Name Lengths

Abdulaziz M. Alqumayzi

(DS-510) – Discovering Statistics Using R

Colorado State University – Global Campus

Dr. Hasan Aljabbouli

November 30, 2020

Name Lengths

Introduction

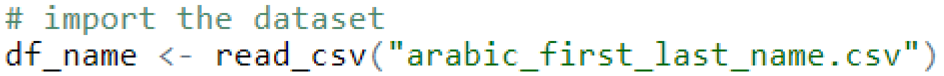
In this activity, we will examine data to research whether individuals with long last names still have long first names? Create and analysis a linear regression model focused on the relationship in the data set between the duration of the first and last names and offer insights into the phenomena within the constraints of the model. We count the letters of alphabets used to name 100 persons from this dataset.

Part I: Exploratory Analysis

Firstly, the " arabic_first_last_name.csv " file was downloaded and imported as in figure 1, as a dataframe named df_name that containing the first and last names of 100 individuals, which is organized into two columns. The first column has the first name of persons and the second column has the last name.

Figure 1

import the data to Rstudio Cloud environment

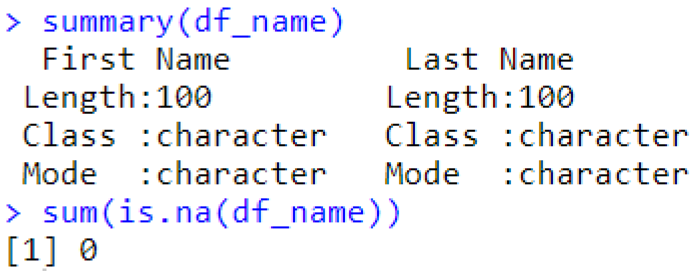


```
# import the dataset
df_name <- read_csv("arabic_first_last_name.csv")
```

Secondly, a quick check on the dataset in figure 2. a summary() function used to show the length of each column and the data type which is character. sum() and is.na() functions used together to return the number of missing values, the dataset does not have any missing values. A new dataframe created called df_name_new that has the lengths of the last and first names in figure 3. The new columns are last_length for the last name length and first_length for the first name length. nchar() function used to calculate the length of the two variables.

Figure 2

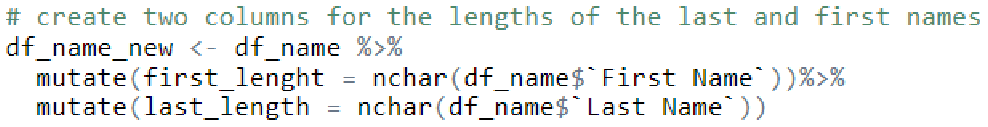
summary(), sum() and is.na() functions code and outputs



```
> summary(df_name)
  First Name         Last Name
 Length:100         Length:100
 Class :character   Class :character
 Mode  :character   Mode  :character
> sum(is.na(df_name))
[1] 0
```

Figure 3

Create new dataframe and two columns for last and first lengths code

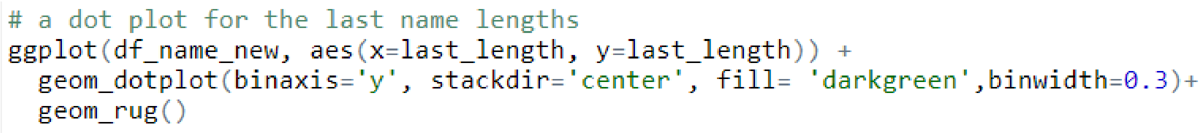


```
# create two columns for the lengths of the last and first names
df_name_new <- df_name %>%
  mutate(first_lenght = nchar(df_name$`First Name`))%>%
  mutate(last_length = nchar(df_name$`Last Name`))
```

A dot plot was created using ggplot2 library from R language for the last name length. The code is in figure 4 and the dot plot shown in figure 5. The most length of last names is 5.

