Exploratory Data Analysis

Abdulaziz M. Alqumayzi

(DS-510) – Discovering Statistics Using R

Colorado State University – Global Campus

Dr. Hasan Aljabbouli

September 25, 2020

Exploratory Data Analysis

Contextualization

The data set chosen is The Titanic data set.

The three questions of the previous critical thinking assignment were:

1- How many people that died and survived in this tragedy?

2- How many males and females in this tragedy?

3- What is the number of survived male and female in this tragedy?

These questions were answered in the previous critical thinking assignment. These questions may lead us to more insight after to do exploratory analysis.

Findings of the preliminary data analysis of the previous critical thinking assignment were:

1. The percentage of survivors in this incident is 0.38
2. The number of females and males in this incident is 466 and 843, respectively
3. The number of females and males who survived this incident were 339 and 161, respectively

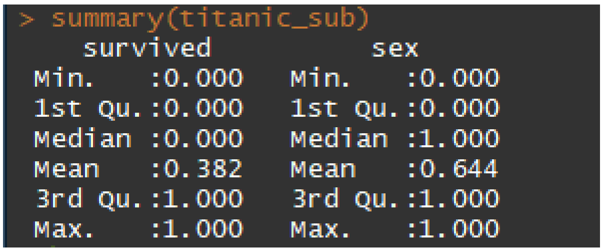
The three research questions were using the same variables. survived and sex variables. Question one used one variable which is survived variable. For the question two, the used variable is sex. Third question used both variables, survived and sex columns. In this exploratory data analysis, survived variable will be the dependent variable and sex variable will be the independent variable.

Statistical Description

The values of the survived and sex variables are 0 and 1, so the expected statistical description will be valued between 0 and 1.

Figure 1

summary function



```
> summary(titanic_sub)
    survived          sex       
 Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :1.000  
 Mean   :0.382   Mean   :0.644  
 3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.000   Max.   :1.000  
```

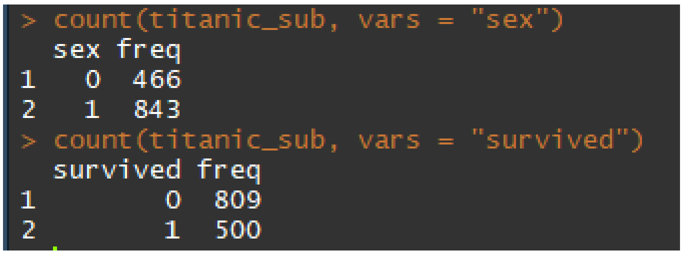
Note. In figure 1, summary() function is to generate the statistical description of the two variables.

Univariate Analysis

In univariate analysis shows the distribution of values in a column.

Figure 2

count of survived and sex columns

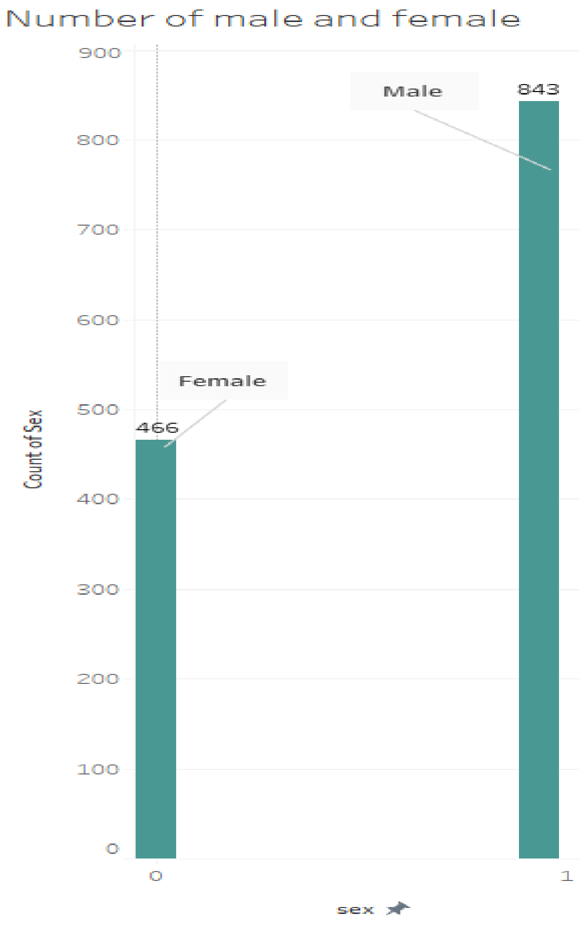


```
> count(titanic_sub, vars = "sex")
  sex freq
1   0  466
2   1  843
> count(titanic_sub, vars = "survived")
  survived freq
1        0  809
2        1  500
```

