R Notebook

#BIS581 #Visualization homework Load some data

titanic <- read.csv("titanic.csv",header=TRUE)

Load libraries

library(ggplot2)  
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

Perform any work you need transform,fix,wrangle your data

str(titanic)

## 'data.frame': 1313 obs. of 3 variables:  
## $ Name : chr "Allen, Miss Elisabeth Walton" "Allison, Miss Helen Loraine" "Allison, Mr Hudson Joshua Creighton" "Allison, Mrs Hudson JC (Bessie Waldo Daniels)" ...  
## $ PClass : chr "1st" "1st" "1st" "1st" ...  
## $ Survived: int 1 0 0 0 1 1 1 0 1 0 ...

summary(titanic)

## Name PClass Survived   
## Length:1313 Length:1313 Min. :0.0000   
## Class :character Class :character 1st Qu.:0.0000   
## Mode :character Mode :character Median :0.0000   
## Mean :0.3427   
## 3rd Qu.:1.0000   
## Max. :1.0000

titanic$Name <- as.character(titanic$Name)  
titanic$Survived <- as.factor(titanic$Survived)  
summary(titanic)

## Name PClass Survived  
## Length:1313 Length:1313 0:863   
## Class :character Class :character 1:450   
## Mode :character Mode :character

#create variable “title” extract second word of each name, apply this to every instance within the dataset and store it in title

titanic$Title <- sapply(titanic$Name, FUN=function(x) {regmatches(x,regexec(',[[:space:]](.\*?)[[:space:]]',x))[[1]][2]})

#create new variable gender, create char vector female\_titles that contains a list of title strings, we check if its in the list, if it is then the gender will be set to female and if not it will be male

Female\_titles <- c("Miss","Mrs","Lady", "Madame","Ms")  
titanic$Gender <- ifelse(titanic$Title %in% Female\_titles, "Female", "Male")

#change the new variables data type we made into factors so we can plot them correctly.

titanic$PClass <- as.factor(titanic$PClass)  
titanic$Gender <- as.factor(titanic$Gender)  
str(titanic)

## 'data.frame': 1313 obs. of 5 variables:  
## $ Name : chr "Allen, Miss Elisabeth Walton" "Allison, Miss Helen Loraine" "Allison, Mr Hudson Joshua Creighton" "Allison, Mrs Hudson JC (Bessie Waldo Daniels)" ...  
## $ PClass : Factor w/ 4 levels "\*","1st","2nd",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Survived: Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 1 2 1 ...  
## $ Title : chr "Miss" "Miss" "Mr" "Mrs" ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 2 1 2 1 2 ...

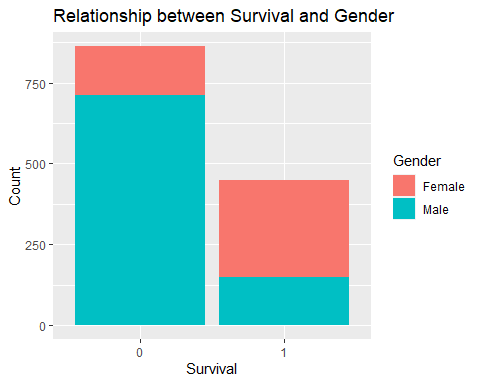
summary(titanic)

## Name PClass Survived Title Gender   
## Length:1313 \* : 1 0:863 Length:1313 Female:450   
## Class :character 1st:322 1:450 Class :character Male :863   
## Mode :character 2nd:279 Mode :character   
## 3rd:711

Using ggplot2, create appropriate plots to answer the following questions:

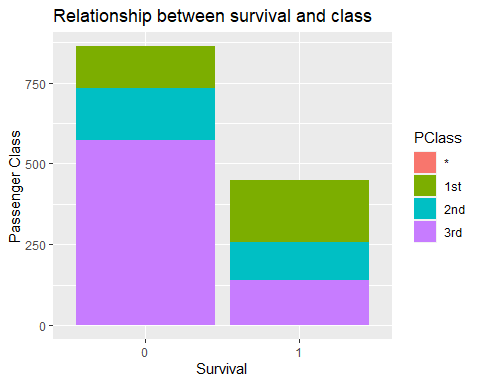
Is there a relationship between survival and gender? -Yes, there appears to be a relationship between survival and gender, with females having a higher chance of survival compared to males. The data shows that males had a significantly higher probability of death than females.

ggplot(data=titanic,mapping=aes(x=Survived,fill=Gender)) +   
 geom\_bar(stat="count") +  
 xlab("Survival") +  
 ylab("Count") +  
 ggtitle("Relationship between Survival and Gender")



Is there a relationship between survival and class? -Yes, there appears to be a relationship between survival and class, it seems the 3rd class has a marginally greater chance of death than the first and second class. We can also see that the 1st class has the highest chance of survival.

ggplot(data=titanic,mapping=aes(x=Survived,fill=PClass)) +   
 geom\_bar(stat = "count") +   
 xlab("Survival") +   
 ylab("Passenger Class") +   
 ggtitle("Relationship between survival and class")



#create new variable married

titanic$martial\_status <- ifelse(titanic$Title %in% c("Mrs", "Lady", "Madame"), "Married", "Not married")  
titanic$martial\_status <- as.factor(titanic$martial\_status)

Is there a relationship between survival and marital status for females? What about males? If there isn’t enough data to answer a question, state so and explain. -Yes it seems as there is a relationship between survival and marital status for females, from the chart we can tell that alot more females died who were not married than married this may be due to saving families before single passengers. -It is not that clear cut for males due to us not having enough data to correctly label the males accurately as married. We can try to use the titles for males. Males don’t use standardized titles in most places so it is not as easy or accurate to do as for females.

ggplot(data = titanic,mapping = aes(x=Survived,fill=martial\_status)) + geom\_bar(stat = "count") +  
 facet\_wrap("Gender") +  
 xlab("survival") +  
 ylab("martial status") +  
 ggtitle("Relationship between survival and martial status for females and males")

