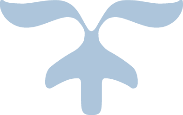


Titanic Dataset

CST8390\_23W Assignment 2



February 13, 2023

Algonquin College

Computer Engineering Technology

Business Intelligence and Data Analytics

Question: Which factors contribute the most towards vehicle-related incidents involving injury?

Hoang, John – 040896360

Kingdom, Neil – 040967309

Prepared for:

Dr. Anu Thomas

Table of Contents

[Introduction 3](#_Toc127282987)

[Data Understanding 3](#_Toc127282988)

[Starting Dataset 3](#_Toc127282989)

[Dataset Breakdown 4](#_Toc127282990)

[Classes Train Set 4](#_Toc127282991)

[Attributes Train Set (Age Updated) 5](#_Toc127282992)

[Classes Test Set 5](#_Toc127282993)

[Attributes Test Set (Updated Age) 5](#_Toc127282994)

[Data Preparation 6](#_Toc127282995)

[Preprocessing – Classification 6](#_Toc127282996)

[Updated Training Dataset on Classification 6](#_Toc127282997)

[Updated Test Dataset on Classification 7](#_Toc127282998)

[Cleaning Up 7](#_Toc127282999)

[Removing Data 7](#_Toc127283000)

[Updating Missing Values 8](#_Toc127283001)

[Modeling and Evaluation 8](#_Toc127283002)

[Discussion of Results 9](#_Toc127283003)

[Conclusion 9](#_Toc127283004)

[References 10](#_Toc127283005)

[Appendix 1 11](#_Toc127283006)

[Training Set Male (Mr.) 11](#_Toc127283007)

[Training Set Female (Miss) 12](#_Toc127283008)

[Training Set Female (Mrs.) 13](#_Toc127283009)

[Training Set Unique 13](#_Toc127283010)

[Appendix 2 14](#_Toc127283011)

[Test Set All Missing Age Instances 14](#_Toc127283012)

[Test Set Male (Mr.) 15](#_Toc127283013)

[Test Set Female (Ms.) 16](#_Toc127283014)

[Test Set Female (Mrs.) 16](#_Toc127283015)

[Test Set Unique 16](#_Toc127283016)

[Test Set Fare Missing 16](#_Toc127283017)

# Introduction

The city of San-Diego provides public statistics regarding information about its citizens on the internet website data.sandiego.gov. One of these datasets is the Traffic Collisions dataset (with people and vehicles involved). In this report, we will attempt to answer a simple question by analysing this dataset. The question is as follows: Which factors contribute the most towards vehicle-related incidents involving injury? Prior to answering these questions, a few hypotheses can be explored. Out of handful of features provided within the dataset, it can be assumed that the following may have some correlation with injury: the time at which the incident occurred, the type of vehicles involved in the incident, the location of the incident (indicated by the San-Diego Police beat), the type of road that the incident occurred on (given by numerous features, such as street number, street name, street suffix, and intersection name), the violation type of the primary collision factor, the number of people killed in the accident, and the hit and run level. We cannot be sure as to which times of day, days of the week, or days of the month, will have the greatest impact on injury, although we can estimate that incidents will be more likely to occur on Fridays, and during the weekend, as these are typically the busiest times of the week. Considering that rush hour occurs around 7am to 10am and 4pm to 6pm, it can be estimated that incidents will be more likely to occur during these times, increasing the risk of injury. Streets which are busier and have a higher level of traffic congestion are estimated to produce more severe injury. The number of people killed in an incident will likely have a strong positive correlation with injury. Finally, hit and runs associated with felonies may have a greater correlation with injury, considering that criminals are less likely to be cautious about their surroundings, and tend to cause more harm than, for instance, a misdemeanor.

# Data Understanding

The San-Diego dataset consists initially of 119, 297 instances and 22 features. Each instance within the dataset represents a single person involed in a vehicle-related incident which occurred within the city of San-Diego, and was reported by the San-Diego Police Department. Thus, the same collision may be reported multiple times for each person involved in an incident. The dataset is updated yearly by the city. As of current date, it contains records from between 2015-2023. As noted on data.sandiego.gov, reports are generally not made for property damage-only incidents that do not involve a hit and run or DUI. Additionally, the dataset only records incidents that occurred on a street, as incidents occurring on the freeway are handled by the California Highway Patrol.

## Starting Dataset

|  |  |
| --- | --- |
| **Attribute** | **Datatype** |
| Passenger Id | Distinct Numeric |
| PClass | Nominal |
| Sex | Nominal Binary |
| Age | Continuous |
| Sibsp | Distinct Numeric |
| Parch | Distinct Numeric |
| Fare | Continuous |
| Embarked | Nominal |
| Cabin | Distinct Numeric |
| Ticket | Distinct Numeric |
|  |  |
| Survived | Nominal Binary |

