Problem Set 1 Report

Minh Duong

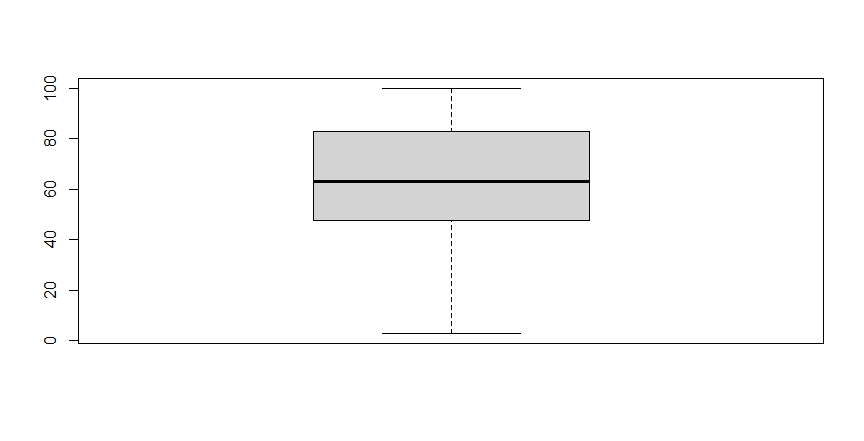
2023 – 02 – 06

1. **Scores**
   1. Summary statistics for scores, including min, max, sample mean, sample variance, sample standard deviation, coefficient of variation, mean absolute deviation, Q1, median, Q3, and IQR. Results are presented on well-formatted table with a title.

|  |  |
| --- | --- |
| **Statistics name** | **Value** |
| Min | 3 |
| Max | 100 |
| Sample Mean | 62.37 |
| Sample Variance | 546.033266 |
| Standard Deviation | 23.3673547 |
| Coefficient of Variance | 0.37465696 |
| Mean Average Deviation | 18.6863 |
| Interquartile Range (IQR) | 35.25 |

**Summary Statistics for Scores**

* 1. Box and whisker plot for score (on next page).

