Titanic data Univariate 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:

Read the titanic dataset which divided into train.csv and test.csv from the following website.

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

Have a look at the data set.

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 ...

Here we have 891 observations and 12 variables in the training data set. 1.Survived (whether the passenger died or survived) If Survived is 0 means passenger died and 1 means passenger Survived. This is the dependent variable.Let us check this variable first.We are converting this variable into factor.

train\_data$Survived = as.factor(train\_data$Survived)

Check the missing values for Survived

sum(is.na(train\_data$Survived) == TRUE)

## [1] 0

There are no missing values for Survived

Table of Passengers survived vs passengers not survived

table(train\_data$Survived)

##   
## 0 1   
## 549 342

Out of 891 passengers, 549 were died and 342 were survived

Proportion of Passengers survived vs passengers not survived

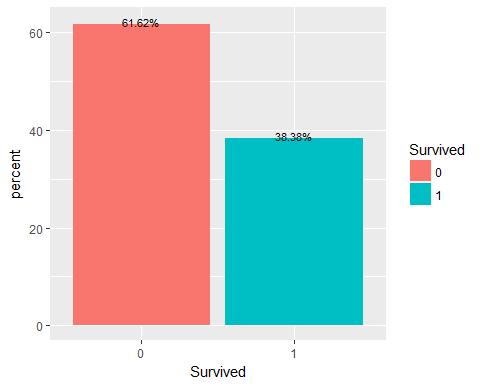
prop.table(table(train\_data$Survived))

##   
## 0 1   
## 0.6161616 0.3838384

So 62% of passengers died and 38% survived.

Data visualization of Survived

# install.packages("ggplot2")  
library(ggplot2)  
ggplot(train\_data, aes(x = Survived, 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 = 'Survived', y = 'percent', fill = 'Survived')



Next we combine both train and test data.

str(test\_data)

## 'data.frame': 418 obs. of 11 variables:  
## $ PassengerId: int 892 893 894 895 896 897 898 899 900 901 ...  
## $ Pclass : int 3 3 2 3 3 3 3 2 3 3 ...  
## $ Name : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 58 5 104 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...  
## $ Age : num 34.5 47 62 27 22 14 30 26 18 21 ...  
## $ SibSp : int 0 1 0 0 1 0 0 1 0 2 ...  
## $ Parch : int 0 0 0 0 1 0 0 1 0 0 ...  
## $ Ticket : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...  
## $ Fare : num 7.83 7 9.69 8.66 12.29 ...  
## $ Cabin : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Embarked : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...

test\_data$Survived <- NA  
Full\_data <- rbind(train\_data, test\_data)

Test data has 418 observations Check the Full data

str(Full\_data)

## 'data.frame': 1309 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : Factor w/ 1307 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/ 929 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/ 187 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 ...

The Full data has 1309 observations and 12 variables.

Look at the first few rows

#install.packages("dplyr")  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

glimpse(Full\_data)

## Observations: 1,309  
## Variables: 12  
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...  
## $ Survived <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,...  
## $ Pclass <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3,...  
## $ Name <fct> Braund, Mr. Owen Harris, Cumings, Mrs. John Bradle...  
## $ Sex <fct> male, female, female, female, male, male, male, ma...  
## $ Age <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, ...  
## $ SibSp <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4,...  
## $ Parch <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1,...  
## $ Ticket <fct> A/5 21171, PC 17599, STON/O2. 3101282, 113803, 373...  
## $ Fare <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, ...  
## $ Cabin <fct> , C85, , C123, , , E46, , , , G6, C103, , , , , , ...  
## $ Embarked <fct> S, C, S, S, S, Q, S, S, S, C, S, S, S, S, S, S, Q,...

