Final project banking data

Trisha Winquist

2023-05-28

```
#**Overall, write a coherent narrative that tells a story with the data as you complete this section.**

####As much as data science is playing a pivotal role everywhere, banking also finds it prominent
#**application. 401 k data  is the top ways of saving money for the future. Its very interesting
#**the difference in age group and the Gender that plays a role. The objective of this project is
#**to check the correlation of Age and Gender and the deferral amount.**

####I used a data set from the bank that wants to use data to determine if they should add some more
#**training documents for people who are unsure of their involvement in the retirement funds. **

#**Summarize the problem statement you addressed.**

#**The dataset has 667 data points and 9 variables. out of the 9 variables or features of this dataset,
#**One is the Gender of the participation, another is the Age of the participants. I am going to need
#** to change the Gender of Male to female as a 1 or a 2 to be consistent with the numeric code.**

library(readxl)
library(latexpdf)
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
```

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/Winquistt/OneDrive - Tyson Online/Desktop/Data Science/Statistics for data science")
getwd()
```

```
## [1] "C:/Users/Winquistt/OneDrive - Tyson Online/Desktop/Data Science/Statistics for data science"
```

```
banking_df <- read_excel("Final project/401K-Contributions.xlsx", sheet="2022", skip = 0)
str(banking_df)
```

```
## tibble [667 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Gender             : chr [1:667] "Male" "Female" "Male" "Male" ...
##  $ Match Group        : chr [1:667] "PRIOR JULY 16" "PRIOR JULY 16" "PRIOR JULY 16" "PRIOR JULY 16"
##  $ Status             : chr [1:667] "ACTIVE" "ACTIVE" "ACTIVE" "ACTIVE" ...
```

```
##  $ Age                : num [1:667] 59 41 51 61 40 53 63 58 33 39 ...
##  $ Years of Service   : num [1:667] 23 16 8 13 14 6 9 12 0 13 ...
##  $ Salary             : num [1:667] 125725 91046 122670 283650 181155 ...
##  $ Deferral           : num [1:667] 20116 10925 28890 25110 21739 ...
##  $ Prior July 16 Match: num [1:667] 2515 1821 2453 5673 3623 ...
##  $ After July 16 Match: num [1:667] 0 0 0 0 0 ...
```

```
nrow(banking_df)
```

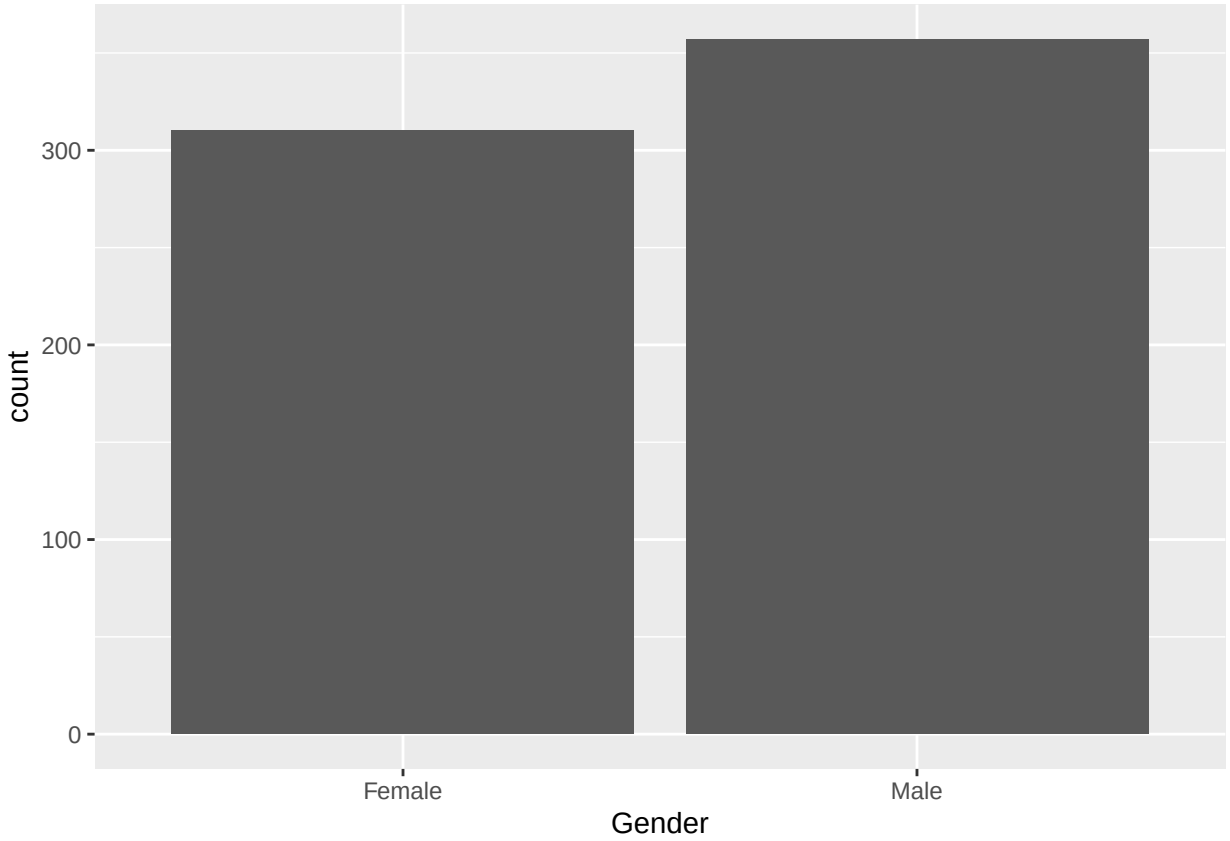
```
## [1] 667
```

