# Business Intelligence and Data Analytics

## Assignment 2

**Group 51**

| | |
|---|---|
| Sebatian, Jules | 040 983 119 |
| Simpson, Skyler | 040 951 590 |

Submitted Sunday October 24th 2021

# Table of Contents

# Description

The sinking of the Titanic is a well recorded historical event. A ship of great proportions, heralded to be completely unsinkable, sank on its maiden voyage, carrying a large number of passengers. There were not enough safety vessels to save all the passengers. Notoriously, Women and Children were prioritized during the rescue effort, and the dispatching of the safety boats. Since this event occurred, data pertaining to the survival of passengers has been scrapped and analysed many times. This document provides no novel analysis. The goal of this document is to predict the survival rate of any given passenger.

The data we have is incomplete, but sufficient for our amateur purposes. For details pertaining to the attributes of this dataset, see the ***Unmodified Data Table*** included in the following section.Certain fields are missing data, and much of the data provides no value to our analysis.

Data of no use will be discarded. Some data will be aggregated and categorized into new nominal categories. After processing the data, it will be analyzed, visualized and then compared to the findings of data-scientists.

# Tables of Attributes

## Unmodified (Raw Data)

| **Name** | **Type** | **Description** | **Relevant** |
|---|---|---|---|
| Passenger ID | numeric | Unique identifier | No |
| Survived | binary | Describes whether survived, or not | Yes |
| Pclass | Nominal (1-3) | Passenger Class (lower number = higher class) | Yes |
| Name | string | Full Name and Title | No |
| Sex | nominal | Male or Female | Yes |
| Age | numeric | Age in years | Yes |
| SibSP | numeric | Number of Siblings | Yes |
| Parch | numeric | Number of Parents | Yes |
| Ticket | string | Ticket ID | No |
| Fare | numeric | Cost of Ticket | No |
| Embark | string | Designates location of embarkment | Yes |
| Cabin | string | Cabin Given to High Class Passengers | No |

## Relevant

| **Name** | **Reason for Inclusion** |
|---|---|
| Pclass | The passenger class |
| Sex | Women were prioritized on safety vessels, and so evaluating how much sex impacted survability is critical. |
| Embark | Character describing the embarkment location.<br>[C = Cherbourg, Q = Queenstown, S = Southampton] |
| Survived | Considered for comparison with predicted value. |
| SibSp | Used to derive Relatives Attribute |
| Parch | Used to derive Relative attribute |
| Age | Used to derive age Group Attribute (Warning: Contains missing fields) |

## Added and Derived

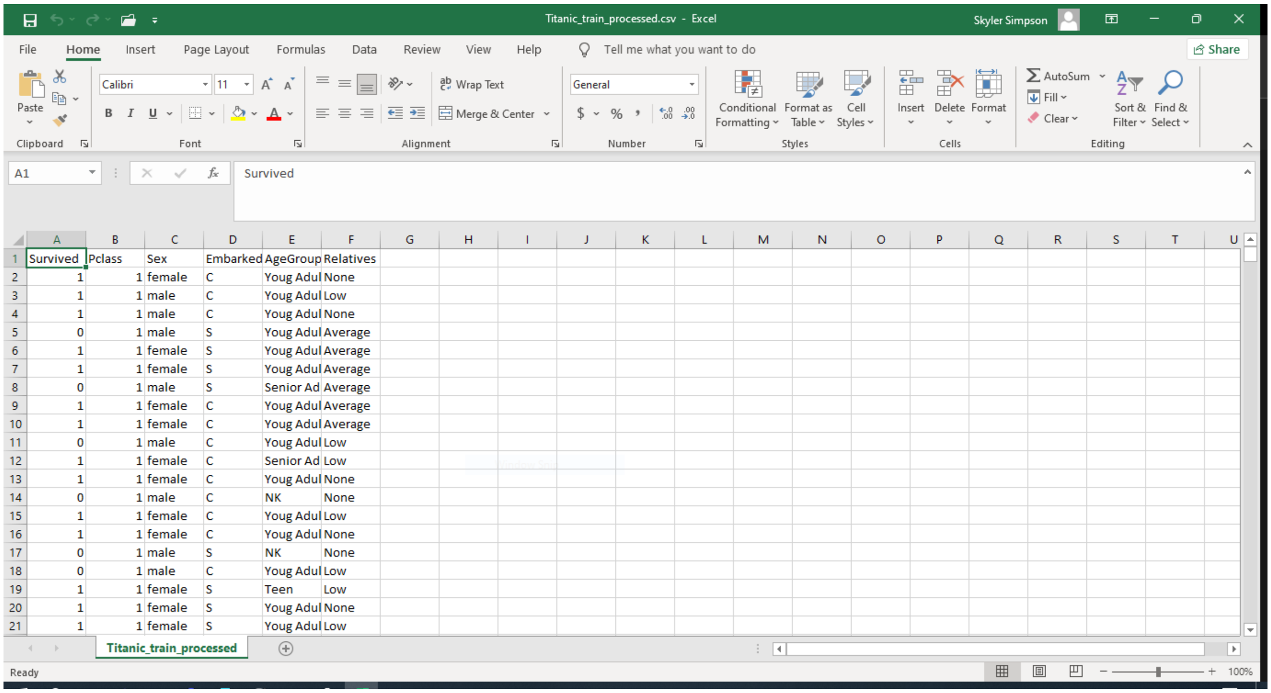
| Name | Derived From | Type | Description |
|---|---|---|---|
| Relative | SibSp and Parch | nominal | Categorization by total number of relatives |
| Age Group | Age | nominal | Categorization by age group |