Figure 4

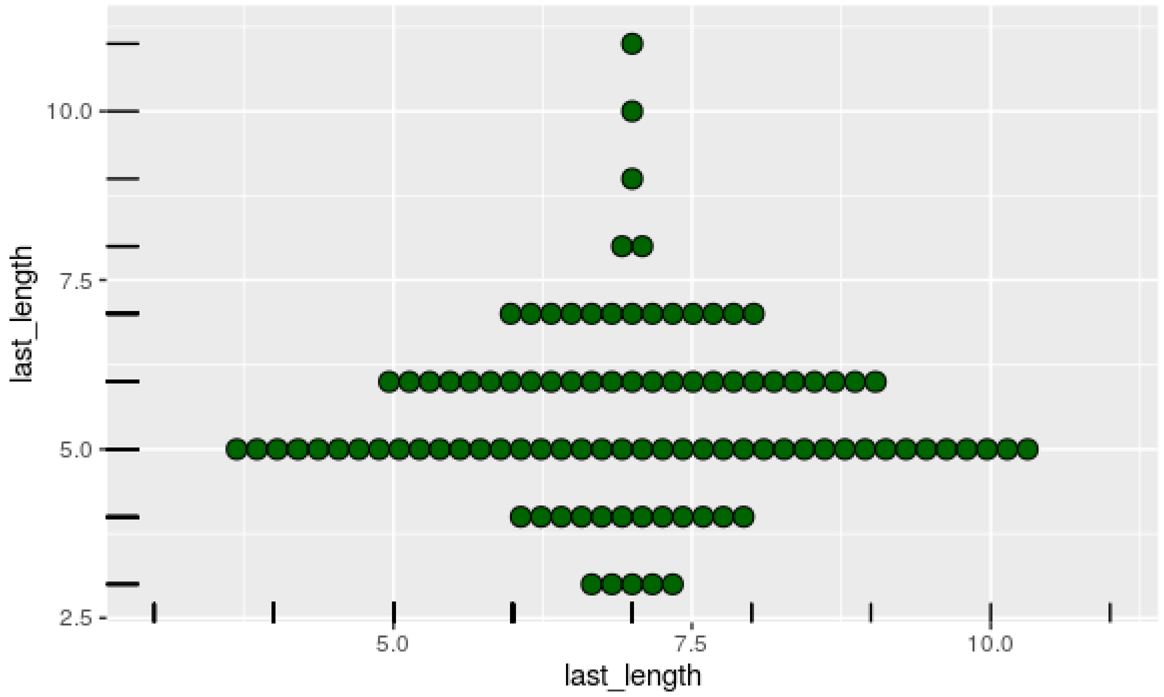
ggplot() function to create last name lengths code



```
# a dot plot for the last name lengths
ggplot(df_name_new, aes(x=last_length, y=last_length)) +
  geom_dotplot(binaxis='y', stackdir='center', fill= 'darkgreen',binwidth=0.3)+
  geom_rug()
```

Figure 5

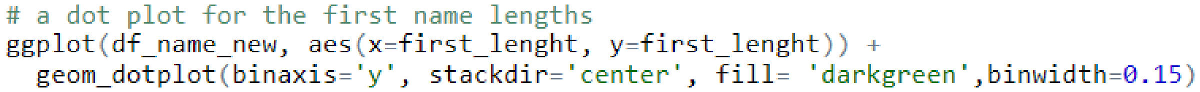
a dot plot for last name lengths



The range of most last names is between 4 and 7. Thirdly, a dot plot created for the first name length. Code in figure 6 and the plot in figure 7. The most length of the first name is 6. Compared to the last name, the distribution of the first name is between 4 and 8, which tends to be normally distributed. On the other side, the distribution of the last name is between 4 and 7 and skewed to the right.

