# Lab 7 part 2 - a solution

November 2020

This worksheet is to practice week 7. Below is a solution.

```
library(ggplot2)
```

1. Load the titanic data set (smaller set)

```
titanic<-read.csv("data/titanic-analysis.csv")
```

Explore the data

```
summary(titanic)
```

```
##   PassengerId       Survived          Sex                 Age       
##  Min.   :  1.0   Min.   :0.0000   Length:891         Min.   : 0.42  
##  1st Qu.:223.5   1st Qu.:0.0000   Class :character   1st Qu.:20.12  
##  Median :446.0   Median :0.0000   Mode  :character   Median :28.00  
##  Mean   :446.0   Mean   :0.3838                      Mean   :29.70  
##  3rd Qu.:668.5   3rd Qu.:1.0000                      3rd Qu.:38.00  
##  Max.   :891.0   Max.   :1.0000                      Max.   :80.00  
##                                                      NA's   :177    
##       Fare       
##  Min.   :  0.00  
##  1st Qu.:  7.91  
##  Median : 14.45  
##  Mean   : 32.20  
##  3rd Qu.: 31.00  
##  Max.   :512.33  
## 
```

The dependent variable is: survived

Check that it is correctly read in, looking at the output of the summary function we can see it has not been correctly interpreted by R. (It would also be possible to check using str(titanic)) In order to fix this it can be changed into a factor:

```
titanic$Survived<-as.factor(titanic$Survived)
```

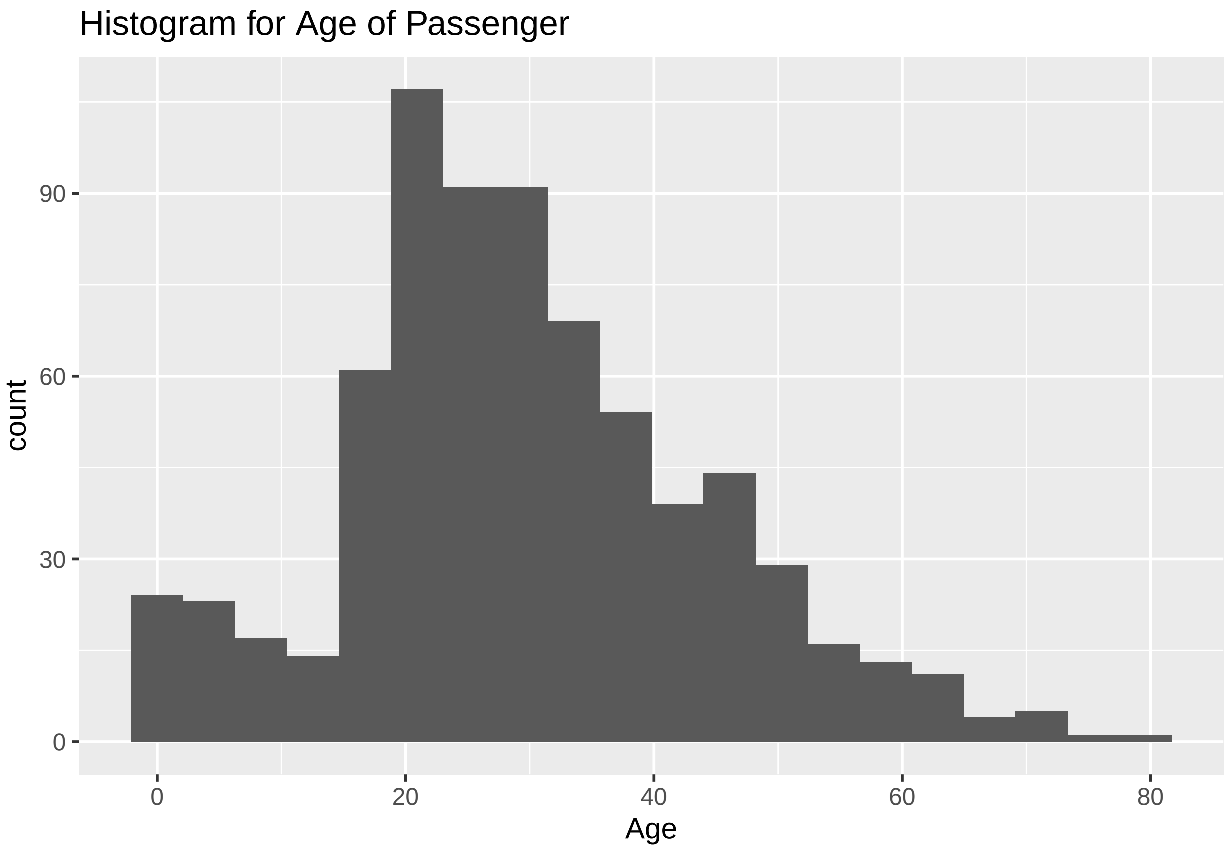
Note that there are 177 rows where the Age is not known.

