Titanic data Bivariate 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'  
colSums(train\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 NA   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 687 0

colSums(is.na(train\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 177   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 0 0

1.Pclass Vs survived

table(train\_data$Pclass, train\_data$Survived)

##   
## 0 1  
## 1 80 136  
## 2 97 87  
## 3 372 119

prop.table(table(train\_data$Pclass, train\_data$Survived))

##   
## 0 1  
## 1 0.08978676 0.15263749  
## 2 0.10886644 0.09764310  
## 3 0.41750842 0.13355780

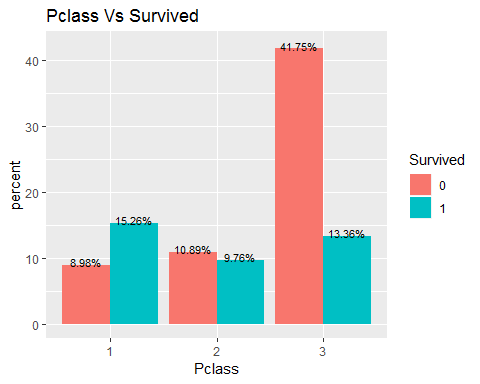
#install.packages("pspearman")  
library(pspearman)  
library(ggplot2)  
spearman.test(train\_data$Survived, train\_data$Pclass)$estimate

## Warning in spearman.test(train\_data$Survived, train\_data$Pclass): Cannot  
## compute exact p-values with ties

## rho   
## -0.3396679

There is no strong correlation between Pclass and Survived Visualize Pclass 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) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Survived', title = 'Pclass Vs Survived')

 Here we can see that majority of first class passengers survived, and most people in 3rd class died. 2.Sex Vs Survived

table(train\_data$Sex, train\_data$Survived)

##   
## 0 1  
## female 81 233  
## male 468 109

prop.table(table(train\_data$Sex, train\_data$Survived))

##   
## 0 1  
## female 0.09090909 0.26150393  
## male 0.52525253 0.12233446

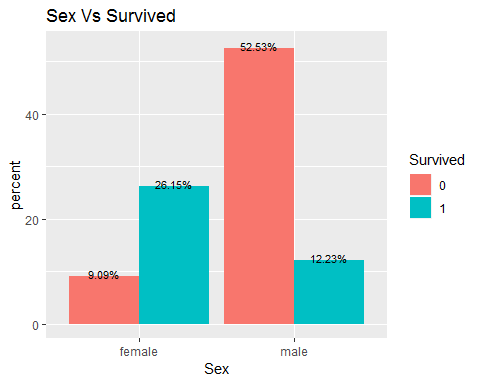
spearman.test(train\_data$Sex, train\_data$Survived)$estimate

## Warning in spearman.test(train\_data$Sex, train\_data$Survived): Cannot  
## compute exact p-values with ties

## rho   
## -0.5433514

There is no strong correlation between Sex and Survived Visualize Sex and Survived

ggplot(train\_data, aes(x = Sex, 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) +  
 labs(x = 'Sex', y = 'percent', fill = 'Survived', title = 'Sex Vs Survived')

 Majority of females were Survived and more males were died. 3.Embarked Vs survived

table(train\_data$Embarked, train\_data$Survived)

##   
## 0 1  
## 0 0  
## C 75 93  
## Q 47 30  
## S 427 219

prop.table(table(train\_data$Embarked, train\_data$Survived))

##   
## 0 1  
## 0.00000000 0.00000000  
## C 0.08417508 0.10437710  
## Q 0.05274972 0.03367003  
## S 0.47923681 0.24579125

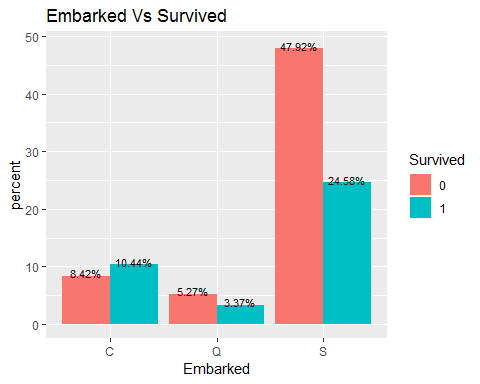
#install.packages("pspearman")  
library(pspearman)  
spearman.test(train\_data$Survived, train\_data$Embarked)$estimate

