Bank EDA

Megan Riley

3/13/2020

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- t
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------- tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggplot2)
library(dplyr)
library(here)
```

```
## here() starts at /Users/zmartygirl/data/MSDSR/Stats6372Project/Stats-6372-Project-Two
```

```
root = here()


bank_20 = read.csv(paste(root,"/data/bank-additional/bank-additional-full.csv", sep = ""), sep=";")

bank_17 = read.csv(paste(root,"/data/bank/bank-full.csv", sep = ""), sep = ";")
```

Summary

Unknown whether we should work with both data sets or if Dr. Turner is good with us choosing one. My vote is for bank_20 if we can choose.

Variable Notes: -Duration is a variable not known until Y is determined, duration is the duration of the call when attempting to sell the term deposit package.
- No NAs, uses unknown in places otherwise -Campaign is # of contacts, minimum 1 b/c it includes this contact in the data, even if hte contact was unsuccessful -pdays needs to be potentially cleaned where 999 should equal NA or potentially switched to a categorical variable -Do not understand some of the later variables, seem to be more socially based.

```
#Dr Turner is heavily requesting summary stats
summary(bank_20)
```

```
##       age                  job            marital
##  Min.   :17.00   admin.     :10422   divorced: 4612
##  1st Qu.:32.00   blue-collar: 9254   married :24928
##  Median :38.00   technician : 6743   single  :11568
##  Mean   :40.02   services   : 3969   unknown :   80
##  3rd Qu.:47.00   management : 2924
##  Max.   :98.00   retired    : 1720
##                  (Other)    : 6156
##                education        default         housing
##  university.degree  :12168   no     :32588   no     :18622
##  high.school        : 9515   unknown: 8597   unknown:  990
##  basic.9y           : 6045   yes    :    3   yes    :21576
##  professional.course: 5243
##  basic.4y           : 4176
##  basic.6y           : 2292
##  (Other)            : 1749
##       loan            contact          month       day_of_week
##  no     :33950   cellular :26144   may    :13769   fri:7827
##  unknown:  990   telephone:15044   jul    : 7174   mon:8514
##  yes    : 6248                     aug    : 6178   thu:8623
##                                    jun    : 5318   tue:8090
##                                    nov    : 4101   wed:8134
##                                    apr    : 2632
##                                    (Other): 2016
##     duration         campaign          pdays          previous
##  Min.   :   0.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.000
##  1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000
##  Median : 180.0   Median : 2.000   Median :999.0   Median :0.000
##  Mean   : 258.3   Mean   : 2.568   Mean   :962.5   Mean   :0.173
##  3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000
##  Max.   :4918.0   Max.   :56.000   Max.   :999.0   Max.   :7.000
##
##         poutcome      emp.var.rate      cons.price.idx  cons.conf.idx
##  failure    : 4252   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
##  nonexistent:35563   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
##  success    : 1373   Median : 1.10000   Median :93.75   Median :-41.8
##                      Mean   : 0.08189   Mean   :93.58   Mean   :-40.5
##                      3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
##                      Max.   : 1.40000   Max.   :94.77   Max.   :-26.9
##
##    euribor3m      nr.employed     y
##  Min.   :0.634   Min.   :4964   no :36548
##  1st Qu.:1.344   1st Qu.:5099   yes: 4640
##  Median :4.857   Median :5191
##  Mean   :3.621   Mean   :5167
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
##
```