It is also valuable to check the data visually

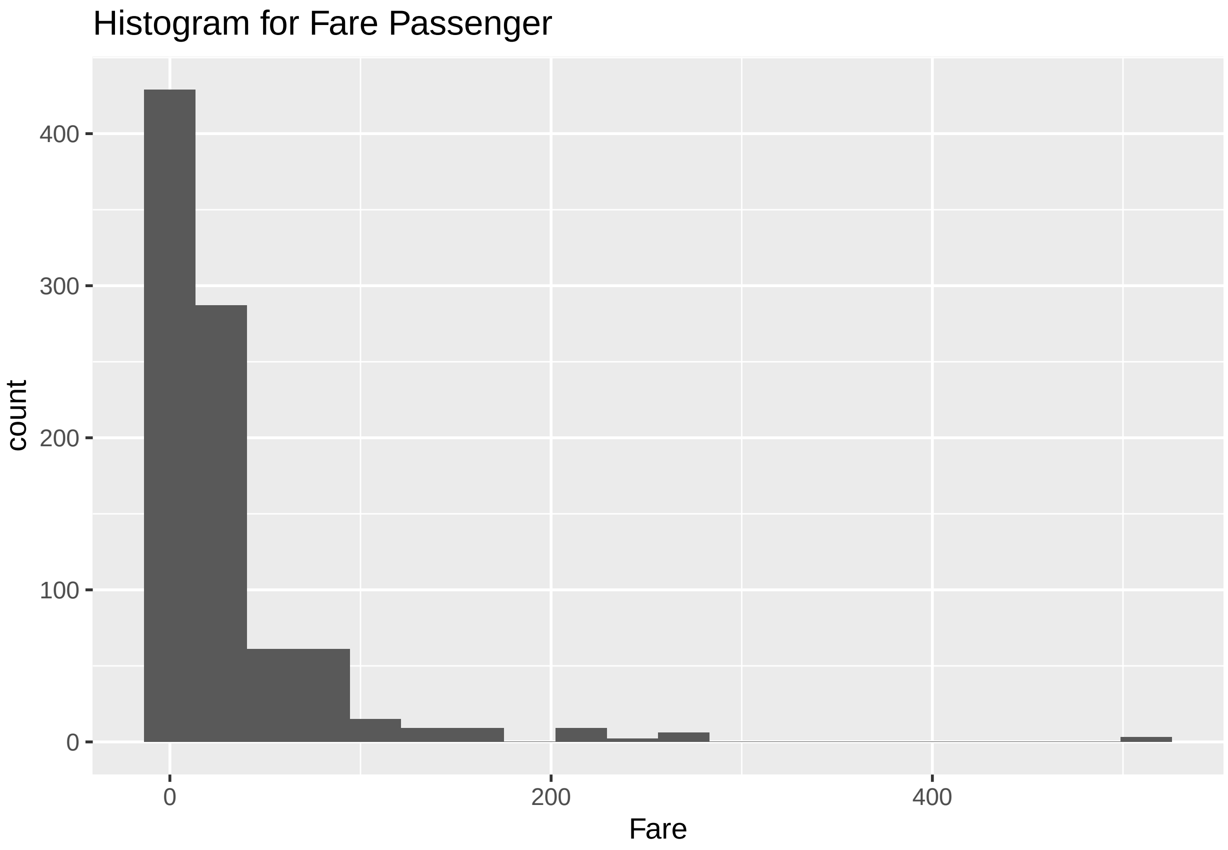
Lets start with Age and Fare

```
ggplot(titanic, aes(x=Age)) + geom_histogram(bins=20) +ggtitle("Histogram for Age of Passenger")
```

