Data Manipulation(Titanic)

Buhari et Fadlulah

3/30/2022

Data Manipulation Using “dplyr”

Description Of Data(Titanic):

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner ‘Titanic’, summarized according to economic status (class), sex, age and survival. A 4-dimensional array resulting from cross-tabulating 2201 observations on 4 variables. The variables and their levels are as follows:

No Name Levels 1 Class 1st, 2nd, 3rd, Crew 2 Sex Male, Female 3 Age Child, Adult 4 Survived No, Yes

Insights from the data using “dplyr”

```
library(dplyr)
library(tidyr)
library(knitr)
library(ggplot2)
library(tidyverse)
```

```
head(Titanic)
```

```
## , , Age = Child, Survived = No
## 
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
## 
## , , Age = Adult, Survived = No
## 
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
## 
## , , Age = Child, Survived = Yes
## 
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
## 
## , , Age = Adult, Survived = Yes
## 
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

```
dim(Titanic)
```

```
## [1] 4 2 2 2
```

```
summary(Titanic)
```

```
## Number of cases in table: 2201 
## Number of factors: 4 
## Test for independence of all factors:
##  Chisq = 1637.4, df = 25, p-value = 0
##  Chi-squared approximation may be incorrect
```

```
df = data.frame(Titanic)
head(df)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

selection from Sex to Freq

```
df1 = select(df, Sex:Freq)
head(df1)
```

```
##      Sex   Age Survived Freq
## 1   Male Child       No    0
## 2   Male Child       No    0
## 3   Male Child       No   35
## 4   Male Child       No    0
## 5 Female Child       No    0
## 6 Female Child       No    0
```

filtering out the Male Sex

```
df2 = filter(df1, Sex == "Male" )
head(df2)
```

```
##    Sex   Age Survived Freq
## 1 Male Child       No    0
## 2 Male Child       No    0
## 3 Male Child       No   35
## 4 Male Child       No    0
## 5 Male Adult       No  118
## 6 Male Adult       No  154
```

```
df3 = filter(df1, Sex == "Male" & Age == "Adult") ## filtering out the male adult
head(df3)
```

```
##    Sex   Age Survived Freq
## 1 Male Adult       No  118
## 2 Male Adult       No  154
## 3 Male Adult       No  387
## 4 Male Adult       No  670
## 5 Male Adult      Yes   57
## 6 Male Adult      Yes   14
```

```
df4 = mutate(df1, dd = Freq * 2 )
head(df4)
```

```
##      Sex   Age Survived Freq dd
## 1   Male Child       No    0  0
## 2   Male Child       No    0  0
## 3   Male Child       No   35 70
## 4   Male Child       No    0  0
## 5 Female Child       No    0  0
## 6 Female Child       No    0  0
```

```
df5 = mutate(df3, cd = Freq^2)
head(df5)
```

```
##    Sex   Age Survived Freq     cd
## 1 Male Adult       No  118  13924
## 2 Male Adult       No  154  23716
## 3 Male Adult       No  387 149769
## 4 Male Adult       No  670 448900
## 5 Male Adult      Yes   57   3249
## 6 Male Adult      Yes   14    196
```

```
df6 = filter(df3, Survived == "Yes" )
head(df6)
```

```
##    Sex   Age Survived Freq
## 1 Male Adult      Yes   57
## 2 Male Adult      Yes   14
## 3 Male Adult      Yes   75
## 4 Male Adult      Yes  192
```

```
Mean = summarize(df3, Mean = mean(Freq))
Mean
```

```
##      Mean
## 1 208.375
```

```
class = filter(df, !Class == "1st") ###class excluding 1st in the df
head(class)
```

```
##   Class    Sex   Age Survived Freq
## 1   2nd   Male Child       No    0
## 2   3rd   Male Child       No   35
## 3  Crew   Male Child       No    0
## 4   2nd Female Child       No    0
## 5   3rd Female Child       No   17
## 6  Crew Female Child       No    0
```

