Investigative Assignment

Name: Satya Sai Koppalu

Plan

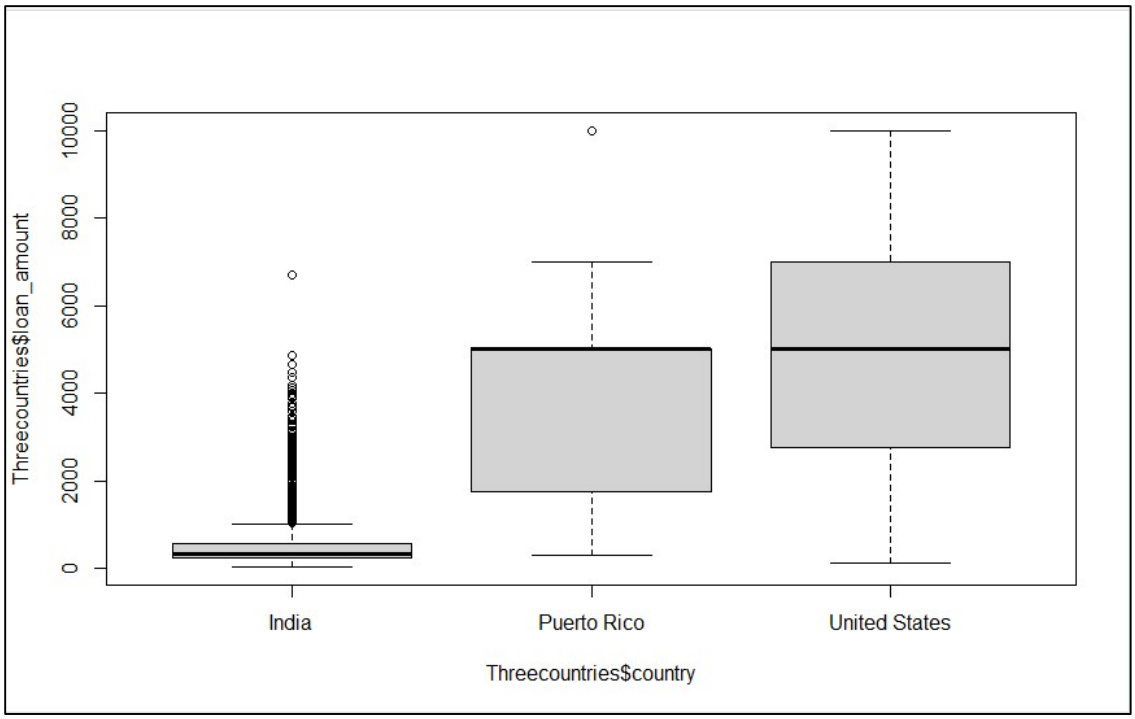
- Question: Compare the Kiva loan amounts of three countries- India, United States and Puerto Rico.
- Response Variable- Y= loan amounts.
- Factor = Country.
- Create a Boxplot displaying the loan amounts by the chosen countries.

Creating the Boxplot

Code:

```
## Creating Boxplots
boxplot(Threecountries$loan_amount ~ Threecountries$country)
```

Output:



After interpreting the boxplot,

- We can see that the United States has the highest range with Puerto Rico having the second highest and India the least among the three countries.
- Puerto Rico and United States share the same median.
- India and Puerto Rico are skewed while United States is symmetric.

Solving for Equal Variance

Code:

```
## Finding the length, mean and standard deviation
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= length)
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= mean)
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= sd)
```

Output:

```
> aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= sd)
        Group.1         x
1         India  738.7518
2   Puerto Rico 2978.4071
3 United States 2970.8238
```

Assumptions:

- Normal Distribution Assumption- India and Puerto Rico have outliers but since the size of samples are large, we can go ahead.
- Random Samples- don't know
- Equal variance does not exist since the value we get by dividing the largest standard deviation by the smallest standard deviation is greater than 2.

F- test

Parameters- Allows to compare three or more population means

- Mu1 = population mean loan amount for country1 (India)
- Mu2 = population mean loan amount for country 2 (United States)
- Mu3 = population mean loan amount for country 3 (Puerto Rico)

Hypotheses-

- Null Hypothesis- Ho= Mu1 = Mu2 = Mu3
- Alternative Hypothesis= Ha= At least one Mu in the three countries is not equal to the rest

One way ANOVA Results-

- P Value is found to be almost 0

Code:

```
## Performing Anova
Kano <-aov(Threecountries$loan_amount ~ Threecountries$country)
anova(Kano)
```

Output:

```
> ## Performing Anova
> Kano <-aov(Threecountries$loan_amount ~ Threecountries$country)
> anova(Kano)
Analysis of Variance Table

Response: Threecountries$loan_amount
                          Df     Sum Sq    Mean Sq F value    Pr(>F)
Threecountries$country     2 1.3882e+10 6941145501  1965.7 < 2.2e-16 ***
Residuals               2990 1.0558e+10    3531150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion-

- With the p-value almost 0, we have very strong evidence that the population mean loan amount for at least one country is different from the rest. Thus, null hypothesis can be rejected.

Tukey HSD Analysis

Code:

```
## Performing TukeyHSD Analysis
TukeyHSD(Kano, conf.level = 0.95)
```

Output:

```
> ## Performing TukeyHSD Analysis
> TukeyHSD(Kano, conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Threecountries$loan_amount ~ Threecountries$country)

$`Threecountries$country`
                              diff       lwr      upr     p adj
Puerto Rico-India         3612.6063 2280.2337 4944.979 0.0000000
United States-India       4492.1029 4323.8415 4660.364 0.0000000
United States-Puerto Rico  879.4966 -455.8798 2214.873 0.2703734
```

- Puerto Rico-India: We are 95% confident that the population mean loan amount for Puerto Rico is 2280.2337 to 4944.979 more than India.
- United States-India: We are 95% confident that the population mean loan amount for United States is 4323.8415 to 4660.364 more than India.
- United States-Puerto Rico: We are 95% Confident that the population mean loan amount for United States is 455.8798 less to 2214.873 more than Puerto Rico.

Limitations

- Equal variance does not exist since the value we get by dividing the largest standard deviation by the smallest standard deviation is greater than 4.
- We do not know if the sample is random.

Kruskal Wallace Test

Hypotheses:

- Null Hypothesis Ho: The population median loan amounts for the three selected countries are the same.
- Alternative Hypothesis Ha: The population median loan amount for at least one of the three selected countries is different from the rest.

Code:

```
## Performing the kruskal Wallace test for Multiple Comparisons
kruskal.test(Threecountries$loan_amount ~ Threecountries$country)
pgirmess::kruskalmc(Threecountries$loan_amount ~ Threecountries$country)
```

Output:

```
> ## Performing the kruskal Wallace test for Multiple Comparisons
> kruskal.test(Threecountries$loan_amount ~ Threecountries$country)

        Kruskal-Wallis rank sum test

data:  Threecountries$loan_amount by Threecountries$country
Kruskal-Wallis chi-squared = 1813.1, df = 2,
p-value < 2.2e-16

> pgirmess::kruskalmc(Threecountries$loan_amount ~ Threecountries$country)
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                          obs.dif critical.dif
India-Puerto Rico       1186.6370    625.54394
India-United States     1400.8059     78.99809
Puerto Rico-United States  214.1689    626.95417
                        difference
India-Puerto Rico             TRUE
India-United States           TRUE
Puerto Rico-United States    FALSE
```

- From the Kruskal Wallace test we observe that the p value is almost 0.
- The chi-squared test statistic is large with an observed value of 1813.1, signifying the population medians loan amounts of at least two countries differ considerably.
- After interpreting the results from the Kruskal Wallace test we observe that the population median loan amounts for India-Puerto Rico and India-United States are different. Whereas Puerto Rico-United States share a similar median.

Assumptions-

- Independent groups- countries
- Independent observations- random samples

Conclusion-

- With a large test statistic value and the p value being almost 0. We have very strong evidence that the population median loan amount is different for at least one of the three selected countries. Therefore, we can reject the null hypothesis.

Determination of the best method to use-

Kruskal Wallace test gives us much more accurate results than compared to the one-way ANOVA test for the following reasons:

- The absence of equal variances in the one-way ANOVA test has the potential to impact the results.
- The one-way ANOVA test uses population mean and has the potential to be affected by outliers. Whereas the Kruskal Wallace test uses the population median and has very little chance of getting affected by outliers. Thus, giving us much more accurate results.

Code:

```
## Importing data from Excel
library(readxl)
Kcountries <- read_excel("C:/Users/satya/Downloads/Kiva_Sample_2021_updated.xlsx")


## Installing tidyverse
install.packages("tidyverse")
library("tidyverse")

## Filtering for the three selected countries from the data
Threecountries<-filter(Kcountries, country =="India" |  country == "United States" | country == "Puerto Rico")

## Creating Boxplots
boxplot(Threecountries$loan_amount ~ Threecountries$country)

## Performing Anova
Kano <-aov(Threecountries$loan_amount ~ Threecountries$country)
anova(Kano)

## Finding the length, mean and standard deviation
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= length)
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= mean)
aggregate(Threecountries$loan_amount, by=list(Threecountries$country), FUN= sd)

## Performing TukeyHSD Analysis
TukeyHSD(Kano, conf.level = 0.95)

## Installing pgrimess
install.packages("pgirmess")
library("pgirmess")

## Performing the kruskal Wallace test for Multiple Comparisons
kruskal.test(Threecountries$loan_amount ~ Threecountries$country)
pgirmess::kruskalmc(Threecountries$loan_amount ~ Threecountries$country)
```