ggplot

Chandana

October 11, 2018

```
library(readxl)
library(ggplot2)

orig_list <- data.frame(readxl::read_excel("titanic3_assignment.xls"))
plist <- orig_list

dim(plist)
```

```
## [1] 1309    7
```

```
summary(plist)
```

```
##      pclass          survived         name                sex
##  Min.   :1.000   Min.   :0.000   Length:1309        Length:1309
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :2.295   Mean   :0.382
##  3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :1.000
##
##       age               fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
##  1st Qu.:21.0000   1st Qu.:  7.896   Class :character
##  Median :28.0000   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:39.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329
##  NA's   :263       NA's   :1
```

```
#1
plist$survived <- as.logical(plist$survived)
summary(plist)
```

```
##      pclass       survived           name                sex
##  Min.   :1.000   Mode :logical   Length:1309        Length:1309
##  1st Qu.:2.000   FALSE:809       Class :character   Class :character
##  Median :3.000   TRUE :500       Mode  :character   Mode  :character
##  Mean   :2.295
##  3rd Qu.:3.000
##  Max.   :3.000
##
##       age               fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
```

```
##  1st Qu.:21.0000   1st Qu.:  7.896   Class :character
##  Median :28.0000   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:39.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329
##  NA's   :263       NA's   :1
```

```
#2
plist$pclass <- ifelse(plist$pclass==1, "First", ifelse(plist$pclass==2,
"Second", "Third"))
summary(plist)
```

```
##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##
##       age              fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
##  1st Qu.:21.0000   1st Qu.:  7.896   Class :character
##  Median :28.0000   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:39.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329
##  NA's   :263       NA's   :1
```

```
unique(plist$pclass)
```

```
## [1] "First"  "Second" "Third"
```

```
#3
plist[!complete.cases(plist$age),]["age"] <- mean(plist$age, na.rm = T)
summary(plist)
```

```
##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##
##       age              fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
##  1st Qu.:22.0000   1st Qu.:  7.896   Class :character
##  Median :29.8811   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:35.0000   3rd Qu.: 31.275
```

```
##  Max.   :80.0000   Max.   :512.329
##                    NA's   :1

#4
plist[!complete.cases(plist$fare),]["fare"] <- mean(plist$fare, na.rm = T)
summary(plist)

##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##       age              fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
##  1st Qu.:22.0000   1st Qu.:  7.896   Class :character
##  Median :29.8811   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:35.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329

#5
set.seed(99)
plist[is.na(plist$embarked), c("embarked")] <- sample(c("S", "C", "Q"),
sum(is.na(plist$embarked)), replace = T)
summary(plist)

##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##       age              fare           embarked
##  Min.   : 0.1667   Min.   :  0.000   Length:1309
##  1st Qu.:22.0000   1st Qu.:  7.896   Class :character
##  Median :29.8811   Median : 14.454   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:35.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329

#6
plist$age_cohort <- ifelse(plist$age<16, "Child", ifelse(plist$age>=16 &
plist$age<60, "Adults", "Elderly"))
summary(plist)

##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
```

```
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##       age              fare           embarked          age_cohort
##  Min.   : 0.1667   Min.   :  0.000   Length:1309        Length:1309
##  1st Qu.:22.0000   1st Qu.:  7.896   Class :character   Class :character
##  Median :29.8811   Median : 14.454   Mode  :character   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:35.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329

#7
plist$embarked <- ifelse(plist$embarked=="Q", "Cobh",
ifelse(plist$embarked=="C", "Cherbourg", "Southampton"))
summary(plist)

##     pclass           survived           name               sex
##  Length:1309        Mode :logical   Length:1309        Length:1309
##  Class :character   FALSE:809       Class :character   Class :character
##  Mode  :character   TRUE :500       Mode  :character   Mode  :character
##
##
##
##       age              fare           embarked          age_cohort
##  Min.   : 0.1667   Min.   :  0.000   Length:1309        Length:1309
##  1st Qu.:22.0000   1st Qu.:  7.896   Class :character   Class :character
##  Median :29.8811   Median : 14.454   Mode  :character   Mode  :character
##  Mean   :29.8811   Mean   : 33.295
##  3rd Qu.:35.0000   3rd Qu.: 31.275
##  Max.   :80.0000   Max.   :512.329

#8
head(plist)

##   pclass survived                                             name    sex
## 1  First     TRUE                    Allen, Miss. Elisabeth Walton female
## 2  First     TRUE                   Allison, Master. Hudson Trevor   male
## 3  First    FALSE                     Allison, Miss. Helen Loraine female
## 4  First    FALSE              Allison, Mr. Hudson Joshua Creighton   male
## 5  First    FALSE Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6  First     TRUE                                Anderson, Mr. Harry   male
##       age     fare    embarked age_cohort
## 1 29.0000 211.3375 Southampton     Adults
## 2  0.9167 151.5500 Southampton      Child
## 3  2.0000 151.5500 Southampton      Child
## 4 30.0000 151.5500 Southampton     Adults
## 5 25.0000 151.5500 Southampton     Adults
## 6 48.0000  26.5500 Southampton     Adults

dim(plist)
```

```
## [1] 1309    8
```

```
table(plist$survived)
```

```
##
## FALSE  TRUE
##   809   500
```

```
table(plist$survived, plist$age_cohort)
```

```
##
##         Adults Child Elderly
##   FALSE    732    49      28
##   TRUE     422    66      12
```

