Intro to Programming Stage 5 Submission

Titanic Data Analysis Summary

Dataset: Titanic_data.csv

Files included in this submission folder:

1- First submission folder (If needed)
2- Python files
    a. percentage_of_survivors.py
    b. gender_analysis.py
    c. np_age_groups.py
    d. passenger_class.py
    e. correlations.py
3- A pdf file: titanic_data_analysis.pdf

Agenda

- ❖ Brief Description of the Analysis
- ❖ Questions
- ❖ Data Wrangling
- ❖ Analyzing Survival data
- ❖ Analyzing Gender and survival variables
- ❖ Analyzing Age and survival variables
- ❖ Analyzing social status (Fare and Pclass) and survival
- ❖ Conclusion (Findings and limitations)
- ❖ References

Brief Description of the Analysis

In this second submission for stage 5 of 'Intro to Programming Nanodegree' I analyze the dataset 'tatanic_data.csv' by applying single variable analysis and multiple variable analysis using both numpy and pandas.

The analysis utilizes matplotlib to visualize the findings from the analysis.

The data wrangling involved in this analysis is limited to handling missing values Additional data wrangling could have been applied but are out of scope of this analysis (i.e. creating a 'FamilyName' field from the 'Name' field)

The analysis will start by posing questions and discussing limitations before indulging to analyzing each variable and the relation between multiple variables.

Finally, In the summary I will try to provide my own subjective conclusion based on the findings of the analysis.

Questions

1- What is the percentage of survivors among all passengers?
2- What is the percentage of Female passenger vs. Male passengers?
3- What is the percentage of Female survivors vs. Male survivors?
4- What is the survival rate for Females vs. Males?
5- How many missing age values are there? What percentage do they constitute?
6- How are the passengers distributed by age groups?
7- How are the survivors distributed by age groups?
8- What is the survival rate for each age group?
9- How are the passengers distributed by passenger class?
10-How are survivors distributed by passenger class?
11-What is the survival rate for each passenger class?
12-Is there any relation between fare and class?
13-Is there any relation between survival and fare?
14-Is there any relation between survival and class?
15- Is there any relation between survival and age?

Data Wrangling

The dataset titanic_data.csv has missing values in the following fields:

1- Age
2- Cabin
3- Embarked

The latter two variables are directly involved in the analysis, while 19.87% of the values under Age field are missing.

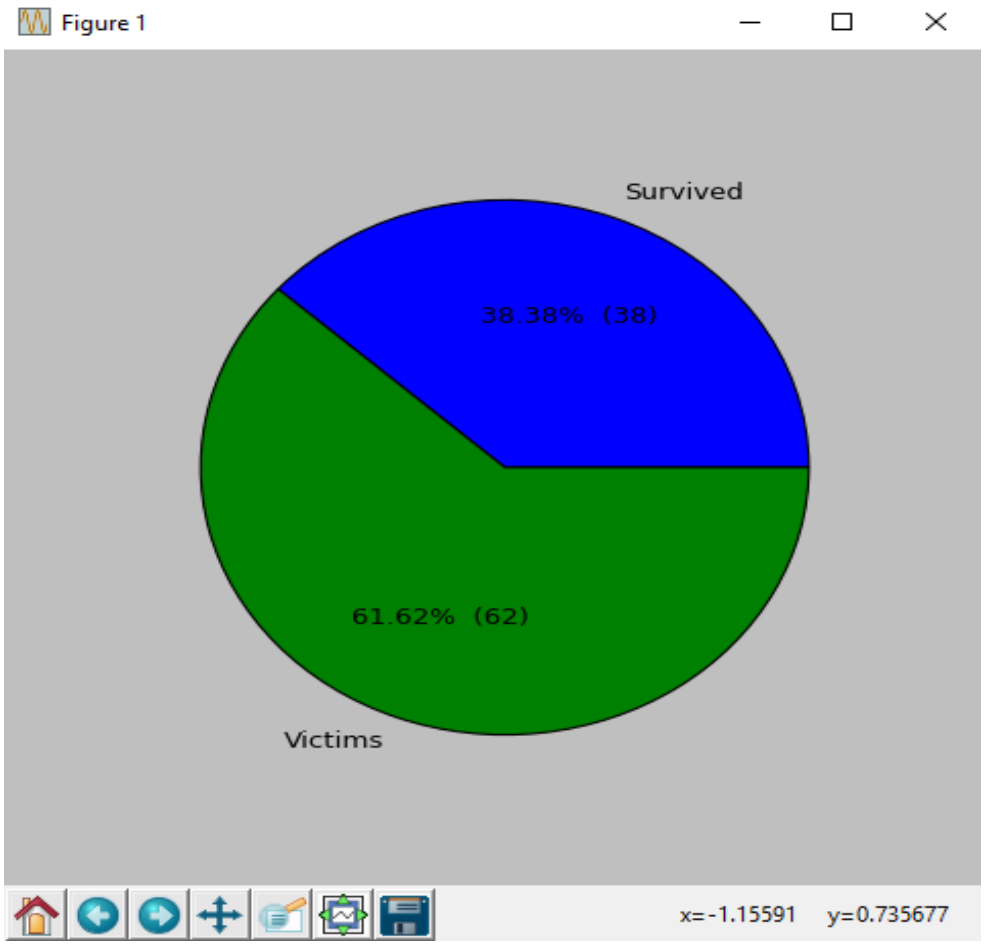
The technique applied to handle missing age values is assigning ‘0’ to each missing field. Since I thought getting the average age is not crucial for my analysis, replacing Nans with zeros had no drawbacks. Otherwise, the other option would have been to use pandas drop.na (as applied in first submission: average_age.py)

Additional details regarding the missing age values will be mentioned when analyzing Age variable.

