Stats with R Assignment 1

Amanda

11/4/2020

```
#input data
Kiva_Example <- read.csv("Kiva_Sample.csv", header = TRUE)

## Install tidyverse
#install.packages("rmarkdown")
#install.packages("tidyverse")
#install.packages("knitr")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.3.0
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("rmarkdown")
library("knitr")
```

Filter the data set to just the three countries that you selected.

The countries I selected were the United States, Puerto Rico and the Vigin Islands.

My initial country filters were not successful, so I ran the indivdual countries one at a time to see how many observarions were in the dataset for each country. Then upon discovering that there were 0 observations for the Virgin Islands I subsituted Israel for the Virgin Islands.

```
UnitedStates <-filter(Kiva_Example, country =="United States") #1069 observations
PuertoRico <- filter(Kiva_Example, country == "Puerto Rico") #10 observations
VirginIslands <- filter(Kiva_Example, country == "Virgin Island") #0 observations

#Used a subsitution country for the Virgin Islands
Israel <- filter(Kiva_Example, country == "Israel") #37 observations
```

Give the data set a name.

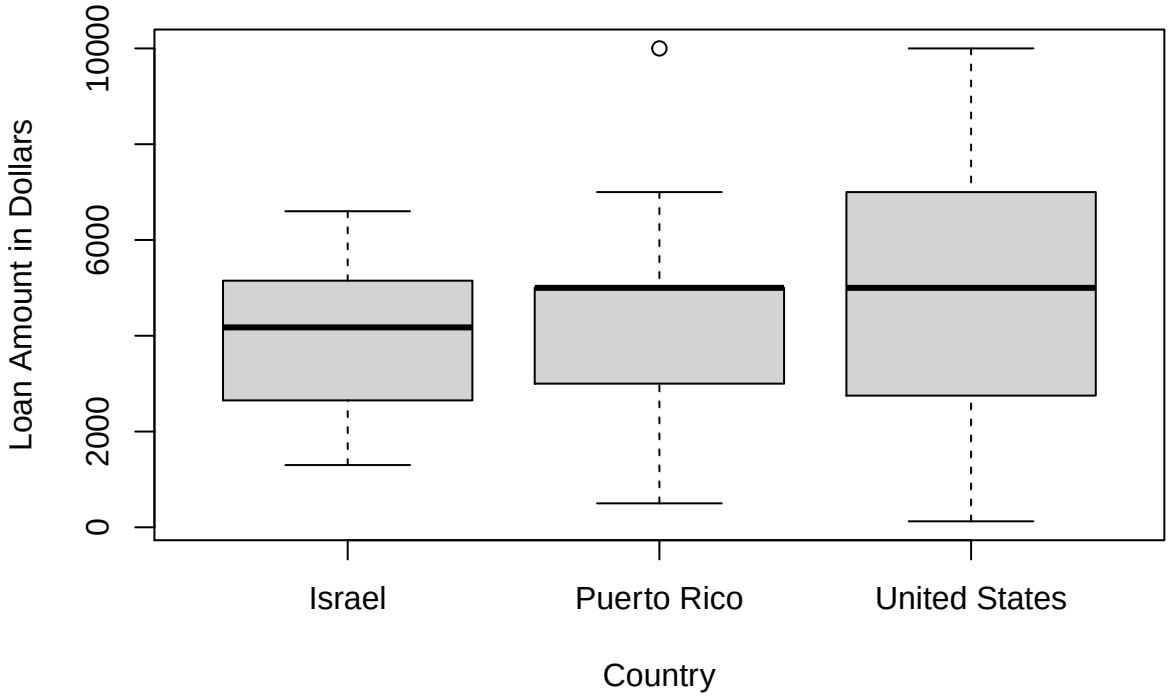
The professor called the data set “small”, but I choose to call the dataset KivaComparison to make it easier to identify.

```
KivaComparison <- filter(Kiva_Example, country == "United States" | country == "Puerto Rico" | country =
KivaComparison$loan_amount <- as.numeric(KivaComparison$loan_amount) #converts from character to numeri
```