Let us have the summary of the data

summary(Full\_data)

## PassengerId Survived Pclass   
## Min. : 1 0 :549 Min. :1.000   
## 1st Qu.: 328 1 :342 1st Qu.:2.000   
## Median : 655 NA's:418 Median :3.000   
## Mean : 655 Mean :2.295   
## 3rd Qu.: 982 3rd Qu.:3.000   
## Max. :1309 Max. :3.000   
##   
## Name Sex Age   
## Connolly, Miss. Kate : 2 female:466 Min. : 0.17   
## Kelly, Mr. James : 2 male :843 1st Qu.:21.00   
## Abbing, Mr. Anthony : 1 Median :28.00   
## Abbott, Mr. Rossmore Edward : 1 Mean :29.88   
## Abbott, Mrs. Stanton (Rosa Hunt): 1 3rd Qu.:39.00   
## Abelson, Mr. Samuel : 1 Max. :80.00   
## (Other) :1301 NA's :263   
## SibSp Parch Ticket Fare   
## Min. :0.0000 Min. :0.000 CA. 2343: 11 Min. : 0.000   
## 1st Qu.:0.0000 1st Qu.:0.000 1601 : 8 1st Qu.: 7.896   
## Median :0.0000 Median :0.000 CA 2144 : 8 Median : 14.454   
## Mean :0.4989 Mean :0.385 3101295 : 7 Mean : 33.295   
## 3rd Qu.:1.0000 3rd Qu.:0.000 347077 : 7 3rd Qu.: 31.275   
## Max. :8.0000 Max. :9.000 347082 : 7 Max. :512.329   
## (Other) :1261 NA's :1   
## Cabin Embarked  
## :1014 : 2   
## C23 C25 C27 : 6 C:270   
## B57 B59 B63 B66: 5 Q:123   
## G6 : 5 S:914   
## B96 B98 : 4   
## C22 C26 : 4   
## (Other) : 271

2.Pclass (passengers socio-economic status (SES) class 1=Upper 2=Middle 3=Lower) First we convert this variable into factor.

Full\_data$Pclass = as.factor(Full\_data$Pclass)

Check there is NA’s in the attributes.

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 263   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0

There is no NA’s in Pclass. Let us check for blank spaces

colSums(Full\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 NA 0 0 0 NA   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 NA 1014 2

There is no blank spaces Table of passengers in 3 classes

table(Full\_data$Pclass)

##   
## 1 2 3   
## 323 277 709

Proportion of passengers in 3 classes

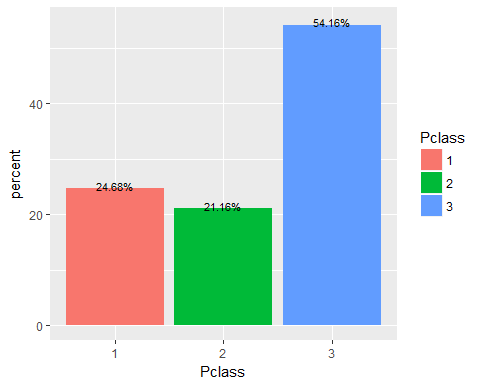
prop.table(table(Full\_data$Pclass))

##   
## 1 2 3   
## 0.2467532 0.2116119 0.5416348

From this we can see that most of the passengers were in the lower level(Pclass=3), more than 50%

Data visualization of Pclass

ggplot(Full\_data, aes(x = Pclass, 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 = 'Pclass', y = 'percent', fill = 'Pclass')

 3. Name (passengers name) Change this variable into character.

Full\_data$Name = as.character(Full\_data$Name)

Check for the missing values

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 263   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0

colSums(Full\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 NA 0 0 0 NA   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 NA 1014 2

There are no missing values for Name. The Name variable contains a Title for each person, So if we separate this from the name we will get a tidy data and we can analyse more accurate.So Make a variable named Title from Name.

Full\_data$Title <- sapply(Full\_data$Name, FUN=function(x) {strsplit(x, split='[,.]')[[1]][2]})  
Full\_data$Title <- sub(' ', '', Full\_data$Title)

Have a look at the Title table

table(Full\_data$Title)

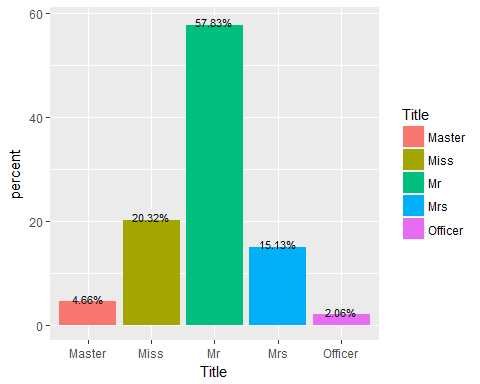
##   
## Capt Col Don Dona Dr   
## 1 4 1 1 8   
## Jonkheer Lady Major Master Miss   
## 1 1 2 61 260   
## Mlle Mme Mr Mrs Ms   
## 2 1 757 197 2   
## Rev Sir the Countess   
## 8 1 1

