P2: Investigate a Dataset

The dataset I chose to work with was the Titanic Dataset.

I chose to tackle this project by asking myself which variables may have had more of an impact on survival rates than others. Consequently, I chose to measure my Dependent Variable, Survival, against three Independent Variables.

- Is there a correlation between Pclass [Independent Variable] & Survival [Dependent Variable] ?
- Is there a correlation between Gender [Independent Variable] & Survival [Dependent Variable] ?
- Is there a correlation between Age [Independent Variable] & Survival [Dependent Variable] ?

I tackled this dataset first by importing the necessary libraries (numpy / pandas / matplotlib / pprint / & pdb -for debugging purposes). I then read the csv file into memory and wrangled the file by choosing the columns I wanted based off of my questions above. The columns I chose to keep were “Passenger Id’, “Survived”, “Pclass”, “Sex” and “Age”.

Next I computed the statistical mean for survivability using hierarchical indexing. I wanted to get a good idea of my data before I dived a bit deeper into the analysis. I thus attached a group by function to my newly designed data frame that allowed me to glimpse at the survivability mean for each pclass, sex and age class.

Moving on, I wanted to examine the dataset for missing and incomplete data (which can influence the outcome of the analysis tremendously). I added an info function to our data frame and found that the only columns that were applicable to us that had any missing values was age (total of 177 missing values). In order to populate the missing ages, I created a function that used the mean age based on the sex and pclass groupings.

Next I created a function called “survivors” that generated an overall sum of the amount of survivors from the sample dataset [342 out of 891] so that I could use this as a benchmark figure for my analyses. I then created three independent data frames with a unique column identifier, each referring to one of the three questions I posed above.

The meat of my analysis came from the three functions I then defined for assessing my data. My function “p_class” allowed me to perform a cross analysis for comparing rates of

survivability with passenger class (which ranged from 1 [referring to first class] to 3 [referring to third class]). My function “sex” allowed me to do a similar cross analysis, however in this case I was comparing survivability with gender (both male & female). Lastly, my function “age” had a similar purpose, comparing survivability with passenger age across the entire sample size. I chose to slice my age data into buckets, ranging from 0 – 3, 4 – 10, & 11 – 17 (taking into account 18 was considered an adult back then). I chose these figures because 0 – 3 represents an infant to most people, 4 – 10 represents a small, helpless child and 11 – 17 a kid who can manage to act independently of their parents but only to a questionable degree. Moreover, the latter two buckets contained the same age count in years (which was 7).

Lastly, I consolidated the return statements to make it easier to read, called the functions and then printed them out.

The following code & data visualizations assisted me in the summary statistics.