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

```
##
##         female male
##   FALSE    127  682
##   TRUE     339  161
```

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

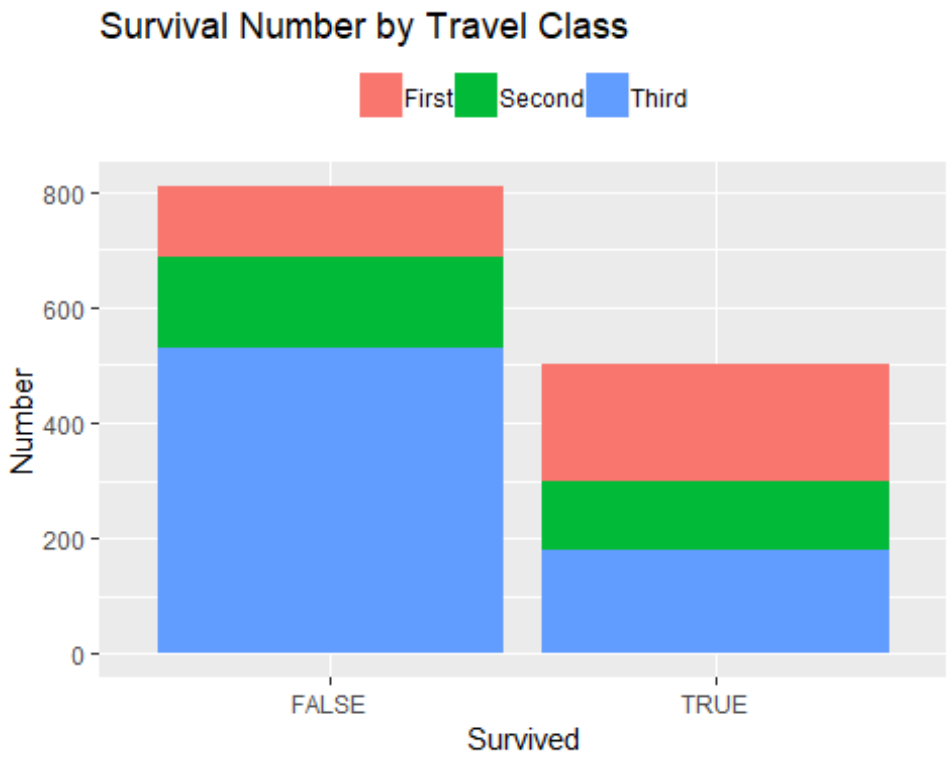
```
##
##         First Second Third
##   FALSE   123    158   528
##   TRUE    200    119   181
```

```
table(plist$survived, plist$embarked)
```

```
##
##         Cherbourg Cobh Southampton
##   FALSE       120   79         610
##   TRUE        151   44         305
```

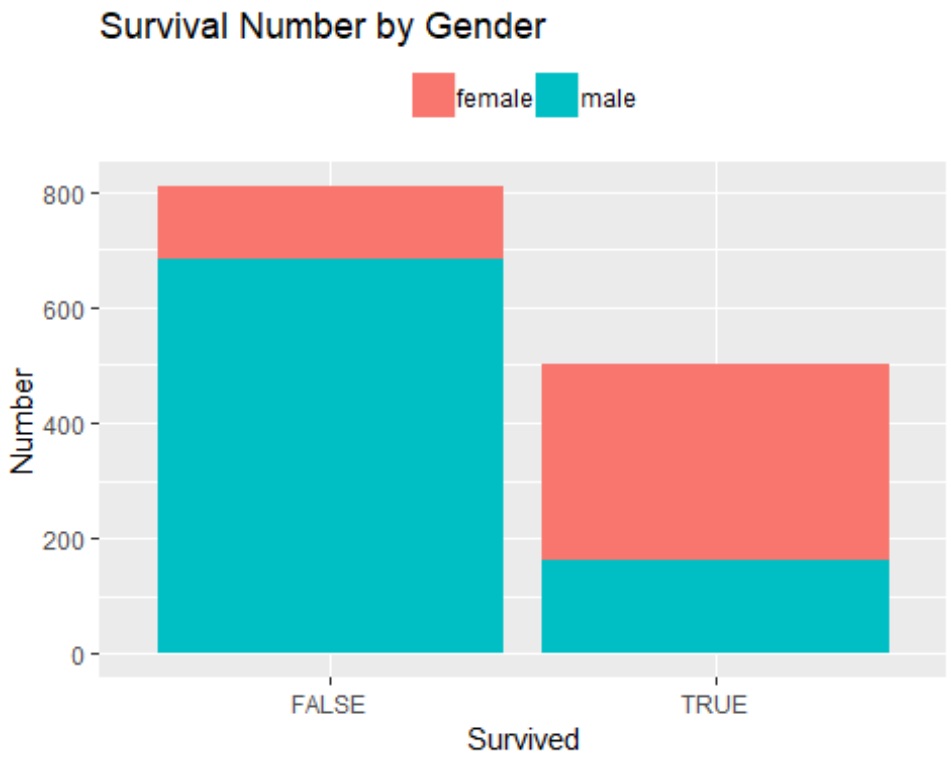
plot 1

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = pclass)) +
  ggtitle("Survival Number by Travel Class") +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



