Problem Set 3

Caroline Robbins

2024-09-18

Task 1 Load kdrama dataset that we have used in Problem Set 1.

```
##Set working directory and bring in necessary libraries.
setwd("C:/Users/crobb/Downloads")
library(tidyverse)
library(tinytex)

##Load the KDrama dataset into the workspace.
kdrama_data <- read.csv("kdrama.csv")
```

1. Create a smaller dataset by selecting only three columns: Name, Aired.On, and Rating variables.

```
small_kdrama <- kdrama_data %>%
  select("Name", "Aired.On", "Rating")
small_kdrama
```

2. Construct a long data set where each row will be a movie per day of the week. Make sure that for each movie you only keep days of the week when it was actually aired. The long dataset should contain only 3 columns: Name, Aired day of the week, and Rating. What would be the dimensions of this new long dataset?

```
##View KDrama data
small_kdrama

##Remove the spaces from the dataset (will help with steps later)
no_space <- data.frame(gsub(" ", "", small_kdrama$Aired.On))
colnames(no_space)
```

```
## [1] "gsub.........small_kdrama.Aired.On."
```

```
##Rename the odd column names produced
rename <- no_space %>%
  rename(Aired.On = gsub.........small_kdrama.Aired.On.)

##Split the days the shows aired into individual variables
adjust_kdrama<-data.frame(small_kdrama,
  str_split_fixed(small_kdrama$Aired.On, ",", 7))

##Make sure all spaces are removed from the dataframe (problematic otherwise)
no_space <- adjust_kdrama %>%
  mutate(across(where(is.character), str_remove_all, pattern = fixed(" ")))

##Select out our variables of interest (Name, Aired.On, Rating, and the new
##Days of the week variables)
select_kdrama <- no_space %>%
  select(Name, Aired.On, Rating, X1, X2, X3, X4, X5, X6, X7)

##Transform the dataset so it is in long form with individual entries for each
##day of the week aired.
longer_kdrama <- select_kdrama %>%
  pivot_longer(cols = starts_with("X"),
             values_to = "Aired_day_of_the_week") %>%
  relocate(Name, .after = Aired_day_of_the_week) %>%
  select(-name)

##Remove any repeated rows from the dataset
unique_kdrama <- unique(longer_kdrama)
class(unique_kdrama$Aired_day_of_the_week)
```

```
## [1] "character"
```

```
##Filter the dataset so any values that have a blank in the Aired_day_of_the_
##week column are no longer included.
fixed_kdrama <- filter(unique_kdrama, Aired_day_of_the_week == "Monday" |
                         Aired_day_of_the_week == "Tuesday" |
                         Aired_day_of_the_week == "Wednesday" |
                         Aired_day_of_the_week == "Thursday" |
                         Aired_day_of_the_week == "Friday" |
                         Aired_day_of_the_week == "Saturday" |
                         Aired_day_of_the_week == "Sunday")

##Remove the old Aired.On variable from the dataset
longest_kdrama <- fixed_kdrama %>%
  select(Name, Aired_day_of_the_week, Rating)

##View the new dataset
longest_kdrama

##Collect the dimensions of the dataset
dim(longest_kdrama)
```

```
## [1] 474   3
```

```
##The dimensions of this dataset are 474 rows and 3 columns.
```

3. Transform this long dataset into a wide one. What are the dimensions of the new wide dataset?

```
wide_kdrama <- longest_kdrama %>%
  pivot_wider(names_from = Name, values_from = Rating)
wide_kdrama
dim(wide_kdrama)
```

```
## [1]   7 251
```

```
##The dimensions of this dataset are 7 rows and 251 columns.
```

Task 2 Load three datasets about Titanic passengers titanic.csv, titanic2.csv and survived.csv from the email with a PS3. You can also access datasets from the shared dropbox folder.

```
##Load the datasets into the workspace.
titanic_data <- read.csv("titanic.csv")
titanic2_data <- read.csv("titanic2.csv")
survived_data <- read.csv("survived.csv")
```

1. Print the first top rows of all three datasets to understand their structure. How many common columns are there between the three datasets? (Please write the response to the question either in the comments section of the R script or as a text field in your R Markdown document.)

```
head(titanic_data)
head(titanic2_data)
head(survived_data)

##Between all three datasets, the common column is PassengerId. Titanic data
##and Survived data both contain the Survived column, but Titanic2 does not.
```