Analysis of Survival Data

Python file: percentage_of_suvivors.py

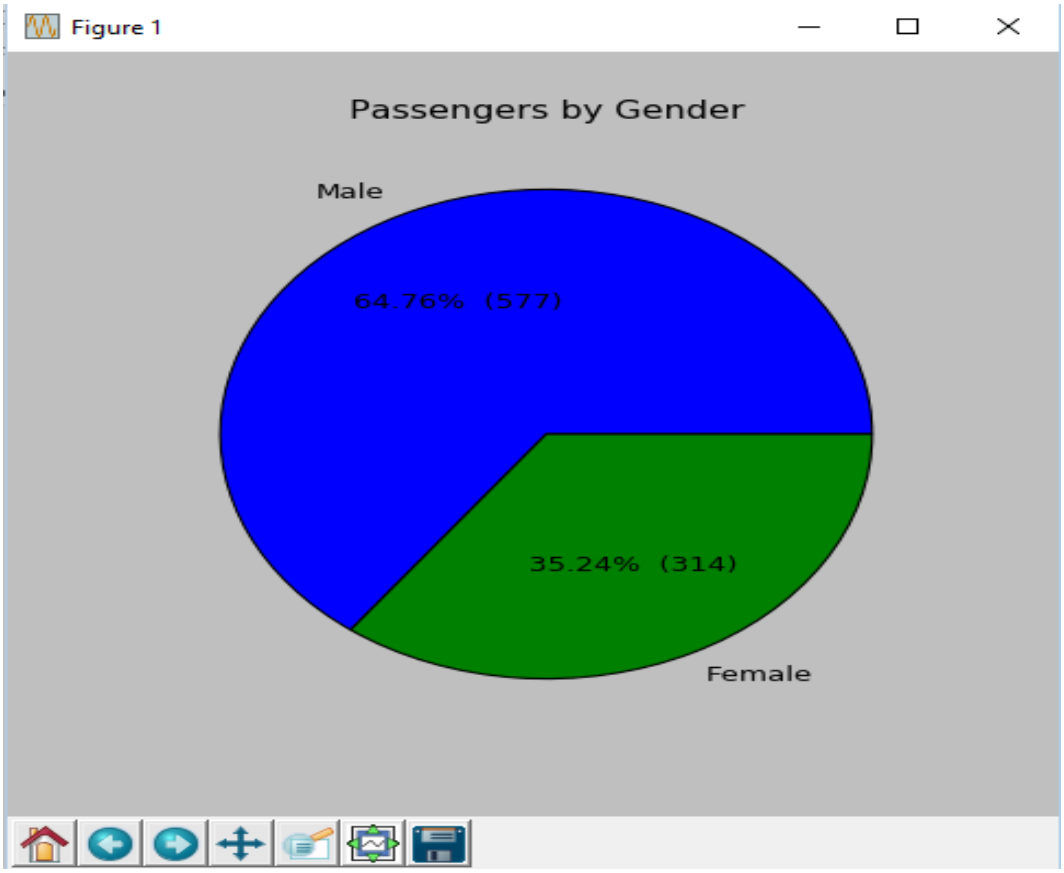
The analysis of survival data of titanic_data.csv dataset shows that 38.38% of the passengers have survived the accident while 61.62% didn't.



Analysis of Gender Data

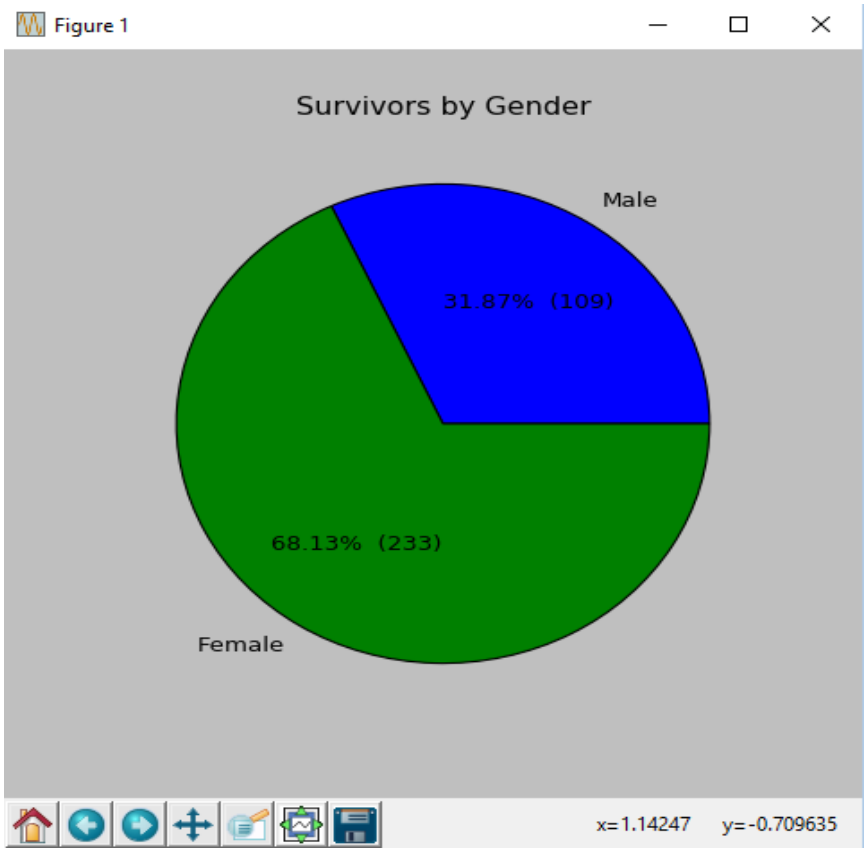
Python file: gender_analysis.py

The gender data in the dataset shows that females constitute 35.24% of total passengers while males constitute 64.76% of the passengers.



Next step in analyzing gender data is finding out the percentage of females and male among survivors. This will help us discover if there is any relation between gender the survival chances.

Analyzing the sex of the passengers who have survived shows that 68.13% of survivors are females while 31.87% are males.



To get a deeper insight into we look at the recorded survival rate for each gender which shows that 74.2% of females have survived while 18.89% of males have survived as shown in the graph below.