```
#Does not look like any NAs in either data set
sapply(bank_20, function(x) sum(is.na(x)))
```

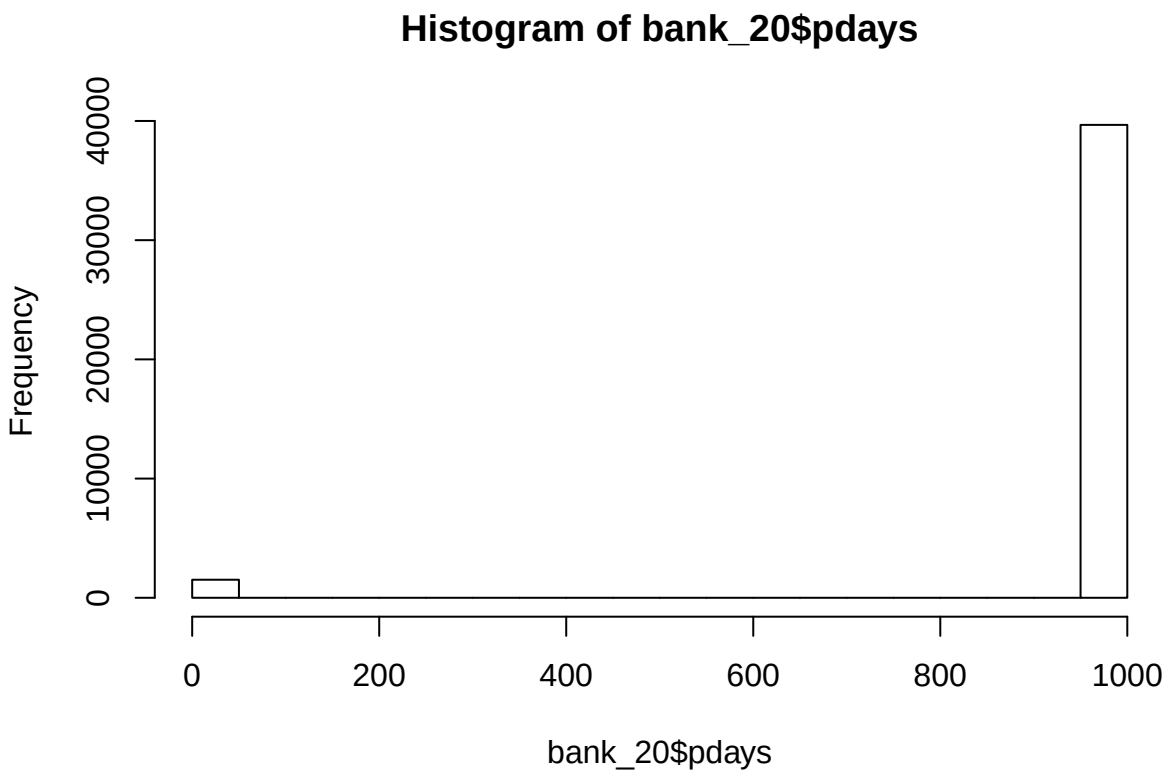
```
##            age            job        marital      education        default
##              0              0              0              0              0
##        housing           loan        contact          month    day_of_week
```

```
##             0             0             0             0             0
##      duration      campaign         pdays      previous      poutcome
##             0             0             0             0             0
##   emp.var.rate cons.price.idx  cons.conf.idx     euribor3m  nr.employed
##             0             0             0             0             0
##             y
##             0
```

```
sapply(bank_17, function(x) sum(is.na(x)))
```

```
##       age       job   marital education   default   balance   housing
##         0         0         0         0         0         0         0
##      loan   contact       day     month  duration  campaign     pdays
##         0         0         0         0         0         0         0
##  previous  poutcome         y
##         0         0         0
```

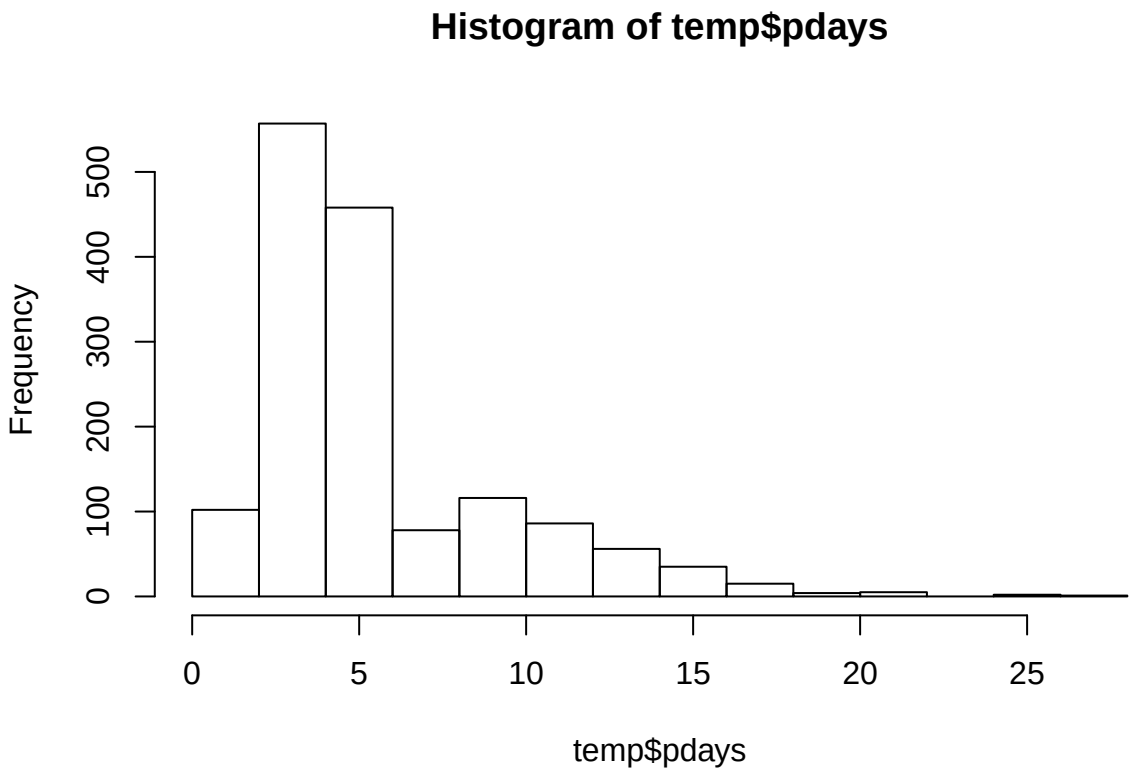
```
#Set any predictions without using Duration or Y