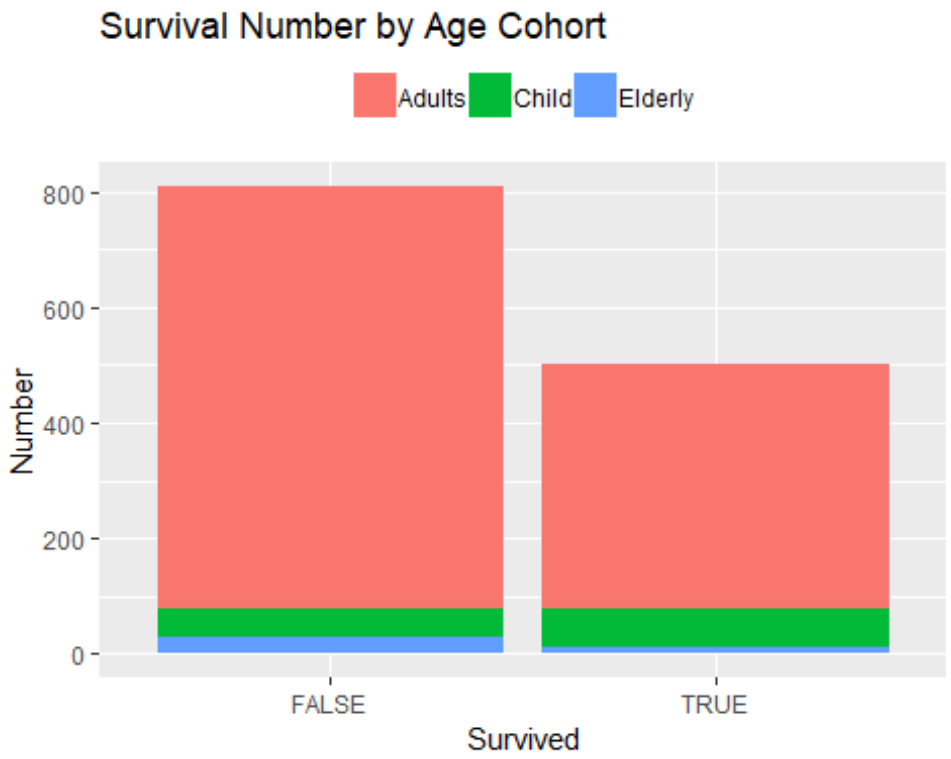
plot2

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = sex)) +
  ggtitle("Survival Number by Gender") +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



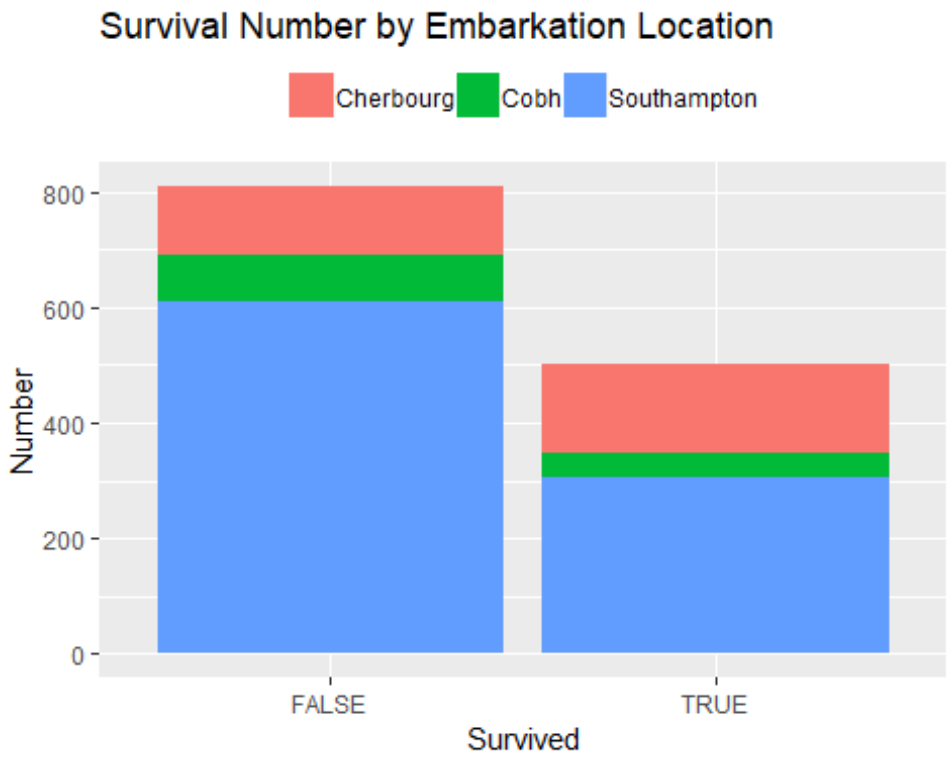
plot 3

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = age_cohort)) +
  ggtitle("Survival Number by Age Cohort") +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