Figure 6

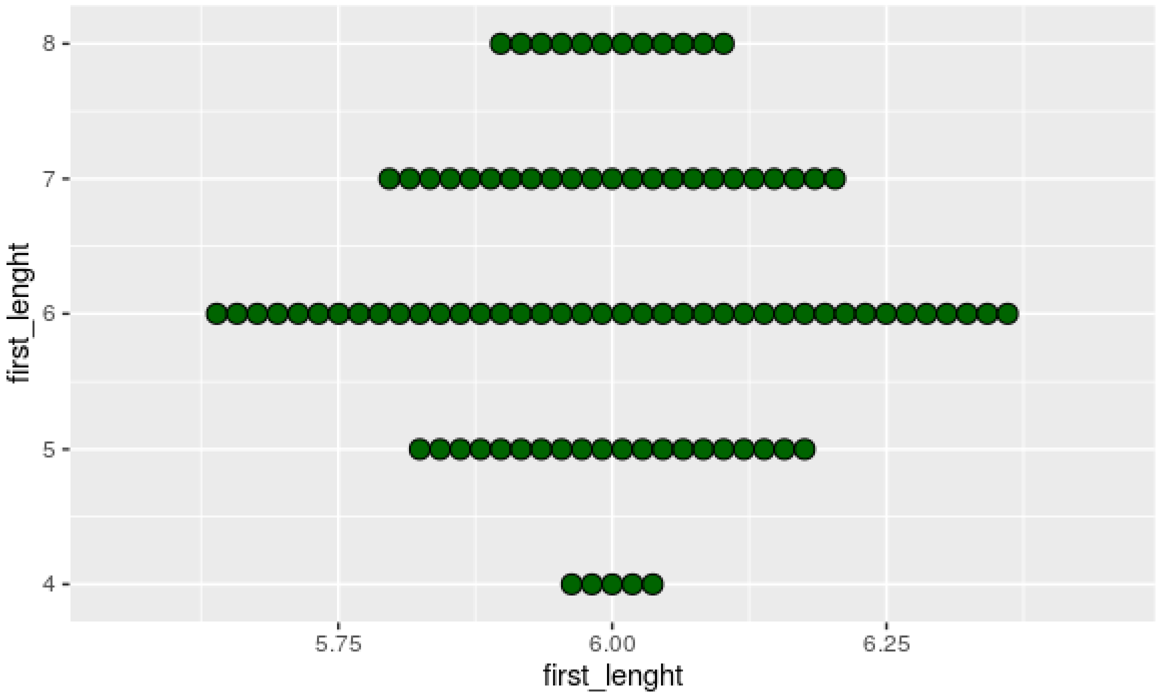
ggplot() function to create first name lengths code



```
# a dot plot for the first name lengths
ggplot(df_name_new, aes(x=first_lenght, y=first_lenght)) +
  geom_dotplot(binaxis='y', stackdir='center', fill= 'darkgreen',binwidth=0.15)
```

Figure 7

a dot plot for first name lengths



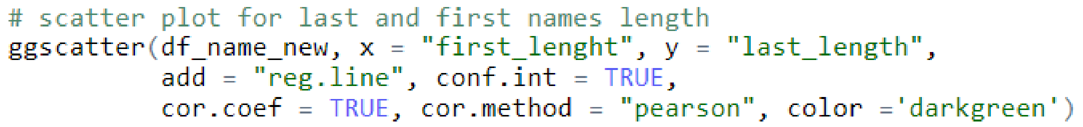
Part II: Regression Model

The regression model created using ggscatter() function from library “ggpubr” which is capable of creating the scatter plot with correlation coefficient. Also, this function provides many capabilities of adding the regression line, confidence interval and correlation method.

In the following figure 8, is the code to create the regression model. First name assigned to the x-axis and last name assigned to the y-axis.