2. Create a new dataset merged df that will be a result of merging titanic2.csv with survived.csv. What are the dimensions of the new merged dataset?

```
merged_df <- left_join(titanic2_data, survived_data, by = "PassengerId")
merged_df
dim(merged_df)
```

```
## [1] 418  12
```

```
##The dimensions of the merged dataframe are 418 rows and 12 columns.
```

3. Do merged df and titanic.csv have an overlap in the passenger list? Type up an answer and show how you derived this answer with an R code.

```
range(merged_df$PassengerId)
```

```
## [1]  892 1309
```

```
##892-1309
range(titanic_data$PassengerId)
```

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

```
##1-891

##The merged dataframe and titanic.csv do not have an overlap in the passenger
##list as the last passenger in the original dataframe is 891 where the merged
##dataframe begins with the next passenger, 892. All of these are unique values
##that do not overlap with each other, as demonstrated by the lack of shared
##values in the ranges.
```

4. Combine merged df and titanic.csv datasets into one dataset full df. What are the dimensions of this new full df dataset? (Hint: here you are asked not to merge, but to combine datasets together. Try exploring how functions rbind and cbind work.)

```
full_df <- rbind(titanic_data, merged_df)
full_df
dim(full_df)
```

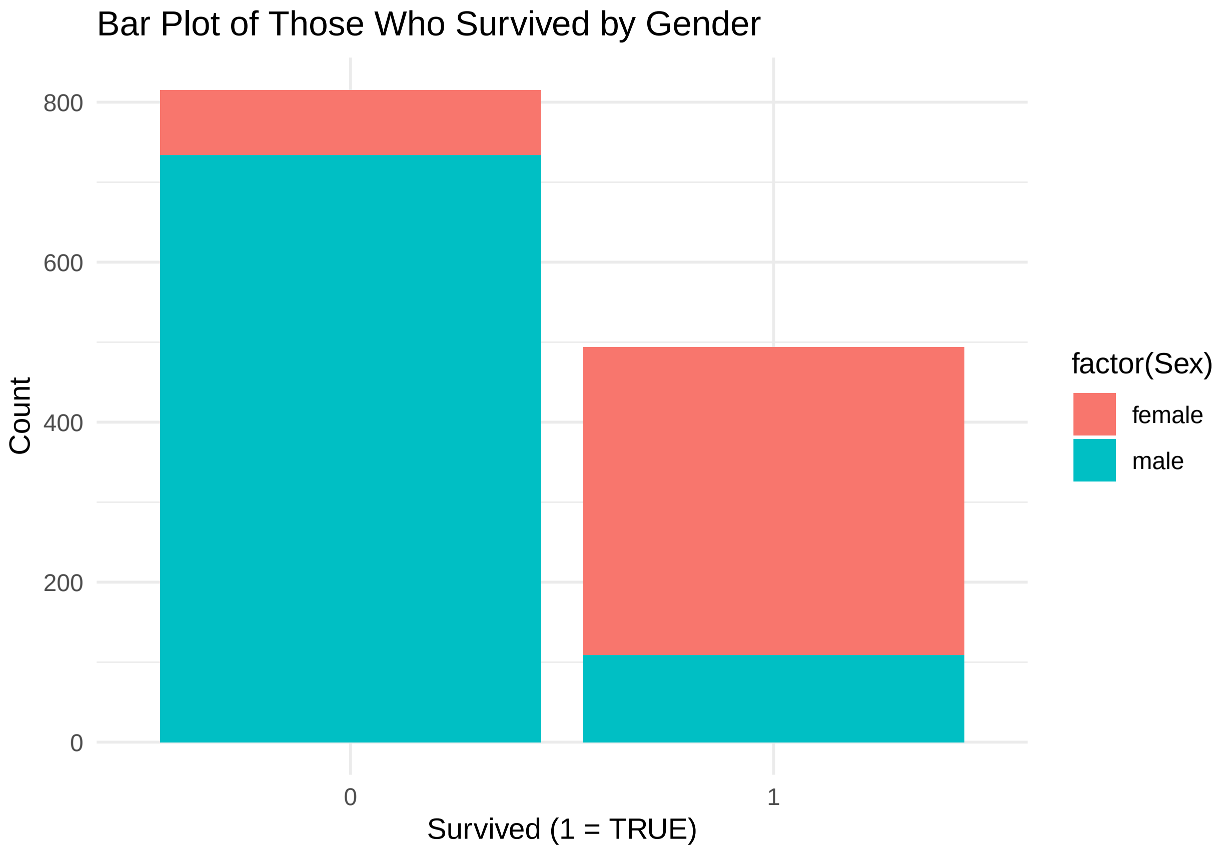
```
## [1] 1309   12
```

```
##This dataset has 1309 rows and 12 columns.
```