```
ncol(banking_df)
```

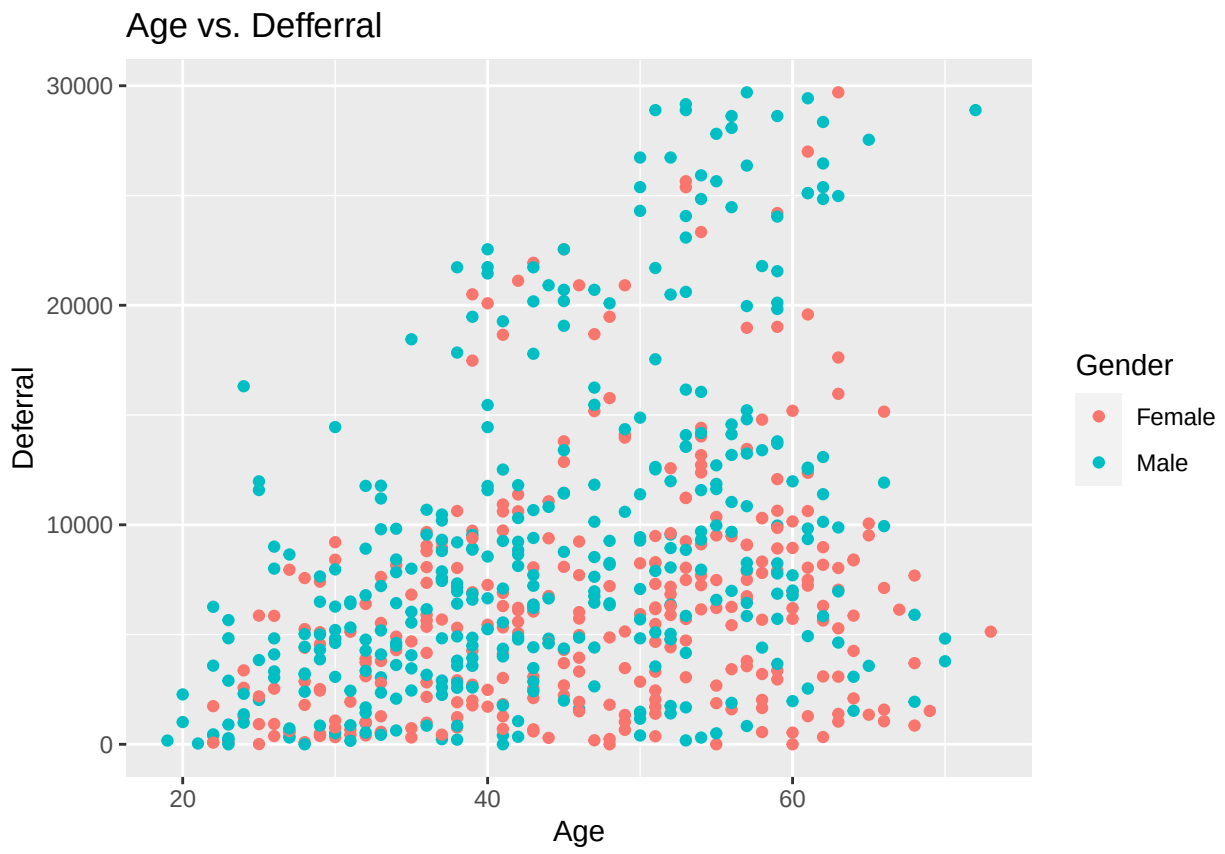
```
## [1] 9
```

