Exercise 7 - R Tableau Integration

Ernesto Nava

03/06/2020

Data Formatting

```
titanic<-read.csv('titanic.csv')
dim(titanic)
```

```
## [1] 891  12
```

```
meanAge<-sum(na.omit(titanic$Age))/length(na.omit(titanic$Age))
meanAge
```

```
## [1] 29.69912
```

```
titanic$Age[is.na(titanic$Age)]<-meanAge
titanic$Age<-round(titanic$Age)

titanic$AgeCat[titanic$Age>=0&titanic$Age<=16]<-"0-16"
titanic$AgeCat[titanic$Age>=17&titanic$Age<=32]<-"17-32"
titanic$AgeCat[titanic$Age>=33&titanic$Age<=48]<-"33-48"
titanic$AgeCat[titanic$Age>=49&titanic$Age<=64]<-"49-64"
titanic$AgeCat[titanic$Age>=65]<-"65 and Above"

titanic$Survived[titanic$Survived==0]<-"Not Survived"
titanic$Survived[titanic$Survived==1]<-"Survived"

titanic$Pclass<-factor(titanic$Pclass)
titanic$AgeCat<-factor(titanic$AgeCat)
titanic$Survived<-factor(titanic$Survived)
titanic$Embarked<-as.character(titanic$Embarked)
titanic$Embarked[titanic$Embarked=="S"]<-"Southampton"
titanic$Embarked[titanic$Embarked=="C"]<-"Cherbourg"
titanic$Embarked[titanic$Embarked=="Q"]<-"Queenstown"
titanic$Embarked<-factor(titanic$Embarked)

titanic=titanic[c(-9,-11)]

View(titanic)

write.csv(titanic, file="titanicNew.csv")
```

Decision Tree Analysis

```
decision_tree<-titanic
SibSpCat=ifelse(decision_tree$SibSp >=3,">=3","<3")
decision_tree<-data.frame(decision_tree,SibSpCat)
decision_tree$SibSpCat<-as.factor(decision_tree$SibSpCat)
ParchCat=ifelse(decision_tree$Parch >=3,">=3","<3")
decision_tree<-data.frame(decision_tree,ParchCat)
decision_tree$ParchCat<-as.factor(decision_tree$ParchCat)

set.seed(1)

test = sample(1:nrow(decision_tree),nrow(decision_tree)/3)
train = -test
training_data = decision_tree[train,]
testing_data = decision_tree[test,]
testing_survived = decision_tree$Survived[test]

library(rpart)
library(rattle)
```

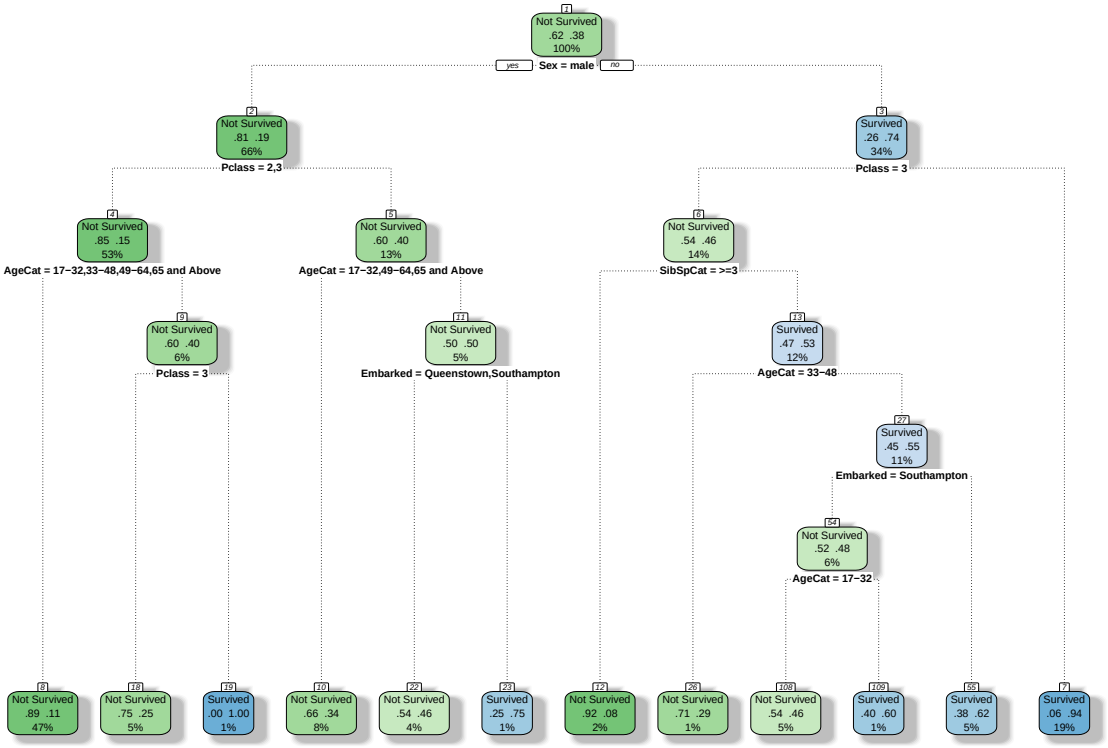
```
## Warning: package 'rattle' was built under R version 4.0.3
```

```
## Loading required package: tibble
```

```
## Loading required package: bitops
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
tree_model=rpart(Survived ~ Pclass + Sex + AgeCat + Embarked + SibSpCat + ParchCat, data = training_data
fancyRpartPlot(tree_model,sub = "decision_tree")
```



