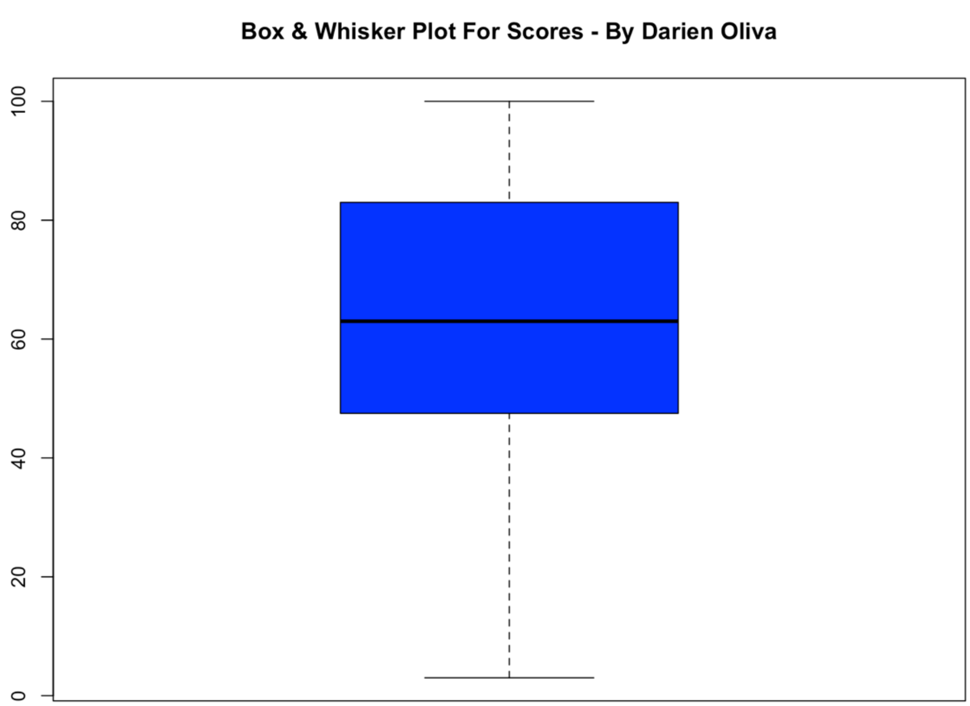
DARIEN OLIVA
ECON 453 – PSET1

1A) Summary Statistics for Scores

Minimum	3
Maximum	100
Sample Mean	62.37
Sample Variance	546.0333
Sample Standard Deviation	23.36735
Coefficient of Variation	37.4657
Mean Absolute Deviation	18.6863
Q1	47.75
Median	63
Q3	83
IQR	35.25

1B) BOX & WHISKER PLOT FOR SCORES



1C)

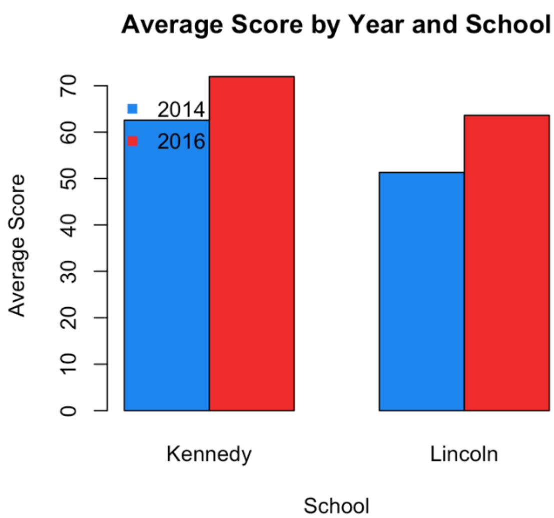
SAMPLE MEAN & STANDARD DEVIATION BY SCHOOL (2 DECIMALS)

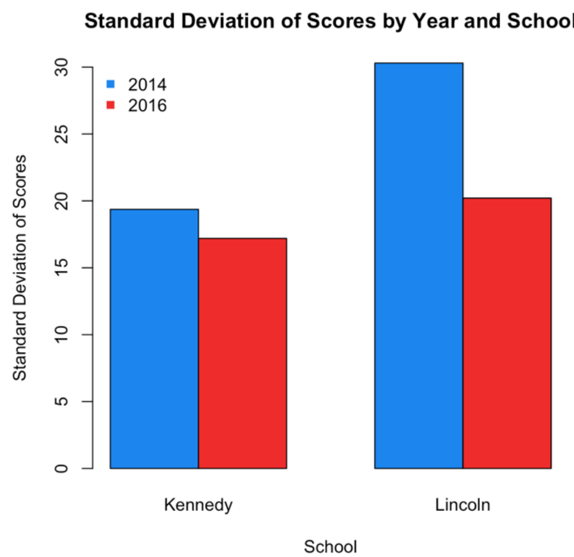
	Lincoln	Kennedy
Sample Mean	57.47	67.27
Sample Standard Deviation	26.36	18.82

