project5_in_R

Isaac Moore

9/4/2017

```r
setwd("~/Google Drive/data_science/general_assembly/Projects/DSI_SM_Project5/r")
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```r
library(RPostgreSQL)
```

```
## Loading required package: DBI
```

```r
library(stringr)
```

```r
# had to add a new table of the test dataframe, since someone had deleted it table form the database.

# test_clean <- read_csv("../data/test.csv")
# dbWriteTable(con, "titanic_test",
#              value = test_clean, append = FALSE, row.names = FALSE)
```

Part 1: Aquire the Data

1. Connect to the remote database

```r
pw <- {
  "gastudents"
}

# loads the PostgreSQL driver
drv <- dbDriver("PostgreSQL")
# creates a connection to the postgres database
# note that "con" will be used later in each connection to the database
con <- dbConnect(drv, dbname = "titanic",
                 host = "dsi.c20gkj5cvu3l.us-east-1.rds.amazonaws.com", port = 5432,
                 user = "dsi_student", password = pw)
rm(pw) # removes the password
```

2. Query the database and aggregate the data

```
train <- dbGetQuery(con, "SELECT * from titanic_train")
test <- dbGetQuery(con, "SELECT * from titanic_test")
```

Part 2: Exploratory Data Analysis

1. Describe the Data

```
df <- train
```

```
colSums(is.na(df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

```
summary(df)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Exploring survival statistics

```
glimpse(df)
```

```
## Observations: 891
## Variables: 12
## $ PassengerId <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Survived    <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,...
## $ Pclass      <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3,...
```

```
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bra...
## $ Sex         <chr> "male", "female", "female", "female", "male", "mal...
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, ...
## $ SibSp       <dbl> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4,...
## $ Parch       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1,...
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "1138...
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, ...
## $ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, ...
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", ...
```

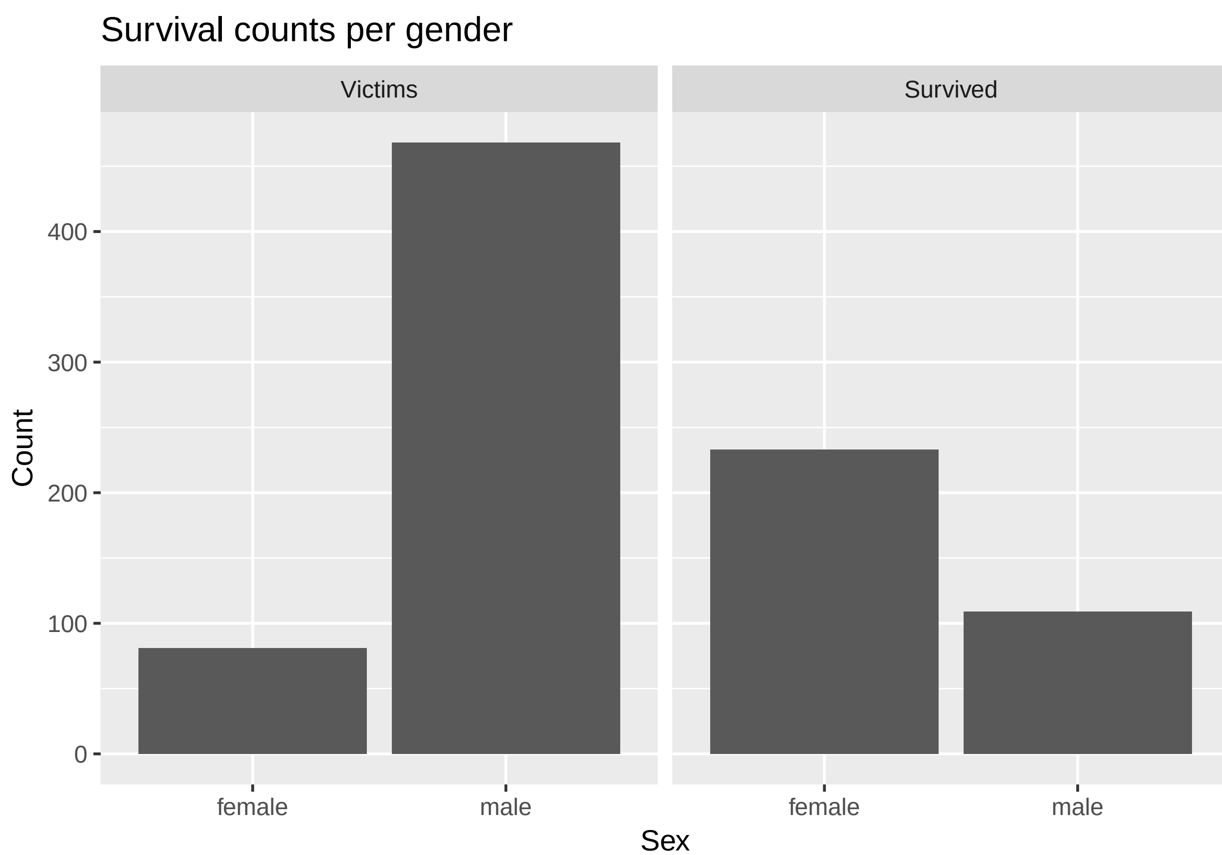
```
df$Survived <- as.factor(df$Survived)
levels(df$Survived) <- c("Victims", "Survived")
table(df$Survived)
```

```
##
##  Victims Survived
##      549      342
```

```
df %>% group_by(Sex, Survived) %>% summarise(n = n())
```

```
## Source: local data frame [4 x 3]
## Groups: Sex [?]
##
##      Sex Survived     n
##    <chr>   <fctr> <int>
## 1 female  Victims    81
## 2 female Survived   233
## 3   male  Victims   468
## 4   male Survived   109
```

