Investigating a Dataset

Quality of Analysis

The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

1. Descriptive
    a. Decide what are the main influences of the probability of a passenger surviving the disaster
    b. Understand how to clean the data whilst preserving as much of the other data points as possible
2. Predictive
    a. What is the probability of an individual surviving in the Titanic disaster based off the person's
        i. Passenger Class Ticket
        ii. Sex
        iii. Age
        iv. Number of Siblings/Spouses Aboard
        v. Number of Parents/Children Aboard
        vi. Passenger Fare
        vii. Port of Embarkation
    b. Given information about the 7 variables mentioned above, make a prediction if the passenger will survive the disaster
        i. See if it is possible to extend this to other disasters such as Costa Concordia or RMS Lusitania

I used a logistic regression on the data based off the variables mentioned above

The data is cleaned by creating several new binary variables for each passenger class as well as a binary variable for each of the port of embarkation. These new variables are named 'pclass1', 'pclass2', 'plcass3', 'embarkedC', 'embarkedQ' and 'embarkedS' where it is True if the passenger satisfies the criteria suggested by the variable name.

The y result vector is examined to ensure that it is not extremely skewed through visual verification of the histogram. This is because if the probability of an individual surviving the Titantic disaster is highly skewed (e.g. all individuals died bar a few), a model that predicts all individuals did not survive will have a high accuracy and low cost even though it is rather meaningless. This helps to determine what will be a feasible measurement of performance.

For this instance, cost error is chosen (if a highly skewed sample is witnessed on the histogram, perhaps an f1 test on precision and recall will be more suitable)

Next, a suitable alpha is chosen so as to ensure that the steps taken are not outsized/ too small with regards to the data. This allows the convergence of the gradient descent algorithm as well as to ensure that it is relatively quick to reach the local optimum

Then a suitable regularisation lambda is chosen through a similar process where theta are calculated with different lambda regularisation factors and checked against a cross validation set which yield the minimal error.

Data Wrangling Phase

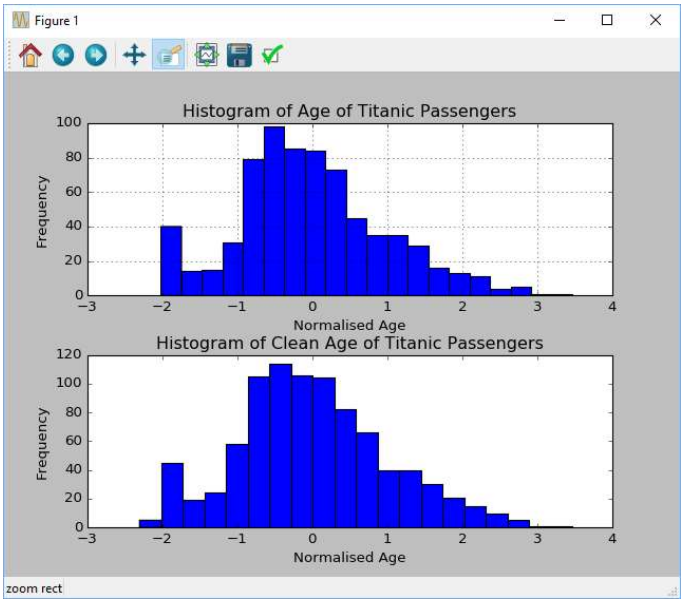
The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

```
In [31]: fhand.count()
Out[31]:
PassengerId     891
Survived        891
Pclass          891
Name            891
Sex             891
Age             714
SibSp           891
Parch           891
Ticket          891
Fare            891
Cabin           204
Embarked        889
dtype: int64
```

Looking at the data set, we noticed that most of the data is complete. However, there are 3 variables in specific that we have incomplete information, namely Age, Cabin and Embarked.