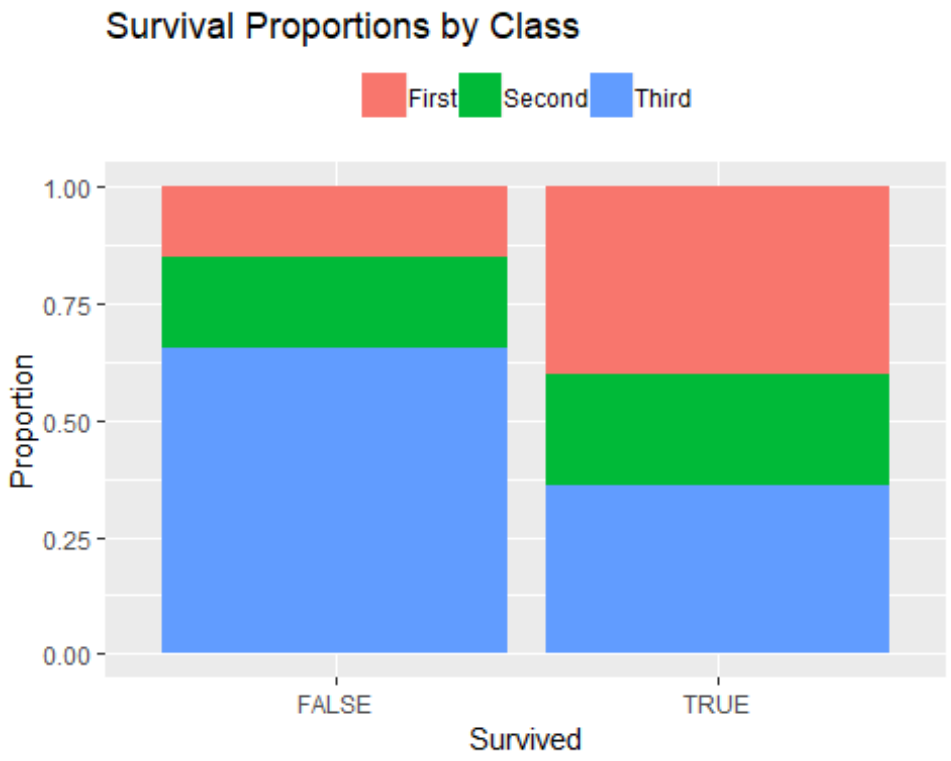
plot 4

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = embarked)) +
  ggtitle("Survival Number by Embarkation Location") +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



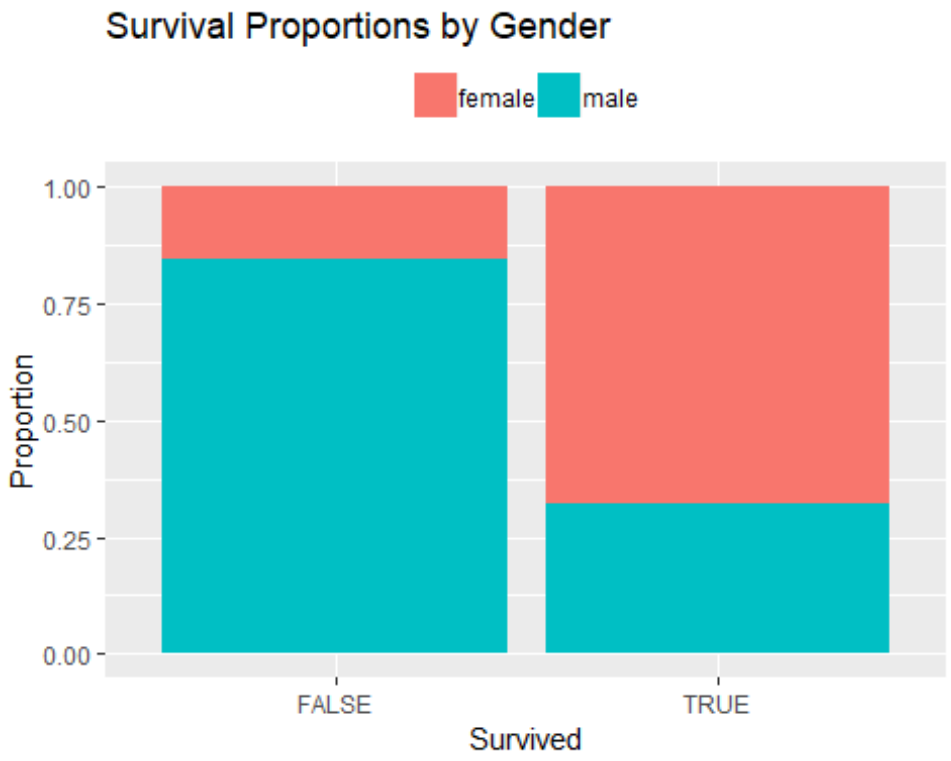
plot 5

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = pclass), position = "fill") +
  ggtitle("Survival Proportions by Class") +
  ylab("Proportion") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