SAMPLE MEAN & STANDARD DEVIATION BY YEAR (2 DECIMALS)

	2014	2016
Sample Mean	56.95	67.79
Sample Standard Deviation	25.93	19.13

Bar Charts (made in R)





1D) Based on these numbers, we can see that the Kennedy School has a higher mean for test scores than the Lincoln School by approximately 10 points. As the max score is 100, and the matter is testing scores, I would say that this is a decently significant difference, and the school with the lower mean should definitely be interested in why this is. For the standard deviation of scores, both schools experienced a decrease of this from the year 2014 to the year 2016, notably a much larger decrease for the Lincoln School. This decreased standard deviation of scores for the Lincoln school could be a reflection on better or more efficient teaching or administration practices.

2A) Summary Statistics for All Numeric Values

Category	Med Exp	Income	Education
Minimum	1	4	0
Maximum	62.231	99	18
Sample Mean	19.188	37.42	10.18
Sample Variance	201.7416	399.4613	22.33754
Sample Standard Deviation	14.20358	19.98653	4.72626
Coefficient of Variation	74.02509	53.40632	46.44302
Mean Absolute Deviation	11.62128	15.87017	3.9391
Q1	8.208	21	6
Median	16.351	34	11
Q3	26.822	48	13
IQR	18.614	27	7

2B) Outliers in the data for medical expenses?

I used option ii from the instruction sheet (data point that is outside [Q1- 1,5•IQR, Q3+1.5•IQR]) and found that there was only 1 outlier for medical expenses, being 62.231.

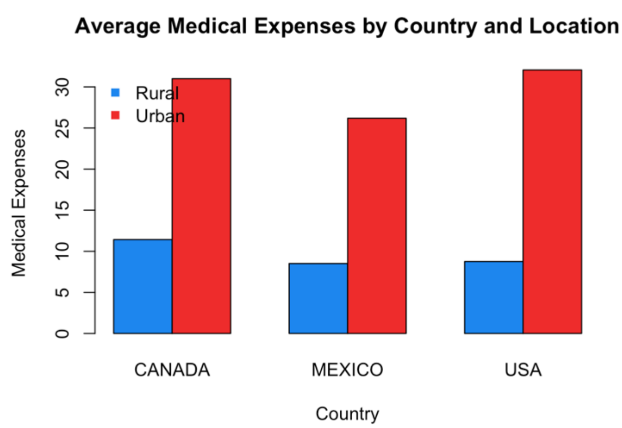
2C)

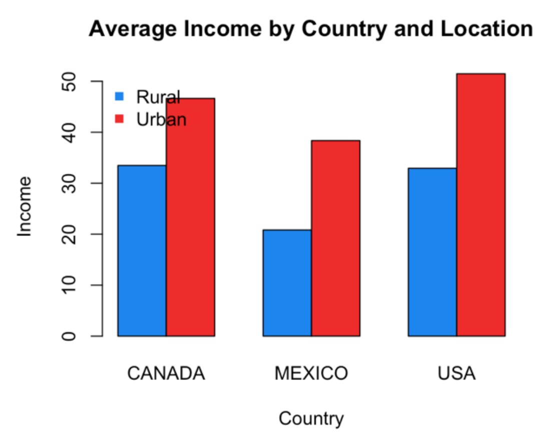
MEDICAL EXP	Sample Mean	Sample Standard Dev.
CANADA URBAN	31.00	8.18
MEXICO URBAN	26.19	15.52
USA URBAN	32.06	12.29
CANADA RURAL	11.43	7.63
MEXICO RURAL	8.51	7.32
USA RURAL	8.76	7.13

