Charles_James_MileStone_4

Charles James

2023-10-08

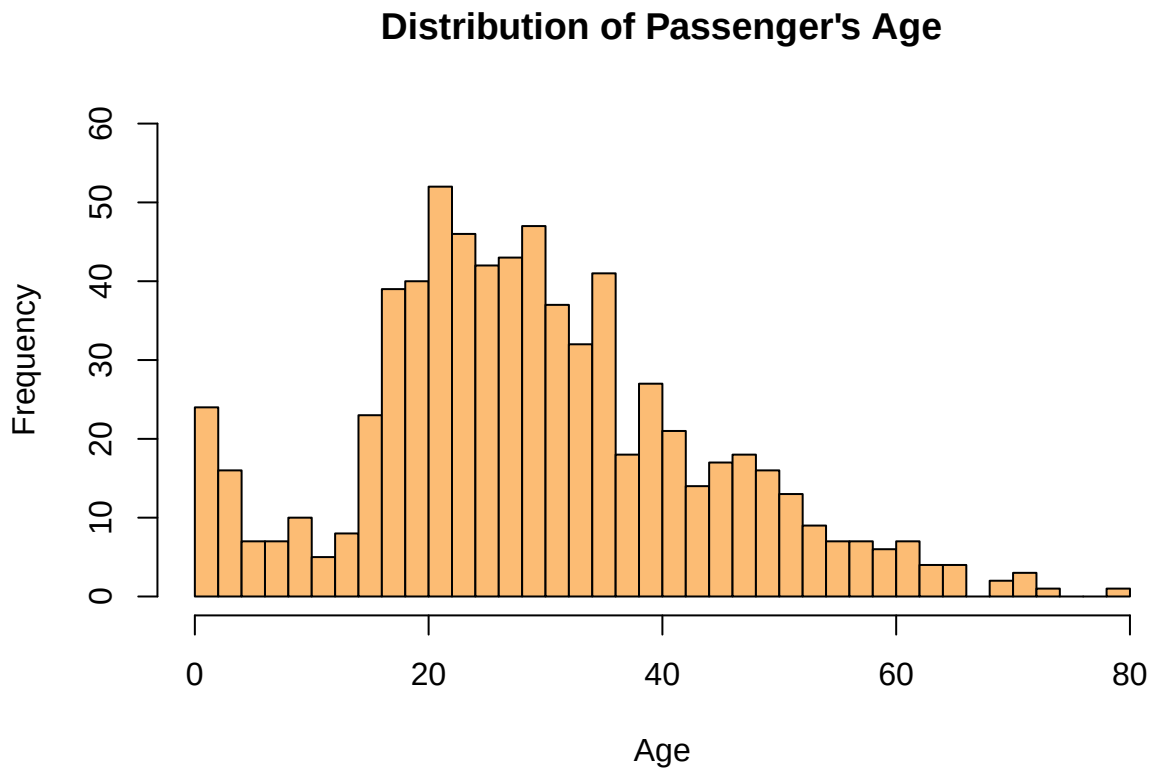
```
titanic = read.csv("titanic.csv")
colnames(titanic)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
dim(titanic)
```

