Stats2 Project2

# Load libraries all in one place using pacman package

library(pacman)  
p\_load("tidyverse", "ggplot2", "caret", "tidyr", "dplyr", "e1071", "class", "gridExtra", "GGally",  
 "summarytools", "magrittr")

## Load the dataset

You can also embed plots, for example:

library(readr)  
bank\_additional\_full <- read.csv("bank-additional-full.csv", sep=";")  
dim(bank\_additional\_full) #count the columns

## [1] 41188 21

summary(bank\_additional\_full)

## age job marital education   
## Min. :17.00 Length:41188 Length:41188 Length:41188   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41188 Length:41188 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41188 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08189   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228   
## y   
## Length:41188   
## Class :character   
## Mode :character   
##   
##   
##

str(bank\_additional\_full)

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : chr "housemaid" "services" "services" "admin." ...  
## $ marital : chr "married" "married" "married" "married" ...  
## $ education : chr "basic.4y" "high.school" "high.school" "basic.6y" ...  
## $ default : chr "no" "unknown" "no" "no" ...  
## $ housing : chr "no" "no" "yes" "no" ...  
## $ loan : chr "no" "no" "no" "no" ...  
## $ contact : chr "telephone" "telephone" "telephone" "telephone" ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ day\_of\_week : chr "mon" "mon" "mon" "mon" ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : chr "no" "no" "no" "no" ...

# Analyzing the data  
print(dfSummary(bank\_additional\_full, graph.magnif = 0.75), method = 'browser')

## Output file written: /var/folders/2g/drgkln6d0md48\_mrm74mc7c00000gn/T//Rtmpd7qrHm/fileeb4e642517e0.html

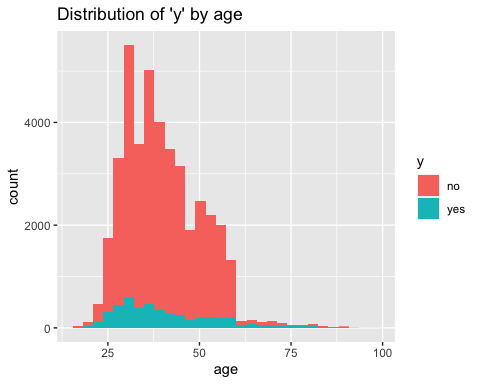
# "browser" means load the thing in a separate browser instead of embedding it in here, cos its big and it will jsut fill your document

Looking at the dfsummary, there doesnt seem to be missing data in terms of just not having values. However, there are some fields that have explicit unknown or non-existent classes that could be considered as ‘missing’. For example, loan and housing have 990 “unknown” values. And ‘default’ has 8597 “unknown” values representing 20.9%

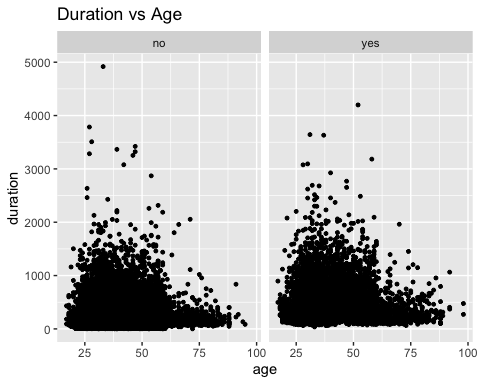
wondering what is ‘pdays’ which has a median of 999 but 27 distinct values and appears to be bimodal

ggplot(bank\_additional\_full) +   
 geom\_histogram(mapping = aes(x=age, fill=y)) +   
 ggtitle("Distribution of 'y' by age")

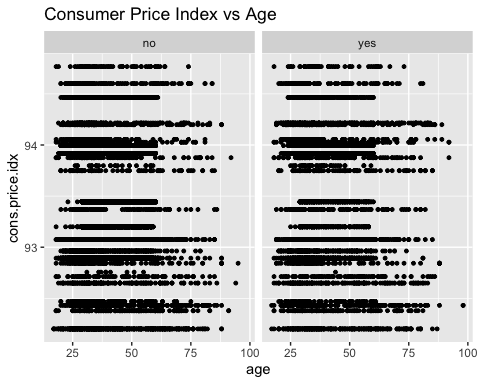
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



df\_yes <- bank\_additional\_full %>%  
 filter(y=="yes")  
#summary(df\_yes)  
  
# Nothing interesting found in the below code so commenting it out  
# ggplot(bank\_additional\_full, aes(x=age, y=emp.var.rate)) +  
# geom\_point(size=1, shape="circle") +  
# ggtitle("Employment Variation Rate vs Age") +   
# facet\_wrap(~ y)  
  
ggplot(bank\_additional\_full, aes(x=age, y=duration)) +  
 geom\_point(size=1, shape="circle") +   
 ggtitle("Duration vs Age") +   
 facet\_wrap(~ y)



ggplot(bank\_additional\_full, aes(x=age, y=cons.price.idx)) +  
 geom\_point(size=1, shape="circle") +   
 ggtitle("Consumer Price Index vs Age") +   
 facet\_wrap(~ y)

