Expantory Data Analysis Practice

Mohamed Ahmed

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: The function dplyer was used filter the countries (US,UK, France and Germany) and then compute the median profit for comanies in each country. The results show that median for these countries ranges between 0.19 to 0.24 with USA having the highest median.

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

2. Find all German companies with negative profit?

Answer: Data was filtered based on location and if profits their were less than 0. Thirteen companies in Germany have negative profits.

```
##  [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. To which business category do most of the Bermuda island companies belong?

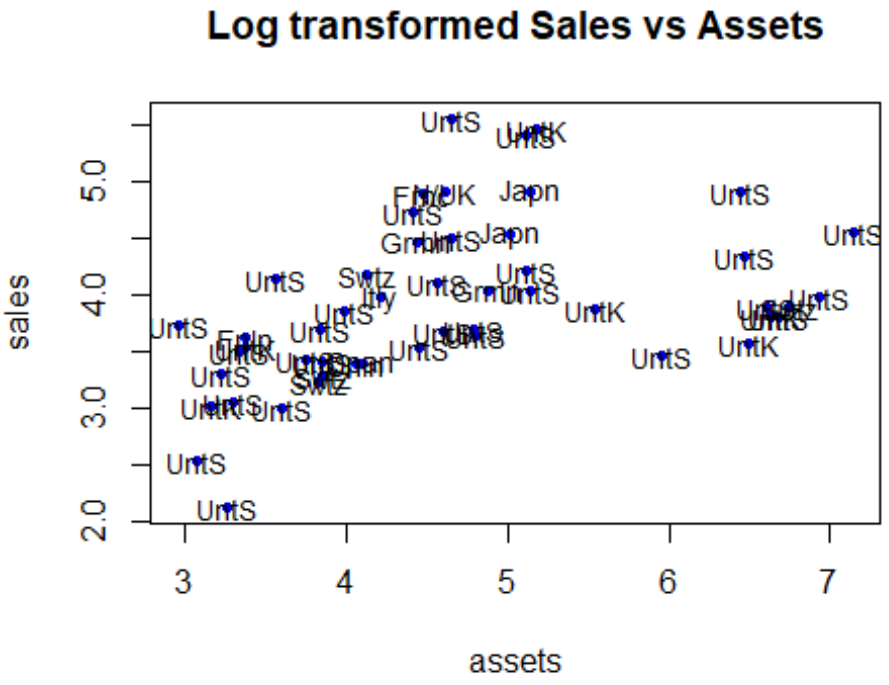
First, we filter the column country to get bermuda rows and then count the category with most companies. Insurance companies have a big market in the Bermuda Islands.

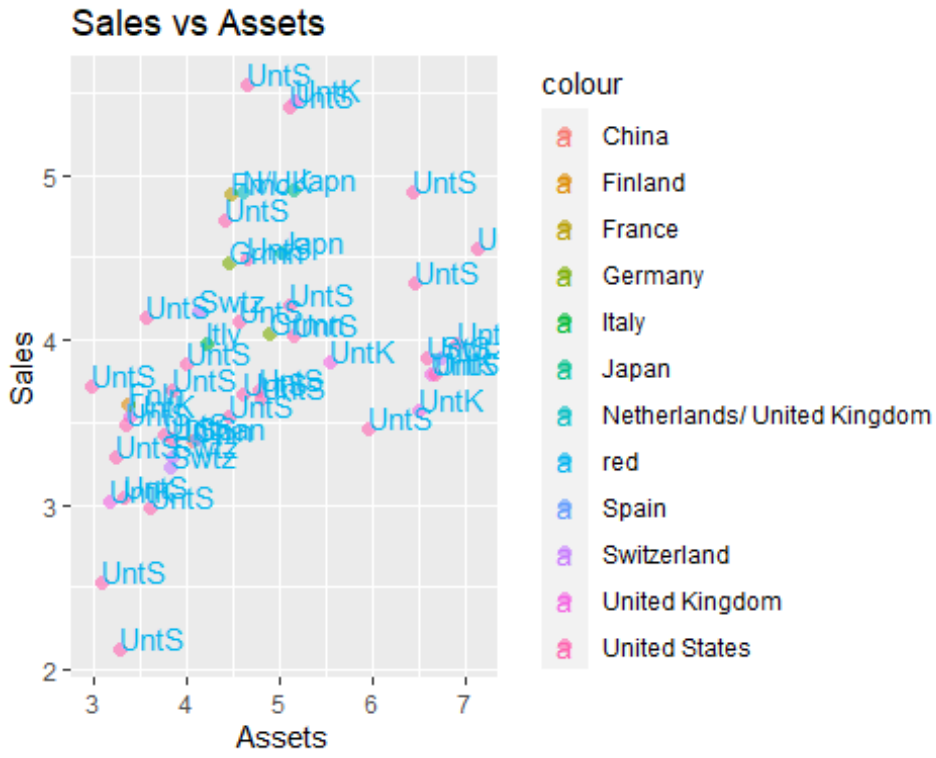
```
##    category  n
## 1 Insurance 10
```

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), 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: To answer this question, the dplyer package was used to manipulate data and extract data of top 50 companies based o their profits. The data was arranged from highest

