Titanic Visualization\_Sonia

#BIS581 #Visualization homework Load some data

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

library(tidyr)  
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

library(plyr)

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

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

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

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 ...

head(titanic)

## Name PClass Survived  
## 1 Allen, Miss Elisabeth Walton 1st 1  
## 2 Allison, Miss Helen Loraine 1st 0  
## 3 Allison, Mr Hudson Joshua Creighton 1st 0  
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels) 1st 0  
## 5 Allison, Master Hudson Trevor 1st 1  
## 6 Anderson, Mr Harry 1st 1

## Converting the variables to factors  
  
titanic$Survived = factor(titanic$Survived)  
titanic$PClass=factor(titanic$PClass)

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 : 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 ...

## Changing the dummy variable (survival) to Survived or Not Survived  
  
titanic$Survived <-revalue(titanic$Survived, c("1"="Survived", "0"="Not Survived"))

table(titanic$Survived)

##   
## Not Survived Survived   
## 863 450

## Creating new variable 'Title' by using the title of the passengers  
  
titanic$Title <- sapply(titanic$Name,FUN=function(x){regmatches(x,regexec(',[[:space:]](.\*?)[[:space:]]',x))[[1]][2]})

## Converting the 'Title' variable to factor  
  
titanic$Title=factor(titanic$Title)

str(titanic)

## 'data.frame': 1313 obs. of 4 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 "Not Survived",..: 2 1 1 1 2 2 2 1 2 1 ...  
## $ Title : Factor w/ 31 levels "Albert","Captain",..: 15 15 17 18 14 17 15 17 18 17 ...

titanic %>% pull(Title) %>% levels()

## [1] "Albert" "Captain" "Col" "Colonel" "Dr" "Eino"   
## [7] "Ernst" "Hilda" "Jenny" "Jonkheer" "Lady" "Madame"   
## [13] "Major" "Master" "Miss" "Mlle" "Mr" "Mrs"   
## [19] "Ms" "Ms " "Nikolai" "Nils" "Oscar" "Rev"   
## [25] "Richard" "Simon" "Sir" "the" "Thomas" "W"   
## [31] "William"

## Trying to create a variable 'Gender' to show the relationship between survival and gender  
  
titanic <- within(titanic, {  
 Gender = "male"  
 Gender[Title %in% c("Albert", "Captain","Col", "Colonel","Dr","Eino",  
 "Ernst","Jonkheer","Major","Master","Mr","Nikolai",  
 "Nils","Oscar","Rev","Richard", "Simon","Sir",  
 "Thomas","W","William")] = "Male"  
 Gender[Title %in% c("Hilda","Jenny", "Lady","Madame", "Miss","Mlle", "Mrs","Ms", "Ms ", "the")] = "Female"  
})

summary(titanic)

## Name PClass Survived Title   
## Length:1313 \* : 1 Not Survived:863 Mr :750   
## Class :character 1st:322 Survived :450 Miss :236   
## Mode :character 2nd:279 Mrs :200   
## 3rd:711 Master : 59   
## Ms : 12   
## (Other): 44   
## NA's : 12   
## Gender   
## Length:1313   
## Class :character   
## Mode :character   
##   
##   
##   
##

**There are 12 missing values in Title. That means 12 passengers of total 1313 passengers have not been assigned titles.**

## Converting the 'Gender' variable to factor  
  
titanic$Gender = factor(titanic$Gender)

## Trying to Omit NA's from data set  
  
titanic\_1 <- na.omit (titanic)

## Trying to check PClass  
  
titanic\_1 %>% pull(PClass) %>% as.factor() %>% levels()

## [1] "\*" "1st" "2nd" "3rd"

## Dropping the class "\*" from PClass  
  
titanic\_1$PClass<-factor(titanic\_1$PClass, levels=c("1st", "2nd", "3rd"))

titanic\_1 %>% pull(PClass) %>% as.factor() %>% levels()

## [1] "1st" "2nd" "3rd"

summary(titanic\_1)

## Name PClass Survived Title Gender   
## Length:1301 1st:317 Not Survived:856 Mr :750 Female:455   
## Class :character 2nd:279 Survived :445 Miss :236 Male :846   
## Mode :character 3rd:705 Mrs :200   
## Master : 59   
## Ms : 12   
## Dr : 8   
## (Other): 36

table(titanic\_1$Gender)

