Investigate a dataset

1. The dataset analysed

- Titanic survival data.
- It consists of multivariate wise demographic data. It also has passenger and crew info who were present during the Titanic wreak.

2. Statement of questions posed

- Which factors influenced the survival metric of a person?
- Who has better survival chances? The ones who travelled with family? Or the ones who travelled alone?

3. Steps taken for analysis based on the questions posed

- The most reasonable features or attributes to consider were — *Gender, Age, SibSp* and *Parch, Pclass*. These features are analysed against the *Survived* attribute.
- **Gender** : This was a simple computation of grouping by Male and Female. Plotted graph in terms of percentage to understand who had better survival rate.
- **Age** : Compared this against six age groups based on biological factors that I believe change in the humans. Roughly based on physical strength or endurance and the maturity levels. So created six age brackets as follows :
    - Age 0 to 8 yrs, 9 to 18 yrs, 19 to 28 yrs, 29 to 40 yrs, 41 to 60 yrs, 61 to 80 yrs.

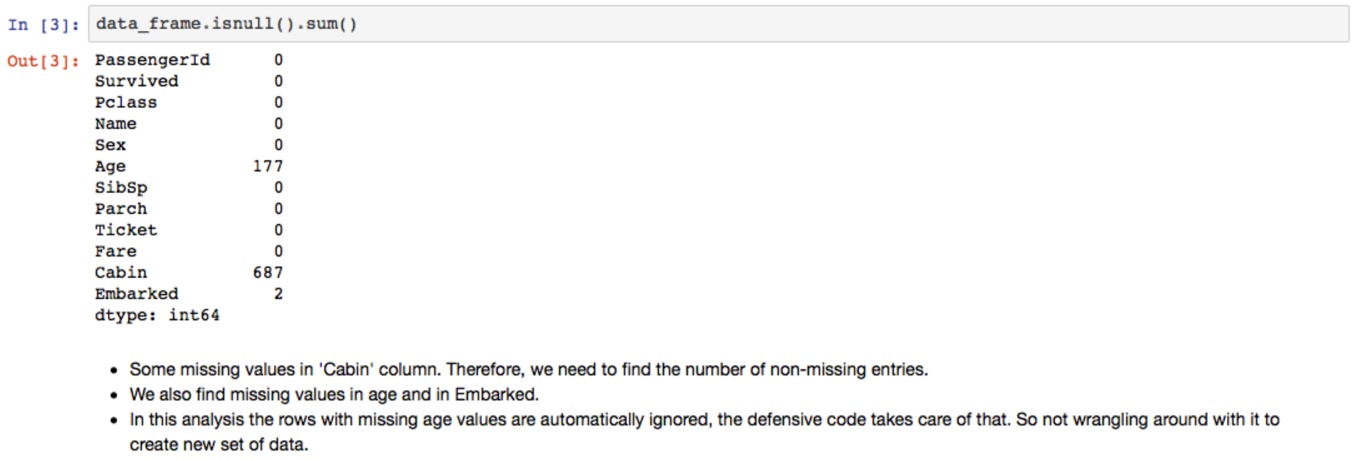
- **Pclass :** Compared against three different classes.
    - First class, second class and third class.

- **SibSp and Parch** : This feature is analysed to answer my second question. Again a simple computation of grouping by based on the accompany present for every passenger. A % plot to show if more alone people survived.

---

4. Documentation of data wrangling.

- No scope of any data wrangling in this dataset. Not a complex structure or any format to be converted.
- However found some NaN values as shown in the screenshot below.



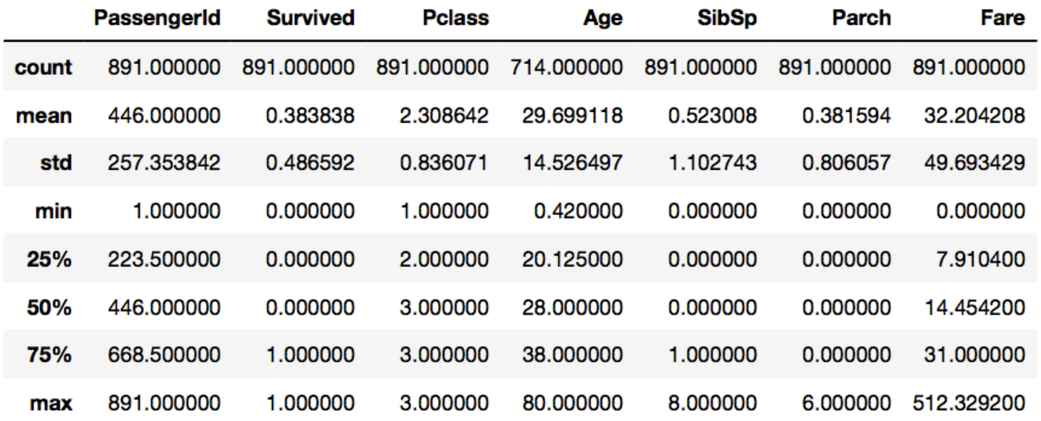
```
In [3]: data_frame.isnull().sum()
```

```
Out[3]: PassengerId      0
        Survived         0
        Pclass           0
        Name             0
        Sex              0
        Age            177
        SibSp            0
        Parch            0
        Ticket           0
        Fare             0
        Cabin          687
        Embarked         2
        dtype: int64
```

