COMP 488 - Machine Learning
Titanic Dataset Analysis

Loyola University Chicago
Jose Luis Rodriguez

September 11, 2017

1 Overview

The challenge is to develop an algorithm using the Titanic dataset to make a prediction about passenger survival. The data set consist of a number of variables (features) describing each passenger (gender, age, ticket class, fare, others). There are two data sets the training set which contains a SURVIVAL variable and the test set to determine the accuracy of the predictive model on unseen data. The model, charts and analysis on this report was generated using python (Version 3.6.1) and the following packages:

- **Pandas:** An open source library providing high-performance, easy-to-use data structures and data analysis tools
- **Matplotlib:** A Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments
- **Seaborn:** A Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics
- **Sckitlearn:** Machine Learning in Python. Simple and efficient tools for data mining and data analysis

Table 1: Titanic Data Set - Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Gender	male or female
age	Age in years	numeric
sibsp	#siblings/spouses	numeric
parch	#parents/children	numeric
ticket	Ticket number	character
fare	Passenger fare	character
cabin	Cabin number	character
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

- **pclass:** A proxy for socio-economic status (SES) as 1st = Upper, 2nd = Middle, 3rd = Lower
- **age:** Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- **sibsp:** The dataset defines family relations for close relatives as Sibling = brother/sister and Spouse = husband/wife
- **parch:** Some children travelled only with a nanny, hence parch=0 for this cases, other family relations Parent = mother/father and Child = daughter/son

1.1 Python Code

This report also provides a python (Version 2.7.10) code that reads the datasets as text files then clean them using native python functions. After that use Numpy to create an array from the text files and set the different data types, columns names and perform some data transformation resulting on a matrix (A similar procedure is perform on the test data) that contains features to be use in the model. Finally the library Sckit-learn is use to create a the predictive model using the training and testing data sets.

2 Data Exploration and Visual Analysis

As part of the data preparation process its necessary to check the shape of both data sets (train, test), to determine the number of observation and columns (instances and features) that are present in each data set. Its also good to check the first and last couple of rows of the data set to have an idea of how the data looks like.

Table 2: Training Data Set Inspection - Shape (Rows, Columns): (891, 12)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

By determine the shape of the data we can get an idea of the number of observations and features present in the data. Then by inspecting the first couple of rows (Table.2) of the training data set preliminary observations of the data can be make this step is important as it will give us an idea of how the data looks like also it helps to determine if there are any issues with the data header (column names) and perform some preliminary data integrity. Moreover we can see that the the training set contains the variable that we are trying to predict (Survived).

Table 3: Testing Data Set Inspection - Shape (Rows, Columns): (418, 11)

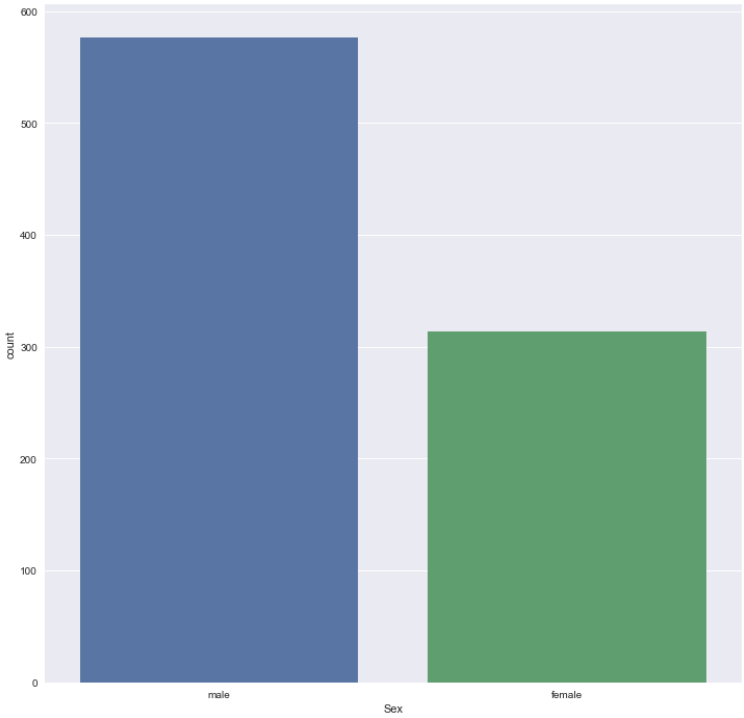
PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

During the data exploration process we can start thinking in what variables may be useful to predict passenger survival, such as the variable Name and Ticket. Also on this step we can start thinking in how to process to deal with missing values (NaN) as well as how to interpret values in the Age column as some of the are given as double instead of integers (whole numbers) which is how age is usually report. Finally from the testing table above (Table.3) we can see that the Survived column (variable) is not present as this is the data set that we are going to use to test the accuracy of the model.

