate of age groups 55-64 is between 14 and 80 ,With a higher overall mortality rate compared to other age groups. However there is an outlier in the age groups 65-74 where the mortality rate is above 100. * The boxplot for age group 35-44, with a median in the center demonstrates that mortality rate is below ~25 for 50% observations and above ~25 for the rest and the distribution is condense and can be predicted with more confidence than other age groups. * The median is at about the center of the boxplot for age groups 55-64.However, the 4th quartile indicates that the rate is wide spread. * For the age group 65-74, the lowest mortality rate is above 20 indicating that for this age group all the countries have mortality rate of more than 20 which is higher than other age groups.


Comparision of Mortality rate
Mortality Rate
0
20
40
60
80
100
25−34
35−44
45−54
55−64
65−74
Age Groups



Question 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.

Answer: Manually calculated the mean absolute deviation of weights variable of chcikwts dataset. The result is same compared to mean absolute deviation derived by using mad function.

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

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

Question 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.

Answer:To answer this question, first the vectors state.name and state.division were added to the state.x77 data set by using cbind function.Per capita income was computed and added as a new variable to the dataset.Then,functions filter,arrange,select functions from *dplyr* package were used to filter the New England state division,arranged the per capita incoome by descending order and selected relevant variables to find the state with minimum per capita income.

```
##      state.name state.division Per_capita_income
## 1 Massachusetts    New England         0.8178535
```

Question 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.

Answer: Using the standard subsetting functions subset relevant variables and filtered vehicles with highway mileage less than 25 miles per gallon and weight over 3500lbs. Then, using *dplyr* arrange function sorted the vehicles by weight and pronted the head of dataset.

```
##     Model Min.Price Max.Price MPG.highway Weight
## 1   Quest      16.7      21.5          23   4100
## 2   Astro      14.7      18.6          20   4025
## 3     Q45      45.4      50.4          22   4000
## 4 Eurovan      16.6      22.7          21   3960
## 5 Stealth      18.5      33.1          24   3805
## 6  Previa      18.9      26.6          22   3785
```

Question 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:

Answer: Using *dplyr* ,select function selected relevant variables from the Cars93 dataset. Then converted to a matrix object. a) Created a vector of means of all variables called Cars.Means using colMeans function. b) Calculated the number of observations first, then using colSDS function from *GMCM* package calculated standard deviation of all columns and finally calculated standard errors of mean by dividing standard deviation with number of observations and created a vector object called Cars.Std. Errors. c) Created a list object of Cars.Means and Cars.Std.Errors.

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   Max.Price    MPG.city MPG.highway  EngineSize      Length
##   0.9069210   1.1438051   0.5827473   0.5528742   0.1075695   1.5141964
##      Weight
##  61.1694186
```

c) create a list object cars.stats

```
## [[1]]
##   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
##
## [[2]]
##   Min.Price   Max.Price    MPG.city MPG.highway  EngineSize      Length
##   0.9069210   1.1438051   0.5827473   0.5528742   0.1075695   1.5141964
##      Weight
##  61.1694186
```

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

Answer:
a) Computed sample means of variables Sepal Length, Sepal Width, Petal Length, Petal Width for all the species Setosa, Versicolor, Virginica using apply function. b) Computed sample means for whole dataset using apply function

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
```

Question 13. Use the data matrix **state.x77** and the `tapply()` function to obtain:

Answer:

```
a) Converted the factor state.region into a data frame and binded factors state.name, state.division, s
b)Using tapply computed max function on illiteracy rate of all states in the division.
c)Using tapply computed length of states in all regions to determine number so states per region.
```

