Titanic

Yafet A Mekonnen

10/27/2021

Introduction

The sinking of the Titanic is one of the most famous shipwrecks in history. On April 15 1912, the mostly known unsinkable Titanic ship sank after colliding with an iceberg. Due to the lack of lifeboats, 1502 out of 2240 passengers and crew lost their life. I am interested in this dataset to see how humans handle life and death situations when they have to make the tough choice of who gets on the lifeboat.

```
titanic <- (read.csv("~/Desktop/titanic.csv"))
```

This dataset comes from a source called Vanderbilt Biostatistics Datasets which is an online community where data scientists go to find useful datasets. The limitation to this dataset is that it includes 1309 passengers data while the titanic ship had 2240 passengers, which means 931 passengers data is missing.

- `pclass`: passenger class; proxy for socio-economic status (1st ~ upper, 2nd ~ middle, 3rd ~ lower)
- `survived`: survival status (0=No, 1=Yes) they are represented as binary since there are only two choice 0 for not surivied and 1 for surivied
- `name`: passenger name
- `sex`: passenger sex (male, female)
- `age`: passenger age in years (fractional if age is less than 1 and if age is estimated, it is in the form xx.5)
- `sibsp`: number of siblings/spouses aboard (includes step-siblings, mistresses and fiances ignored)
- `parch`: number of parents/children aboard (parent only considers mother or father and child includes stepchildren)
- `ticket`: ticket number
- `fare`: passenger fare (in pre-1970 British pounds)
- `cabin`: cabin number
- `embarked`: port of embarkation (C=Cherbourg, Q=Queenstown, S=Southampton)
- `boat`: lifeboat number (if passenger boarded one)
- `body`: body identification number
- `home.dest`: passenger home/destination

```
titanic_new <- subset(titanic,select = -c(name,sibsp,parch,ticket,cabin,boat,body,home.dest,embarked))
```

I chose to remove columns like name,sibsp,parch,ticket,cabin,boat,body,home.dest and embarked since they are not going to be useful for the purpose of the visualization

```
colSums(is.na(titanic_new)) # see any missing vaules
```

```
##   pclass survived      sex      age     fare
##        0        0        0      263        1
```

Above we can clearly see that age column got 263 missing vaules and fare column has 1 missing vaule

```
titanic_new$fare[is.na(titanic_new$fare)] <-mean(na.omit(titanic_new$fare))
```

Replace the missing vaule in fare with the mean

```
titanic_new$age[is.na(titanic_new$age)] <-median(na.omit(titanic_new$age))
```

Replace the missing vaules in age with the median vaule

```
colSums(is.na(titanic_new)) # see any missing vaules
```

```
##   pclass survived      sex      age     fare
##        0        0        0        0        0
```

```
titanic_new$survived[titanic_new$survived == "1"] <- "Alive"
titanic_new$survived[titanic_new$survived == "0"] <- "Deceased"
```

Replace the vaules for survived column from 1 to Alive and from 0 to Deceased.

```
max_age <- max(titanic_new$age)
max_age
```

```
## [1] 80
```

```
titanic_new$age[titanic_new$age < 19 ] <- "Young"

titanic_new$age[titanic_new$age >= 19 & titanic_new$age <= 50 ] <- "Adult"

titanic_new$age[titanic_new$age >= 51 & titanic_new$age <= max_age ] <- "Older"
```

Breaking down the age column by separating into three different categories which is the age less than 19 to be called Young, the age between 19 and 50 to be Adult and above 50 to be Older age group.