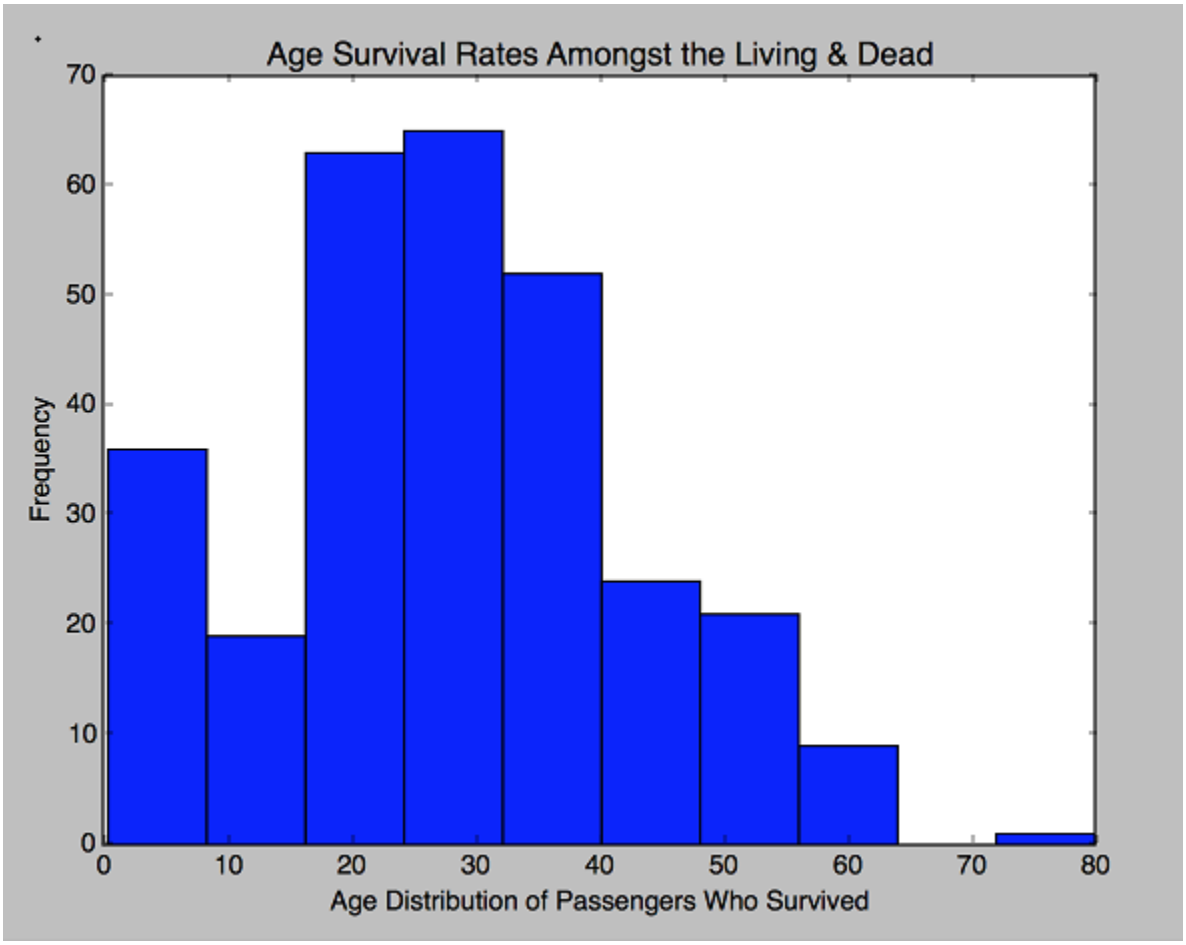
Source Code

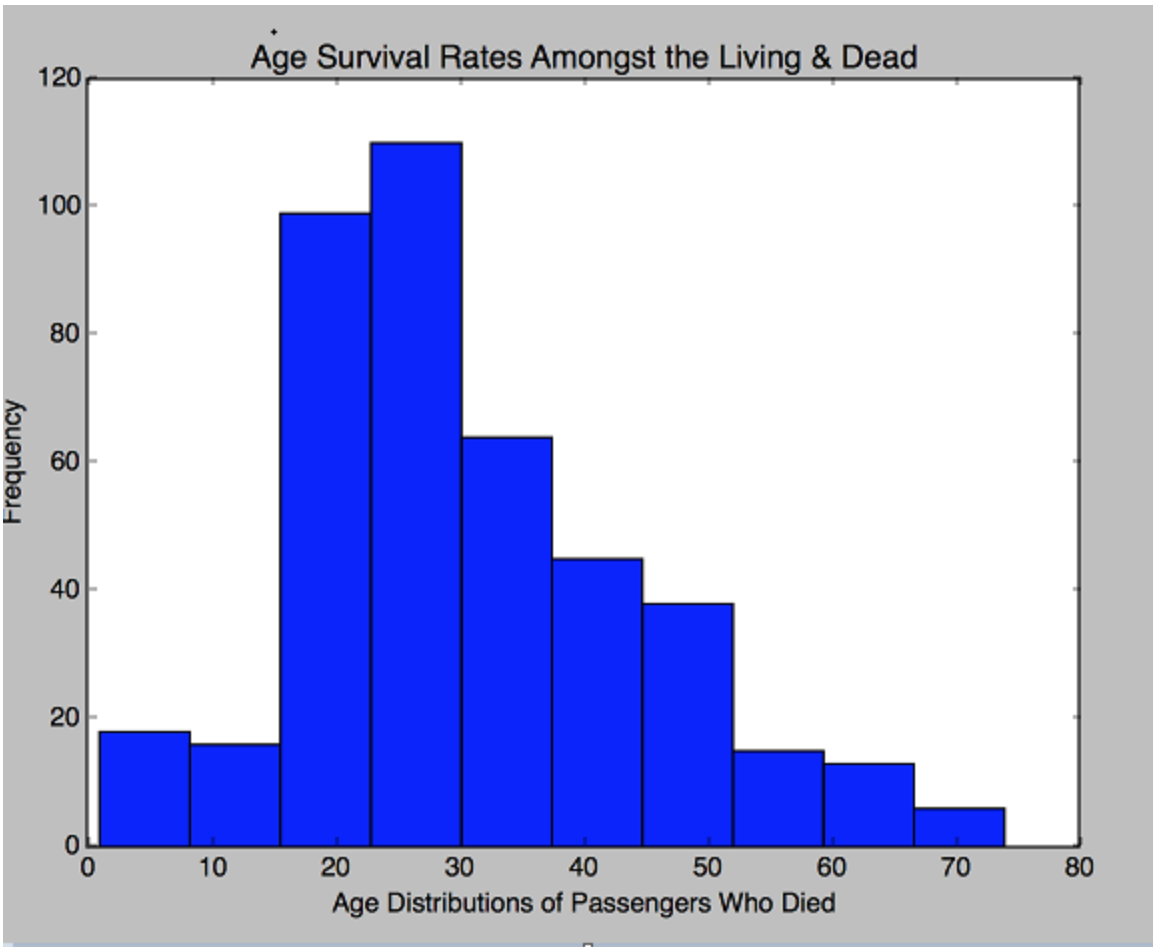
```
# import the relevant libraries

import numpy as np
import pandas as pd
from pandas import Series, DataFrame
import matplotlib.pyplot as plt
import pprint
import pdb


# read csv file into memory & sort columns by columns desired

df1 = pd.read_csv('titanic_raw_data.csv')
keep_cols = ["PassengerId", "Survived", "Pclass", "Sex", "Age"]
new_df = df1[keep_cols]
pprint.pprint(new_df)


# compute statistical mean for survivability using heirarchical indexing

stats_df = new_df.groupby(['Pclass', 'Sex', 'Age'])[['Survived']].mean()
print(stats_df)


# find missing & incomplete data

print(new_df.info())

# find mean age based on sex & pclass

missing_ages = new_df[new_df['Age'].isnull()]
mean_ages = new_df.groupby(['Pclass', 'Sex'])['Age'].mean()
```

```
def remove_na_ages(row):
    if pd.isnull(row['Age']):
        return mean_ages[row['Pclass'], row['Sex']]
    else:
        return row['Age']

new_df['Age'] = new_df.apply(remove_na_ages, axis=1)

print(new_df['Age'])



# create a function for total # of survivors

def survivors():
    survivors = new_df["Survived"].sum()
    return survivors

survivors = survivors()
