Titanic Data Analysis

Isha Doshi

2022-11-27

Loaded data and performed sanity checks.

```r
titanic <-read.delim("titanic.csv.bz2",sep = ",")

dim(titanic)
```

```
## [1] 1309   14
```

```r
names(titanic)
```

```
##  [1] "pclass"    "survived"  "name"      "sex"       "age"       "sibsp"    
##  [7] "parch"     "ticket"    "fare"      "cabin"     "embarked"  "boat"     
## [13] "body"      "home.dest"
```

```r
head(titanic)
```

```
##   pclass survived                                            name    sex
## 1      1        1                   Allen, Miss. Elisabeth Walton female
## 2      1        1                  Allison, Master. Hudson Trevor   male
## 3      1        0                    Allison, Miss. Helen Loraine female
## 4      1        0            Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                             Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                    St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                    New York, NY
```

```
any(is.na(titanic))
```

```
## [1] TRUE
```

Checking for missing values

```
missings <- sum(is.na(titanic))
cat("The total number of missing values are", missings,"\n")
```

```
## The total number of missing values are 1452
```

```
any(is.na(titanic$pclass))
```

```
## [1] FALSE
```

```
any(is.na(titanic$survived))
```

```
## [1] FALSE
```

```
any(is.na(titanic$sex))
```

```
## [1] FALSE
```

```
any(is.na(titanic$age))
```

```
## [1] TRUE
```

```
cat("There are", sum(is.na(titanic$age)),"missing values in age" )
```

```
## There are 263 missing values in age
```

According to me, the variables pclass, sex, age and survived would be the most important ones to describe survival. Based on the story, women and children from higher class would have highest chance of survival, men from lower class having the least chance of survival.

Creating a new variable child, that is 1 if the passenger was younger than 14 years old.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
titanic <- titanic %>% mutate(data=titanic, child = case_when(age < 14 ~ 1,
  age >= 14 ~ 0))
head(titanic)
```

```
##   pclass survived                                            name    sex
## 1      1        1                   Allen, Miss. Elisabeth Walton female
## 2      1        1                  Allison, Master. Hudson Trevor   male
## 3      1        0                    Allison, Miss. Helen Loraine female
## 4      1        0            Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                             Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare    cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest data.pclass data.survived
## 1                    St Louis, MO           1             1
## 2 Montreal, PQ / Chesterville, ON           1             1
## 3 Montreal, PQ / Chesterville, ON           1             0
## 4 Montreal, PQ / Chesterville, ON           1             0
## 5 Montreal, PQ / Chesterville, ON           1             0
## 6                    New York, NY           1             1
##                                         data.name data.sex data.age data.sibsp
## 1                   Allen, Miss. Elisabeth Walton   female  29.0000          0
## 2                  Allison, Master. Hudson Trevor     male   0.9167          1
## 3                    Allison, Miss. Helen Loraine   female   2.0000          1
## 4            Allison, Mr. Hudson Joshua Creighton     male  30.0000          1
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)   female  25.0000          1
## 6                             Anderson, Mr. Harry     male  48.0000          0
##   data.parch data.ticket data.fare data.cabin data.embarked data.boat data.body
## 1          0       24160  211.3375         B5             S         2        NA
## 2          2      113781  151.5500    C22 C26             S        11        NA
## 3          2      113781  151.5500    C22 C26             S                  NA
## 4          2      113781  151.5500    C22 C26             S                 135
## 5          2      113781  151.5500    C22 C26             S                  NA
## 6          0       19952   26.5500        E12             S         3        NA
##                    data.home.dest child
## 1                    St Louis, MO     0
## 2 Montreal, PQ / Chesterville, ON     1
## 3 Montreal, PQ / Chesterville, ON     1
## 4 Montreal, PQ / Chesterville, ON     0
## 5 Montreal, PQ / Chesterville, ON     0
## 6                    New York, NY     0
```

e need to convert pclass into categorical because it consists of discrete values and not continuous ones. It only consists of 0, 1, 2, and 3