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```

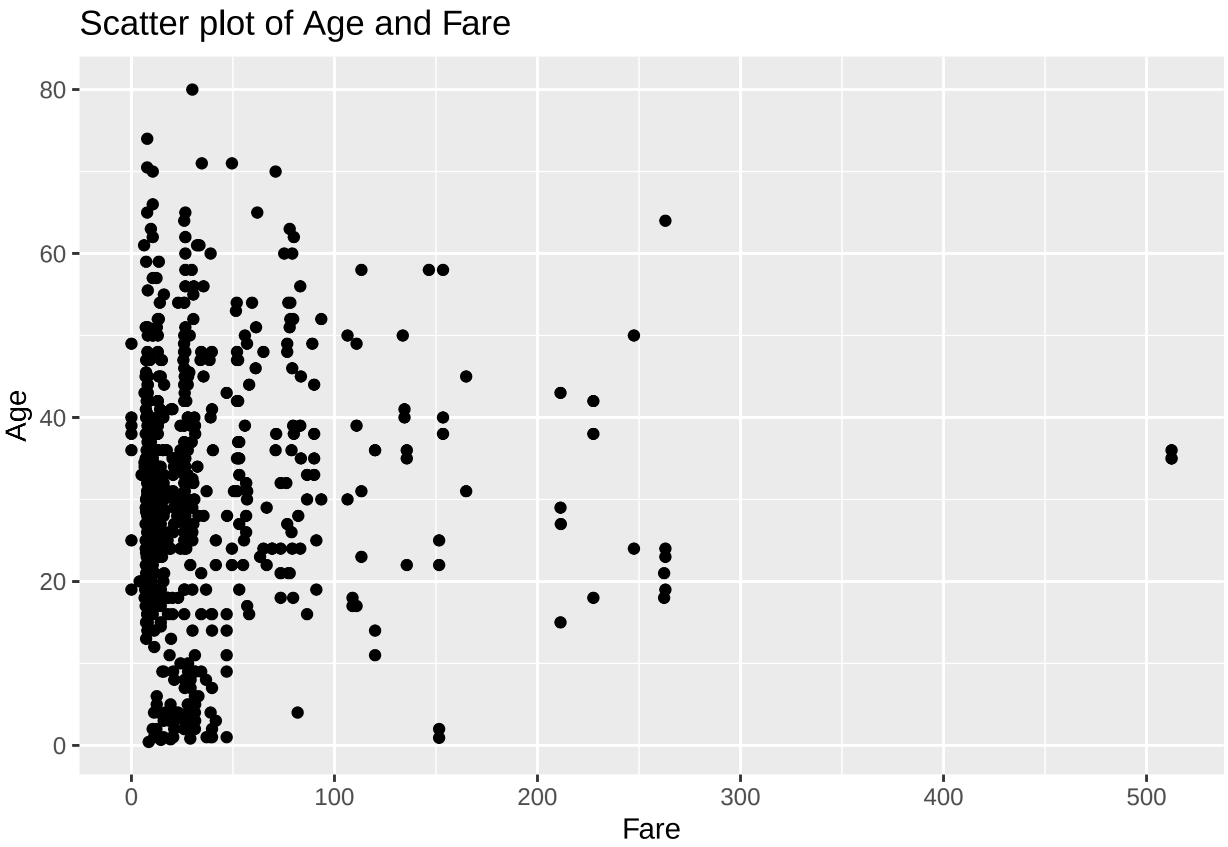


```
ggplot(titanic, aes(x=Fare)) + geom_histogram(bins=20) +ggtitle("Histogram for Fare Passenger")
```



```
ggplot(titanic, aes(x=Fare, y=Age)) + geom_point() +ggtitle("Scatter plot of Age and Fare")
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```

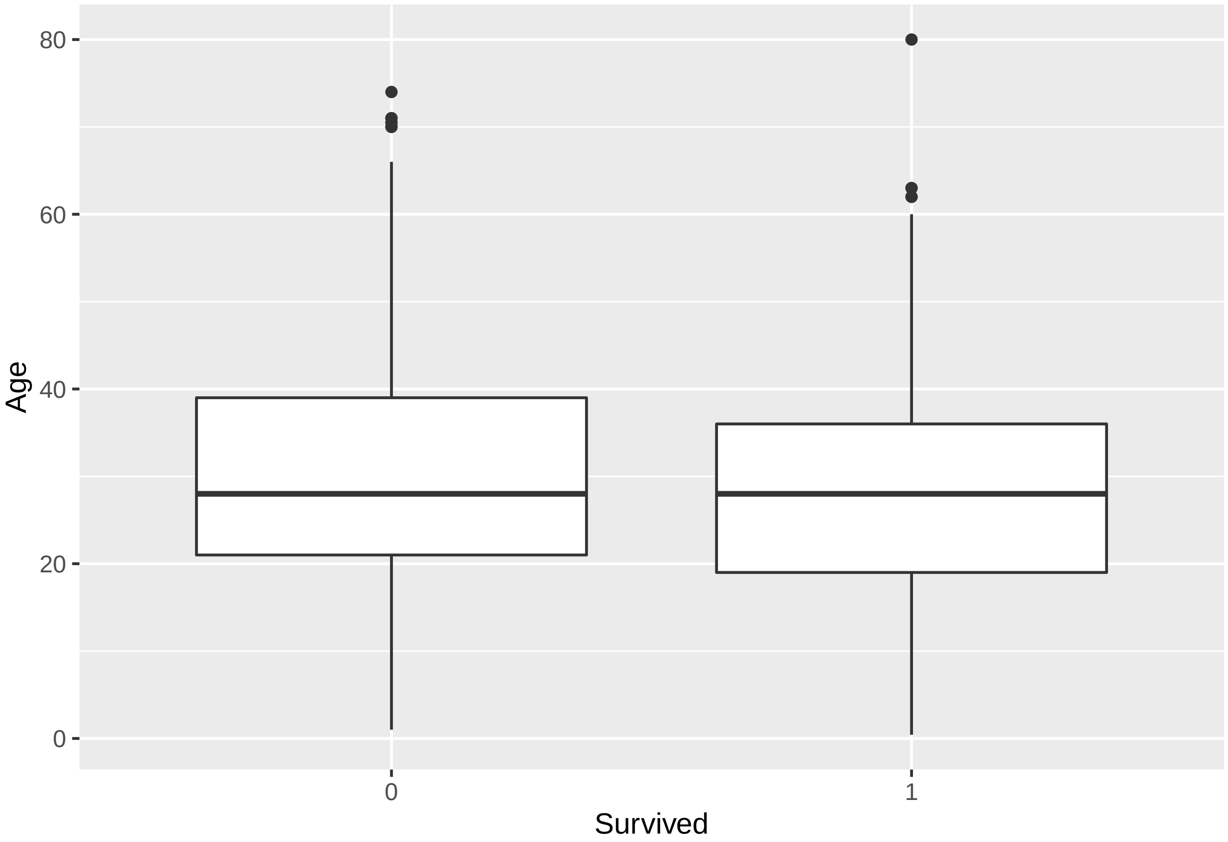