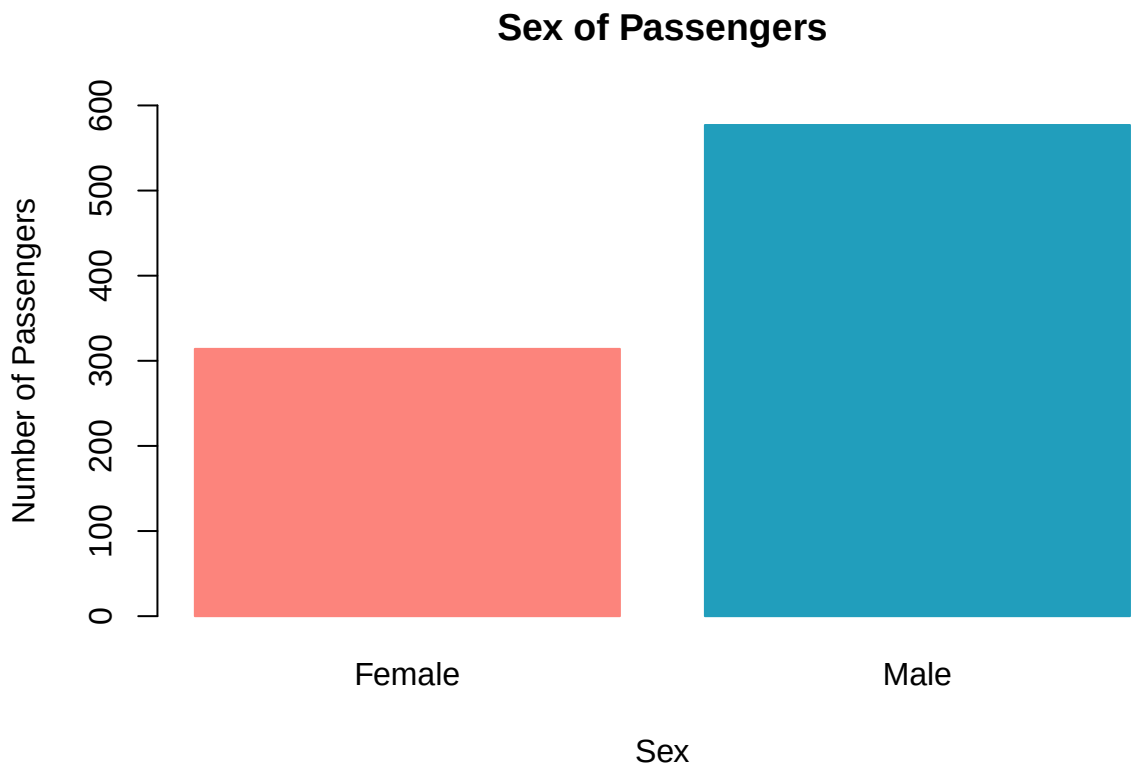
```
## [1] 891  12
```

```r
hist(titanic$Age,
     breaks = 50,
     main = "Distribution of Passenger's Age",
     xlab = "Age",
     ylab = "Frequency",
     ylim = c(0,60),
     col="#fcbc74",
     )
```



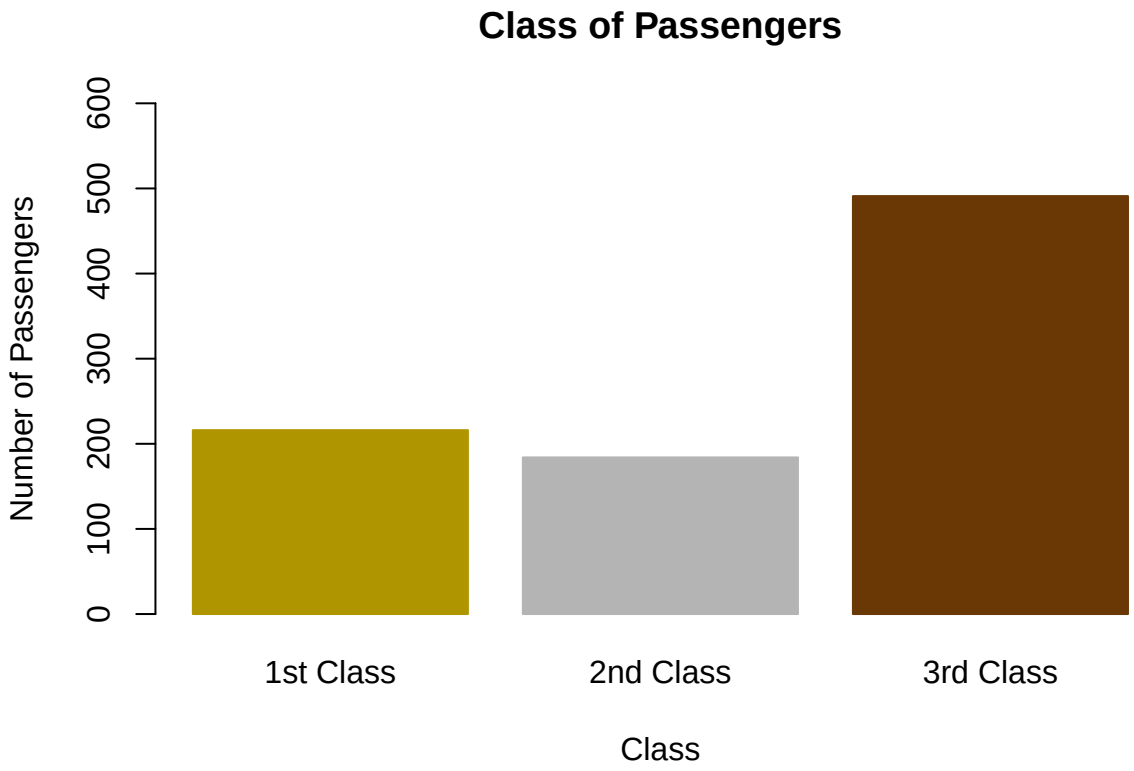
The distribution of age with breaks equal to 50 slightly resembles the normal distribution with the bulk of the data being centered in the middle. It shows that most of the passengers were between 16 and 40. There were a few outliers older though, as some of the passengers were older than 60. There was even a man on board that was 80.

```
barplot(table(titanic$Sex),
        col=c("#fc847c","#219ebc"),
        border=c("#fc847c","#219ebc"),
        main="Sex of Passengers",
        xlab="Sex",
        ylab="Number of Passengers",
        ylim=c(0,600),
        names.arg = c("Female", "Male"),
        beside=TRUE
        )
```