print("\tThe amount of people who survived were " + str(survivors) + " people.")


# create a function that compares rate of survival with class


def p_class(new_df):
    grouped_by_p_class = new_df.groupby((['Pclass'])['Survived' == 1]).sum()
    return grouped_by_p_class


# create a function that compares rate of survival with sex

def sex(new_df):
    grouped_by_sex = new_df.groupby((['Sex'])['Survived' == 1]).sum()
    return grouped_by_sex



# create a function that compares rate of survival with age

def age(new_df):
    grouped_by_age_young = new_df.groupby([new_df['Age'] <= 3]).sum()['Survived'][1]
    grouped_by_age_middle = new_df.groupby([(new_df['Age'] >= 4) & (new_df['Age'] <=
10)]).sum()['Survived'][1]
    grouped_by_age_old = new_df.groupby([(new_df['Age'] >= 11) & (new_df['Age'] <= 17)]).sum()['Survived'][1]

    return (grouped_by_age_young, grouped_by_age_middle, grouped_by_age_old)


grouped_total_age = ['grouped_by_age_young', 'grouped_by_age_middle', 'grouped_by_age_old']

# call the functions & print

grouped_by_p_class = p_class(new_df)
grouped_by_sex = sex(new_df)
grouped_total_age = age(new_df)


print(grouped_by_p_class, grouped_by_sex, grouped_total_age)
```