It is better to reduce the number of titles to create a better and more substantial Titles that can be used for prediction. Ms. is usually used for younger married women. I join this one with Miss. Mlle stands for Mademoiselle in French so join this category with Miss and also Dona and Lady i put it with Miss. I assume that Mme stands for Madame, and I will join Madame with Mrs. I put Lady is also with Mrs. For the titles Captis Captain, col is colonel, Major, Sir, the Countess, Don, Dr, Jonkheer, Rev, I will create one new category Officer.

Full\_data$Title[Full\_data$Title %in% c('Mlle', 'Ms', 'Dona', 'Lady')] <- 'Miss'  
Full\_data$Title[Full\_data$Title %in% c('Mme')] <- 'Mrs'  
Full\_data$Title[Full\_data$Title %in% c('Capt', 'Col', 'Don', 'Dr', 'Jonkheer', 'Rev', 'the Countess', 'Major', 'Sir')] <- 'Officer'

Now visualize the new Title variable

ggplot(Full\_data, aes(x = Title, 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 = 'Title', y = 'percent', fill = 'Title')

 4.Sex (Female or Male) There are no missing values in Sex. Have a look at the Sex table.

table(Full\_data$Sex)

##   
## female male   
## 466 843

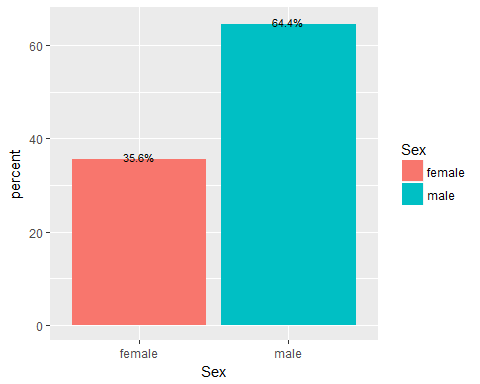
Check the proportion of male and female

prop.table(table(Full\_data$Sex))

##   
## female male   
## 0.3559969 0.6440031

So 64% male and 36% of female were travelled. Now visualize it

ggplot(Full\_data, aes(x = Sex, 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 = 'Sex', y = 'percent', fill = 'Sex')

 5.Age ( Passenger’s Age) Check for the missing values

sum(is.na(Full\_data$Age) == TRUE)

## [1] 263

For an age variable containing a missing value, assign the mean age value for each title not containing a missing value.So first compute the mean age for Mr,Mrs,Miss,Master and Officer and assign this mean to age value that corresponds to. First find the mean of all Name whose Title is Mr.

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

Now same procedure with Mrs

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

Now with Master

mean\_master = mean(Full\_data$Age[Full\_data$Title == 'Master' & !is.na(Full\_data$Age)])  
Full\_data$Age[Full\_data$Title == 'Master' & is.na(Full\_data$Age)]=mean\_master

Now with Miss

mean\_miss = mean(Full\_data$Age[Full\_data$Title == 'Miss' & !is.na(Full\_data$Age)])  
Full\_data$Age[Full\_data$Title == 'Miss' & is.na(Full\_data$Age)]=mean\_miss

Now with Officer

mean\_officer = mean(Full\_data$Age[Full\_data$Title == 'Officer' & !is.na(Full\_data$Age)])  
Full\_data$Age[Full\_data$Title == 'Officer' & is.na(Full\_data$Age)]=mean\_officer

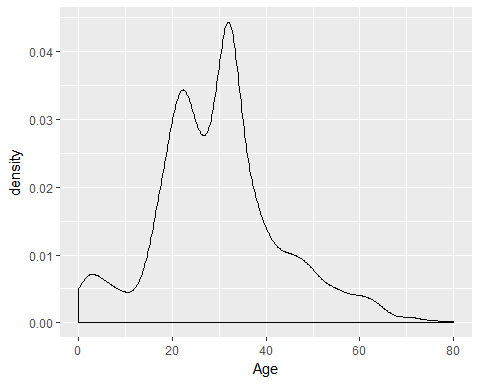
Have a look at the summary of Age

summary(Full\_data$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.17 22.00 30.00 29.90 36.00 80.00

Now visualize the data

ggplot(Full\_data, aes(x=Age)) +  
 geom\_density(aes(fill = Age), alpha = 0.5)

 Here we can see the mean age of passengers were 29.90 and most people who travelled were betwee 30 and 35.Highest Age is 80.

