# Overview of Data

**Jaime: can you add a description of data we are using and what we hope to achieve (logistic regression to model the response variable which is binary for survival)**

# Exploratory Data Analysis

While reviewing the provided Train and Test data sets it became clear that there were multiple observations that did not contain a value for Age. The only other variable with a missing value was Fare, that was isolated to the provided Test set. Figure 1 below provides a visualization of the null values among the data for both the Train and Test data sets.

|  |
| --- |
|  |
| ***Figure 1: Null Values Within Data Sets*** |

Given that Age is likely an important variable needed to determine survival, imputation was conducted on the column using the Mice Package in R.

|  |
| --- |
|  |
| ***Figure 2: Imputation for Age*** |

|  |
| --- |
|  |
| ***Figure 2: Replacement for Fare*** |

# Derived Variables

**Arturo:** can you write something about the variables you derived? Mine will be here too like ‘title’.

Title

AgeBin

# Kaggel Results

Manny: Write up on results from Kaggle and meaning.

"Survived" "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare" "Embarked" "Title" "AgeBin" "Family"

Care was taken not to remove too many data points given that the data set is broad, and the more information removed from the model the harder more extreme Sale Price might be to predict.

# Logistic Regression within Train Data

All of us

"Survived" "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare" "Embarked" "Title" "AgeBin" "Family"

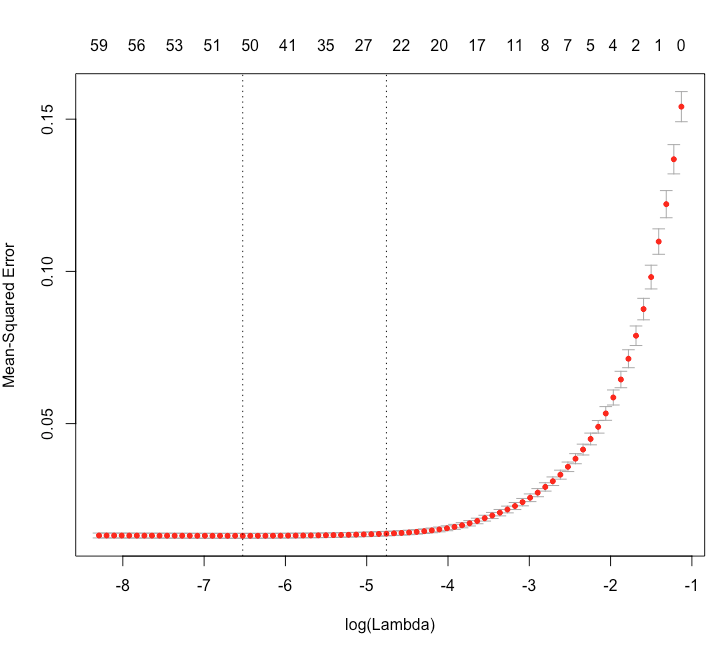
Filler filler filler

filler

filler

filler

Old picture but maybe something like this in this report…



***Figure 7: LASSO Cross Validation Fit Plot***

# Conclusion

All of use: filler

filler

filler

filler

filler

Appendix

# Data Dictionary

The data was provided by Kaggle and corresponds to the passengers who sailed on the titanic.

The data consists of two sets, a training set containing 891 observations and a testing set containing 418 observations. The observations consist of 12 base variables which include information from the size of families to amenities and the quality of many aspects of the cabin on the ship. Important variables that will recur in this analysis include:

1. PassengerID: Unique identifery of passengers on board
2. Survived: Survival; 0 = No, 1 = Yes
3. Pclass: Ticket class; 1 = 1st, 2 = 2nd, 3 = 3rd
4. Name: Passenger’s full name and title
5. Sex: Sex; ‘male’ or ‘female’
6. Age : Age in years
7. Sibsp: # of siblings / spouses aboard the Titanic
8. Parch: # of parents / children aboard the Titanic
9. Ticket: Ticket number
10. Fare: Passenger fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation; C = Cherbourg, Q = Queenstown, S = Southampton

The following variables were derived from the core table:

1. Title: Status or title of the passenger
2. AgeBin: Categories corresponding to passenger’s age
3. Family: The sum of #of siblings/spouses and # of parents/children

(Where did the data come from? How big is it? How many observations? Where can we find out more? What are the specific variables that we need to know to understand with respect to your analysis?)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Variables** | **Description** | **Key** | **Notes** |
| 1 | ID | Unique numeric identifier corresponds to passengers | Numeric:  [1 – 891]:train;  [892 –1309]:test | Unique identifier for passengers |
| 2 | Survived | Notes whether or not a person survived | Numeric; 0 = Did not survive;  1 = Survived |  |
| 3 | Pclass | Class of ticket belonging to passenger | Numeiric: 1 = 1st, 2 = 2nd, 3 = 3rd | Should be categorical |
| 4 | Sex | Sex of passenger | Character; ‘male’ or ‘female’ |  |
| 5 | Age | Age of passenger in years | Numeric | Age of infants less than 1 year old was guessed and reported as a decimal; Ex: 0.75.  [0.17 – 80].  177 NAs in train  86 NAs in test |
| 6 | Sibsp | Number of siblings or spouses aboard the titanic | Integer | [0 -8]; no 6 or 7 |
| 7 | Parch | Number of parents or children aboard the titanic | Integer | [0-6] |
| 8 | Ticket | Ticket number | Character | Many are numbers, some contain descriptions |
| 9 | Fare | Passenger Fare / Ticket Price | Numeric | 1 NA in test  [0 – 512.33]; presumed to be in American dollars |
| 10 | Cabin | Passenger’s Cabin number | Character; usually in the form of *LETTER##,* EX: A55, | Almost Unique for every passenger. One passenger can be in multiple cabins; 827 blanks for train; 327 blanks for test. |
| 11 | Embarked | Port of Embarkation; | Character;  C = Cherbourg,  Q = Queenstown,  S = Southampton | 2 blanks for train |
| 12 | Title | Extracted from name and corresponds to rank or marital status | Character; Ex: Mr., Mrs., Miss, | Some titles were uncommon, such as Countess, Captain; these are placed under ‘uncommon’ |
| 13 | AgeBin | Ages of passengers by increments of 10 years | Character;  (0-10], (10-20],…(80,90], ‘unknown’ | Missing values are given their own category, they are not designated NA |
| 14 | Family | Sum of: number of parents/children and number of siblings/spouses | Numeric |  |

# Kaggel Score