#pdays- about 40k of the 41k are at level 999, no previous contact
#could bin this data
hist(bank_20$pdays)
```



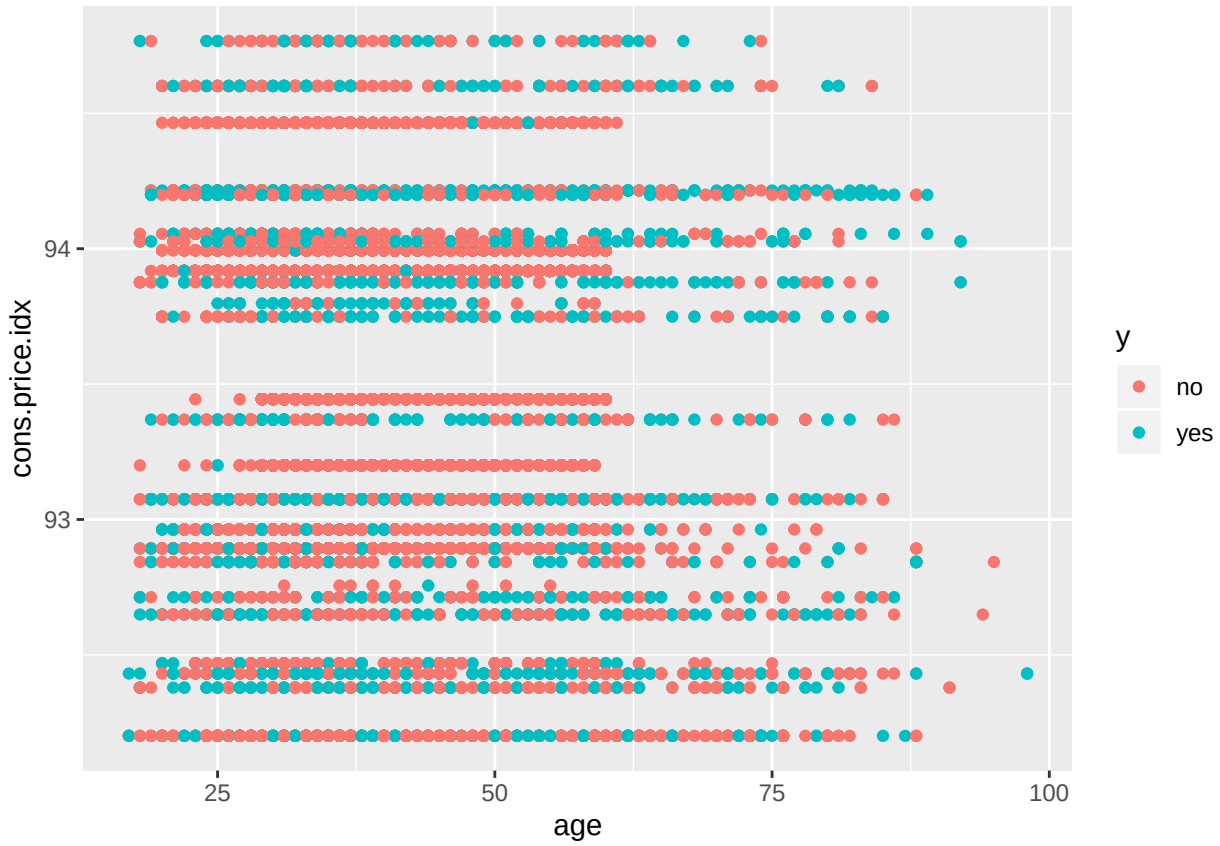

