Titanic answer sheet

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:

```
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(ggplot2)
#*****************************************************************************
#**********************Session 4 Lab*****************************************
#*****************************************************************************

###############################################################################
##                        TITANIC EXERCISE                                   ##
###############################################################################

# set working directory and read titanic.csv

# read more about the data here

# survival- Survival 0 = No, 1 = Yes
# pclass- Ticket class  1 = 1st, 2 = 2nd, 3 = 3rd
# sex-  Sex
# age- Age in years
# sibsp- # of siblings / spouses aboard the Titanic
# parch- # of parents / children aboard the Titanic
# ticket-   Ticket number
# fare- Passenger fare
# cabin- Cabin number
# embarked- Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton
```

```r
# 1. Let's start with setting the working diretory

setwd("~/Google Drive/MGT 585/class material/S4-Visualization/data")

# 2. read the data into titanic object

titanic <- read.csv("titanic.csv")

# 3. explore your dataset using 5 functions: dim(), str(), colnames(), head() and tail

dim(titanic)
```

```
## [1] 891  12
```

```r
str(titanic)
```

```
## '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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```r
colnames(titanic)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```r
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
```

```
## 6                                   Moran, Mr. James   male  NA     0     0
##              Ticket    Fare Cabin Embarked
## 1         A/5 21171  7.2500              S
## 2          PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
tail(titanic)
```

```
##     PassengerId Survived Pclass                                      Name    Sex
## 886         886        0      3     Rice, Mrs. William (Margaret Norton) female
## 887         887        0      2                     Montvila, Rev. Juozas   male
## 888         888        1      1              Graham, Miss. Margaret Edith female
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female
## 890         890        1      1                     Behr, Mr. Karl Howell   male
## 891         891        0      3                       Dooley, Mr. Patrick   male
##     Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886  39     0     5     382652 29.125              Q
## 887  27     0     0     211536 13.000              S
## 888  19     0     0     112053 30.000   B42        S
## 889  NA     1     2 W./C. 6607 23.450              S
## 890  26     0     0     111369 30.000  C148        C
## 891  32     0     0     370376  7.750              Q
```

```
# Write the number of rows and columns here:
# Note factor, numeric and integer columns

# 4. make sure nominal variables are factors which includes Survived, Pclass and Sex

titanic$Survived<-as.factor(titanic$Survived)
titanic$Pclass<-as.factor(titanic$Pclass)
titanic$Sex<-as.factor(titanic$Sex)

# 5. use table() to see the distribution of Survived, Pclass and Sex

table(titanic$Survived)
```

```
##
##   0   1
## 549 342
```

```
table(titanic$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

```
table(titanic$Sex)
```

```
##
## female   male
##    314    577
```

```
# 6. use summary() and sd() to see the summary statistics of Fare, SibSp and Age

summary(titanic$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.91   14.45   32.20   31.00  512.33
```

```
sd(titanic$Fare)
```

```
## [1] 49.69343
```

```
summary(titanic$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

```
sd(titanic$Age)
```

```
## [1] NA
```

```
sd(!is.na(titanic$Age))
```

```
## [1] 0.3992104
```

```
summary(titanic$SibSp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.523   1.000   8.000
```

```
sd(titanic$SibSp)
```

```
## [1] 1.102743
```

```
################################################################################

# OBJECTIVE: we will answer the question: who survived using visualization

# import ggplot

library(ggplot2)

# Q1) Does age play a role?

# Let's try to imagine how the graph should look like
```