```
titanic$pclass <- factor(titanic$pclass)
```

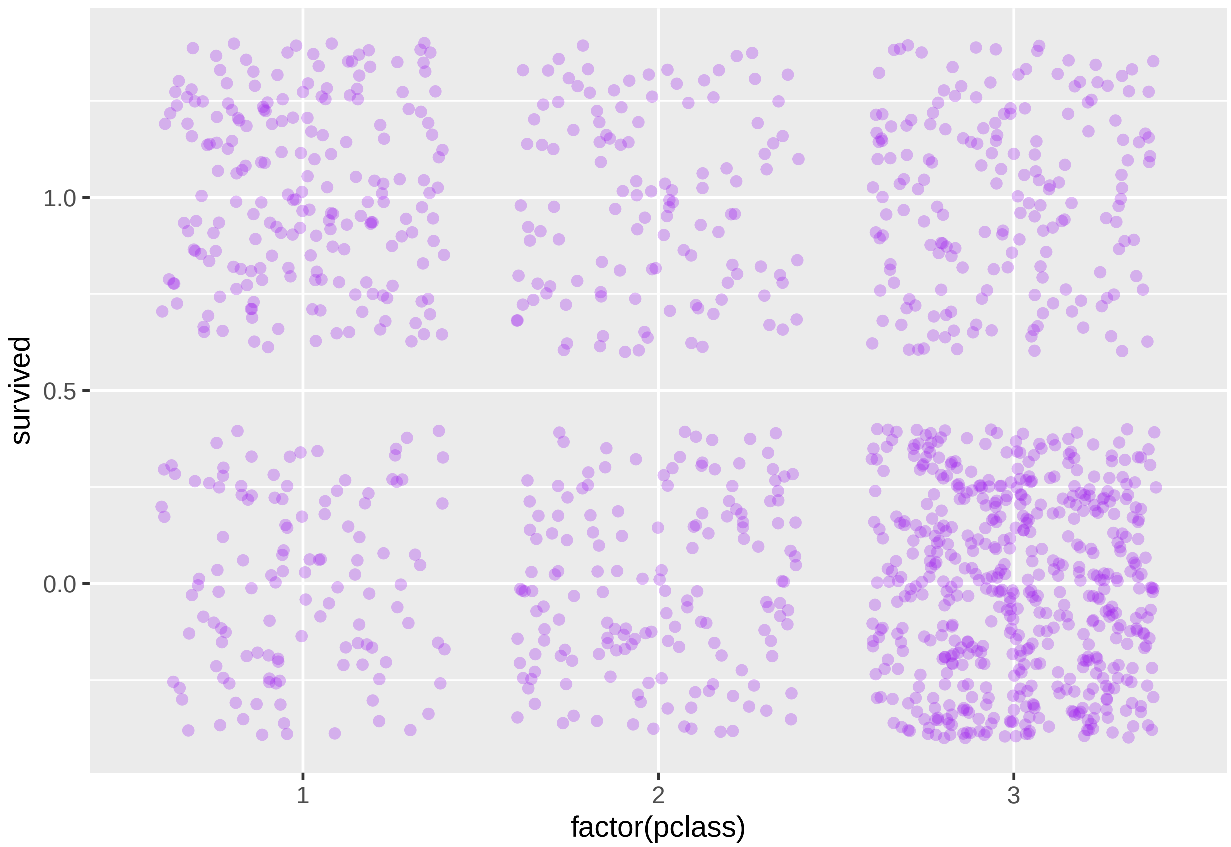
Estimating a multiple logistic regression model

survival vs pclass

```
library(ggplot2)
model_pclass <- glm(survived ~ factor(pclass), data=titanic, family=binomial())
summary(model_pclass)
```