Note. In figure 2, count() function returns every value in a column and frequencies of that value.

Figure 3

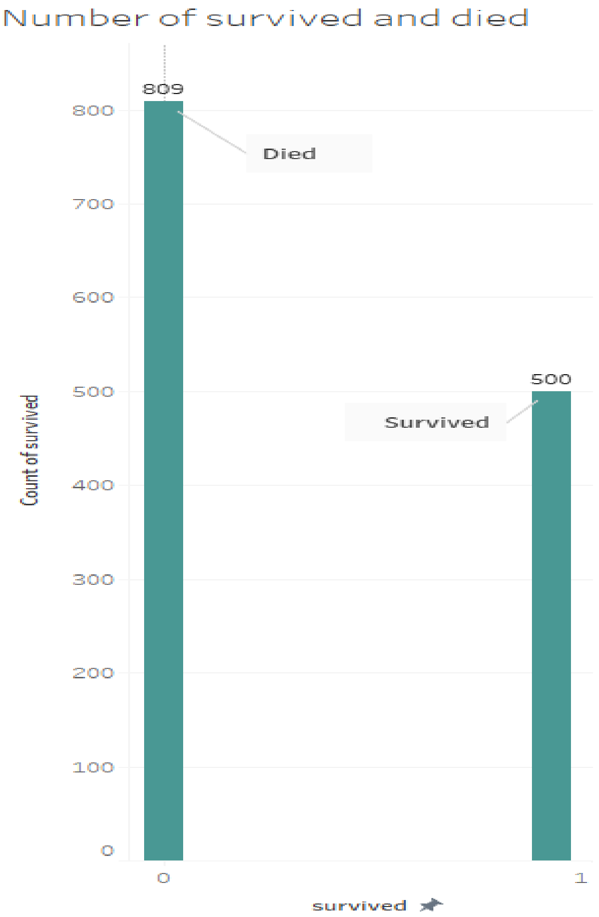
tableau histogram for column sex



Note. In figure 3, a Tableau histogram chart shows the number of males and females in the Titanic.

Figure 4

tableau histogram for column survived



Note. In figure 4, a Tableau histogram chart shows the number of survived and died in the Titanic.

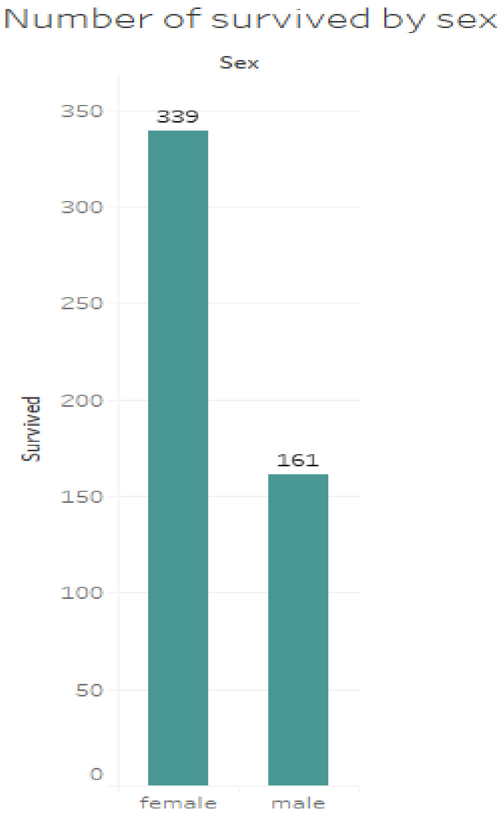
Summarization of the two histogram charts. The first histogram chart is the number of male and female in the Titanic were 843 males and 466 females. The sex column has two values 0 and 1, so the histogram seems to be a bar chart. Second, the number of survived and died in the Titanic was 500 people that survived and 809 died.