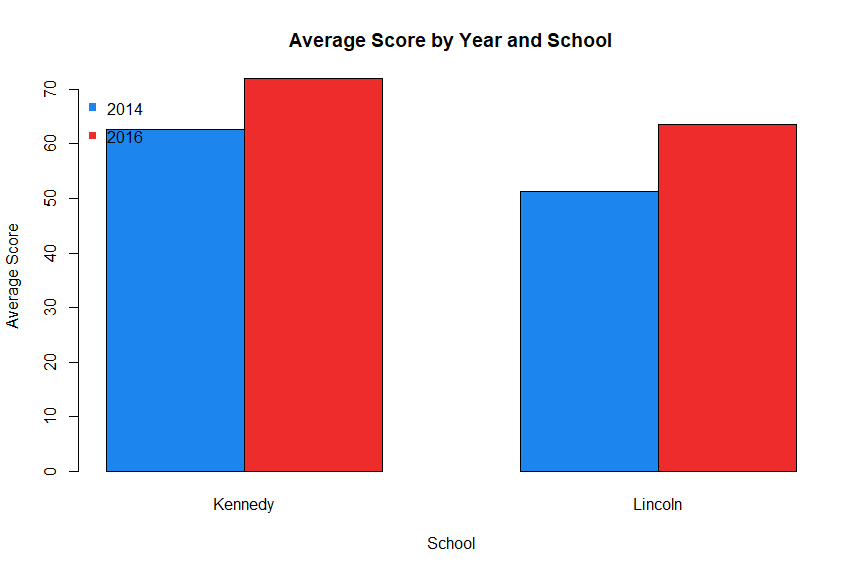


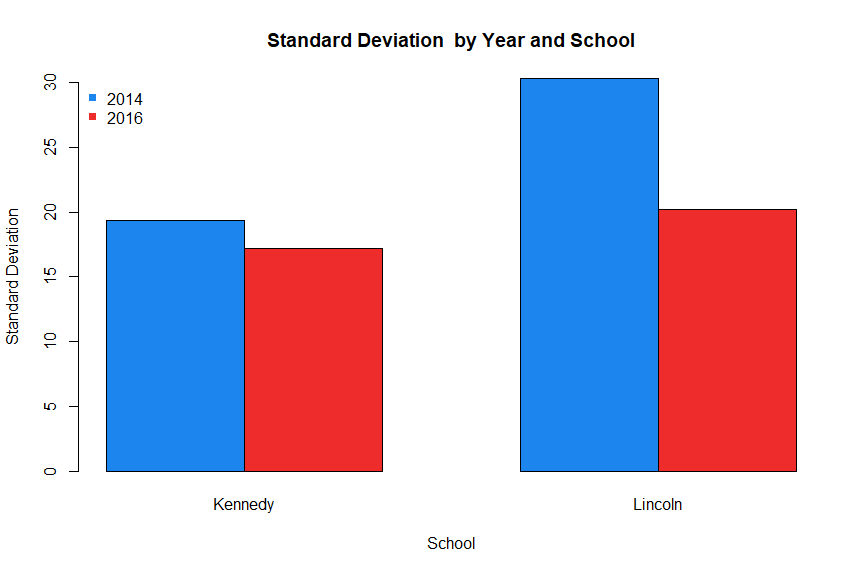
**Box and whisker plot for Scores**

* 1. Summary for scores by year and by school, including sample mean and sample standard deviation. Statistics are presented on a well-formatted table and a bar chart diagram.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **School** | **Sample Mean** | **Sample Standard Deviation** |
| 2014 | Kennedy | 62.58 | 19.36765 |
| 2016 | Kennedy | 71.96 | 17.19748 |
| 2014 | Lincoln | 51.32 | 30.30285 |
| 2016 | Lincoln | 63.62 | 20.20799 |

**Summary Statistic for Scores, aggregated by Year and School**

****

****

* 1. School Comparison:

Based on the numerical data we get, in both years Kennedy High School has a better average score than Lincoln High School, which can indicate a better academic achievement of the students in Kennedy High School. It is also worth noting here that both schools have larger average scores in 2016 than in 2014, which suggests an improvement over the scores from 2014 to 2016 in both schools. Regarding the Standard Deviation, the scores in Lincoln High School have a larger Standard Deviation, which suggests a higher dispersion in the scores of the school. It also means that there is a bigger difference in the academic performance of students at Lincoln High school than that of Kennedy School. In terms of trend, both schools features lower standard deviation in 2016, indicating that the students became to have a more similar performances to each other.

1. **Medical Expenses**
   1. Summary statistics for all numeric variables, including min, max, sample mean, sample variance, sample standard deviation, coefficient of variation, mean absolute deviation, Q1, median, Q3, and IQR. The results are presented on a well-formatted table with a title.

|  |  |
| --- | --- |
| **Statistics name** | **Values** |
| Medical Expenses - Min | 1 |
| Medical Expenses - Max | 62.231 |
| Medical Expenses - Sample Mean | 19.18752 |
| Medical Expenses - Sample Variance | 201.7416 |
| Medical Expenses - sample standard deviation | 14.20358 |
| Medical Expenses - coefficient of variation | 0.740251 |
| Medical Expenses - mean absolute deviation | 11.62128 |
| Medical Expenses - Q1 | 8.208 |
| Medical Expenses - median | 16.351 |
| Medical Expenses - Q3 | 26.822 |
| Medical Expenses - IQR | 18.614 |
| Income - Min | 4 |
| Income - Max | 99 |
| Income - Sample Mean | 37.42353 |
| Income - Sample Variance | 399.4613 |
| Income - sample standard deviation | 19.98653 |
| Income - coefficient of variation | 0.534063 |
| Income - mean absolute deviation | 15.87017 |
| Income - Q1 | 21 |
| Income - median | 34 |
| Income - Q3 | 48 |
| Income - IQR | 27 |
| Education - Min | 0 |
| Education - Max | 18 |
| Education - Sample Mean | 10.17647 |
| Education - Sample Variance | 22.33754 |
| Education - sample standard deviation | 4.72626 |
| Education - coefficient of variation | 0.46443 |
| Education - mean absolute deviation | 3.9391 |
| Education - Q1 | 6 |
| Education - median | 11 |
| Education - Q3 | 13 |
| Education - IQR | 7 |