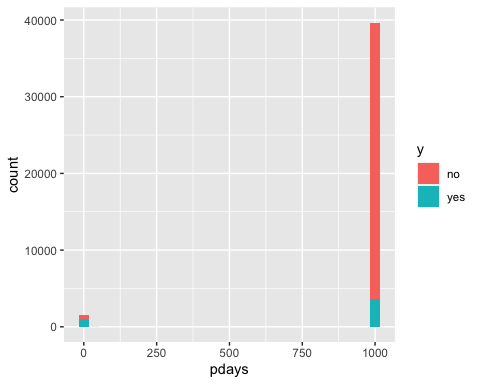


# ggplot(bank\_additional\_full, aes(x=age, y=education)) +  
# geom\_point(size=1, shape="circle") +   
# ggtitle("Education vs Age") +   
# facet\_wrap(~ y)

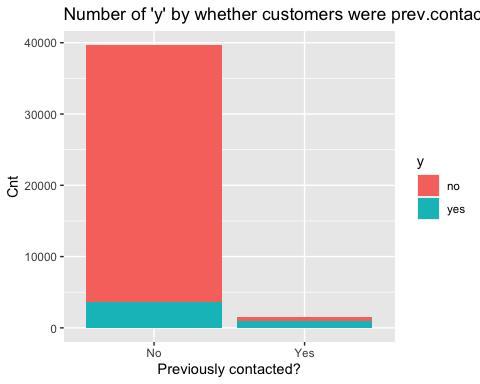
Duration vs Age: The duration of last contact (in seconds) was longer for ages 25-50. And it was understandably longer for “yes” vs for “no”

#Analyzing pdays  
ggplot(bank\_additional\_full) + geom\_histogram(mapping = aes(x=pdays, fill=y))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

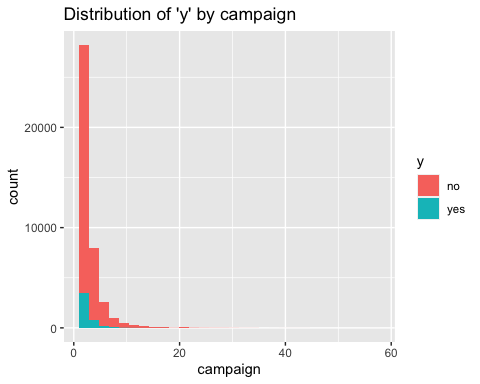
 - Analyzing ‘pdays’ ie., number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) - Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe

#prevly\_Cntctd Yes/No. TO see the distribution or 'Y' on first time contact vs. a follow up  
bank\_additional\_full$prevly\_Cntctd <- case\_when(bank\_additional\_full$pdays==999 ~ "No", !bank\_additional\_full$pdays==999 ~ "Yes")  
  
ggplot(bank\_additional\_full) +   
 geom\_bar(mapping = aes(x=prevly\_Cntctd, fill = y)) +   
 ggtitle("Number of 'y' by whether customers were prev.contacted or not") +  
 ylab("Cnt") + xlab("Previously contacted?")

 Same observation here as above: Most folks had no previous campaign but if they did, it looks like most who had a previous campaign decided to subscribe / likely to say “Yes”

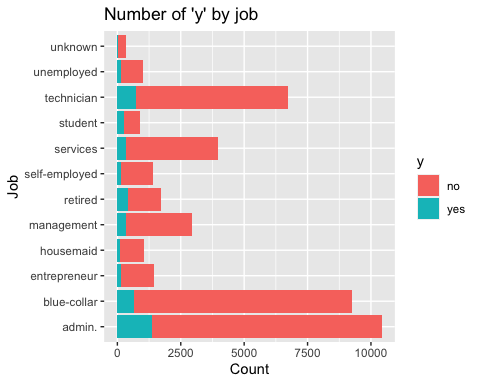
#Analysing campaign  
ggplot(bank\_additional\_full) +   
 geom\_histogram(mapping = aes(x=campaign, fill=y)) +  
 ggtitle("Distribution of 'y' by campaign")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 - Just visually, when we decided to stop contacting a person it didnt affect our closing ratio which still dropped off precipitously

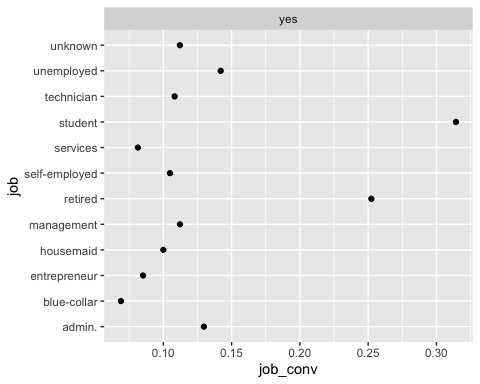
* Ideally, the campaign would stop contacting people who are less likely to subscribe, and keep contacting people if they are more likely to subscribe. Then we should see the ratio of Yes to No go up as the number of no contacts goes up. Instead, it looks like the ratio stays the same and the number of Yes’s drops proportionately with the number of No’s.