## Irrelevant and Removed

| Name | Reason for Exclusion |
|---|---|
| Name | Names don't provide relevant statistical data<br>It is worth noting that the Title included in the name may provide some use, however. |
| Age | Age has been abstracted to the nominal category “Age Group” |
| SibSp | Removed after being used to create “Relative” |
| Parch | Removed after being used to create “Relative” |
| Ticket | Ticket information is used to designate tickets from each other. They provide no information on survivability. |
| Fare | The price paid to partake does not inform the treatment of the passenger; it is the passenger class that informs precedence of treatment. |
| Cabin | If a cabin is reserved for an individual, it is recorded here. While having a cabin correlates with the highest passenger class, this category does not itself inform survivability. |
| Passenger ID | A unique identifier does not provide information on survivability |

# Visualizations and Screenshots

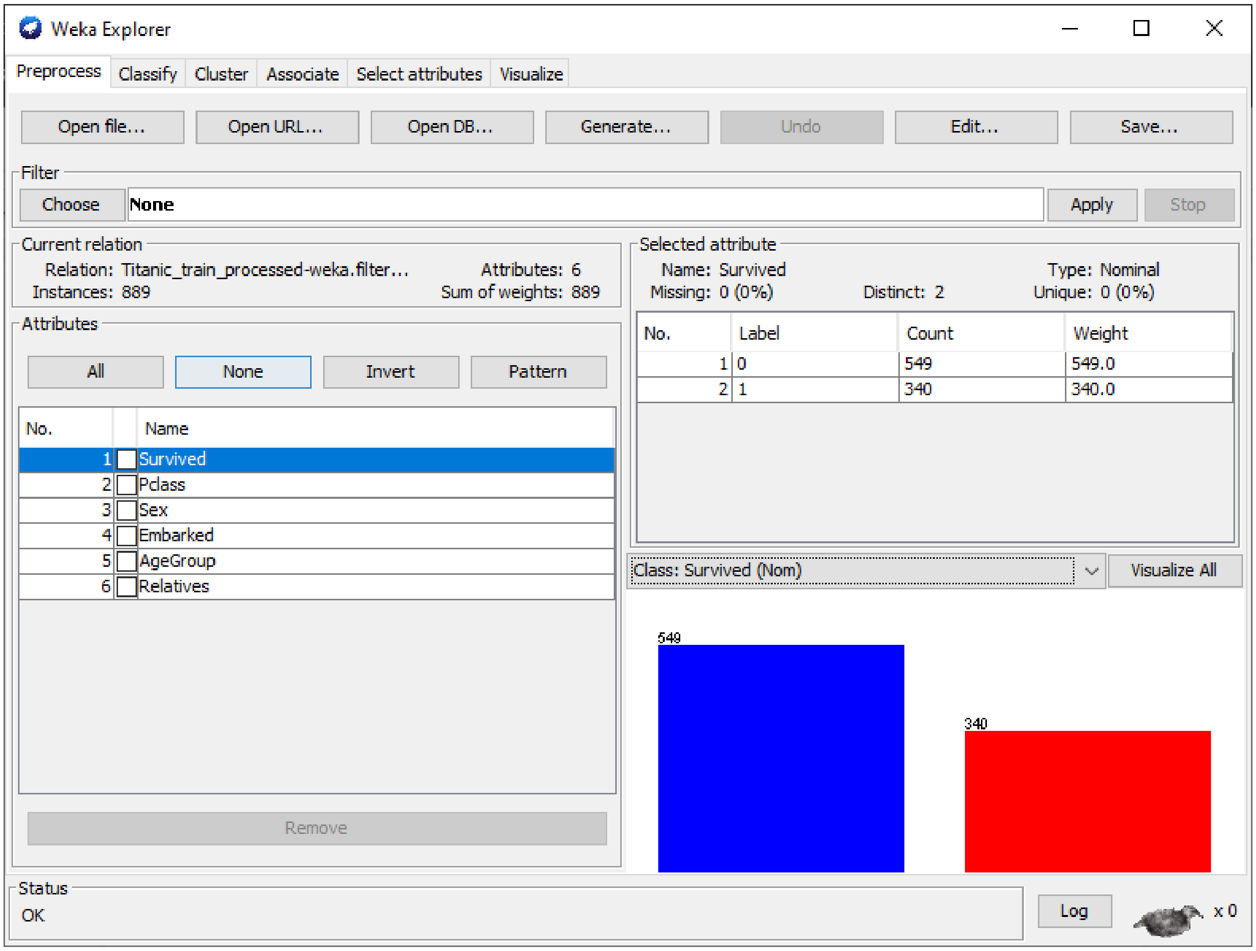
## Titanic_train_processed.csv



|  | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Survived | Pclass | Sex | Embarked | AgeGroup | Relatives |
| 2 | 1 | 1 | female | C | Youg Adul | None |
| 3 | 1 | 1 | male | C | Youg Adul | Low |
| 4 | 1 | 1 | male | C | Youg Adul | None |
| 5 | 0 | 1 | male | S | Youg Adul | Average |
| 6 | 1 | 1 | female | S | Youg Adul | Average |
| 7 | 1 | 1 | female | S | Youg Adul | Average |
| 8 | 0 | 1 | male | S | Senior Ad | Average |
| 9 | 1 | 1 | female | C | Youg Adul | Average |
| 10 | 1 | 1 | female | C | Youg Adul | Average |
| 11 | 0 | 1 | male | C | Youg Adul | Low |
| 12 | 1 | 1 | female | C | Senior Ad | Low |
| 13 | 1 | 1 | female | C | Youg Adul | None |
| 14 | 0 | 1 | male | C | NK | None |
| 15 | 1 | 1 | female | C | Youg Adul | Low |
| 16 | 1 | 1 | female | C | Youg Adul | None |
| 17 | 0 | 1 | male | S | NK | None |
| 18 | 0 | 1 | male | C | Youg Adul | Low |
| 19 | 1 | 1 | female | S | Teen | Low |
| 20 | 1 | 1 | female | S | Youg Adul | None |
| 21 | 1 | 1 | female | S | Youg Adul | Low |

