Learning excel pivot tables by exploring the Kaggle titanic dataset

Bishoy Sharobim

27/08/2017

CONTENTS

1 Motivations 1
2 Foreword 1
3 Questions 2
4 Solutions 2
5 Conclusion 6
6 Appendix 6
7 References 9

1 MOTIVATIONS

In the first Minerva Collective mini-hackathon meet up that I went to, I was placed into a group that was to work on data from the organization that runs the peer support system across Australian schools. Some team members wasted no time and upon copying the datasets, they immediately dived into deep programming territory using Jupiter notebooks, python and the like.

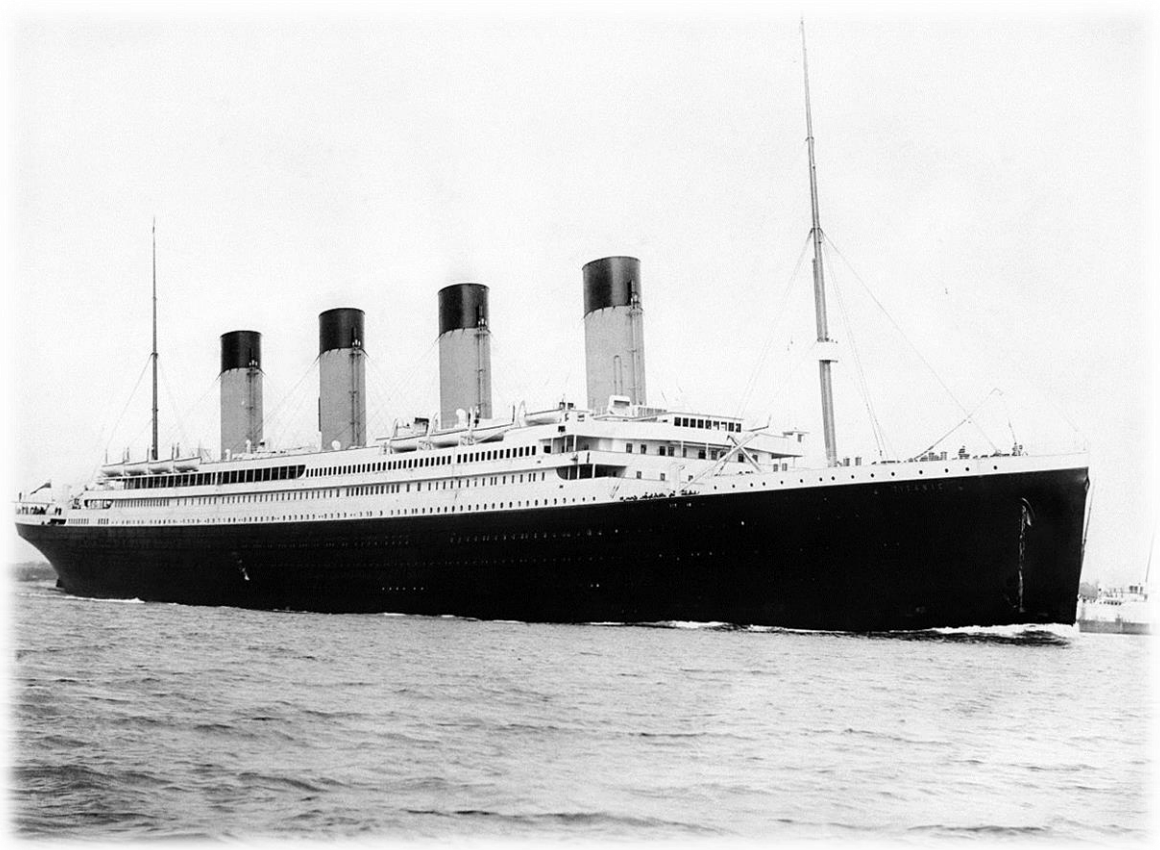
Who derived the best results in the tight timeframe? A team member who resorted to Ms Excel and pivot tables.

The purpose of this project is to help me learn Excel's pivot tables and its other powerful functions by analysing the famous titanic dataset from kaggle. Learning this I feel would grant me a stronger grounding of data analysis techniques and processes, paving the way for me to understand the dynamics behind the use complicated programming languages.

On this journey I was thrilled to find out hidden information that unravelled the biggest shipping tragedy in human history.

