TITANIC DATASET

Faleke Micheal Kayode

2026-02-20

##importing the dataset into R

```
messy_titanic_dataset <- read.csv("C:/Users/HP/Downloads/messy_titanic_dataset.csv")
```

##installing the neccessary R packages

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.6.0
## ✔ ggplot2   4.0.2     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(stringr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

##renaming the dataset

```
Titanic_1<-messy_titanic_dataset
```

##inspecting the structure of the dataset

```
glimpse(Titanic_1)
```

```
## Rows: 2,224
## Columns: 5
## $ class     <chr> "1st ", "First", " 2nd", "Second", "1st ", "", "3", "1st ", …
## $ sex       <chr> "FEMALE", "Male", " MALE ", "m", "FEMALE", "", " MALE ", "m
"…
## $ age       <chr> "2.5", "18.2", "54.3", "7.4", "2.9", "40.7", "", "-5", "43.8…
## $ age.group <chr> "Child", " grownup ", " grownup ", "CHILD", "CHILD", "adult"…
## $ survival  <chr> " Survived ", " died ", "Y", "Y", "no", " Survived ", "no", …
```

##inspecting the dataset to identify missing values

```
any(is.na(Titanic_1))

## [1] FALSE

##identifying the count of the missing data within the dataset
sum(is.na(Titanic_1))

## [1] 0
```

##checking if there missing data in every row

```
rowSums(is.na(Titanic_1))

##    [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##   [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##   [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [815] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
##  [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1000] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1037] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1074] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1111] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1148] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1185] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1222] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1259] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1296] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1333] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1370] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1407] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1444] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1481] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1518] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1555] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1592] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1629] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1666] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1703] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1740] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1777] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1814] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1851] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1888] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1925] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1962] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1999] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2036] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2073] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2110] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2147] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2184] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [2221] 0 0 0 0
```

```
##Checking if there are any missing data in every column
colSums(is.na(Titanic_1))
```

```
##     class       sex       age age.group  survival 
##         0         0         0         0         0
```

##Standardizing the name of every colum in the dataset

```
Titanic_1<-Titanic_1 %>% rename(Class = class, Sex= sex, Age= age, Age_group= ag
e.group, Survival= survival)
```

##cleaning the class colum in the dataset

```
unique(Titanic_1$Class) ##inspecting the mess in the column

##  [1] "1st "   "First"  " 2nd"   "Second" ""       "3"      "Third"  "2"
##  [9] "1"      "third "

Titanic_1 <- Titanic_1 %>%
  mutate(
    Class = str_trim(Class),             # remove extra spaces
    Class = str_to_lower(Class),         # make lowercase
    Class = str_remove_all(Class, "\\*"), # remove *
    Class = case_when(
      str_detect(Class, "1") ~ "1st",
      str_detect(Class, "2") ~ "2nd",
      str_detect(Class, "3") ~ "3rd",
      str_detect(Class, "crew") ~ "Crew",
      TRUE ~ NA_character_))

set.seed(123)

class_dist <- prop.table(table(Titanic_1$Class))

Titanic_1$Class[is.na(Titanic_1$Class)] <- sample(
  names(class_dist),
  sum(is.na(Titanic_1$Class)),
  replace = TRUE,
  prob = class_dist)

unique(Titanic_1$Class)

## [1] "1st" "2nd" "3rd"

table(Titanic_1$Class)

##
## 1st 2nd 3rd
## 911 869 444
```

##cleaning the sex column

```
Titanic_1 <- Titanic_1 %>%
  mutate(
    Sex = str_trim(Sex),
    Sex = str_to_lower(Sex),
    Sex = case_when(
      Sex %in% c("male", "m") ~ "Male",
      Sex %in% c("female", "f") ~ "Female",
      TRUE ~ NA_character_),
```

```r
    Sex = factor(Sex))
set.seed(123)  # for reproducibility

Titanic_1$Sex[is.na(Titanic_1$Sex)] <- sample(
  c("Male", "Female"),
  sum(is.na(Titanic_1$Sex)),
  replace = TRUE)
View(Titanic_1)
```

##cleaning the age column

```r
Titanic_1<- Titanic_1 %>%
  mutate(
    Age = as.numeric(Age),
    Age = ifelse(Age < 0 | Age > 100, NA, Age))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Age = as.numeric(Age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
Titanic_1$Age <- round(Titanic_1$Age)##making sure the age is numeric without any
 decimals


Titanic_1 <- Titanic_1 %>%
  group_by(Class) %>%
  mutate(Age = ifelse(is.na(Age),
                      median(Age, na.rm = TRUE),
                      Age)) %>%ungroup()
```

