Assignment Cost_of_living

Oleg Sechovcov

2024-10-27

Assignment Cost_of_living

```
library(e1071)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
setwd("C:/Users/olegs/Documents/SDU - Copy/SDU/R/Data")
data <- read.csv("Cost_of_living.txt")
```

Introduction

First step, we need to load the data and see what we have.I will be presenting the first 6 rows of the data set.

```
head(data, 6)
```

```
##        City     Country Cost.of.Living.Index Rent.Index
## 1    Zurich Switzerland               128.29      61.66
## 2     Basel Switzerland               125.54      45.76
## 3  Lausanne Switzerland               124.02      50.64
## 4    Geneva Switzerland               118.98      68.47
## 5      Bern Switzerland               116.03      40.52
## 6 Stavanger      Norway               102.27      36.10
##   Cost.of.Living.Plus.Rent.Index Groceries.Index Restaurant.Price.Index
## 1                          96.42          127.96                 124.73
## 2                          87.38          124.99                 123.11
## 3                          88.92          127.26                 123.61
```

```
## 4                          94.82          112.88                 119.58
## 5                          79.91          107.58                 115.56
## 6                          70.62           90.99                 112.45
##   Local.Purchasing.Power.Index
## 1                       126.90
## 2                       121.47
## 3                       110.52
## 4                       111.16
## 5                       131.89
## 6                        87.58
```

The data set contains 8 columns and 440 rows. The Columns are: City, Country, Cost.of.Living.Index, Rent.Index, Cost.Of.Living.Plus.Rent.Index, Groceries.Index, Restaurant.Price.Index, Local.Purchasing.Power.Index.

The data shows the cost of living in different cities around the world. This data set is useful for people who are planning to move to another city or country and want to know the cost of living in that place. For example, if an SDU student wants to take a semester in another county, they can use this data set to compare the cost of living in different cities and choose the one that fits their budget.

Data Analysis

Now, let's do some data analysis. I will start by checking the summary of the data set.

```
summary(data)
```

```
##      City              Country          Cost.of.Living.Index   Rent.Index
##  Length:440         Length:440         Min.   : 19.77       Min.   :  3.46
##  Class :character   Class :character   1st Qu.: 37.09       1st Qu.: 10.62
##  Mode  :character   Mode  :character   Median : 52.45       Median : 20.17
##                                        Mean   : 54.82       Mean   : 23.75
##                                        3rd Qu.: 70.67       3rd Qu.: 32.50
##                                        Max.   :128.29       Max.   :115.58
##  Cost.of.Living.Plus.Rent.Index Groceries.Index  Restaurant.Price.Index
##  Min.   : 12.38                 Min.   : 19.66   Min.   : 10.66
##  1st Qu.: 24.59                 1st Qu.: 30.82   1st Qu.: 30.21
##  Median : 37.75                 Median : 44.77   Median : 47.55
##  Mean   : 39.96                 Mean   : 47.50   Mean   : 51.21
##  3rd Qu.: 52.09                 3rd Qu.: 61.03   3rd Qu.: 70.83
##  Max.   :103.02                 Max.   :127.96   Max.   :124.73
##  Local.Purchasing.Power.Index
##  Min.   :  2.36
##  1st Qu.: 42.27
##  Median : 66.64
##  Mean   : 70.80
##  3rd Qu.: 95.52
##  Max.   :163.27
```

This summary shows the minimum, maximum, median, and mean values of each column in the data set. For example, the minimum value of the Cost.of.Living.Index column is 33.19, the maximum value is 141.25, the median value is 74.58, and the mean value is 74.57.