*View of data after unused attributes have been purged, and new attributes have been included.*

## Distribution of the Class Attribute (Survived)



Filter: None

Current relation

Relation: Titanic_train_processed-weka.filter... Attributes: 6

Instances: 889 Sum of weights: 889

Attributes

| No. | | Name |
|---|---|---|
| 1 | □ | Survived |
| 2 | □ | Pclass |
| 3 | □ | Sex |
| 4 | □ | Embarked |
| 5 | □ | AgeGroup |
| 6 | □ | Relatives |

Selected attribute

Name: Survived Type: Nominal

Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

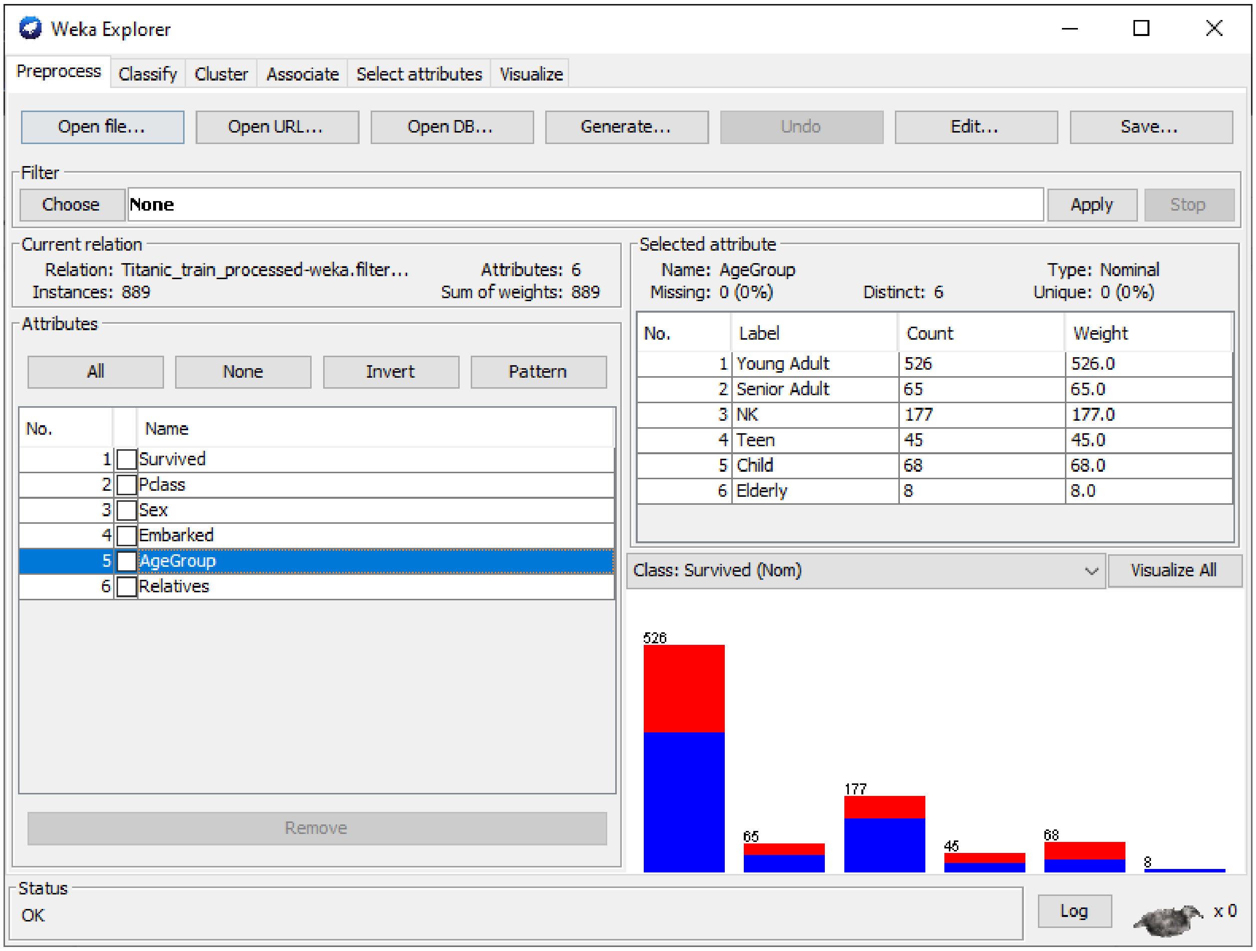
| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | 0 | 549 | 549.0 |
| 2 | 1 | 340 | 340.0 |

Class: Survived (Nom)

Status

OK

# Distribution of the Age Group Attribute



Filter: None

Current relation
Relation: Titanic_train_processed-weka.filter... Attributes: 6
Instances: 889 Sum of weights: 889

Attributes

| No. | Name |
|---|---|
| 1 | □ Survived |
| 2 | □ Pclass |
| 3 | □ Sex |
| 4 | □ Embarked |
| 5 | □ AgeGroup |
| 6 | □ Relatives |

Selected attribute
Name: AgeGroup Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Young Adult | 526 | 526.0 |
| 2 | Senior Adult | 65 | 65.0 |
| 3 | NK | 177 | 177.0 |
| 4 | Teen | 45 | 45.0 |
| 5 | Child | 68 | 68.0 |
| 6 | Elderly | 8 | 8.0 |

Class: Survived (Nom)