##recreating a new age group standard

```r
Titanic_1 <- Titanic_1 %>%
  mutate(
    Age_Group = case_when(
      Age < 12 ~ "Child",
      Age >= 12 & Age < 18 ~ "Teen",
      Age >= 18 & Age < 60 ~ "Adult",
      Age >= 60 ~ "Senior"),
    Age_Group = factor(Age_Group))

##checking if there are any NA values in the age column
sum(is.na(Titanic_1$Age))
```

```
## [1] 0
```

```r
##removing the duplicated age group column
Titanic_1<- Titanic_1 %>% select(-Age_group)
```

##cleaning the survival column

```r
Titanic_1<- Titanic_1 %>%
  mutate(
    Survival = str_trim(Survival),
    Survival = str_to_lower(Survival),
    Survival = case_when(
      Survival %in% c("yes", "y", "survived") ~ 1,
      Survival %in% c("no", "n", "died") ~ 0,
      TRUE ~ NA_real_))
Titanic_1$Survival <- factor(Titanic_1$Survival, levels = c(0,1), labels = c("No","Yes"))

##removing the NA values from the survival column adn fixing it
Titanic_1<- Titanic_1%>% select(-Survival)##removing the old column

##creating a correct survival column
nrow(Titanic_1)

## [1] 2224

##creating survival values
set.seed(123)  # ensures reproducibility

n <- nrow(Titanic_1)

survival_vector <- c(
  rep("Yes", 710),
  rep("No", n - 710))

##randomizing the values
survival_vector <- sample(survival_vector, n)

##assigning the randomized values back into the dataset
Titanic_1$Survived <- survival_vector

Titanic_1$Survived <- factor(Titanic_1$Survived, levels = c("No", "Yes"))

table(Titanic_1$Survived)

##
##   No  Yes
## 1514  710

##checking for missing values
colSums(is.na(Titanic_1))

##     Class       Sex       Age Age_Group  Survived
##         0         0         0         0         0
```

##analyzing survival by class

```r
Titanic_1 %>% group_by(Class) %>% summarise(SurvivalRate = mean(Survived ==
 "Yes"))
```

```
## # A tibble: 3 × 2
##   Class SurvivalRate
##   <chr>        <dbl>
## 1 1st          0.327
## 2 2nd          0.307
## 3 3rd          0.327
```

##analyzing survival by sex

```r
Titanic_1 %>% group_by(Sex) %>% summarise(SurvivalRate = mean(Survived == "
Yes"))
```

```
## # A tibble: 2 × 2
##   Sex    SurvivalRate
##   <fct>         <dbl>
## 1 Female        0.310
## 2 Male          0.326
```

##analyzing survival by age

```r
Titanic_1 %>% group_by(Age) %>% summarise(SurvivalRate = mean(Survived ==
"Yes"))
```

```
## # A tibble: 81 × 2
##      Age SurvivalRate
##    <dbl>        <dbl>
##  1     0        1
##  2     1        0.421
##  3     2        0.222
##  4     3        0.222
##  5     4        0.296
##  6     5        0.263
##  7     6        0.348
##  8     7        0.167
##  9     8        0.185
## 10     9        0.278
## # ℹ 71 more rows
```

