HARVARD EXTENSION SCHOOL

EXT CSCI E-106 Model Data Class Group Project Template

Author: Dinesh Bedathuru Author: Brian Calderon
Author: Jeisson Hernandez Author: Hao Fu Author: Derek Rush
Author: Jeremy Tajonera Author: Catherine Tully

06 May 2025

In this project, our aim is to classify the probability of a passenger surviving the Titanic crash of 1912. We used a variety of linear and non-linear models to deduce the most accurate model and provide long-term stability in our predictions.

Table of contents

List of Figures

List of Tables

Classify whether a passenger on board the maiden voyage of the RMS Titanic in 1912 survived given their age, sex and class. Sample-Data-Titanic-Survival.csv to be used in the Final Project

Variable	Description
pclass	**Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)**
survived	**Survival (0 = No; 1 = Yes)**
name	**Name**
sex	**Sex**
age	**Age**
sibsp	**# of siblings / spouses aboard the Titanic**
parch	**# of parents / children aboard the Titanic**
ticket	**Ticket number**
fare	**Passenger fare**
cabin	**Cabin number**
embarked	**Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)**
boat	**Lifeboat ID, if passenger survived**
body	**Body number (if passenger did not survive and body was recovered**
home.dest	**The intended home destination of the passenger**

1 Instructions:

0. Join a team with your fellow students with appropriate size (Up to Nine Students total) If you have not group by the end of the week of April 11 you may present the project by yourself or I will randomly assign other stranded student to your group. I will let know the final groups in April 11.

1. Load and Review the dataset named "Titanic_Survival_Data.csv" 2. Create the train data set which contains 70% of the data and use set.seed (15). The remaining 30% will be your test data set.

3. Investigate the data and combine the level of categorical variables if needed and drop variables as needed. For example, you can drop id, Latitude, Longitude, etc.

4. Build appropriate model to predict the probability of survival.

5. Create scatter plots and a correlation matrix for the train data set. Interpret the possible relationship between the response.

6. Build the best models by using the appropriate selection method. Compare the performance of the best logistic linear models.

7. Make sure that model assumption(s) are checked for the final model. Apply remedy measures (transformation, etc.) that helps satisfy the assumptions.

8. Investigate unequal variances and multicollinearity.

9. Build an alternative to your model based on one of the following approaches as applicable to predict the probability of survival: logistic regression, classification Tree, NN, or SVM. Check the applicable model assumptions. Explore using a negative binomial regression and a Poisson regression.

10. Use the test data set to assess the model performances from above.

11. Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model.

12. Create a model development document that describes the model following this template, input the name of the authors, Harvard IDs, the name of the Group, all of your code and calculations, etc..

Due Date: May 12 2025 1159 pm hours EST Notes No typographical errors, grammar mistakes, or misspelled words, use English language All tables need to be numbered and describe their content in the body of the document All figures/graphs need to be numbered and describe their content All results must be accurate and clearly explained for a casual reviewer to fully understand their purpose and impact Submit both the RMD markdown file and PDF with the sections with appropriate explanations. A more formal.

2 Executive Summary

This section will describe the model usage, your conclusions and any regulatory and internal requirements. In a real world scneario, this section is for senior management who do not need to know the details. They need to know high level (the purpose of the model, limitations of the model and any issues).

3 I. Introduction (5 points)

This section needs to introduce the reader to the problem to be resolved, the purpose, and the scope of the statistical testing applied. What you are doing with your prediction? What is the purpose of the model? What methods were trained on the data, how large is the test sample, and how did you build the model?

The Titanic was a British-registered ship that set sail on its maiden voyage on April 10th, 1912 with 2,240 passengers and crew on board. On April 15th, 1912, the ship struck an iceberg, split in half, and sank to the bottom of the ocean (National Oceanic and Atmospheric Administration (NOAA), 2023). In this report, we are going to analyze the data in the Titanic.csv file and use it to determine the best model for predicting whether someone on board would live or die. By creating this model, we hope to understand what factors a passenger could have taken into account in order to reduce their risk of death during the trip. We cleaned the data and split into into a train/test split in order to properly train our models. We created simple linear models, multivariate linear models, logistic models (both binomial and poisson), a regression tree, and a neural network model. The train sample size was 916 data points (70.03%) and the test sample size was 392 data points (29.97%). We built the models after examining the data and determining which predictor variables we thought would be most relevant for survival rate. Once we had our variables and training data, we created the models and examined the performance of the models on both training and testing data to determine if they were robust. We also examined if the model assumptions appeared to hold for each model.