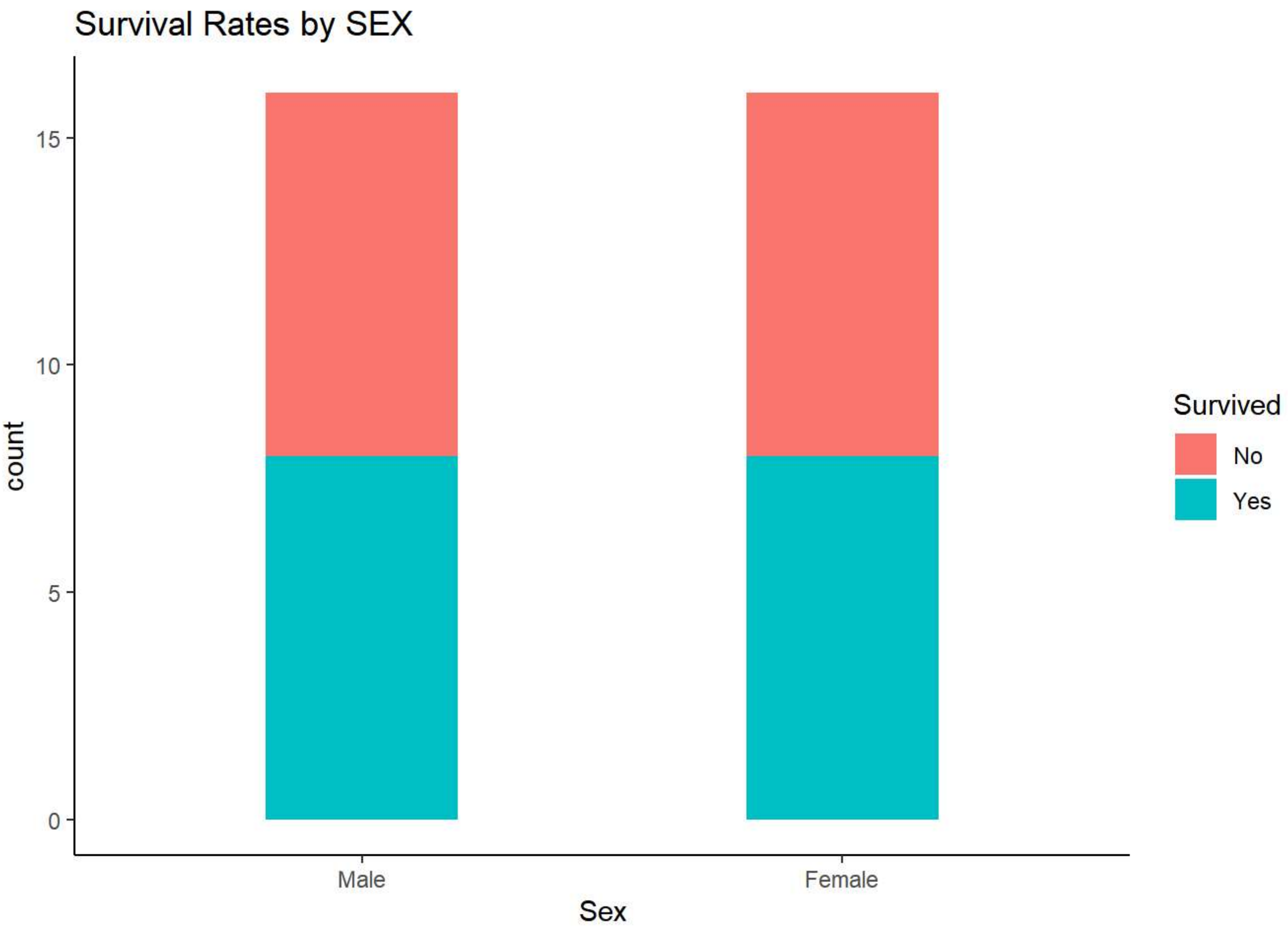
```
child = filter(df, Age == "Child")
head(child)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

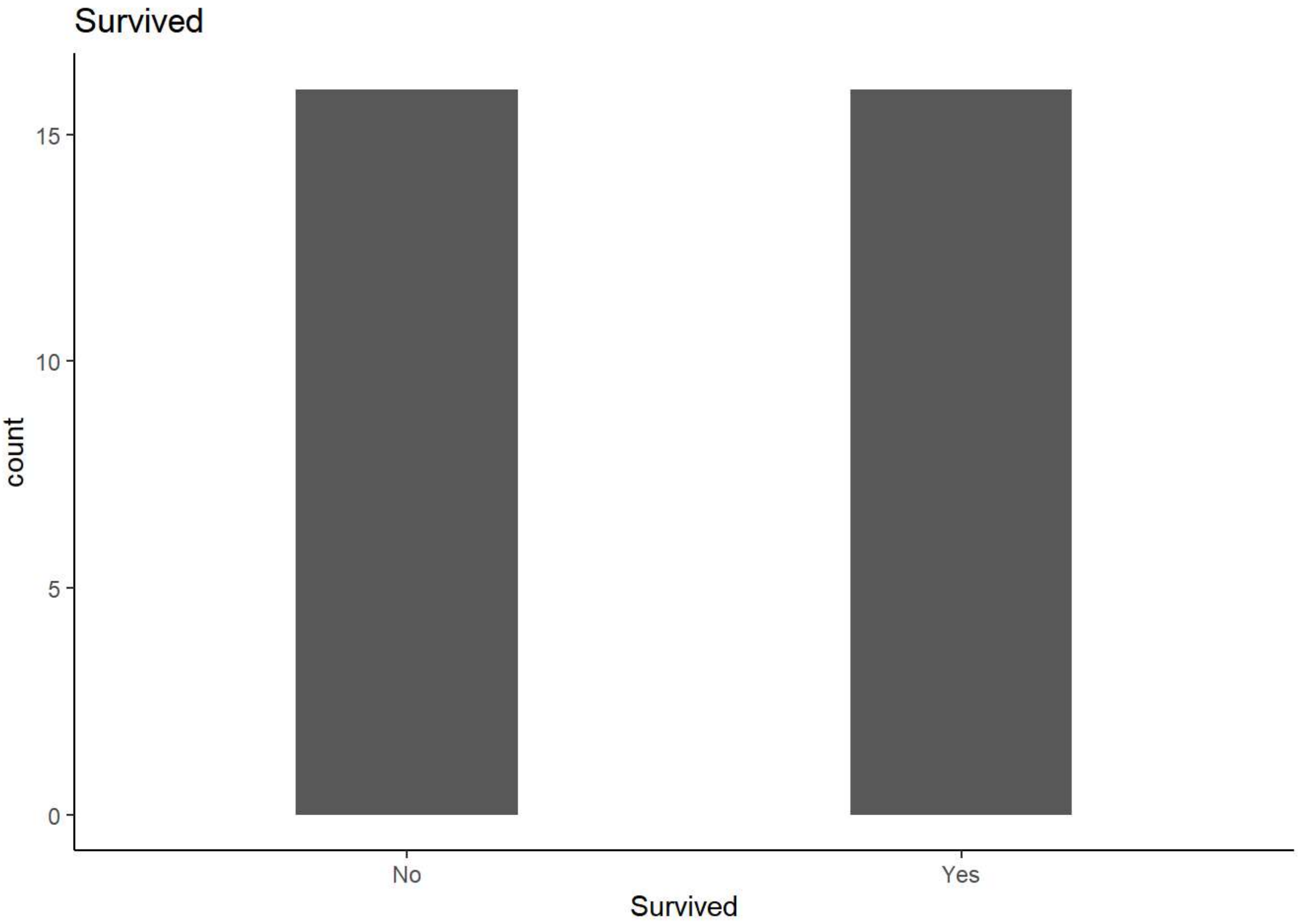
```
Freq_35plus = filter(df, Freq > 35)