to lowest, the top 50 companies were selected along with necessary columns for plotting. For plotting, I assumed that data need to be normalized so I applied log transformation. sales vs assets were plotted using base R and ggplot to visualize some of the data we are interested in





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 dollar.

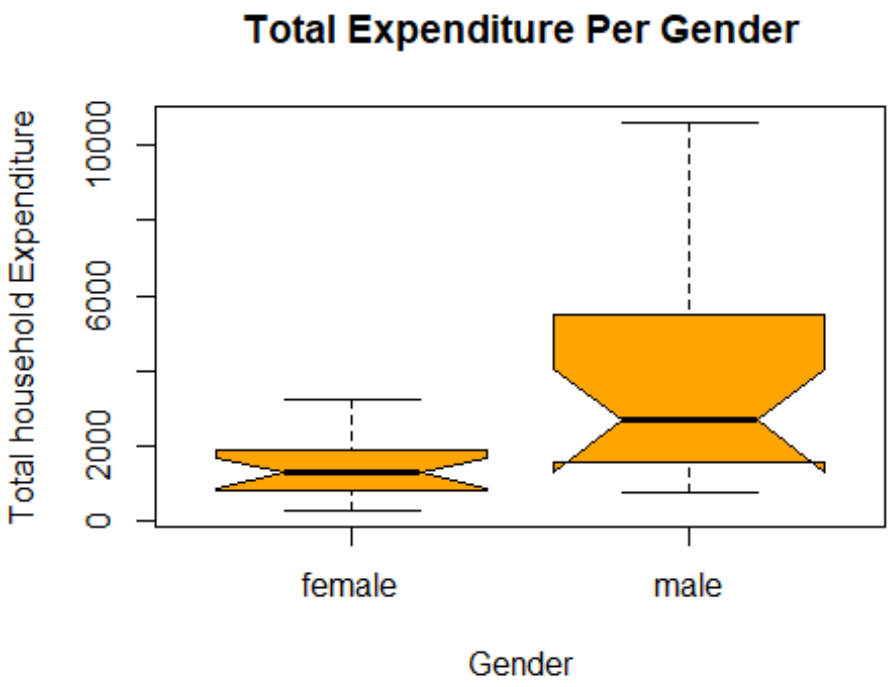
Answer: Using dyplyer package, companies were grouped by country then the average value of sales for companies in each country was computed. For the second part of the question, companies with five billion profits and above were filtered,and then counted by country.

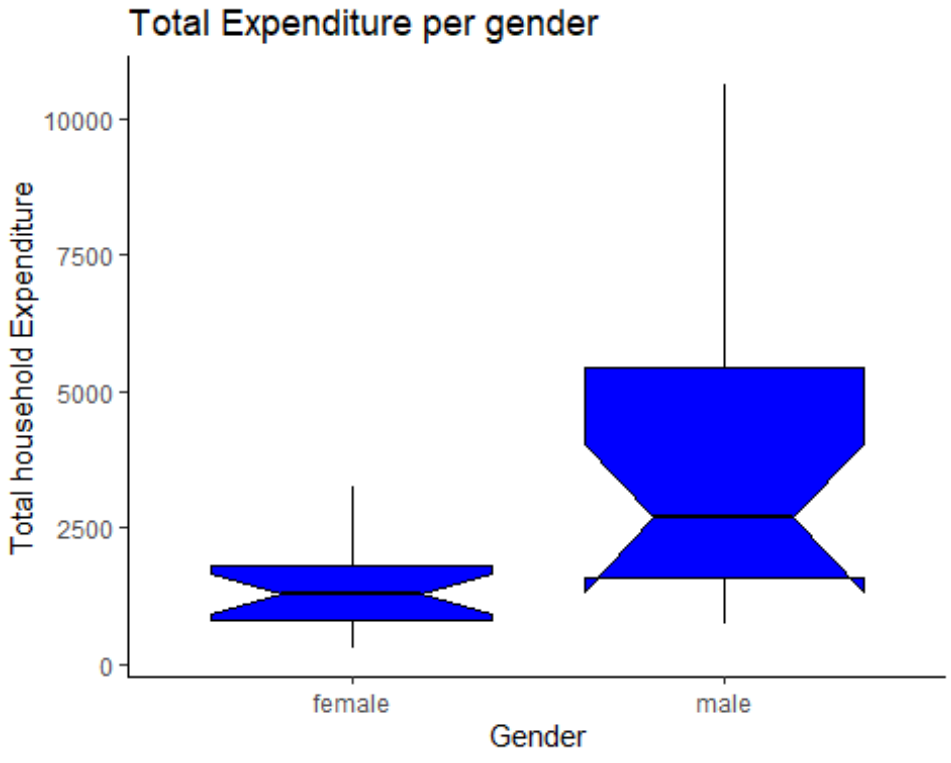
```
## # A tibble: 6 x 2
##   country                    Mean
##   <fct>                     <dbl>
## 1 Africa                     6.82
## 2 Australia                  5.24
## 3 Australia/ United Kingdom 11.6
## 4 Austria                    4.14
## 5 Bahamas                    1.35
## 6 Belgium     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

