Titanic_Disaster_Survivers

Pranjal Vijay

January 15, 2018

Loading the Packages

```
library(ggplot2)
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
```

Loading the Data

```
Traindataset<-read.csv("C:/Users/ddddd/Titanic_survived/train.csv", stringsAsFactors = F)
Testdataset<-read.csv("C:/Users/ddddd/Titanic_survived/test.csv", stringsAsFactors = F)
```

Research Question:

According to the provided data of Titanic Disaster we are asked to 0find out the information of people who survived . Observing the given data sets , I found these variables more impactable: Survived , Sex and Age I am going to calculate here that who survived more : (i) Males or Females (ii) People below or above average age

Starting of Solution

```
Traindataset %>%
  select(Age
, Sex
) %>%
  str()
```

```
## 'data.frame':    891 obs. of  2 variables:
##  $ Age: num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ Sex: chr  "male" "female" "female" "female" ...
```

```
Testdataset %>%
  select(Age
, Sex
) %>%
  str()
```

```
## 'data.frame':    418 obs. of  2 variables:
##  $ Age: num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ Sex: chr  "male" "female" "male" "male" ...
```

Combining the data sets

```
Testdataset$Survived <- NA
united_data <- rbind(Traindataset, Testdataset)
str(united_data)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. La
ina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ 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" ...
```

Showing the Missing Values

```
sapply(united_data, function(x) {sum(is.na(x))})
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

EDA of NA’s

```
NA_value <- as.data.frame(sort(sapply(united_data, function(x) sum(is.na(x))),decreasing  = F))

colnames(NA_value)[1] <- "missingvaluesPercentage"

NA_value$Survived <- rownames(NA_value)

ggplot(NA_value[NA_value$missingvaluesPercentage>0,],aes(reorder(Survived,-missingvaluesPercentage),missingvaluesPercentage,
fill= Survived)) +geom_bar(stat="identity") +theme_minimal(base_family = "Ubuntu Condensed") +theme(axis.text.x =element_tex
t(angle = 360, hjust = 0.5), legend.position = "left") + ylab("Na values") +xlab("Variables having NA") + ggtitle("Na(missin
g) values")
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font family not found in Windows font database

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font family not found in Windows font database

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font family not found in Windows font database
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font family not found in Windows font database
```

Removing NA values

```
united_data %>%
  filter(Age != "NA") %>%
  group_by(Age) %>%
  summarise(count = n())
```

```
## # A tibble: 98 x 2
##      Age count
##    <dbl> <int>
##  1  0.17     1
##  2  0.33     1
##  3  0.42     1
##  4  0.67     1
##  5  0.75     3
##  6  0.83     3
##  7  0.92     2
##  8  1.00    10
##  9  2.00    12
## 10  3.00     7
## # ... with 88 more rows
```

```
united_data %>%
  filter(Sex != "NA") %>%
  group_by(Sex) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##      Sex count
##    <chr> <int>
## 1 female   466
## 2   male   843
```

```
united_data %>%
  filter(Survived != "NA") %>%
  group_by(Survived) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   Survived count
##      <int> <int>
## 1        0   549
## 2        1   342
```

Looking for the conclusion

```
united_data %>%
filter(Survived != "NA", Sex != "NA") %>%
  group_by(Survived, Sex) %>%
  summarise(count=n())
```

```
## # A tibble: 4 x 3
## # Groups:   Survived [?]
##   Survived    Sex count
##      <int>  <chr> <int>
## 1        0 female    81
## 2        0   male   468
## 3        1 female   233
## 4        1   male   109
```

Conclusion 1: Females survived less than males.

Now move on to the next variable ‘Age’

```
united_data %>%
filter(Survived != "NA", Age != "NA") %>%
  group_by(Survived, Age) %>%
  summarise(count=n())
```

```
## # A tibble: 142 x 3
## # Groups:   Survived [?]
##    Survived   Age count
##       <int> <dbl> <int>
##  1        0     1     2
##  2        0     2     7
##  3        0     3     1
##  4        0     4     3
##  5        0     6     1
##  6        0     7     2
##  7        0     8     2
##  8        0     9     6
##  9        0    10     2
## 10        0    11     3
## # ... with 132 more rows
```

```
united_data %>%
  filter(!(is.na(Survived))) %>%
  summarise(Survivedmean = mean(Survived), Survivedmedian = median(Survived), Survivedsd = sd(Survived), Survivedmin = min(S
urvived), Survivedmax = max(Survived))
```

```
##   Survivedmean Survivedmedian Survivedsd Survivedmin Survivedmax
## 1    0.3838384              0  0.4865925           0           1
```

It can be seen that Average Survived is 0.38 Minimum Survived is 0 and Maximum age is 1

```
united_data <- united_data %>%
    filter(!(is.na(Survived))) %>%
  mutate(Positive_Sur = ifelse(Survived > 0, "Positive_Sur", "Negative_Sur"))
united_data %>%
  group_by(Positive_Sur) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   Positive_Sur count
##          <chr> <int>
## 1 Negative_Sur   549
## 2 Positive_Sur   342
```

```
united_data %>%
  filter(!(is.na(Age))) %>%
  summarise(Agemean = mean(Age), Agemedian = median(Age), Agesd = sd(Age),
            Agemin = min(Age), Agemax = max(Age))
```