2 FOREWORD

On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.



One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

3 QUESTIONS

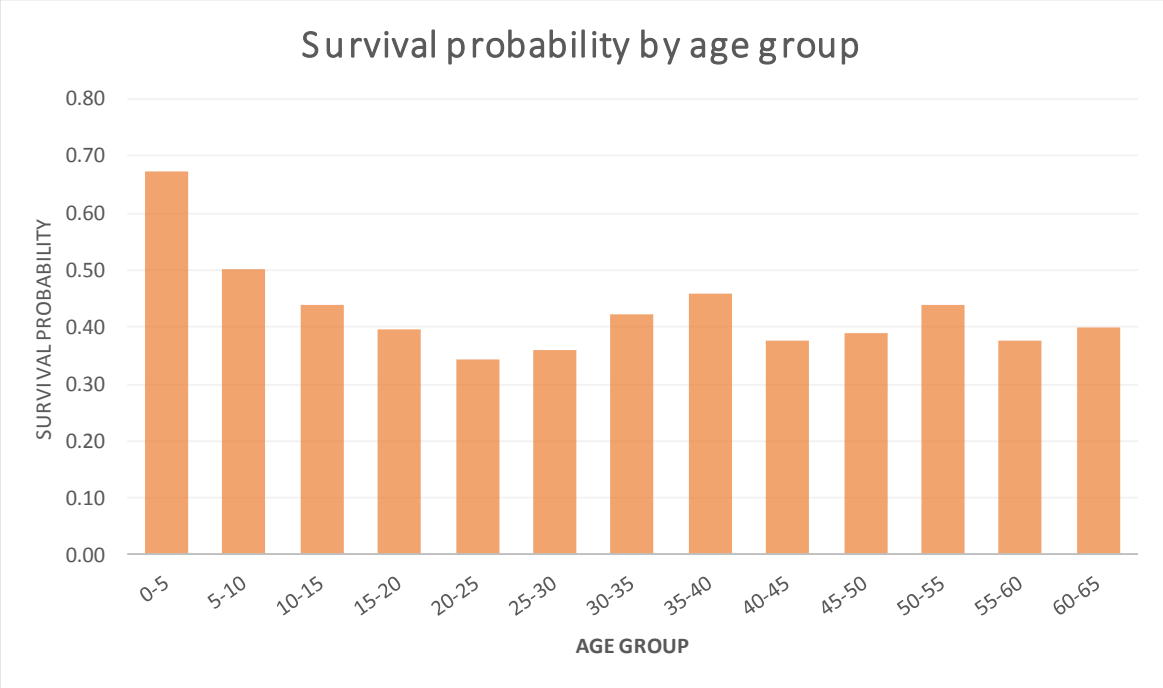
1) Find the amount of passengers in each 5 yr age range. Which age group has the highest survival probability? And which the smallest?
2) How much more likely is a male to survive than a female?
3) How many of each class were on board the titanic? Which class had the most survivors? What was the survival probability for each class?
4) How many families were on board the titanic? Which family had the most survivors?
5) Why wouldn't it make sense to talk about the number of survivors according to a particular category i.e. class, sex or age as an indicator of survivability?
6) Using the information from 1, 2 and 3, predict which passengers in "test.csv" survived or not.

4 SOLUTIONS

1) Find the amount of passengers in each 5 yr age range. Which age group has the highest survival probability? And which the smallest?

Using excel pivot tables, I summed up the number of passengers for each existing age and the corresponding number of survivors for each age. I then used the grouping feature on excel for pivot tables to aggregate these results for 5 year intervals.

The result is summarized as follows in the below column graph...



I only considered age brackets for where there were at least 10 passengers on board in that particular age range. This is because calculating information using such little data for a particular age group is unfairly representative of that particular age group.

For example there was only 1 person over the age of 85, who survived. It is unreasonable to say that all passengers over 85 in similar shipping catastrophes have a survival probability of 1.

Hence the age bracket that has the highest survival chances are the infants of 0 – 5. The one that had the smallest survival probability are passengers between the age of 20 and 25.

