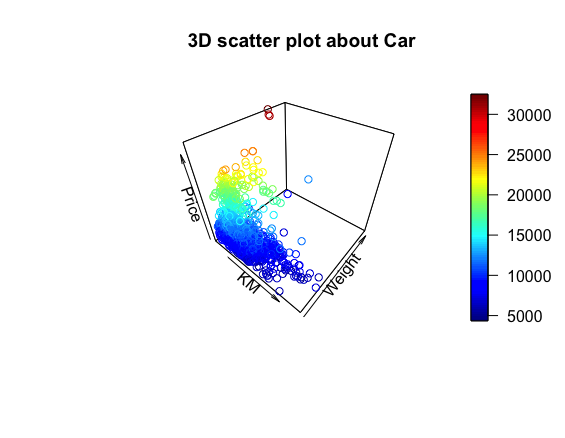
**HW1 (산업공학과 2014002542 황보성훈)**

**PART #1**



**PART #2**

**(1) 5-fold cross-validation**

**(Model 1) Price =**

> modelcv1

Linear Regression

1436 samples

3 predictor

No pre-processing

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 1149, 1148, 1150, 1149, 1148

Resampling results:

RMSE Rsquared MAE

3531.778 0.3415816 2187.706

Tuning parameter 'intercept' was held constant at a value of TRUE

**(Model 2) Price =**

> modelcv2

Linear Regression

1436 samples

4 predictor

No pre-processing

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 1148, 1149, 1149, 1149, 1149

Resampling results:

RMSE Rsquared MAE

3458.01 0.3423959 2141.731

Tuning parameter 'intercept' was held constant at a value of TRUE

**(Model 3) Price =**

> modelcv3

Linear Regression

1436 samples

5 predictor

No pre-processing

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 1149, 1149, 1149, 1148, 1149

Resampling results:

RMSE Rsquared MAE

2480.184 0.5360121 1915.162

Tuning parameter 'intercept' was held constant at a value of TRUE

**(2) randomized 80%training-20%testing validation**

**(Model 1) Price =**

> sqrt(mean(error1 ^ 2))

[1] 2720.084

> tot\_error <- test$Price - mean(test$Price)

> sstot\_error <- sum(tot\_error^2)

> (rsquared\_error1 <- 1 - sum(error1^2)/sstot\_error)

[1] 0.3622016

**(Model 2) Price =**

> sqrt(mean(error2 ^ 2))

[1] 2662.329

> (rsquared\_error2 <- 1 - sum(error2^2)/sstot\_error)

[1] 0.3889982

**(Model 3) Price =**

> sqrt(mean(error3 ^ 2))

[1] 2378.588

> (rsquared\_error3 <- 1 - sum(error3^2)/sstot\_error)

[1] 0.5122947

🡪 **From (1), (2), I decided that Model 3 is the best model.**

**PART #3**

> str(titanic\_train2)

'data.frame': 891 obs. of 8 variables:

$ Survived: int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Sex : chr "male" "female" "female" "female" ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Embarked: chr "S" "C" "S" "S" ...

**🡪 From the structure of data set, I decided to covert ‘Sex’, ‘Pclass’, ‘Embarked’ to dummy variable.**

**[Result]**

> head(titanic\_train4)

Survived Age SibSp Parch Fare Sexmale Pclass2 Pclass3 EmbarkedC EmbarkedQ EmbarkedS