## Dataset Breakdown

## Classes Train Set

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Survivor | Survived (1) | 549 |
| Non-Survivor (0) | 340 |
| Gender | Female (1) | 312 |
| Male (1) | 577 |
| Embarked | C | 167 |
| S | 644 |
| Q | 78 |
| Siblings and or Spouse | 0 | 606 |
| 2 | 237 |
| 3 | 16 |
| 4 | 18 |
| 5 | 5 |
| 8 | 7 |
| Parent and or Children | 0 | 676 |
| 1 | 118 |
| 2 | 80 |
| 3 | 5 |
| 4 | 4 |
| 5 | 5 |
| 6 | 1 |

## Attributes Train Set (Age Updated)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Minimum** | **Maximum** | **Means** | **StdDev** |
| Age | 0.42 | 80 | 29.656 | 14.03 |
| SibSp | 0 | 8 | 0.5224 | 1.104 |
| Parch | 0 | 6 | 0.382 | 0.807 |
| Fare | 0 | 512.329 | 32.097 | 49.698 |

## Classes Test Set

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Gender | Female (1) | 152 |
| Male (0) | 266 |
| Embarked | C | 102 |
| S | 270 |
| Q | 46 |
| Siblings and or Spouse | 0 | 283 |
| 1 | 110 |
| 2 | 14 |
| 3 | 4 |
| 4 | 4 |
| 5 | 1 |
| 8 | 2 |
| Parent and or Children | 0 | 324 |
| 1 | 52 |
| 2 | 33 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |
| 6 | 1 |
| 9 | 2 |

## Attributes Test Set (Updated Age)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Minimum** | **Maximum** | **Means** | **StdDev** |
| Age | 0.17 | 76 | 29.917 | 13.891 |
| SibSp | 0 | 8 | 0.447 | 0.897 |
| Parch | 0 | 9 | 0.392 | 0.981 |
| Fare | 0 | 512.329 | 35.599 | 55.843 |

# Data Preparation

## Preprocessing – Classification

Since the name of the passengers provides for a title, it is useful in categorizing the passengers into classes of Mr., Ms. (Ms., Miss and Mlle), Mrs. (Mrs. and Mme.), Master, Capt., Major, Col., Lady, Rev can be its own unique class. Although, Masters does appear to be 12 years of age or under.

It is also useful to categorized the age group. Here, passenger age less than 2 classified as baby, age of less than 12 as a child, age of less than 25 as youth, age of less than or equal to 60 as adult and older than 60 as seniors.

For the Siblings and Spouse and Parent and Children fields the sum is categorized as Relative with none for no relatives on board, few for less than 3 relatives on board and 3 or greater as having many relatives on board the Titanic.

Since the Fare is a continuous type it is possible to perform classification by binning using one of two techniques equal width and equal frequency. On equal width binning, the value of Fare is arranged in ascending order which allows for a quick observation on the number of bins required to divide the features into roughly equal classes. It would seem that to utilized equal width binning it is required to have about 50 identifiable classes. Given that the range of the feature is about 500 starting at 0 to 512.33 and 870 instances are under 100 inclusively of 768 instances being under 30 the bins would have to be in bins of 10 instances. This would make the category’s classes thinly distributed and varies widely with many classes in between being empty.

Equal frequency binning makes more sense in categorizing the Fare feature since it would ensure that each classes contain a number of the ranges of the feature. This is done by arranging the feature values in ascending order and then numbering each instances in orders of 10 from the lowest value fields to the highest value fields.

## Updated Training Dataset on Classification

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Age Group | BABY | 14 |
| CHILD | 68 |
| YOUTH | 250 |
| ADULT | 536 |
| SENIOR | 21 |
| Relative | NONE | 535 |
| FEW | 263 |
| MANY | 91 |
| Title | UNIQUE | 63 |
| Ms. | 184 |
| Mrs. | 125 |
| Mr. | 517 |
| Equal Frequency | 1 | 89 |
| 2 | 89 |
| 3 | 89 |
| 4 | 89 |
| 5 | 89 |
| 6 | 89 |
| 7 | 89 |
| 8 | 89 |
| 9 | 89 |
| 10 | 88 |

## Updated Test Dataset on Classification

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Age Group | BABY | 8 |
| CHILD | 18 |
| YOUTH | 144 |
| ADULT | 237 |
| SENIOR | 11 |
| Relative | NONE | 253 |
| FEW | 131 |
| MANY | 34 |
| Title | UNIQUE | 27 |
| Ms. | 79 |
| Mrs. | 72 |
| Mr. | 240 |
| Equal Frequency | 1 | 42 |
| 2 | 42 |
| 3 | 42 |
| 4 | 42 |
| 5 | 42 |
| 6 | 42 |
| 7 | 42 |
| 8 | 42 |
| 9 | 41 |
| 10 | 41 |

