MSDS660week1assignment

Ahmad Alqurashi

8/31/2021

Introduction

In week 1 assignment, we are using R to perform statistical methods and analyze data to interpret relationship between two or more variables. R language is a programming language for statistical computing and visualizing data. Moreover, R language is supported by several packages that help statisticians to use variety of methods and models that make analytics easy. In this assignment, we are using data collected from a survey -wave6- that conducted by Institute for Social Research. Furthermore, the dataset contains many variables such as happiness, income, tv watch time, importance of politics, etc. For this analysis, we are going to propose a hypothesis stating that the importance of family increases among people who consider religion is important. To make the analysis and test the hypothesis, we need two variables from the data: Importance of family V4 and Importance of religion V9.

Methods and Results

To make this analysis, I am going to use two packages: data.table and ggplot2. data.table library is a package that helps to convert dataset to data.table format to enable handling data. ggplot2 is a package that helps to visualize data to several plots type such as histogram, boxplot, etc. Below shows code and results.

- we import needed packages.

```
library(data.table)
library(ggplot2)
```

- load wave 6 dataset to R.

```
load("WV6_Data_R_v20201117.rdata")
```

- converting dataset to data.table format.

```
df <- `WV6_Data_R_v20201117`
```

Here, we perform Exploratory Data Analysis.

- Summary function for both variables shows us the min, max, mean, median, Null values. this helps us to get insight of the data

```
summary(df$V4)
```

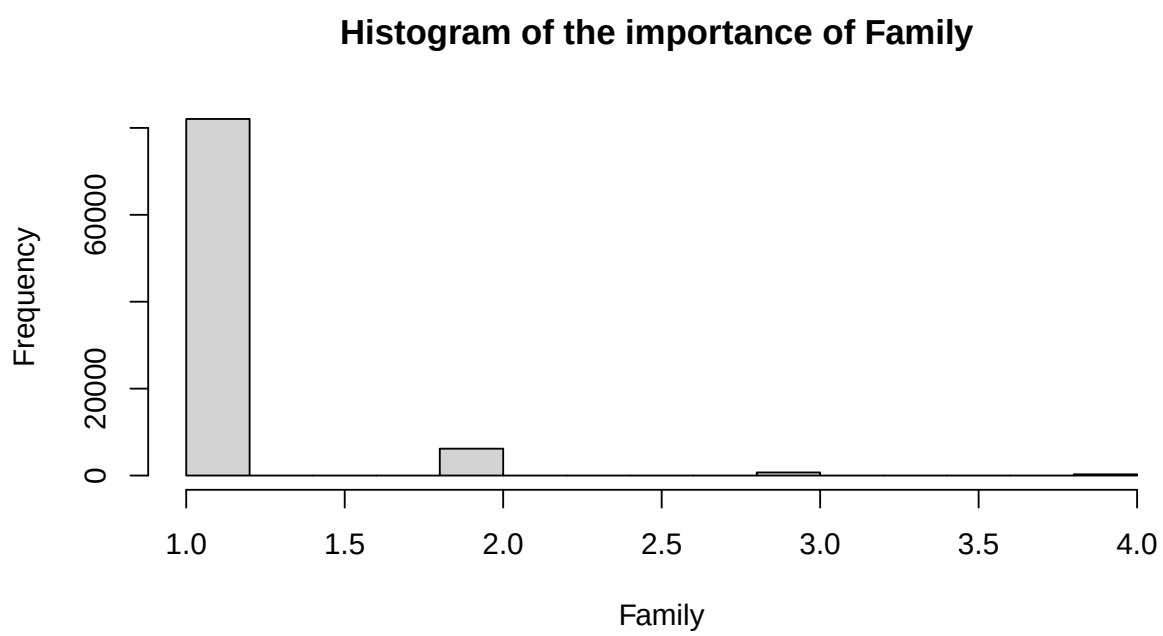
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   1.000   1.095   1.000   4.000     341
```