- Some missing values in 'Cabin' column. Therefore, we need to find the number of non-missing entries.
- We also find missing values in age and in Embarked.
- In this analysis the rows with missing age values are automatically ignored, the defensive code takes care of that. So not wrangling around with it to create new set of data.

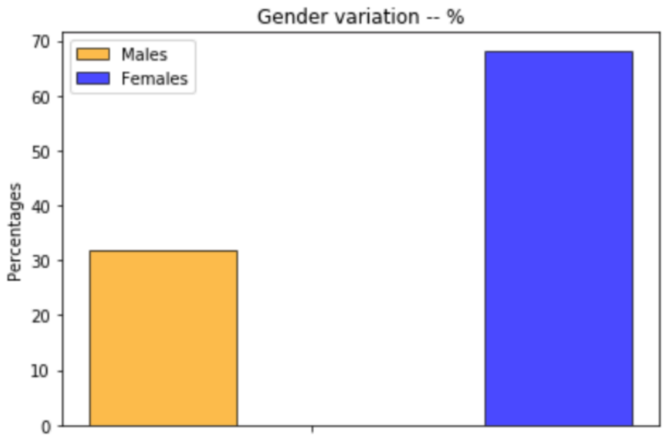
5. Summary statistics and plots

- Summary of features to get a bigger picture.



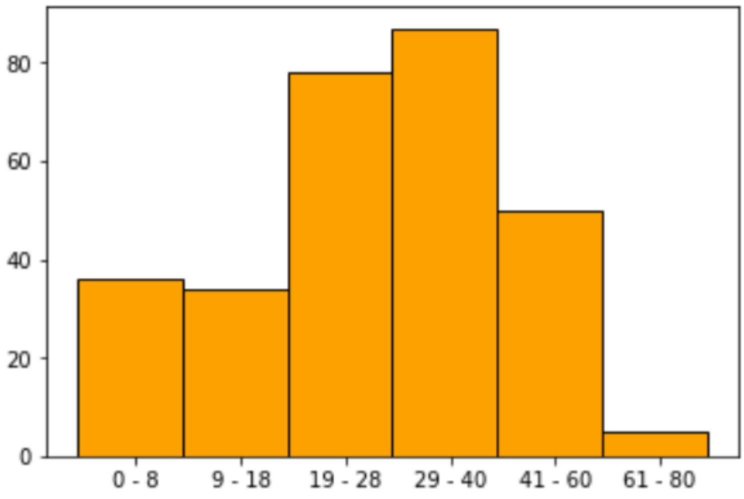
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

- Gender variation



- More females survived than the males.
- Therefore survival rate was higher in females than the male gender.