Status
OK

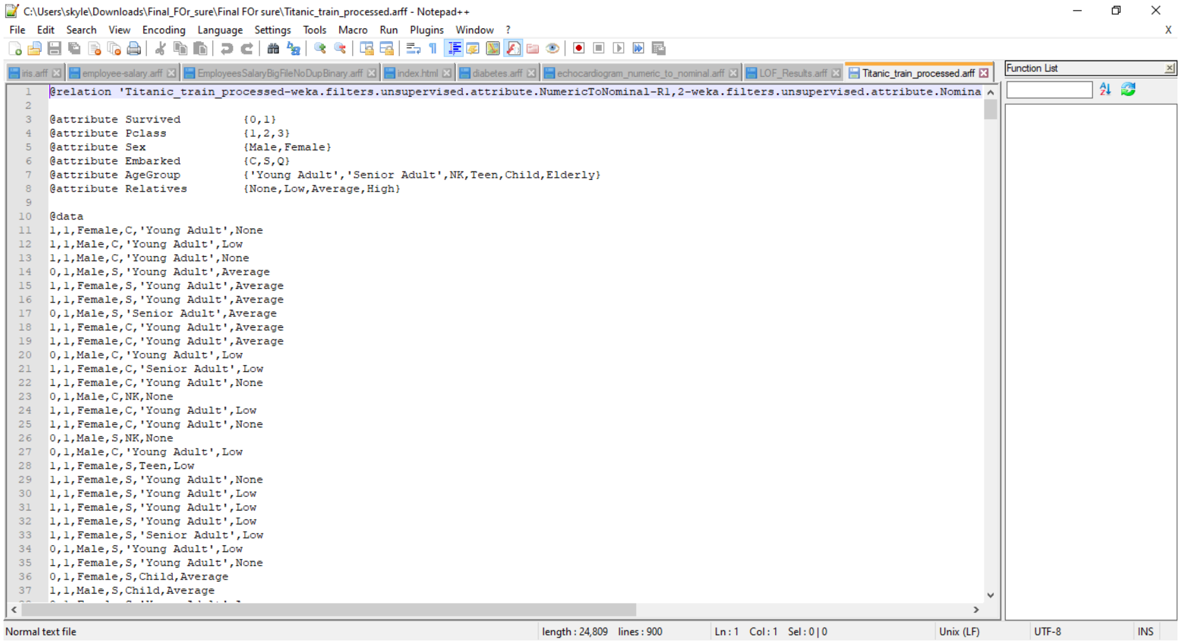
*Visualisation of the distribution of data the nominal age categories. Not Known is the 3rd most popular category.*

**Consider:** The majority of the passengers were categorised by age as Young Adults.

## Age Group Interpretation

| Name | Minimum Age (years) | Maximum Age (years) |
|---|---|---|
| Child | 0 | 12 |
| Teen | 12+ | 18 |
| Young Adult | 18+ | 50 |
| Senior Adult | 50+ | 65 |
| Elderly | 65+ | - |
| NK (Not Known) | ? | ? |

## Titanic_train_processed.arff


C:\Users\skyle\Downloads\Final_FOr_sure\Final FOr sure\Titanic_train_processed.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
iris.arff
employee-salary.arff
EmployeesSalaryBigFileNoDupBinary.arff
index.html
diabetes.arff
echocardiogram_numeric_to_nominal.arff
LOF_Results.arff
Titanic_train_processed.arff
Function List
Normal text file
length : 24,809 lines : 900
Ln : 1 Col : 1 Sel : 0 | 0
Unix (LF)
UTF-8
INS

```
@relation 'Titanic_train_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1,2-weka.filters.unsupervised.attribute.Nomina

@attribute Survived          {0,1}
@attribute Pclass            {1,2,3}
@attribute Sex               {Male,Female}
@attribute Embarked          {C,S,Q}
@attribute AgeGroup          {'Young Adult','Senior Adult',NK,Teen,Child,Elderly}
@attribute Relatives         {None,Low,Average,High}

@data
1,1,Female,C,'Young Adult',None
1,1,Male,C,'Young Adult',Low
1,1,Male,C,'Young Adult',None
0,1,Male,S,'Young Adult',Average
1,1,Female,S,'Young Adult',Average
1,1,Female,S,'Young Adult',Average
0,1,Male,S,'Senior Adult',Average
1,1,Female,C,'Young Adult',Average
1,1,Female,C,'Young Adult',Average
0,1,Male,C,'Young Adult',Low
1,1,Female,C,'Senior Adult',Low
1,1,Female,C,'Young Adult',None
0,1,Male,C,NK,None
1,1,Female,C,'Young Adult',Low
1,1,Female,C,'Young Adult',None
0,1,Male,S,NK,None
0,1,Male,C,'Young Adult',Low
1,1,Female,S,Teen,Low
1,1,Female,S,'Young Adult',None
1,1,Female,S,'Young Adult',Low
1,1,Female,S,'Young Adult',Low
1,1,Female,S,'Young Adult',Low
1,1,Female,S,'Senior Adult',Low
0,1,Male,S,'Young Adult',Low
1,1,Female,S,'Young Adult',None
0,1,Female,S,Child,Average
1,1,Male,S,Child,Average
```