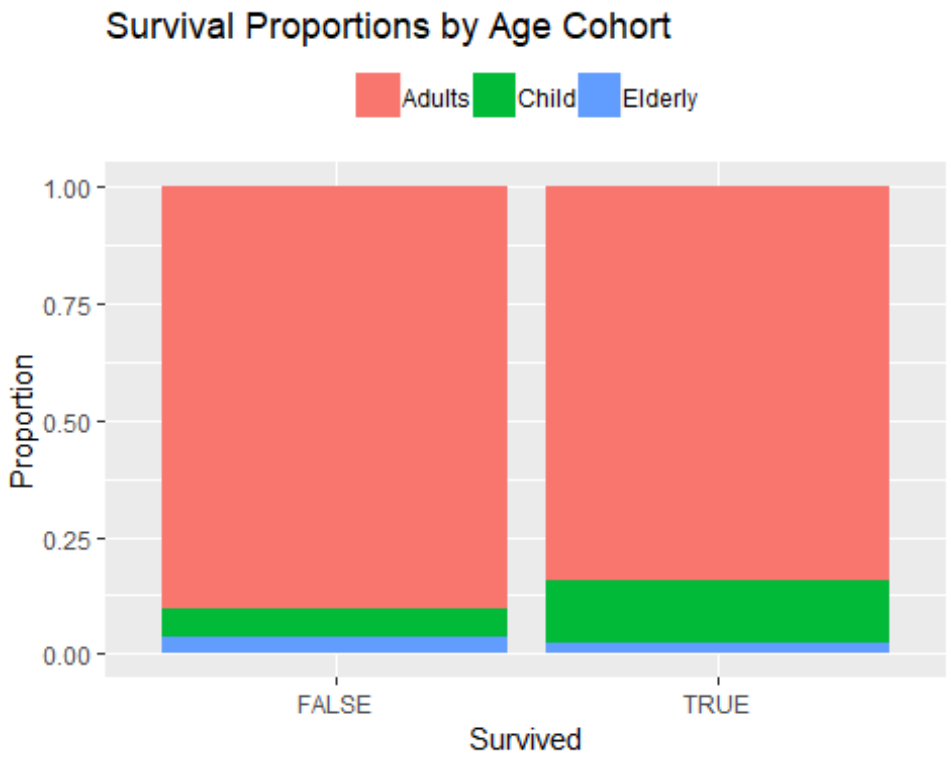
plot 6

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = sex), position = "fill") +
  ggtitle("Survival Proportions by Gender") +
  ylab("Proportion") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



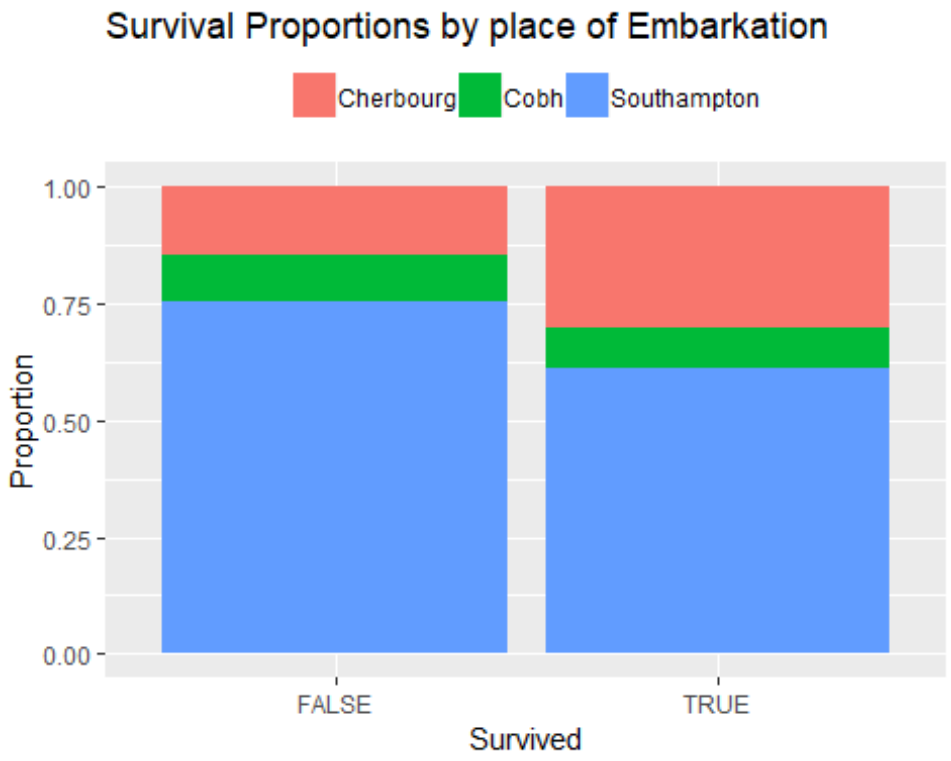
plot 7

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = age_cohort), position = "fill") +
  ggtitle("Survival Proportions by Age Cohort") +
  ylab("Proportion") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