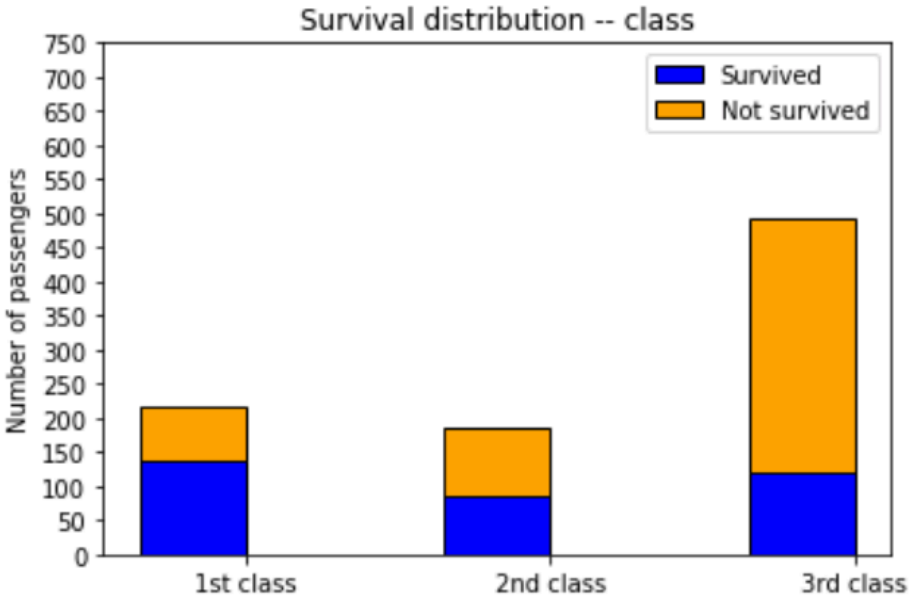
- Age variation



- Most of the passengers in the age-group 19-40 survived.
- Very few survived from age-group 61-80 and also very few survived from 0-18 age groups.

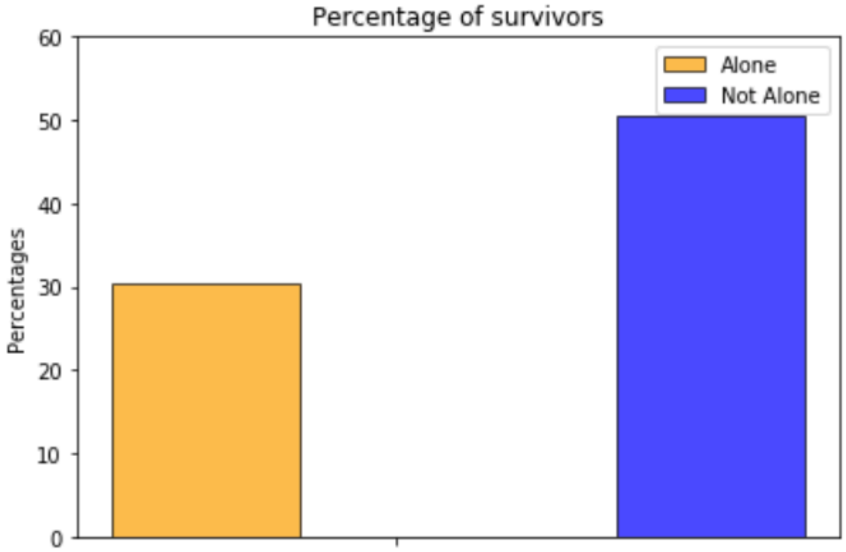
This is an indication of strength and endurance during the crisis. Mostly the middle aged groups were strong enough. Not concluding this. This might be possible reason.

- PClass variation



- The survival rate of a first class passenger is the most.
- But a third class has better survival than 2nd class.

- Passengers without any accompanies vs Passengers with accompanies.



- More chances of survival when you were accompanied by someone, related to family.

6. Conclusions

Overall observations.

- The eldest passenger is aged 80 years whereas the youngest one is 4.2 months old.
- The standard deviation is significantly high to understand that passengers of all age groups were present.
- The mean age is about 30 years.
- Some of the passengers didn't pay any fare at all!

Factors which affected survival.

- Survival rate was more in females, mainly because men took risk protecting them. Maybe.
- Most of the passengers in the age-group 19-40 survived.

- Very few survived from age-group 61-80 and also very few survived from 0-18 age groups.
- This is an indication of strength and endurance during the crisis. Mostly the middle aged groups were strong enough. Not concluding this. This might be possible reason.
- The survival rate of a first class passenger is the most.
- But a third class has better survival than 2nd class.

Survival based on accompany of the passengers. (Alone or Not Alone)

- More chances of survival when you were accompanied by someone, related to family.
- This indicates how accompany is important morally to be brave enough to survive during such times.
- Those who travelled alone probably couldn't reach out to others.
- So travelling with the family ups your chances of survival as they can have better understanding on how to escape from the havoc.

Major limitations while analyzing this dataset

- It was filled with some missing values, so it would have been better with full data.
- I was particularly interested to understand how cabin data could affect survival, unfortunately not sufficient info.
- I think there's a possibility to predict those missing values. If we had ALL the passenger data. In the actual RMS Titanic, there were 2,224 people. Probably access to the whole dataset will help us predict missing values too! 891 rows aren't sufficient. We can use ML and correlation techniques to predict such missing value data based on similar features. This might not lead to correct info but it will definitely give better answer to the questions posed.
- Though most features have a strong data value proposition, more variables which are fuzzy like Tickets and Fare values could also have been analyzed. For example :- Understanding how ticket numbers were generated and grouping them based on a pattern, and then finding its correlation with the fare values. This would have been a better analysis. So ticket values could have been more meaningfully concrete.