I decide to drop the Cabin variable as less than a third of the data is available and hence I will not be able to make meaningful inferences and the cost of omitting other variables at its expenses is too much to bear

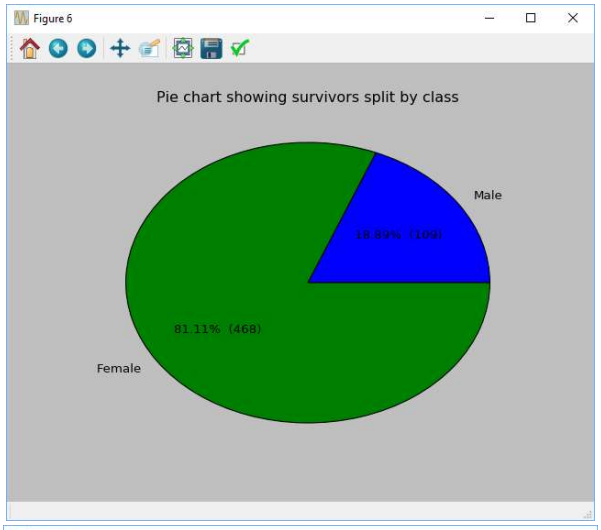
For each of the NaN value in normalised Age, I replaced it with a randomly generated normally distributed variable. As the original distribution has a rather Gaussian-like distribution, I decide to use replace all the NaN values in the normalised Age variable to something that has similar distribution.

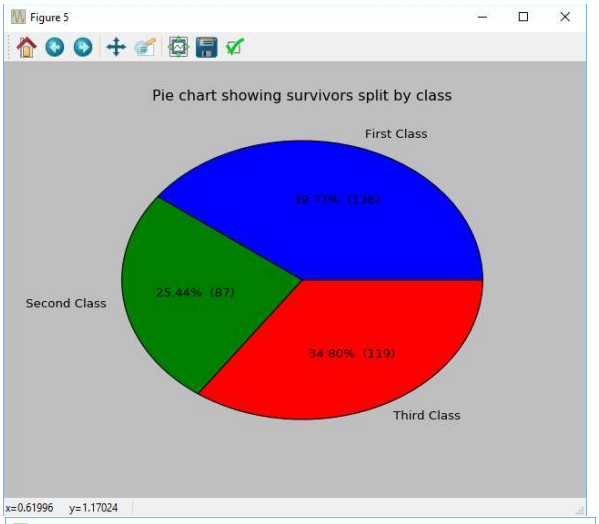


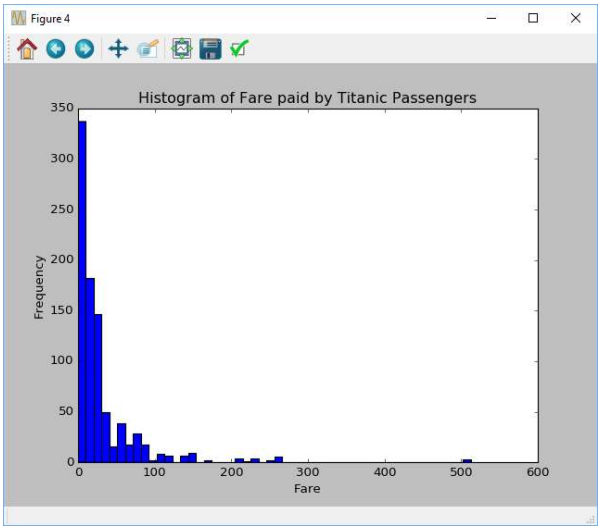
The issue with Port of Embarkation occurs for 2 passengers and given the it is less than 0.5% of the sample, I decide that it will be much easier to remove them from the sample through the use of dropna() than to make any assumptions about their port of embarkation.

