Main

Megan Riley

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

```
## -- Attaching packages -------------------------------------------------------------------------- tid
```

```
## 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 ----------------------------------------------------------------------------- tidyverse_
## 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
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
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 = ";")
```

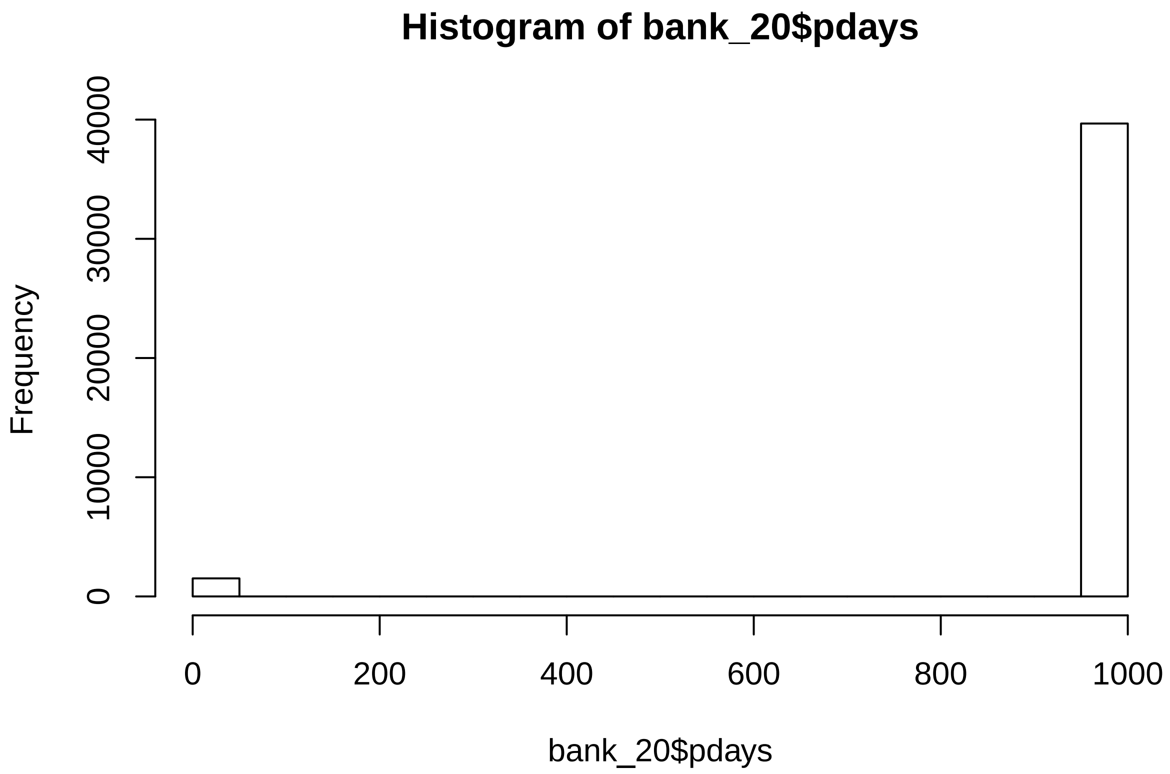
```
summary(bank_20)
```

```
##       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
```

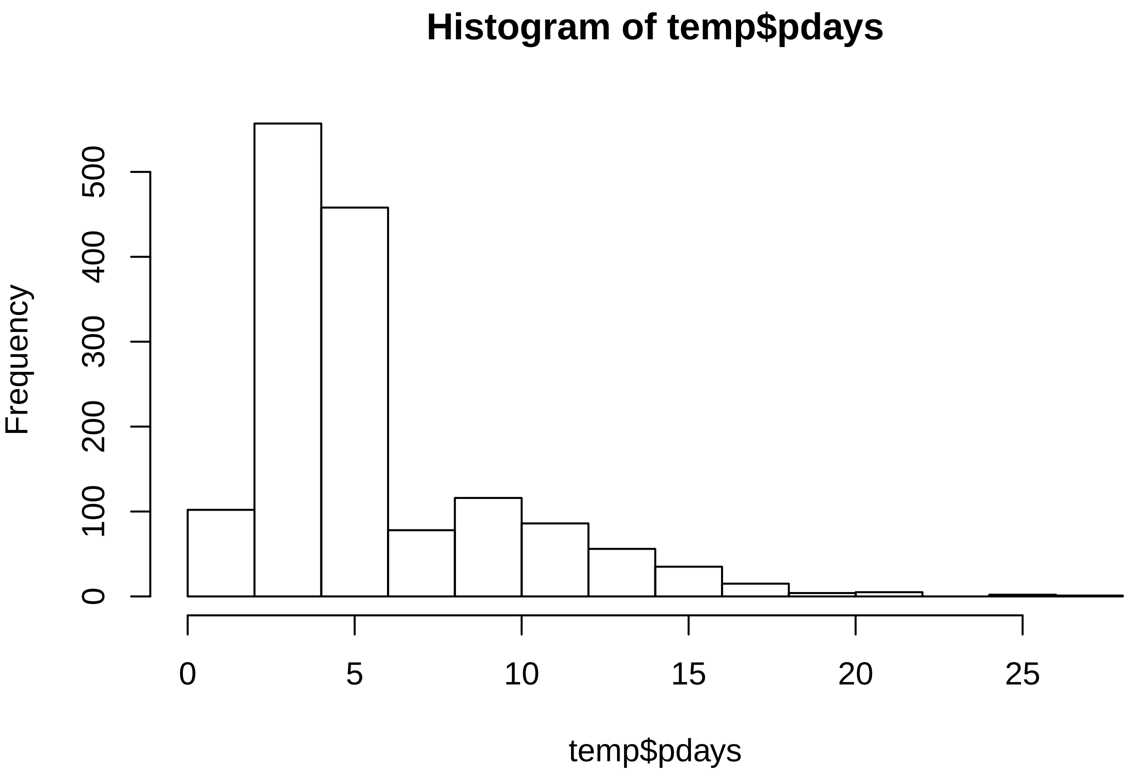
```
clean_bank_20 = as.data.frame(bank_20)
#pdays- about 40k of the 41k are at level 999, no previous contact
#could bin this data
hist(bank_20$pdays)
```



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

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

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



```
summary(temp$pdays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   3.000   6.000   6.015   7.000  27.000
```

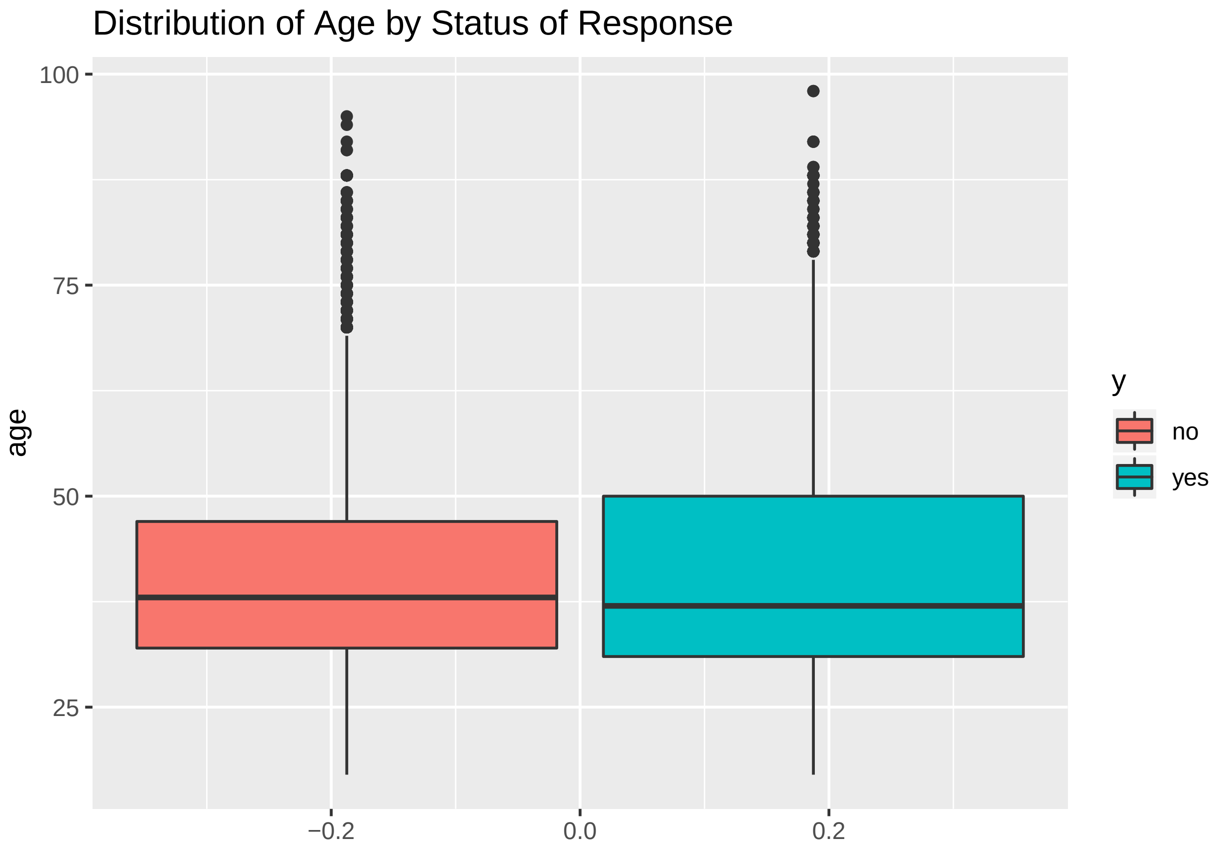
```
#within 5 days, 10 , 15, 30 and never

clean_bank_20$newpdays = case_when(bank_20$pdays == 999 ~ "Never",
                     bank_20$pdays >= 15 ~ "Within 30 Days",
                     bank_20$pdays >= 10 & bank_20$pdays < 15 ~ "Within 15 Days",
                     bank_20$pdays >= 5 & bank_20$pdays < 10 ~ "Within 10 Days",
                     bank_20$pdays < 5 ~ "Within 5 Days")

#clean_bank_20 = dplyr::select(clean_bank_20, -pdays)


#Dr Turner's other suggestion
#Set up a categorical variable to turn the continuous variable on or off.
#any use of this would have to be both in tandem
#ie y ~ altpdays_cat*altpdays_cont
alt_pdays_cat = ifelse(bank_20$pdays == 999, 0, 1)
#remains the same as original pdays,

alt_clean_bank_20 = bank_20
alt_clean_bank_20$pdays_cat = alt_pdays_cat
```

```
#Currently produces a train set of 52 n / 48 y
#90/10 yes train test split
set.seed(4567)
```

