CS544, Fundamentals of Analysis Homework 3

Jefferson Parker

[japarker@bu.edu](mailto:japarker@bu.edu)

For all problems, see associated section in the R code file for support.

Part 1, Primes Data.

The difference of successive values is:

primesdiff;

[1] 1 2 2 4 2 4 2 4 6 2 6 4 2 4 6 6 2 6 4 2 6 4 6 8 4 2 4 2 4 14 4 6 2 10 2 6 6 4 6

[40] 6 2 10 2 4 2 12 12 4 2 4 6 2 10 6 6 6 2 6 4 2 10 14 4 2 4 14 6 10 2 4 6 8 6 6 4 6 8 4

[79] 8 10 2 10 2 6 4 6 8 4 2 4 12 8 4 8 4 6 12 2 18 6 10 6 6 2 6 10 6 6 2 6 6 4 2 12 10 2 4

[118] 6 6 2 12 4 6 8 10 8 10 8 6 6 4 8 6 4 8 4 14 10 12 2 10 2 4 2 10 14 4 2 4 14 4 2 4 20 4 8

[157] 10 8 4 6 6 14 4 6 6 8 6 12 4 6 2 10 2 6 10 2 10 2 6 18 4 2 4 6 6 8 6 6 22 2 10 8 10 6 6

[196] 8 12 4 6 6 2 6 12 10 18 2 4 6 2 6 4 2 4 12 2 6 34 6 6 8 18 10 14 4 2 4 6 8 4 2 6 12 10 2

[235] 4 2 4 6 12 12 8 12 6 4 6 8 4 8 4 14 4 6 2 4 6 2 6 10 20 6 4 2 24 4 2 10 12 2 10 8 6 6 6

[274] 18 6 4 2 12 10 12 8 16 14 6 4 2 4 2 10 12 6 6 18 2 16 2 22 6 8 6 4 2 4

The frequency of difference values is:

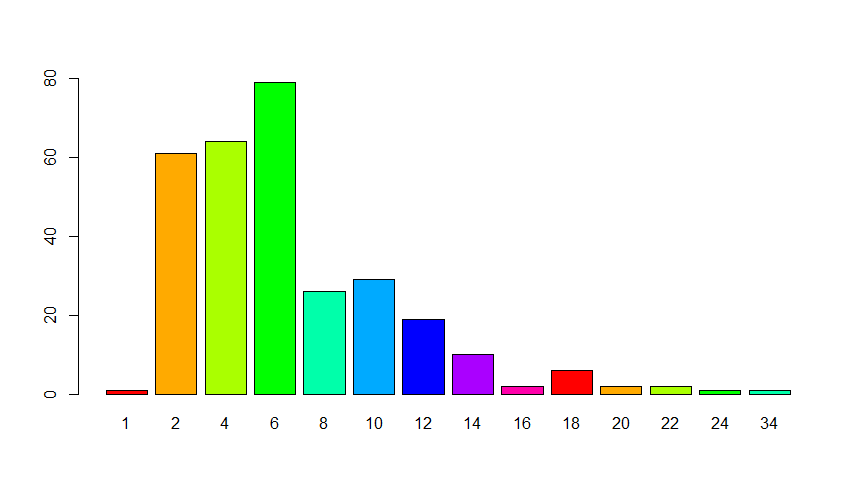
table(primesdiff);

primesdiff

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 of the differences is:



Part 2, Coins Data

1. The number of coins for each denomination is:

table(coins$value);

0.01 0.05 0.1 0.25

203 59 42 67

1. The value of the coins per denomination is (total column):

coins.df;