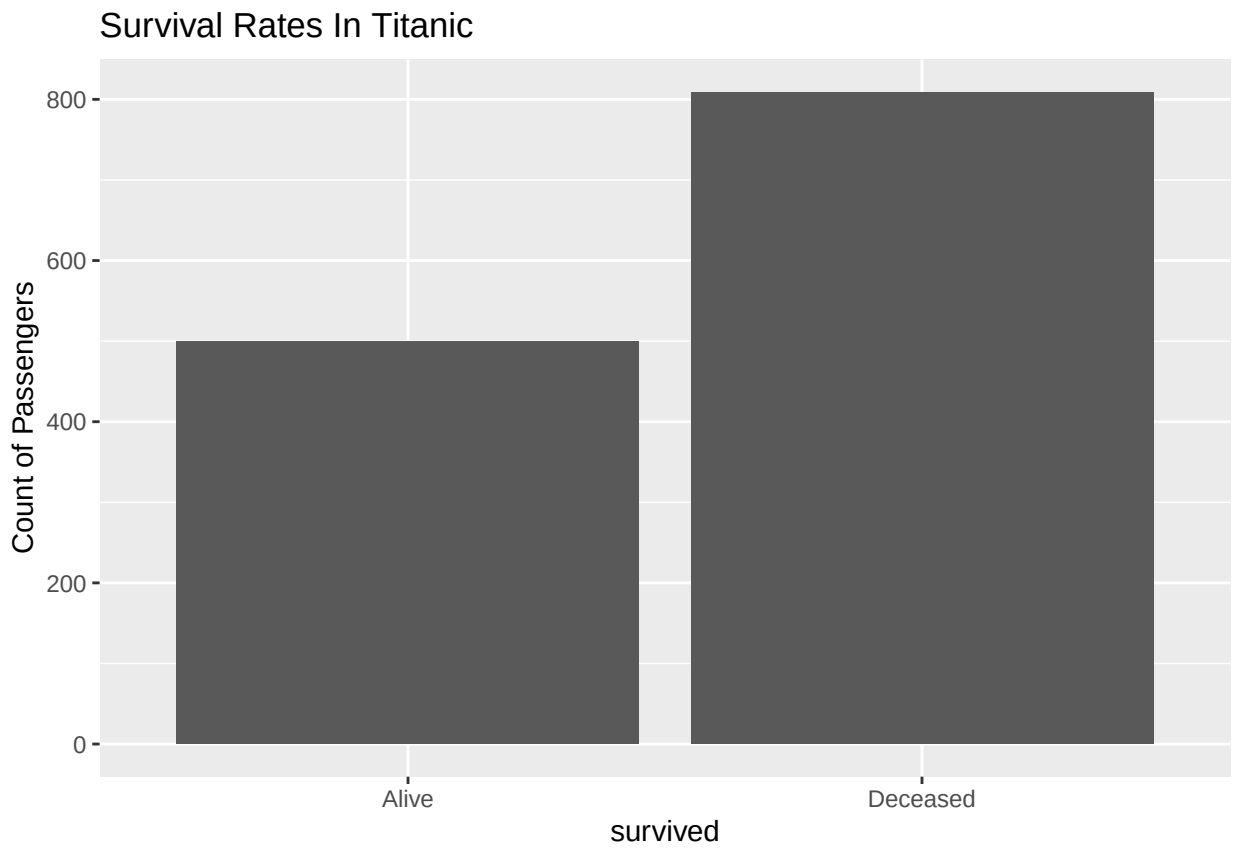
```
titanic_new$pclass[titanic_new$pclass == "1"] <- "Upper Class"

titanic_new$pclass[titanic_new$pclass == "2"] <- "Middle Class"

titanic_new$pclass[titanic_new$pclass == "3"] <- "Lower Class"
```

Changing the socio-economic status from 1 to Upper Class, 2 to Middle Class and 3 to the Lower class. This will be helpful for better understanding the graphs below and creating improved visualization.

```
head(titanic_new)
```

```
##        pclass survived    sex   age     fare
## 1 Upper Class    Alive female Adult 211.3375
## 2 Upper Class    Alive   male Young 151.5500
## 3 Upper Class Deceased female Young 151.5500
## 4 Upper Class Deceased   male Adult 151.5500
## 5 Upper Class Deceased female Adult 151.5500
## 6 Upper Class    Alive   male Adult  26.5500
```

```
ggplot(titanic_new, aes(x=survived)) + geom_bar() + labs(y = "Count of Passengers", title = "Survival Ra
```



Above it can be seen clearly that the majority of the passengers did not survive the shipwrecks

```
table(titanic_new$survived)
```

```
##
##    Alive Deceased
##      500      809
```

```
prop.table(table(titanic_new$survived))*100
```