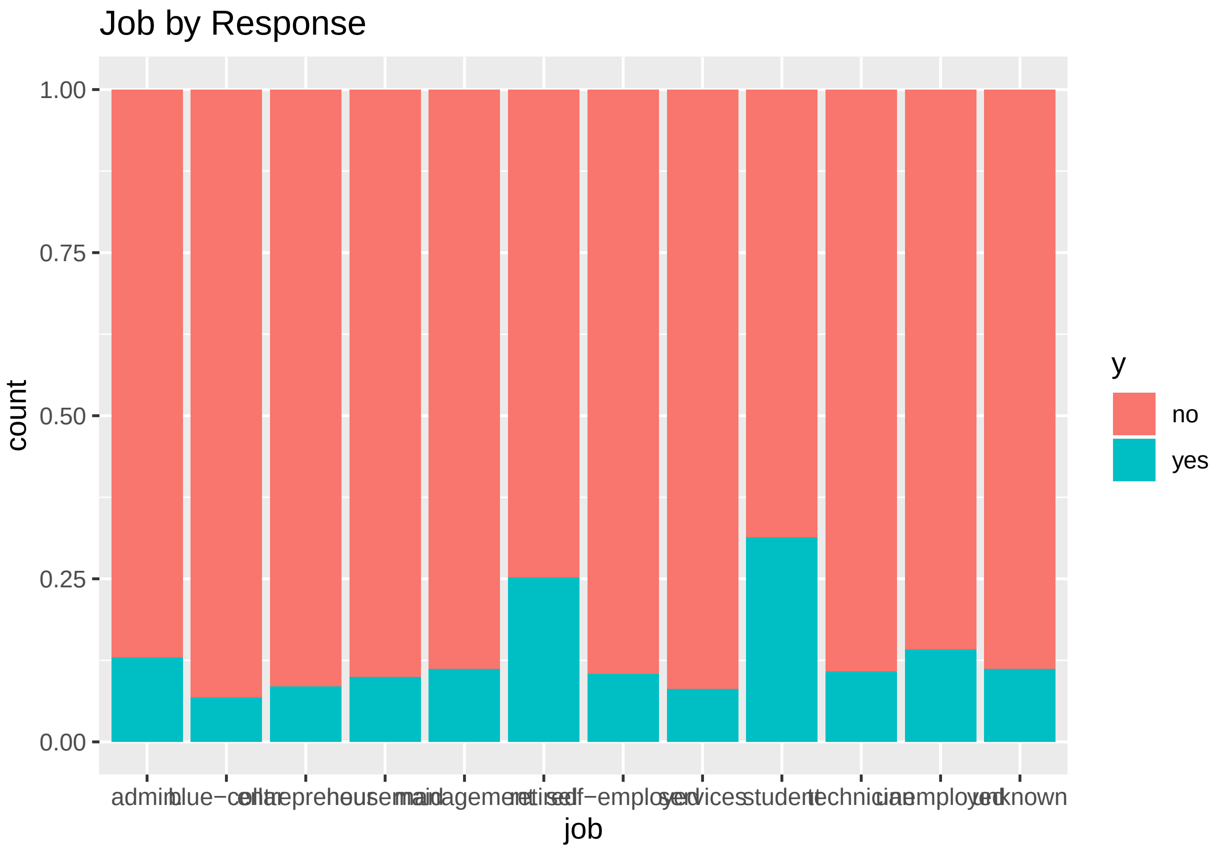
```
yes_indices = which(clean_bank_20$y == "yes")
yes_train_indices = sample(yes_indices, length(yes_indices) * .9)
no_indices = which(clean_bank_20$y == "no")
#
no_train_indices = sample(no_indices, length(yes_indices))
train_indices = c(no_train_indices,yes_train_indices)

balanced_train_bank_20 = clean_bank_20[train_indices,]

test_bank_20 = clean_bank_20[-train_indices,]
summary(balanced_train_bank_20$default)
```

```
##      no unknown     yes
##    7349    1466       1
```

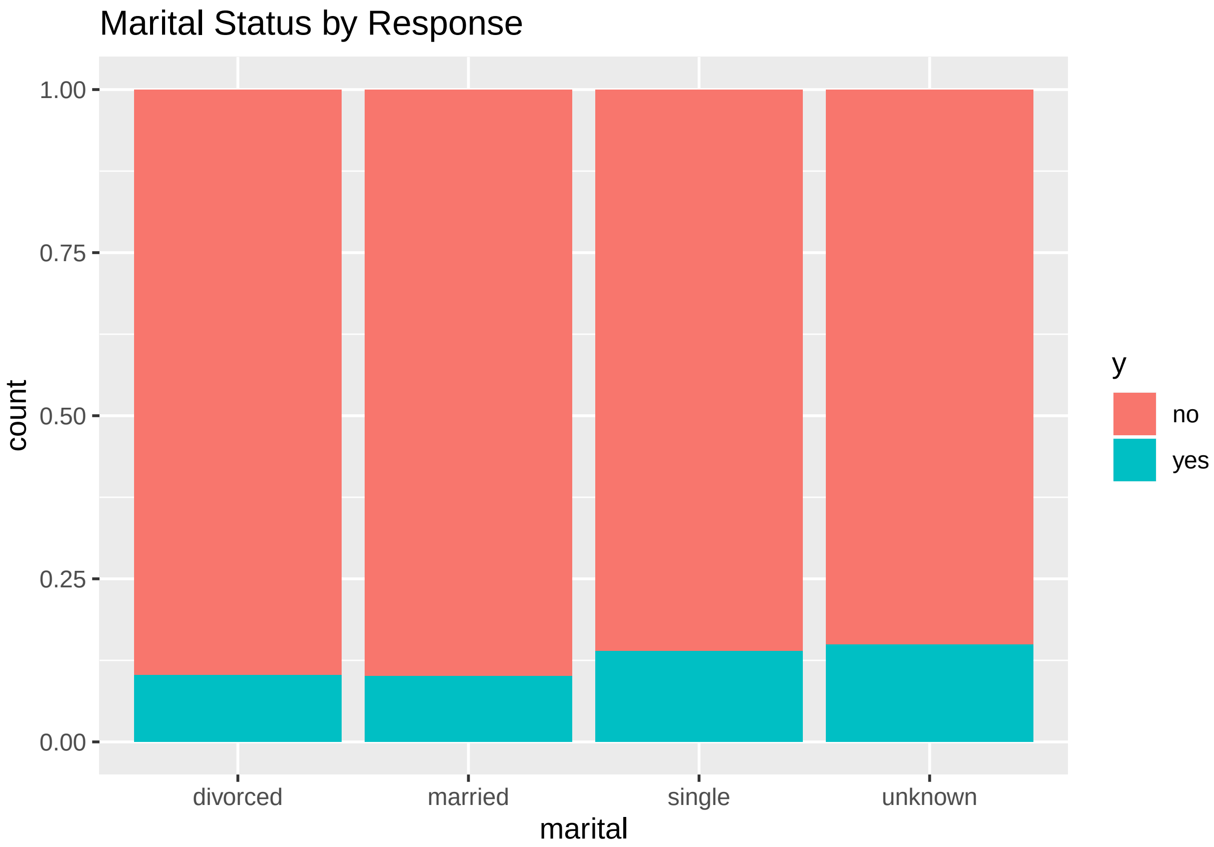
```
#Age
clean_bank_20 %>% ggplot(aes(y= age,fill = y)) + geom_boxplot() + ggtitle("Distribution of Age by Statu
```



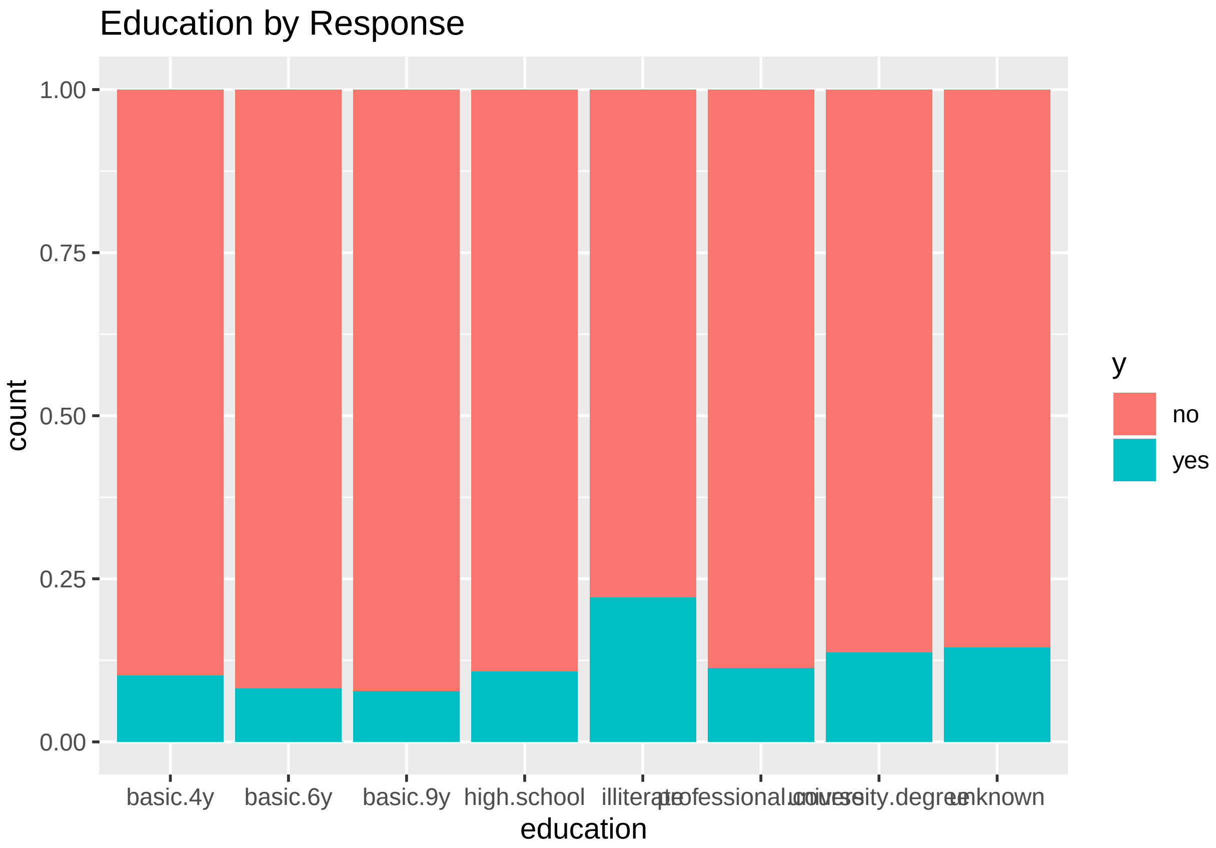
```
#Job
clean_bank_20 %>% ggplot(aes(x = job, fill = y)) + geom_bar(position = "fill") + ggtitle("Job by Respon
```