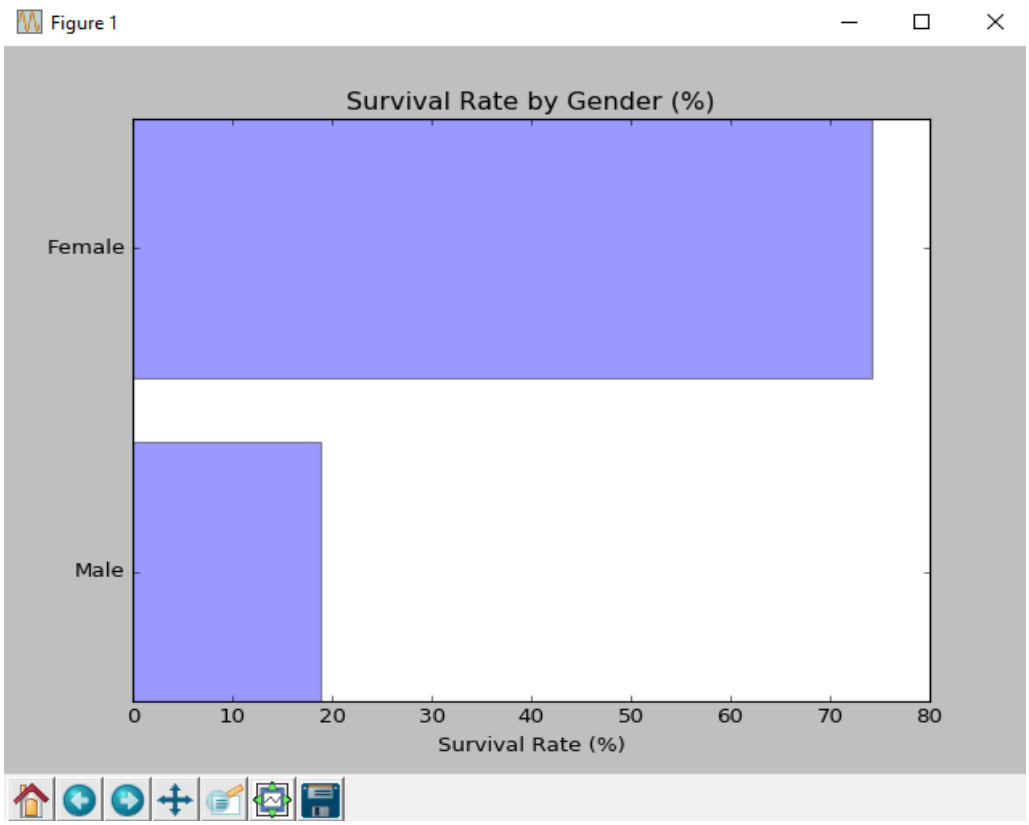


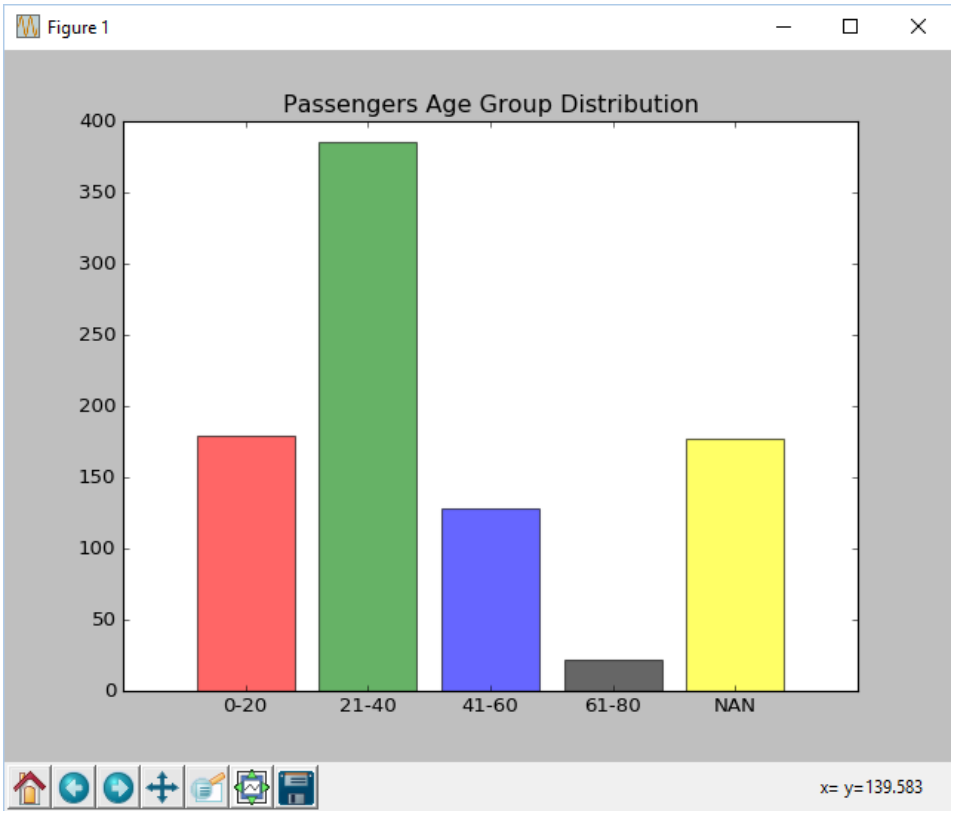
Figure 1
Survival Rate by Gender (%)
Female
Male
0
10
20
30
40
50
60
70
80
Survival Rate (%)

