Titanic-Survival-1912-Statistic

Thamma Tharasombat

```
library(titanic)
```

```
## Warning: package 'titanic' was built under R version 4.2.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.2.2
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.2.2
```

```
## Loading required package: Rcpp
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.1, built: 2022-11-18)
## ## Copyright (C) 2005-2023 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(ggplot2)
```

View dataframe

First, lets view the dataset.

```
glimpse(titanic_train)
```

```
## Rows: 891
## Columns: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1~
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3~
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
## $ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, ~
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0~
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0~
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,~
## $ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "C~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

The dataset contains information about 891 passengers and 12 columns.

Drop NA

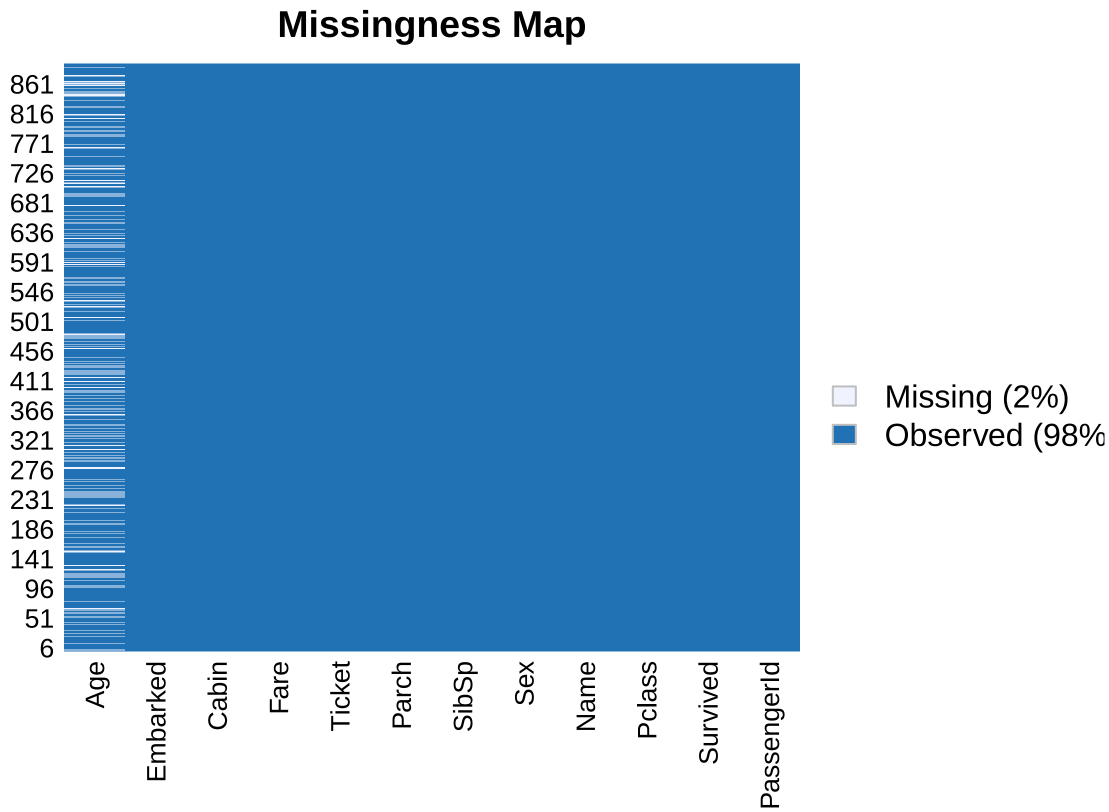
Count total missing values