4 II. Description of the data and quality

Here you need to review your data, the statistical test applied to understand the predictors and the response and how are they correlated. Extensive graph analysis is recommended. Is the data continuous, or categorical, do any transformation needed? Do you need dummies?

4.1 Loading the data

```
odata <- read.csv("../data/Titanic_Survival_Data.csv")
cat("Size of entire data set:", nrow(odata), "\n")
```

```
Size of entire data set: 1310
```

4.2 Removing un-needed columns

Name: Removing because names have no inference on surivival (inference)

ticket: Ticket No. will also likely not have an influence in survival

boat: This is highly correlated to the survival dependant variable since people who made it on a boat likely survived

body: This is highly correlated to the survival dependant variable since people who's body was recovered did not survive.

home.dest: The destination likely has nothing to do with the survival

```
data.clean = odata[, !(names(odata) %in% c("name", "ticket", "boat","body","home.dest"))]
```

4.3 Data Augmentation

We extracted the deck letter from the cabin since it could potentially correlate to the survival.

```
#Extract deck letter from cabin
data.clean$deck <- substr(data.clean$cabin, 1,1)
# Remove cabin col:
data.clean$cabin <- NULL
```

4.4 Initial Check for Missing values

We see that age and deck have the most amount of missing data, therefore we proceed to impute them.

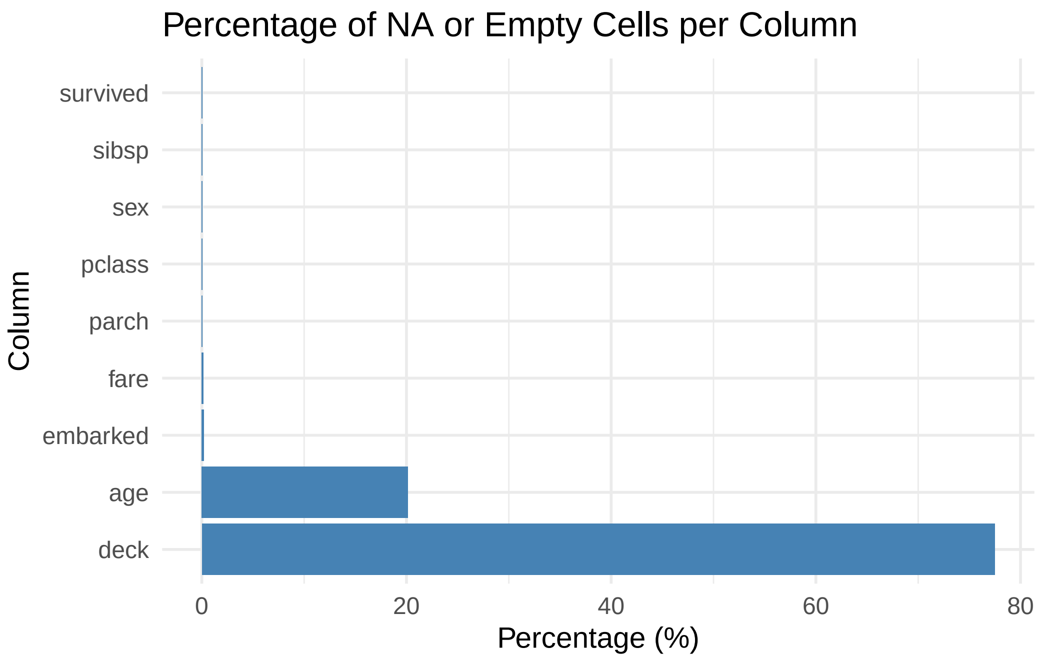


Figure 1: Percentage of Missing Values

4.5 Imputing data

Below we impute Age using the median value in that column.