Plots 1& 2:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df1 = pd.read_csv('titanic_raw_data.csv')
keep_cols = ['Age', 'Survived']
new_df = df1[keep_cols]

survived = new_df[new_df['Survived'] == 1]
dead = new_df[new_df['Survived'] == 0]

def hist(column_data, xlabel):
    plt.xlabel(xlabel)
    plt.ylabel('Frequency')
    plt.title('Age Survival Rates Amongst the Living & Dead')
    plt.hist(column_data)
    plt.show()

hist(survived['Age'].dropna(), 'Age Distribution of Passengers Who Survived')
hist(dead['Age'].dropna(), 'Age Distributions of Passengers Who Died')
```

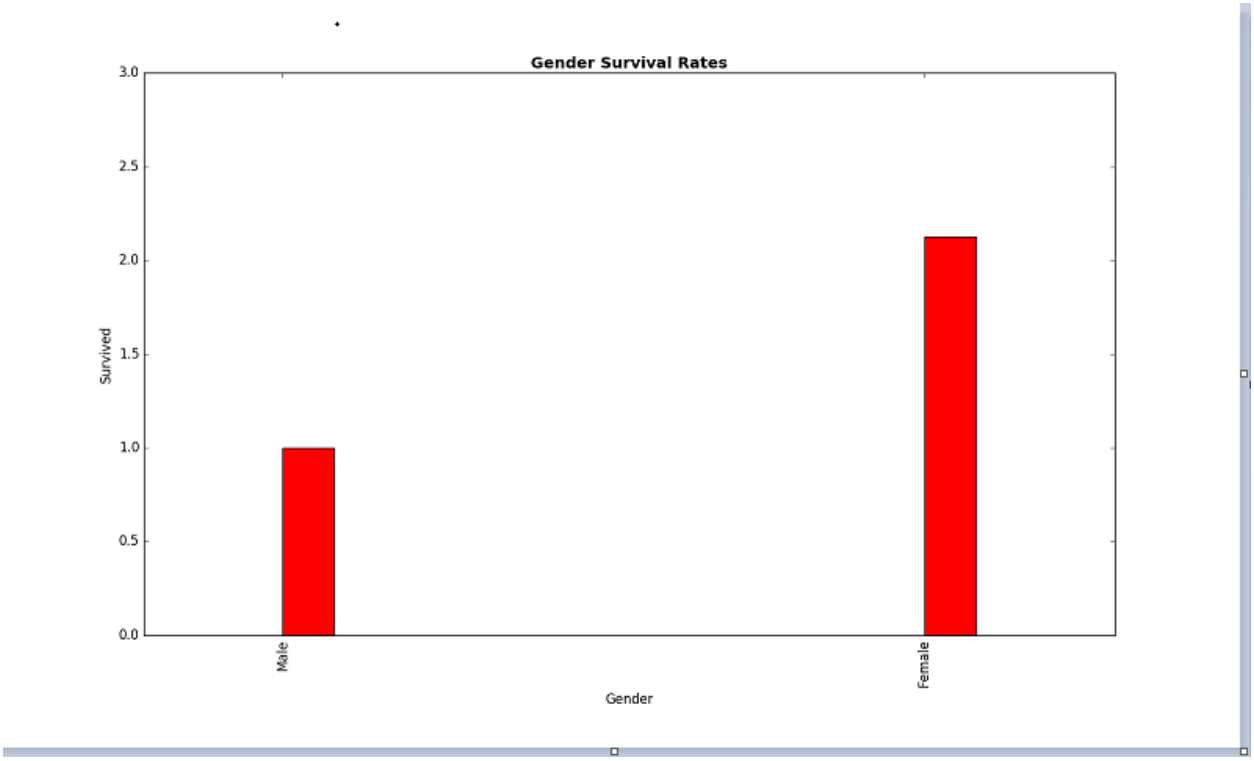




Plot 3:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df1 = pd.read_csv('titanic_raw_data.csv')
keep_cols = ["PassengerId", "Survived", "Pclass", "Sex", "Age"]
new_df = df1[keep_cols]

men = new_df.groupby([new_df['Sex'] == 'male']).sum()['Survived'][1]
women = new_df.groupby([new_df['Sex'] == 'female']).sum()['Survived'][1]
gender = np.array([men, women])

plt.xlabel('Gender')
plt.ylabel('Survived')
plt.title('Gender Survival Rates', fontweight='bold')
labels = ['Male', 'Female']
plt.xticks(gender, labels, rotation='vertical')
plt.ylim(0, 3)
plt.margins(0.2)
plt.subplots_adjust(bottom=0.15)
plt.bar(men, height=1, width=10, color='red')
```

