Assignment - 1

Session 6 – Visualization

and Plotting

1. Import the Titanic Dataset from the link Titanic Data Set.

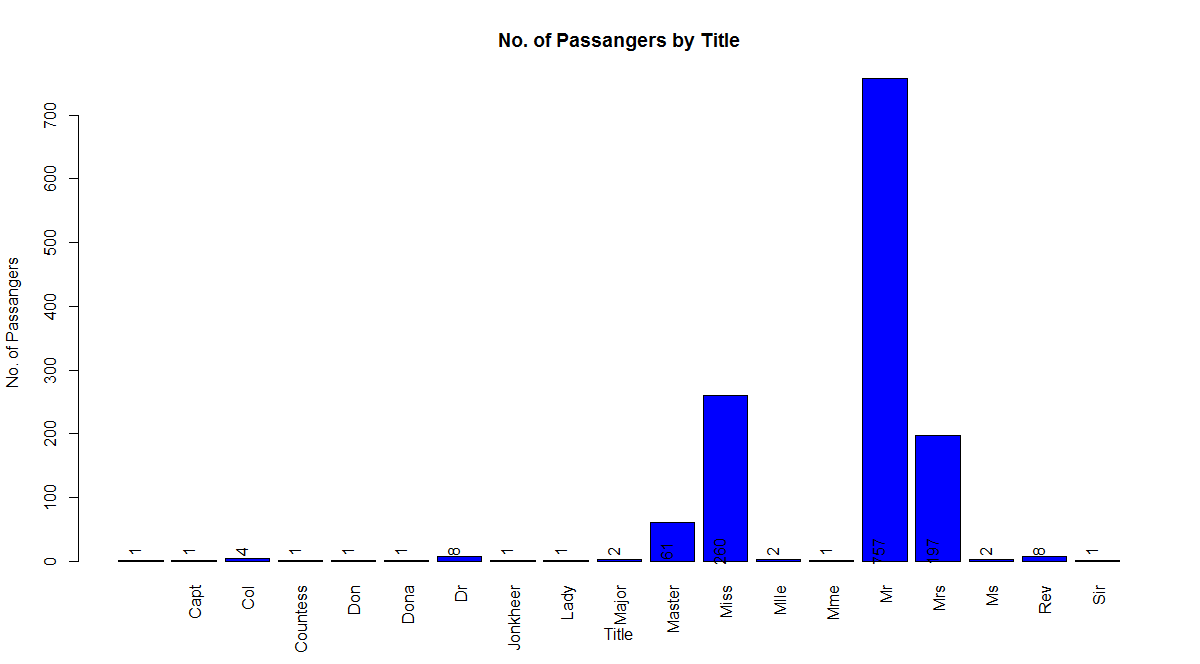
Perform the following:

a. Preprocess the passenger names to come up with a list of titles that represent families

and represent using appropriate visualization graph.

Ans:

|  |
| --- |
| > str(titanic)  'data.frame': 1310 obs. of 15 variables:  $ pclass : int 1 1 1 1 1 1 1 1 1 1 ...  $ survived : int 1 1 0 0 0 1 1 0 1 0 ...  $ Family : Factor w/ 876 levels "","Abbing","Abbott",..: 17 18 18 18 18 22 26 26 29 32 ...  $ Name : Factor w/ 1143 levels ""," Capt. Edward Gifford",..: 150 53 186 571 1031 545 224 872 993 797 ...  $ sex : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...  $ age : num 29 0.917 2 30 25 ...  $ sibsp : int 0 1 1 1 1 0 1 0 2 0 ...  $ parch : int 0 2 2 2 2 0 0 0 0 0 ...  $ ticket : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...  $ fare : num 211 152 152 152 152 ...  $ cabin : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...  $ boat : Factor w/ 28 levels "","1","10","11",..: 13 4 1 1 1 14 3 1 28 1 ...  $ body : int NA NA NA 135 NA NA NA NA NA 22 ...  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...  > View(titanic)  > psych::describe(titanic)  vars n mean sd median trimmed mad min max range skew kurtosis  pclass 1 1309 2.29 0.84 3.00 2.37 0.00 1.00 3.00 2.00 -0.60 -1.32  survived 2 1309 0.38 0.49 0.00 0.35 0.00 0.00 1.00 1.00 0.49 -1.77  Family\* 3 1310 429.34 255.47 412.00 427.46 327.65 1.00 876.00 875.00 0.07 -1.22  Name\* 4 1310 571.17 320.23 584.00 571.12 405.49 1.00 1143.00 1142.00 -0.02 -1.14  sex\* 5 1310 2.64 0.48 3.00 2.68 0.00 1.00 3.00 2.00 -0.62 -1.57  age 6 1046 29.88 14.41 28.00 29.39 11.86 0.17 80.00 79.83 0.41 0.13  sibsp 7 1309 0.50 1.04 0.00 0.27 0.00 0.00 8.00 8.00 3.84 19.93  parch 8 1309 0.39 0.87 0.00 0.18 0.00 0.00 9.00 9.00 3.66 21.42  ticket\* 9 1310 465.25 278.23 461.00 465.88 379.55 1.00 930.00 929.00 -0.01 -1.33  fare 10 1308 33.30 51.76 14.45 21.57 10.24 0.00 512.33 512.33 4.36 26.87  cabin\* 11 1310 23.02 47.81 1.00 10.08 0.00 1.00 187.00 186.00 2.10 3.15  embarked\* 12 1310 3.49 0.82 4.00 3.61 0.00 1.00 4.00 3.00 -1.13 -0.49  boat\* 13 1310 5.97 8.00 1.00 4.27 0.00 1.00 28.00 27.00 1.43 0.65  body 14 121 160.81 97.70 155.00 160.34 130.47 1.00 328.00 327.00 0.09 -1.28  home.dest\* 15 1310 113.07 124.55 53.50 98.79 77.84 1.00 370.00 369.00 0.59 -1.19  se  pclass 0.02  survived 0.01  Family\* 7.06  Name\* 8.85  sex\* 0.01  age 0.45  sibsp 0.03  parch 0.02  ticket\* 7.69  fare 1.43  cabin\* 1.32  embarked\* 0.02  boat\* 0.22  body 8.88  home.dest\* 3.44  > colnames(titanic) <- c("PassengerId","Survived","Pclass","Name",  + "Sex","Age","SibSp","Parch","Ticket","Fare",  + "Cabin","Embarked")  > titanic = titanic [,-13]  > titanic$Pclass <- as.factor(titanic$Pclass)  > titanic$SibSp <- as.factor(titanic$SibSp)  > titanic$Parch <- as.factor(titanic$Parch)  > titanic$Survived <- as.factor(titanic$Survived)  > str(titanic)  'data.frame': 1310 obs. of 14 variables:  $ PassengerId: int 1 1 1 1 1 1 1 1 1 1 ...  $ Survived : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...  $ Pclass : Factor w/ 876 levels "","Abbing","Abbott",..: 17 18 18 18 18 22 26 26 29 32 ...  $ Name : Factor w/ 1143 levels ""," Capt. Edward Gifford",..: 150 53 186 571 1031 545 224 872 993 797 ...  $ Sex : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...  $ Age : num 29 0.917 2 30 25 ...  $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 1 2 2 2 2 1 2 1 3 1 ...  $ Parch : Factor w/ 8 levels "0","1","2","3",..: 1 3 3 3 3 1 1 1 1 1 ...  $ Ticket : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...  $ Fare : num 211 152 152 152 152 ...  $ Cabin : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...  $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...  $ NA : int NA NA NA 135 NA NA NA NA NA 22 ...  $ NA.1 : Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...  > titanic$Name <- as.character(titanic$Name)  > titanic$SubTitle <- gsub("\\..\*", "", titanic$Name)  > titanic$Title <- gsub(".\*\\ ", "", titanic$SubTitle)  > table(titanic$Title) # Count of Titles  Capt Col Countess Don Dona Dr Jonkheer Lady Major  1 1 4 1 1 1 8 1 1 2  Master Miss Mlle Mme Mr Mrs Ms Rev Sir  61 260 2 1 757 197 2 8 1  > Title <- barplot(table(titanic$Title),  + main = "No. of Passangers by Title", xlab = "Title",  + ylab = "No. of Passangers", col = "Blue", las =3)  > text(Title, 0,table(titanic$Title), pos = 3, srt = 90) |
|  |
| |  | | --- | | > | |



b. Represent the proportion of people survived from the family size using a graph.