For the following questions you will work with the newly constructed full df (you might want to use ggplot for these tasks):

5. Create a bar plot showing the survival rate (Survived) by gender (Sex). How can you interpret the plot?

```
ggplot(full_df, aes(x = factor(Survived), fill = factor(Sex))) +
  geom_bar() +
  labs(
    title = "Bar Plot of Those Who Survived by Gender",
    x = "Survived (1 = TRUE)",
    y="Count"  ) +
  theme_minimal()
```

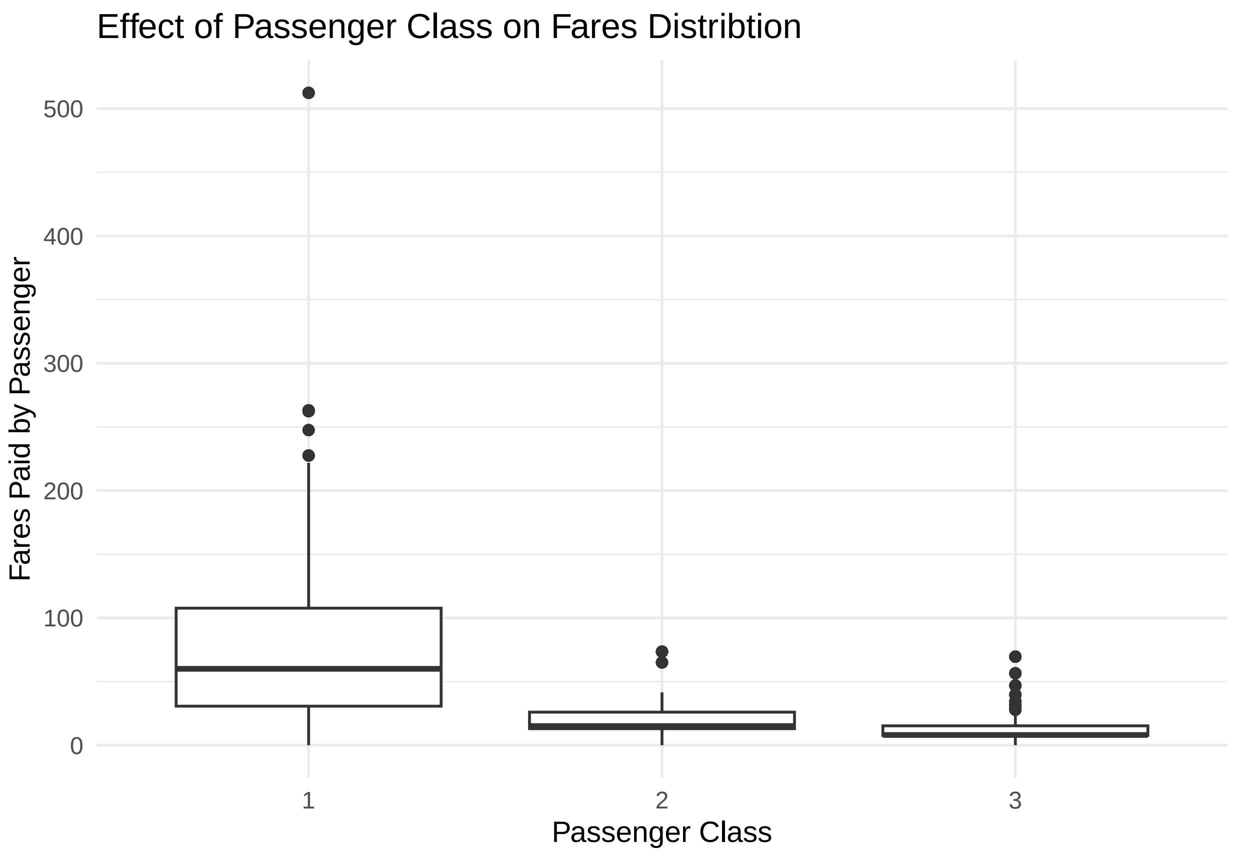


```
##Of the individuals who survived the sinking of the Titanic, the vast
##majority were women.
```