```
#**Summarize how you addressed this problem statement** (the data used and the methodology employed,
#**including a recommendation for a model that could be implemented)**
#** In this project, I first analyzed the data and looked for any clean up. At first I didnt think
#** I needed to clean up any of the data but I wanted to make the Gender into either a 1 for males
#** or 0 for females. I wanted to know what the correlation to a male or female putting money into
#** their retirement fund. or if the age between male and females had any correlation. I created
#**a boxplot to show the contribution rate by age and gender **

banking_df$Gender[banking_df$Gender==1]<-"male"
banking_df$Gender[banking_df$Gender==0]<-"female"
#Create a new column with binary values
banking_df$GenderBinary <- ifelse(banking_df$Gender=="Female",1,0)

#create a box plot to show contribution rate by age and gender

ggplot(banking_df, aes(x=Age, y= Gender)) + geom_point() + geom_boxplot() + ggtitle("Age VS. Gender")
```


```
#create a geom_bar plot for the number of records for each Gender
ggplot(banking_df, aes(Gender)) + geom_bar()
```




```
ggplot(banking_df, aes(x=Age, y=Deferral, col=Gender)) + geom_point() + ggtitle("Age vs. Defferral") +
  xlab("Age") + ylab("Deferral")
```




```
head(banking_df)
```

```
## # A tibble: 6 x 10
##   Gender `Match Group` Status   Age `Years of Service`  Salary Deferral
##   <chr>  <chr>         <chr>  <dbl>              <dbl>   <dbl>    <dbl>
## 1 Male   PRIOR JULY 16 ACTIVE    59                 23 125725.   20116.
## 2 Female PRIOR JULY 16 ACTIVE    41                 16  91046.   10925.
## 3 Male   PRIOR JULY 16 ACTIVE    51                  8 122670.   28890
## 4 Male   PRIOR JULY 16 ACTIVE    61                 13 283650    25110
## 5 Male   PRIOR JULY 16 ACTIVE    40                 14 181155.   21739.
## 6 Male   AFTER JULY 16 ACTIVE    53                  6 206107.   20611.
## # i 3 more variables: `Prior July 16 Match` <dbl>, `After July 16 Match` <dbl>,
## #   GenderBinary <dbl>
```

```
tail(banking_df)
```

```
## # A tibble: 6 x 10
##   Gender `Match Group` Status             Age `Years of Service` Salary Deferral
##   <chr>  <chr>         <chr>            <dbl>              <dbl>  <dbl>    <dbl>
## 1 Male   PRIOR JULY 16 TERMINATED = ZE~    68                  7 21527.    1937.
## 2 Male   AFTER JULY 16 TERMINATED = ZE~    55                  2  6251.     500.
## 3 Female AFTER JULY 16 TERMINATED = ZE~    30                  1 19072.     763.
## 4 Male   PRIOR JULY 16 ACTIVE             39                  7 80220.    2594.
## 5 Female PRIOR JULY 16 TERMINATED          41                  9 85042.     701.
## 6 Male   AFTER JULY 16 TERMINATED = ZE~    27                  1  9031.     722.
## # i 3 more variables: `Prior July 16 Match` <dbl>, `After July 16 Match` <dbl>,
## #   GenderBinary <dbl>
```

```r
Avg_age <- aggregate(banking_df$Deferral, list(banking_df$Age), FUN = mean)
Avg_age
```

```
##    Group.1          x
## 1       19   168.4210
## 2       20  1643.2569
## 3       21    37.0327
## 4       22  1800.4641
## 5       23  1855.4358
## 6       24  4022.9285
## 7       25  4798.9217
## 8       26  4197.1458
## 9       27  2738.0217
## 10      28  3352.9204
## 11      29  3678.3270
## 12      30  5105.8630
## 13      31  2985.2499
## 14      32  3938.6675
## 15      33  4571.8873
## 16      34  5440.7834
## 17      35  5504.2211
## 18      36  5904.0888
## 19      37  6255.5955
## 20      38  5789.3453
## 21      39  7480.1679
## 22      40 11355.3919
## 23      41  6528.0266
## 24      42  7890.0530
## 25      43  7742.6568
## 26      44  7991.6997
## 27      45 11345.8886
## 28      46  5978.4926
## 29      47 10119.9899
## 30      48  8302.2791
## 31      49  8555.3030
## 32      50  9440.4060
## 33      51  8242.3023
## 34      52  8561.0339
## 35      53 13152.2991
## 36      54 12636.7307
## 37      55  9654.7066
## 38      56 11921.3435
## 39      57 11189.4303
## 40      58  8017.1886
## 41      59 12855.1348
## 42      60  6855.8388
## 43      61 13238.9220
## 44      62 12244.8785
## 45      63 10430.6914
## 46      64  4800.7098
## 47      65 10407.4771
## 48      66  7794.2712
## 49      67  6133.3580
## 50      68  4015.2452
```