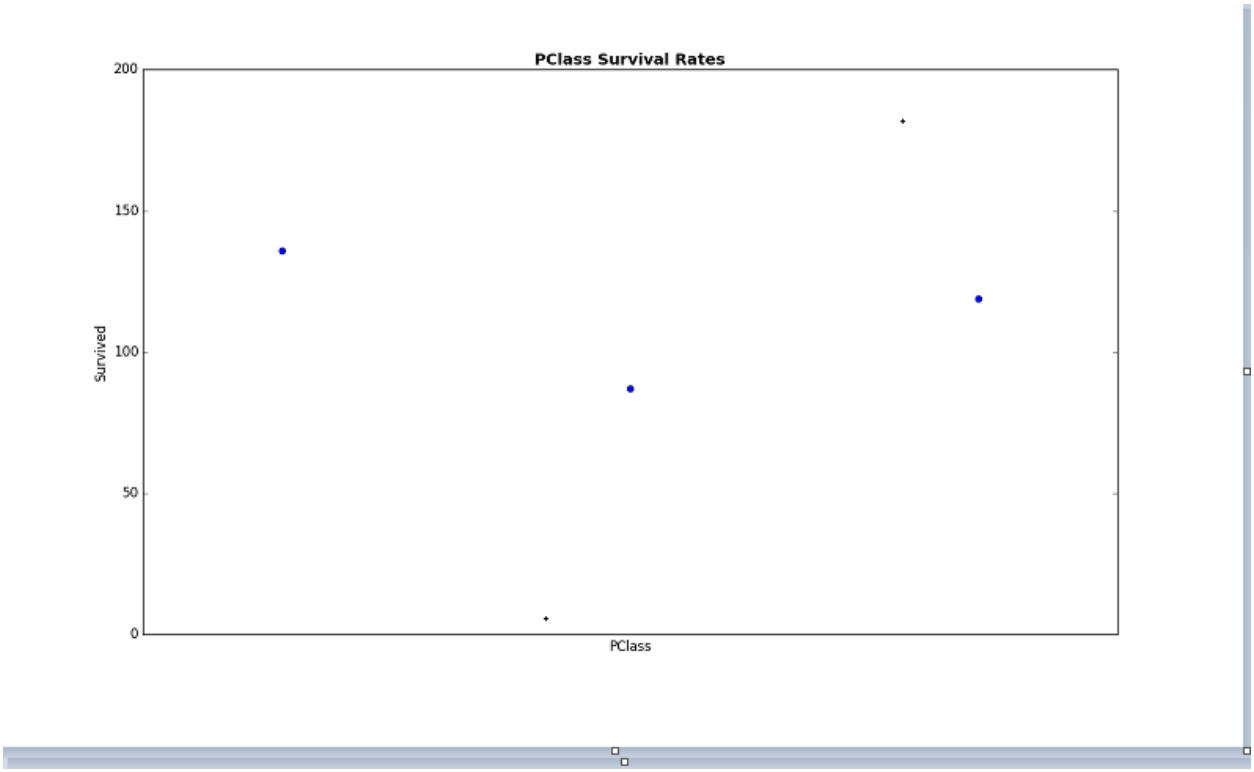
```
plt.bar(women, height=2.125, width=10, color='red')
plt.show()
```