**Summary Statistics for Scores**

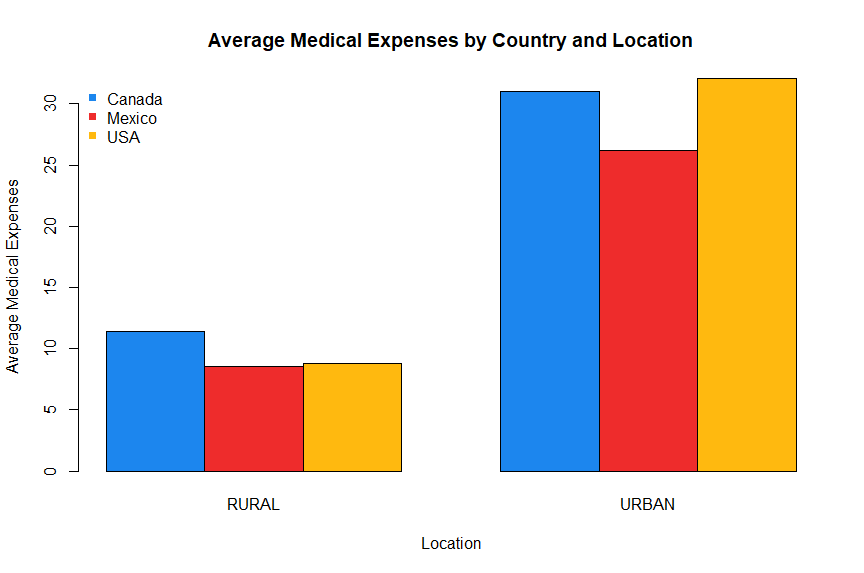
* 1. Outlier identification for medical expenses:

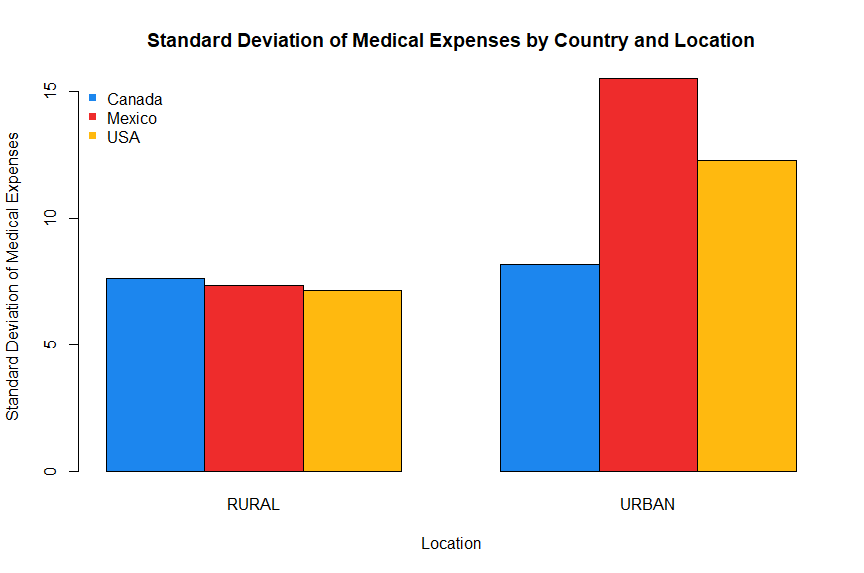
Based on the data we got (in the R code), the Interquartile Range (IQR) of medical expenses is 18.164 (unit: $100). Based on the definition that the outliers are data points outside the range of , we get the outlier is the value 62.231 of row 27 (or row 28 if we do take into account the first row containing the names of the fields).

* 1. Summary statistics for all numeric variables, including Sample Mean and Sample Standard Deviation, by country and by location (urban/rural). The results are presented in well-formatted tables and bar chart diagrams.