For deck we use KNN to impute the missing deck values.

After imputing these two columns we can see that the largest amount of missing data is ~0.2% which is quite small and can be removed.

```
# ---- Age----
#Replace NAs in age column with Median value
median_age <- median(data.clean$age, na.rm = TRUE)
data.clean <- data.clean %>%
  mutate(age = ifelse(is.na(age), median_age, age))

# ---- deck----
# For deck, since its a category, we decided to use KNN  to impute the column:

# Install if not already installed
# install.packages("VIM")
library(VIM)
```

```
Warning: package 'VIM' was built under R version 4.4.3
```

```
Loading required package: colorspace
```

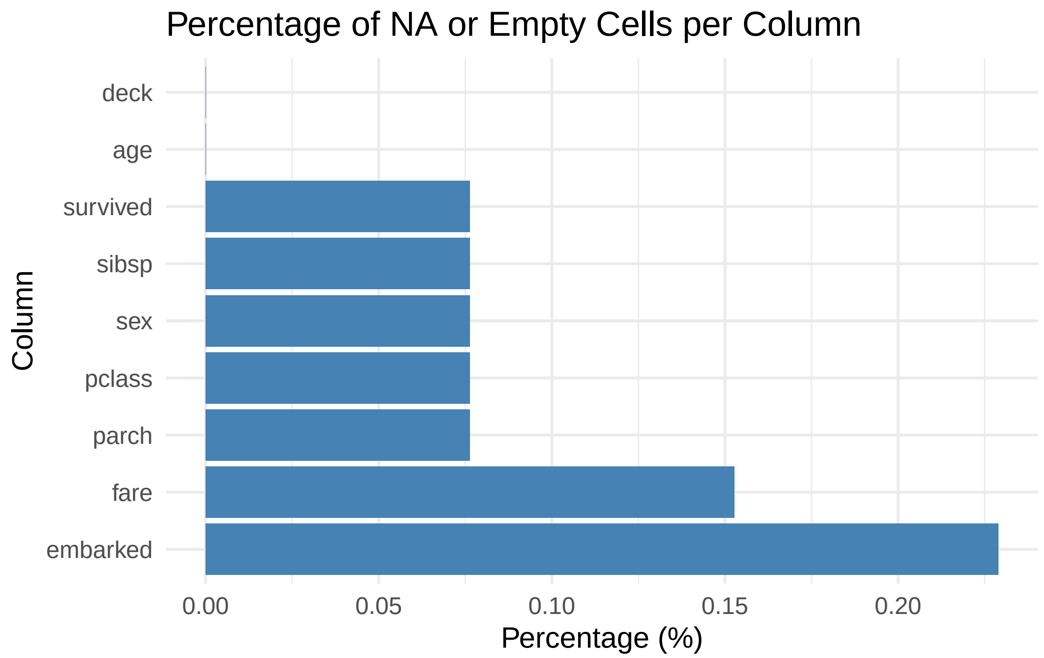
```
Loading required package: grid
```

```
VIM is ready to use.
```

```
Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues


Attaching package: 'VIM'


The following object is masked from 'package:datasets':

    sleep
```

```
# Replace "" with NA in the 'deck' column
data.clean$deck[data.clean$deck == ""] <- NA

# Convert 'cabin' to factor
data.clean$deck <- as.factor(data.clean$deck)

# Apply kNN imputation just to Cabin column
data.clean <- kNN(data.clean, variable = "deck", k = 5)

# Check that NAs were imputed
# sum(is.na(data.clean$deck))        # Original
# sum(is.na(data.clean.imputed$deck)) # After

# Remove indicator col:
data.clean$deck_imp <- NULL
```

```
################################################################################
#          Check for Missing values after Imputation                           #
################################################################################
# Function to calculate % of NA or empty strings per column
#| fig-cap: "Percentage of Missing Values after Imputation"
plot_missing_barchart(data.clean)
```