```
temp = bank_20 %>% filter(pdays != 999)
dim(temp)
```

```
## [1] 1515   21
```

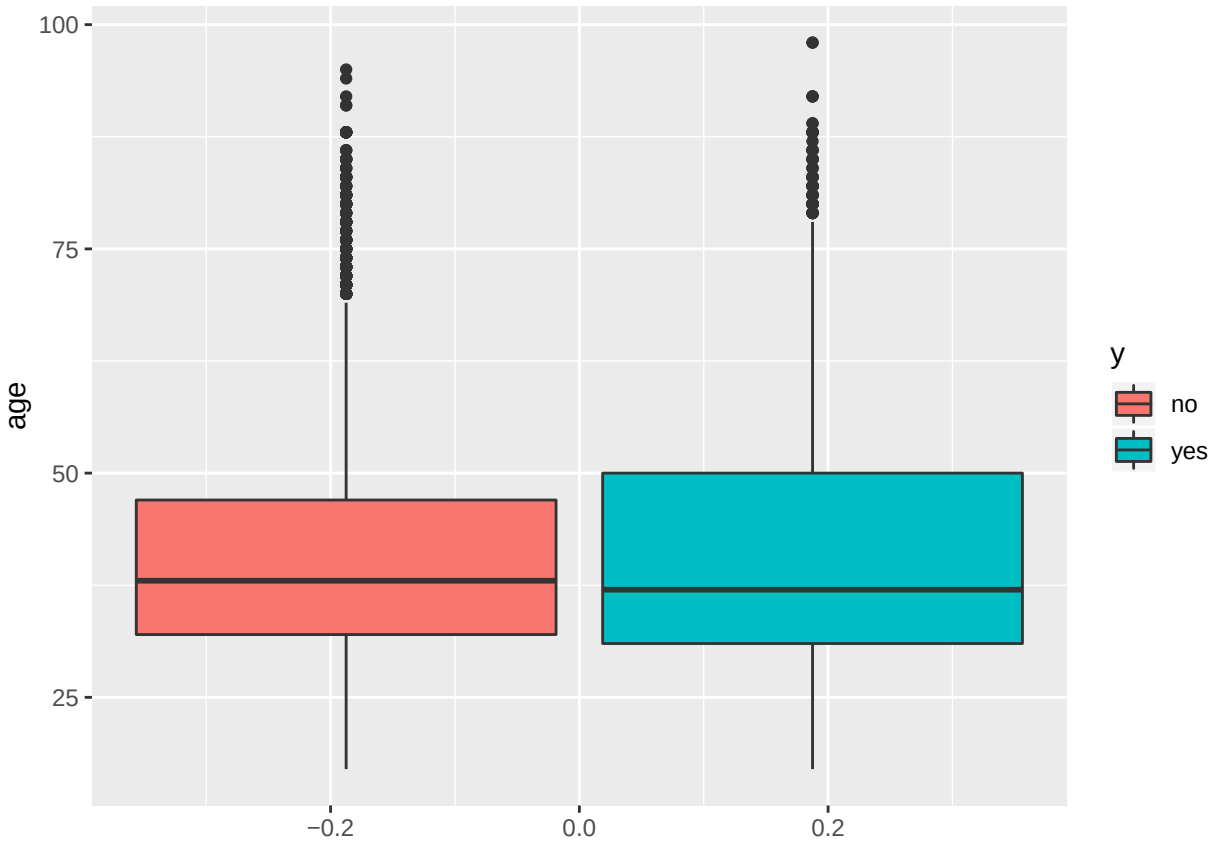
```r
hist(temp$pdays)
```



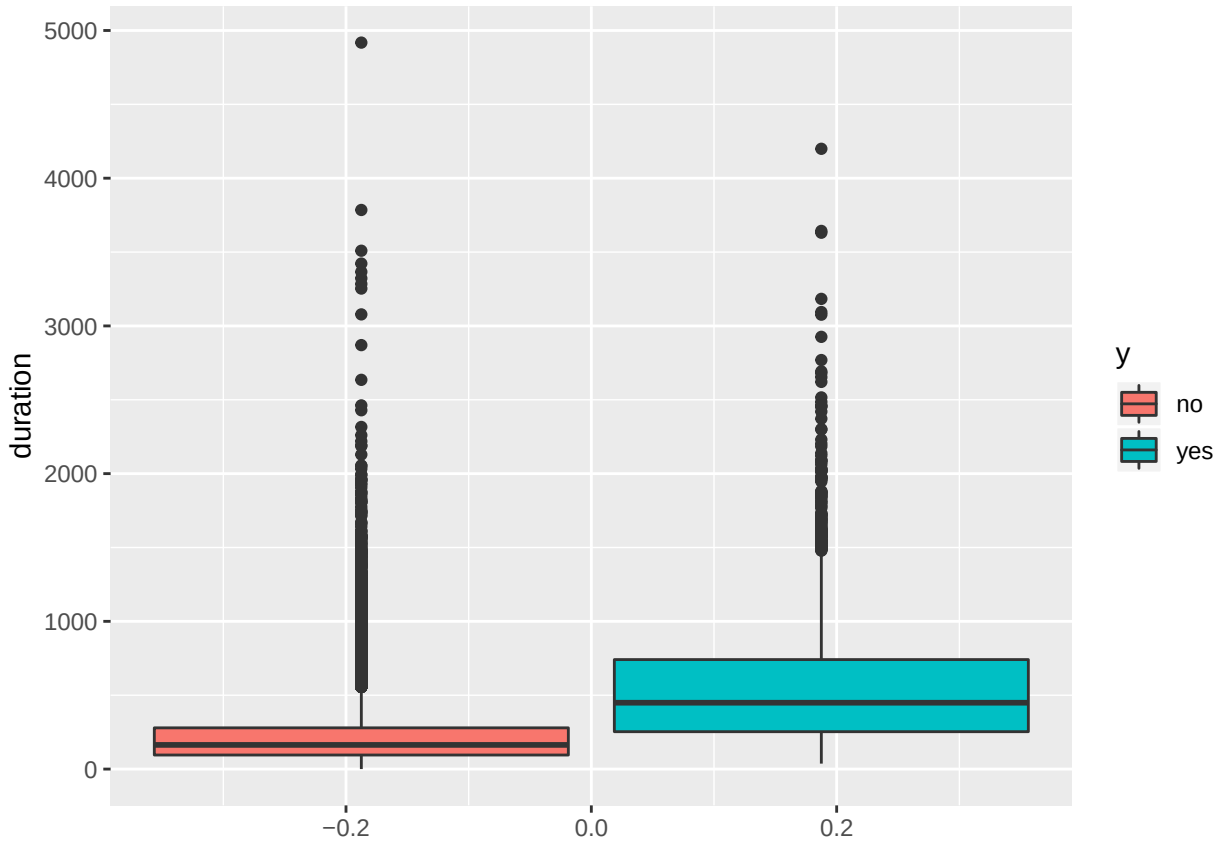