```
sum(is.na(titanic_train))
```

```
## [1] 177
```

Plot missing map

```
missmap(titanic_train)
```



## Drop na

```
titanic_df <- na.omit(titanic_train)
```

Visualizing Data

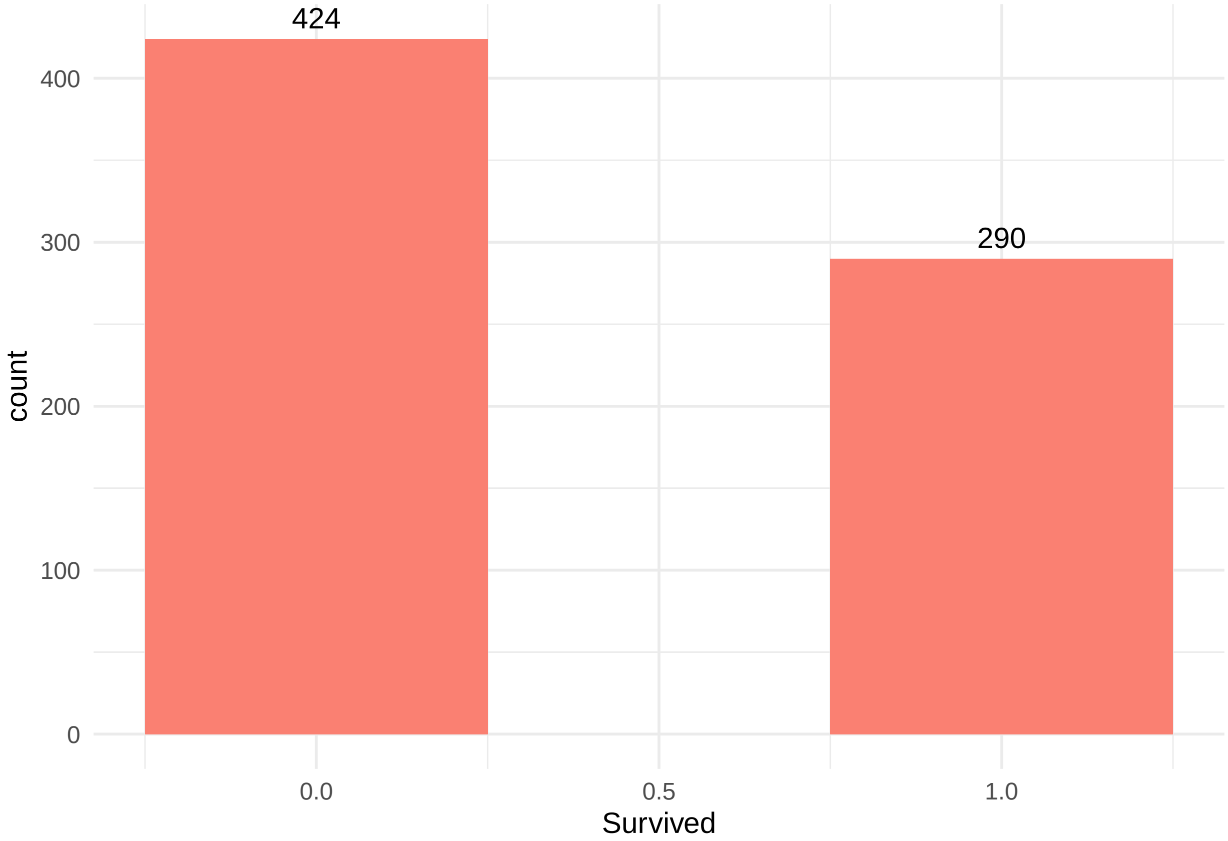
Check the structure of the data

```
str(titanic_df)
```

```
## 'data.frame':    714 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 7 8 9 10 11 ...
##  $ Survived   : int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 1 3 3 2 3 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 54 2 27 14 4 ...
##  $ SibSp      : int  1 1 0 1 0 0 3 0 1 1 ...
##  $ Parch      : int  0 0 0 0 0 0 1 2 0 1 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
##   ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
```

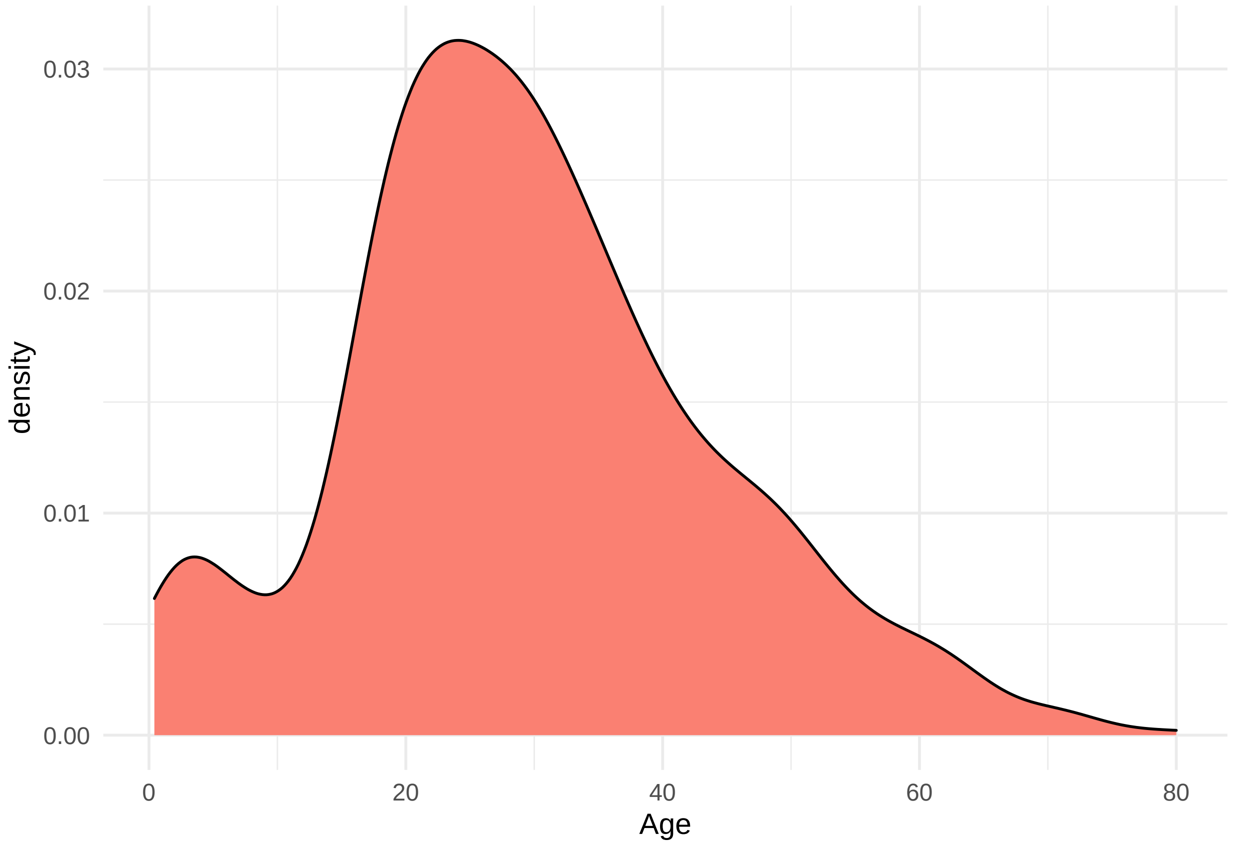
Survived count