Var1 Freq total

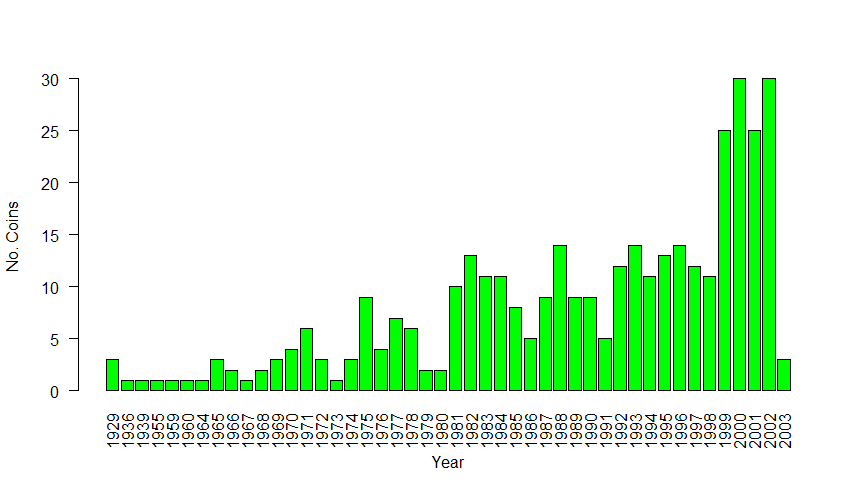
1 0.01 203 2.03

2 0.05 59 2.95

3 0.10 42 4.20

4 0.25 67 16.75

1. The total value of all coins is$25.93
2. The barplot of coins by year is:



Part 3, South Data

1. The stem plot is shown below. From this, we can interpret that most states had between 10 – 14 murders in the measured time.

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 five number summary is:

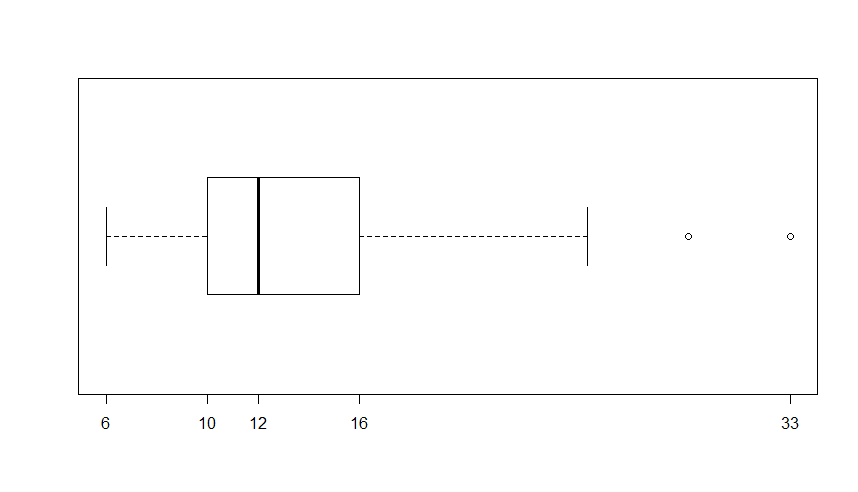
fivenum(south);

[1] 6 10 12 16 33

The lower and upper ends of the outlier range are 1 and 25 respectively.

The outlier values are 25, 29 and 33.

1. The horizontal boxplot labeled with five number summary is:



Part 4, pi2000 Data

1. Each digit 0 – 9 appears the following number of times:

table(pi2000);

pi2000

0 1 2 3 4 5 6 7 8 9

181 213 207 189 195 205 200 197 202 211

1. The percentage of the frequencies of each digit is:

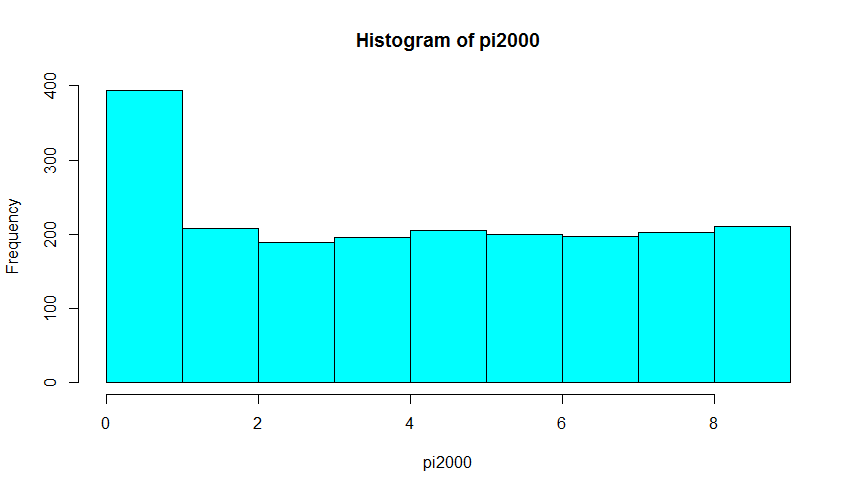
prop.table(table(pi2000)) \* 100;

pi2000

0 1 2 3 4 5 6 7 8 9

9.05 10.65 10.35 9.45 9.75 10.25 10.00 9.85 10.10 10.55

1. The histogram of the data is:



Part 5, Sports Data

1. – d) The data table is:

sports

Sport

Gender NFL NBA NHL

Men 25 10 15

Women 20 40 30

1. The marginal distributions by gender and sport are:

margin.table(sports, 1);