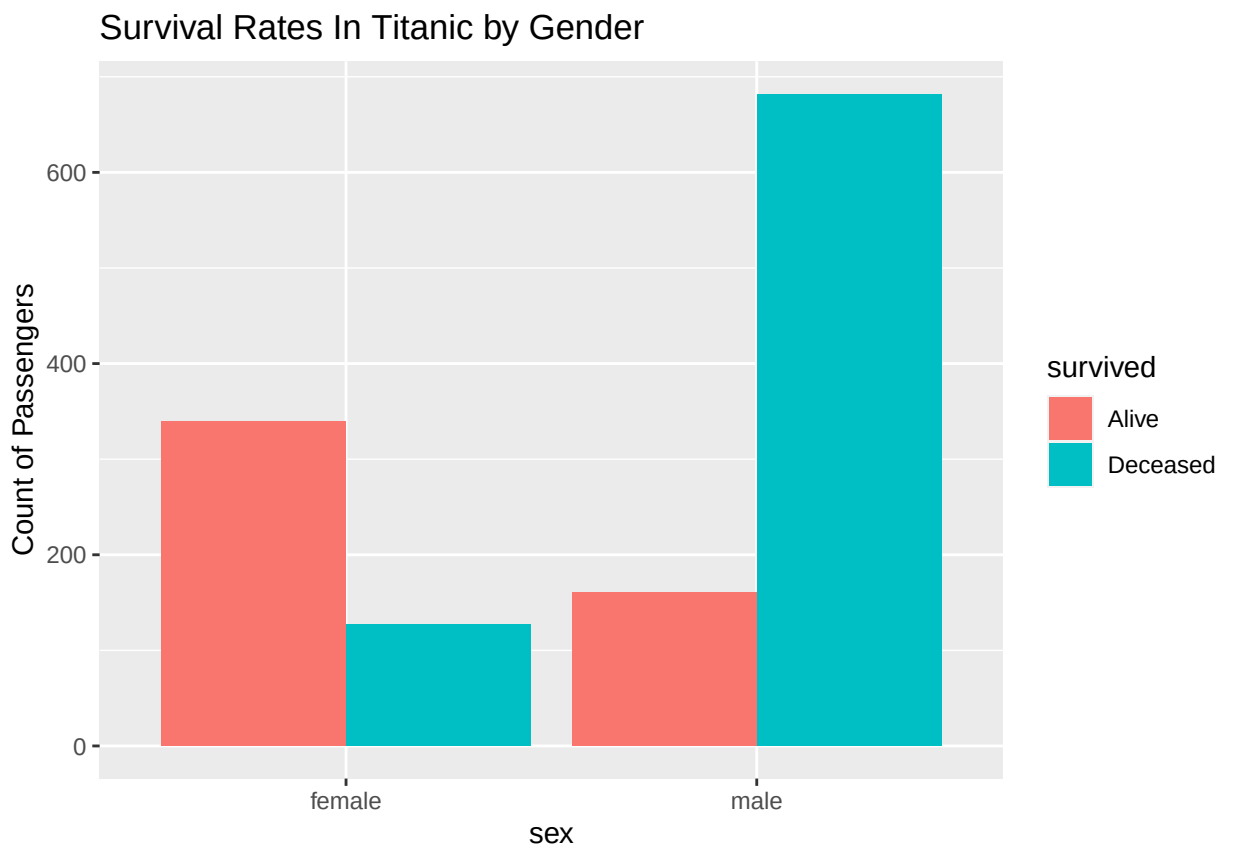
```
##
##    Alive Deceased
##  38.1971  61.8029
```

```
nrow(titanic_new)
```

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

Only 38% of the passengers which is 500 passengers survived the shipwrecks and the rest 61.8 % which is 809 did not survive.

```
ggplot(titanic_new, aes(x=sex, fill=survived)) + geom_bar(position = "dodge") + labs(y = "Count of Passe
```



The graph above indicates that most male passengers did not survive and most females survived this is because only females were allowed to board the lifeboats first.

```
table(titanic_new$survived , titanic_new$sex)
```