```
##
## Call:
## glm(formula = survived ~ factor(pclass), family = binomial(),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3896  -0.7678  -0.7678   0.9791   1.6525
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.4861     0.1146   4.242 2.21e-05 ***
## factor(pclass)2  -0.7696     0.1669  -4.611 4.02e-06 ***
## factor(pclass)3  -1.5567     0.1433 -10.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1306  degrees of freedom
## AIC: 1619.3
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(titanic,aes(factor(pclass),survived))+geom_jitter(col="purple",alpha=0.3)
```



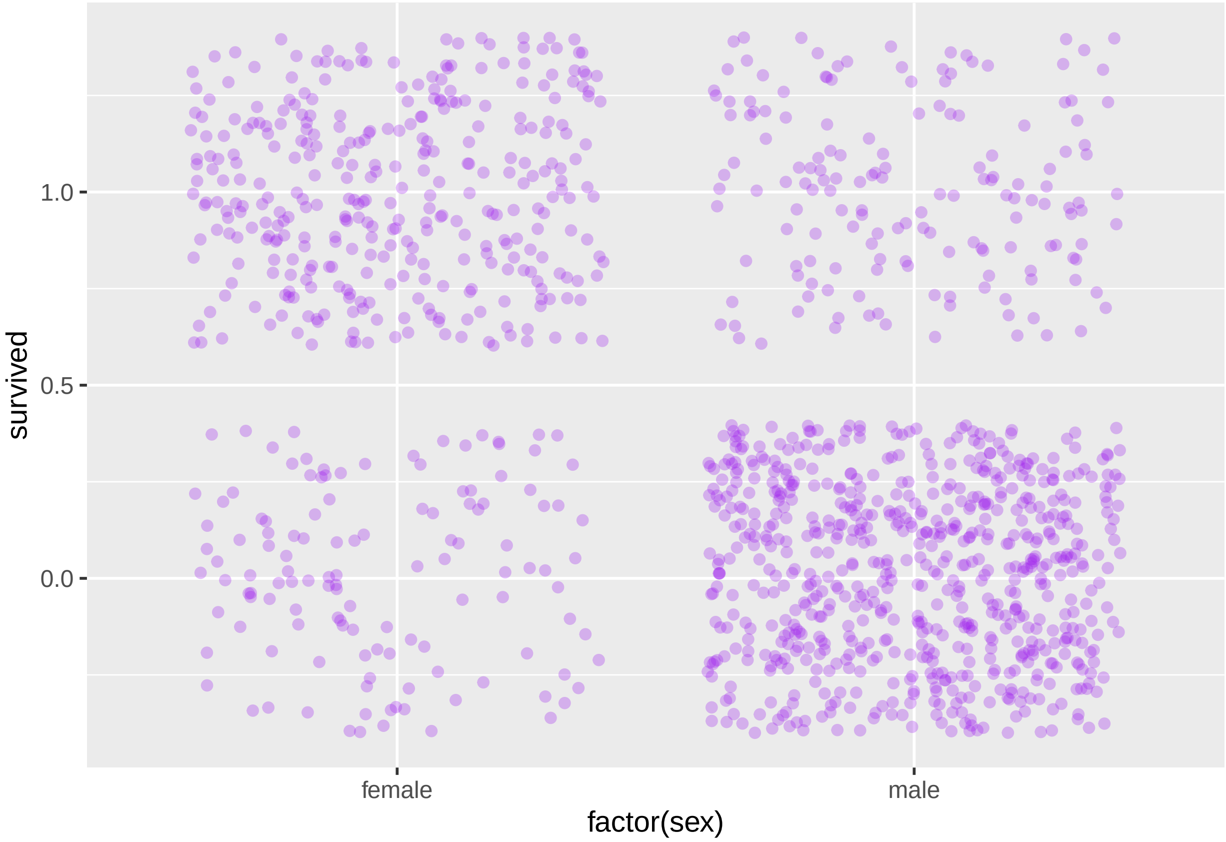
survival vs sex

```
model_sex <- glm(survived ~ factor(sex), data=titanic, family = binomial())
summary(model_sex)
```