Bivariate Analysis

The bivariate data analysis will be on the two variables. The dependent variable is survived, and the independent variable is sex. Both variables have binary values, sex has male and female values and survived has 0 and 1.

Figure 5

tableau bar for number of survived by sex



Note. In figure 5, a Tableau bar chart shows the relationship between the survived column and the sex column.

The bar chart in figure 5 shows the relationship between the dependent variable survived and the independent variable sex. The first bar shows the number of females that survived in the Titanic and they were 339 females. The second bar shows the number of males that survived which is 161.

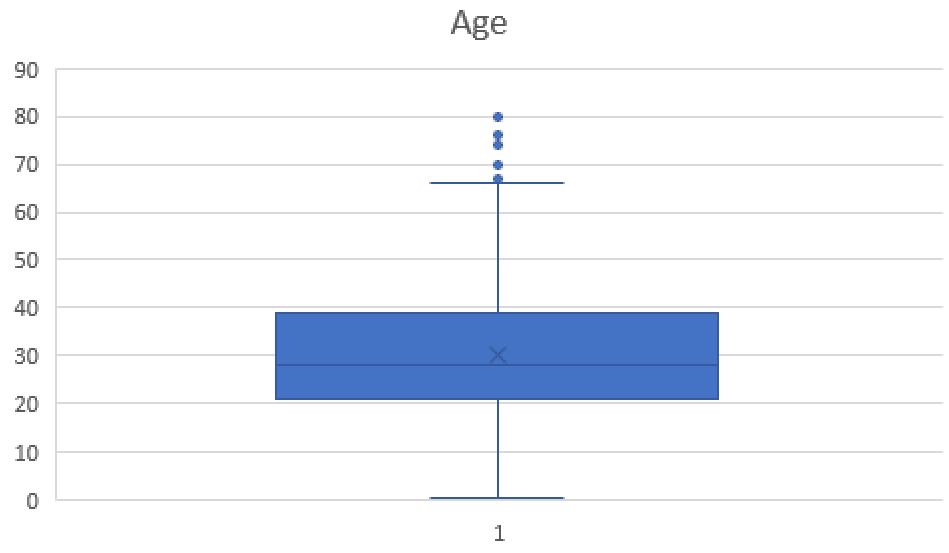
Outliers

In this exploratory data analysis, especially the analysis of the dependent variable survived and independent variable sex there are no outlier data points in both variables. As result has shown previously in figure 1, there were only two values in both variables 0 and 1.

Which means there are no outliers values in the two variables. The other variables will be shown in the following figures using Excel software.