## Cleaning Up

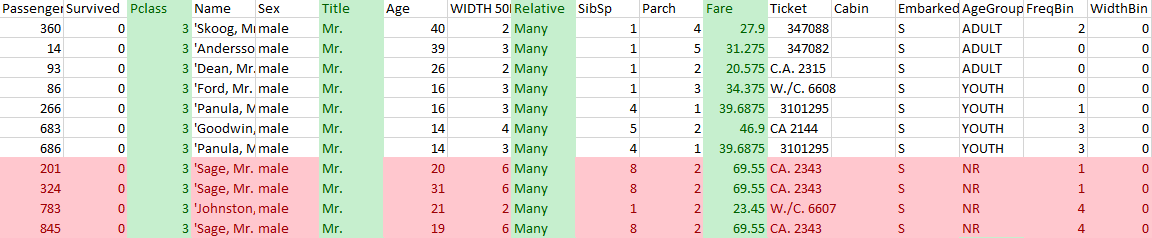
## Removing Data

The first step is to remove highly correlated data between the features and the survivor class. Since passenger id and name classes has a 1 to 1 correlation to survivor and non-survivor and relevant information from the name has already been extracted, it’s remove from the dataset. Onward, it’s not necessary to remove duplicates because it had already been verified by the passenger Id and names that all instances are different. Although cabin identifier can be useful in identifying the location of the passenger room, it is missing a significant amount of data on about 80% of the records it’s remove for the reason that improvising a method to fill in the values introduces overfitting of the dataset. Ticket id has some relevancy in the dataset that provides for the venue of purchasing the ticket but over 74%, 660 instances, does not have an identifier code attached to it. Attempts to correlate the identifier code to the ticket without identifiers yielded no significant contributions to the method use on issuance of the tickets. Although a pattern exists but it could not be reasonably identified and over 74% of the data is missing the ticket class is removed.

## Updating Missing Values

The age feature has about 20% of the data missing which is manageable on formulating a solution to resolve a value to the passengers that has a missing age field. Here the idea is to use a random number generator on the range of the age of the passengers that have correlating features. The simplest approach is to use excel spreadsheet filtering tools that can separately identify classes in the categories keeping in tack all other data [1].

The method used is to first identify the missing values by the title class that was previously produce. The result is then filtered by relatives, a category that was also previously produced from the total of spouse, siblings, children and parent of each passenger. An additional passenger class feature is applied on narrowing the range of ages on the missing age value. On the gathered results, the age range is established by the range of cost to the passenger’s fare. The excel expression formula “=RandBetween (lower value, higher value)” is used with the range to resolve the missing age value. For example, a title of “Mr” is crossed filtered by relative with “None” then by passenger class of “First Class” then finally observing the range of ages between the cost of the passenger fare the minimum age and the maximum age provides the inputs to the random number generator. A visual example is provided below where the highlighted green cells shows the features used on filtering the data and the red cells rows shows the instances with missing age value. A complete list of the missing age value and the corresponding retrofitted value for the Train set is provided in “Appendix 1” and the test set in “Appendix 2”.



Similar filtering of dataset is done with the test set to resolve the missing age value and the missing fare value.

# Modeling and Evaluation

a. Total instances in the test file: 418  
b. Number of persons predicted to survive (1): 136  
c. Number of persons predicted not to survive (0): 282  
d. Percentage of predicted survival: 32.5%

