The Titanic

The Titanic was a famous ship, built in 1909, that set sail for its maiden voyage on April 10, 1912 and sank five days later, on April 15. Despite being known from the start as the "Unsinkable Ship", a collision with an iceberg led to the major disaster, killing more than 1500 people. But shouldn't the giant have survived that impact? Before its departure, there was a coal fire below the deck, which wasn't completely handled by the time the ship set sail. Some historians believe that this fire was getting out of control during the trip, forcing the crew to attempt a full-speed crossing. Because of how fast they were moving it was impossible to avoid the collision.

In January 2017, new photos were realized, showing the ship before the tragic accident. According to the Houston Chronicle newspaper: "[…] what drained resources going into the Titanic's building was its sister ship, the Olympic, which was also being built nearby. This led to a reduction in skilled labor and quality materials, including steel that was less than ideal […]". And that's what the pictures showed; the structure and material of the ship were the reasons for the accelerated sinking and, therefore, the impossibility of an effective rescue.

Another one of the Titanic flaws was the capacity of the lifeboats, since they could accommodate less than half its passengers. An aggravating element was the chaos, which resulted in many lifeboats being launched under-filled, some of them with only a handful of passengers. These are some of the reasons why only 705 people survived the tragic event.

There are many mysteries about the RMS Titanic, and until today there are new evidences and lost objects being found. In October 2017, a letter written by a Titanic passenger during the trip was sold at an auction to a collector. The passenger was Alexander Oskar Holverson, a 42 year old salesman who was travelling in first class with his wife. He wrote a letter to his mother, and the content was shared with the public by The New York Times: "[…] Mr. Holverson describes a 'giant' ship 'fitted up like a palatial hotel'. He also mentions seeing the millionaire John Jacob Astor sitting on a deck of the vessel […]. In the most poignant mention, Mr. Holverson writes, 'If all goes well we will arrive in New York Wednesday A.M' […]". The salesman died, along with many others.

The tragedy of the Titanic became a well-known story that composed countless books and movies, including the 1997 success starring Leonardo DiCaprio and Kate Winslet. It's important to remember and honor the victims by keeping the story alive, and continuing studies and researches about what happened. Analyzing datasets about the ship and its passengers is the contribution that data scientists can offer.

Working with the Titanic dataset

I was looking at the available data in R, and Titanic caught my attention, but when I checked it out, there were three things about it that I didn't like: there weren't any quantitative variables, the dataset worked with age groups (Child and Adult) instead of the actual age, and it was in a format that used a frequency column instead of having one row for each person. So I decided to stick with the interesting topic, but to look for another dataset about it. Finally, I got this one from kaggle, which contained 888 rows and 8 columns. There

were four categorical variables (Survived, Pclass, Sex, and Name) and four quantitative (Age, Siblings/Spouses Aboard, Parents/Children Aboard and Fare).

The cleaning was done using dplyr: with the function mutate, I created a new variable called “Family_Members_Aboard”, which was the sum of Siblings/Spouses Aboard and Parents/Children Aboard; using rename, I changed the name of Pclass to Class; with select, I specified which variables I wanted to keep and in what order; using arrange I ordered the rows alphabetically by names; finally, I changed the values of Survive from 1 and 0 to Yes and No, using for loops and if/else conditions.

When the data was ready, I started creating some exploratory graphs to look at each variable. The pie chart for the Survived variable tells us that more people died than survived. The bar chart for Class shows that classes 1 and 2 had a similar number of people, but class 3 had a lot more than the other two. The pie chart for Sex tells us that there are a lot more men than women. The frequency polygon for Age forms an approximately normal distribution, centered at 25 or 30, which means that there are more adults, and not many children and elderly. The density plot for Family_Members_Aboard is right-skewed and from it we can see that most people had 0, 1 or 2 family members aboard. The histogram for Fare is also right-skewed, like most plots about money are, showing the big concentration of data for low values and including just a few higher values, which represent the privileged.

To quantify the relationship between Class and Fare I composed a linear model. First, I created a scatterplot for the two variables and included a linear line. This plot suggests a negative, weak linear association, meaning that to be in the first class people needed to pay more than to be in the second, and so on. By creating a model with the lm function and printing its summary, we get a lot of information. From the estimates for the intercept and the slope of the linear line, we can find the formula y=107.606-32.661x, which tells us that if x were to be 0, y would get the value of 107.606 (this is theoretical, because Class will never get the value 0), and also that when Class increases in 1 unit, Fare decreases in 32.661 units on average. The three asterisks indicate significance and are probably there because of the low p-value that is indicated in the last line of the output. We also need to look at the Adjusted R-squared, which is 0.3005, indicating that approximately 30% of the variation in the observations can be explained by this model.

For my final R visualization, I created bar charts for Survived in each class, colored by Sex. For all three classes there is a pattern where men died more and women survived more. It is also possible to look just at the blue bars, and see that for all classes, the number of men who died is bigger than the number of men who survived, but that the difference between these numbers is a lot more accentuated for the third class, pointing out that the variable Class may be associated with Survived. Looking at the pink bars, we see that for classes 1 and 2, the count of women who died is lower than the count of women who survived, but for class 3, the count is roughly the same, reaffirming that there could be a relationship between Class and Survived.

To look at the relationship between Age and Survived, I created boxplots in Tableau. The box that represents people who died is a bit higher in the age distribution, which means that older people died more in the Titanic. By moussing over the boxes, we can see the five

number summary of each age distribution, and confirm that the first and third quartiles are higher for the No category. Tableau also shows the actual data points inside the boxplots, so I was able to color by Sex and increase the size according to Fare. The dots at the Yes category are generally bigger than the ones at the No category, which would indicate an association between Fare and Survived. But if we look at the colors and sizes, the points that represent women are usually bigger, and we also know that the group that did survive is mostly composed by women. This could mean that since Sex is related to Survived and also to Fare, it looks like Fare and Survived are associated, when really they aren't. If someone wants to look at a particular victim, it is possible to mouse over the points and see the name, age, fare and class of each person, and also how many family members were aboard.

In short, I found it very interesting to work with the Titanic dataset and am happy about the visualizations I created for the project. I liked using facet_grid to add a third variable to my R visualization, since it showed me the relationship for Sex and Survived inside each category of Class. Also, this was my first time creating a boxplot in Tableau, and I was glad to know that it shows not only the boxes, but the actual data points, allowing me to see both the five number summaries and the complete information about the victims when moussing over. My only regret is not having enough time to explore more relationships between pairs of variables; that would have been interesting, since everything seems to be connected in this dataset.

Bibliography

HOUSTON CHRONICLE. **Newly-released photos of Titanic show ship before its tragic demise.** Available at: <https://www.chron.com/news/nation-world/article/Newly-released-photos-of-Titanic-show-ship-before-10866647.php>

THE NEW YORK TIMES. **Titanic Letter Sells for Record Price at Auction in England.** Available at: <https://www.nytimes.com/2017/10/22/world/europe/uk-titanic-letter-record.html?rref=collection%2Ftimestopic%2FTitanic>

HISTORY. **Titanic**. Available at: < https://www.history.com/topics/early-20th-century-us/titanic>