Figure 6

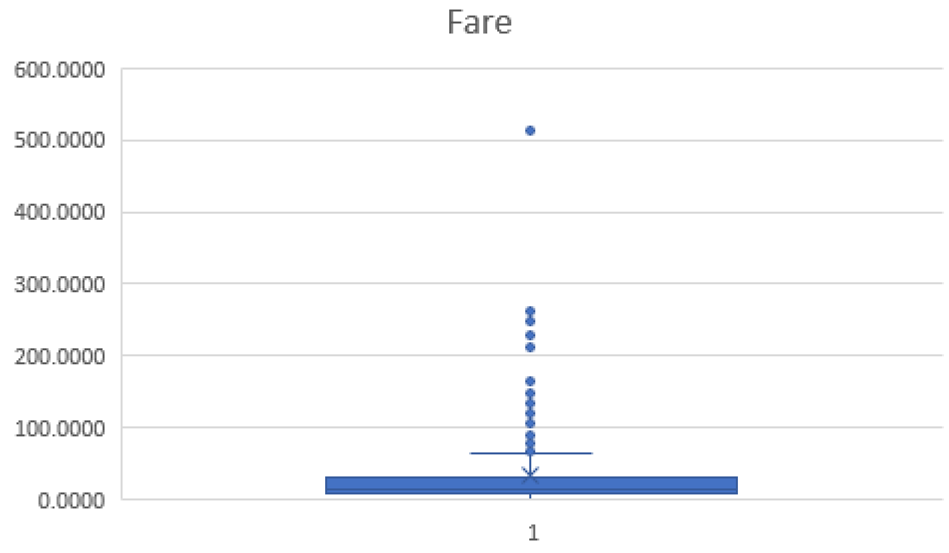
excel boxplot of age



Note. In figure 6, the boxplot shows there are outliers above the upper whisker.

Figure 7

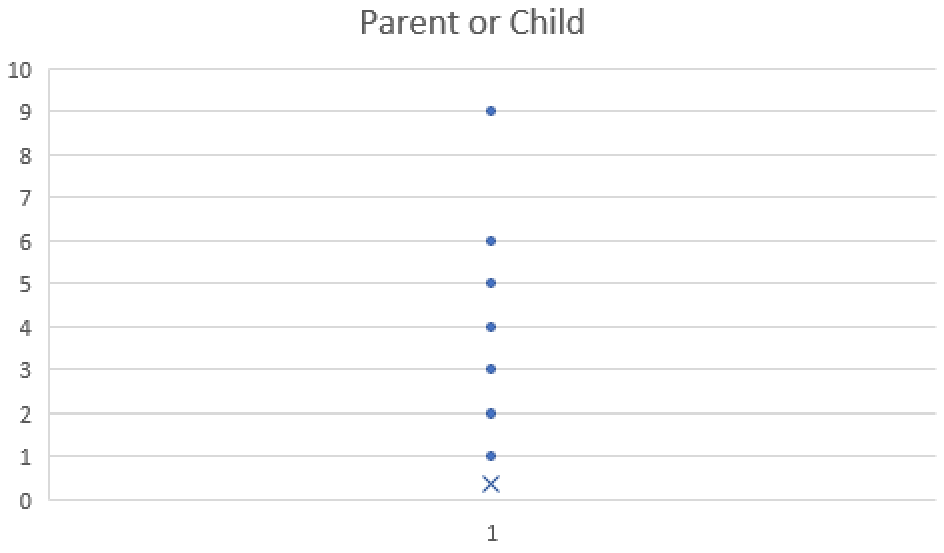
excel boxplot of fare



Note. In figure 7, the boxplot shows an extremely outlier and a lot of outliers above the upper whisker.