```r
bank_20 %>% ggplot(aes(x = age, y = cons.price.idx, color = y)) + geom_point()
```




```
bank_20 %>% ggplot(aes(y = age, fill = y)) + geom_boxplot()
```




```
bank_20 %>% ggplot(aes(y = duration, fill = y)) + geom_boxplot()
```



Predicting Duration, in order to predict calls.

Basically if we know there is a relationship between duration and the result we are predicting, we can use information that explains duration to therefore explain response.

```
duration_model = lm(duration ~ ., data = bank_20)
```

Uneven Split in outcomes

Yes happens about 10% of the time, where no is the response 90% of the time. This unbalance makes it difficult to predict. - Can balance the train/test split, what else have we learned about predicting unbalanced outcomes?

```
yes_answer = bank_20 %>% filter(y == "yes")

no_answer_all = bank_20 %>% filter(y == "no")
no_indices = sample(dim(no_answer_all)[1], dim(yes_answer)[1])
no_answer = no_answer_all[no_indices,]

balanced_bank_20 = rbind(yes_answer, no_answer)

balanced_indices = sample(dim(balanced_bank_20)[1], round(dim(balanced_bank_20)[1] * .1 ))
balanced_test = balanced_bank_20[balanced_indices,]
balanced_train = balanced_bank_20[-balanced_indices,]
```