```
##
## Call:
## glm(formula = survived ~ factor(sex), family = binomial(), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6124  -0.6511  -0.6511   0.7977   1.8196
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.9818     0.1040   9.437   <2e-16 ***
## factor(sex)male  -2.4254     0.1360 -17.832   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1368.1  on 1307  degrees of freedom
## AIC: 1372.1
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
ggplot(titanic,aes(factor(sex),survived))+geom_jitter(col="purple",alpha=0.3)
```



survival vs age

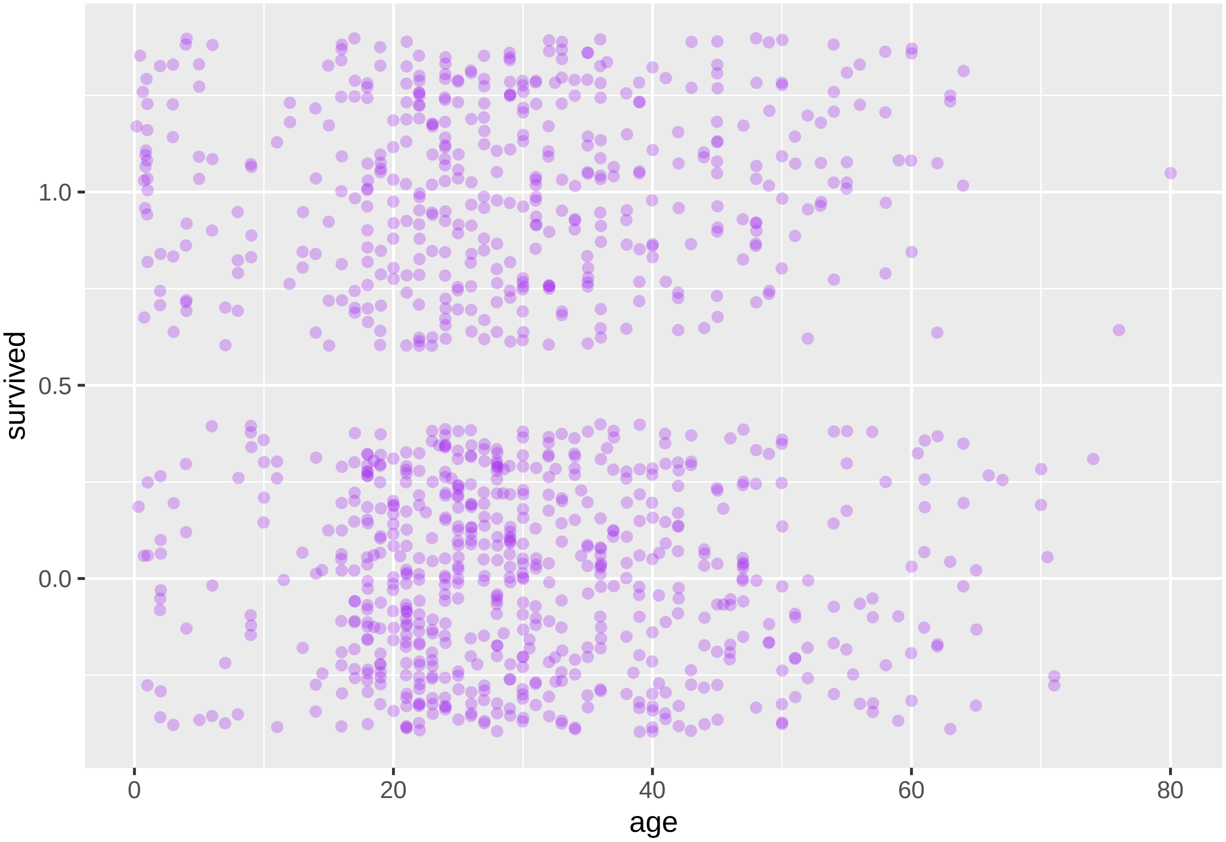
```
model_age <- glm(survived ~ age, data=titanic, family = binomial())
summary(model_age)
```

```
##
## Call:
## glm(formula = survived ~ age, family = binomial(), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1189  -1.0361  -0.9768   1.3187   1.5162
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.136531   0.144715  -0.943   0.3455
## age         -0.007899   0.004407  -1.792   0.0731 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1411.4  on 1044  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1415.4
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(titanic,aes(age,survived))+geom_jitter(col="purple",alpha=0.3)
```

```
## Warning: Removed 263 rows containing missing values (geom_point).
```



survival vs age, pclass, and sex

```
model<-glm(survived~age+factor(sex)+factor(pclass), data=titanic, family=binomial())
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ age + factor(sex) + factor(pclass),
##     family = binomial(), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6399  -0.6979  -0.4336   0.6688   2.3964
```