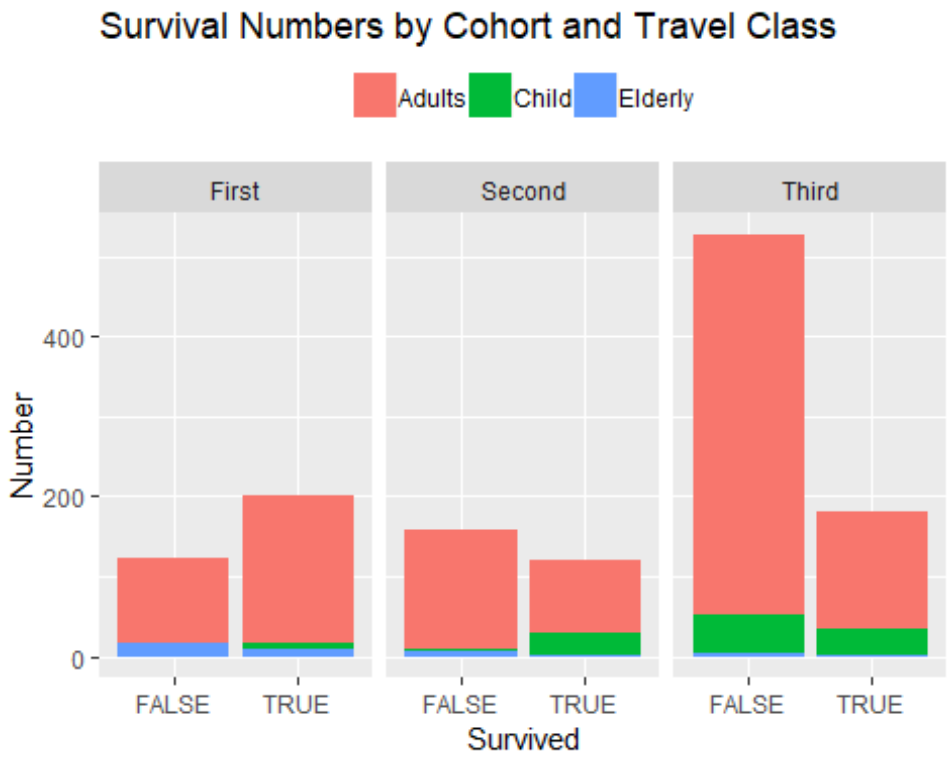
plot 8

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = embarked), position = "fill") +
  ggtitle("Survival Proportions by place of Embarkation") +
  ylab("Proportion") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



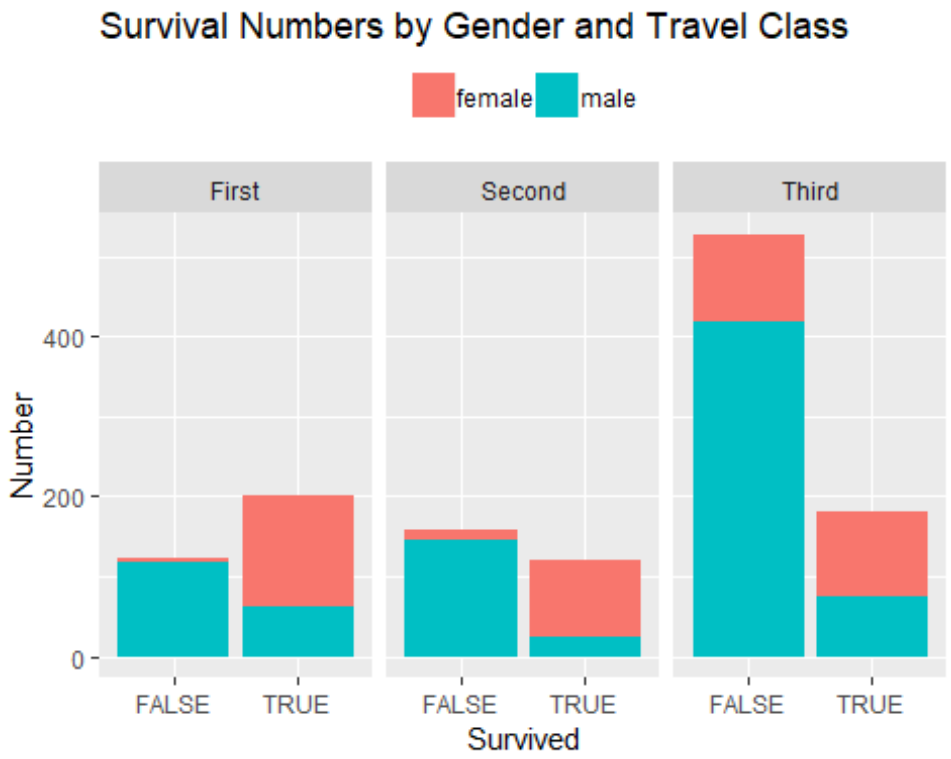
plot 9

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = age_cohort)) +
  ggtitle("Survival Numbers by Cohort and Travel Class") +
  facet_grid(~ pclass) +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



