Jessie Wong
STA4155 - EMWA
Professor Yue
September 14, 2022

R HW #1

Question 1: Produce a scatterplot between the Cost of Living Index and EACH of the other index variables. As a result, there should be 4 scatterplots in total. Examine the relationship shown in each scatterplot in terms of its form, strength, and direction.

Answer:
For Rent and Cost of Living, the scatterplot indicates a moderate positive correlation: when rent increases, so does the cost of living. The direction of the association in the scatterplot is positive. The form of the association is curved because the outliers at the value of 140 units of Rent.Index pulled the graph downwards. At the start of the graph, the points are close together and is forming a linear line, but when Rent is at 60 units, the points start to disperse in opposite directions, leading to outliers and a curved form.

For Groceries and Cost of Living, the scatterplot has a strong positive correlation, and it's clear that as grocery prices increase, so does the Cost of Living. The direction of the association is positive. The majority of points are cluttered together, forming an almost perfect correlation. There are also outliers at the end, but they form the rest of the linear line and doesn't go in other directions. The form of the association is straight. Overall, the graph depicts a strong positive correlation between Groceries and Cost of Living.

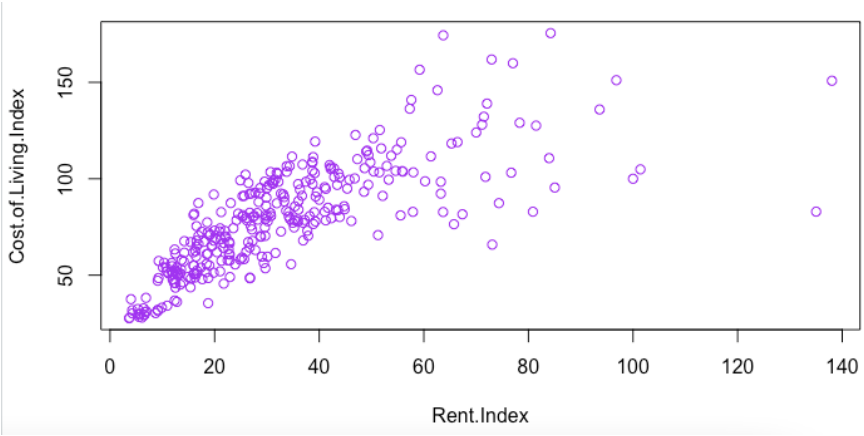
For Restaurant Prices and Cost of Living, the scatterplot's strength and direction of association is a strong positive correlation, and it's clear that the direction of association is positive. The points are accumulated together to form a positive straight linear line from 0 - 100 units in Restaurant Price, and although there are outliers, they are still within the linear line and doesn't change the direction, or form, of the line.

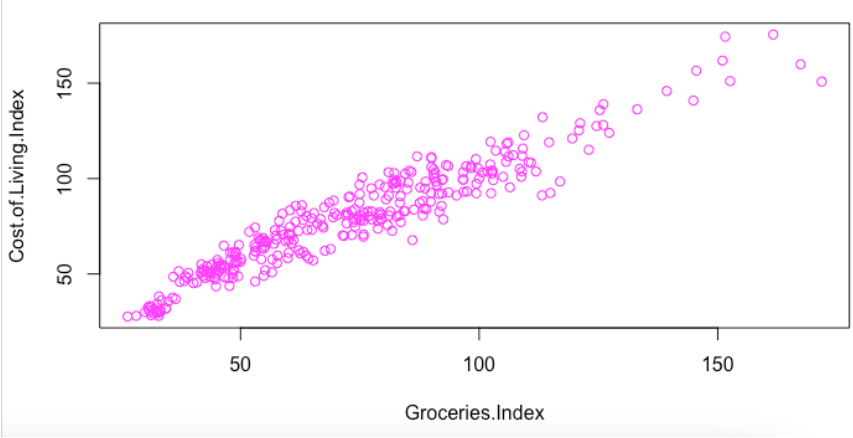
For Local Purchasing Power and Cost of Living, the scatterplot's direction and strength of association depicts a weak positive correlation because many points are scattered everywhere. There is some evidence that an increase in Local Purchasing Power leads to an increase in the Cost of Living. The form of association is slightly curved downwards from 100-150 units on the x-axis. There is also clustering of points in the bottom left and in the middle part of the graph. In conclusion, there seems to be a weak positive correlation between Local Purchasing Power and Cost of Living.

