Titanic Data Analysis

Ahmed Ashraf Mohamed

Contents

1 Preliminary Look at the data **1**

2 Summary of the Data **2**

2.1 Grouped by Sex 3

1 Preliminary Look at the data

The

Variable	Definition	Key
survival	Survival	0 = No, 1 = yes
pclass	ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	sex	
age	Age in year	
sibsp	Number of siblings/spouses aboard the titanic	
parch	Number of parents/children aboard the Titanic	
ticket	ticket number(unique)	
fare	Passenger fare	
cabin	Cabin number	
embarked	port of embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
train <- read_csv(file = "data/train.csv")
```

```
## Rows: 891 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
train <- train %>%
  mutate(Sex = as_factor(Sex),Embarked = as_factor(Embarked))

head(train)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <fct> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  C123
## 5           5        0      3 Allen,~ male     35     0     0 373450  8.05 <NA>
## 6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 <NA>
## # ... with 1 more variable: Embarked <fct>
```

```
tail(train)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <fct> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1         886        0      3 "Rice,~ fema~    39     0     5 382652 29.1  <NA>
## 2         887        0      2 "Montv~ male     27     0     0 211536 13    <NA>
## 3         888        1      1 "Graha~ fema~    19     0     0 112053 30    B42
## 4         889        0      3 "Johns~ fema~    NA     1     2 W./C.~ 23.4  <NA>
## 5         890        1      1 "Behr,~ male     26     0     0 111369 30    C148
## 6         891        0      3 "Doole~ male     32     0     0 370376  7.75 <NA>
## # ... with 1 more variable: Embarked <fct>
```

2 Summary of the Data

```
summary(train)
```

```
##   PassengerId       Survived          Pclass          Name          
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891        
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character  
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309                     
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                     
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000                     
##                                                                     
##      Sex           Age            SibSp           Parch       
##  male  :577   Min.   : 0.42   Min.   :0.000   Min.   :0.0000  
##  female:314   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000  
##               Median :28.00   Median :0.000   Median :0.0000  
##               Mean   :29.70   Mean   :0.523   Mean   :0.3816  
##               3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000  
##               Max.   :80.00   Max.   :8.000   Max.   :6.0000  
##               NA's   :177                                     
##     Ticket               Fare            Cabin           Embarked  
##  Length:891         Min.   :  0.00   Length:891         S   :644  
##  Class :character   1st Qu.:  7.91   Class :character   C   :168
##  Mode  :character   Median : 14.45   Mode  :character   Q   : 77
##                     Mean   : 32.20                      NA's:  2
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

2.1 Grouped by Sex

```
head(train)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex      Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <fct> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Braund~ male      22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cuming~ fema~     38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heikki~ fema~     26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futrel~ fema~     35     1     0 113803 53.1  C123
## 5           5        0      3 Allen,~ male      35     0     0 373450  8.05 <NA>
## 6           6        0      3 Moran,~ male      NA     0     0 330877  8.46 <NA>
## # ... with 1 more variable: Embarked <fct>
```

```
train %>%
  group_by(Sex) %>%
  summarise(Age_mean = mean(Age,na.rm=TRUE),
            age_sd = sd(Age,na.rm=T),
            surival_mean = mean(Survived,na.rm =T),
            surival_sd = sd(Survived,na.rm = T))
```

```
## # A tibble: 2 x 5
##   Sex    Age_mean age_sd surival_mean surival_sd
##   <fct>     <dbl>  <dbl>        <dbl>      <dbl>
## 1 male       30.7   14.7        0.189      0.392
## 2 female     27.9   14.1        0.742      0.438
```