decision_tree

```
tree_predict=predict(tree_model,testing_data,type = "class")

mean(tree_predict != testing_survived)
```

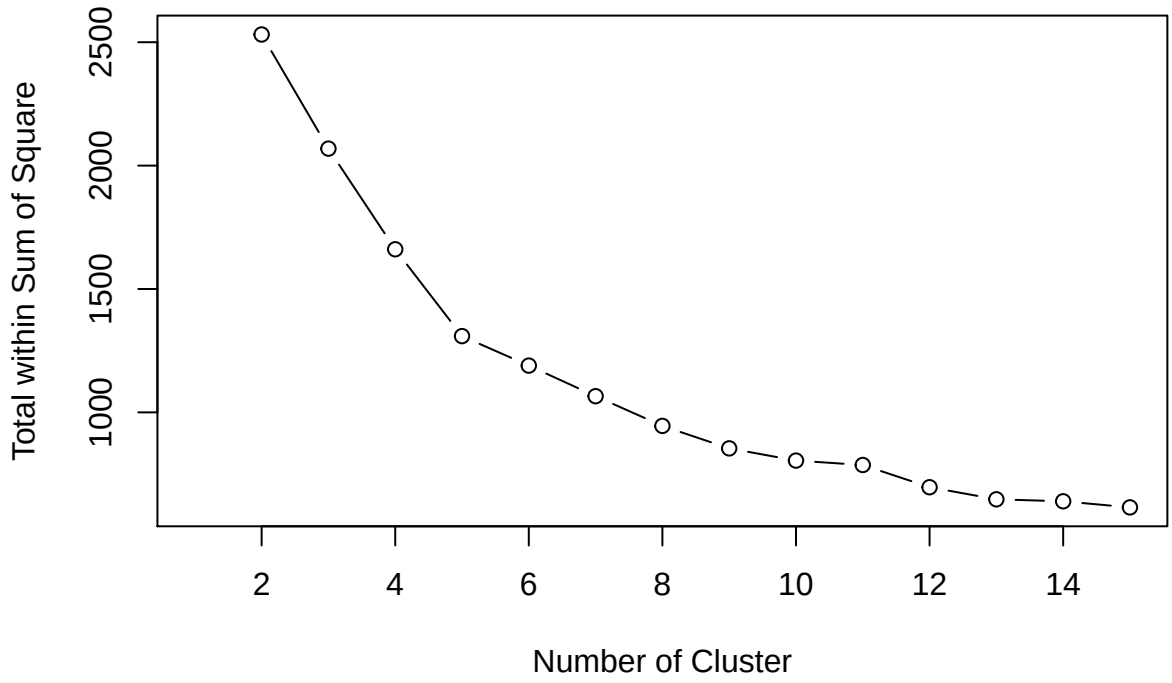
```
## [1] 0.1750842
```

K-means Clustering

```
titanicNew<-read.csv("titanicNew.csv")
titanicUpdated<-titanicNew
SurvivedNum<-ifelse(titanicUpdated$Survived=="Not Survived",0,1)
titanicUpdated<-data.frame(titanicUpdated,SurvivedNum)

SexN<-ifelse(titanicUpdated$Sex=="male",1,0)
titanicUpdated<-data.frame(titanicUpdated,SexN)

EmbarkedN<-ifelse(titanicUpdated$Embarked=="Southampton",1,ifelse(titanicUpdated$Embarked=="Cherbourg",1
titanicUpdated<-data.frame(titanicUpdated,EmbarkedN)

write.csv(titanicUpdated, file="titanicUpdated.csv")

titanic.scaled<-scale(data.frame(titanic$Age,titanic$Parch,titanic$SibSp,titanic$Fare))