##   
## Female Male   
## 455 846

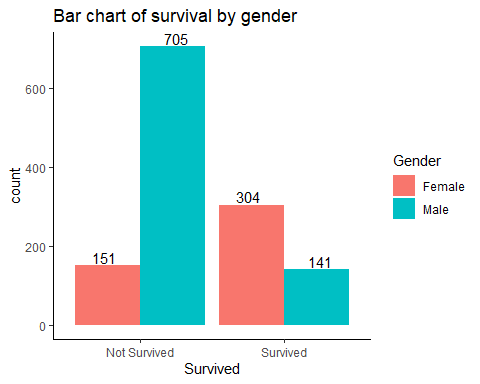
**Now we have all variables to answer the following questions.**

library(ggplot2)

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

**1. Is there a relationship between survival and gender?**

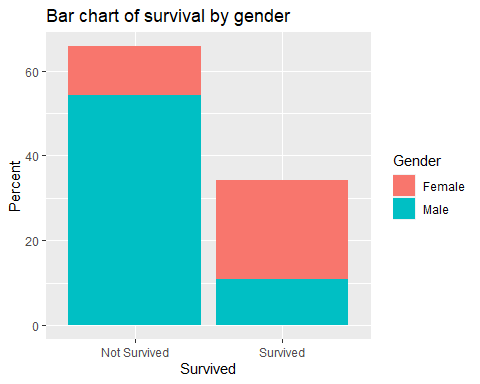
ggplot(titanic\_1, aes(x = Survived, fill=Gender)) +geom\_bar(position = position\_dodge()) +geom\_text(stat='count', aes(label=stat(count)), position = position\_dodge(width=1), vjust=-0.2)+theme\_classic()+labs(title="Bar chart of survival by gender")



## in percentage terms  
  
titanic\_12 <- titanic\_1 %>% group\_by(Survived, Gender) %>% dplyr::summarize(count = n()) %>%mutate(Percent=count/sum(count))

## `summarise()` has grouped output by 'Survived'. You can override using the `.groups` argument.

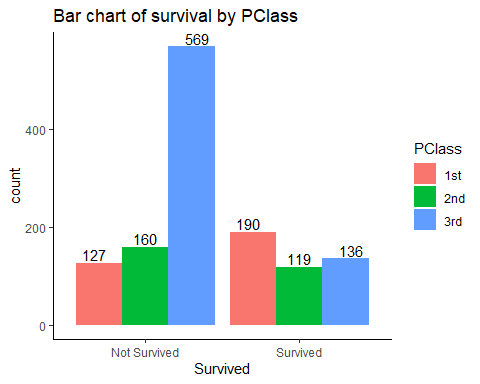
ggplot(titanic\_12, aes(x= Survived, y=Percent\*100, fill = Gender)) +geom\_bar(stat="identity") + labs(title="Bar chart of survival by gender", y="Percent")



We created two graphs for showing the relationship between survival and gender. The first one is based on count and the second one is based on percentage. From the above graphs, we can infer that females are more likely to survive than males.But there are some titles i.e ‘Captain’, ‘Master’, ‘Colonel’,‘Major’,‘Dr’ that are gender neutral. We assumed these gender neutral title as Male. If we assume these title as female, then results may be slightly different. This shows that there is a relationship between survival and gender.

**2. Is there a relationship between survival and class?**

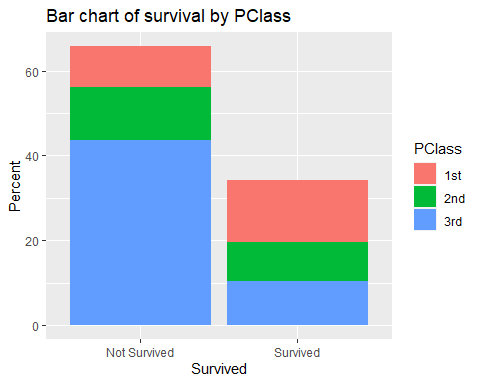
ggplot(titanic\_1, aes(x = Survived, fill=PClass)) +geom\_bar(position = position\_dodge()) +geom\_text(stat='count', aes(label=stat(count)), position = position\_dodge(width=1), vjust=-0.2)+theme\_classic()+labs(title="Bar chart of survival by PClass")



## in percentage terms  
  
titanic\_13 <- titanic\_1 %>% group\_by(Survived, PClass) %>% dplyr::summarize(count = n()) %>%mutate(Percent=count/sum(count))

## `summarise()` has grouped output by 'Survived'. You can override using the `.groups` argument.

ggplot(titanic\_13, aes(x= Survived, y=Percent\*100, fill = PClass)) +geom\_bar(stat="identity") + labs(title="Bar chart of survival by PClass", y="Percent")



From the two figures above, we can infer that the number of surviving passengers in the first class was higher and the number of non-surviving passengers in the third class was higher. This shows that there is a relationship between survival and class.