4.6 Dummifying Columns:

We dummify pclass, sex, embarked and deck. We leave sibsp and parch as continuous variables as we observed that dummifying these columns leads to smaller significance, whilst leaving them as continuous maximizes their contributions to the models explanatory power.

```r
# Dummifying pclass:
data.clean$pclass_1 = ifelse(data.clean$pclass == 1, 1, 0)
data.clean$pclass_2 = ifelse(data.clean$pclass == 2, 1, 0)

# Dummifying sex:
data.clean$sex_M = ifelse(data.clean$sex == 'male', 1, 0)

# Dummifying embarked:
data.clean$embarked_C = ifelse(data.clean$embarked == 'C', 1, 0)
data.clean$embarked_Q = ifelse(data.clean$embarked == 'Q', 1, 0)

# Dummifying deck:
data.clean$deck_A = ifelse(data.clean$deck == 'A', 1, 0)
data.clean$deck_B = ifelse(data.clean$deck == 'B', 1, 0)
data.clean$deck_C = ifelse(data.clean$deck == 'C', 1, 0)
data.clean$deck_D = ifelse(data.clean$deck == 'D', 1, 0)
data.clean$deck_E = ifelse(data.clean$deck == 'E', 1, 0)
data.clean$deck_F = ifelse(data.clean$deck == 'F', 1, 0)
data.clean$deck_G = ifelse(data.clean$deck == 'G', 1, 0)

# Removing Dummified cols:
data.clean = subset(data.clean, select  = -c(pclass, sex, embarked,deck))
```

4.7 Remove NA rows

Below we remove NA rows, which turned out to be only 2 after proper cleaning and imputation.

```r
data.clean = na.omit(data.clean)

cat(nrow(odata) - nrow(data.clean),'rows were removed from original dataset')
```

```
2 rows were removed from original dataset
```

4.8 Divide into Test / Train

Finally we divide into 70% training data and 30% test data.

```r
train_indices = sample(1 : nrow(data.clean), size = 0.70*nrow(data.clean), replace = FALSE)
train = data.clean[train_indices,]
test = data.clean[-train_indices,]
cat("We are using:", nrow(train)/nrow(data.clean) * 100, '% of the data for training')
```

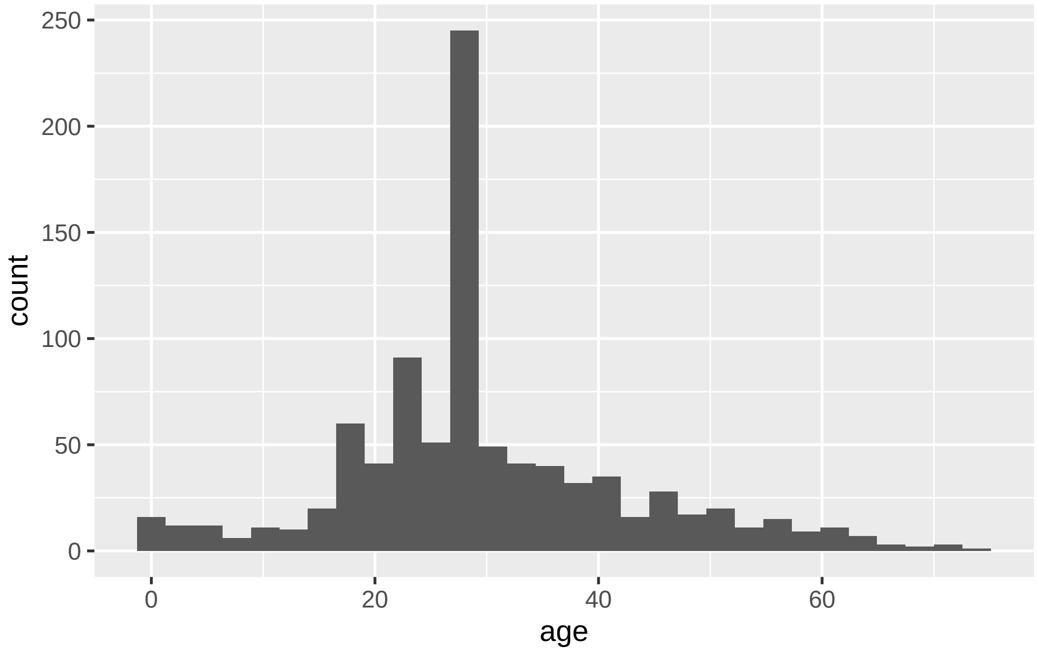
```
We are using: 69.95413 % of the data for training
```