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Location** | **Medical Expenses - Sample Mean** | **Medical Expenses - Sample Standard Deviation** |
| CANADA | RURAL | 11.43058 | 7.629464 |
| MEXICO | RURAL | 8.512818 | 7.324204 |
| USA | RURAL | 8.7616 | 7.132344 |
| CANADA | URBAN | 31.00309 | 8.175206 |
| MEXICO | URBAN | 26.19179 | 15.517216 |
| USA | URBAN | 32.06493 | 12.288999 |

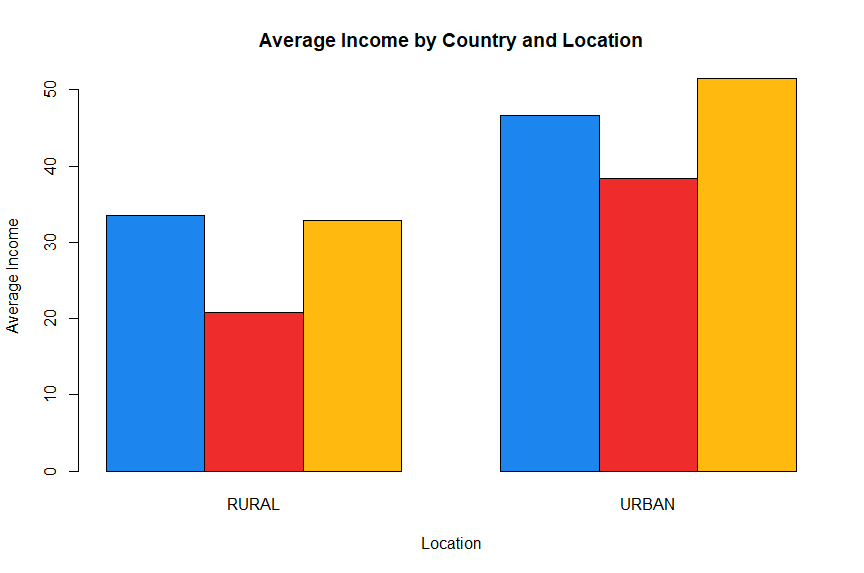
**Summary Statistics for Medical Expenses**

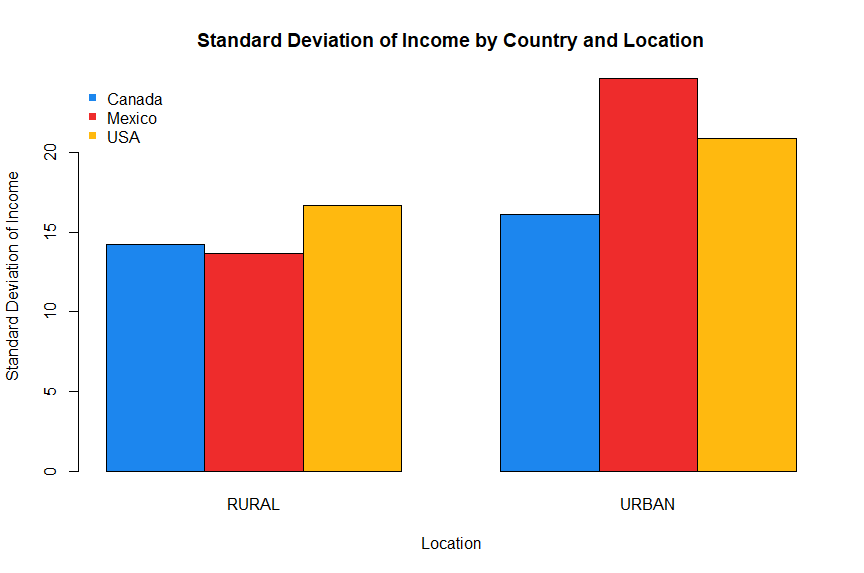
****

****

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Location** | **Income - Sample Mean** | **Income - Sample Standard Deviation** |
| CANADA | RURAL | 33.47368 | 14.23754 |
| MEXICO | RURAL | 20.81818 | 13.65883 |
| USA | RURAL | 32.93333 | 16.69246 |
| CANADA | URBAN | 46.63636 | 16.13241 |
| MEXICO | URBAN | 38.35714 | 24.65019 |
| USA | URBAN | 51.46667 | 20.87674 |