Figure 8

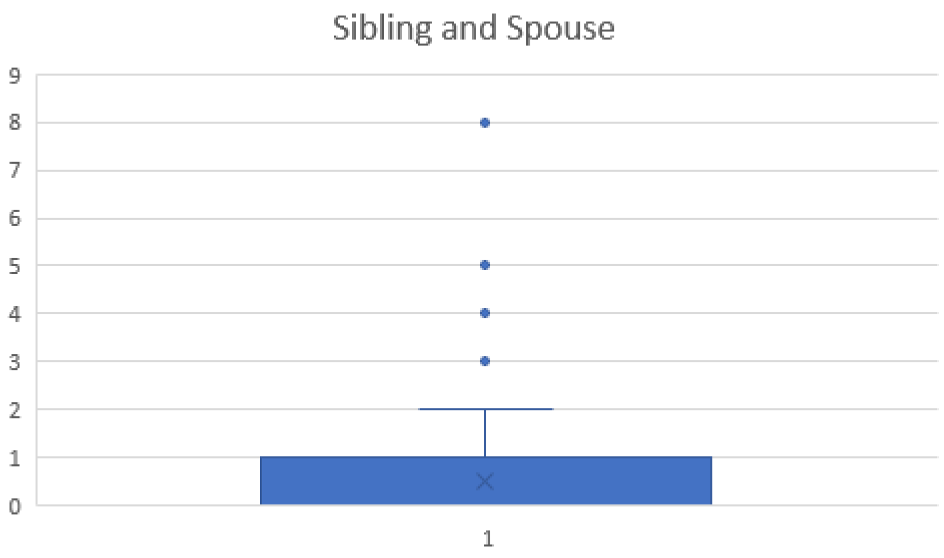
excel boxplot of parent or child



Note. In figure 8, This boxplot shows unusual and unique values. This variable can be used as a category.

Figure 9

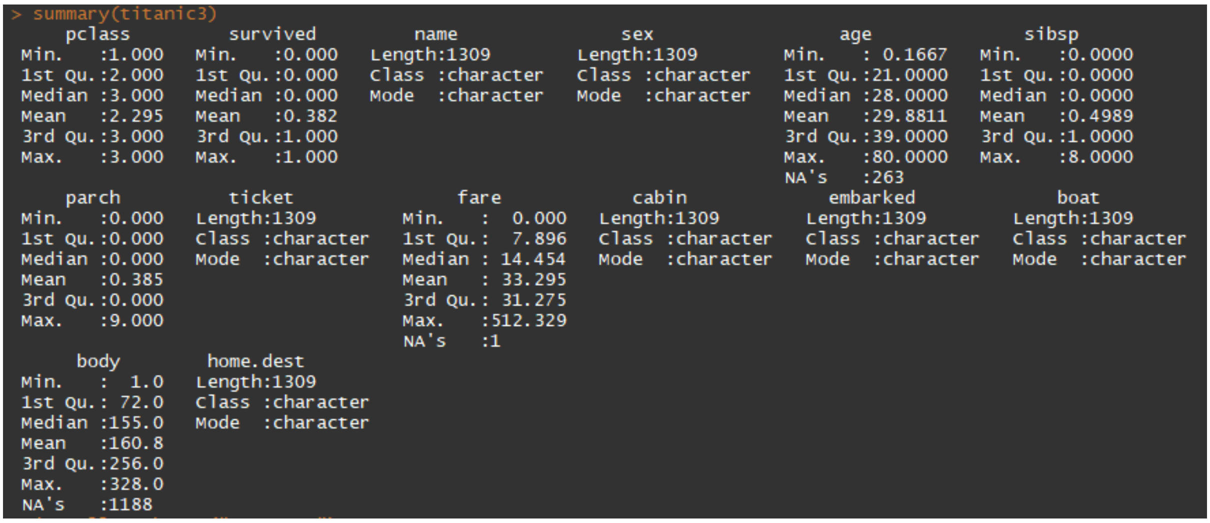
excel boxplot of sibling and spouse



Note. In figure 9, the boxplot shows above 2 siblings are outliers.