```
## 51      69  1518.9984
## 52      70  4300.2046
## 53      72 28890.0000
## 54      73  5130.6204
```

```
Age_df <- mean(banking_df$Age)
Age_df
```

```
## [1] 45.21139
```

```
YearsOfService_df <- mean(banking_df$`Years of Service`)
YearsOfService_df
```

```
## [1] 7.475262
```

```
#**The boxplot shows that males participate more than females in the 401 k retirement contribution.
#*Its interesting that males start participating at an earlier age and have the retirement funds needed
#*to maybe retire ealier?
#*A boxplot is a standardized way of displaying the distribution of data based on a five number
#*summary ("minimum", first quartile [Q1], median, third quartile [Q3] and "maximum"). It can
#*tell you about your outliers and what their values are. Boxplots can also tell you if your data
#*is symmetrical, how tightly your data is grouped and if and how your data is skewed. A boxplot is
#*a graph that gives you a good indication of how the values in the data are spread out.THe data
#*indicates that males also start contributing to their retirement at an earlier age.The Scatter
#* plot graph shows a nice visual of males contributing to their 401 k's more **

#**Summarize the interesting insights that your analysis provided. **
#*
#*What I found Interesting with the data, you would think that the average of each age group had no
#*determination of how much money people #deffered. For instance the average 18 year old deferred more
#*money than the average age groups between 19-24 and was better than other average #age groups.
#*There was a small correlation to the age group to how much people deferred. **

##Correlation

cor(banking_df$Age, banking_df$`Years of Service`)
```

```
## [1] 0.4148268
```

```
##Shapiro-wilk normality test for defferal
library(ggpubr)
shapiro.test(banking_df$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  banking_df$Age
## W = 0.97748, p-value = 1.319e-08
```

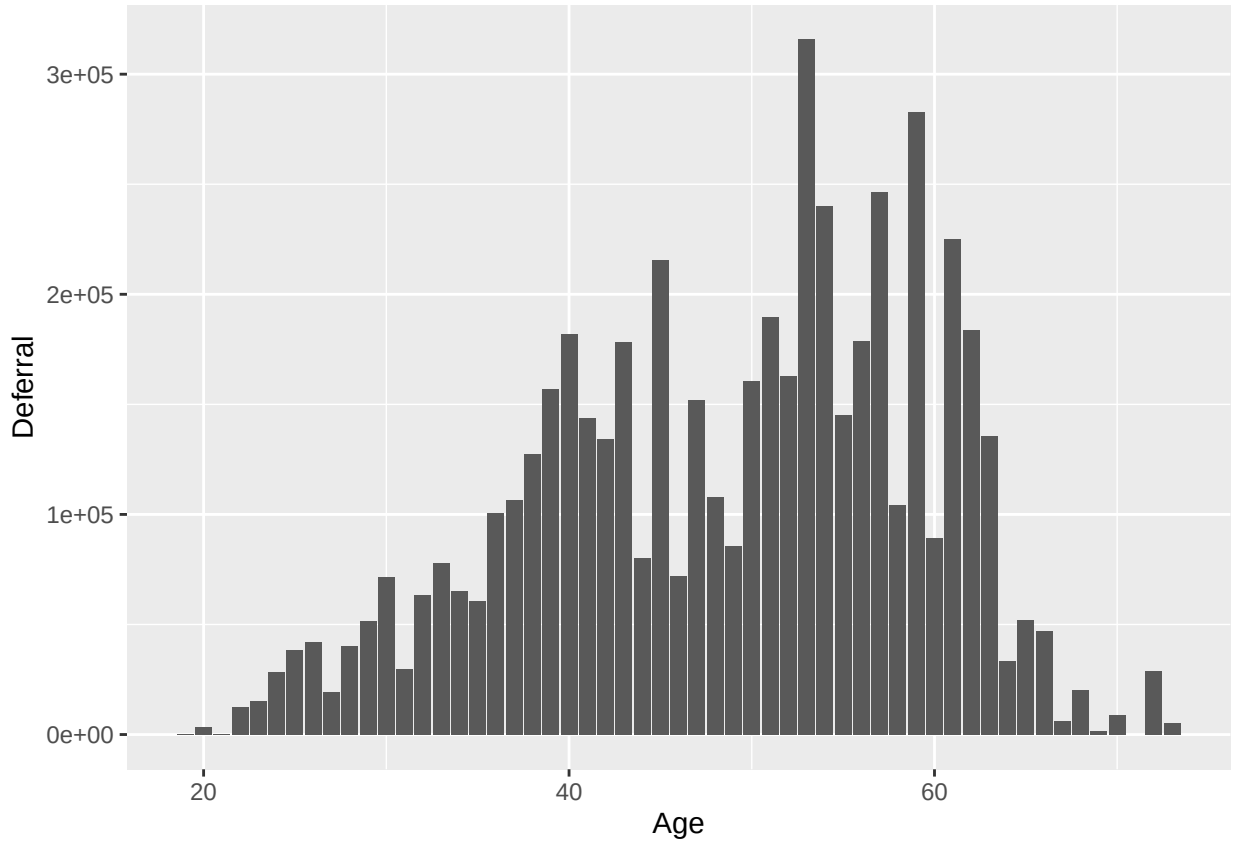
```
shapiro.test(banking_df$Deferral)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  banking_df$Deferral
## W = 0.87507, p-value < 2.2e-16
```