One sample t-test for Cost of Living Plus Rent for the whole world

Check CLT conditions

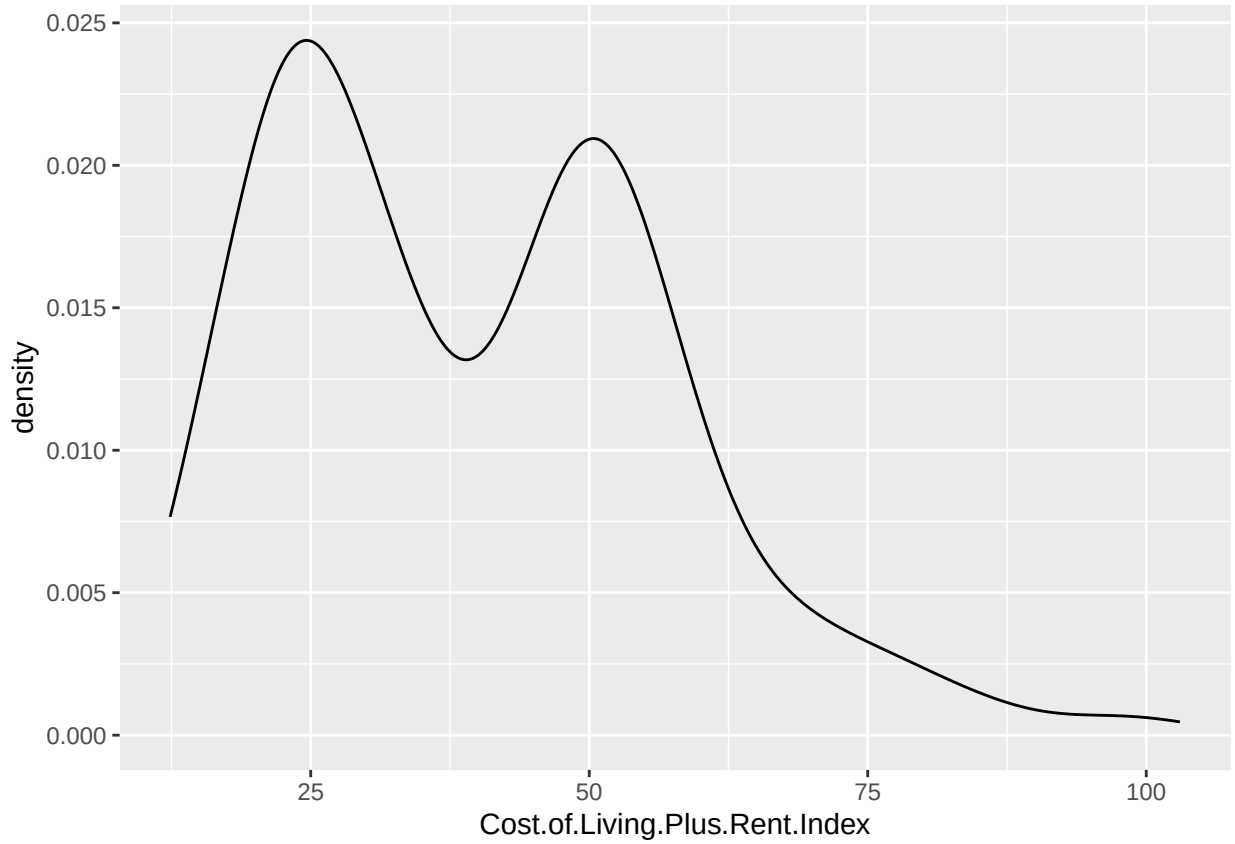
- Data is independent - YES
- Distribution is not strongly skewed - YES

```
skewness_value <- skewness(data$Cost.of.Living.Plus.Rent.Index)
skewness_value
```

```
## [1] 0.6359417
```

The skewness value is not above 1 or below -1, so the distribution is not strongly skewed. Below is a density of the Cost of Living Plus Rent Index. But as we can see, the graph shows us that it is strongly skewed. The assumption is that the data set is 8 column with 440 rows, which makes the skewness less important.

```
ggplot(data, aes(x=Cost.of.Living.Plus.Rent.Index)) +
  geom_density()
```



Set-up hypothesis Now, we want to check if the mean Cost of Living Plus Rent Index in the world is 40 Before checking it for each country, we will check it for the whole world.

$$H_0 : \mu = 40$$

$$H_1 : \mu \neq 40$$

Assume threshold values (alpha significance level) We will assume the alpha significance level to be 0.05.

$$\alpha = 0.05$$

```
H0 <- 40
```

Calculate

- Point estimate

To calculate the point estimate, we will find the mean of the Cost of Living Plus Rent Index column.The formula is:

$$\bar{x} = \frac{\sum x}{n}$$

```
pe_Cost_Of_Living_Plus_Rent <- mean(data$Cost.of.Living.Plus.Rent.Index)
pe_Cost_Of_Living_Plus_Rent
```

```
## [1] 39.95932
```

- Standard error To calculate the standard error, we will use the formula: $SE = \frac{s}{\sqrt{n}}$

```
sd_Cost_Of_Living_Plus_Rent <- sd(data$Cost.of.Living.Plus.Rent.Index)
SE_Cost_Of_Living_Plus_Rent <- sd_Cost_Of_Living_Plus_Rent/sqrt(nrow(data))
SE_Cost_Of_Living_Plus_Rent
```

```
## [1] 0.8460371
```

This means that that if we kept repeatedly sampling from the population, the mean of those samples would typically differ from the sample mean by about 0.846.