Gender

Men Women

50 90

margin.table(sports, 2);

Sport

NFL NBA NHL

45 50 45

1. The data table with margins is:

addmargins(sports);

Sport

Gender NFL NBA NHL Sum

Men 25 10 15 50

Women 20 40 30 90

Sum 45 50 45 140

1. The proportional data for gender and sport are:

prop.table(sports, 1);

Sport

Gender NFL NBA NHL

Men 0.5000000 0.2000000 0.3000000

Women 0.2222222 0.4444444 0.3333333

prop.table(sports, 2);

Sport

Gender NFL NBA NHL

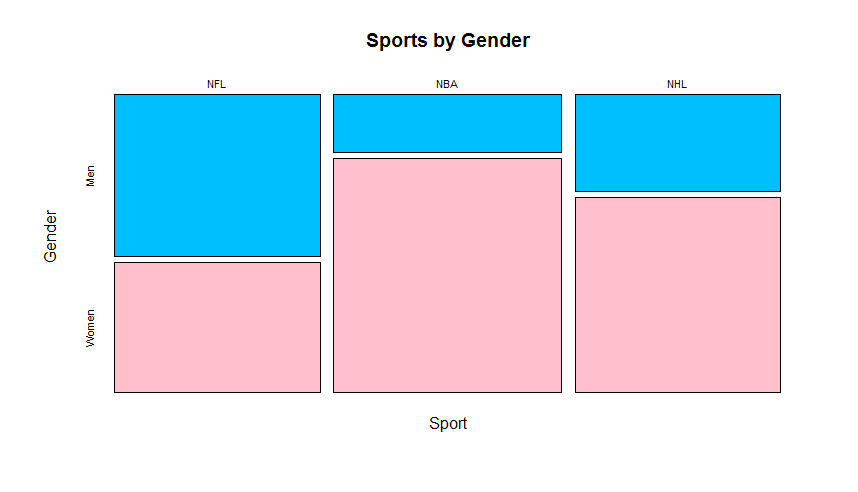
Men 0.5555556 0.2 0.3333333

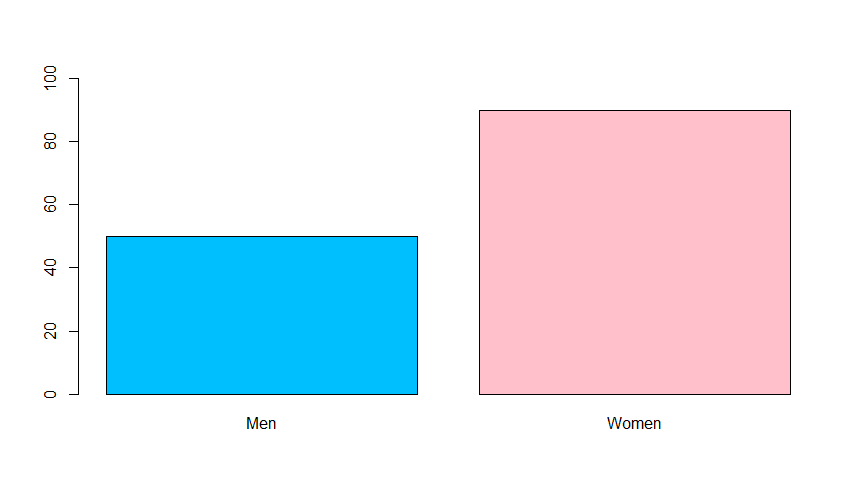
Women 0.4444444 0.8 0.6666667

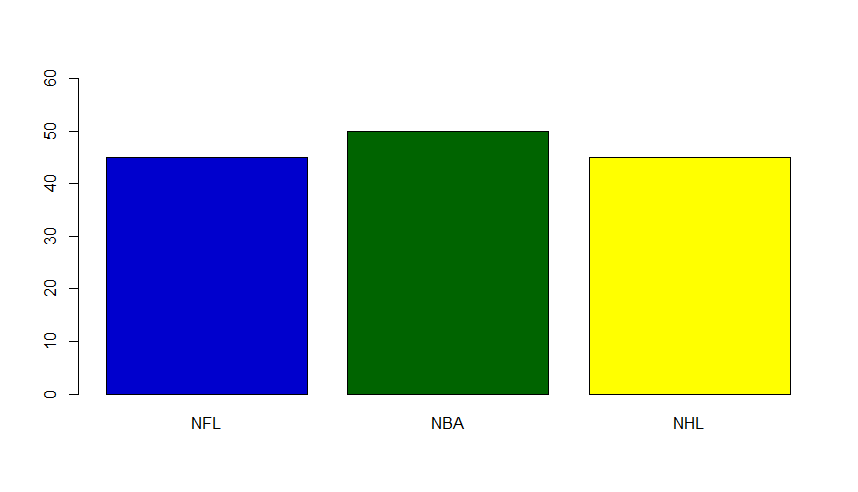
Men prefer NFL (50%) as much as NBA (20%) and NHL (30%) combined. Women prefer NBA (44%), then NHL (33%) and NFL (22%) in decreasing order.