4.9 EDA

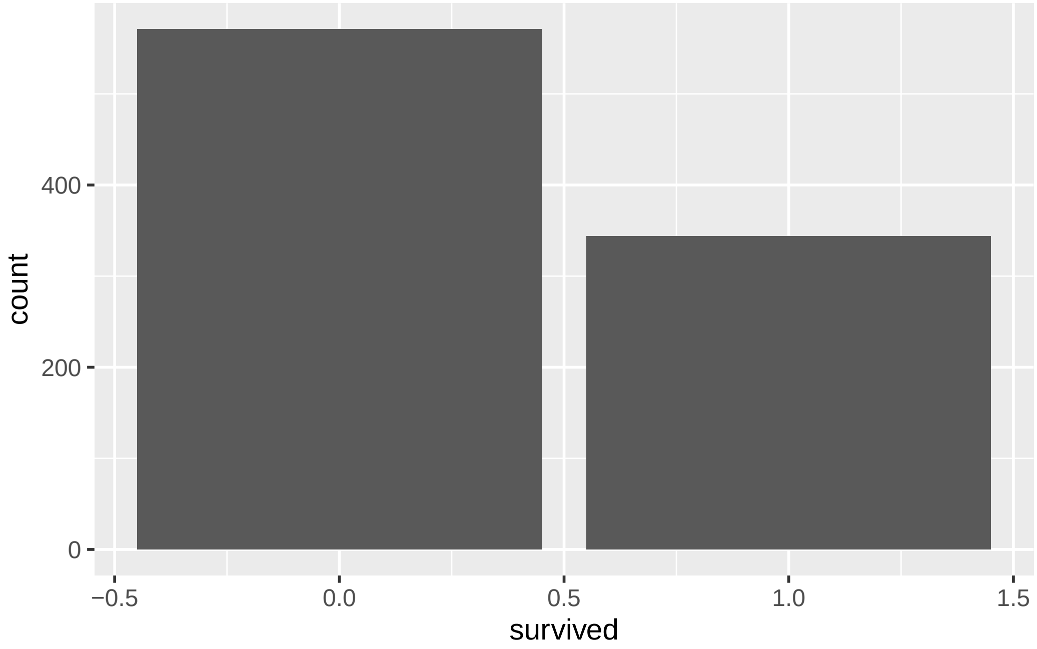
Using the training data set we use a variety of method to draw some initial conclusions:

- Histogram: Showing that more people in their late teens up to late thirties survived.