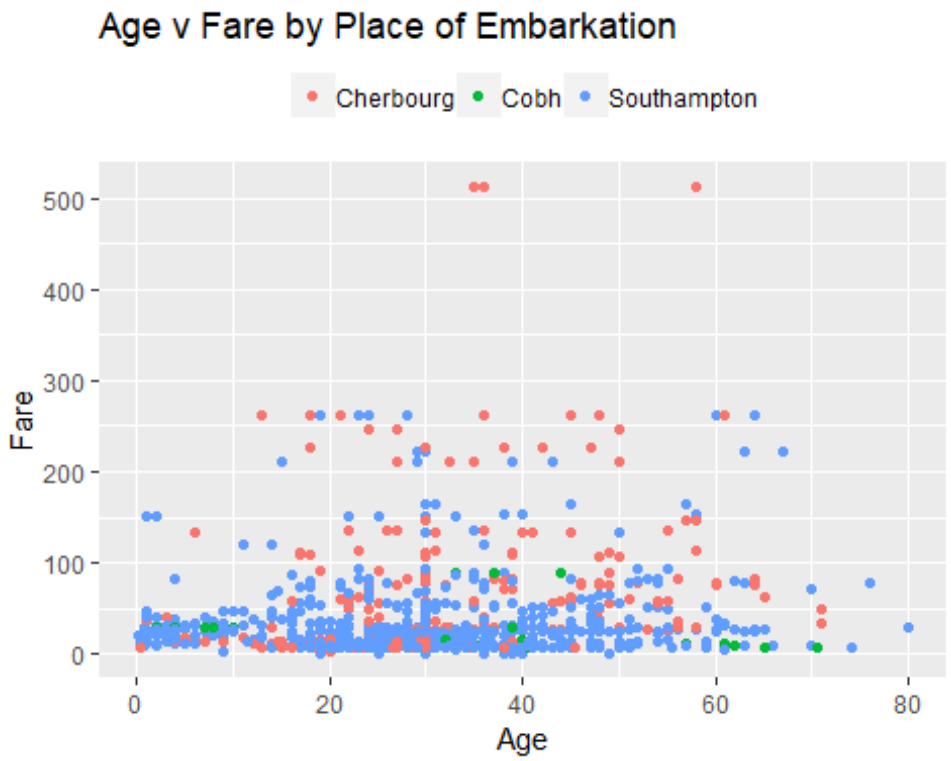
plot 10

```
ggplot(plist, aes(x = survived)) +
  geom_bar(aes(fill = sex)) +
  ggtitle("Survival Numbers by Gender and Travel Class") +
  facet_grid(~ pclass) +
  ylab("Number") +
  xlab("Survived") +
  theme(legend.position = "top", legend.title = element_blank())
```



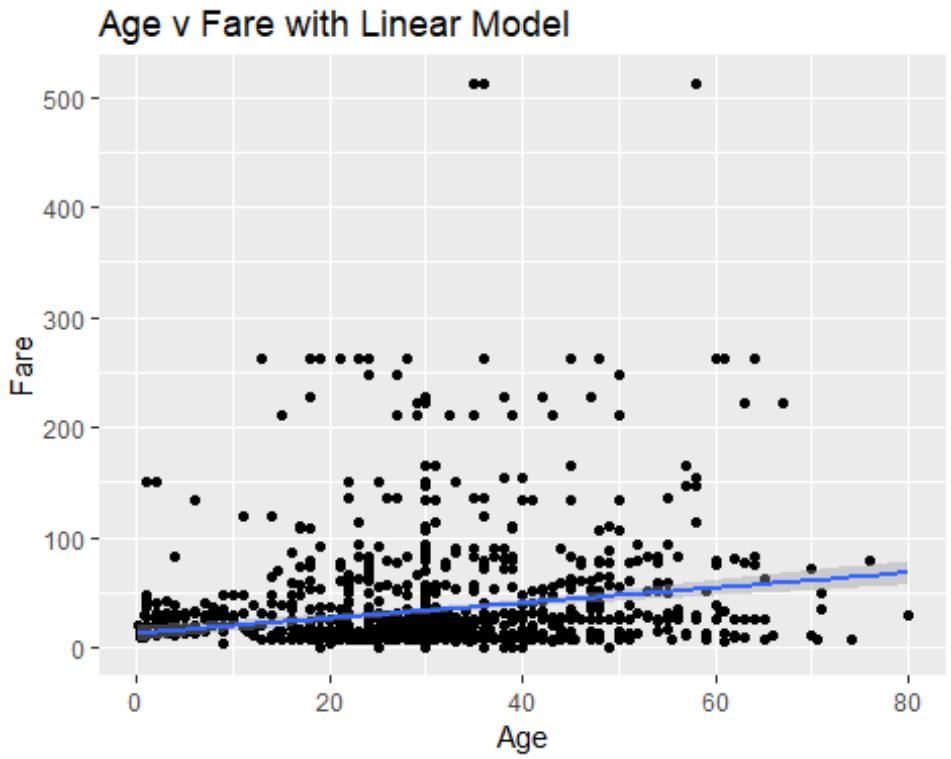
plot 11

```
ggplot(plist, aes(x = age, y = fare)) +
  geom_point(aes(color = embarked)) +
  ggtitle("Age v Fare by Place of Embarkation") +
  ylab("Fare") +
  xlab("Age") +
  theme(legend.position = "top", legend.title = element_blank())
```