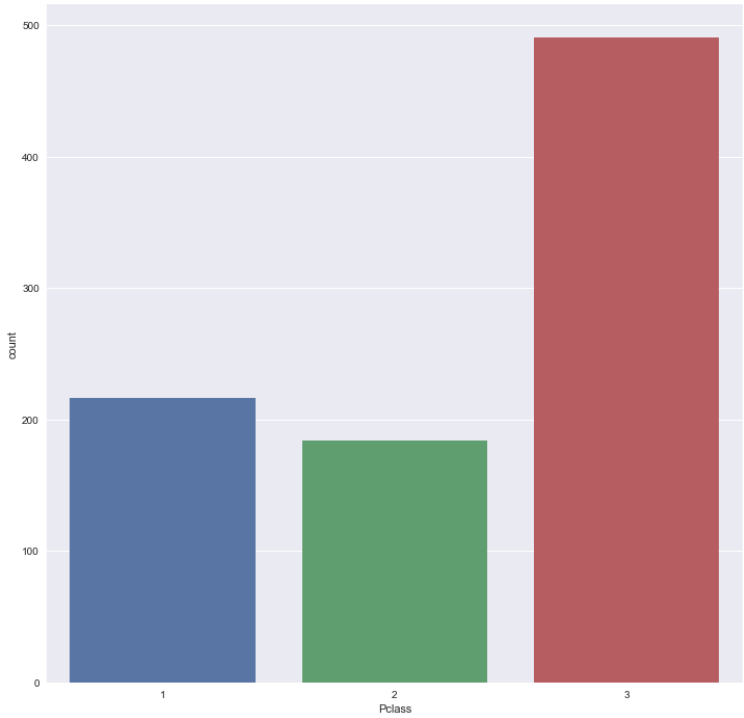
2.1 Data Visualization

When exploring a data set often its useful to try to visualize the variables and possible relationships between variables in the data. For this purpose we can use simple plots such as barplots, histograms and scatter plots if possible to help us identify some patterns in the data in a more visual way. These observation will be helpful later when we are determine what variables would be more helpful when trying to predict passenger survival.

