Sean Toon

IT-460-18500-M01 Machine Learning 2024 C-5 (Sep - Oct)

9/28/2024

Module Three Lab and Report – R Data Structures and Visualization

The data set used for this lab is a training set that consists of passenger features related to the disaster of the Titanic. Some of the information that was included is name, age, gender, socio-economic class and whether the passenger survived the wreck or not.

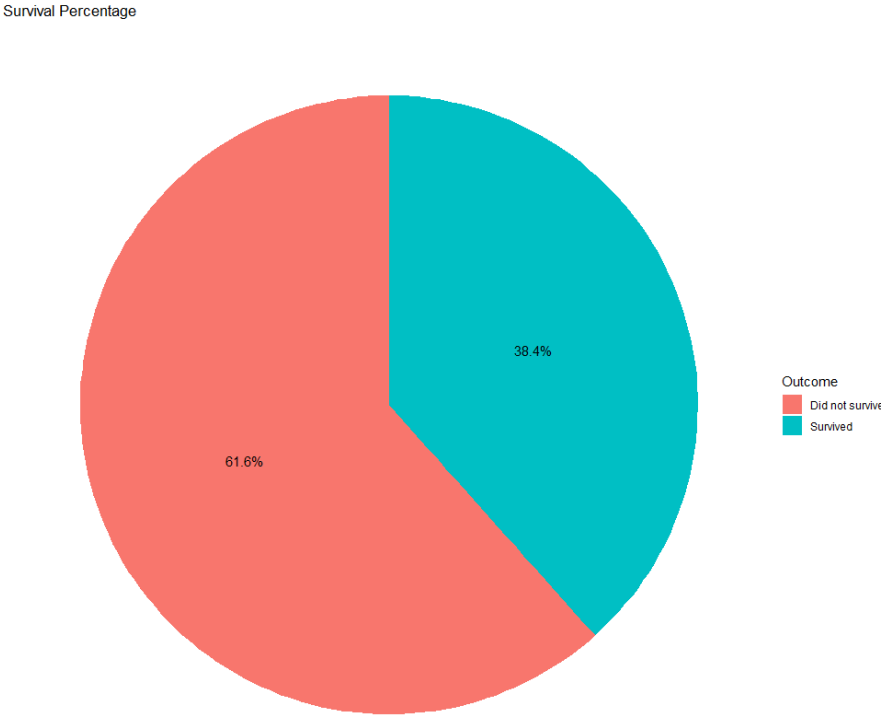
Firstly, I loaded the data in R just to get another idea about how the data was stored in the csv files and to make sure to use the correct variables further along the lab. Before creating any visualizations, I made a few manual calculations to get some more practice with R and to confirm the data that was provided.

```
> library(train.csv)
Error in library(train.csv) : there is no package called 'train.csv'
> library(train)
Error in library(train) : there is no package called 'train'
> train <- read.csv("\\Users\\toons\\Downloads\\train.csv")
> summary(train)
  PassengerId       Survived          Pclass          Name               Sex                 Age
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00
                                                                                        NA's   :177
     SibSp           Parch           Ticket               Fare           Cabin              Embarked
 Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00   Length:891         Length:891
 1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   Class :character
 Median :0.000   Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Mode  :character
 Mean   :0.523   Mean   :0.3816                      Mean   : 32.20
 3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00
 Max.   :8.000   Max.   :6.0000                      Max.   :512.33

> summary(train$Survived)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
> num_survivors <- sum(train$Survived)
> print(num_survivors)
[1] 342
> total_passengers <- nrow(train)
> print(total_passengers)
[1] 891
> survival_rate <- num_survivors / total_passengers
> print(survival_rate)
[1] 0.3838384
> survival_percentage <- survival_rate * 100
> print(survival_percentage)
[1] 38.38384
```