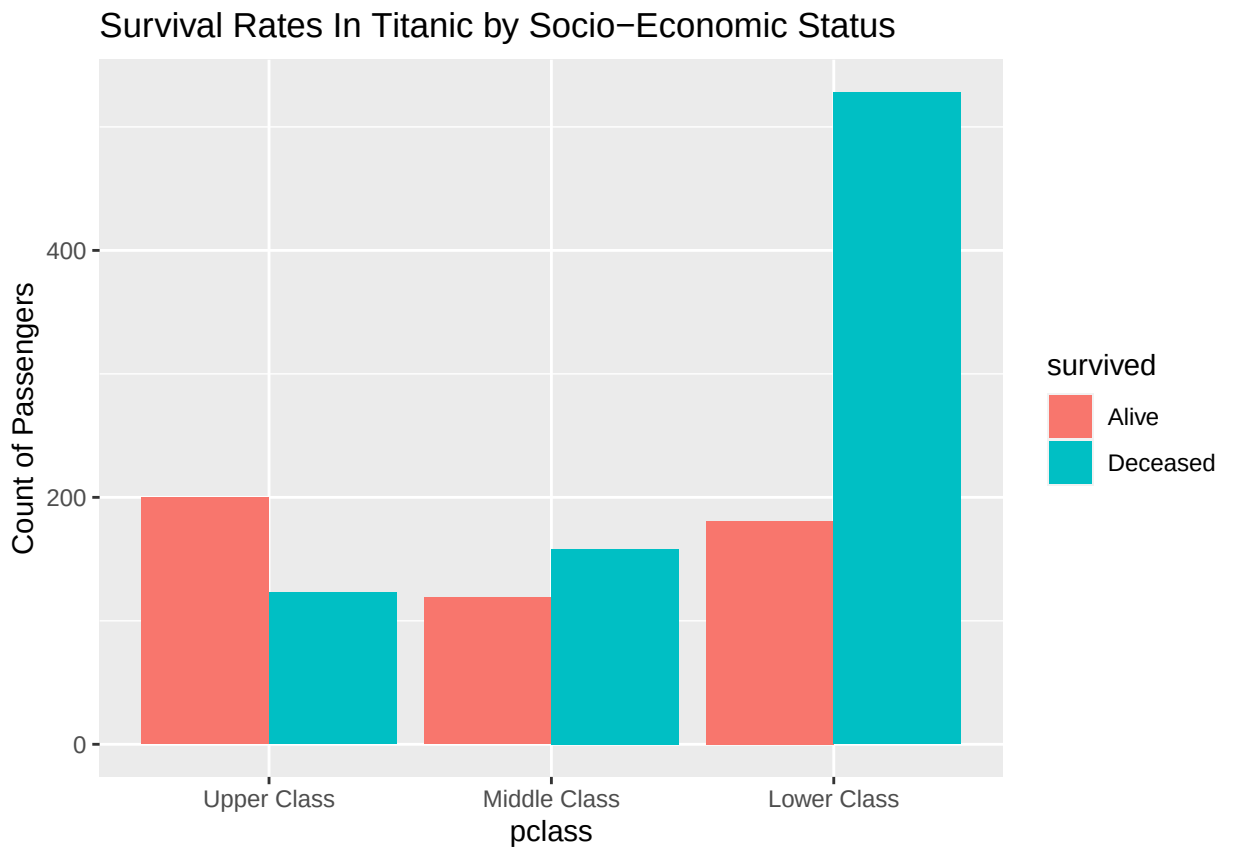
```
##
##            female male
##   Alive       339  161
##   Deceased    127  682
```

```
prop.table(table(titanic_new$survived , titanic_new$sex))*100
```

```
##
##                female      male
##   Alive    25.897632 12.299465
##   Deceased  9.702063 52.100840
```

Above you can see that 682 males did not survive which is 52% and 161 which is 12 % male survive where 127 females which are 9 % did not survive and 339 females which are 25% survive.

```
titanic_new$pclass <- factor(titanic_new$pclass, levels = c("Upper Class","Middle Class","Lower Class"))

ggplot(titanic_new, aes(x = pclass, fill = survived)) + geom_bar(position = "dodge") + labs(y = "Count
```



If you look at the socio-economic class the majority group where there is a high death rate is in a lower class. It also shows that most of the passengers are in were in the lower class. Upper class passengers have a higher survival rate.

```
table(titanic_new$survived , titanic_new$pclass)
```