- Degree of freedom To calculate the degree of freedom, we will use the formula: $d = n - 1$

```
df_Cost_Of_Living_Plus_Rent <- nrow(data)-1
df_Cost_Of_Living_Plus_Rent
```

```
## [1] 439
```

Having 439 degrees of freedom means that we have 439 independent pieces of information to estimate the population parameter.

- t_statistic To calculate the t-statistic, we will use the formula: $t = \frac{\bar{x}-\mu}{SE}$

```
t_statistic_Cost_Of_Living_Plus_Rent <-  (pe_Cost_Of_Living_Plus_Rent - H0)/SE_Cost_Of_Living_Plus_Rent
t_statistic_Cost_Of_Living_Plus_Rent
```

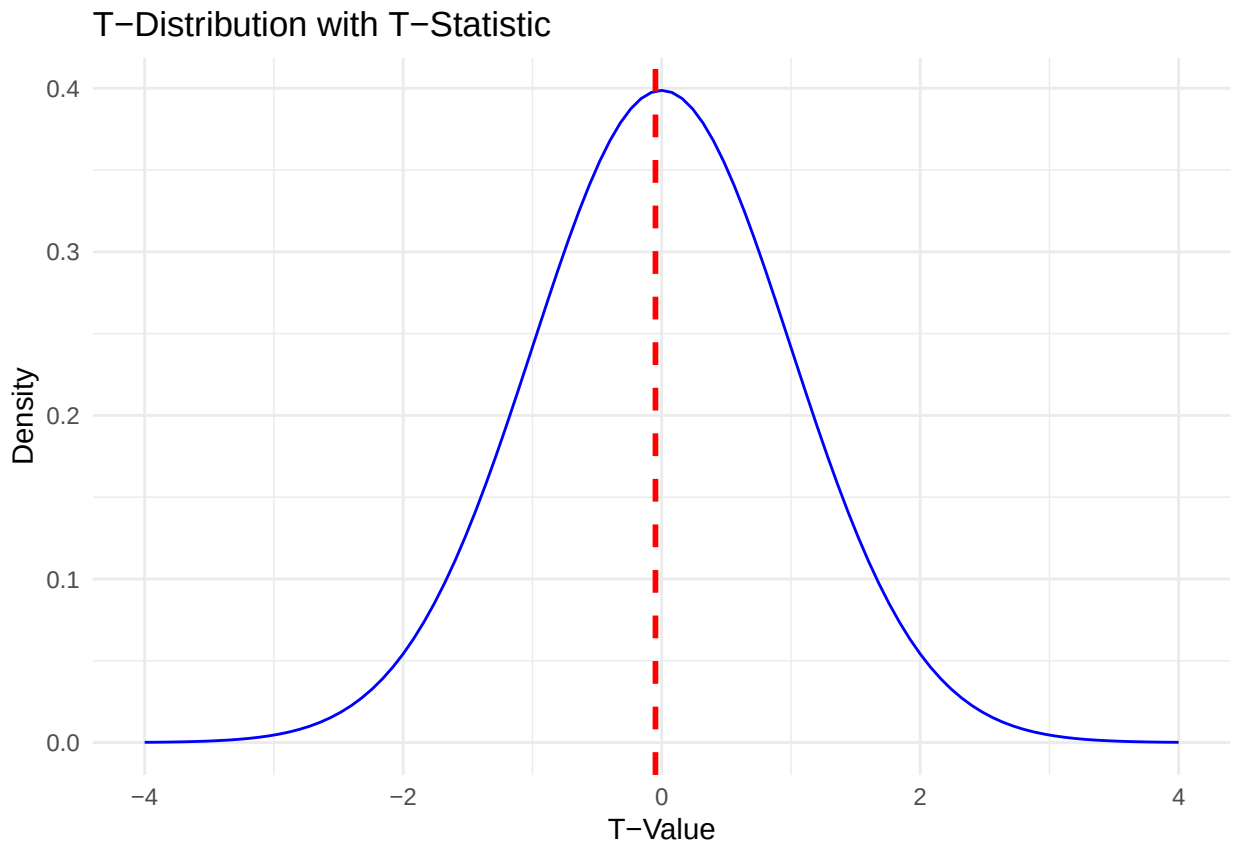
```
## [1] -0.04808515
```

The t-statistic is -0.04808515 which means that the sample mean is 0.04808515 standard errors below the hypothesized population mean.Vizualization of the t-distribution is shown below.

```r
x_values <- seq(-4, 4, length = 100)

ggplot(data.frame(x = x_values), aes(x = x)) +
  stat_function(fun = dt, args = list(df = df_Cost_Of_Living_Plus_Rent), color = "blue") +
  geom_vline(xintercept = t_statistic_Cost_Of_Living_Plus_Rent, color = "red", linetype = "dashed", size
  labs(
    title = "T-Distribution with T-Statistic",
    x = "T-Value",
    y = "Density"
  ) +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



* p-value

```r