```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        3.522074   0.326702  10.781  < 2e-16 ***
## age               -0.034393   0.006331  -5.433 5.56e-08 ***
## factor(sex)male   -2.497845   0.166037 -15.044  < 2e-16 ***
## factor(pclass)2   -1.280570   0.225538  -5.678 1.36e-08 ***
## factor(pclass)3   -2.289661   0.225802 -10.140  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  982.45  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 992.45
##
## Number of Fisher Scoring iterations: 4
```

Based on the AIC score, we can interpret that multiple logistic regression with all three variables is better than survival vs every variable. Women had a better chance of survival than men. People from 1st class had a higher chance of survival than people from 2nd and 3rd class. A lot of people from 3rd class didn’t survive. Younger people (below the age of 35) had a better chance of survival.

More young men (18-35) died than children and old men. Although they appear to be more likely to survive than older men.

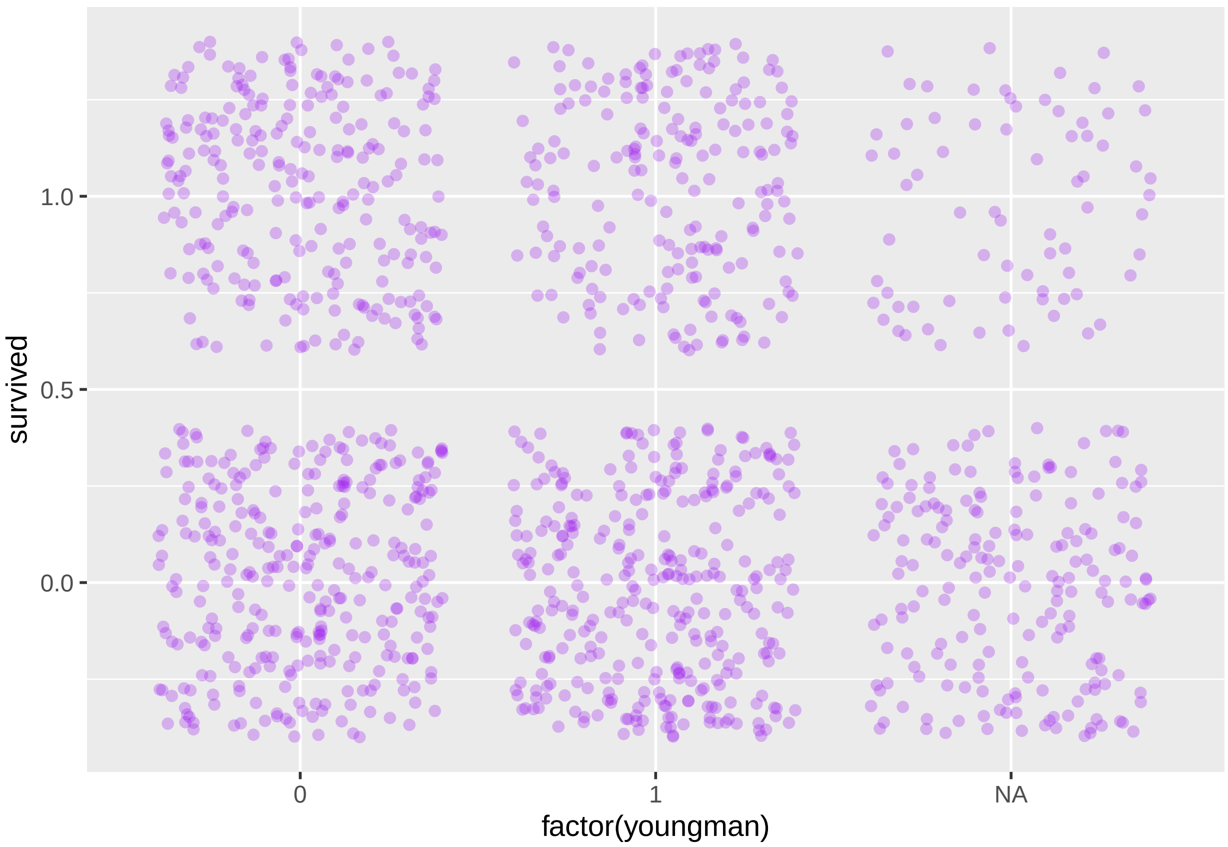
```
titanic <- titanic %>% mutate(data=titanic, youngman = case_when(age > 18 & age < 35 ~ 1,
  age <=18 ~ 0,
  age >=35 ~ 0))
head(titanic)
```

