tidyverse

Subhalaxmi Rout

3/14/2020

Contents

1. Introduction

This is the dataset of `titanic`, I have chosen from `Kaggle`. This data set has below columns.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
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

2. Load library

```
#install.packages("tidyverse")
#install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t
```

```
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.3
```

```
## -- Conflicts ------------------------------------------------------------------------------------- tidyve
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

3. Load data to R

```
# get the data from Git repository
url <- "https://raw.githubusercontent.com/SubhalaxmiRout002/tidyverse/master/titanic.csv"
# read the csv file
titanic_data <- read.csv(url, stringsAsFactors = FALSE)
# view fisrt 6 rows of data
head(titanic_data)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

4. Clean data

```
# remove unwanted column
titanic_data <- titanic_data %>% select(-SibSp,-Ticket,-Fare,-Cabin,-Embarked,-Parch)

# remove where name in NA
titanic_data <- titanic_data %>% filter(!is.na(Name))

# remove duplicates from data, if present any
titanic_data <- unique(titanic_data)

# rename column
titanic_data <- titanic_data %>% rename(Class_Type = Pclass)

# view data
head(titanic_data,5)
```

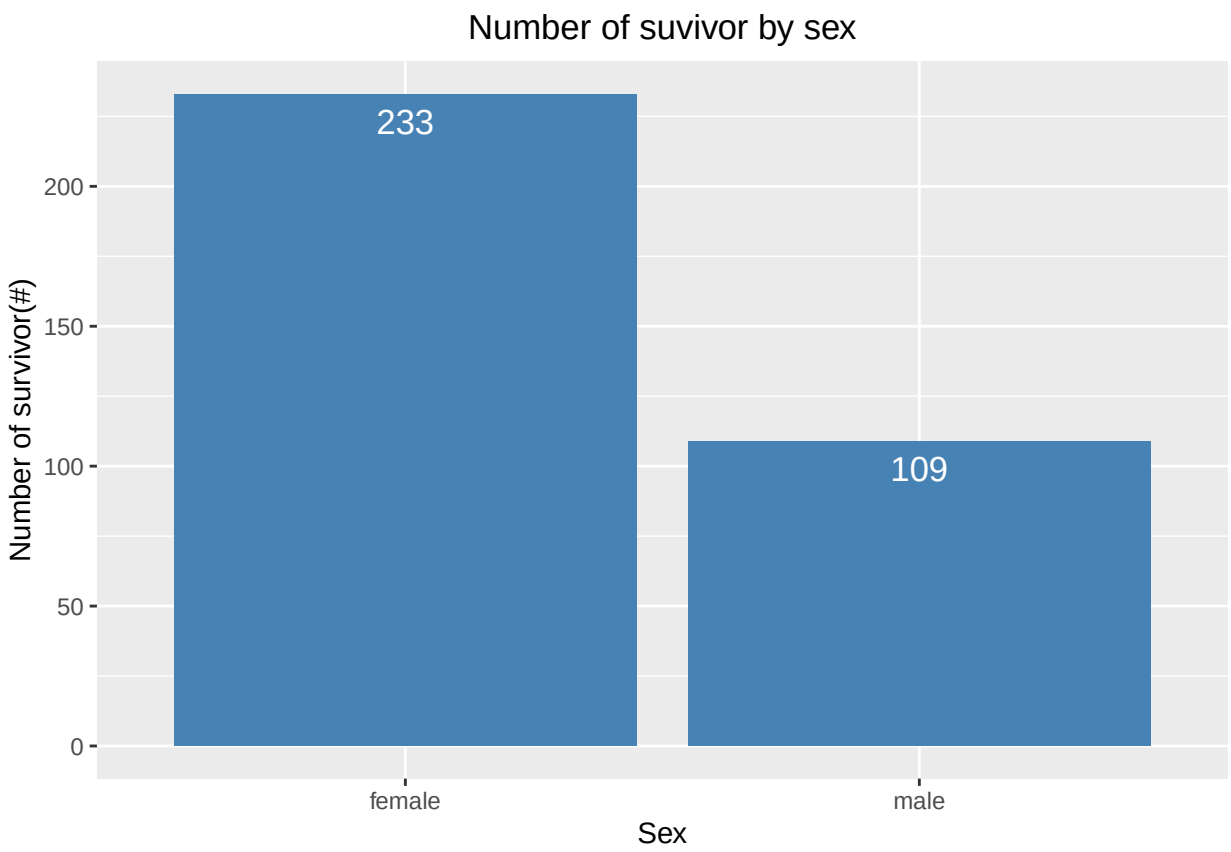
```
##   PassengerId Survived Class_Type
## 1           1        0          3
## 2           2        1          1
## 3           3        1          3
## 4           4        1          1
## 5           5        0          3
##                                                  Name    Sex Age
## 1                             Braund, Mr. Owen Harris   male  22
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38
## 3                              Heikkinen, Miss. Laina female  26
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35
## 5                            Allen, Mr. William Henry   male  35
```