Plot 4:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df1 = pd.read_csv('titanic_raw_data.csv')
keep_cols = ["PassengerId", "Survived", "Pclass", "Sex", "Age"]
new_df = df1[keep_cols]

first_class = new_df.groupby([new_df['Pclass'] == 1]).sum()['Survived'][1]
second_class= new_df.groupby([new_df['Pclass'] == 2]).sum()['Survived'][1]
third_class = new_df.groupby([new_df['Pclass'] == 3]).sum()['Survived'][1]
p_class = np.array([first_class, second_class, third_class])

plt.xlabel('PClass')
plt.ylabel('Survived')
plt.title('PClass Survival Rates', fontweight = 'bold')
labels = ['First Class', 'Second Class', 'Third Class']
plt.xticks(p_class, labels, rotation='vertical')
plt.ylim(0, 200)
plt.margins(0.2)
plt.subplots_adjust(bottom=0.15)
```

```
plt.plot(p_class, linestyle='', marker='o', color='blue')
plt.show()
```



As a result of the accompanying code & data visualizations, the following assumptions could be made from the sample.

Our univariate analysis included two histograms (both moderately bell-shaped) of age survivability rates between the living and the dead for our passengers. Each plot demonstrated that those between ages 20 & 40 had the highest rates of either living or dying. Chances of survival dropped off for both plots once age fell above or below that range. I wonder if this age group had the highest chance of survival because people around this time of their lives are usually at their peak in physical fitness levels, thus contributing towards their better survival rates. On the flip side of this, perhaps they also had the highest chance of dying because when people are young they also tend to be less wealthy, thus many I'm assuming couldn't afford to purchase first class (or even second class tickets), which most likely had easier and swifter access to the lifeboats.

With regards to Pclass, it appeared you had a 63% chance of survival if you were a first class passenger (136 out of 216). You had a 47% chance of survival if you were second class

passenger (87 out of 184). Lastly, you had a 24% chance of survival if you were a third class passenger (119 out of 491). [As a side note, I decided against including any type of connecting linestyle in my plot (connected, dashed, etc.) in order to avoid misleading my readers into thinking the survivors were absolute across Pclass.]

Moving on to gender, passengers who were female had a 74% chance of survival (233 / 314), while passengers who were male had a 19% chance of survival (109 / 577). [My accompanying bar chart demonstrates the difference between male and female survivors as a ratio of total survivors (approx. 2 to 1), NOT as a percentage of their associated sex.].

I wasn't surprised by these findings at all, as women and children are usually the first to be evacuated out of most high stress situations, thus accounting for a majority of the limited lifeboats. Moreover, given that society was rather patriarchal back then, if you were a woman or a child you likely had a higher chance of being in first class as your wealthy husband most likely purchased your tickets for you.

Last but not least comes age. While passengers under the age of 18 only accounted for 17.7% of the total passengers in our sample, there were still some interesting observations. Children in the youngest bucket (3 & under) accounted for 6% of total survivors, while accounting for 33% of survivors (20 out of 61) amongst children 17 and under (despite being the smallest bucket in years). Children in the middle bucket (4 to 10) accounted for 5% of total survivors while accounting for 30% or survivors (18 out of 61), amongst children 17 and under. Finally, children in the oldest bucket (11 – 17) accounted for 6.7% of total survivors, accounting for 38% of survivors (23 out of 61) amongst children 17 and under.

What surprised me about this was that the survivability rate in percentage terms was higher amongst kids in the oldest bucket than that of the other two younger, more vulnerable buckets who I thought would have accounted for a larger majority of survivors (given that there were limited lifeboats on deck). I guess when life or death is on the line, it's a dog eat dog world out there, and a lot of the older kids pushed through the lines in order to snag the limited seats.

As to the limitations of the dataset, the fact that we had missing data definitely skewed our analysis a bit. Our age column had 714 out of 891 non-null values, meaning 177 were missing. (53 female & 124 male). We accounted for this lack of information by generating ages based on the mean of ages, however we are ultimately unsure as to how much this may have skewed our final analysis in the end. Moreover, the size of the sample data could also impact the results as we are unsure if this is a random sample or if the selection of the data is biased or unbiased.

Finally, as with most other datasets, the more information we have, the better it can be analyzed. The dataset does not distinguish between passenger and crew, yet we know that a

mixture of both of these classes survived the boat's collision with the iceberg. I believe if we had this data from the beginning, we would have been able to have yet an even deeper understanding of survivability rates amongst those aboard the infamous RMS Titanic.