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

6. 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 Hon Kong dollars, 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 vehicles The aim of the survey was to investigate how the division of household expenditure between 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 conclusion.

Answer: First, we calculated the total expenditure to visualize how much each gender was spending overall. The second step in our analysis was to visualize how much each gender spends on each commodity.


Total Expenditure Per Gender
Total household Expenditure
0
2000
6000
10000
female
male
Gender

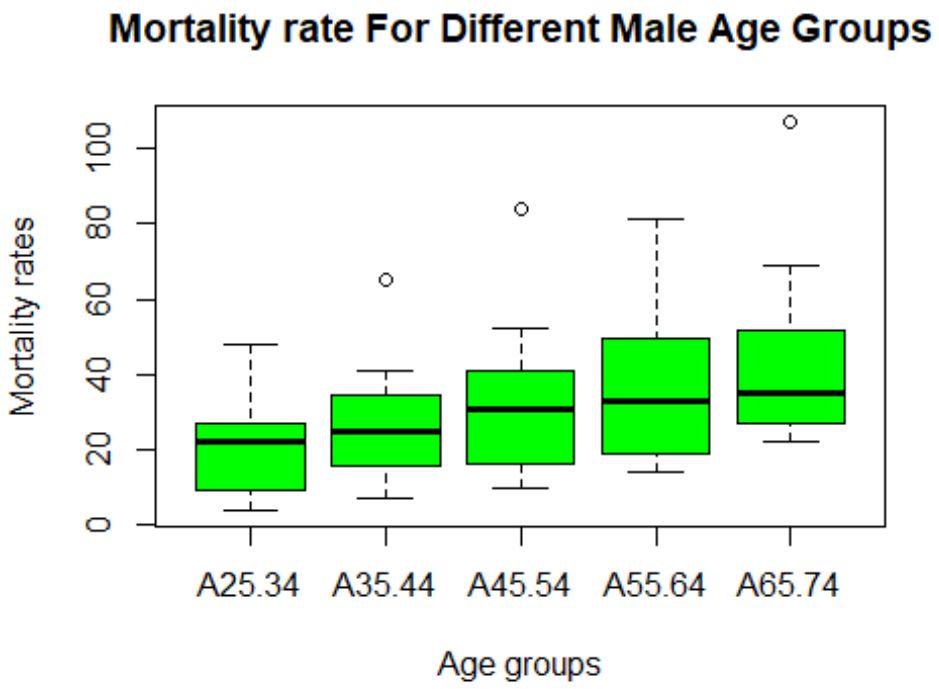

Total Expenditure per gender
Total household Expenditure
0
2500
5000
7500
10000
female
male
Gender

