My Report

Li-Hsun Chang

2024-03-01

Table of contents

Dataset

```
data <- read.csv("titanic.csv")
head(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
```

Summary

```
data[data == ""] <- NA
data$Survived <- as.factor(data$Survived)
data$Pclass <- as.factor(data$Pclass)
data$Sex <- as.factor(data$Sex)
data$SibSp <- as.factor(data$SibSp)
data$Parch <- as.factor(data$Parch)
data$Embarked <- as.factor(data$Embarked)
summary(data)
```

```
  PassengerId    Survived Pclass      Name              Sex
 Min.   :  1.0   0:549    1:216   Length:891         female:314
 1st Qu.:223.5   1:342    2:184   Class :character   male  :577
 Median :446.0            3:491   Mode  :character
 Mean   :446.0
 3rd Qu.:668.5
 Max.   :891.0

      Age        SibSp   Parch      Ticket               Fare
 Min.   : 0.42   0:608   0:678   Length:891         Min.   :  0.00
 1st Qu.:20.12   1:209   1:118   Class :character   1st Qu.:  7.91
 Median :28.00   2: 28   2: 80   Mode  :character   Median : 14.45
 Mean   :29.70   3: 16   3:  5                      Mean   : 32.20
 3rd Qu.:38.00   4: 18   4:  4                      3rd Qu.: 31.00
 Max.   :80.00   5:  5   5:  5                      Max.   :512.33
 NA's   :177     8:  7   6:  1
```

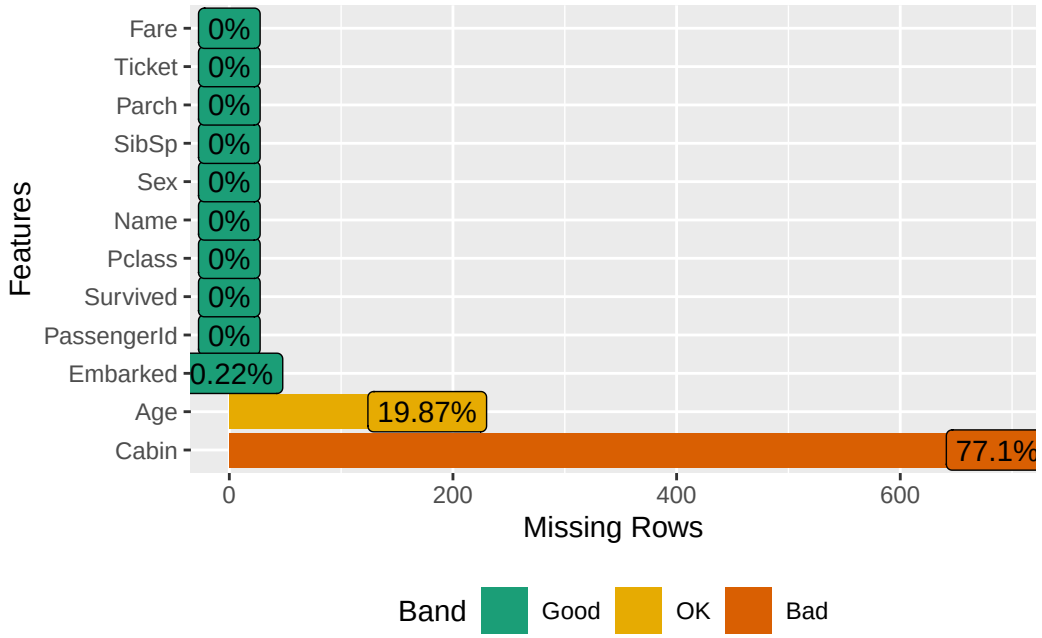
```
   Cabin             Embarked
Length:891         C    :168
Class :character   Q    : 77
Mode  :character   S    :644
                   NA's:  2
```

Missing Values

```
library(DataExplorer)
```

```
Warning: package 'DataExplorer' was built under R version 4.3.2
```

```
plot_missing(data)
```



```r
colSums(is.na(data))
```

```
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         177
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0         687           2
```

