Data Collection And Preliminary Data Analysis

Abdulaziz M. Alqumayzi

(DS-510) – Discovering Statistics Using R

Colorado State University – Global Campus

Dr. Hasan Aljabbouli

September 18, 2020

Data Collection And Preliminary Data Analysis

Contextualization

The data set chosen is The Titanic. The RMS Titanic, a luxury steamship, sank in the early hours of April 15, 1912, off the coast of Newfoundland in the North Atlantic after sideswiping an iceberg during its maiden voyage. Of the 2,240 passengers and crew on board, more than 1,500 lost their lives in the disaster. Titanic has inspired countless books, articles, and films (including the 1997 “Titanic” movie starring Kate Winslet and Leonardo DiCaprio), and the story of the ship has entered the public consciousness as a cautionary tale about the perils of human hubris. This data source provided in the blackboard and downloaded from the link that was provided in it. The Titanic data set has 1309 observations and 14 variables.

The story behind chosen this Titanic data set is to make an investigation about the survived people compared to the dead ones in this tragedy. The average age of the survived people. And the old and children that survived in this tragedy. In addition, type of gender that survived more. Also, the number of boats were in Titanic and the number of survived per boat.

This Titanic data set provide information of the passengers in this tragedy. Not all the passenger's information existed in this Titanic data set. The passengers and crew of the Titanic were 2,240 and only 1309 recorded in this data set. This sample is more than enough to answer some curiosity of this tragedy.

The data dictionary of the Titanic data set in Table 1.

Table 1

The Titanic dictionary

Variable	Definition	Key
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
survived	Survival	0 = No , 1 = Yes
name	Name	
sex	Sex	
age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton
boat	Boat number	
body	Body number	
home.dest	home destination	

Note. Adapted from Titanic: Machine Learning from Disaster by kaggle.com, 2020, https://www.kaggle.com/c/titanic/data

The Titanic variable notes:

1. pclass: A proxy for socio-economic status (SES):
    a. 1st = Upper
    b. 2nd = Middle
    c. 3rd = Lower
2. age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

3. sibsp: The dataset defines family relations in this way:
    a. Sibling = brother, sister, stepbrother, stepsister
    b. Spouse = husband, wife (mistresses and fiancés were ignored)
4. parch: The dataset defines family relations in this way:
    a. Parent = mother, father
    b. Child = daughter, son, stepdaughter, stepson
    c. Some children travelled only with a nanny, therefore parch=0 for them.

From the exploration of this Titanic data set, it clear that this data set is not good quality. There are a lot of missing values from an only visual assessment. Whet the programmatic assessment applied, the assessment will be more accurate. The types of data in the Titanic data set are mixed between qualitative type and quantitative type. pclass variable is qualitative data type and has three categories. Each category describes an ordinal level of cabin location in Titanic. In addition, the embarked variable is a qualitative nominal data type described as letters, and each letter indicates to the port of embarkation. Variables survived and sex are qualitative data types with binary values. For the survived variable, number 0 means that he or she died, number 1 means he or she survived. In contrast, the sex variable is binary value with a string representation of male and female. The variable age, sibsp, parch and fare are quantitative data types. They should be three of them discrete data and one is continuous, respectively. specifically, the age variable should be discrete, but in this Titanic data set, they put fraction numbers. The variables ticket, boat and body are nominal data, even though their values contain numbers they still qualitative data type because when the values are added or subtracted to each other will not give a useful result. The last to variables cabin and home.dest are clearly qualitative variables.

The Titanic data set captured from this data set that data need a lot of cleaning due to missing values. Applying machine learning needs a tidy data that each variable forms a

column, each observation forms a row and each type of observational unit forms a table to make the prediction very accurate. In this case, we only will describe the data we have. Most of the values of the variables are consistent. There are variables that have more than one value such as name and home.dest. This problem can be solved by regular expression codes integrated with R. The age variable had fraction numbers that are not correct. Also, the fare unit did not mention so we will not assume which unit of currency it is.

Summarize three observations from a cursory visual inspection of the data. The first observation is Andrews, Mr. Thomas Jr. He was in the upper class and his ticket was 112050. The fare is 0 which indicates that maybe he was part of the crew or data entry error. His cabin was A36 and he is from Belfast, NI. He died in this tragedy at the age of 39 years old. The second observation is Sage, Master. William Henry. A 14 or 15 years old boy from class 3, which is the lower class. His ticket number is CA. 2343 and there is no cabin data. The boy William was the only one found from his relatives. His relative's numbers were 8 in the Titanic and only the body of the boy was found. All his family died, and no body found except his body. The third observation is Baclini, Miss. Marie Catherine and she is 5 years old. Also, she was in the lower class. The good news she and her sisters survived in this tragedy in boat C. the fare was 19.2583 and the cabin number not mentioned.