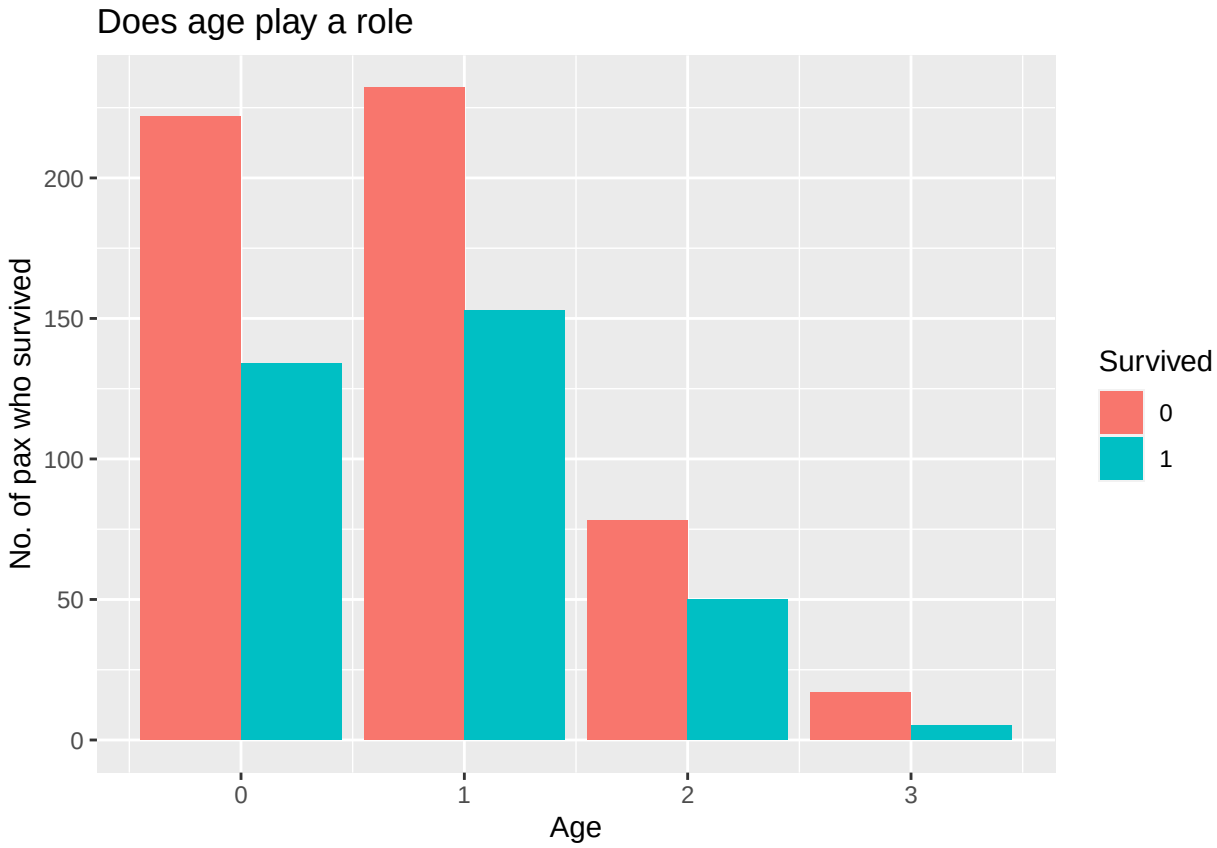
```r
# It should be a bar plot with age on the x axis
# No. of passengers who survive on the y axis
# the grouping variable (fill) is Survived

# I have created a categorical variable for you from the countuous age column
# It takes four values from 0-3
# age 0-20 are group 0
# age 21- 40 are group 1
# age 41 to 60 are group 2
# age 61 and above are group 3

titanic$age_cat<-0
titanic$age_cat[titanic$Age>20]<-1
titanic$age_cat[titanic$Age>40]<-2
titanic$age_cat[titanic$Age>60]<-3

# create a data frame with three columns using group_by and summarise()
# group_by should be on two columns of age_cat and Survived
# in the summariz() use n() to count the number of rows

titanic0 <- titanic %>%
  group_by(age_cat,Survived) %>%
  summarise(totalSurv = n())
```

```
## `summarise()` regrouping output by 'age_cat' (override with `.groups` argument)
```

```r
titanic0
```

```
## # A tibble: 8 x 3
## # Groups:   age_cat [4]
##   age_cat Survived totalSurv
##     <dbl> <fct>        <int>
## 1       0 0              222
## 2       0 1              134
## 3       1 0              232
## 4       1 1              153
## 5       2 0               78
## 6       2 1               50
## 7       3 0               17
## 8       3 1                5
```