Anlalysis of Age data

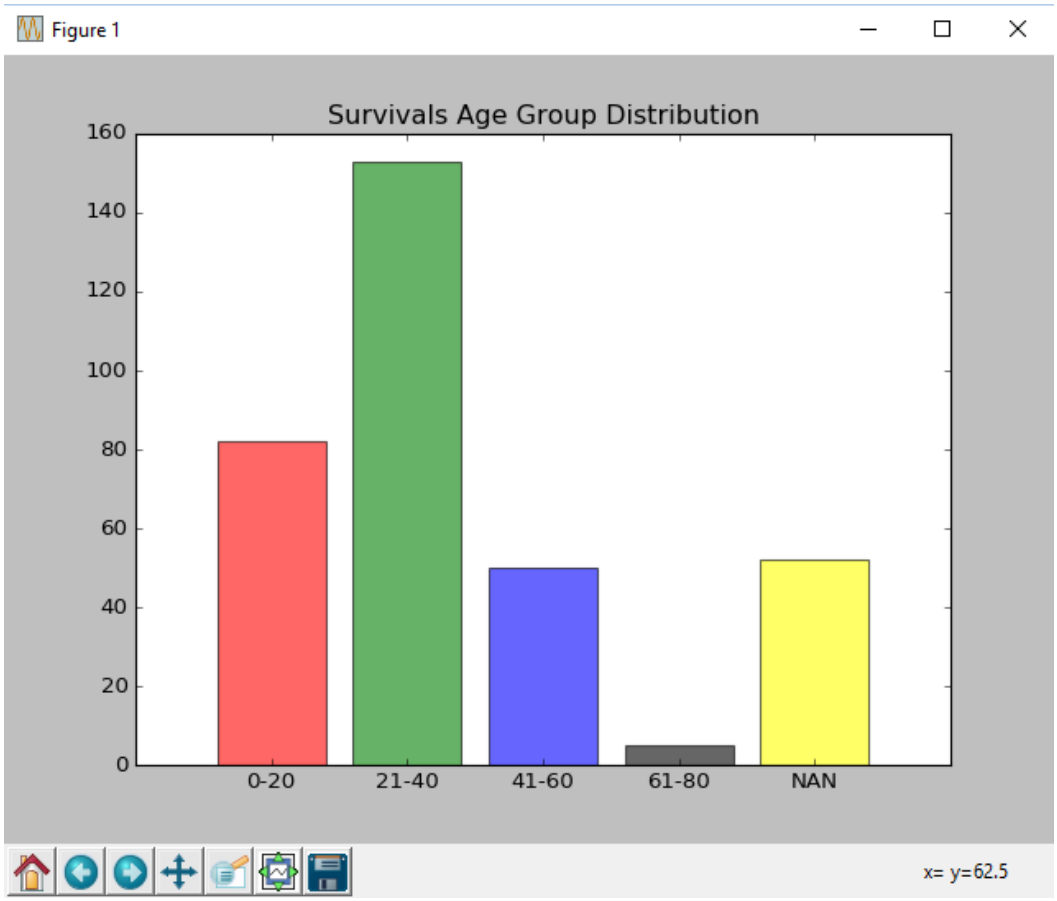
Python File: np_age_groups.py

Analyzing age data is a bit different since not all passengers in the dataset have age data provided, so we need to investigate to find out how many missing age values are there and what percentage they constitute from the 891 available passenger data.

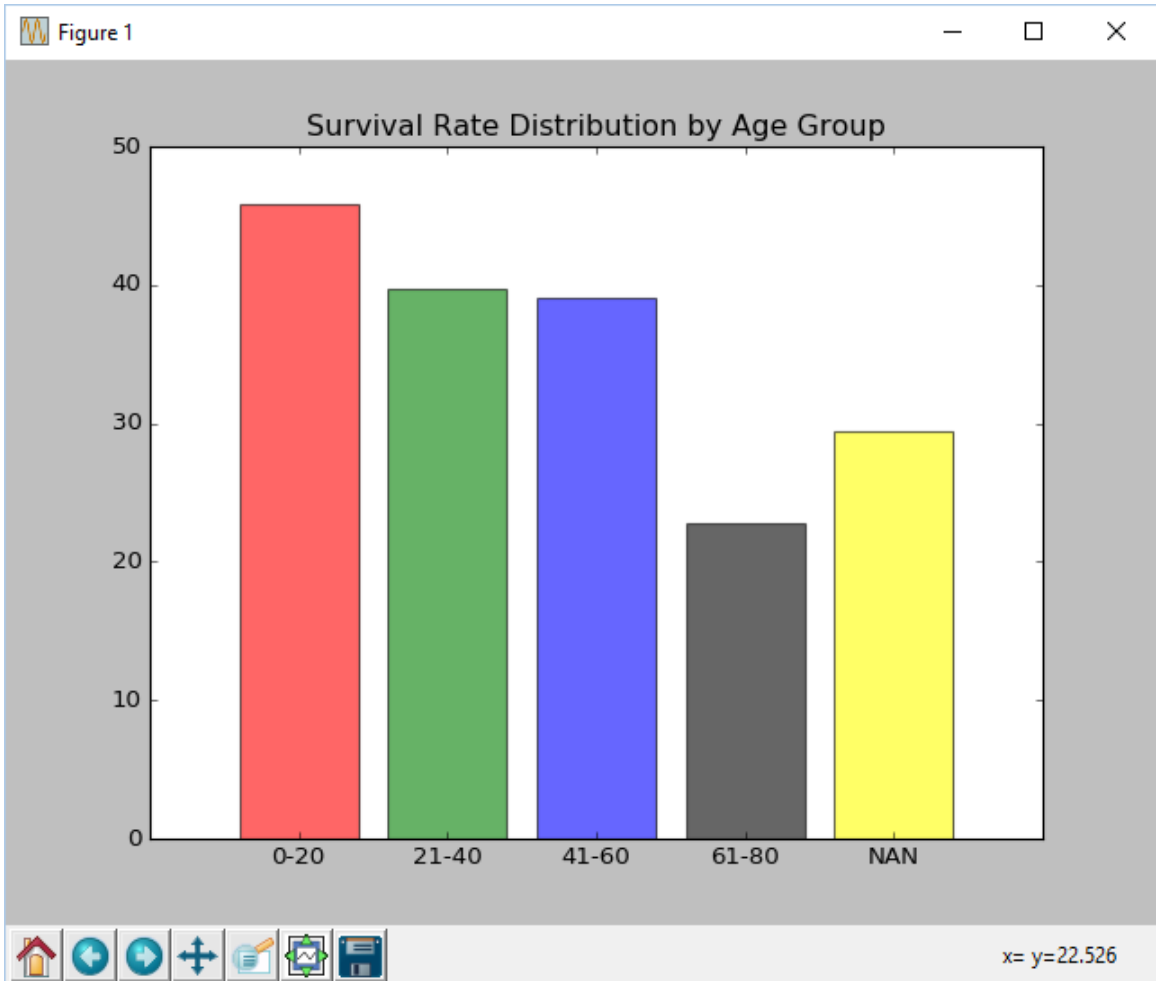
Analyzing missing values reveals that the number of missing age values are 177(19.87%) of which 52(5.85%) are survivals. The below bar chart shows the distribution of passengers by age group. We find that passengers in the 21-40 years intervals constitute the majority of available age data while passenger in the 61-80 years age group are a minority.