## Warning in spearman.test(train\_data$Survived, train\_data$Embarked): Cannot  
## compute exact p-values with ties

## rho   
## -0.1601964

There is no strong correlation between Embarked and Survived Visualize Embarked 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) +  
 labs(x = 'Embarked', y = 'percent', fill = 'Survived', title = 'Embarked Vs Survived')

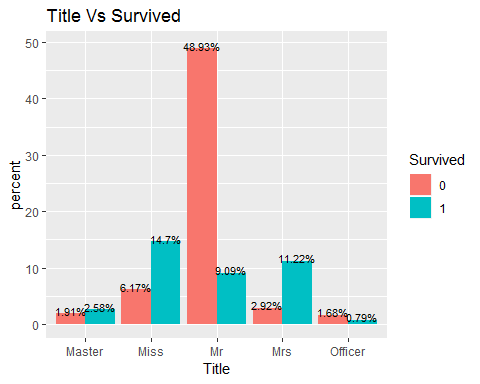
 Passengers who embarked at Southampton suvived more and died more too. 4.Name Vs Survived Make a variable named Title from Name.

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'

ggplot(train\_data, aes(x = Title, 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) +   
 labs(x = 'Title', y = 'percent', fill = 'Survived', title = 'Title Vs Survived')

 5.Age Vs Survived For an age variable containing a missing value, assign the mean age value for each title not containing a missing value.

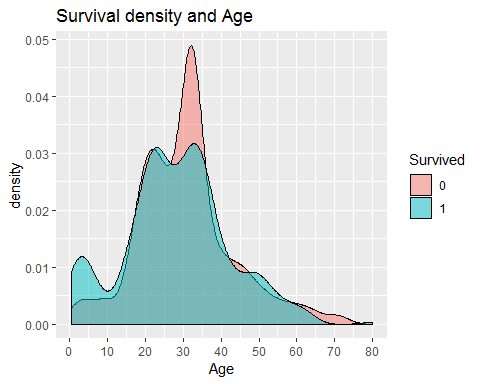
mean\_mr = mean(train\_data$Age[train\_data$Title == 'Mr' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Mr' & is.na(train\_data$Age)]=mean\_mr  
  
mean\_mrs = mean(train\_data$Age[train\_data$Title == 'Mrs' & !is.na(train\_data$Age)])  
train\_data$Age[train\_data$Title == 'Mrs' & is.na(train\_data$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

summary(train\_data$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.42 22.00 30.00 29.77 35.79 80.00

Visualize Age with Survived

ggplot(train\_data, aes(x = Age, fill = Survived)) +  
geom\_density(alpha=0.5, aes(fill=factor(Survived))) + labs(title="Survival density and Age") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

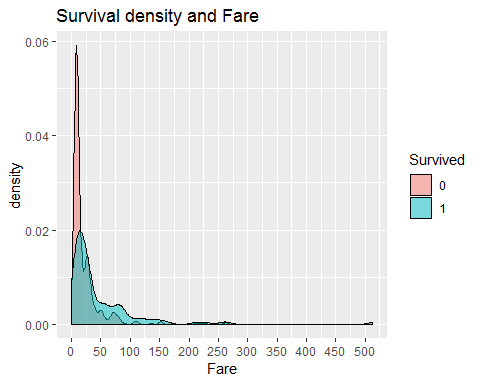
 Ages between 30 and 35 mostly survived 6.Fare Vs Survived

colSums(is.na(train\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 0 0   
## Title   
## 0

Visualize Fare with Survived

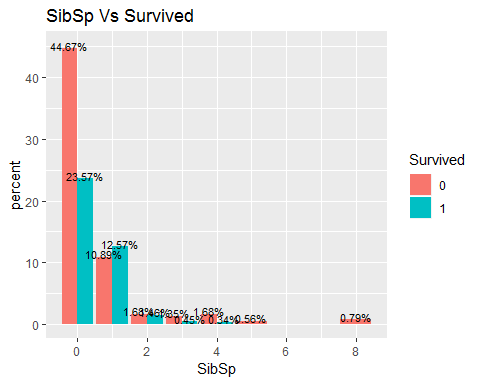
ggplot(train\_data, aes(x = Fare, fill = Survived)) +  
geom\_density(alpha=0.5, aes(fill=factor(Survived))) + labs(title="Survival density and Fare") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 7.Sibsp Vs Survived

prop.table(table(train\_data$SibSp, train\_data$Survived))