Figure 1: Barplots of Passenger Count by Gender and PClass



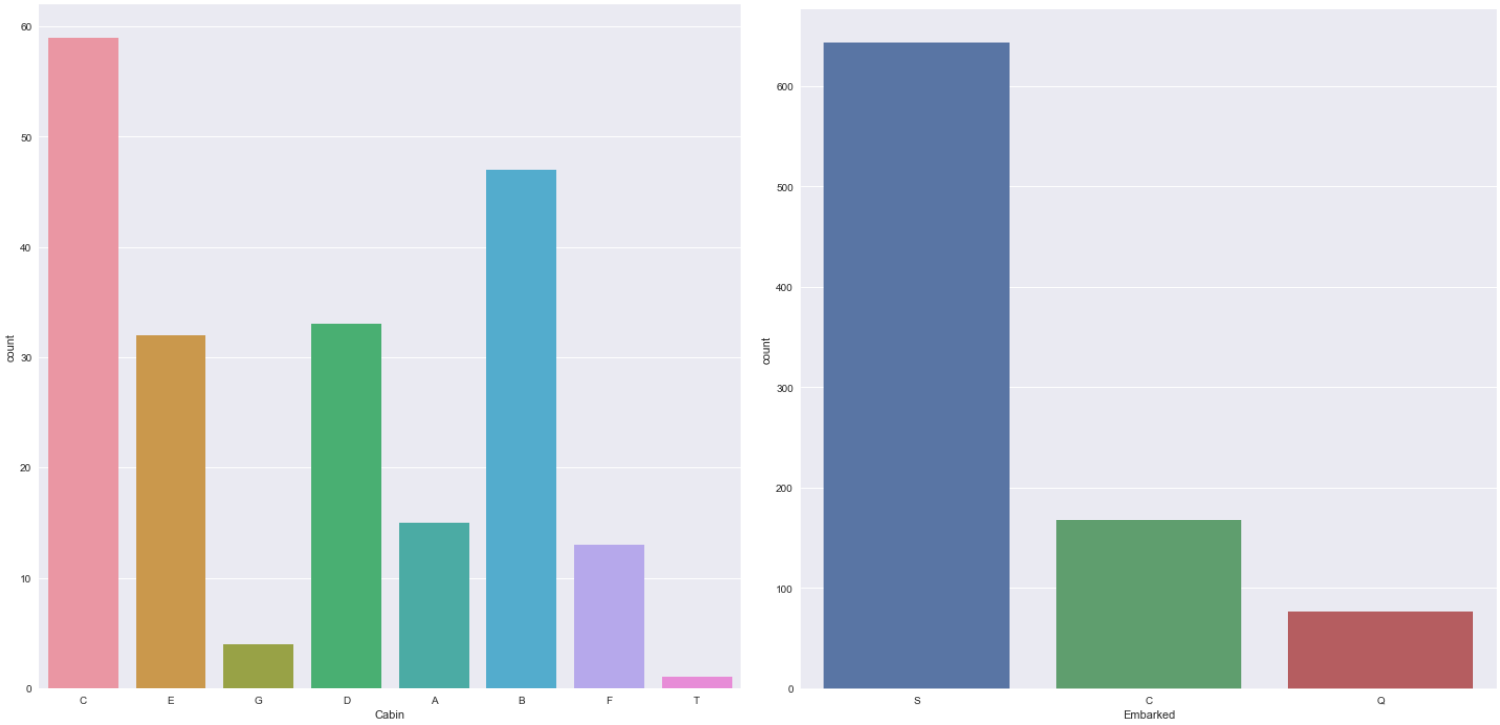
(a) Passengers Count by Gender



(b) Passengers Count by PClass

From the passenger count by gender plot (Figure.1a) we can see how there are almost twice the number of men than women. There may be a number of explanations for why there are more men in the ship such as most crew members on the Titanic were men or in that time it was most likely for a men to travel alone than a women. Now from the passenger count by Pclass (Figure.1b) we can determine that most of the passengers in the training data were in PClass number 3 or lower class again we can speculate that this is due to the number of crew members and people migrating to The United States.

Figure 2: Barplots of Passenger Count by Cabin and Port of Embarked



(a) Passengers Count by Cabin

(b) Passengers Count by Port of Embarked

By looking at the passenger count by Cabin (Figure.2a) pair with the Titanic deck layout we can have an idea of passengers proximity to the upper level. From the second plot (Figure.2b) showing the count by port of embarked, we can determine that Southampton port (S) is where most passengers boarded the Titanic.

2.2 Variable Interaction

The previous plots gives us an idea of variables (Gender, Pclass, Cabin and Embarked) distribution and behavior. Now we can take that same approach one step further to examine how the different independent variables (features – Pclass, Sex, Age, SibSp, Parch, Fare, Cabin and Embarked) interact with the dependent variable (target – Survived). In Figure 3, there are four different plots, starting with figure.3a the interaction between SURVIVED and FARE where we can see that as fare prices are higher survival rate increase. In contrasts figure.3b the interaction between SURVIVED and AGE shows survival rate among younger passengers is greater. By looking at the mean survival rate we can see some interesting

patterns such as in figure.3c mean(Survived) by Gender we can see that survival mean rate among female passenger is almost 3 times mens survival rate. Finally in figure.3d mean(Survived) by Gender and PClass shows the mean survival rate of passengers by gender and the three different ticket classes (1=Upper Class, 2=Middle Class and 3=Lower Class)

3 Data Preparation and Analysis

During the data preparation process it was necessary to transform some variables from character categorical variables to binary/numeric variables in order to make them part of the model. The Sex column (feature) was transform from character string male and female to 0, 1. The Embarked Column with entries (S,C,Q) was also transform to values 0,1,2 as well as the cabin column. These type of data transformations allows to perform a more in-depth data analysis and get some more insights about the data. Its important to note that some variables dont any any numeric meaning such as PassengerId that represent unique passengers ids.