# Discussion of Results

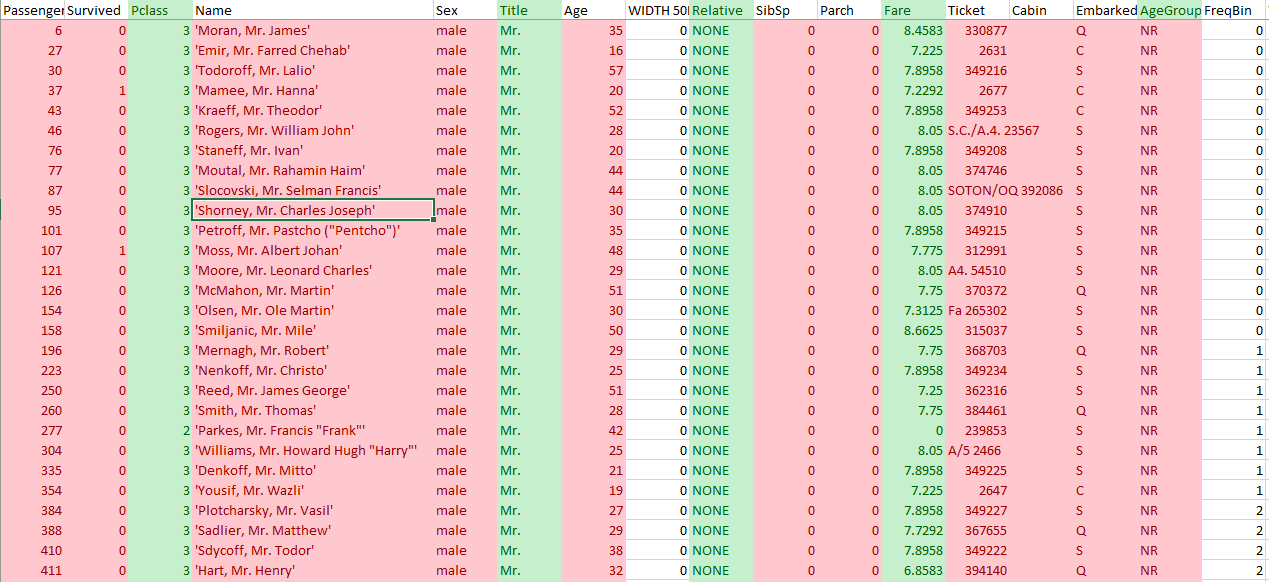
# Conclusion

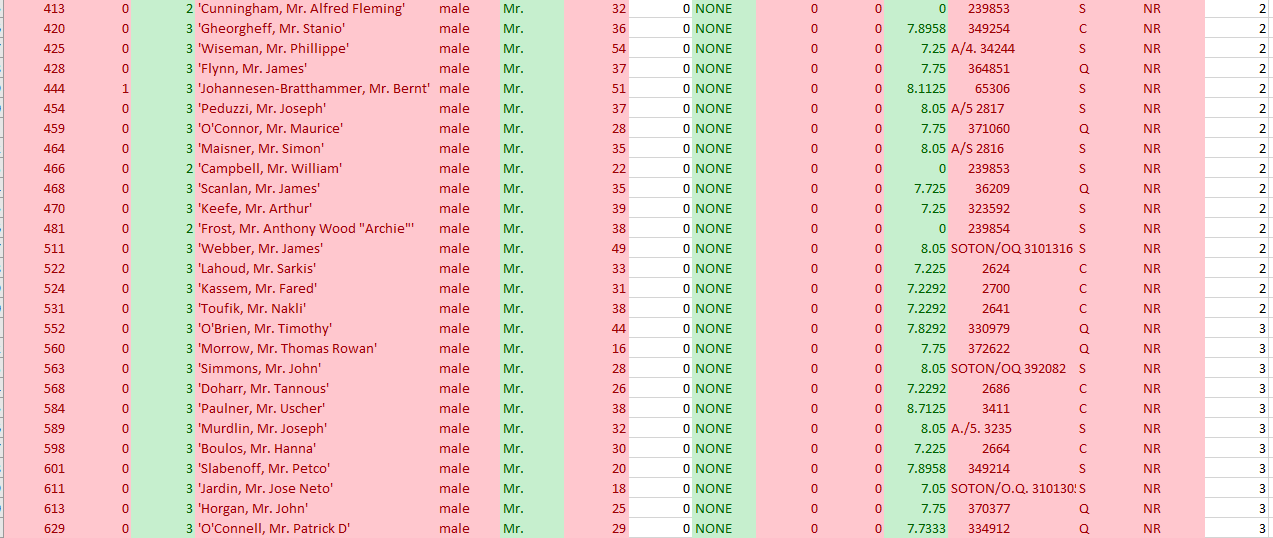
# References

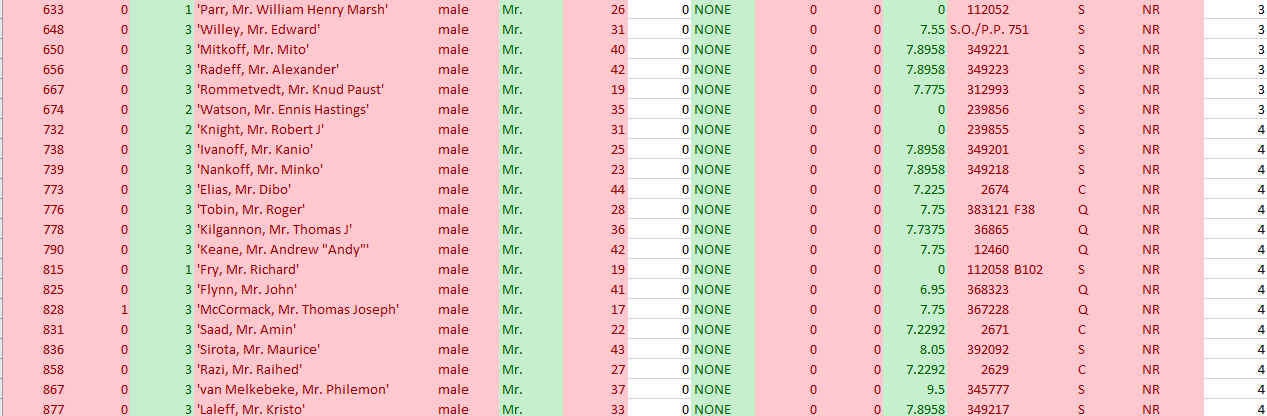
|  |  |
| --- | --- |
| [1] | N. Donges, "Predicting the Survival of Titanic Passengers," Towards Data Science, 14 May 2018. [Online]. Available: https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8. [Accessed 13 February 2023]. |
| [2] | I. H. Witten, "The University of Waikato," 2013. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class3-DataMiningWithWeka-2013.pdf. [Accessed 13 February 2023]. |
| [3] | T. Facts, D. Fowler and H. i. Numbers, "Titanic Survivor," Titanic Facts, 2023. [Online]. Available: https://titanicfacts.net/titanic-survivors/. [Accessed 13 February 2023]. |