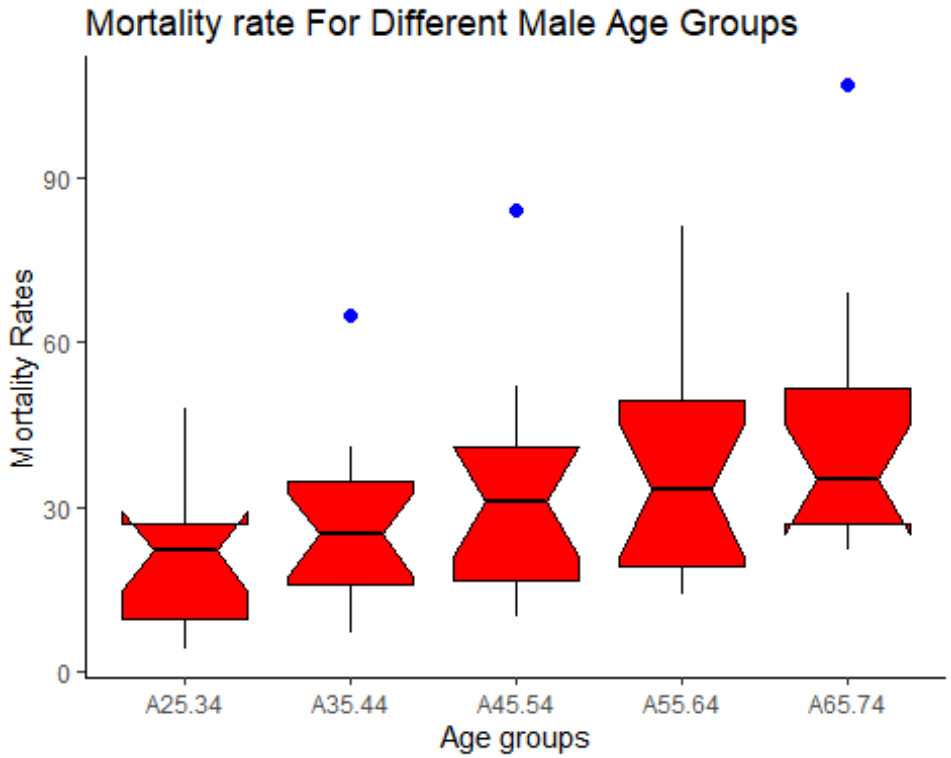
Conclusion: By analyzing our graphs, We can see that, overall, males spend more on all categories than females. Males spend on average a total of 3688 on all commodities. On the other hand, females spend on average a total of 1507 on all commodities. When we checked how much each gender spends on each commodity, we saw that females spend insignificant amount of money on food, services, and goods compared to males; however, both genders spend significant amounts on housing. Males spend about 20000 on housing, and females spend only 5000 less than males.

7. Mortality rates per 100,000 from male suicides for a number of age groups and a number of countries are given in Table 2.3. Construct side-by-side box plots for the data from different age groups, and comment on what the graphic tells us about the data.

```
##                A25.34 A35.44 A45.54 A55.64 A65.74
## Canada             22     27     31     34     24
## Israel              9     19     10     14     27
## Japan              22     19     21     31     49
## Austria            29     40     52     53     69
## France             16     25     36     47     56
## Germany            28     35     41     49     52
## Hungary            48     65     84     81    107
## Italy               7      8     11     18     27
## Netherlands         8     11     18     20     28
## Poland             26     29     36     32     28
## Spain               4      7     10     16     22
## Sweden             28     41     46     51     35
## Switzerland        22     34     41     50     51
```

```
## UK                  10      13      15      17      22
## USA                 20      22      28      33      37
```





Conclusion: Mortality rate keep on increasing as age increases. The group A65.74 has the highest Mortality rate.