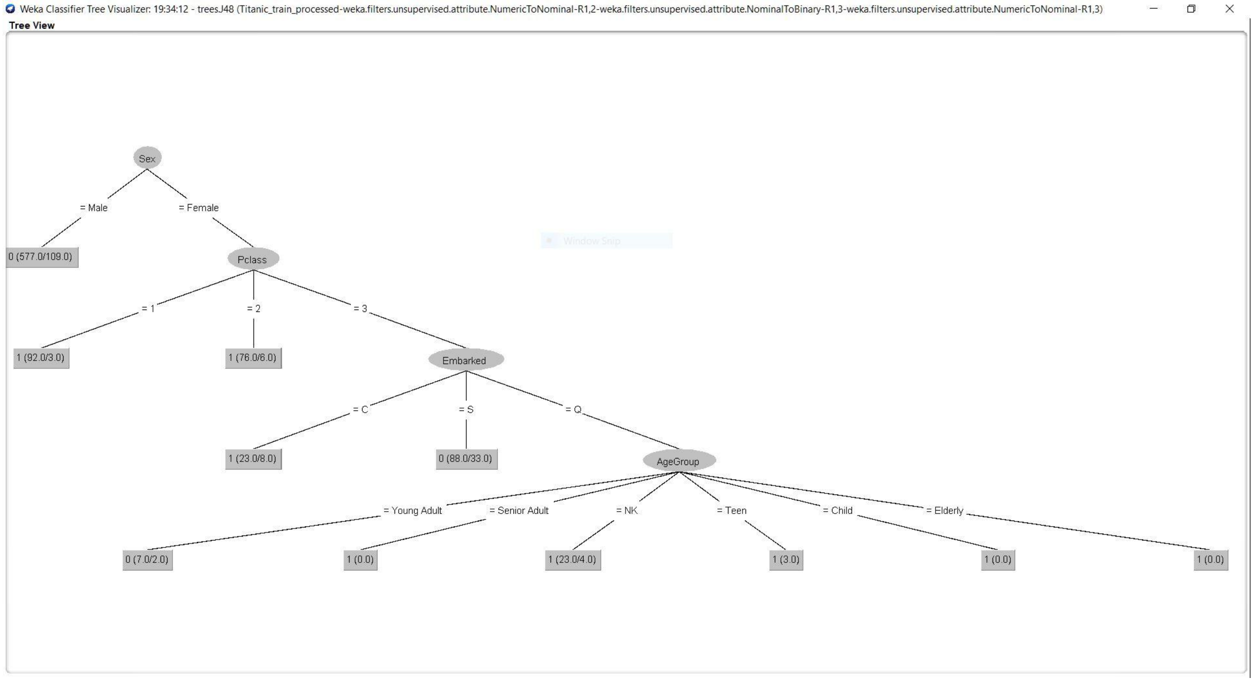
# Results

## Confusion Matrix

### Minimum Number of Objects Equals Two (2)

```
a      b      <--    classified as
519    30     |      a = 0
140    200    |      b = 1
```

# Decision Tree



Weka Classifier Tree Visualizer: 19:34:12 - trees.J48 (Titanic_train_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1,2-weka.filters.unsupervised.attribute.NominalToBinary-R1,3-weka.filters.unsupervised.attribute.NumericToNominal-R1,3)

## Interpretation of the Tree

**How to read the tree:**

- If your Sex is Male,
  - You are predicted to die
- If your Sex is Female,
  - If you are in the Highest Passenger Class,
    - you are predicted to live
  - If you are in the Middle passenger Class,
    - you are predicted to live
  - If you are in the Lowest Passenger Class,
    - If you embarked from C,
      - You are predicted to live
    - If you embarked from S,
      - You are predicted to die
    - If you embarked from Q,
      - If you are a Young Adult,
        - You are predicted to die
      - Otherwise,
        - You are predicted to live

## Explanation of Anomalous Results

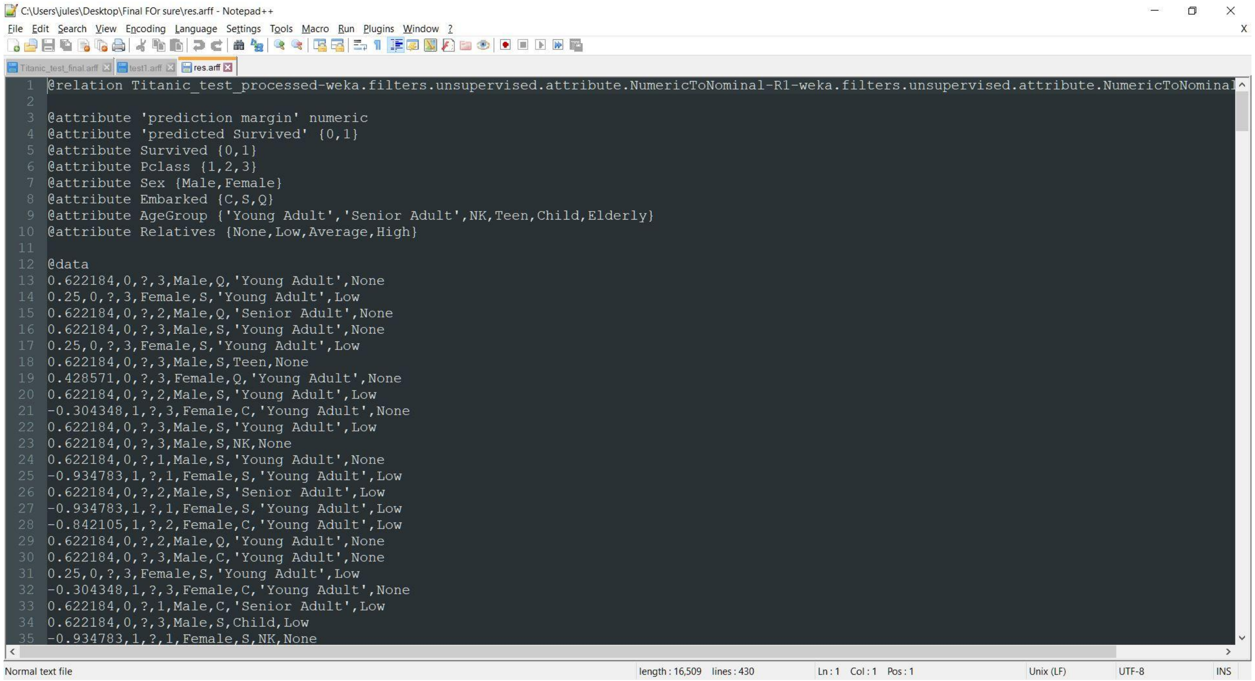
The first step in the classification was done by branching out on the basis of sex. The model predicted all the men to die in the crash, 577 of which were correctly classified and 109 were incorrectly classified. We believe this is due to the sample data which was selected for interpretation.