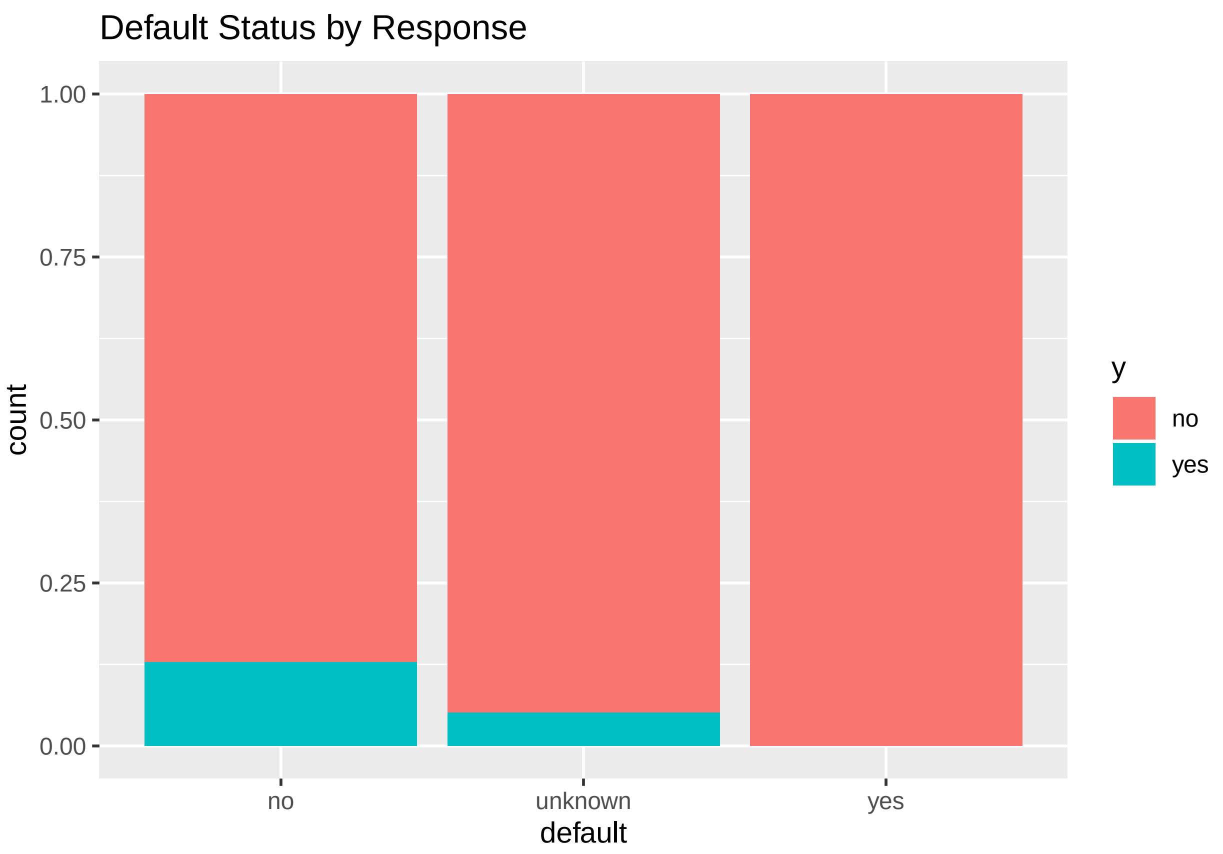
```
#Marital
clean_bank_20 %>% ggplot(aes(x = marital, fill = y)) + geom_bar(position = "fill") + ggtitle("Marital S
```



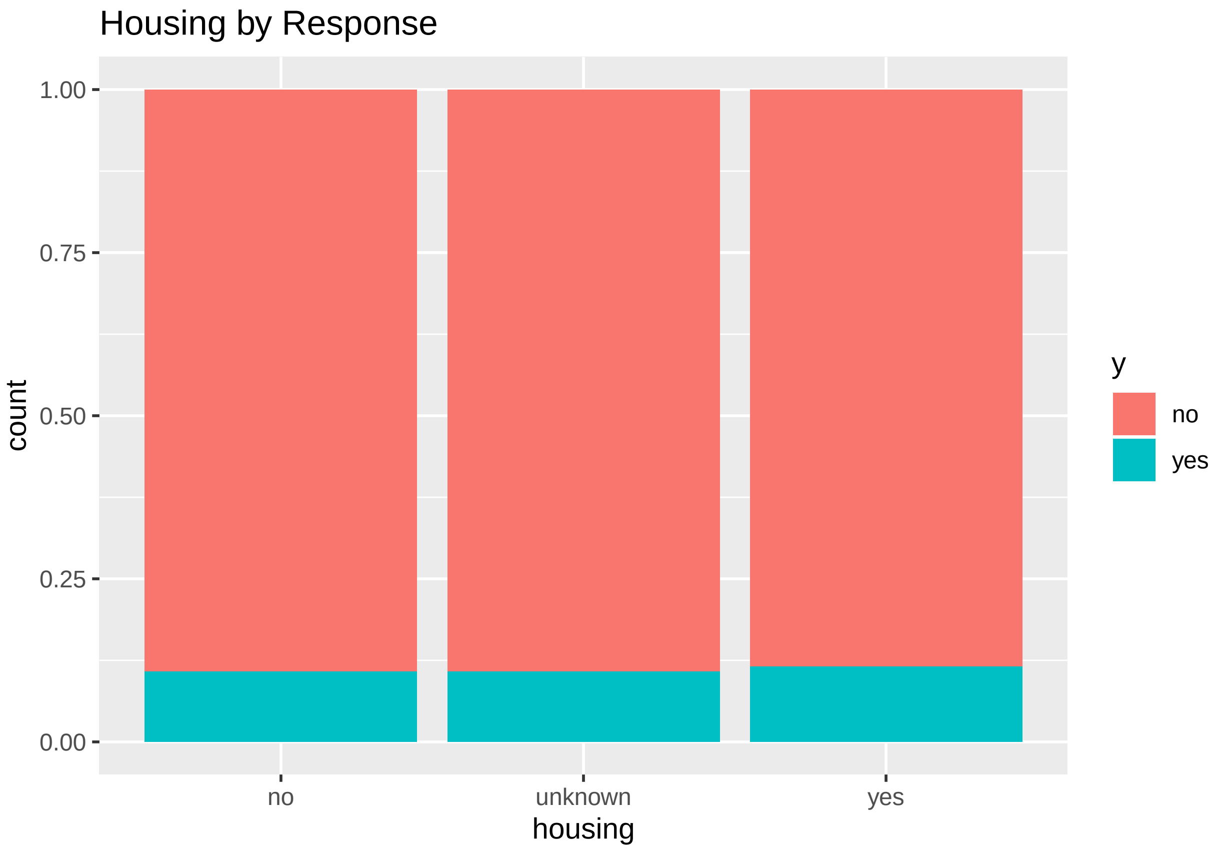
```
#Education
clean_bank_20 %>% ggplot(aes(x = education, fill = y)) + geom_bar(position = "fill")  + ggtitle("Educat
```



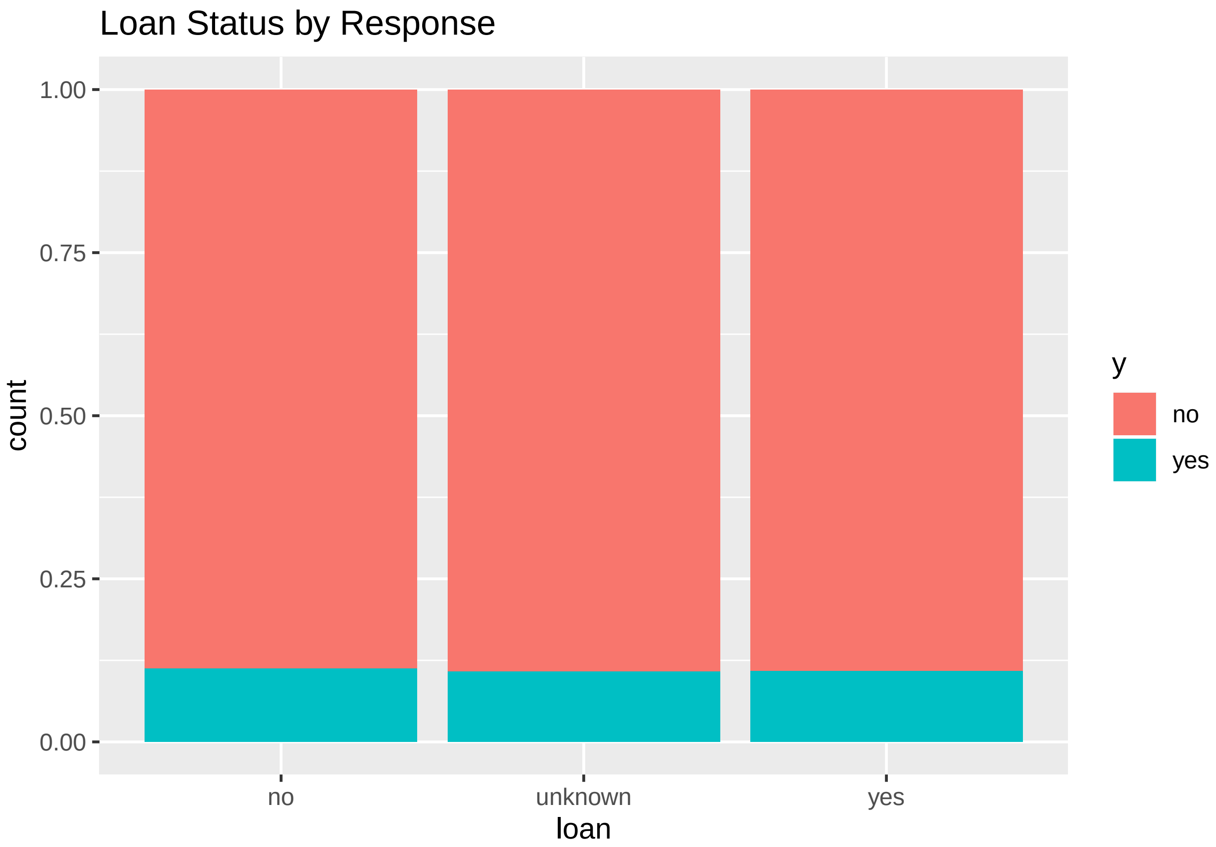
```
#Default
clean_bank_20 %>% ggplot(aes(x = default, fill = y)) + geom_bar(position = "fill")  + ggtitle("Default
```