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 data set . Use the R function `mad()` to verify your answer.

The Median Absolute Deviation was calculated using $1.4826 \cdot median|x - \hat{\mu}|$, then the mad function was used to verify my 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.

Answer: The matrix state.x77, the factor state.division, and the vector state.name were combined in one data frame. The state with the minimum per capita income in Division was filtered.

```
##       Income State    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.

Answer: Vehicles with less than 25 mileage and weight more than 3500 Ibs were filtered along with columns that we were asked to print (Model, Min.Price, Max.Price, MPG.highway, Weight).

```
##          Model Min.Price Max.Price MPG.highway Weight
## 16 Lumina_APV      14.7      18.0          23   3715
## 17      Astro      14.7      18.6          20   4025
## 26    Caravan      13.6      24.4          21   3705
## 28    Stealth      18.5      33.1          24   3805
## 36   Aerostar      14.5      25.3          20   3735
## 48        Q45      45.4      50.4          22   4000
## 49      ES300      27.5      28.4          24   3510
## 50      SC300      34.7      35.6          23   3515
## 56        MPV      16.6      21.7          24   3735
## 63   Diamante      22.4      29.9          24   3730
## 66      Quest      16.7      21.5          23   4100
## 70 Silhouette      19.5      19.5          23   3715
## 87     Previa      18.9      26.6          22   3785
## 89    Eurovan      16.6      22.7          21   3960
```

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

The dplyr package was used to select the variables needed to compute the mean.Lapply function was used to compute the mean for each variable.

```
##   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

Use the standard errors of means formula to compute standard errors of means for each selected column.

```
##   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
```

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

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

Computed sample means of the variables and for all the species using the apply function.

```
##          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.

Computed sample means of the variables and for all the species using the apply function.

```
## 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

We combined the matrix and the factor into one data frame and calculated the average income for each 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

We combined the and the factor into one data frame and calculated The maximum illiteracy rates for states in each of the nine divisions.

```
##        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
```

