Problem Set 3

Bryson Lyons

2024-09-21

Task 1

Select Kdrama

```
kdrama_select<- select(kdrama, c(Name, `Aired On`, Rating))
```

Long Data

The long data set has 474 rows and 3 columns.

```
kdrama_select<- kdrama_select %>%
  separate_rows(`Aired On`, sep = ", ")

dim(kdrama_select)
```

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

Wide Data

The wide data set has 250 rows and 8 columns.

```
kdrama_wide <- kdrama_select %>%
  pivot_wider(names_from = `Aired On`, values_from = Rating)

dim(kdrama_wide)
```

```
## [1] 250   8
```

Task 2

Data Headers

The PassengerID column is the only column included in all three datasets.

```r
head(titanic)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  C123
## 5           5        0      3 Allen,~ male     35     0     0 373450  8.05 <NA>
## 6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 <NA>
## # i 1 more variable: Embarked <chr>
```

```r
head(titanic2)
```

```
## # A tibble: 6 x 11
##   PassengerId Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin Embarked
##         <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>
## 1         892      3 Kelly,~ male   34.5     0     0 330911  7.83 <NA>  Q
## 2         893      3 Wilkes~ fema~  47       1     0 363272  7    <NA>  S
## 3         894      2 Myles,~ male   62       0     0 240276  9.69 <NA>  Q
## 4         895      3 Wirz, ~ male   27       0     0 315154  8.66 <NA>  S
## 5         896      3 Hirvon~ fema~  22       1     1 31012~ 12.3  <NA>  S
## 6         897      3 Svenss~ male   14       0     0 7538    9.22 <NA>  S
```

```r
head(survived)
```

```
## # A tibble: 6 x 2
##   PassengerId Survived
##         <dbl>    <dbl>
## 1         892        0
## 2         893        1
## 3         894        0
## 4         895        0
## 5         896        1
## 6         897        0
```

Merging Data

The new, merged dataset has 418 rows and 12 columns.

```r
merged_df<- titanic2  %>% left_join( survived, by =c( "PassengerId" = "PassengerId"))

dim(merged_df)
```

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

Overlap

There does not seem to be an overlap of PassengerId between the two datasets. The "tail" of titanic and the "head" of merged_df do not include any of the same Passenger IDs.

```
tail(titanic)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1         886        0      3 "Rice,~ fema~    39     0     5 382652 29.1  <NA>
## 2         887        0      2 "Montv~ male     27     0     0 211536 13    <NA>
## 3         888        1      1 "Graha~ fema~    19     0     0 112053 30    B42
## 4         889        0      3 "Johns~ fema~    NA     1     2 W./C.~ 23.4  <NA>
## 5         890        1      1 "Behr,~ male     26     0     0 111369 30    C148
## 6         891        0      3 "Doole~ male     32     0     0 370376  7.75 <NA>
## # i 1 more variable: Embarked <chr>
```

```
head(merged_df)
```

```
## # A tibble: 6 x 12
##   PassengerId Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin Embarked
##         <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>
## 1         892      3 Kelly,~ male   34.5     0     0 330911  7.83 <NA>  Q
## 2         893      3 Wilkes~ fema~  47       1     0 363272  7    <NA>  S
## 3         894      2 Myles,~ male   62       0     0 240276  9.69 <NA>  Q
## 4         895      3 Wirz, ~ male   27       0     0 315154  8.66 <NA>  S
## 5         896      3 Hirvon~ fema~  22       1     1 31012~ 12.3  <NA>  S
## 6         897      3 Svenss~ male   14       0     0 7538    9.22 <NA>  S
## # i 1 more variable: Survived <dbl>
```

Combining Datasets

The new dataset has 1309 rows and 12 columns.

