



# A trip by sea what fun it can be a voyage so beautiful, you'll see.

Poem by unknown source

*This blog post is part of my “Data Scientists Nanodegree” of Udacity. Detailed analysis with all required code is posted in my github repository.*

A sea voyage can be a very adventurous and beautiful undertaking. However, there have been tragic accidents in the past (see e.g. Costa Concordia disaster 2012, [https://en.wikipedia.org/wiki/Costa\\_Concordia\\_disaster](https://en.wikipedia.org/wiki/Costa_Concordia_disaster)). So that a vacation does not become a disaster, I would like to examine in this blog post one of the biggest accidents closer and make recommendations what possible risk factors are.

So I decided to take a closer look at one of the biggest accidents in seafaring. The Titanic disaster.

Hence, I was interestested in using Titanic data from Kaggle to better understand what my chances are to survive a shipping accident:

- **Question 1:** How old were the passengers of the titanic. How old was the oldest passenger?
- **Question 2:** How about the passenger distribution by gender? Where more male or more female passenger on board?
- **Question 3:** What is the most influential factor if a passenger died or survived?

## Passenger Structure (Q1 + Q2)

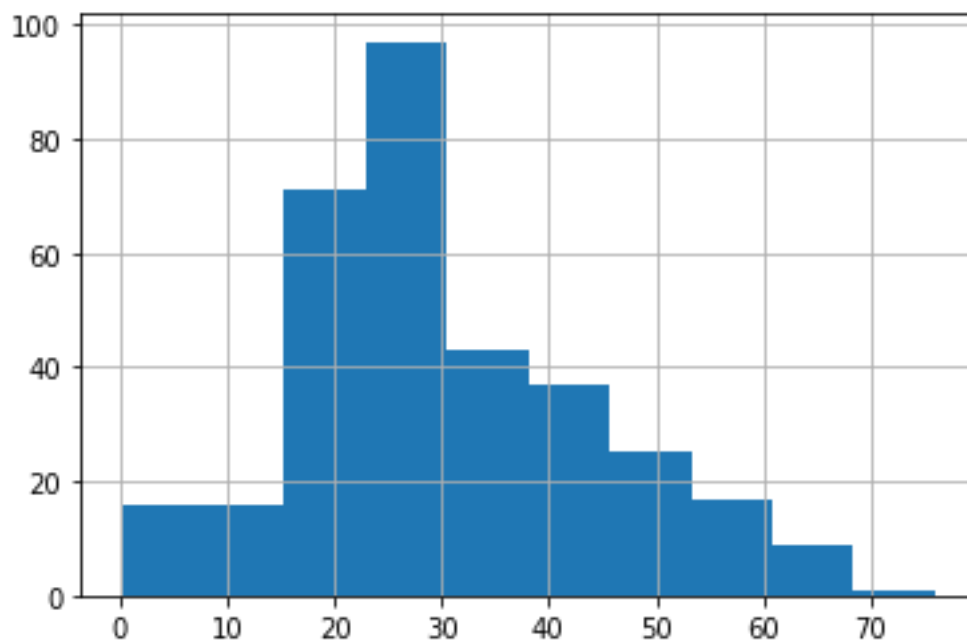
In a first step I decided to investigate the passenger structure to get a better understanding of the age and the passenger characteristics. For loaded all passengers in my workspace and printed the key characteristics of may dataset:

:

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

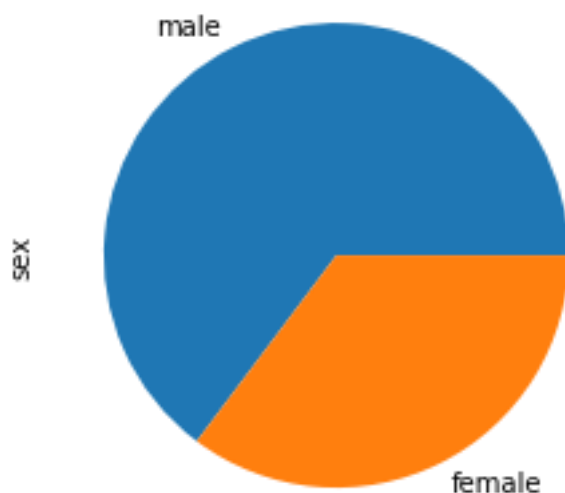
As we can see the mean age of the passengers was about 30 years with a sd of 14.5 years. The interquartile range is between 20 and 38 years, indicating that 50% of the passengers were in this age range. The oldest passenger was 80 years old.

This gets even more obvious if we take a look at the passenger distribution:



In a next step I investigated if there might be influence of the gender to the survival of the catastrophe. Everyone knows the famous sentence "*women and children first*". I wanted to know if this holds true in case of an accident.

Hence I investigated if there were more female or male passengers on board and compared the numbers with the survival rates of male or females.



As we can see there were more male passengers. In a next step, I plotted the survival rates by gender, and there was a clear gender bias towards female passengers' survival.

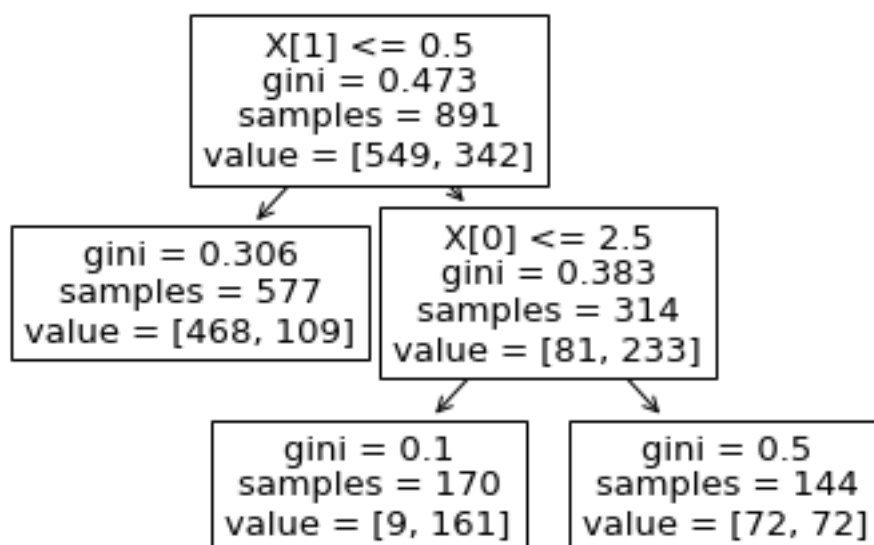
Female Rate: 0.34763948497854075  
Male Rate: 0.2329059829059829

Without going into detail I found strong evidence that I as a male would probably have serious problems in case of an accident because females are rescued first. The male passengers that survived were probably male children. So bad news for you guys...

### Risk Factors (Q3)

In a last step I investigated possible risk factors on survival. For that purpose I build a decision tree and investigated the most influential features.

As you can see in the tree plot gender, but also position in the ship were crucial factors.



As conclusion, I would recommend male passengers to invest some more money on their ticket to get a higher position in the ship and hence have a higher chance to survive an accident. To change gender is something I would not recommend, but could also influence your survival rates positively dear male folks.

Have a nice day and thanks for your attention