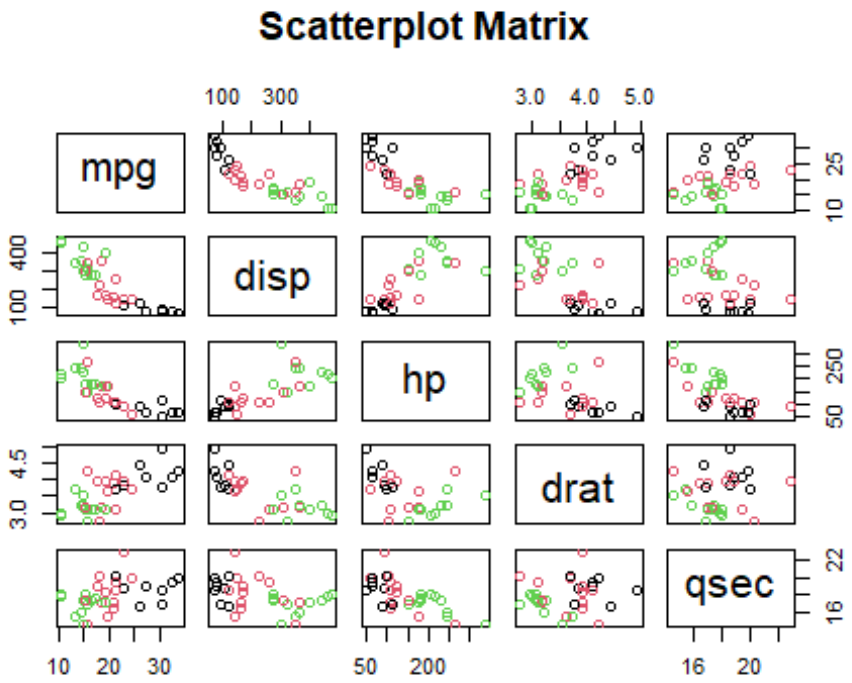
c) The number of states in each region

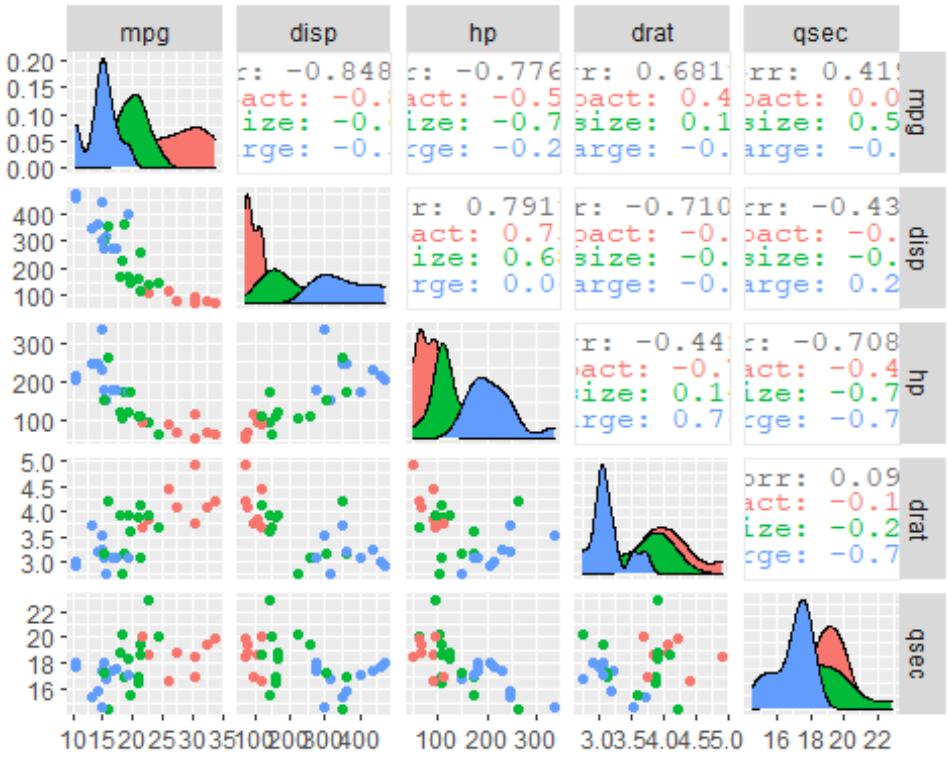
We combined the matrix and the factor into one data frame and then calculated the mean.

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

14. Using the data frame , 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.

The variable was appended to the data frame then the scatter plot matrix was created. In the scatter plot matrix, the variable in each row serves as Y axis and the variable in each serves as the x axis.





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)
## 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
```

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 data set) is equal to 240 against the two-sided alternative. `echo = T`

We created the a function called ttest. using the ttest function we calculated the T test using this formula $t = (x - \mu)/(s/\sqrt{n})$ and the P value for the T test. Our null hypothesis was that sample mean is 240 and the alternative hypothesis is sample mean is not equal to 240.

```
# loading packages
library(stats)

# define function ttest
ttest = function(y, mu0, alpha){


  sample.mean = mean(y$weight)
  sample.size = nrow(y)
  SD = sqrt(var(y$weight))
  z_1 = qt(alpha/2,sample.size-1)
  z_2 = qt(1-alpha/2,(sample.size-1))

  # calculate T and p value
  T =(sample.mean-mu0)/(SD/sqrt(sample.size))
  P =2*(1-pt(T,df=sample.size-1))
  return (list(P,T))
}
print("T value & P value")

## [1] "T value & P value"

t_test <- ttest(y=chickwts,mu0 = 240,alpha = 0.05)
t_test

## [[1]]
## [1] 0.02444107
##
## [[2]]
## [1] 2.299879
```

Conclusion: Since this $P.value = 0.0244$ is less than our chosen $alpha = 0.05$ value, we can reject the null hypothesis. therefore, we ave enough evidence to conclude that the theoretical mean for the weight is not a good approximation of the mean.