MET CS544 Foundations of Analytics
Homework#3

Problem#1

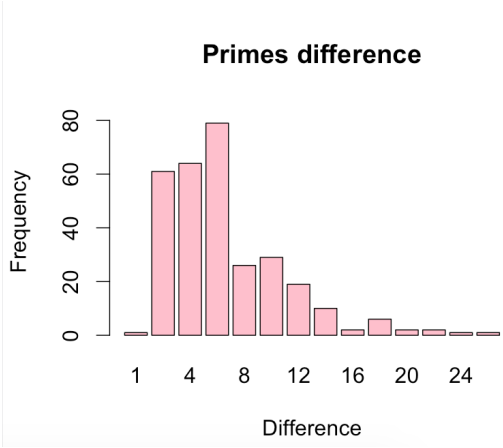
Use the primes (UsingR) dataset. Use the diff function to compute the differences between successive primes. Show the frequencies of these differences. Show the barplot of these differences.

Solution:

The frequency of the differences are given by the following figure which is calculated from the diff function

```
> prime.bar
prime.difference
 1  2  4  6  8 10 12 14 16 18 20 22 24 34
 1 61 64 79 26 29 19 10  2  6  2  2  1  1
```

The barplot which shows the frequencies of the difference is shown below



Problem#2

Use the coins (UsingR) dataset. Do not use explicit loops for any calculations. Do not hard code the denominations in the solution. The solution should work for any denominations.

a) How many coins are there of each denomination?
b) What is the total value of the coins for each denomination?
c) What is the total value of all the coins?
d) Show the barplot for the number of coins by year.

Solution:

a) This shows the number of coins for each denominations.

```
> table(coins$value)

0.01 0.05  0.1 0.25
 203   59   42   67
```

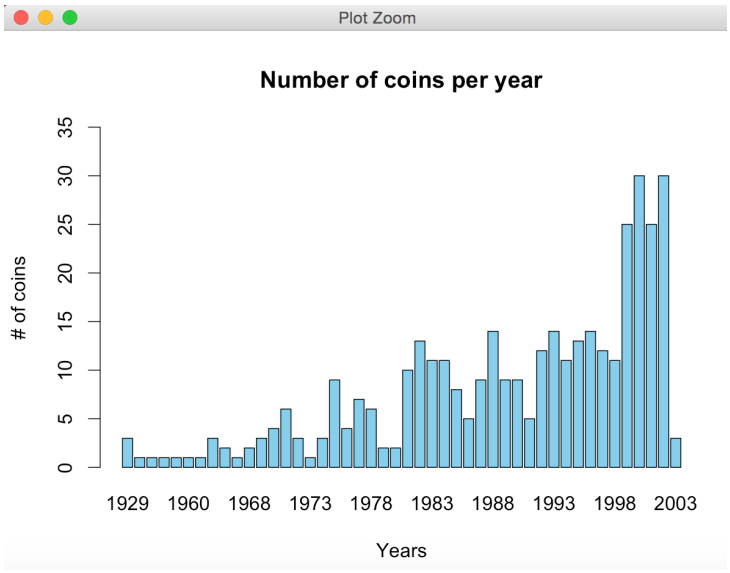
b) The total value of each denominations are

```
> Totals
  Denomination Total_counts
1         0.01         2.03
2         0.05         2.95
3          0.1         4.20
4         0.25        16.75
```

c) The total value of the coins are

```
> Total.denom.value
[1] 25.93
```

d) The barplot showing the #of coins by year



Problem#4

Use the pi2000 (UsingR) dataset.

a) How many times each of the digits 0 to 9 occur in this dataset?
b) Show the percentages of their frequencies.
c) Show the histogram of the data.

