Titanic Dataset: data pre-processing, predictive analysis, regressions, association

Main objective:

To analyse to which extent demographic characteristics of Titanic's passengers affected their chances of surviving.

Research phases:

- Data pre-processing
- Data preparation
- Predictive analysis
- Regressions
- Association methods

Original variables:

- Passenger ID
- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

SibSp

Parch

'Fam' Variable → Dummy

1 — They have family

0 — They don't have family

```
##   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
##
```

Summary of our original data with demographic characteristics

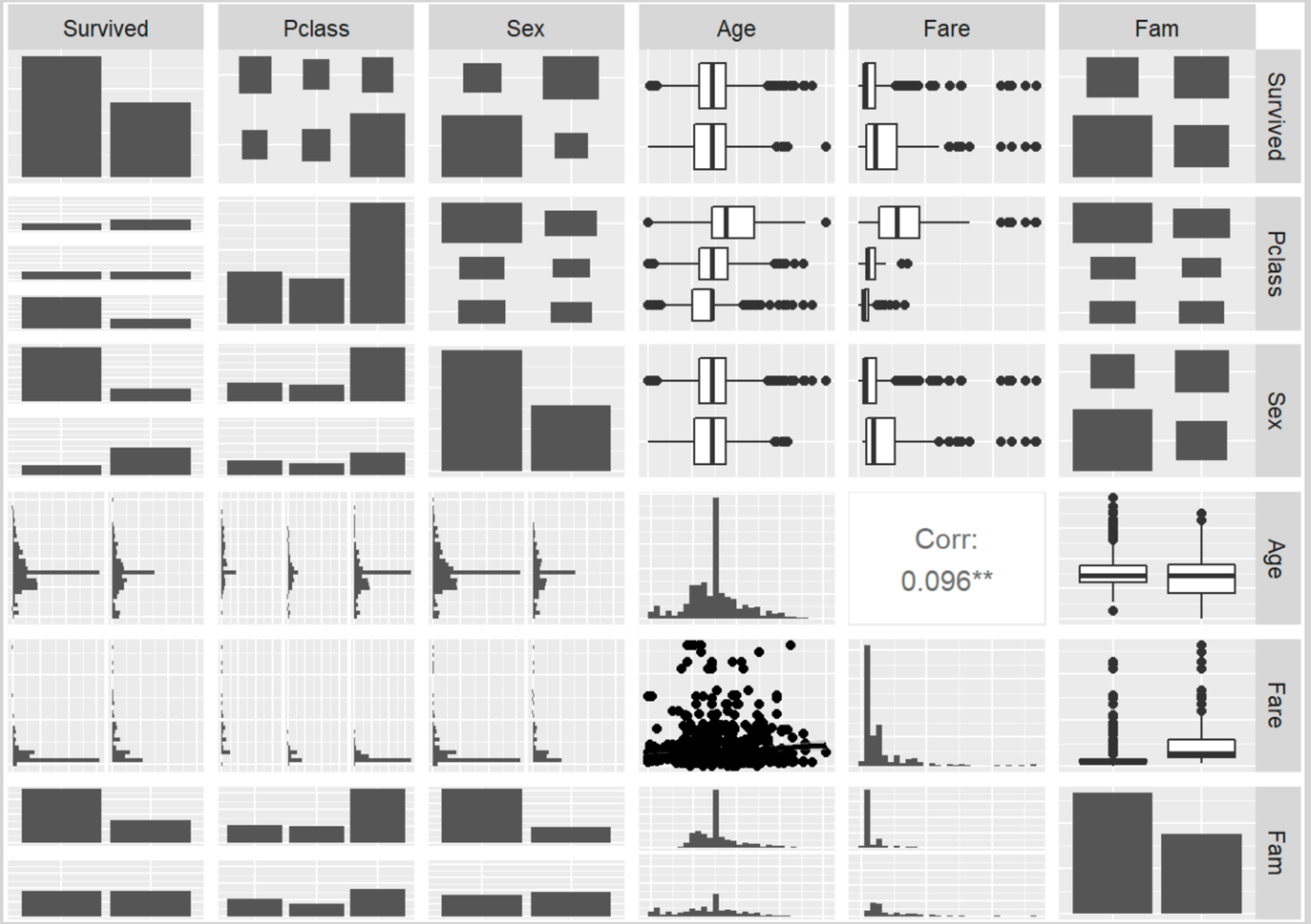
Data pre-processing and initial observations

1:

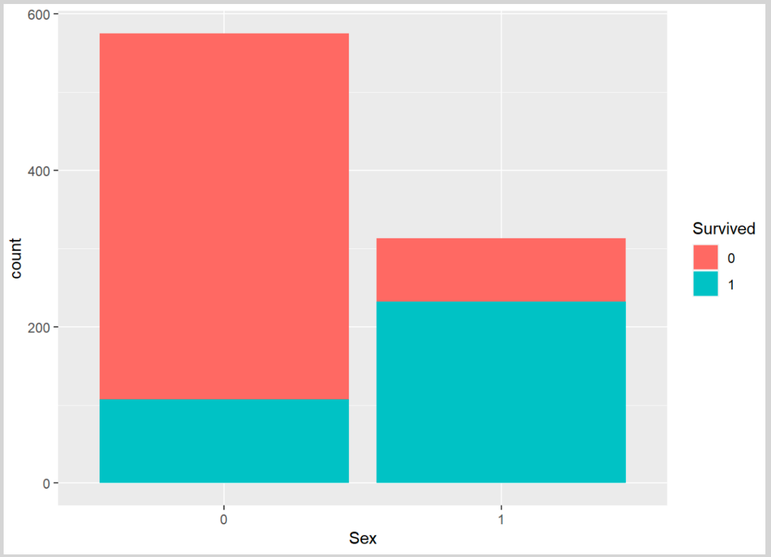
We prepare our variables, normalizing them and dealing with outliers and missing values.