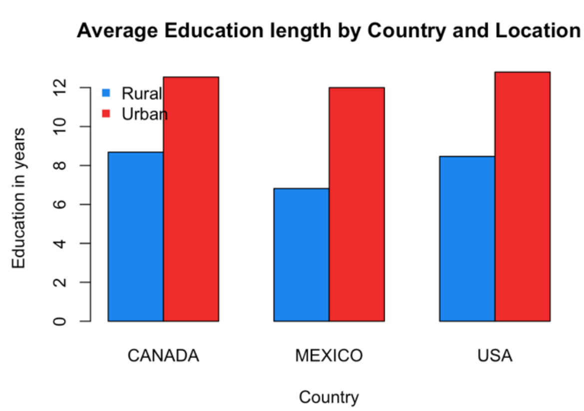
INCOME	Sample Mean	Sample Standard Dev.
CANADA URBAN	46.64	16.13
MEXICO URBAN	38.36	24.65
USA URBAN	51.57	20.88
CANADA RURAL	33.47	17.24

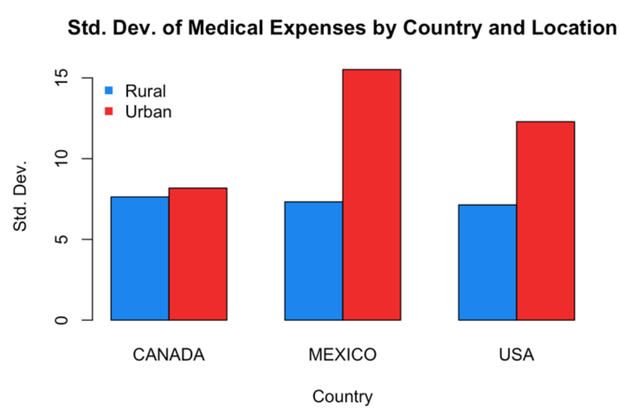
MEXICO RURAL	20.82	13.66
USA RURAL	32.93	16.69

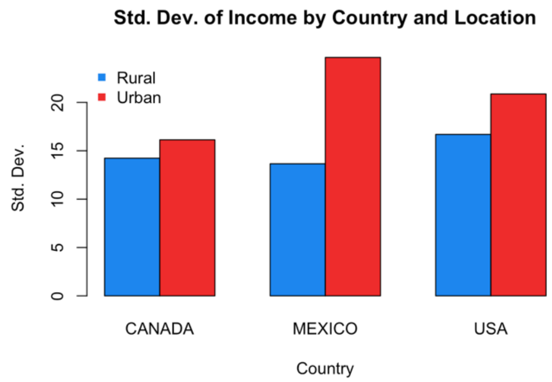
EDUCATION	Sample Mean	Sample Standard Dev.
CANADA URBAN	12.55	3.91
MEXICO URBAN	12.00	5.67
USA URBAN	12.80	4.13
CANADA RURAL	8.68	3.68
MEXICO RURAL	6.82	4.45
USA RURAL	8.47	3.81