Solution:

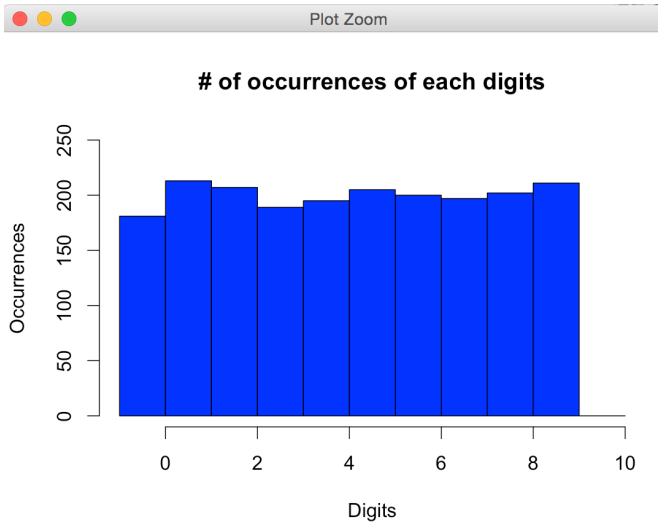
a) The number of occurrences are shown here

```
> pi.data
   digits occurrence
1       0        181
2       1        213
3       2        207
4       3        189
5       4        195
6       5        205
7       6        200
8       7        197
9       8        202
10      9        211
```

b) The percentage of frequencies are

```
> pi.digit.percentage
   pi.data.digits pi.data.percentage
1               0               9.05
2               1              10.65
3               2              10.35
4               3               9.45
5               4               9.75
6               5              10.25
7               6              10.00
8               7               9.85
9               8              10.10
10              9              10.55
```

c) The histogram of the data pi2000 is



Problem#5

Suppose that a football (NFL), basketball (NBA), and hockey (NHL) games are being shown at the same time. Consider the two-way summarized data shown below showing the preferences of men and women what sport they wish to watch.

a) Using cbind, create the matrix for the above data.
b) Set the row names for the data.
c) Set the column names for the data.
d) Now, add the dimension variables Gender and Sport to the data.
e) Show the marginal distributions for the Gender and the Sport.
f) Show the result of adding margins to the data.
g) Show the proportional data separately for Gender and Sport. Interpret the results.
h) Using appropriate colors, show the mosaic plot for the data. Also show the barplot for Gender and Sport separately with the bars side by side. Add legend to the plots.

Solution:

a) The matrix is built and the

```
> games.data
      a  b  c
[1,] 25 10 15
[2,] 20 40 30
```

b & c) Naming the rows and columns is done

```
> games.data
      NFL NBA NHL
Men    25  10  15
Women  20  40  30
>
```

d) Additional dimensions variable is set

```
> games.data
       Sport
Gender  NFL NBA NHL
  Men    25  10  15
  Women  20  40  30
>
```

e) Marginal distributions for both the gender and sport is found

```
> margin.table(games.data,1)
Gender
  Men Women
   50    90
> margin.table(games.data,2)
Sport
NFL NBA NHL
 45  50  45
```

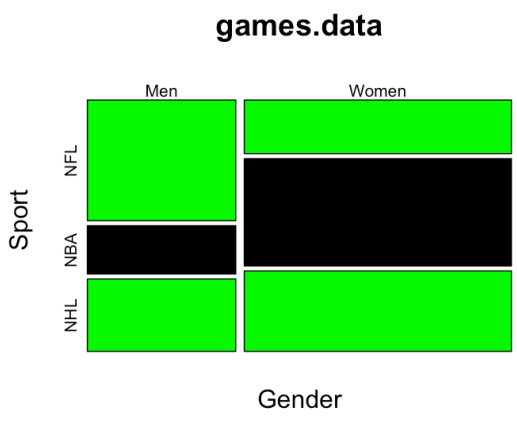
f) Adding the marginal values to the data and the result is shown

```
> addmargins(games.data)
       Sport
Gender  NFL NBA NHL Sum
  Men    25  10  15  50
  Women  20  40  30  90
  Sum    45  50  45 140
```