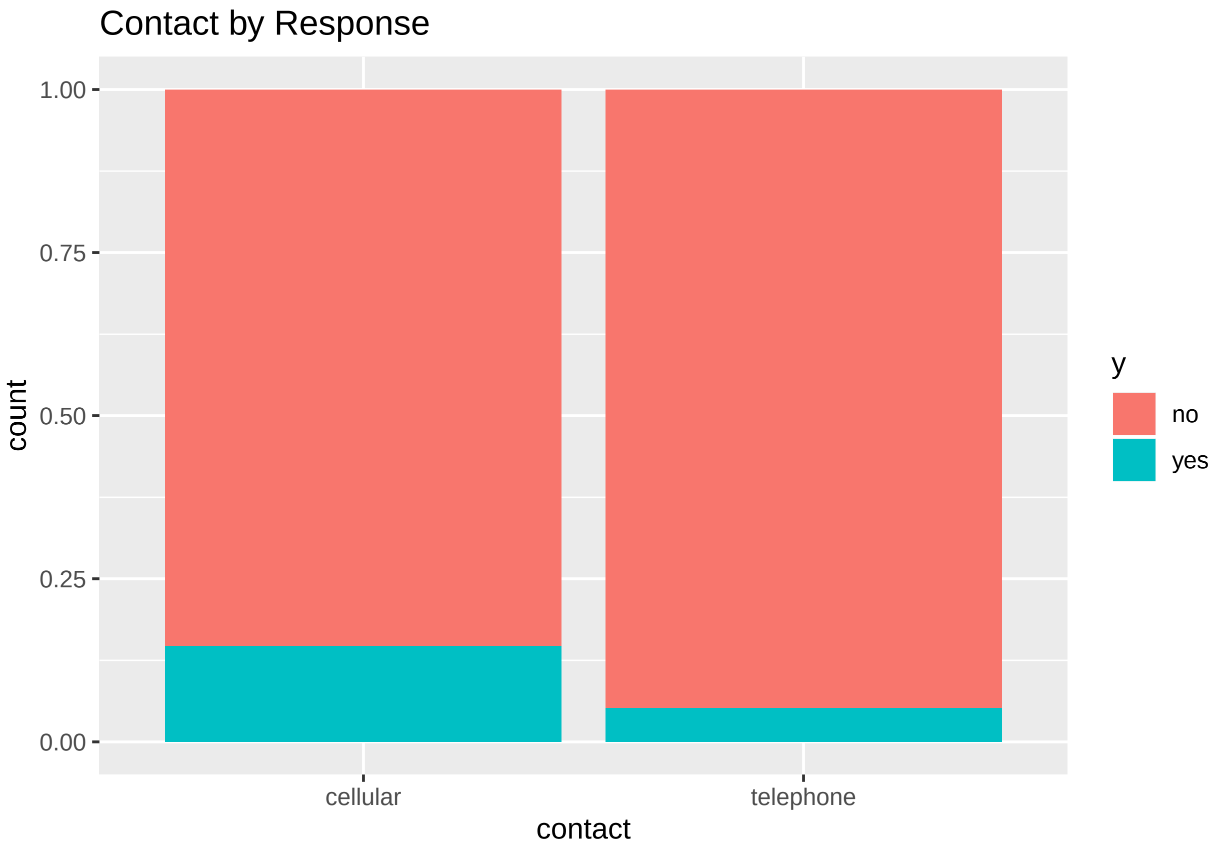
```
#Housing
clean_bank_20 %>% ggplot(aes(x = housing, fill = y)) + geom_bar(position = "fill") + ggtitle("Housing by
```



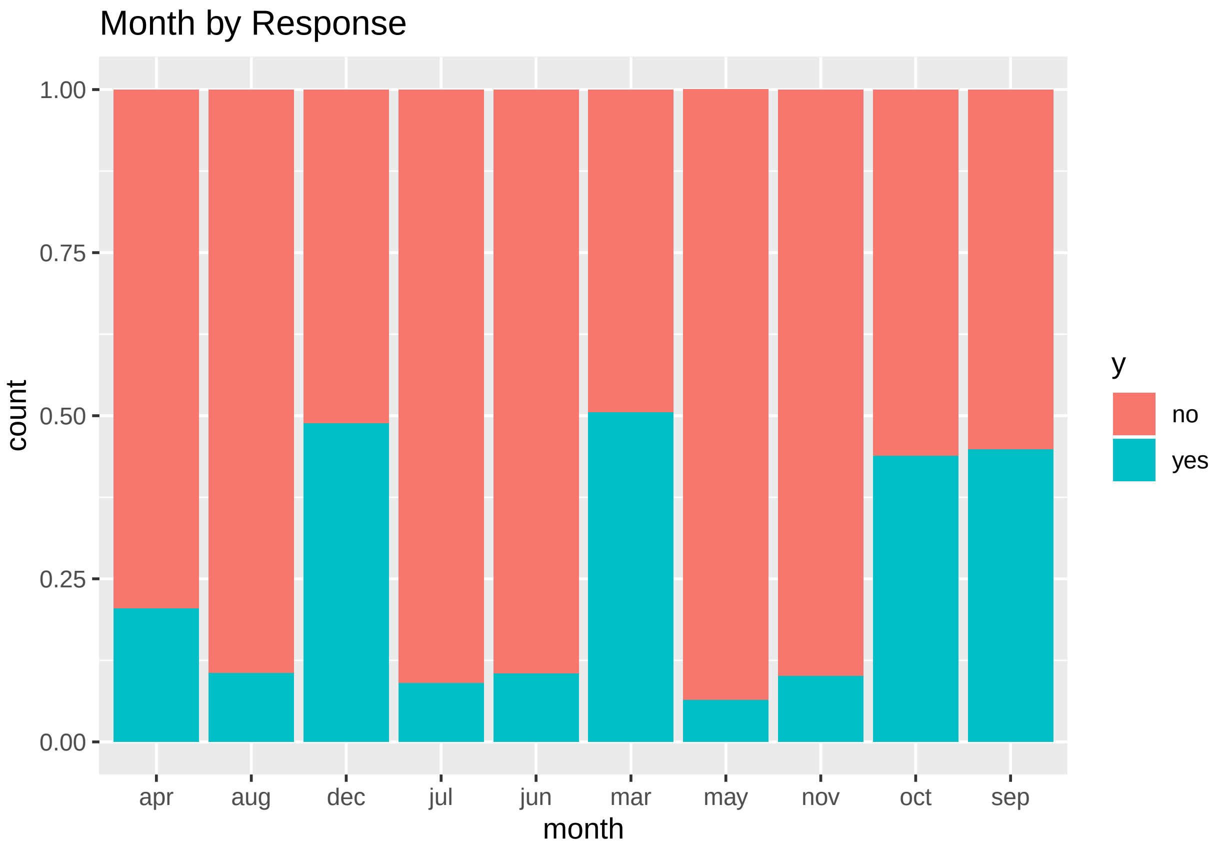
```
#Loan
clean_bank_20 %>% ggplot(aes(x = loan, fill = y)) + geom_bar(position = "fill") + ggtitle("Loan Status b
```



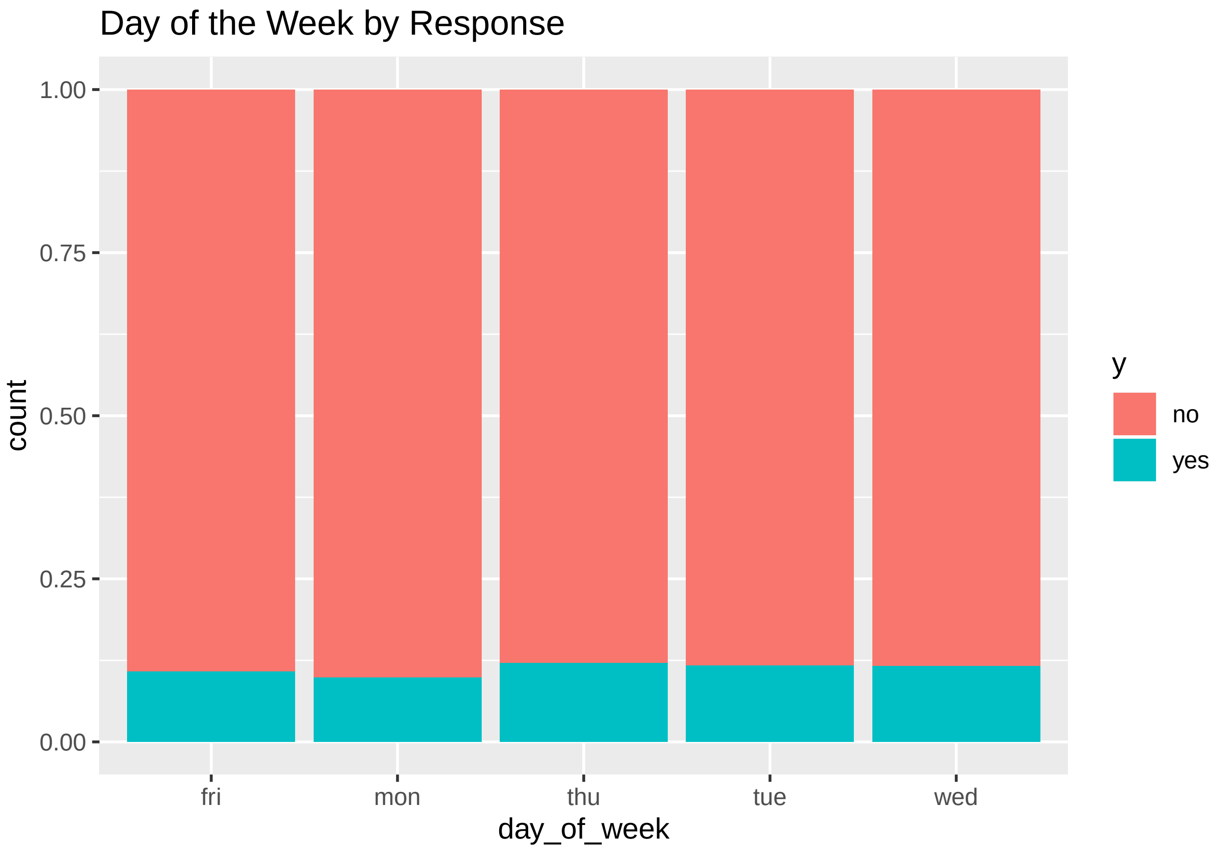
```
#Contact
clean_bank_20 %>% ggplot(aes(x = contact, fill = y)) + geom_bar(position = "fill") + ggtitle("Contact by
```