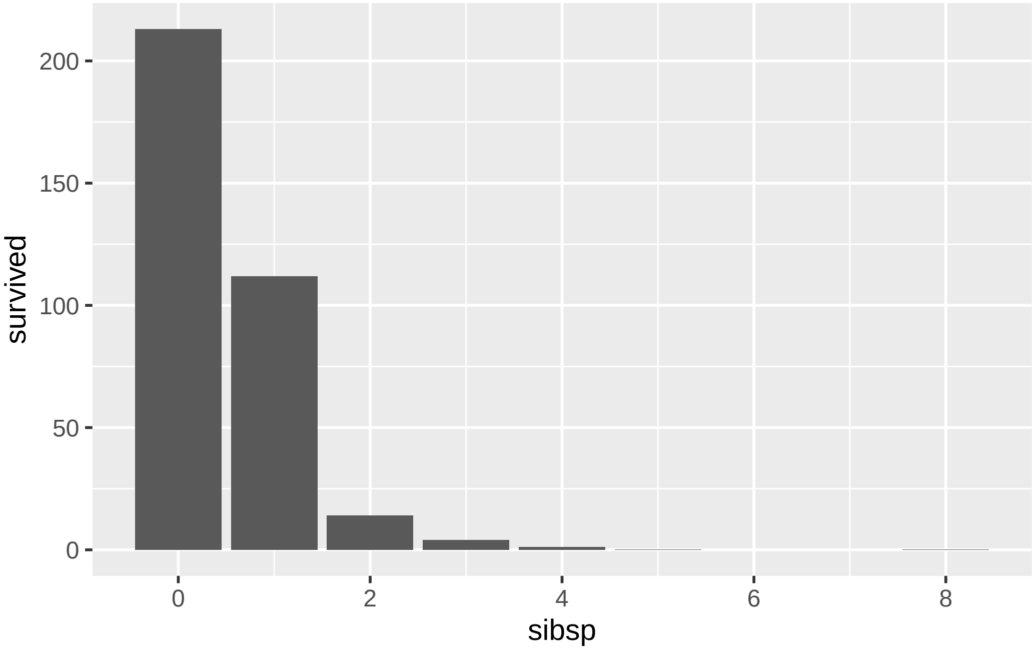
```
# Histogram showing that more people in their late teens up to late thirties survived.
ggplot(train, aes(age)) +
  geom_histogram(bins=30)
```



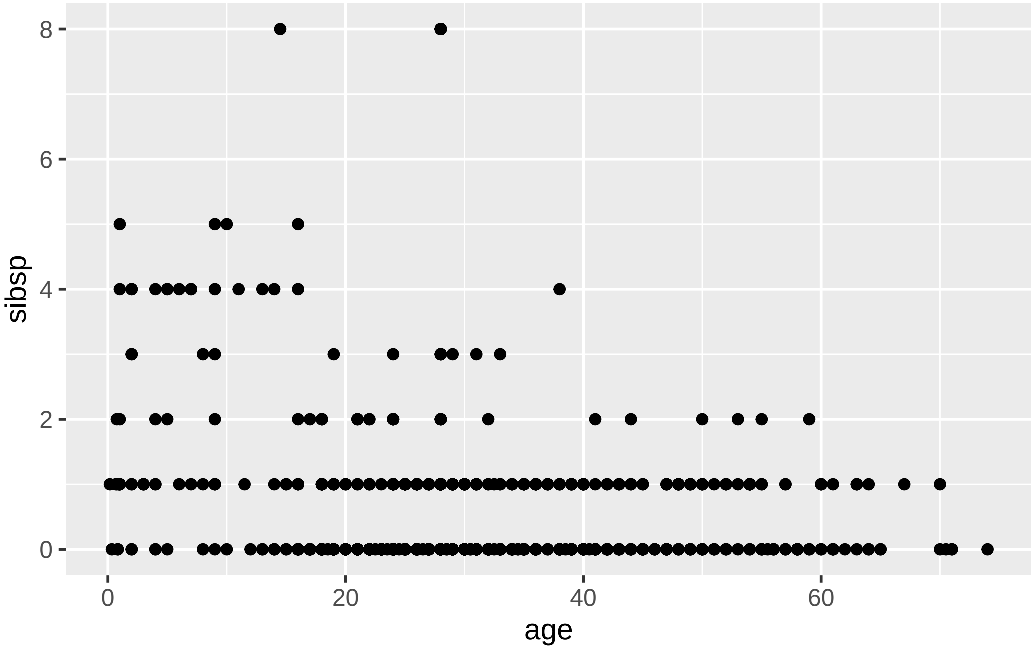
```
# Bar chart showing that more females survived than males.
ggplot(train, aes(survived)) +
  geom_bar()
```