From the the plots so far we can see that Age and Fare seem acceptable in ranges. We are not requiring distributional assumptions. And we can also see that there is no correlation (evident) between Age and Fare.

Now lets look at the Age and the Fare vs the Survival

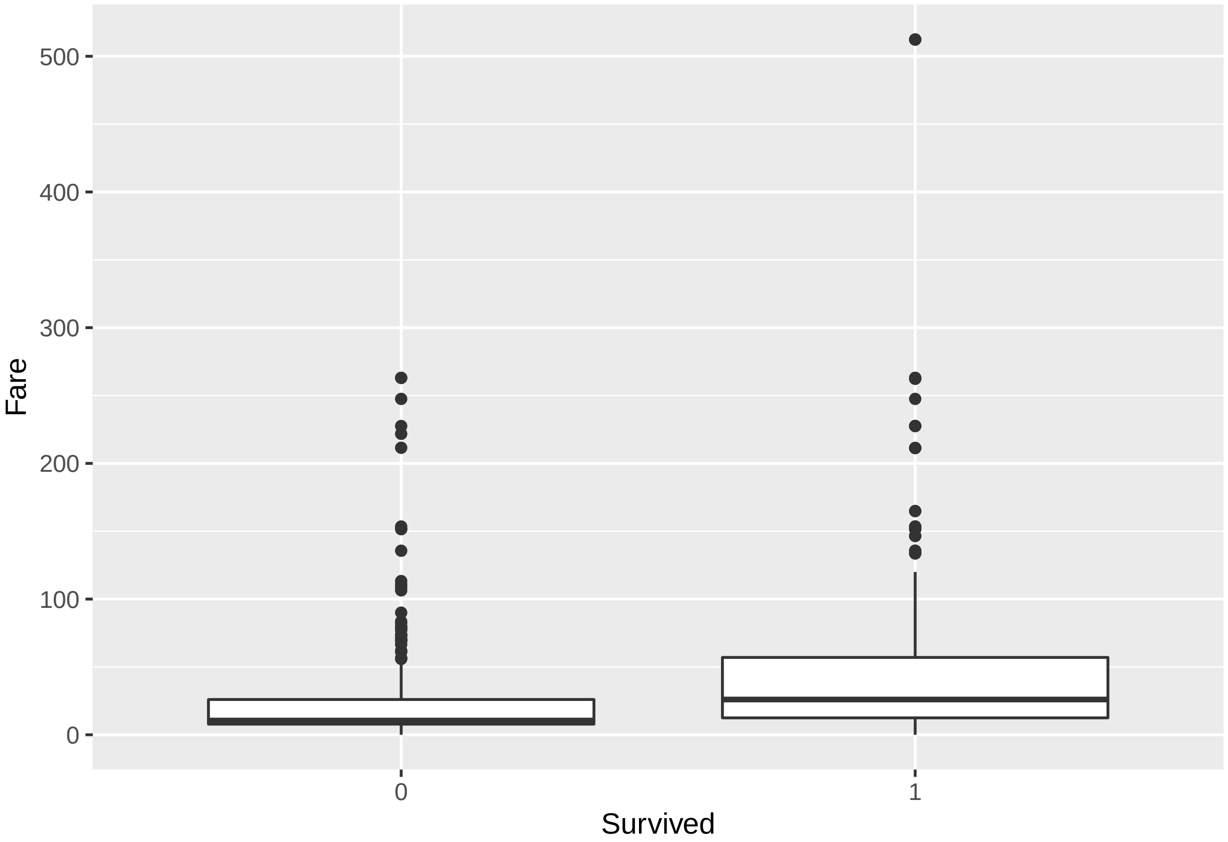
```
ggplot(titanic, aes(x=Survived, y=Age)) + geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```