Table 4: Training Data Set - Summary Statistics

	Survived	**Pclass**	**Sex**	**Age**	**SibSp**	**Parch**	**Fare**	**Cabin**	**Embarked**
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	0.352413	29.699118	0.523008	0.381594	32.204208	-0.188552	0.359147
std	0.486592	0.836071	0.477990	13.002015	1.102743	0.806057	49.693429	1.815143	0.638707
min	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.000000	-1.000000	-1.000000
25%	0.000000	2.000000	0.000000	22.000000	0.000000	0.000000	7.910400	-1.000000	0.000000
50%	0.000000	3.000000	0.000000	29.699118	0.000000	0.000000	14.454200	-1.000000	0.000000
75%	1.000000	3.000000	1.000000	35.000000	1.000000	0.000000	31.000000	-1.000000	1.000000
max	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.329200	7.000000	2.000000

- Survived: Possible values 1 or 0 and mean of 0.38, which suggest that 38% of the passenger survived
- Pclass: With three possible values (1,2 and 3) and mean of 2.3086 suggest that most passengers belonged to 2 and 3 Pclass
- Sex: Two possible values 1 or 0 and mean of 0.3524 which suggest that the number of 0's (men) may be higher

- Age: From the mean of 29.69 we can infer that the passengers were relatively young.
- SibSp: The max value is 8 (large family) and mean of 0.5230 suggesting that most passengers travel alone or with one family member.
- Parch: The max value is 6 and mean of 0.3815 again suggesting that most passengers travel alone.
- Fare: This variable is interesting because the mean 32.2042 and $25\%, 50\%$ and 75% percentiles seem to be reasonable spread out but the max value of 512.3292 which is significantly far from the mean a more in-depth analysis is necessary for this variable.
- Embarked: Three possible values $0, 1, 2$ (S,C,Q) and mean of 0.359147 below 1 it supports the visual analytics that most people embarked from port S

3.1 Correlation Analysis

After looking at the summary statistics and doing some general observation its time to do some more in depth analysis of the relationship between the dependent and independent variables. In this case the relationship between the variable SURVIVED and the other variables, by looking at the Pearson Correlation table showing the strength (positive, negative) of the relation between the variables in the data set. By looking at the first column we can identify a couple of variable with significantly correlated coefficient such as (Sex, Pclass, Cabin and Fare), it may be hard to find correlations in the table hence Figure.4 shows the table using a Heatmap allowing for a visual representation of the correlation table.

Table 5: Titanic Train Data - Correlation Table

	Survived	**Pclass**	**Sex**	**Age**	**SibSp**	**Parch**	**Fare**	**Cabin**	**Embarked**
Survived	1.000000	-0.338481	0.543351	-0.069809	-0.035322	0.081629	0.257307	0.270310	0.101849
Pclass	-0.338481	1.000000	-0.131900	-0.331339	0.083081	0.018443	-0.549500	-0.562800	0.050992
Sex	0.543351	-0.131900	1.000000	-0.084153	0.114631	0.245489	0.182333	0.114229	0.111249
Age	-0.069809	-0.331339	-0.084153	1.000000	-0.232625	-0.179191	0.091566	0.167017	0.001932
SibSp	-0.035322	0.083081	0.114631	-0.232625	1.000000	0.414838	0.159651	-0.054525	-0.058008
Parch	0.081629	0.018443	0.245489	-0.179191	0.414838	1.000000	0.216225	0.036944	-0.076625
Fare	0.257307	-0.549500	0.182333	0.091566	0.159651	0.216225	1.000000	0.376888	0.058462
Cabin	0.270310	-0.562800	0.114229	0.167017	-0.054525	0.036944	0.376888	1.000000	0.007110
Embarked	0.101849	0.050992	0.111249	0.001932	-0.058008	-0.076625	0.058462	0.007110	1.000000

4 Model Selection

Per the preliminary analysis (visual and descriptive) we have and idea of what columns (features) to use to create a model, now the challenge is to predict passenger survival identify in the training data set as SURVIVED. This a discrete variable (yes/no). We can use regression analysis to be more precise logistic regression method to classify passenger survival given the different features present in the training data set.