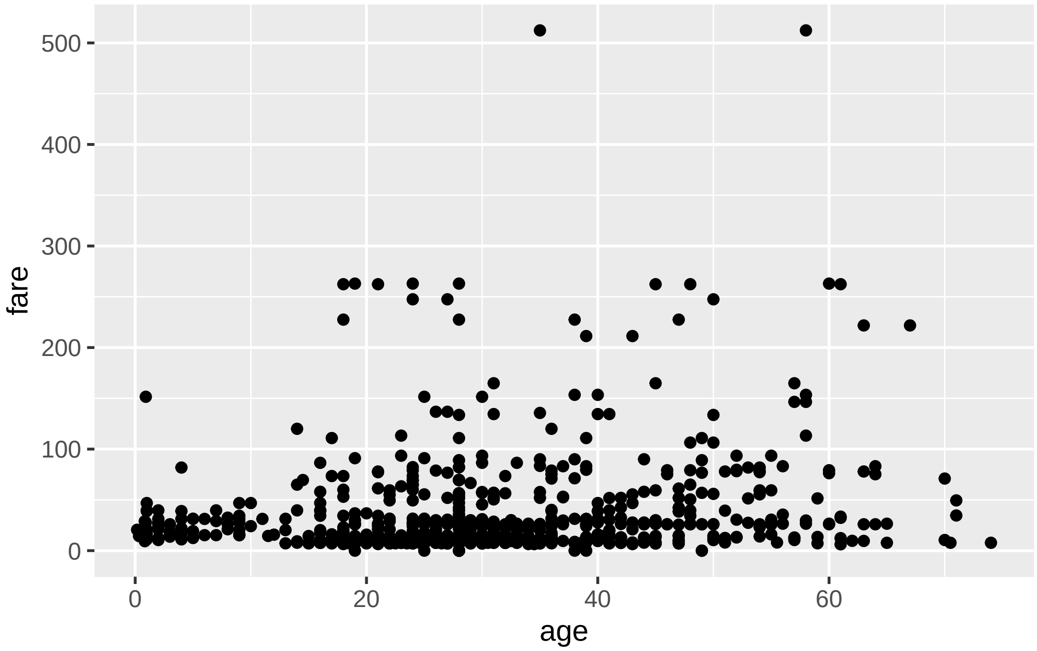
```
# Bar chart showing that a higher number of people survived when they had less siblings on board.
ggplot(train, aes(sibsp, survived)) +
  geom_bar(stat='identity')
```



```
ggplot(train, aes(age, sibsp)) +
  geom_point()
```



```
ggplot(train, aes(age, fare)) +
  geom_point()
```



```
#install.packages("psych")
library(psych) # Reference 4 to understand how this works
```

```
Warning: package 'psych' was built under R version 4.4.3
```

```
Attaching package: 'psych'
```

```
The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
tetrachoric(train[, c("survived", "sex_M")])
```

```
Call: tetrachoric(x = train[, c("survived", "sex_M")])
tetrachoric correlation
         srvvd sex_M
survived  1.00
sex_M    -0.75  1.00

 with tau of
survived    sex_M
    0.32    -0.37
```

```
tetrachoric(train[, c("survived", "pclass_1")])
```

```
Call: tetrachoric(x = train[, c("survived", "pclass_1")])
tetrachoric correlation
         srvvd pcl_1
survived 1.00
```

```
pclass_1 0.46  1.00

 with tau of
survived pclass_1
    0.32     0.72
```

```
tetrachoric(train[, c("survived", "pclass_2")])
```

```
Call: tetrachoric(x = train[, c("survived", "pclass_2")])
tetrachoric correlation
         srvvd pcl_2
survived 1.00
pclass_2 0.08  1.00

 with tau of
survived pclass_2
    0.32     0.83
```

```
#install.packages("rcompanion")
library(rcompanion) # Reference 4 to understand how this works.
```

```
Warning: package 'rcompanion' was built under R version 4.4.3
```

```
Attaching package: 'rcompanion'
```

```
The following object is masked from 'package:psych':

    phi
```