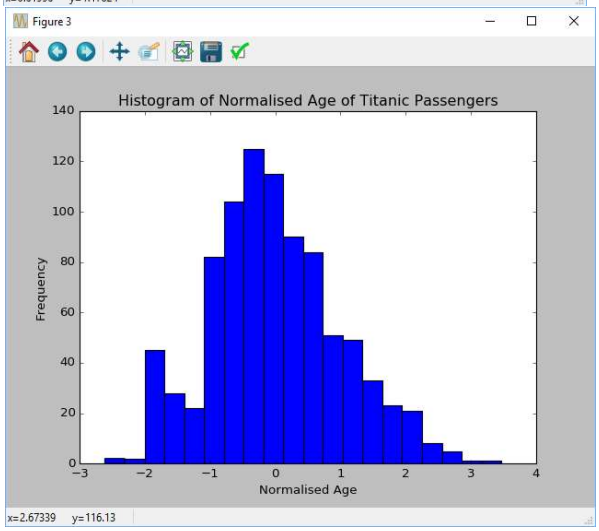
Exploration Phase

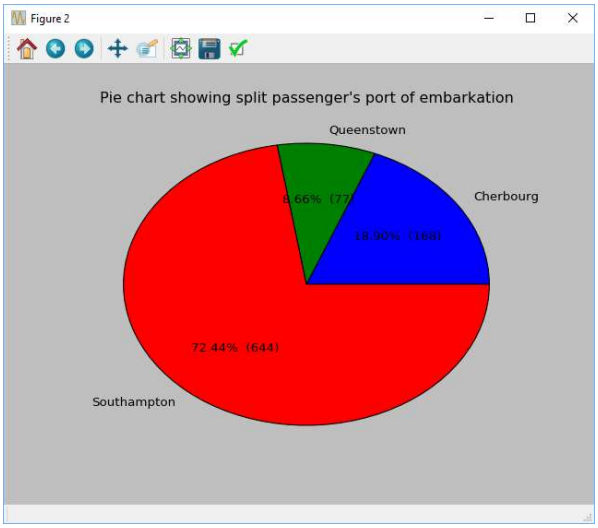
The project investigates the stated question(s) from multiple angles. At least three variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