When it comes to the distribution of surviving passengers by their age group we are faced by a similar data trend as the below bar chart shows which does not allow us to reach any conclusion that indicates that a certain age group had better chances for survival.



To further analyze this we look at survival rates for each age group to find that survival rate amongst age group 0-20 is the highest while it the lowest for age group 61-80 as revealed in the below chart.

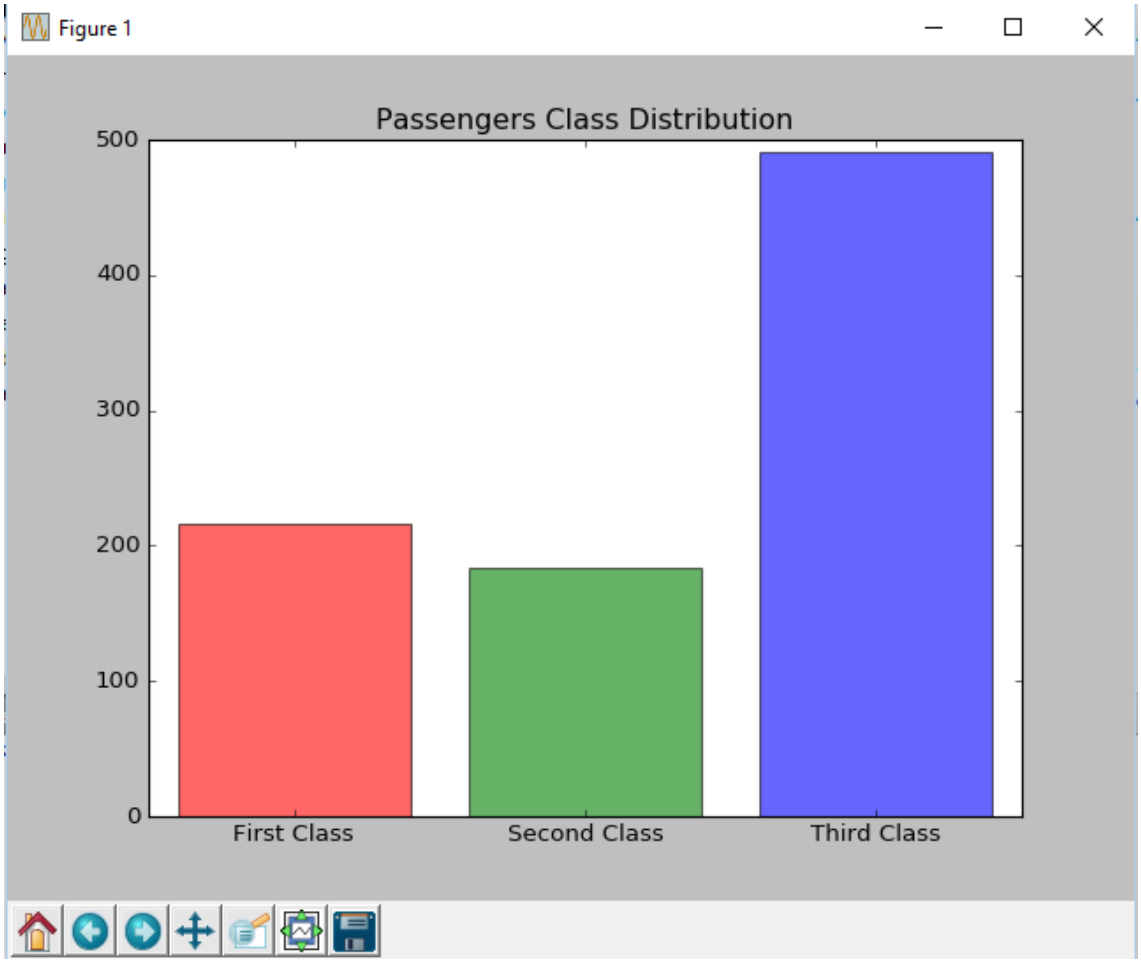


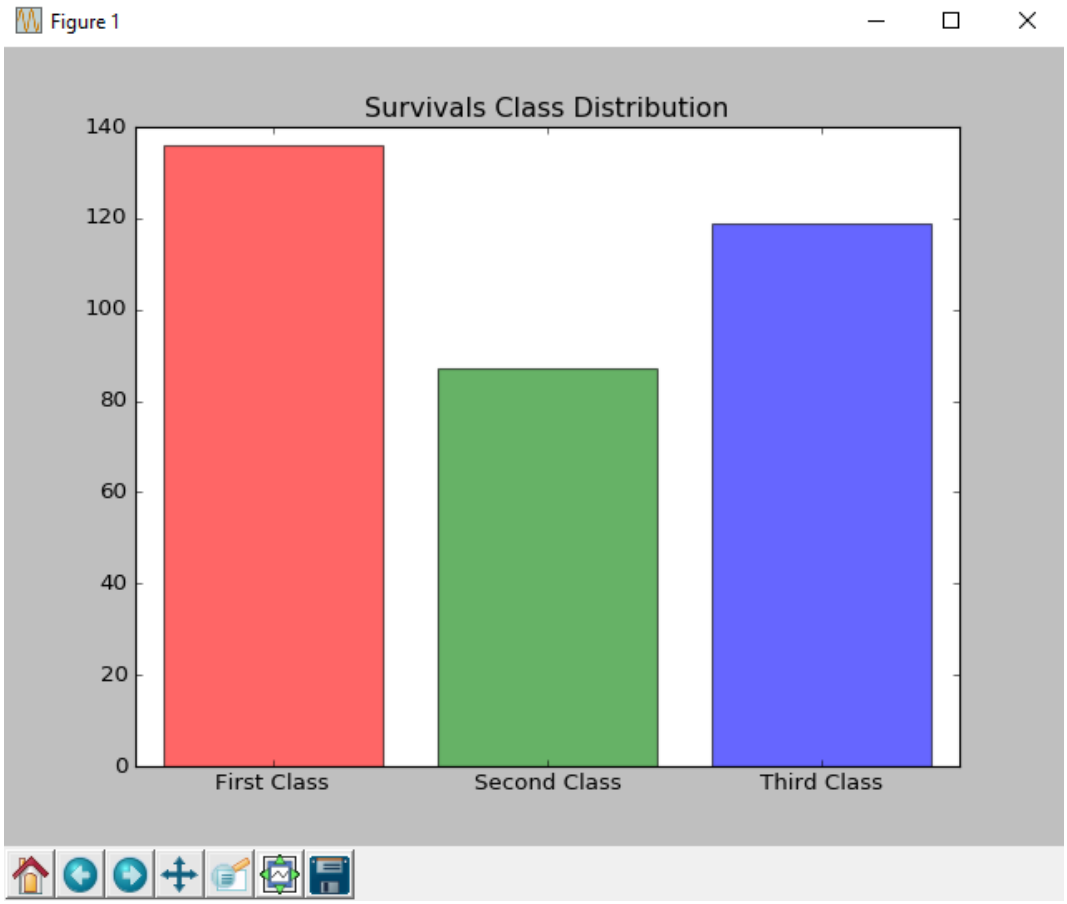
Anlaysis of Social Status Data

Python File: passneger_class.py

Given the strong and expected correlation between fare and class as will be shown in the next section, we can assume that class data can be an indication of social status.