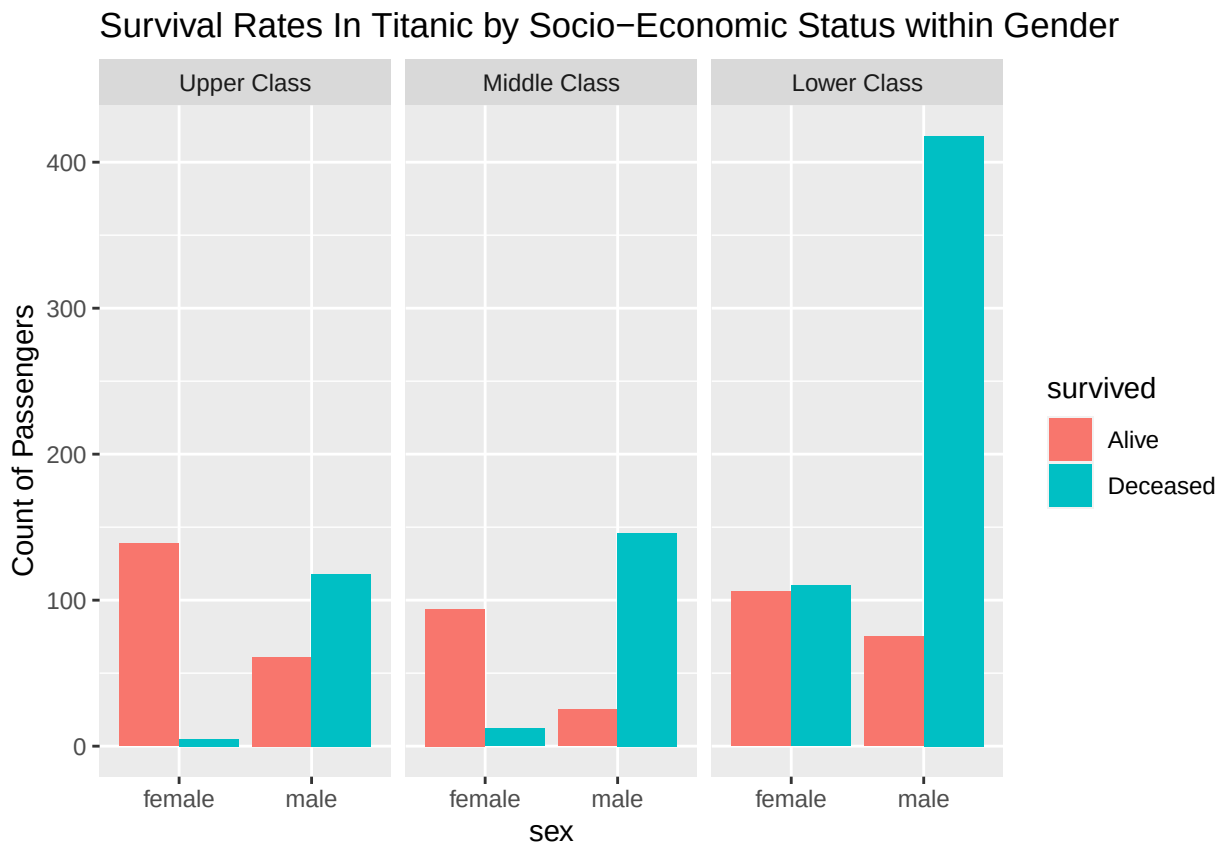
```
##
##            Upper Class Middle Class Lower Class
##   Alive            200          119         181
##   Deceased         123          158         528
```

```
prop.table(table(titanic_new$survived , titanic_new$pclass))*100
```

```
##
##            Upper Class Middle Class Lower Class
##   Alive      15.278839     9.090909   13.827349
##   Deceased    9.396486    12.070283   40.336134
```