Between genders, women prefer NBA (80%) and NHL (67%) much more than men (20% and 33% respectively) while both genders like NFL almost equally (56% for men and 44% for women).

1. The mosaic and bar plots for the data are:

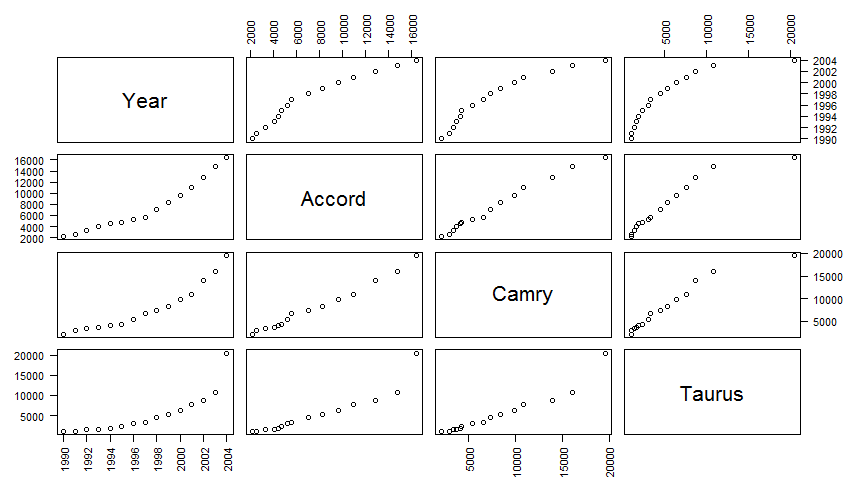






Part 6, Midsize Data

1. The pairwise plot for all variables is:



1. From the data we can see that:
   1. Prices for all models increased more rapidly from 1990 ~1996.
   2. The Taurus had a significant price increase relative to other models in the 2004 model year.
   3. The Camry and Accord increased in price at almost equal rates.
   4. The rate of price increase slowed more dramatically for the Camry and Accord after 1998, while the Taurus price slowdown was more gradual.

Part 7, MLBattend Data

1. – b) The data frame for the wins of the indicated teams is:

wins;