The analysis of passenger class shows that most of the passengers are third class passengers, while the distribution of survivors by passenger class shows that the majority of survivors are first class passengers.





The survival rate among each class reveals a clear inclination towards higher classes when it comes to survival chances. First class passengers have the highest survival rate and third class passengers have the lowest survival rate, while the second class comes in between as shown in the graph below.

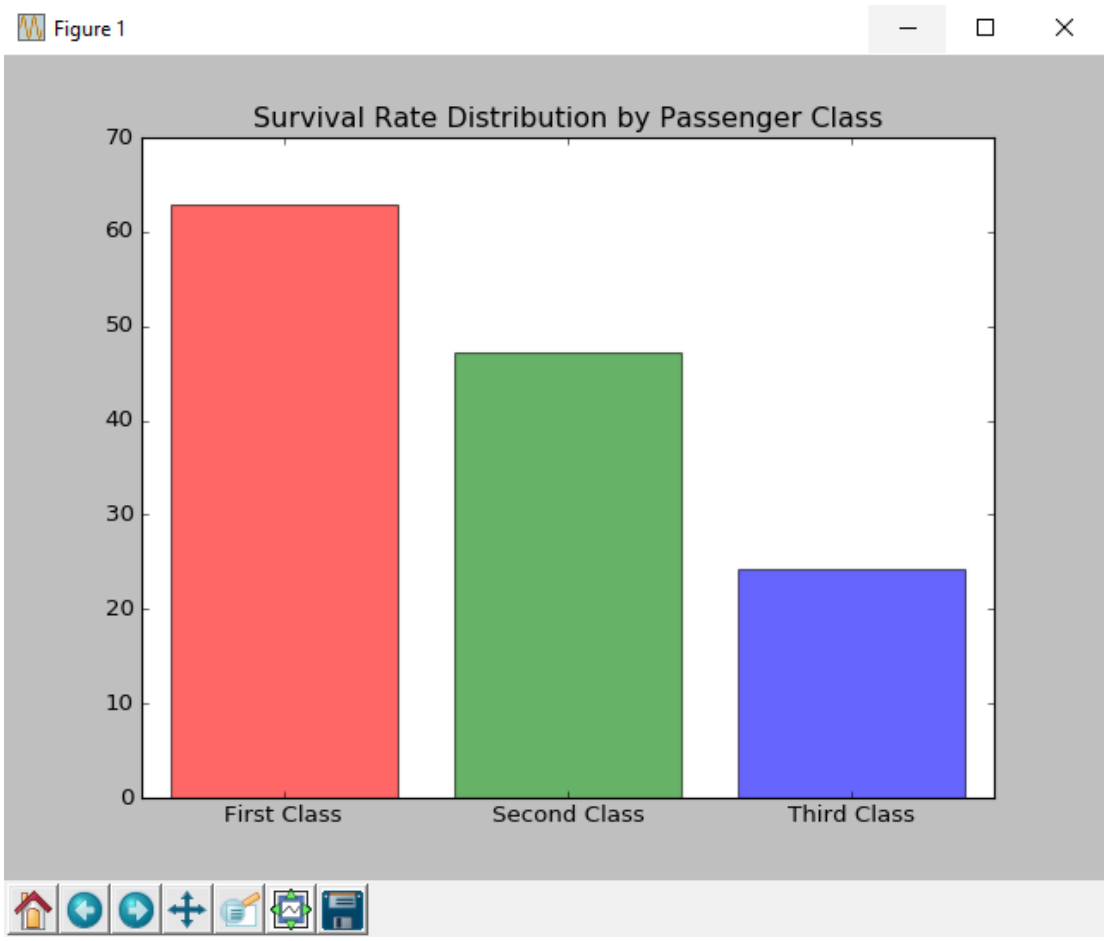


Figure 1
Survival Rate Distribution by Passenger Class
70
60
50
40
30
20
10
0
First Class
Second Class
Third Class