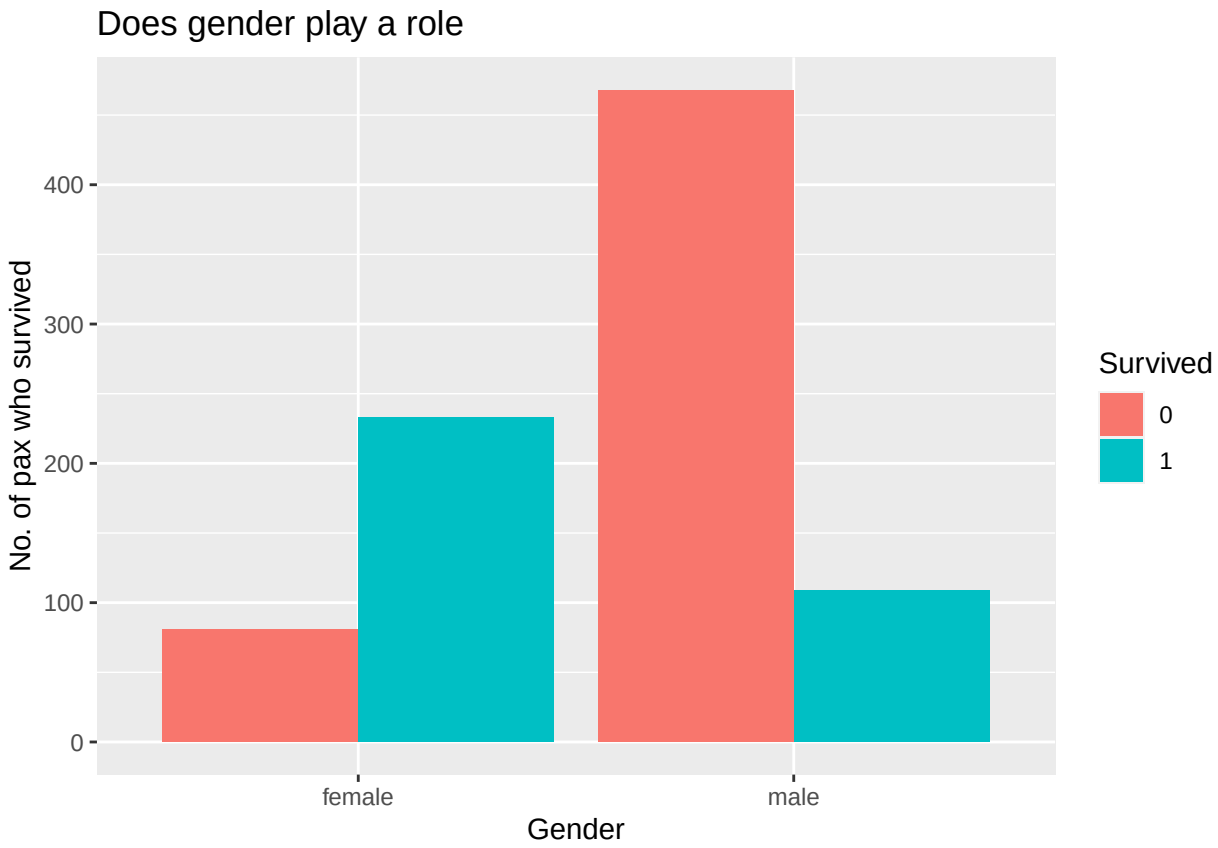
```r
# now using the new object that you just created, plot a bar graph where
# x axis is age
# y axis is number of passengers who survived
# grouping variable (fill) is Survived

ggplot(titanic0, aes(fill=as.factor(Survived), y=totalSurv, x=age_cat)) +
  geom_bar(position="dodge", stat="identity") + xlab("Age") + ylab("No. of pax who survived") +ggtitle(
```




```
# You will notice that I have use more functions
# xlab() is label of x-axis
# ylab() is label of y-axis
# ggtitle() is the main title of the chart
# lab() specifies the label for the grouping variable

################################################################################

# Q2) Does gender play a role?

# Follow similar pattern as the last question, replace age_cat with Sex

titanic1 <- titanic %>%
  group_by(Sex,Survived) %>%
  summarise(totalSurv = n())
```

```
## `summarise()` regrouping output by 'Sex' (override with `.groups` argument)
```