colnames(titanic.scaled)
```

```
## [1] "titanic.Age"   "titanic.Parch" "titanic.SibSp" "titanic.Fare"
```

```
totwss<-vector()
btwss<-vector()
for (i in 2:15) {
  set.seed(1234)
  temp<-kmeans(titanic.scaled,centers = i)
  totwss[i]<-temp$tot.withinss
  btwss[i]<-temp$betweenss
}

plot(totwss,xlab = "Number of Cluster",type = "b", ylab = "Total within Sum of Square")
```




```
plot(btwss,xlab = "Number of Cluster",type = "b", ylab = "Total Between Sum of Square")
```

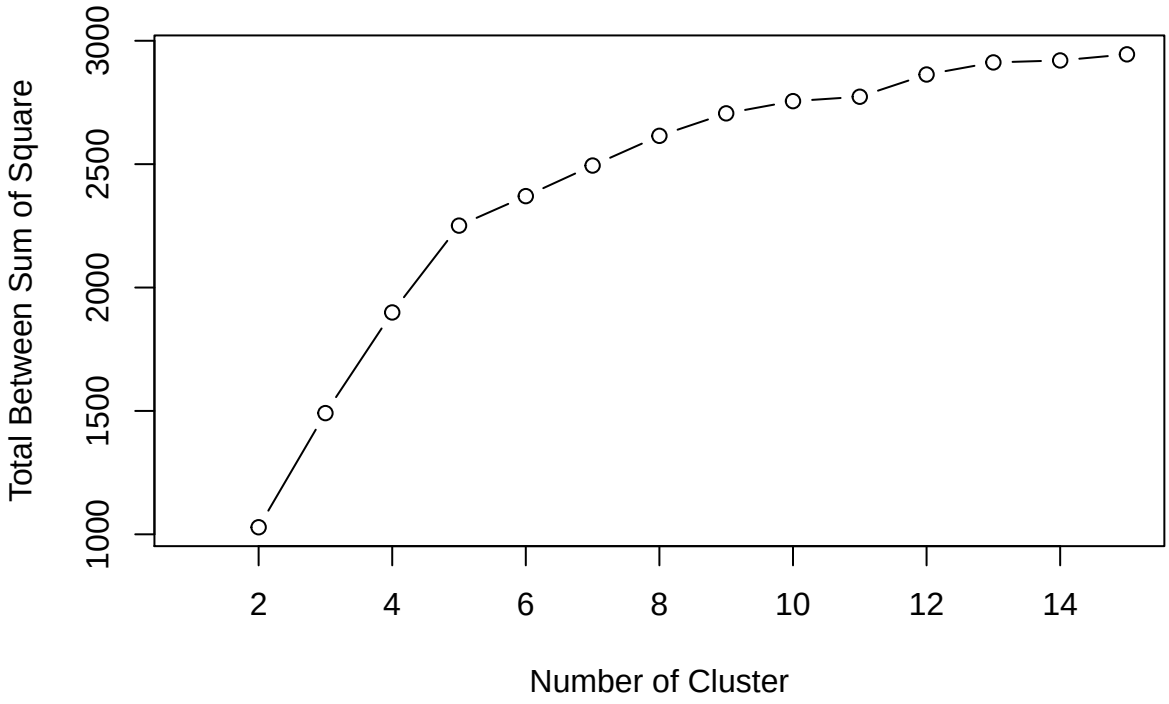


Tableau Integration

```
#install.packages("Rserve")
library(Rserve)
```

```
## Warning: package 'Rserve' was built under R version 4.0.3
```

```
Rserve()
```

```
## Starting Rserve...
##  "C:\Users\navaa\DOCUME~1\R\win-library\4.0\Rserve\libs\x64\Rserve.exe"
```