head(Freq_35plus)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Adult       No  118
## 2   2nd   Male Adult       No  154
## 3   3rd   Male Adult       No  387
## 4  Crew   Male Adult       No  670
## 5   3rd Female Adult       No   89
## 6   1st   Male Adult      Yes   57
```

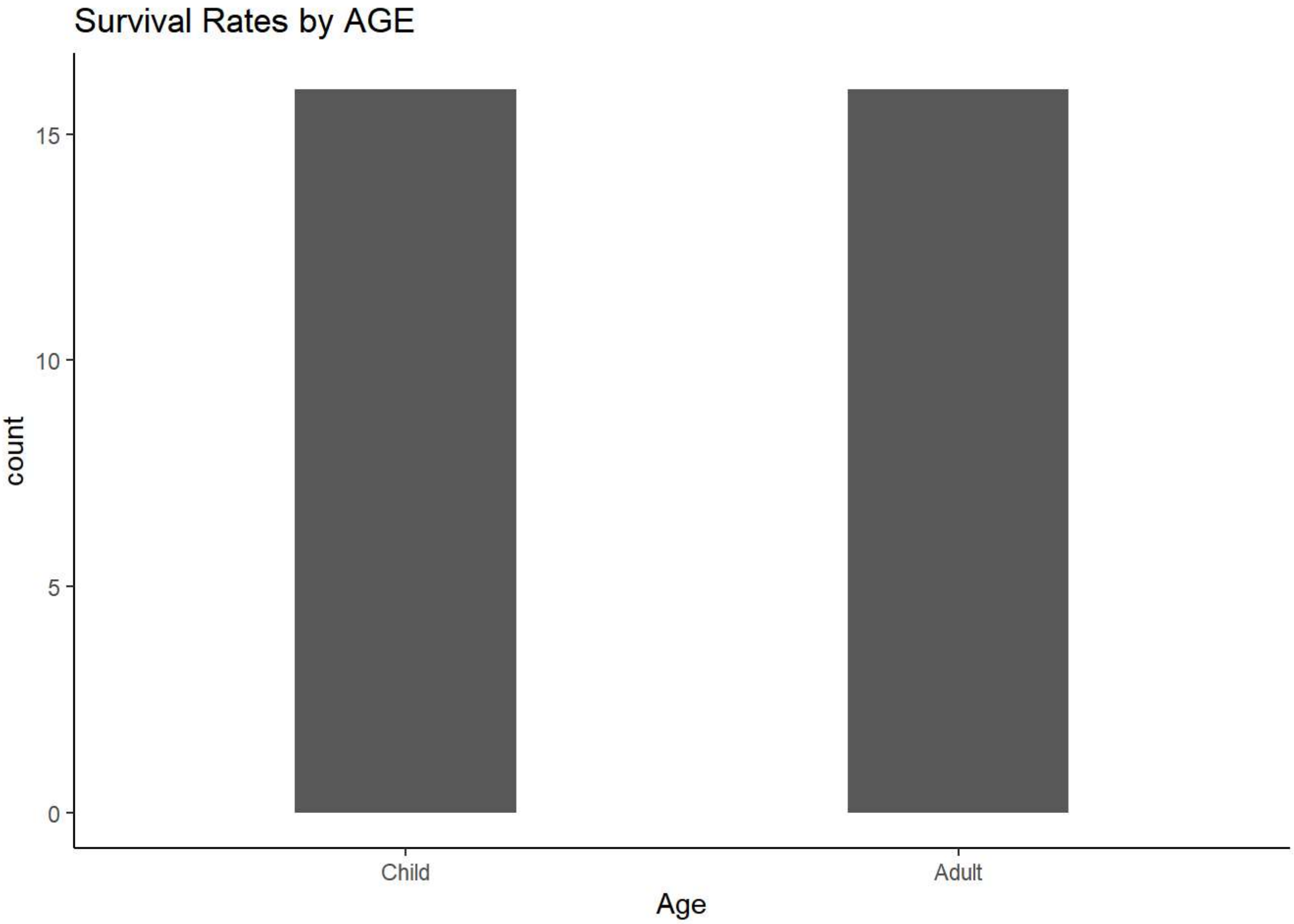
```
df%>%
  ggplot(aes(x = Sex, fill = Survived)) +
  geom_bar(width = 0.4) +
  theme_classic() +
  labs(title = "Survival Rates by SEX")
```



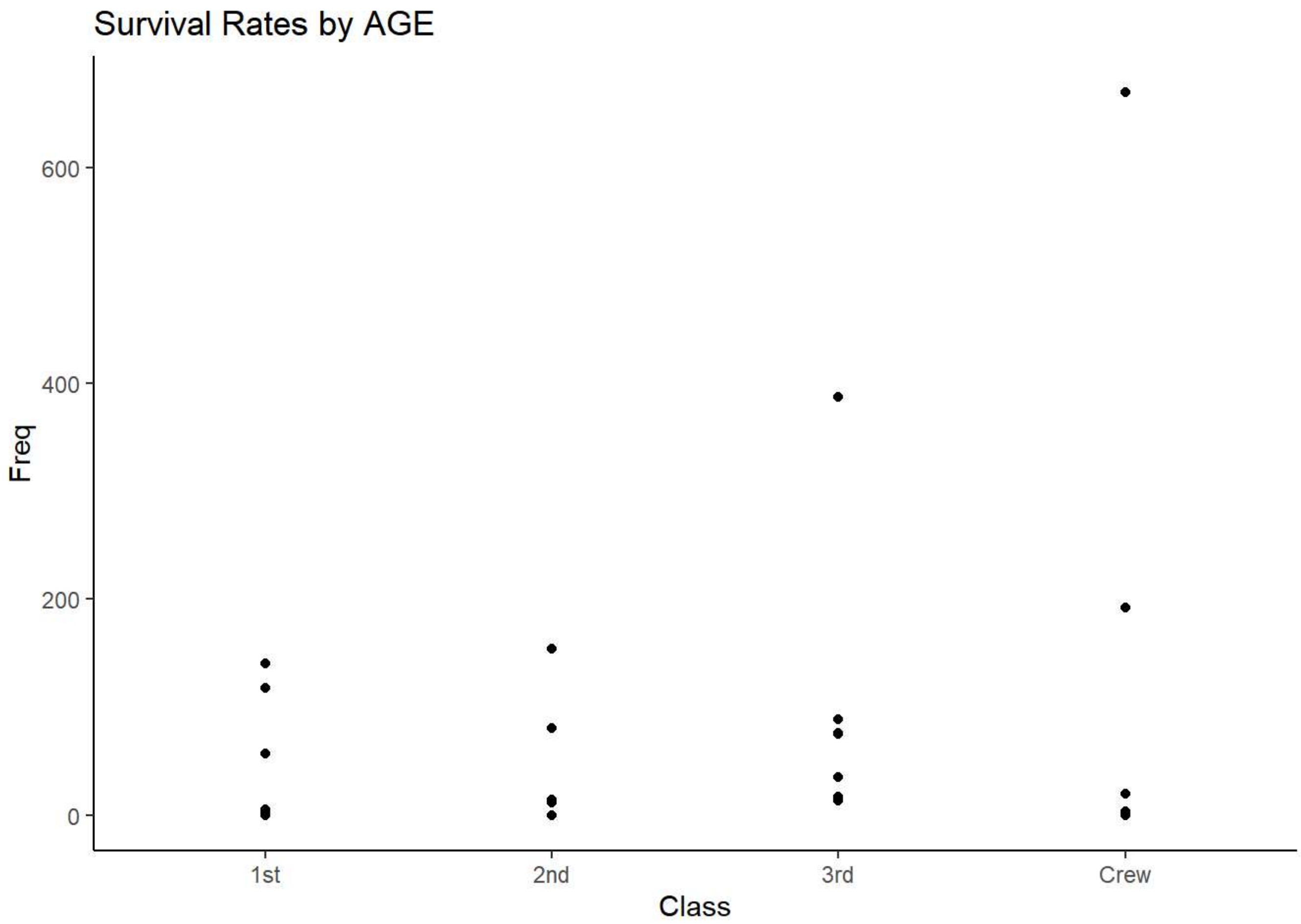