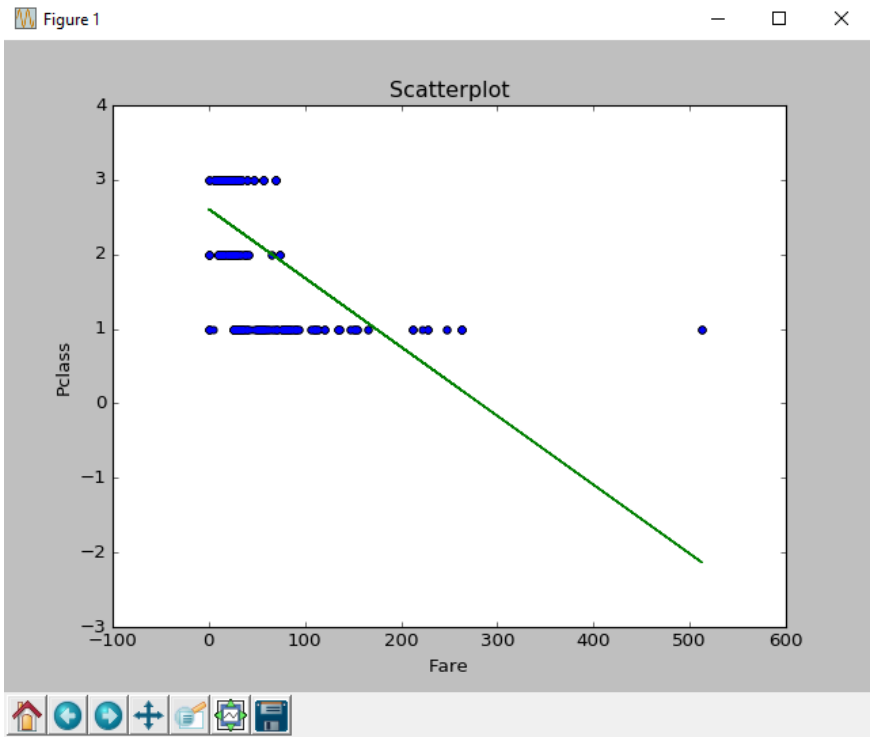
Correlation Analysis

Python Files: correlations.py

Correlation analysis have been applied to test the following relations:

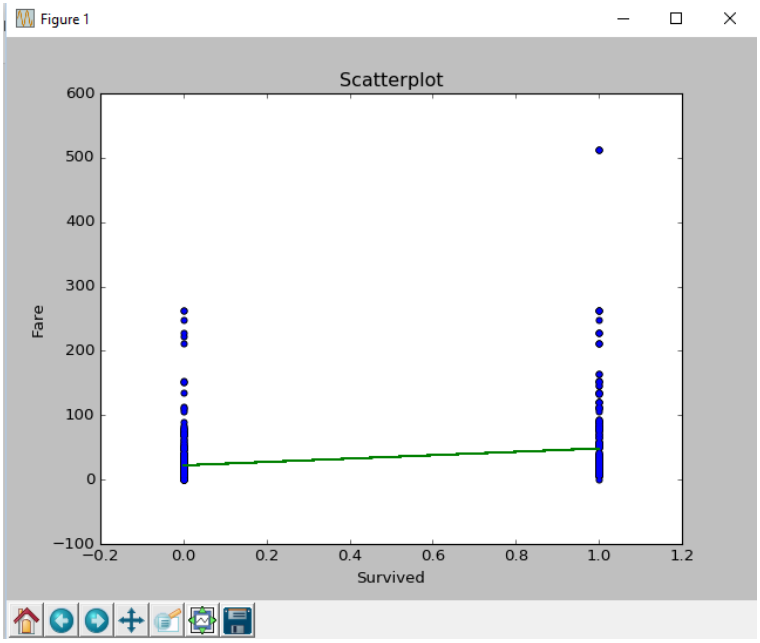
1- Fare and passenger class
2- Survival and fare
3- Survival and passenger class
4- Survival and age

To confirm the conventional relationship between fare and passenger class we applied a correlation analysis. The resulting correlation is -0.549 which indicate a strong negative correlation between fare and class, such that the higher the fare the lower is the passenger class number(Where lower passenger class number implies higher social status since passenger class 1 is first class, 2 is second class, and 3 is third class)

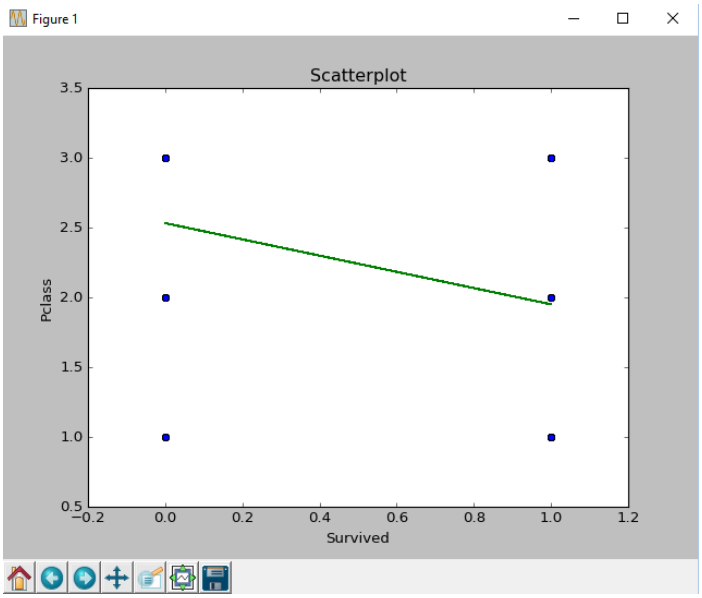


After confirming the relation between fare and class we test the correlation between each of the two variable in one side and survival in the other.

The correlation between fare and survival is 0.257 which is considerable but not strong enough to jump to a conclusion based on that.