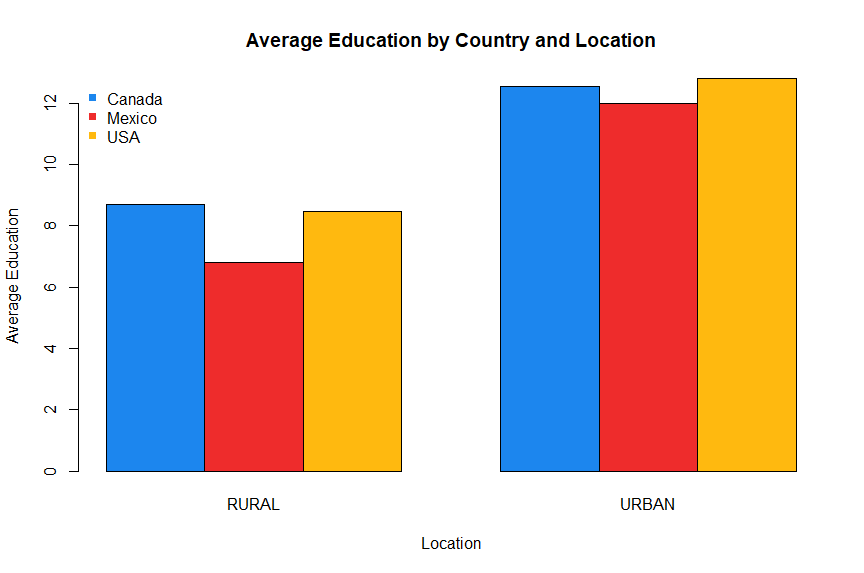
**Summary Statistics for Income**

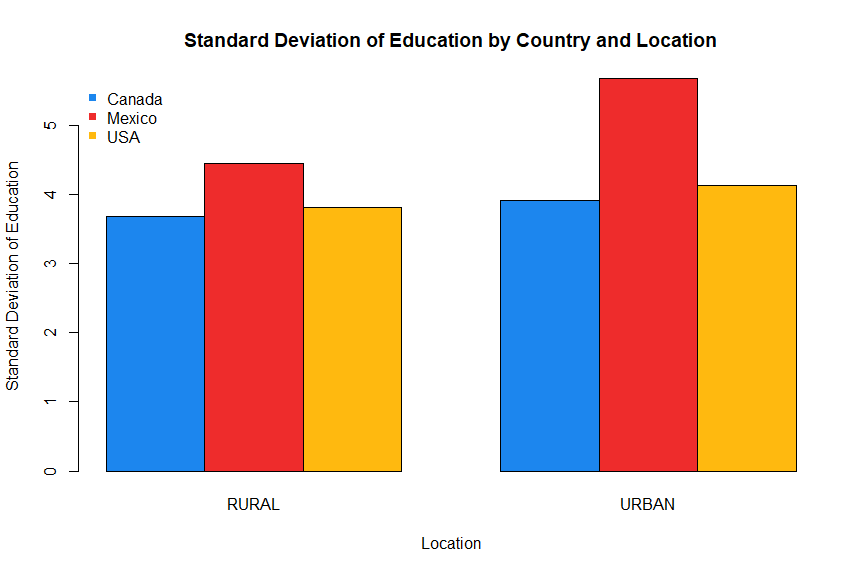
****

****

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Location** | **Education - Sample Mean** | **Education - Sample Standard Deviation** |
| CANADA | RURAL | 8.684211 | 3.682581 |
| MEXICO | RURAL | 6.818182 | 4.445631 |
| USA | RURAL | 8.466667 | 3.814758 |
| CANADA | URBAN | 12.54546 | 3.908034 |
| MEXICO | URBAN | 12 | 5.670436 |
| USA | URBAN | 12.8 | 4.126569 |