Research Questions

Three questions for this Titanic data set comes in mind. How many people that died and survived in this tragedy? The ratio between the people survived and died will show the meaning of the disaster that happened in Titanic. The second question is how many males and females in this tragedy? This information will lead us to other questions. The third question is what is the number of survived male and female in this tragedy? this question will give us two information on the ratio of males survived and how many females survived.

The importance of these questions to show the behaviors of humans in tragedies like the Titanic tragedy. When a similar tragedy happened in the future, did the same behavior will happen? We do not know yet the answers, but the answers will show how humans decide to be sacrificed and made others survive. These answers will show us is there a bias in tragedies between gender? is there a way to eliminate this bias? Answers to these questions will give us insights into human behaviors in tragedies.

Variables that will help us to answer the questions are survived and sex variables. These two variables are enough to answer the three questions of the research. There are other variables to answer another question to find deep answers, but we will stick to the two variables that answer the research questions. the survived variable will answer the first question. The second question will be answered by the sex variable. And the last question, the two variables survived and sex will be used to answer this question.

The outcome of the data analysis should answer the three questions clearly. data visualization one effective way to communicate with anyone. The outcome from the first question will be a bar chart, a pie chart, or both charts together with the ratio between the survived and dead people of this tragedy and text report if it needs more clarification. The second outcome will be from the second question to answer how many males and females in this tragedy. Communicating this outcome will be as the previous outcome, a pie chart, bar chart, or both of the charts with text reports if they need more clarification. The last outcome again will be communicated by the same charts, pie chart and bar chart with the ratio of the survived male and the survived female. This outcome, which is the last one will use the variables survived and sex together to answer the question and use them in one chart.

The reason behind to visualize the outcome using the pie chart or bar chart because the variables survived and sex are qualitative data type, to be specific binary data. The pie

chart is very effective for comparing two to five values. And the two variables survived and sex has two values which make it possible and effective to use a pie chart. The problem with the pie chart is the count of values. It shows ratio effectively but not value count. So, the bar chart will be more effective to show the count of survived and sex variables.

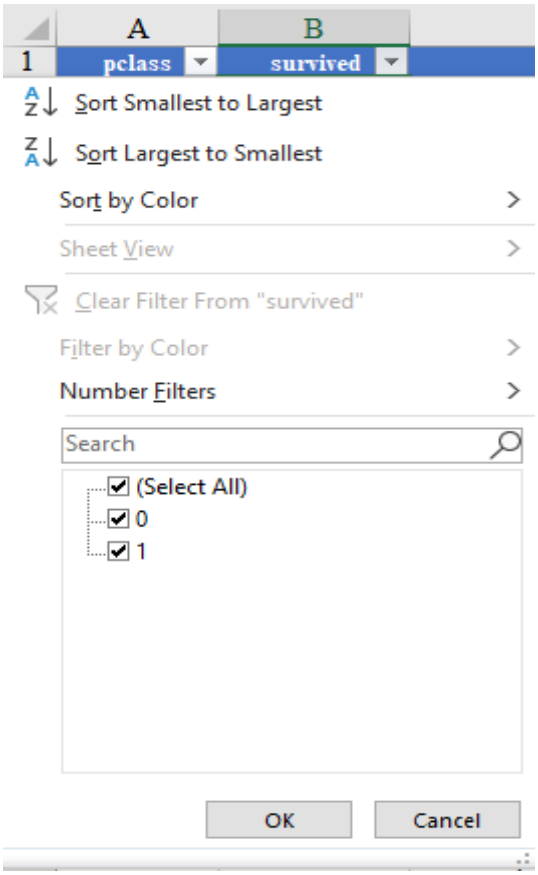
The analysis of this Titanic data set will focus on survived and sex variables. the data analysis starts with count the values for both variables and makes sure there are only two values. For the survived variable, it must have 0 or 1 values. In contrast, the sex variable must have male and female values. If both values are consistent, we can start to analyze the data and draw the charts to answer the questions and make our conclusion. But If the values of the two variables are inconsistent, the values need to be modified before do the analysis and draw the visualization.

Data Preparation and Distribution Analysis

The columns that will be used to answers the question are survived and sex columns. Before the start of the data analysis, values of the column survived must have two values which are 0 and 1. Identify these two values are very easy by excel software. First, convert the data to the table, then use the down arrow of the survived column to show all values of this column. This process will be repeated to the sex column.