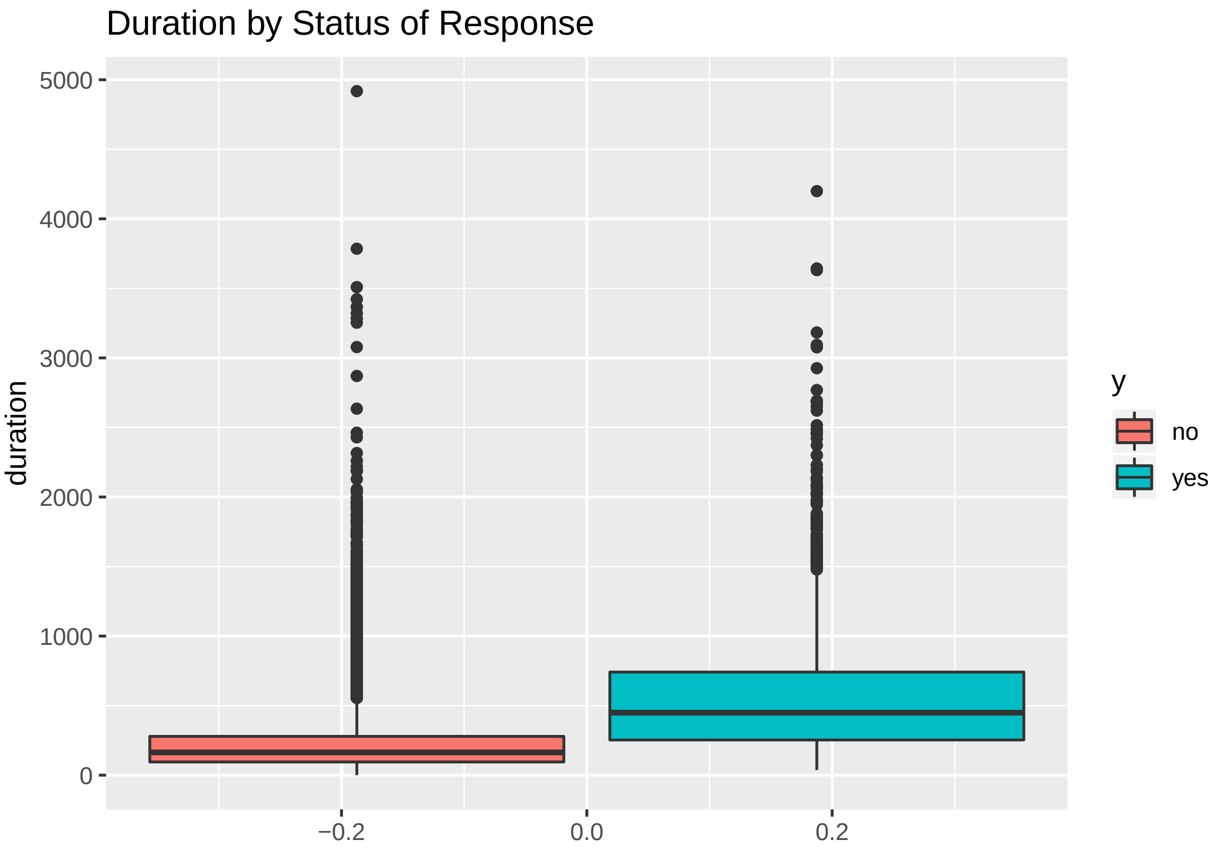
```
#Month
clean_bank_20 %>% ggplot(aes(x = month, fill = y)) + geom_bar(position = "fill")+ ggtitle("Month by Res
```



```
#Day_of_week
clean_bank_20 %>% ggplot(aes(x = day_of_week, fill = y)) + geom_bar(position = "fill") + ggtitle("Day of
```

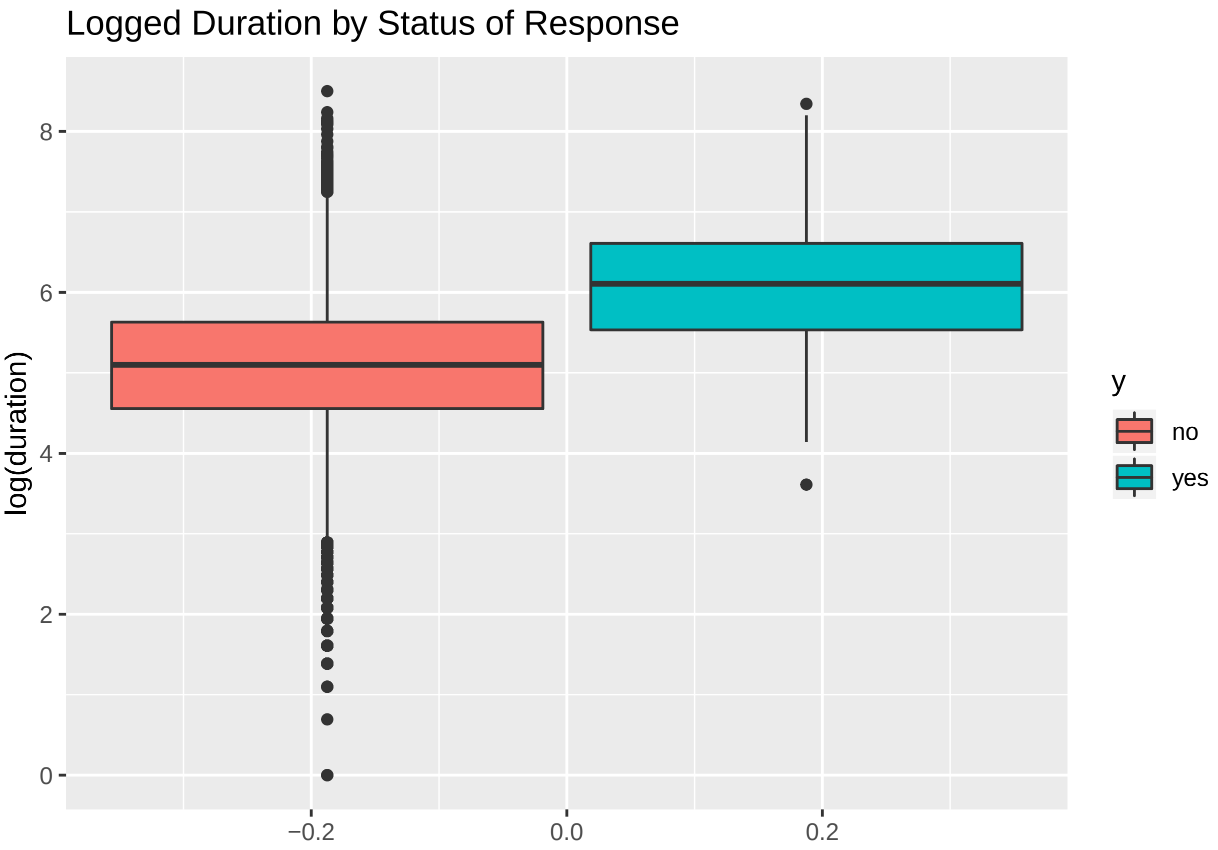


```
#duration
clean_bank_20 %>% ggplot(aes(y = duration, fill = y)) + geom_boxplot() + ggtitle("Duration by Status of
```

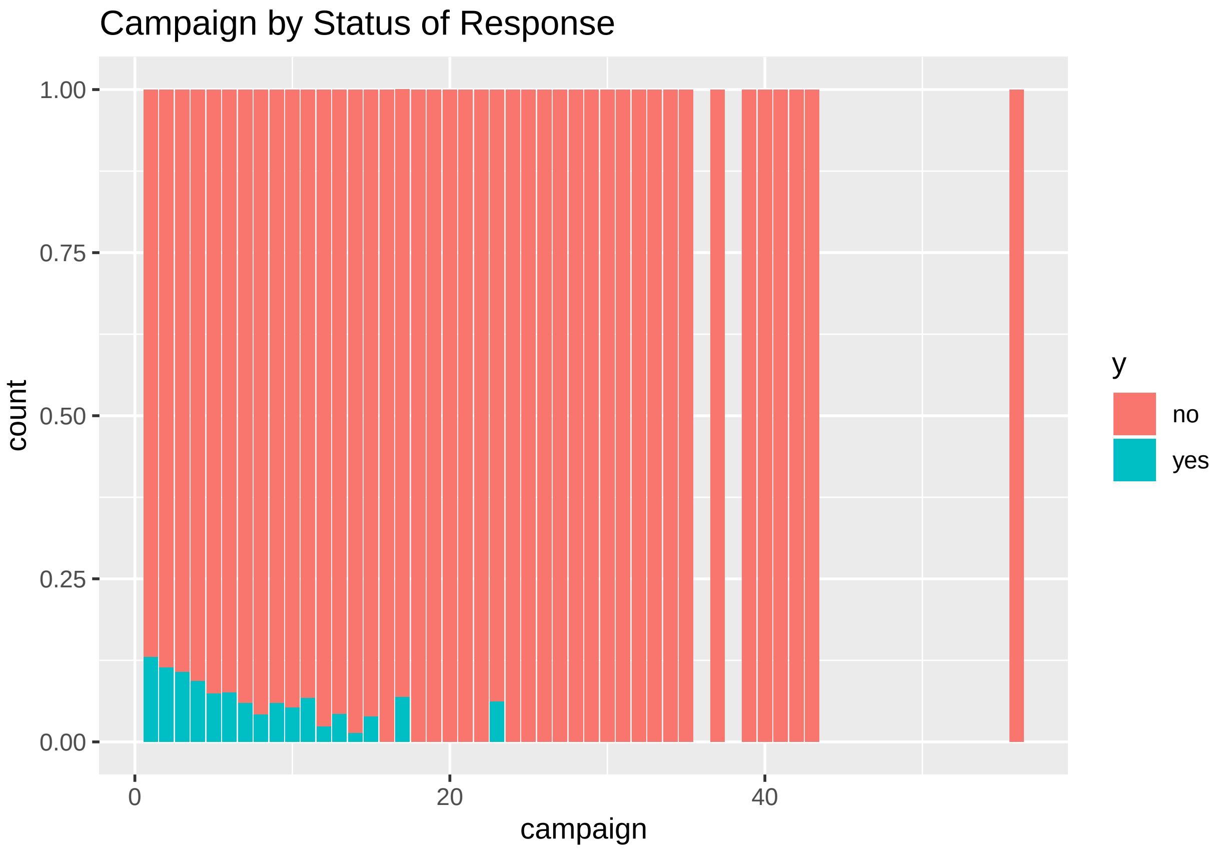


```
clean_bank_20 %>% ggplot(aes(y = log(duration), fill = y)) + geom_boxplot() + ggtitle("Logged Duration
```