There is no difference in the median ages (from the plot) between survivors and perished.

```
ggplot(titanic, aes(x=Survived, y=Fare)) + geom_boxplot()
```



Now, outliers aside, there is a difference in the median Fare and their survival status…

3. Explore the data and suggest a model using Age, Gender and fare. As the dependent variable is binary, this is where we can use Logistic Regression.

```
titanic.glm<-glm(titanic$Survived~titanic$Age+titanic$Sex+titanic$Fare, family = "binomial")
summary(titanic.glm)
```

```
##
## Call:
## glm(formula = titanic$Survived ~ titanic$Age + titanic$Sex +
##     titanic$Fare, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4107  -0.6376  -0.5875   0.7900   2.0342
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.934841   0.239101   3.910 9.24e-05 ***
## titanic$Age      -0.010570   0.006498  -1.627    0.104
## titanic$Sexmale -2.347599   0.189956 -12.359  < 2e-16 ***
## titanic$Fare      0.012773   0.002696   4.738 2.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 716.07  on 710  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 724.07
##
## Number of Fisher Scoring iterations: 5
```

From this we can see that the Fare and the gender are significant but not Age.

```
exp(coef(titanic.glm))
```

```
##     (Intercept)     titanic$Age titanic$Sexmale    titanic$Fare
##      2.54680790      0.98948613      0.09559845      1.01285487
```

The odds ratios show us that being Male lowers survival chances, with every increase in fare there is an increase in survival odds and with every increase in year (Age) there is a decrease in survival odds.

Lets simplify the model - we can do this manually or using step. Note that you may get different models…

```
titanic2.glm<-glm(titanic$Survived~titanic$Sex+titanic$Fare, family = "binomial")
summary(titanic2.glm)
```

```
##
## Call:
## glm(formula = titanic$Survived ~ titanic$Sex + titanic$Fare,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2082  -0.6208  -0.5824   0.8126   1.9658
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.647100   0.148502   4.358 1.32e-05 ***
## titanic$Sexmale -2.422760   0.170515 -14.208  < 2e-16 ***
## titanic$Fare     0.011214   0.002295   4.886 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  884.31  on 888  degrees of freedom
## AIC: 890.31
##
## Number of Fisher Scoring iterations: 5
```

Now there is a dilemma….use the more complex model (but with a variable that has missing data) or a simpler model with a higher AIC?

I am going to use the simpler model, as it will be easier to explain and all the coefficients are significants.

4. What are the odds ratios for survival

```
exp(coef(titanic2.glm))
```

```
##     (Intercept) titanic$Sexmale    titanic$Fare
##      1.90999419      0.08867652      1.01127719
```

```
exp(cbind(OR=coef(titanic2.glm), confint(titanic2.glm)))
```

```
## Waiting for profiling to be done...
```

```
##                         OR      2.5 %    97.5 %
## (Intercept)     1.90999419 1.43146595 2.5639737
## titanic$Sexmale 0.08867652 0.06317085 0.1233137
## titanic$Fare    1.01127719 1.00699425 1.0160904
```

We can see that being male lowers your chances of surviving, where as the more expensive your ticket the higher the changes of surviving.

