Titanic data Multivariate Analysis

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

train\_data <- read.csv(file="https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.csv",header=T,sep=",")  
str(train\_data)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ 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 ...  
## $ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

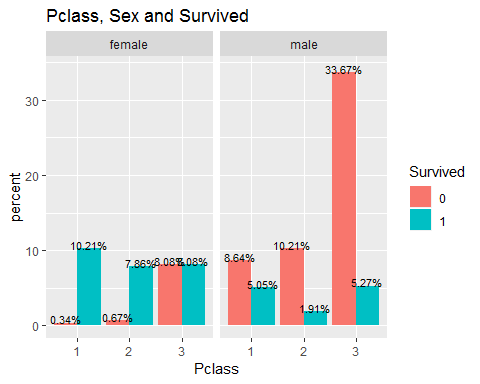
train\_data$Survived = as.factor(train\_data$Survived)  
train\_data$Pclass = as.factor(train\_data$Pclass)  
train\_data$Name = as.character(train\_data$Name)  
train\_data$Ticket = as.character(train\_data$Ticket)  
train\_data$Embarked[train\_data$Embarked == ""] = 'S'  
train\_data$Title <- sapply(train\_data$Name, FUN=function(x) {strsplit(x, split='[,.]')[[1]][2]})  
train\_data$Title <- sub(' ', '', train\_data$Title)  
table(train\_data$Title)

##   
## Capt Col Don Dr Jonkheer   
## 1 2 1 7 1   
## Lady Major Master Miss Mlle   
## 1 2 40 182 2   
## Mme Mr Mrs Ms Rev   
## 1 517 125 1 6   
## Sir the Countess   
## 1 1

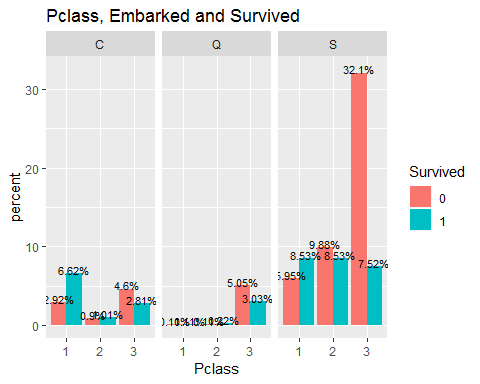
train\_data$Title[train\_data$Title %in% c('Mlle', 'Ms', 'Dona', 'Lady')] <- 'Miss'  
train\_data$Title[train\_data$Title %in% c('Mme')] <- 'Mrs'  
train\_data$Title[train\_data$Title %in% c('Capt', 'Col', 'Don', 'Dr', 'Jonkheer', 'Rev', 'the Countess', 'Major', 'Sir')] <- 'Officer'  
mean\_mr = mean(train\_data$Age[train\_data$Title == 'Mr' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Mr' & is.na(train\_data$Age)]=mean\_mr  
  
mean\_mrs = mean(train\_data$Age[train\_data$Title == 'Mrs' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Mrs' & is.na(train\_data$Age)]=mean\_mrs  
  
mean\_master = mean(train\_data$Age[train\_data$Title == 'Master' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Master' & is.na(train\_data$Age)]=mean\_master  
  
mean\_miss = mean(train\_data$Age[train\_data$Title == 'Miss' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Miss' & is.na(train\_data$Age)]=mean\_miss  
  
mean\_officer = mean(train\_data$Age[train\_data$Title == 'Officer' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Officer' & is.na(train\_data$Age)]=mean\_officer

Pclass, Sex and Survived

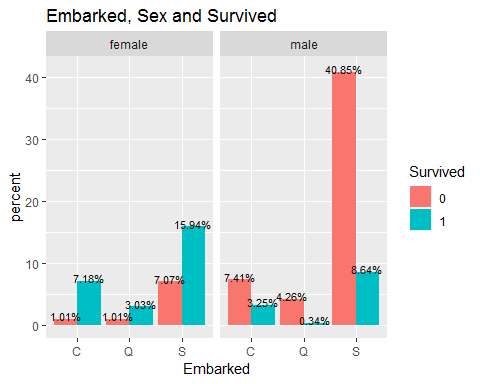
library(ggplot2)  
ggplot(train\_data, aes(x = Pclass, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Sex) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Survived', title = 'Pclass, Sex and Survived')+ theme\_grey()

 Pclass, Embarked and Survived

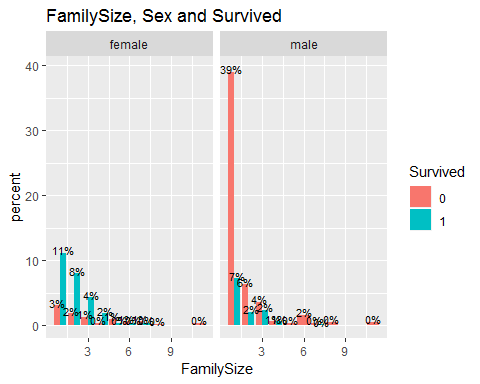
ggplot(train\_data, aes(x = Pclass, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Embarked) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Survived', title = 'Pclass, Embarked and Survived')+ theme\_grey()