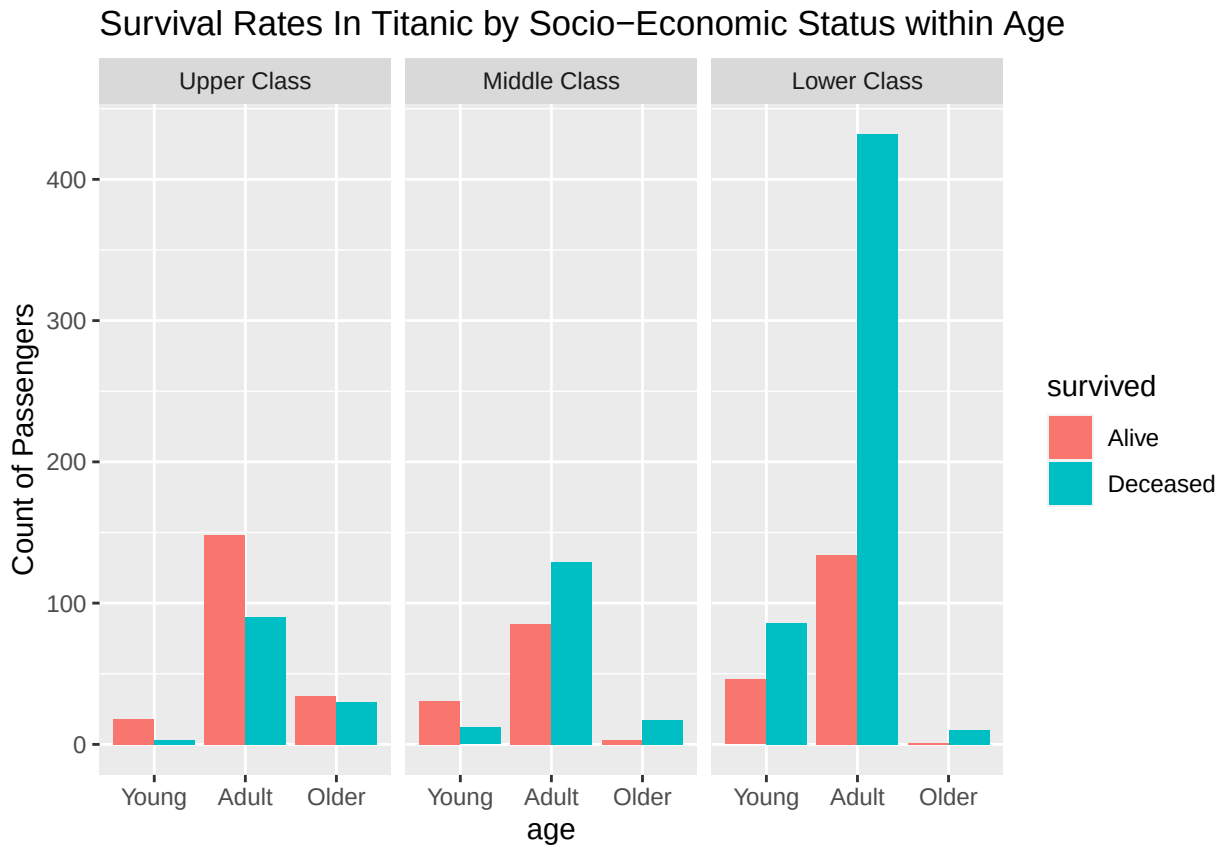
Above it shows that 123 passengers in upper class which is 9% did not survive and 200 which is 15% survive. In the second class, 158 passengers which are 12% did not survive and 119 which is 9% survive. In the third class where the majority of passengers were located 528 which is 40% did not survive and 181 which is 13% survived.

```
ggplot(titanic_new, aes(x=sex, fill = survived)) + facet_wrap(~pclass) + geom_bar(position = "dodge")
```



The above graph indicates that males that are in the lower class are the majority group who did not survive and females in upper class are the ones that survived the most.

```
titanic_new$age <- factor(titanic_new$age, levels = c("Young","Adult","Older"))

ggplot(titanic_new, aes(x=age, fill = survived)) + facet_wrap(~pclass) + geom_bar(position = "dodge")
```



It can be seen that Adults in all socio-economic class is the age group that did not survive. Adult age group also has the highest count of passengers.

```
table(titanic_new$survived , titanic_new$age)
```

```
##
##            Young Adult Older
##   Alive       95   367    38
##   Deceased   101   651    57
```

```
prop.table(table(titanic_new$survived , titanic_new$age))*100
```

```
##
##                Young     Adult     Older
##   Alive     7.257448 28.036669  2.902979
##   Deceased  7.715814 49.732620  4.354469
```

