Data Wrangling Project 2

Vamsi K Mamidi

September 26, 2017

### Loading Packages

library (tidyr)  
library (readr)   
library (dplyr)

## Load the data in RStudio

titanic\_original <- read\_csv("titanic\_original.csv",col\_names = TRUE)

## Parsed with column specification:  
## cols(  
## pclass = col\_integer(),  
## survived = col\_integer(),  
## name = col\_character(),  
## sex = col\_character(),  
## age = col\_double(),  
## sibsp = col\_integer(),  
## parch = col\_integer(),  
## ticket = col\_character(),  
## fare = col\_double(),  
## cabin = col\_character(),  
## embarked = col\_character(),  
## boat = col\_character(),  
## body = col\_integer(),  
## home.dest = col\_character()  
## )

### Output

glimpse(titanic\_original)

## Observations: 1,310  
## Variables: 14  
## $ pclass <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...  
## $ survived <int> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1...  
## $ name <chr> "Allen, Miss. Elisabeth Walton", "Allison, Master. H...  
## $ sex <chr> "female", "male", "female", "male", "female", "male"...  
## $ age <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, ...  
## $ sibsp <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0...  
## $ parch <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1...  
## $ ticket <chr> "24160", "113781", "113781", "113781", "113781", "19...  
## $ fare <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26...  
## $ cabin <chr> "B5", "C22 C26", "C22 C26", "C22 C26", "C22 C26", "E...  
## $ embarked <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S", "C", "C...  
## $ boat <chr> "2", "11", NA, NA, NA, "3", "10", NA, "D", NA, NA, "...  
## $ body <int> NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124, NA, NA...  
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "...

## Port of embarkation

titanic\_original %>%filter(is.na(embarked))

## # A tibble: 3 x 14  
## pclass survived name sex age  
## <int> <int> <chr> <chr> <dbl>  
## 1 1 1 Icard, Miss. Amelie female 38  
## 2 1 1 Stone, Mrs. George Nelson (Martha Evelyn) female 62  
## 3 NA NA <NA> <NA> NA  
## # ... with 9 more variables: sibsp <int>, parch <int>, ticket <chr>,  
## # fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <int>,  
## # home.dest <chr>

titanic\_original <- titanic\_original %>% replace\_na(list(embarked = "S"), embarked)

### Output:

titanic\_original %>%select(embarked)%>%unique()

## # A tibble: 3 x 1  
## embarked  
## <chr>  
## 1 S  
## 2 C  
## 3 Q

## AGE

Calculated Mean, Meadian, Min, Max after removing the missing values

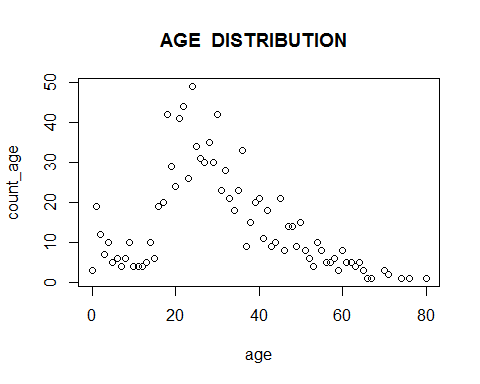
titanic\_original%>%filter(!is.na(age))%>%summarise(mean=mean(age),max=max(age),min=min(age),median=median(age))

## # A tibble: 1 x 4  
## mean max min median  
## <dbl> <dbl> <dbl> <dbl>  
## 1 29.88113 80 0.1667 28

The below plot is a scattered plot of ages vs Frequency, after removing the missing ages. we can see that the propability of a missing age value around the mean is more compared with other age groups.

