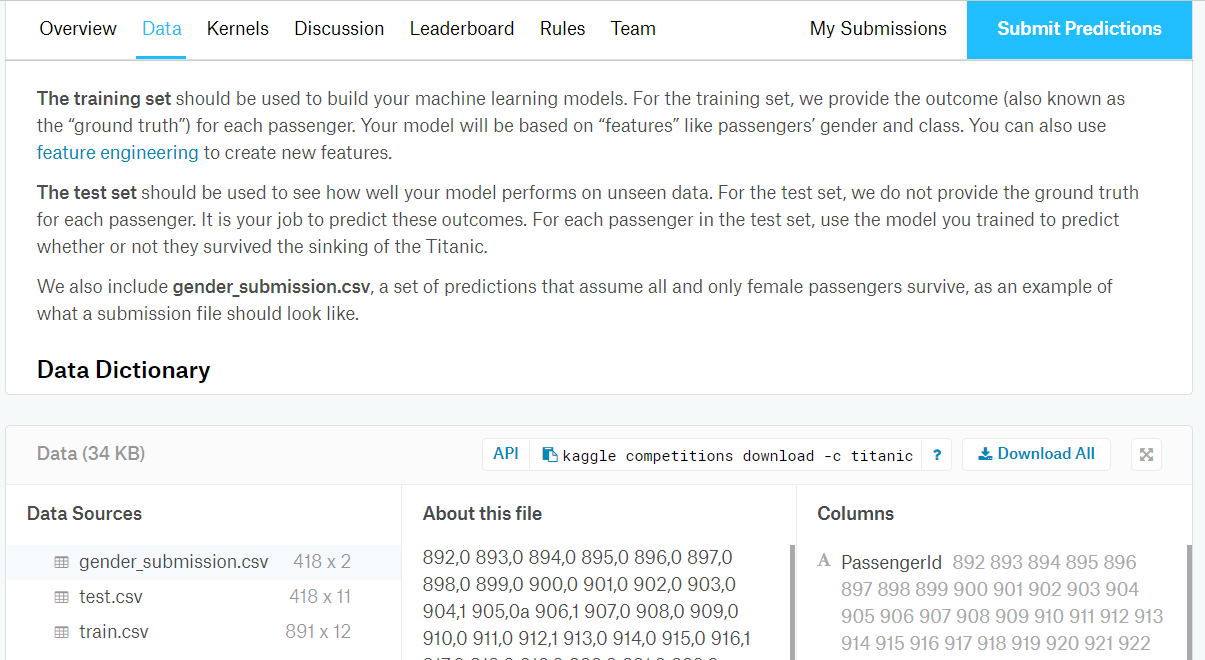
# Guess who would survive the Titanic with the Key Influencers visual!

Hi I am Bertalan Ronai from Hungary. <https://www.linkedin.com/in/ronaibertalan/>

The following tutorial is going to describe my Power BI report that is using the new Key Influencers visual to enter in the Titanic machine learning competition using only Power BI.   
No peeking at Azure Machine Learning Studio is allowed this time!

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew.