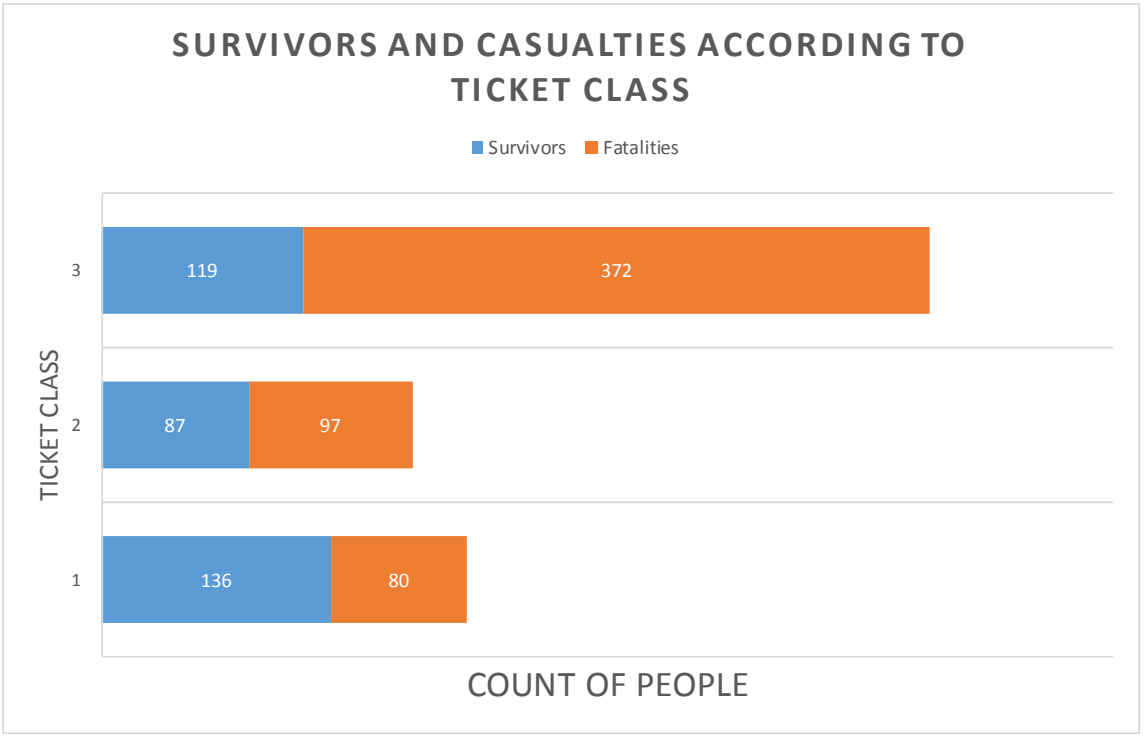
2) How much more likely is a male to survive than a female?

In the train dataset, this was the statistics of survivors according to each gender…

Sex	passengers count	Survivors count	Survival probability
female	314	233	0.742038217
male	577	109	0.188908146

Hence the survival probability of a female is 233/314 = 74% VS 109/577 = 19% for males. A female was 55% more likely to survive than a male confirming the data description in the foreword.

3) How many of each class were on board the titanic? Which class had the most survivors? What was the survival probability for each class?



Examining this above bar graph, we see that

- Class 1 had 216 individuals
- Class 2 had 184 individuals
- Class 3 had 491 individuals

From the data we see that the class that had the most survivors was class 3. However this does not indicate that a passenger who bought a 3rd class ticket are more likely to survive because we must take into account the overall amount of individuals who bought a particular ticket class.

Hence, the survival probability of each class is given as follows…

Class	Individuals	Survivors	Fatalities	Survival probablity
1	216	136	80	0.63
2	184	87	97	0.47
3	491	119	372	0.24

The 3rd class in fact had the lowest survival rate confirming the data description in the foreword, followed by the 2nd class, then 1st class.

4) How many families were on board the titanic? Which family had the most survivors?

[Note I realised that this whole questions is flawed - unfortunately and humorously. That's because the test dataset will contain some missing family members. I'll leave the answer in the appendix to proudly display my mistakes and as a source of education. I'm still proud of my thought processes in this question though.]

5) Why wouldn't it make sense to talk about the number of survivors according to a particular category i.e. class, sex or age as an indicator of survivability?

That's because within each of those categories, there would be an unequal amount of passengers for each possibility. For example more people with class 3 ticket survived than those with class 2 tickets. Does that imply the chances of survival are higher for those with class 3 tickets? No.