6. Create a boxplot to show the distribution of fares (Fares) paid by passengers in each passenger class (Pclass). How can you interpret the plot?

```
ggplot(full_df, aes(x = factor(Pclass), y = Fare)) +
  geom_boxplot() +
  labs(
    title = "Effect of Passenger Class on Fares Distribtion",
    x = "Passenger Class",
    y = "Fares Paid by Passenger"
  ) +
  theme_minimal()
```

```
## Warning: Removed 1 row containing non-finite outside the scale range (`stat_boxplot()`).
```

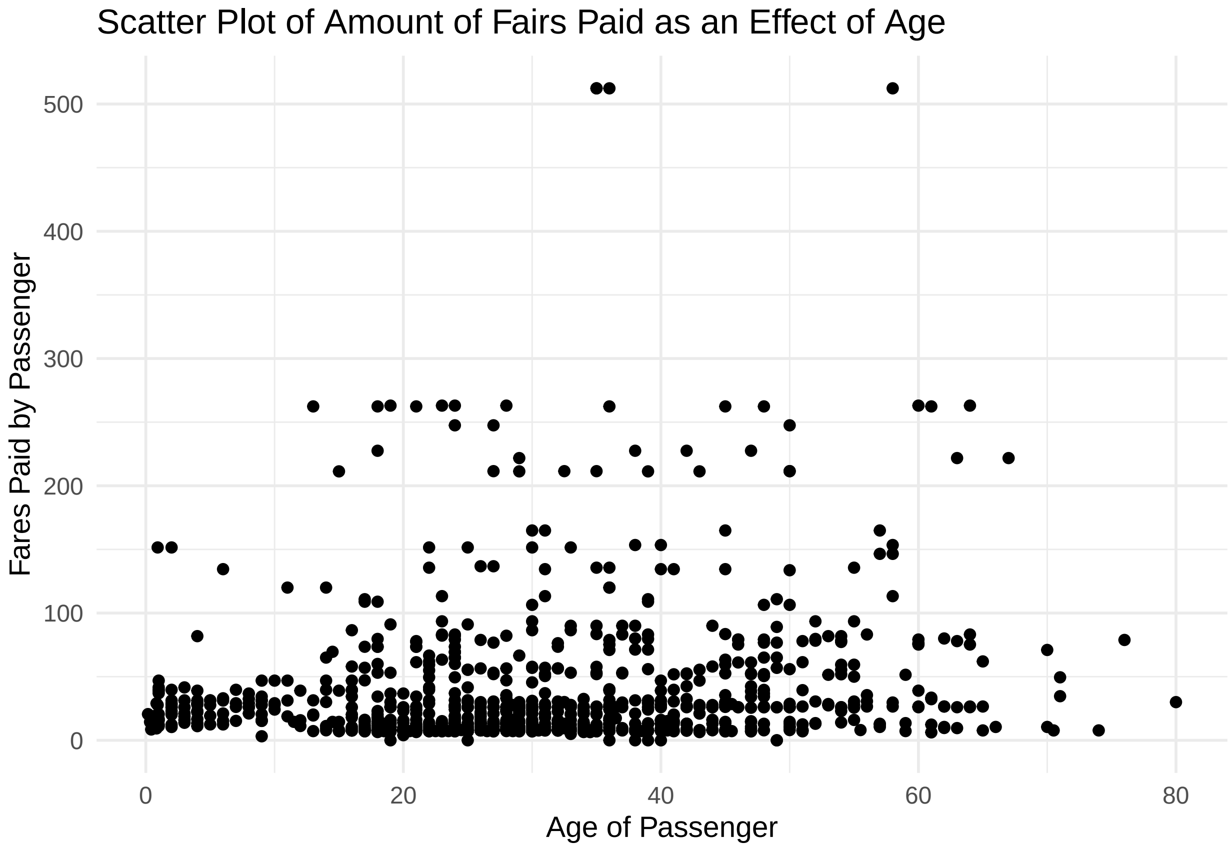


```
##Passengers in the first class paid the largest range of fares for the trip,
##including several outlying high costs. Comparatively, the second and third
##classes paid much less with lower levels of distribution and outliers on the
##higher ends of payment.
```