1 0 22 1 0 7.2500 1 0 1 0 0 1

2 1 38 1 0 71.2833 0 0 0 1 0 0

3 1 26 0 0 7.9250 0 0 1 0 0 1

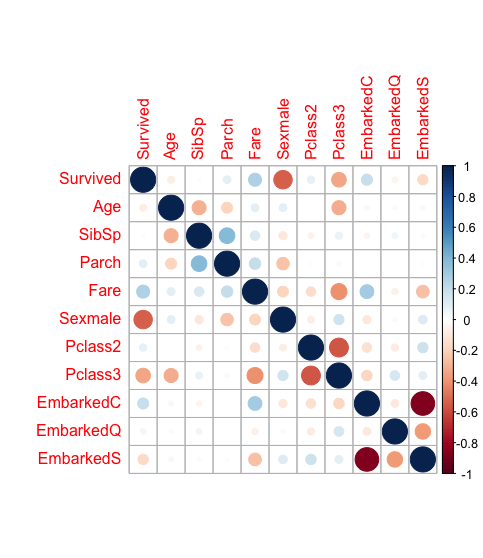
4 1 35 1 0 53.1000 0 0 0 0 0 1

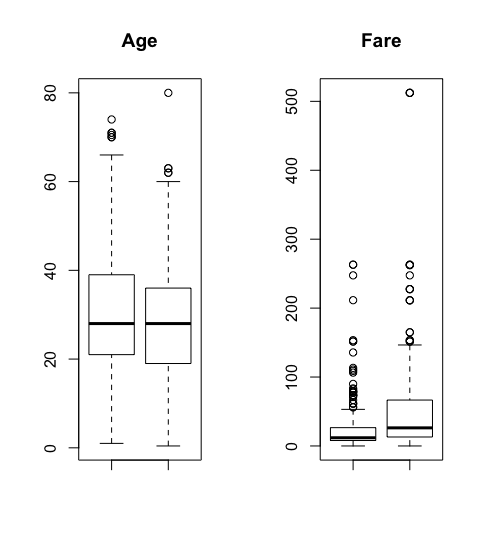
5 0 35 0 0 8.0500 1 0 1 0 0 1

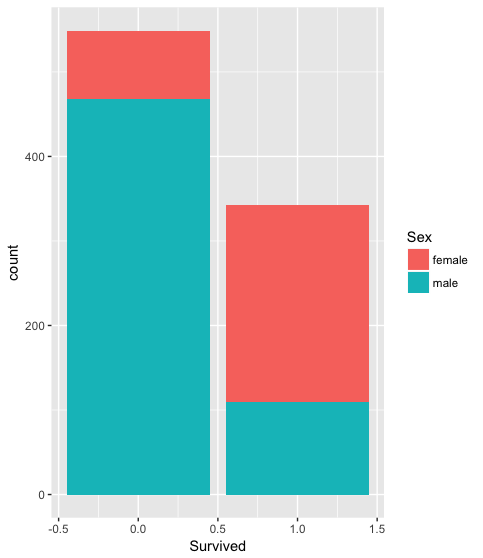
6 0 NA 0 0 8.4583 1 0 1 0 1 0

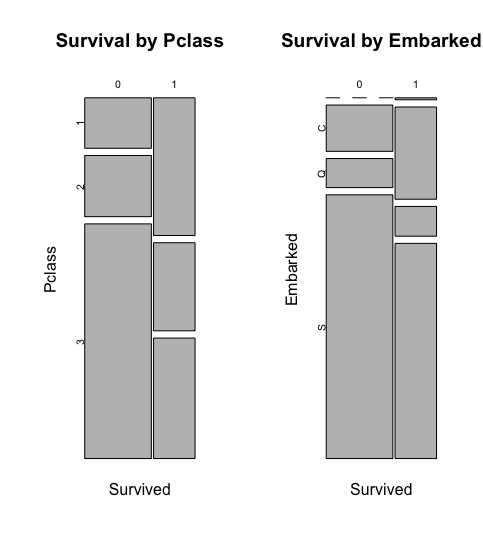
**[EDA]**

1. summary(titanic\_train4)를 통해서 Age의 NA가 177개임을 확인할 수 있다.
2. Correlation plot



1. Box plot을 통한 Age, Fare에 따른 생존 데이터 분포 확인
2. 성별에 따른 생존 데이터 분포 확인



1. Pclass와 Embarked에 따른 생존 데이터 분포 확인