p_value_Cost_Of_Living_Plus_Rent <- 2*pt(-abs(t_statistic_Cost_Of_Living_Plus_Rent), df_Cost_Of_Living_
p_value_Cost_Of_Living_Plus_Rent
```

```
## [1] 0.9616703
```

The p-value is 0.96 which is greater than the alpha significance level of 0.05. This means that we fail to reject the null hypothesis. There is not enough evidence to suggest that the mean Cost of Living Plus Rent Index in the world is different from 40.

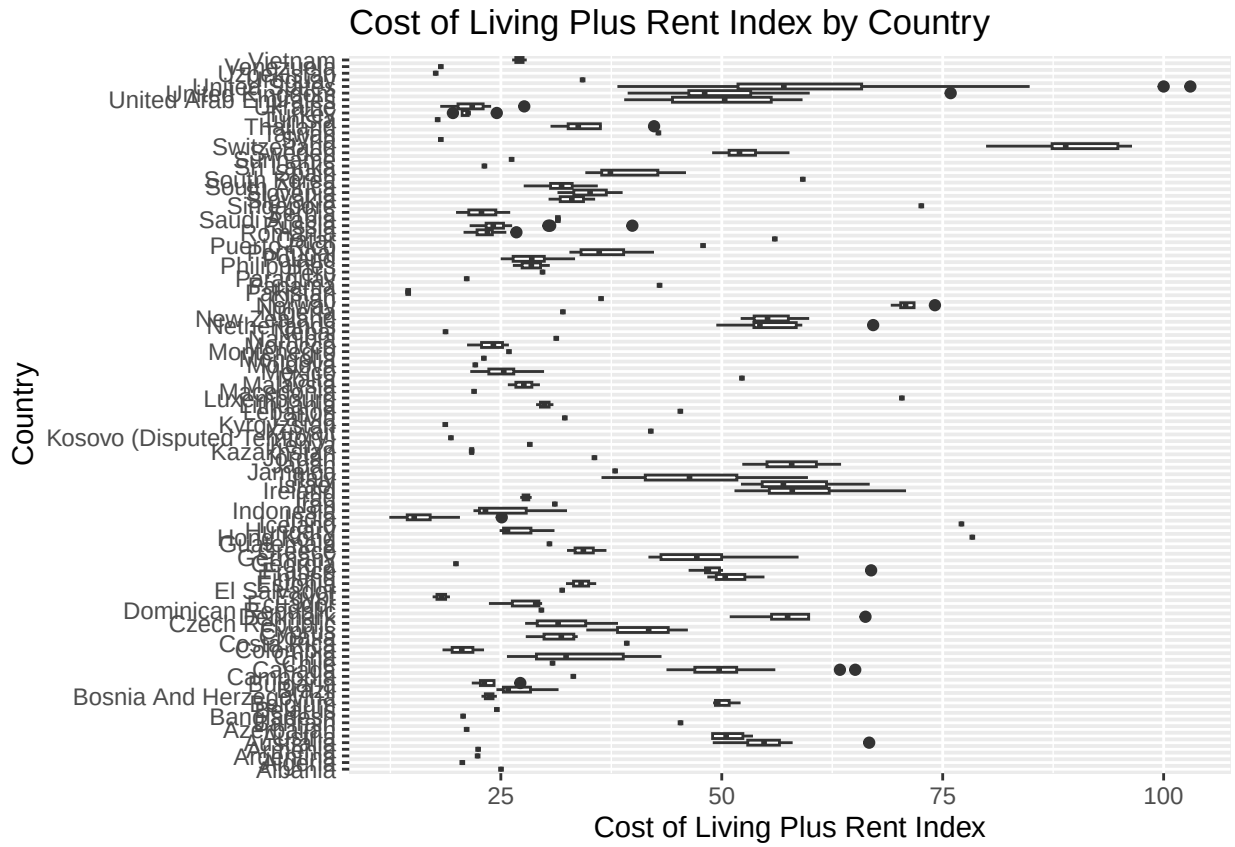
Conclusion The findings imply that as students looking to move to different cities may find the average cost of living plus rent to be manageable if it aligns with this mean.

While the overall average is stable, it's essential to conduct further analyses on specific regions or cities to understand local cost dynamics, which could differ significantly from the average.

Analysis of Variance (ANOVA)

Visualizing the data Now, let's perform an analysis of variance (ANOVA) to compare the Cost of Living Plus Rent Index among different countries. We will use the ANOVA test to determine if there are any significant differences in the Cost of Living Plus Rent Index between the countries. Visualizing the data will help us understand the distribution of the Cost of Living Plus Rent Index among different countries.