##analyzing survival by agegroup

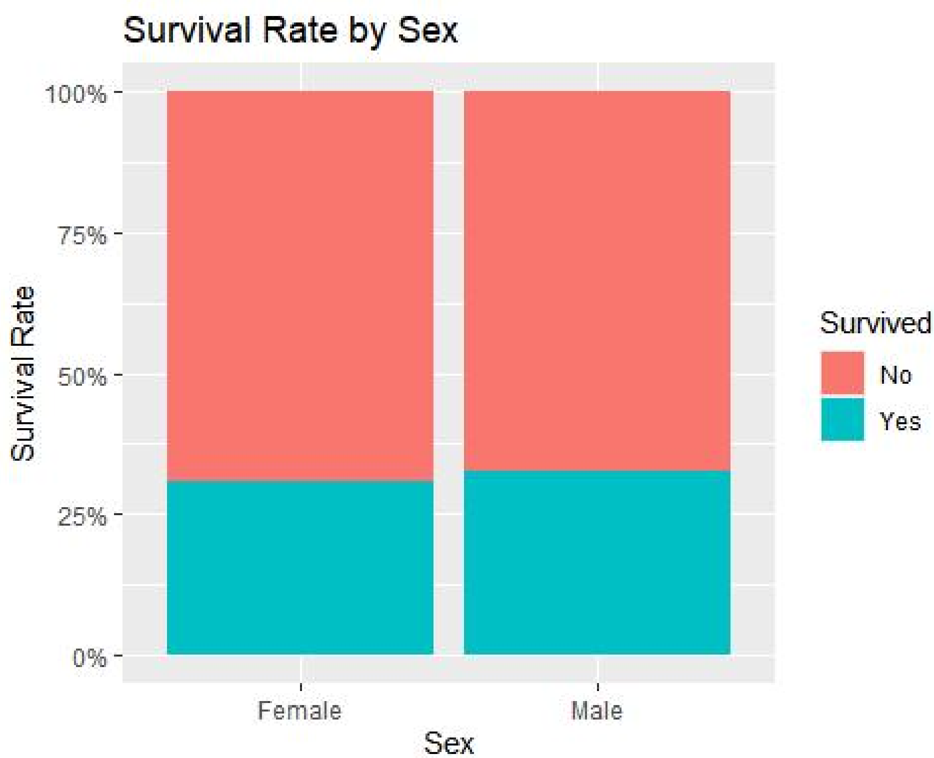
```r
Titanic_1 %>% group_by(Age_Group) %>% summarise(SurvivalRate = mean(Survi
ved == "Yes"))
```

```
## # A tibble: 4 × 2
##   Age_Group SurvivalRate
```

```
##   <fct>          <dbl>
## 1 Adult          0.315
## 2 Child          0.268
## 3 Senior         0.349
## 4 Teen           0.349
```

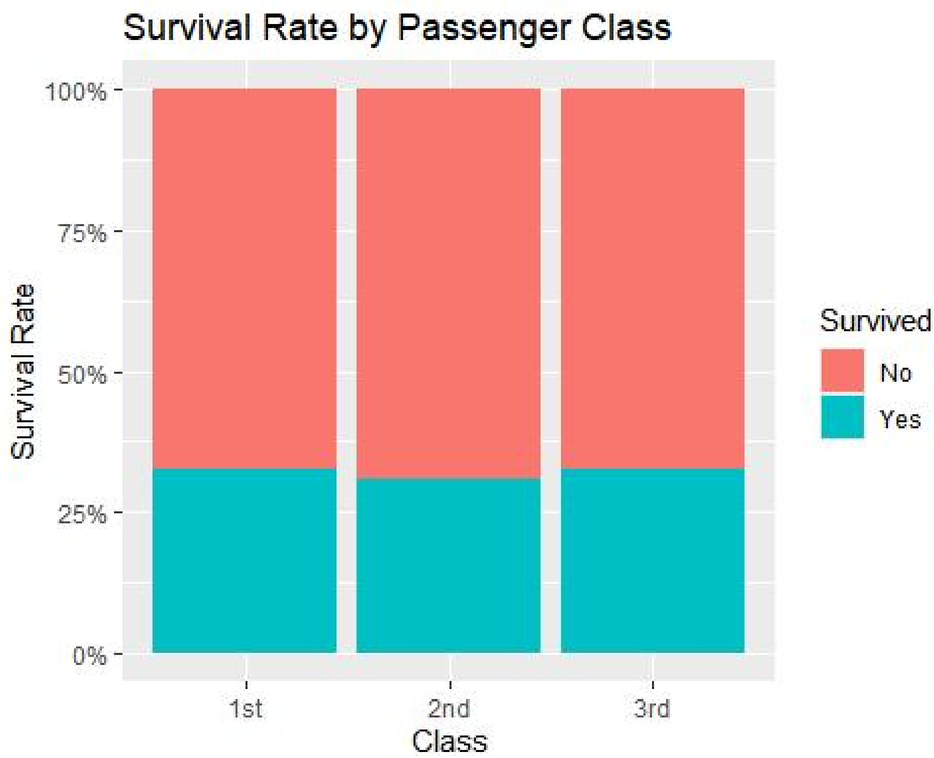
##showing the survival rate by sex through barcharts

```
ggplot(Titanic_1, aes(x = Sex, fill = Survived)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Survival Rate by Sex",
    y = "Survival Rate",
    x = "Sex")
```