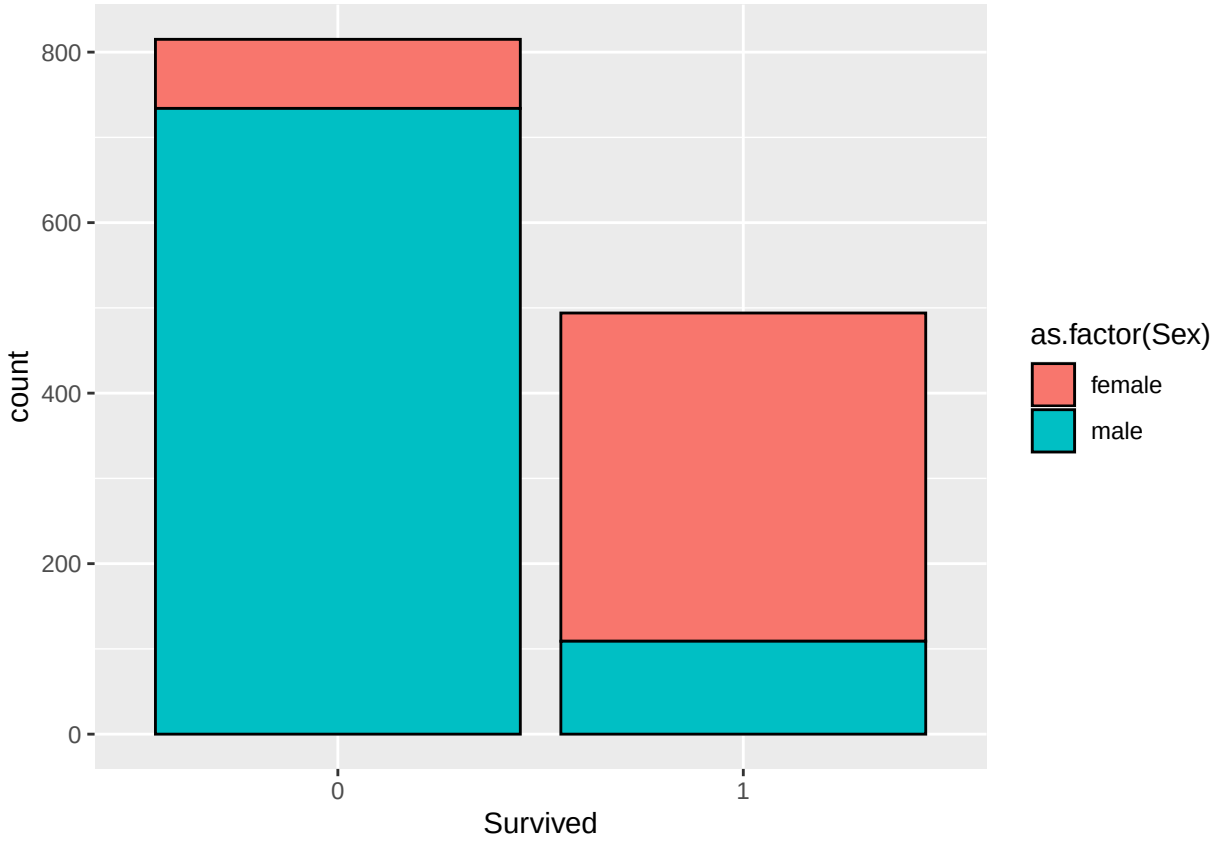
```
full_df <- rbind(titanic, merged_df)
dim(full_df)
```

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

Survival Bar Plot

The "Survived" variable is binary, with a 1 indicating the passenger survived and a 0 indicating they did not. A couple of immediate observations from this plot include: 1. From our dataset, there were less survivors than casualties. 2. Of those survivors, most of them were women. 3. Most men in the dataset lost their life.

```
ggplot(full_df, aes(as.factor(Survived), fill=as.factor(Sex)))+
  geom_bar(color="black")+
  labs(x="Survived")
```