```
ggplot(data, aes(x = Country, y = Cost.of.Living.Plus.Rent.Index)) +
  geom_boxplot() +
  coord_flip() +  # Flip coordinates for better visibility
  labs(title = "Cost of Living Plus Rent Index by Country",
       x = "Country",
       y = "Cost of Living Plus Rent Index")
```



Set-up hypothesis

We will set up the null and alternative hypotheses for the ANOVA test. The null hypothesis states that the means of the Cost of Living Plus Rent Index are equal across all countries, while the alternative hypothesis states that at least one mean is different.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10}$$

$$H_1 : \text{At least one mean is different}$$

```
H0 <- 0
```

```
alpha <- 0.05
```

Set-up threshold values (alpha significance level)

ANOVA Test Now, let's perform the ANOVA test to compare the Cost of Living Plus Rent Index among different countries.

```
anova_results <- aov(Cost.of.Living.Plus.Rent.Index ~ Country, data = data)
summary(anova_results)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Country     107 122200  1142.1   23.61 <2e-16 ***
## Residuals   332  16060    48.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test results show that the p-value is less than 2.2e-16, which is less than the alpha significance level of 0.05. This means that we reject the null hypothesis and conclude that there is a significant difference in the Cost of Living Plus Rent Index among different countries.

Post-Hoc Analysis To further investigate which countries have significantly different Cost of Living Plus Rent Index values, we can perform a post-hoc analysis using the Tukey HSD test. Below are the results of the top 6 Tukey HSD test for the Cost of Living Plus Rent Index among different countries.

```
tukey_results <- TukeyHSD(anova_results)
head(tukey_results$Country)
```

```
##                        diff        lwr      upr     p adj
## Algeria-Albania    -4.3800 -47.517077 38.75708 1.0000000
## Argentina-Albania  -2.6500 -45.787077 40.48708 1.0000000
## Armenia-Albania    -2.5900 -45.727077 40.54708 1.0000000
## Australia-Albania  30.2440  -1.747312 62.23531 0.1127634
## Austria-Albania    25.8375  -8.265353 59.94035 0.7121222
## Azerbaijan-Albania -3.8900 -47.027077 39.24708 1.0000000
```

The Tukey HSD test results show that there are significant differences in the Cost of Living Plus Rent Index among different countries.

Conclusion The is a significant difference in the Cost of Living Plus Rent Index among different countries. This information can be useful for students or individuals planning to move to a new country and want to compare the cost of living in different countries. It is essential to consider the Cost of Living Plus Rent Index when making decisions about moving to a new country.