```
##   pclass survived                                            name    sex
## 1      1        1                   Allen, Miss. Elisabeth Walton female
## 2      1        1                  Allison, Master. Hudson Trevor   male
## 3      1        0                    Allison, Miss. Helen Loraine female
## 4      1        0            Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                             Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest data.pclass data.survived
## 1                    St Louis, MO           1             1
## 2 Montreal, PQ / Chesterville, ON           1             1
## 3 Montreal, PQ / Chesterville, ON           1             0
## 4 Montreal, PQ / Chesterville, ON           1             0
## 5 Montreal, PQ / Chesterville, ON           1             0
## 6                    New York, NY          1            1
##                                          data.name data.sex data.age data.sibsp
## 1                  Allen, Miss. Elisabeth Walton   female  29.0000          0
## 2                 Allison, Master. Hudson Trevor     male   0.9167          1
## 3                   Allison, Miss. Helen Loraine   female   2.0000          1
## 4           Allison, Mr. Hudson Joshua Creighton     male  30.0000          1
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)   female  25.0000          1
## 6                            Anderson, Mr. Harry     male  48.0000          0
##   data.parch data.ticket data.fare data.cabin data.embarked data.boat data.body
## 1          0       24160  211.3375         B5             S         2        NA
## 2          2      113781  151.5500    C22 C26             S        11        NA
## 3          2      113781  151.5500    C22 C26             S                  NA
## 4          2      113781  151.5500    C22 C26             S                 135
## 5          2      113781  151.5500    C22 C26             S                  NA
## 6          0       19952   26.5500        E12             S         3        NA
##                      data.home.dest data.data.pclass data.data.survived
## 1                   St Louis, MO                1                  1
## 2 Montreal, PQ / Chesterville, ON                1                  1
## 3 Montreal, PQ / Chesterville, ON                1                  0
## 4 Montreal, PQ / Chesterville, ON                1                  0
## 5 Montreal, PQ / Chesterville, ON                1                  0
## 6                   New York, NY                1                  1
##                                     data.data.name data.data.sex data.data.age
## 1                  Allen, Miss. Elisabeth Walton        female       29.0000
## 2                 Allison, Master. Hudson Trevor          male        0.9167
## 3                   Allison, Miss. Helen Loraine        female        2.0000
## 4           Allison, Mr. Hudson Joshua Creighton          male       30.0000
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)        female       25.0000
## 6                            Anderson, Mr. Harry          male       48.0000
##   data.data.sibsp data.data.parch data.data.ticket data.data.fare
## 1               0               0            24160       211.3375
## 2               1               2           113781       151.5500
## 3               1               2           113781       151.5500
## 4               1               2           113781       151.5500
## 5               1               2           113781       151.5500
## 6               0               0            19952        26.5500
##   data.data.cabin data.data.embarked data.data.boat data.data.body
## 1              B5                  S              2             NA
## 2         C22 C26                  S             11             NA
## 3         C22 C26                  S                            NA
## 4         C22 C26                  S                           135
## 5         C22 C26                  S                            NA
## 6             E12                  S              3             NA
##               data.data.home.dest data.child child youngman
## 1                   St Louis, MO          0     0        1
## 2 Montreal, PQ / Chesterville, ON          1     1        0
## 3 Montreal, PQ / Chesterville, ON          1     1        0
## 4 Montreal, PQ / Chesterville, ON          0     0        1
## 5 Montreal, PQ / Chesterville, ON          0     0        1
## 6                   New York, NY          0     0        0
```

```r
model_youngMan <- glm(survived ~ factor(youngman), data=titanic, family=binomial())
summary(model_youngMan)
```

```
##
## Call:
## glm(formula = survived ~ factor(youngman), family = binomial(),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0716  -1.0716  -0.9744   1.2871   1.3950
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.25415    0.08692  -2.924  0.00346 **
## factor(youngman)1 -0.24410    0.12621  -1.934  0.05310 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1410.9  on 1044  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1414.9
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(titanic,aes(factor(youngman),survived))+geom_jitter(col="purple",alpha=0.3)
```



The survivors' accounts is kind of accurate. Women definitely had a better chance of survival. More men from higher classes survived than the lower class. Not a lot of young children survived, a lot of people from the age of 18-35 could not survive either. There were also multiple entries with missing age, which could have affected our analysis.