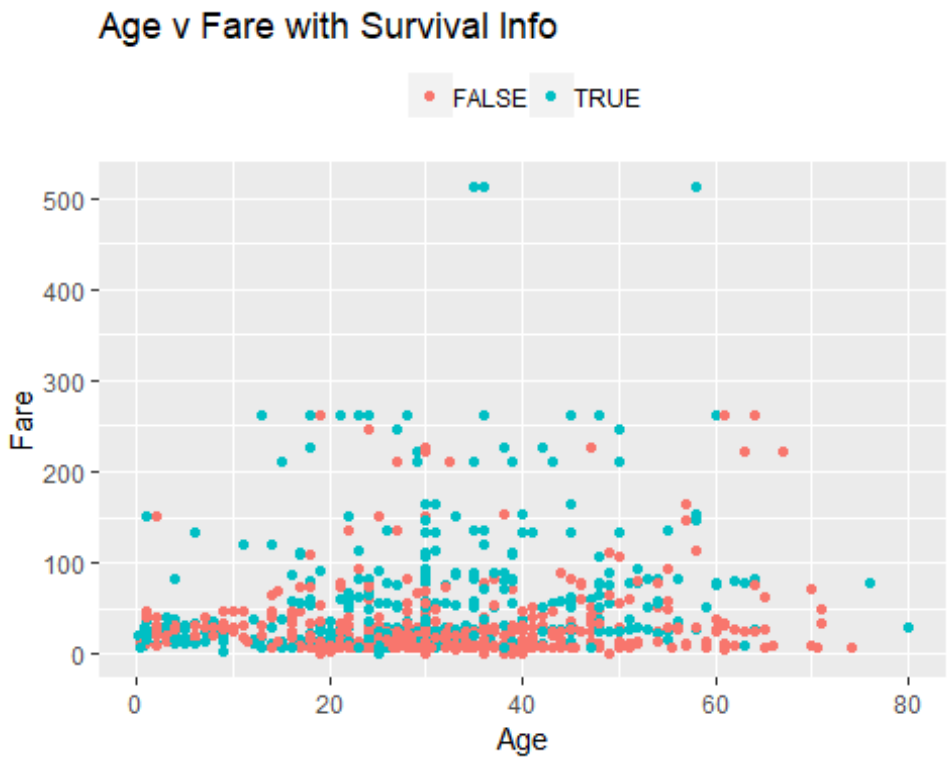
plot 12

```
ggplot(plist, aes(x = age, y = fare)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ggtitle("Age v Fare with Linear Model") +
  ylab("Fare") +
  xlab("Age")
```



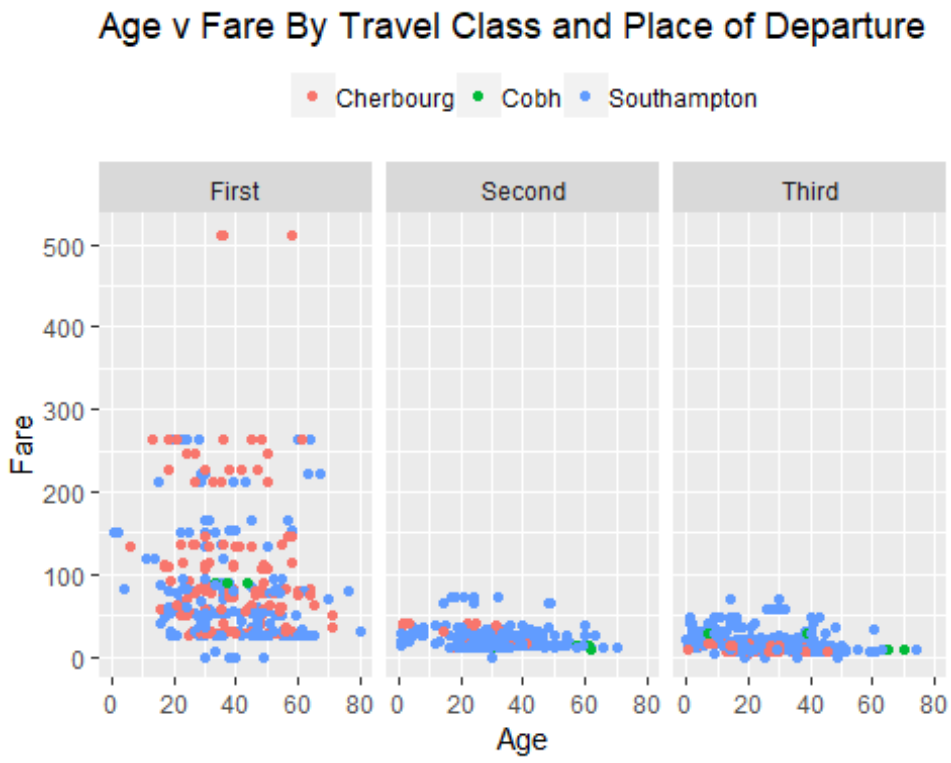
plot 13

```
ggplot(plist, aes(x = age, y = fare)) +
  geom_point(aes(color = survived)) +
  ggtitle("Age v Fare with Survival Info") +
  ylab("Fare") +
  xlab("Age") +
  theme(legend.position = "top", legend.title = element_blank())
```



plot 11

```
ggplot(plist, aes(x = age, y = fare)) +
  geom_point(aes(color = embarked)) +
  ggtitle("Age v Fare By Travel Class and Place of Departure") +
  facet_grid(~ pclass) +
  ylab("Fare") +
  xlab("Age") +
  theme(legend.position = "top", legend.title = element_blank())
```


Age v Fare By Travel Class and Place of Departure
Cherbourg
Cobh
Southampton
First
Second
Third
Fare
500
400
300
200
100
0
0
20
40
60
80
Age