```
df%>%
ggplot(aes(x = Survived)) +
geom_bar(width = 0.4) +
theme_classic() + labs(title = "Survived")
```



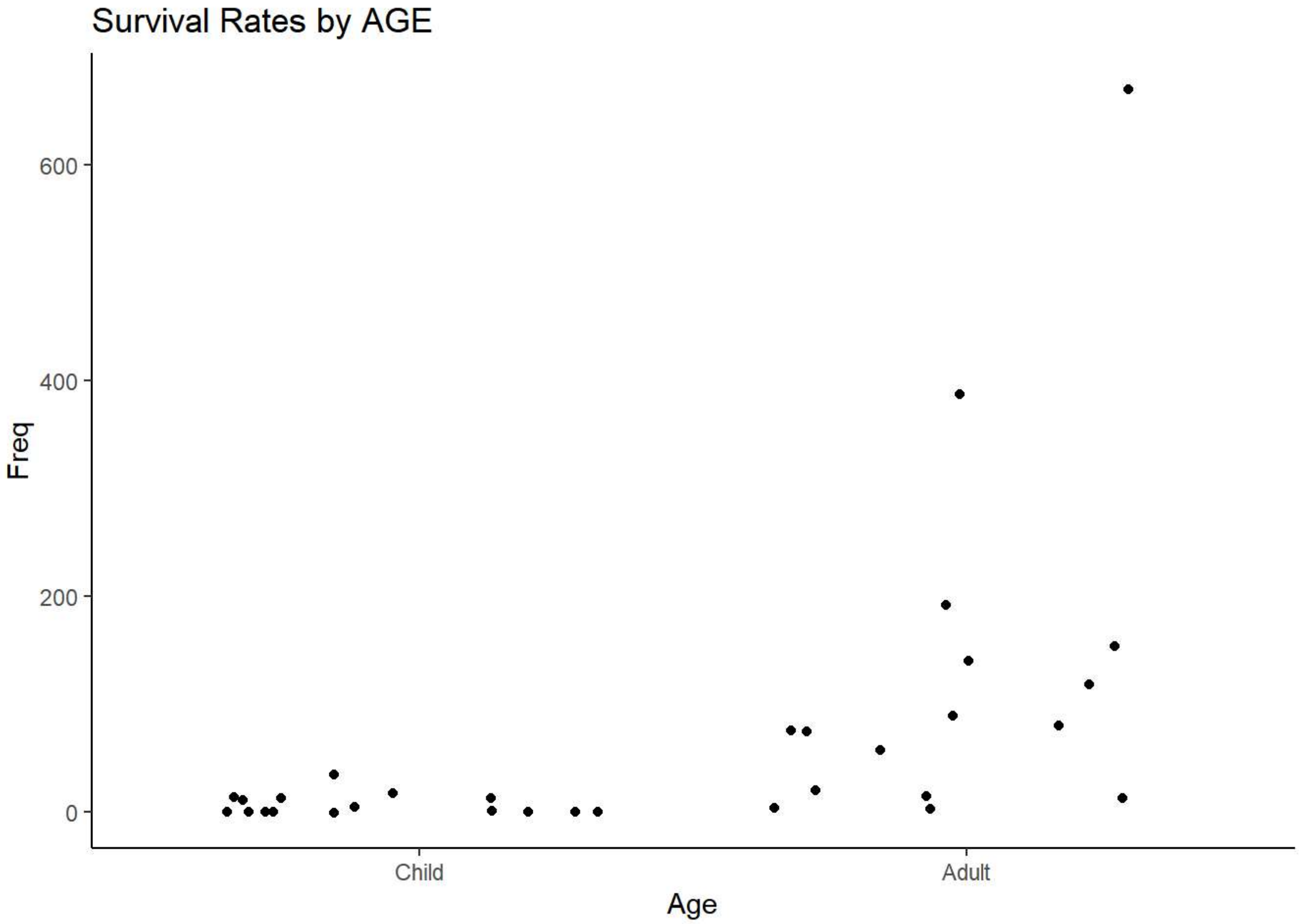
```
df%>%
  ggplot(aes(x = Age)) +
  geom_bar(width = 0.4) +
  theme_classic() +