##   
## 0 1  
## 0 0.446689113 0.235690236  
## 1 0.108866442 0.125701459  
## 2 0.016835017 0.014590348  
## 3 0.013468013 0.004489338  
## 4 0.016835017 0.003367003  
## 5 0.005611672 0.000000000  
## 8 0.007856341 0.000000000

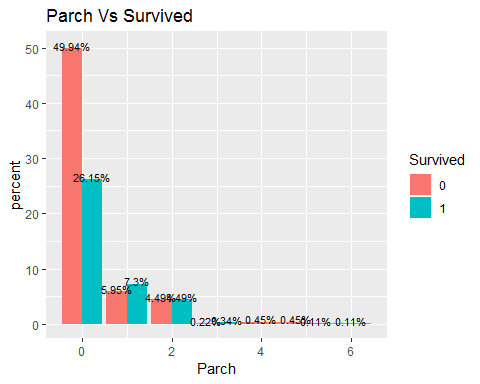
ggplot(train\_data, aes(x = SibSp, fill = Survived)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2),'%')),   
 stat = 'count',   
 position = position\_dodge(.9),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

 7.Parch Vs Survived

prop.table(table(train\_data$Parch, train\_data$Survived))

##   
## 0 1  
## 0 0.499438833 0.261503928  
## 1 0.059483726 0.072951740  
## 2 0.044893378 0.044893378  
## 3 0.002244669 0.003367003  
## 4 0.004489338 0.000000000  
## 5 0.004489338 0.001122334  
## 6 0.001122334 0.000000000

ggplot(train\_data, aes(x = Parch, 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) +   
 labs(x = 'Parch', y = 'percent', fill = 'Survived', title = 'Parch Vs Survived')

 we add SibSp and Parch named FamilySize.

train\_data$FamilySize <- train\_data$SibSp + train\_data$Parch +1

Have a look at this table

table(train\_data$FamilySize)

##   
## 1 2 3 4 5 6 7 8 11   
## 537 161 102 29 15 22 12 6 7

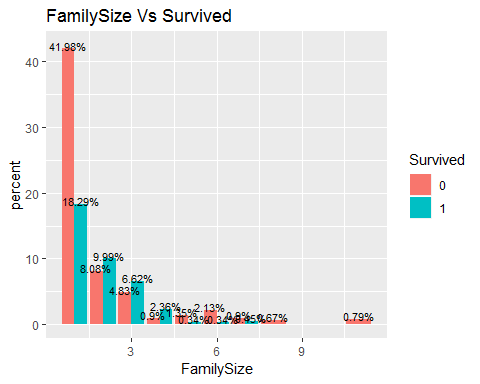
Let us look at the proportion

prop.table(table(train\_data$FamilySize, train\_data$Survived))

##   
## 0 1  
## 1 0.419753086 0.182940516  
## 2 0.080808081 0.099887767  
## 3 0.048260382 0.066217733  
## 4 0.008978676 0.023569024  
## 5 0.013468013 0.003367003  
## 6 0.021324355 0.003367003  
## 7 0.008978676 0.004489338  
## 8 0.006734007 0.000000000  
## 11 0.007856341 0.000000000

Visualize this data

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, 2),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +  
 labs(x = 'FamilySize', y = 'percent', fill = 'Survived', title ='FamilySize Vs Survived')

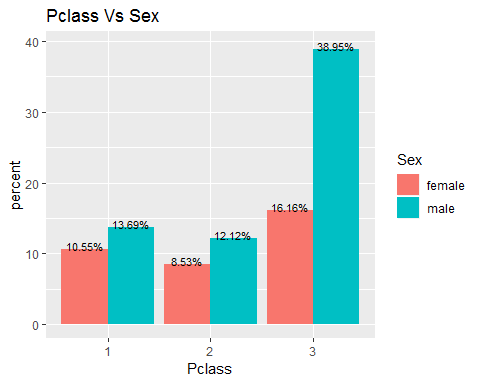
 This shows that families with a family size bigger than 2 but less than 5 have a more than 50% to survive, in contrast to families with 1 member or more than 5 members. Pclass and sex

prop.table(table(train\_data$Pclass, train\_data$Sex))