7. Create a scatter plot to visualize the relationship between passengers' ages (Age) and the fares they paid (Fare). Do you think there is an obvious correlation between Age and Fare?

```
ggplot(full_df, aes(x = Age, y = Fare)) +
  geom_point() +
  labs(
    title = "Scatter Plot of Amount of Fairs Paid as an Effect of Age",
    x = "Age of Passenger",
    y = "Fares Paid by Passenger"
  ) +
  theme_minimal()
```

```
## Warning: Removed 264 rows containing missing values or values outside the scale range (`geom_point()
```

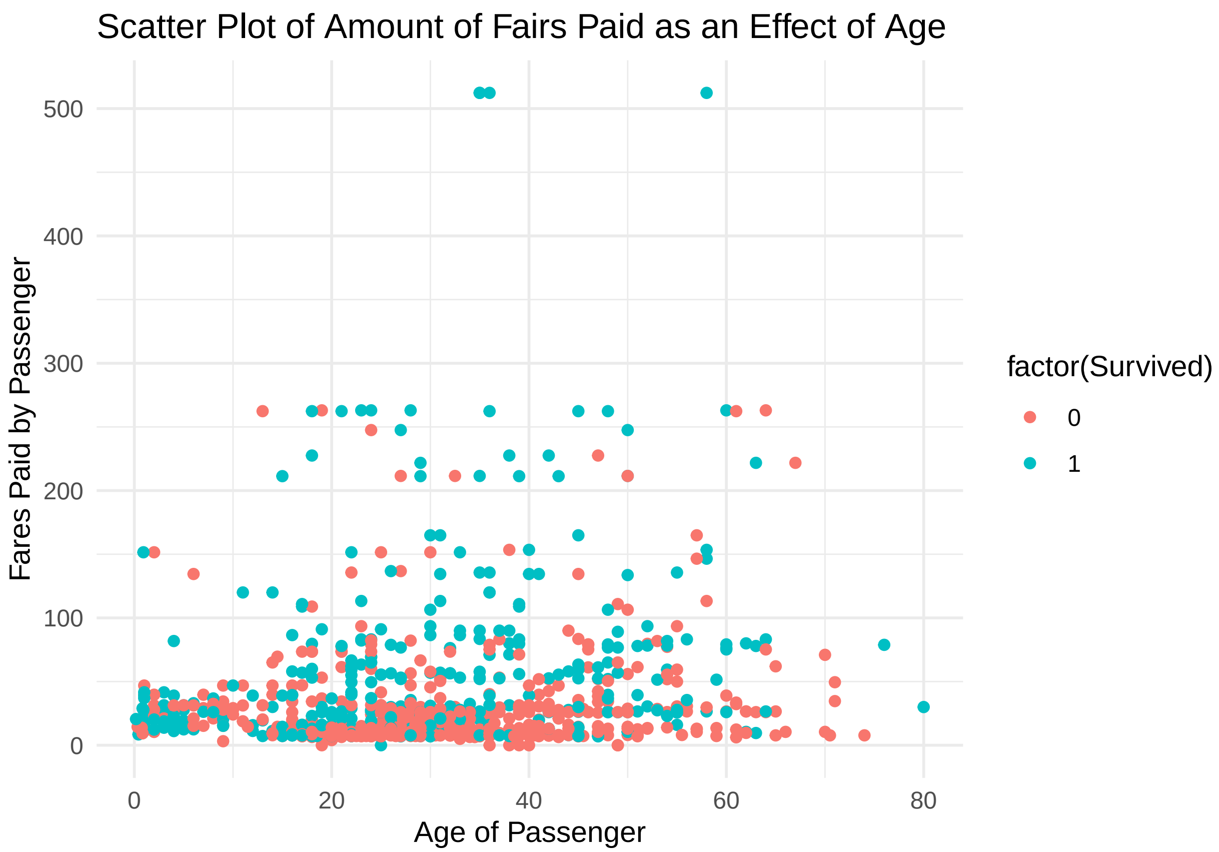


```
##Based on the initial scatterplot, there does not appear to be a correlation
##between age and amount of fares paid per passenger. The rate appears to be
##steady across age groupings.
```

8. On the same scatter plot use different colors to differentiate between passengers who survived and those who did not. How can you interpret the plot?

```r
ggplot(full_df, aes(x = Age, y = Fare, color = factor(Survived))) +
  geom_point() +
  labs(
    title = "Scatter Plot of Amount of Fairs Paid as an Effect of Age",
    x = "Age of Passenger",
    y = "Fares Paid by Passenger"
  ) +
  theme_minimal()
```

```
## Warning: Removed 264 rows containing missing values or values outside the scale range (`geom_point()
```