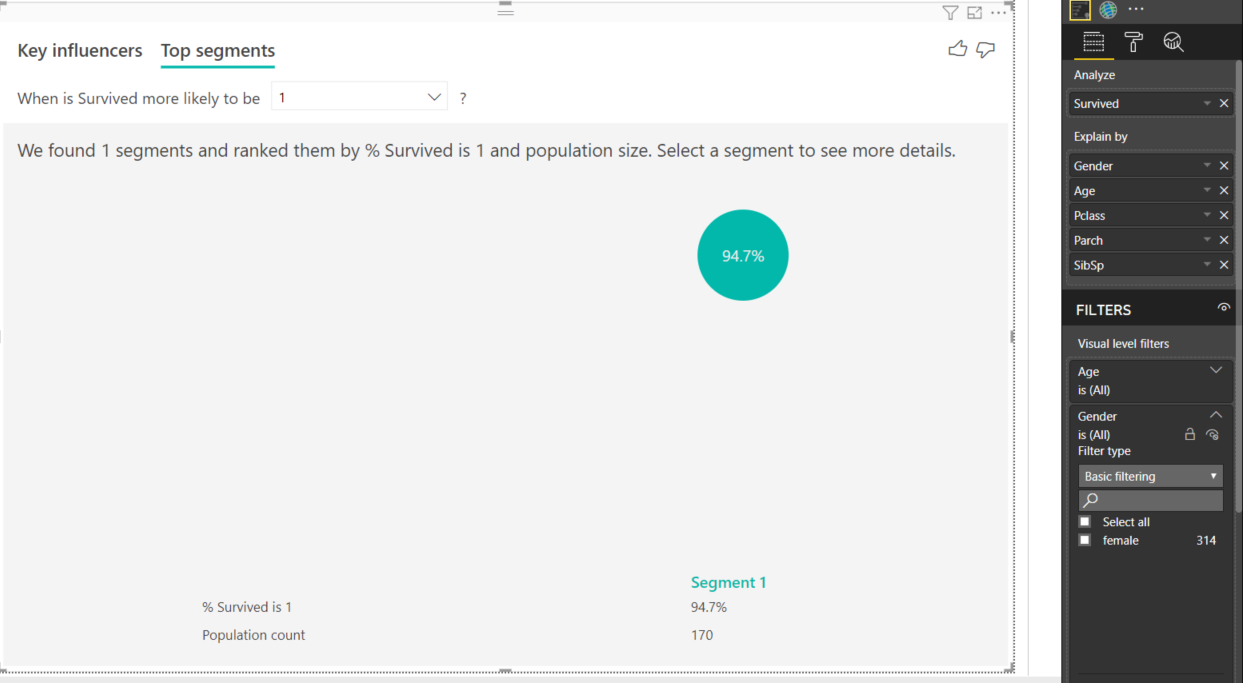
You can read more about the competition. <https://www.kaggle.com/c/titanic> And download the csv-files from here: <https://www.kaggle.com/c/titanic/data>   
We are going to use the test.csv , train.csv and gender\_submission.csv files.



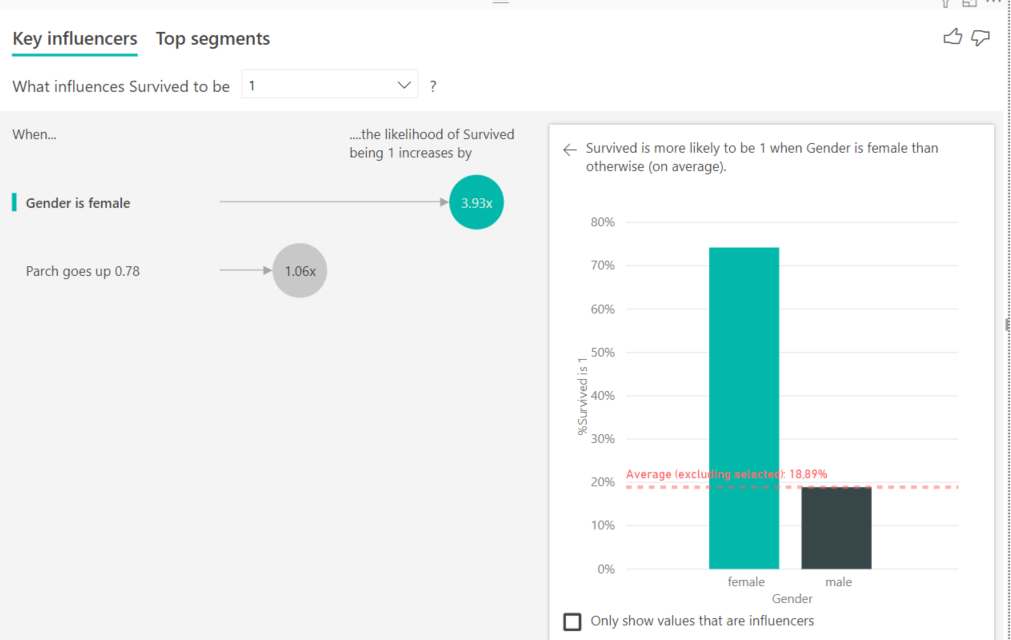
You are training your machine learning model on one dataset and test it on a test dataset.