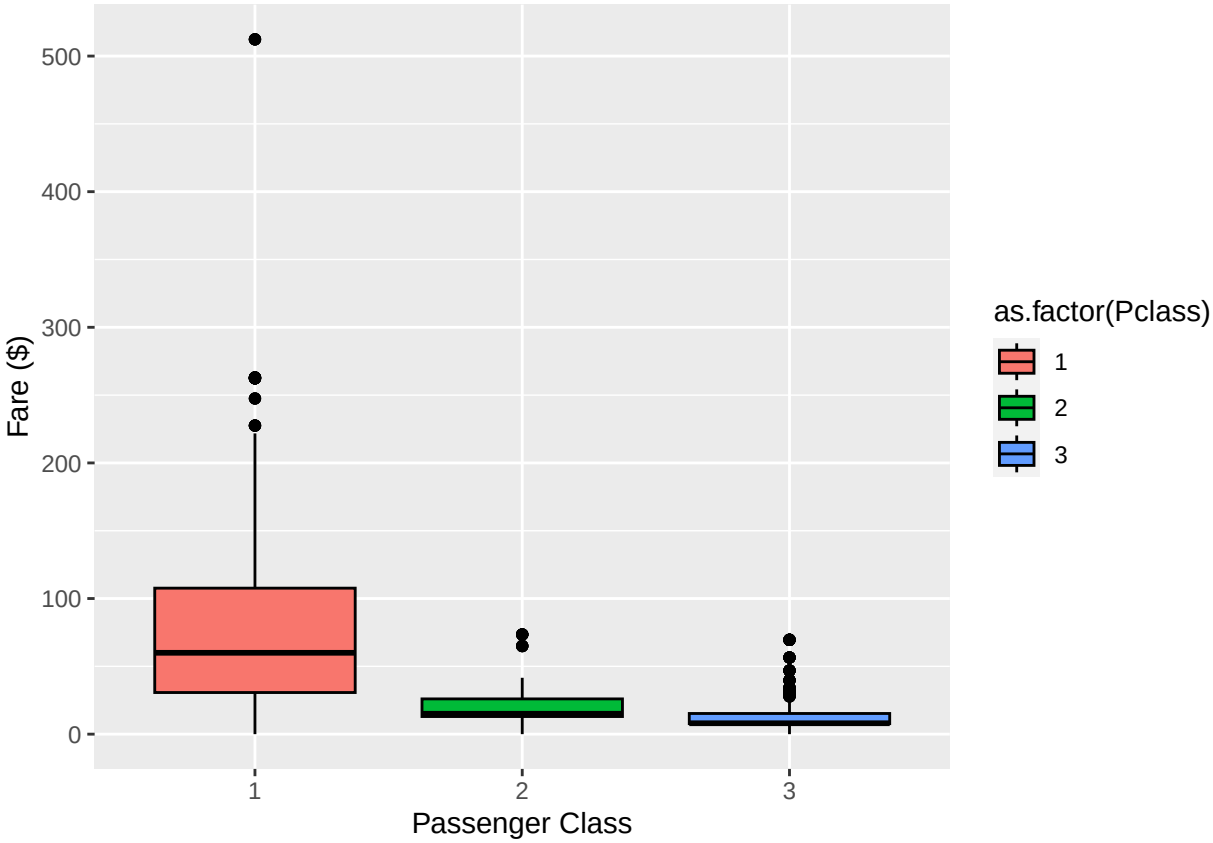


Fare Boxplot

From our boxplot, we can see that there was not a very wide distribution of fares between the cheaper 2nd and 3rd passenger classes. The first class spent more money than the other classes and had a much wider distribution. The maximum fare spent was north of $500 dollars. The median (black line in the box) fare spent by the first class was a bit more than $50.

```
ggplot (full_df,  aes( x=as.factor(Pclass),  y=Fare, fill=as.factor(Pclass) ))+
   geom_boxplot(  color="black" ) +
  labs(x= "Passenger Class", y= "Fare ($)")
```

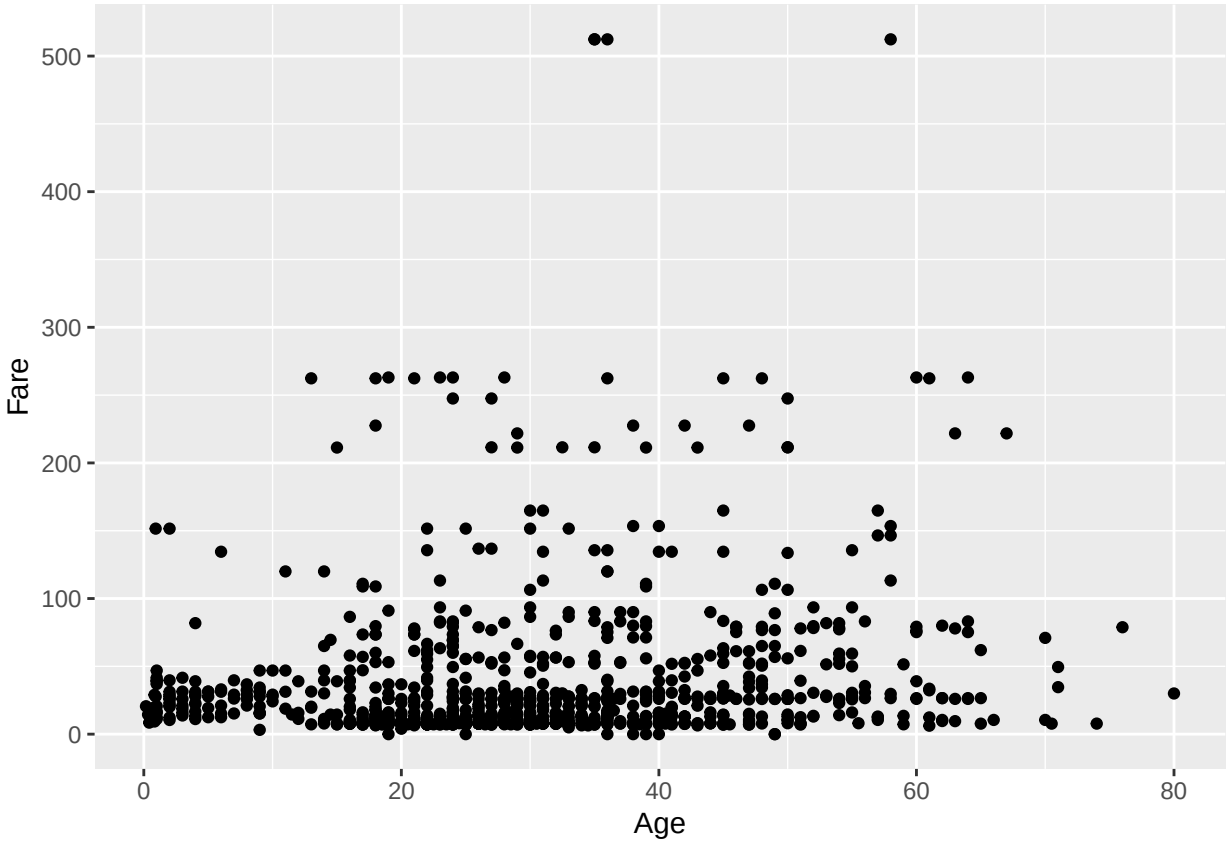
```
## Warning: Removed 1 rows containing non-finite values (‘stat_boxplot()‘).
```