##   
## female male  
## 1 0.10549944 0.13692480  
## 2 0.08529742 0.12121212  
## 3 0.16161616 0.38945006

visualize Pclass and Sex

ggplot(train\_data, aes(x = Pclass, fill = Sex)) +  
 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) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Sex', title = 'Pclass Vs Sex')

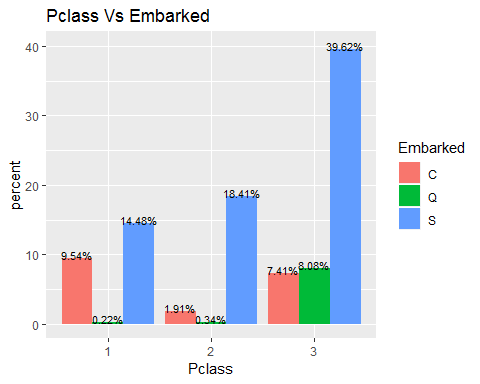
 More passengers travelled in class 3 and most of them were males, almost 39%. Pclass Vs Embarked

prop.table(table(train\_data$Pclass, train\_data$Embarked))

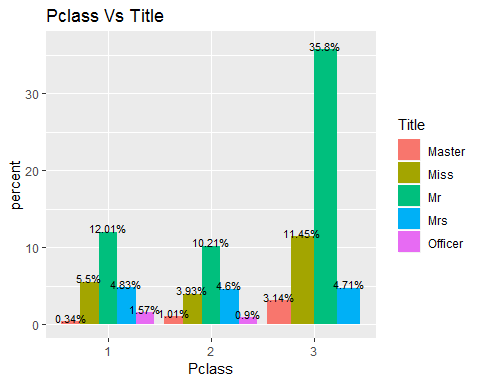
##   
## C Q S  
## 1 0.000000000 0.095398429 0.002244669 0.144781145  
## 2 0.000000000 0.019079686 0.003367003 0.184062851  
## 3 0.000000000 0.074074074 0.080808081 0.396184063

Visualize it

ggplot(train\_data, aes(x = Pclass, fill = Embarked)) +  
 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) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Embarked', title = 'Pclass Vs Embarked')

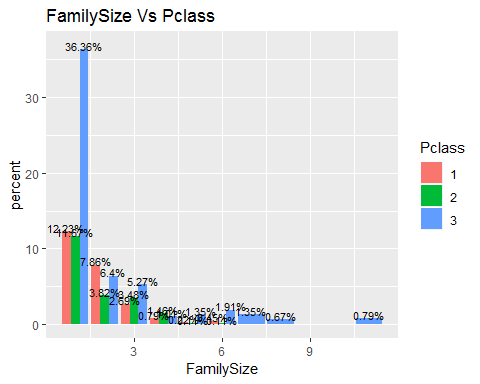
 Most of the passengers embarked to Southampton and most of them were travelled in class 3. Pclass Vs Title

ggplot(train\_data, aes(x = Pclass , fill = Title)) +  
 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) +  
 labs(x = 'Pclass', y = 'percent', fill = 'Title', title = 'Pclass Vs Title')



Pclass Vs FamilySize

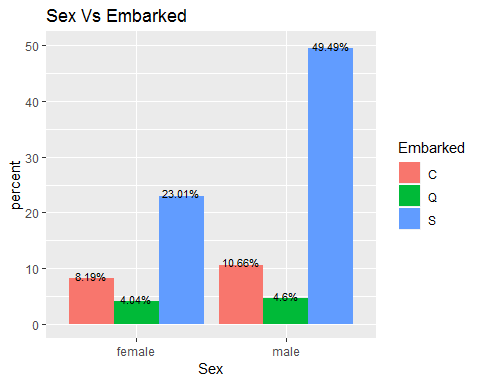
ggplot(train\_data, aes(x = FamilySize, fill = Pclass)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2),'%')),  
 stat = 'count',   
 position = position\_dodge(.9),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



Sex and Embarked

Visualization

ggplot(train\_data, aes(x = Sex, fill = Embarked)) +  
 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) +  
 labs(x = 'Sex', y = 'percent', fill = 'Embarked', title = 'Sex Vs Embarked')

 Most of the male and female are embarked to Southampton and only less people to Queenstown Sex Vs FamilySize

ggplot(train\_data, aes(x = FamilySize, fill = Sex)) +  
 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) +  
 labs(x = 'FamilySize', y = 'percent', fill = 'Sex', title = 'FamilySize Vs Sex ')