1. SibSp (Number of siblings / spouses aboard the Titanic) Check there is missing values

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0   
## Title   
## 0

There are no missing values. See the table of SibSp

table(Full\_data$SibSp)

##   
## 0 1 2 3 4 5 8   
## 891 319 42 20 22 6 9

See the proportion

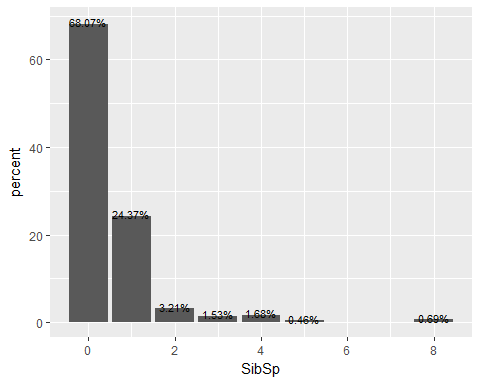
prop.table(table(Full\_data$SibSp))

##   
## 0 1 2 3 4 5   
## 0.680672269 0.243697479 0.032085561 0.015278839 0.016806723 0.004583652   
## 8   
## 0.006875477

Here we can see that 68% without siblings or spouses.

Visualize this data

ggplot(Full\_data, aes(x = SibSp, fill = SibSp)) +  
 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 = 'SibSp')

 7.Parch (Number of parents / children aboard the Titanic) Check for the missing values

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0   
## Title   
## 0

There are no missing values Check the table for Parch

table(Full\_data$Parch)

##   
## 0 1 2 3 4 5 6 9   
## 1002 170 113 8 6 6 2 2

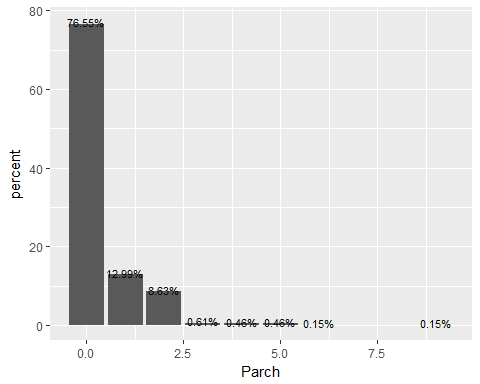
Check with proportion

prop.table(table(Full\_data$Parch))

##   
## 0 1 2 3 4 5   
## 0.765469824 0.129870130 0.086325439 0.006111536 0.004583652 0.004583652   
## 6 9   
## 0.001527884 0.001527884

77% passengers were without parents or children Now visualize this data

ggplot(Full\_data, aes(x = Parch, fill = Parch)) +  
 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 = 'Parch')

 8.Ticket(Passenger’s ticket number) First we change it to character

Full\_data$Ticket <- as.character(Full\_data$Ticket)

Check with missing data

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0   
## Title   
## 0

colSums(Full\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 NA 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 NA 1014 2   
## Title   
## 0

There are no missing values.

actual\_mode <- table(Full\_data$Ticket)  
names(actual\_mode)[actual\_mode == max(actual\_mode)]

## [1] "CA. 2343"

The ticket number CA.2343 is the most occured one 11 times.

1. Fare(Passenger fare) Check the missing values

colSums(is.na(Full\_data))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 0 0   
## Title   
## 0

One missing value for the fare.

Let’s check the summaryof fare

summary(Full\_data$Fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 7.896 14.454 33.295 31.275 512.329 1

Maximun fare is 512.329 and mean of fare is 33.295 Let us find the details of fare with NA

#install.packages("kableExtra")  
library(kableExtra)  
kable(Full\_data[which(is.na(Full\_data$Fare)),c('Title', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked')])