6 cheapest countries to live in

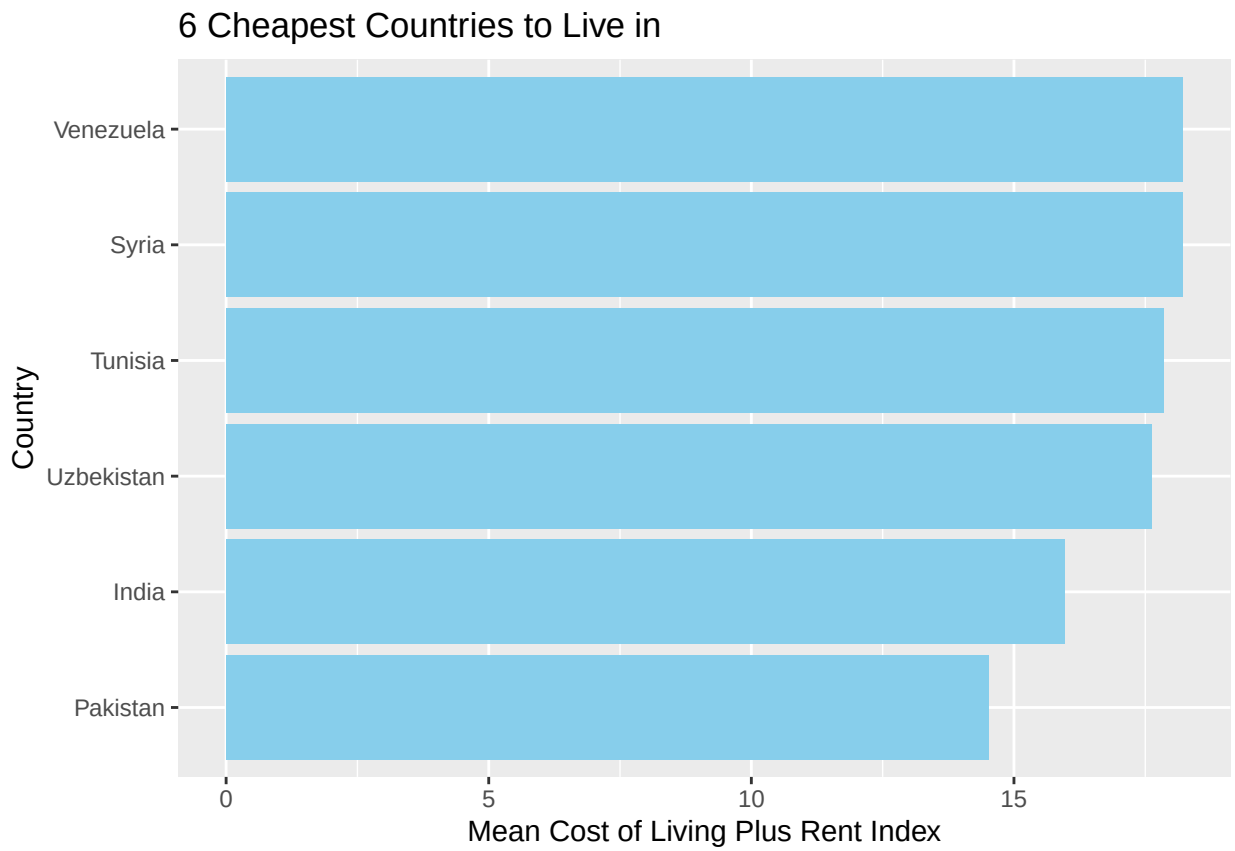
To find the 6 cheapest countries to live in, we will calculate the mean Cost of Living Plus Rent Index for each country and sort the countries in ascending order based on the mean Cost of Living Plus Rent Index. The top 6 countries with the lowest mean Cost of Living Plus Rent Index values will be considered the cheapest countries to live in.

```
cheapest_countries <- data %>%
  group_by(Country) %>%
  summarize(mean_Cost_of_Living_Plus_Rent = mean(Cost.of.Living.Plus.Rent.Index, na.rm = TRUE)) %>%
  arrange(mean_Cost_of_Living_Plus_Rent) %>%
  head(6)
cheapest_countries
```

```
## # A tibble: 6 x 2
##   Country    mean_Cost_of_Living_Plus_Rent
##   <chr>                              <dbl>
## 1 Pakistan                            14.5
## 2 India                               16.0
## 3 Uzbekistan                          17.6
## 4 Tunisia                             17.8
## 5 Syria                               18.2
## 6 Venezuela                           18.2
```

Visualizing the 6 cheapest countries to live in based on the mean Cost of Living Plus Rent Index.

```
ggplot(cheapest_countries, aes(x = reorder(Country, mean_Cost_of_Living_Plus_Rent), y = mean_Cost_of_Li
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "6 Cheapest Countries to Live in",
       x = "Country",
       y = "Mean Cost of Living Plus Rent Index")
```



Conclusion

The 6 cheapest countries to live in based on the mean Cost of Living Plus Rent Index are: Venezuela, Syria, Tunisia, Uzbekistan, India,and Pakistan These countries have the lowest cost of living plus rent index values, making them affordable options for students looking to live in a cost-effective environment.