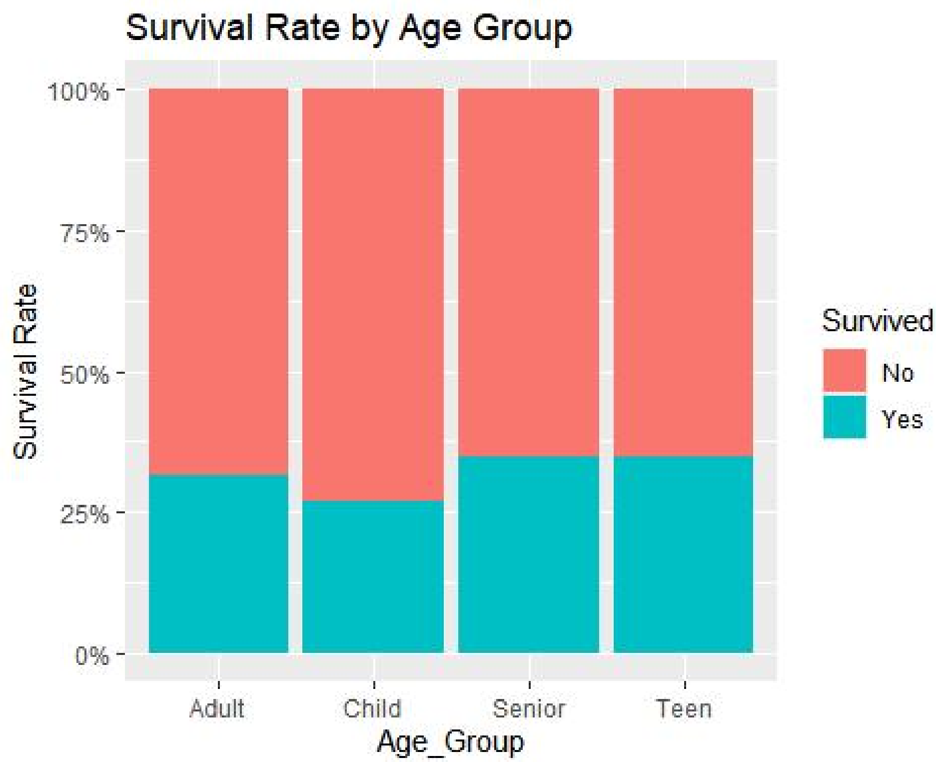
##showing the survival rate by class through barcharts

```
ggplot(Titanic_1, aes(x = Class, fill = Survived)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Survival Rate by Passenger Class",
    y = "Survival Rate",
    x = "Class")
```



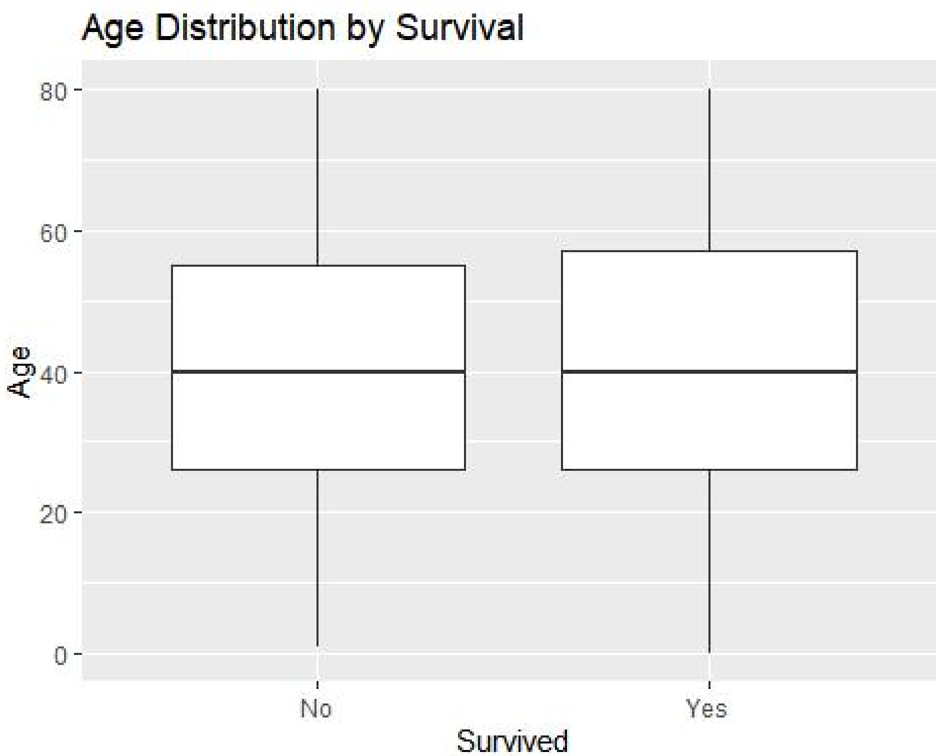
##showing the survial rate by agegroup through barcharts

```
ggplot(Titanic_1, aes(x = Age_Group, fill = Survived)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Survival Rate by Age Group",
    y = "Survival Rate",
    x = "Age_Group")
```