**Summary Statistics for Education**

****

****

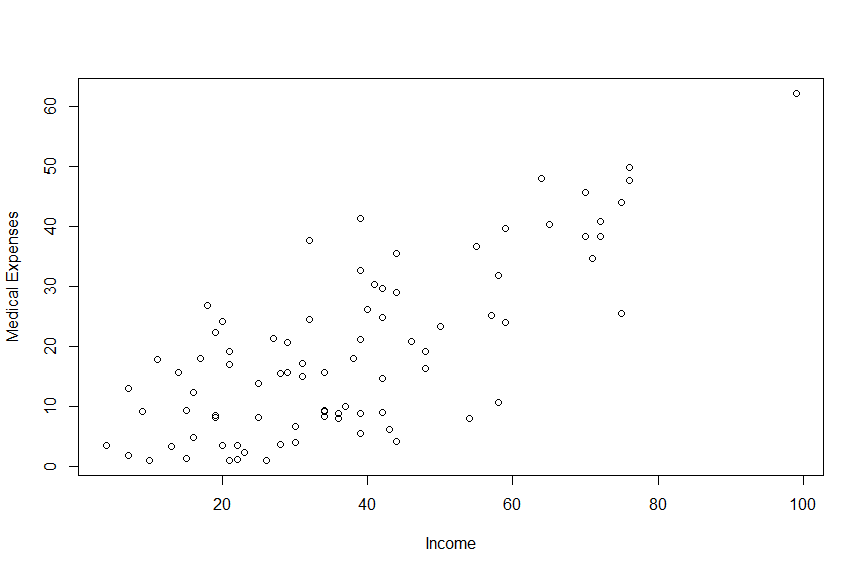
* 1. Comparison among countries and locations:

Firstly, for medical expenses, the average values in urban households are significantly higher than that of rural households (in respective countries). Canada and USA have a higher average value than Mexico. Regarding the standard deviation, the three countries have a relatively similar standard deviation for the rural areas. It is noteworthy that for the urban areas, Mexico has a significantly higher standard deviation compared to the other two countries, indicating highly differentiated medical expenses among their urban households.

A similar fashion can be found in the income data. The urban areas have significantly higher average incomes than their rural counterparts. Canada and USA have a higher average income than Mexico. However, Mexico has a significantly higher standard deviation, indicating highly dispersed income data points in their urban citizens.

In terms of education, the urban areas also have higher average education than the rural areas. Canada and USA also lead in the average education. Mexico has a significantly high standard deviation in both its rural and urban areas, which implies a relatively large difference in the level of education among Mexican households.

* 1. Medical Expenses and Income



**Scatter plot of Medical Expenses and Income**

Scatter plot inference: From the Scatter plot, we can infer that the majority of medical expenses are less than $5,000, and the majority of the income is less than $80,000. There is one outlier with medical expenses of approximately $6,000 and income of nearly $100,000, which has been reported in part B. We can also observe a relatively positively linear relationship between medical expenses and income, where the income increases with medical expenses and vice versa. This relationship will be further explored in part F below.

* 1. Sample correlations between all numeric variables and present them in a table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Medical Expenses** | **Income** | **Education** |
| **Medical Expenses** | 1 | 0.748268 | 0.689534 |
| **Income** | 0.748268 | 1 | 0.684411 |
| **Education** | 0.689534 | 0.684411 | 1 |

**Sample Correlations among Medical Expenses, Income, and Education**

Firstly, regarding the medical expenses and income, we can observe that they have a correlation of 0.75, which is positive and relatively close to 1, indicating a strong positive linear association between the two variables. It means that as income increases, we can observe a proportional increase in medical expenses and vice versa.

In terms of medical expenses and education, we can also see a relatively strong positive linear correlation between the two variables, with a correlation of 0.69, which is positive and close to 1. This also indicates an observation that as medical expenses increase, there is also an increase in education and vice versa.

Lastly, the relationship between income and education also follows a similar fashion, with a correlation of 0.68, implying a strong positive linear association, as it is positive and relatively close to 1. It means that income and education are observed to grow proportionally to each other.

In conclusion, the correlation values suggest a strong positive linear association among the three variables.

Addendum

Project Repository: <https://github.com/MykeDuong/econ453>

R Script code:

# Minh Duong, ECON 453, pset 1

# Packages Install && Import

library(readxl)

# Clear workspace

rm(list = ls())

# Relative directory

# Put data to the data directory inside the project directory

setwd(".")

getwd()

# Problem 1

data1<- read\_excel("data/pset1\_data.xlsx", sheet="scores")

summary(data1)

## 1A

# Add necessary summary statistics:

summary(data1)

data1\_summary <- as.data.frame(

apply(data1, 2, summary)

)

data1\_summary

data1\_summary = rbind(

min(data1$score),

max(data1$score),

mean(data1$score),

var(data1$score),

sd(data1$score),

sd(data1$score) / mean(data1$score),

mean(abs(data1$score - mean(data1$score))),

IQR(data1$score)