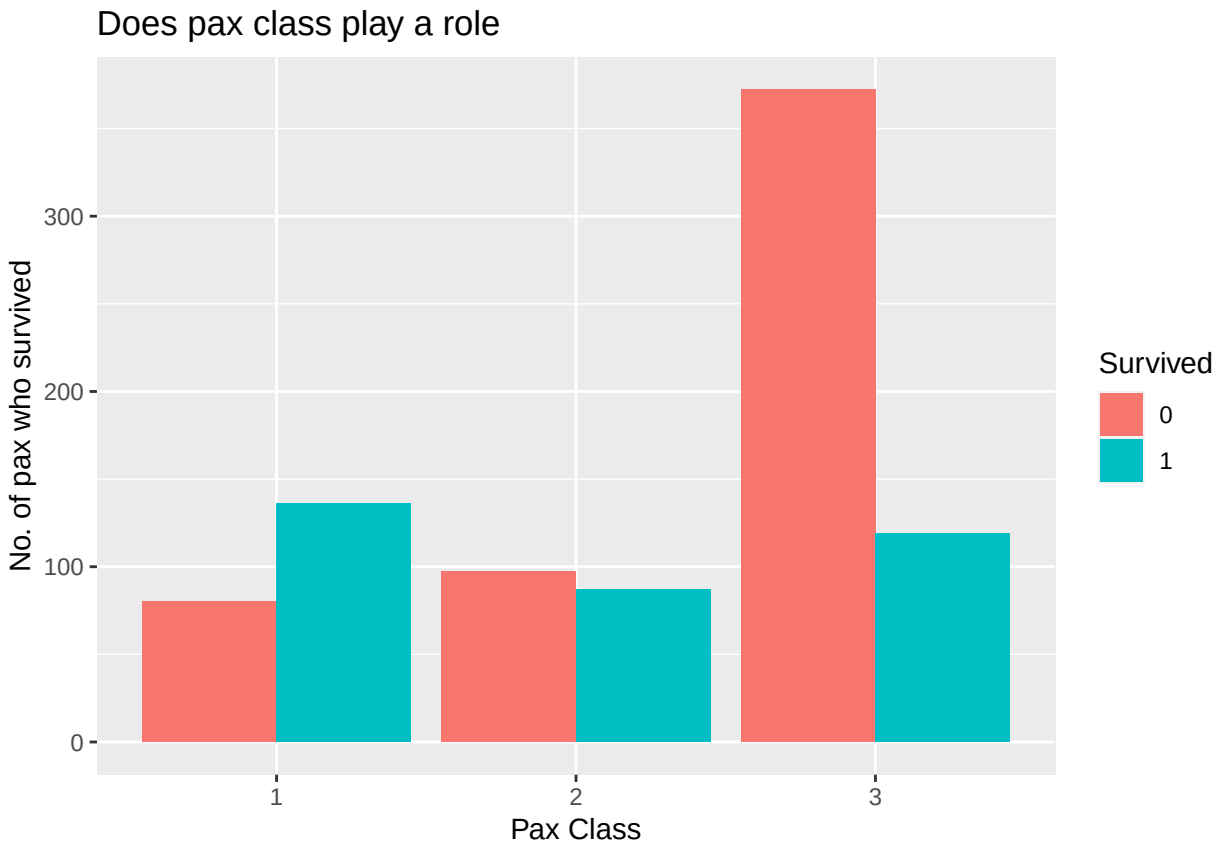
```
ggplot(titanic1, aes(fill=as.factor(Survived), y=totalSurv, x=Sex)) +
  geom_bar(position="dodge", stat="identity") + xlab("Gender") + ylab("No. of pax who survived") +ggtitl
```




```
##############################################################################

# Q3) Does passenger's class play a role?

# Follow similar pattern as the last question, replace age_cat with Pclass

titanic2 <- titanic %>%
  group_by(Pclass,Survived) %>%
  summarise(totalSurv = n())
```

```
## `summarise()` regrouping output by 'Pclass' (override with `.groups` argument)
```