---

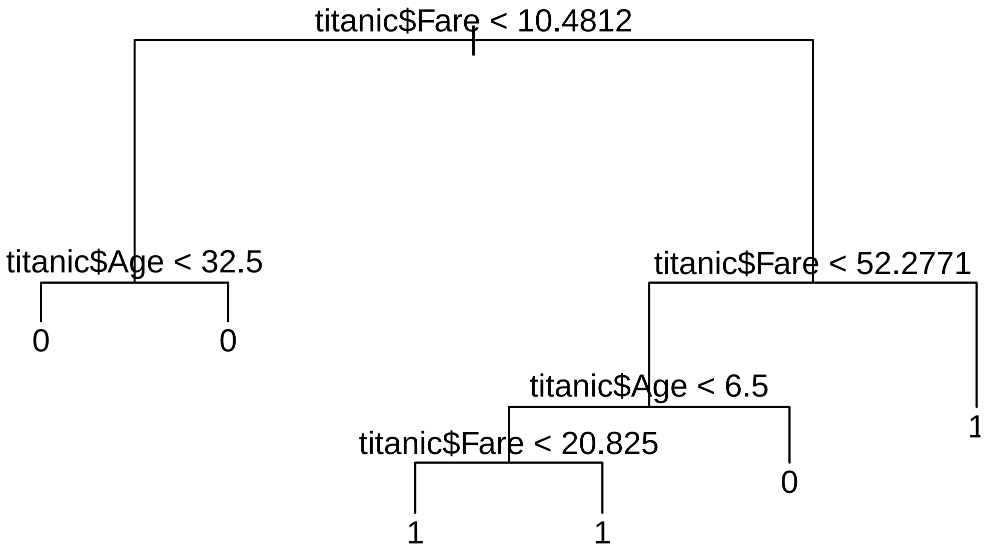
OPTIONAL

Lets start with a model with interactions. I am using a tree to explore the structure.

```
library(tree)
titanic.tree<-tree(titanic$Survived~titanic$Age+titanic$Sex+titanic$Fare)
```

```
## Warning in tree(titanic$Survived ~ titanic$Age + titanic$Sex + titanic$Fare):
## NAs introduced by coercion
```

```
plot(titanic.tree)
text(titanic.tree)
```



This tree structure points to potential interaction between fare and age. But as we have few explanatory variables to begin with lets put all the iteractions in.

```
titanic.i.glm<-glm(titanic$Survived~titanic$Age*titanic$Sex*titanic$Fare, family = "binomial")
summary(titanic.i.glm)
```

```
## 
## Call:
## glm(formula = titanic$Survived ~ titanic$Age * titanic$Sex * 
##     titanic$Fare, family = "binomial")
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.8283  -0.6795  -0.5567   0.8279   2.3043  
## 
## Coefficients:
##                                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               1.1231467  0.4671529   2.404 0.016206
## titanic$Age                              -0.0321516  0.0180072  -1.785 0.074183
## titanic$Sexmale                          -2.0077288  0.5835864  -3.440 0.000581
## titanic$Fare                             -0.0152546  0.0129267  -1.180 0.237968
## titanic$Age:titanic$Sexmale               0.0067298  0.0212314   0.317 0.751263
## titanic$Age:titanic$Fare                  0.0015767  0.0005771   2.732 0.006295
## titanic$Sexmale:titanic$Fare              0.0234538  0.0145952   1.607 0.108065
## titanic$Age:titanic$Sexmale:titanic$Fare -0.0015390  0.0006070  -2.536 0.011224
##
## (Intercept)                              *
## titanic$Age                              .
## titanic$Sexmale                          ***
## titanic$Fare
## titanic$Age:titanic$Sexmale
## titanic$Age:titanic$Fare                 **
## titanic$Sexmale:titanic$Fare
## titanic$Age:titanic$Sexmale:titanic$Fare *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 695.32  on 706  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 711.32
##
## Number of Fisher Scoring iterations: 7
```

The Deviance improvement from the simple model is not too great. I would propose the first or second model.