Since, all men were predicted to die according to the tree, it only branched under the women category. The women category was further classified on the basis of the class that the passengers were travelling in.

We have a high certainty (97%) that all the women in the highest passenger class survived, due to being given precedence during the evacuation. We have a similar certainty (93%) that women in the middle passenger class also survived. In the lowest class, according to our model, the predicted survival rate seems to have depended on the embarkment location.

According to the model, the passengers from Cherbourg had a higher rate of survivability compared to Southampton. For the passengers from Queenstown Age Group was also uniquely a factor.

# res.arff



C:\Users\jules\Desktop\Final FOr sure\res.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Titanic_test_final.arff
test1.arff
res.arff

```
@relation Titanic_test_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1-weka.filters.unsupervised.attribute.NumericToNominal

@attribute 'prediction margin' numeric
@attribute 'predicted Survived' {0,1}
@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {Male,Female}
@attribute Embarked {C,S,Q}
@attribute AgeGroup {'Young Adult','Senior Adult',NK,Teen,Child,Elderly}
@attribute Relatives {None,Low,Average,High}

@data
0.622184,0,?,3,Male,Q,'Young Adult',None
0.25,0,?,3,Female,S,'Young Adult',Low
0.622184,0,?,2,Male,Q,'Senior Adult',None
0.622184,0,?,3,Male,S,'Young Adult',None
0.25,0,?,3,Female,S,'Young Adult',Low
0.622184,0,?,3,Male,S,Teen,None
0.428571,0,?,3,Female,Q,'Young Adult',None
0.622184,0,?,2,Male,S,'Young Adult',Low
-0.304348,1,?,3,Female,C,'Young Adult',None
0.622184,0,?,3,Male,S,'Young Adult',Low
0.622184,0,?,3,Male,S,NK,None
0.622184,0,?,1,Male,S,'Young Adult',None
-0.934783,1,?,1,Female,S,'Young Adult',Low
0.622184,0,?,2,Male,S,'Senior Adult',Low
-0.934783,1,?,1,Female,S,'Young Adult',Low
-0.842105,1,?,2,Female,C,'Young Adult',Low
0.622184,0,?,2,Male,Q,'Young Adult',None
0.622184,0,?,3,Male,C,'Young Adult',None
0.25,0,?,3,Female,S,'Young Adult',Low
-0.304348,1,?,3,Female,C,'Young Adult',None
0.622184,0,?,1,Male,C,'Senior Adult',Low
0.622184,0,?,3,Male,S,Child,Low
-0.934783,1,?,1,Female,S,NK,None
```

Normal text file    length : 16,509    lines : 430    Ln : 1    Col : 1    Pos : 1    Unix (LF)    UTF-8    INS

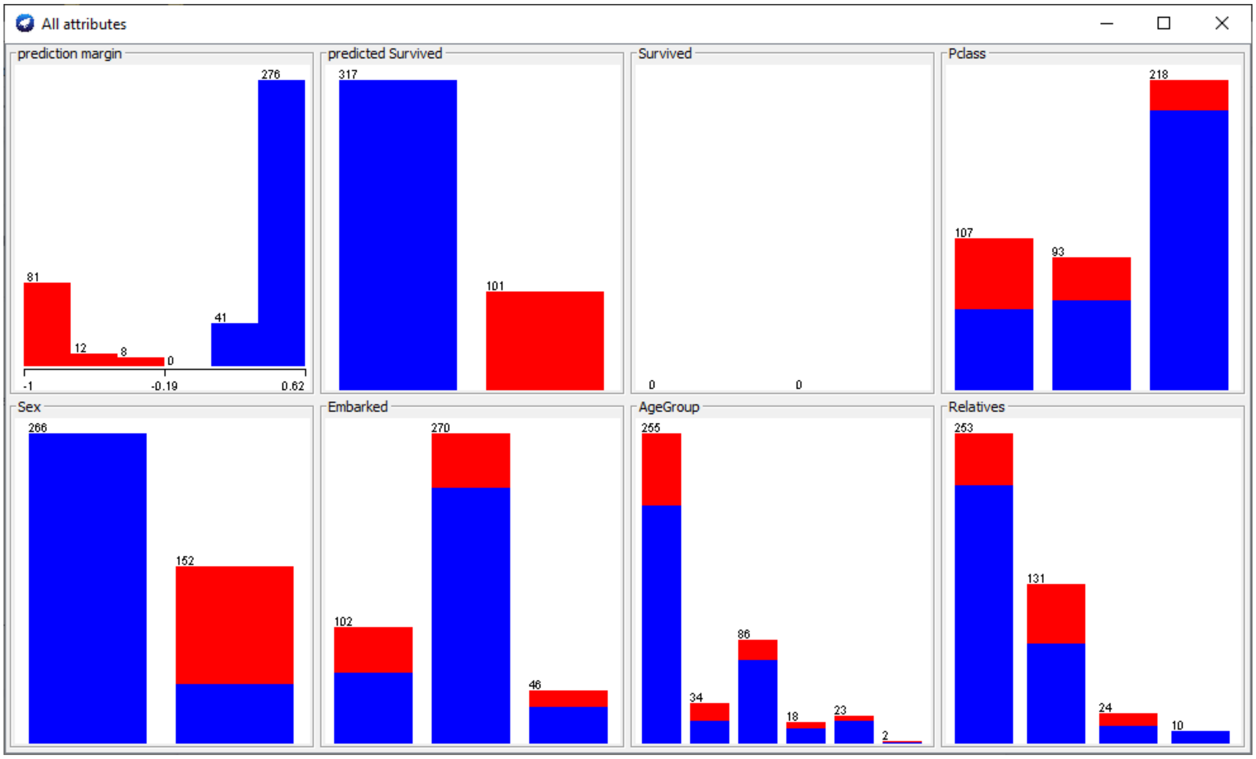
# Findings

a. Total instances in the test file: **418**
b. Number of persons predicted to survive (1): **101**
c. Number of persons predicted not to survive (0): **317**
d. Percentage of predicted survival: **24.162%**