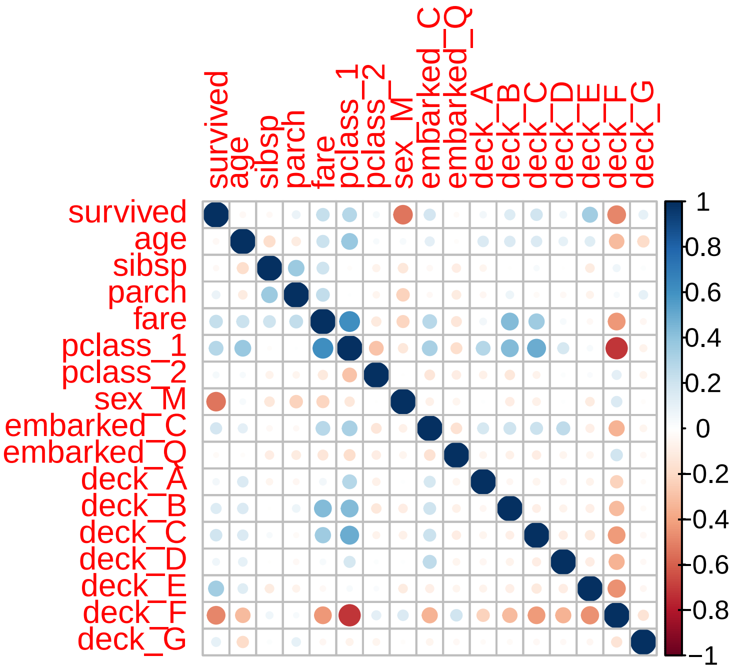
```
cramerV(train$survived, train$sex)
```

```
Cramer V
  0.5345
```

```
cor_matrix <- cor(train)#[,1]

library(corrplot)
```

```
corrplot 0.95 loaded
```

```
corrplot(cor_matrix, method = "circle")
```



5 III. Model Development Process

Build an appropriate model to predict probability of survival. And of course, create the train data set which contains 70% of the data and use set.seed (1023). The remaining 30% will be your test data set. Investigate the data and combine the level of categorical variables if needed and drop variables. For example, you can passenger name, cabin, etc..

6 IV. Model Performance Testing

Use the test data set to assess the model performances. Here, build the best model by using appropriate selection method. You may compare the performance of the best two logistic or other classification model selected. Apply remedy measures as applicable (transformation, etc.) that helps satisfy the assumptions of your particular model. Deeply investigate unequal variances and multicollinearity if warranted.

7 V. Challenger Models

Build an alternative model based on one of the following approaches to predict survival as applicable:logistic regression, decision tree, NN, or SVM, Poisson regression or negative binomial. Check the applicable model assumptions. Apply in-sample and out-of-sample testing, back testing and review the comparative goodness of fit of the candidate models. Describe step by step your procedure to get to the best model and why you believe it is fit for purpose.

8 VI. Model Limitation and Assumptions

Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model. Validate your models using the test sample. Do the residuals look normal? Does it matter given your technique? How is the prediction performance using Pseudo R^2, SSE, RMSE? Benchmark the model against alternatives. How good is the relative fit? Are there any serious violations of the logistic model assumptions? Has the model had issues or limitations that the user must know? (Which assumptions are needed to support the Champion model?)

9 VII. Ongoing Model Monitoring Plan

How would you picture the model needing to be monitored, which quantitative thresholds and triggers would you set to decide when the model needs to be replaced? What are the assumptions that the model must comply with for its continuous use?

In order to maintain the effectiveness of the model, we would need to continue to test it on new data. Since the Titanic was a rare event, we do not have a lot of new data to test on the model, but we can still be prepared in case new data were to become available. The first step in monitoring the model is to determine specific thresholds that we expect the model to stay above. We would want the model to maintain certain R^2, RMSE, and MAE values in order to determine that the model is working correctly. One of the biggest concerns with our model is data drift. Since the Titanic sank over 100 years ago, the data that we are using from the model may not align with today relevant to ship travel today.

10 VIII. Conclusion

Summarize your results here. What is the best model for the data and why?

11 References

12 Appendix

A:

National Oceanic and Atmospheric Administration (NOAA). (2023). *RMS titanic – history and significance.* https://www.noaa.gov/office-of-general-counsel/gc-international-section/rms-titanic-history-and-significance