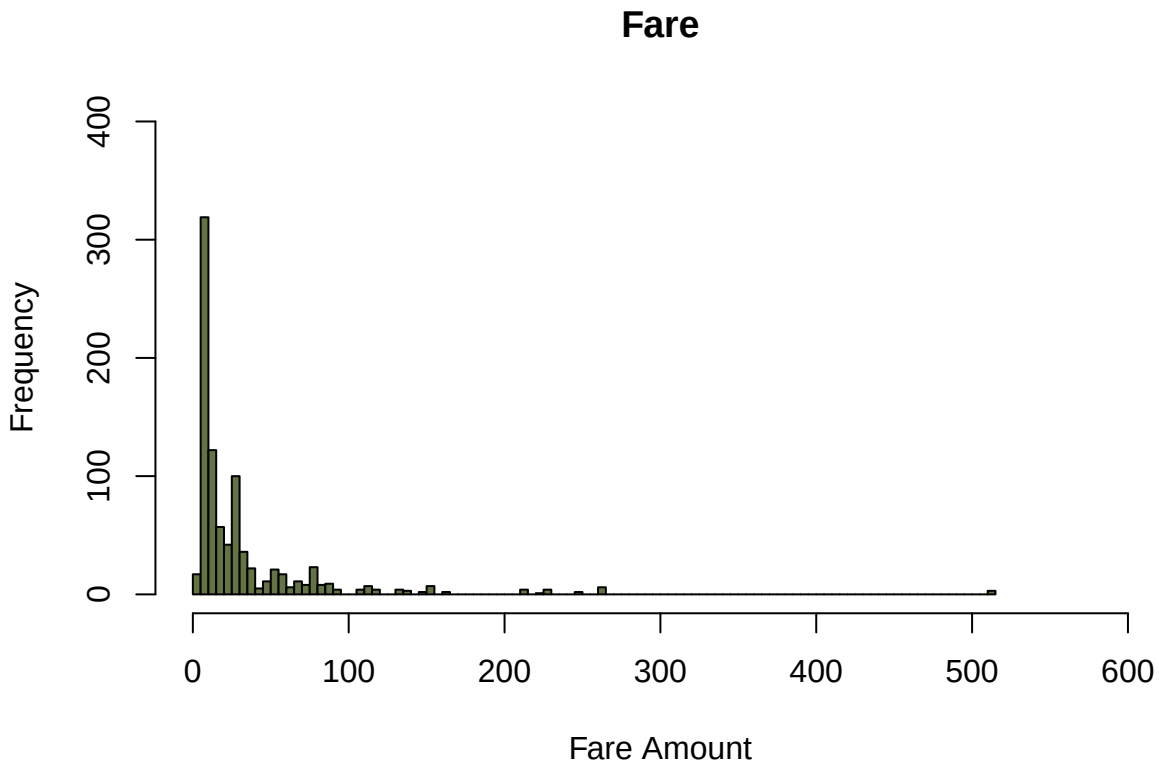
This bar plot shows the passengers count by sex. We can see that there were nearly twice as many men on board than women. A reason for this could be because of societal norms and gender roles at the time. Men were mostly the breadwinners so they had the means to afford the ticket.

```
barplot(table(titanic$Pclass),
        col=c("#AF9500","#B4B4B4","#6A3805"),
        border=c("#AF9500","#B4B4B4","#6A3805"),
        main="Class of Passengers",
        xlab="Class",
        ylab="Number of Passengers",
        ylim=c(0,600),
        #beside=TRUE
        names.arg = c("1st Class", "2nd Class", "3rd Class"),
        beside=TRUE
        )
```