)

# Provide Row names

rownames(data1\_summary) <- c(

"Min",

"Max",

"Sample Mean",

"Sample Variance",

"Standard Deviation",

"Coefficient of Variance",

"Mean Average Deviation",

"Interquartile Range (IQR)"

)

colnames(data1\_summary) <- c(

"Value"

)

# Report the table:

data1\_summary

write.csv(data1\_summary, "exports/data1\_summary.csv", row.names = TRUE)

# 1B

boxplot(data1$score)

# 1C

# Sample Mean, Sample SD aggregated by year & school

aggregated\_data1 = aggregate(

data1$score,

list(

Year = data1$year,

School = data1$school

),

FUN = function(x) c(

"Sample Mean" = mean(x),

"Sample SD" = sd(x)

)

)

aggregated\_data1

write.csv(aggregated\_data1, "exports/aggregated\_data1\_by\_school\_and\_year.csv", row.names = TRUE)

# Bar Chart Draw

# Sample Mean Chart

barplot(

x[,"Sample Mean"] ~ Year + School,

data = aggregated\_data1,

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")

)

# Standard Deviation Chart

barplot(

x[,"Sample SD"] ~ Year + School ,

data = aggregated\_data1,

beside = T,

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

main = "Standard Deviation by Year and School",

ylab = "Standard Deviation",

xlab = "School"

)

legend(

"topleft",

c("2014", "2016"),

pch = 15,

bty = "n",

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

)

# Question 2

data2<- read\_excel("data/pset1\_data.xlsx", sheet="medical\_expenses");

# 2A

summary(data2)

data2\_summary <- as.data.frame(

apply(data2, 2, summary)

)

data2\_summary

data2\_summary = rbind(

# Medical Experience

min(data2$medicalexpn), # Min

max(data2$medicalexpn), # Max

mean(data2$medicalexpn),# Mean

var(data2$medicalexpn), # Variance

sd(data2$medicalexpn), # Standard Deviation

sd(data2$medicalexpn) / mean(data2$medicalexpn), # Co. of Variation

mean(abs(data2$medicalexpn - mean(data2$medicalexpn))), # Mean abs. Deviation

quantile(data2$medicalexpn, 0.25), # 1st Quartile

quantile(data2$medicalexpn, 0.5), # Median

quantile(data2$medicalexpn, 0.75), # 3rd Quartile

IQR(data2$medicalexpn), # Interquartile Range

# Income

min(data2$income),

max(data2$income),

mean(data2$income),

var(data2$income),

sd(data2$income),

sd(data2$income) / mean(data2$income),

mean(abs(data2$income - mean(data2$income))),

quantile(data2$income, 0.25),

quantile(data2$income, 0.5),

quantile(data2$income, 0.75),

IQR(data2$income),

# Education

min(data2$education),

max(data2$education),

mean(data2$education),

var(data2$education),

sd(data2$education),

sd(data2$education) / mean(data2$education),

mean(abs(data2$education - mean(data2$education))),

quantile(data2$education, 0.25),

quantile(data2$education, 0.5),

quantile(data2$education, 0.75),

IQR(data2$education)

)

# Provide Row names

rownames(data2\_summary) <- c(

# Medical Experience

"Medical Expenses - Min",

"Medical Expenses - Max",

"Medical Expenses - Sample Mean",

"Medical Expenses - Sample Variance",

"Medical Expenses - sample standard deviation",

"Medical Expenses - coefficient of variation",

"Medical Expenses - mean absolute deviation",

"Medical Expenses - Q1",

"Medical Expenses - median",

"Medical Expenses - Q3",

"Medical Expenses - IQR",

# Income

"Income - Min",

"Income - Max",

"Income - Sample Mean",

"Income - Sample Variance",

"Income - sample standard deviation",

"Income - coefficient of variation",

"Income - mean absolute deviation",

"Income - Q1",

"Income - median",

"Income - Q3",

"Income - IQR",

# Education

"Education - Min",

"Education - Max",

"Education - Sample Mean",

"Education - Sample Variance",

"Education - sample standard deviation",

"Education - coefficient of variation",

"Education - mean absolute deviation",

"Education - Q1",

"Education - median",

"Education - Q3",

"Education - IQR"