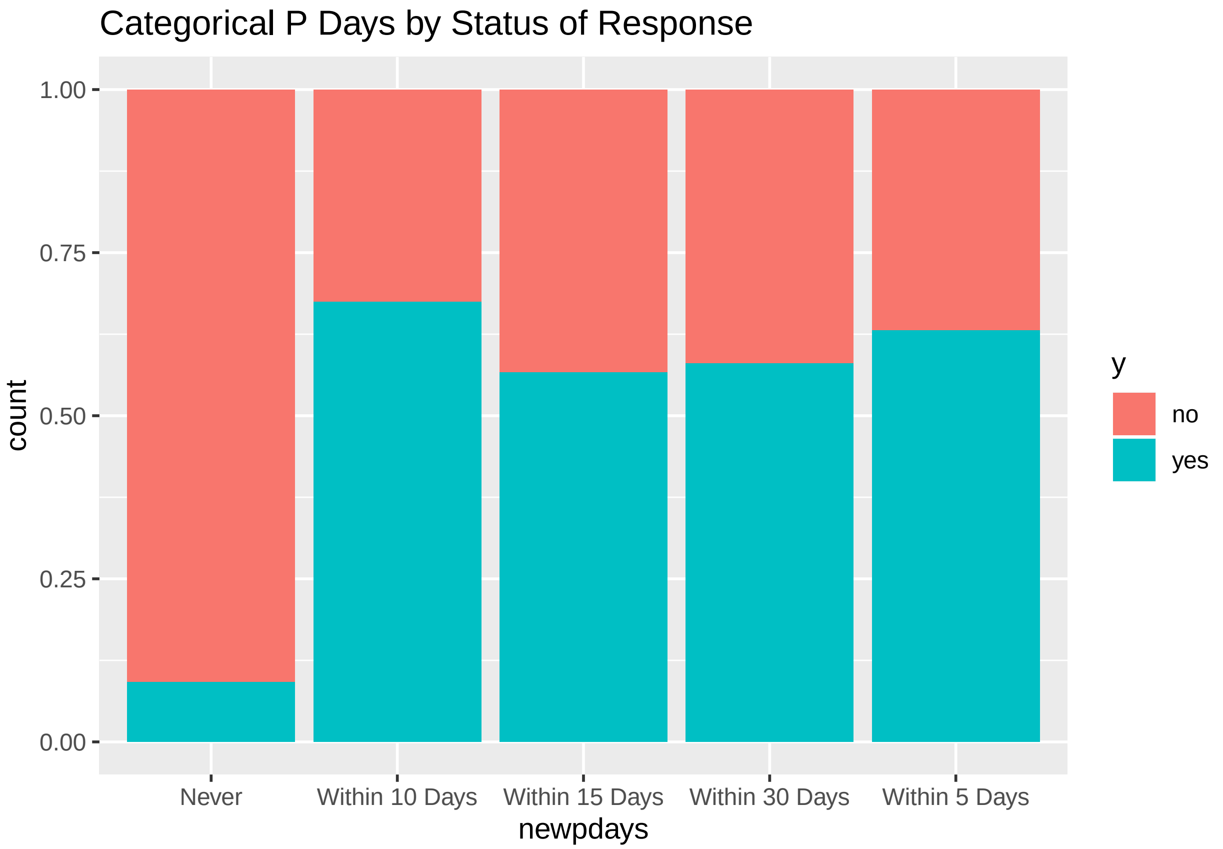
```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



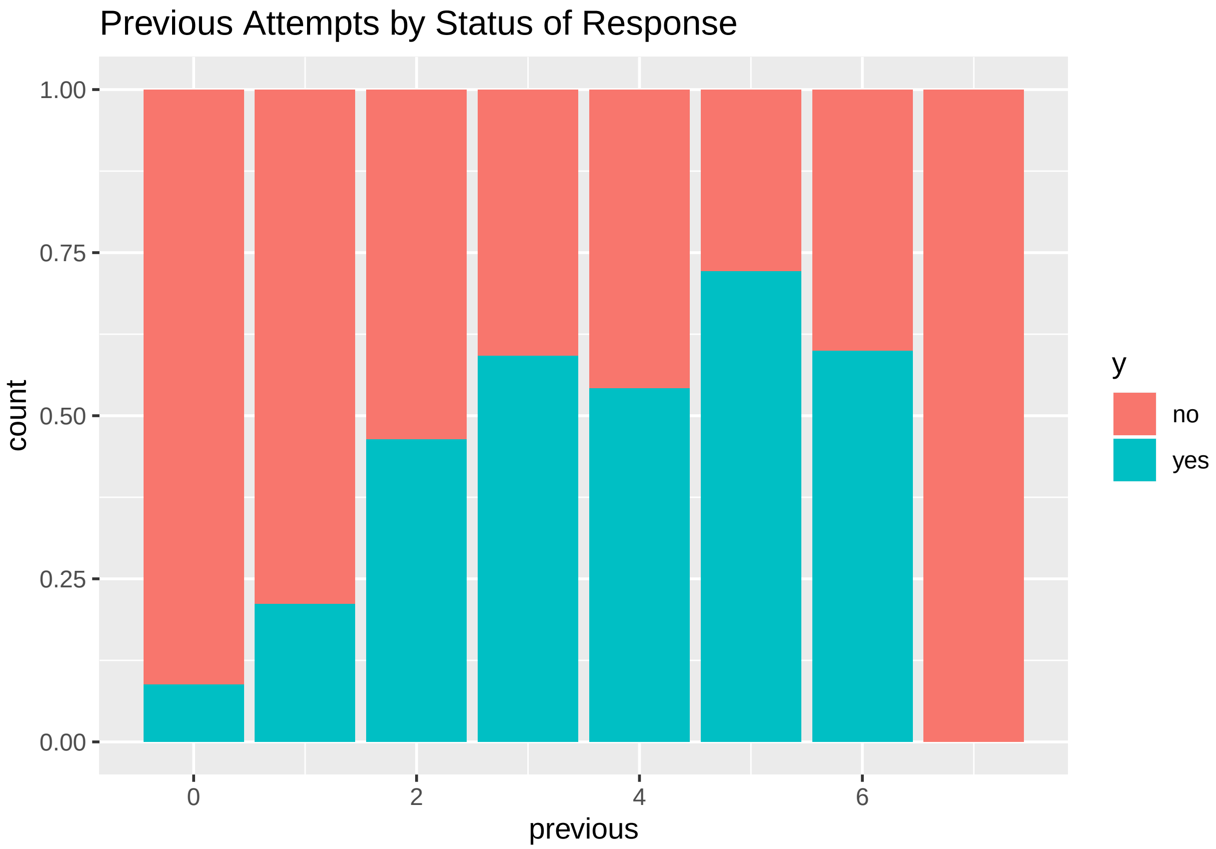
```
#Campaign