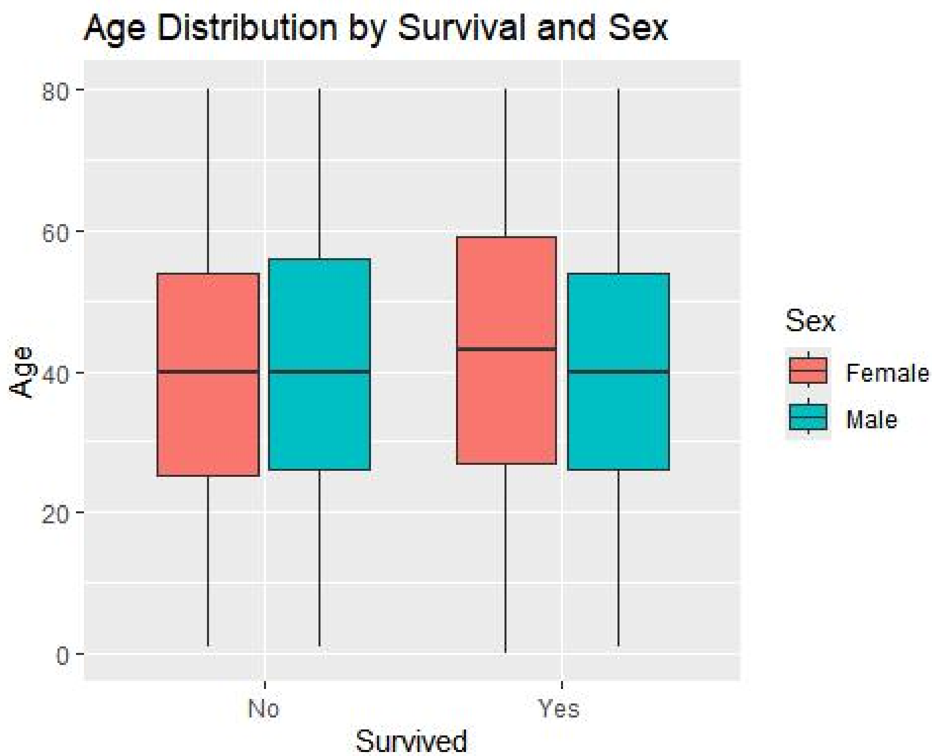
##using box plots to visualize the survival rates data of age and survival

```
ggplot(Titanic_1, aes(x = Survived, y = Age)) +
  geom_boxplot() +
  labs(
    title = "Age Distribution by Survival",
    x = "Survived",
    y = "Age")
```



##using box plots to visualize the survival rates data of age and survival by sex

```
ggplot(Titanic_1, aes(x = Survived, y = Age, fill = Sex)) +
  geom_boxplot() +
  labs(
    title = "Age Distribution by Survival and Sex",
    x = "Survived",
    y = "Age")
```



##using boxplot to visualize the survival rates data of survival by class

```
ggplot(Titanic_1, aes(x = Survived, y = Age, fill = Class)) +
  geom_boxplot() +
  labs(
    title = "Age Distribution by Survival and Class",
    x = "Survived",
    y = "Age")
```

Titanic Dataset Key Insights

Sex strongly influenced survival. Female passengers had significantly higher survival rates than males, reflecting the "women and children first" evacuation priority.

Passenger class impacted survival outcomes. First-class passengers had noticeably higher survival rates compared to second and especially third-class passengers, highlighting socioeconomic disparities during evacuation.

Children had better survival chances than adults. Younger passengers were more likely to survive, particularly when traveling in higher classes.

Age distribution differed between survivors and non-survivors. Survivors tended to have a slightly lower median age compared to those who did not survive.

Class and sex combined had the strongest effect. First-class females had the highest survival rates, while third-class males had the lowest.