That's because there were more people with class 3 tickets than class 2 tickets on board the ship. Taking into account the number of total number of people on board for each ticket class, the survival probability of someone with class 3 tickets is LESS than for someone with class 1 tickets as I discovered in Q3.

6) Using the information from 1, 2 and 3, predict which passengers in "test.csv" survived or not.

Using Vlookup, I searched for the survival probabilities of each passenger in the test data in the values calculated through the training data. I did this according to the sex, age and class of the passenger.

For example, for the first passenger – Mr James Kelly – his survival probability based on the fact that he is a male utilised the formula …

= VLOOKUP([Sex], 'Pivot tables'!$S$8:$V$10, 4)

This formula would go to the sex variable for his row, then go to the table I included in Q2 and return the value in the 4th column i.e. 0.19. I repeated this according to the two other factors.

I only used these three factors because these seem to be the most influential factors in determining whether or not a passenger survived. Some factors are very clearly irrelevant for example the ticket number of the customer or the fare they paid. These two would play a very minimal role in determining whether or not a passenger survived.

I then determined the average of those three survival probabilities determined. Using an if statement, if this average survival probability was higher than 0.5 that means the passenger survived. I showed this as 1. If this was below 0.5, this means the passenger died, which is shown as 0.

Survival probability (sex)	Survival probability (age)	Survival probability (Pclass)	Average survival probabilty	Survived
0.19	0.42	0.24	0.32	0
0.74	0.39	0.24	0.46	0
0.19	0.40	0.47	0.35	0
0.19	0.36	0.24	0.26	0
0.74	0.34	0.24	0.44	0
0.19	0.44	0.24	0.29	0
0.74	0.42	0.24	0.47	0
0.19	0.36	0.47	0.34	0
0.74	0.40	0.24	0.46	0
0.19	0.34	0.24	0.26	0
0.19	0.68	0.24	0.37	0

I compared my results to the solutions provided from gender_submission.csv to see if I was accurate in predicting of whether or not each specific individual survived. I achieve an accuracy of 89% for which I am very pleased about.

5 Conclusion

Through this very fun and exciting analysis performed on the titanic dataset I was able to successfully train myself in learning more about and the use of excel pivot tables. Not only that, I've also now participated in my first ever Kaggle competition. I am quite glad about this as kaggle is the biggest data science competition environment that exists and I was able to be a part of this large and bright community. I am also very glad about achieving an accuracy of 89%. It goes to show that even simple tools like Excel and it's pivot tables have their place in the data science world.

6 Appendix

Excel functions and features used

Excel functions used were Vlookup, Average, If statement, Sum, Countif, Countifs, Left, Find, & and Floor. Other features included pivot tables, the ability to further summarise the pivot tables by grouping the first variable and the powerful graphical features of Excel.

The failed question 5

5) How many families were on board the titanic? Which family had the most survivors?

First off I used the LEFT and FIND functions in a formula that looks like this as away =LEFT(D2, FIND(",", [Name],1) - 1) to isolate the surnames of the passengers. I stored these values into a column with the heading of "Surname".

To work out the number of families on board I had to use to use a complex degree of COUNTIFS functions and add them together. I also had to use union and intersection theories. I explain this below.

For two passengers let…

- A be the event that their surnames, ticket class and port of embarkation match
- B be the event that their recorded number of siblings on board match
- C be the event that their recorded number of parents on board match

So in a logical way, I wanted the surnames, ticket classes and port of embarkation to match indefinitely. Additionally I wanted EITHER the recorded number of siblings on board OR the recorded number of parents on board to match. If an individual has at least one sibling on board, the "SibSp" variables will match for all the entries associated with the siblings on board. Likewise if an individual has at least one parent or child on board.