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
##      2.871174      1.539944      2.165812      5.299430
```

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

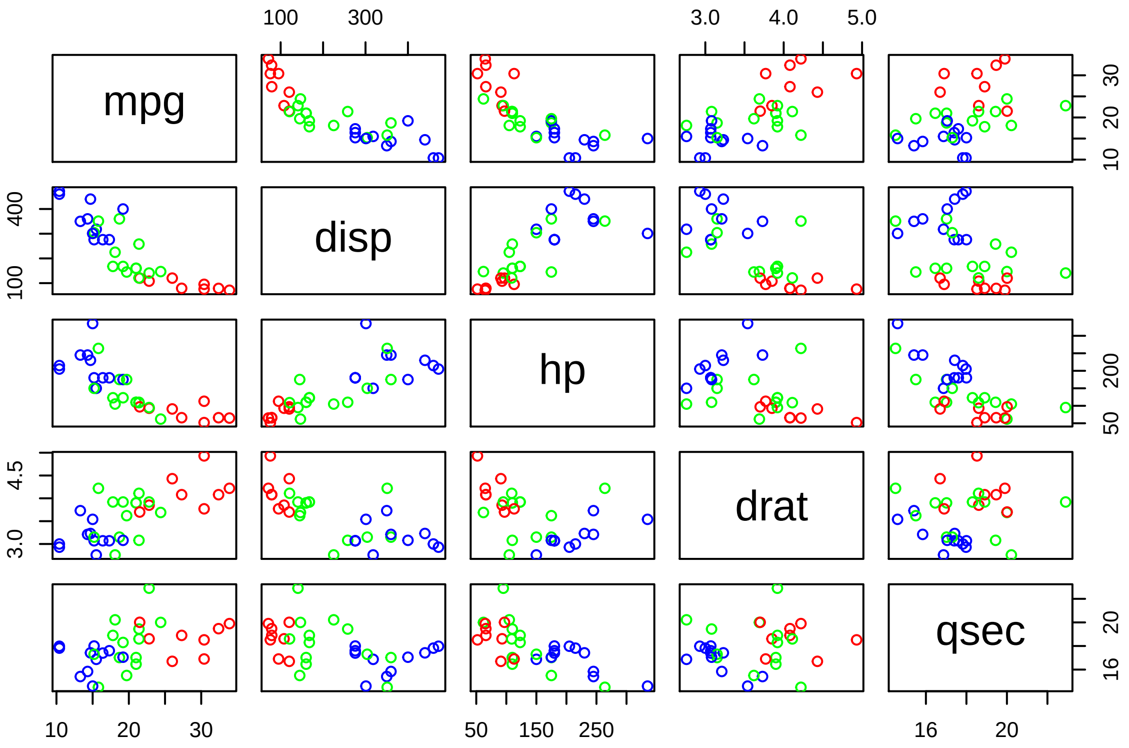
```
##        New England    Middle Atlantic     South Atlantic East South Central
##                1.3                1.4                2.3                2.4
## West South Central East North Central West North Central           Mountain
##                2.8                0.9                0.8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

c) The number of states in each region

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

Question 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.

Answer: Categorized the cars by car weight as Compact,Midsize and Large and saved as a vector carsize. Selected the relevant variables and added carsize variable to dataset.Defined vector, col for assigning different colors to cars by carsize.Plotted Scattor plot matrix using pairs function on the newly created dataset without carsize variable.



Question 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.

Answer 15: Applied aov function of weight and feed column of chickwts dataset and printed the ANOVA table.

```
## Call:
##    aov(formula = weight ~ feed, data = chickwts)
##
## Terms:
##                      feed Residuals
## Sum of Squares   231129.2  195556.0
## Deg. of Freedom         5        65
##
## Residual standard error: 54.85029
## Estimated effects may be unbalanced
```

Question 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.*

Answer 16:a) Defined a R function ttest() with variables x,y,mu(mean),alternative parameters that returns a list object containing t-statistic(T) value and two-sided p-value(P).t-statistic(T) value is computed using t.test fucntion and return the statistics and two-sided p-value(P) is coputed using t.test function and returns p.value. b) Tested the hypothesis mu =240 and mu=0. Afer applying the ttest function it is clear that when compared to the two sided test that the hypothetical mean is not a good approximation of the mean because of the low t statistic, high p value, and the mean is outside of the 95% CI of the two sided test. c) Then I verified the results using the inbuilt function t.test in both cases. And the results were observed to similar.

```
# creating function ttest()
ttest <- function(x,y=NULL,mu=0,alternative = 'two.sided'){
    T <- t.test(x=x,y=y,mu=mu,alternative=alternative)$statistic
    P <- t.test(x=x,y=y,mu=mu,alternative=alternative)$p.value
    result=list(T=T,P=P)
    return(result)
}

cat('T Test at mu = 240\n')
```

```
## T Test at mu = 240
```

```
ttest(x=chickwts$weight,mu=240,alternative='two.sided')
```

```
## $T
##        t
## 2.299879
##
## $P
## [1] 0.02444107
```

```
# verifying the results using t.test 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
```

```
cat(’T Test at mu = 0\n’)
```

```
## T Test at mu = 0
```

```
ttest(x=chickwts$weight,mu=0,alternative=’two.sided’)
```

```
## $T
##        t
## 28.20202
##
## $P
## [1] 5.919394e-40
```

```
# verifying the results using t.test function
t.test(chickwts$weight,mu=0)
```

```
##
##  One Sample t-test
##
## data:  chickwts$weight
## t = 28.202, df = 70, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  242.8301 279.7896
## sample estimates:
## mean of x
##  261.3099
```