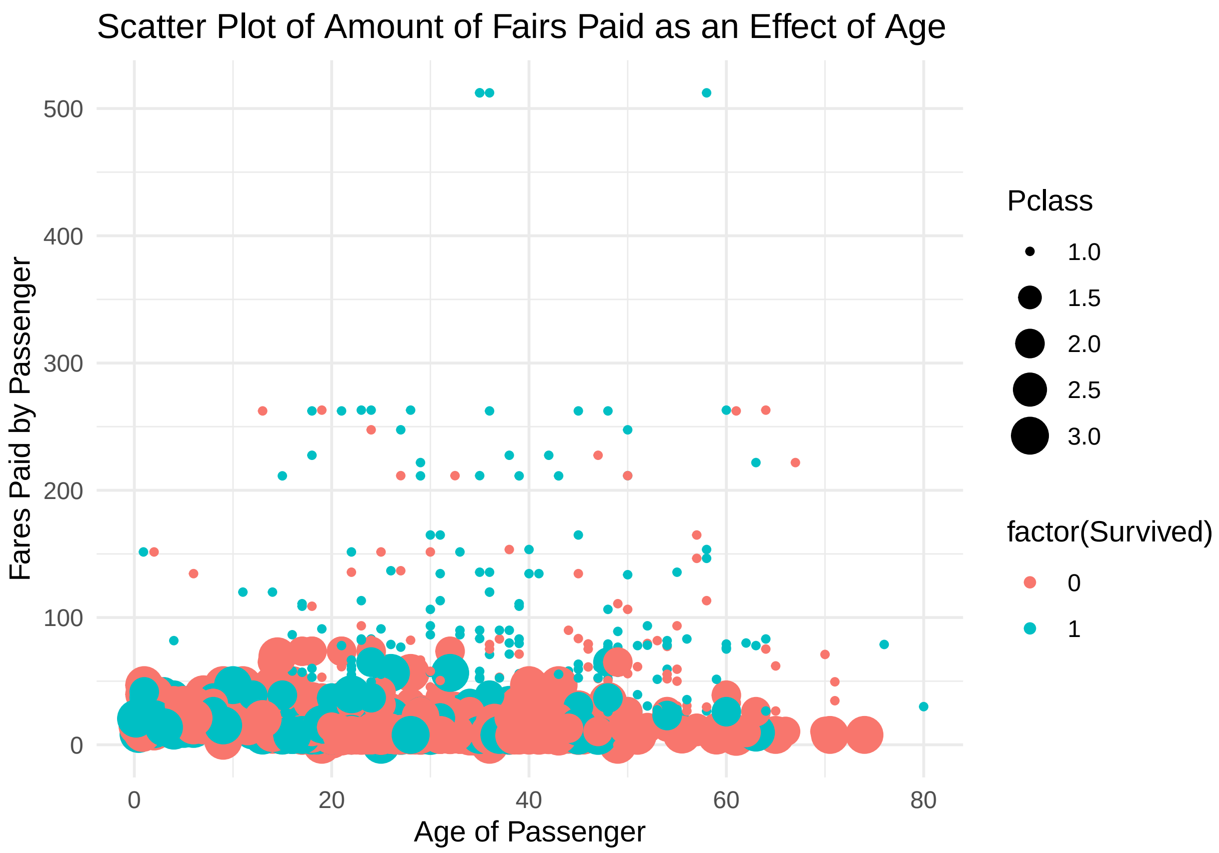


```
##The majority of individuals who paid more in terms of fares were more likely
##to survive when compared to their counterparts who paid less. Additionally,
##more individuals from low to middle age ranges were likely to survive.
```

9. On the same scatter plot that you got in the previous point now additionally use different point sizes to differentiate between passengers in each passenger class (Pclass). How can you interpret this new result?

```
ggplot(full_df, aes(x = Age, y = Fare, color = factor(Survived))) +
  geom_point(aes(size = Pclass)) +
  labs(
    title = "Scatter Plot of Amount of Fairs Paid as an Effect of Age",
    x = "Age of Passenger",
    y = "Fares Paid by Passenger"
  ) +
  theme_minimal()
```

```
## Warning: Removed 264 rows containing missing values or values outside the scale range (`geom_point()
```



```
##While the full trends of this graph are not entirely observable due to the
##overlapping datapoints, there is a clear distinction of how those in lower
##classes (higher number values) were more likely to not survive when compared
##to those in higher classes (lower number values).
```