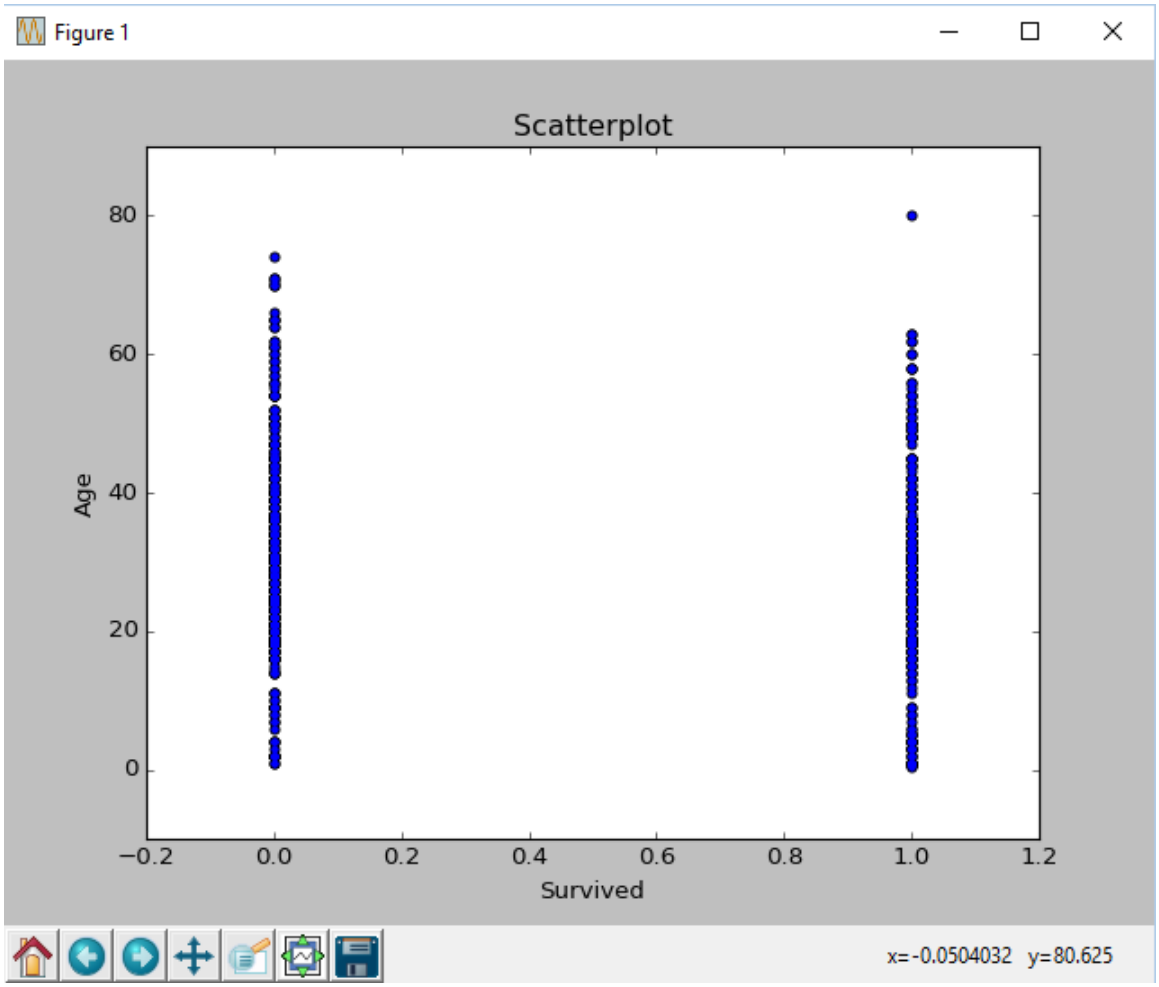


On the other hand, the correlation between class and survival is -0.338 which is a stroger correlation that the one existing between fare and survival as the scatter plots reveal.



The analysis that have been done on age data could not reveal any obvious relation between age and survival so we test the correlation between the 2 variable to check if any meaninful correlation can be detected.

The correlation analysis between age and survival gave a result which is negative 0.077. The very low correlation between age and survival will prevent us from reaching any additional insights from what we had before.



Summary

In the above analysis of the titanic data we investigated several factors and explored what impact each factor has on passenger fate in terms of survivorship.

As we explored gender data we found that the survival rates for female passengers is significantly higher than male passenger in such a way that the survival chances for female passengers is more than double that of male passengers. This might lead us to a conclusion that rescue efforts were aimed at evacuating female passengers before male passenger.

The above leads us to investigate the age factor and whether any age group

Had better survival rates than others. The convention is that rescue efforts makes a priority to save women, children and elderly before the rest of passengers. We Investigated survival data among 4 age groups of 20-year intervals, and kept record of the missing age values which constitute around 20% of total passengers.

The analysis showed that the youngest age group (newly born till 20 years old) had higher survival rate than other age groups with a little less the 50% surviving. In the other hand, passengers who are older than 60 years had the lowest survival rate. This would support the notion mentioned above regarding the priorities of rescue missions when it comes to children but not when it comes to elderly. This could be attributed to the fact that elderly people have less ability to adapt with extreme conditions and are highly subject to sudden health failures. Yet, this conclusion cannot be confirmed solely from the analysis performed. To further investigate the relation between age and survival we performed a correlation analysis which in turn did not reveal any significant correlation, thus investigating

the age variable did not allow us to confidently reach a theory regarding how passengers' age has affected their survival chances.

The next step was investigating the relationship between social status and survival. The two variables that help us identify the social status of passengers are passenger class and fare. The analysis of social status data was more consistent than other factors, such as the strong positive relation between social class and survival is evident in all the analysis performed, such as higher class passenger (i.e. first class passengers) have better chances of survival. This was also supported by the correlation analysis performed between passenger class and surviving, and between fare and surviving.

Limitations:

As mentioned more than once in this paper the fact that the dataset is missing around 20% of age data made it more difficult to draw any solid conclusions. The dataset has missing values in other variable yet those were not part of the analysis.

When it comes to fare data, showed that some passengers did not pay any fare (0). The might indicate that those passengers are among the ship staff, which could have helped if a staff Boolean variable staff is present in the dataset. That said, those passengers constitute only 1.6% which makes their impact minimal, yet existing. This was a factor that made the passenger class the more obvious choice when it comes to investigating the relationship between social status and survival.

The description available for the port of embarkation data is not enough to make any meaningful analysis around this variable.

Thank you

References:

http://stackoverflow.com/questions/6170246/how-do-i-use-matplotlib-autopct

http://stackoverflow.com/questions/20995196/python-pandas-counting-and-summing-specific-conditions

http://stackoverflow.com/questions/5124376/convert-nan-value-to-zero

http://stackoverflow.com/questions/10660435/pythonic-way-to-create-a-long-multi-line-string

http://stackoverflow.com/questions/9560207/how-to-count-values-in-a-certain-range-in-a-numpy-array

http://stackoverflow.com/questions/18974928/how-to-create-custom-legend-in-matplotlib-based-on-the-value-of-the-barplot

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153382/

http://matplotlib.org/examples/api/barchart_demo.html

http://stackoverflow.com/questions/19068862/how-to-overplot-a-line-on-a-scatter-plot-in-python