5. Analysis

```
# find number of survivor ratio
row <- nrow(titanic_data)

# add new colum Survived_Ration in to dataset
titanic_data <- titanic_data %>% mutate(Survived_Ratio = Survived/row)
```

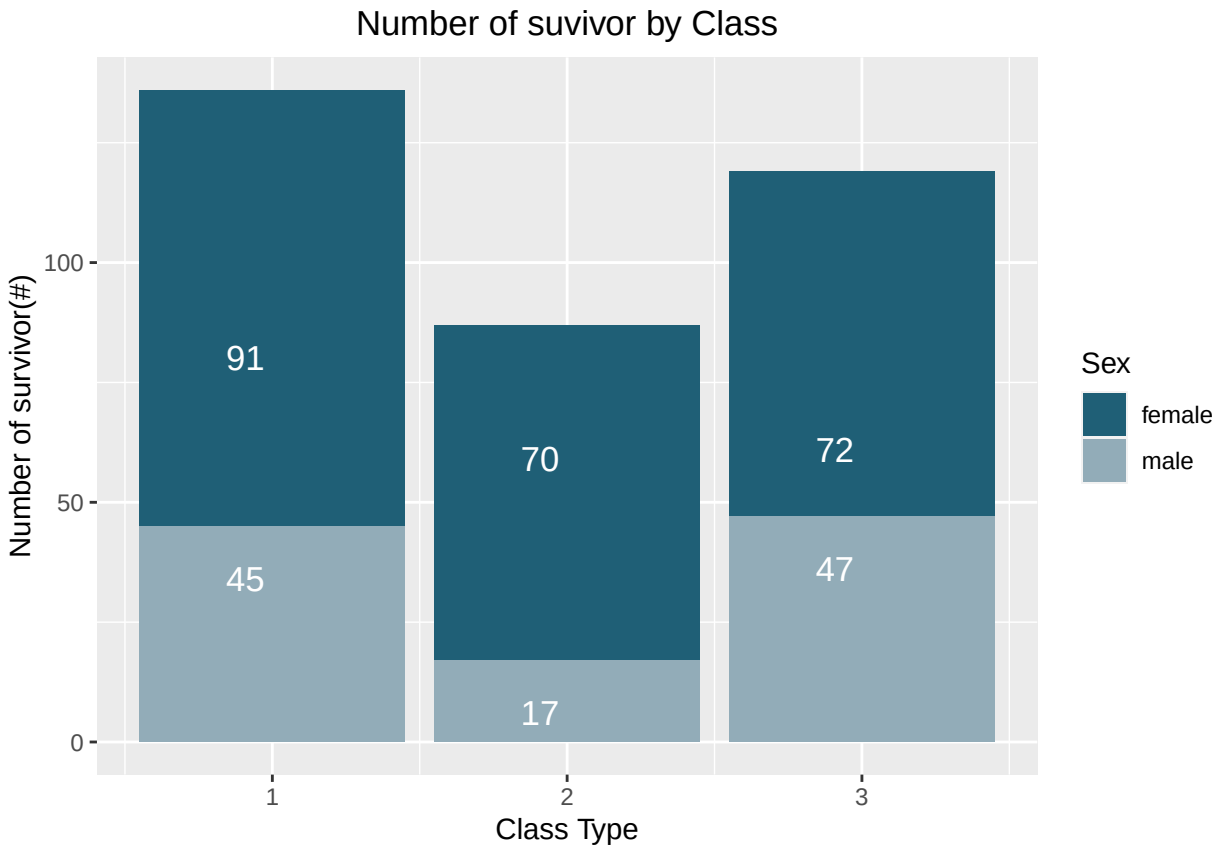
5.1 survivor group by sex

```
# number of survivor group by sex
survivor_sex <- titanic_data %>%  filter(Survived == 1) %>% group_by(Sex) %>% count(Survived)


# draw graph
ggplot(survivor_sex) + geom_bar(aes(x = survivor_sex$Sex, y = survivor_sex$n), stat = "identity", fill =
```



5.2 survivor group class type

```
# number of survivor group by class type
survivor_class <- titanic_data %>%  filter(Survived == 1) %>% group_by(Class_Type,Sex) %>% count(Survive

# draw graph
ggplot(data = survivor_class, aes(x = survivor_class$Class_Type, y = survivor_class$n, fill = survivor_
```



6. Conclusion

From Plot 4.1 and 4.2 we found:

- The number of Surivior is high in female.
- Highest survivor is in female and Class 1 type.

I have used, select(), filter(), mutate(), rename() functions of `tidyverse` package to clean and manupulate data.