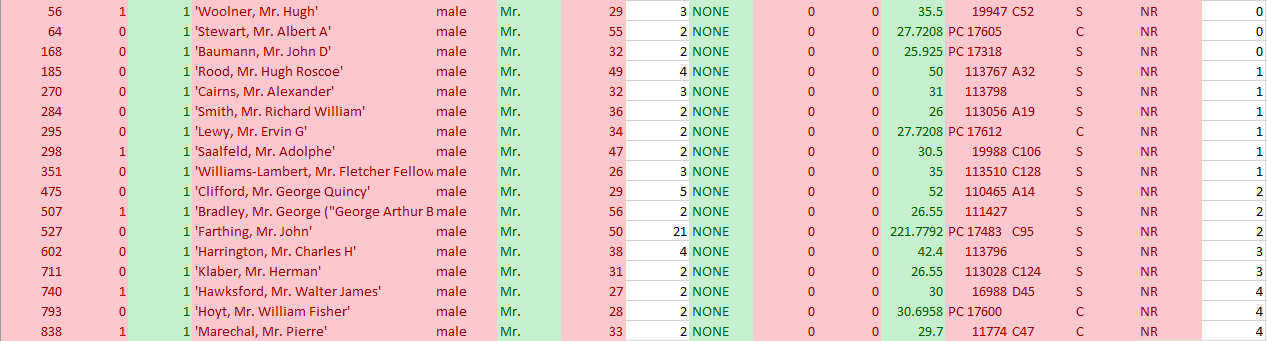
# Appendix 1

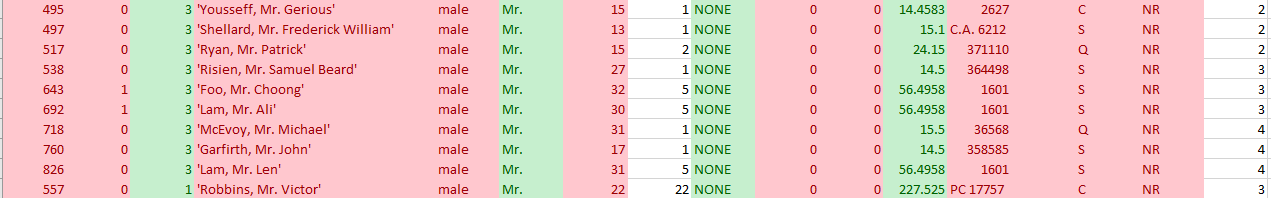
### Training Set Male (Mr.)