)

colnames(data2\_summary) <- c("Values")

data2\_summary

write.csv(data2\_summary, "exports/data2\_summary.csv", row.names = TRUE)

# 2B

# Outliers not in the range [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR]

IQR\_med = quantile(data2$medicalexpn, 0.75) - quantile(data2$medicalexpn, 0.25)

IQR\_med

# get the lower and higher bound

low\_med = quantile(data2$medicalexpn, 0.25) - 1.5 \* IQR\_med

high\_med = quantile(data2$medicalexpn, 0.75) + 1.5 \* IQR\_med

# identify the outliers

# Outlier value(s)

data2$medicalexpn[

which(data2$medicalexpn < low\_med | data2$medicalexpn > high\_med)

]

# Outlier Record(s) (Observations(s))

row.names(data2)[

which(data2$medicalexpn < low\_med | data2$medicalexpn > high\_med)

]

# => The outlier is the Value 62.231 of row 27

# 2C

aggregated\_medicalexpn = aggregate(

data2$medicalexpn,

list(

Country = data2$country,

Location = data2$location

),

FUN = function(x) c(

"Sample Mean" = mean(x),

"Sample SD" = sd(x)

)

)

write.csv(

aggregated\_medicalexpn,

"exports/aggregated\_data2\_medicalexpn.csv",

row.names = TRUE

)

# Bar Chart Draw

# Sample Mean Chart

barplot(

x[,"Sample Mean"] ~ Country + Location,

data = aggregated\_medicalexpn,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Average Medical Expenses by Country and Location",

ylab = "Average Medical Expenses",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

# Standard Deviation Chart

barplot(

x[,"Sample SD"] ~ Country + Location,

data = aggregated\_medicalexpn,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Standard Deviation of Medical Expenses by Country and Location",

ylab = "Standard Deviation of Medical Expenses",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

aggregated\_income = aggregate(

data2$income,

list(

Country = data2$country,

Location = data2$location

),

FUN = function(x) c(

"Sample Mean" = mean(x),

"Sample SD" = sd(x)

)

)

write.csv(

aggregated\_income,

"exports/aggregated\_data2\_income.csv",

row.names = TRUE

)

# Bar Chart Draw

# Sample Mean Chart

barplot(

x[,"Sample Mean"] ~ Country + Location,

data = aggregated\_income,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Average Income by Country and Location",

ylab = "Average Income",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

# Standard Deviation Chart

barplot(

x[,"Sample SD"] ~ Country + Location,

data = aggregated\_income,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Standard Deviation of Income by Country and Location",

ylab = "Standard Deviation of Income",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

aggregated\_education = aggregate(

data2$education,

list(

Country = data2$country,

Location = data2$location

),

FUN = function(x) c(

"Sample Mean" = mean(x),

"Sample SD" = sd(x)

)

)

write.csv(

aggregated\_education,

"exports/aggregated\_data2\_education.csv",

row.names = TRUE

)

# Bar Chart Draw

# Sample Mean Chart

barplot(

x[,"Sample Mean"] ~ Country + Location,

data = aggregated\_education,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Average Education by Country and Location",

ylab = "Average Education",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

# Standard Deviation Chart

barplot(

x[,"Sample SD"] ~ Country + Location,

data = aggregated\_education,

beside = T,

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1"),

main = "Standard Deviation of Education by Country and Location",

ylab = "Standard Deviation of Education",

xlab = "Location"

)

legend(

"topleft",

c("Canada", "Mexico", "USA"),

pch = 15,

bty = "n",

col = c("dodgerblue2", "firebrick2", "darkgoldenrod1")

)

# 2E - Draw a scatter plot of medical expenses (on y-axis) and income (on x-axis).

plot(medicalexpn ~ income, data = data2)

plot(

data2$income,

data2$medicalexpn,

xlab = "Income",

ylab = "Medical Expenses"

)

# 2F - Calculate sample correlations between all numeric variables and present them in a table.

correlation = cor(data2[,c(3, 4, 5)])

correlation

write.csv(

correlation,

"exports/data2\_correlation.csv",

row.names = TRUE

)