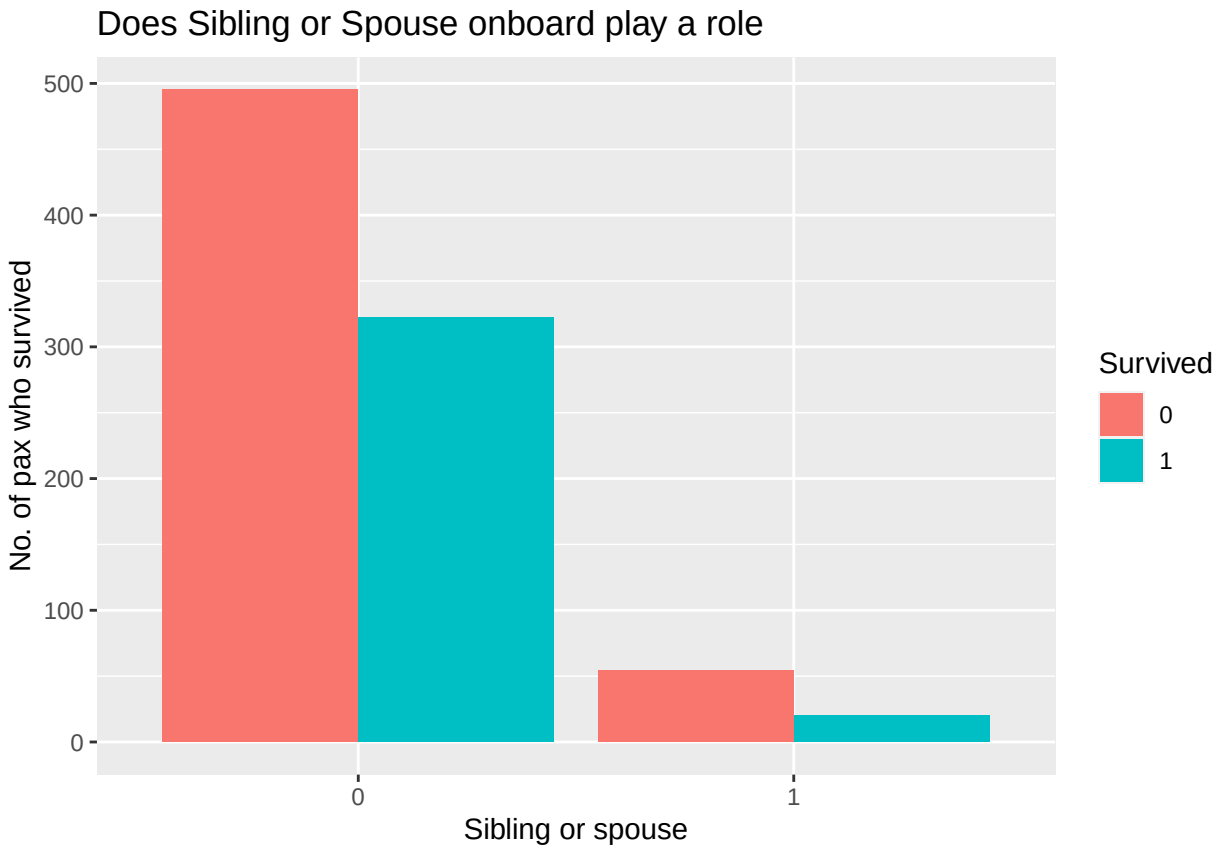
```
ggplot(titanic2, aes(fill=as.factor(Survived), y=totalSurv, x=Pclass)) +
  geom_bar(position="dodge", stat="identity") + xlab("Pax Class") + ylab("No. of pax who survived") +gg
```




```
#################################################################################

# Q4) Does number of siblings or spouse onboard play a role?

# Follow similar pattern as the last question, replace age_cat with SibSp_cat

titanic$SibSp_cat<-0
titanic$SibSp_cat[titanic$SibSp>1]<-1

titanic$SibSp_cat<-as.factor(titanic$SibSp_cat)

titanic3 <- titanic %>%
  group_by(SibSp_cat,Survived) %>%
  summarise(totalSurv = n())
```

```
## `summarise()` regrouping output by 'SibSp_cat' (override with `.groups` argument)
```