  labs(title = "Survival Rates by AGE")
```



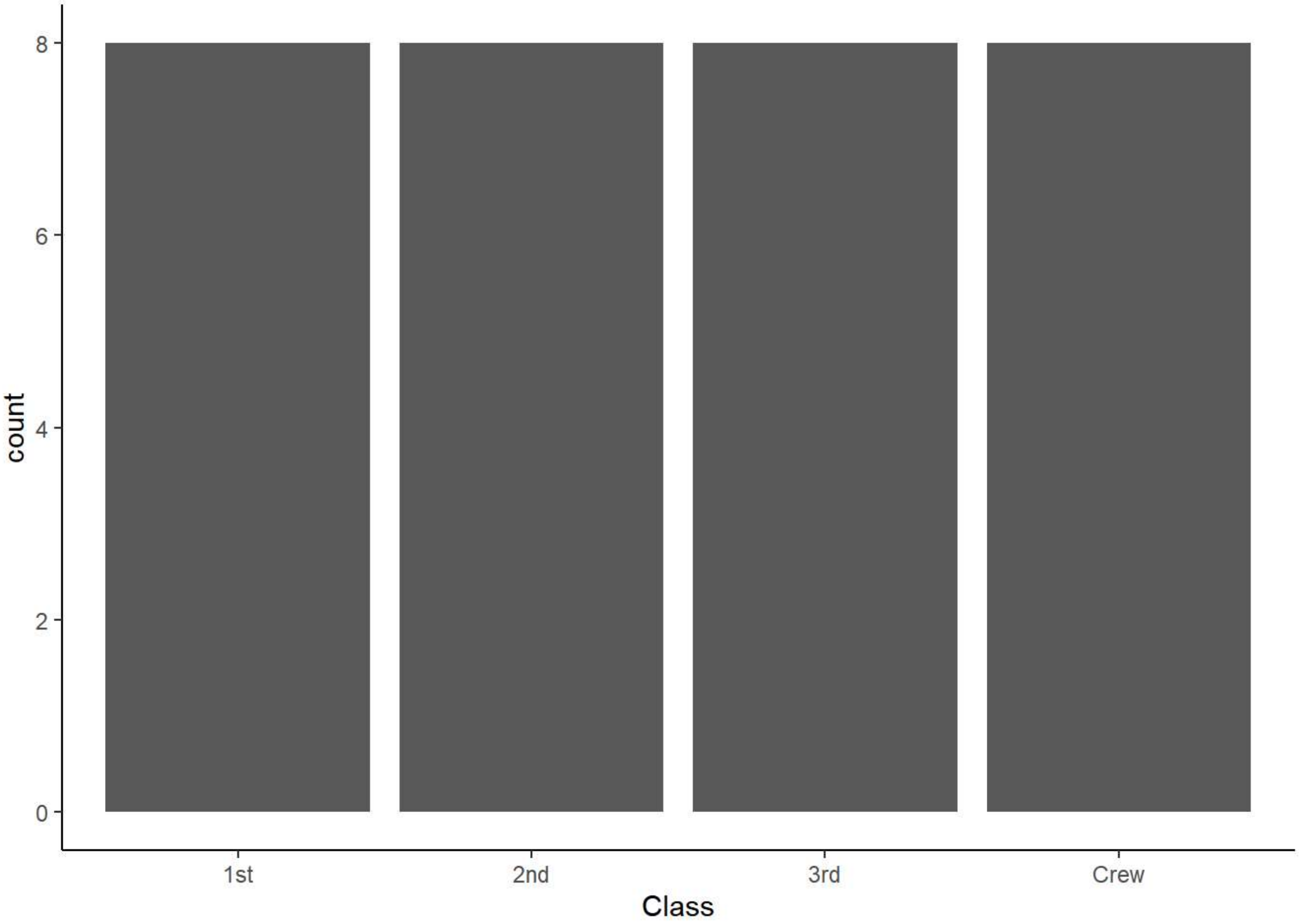
```
df%>%
ggplot(aes(y = Freq, x= Class)) +
geom_point() +
theme_classic() +
labs(title = "Survival Rates by AGE")
```



```
df%>%
ggplot(aes(x = Age, y=Freq)) +
geom_jitter() +
theme_classic() +
labs(title = "Survival Rates by AGE")
```



```
df%>%
  group_by(Class)%>%
  ggplot(aes(x = Class))+
  geom_bar()+
  theme_classic()
```


8
6
4
2
0
count
1st
2nd
3rd
Crew
Class

```
```