---

5. Load the full Titanic data

```
titanic.all<-read.csv("data/titanic-all-cols.csv")
```

Exploring the new variables added

```
summary(titanic.all)
```

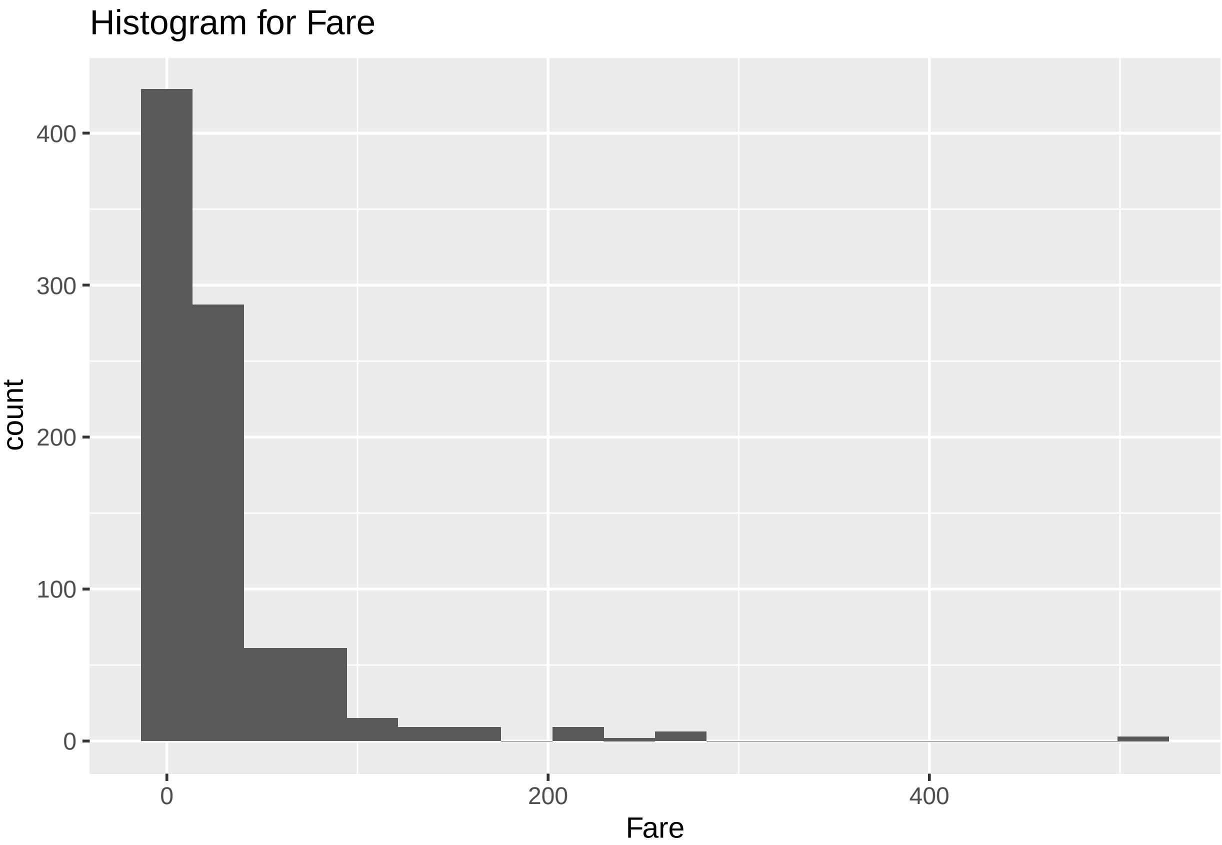
```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

make sure the variables are defined appropriately

```
titanic.all$Survived<-as.factor(titanic.all$Survived)
titanic.all$Pclass<-titanic.all$Pclass
```

Use plots to explore

```
ggplot(titanic.all, aes(x=Fare)) + geom_histogram(bins=20) +ggtitle("Histogram for Fare")
```



We can also look at the relation between the categorical explanatory variables and the dependent variable. For example:

```
table(titanic.all$Survived, titanic.all$Pclass)
```

```
##
##         1   2   3
##   0  80  97 372
##   1 136  87 119
```

We can see that there are more survivors (1) relatively in 1st class compared to others. This will be useful as an explanatory variable.

Now we can start with a model (a large one to begin with)

```
titanic.all.glm<-glm(titanic.all$Survived~titanic.all$Pclass+titanic.all$Sex+ titanic.all$Age +
                       titanic.all$SibSp+ titanic.all$Parch + titanic.all$Fare + titanic.all$Embarked,fa
summary(titanic.all.glm)
```

```
##
## Call:
## glm(formula = titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp + titanic.all$Parch +
##     titanic.all$Fare + titanic.all$Embarked, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7233  -0.6439  -0.3772   0.6288   2.4457
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           17.894850 607.855474   0.029  0.97651
## titanic.all$Pclass    -1.199251   0.164619  -7.285 3.22e-13 ***
## titanic.all$Sexmale   -2.638476   0.222256 -11.871  < 2e-16 ***
## titanic.all$Age       -0.043350   0.008232  -5.266 1.39e-07 ***
## titanic.all$SibSp     -0.363208   0.129017  -2.815  0.00487 **
## titanic.all$Parch     -0.060270   0.123900  -0.486  0.62666
## titanic.all$Fare       0.001432   0.002531   0.566  0.57165
## titanic.all$EmbarkedC -12.257443 607.855250  -0.020  0.98391
## titanic.all$EmbarkedQ -13.080988 607.855452  -0.022  0.98283
## titanic.all$EmbarkedS -12.658656 607.855228  -0.021  0.98339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.34  on 704  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 652.34
##
## Number of Fisher Scoring iterations: 13
```

Lets use a step function to simplify this time…