```
summary(df$V9)
```

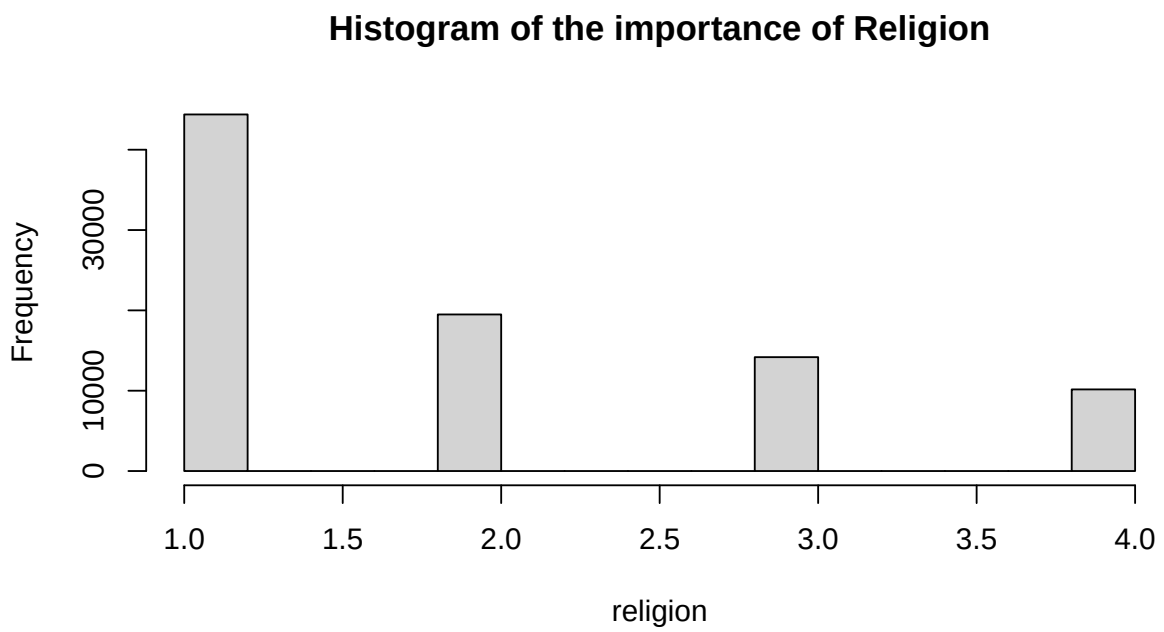
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   1.000   1.888   3.000   4.000    1348
```

- It appears that there is no negative value that would skew our data. Also, missing values are small portion of data. Therefore no cleaning needed
- below code plots two histogram of both variables.

```
hist(df$V4, main = "Histogram of the importance of Family ", xlab ="Family" )
```

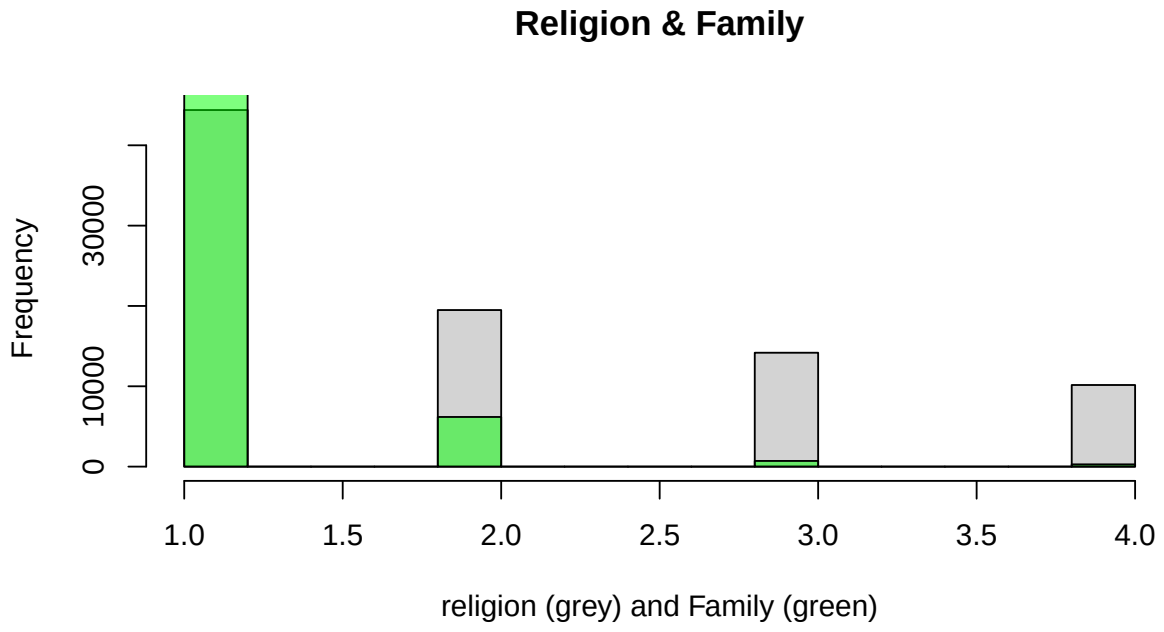



```
hist(df$V9, main = "Histogram of the importance of Religion", xlab ="religion")
```



- this shows a histogram where two variables are overlabing which help us to see result visually.

```
hist(df$V9, main = "Religion & Family", xlab = "religion (grey) and Family (green)")
hist(df$V4, add = TRUE, col = rgb(0, 1, 0, 0.5))
```



+ the results seems that most of people consider family is very important. on the other hand majority o

Conclusion

In conclusion, based on the above analysis, the importance of family does not relate with importance of religion. Thus, we reject null hypothesis. I suggest further analysis with more data to determine whether there is a relation or not.