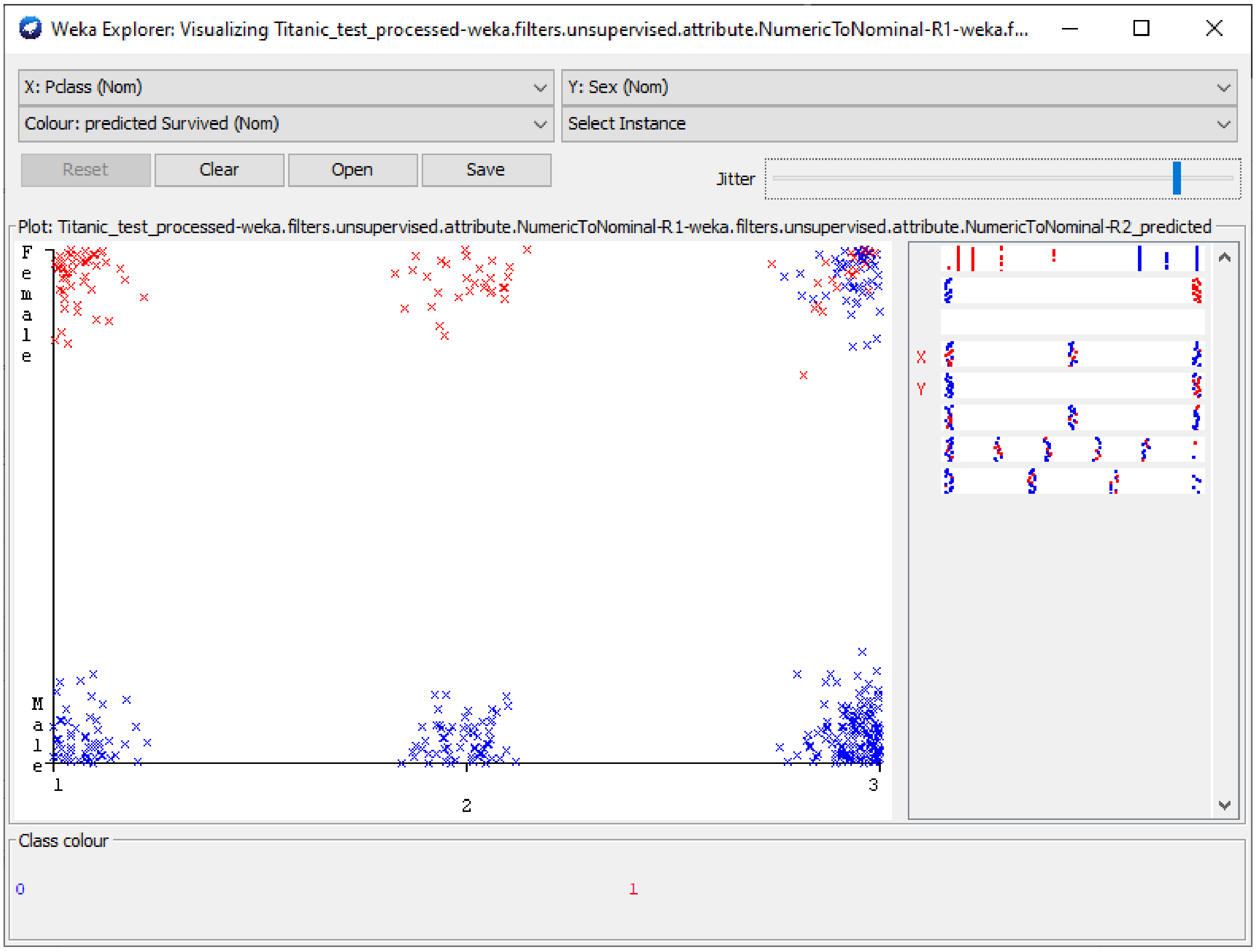
## Table of Expectations

| | **Our Result** | **Expectation (True Value)** |
|---|---|---|
| **Passengers Survived (%)** | 24.162 | 37 |
| **First Class Survived (%)** | 49.504 | 61 |
| **Second Class Survived (%)** | 29.703 | 42 |
| **Third Class Survived (%)** | 20.792 | 24 |
| **Male Survived (%)** | 0 | 20 |
| **Female Survived (%)** | 100 | 75 |

## Predicted Survival Rate (Visualized)



# Overview of Sex and Class Upon Predicted Survival



Weka Explorer: Visualizing Titanic_test_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1-weka.f...

X: Pclass (Nom)

Y: Sex (Nom)

Colour: predicted Survived (Nom)

Select Instance

Plot: Titanic_test_processed-weka.filters.unsupervised.attribute.NumericToNominal-R1-weka.filters.unsupervised.attribute.NumericToNominal-R2_predicted

Class colour

*Red means the passenger is predicted to have survived, and Blue predicts the passenger died. The above graph includes an elevated jitter for easier viewing.*

## Summary of Findings

Our resulting predictions suggest that the primary correlation for survival was sex, and the second corelation was class. Reiterated, ***all men were predicted to die irrespective of class, and women were predicted to survive on the basis of their class.***

## Discussion

Before we mention differences, we should first mention that the numbers which we present include a certain chance of deviation.