Scatter Plot

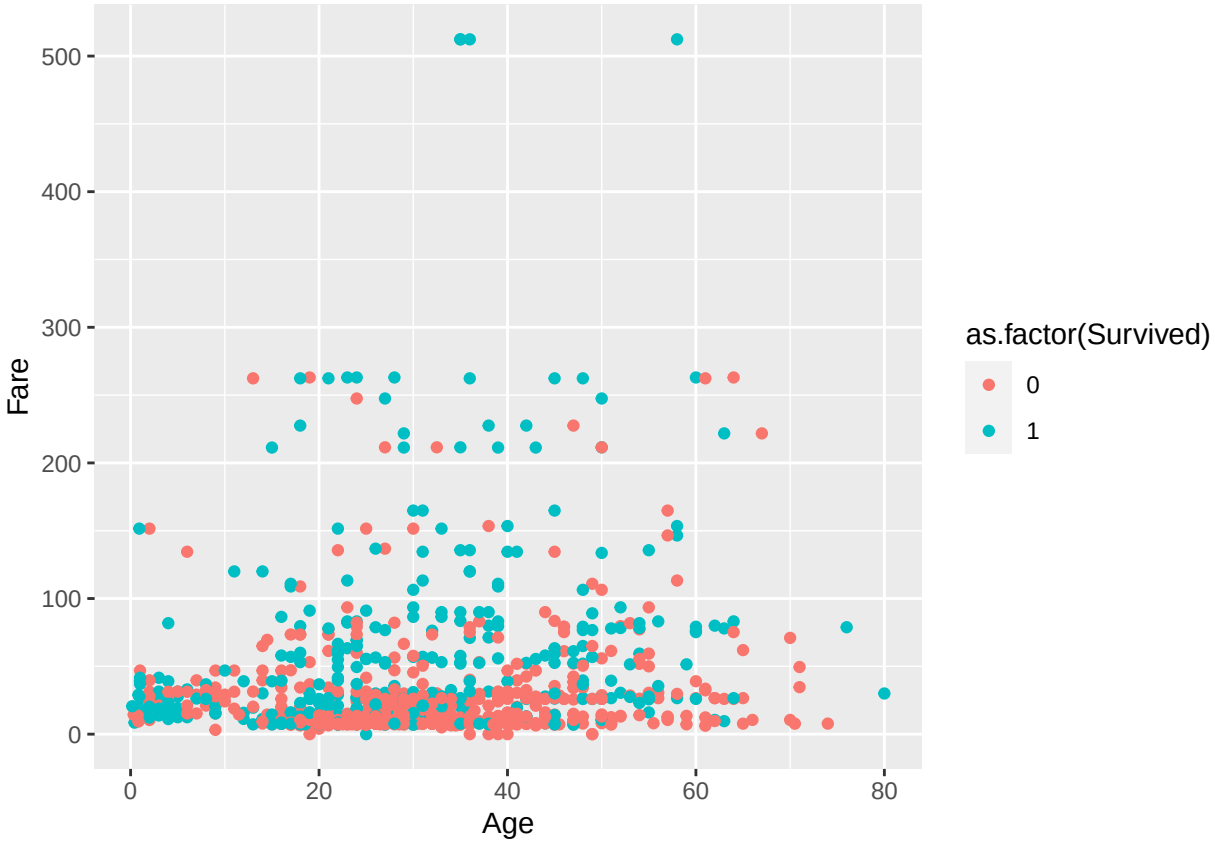
There seems to be a slight positive relationship with age and fare among our dataset.

```
ggplot(full_df, aes(x=Age, y= Fare))+
  geom_point()
```