Hence, I did….

$$A \cap [B \cup C]$$
$$= A \cap [B + C - B \cap C]$$
$$= A \cap B + A \cap C - A \cap B \cap C$$

Hence I did

= COUNTIFS([Surname], E2, [Embarked], P2, [Pclass], C2, [SibSp], K2)

+ COUNTIFS([Surname], E2, [Embarked], P2, [Pclass], C2, [Parch], L2)

- COUNTIFS([Surname], E2, [Embarked], P2, [Pclass], C2, [SibSp], K2, [Parch], L2)

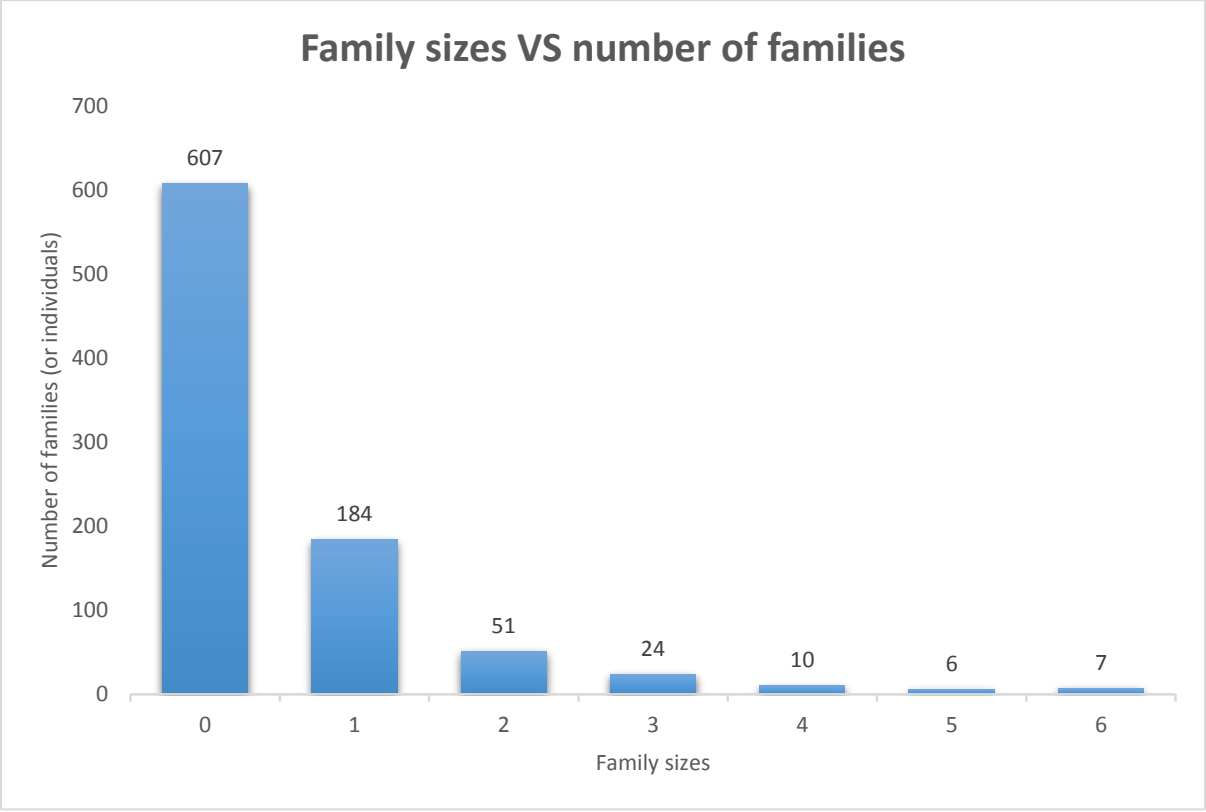
- 1

to work out the number of families on board per passenger. The – 1 is to remove the count of the individual themselves in the answer.

This gave me something like this…

Surname	Boarded family members (1)
Braund	1
Cumings	0
Heikkinen	0
Futrelle	1
Allen	0
Moran	2
McCarthy	0
Palsson	2
Johnson	3
Nasser	1
Sandstrom	0
Bonnell	0
Saundercock	0
Andersson	1
Vestrom	0

Then I created another pivot table demonstrating the count of the number of families of varying sizes. I present this in the graph below…



Thus there were 607 individuals, and 282 families.

I did not factor in whether or not passengers had the same ticket number into this methodology because I believe it would skew the results. I thought that same family members wouldn't necessarily have the same ticket, although 74% did.

I acted under the assumption that if you find records where the port of embarkation, surname, the ticket class match, as well as the number of siblings on board OR the number of parents/children on board, that would be a sufficient indication that those two entries are for two family members.

7 REFERENCES

- www.kaggle.com/c/titanic