Data Visualization

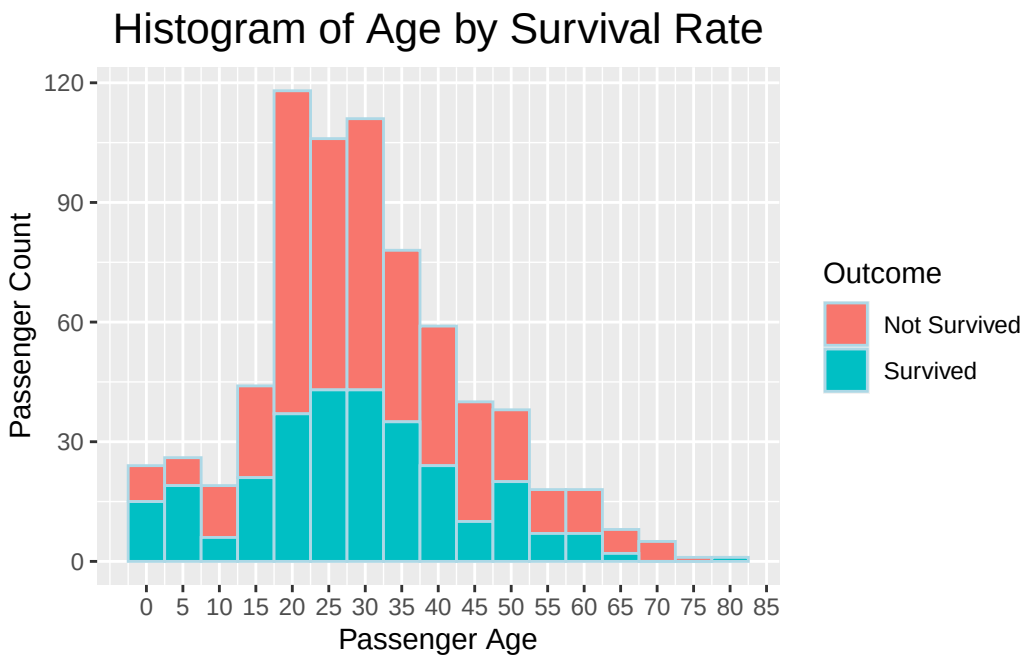
```r
library(ggplot2)
library(dplyr)
```

```
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
```

Histogram of Age by Survival Rate

```r
data %>%
  ggplot() +
  geom_histogram(aes(x = Age, fill = Survived), binwidth = 5, color = "light blue") +
  theme(plot.title = element_text(hjust = 0.5, size = 16)) +
  ggtitle("Histogram of Age by Survival Rate") +
  scale_x_continuous(name = "Passenger Age", breaks = 5*c(0:18)) +
  scale_y_continuous(name = "Passenger Count") +
  scale_fill_discrete(name = "Outcome", labels = c("Not Survived", "Survived"))
```

```
Warning: Removed 177 rows containing non-finite values (`stat_bin()`).
```



Bar chart of Pclass by Survival Rate

```
pclass <- data %>%
  group_by(Pclass) %>%
  summarise(Count = n())
pclass_ratio <- data %>%
  group_by(Pclass, Survived) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = round(Count/sum(Count)*100))
```

```
`summarise()` has grouped output by 'Pclass'. You can override using the
`.groups` argument.
```

```
data %>%
  ggplot() +
  geom_bar(aes(x = Pclass, fill = Survived)) +
  geom_text(data = pclass,
            aes(x = Pclass, y = Count, label = Count),
            position = position_dodge(width = 0.9),
            vjust = -0.25,
```

```
          fontface = "bold") +
geom_label(data = pclass_ratio,
           aes(x = Pclass, y = Count, label = paste0(Percentage, "%"), group = Survived), p
theme_bw() +
theme(plot.title = element_text(hjust = 0.5, size = 18, color = "#054354")) +
ggtitle("Bar chart of Pclass by Survival Rate") +
scale_x_discrete(name= "Pclass") +
scale_y_continuous(limits = c(0, 510), name = "Passenger Count") +
scale_fill_brewer(name = "Survival Rate", labels = c("Not Survived", "Survived"), palette =
```