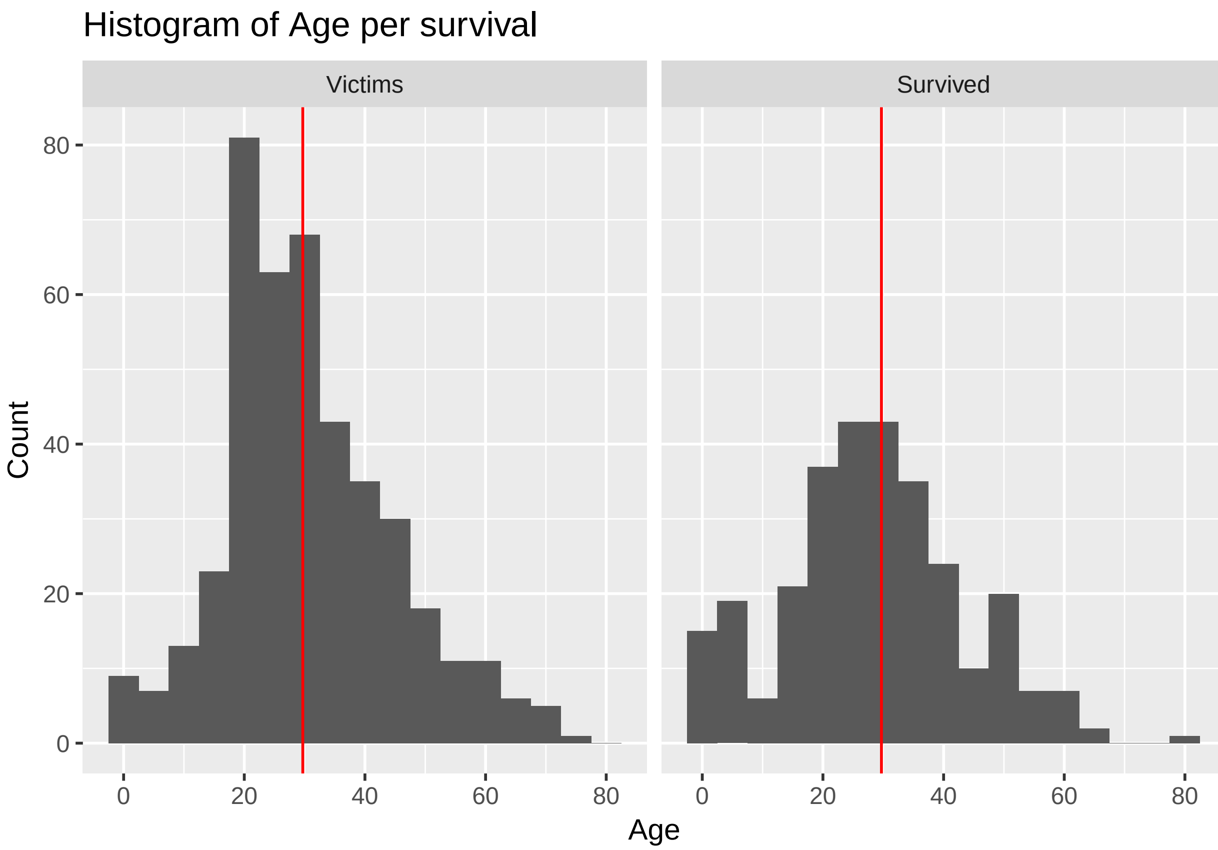
```
df %>% group_by(Sex, Survived) %>%
        summarise(n = n()) %>%
        ggplot(aes(x = Sex, y = n)) +
                geom_bar(stat = "identity") +
                facet_grid(.~Survived) +
                labs(title = "Survival counts per gender", x = "Sex", y = "Count")
```