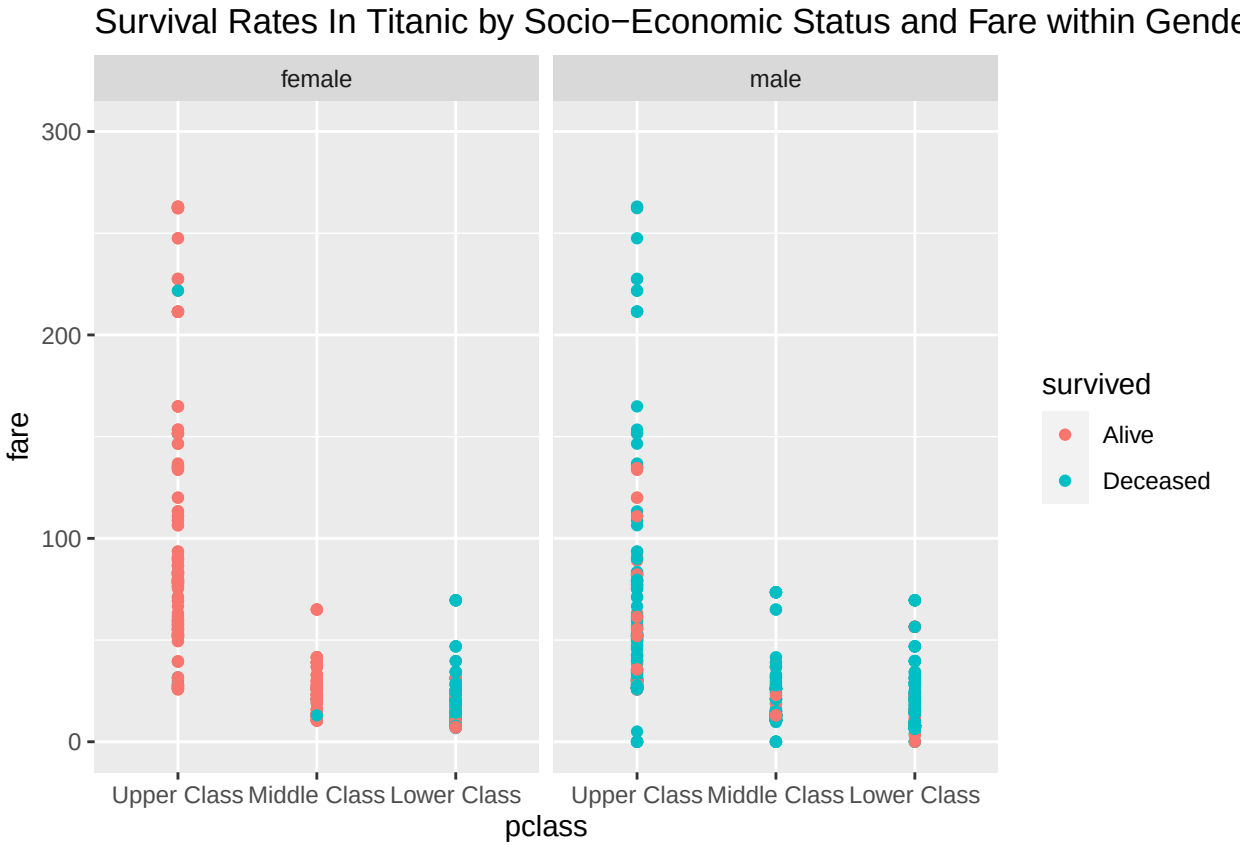
Above it shows that 101 passengers in Young age group which is 7% did not survive and 95 which is 7% survive. In the Adult age group, where the majority of passengers were located 651 passengers which are 49% did not survive and 367 which is 28% survive. In the Old age group 57 which is 4% did not survive and 38 which is 2.9% survived.

```
max(titanic_new$fare)
```

```
## [1] 512.3292
```

```
ggplot(titanic_new) + geom_point(mapping = aes(x=pclass, y=fare, color=survived)) + facet_wrap(~sex) + y
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



This graph looks at the price each passenger paid by looking at their socio-economic class and within different gender There was an outliner where is a passenger that paid 512 pounds I chose not to include that in the graph and set intervals between 0 to 300 pounds. Females who paid the highest amount for the ticket have the highest probability of surviving but females who paid the cheapest amount which was placed in the lower class have a lower survival rate. Males overall have the highest death rate does not matter how much they paid for their ticket.