clean_bank_20 %>% ggplot(aes(x = campaign, fill = y)) + geom_bar(position = "fill") + ggtitle("Campaign
```



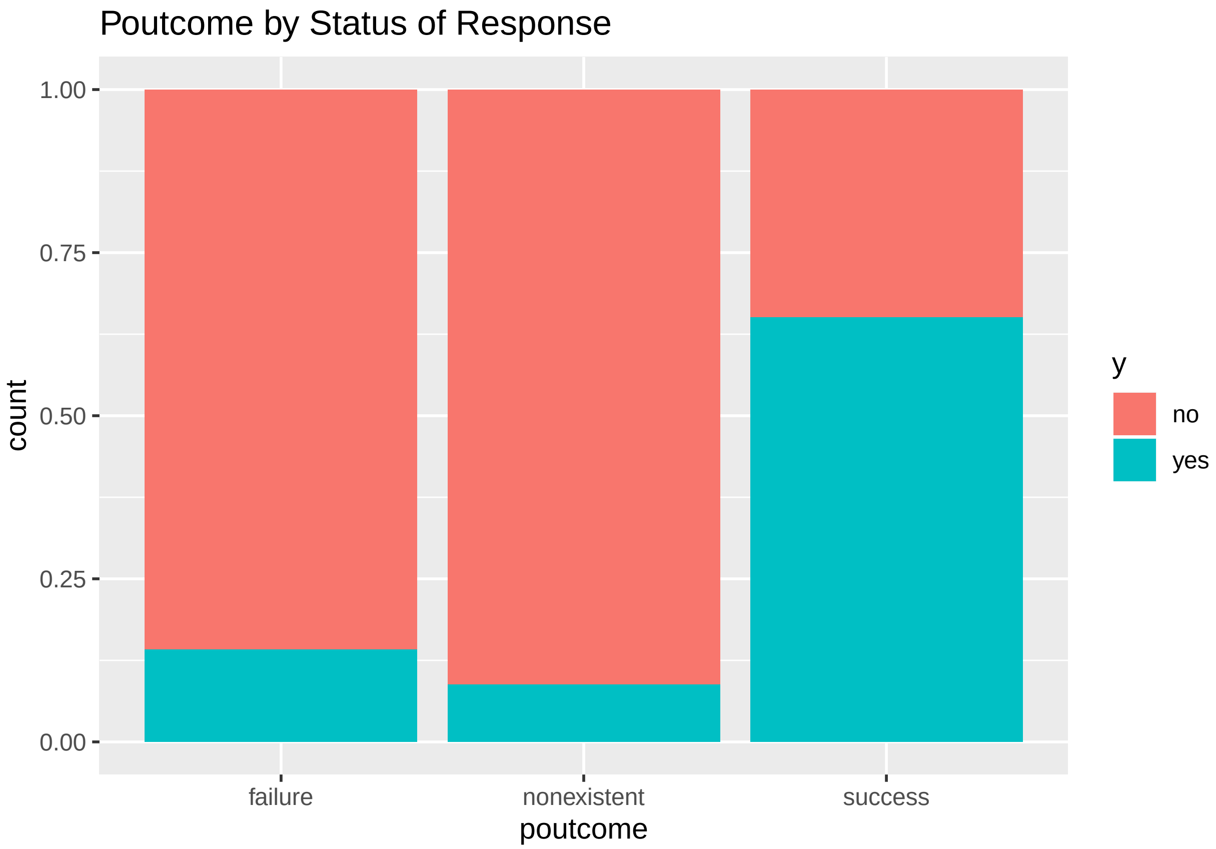
```
#newPdays
clean_bank_20 %>% ggplot(aes(x = newpdays, fill = y)) + geom_bar(position = "fill") + ggtitle("Categori
```



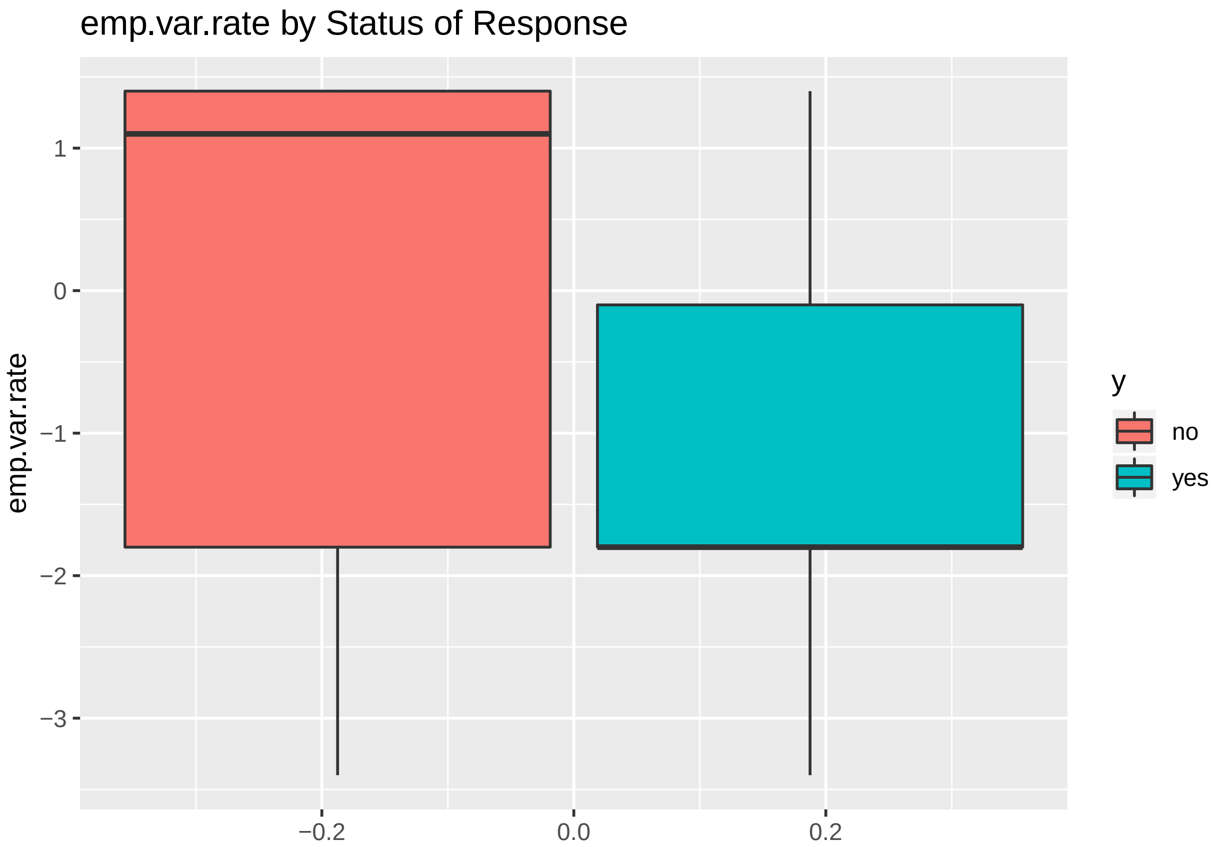
```
#Previous
clean_bank_20 %>% ggplot(aes(x = previous, fill = y)) + geom_bar(position = "fill") + ggtitle("Previous
```



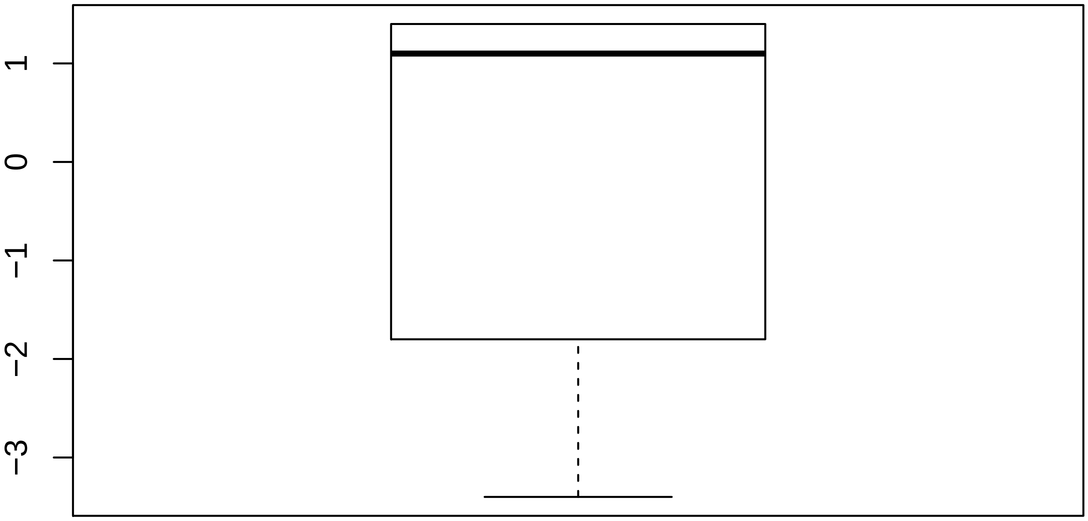
```
#poutcome
clean_bank_20 %>% ggplot(aes(x = poutcome, fill = y)) + geom_bar(position = "fill")+ ggtitle("Poutcome b
```



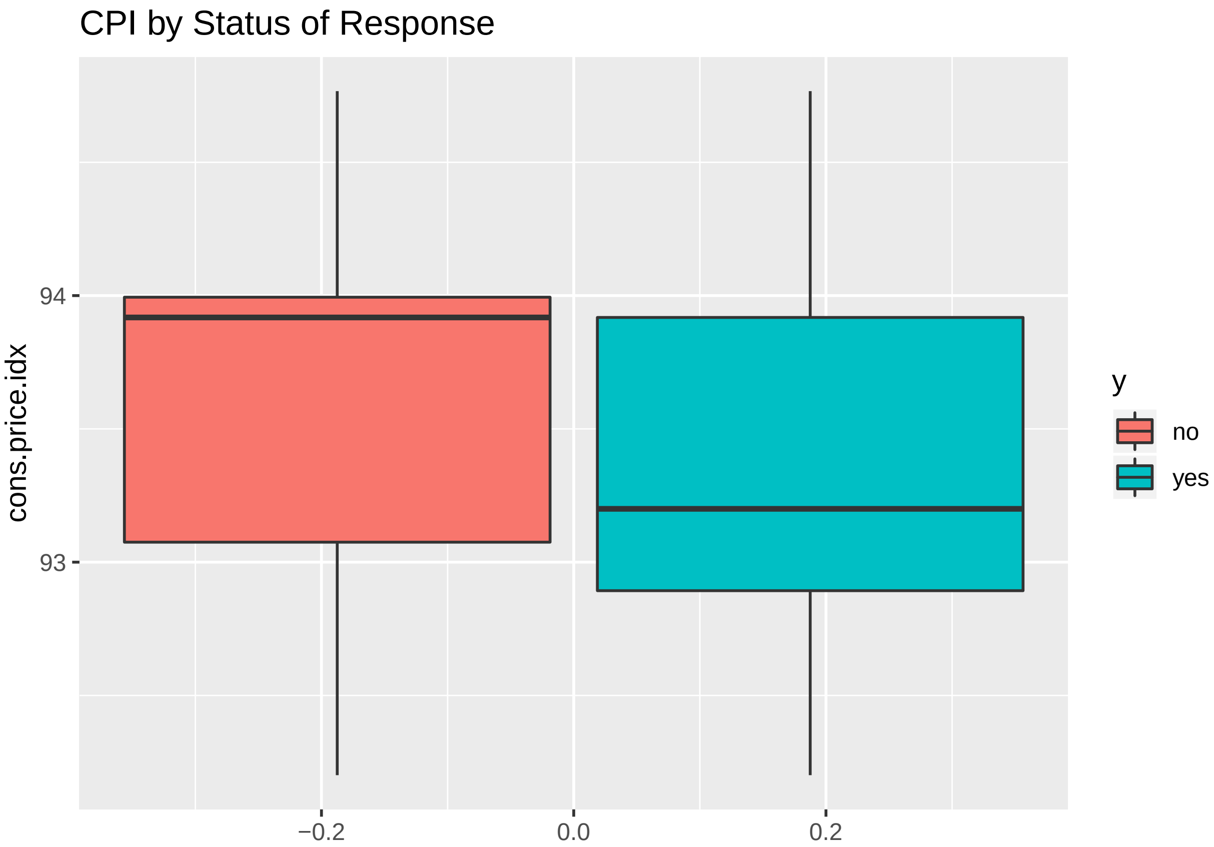
```
#emp.var.rate
clean_bank_20 %>% ggplot(aes(y = emp.var.rate, fill = y)) + geom_boxplot() + ggtitle("emp.var.rate by St
```



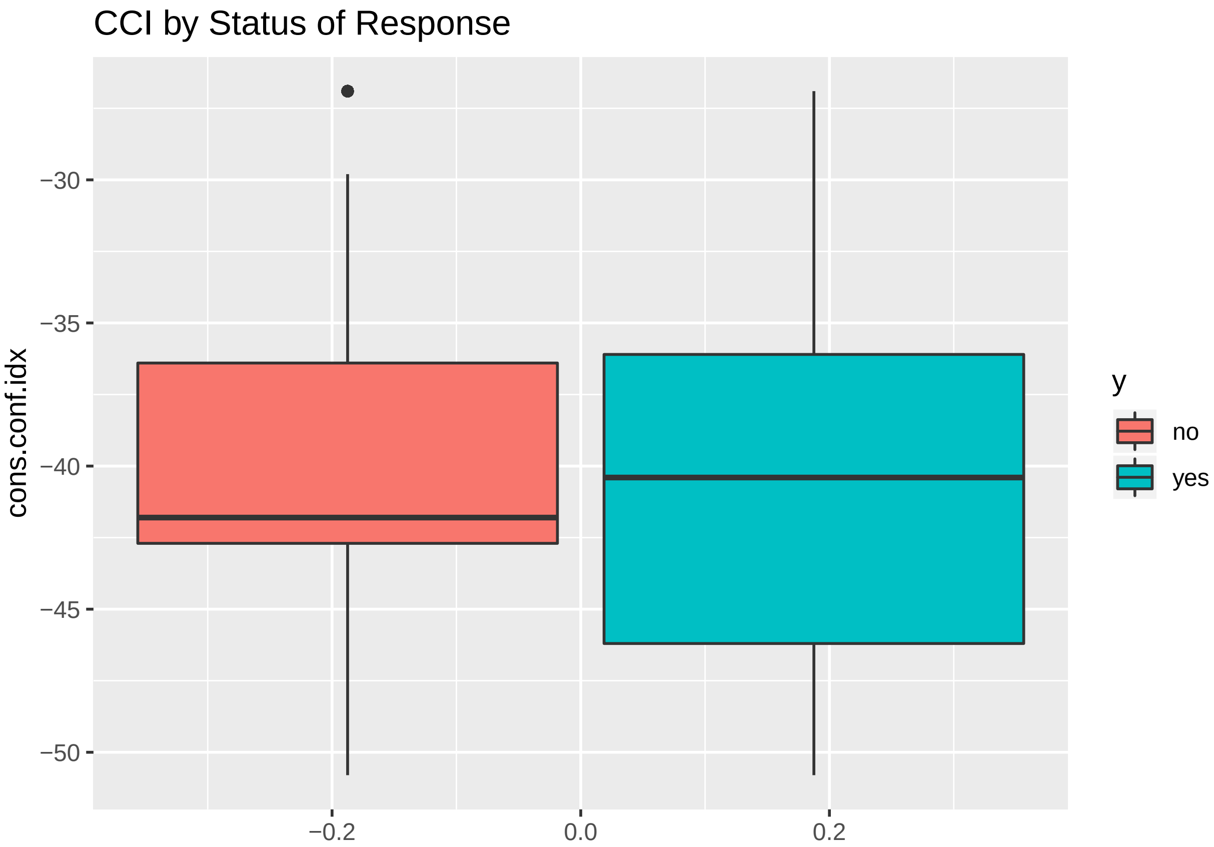
```
boxplot(clean_bank_20$emp.var.rate)
```



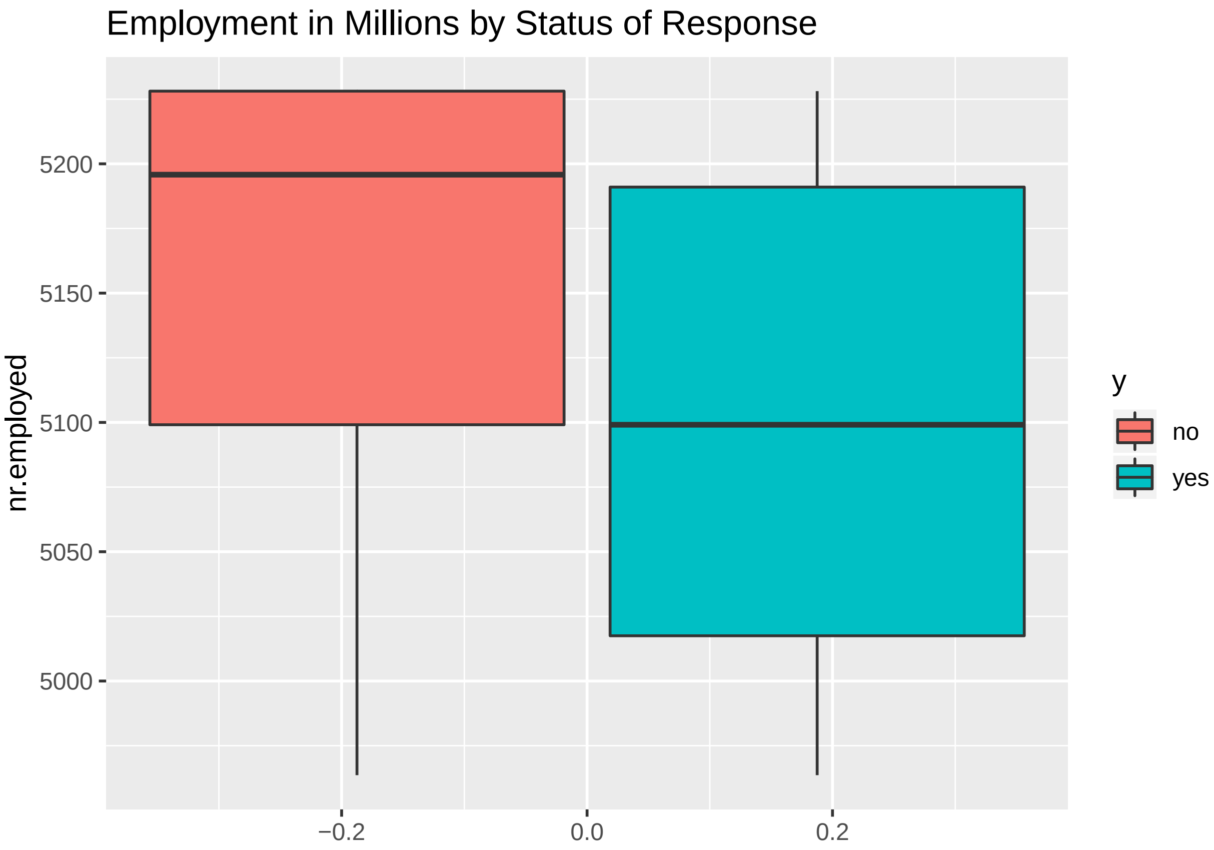
```
#cons.price.idx
clean_bank_20 %>% ggplot(aes(y = cons.price.idx, fill = y)) + geom_boxplot() + ggtitle("CPI by Status o
```