```
##    Agemean Agemedian   Agesd Agemin Agemax
## 1 29.69912        28 14.5265   0.42     80
```

It can be seen that Average age is 29 Minimum age is 14 and Maximum age is 80.

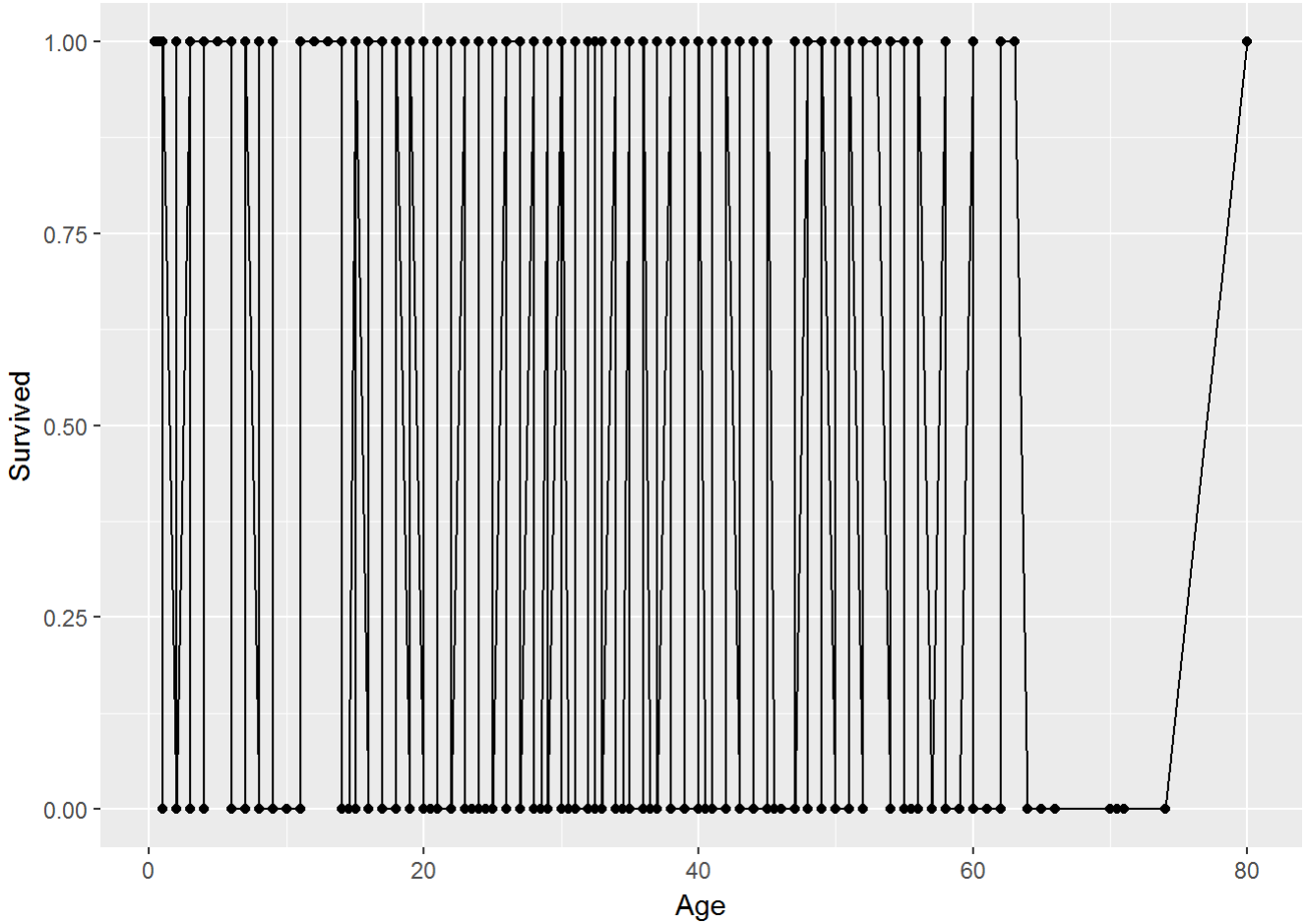
```
united_data <- united_data %>%
    filter(!(is.na(Age))) %>%
   mutate(HighAge = ifelse(Age >= 28, "HighAge", "LowAge"))
united_data %>%
  group_by(HighAge) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   HighAge count
##     <chr> <int>
## 1 HighAge   377
## 2  LowAge   337
```

```
united_data %>%
  filter(!(is.na(Age)), !(is.na(Survived)), Positive_Sur != "Negative_Sur") %>%
  group_by(HighAge,Positive_Sur) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 3
## # Groups:   HighAge [?]
##   HighAge Positive_Sur count
##     <chr>        <chr> <int>
## 1 HighAge Positive_Sur   149
## 2  LowAge Positive_Sur   141
```

```
ggplot(data = united_data, aes(x = Age, y = Survived)) +
  geom_line() +
  geom_point()
```



```
write.table(united_data, "united_data.csv", row.name=FALSE)
```

Coclusion:-

It can be seen that No. of people have age more than 28 are More survived than no. of people have age equals to or less than 28.