```
head(banking_df)
```

```
## # A tibble: 6 x 10
##   Gender `Match Group` Status   Age `Years of Service`  Salary Deferral
##   <chr>  <chr>         <chr>  <dbl>              <dbl>   <dbl>    <dbl>
## 1 Male   PRIOR JULY 16 ACTIVE    59                 23 125725.   20116.
## 2 Female PRIOR JULY 16 ACTIVE    41                 16  91046.   10925.
## 3 Male   PRIOR JULY 16 ACTIVE    51                  8 122670.   28890
## 4 Male   PRIOR JULY 16 ACTIVE    61                 13 283650    25110
## 5 Male   PRIOR JULY 16 ACTIVE    40                 14 181155.   21739.
## 6 Male   AFTER JULY 16 ACTIVE    53                  6 206107.   20611.
## # i 3 more variables: `Prior July 16 Match` <dbl>, `After July 16 Match` <dbl>,
## #   GenderBinary <dbl>
```

```
ggplot(data=banking_df,aes(x=Age,y=Deferral,group=1)) + geom_col()
```




```
#**I would like to do a prediction model for the 401 k deferrals. **
library(foreign)

banking_glm <- glm(as.factor(Deferral)~ Age
                   + GenderBinary + `Years of Service`, data=banking_df, family = binomial)
summary(banking_glm)
```

```
##
## Call:
## glm(formula = as.factor(Deferral) ~ Age + GenderBinary + `Years of Service`,
##     family = binomial, data = banking_df)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1832   0.1204   0.1321   0.1461   0.1754
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.85470    1.59355   2.419   0.0156 *
## Age                 0.02350    0.03975   0.591   0.5545
## GenderBinary       -0.20440    0.82893  -0.247   0.8052
## `Years of Service` -0.01155    0.05772  -0.200   0.8415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.478  on 666  degrees of freedom
## Residual deviance: 68.078  on 663  degrees of freedom
## AIC: 76.078
##
## Number of Fisher Scoring iterations: 7
```

```
#**We can interpret the negative binomial regression coefficient as follows: for a one unit
#* change in the predictor variable,the difference in the logs of expected counts of the
#* response variable is expected to change by the respective regression coefficient,
#* given the other predictor variables in the model are held**

#**Summarize the implications to the consumer (target audience) of your analysis.**
#*The intent of this project was to find out who needs more training to add money to their
#*401 k retirement folders.The bank has a training #program but may only target new hires
#*and not focus on the different factors like Gender and Years of service along with age. The average
#*age #of participation is 45.2 years with the average years of
#* service of 7.47 years **

#**Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
#*I think there are many ways to look at this simple data. One would be to help with a training
#*program to help the team understand the full potential of their retirement funds.
#*If I had more experience I would like to take each year and try to figure out if the covid year
#*had anything to do with retirement funds. In future we can look into the benifits that correlate with
#* implementation of the deferal program
```