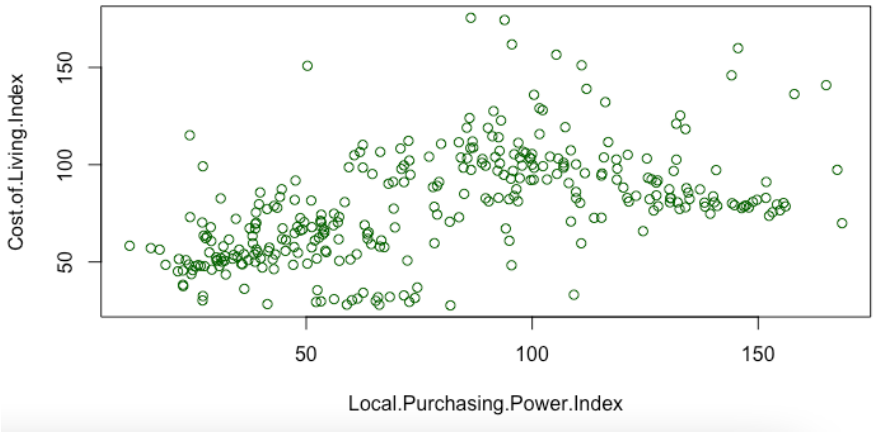
R code:
```
plot(cost_of_living$Rent.Index , cost_of_living$Cost.of.Living.Index, col="purple",
xlab="Rent.Index", ylab="Cost.of.Living.Index")
```

```
plot(cost_of_living$Groceries.Index , cost_of_living$Cost.of.Living.Index, col="magenta", xlab="Groceries.Index", ylab="Cost.of.Living.Index")

plot(cost_of_living$Restaurant.Price.Index , cost_of_living$Cost.of.Living.Index, col="brown", xlab="Restaurant.Price.Index", ylab="Cost.of.Living.Index")

plot(cost_of_living$Local.Purchasing.Power.Index , cost_of_living$Cost.of.Living.Index, col="dark green", xlab="Local.Purchasing.Power.Index", ylab="Cost.of.Living.Index")
```

R output:







Question 2. Compute the correlation coefficients for all the scatterplots obtained above.

Answer:
The correlation coefficient for Rent and Cost of Living is 0.7722926
The correlation coefficient for Groceries and Cost of Living is 0.9538616
The correlation coefficient for Restaurant Price and Cost of Living is 0.9493554
The correlation coefficient for Local Purchasing Power and Cost of Living is 0.525902

R code:

```
cor(cost_of_living$Rent.Index , cost_of_living$Cost.of.Living.Index)
cor(cost_of_living$Groceries.Index , cost_of_living$Cost.of.Living.Index)
cor(cost_of_living$Restaurant.Price.Index , cost_of_living$Cost.of.Living.Index)
cor(cost_of_living$Local.Purchasing.Power.Index , cost_of_living$Cost.of.Living.Index)
```

R output:

```
> cor(cost_of_living$Rent.Index , cost_of_living$Cost.of.Living.Index)
[1] 0.7722926
> cor(cost_of_living$Groceries.Index , cost_of_living$Cost.of.Living.Index)
[1] 0.9538616
> cor(cost_of_living$Restaurant.Price.Index , cost_of_living$Cost.of.Living.Index)
[1] 0.9493554
> cor(cost_of_living$Local.Purchasing.Power.Index , cost_of_living$Cost.of.Living.Index)
[1] 0.525902
```

Question 3. Verify the conditions for EACH correlation coefficient computed above

Answer:
For the Rent Index vs. Cost of Living, the quantitative variable condition applies because the Rent Index and Cost of Living have quantitative variables. For the linearity condition, the scatterplot shows a positive linear line. However, there is evidence of a curve towards the end of the scatterplot, so this graph isn't straight enough to be linear. For the outlier condition, there are several outliers that created a slight downward curve. Therefore, the outlier and linearity conditions are unfulfilled, and we can't use correlation in this relationship.

For the Groceries Index vs. Cost of Living, the quantitative variable condition applies here because the Groceries Index and Cost of Living contain quantitative variables. For the linearity condition, the graph shows a straight, positive linear line, so this condition is verified. For the outlier condition, although there are several outliers, it doesn't distort the correlation, but instead, continued in the direction of a positive linear line. Thus, in this relationship, the outlier condition is verified. Therefore, we can use correlation in this relationship.

For the Restaurant Price Index vs. Cost of Living, the quantitative variable condition applies here because the Price Index and Cost of Living contain quantitative variables. For the linearity condition, the graph shows a straight, positive linear line, so this condition is verified. For the outlier condition, although there are several outliers, it doesn't distort the correlation, but instead, continued in the direction of a positive linear line. Thus, in this relationship, the outlier condition is verified. Therefore, we can use correlation in this relationship.

For Local Purchasing Power vs. Cost of Living, the quantitative variable condition applies here because the Local Purchasing Power Index and Cost of Living contain quantitative variables. For the linearity condition, the graph shows a weak positive correlation that's curved downwards, which violates the linearity condition. For the outlier condition, the outliers are distorting the correlation by slightly curving downwards. Therefore, because the outlier and linearity conditions aren't verified, we can't use correlation in this case.

R code: N/A

R output: N/A

Question 4: Fit a linear regression model between the Cost of Living Index and each of the other index variables. As a result, there should be 4 regression models in total. Interpret the resulting estimated slope in each model.

Answer:
The estimated slope for Rent and Cost of Living is 1.025. This means that for every unit of increase in Rent, the Cost of Living is expected to increase by 1.025 units.

The estimated slope for Groceries and Cost of Living is 0.9529. This means that for every unit of increase in Groceries price, the Cost of Living is expected to increase by 0.9529 units.