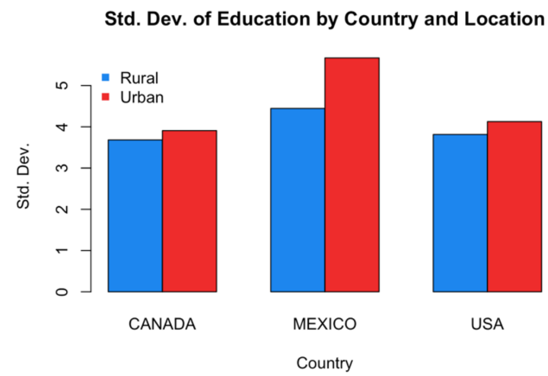






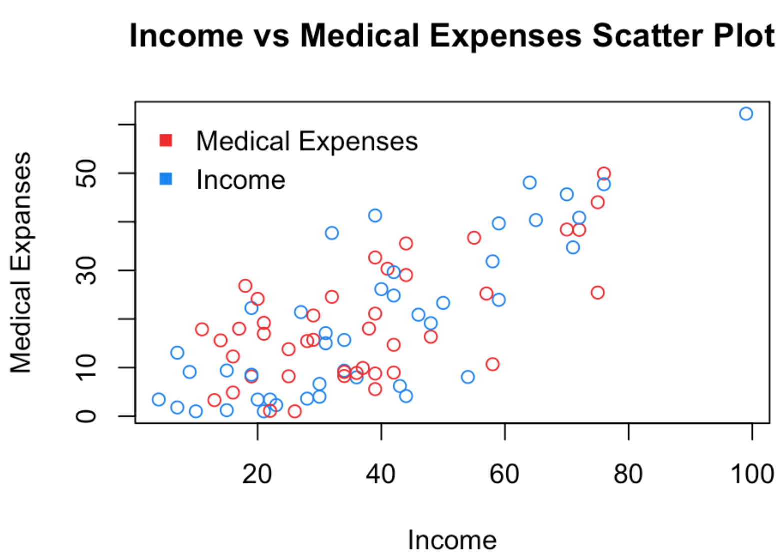






2D) Based on the generated sample statistics and the bar charts produced in R, we can see how the different countries and the different types of households compare. First, we can see that in urban househoulds the average medical expenses are much higher, which could be due to there being greater access to medical care in urban areas, and that the urban professionals are more expensive. Next, we can see that the standard deviations for the categories such as average medical expenses, income, and education, that Mexico has the highest, meaning they have the greatest variabilities in their data. Canada has the lowest stanadrd deviations or variabilities in these categories, and the USA is inbetween. Also important to note, that the rural averages are lower for every single category, and in every country, which is probablu due to the much different lifestsyles held by those in rural locations.

2E) Scatter Plot for Income vs Medical Expenses



2F) <u>Sample Correlations for all numeric values</u> (2 decimals)

	Medical Exp.	Income	Education
Medical Exp.	1.00	0.75	0.69

Income	0.75	1.0	0.68
Education	0.69	0.69	1.00

Medical expenses and income have a moderately to strong, positive correlation of 0.75. Medical expenses and education have a slightly lower correlation, however still in the moderately good category being at 0.69. The magnitude is important because it shows us that there is definitely a significant positive correlation, so the categories here are related, and move in the same direction.

ADDEDNDUM

DARIEN OLIVA _ PSET1 _R-Code

1A)

```
> #Darien Oliva - Econ 453 - PSET1
> min(pset1_data$score)
[1] 3
> max(pset1_data$score)
[1] 100
> mean(pset1_data$score)
[1] 62.37
> var(pset1_data$score)
[1] 546.0333
> sd(pset1_data$score)
[1] 23.36735
> CV <- sd(pset1_data$score) / mean(pset1_data$score) * 100
> CV
[1] 37.4657
> X <- pset1_data$score
> mean(abs(X-mean(X)))
[1] 18.6863
> Q1 <- quantile(X,0.25)
> Q1
47.75
> median(X)
[1] 63
> Q3 <- quantile(X,0.75)
> Q3
 83
> IQR <- quantile(X,0.75)-quantile(X,0.25)
> IQR
35.25
```

NOTE: COULD HAVE ALSO USED > summary(pset1_data) or > summary(X) for most of these.

1B) > boxplot(X, col = "blue", main="Box & Whisker Plot For Scores - By Darien Oliva")

1C)

```
> aggregate.data.frame(x=pset1_data$score,by=list(pset1_data$year),FUN =
mean)
  Group.1     x
1    2014 56.95
2    2016 67.79
>aggregate.data.frame(x=pset1_data$score,by=list(pset1_data$school),FUN =
mean)
  Group.1     x
1 Kennedy 67.27
2 Lincoln 57.47

> aggregate.data.frame(x=pset1_data$score,by=list(pset1_data$year),FUN = sd)
  Group.1        x
1    2014 25.92623
2    2016 19.13287
> aggregate.data.frame(x=pset1_data$score,by=list(pset1_data$school),FUN =
sd)
  Group.1        x
1 Kennedy 18.82179
2 Lincoln 26.35939
```

OR COULD USE:

```
mean_S=aggregate(pset1_data$score, list(pset1_data$school,pset1_data$year),
FUN=mean)
>
> mean_S
  Group.1 Group.2     x
1 Kennedy    2014 62.58
2 Lincoln    2014 51.32
3 Kennedy    2016 71.96
4 Lincoln    2016 63.62

> mean_Ss=tapply(pset1_data$score,  pset1_data[,c(1,2)], mean)
> barplot(mean_Ss, beside=T,
+         col=c("dodgerblue2","firebrick2"),
+         main = "Average Score by Year and School", ylab = "Average Score",
+         xlab = "School")
> legend("topleft", c("2014","2016"), pch=15, bty="n",
+        col=c("dodgerblue2","firebrick2"))

> sd_S=aggregate(pset1_data$score, list(pset1_data$school,pset1_data$year),
FUN=sd)
> sd_S
```

```
  Group.1 Group.2        x
1 Kennedy    2014 19.36765
2 Lincoln    2014 30.30285
3 Kennedy    2016 17.19748
4 Lincoln    2016 20.20799
```

```
> sd_Ss=tapply(pset1_data$score,  pset1_data[,c(1,2)], sd)
> barplot(sd_Ss, beside=T,
+ col=c("dodgerblue2","firebrick2"),
+ main = "Standard Deviation of Scores by Year and School", ylab = "Standard
Deviation of Scores", xlab = "School")
> legend("topleft", c("2014","2016"), pch=15, bty="n",
+ col=c("dodgerblue2","firebrick2"))
```