More females survived, than perished, in our training dataset.

```
df %>% ggplot(aes(x = Age)) +
                geom_histogram(binwidth = 5) +
                facet_grid(.~Survived) +
                geom_vline(xintercept = mean(df$Age, na.rm = T),colour = "red", show.legend = TRUE) +
                labs(title = "Histogram of Age per survival", y = "Count", x = "Age")
```

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



If you are younger you were more likely to survive.

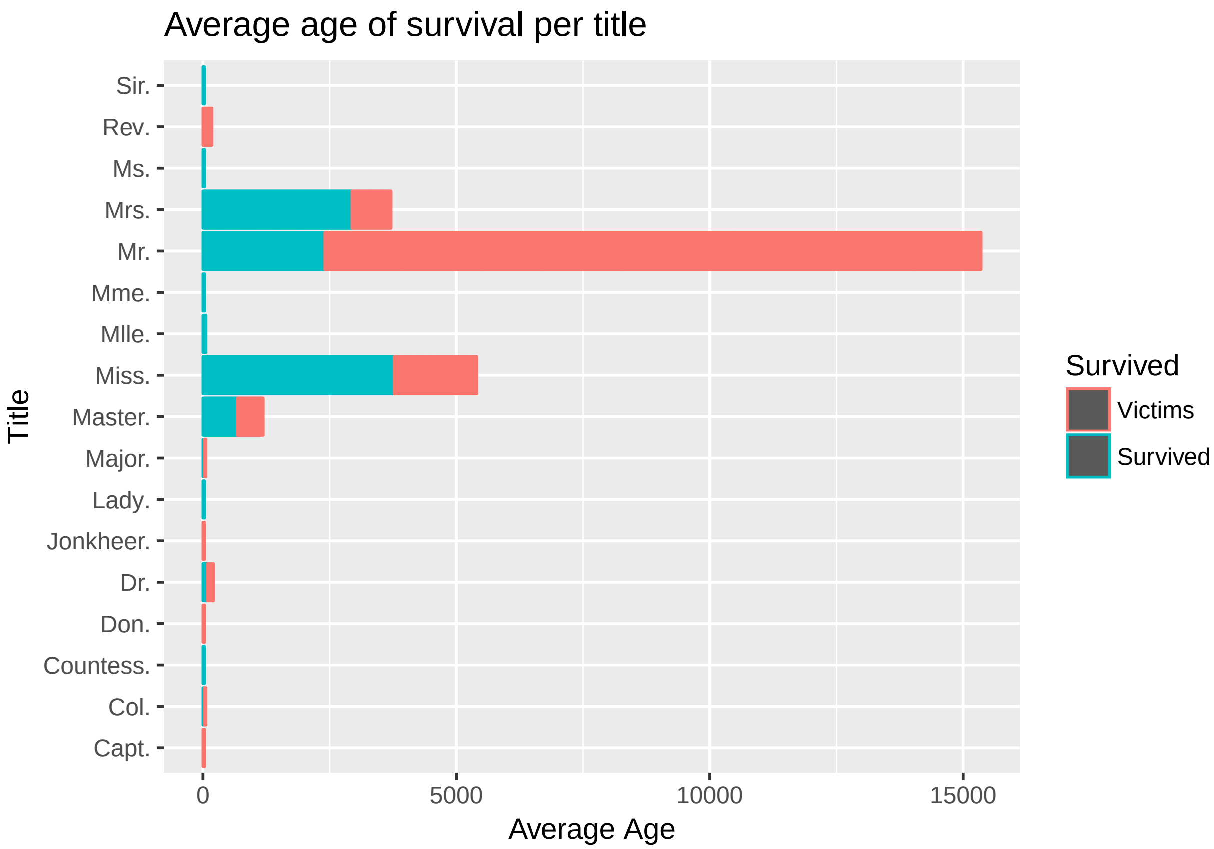
```
df %>% group_by(Sex, Pclass) %>%
        summarise(price = mean(Fare)) %>%
        ggplot(aes(y = price, x = Pclass, col = factor(Sex))) +
                geom_bar(stat = "identity", position = "dodge") +
                labs(title = "Average prices paid per class for Male/Female", x = "Passenger Class", y =
```