 Embarked, Sex and Survived

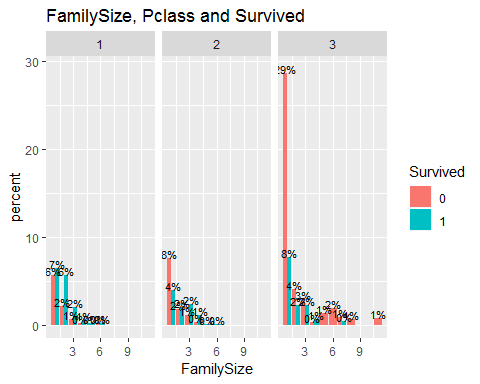
ggplot(train\_data, aes(x = Embarked, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Sex) +  
 labs(x = 'Embarked', y = 'percent', fill = 'Survived', title = 'Embarked, Sex and Survived')+ theme\_grey()

 FamilySize, Sex and Survived

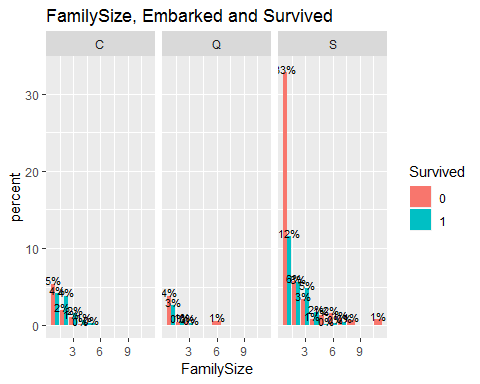
train\_data$FamilySize <- train\_data$SibSp + train\_data$Parch +1  
ggplot(train\_data, aes(x = FamilySize, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 0),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Sex) +  
 labs(x = 'FamilySize', y = 'percent', fill = 'Survived', title = 'FamilySize, Sex and Survived')+ theme\_grey()

 FamilySize, Pclass and Survived

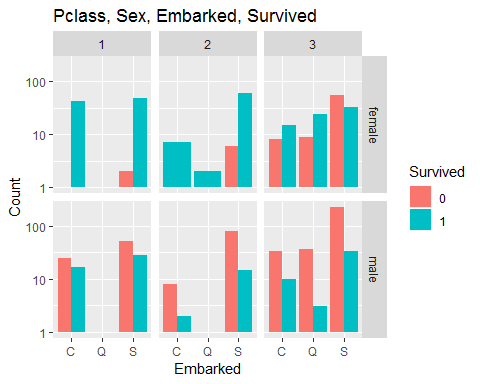
ggplot(train\_data, aes(x = FamilySize, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 0),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Pclass) +  
 labs(x = 'FamilySize', y = 'percent', fill = 'Survived', title = 'FamilySize, Pclass and Survived')+ theme\_grey()

 FamilySize, Embarked and Survived

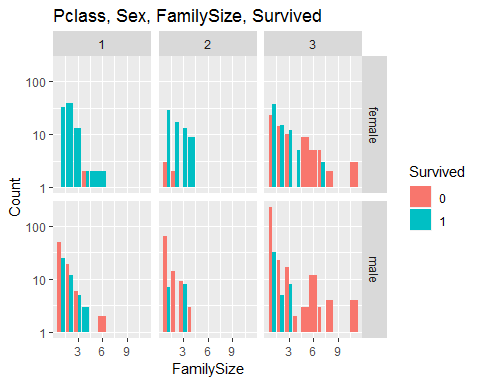
ggplot(train\_data, aes(x = FamilySize, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 0),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) + facet\_grid(.~Embarked) +  
 labs(x = 'FamilySize', y = 'percent', fill = 'Survived', title = 'FamilySize, Embarked and Survived')+ theme\_grey()

 Pclass, Sex, Embarked, Survived

ggplot(train\_data, aes(x = Embarked, fill = Survived)) +  
 geom\_bar(position = "dodge") +  
 facet\_grid(Sex ~ Pclass) +  
 scale\_y\_log10() +  
labs(x = 'Embarked', y = 'Count', fill = 'Survived', title = 'Pclass, Sex, Embarked, Survived')



ggplot(train\_data, aes(x = FamilySize, fill = Survived)) +  
 geom\_bar(position = "dodge") +  
 facet\_grid(Sex ~ Pclass) +  
 scale\_y\_log10() +  
labs(x = 'FamilySize', y = 'Count', fill = 'Survived', title = 'Pclass, Sex, FamilySize, Survived')



Age, Fare, Pclass, Embarked, Survived, Sex

ggplot(train\_data, aes(Age, Fare, color = Survived, shape = Sex)) +  
 geom\_point() +  
 scale\_y\_log10() +  
 facet\_grid(Pclass ~ Embarked)

## Warning: Transformation introduced infinite values in continuous y-axis