If you haven’t already, then please watch this video from [@PATRICKDBA](https://twitter.com/patrickdba?lang=en) and [@JustynaLucznik](https://twitter.com/JustynaLucznik?lang=en) !

<https://www.youtube.com/watch?v=NTvg_3fcoZo> Embed this video!



The Key Influencers visual can find segments in your data and identify key influencers

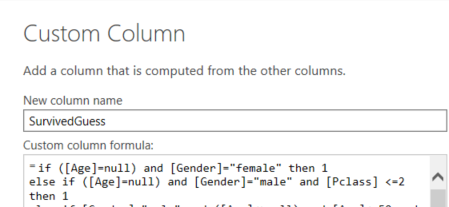


## Short Detour about importing CSV files

When you import from a csv file you might see that you have a hardcoded columns argument in the query editor like this one:  
= Csv.Document(File.Contents("C:\Anaconda\datacamp\_facebook\_live\_titanic\data\train.csv"),[Delimiter=",", Columns=12, Encoding=1250, QuoteStyle=QuoteStyle.None])

What happens when you replace the original csv file with a new csv that you csv that you have exported from a source and the new one has 13 columns? You are right, it won’t import the new column! So you should manually remove the Columns =12 part from the query above to make it independent of the number of columns.

Based on the segments the visual created I created a calculated column that evaluates into 1 or 0 (true or false)



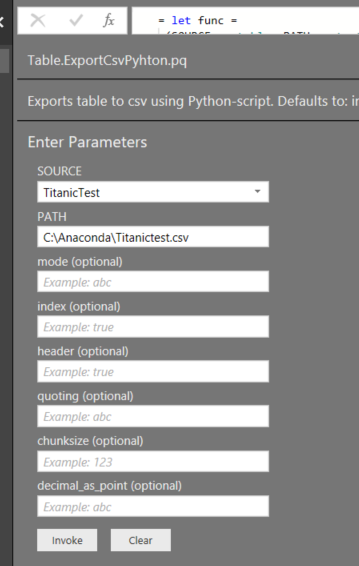
Check what happens if you don’t put a parenthesis around the logical test  
such as HasCabin=true or a[Age]=null!

To create the Test query, I have copied the Titanic Train query that has the Survived column to create a TitanicTest query/table that does not have an original survived column, only the new SurvivedGuess calculated column. I can send Kaggle a sumbsission csv containing two columns, the PassengerId and Survived (containing my guesses).

I have used Imke Feldmann’s Power query script that uses Python to write a csv.

By the way, I contributed a tiny bit to that script by adding a decimal separator argument, check it out! Read more about this great function here: <https://www.thebiccountant.com/2018/12/29/export-data-from-power-bi-to-csv-using-python/>

Here is how I called this function:

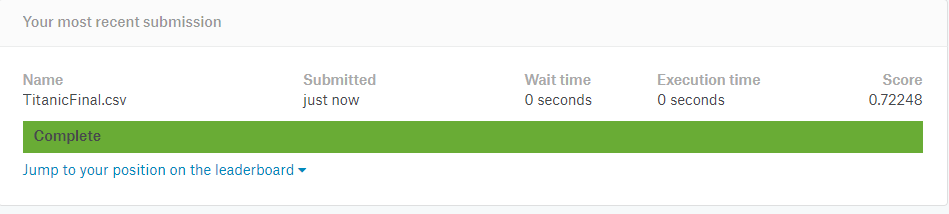


Sorry, I won’t go into the details of how to use Python in Power BI, only to that extent to say that you should install Python regardless of having Anaconda installed or not.

We can read the gender\_submission csv to see how a submission should look like! It has only 418 rows beside the headers. You could also create a query that I called CsvTestBeforeSubmit that reads your csv before submitting.

Alright! The only thing left now is to enter the Kaggle competition and reach a high score.

Try to reach a higher score than mine! Did you know that if you predict that only the women would survive then your score would be 76,55% ? Try to reach a higher number by predicting the clusters of women and men who would survive.



I have asked my penpal, Alex Powers to create a walkthrough video in which he would build upon this idea to make it even better.