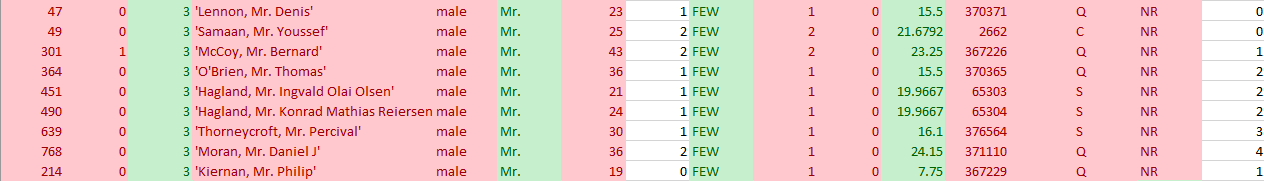


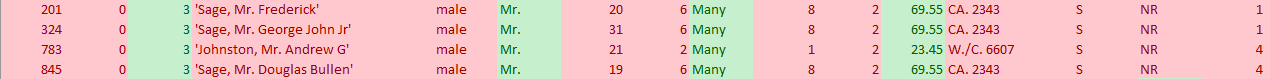




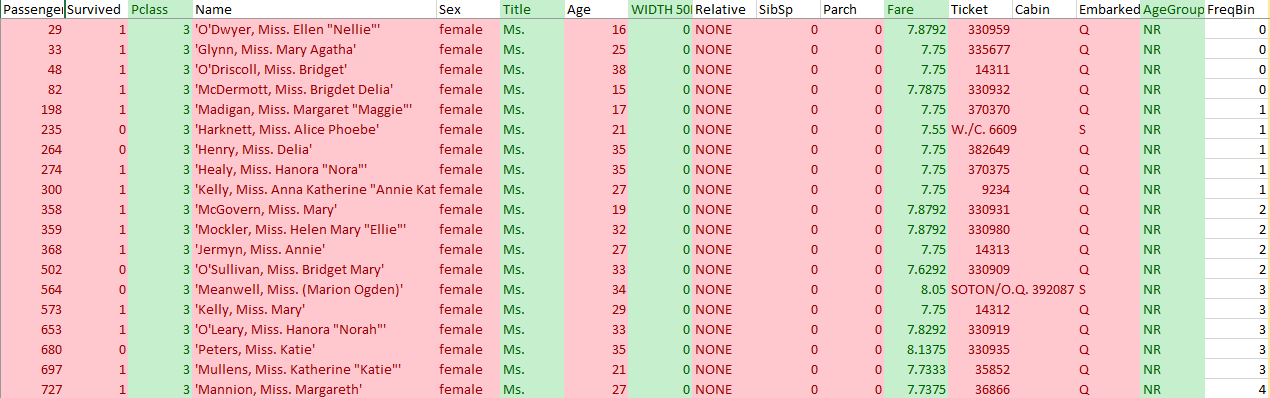








### Training Set Female (Miss)

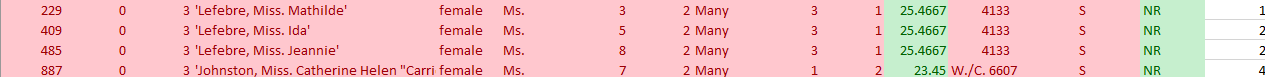












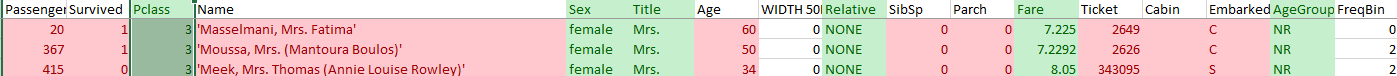


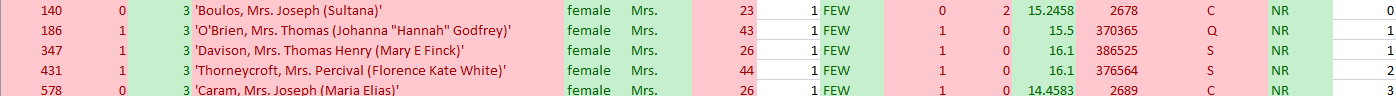