Title

Survived

Pclass

Age

SibSp

Parch

Ticket

Fare

Embarked

1044

Mr

NA

3

60.5

0

0

3701

NA

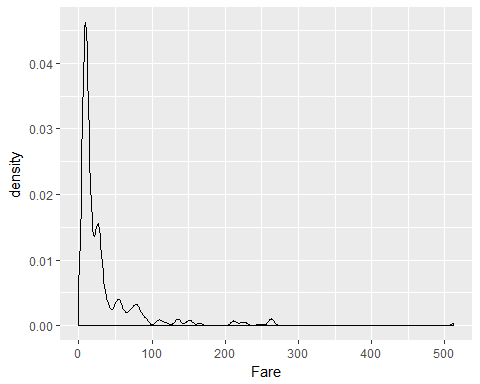
S

From this we can understand that the missing fare is a passenger who embarked to Southampton and ttravelled in class 3 by alone. So i took the mean fare of such class.

Full\_data$Fare[is.na(Full\_data$Fare)] = mean(Full\_data$Fare[which(Full\_data$Embarked == 'S' & Full\_data$Pclass == '3' & Full\_data$SibSp == 0 & Full\_data$Parch == 0)], na.rm = TRUE)

Now visualize the Fare variable.

ggplot(Full\_data, aes(x=Fare)) +  
 geom\_density(aes(fill = Fare), alpha = 0.5)



10.Cabin (Passenger’s cabin number) Check the missing values

colSums(Full\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 NA 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 1014 2   
## Title   
## 0

Since there are too much information is missing, i am ignoring this variable. 11. Embarked (Port of Embarkation, C = Cherbourg, Q = Queenstown, S = Southampton) Check the missing values first

colSums(Full\_data =="")

## PassengerId Survived Pclass Name Sex Age   
## 0 NA 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 1014 2   
## Title   
## 0

There are 2 missing values Check the table

table(Full\_data$Embarked)

##   
## C Q S   
## 2 270 123 914

Most passengers embarked at Southampton. Let’s check that 2 rows who missed the embarked

kable(Full\_data[which(Full\_data$Embarked ==''),c('Title', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked')])

Title

Survived

Pclass

Age

SibSp

Parch

Ticket

Fare

Embarked

62

Miss

1

1

38

0

0

113572

80

830

Mrs

1

1

62

0

0

113572

80

From this data, we can understand that the missing Embarked passengers are females travelled alone in class 1. Have a look at the rows which has the same characters.

table(Full\_data[which(Full\_data$Pclass == '1' & Full\_data$SibSp == 0 & Full\_data$Parch == 0, Full\_data$Sex == 'female'), c('Embarked')])

##   
## C Q S   
## 2 65 0 93

Here most people in this catagory embarked at Southampton, so i put S to the missing place.

Full\_data$Embarked[Full\_data$Embarked == ""] = 'S'

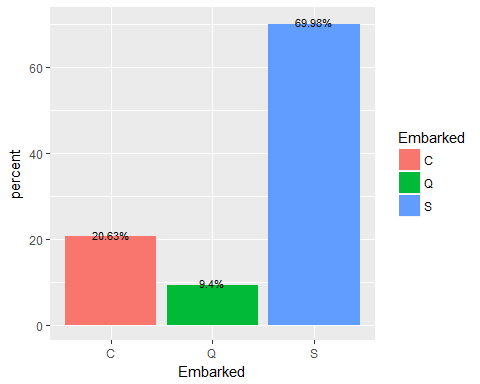
Now look at the proportion

prop.table(table(Full\_data$Embarked))

##   
## C Q S   
## 0.00000000 0.20626432 0.09396486 0.69977082

So most people embarked to Southampton, which is 70%. Let us visualize it

ggplot(Full\_data, aes(x = Embarked, 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 = 'Embarked', y = 'percent', fill = 'Embarked')

 Next if we add SibSp and Parch, we will get more information about family. So we are going to make another column named FamilySize.

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

Have a look at this table

table(Full\_data$FamilySize)

##   
## 1 2 3 4 5 6 7 8 11   
## 790 235 159 43 22 25 16 8 11

Let us look at the proportion

prop.table(table(Full\_data$FamilySize))

##   
## 1 2 3 4 5 6   
## 0.603514133 0.179526356 0.121466769 0.032849503 0.016806723 0.019098549   
## 7 8 11   
## 0.012223071 0.006111536 0.008403361

So 60% passengers were travelled alone. Visualize this data

ggplot(Full\_data, aes(x = FamilySize, fill = FamilySize)) +  
 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 = 'FamilySize')