> x <- table(titanic$Survived, titanic$Title) # table for survived and died

> x # 0 for survived and 1 for died

Capt Col Countess Don Dona Dr Jonkheer Lady Major Master Miss Mlle Mme Mr

0 0 1 2 0 1 0 4 1 0 1 30 84 0 0 634

1 0 0 2 1 0 1 4 0 1 1 31 176 2 1 123

Mrs Ms Rev Sir

0 42 1 8 0

1 155 1 0 1

> p <- x[1,] # number of passengers survived

> p

Capt Col Countess Don Dona Dr Jonkheer Lady

0 1 2 0 1 0 4 1 0

Major Master Miss Mlle Mme Mr Mrs Ms Rev

1 30 84 0 0 634 42 1 8

Sir

0

> prop <- round(p\*100/sum(p),1) # proportion of passangers survived

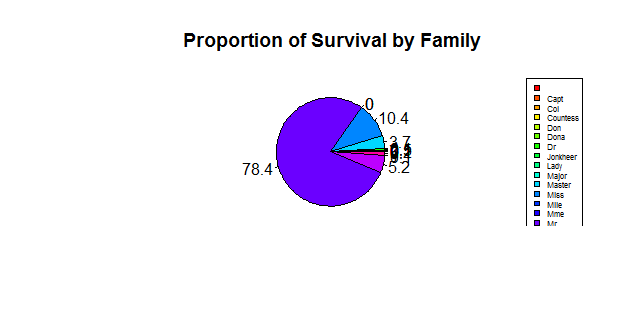
> legend("topright", names(p), cex= 0.5, fill = rainbow(length(p)))

> legend("topright", names(prop), cex= 0.5, fill = rainbow(length(prop)))

> pie(prop, labels = prop, main = " Proportion of Survival by Family",

+ col = rainbow(length(prop)), cex = 1)

> legend("topright", names(prop), cex= 0.5, fill = rainbow(length(prop)))



> barplot(prop, # for percentage of passangers

+ main = "No. of Passangers by Title", xlab = "Title",

+ ylab = "No. of Passangers", col = c("Blue","Red"),

+ legend = rownames(prop), ylim=c(0, 100), las = 3)

> barplot(prop, # for percentage of passangers

+ main = "No. of Passangers by Title", xlab = "Title",

+ ylab = "No. of Passangers", col = c("Blue","Red"),

+ legend = rownames(prop), ylim=c(0, 100), las = 3)