Create a graphic to explore the business question

Looking at the boxplot below we can infer that: 1. Loans in Israel typically fall within in $2500 to $5000. 2. Loans in Puerto Rico typically fall within a similar range to Israel, however there is a single outlier for $10,000. 3. Loans in the U.S. typically see a wider range and higher amounts up to $10,000 are part of the upper range.

```
boxplot(KivaComparison$loan_amount ~ KivaComparison$country, xlab = "Country", ylab = "Loan Amount in D
```



Is there a difference in the loan amounts for a crowdsourced microlending organization (KIVA) between the three countries?

To answer this question with real numbers, I ran the summary statisitics on the combine country data using the KivaComparison dataset. Then I also ran the summary statistics on the indivdidual countries.

```r
#The summary statistics for all groups
summary(KivaComparison$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     125    2750    5000    5045    6581   10000
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 125    2750    5000    5045    6581   10000

#The summary statistics for the United States
UnitedStates$loan_amount <- as.numeric(UnitedStates$loan_amount)
summary(UnitedStates$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     125    2750    5000    5089    7000   10000
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 125    2750    5000    5089    7000   10000

#The summary statistics for the Puerto Rico
PuertoRico$loan_amount <- as.numeric(PuertoRico$loan_amount)
summary(PuertoRico$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500    3500    5000    4600    5000   10000
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 500    3500    5000    4600    5000   10000

#The summary statistics for the Israel
Israel$loan_amount <- as.numeric(Israel$loan_amount)
summary(Israel$loan_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1300    2650    4175    3907    5150    6600
```

```
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#1300    2650    4175    3907    5150    6600
```

Is there a difference in the loan amounts for a crowdsourced microlending organization (KIVA) between the three countries?

When looking at the summary statisitics for the United States loan amounts we can see there is a strong resemblance to the summary statistics for all three countries. This is because out of the 1116 observations, 1069 of them are for loans in the United States.

1. So the first observation is that there are more loans in the United States than either Puerto Rico or Israel.
2. The United States has the largest range of loan amounts from $175 to $10,000.
3. The United States has the smallest loan amount of $175. The minimum loan given for Puerto Rico was $500 and for Israel was $1300.
4. The United States and Israel tied for the highest loan amount of $10,000 and matched median loan amounts of $5,000.

Conduct a one way ANOVA test.

Make sure to include the parameters, hypothesis, assumptions, test statistic, p-value and conclusion.

#Hypothesis Null Hypothesis: That loan amounts in all three countries, the United States, Puerto Rico and Israel are equal. Alternative Hypothesis: The loan amounts in all three countries, The United States, Puerto Rico and Israel are not equal.

#Assumptions 1. Normal Distribution: We are not able to assume a normal distribution of data since there are only 10 Puerto Rico loans, which is under the 30 observation threshold. 2. Outliers: By looking at the boxplot we see there is a single outlier in the Puerto Rico dataset. 3. Randomization: Since the dataset is all loans, the indivdual amounts are randomized.
4. Independence: The loans in one country are independent from the loans in any other country.

‘‘‘r
#Anova

```
LoanAmount <-aov(KivaComparison$loan_amount ~ KivaComparison$country)
anova(LoanAmount)
```

```
## Analysis of Variance Table
##
## Response: KivaComparison$loan_amount
##                           Df     Sum Sq  Mean Sq F value  Pr(>F)
## KivaComparison$country    2   51891424 25945712  3.0153 0.04943 *
## Residuals              1113 9577066848  8604732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interputating Anova Results

Test statistic 3.0153 P-value .04943

conclusion: With a p-value under .05, we have evidence that at least one of the population mean loan amounts is different from the others.

Limitations

Due to the assumptions of a normal distribution not being properly met for Puerto Rico