```
ggplot(titanic_df, aes(x = Survived)) +
    geom_bar(width=0.5, fill = "salmon") +
    geom_text(stat='count', aes(label=stat(count)), vjust=-0.5) +
    theme_minimal()
```



# Age Density

```
ggplot(titanic_df, aes(x = Age)) +
    geom_density(fill="salmon")+
    theme_minimal()
```



Most of the passengers were in age group of 20 to 40.

Split Data

```
set.seed(42)

n <- nrow(titanic_df)
id <- sample(1:n, size= n*0.7)
train_data <- titanic_df[id, ]
test_data <- titanic_df[-id, ]
```

Create model

```
model <- glm(Survived ~ Pclass + Age + Sex + SibSp, data = train_data, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex + SibSp, family = "binomial",
##     data = train_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9097  -0.6223  -0.3378   0.6033   2.4133
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.21585    0.69789   8.907  < 2e-16 ***
## Pclass      -1.49165    0.17867  -8.349  < 2e-16 ***
## Age         -0.05004    0.01031  -4.851 1.23e-06 ***
## Sexmale     -2.82454    0.27045 -10.444  < 2e-16 ***
## SibSp       -0.40560    0.15504  -2.616  0.00889 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 673.56  on 498  degrees of freedom
## Residual deviance: 424.77  on 494  degrees of freedom
## AIC: 434.77
##
## Number of Fisher Scoring iterations: 5
```

Train Data

```
train_prob_su <- predict(model,type = "response")
train_data$pred_su <- ifelse(train_prob_su >= 0.5,1,0)
```

Confuionmetric

```
conM <- table(train_data$pred_su, train_data$Survived,
              dnn = c("Predicted","Actuall"))
```

Model evaluation

```
train_a <- (conM[1,1] + conM[2,2])/ sum(conM)
train_p <- conM[2,2] / (conM[2,2] + conM[2,1])
train_r <- conM[2,2]/(conM[1,2]+conM[2,2])

cat("Acculacy:",train_a,"\nPrecision:",train_p,"\nRecall",train_r)
```

```
## Acculacy: 0.8136273
## Precision: 0.7853403
## Recall 0.7425743
```

Train F1

```
Train_f1 <- 2*((train_p*train_r)/(train_p+train_r))

cat("Train F1 :", Train_f1)
```

```
## Train F1 : 0.7633588
```

Test Data

```
test_prob_su <- predict(model, newdata = test_data, type = "response")
test_data$pred_su <- ifelse(test_prob_su >= 0.5,1,0)
```

Confuionmetric

```
conM1 <- table(test_data$pred_su, test_data$Survived,
              dnn = c("Predicted","Actuall"))
```

Model evaluation