1D)   N/A

2A)

```
> pset1_data <- read_excel("Downloads/pset1_data.xlsx",
+     sheet = "medical_expenses")
> View(pset1_data)
> summary(pset1_data$medicalexpn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   8.208  16.351  19.188  26.822  62.231
> summary(pset1_data$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   21.00   34.00   37.42   48.00   99.00
> summary(pset1_data$education)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   11.00   10.18   13.00   18.00
> var(pset1_data$medicalexpn)
[1] 201.7416
> var(pset1_data$income)
[1] 399.4613
> var(pset1_data$education)
[1] 22.33754
> sd(pset1_data$medicalexpn)
[1] 14.20358
> sd(pset1_data$income)
[1] 19.98653
> sd(pset1_data$education)
[1] 4.72626
> CV <- sd(pset1_data$medicalexpn) / mean(pset1_data$medicalexpn) * 100
> CV
[1] 74.02509
```

```
>
> CV <- sd(pset1_data$medicalexpn) / mean(pset1_data$medicalexpn) * 100
> CV
[1] 74.02509
> CV <- sd(pset1_data$income) / mean(pset1_data$income) * 100
> CV
[1] 53.40632
> CV <- sd(pset1_data$education) / mean(pset1_data$education) * 100
> CV
[1] 46.44302
> mean(abs(pset1_data$medicalexpn-mean(pset1_data$medicalexpn)))
[1] 11.62128
> mean(abs(pset1_data$income-mean(pset1_data$income)))
[1] 15.87017
> mean(abs(pset1_data$education-mean(pset1_data$education)))
[1] 3.9391
> IQR <- quantile(pset1_data$medicalexpn,0.75)-
quantile(pset1_data$medicalexpn,0.25)
> IQR

18.614
> IQR <- quantile(pset1_data$income,0.75)-quantile(pset1_data$income,0.25)
> IQR
27
> IQR <- quantile(pset1_data$education,0.75)-
quantile(pset1_data$education,0.25)
> IQR
  7
```

2B)

```
> IQR_med=quantile(pset1_data$medicalexpn,0.75)-
quantile(pset1_data$medicalexpn,0.25)

> IQR_med

   75%

18.614

> low_med=quantile(pset1_data$medicalexpn,0.25)-1.5*IQR_med

> low_med

    25%

-19.713
```

```
> high_med=quantile(pset1_data$medicalexpn,0.75)+1.5*IQR_med

> high_med

   75%

54.743

> pset1_data$medicalexpn[which(pset1_data$medicalexpn <low_med |
pset1_data$medicalexpn > high_med)]

[1] 62.231
```