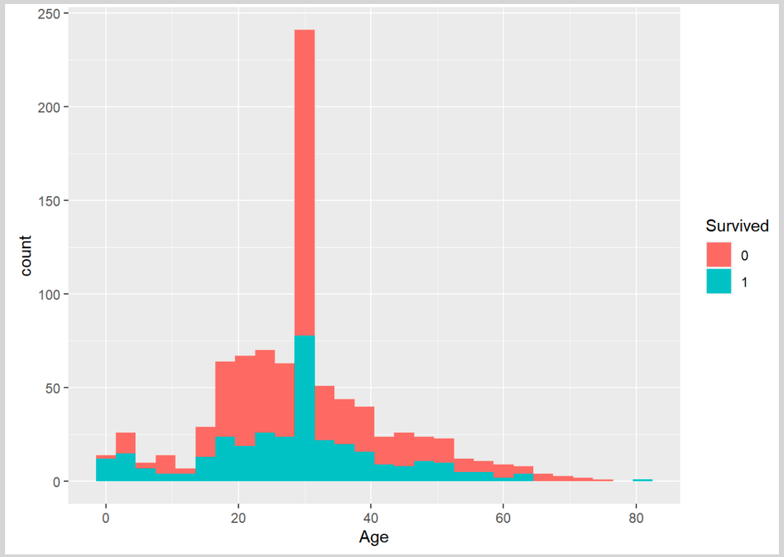
2:

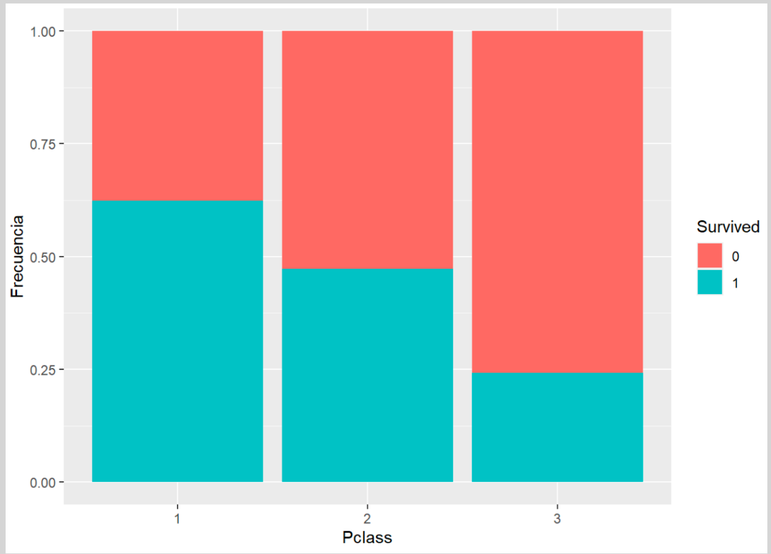
We observe how different demographic variables correlate with the variable ‘Survived’, which states whether that person did or did not survive the Titanic accident.



Correlation Analysis







Sex

Age

Pclass

Variable distribution in terms of survival

With some variables, we perceive indications of a potential relation between such characteristic and survival chances.

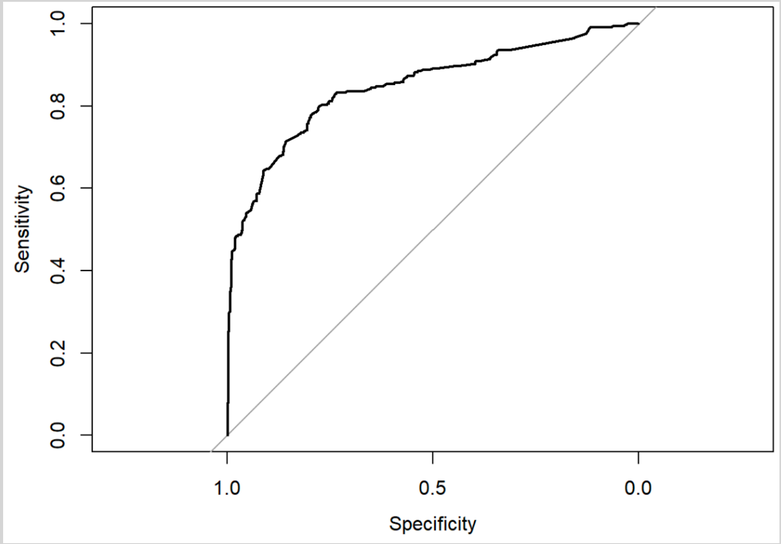
Regressions

```
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Fam, family = binomial,
##     data = df_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6326  -0.6498  -0.4265   0.6237   2.4271
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.939319   0.352065   2.668  0.00763 **
## Age         -0.033740   0.007533  -4.479 7.50e-06 ***
## Sex1         2.639080   0.194296  13.583  < 2e-16 ***
## Pclass2     -1.095095   0.259563  -4.219 2.45e-05 ***
## Pclass3     -2.312422   0.244644  -9.452  < 2e-16 ***
## Fam1        -0.077492   0.188340  -0.411  0.68074
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1180.89  on 887  degrees of freedom
## Residual deviance:  801.61  on 882  degrees of freedom
## AIC: 813.61
##
## Number of Fisher Scoring iterations: 5
```

Logistic regression:

With relevant variables (sex, age, Pclass as proxy of social class and Fam)

We evaluate how much do variances in these variables affect a person's chance of survival.



CROSS VALIDATION

To evaluate the model's effectiveness.

AUC (r) = 0.8479

Association methods

ARULES:

We get to understand how different variables relate to each other