### Training Set Female (Mrs.)













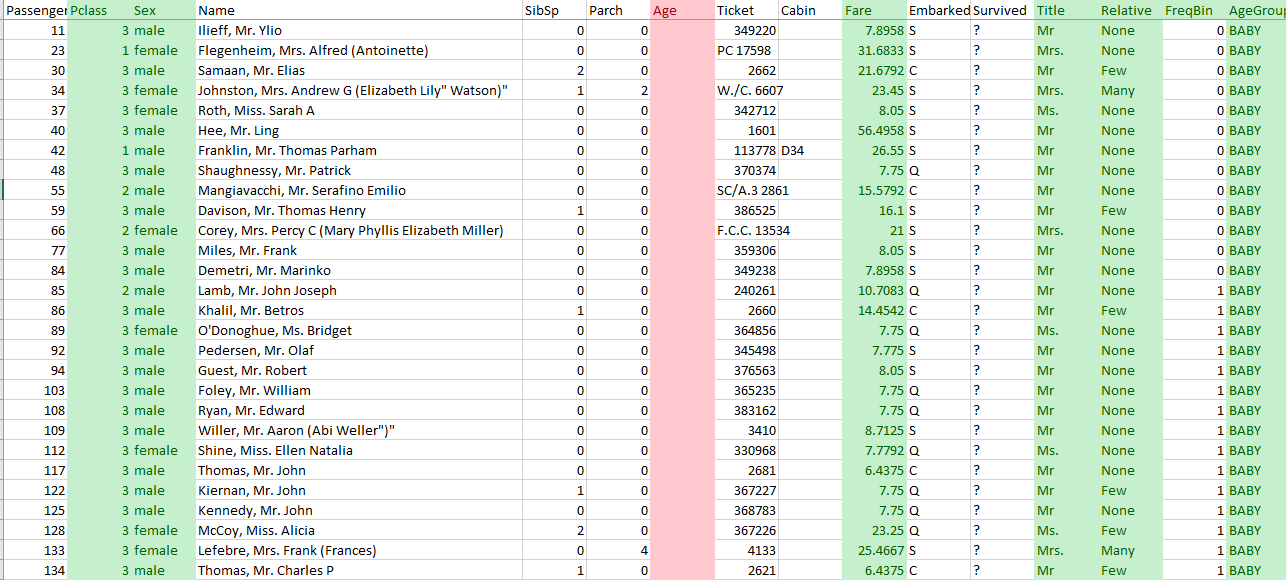
### Training Set Unique

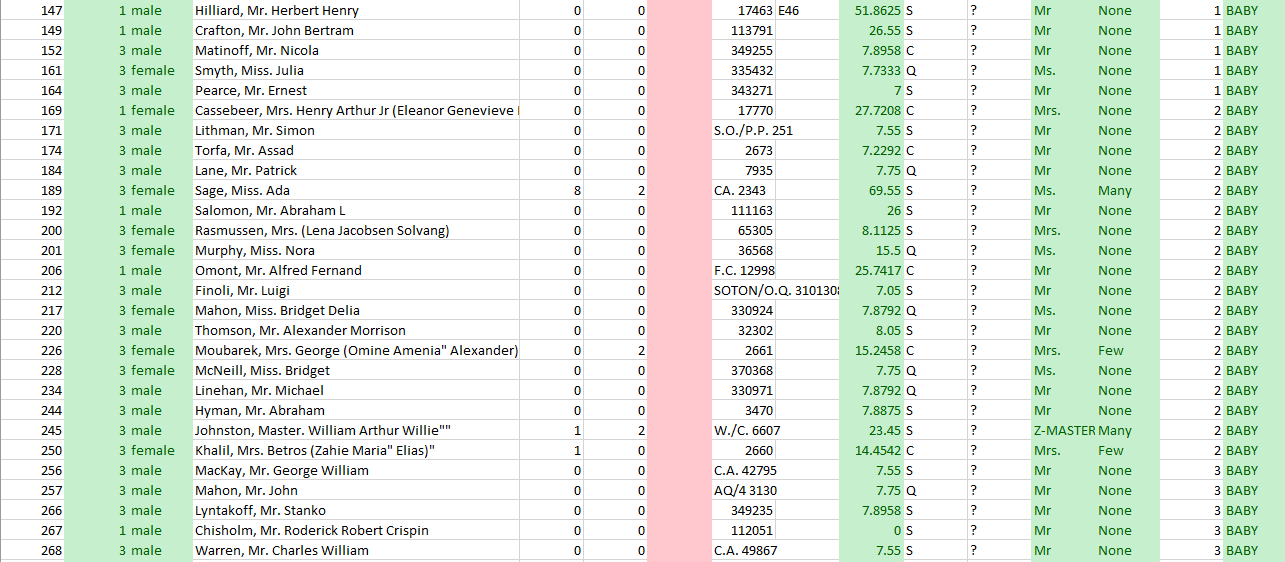


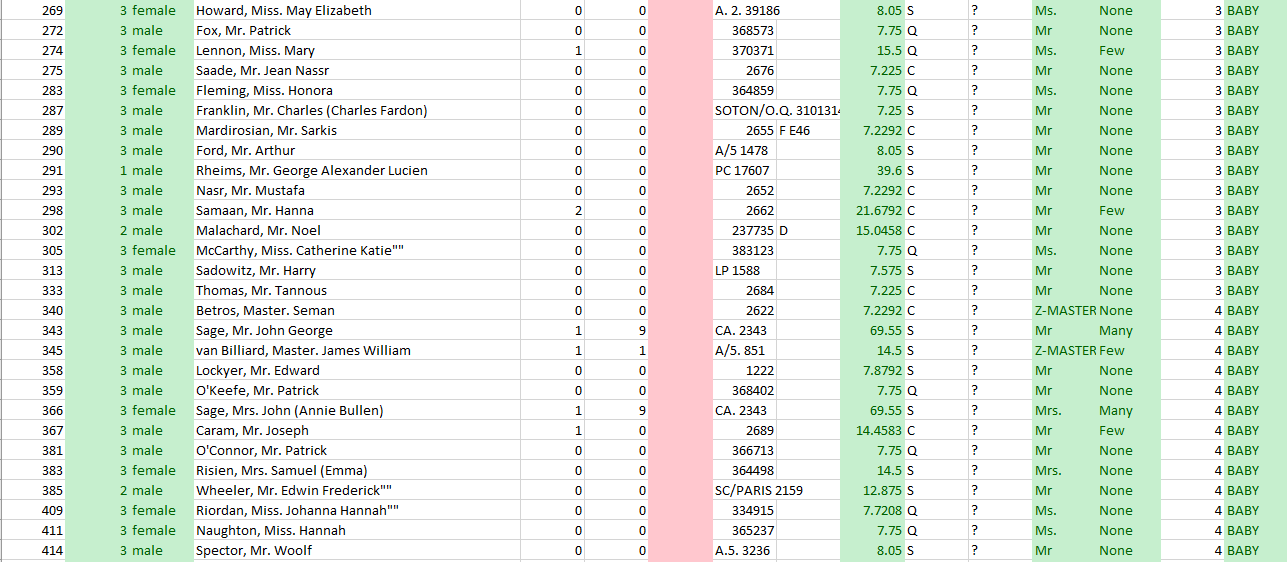


# Appendix 2

### Test Set All Missing Age Instances

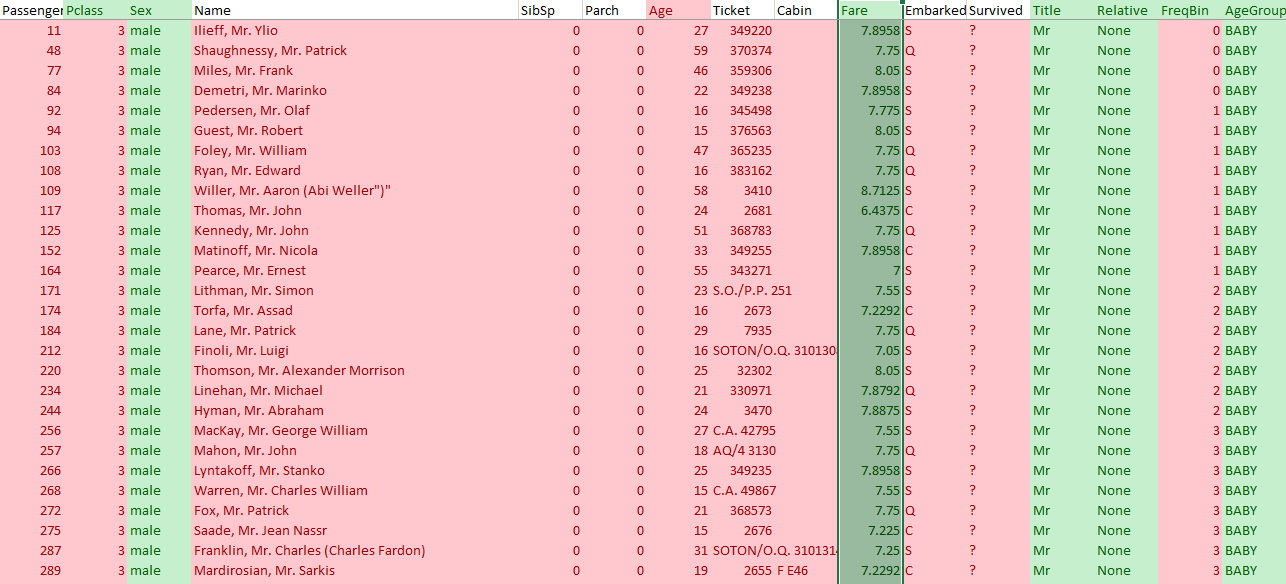


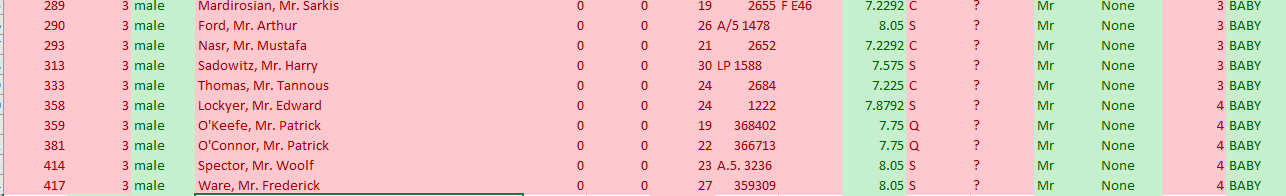






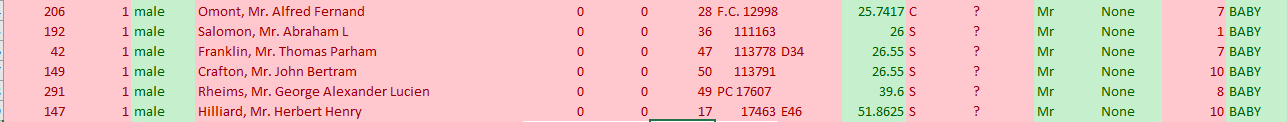