2C)

```
> medexpmean=aggregate(pset1_data$education,
list(pset1_data$country,pset1_data$location), FUN=mean)

> medexpmean

  Group.1 Group.2         x

1  CANADA   RURAL 11.430579

2  MEXICO   RURAL  8.512818

3     USA   RURAL  8.761600

4  CANADA   URBAN 31.003091

5  MEXICO   URBAN 26.191786

6     USA   URBAN 32.064933

> barplot(x~Group.2+Group.1 ,data=medexpmean, beside=T,

+ col=c("dodgerblue2","firebrick2"),

+ main = "Average Medical Expenses by Country and Location", ylab = "Medical
Expenses", xlab = "Country")

> legend("topleft",c("Rural","Urban"), pch=15, bty="n",

+ col=c("dodgerblue2","firebrick2"))
```

```
> incomemean=aggregate(pset1_data$income,
list(pset1_data$country,pset1_data$location), FUN=mean)

> incomemean

  Group.1 Group.2        x

1  CANADA   RURAL 33.47368

2  MEXICO   RURAL 20.81818

3     USA   RURAL 32.93333

4  CANADA   URBAN 46.63636

5  MEXICO   URBAN 38.35714

6     USA   URBAN 51.46667

> barplot(x~Group.2+Group.1 ,data=incomemean, beside=T,

+ col=c("dodgerblue2","firebrick2"),

+ main = "Average Income by Country and Location", ylab = "Income", xlab =
"Country")

> legend("topleft",c("Rural","Urban"), pch=15, bty="n",

+ col=c("dodgerblue2","firebrick2"))

> educationmean=aggregate(pset1_data$education,
list(pset1_data$country,pset1_data$location), FUN=mean)

> educationmean

  Group.1 Group.2         x

1  CANADA   RURAL  8.684211

2  MEXICO   RURAL  6.818182

3     USA   RURAL  8.466667

4  CANADA   URBAN 12.545455

5  MEXICO   URBAN 12.000000
```

```
6     USA   URBAN 12.800000

> barplot(x~Group.2+Group.1 ,data=educationmean, beside=T,

+ col=c("dodgerblue2","firebrick2"),

+ main = "Average Education length by Country and Location", ylab =
"Education in years", xlab = "Country")

> legend("topleft",c("Rural","Urban"), pch=15, bty="n",

+ col=c("dodgerblue2","firebrick2"))

> medexpsd=aggregate(pset1_data$medicalexpn,
list(pset1_data$country,pset1_data$location), FUN=sd)

> medexpsd

  Group.1 Group.2         x

1  CANADA   RURAL  7.629464

2  MEXICO   RURAL  7.324204

3     USA   RURAL  7.132344

4  CANADA   URBAN  8.175206

5  MEXICO   URBAN 15.517216

6     USA   URBAN 12.288999

> barplot(x~Group.2+Group.1 ,data=medexpsd, beside=T,

+ col=c("dodgerblue2","firebrick2"),

+ main = "Std. Dev. of Medical Expenses by Country and Location", ylab =
"Std. Dev.", xlab = "Country")

> legend("topleft",c("Rural","Urban"), pch=15, bty="n",

+ col=c("dodgerblue2","firebrick2"))

> incomesd=aggregate(pset1_data$income,
list(pset1_data$country,pset1_data$location), FUN=sd)

> incomesd
```

```
  Group.1 Group.2        x
1  CANADA   RURAL 14.23754
2  MEXICO   RURAL 13.65883
3     USA   RURAL 16.69246
4  CANADA   URBAN 16.13241
5  MEXICO   URBAN 24.65019
6     USA   URBAN 20.87674
> barplot(x~Group.2+Group.1 ,data=incomesd, beside=T,
+ col=c("dodgerblue2","firebrick2"),
+ main = "Std. Dev. of Income by Country and Location", ylab = "Std. Dev.", xlab = "Country")
> legend("topleft",c("Rural","Urban"), pch=15, bty="n",
+ col=c("dodgerblue2","firebrick2"))
> educationsd=aggregate(pset1_data$education, list(pset1_data$country,pset1_data$location), FUN=sd)
> educationsd
  Group.1 Group.2        x
1  CANADA   RURAL 3.682581
2  MEXICO   RURAL 4.445631
3     USA   RURAL 3.814758
4  CANADA   URBAN 3.908034
5  MEXICO   URBAN 5.670436
6     USA   URBAN 4.126569
> barplot(x~Group.2+Group.1 ,data=educationsd, beside=T,
+ col=c("dodgerblue2","firebrick2"),
```

```
+ main = "Std. Dev. of Education by Country and Location", ylab = "Std.
Dev.", xlab = "Country")

> legend("topleft",c("Rural","Urban"), pch=15, bty="n",

+ col=c("dodgerblue2","firebrick2"))
```

2D) N/A

2E)
```
> plot(pset1_data$income,pset1_data$medicalexpn, main="Income vs Medical
Expenses Scatter Plot",xlab = "Income",ylab = "Medical Expanses",
col=c("dodgerblue2","firebrick2"))

> legend("topleft",c("Medical Expenses","Income"), pch=15, bty="n",

+ col=c("firebrick2","dodgerblue2"))
```

2F)
```
> cor(pset1_data[,c(3,4,5)])

            medicalexpn    income education

medicalexpn   1.0000000 0.7482683 0.6895336

income        0.7482683 1.0000000 0.6844115

education     0.6895336 0.6844115 1.0000000

> cor(cbind(pset1_data$medicalexpn,pset1_data$income,pset1_data$education))

          [,1]      [,2]      [,3]

[1,] 1.0000000 0.7482683 0.6895336

[2,] 0.7482683 1.0000000 0.6844115

[3,] 0.6895336 0.6844115 1.0000000
```