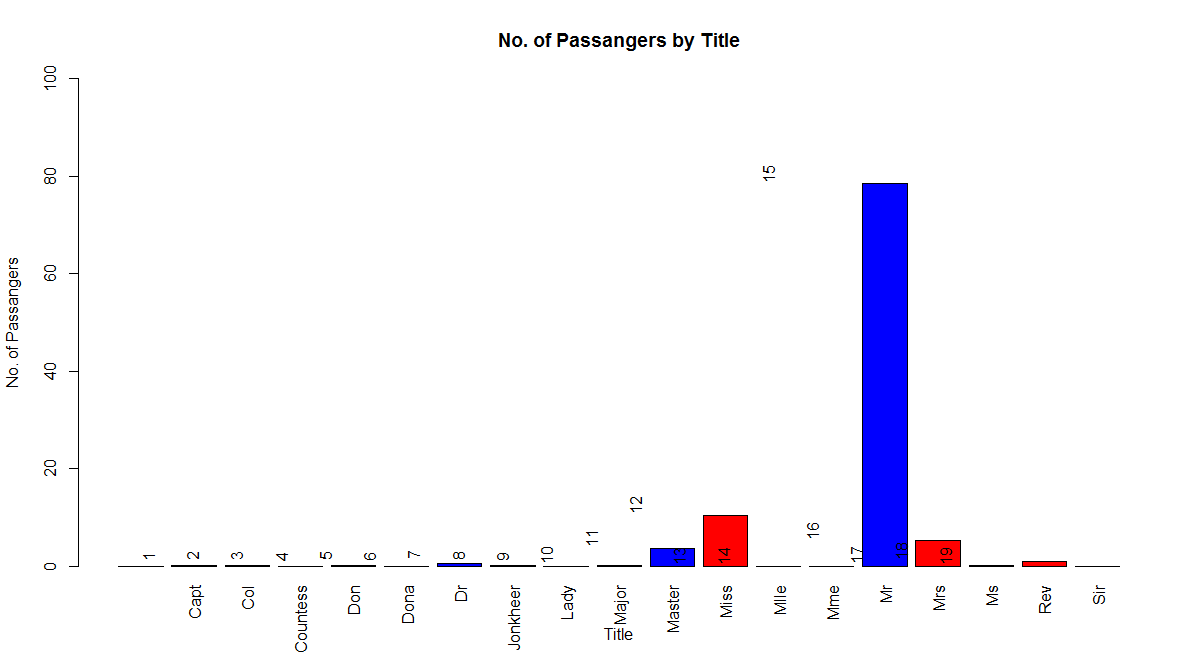
> barplot(prop, # for percentage of passangers

+ main = "No. of Passangers by Title", xlab = "Title",

+ ylab = "No. of Passangers", col = c("Blue","Red"),

+ legend = rownames(prop), ylim=c(0, 100), las = 3)

> text(prop, pos = 3, srt = 90)



c. Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputation.

Ans:

|  |
| --- |
| > library(readr)  > titanic <- within(titanic,  + {  + Age <- NA  + Age[Age>=0 & Age<=25] <- "Low"  + Age[Age>=26 & Age<=40] <- "Middle"  + Age[Age>=41] <- "High"  + })  > head(titanic)  PassengerId Survived Pclass Name Sex Age SibSp  1 1 1 Allen Miss. Elisabeth Walton female 29.0000 0  2 1 1 Allison Master. Hudson Trevor male 0.9167 1  3 1 0 Allison Miss. Helen Loraine female 2.0000 1  4 1 0 Allison Mr. Hudson Joshua Creighton male 30.0000 1  5 1 0 Allison Mrs. Hudson J C (Bessie Waldo Daniels) female 25.0000 1  6 1 1 Anderson Mr. Harry male 48.0000 0  Parch Ticket Fare Cabin Embarked NA NA.1 SubTitle Title  1 0 24160 211.3375 B5 S NA St Louis, MO Miss Miss  2 2 113781 151.5500 C22 C26 S NA Montreal, PQ / Chesterville, ON Master Master  3 2 113781 151.5500 C22 C26 S NA Montreal, PQ / Chesterville, ON Miss Miss  4 2 113781 151.5500 C22 C26 S 135 Montreal, PQ / Chesterville, ON Mr Mr  5 2 113781 151.5500 C22 C26 S NA Montreal, PQ / Chesterville, ON Mrs Mrs  6 0 19952 26.5500 E12 S NA New York, NY Mr Mr  > count <- table(titanic$Age, titanic$Title)  > count    Capt Col Countess Don Dona Dr Jonkheer Lady Major Master Miss Mlle Mme Mr Mrs Ms  0.1667 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0  0.3333 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  0.4167 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  0.6667 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  0.75 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0  0.8333 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0  0.9167 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0  1 0 0 0 0 0 0 0 0 0 0 5 5 0 0 0 0 0  2 0 0 0 0 0 0 0 0 0 0 5 7 0 0 0 0 0  3 0 0 0 0 0 0 0 0 0 0 4 3 0 0 0 0 0  4 0 0 0 0 0 0 0 0 0 0 5 5 0 0 0 0 0  5 0 0 0 0 0 0 0 0 0 0 1 4 0 0 0 0 0  6 0 0 0 0 0 0 0 0 0 0 4 2 0 0 0 0 0  7 0 0 0 0 0 0 0 0 0 0 3 1 0 0 0 0 0  8 0 0 0 0 0 0 0 0 0 0 3 3 0 0 0 0 0  9 0 0 0 0 0 0 0 0 0 0 5 5 0 0 0 0 0  10 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0 0  11 0 0 0 0 0 0 0 0 0 0 2 1 0 0 1 0 0  11.5 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  12 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0  13 0 0 0 0 0 0 0 0 0 0 3 2 0 0 0 0 0  14 0 0 0 0 0 0 0 0 0 0 0 3 0 0 4 1 0  14.5 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0  15 0 0 0 0 0 0 0 0 0 0 0 4 0 0 1 1 0  16 0 0 0 0 0 0 0 0 0 0 0 7 0 0 11 1 0  17 0 0 0 0 0 0 0 0 0 0 0 6 0 0 12 2 0  18 0 0 0 0 0 0 0 0 0 0 0 14 0 0 19 6 0  18.5 0 0 0 0 0 0 0 0 0 0 0 1 0 0 2 0 0  19 0 0 0 0 0 0 0 0 0 0 0 6 0 0 19 4 0  20 0 0 0 0 0 0 0 0 0 0 0 6 0 0 17 0 0  20.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  21 0 0 0 0 0 0 0 0 0 0 0 10 0 0 31 0 0  22 0 0 0 0 0 0 0 0 0 0 0 15 0 0 21 7 0  22.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  23 0 0 0 0 0 0 1 0 0 0 0 7 0 0 15 3 0  23.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  24 0 0 0 0 0 0 0 0 0 0 0 9 2 1 26 9 0  24.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  25 0 0 0 0 0 0 0 0 0 0 0 3 0 0 28 3 0  26 0 0 0 0 0 0 0 0 0 0 0 4 0 0 21 5 0  26.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  27 0 0 0 0 0 0 0 0 0 0 0 5 0 0 19 5 0  28 0 0 0 0 0 0 0 0 0 0 0 4 0 0 22 4 1  28.5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0  29 0 0 0 0 0 0 0 0 0 0 0 3 0 0 18 9 0  30 0 0 0 0 0 0 0 0 0 0 0 13 0 0 22 4 0  30.5 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0  31 0 0 0 0 0 0 0 0 0 0 0 5 0 0 13 5 0  32 0 0 0 0 0 0 1 0 0 0 0 1 0 0 20 2 0  32.5 0 0 0 0 0 0 0 0 0 0 0 1 0 0 3 0 0  33 0 0 0 1 0 0 0 0 0 0 0 3 0 0 12 5 0  34 0 0 0 0 0 0 0 0 0 0 0 1 0 0 12 3 0    Rev Sir  0.1667 0 0  0.3333 0 0  0.4167 0 0  0.6667 0 0  0.75 0 0  0.8333 0 0  0.9167 0 0  1 0 0  2 0 0  3 0 0  4 0 0  5 0 0  6 0 0  7 0 0  8 0 0  9 0 0  10 0 0  11 0 0  11.5 0 0  12 0 0  13 0 0  14 0 0  14.5 0 0  15 0 0  16 0 0  17 0 0  18 0 0  18.5 0 0  19 0 0  20 0 0  20.5 0 0  21 0 0  22 0 0  22.5 0 0  23 0 0  23.5 0 0  24 0 0  24.5 0 0  25 0 0  26 0 0  26.5 0 0  27 1 0  28 1 0  28.5 0 0  29 0 0  30 1 0  30.5 0 0  31 0 0  32 0 0  32.5 0 0  33 0 0  34 0 0  [ reached getOption("max.print") -- omitted 46 rows ]  > library(ggplot2)  > p <- ggplot(data = titanic,  + mapping = aes(Title, fill = Age))  > p + geom\_bar (position = "stack") + theme (axis.text.x = element\_text(angle = 90)) + labs(title = "Counts of Title with Age Groups")  > library(mice)  > library(dplyr)  > library(dplyr)  > str(titanic)  'data.frame': 1310 obs. of 16 variables:  $ PassengerId: int 1 1 1 1 1 1 1 1 1 1 ...  $ Survived : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...  $ Pclass : Factor w/ 876 levels "","Abbing","Abbott",..: 17 18 18 18 18 22 26 26 29 32 ...  $ Name : chr " Miss. Elisabeth Walton" " Master. Hudson Trevor" " Miss. Helen Loraine" " Mr. Hudson Joshua Creighton" ...  $ Sex : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...  $ Age : num 29 0.917 2 30 25 ...  $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 1 2 2 2 2 1 2 1 3 1 ...  $ Parch : Factor w/ 8 levels "0","1","2","3",..: 1 3 3 3 3 1 1 1 1 1 ...  $ Ticket : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...  $ Fare : num 211 152 152 152 152 ...  $ Cabin : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...  $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...  $ NA : int NA NA NA 135 NA NA NA NA NA 22 ...  $ NA.1 : Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...  $ SubTitle : chr " Miss" " Master" " Miss" " Mr" ...  $ Title : chr "Miss" "Master" "Miss" "Mr" ...  > dat <- titanic[,-13]  > str(dat)  'data.frame': 1310 obs. of 15 variables:  $ PassengerId: int 1 1 1 1 1 1 1 1 1 1 ...  $ Survived : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...  $ Pclass : Factor w/ 876 levels "","Abbing","Abbott",..: 17 18 18 18 18 22 26 26 29 32 ...  $ Name : chr " Miss. Elisabeth Walton" " Master. Hudson Trevor" " Miss. Helen Loraine" " Mr. Hudson Joshua Creighton" ...  $ Sex : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...  $ Age : num 29 0.917 2 30 25 ...  $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 1 2 2 2 2 1 2 1 3 1 ...  $ Parch : Factor w/ 8 levels "0","1","2","3",..: 1 3 3 3 3 1 1 1 1 1 ...  $ Ticket : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...  $ Fare : num 211 152 152 152 152 ...  $ Cabin : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...  $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...  $ NA.1 : Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...  $ SubTitle : chr " Miss" " Master" " Miss" " Mr" ...  $ Title : chr "Miss" "Master" "Miss" "Mr" ...  > dat <- dat %>% mutate(Age = as.factor(Age),Title = as.factor(Title)) # convert as factor  > str(dat) # Check the data set  'data.frame': 1310 obs. of 15 variables:  $ PassengerId: int 1 1 1 1 1 1 1 1 1 1 ...  $ Survived : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...  $ Pclass : Factor w/ 876 levels "","Abbing","Abbott",..: 17 18 18 18 18 22 26 26 29 32 ...  $ Name : chr " Miss. Elisabeth Walton" " Master. Hudson Trevor" " Miss. Helen Loraine" " Mr. Hudson Joshua Creighton" ...  $ Sex : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...  $ Age : Factor w/ 98 levels "0.1667","0.3333",..: 45 7 9 46 39 71 88 60 76 95 ...  $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 1 2 2 2 2 1 2 1 3 1 ...  $ Parch : Factor w/ 8 levels "0","1","2","3",..: 1 3 3 3 3 1 1 1 1 1 ...  $ Ticket : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...  $ Fare : num 211 152 152 152 152 ...  $ Cabin : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...  $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...  $ NA.1 : Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...  $ SubTitle : chr " Miss" " Master" " Miss" " Mr" ...  $ Title : Factor w/ 19 levels "","Capt","Col",..: 12 11 12 15 16 15 12 15 16 15 ...  > # Now the data set is ready for imputation  > # using library mice. called earlier  > init = mice(dat, maxit=0)  > meth = init$method  > predM = init$predictorMatrix  > # below variable are not required for predicting the age  > predM[, c("PassengerId","Name", "Age","Ticket","Cabin", "Embarked")]=0  > # specify method for imputing the missing value  > meth[c("Age")]="norm"  > set.seed(1)  > imputed = mice(dat, method=meth, predictorMatrix=predM, m=5)  iter imp variable  1 1 PassengerId |
| imputed <- complete(imputed)  # check for missings in the imputed dataset  # Title and Age Group after imputation  p <- ggplot(data = imputed,  p + geom\_bar(position = "stack")+theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Counts of Title with Age Groups") |