#Analyzing job  
ggplot(bank\_additional\_full) +   
 geom\_bar(mapping = aes(x=job, fill = y)) +   
 coord\_flip() + #Added coord flip here to make it more readable  
 ggtitle("Number of 'y' by job") +  
 ylab("Count") +   
 xlab("Job")



“y” - has a client subscribed a term deposit? : admin, technician and blue collar jobs are the top 3 subscribers by volume

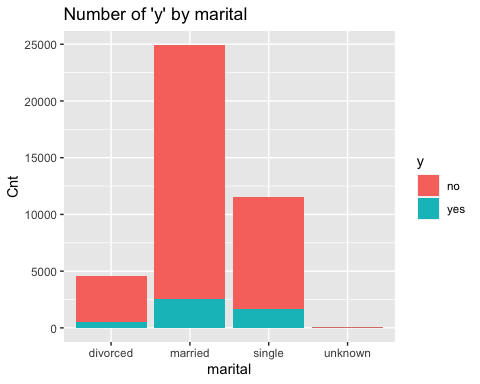
df <- bank\_additional\_full %>%  
 group\_by(job) %>%  
 count(y) %>%  
 mutate(job\_conv = n/sum(n)) %>%  
 filter(y == "yes")  
  
ggplot(df, aes(x=job, y=job\_conv)) +  
 geom\_point() +  
 coord\_flip() +  
 facet\_wrap(~ y)



Above, I looked at the ratio of “yes” vs “no” and see that students and retired persons convert at much higher rates than those of other professions. And ‘blue collar’ has the lowest conversion rate

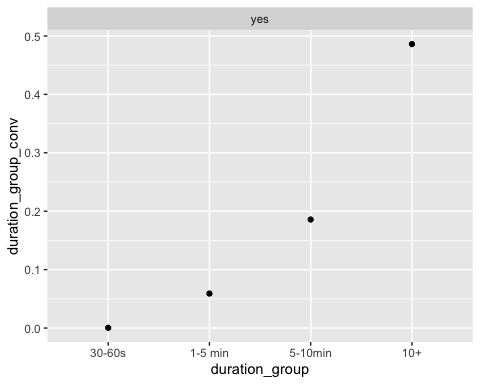
So, if they were to want to improve the cost effectiveness of their campaigns they might want to target more ‘students’ and ‘retirees’

#Analyzing marital  
ggplot(bank\_additional\_full) +   
 geom\_bar(mapping = aes(x=marital, fill = y)) +   
 ggtitle("Number of 'y' by marital") +  
 ylab("Cnt") +   
 xlab("marital")



* More ‘married’ people are represented in the campaign
* Visually looking, conversion rate seems to be higher for ‘single’ people

#Analyzing duration  
  
bank\_additional\_full$duration\_group <-   
 cut(bank\_additional\_full$duration,   
 breaks = c(5,30,60,300,600,Inf),   
 labels = c("5-30s", "30-60s", "1-5 min", "5-10min", "10+"))  
  
df <- bank\_additional\_full %>%  
 group\_by(duration\_group) %>%  
 count(y) %>%  
 mutate(duration\_group\_conv = n/sum(n)) %>%  
 filter(y == "yes")  
  
ggplot(df, aes(x=duration\_group, y=duration\_group\_conv)) +  
 geom\_point() +  
 facet\_wrap(~ y)



* Looking above, clearly conversion rate goes up the longer the most recent call

# HAVING TROUBLE BELOW NEED TO CONTINUE TOMORROW AND FIX IT

# {r} # #Running logistic regression on full data set # simple.log<-glm (y~.,family="binomial",data=bank\_additional\_full) # summary(simple.log) # #

# {r} # set.seed(123) # df <- bank\_additional\_full %>% drop\_na() # train\_idx <- createDataPartition(df$y, p=0.75, list = FALSE) # train <- bank\_additional\_full[train\_idx, ] # test <- bank\_additional\_full[-train\_idx, ] # # fit <- train(y ~ ., data = train, # method = "glm", # family = "binomial", # trControl = trainControl(method="cv", number = 10)) # summary(fit) #