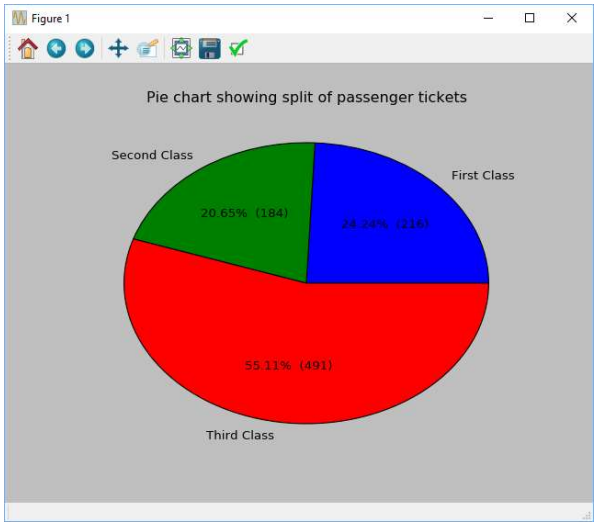












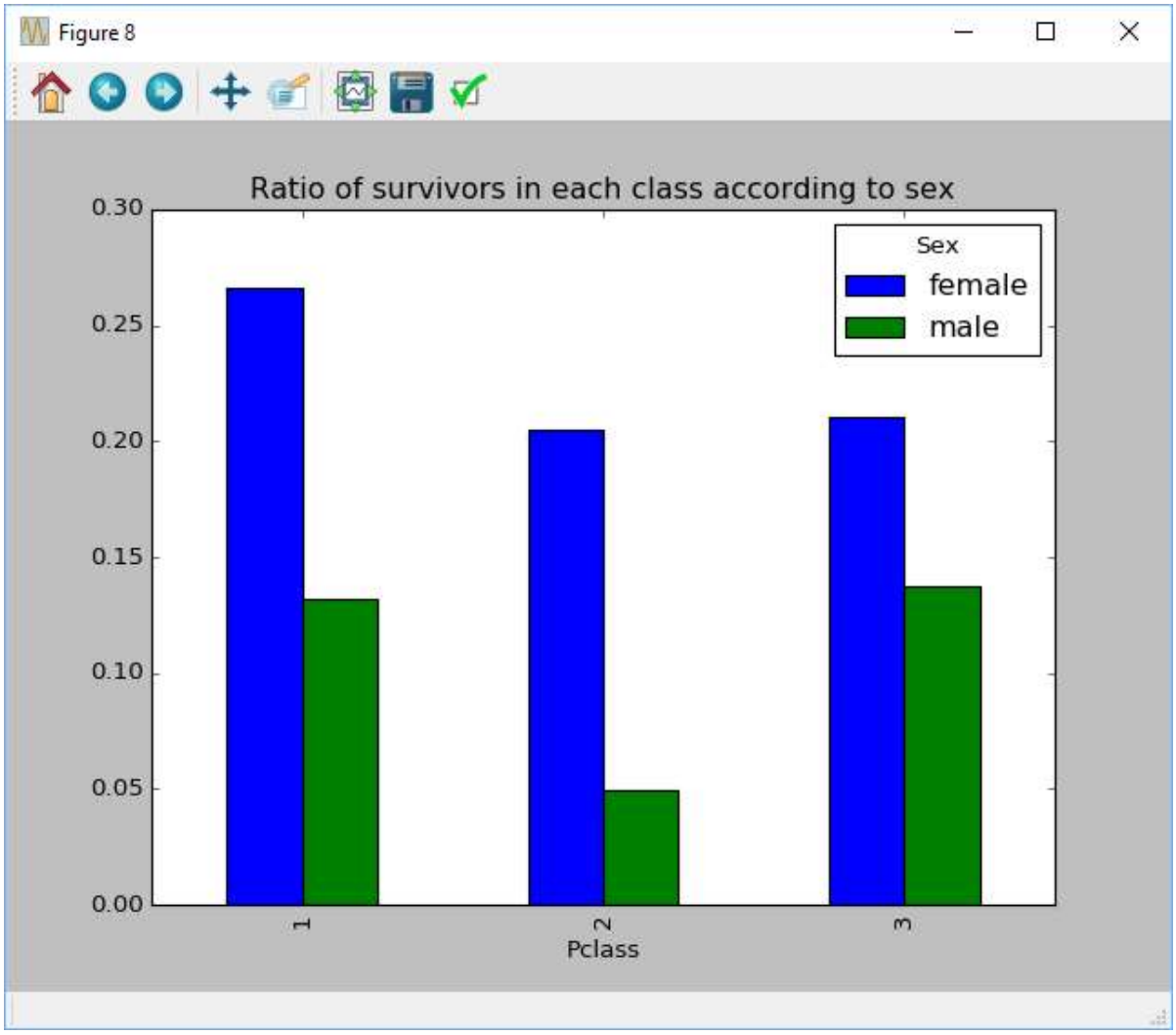


Figure 8
Ratio of survivors in each class according to sex
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Sex
female
male
1
2
3
Pclass

The project's visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data.

At least two kinds of plots should be created as part of the explorations.

Null hypothesis states that the mean probability of survival of a male is the same as the probability of survival of a female passenger
Alternate hypothesis states that the mean probability of survival of a male is different from the probability of a female passenger
$H_0 : \mu_1 = \mu_0$
$H_1 : \mu_1 =/= \mu_0$

The t statistic of the difference in mean probability of male survivors and female survivors is -3.2298 with a degree of freedom of 889

```
In [18]: [tstat,ddof] = ttest(X,'sex','survived')

In [19]: print[tstat,ddof]
[-3.229803384904498, 889]
```

As t_{stat} is less than that of $t_{critical}$ (99.5% with degree of freedom of 1000 = -2.813), we will reject the null hypothesis and can conclude that the probability of survival for a male passenger is significantly lower than that of a female

```
In [29]: print X.groupby(['sex'])['survived'].mean()
sex
0.0    0.742038
1.0    0.188908
Name: survived, dtype: float64
```

We can see that there is a significant difference in the probability of survival due to sex

```
In [30]: print X.groupby(['sex'])['survived'].count()
sex
0.0    314
1.0    577
Name: survived, dtype: int64
```