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

7 90 95 57 88 86

8 88 83 74 92 101

9 97 97 74 98 101

10 90 99 86 95 90

11 102 91 85 79 84

12 100 83 84 92 91

13 59 59 60 63 59

14 94 89 83 88 89

15 98 78 92 91 90

16 85 86 104 79 81

17 83 81 84 95 75

18 73 95 87 73 86

19 67 78 98 73 80

20 54 89 88 94 65

21 87 83 59 77 67

22 76 88 79 86 77

23 67 84 84 93 78

24 89 73 75 63 70

25 85 80 85 81 97

26 63 54 53 58 54

27 71 86 60 78 69

28 88 85 53 90 67

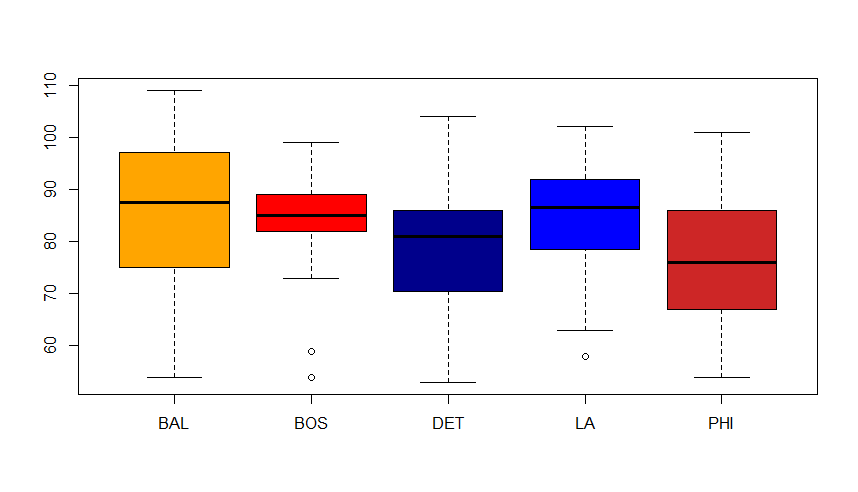
29 98 78 79 88 68

30 79 92 65 83 75

31 78 94 69 77 77

32 74 85 79 86 65

1. The boxplot of the win data for the indicated teams is:



1. Some interpretations of the data are:
   1. The BOS (Boston Red Sox) were generally consistent at winning across the measured years winning at a rate of almost 90 games per season. However there were at least two years which were uncharacteristically low.
   2. The PHL (Philadelphia Phillies) had the lowest average number of wins and overall worst performance of all teams measured, with a mean in the mid-70’s and a third quartile value below most other teams’ mean value.
   3. The BAL (Baltimore Orioles) had the best overall performance with highest mean wins, but they also had the most inconsistent performance with their wins per season having the widest range and interquartile range.
   4. Compared to their average, the DET (Detroit Tigers) had a wider spread of below average winning seasons (wider first quartile) than above average (third quartile).
   5. Similarly, the LA (LA Dodgers) had a wider first quartile range but they averaged more wins per season than almost all other teams.

Part 8, Congress Data

1. The number of members of each house of Congress, by party, is:

table(c(house$Party, senate$Party));

Democratic Independent Republican

239 2 289

1. The top 10 states by decreasing representation in the House is:

sort(table(house$State), decreasing = TRUE)[1:10];

California Texas Florida New York Illinois

52 36 27 27 18

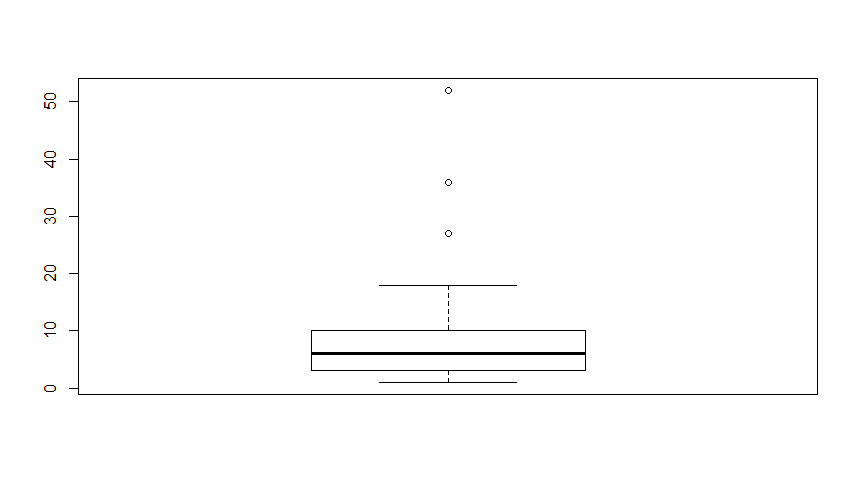
Pennsylvania Ohio Michigan

18 16 14

Georgia North Carolina

13 13

1. The box plot of house members is:



The outlier states are:

statedf[statedf$Freq %in% outvals, ];

Var1 Freq

5 California 52

9 Florida 27

31 New York 27

42 Texas 36

1. The average number of years served by party in the House and Senate are:

House:

[1] "Republican 7.814346"

[1] "Democratic 10.699482"

Senate:

[1] "Democratic 9.521739"

[1] "Independent 7.000000"

[1] "Republican 10.442308"