There are many reasons why our results might differ from our expected results.

Our tested dataset is very small, and so the predicted outcome of each passenger has a greater outcome on the numbers as a whole.

If we followed the same procedure as the data scientists that interpreted the data initially, we could isolate the difference in our results as being from the selection of data. We do not know the process which generated the results to which we are comparing.

We also possibly introduced new errors: we aggregated data(See table of Added and Derived Attributes above) to generate new nominal attributes , which could affect the accuracy of our results: in this specific case, all conversions of numeric to nominal lead to a loss of data (see precision).

We are left with the belief that a more accurate model could have been generated using the more specific numeric fields. To test this idea, we suggest a way to solve for the missing age data and include our findings below, in the Additional Analysis Section.

# Additional Analysis

## Check impact of missing Age values

We noticed that a great deal of our data was missing age data. As previously mentioned in this document, the third most populated age group was Not Known (NK).

To attempt to predict the missing age values, we extracted the titles from the Raw Data Names. Titles imply certain aspects about social status that correlate with age, we used central tendencies to predict numerical missing age data.

We also opted to use the original numeric SibSP and Parch Attributes, in preference to the ordinal Relatives Attribute.

After performing these transformations, we attempted the process of retraining the data based on the populated numeric age attribute, rather than the nominal age-group attribute to see how it would impact our findings.

## Findings

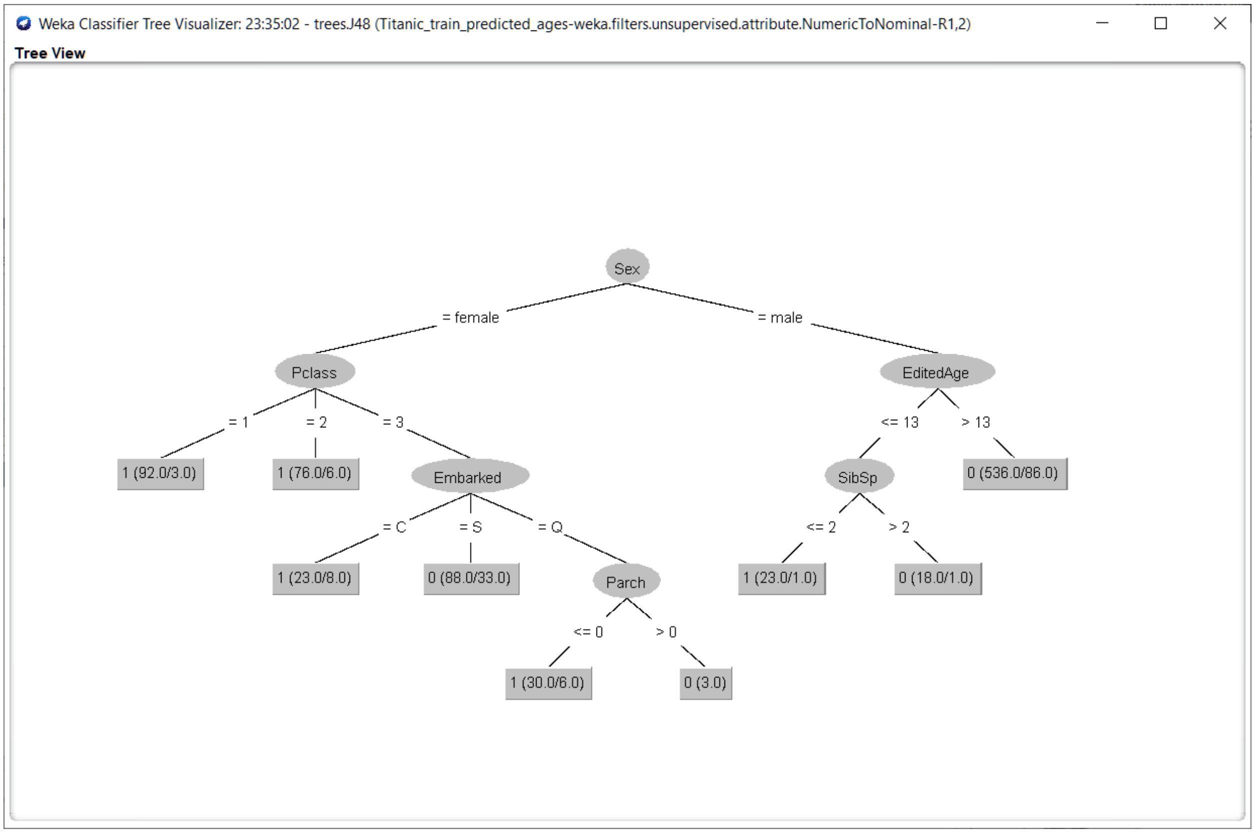
| | |
|---|---|
| Total instances in the test file: | **418** |
| b. Number of persons predicted to survive (1): | **127** |
| c. Number of persons predicted not to survive (0): | **291** |
| d. Percentage of predicted survival: | **30.383%** |

## Table of Expectations

| | **Our Result** | **Expectation (True Value)** |
|---|---|---|
| **Passengers Survived (%)** | 30.383 | 37 |
| **First Class Survived (%)** | 39.370 | 61 |
| **Second Class Survived (%)** | 23.622 | 42 |
| **Third Class Survived (%)** | 24.409 | 24 |
| **Male Survived (%)** | 12.599 | 20 |
| **Female Survived (%)** | 87.402 | 75 |

*Files pertaining to these findings will also be included alongside this document.*

# Decision Tree for Adjusted Model



Weka Classifier Tree Visualizer: 23:35:02 - trees.J48 (Titanic_train_predicted_ages-weka.filters.unsupervised.attribute.NumericToNominal-R1,2)

*Consider that the requirements for predicting life and death in this model are more complex for males than in the previous model.*