4.1 Predictive Model: Logistic Regression

The first step in creating the model would be to split the train data in two in order to test the accuracy of the model. The sklearn package has a model selection method that helps with this task by splitting arrays or matrices into random train and test subsets. After splitting the training data in two subsets, we can create the logistic regression model by fitting the relevant features (independent variables) and to the target testing variable. The following formula is derive from the logistic regression model:

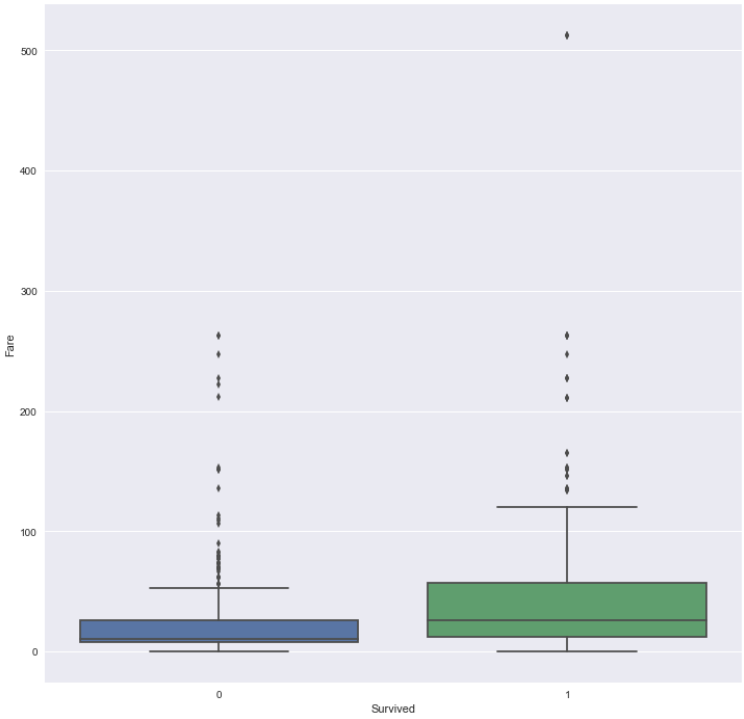
$$\begin{aligned} \text{SURVIVED} =& 1.42981371 - (0.8420 * Pclass) + (2.6601 * Sex) \\ & - (0.0354 * Age) - (0.2276 * SibSp) - (0.1472 * Parch) \\ & + (0.0033 * Fare) + (0.1704 * Cabin) + (0.2498 * Embarked) \end{aligned} \tag{1}$$

The model has an accuracy of 0.7892 meaning that the model explains about 78% of the variation in the data when using most variables but (NAME and TICKET). To improve the model in a future work feature selection could be use to identify the variables that have more influence over the SURVIVED variable.

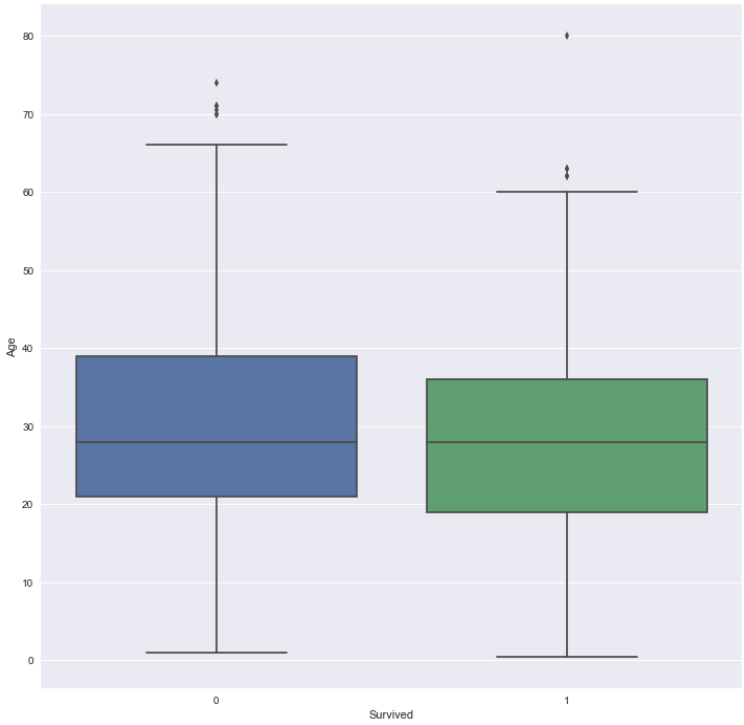
5 Reference

Generalized Linear Models. 1.1. Generalized Linear Models – Scikit-Learn 0.19.0 Documentation, http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