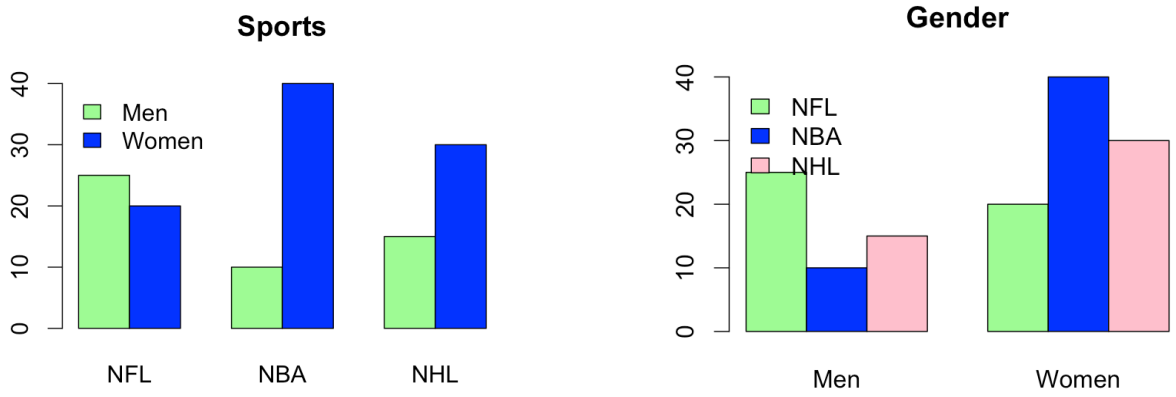
g) Proportional data is given by this plot for each gender and the sport

```
> prop.table(games.data,1)
       Sport
Gender        NFL       NBA       NHL
  Men   0.5000000 0.2000000 0.3000000
  Women 0.2222222 0.4444444 0.3333333
> prop.table(games.data,2)
       Sport
Gender        NFL NBA       NHL
  Men   0.5555556 0.2 0.3333333
  Women 0.4444444 0.8 0.6666667
>
```

h) Mosaic plot is generated to show visually the distribution of data in the game dataset.



The bar plots for gender and sport are given separately as



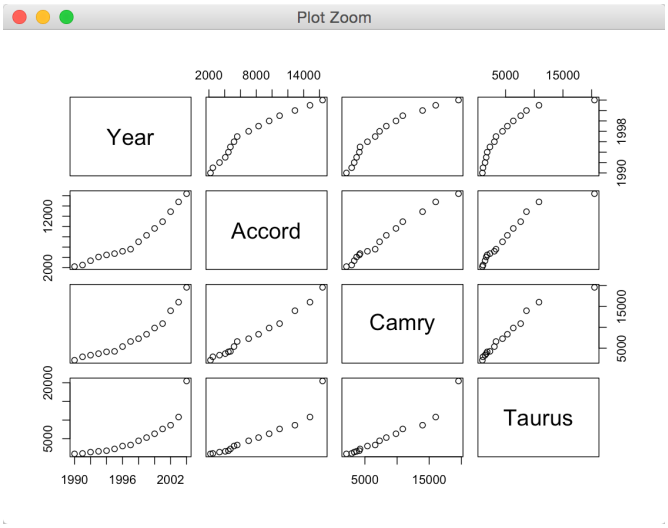
Problem#6

Use the midsize (UsingR) dataset.

a) Show the pair wise plots for all the variables.
b) Provide at least 4 interpretations of the results

Solution:

a) The pairwise plot is drawn for all the four attributes in the midsize dataset and shown below



b) The interpretations from the pair plot are

1. There exists a positive correlation between the Taurus and Accord.
2. There exists a high positive correlation between the Accord and Camry.
3. The correlation of Camry with Year is not better than the correlation it has with Accord.
4. There is no negative correlation between any attributes.

Problem#7

Use the MLBattend (UsingR) dataset.

a) Extract the wins for the teams BAL, BOS, DET, LA, PHI into the respective vectors.
b) Create a data frame of five columns using these vectors. Use the team names for the columns
c) Show the boxplot of the data frame.
d) Provide at least 5 interpretations of the results.

Solution:

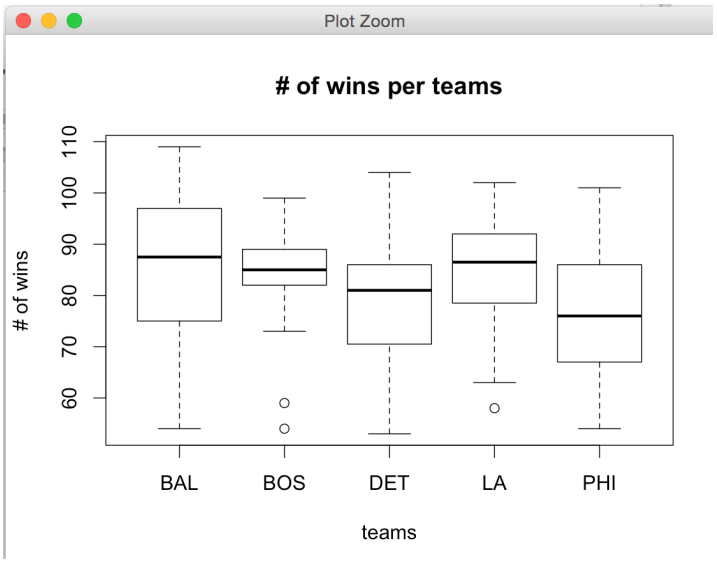
a) The wins are extracted for each team

```
> (team.BAL.wins)
 [1] 109 108 101  80  97  91  90  88  97  90 102 100  59  94  98  85  83  73  67  54  87  76  67  89  85  63  71
[28]  88  98  79  78  74
> (team.BOS.wins)
 [1] 87 87 85 85 89 84 95 83 97 99 91 83 59 89 78 86 81 95 78 89 83 88 84 73 80 54 86 85 78 92 94 85
> (team.DET.wins)
 [1]  90  79  91  86  85  72  57  74  74  86  85  84  60  83  92 104  84  87  98  88  59  79  84  75  85  53  60
[28]  53  79  65  69  79
> (team.LA.wins)
 [1]  85  87  89  85  95 102  88  92  98  95  79  92  63  88  91  79  95  73  73  94  77  86  93  63  81  58  78
[28]  90  88  83  77  86
> (team.PHI.wins)
 [1]  63  73  67  59  71  80  86 101 101  90  84  91  59  89  90  81  75  86  80  65  67  77  78  70  97  54  69
[28]  67  68  75  77  65
```

b) A data frame is created and the columns represent the wins for each teams.

```
> head(team.data)
  BAL BOS DET  LA PHI
1 109  87  90  85  63
2 108  87  79  87  73
3 101  85  91  89  67
4  80  85  86  85  59
5  97  89  85  95  71
6  91  84  72 102  80
```

c) The box plot is done for all the five teams and shown below



d) The interpretations of the data are

1. BAL has the highest maximum wins of all other teams used in this study.
2. BOS has the high agreement with the # of wins.(Compact)
3. LA has the lowest minimum value in the teams compared.
4. BAL has the low agreement with the # of wins.
5. DET has the highest median with itself compared to other teams themselves.

Problem#3

Use the south (UsingR) dataset.

a) Show the stem plot of the data. What do you interpret from this plot?
b) Show the five number summary of the data. Calculate the lower and upper ends of the outlier ranges. What are the outliers in the data?
c) Show the horizontal boxplot of the data along with the appropriate labels on the plot.

Solution:

a) The stem plot is drawn for the south dataset.

```
> south
 [1] 12 10 10 13 12 12 14  7 16 18  8 29 12 14 33 10  6 18 11 25  8 16 14 11 10 20 14 11 12 13
> sort(south)
 [1]  6  7  8  8 10 10 10 10 11 11 11 12 12 12 12 12 13 13 14 14 14 14 16 16 18 18 20 25 29 33
> stem(south)

  The decimal point is 1 digit(s) to the right of the |

  0 | 6788
  1 | 000011122222334444
  1 | 6688
  2 | 0
  2 | 59
  3 | 3
```

1. The number of data values in the range from 10-19 is higher than the other data ranges.
2. There is no data in the range 20-24 and it has 2 values in 25-29 range (25,29)
3. data 12 is the mode of the dataset as it has the highest occurrence.

b) The five number summary of the data is

```
> fivenum(south)
[1]  6 10 12 16 33
> summary(south)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   10.25   12.00   13.97   15.50   33.00
```

The quantile ranges are given by this result and the inter quantile range is also calculated.

```
> lowerq
  25%
10.25
> middleq
50%
 12
> upperq
 75%
15.5
> iqr
 75%
5.25
```

The threshold of the data is also calculated for finding the outliers in the data.

```
> threshold.upper
   75%
23.375
> threshold.lower
  25%
2.375
```

```
> data.upper
[1] 25 29 33
> data.lower
numeric(0)
```

The upper and the lower end outliers are found in the data and it turns out that there is no outliers in the lower end of the data.

c) The horizontal boxplot is drawn for the south dataset and is shown below.