**3. Is there a relationship between survival and marital status for females?**

Here we classified the title with ‘Mrs’, ‘Madame’ as females who are married and the title with ‘Miss’ ‘Ms’, or ’Ms’as females who are not married. There are some females with other titles but we have ignored them in the new data set because it is difficult to tell whether they are married by looking at their titles.

titanic\_14<- within(titanic\_1,{  
 Marital='Married'  
Marital[Title %in% c('Mrs','Madame')]='Married'  
Marital[Title %in% c('Miss', 'Ms','Ms')]='Unmarried'  
})

titanic\_14$Marital = factor(titanic\_14$Marital)

summary(titanic\_14)

## Name PClass Survived Title Gender   
## Length:1301 1st:317 Not Survived:856 Mr :750 Female:455   
## Class :character 2nd:279 Survived :445 Miss :236 Male :846   
## Mode :character 3rd:705 Mrs :200   
## Master : 59   
## Ms : 12   
## Dr : 8   
## (Other): 36   
## Marital   
## Married :201   
## Unmarried:248   
## NA's :852   
##   
##   
##   
##

titanic\_14 <- filter(titanic\_14, Gender=="Female")

summary(titanic\_14)

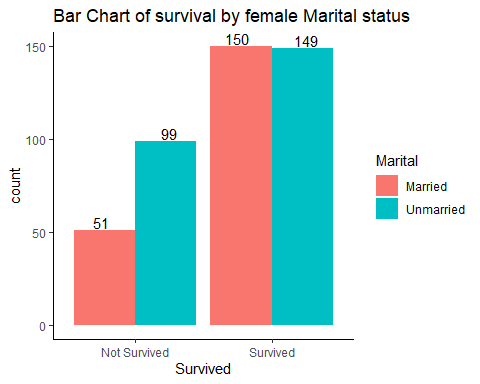
## Name PClass Survived Title Gender   
## Length:455 1st:137 Not Survived:151 Miss :236 Female:455   
## Class :character 2nd:107 Survived :304 Mrs :200 Male : 0   
## Mode :character 3rd:211 Ms : 12   
## Hilda : 1   
## Jenny : 1   
## Lady : 1   
## (Other): 4   
## Marital   
## Married :201   
## Unmarried:248   
## NA's : 6   
##   
##   
##   
##

##Omitting NA's from the data set  
  
titanic\_14<- na.omit(titanic\_14)

summary(titanic\_14)

## Name PClass Survived Title Gender   
## Length:449 1st:133 Not Survived:150 Miss :236 Female:449   
## Class :character 2nd:107 Survived :299 Mrs :200 Male : 0   
## Mode :character 3rd:209 Ms : 12   
## Madame : 1   
## Albert : 0   
## Captain: 0   
## (Other): 0   
## Marital   
## Married :201   
## Unmarried:248   
##   
##   
##   
##   
##

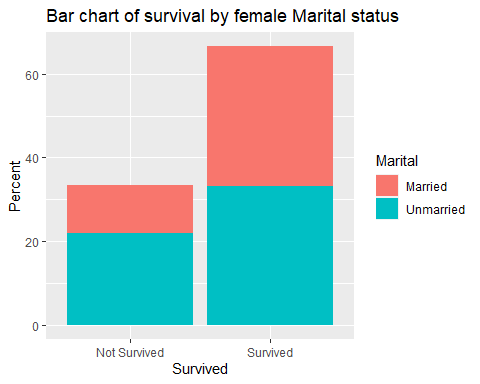
ggplot(titanic\_14, aes(x = Survived, fill=Marital)) +geom\_bar(position = position\_dodge()) +geom\_text(stat='count', aes(label=stat(count)), position = position\_dodge(width=1), vjust=-0.2)+theme\_classic()+labs(title="Bar Chart of survival by female Marital status")



##In terms of percentage  
  
titanic\_14 <- titanic\_14 %>%group\_by(Survived, Marital) %>% dplyr::summarize(count = n()) %>%mutate(Percent=count/sum(count))

## `summarise()` has grouped output by 'Survived'. You can override using the `.groups` argument.

ggplot(titanic\_14, aes(x= Survived, y=Percent\*100, fill = Marital)) +geom\_bar(stat="identity") + labs(title="Bar chart of survival by female Marital status", y="Percent")



From the above diagrams, we can see that 150 married female passengers and 149 unmarried female passengers survived. 51 married female passengers and 99 unmarried female passengers died.This shows that there is very weak relationship between survival and female marital status.

**What about males? If there isn’t enough data to answer a question, state so and explain.**

I think there isn’t enough data to answer the question for male. There are 21 titles for male and it’s not possible to infer whether they are married by looking at their titles because none of them are used to infer their marital status. Even by making some assumptions I think it’s not enough to answer the question.