```
step(titanic.all.glm)
```

```
## Start:  AIC=652.34
## titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp + titanic.all$Parch +
##     titanic.all$Fare + titanic.all$Embarked
##
##                        Df Deviance    AIC
## - titanic.all$Embarked  3   635.81 649.81
## - titanic.all$Parch     1   632.58 650.58
## - titanic.all$Fare      1   632.68 650.68
## <none>                      632.34 652.34
## - titanic.all$SibSp     1   640.91 658.91
## - titanic.all$Age       1   662.75 680.75
## - titanic.all$Pclass    1   686.64 704.64
## - titanic.all$Sex       1   808.42 826.42
##
## Step:  AIC=649.81
## titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp + titanic.all$Parch +
##     titanic.all$Fare
##
##                       Df Deviance    AIC
## - titanic.all$Parch   1   636.07 648.07
## - titanic.all$Fare    1   636.62 648.62
## <none>                    635.81 649.81
## - titanic.all$SibSp   1   645.25 657.25
## - titanic.all$Age     1   667.62 679.62
## - titanic.all$Pclass  1   695.26 707.26
## - titanic.all$Sex     1   815.18 827.18
##
## Step:  AIC=648.07
## titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp + titanic.all$Fare
##
##                       Df Deviance    AIC
## - titanic.all$Fare    1   636.72 646.72
## <none>                    636.07 648.07
## - titanic.all$SibSp   1   647.23 657.23
## - titanic.all$Age     1   667.86 677.86
## - titanic.all$Pclass  1   699.21 709.21
## - titanic.all$Sex     1   820.07 830.07
##
## Step:  AIC=646.72
## titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp
##
##                       Df Deviance    AIC
## <none>                    636.72 646.72
## - titanic.all$SibSp   1   647.29 655.29
## - titanic.all$Age     1   669.44 677.44
## - titanic.all$Pclass  1   742.29 750.29
## - titanic.all$Sex     1   823.84 831.84
```

```
##
## Call:  glm(formula = titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp, family = "binomial")
##
## Coefficients:
##         (Intercept)   titanic.all$Pclass  titanic.all$Sexmale
##             5.60085             -1.31740             -2.62348
##     titanic.all$Age    titanic.all$SibSp
##            -0.04438             -0.37612
##
## Degrees of Freedom: 713 Total (i.e. Null);  709 Residual
##   (177 observations deleted due to missingness)
## Null Deviance:       964.5
## Residual Deviance: 636.7     AIC: 646.7
```

The model is suggests is

```
titanic.step.glm<-glm(titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
    titanic.all$Age + titanic.all$SibSp, family = "binomial")
summary(titanic.step.glm)
```

```
##
## Call:
## glm(formula = titanic.all$Survived ~ titanic.all$Pclass + titanic.all$Sex +
##     titanic.all$Age + titanic.all$SibSp, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7714  -0.6445  -0.3836   0.6276   2.4585
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           5.600846   0.543441  10.306  < 2e-16 ***
## titanic.all$Pclass   -1.317398   0.140900  -9.350  < 2e-16 ***
## titanic.all$Sexmale -2.623483   0.214524 -12.229  < 2e-16 ***
## titanic.all$Age     -0.044385   0.008155  -5.442 5.26e-08 ***
## titanic.all$SibSp   -0.376119   0.121080  -3.106  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 636.72  on 709  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 646.72
##
## Number of Fisher Scoring iterations: 5
```

From these coefficients estimates we can see that: - the higher the travel class the lower the logit for survival - Males have lower survival chances - the higher the number of siblings or spouses also point to lower survival chances - Age also makes a difference, the higher the lower the survival chances

```
exp(coef(titanic.step.glm))
```

```
##         (Intercept)  titanic.all$Pclass titanic.all$Sexmale     titanic.all$Age
##         270.6553299           0.2678313           0.0725497           0.9565859
##   titanic.all$SibSp
##           0.6865205
```

The odds ratio (obviously) paint the same picture. Survival odds are smaller for higher travel class, Male, Age and the higher the number of siblings or spouses.

Other approaches: - Dont use Age as it is missing - Treat Travel class as a category (losing the ordinal relation between 1,2,3) - Add the interactions to the minimal adequate model - Use Tree to see what variables are important to differentiating between survival and not, and see if there are interactions.