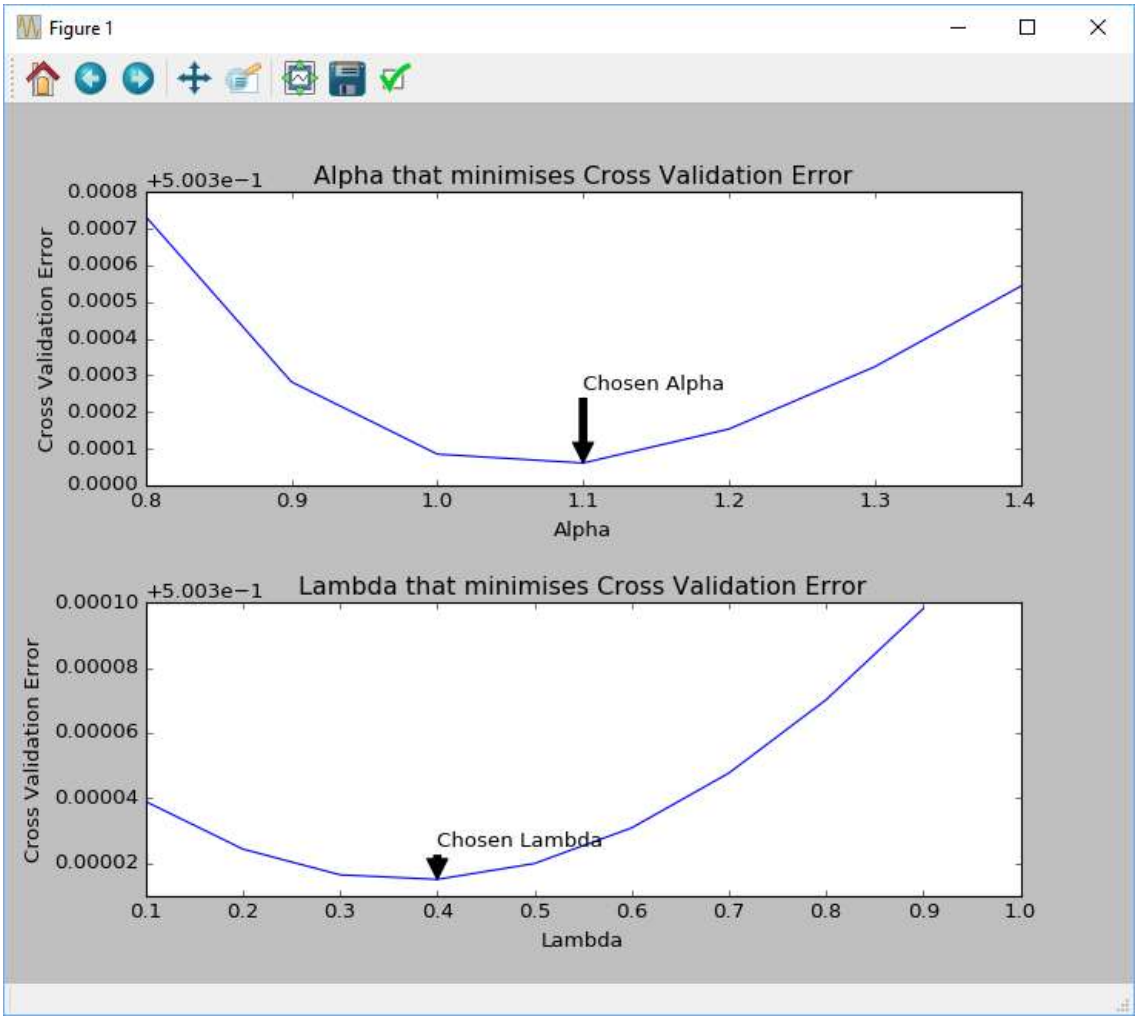
Looking at the sample size, we can see that the count is reasonable for us to come to the conclusion that sex does make a significant contribution to the probability of survival of a passenger.

```
In [31]: print X.groupby(['pclass1','pclass2','pclass3','sex'])['survived'].mean()*100
pclass1  pclass2  pclass3  sex
0.0      0.0      1.0      0.0    50.000000
                           1.0    13.544669
         1.0      0.0      0.0    92.105263
                           1.0    15.740741
1.0      0.0      0.0      0.0    96.808511
                           1.0    36.885246
Name: survived, dtype: float64
```