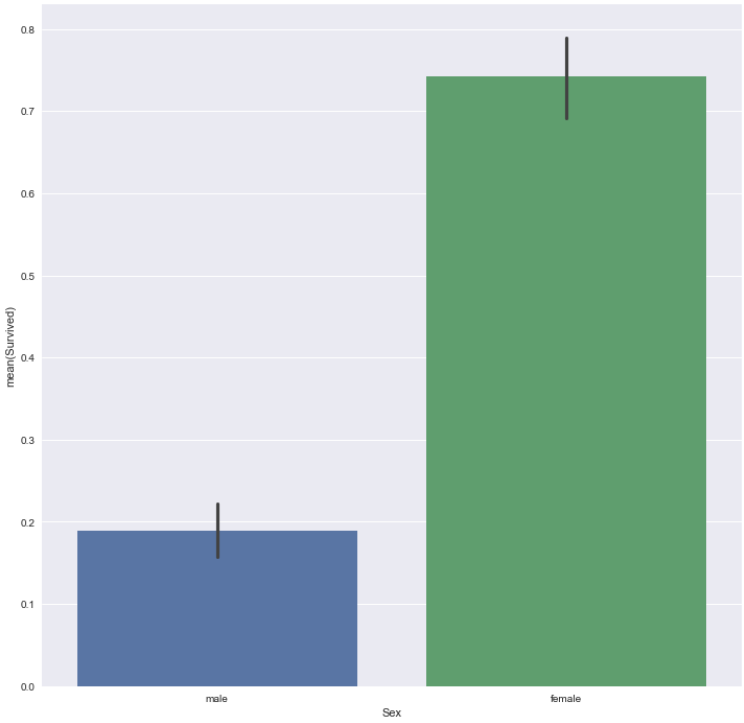
Figure 3: Variable Interaction with Target Variable - Survived



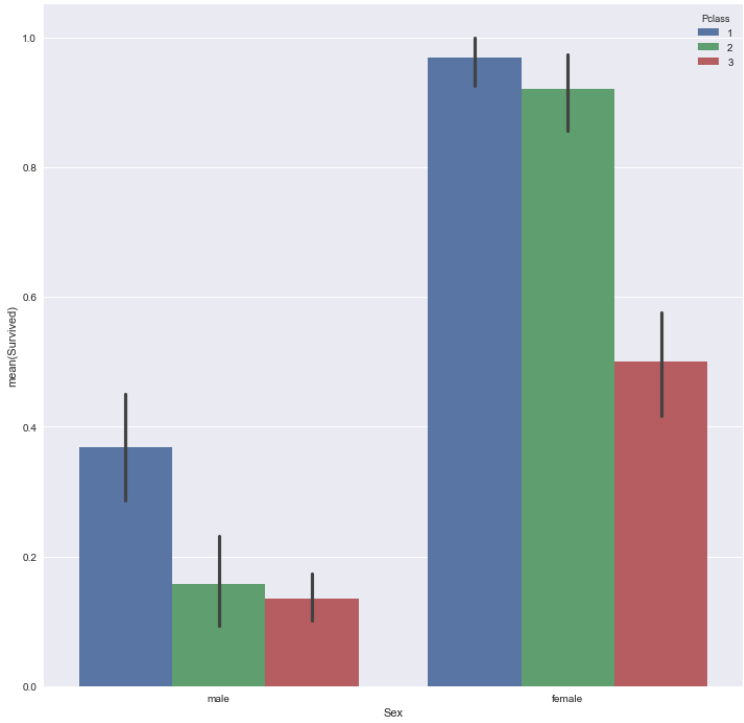
(a) Boxplot - Survived by Fare Price



(b) Boxplot - Survived by Age



(c) Barplot - mean(Survived) by Gender



(d) Barplot - mean(Survived) by Gender and PClass

Figure 4: Titanic Train Data - Correlation Heatmap