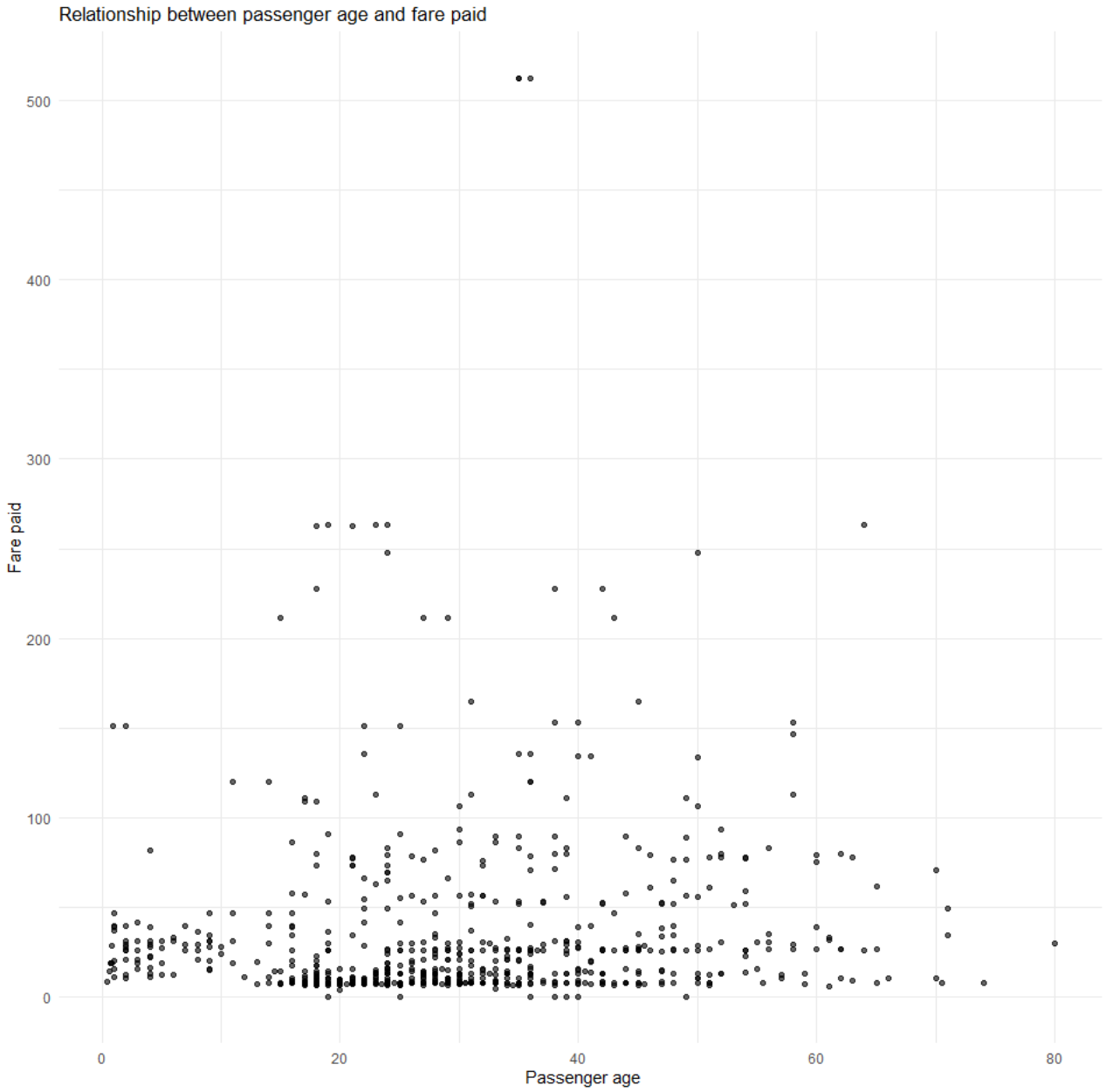
Visual for the above survival percentage commands and the commands used to generate the chart:

```
> >  ggplot(survival_data, aes(x = "", y = Count, fill = Outcome)) +
+ geom_bar(width = 1, stat = "identity") +
+ coord_polar("y") +
+ labs(title = "Survival Percentage") +
+ theme_void()+
+ geom_text(aes(label = paste0(round(Count / total_passengers * 100, 1), "%")), position = position_stack(vjust = 0.5))
>
```



Lastly, I showed the relationship between age of passenger and fare paid with the following command in a plot:

```
> ggplot(train, aes(x = Age, y = Fare)) + geom_point(alpha = 0.6) + labs(title = "Relationship between passenger age and fare paid", x = " Passenger age", y = "Fare paid") + theme_minimal()
Warning message:
Removed 177 rows containing missing values or values outside the scale range
(`geom_point()`).
>
```


Relationship between passenger age and fare paid
Fare paid
500
400
300
200
100
0
0
20
40
60
80
Passenger age