Lastly we can see that probability of survival is significantly affected by the passenger class as well. We can see that a 3rd class female passenger has about 50% probability of survival compared to

96.8% of a 1st class female passenger. This is rather obvious across the sample as we can see that probability of survival of an average 1st class of passenger is almost twice as much as that of an average 3rd class passenger. The step up from 3rd class to 2nd class is significant for that of a female passenger and looks negligible for a male passenger. In a simplistic sense, the marginal gain of moving up a class from 3rd class looks rather significant for a female passenger whilst the marginal gain for male passenger only looks significant when he moves 2 classes up to the 1st class cabin.

Looking at the plot below we can see that the chosen alpha and the chosen lambda minimises the cross validation set error and hence will be the optimal one to use to train the model.



Conclusions Phase

The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

Communication Phase

Reasoning is provided for each analysis decision, plot, and statistical summary. Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.

Data Limitations:

1. Repopulation of age data
    a. There is an assumption of the distribution of the passengers on board Titanic.
        i. This assumption of Gaussian distribution might not hold and hence introduce certain amount of noise
        ii. However, with this assumption, we are able to retrieve 177 other data points passenger class, sex, sibling-spouse count, parent-child count, port of embarkation and fare
        iii. Furthermore, looking at the histogram distribution, it seems that the distribution in general looks relatively unchanged.
    b. There is no need to do a t test because distribution of the "cleaned" data is based off the original distribution of age and has the same mean of 0 thus, a t test to see if the data's mean are different will be negative by definition.
2. Cabin Data
    a. The amount of data missing is much more than the amount available and hence is almost impossible for us to make any good assumption to adjust the data
3. Port of Embarkation
    a. The port of embarkation has 2 missing values and they are dropped from the data set as a sample of 889 vs 891 is minimal and the 2 samples consist only 0.22% of the full data

Observations of the data and results of the predictive logistic regression

1. Three variables in particular drives the probability of survival in the Titanic disaster.
    a. Sex
        i. The sex variable increases the probability of survival of a passenger

```
In [56]: X.groupby(['sex'])['survived'].mean()
Out[56]:
sex
0.0    0.742038
1.0    0.188908
Name: survived, dtype: float64
```

    b. Passenger Class
        i. A first class passenger is passenger is almost thrice as likely to survive the disaster as compared to a third class passenger

```
In [57]: X.groupby(['pclass1','pclass2','pclass3'])['survived'].mean()
Out[57]:
pclass1  pclass2  pclass3
0.0      0.0      1.0        0.242363
         1.0      0.0        0.472826
1.0      0.0      0.0        0.629630
Name: survived, dtype: float64
```

c. Age

i. The average of the survivors of the Titanic disaster is lower than that of those that did not survive the disaster as seen below.

```
In [62]: print X.groupby(['survived'])['age'].mean()
survived
0.0    0.034859
1.0   -0.080336
Name: age, dtype: float64
```

2. The most likely survivor of the disaster as seen by the predictive logistic regression has a profile of:
   a. First Class ticket holder
   b. Female
   c. Young
   d. No siblings or spouse on board
   e. Have parent/ children on board
   f. Embarked at Queenstown
   g. (Fare is relatively insignificant perhaps due to its strong correlation with the class of ticket a passenger holds)

```
In [60]: theta
Out[60]:
                Theta
pclass1      1.664406
pclass2      1.008945
pclass3     -0.049632
sex         -2.666017
age         -0.312344
sibsp       -0.322988
parch        0.100497
embarkedC    0.560160
embarkedQ    1.143073
embarkedS    0.347543
fare        -0.021537
```