The estimated slope for Restaurant Price and Cost of Living is 0.8033. This means that for every unit of increase in Restaurant price, the Cost of Living is expected to increase by 0.8033 units.

The estimated slope for Local Purchasing Power and Cost of Living is 0.3762. This means that for every unit of increase in Local Purchasing Power, the Cost of Living is expected to increase by 0.3762 units.

R code:

```
R <- lm(Cost.of.Living.Index ~ Rent.Index, data=cost_of_living)
R
G <- lm(Cost.of.Living.Index ~ Groceries.Index, data=cost_of_living)
G
R1 <- lm(Cost.of.Living.Index ~ Restaurant.Price.Index, data=cost_of_living)
R1
L <- lm(Cost.of.Living.Index ~ Local.Purchasing.Power.Index , data=cost_of_living)
L
```

R output:

```
Coefficients:
(Intercept)   Rent.Index
     45.233        1.025
```

```
Coefficients:
    (Intercept)  Groceries.Index
         9.2178           0.9529
```

```
Coefficients:
           (Intercept)  Restaurant.Price.Index
               24.6636                  0.8033
```

```
Coefficients:
                (Intercept)  Local.Purchasing.Power.Index
                    48.9974                        0.3762
```

Question 5. Based on the correlation coefficients and the regression models obtained above, which item would be the best predictor of the overall cost in these cities? Which would be the worst? Explain.

Answer:
The Groceries Index is the best predictor of the overall cost of these cities, because it has the strongest correlation coefficient of about 0.95, and has a relatively high slope value in the regression model compared to other graphs' regression models and correlation coefficients. It's safe to say that there is a strong positive relationship between Groceries Prices and Cost of Living. Although the Rent Index has a higher slope value than the Groceries Index, the correlation coefficient of the Rent Index is smaller than the Groceries Index, which is why it's not the best predictor of the overall cost.

The worse predictor of the overall cost is Local Purchasing Power because it has a weak positive correlation of about 0.5259 and a low slope value of 0.3762, and therefore, doesn't have a strong relationship with the Cost of Living variable.

R code: N/A

R output: N/A

Question 6. Find the cost of living as predicted by the Groceries Index and its residual for Beijing, China. (Hint: Find the row index of Beijing in the dataset, and then use that index to extract the corresponding fitted value and residual from the regression result.)

Answer:
The row index for Beijing, China is in the 172nd position. The corresponding fitted value is 88.85556, and the residual is -11.66556.

R code:
```
which(cost_of_living$City == 'Beijing, China')
G$fitted.values[172]
G$residuals[172]
```

R output:

```
> which(cost_of_living$City == 'Beijing, China')
[1] 172
> #positioned in 172th place
> G$fitted.values[172]
     172
88.85556
> G$residuals[172]
      172
-11.66556
```

Appendix:

```
read.table('Cost_of_Living_2013.txt', sep = '\t', header = TRUE)
cost_of_living <- read.table('Cost_of_Living_2013.txt', sep = '\t', header = TRUE)
names(cost_of_living)

read.table('Cost_of_Living_2013.txt', sep = '\t', header = TRUE)
cost_of_living <- read.table('Cost_of_Living_2013.txt', sep = '\t', header = TRUE)
names(cost_of_living)

plot(cost_of_living$Rent.Index , cost_of_living$Cost.of.Living.Index, col="purple",
xlab="Rent.Index", ylab="Cost.of.Living.Index")

plot(cost_of_living$Groceries.Index , cost_of_living$Cost.of.Living.Index, col="magenta",
xlab="Groceries.Index", ylab="Cost.of.Living.Index")

plot(cost_of_living$Restaurant.Price.Index , cost_of_living$Cost.of.Living.Index, col="brown",
xlab="Restaurant.Price.Index", ylab="Cost.of.Living.Index")

plot(cost_of_living$Local.Purchasing.Power.Index , cost_of_living$Cost.of.Living.Index,
col="dark green", xlab="Local.Purchasing.Power.Index", ylab="Cost.of.Living.Index")

cor(cost_of_living$Rent.Index , cost_of_living$Cost.of.Living.Index)

cor(cost_of_living$Groceries.Index , cost_of_living$Cost.of.Living.Index)

cor(cost_of_living$Restaurant.Price.Index , cost_of_living$Cost.of.Living.Index)

cor(cost_of_living$Local.Purchasing.Power.Index , cost_of_living$Cost.of.Living.Index)

R <- lm(Cost.of.Living.Index ~ Rent.Index, data=cost_of_living)
R
G <- lm(Cost.of.Living.Index ~ Groceries.Index, data=cost_of_living)
G
```

```
R1 <- lm(Cost.of.Living.Index ~ Restaurant.Price.Index, data=cost_of_living)
R1
L <- lm(Cost.of.Living.Index ~ Local.Purchasing.Power.Index , data=cost_of_living)
L

which(cost_of_living$City == 'Beijing, China')
G$fitted.values[172]
G$residuals[172]
```