### Test Set Male (Mr.)









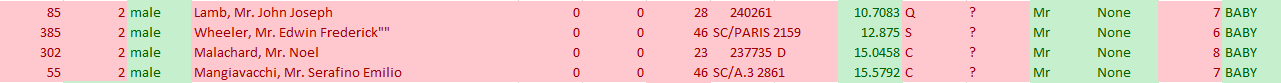




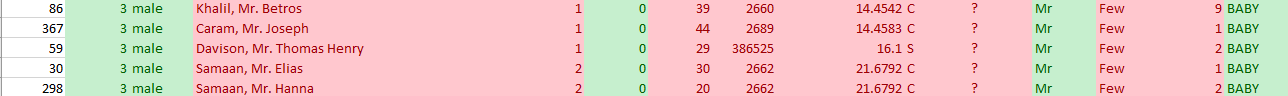






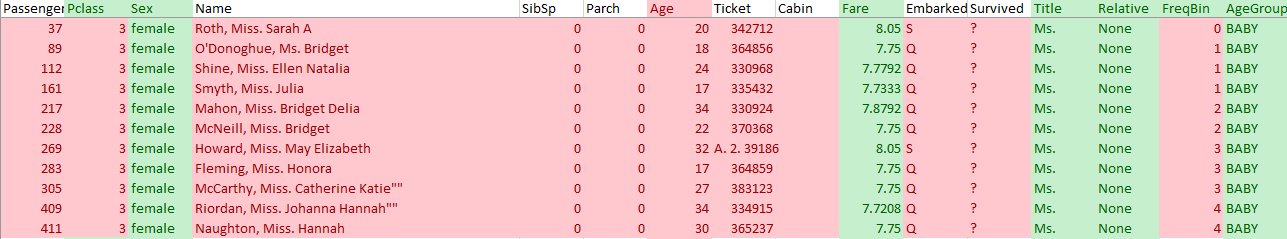








### Test Set Female (Ms.)







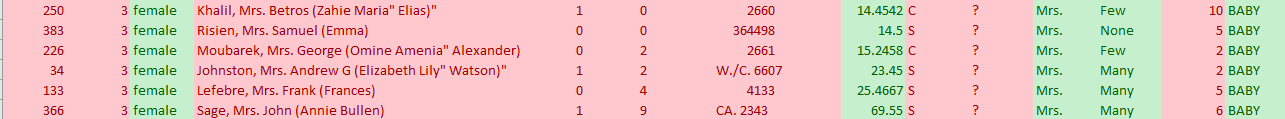




### Test Set Female (Mrs.)















### Test Set Unique





### Test Set Fare Missing