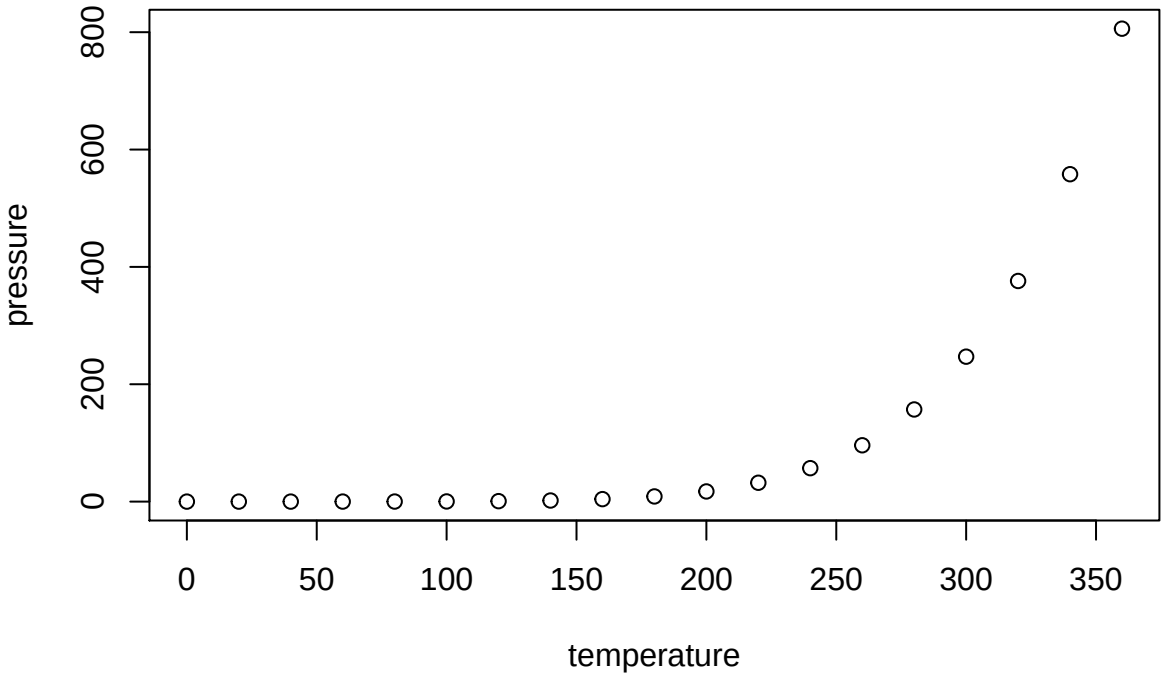
```
ggplot(titanic3, aes(fill=as.factor(Survived), y=totalSurv, x=SibSp_cat)) +
  geom_bar(position="dodge", stat="identity") + xlab("Sibling or spouse") + ylab("No. of pax who survive
```




```
################################################################################
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.