The Class variable is categorical and shows what kind of ticket the person purchased, 1st Class, 2nd Class or 3rd Class Tickets. From the plot we can see that the majority of passengers were in 3rd Class. Interestingly 1st and 2nd Class have almost the same amount of passengers. I would have assumed less people to be in 1st Class.

```
hist(titanic$Fare,
     breaks = 75,
     main = "Fare",
     xlab = "Fare Amount",
     ylab = "Frequency",
     ylim = c(0,400),
     xlim = c(0,600),
     col="#637343",
     )
```



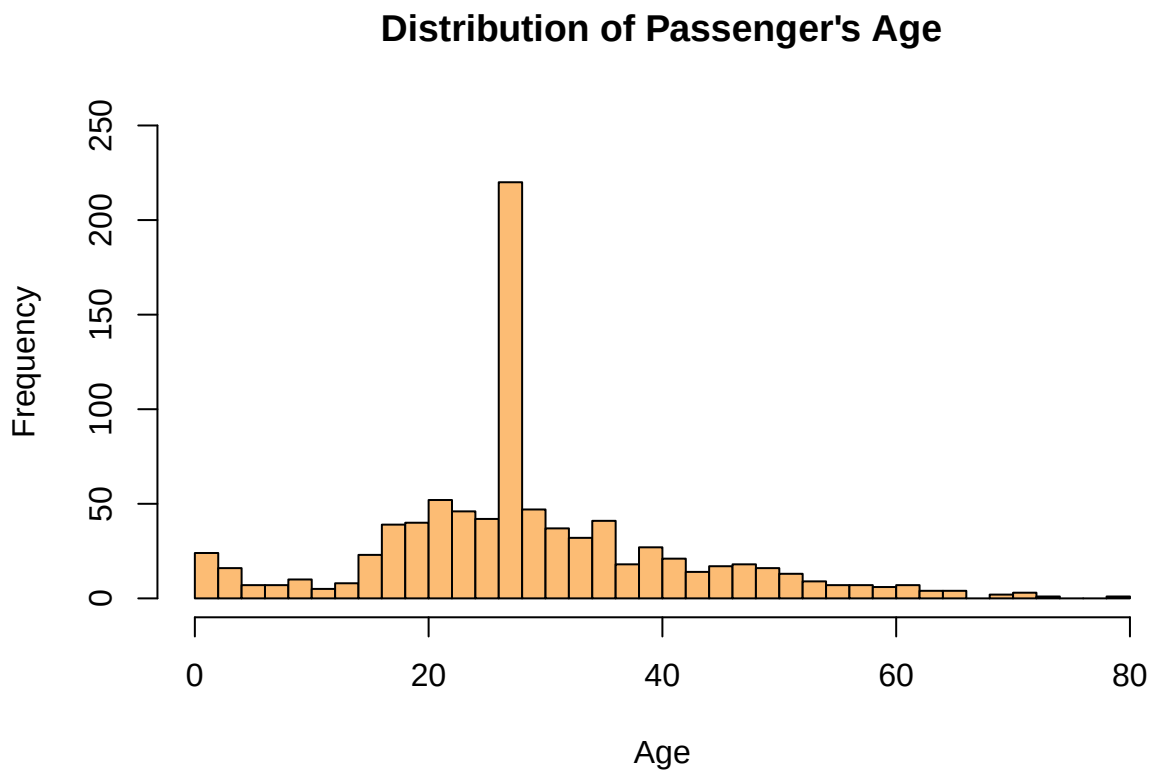
The distribution of fare is skewed to the right heavily. The majority of passengers paid between $10 and $40. There were some passengers on board who paid $0 - $5. The outlier fare amount is approximately $520. The next highest value is approximately $270. This is an extreme outlier that strongly deviates from the data.

Cleaning the Data

```
# get the median of the 714 passengers
median_age <- median(titanic$Age, na.rm=TRUE)

# use the logical vector median_age to replace values in the titanic df
missing_age <- is.na(titanic$Age)
titanic$Age[missing_age] <- median_age


hist(titanic$Age,
     breaks = 50,
     main = "Distribution of Passenger's Age",
     xlab = "Age",
     ylab = "Frequency",
     ylim = c(0,250),
     col="#fcbc74",
     )
```



The original data set has missing Age values for 177 of the 891 passengers. I replaced the missing values with the median age which is 28 years old. This changed the shape of the histogram. Now the data contains approximately 220 passengers around the age of 28. We can see this easily because of the huge peak in the data. .