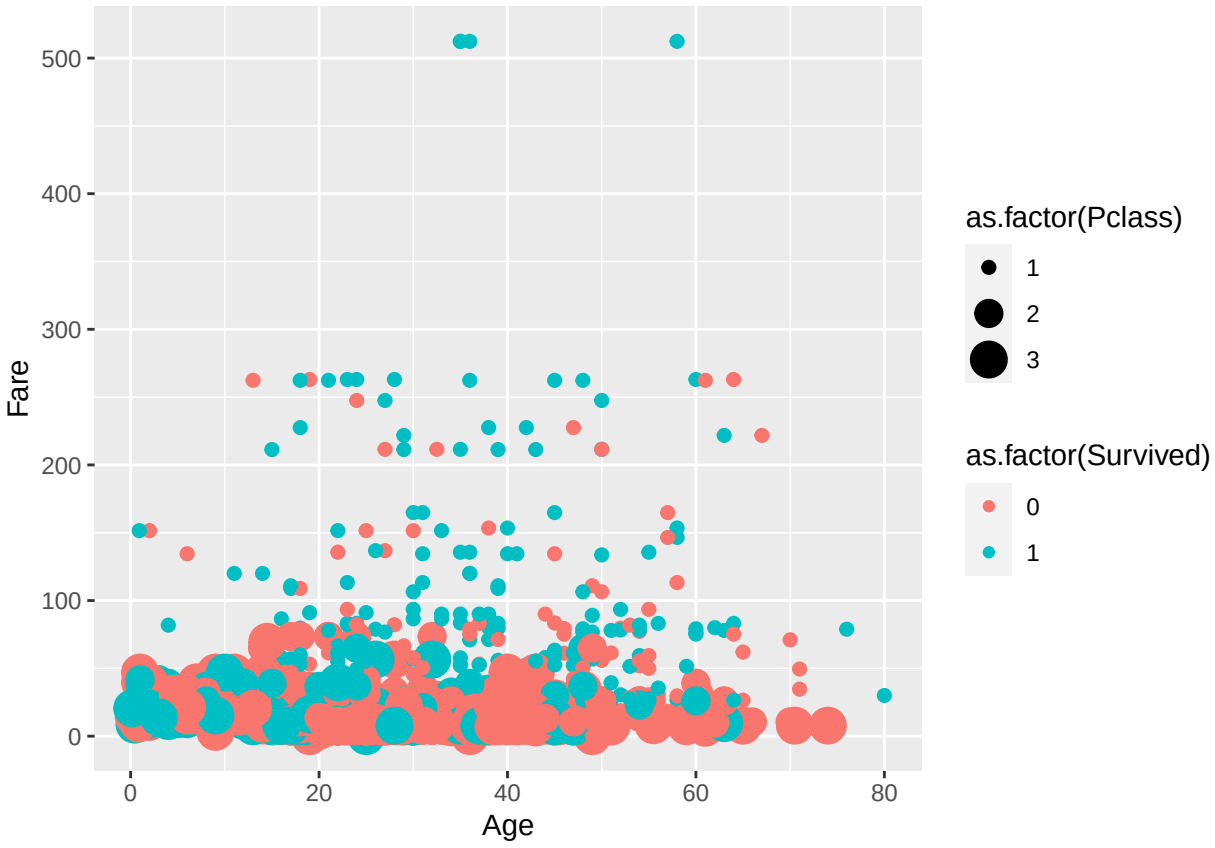
It seems that those who spent more than $100 on their fare had a better chance to survive. Passengers between the age of 20 and 60 who paid the lowest fare seem to be a significant portion of the casualties.

```
ggplot(full_df, aes(x=Age, y= Fare))+
  geom_point(aes(col=as.factor(Survived)))
```



The addition of size by passenger class kills most of the interpretability for any passengers who spent less than $100 on their fare. Every passenger above the low fare mass were all first class. This addition doesn't give any new findings, as we already knew those who spent more money had a better chance to survive, so including passenger class is somewhat redundant.

```
ggplot(full_df, aes(x=Age, y= Fare))+
  geom_point(aes(col=as.factor(Survived), size=as.factor(Pclass)))
```


Fare
500
400
300
200
100
0
0
20
40
60
80
Age
as.factor(Pclass)
1
2
3
as.factor(Survived)
0
1