### AGE DISTRIBUTION

titanic\_original%>%filter(!is.na(age))%>%mutate(age=round(age,0))%>%group\_by(age)%>%summarise(count\_age=n())%>%arrange(desc(count\_age))%>%with(plot(age,count\_age,main = "AGE DISTRIBUTION "))



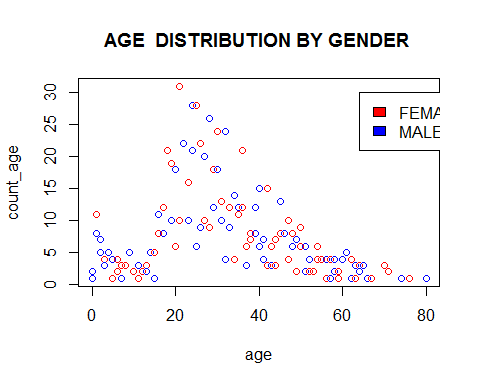
### AGE DISTRIBUTION by Gender

There are other ways to determine the missing age values which require more detail analysis by including other features like gender. For example we can calculate the mean age for both male and Female and we can replace the male missing age value with male mean age and female missing age with female mean age

titanic\_original%>%filter(!is.na(age))%>%group\_by(sex)%>%summarise(mean=mean(age),max=max(age),min=min(age),median=median(age))

## # A tibble: 2 x 5  
## sex mean max min median  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 female 28.68707 76 0.1667 27  
## 2 male 30.58523 80 0.3333 28

titanic\_original%>%filter(!is.na(age))%>%mutate(age=round(age,0))%>%group\_by(age,sex)%>%summarise(count\_age=n())%>%arrange(desc(count\_age))%>%with(plot(age,count\_age,col=c("red","blue"),main = "AGE DISTRIBUTION BY GENDER "))  
legend(x=64,y=30,legend=c("FEMALE","MALE"),fill=c("red","blue"))



I assume that by replacing the missing values depending on gender mean ages is better than just replacing all the missing age values with overall age mean, because as we can see that the distribution of age for male and female are different.

### Replacing missing age values with gender age mean

titanic\_original<-titanic\_original%>%mutate(age=ifelse(is.na(age) & sex=="male" , 30.58523, ifelse(is.na(age) & sex=="female" , 28.68707,age)))

## Lifeboat

Fill these empty slots with a dummy value e.g. the string ‘None’ or ‘NA’

titanic\_original<-titanic\_original%>%mutate(boat=ifelse(is.na(boat) ,"None",boat))

### Output:

titanic\_original%>%group\_by(boat)%>%summarise(count\_boat=n())%>%arrange(desc(count\_boat))

## # A tibble: 28 x 2  
## boat count\_boat  
## <chr> <int>  
## 1 None 824  
## 2 13 39  
## 3 C 38  
## 4 15 37  
## 5 14 33  
## 6 4 31  
## 7 10 29  
## 8 5 27  
## 9 3 26  
## 10 11 25  
## # ... with 18 more rows

## Cabin

Create a new column has\_cabin\_number which has 1 if there is a cabin number, and 0 otherwise.

titanic\_original<-titanic\_original%>%mutate(has\_cabin\_number=ifelse(is.na(cabin) ,0,1))

### Output:

titanic\_original%>%select(cabin,has\_cabin\_number,survived)

## # A tibble: 1,310 x 3  
## cabin has\_cabin\_number survived  
## <chr> <dbl> <int>  
## 1 B5 1 1  
## 2 C22 C26 1 1  
## 3 C22 C26 1 0  
## 4 C22 C26 1 0  
## 5 C22 C26 1 0  
## 6 E12 1 1  
## 7 D7 1 1  
## 8 A36 1 0  
## 9 C101 1 1  
## 10 <NA> 0 0  
## # ... with 1,300 more rows

### Table for Survied Vs has Cabin number

titanic\_original%>%filter(! is.na(survived))%>%group\_by(survived,has\_cabin\_number)%>%summarise(count\_survied=n())

## # A tibble: 4 x 3  
## # Groups: survived [?]  
## survived has\_cabin\_number count\_survied  
## <int> <dbl> <int>  
## 1 0 0 707  
## 2 0 1 102  
## 3 1 0 307  
## 4 1 1 193

# Create a CSV file

write.csv(titanic\_original, "titanic\_clean.csv")