Figure 10

summary function for all variables



```
> summary(titanic3)
     pclass          survived          name                sex                  age               sibsp       
 Min.   :1.000   Min.   :0.000   Length:1309        Length:1309        Min.   : 0.1667   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character   1st Qu.:21.0000   1st Qu.:0.0000  
 Median :3.000   Median :0.000   Mode  :character   Mode  :character   Median :28.0000   Median :0.0000  
 Mean   :2.295   Mean   :0.382                                         Mean   :29.8811   Mean   :0.4989  
 3rd Qu.:3.000   3rd Qu.:1.000                                         3rd Qu.:39.0000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.000                                         Max.   :80.0000   Max.   :8.0000  
                                                                       NA's   :263                       
     parch          ticket                fare            cabin              embarked              boat          
 Min.   :0.000   Length:1309        Min.   :  0.000   Length:1309        Length:1309         Length:1309        
 1st Qu.:0.000   Class :character   1st Qu.:  7.896   Class :character   Class :character    Class :character   
 Median :0.000   Mode  :character   Median : 14.454   Mode  :character   Mode  :character    Mode  :character   
 Mean   :0.385                      Mean   : 33.295                                                             
 3rd Qu.:0.000                      3rd Qu.: 31.275                                                             
 Max.   :9.000                      Max.   :512.329                                                             
                                    NA's   :1                                                                   
      body          home.dest        
 Min.   :  1.0   Length:1309        
 1st Qu.: 72.0   Class :character   
 Median :155.0   Mode  :character   
 Mean   :160.8                      
 3rd Qu.:256.0                      
 Max.   :328.0                      
 NA's   :1188                       
```

Note. In figure 10, the summary() function of all variables to assess the variables programmatically. 1st and 3rd quartile values can be used to identify the upper and lower whiskers. The values above the upper whisker and values less than the lower whisker are outliers. For example, age 1st quartile (Q1) is 21 and 3rd quartile (Q3) is 39. The interquartile

(IQR) is Q3-Q1 which is 18. The lower whisker (Q1-1.5*IQR) is -6 and the upper whisker (Q3+1.5IQR) is 66.

Reporting

In this critical thinking 2, It started by research questions and throughout the tasks, I did to answer these questions. variables that answer the research questions were survived and sex variables survived variable chosen to be the dependent variable and sex variable for the independent variable. These variables have two values. the survived variable has 0 and 1 values, and the sex variable has male and female. The sex variable its values changed to 0 for female value and 1 for male for value in the previous critical thinking 1. So, it shows 0 and 1 instead of male and female that will use in Tableau for this critical thinking 2.

The statistical description of the two variables expected values to be between 0 and 1. Which is what the results are shown in figure 1. The minimum is 0 and the maximum is 1 which indicates that values are consistent. Function summary() from the R library base that used to summarize the statistical description of these variables.

The Univariate analysis of the survived and sex variables is to show the distributions of values for each variable. The software used to visualize the histogram charts was Tableau. Data was opened in Tableau and the software used to visualize the charts. There were two data imported, the original file and the cleaned file of the previous critical thinking 1. In figure 2, shows the frequency number of each value in each variable. Figures 3 and 4 show the values of each variable in a histogram chart. The histogram chart seems like a bar chart due to the only two values in the variables. The sex variable in figure 3 shows that the number of males is 843 and 466 for females. The survived variable in figure 4 shows the number of survived and died which is 500 and 809, respectively.

The bivariate analysis of the survived and sex variables is to show the relationship between the two variables. In this bivariate analysis, Tableau was used to visualize the bar chart for the dependent variable survived and the independent variable sex. This used the original data source by opening the data from Tableau software. In figure 5 shows the relationship between the two variables in a bar chart. Bar chart one the effective chart to show the relationship between quantitative versus qualitative data types. The bar chart in figure 5 shows that the number of survived females is 339 and 161 for males. A huge difference between the survival between females and males in this Titanic tragedy.

Outliers or anomalies in these two survived and sex variables are none. The variables survived and sex has two values, 0 or 1 which all values were verified as shown in figure 2. The other variables were assessed visually using excel as shown in figures 6, 7, 8, and 9 and assessed programmatically using the summary function of R code. Age in figure 6 shows that there are outliers above the upper whisker which is 66. The fare in figure 7 shows a lot and extreme outliers above the upper whisker which could be 70. The boxplot of parent and child that are shown in figure 8 seems not useful to identify the outliers. Therefore, it can be changed to a category. The sibling and spouse in figure 9 clearly identified the outliers above number 2.

References

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