Figure 1

survived column and its values

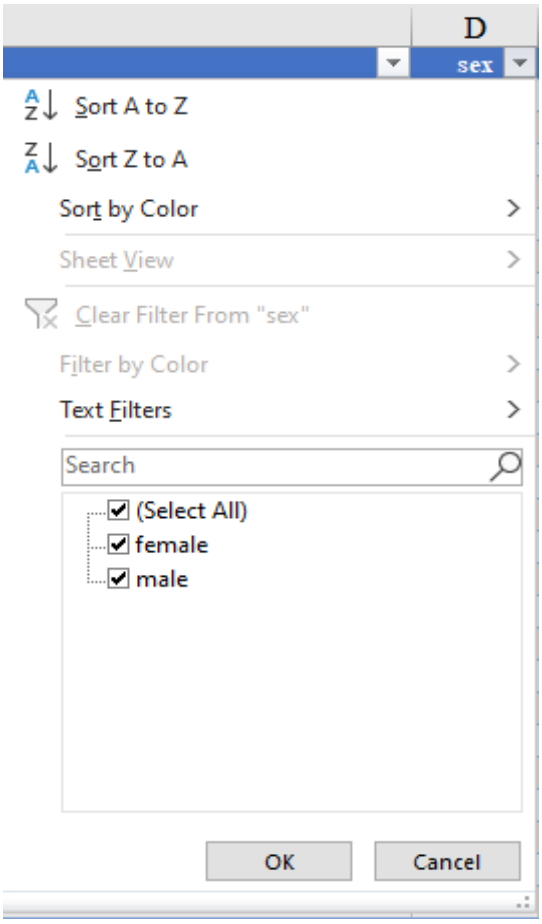


Note. This figure 1 shows the values of the column survived. The survived column contains two values 0 and 1 which is a consistent column.

The other column sex must have two values, male and female.

Figure 2

sex column and its values



Note. The figure 2 shows that sex column has two values, male and female.

The two columns are consistent. Now we can start the data analysis to answer the three questions. First, the libraries needed for this analysis were imported. The Titanic data set file is excel format, so the data has been imported to Rstudio using library readxl. Secondly, the variables that needed for this analysis are survived and sex columns, so a script generated to subset these two data into a new data frame. The data type of sex column is character type and has two values male and female. The two values were changed to 0 and 1, 0 for the female, and 1 for the male. After that, the data type was changed from character type into a numeric type. Then SQL queries were used to answer the research questions.

Figure 3

R code to assign a new data frame

```r
# generate a subset of the data
titanic_sub <- titanic3[,c("survived","sex")]
View(titanic_sub)
```

Note. This code in figure 3 assigns a subset of the main file into a new data frame. Two columns in this data frame are survived and sex.

Figure 4

R code to modify a value of a column

```r
# convert male and female values to 0 and 1
titanic_sub$sex[titanic_sub$sex == "female"] <- 0
titanic_sub$sex[titanic_sub$sex == "male" ] <- 1
```

Note. In figure 4 there are two codes. The first one is to change all values from female into 0 in column sex. The second code is to change the values from male to 1.

Figure 5

R code to modify the data type

```r
# change sex data type from character into numeric
titanic_sub$sex <- as.numeric(titanic_sub$sex)
```

Note. The code in figure 5 shows how to change data type into numeric.

Figure 6

R code script for a query the average of survived

```r
# the insight for question one
sqldf("select avg(survived) from titanic_sub")
```

Note. In figure 6, SQL query to answer the first research question.

Figure 7

R code script for queries the number of sexes

```
# the insight for question two
# first code to show insight the number of females
sqldf("select count(sex) from titanic_sub where sex = 0")
# second code to show insight the number of males
sqldf("select count(sex) from titanic_sub where sex = 1")
```

Note. In figure 7, the first code is to show how many females in the Titanic data set. In the second code, it shows how many males in the Titanic data set. These two codes answers the second research question.

Figure 8

R code script for queries the number of male and female survived

```
# the insight for question three
# the number of survived male
sqldf("select count(survived) from titanic_sub where sex = 1 and survived = 1")
# the number of survived female
sqldf("select count(survived) from titanic_sub where sex = 0 and survived = 1")
```

Note. In figure 8, the first code answers the third research question of how many males survived and the second for how many females survived.

Reporting

In this report, I will talk about an analysis of the famous incident of luxury steamship Titanic that sank in the early hours of April 15, 1912, off the coast of Newfoundland in the North Atlantic after sideswiping an iceberg during its maiden voyage. The Titanic dataset was obtained from Kaggle. A dataset contains many variables. In this data analysis, three research questions were focused on:

1. Percentage of survivors in this incident
2. The number of females and males in this incident

3. The number of females and males' survivors of this disaster

The work on this report was done using R and Excel. Starting from collecting the necessary data in the Rstudio environment and then modifying some of the data to extract important information to answer the questions of this research.

The answers to the research questions were found as follows:

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

In conclusion, the survivors were less than half and more than a quarter of the passengers of the Titanic incident, it was a sad disaster. The number of females on this Titanic ship's voyage is almost half the number of males. While the number of male survivors is much less than the number of female survivors. From these insights, it appears that males sacrificed their lives maybe to keep their families and females alive. This insight may show us that males may sacrifice again to keep females a life in disasters. To predict that will happen in the future, more predictive analysis needed, and this was only descriptive of a tragedy.

References

Berndtsson, M., Hansson, J., Olsson, B., & Lundell, B. (2008). *Thesis Projects: A Guide for Students in Computer Science and Information Systems*. London: Springer.

History.com Editors, H. (2009, November 09). Titanic. Retrieved September 16, 2020, from https://www.history.com/topics/early-20th-century-us/titanic