```
test_a <- (conM1[1,1] + conM1[2,2])/ sum(conM1)
test_p <- conM1[2,2] / (conM1[2,2] + conM1[2,1])
test_r <- conM1[2,2]/(conM1[1,2]+conM1[2,2])

cat("Acculacy:",test_a,"\nPrecision:",test_p,"\nRecall",test_r)
```

```
## Acculacy: 0.8
## Precision: 0.7848101
## Recall 0.7045455
```

Test F1

```
Test_f1 <- 2*((test_p*test_r)/(test_p+test_r))

cat("Train F1 :", Test_f1)
```

```
## Train F1 : 0.742515
```

Plot model

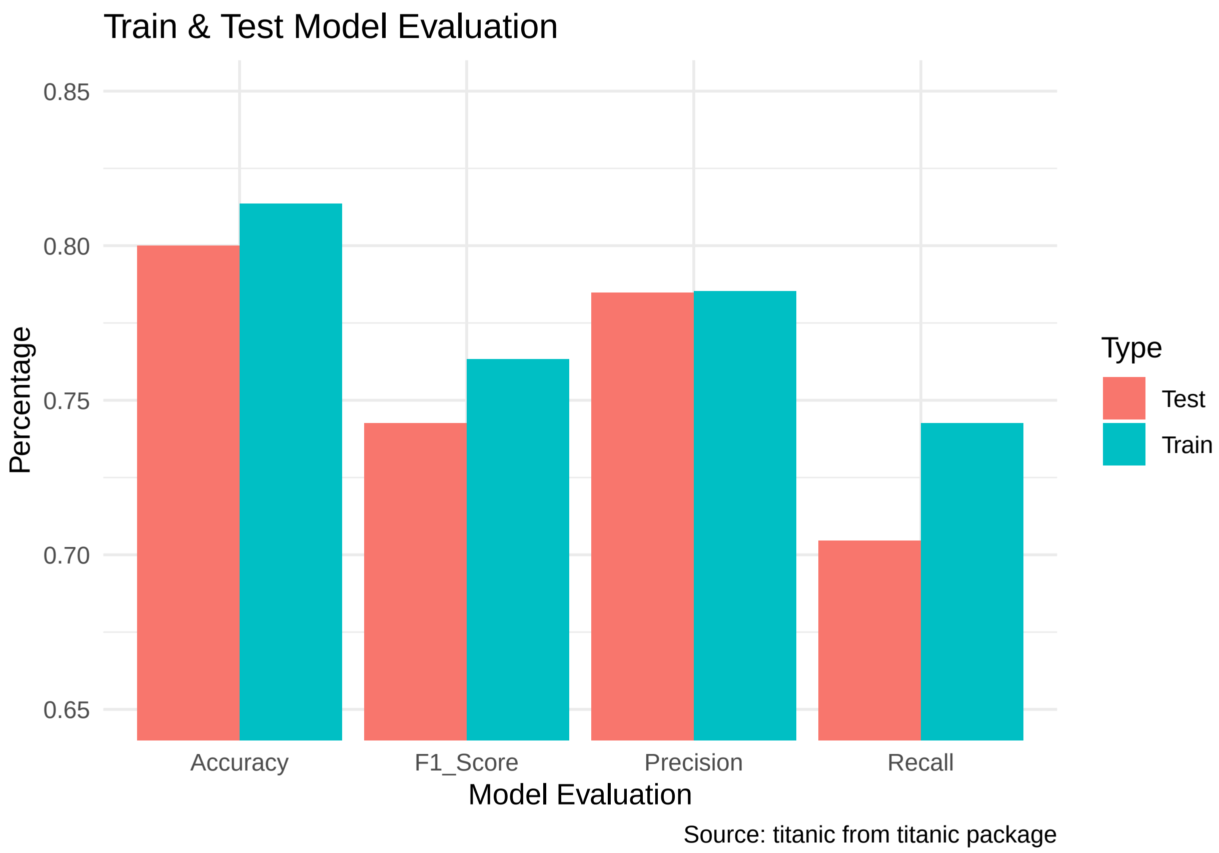
```
sum_model <- data.frame(
    Type = c('Train', 'Test'),
    Accuracy = c(train_a, test_a),
    Precision = c(train_p, test_p),
    Recall = c(train_r, test_r),
    F1_Score = c(Train_f1, Test_f1))
```

Turn to data for plot

```
final <- sum_model %>%
    pivot_longer(-Type ,
                 names_to = "me_type",
                 values_to = "percent")
```

Plot data

```
ggplot(final,aes(me_type, percent, fill = Type)) +
    geom_bar(stat='identity', position = 'dodge') +
    coord_cartesian(ylim = c(0.65, 0.85)) +
    theme_minimal() +
    labs(title = "Train & Test Model Evaluation",
         x ="Model Evaluation", y="Percentage",
         caption = "Source: titanic from titanic package")
```


Train & Test Model Evaluation
0.85
0.80
0.75
0.70
0.65
Percentage
Accuracy
F1_Score
Precision
Recall
Model Evaluation
Type
Test
Train
Source: titanic from titanic package