Figure 8

ggscatter() function create scatter plot and correlation coefficient code

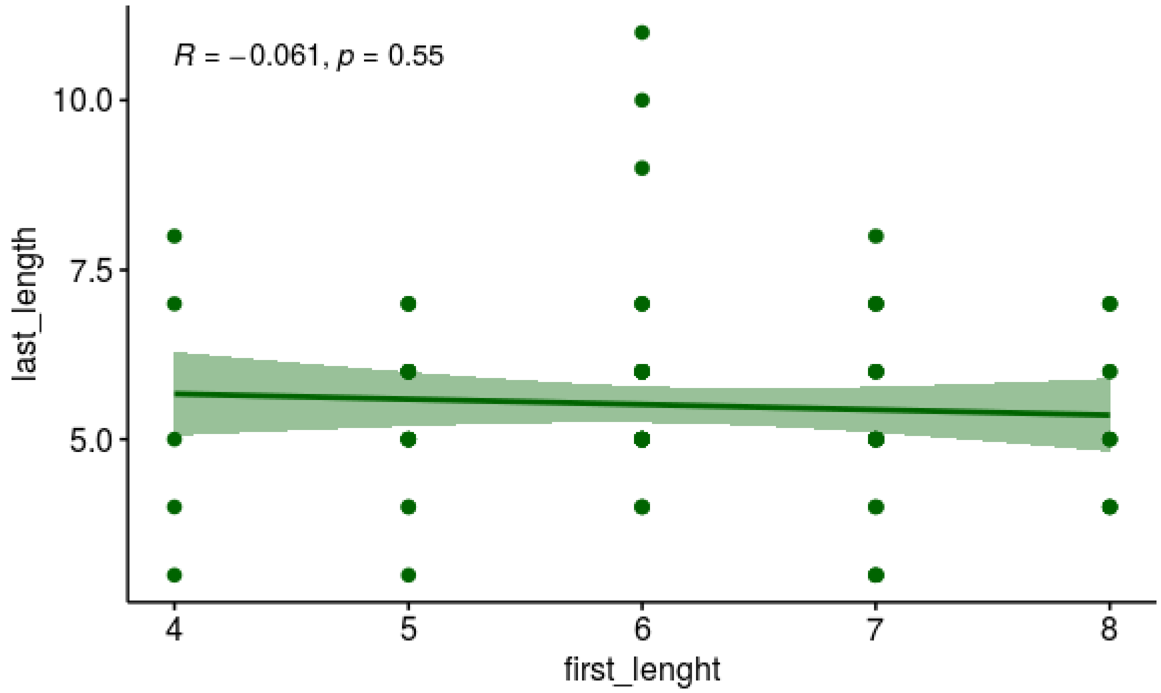


```
# scatter plot for last and first names length
ggscatter(df_name_new, x = "first_lenght", y = "last_length",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson", color ='darkgreen')
```

The scatter plot looks like a category plot, tend to be strip plot rather than scatter plot. This plot not consistent and did not show any positive or negative pattern. We cannot say people with long last names tend to have long first names. The plot in figure 9 shows who has 6 length first name is having the longest length. Correlation coefficient R is -0.061, this value shows weak relationship between the two variables and provide an evidence that there are no relationship between first and last names.

Figure 9

scatter plot and correlation coefficient



Part III: Reporting

My name is Abdulaziz Alqumayzi, unfortunately, my name did not exist in both last and first names. But my child Mohammed exists in the last name. The dataset has good quality, free from missing values, duplicates, and outliers. Yet, the two variables did not provide any correlation or pattern between the last and first variables to make further prediction analysis.

The analysis of this dataset started by creating new columns that have the length of the last and first names. Then, plot the two variables individually and gather to have insights about the two variables. After that, the ggscatter() function used to plot and show the correlation of the two variable.

Last name variable has skewed to the right distribution, first name variable has a normal distribution. The correlation coefficient is near to zero which provides a weak relationship between the two variables.

References

Field, A. P., Miles, J., & Field, Z. (2014). *Discovering statistics using R*. London, UK: Sage.

Peck, R., Short, T., & Olsen, C. (2020). *Introduction to statistics and data analysis*. Boston, MA: Cengage.