Females on average paid more than males, especially in first class.

```
df$title <- str_extract(df$Name, regex("[A-Z]\\w+\\."))
df$title[is.na(df$title)] <- "Other"
table(df$title)
```

```
##
##     Capt.      Col. Countess.      Don.       Dr. Jonkheer.      Lady.
##         1         2         1         1         7         1         1
##    Major.   Master.     Miss.     Mlle.      Mme.       Mr.      Mrs.
##         2        40       182         2         1       517       125
##       Ms.      Rev.      Sir.
##         1         6         1
```

```
ggplot(df, aes(x = title, y = mean(Age, na.rm = T), col = Survived)) +
        geom_bar(stat = "identity", position = "stack") +
        labs(title = "Average age of survival per title", y = "Average Age", x = "Title") +
        coord_flip()
```



Unmarried women (Miss.) had a better survival rate (per average age) vs married women (Mrs.)

Part 3: Data Wrangling

1. Create Dummy Variables for Sex

I will convert the Sex column to a factor, which will work better in R

```
df$Sex <- as.factor(df$Sex)
df$Pclass <- as.factor(df$Pclass)
df$Embarked <- as.factor(df$Embarked)
```

Fill NA values…

```
df$Age[is.na(df$Age)] <- mean(df$Age, na.rm = T)# Filling with the mean Age
df$Cabin[is.na(df$Cabin)] <- "???" # too many to drop the columns, filling with '???'
df <- na.omit(df)
colSums(is.na(df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
##       title
##           0
```

Part 4: Logistic Regression and Model Validation

1. Define the variables that we will use in our classification analysis

We will be using the *Pclass + Sex + Age + Parch + Fare + Embarked* columns from the dataframe to predict who survived on the Titanic

2. Transform "Y" into a 1-Dimensional Array for SciKit-Learn

No need to perform for logistic regression in R, you are able to specify our dependent and independent variables in the call to formulate the model.

3. Conduct the logistic regression

```
model <- glm(Survived ~ Pclass + Sex + Age + Parch + Fare + Embarked, family=binomial(link='logit'), da
```

4. Examine the coefficients to see our correlations

```
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Parch + Fare +
##     Embarked, family = binomial(link = "logit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5108  -0.6606  -0.4016   0.6322   2.4743
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.936571   0.464043   8.483  < 2e-16 ***
## Pclass2     -0.912829   0.292904  -3.116  0.00183 **
## Pclass3     -2.205711   0.292670  -7.537 4.83e-14 ***
## Sexmale     -2.647906   0.196636 -13.466  < 2e-16 ***
## Age         -0.035023   0.007617  -4.598 4.27e-06 ***
## Parch       -0.202220   0.116487  -1.736  0.08256 .
## Fare         0.001159   0.002282   0.508  0.61146
## EmbarkedQ   -0.048027   0.374525  -0.128  0.89796
## EmbarkedS   -0.529483   0.236918  -2.235  0.02543 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  793.85  on 880  degrees of freedom
## AIC: 811.85
##
## Number of Fisher Scoring iterations: 5
```

6. Test the Model by introducing a Test or Validaton set

```
test$Sex <- as.factor(test$Sex)
test$Pclass <- as.factor(test$Pclass)
test$Embarked <- as.factor(test$Embarked)
```

```
test_sub <- test %>% select(Pclass, Sex, Age, Parch, Fare, Embarked)
preds <- predict(model,test_sub,type='response')
preds[1:10]
```

```
##          1          2          3          4          5          6
## 0.10300383 0.39255581 0.13795547 0.08454749 0.56042736 0.12717657
##          7          8          9         10
## 0.65497764 0.22568340 0.75191062 0.10395641
```