```
#con.conf.idx
clean_bank_20 %>% ggplot(aes(y = cons.conf.idx, fill = y)) + geom_boxplot() + ggtitle("CCI by Status of
```



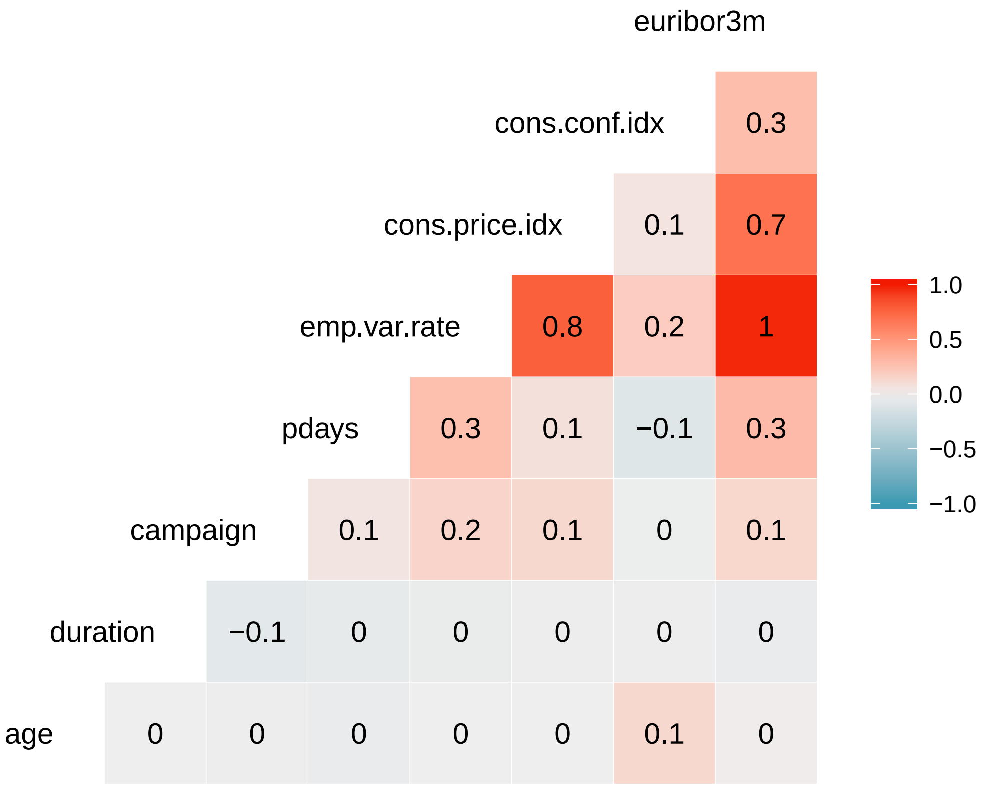
```
#nr Employed
clean_bank_20 %>% ggplot(aes(y = nr.employed, fill = y)) + geom_boxplot() + ggtitle("Employment in Milli
```



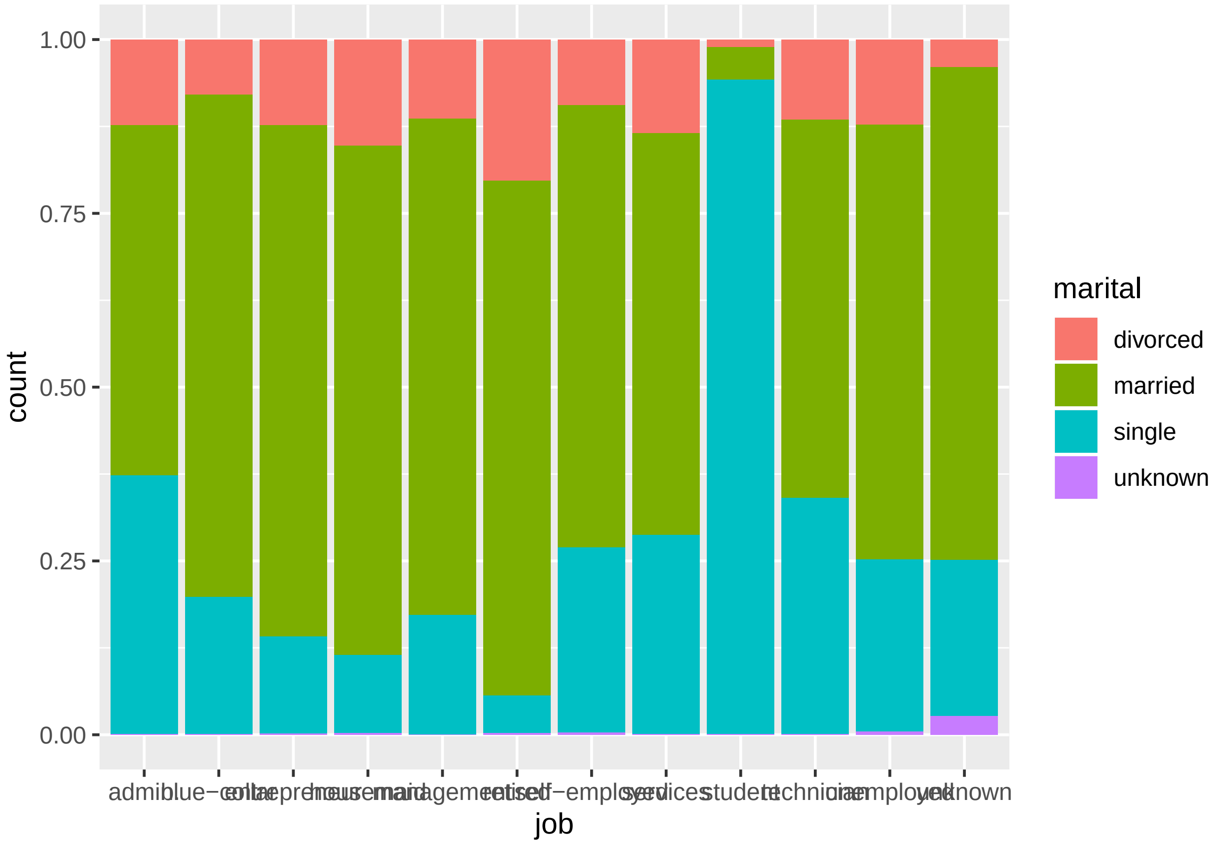
Multicolinearity and Interactions

```
#Multicoliniearity in the continuous variables.
cont_bank = clean_bank_20[,c(1,11,12,13,15,16,17,18,19)]
ggcorr(cont_bank, label = TRUE, hjust = 1  )
```

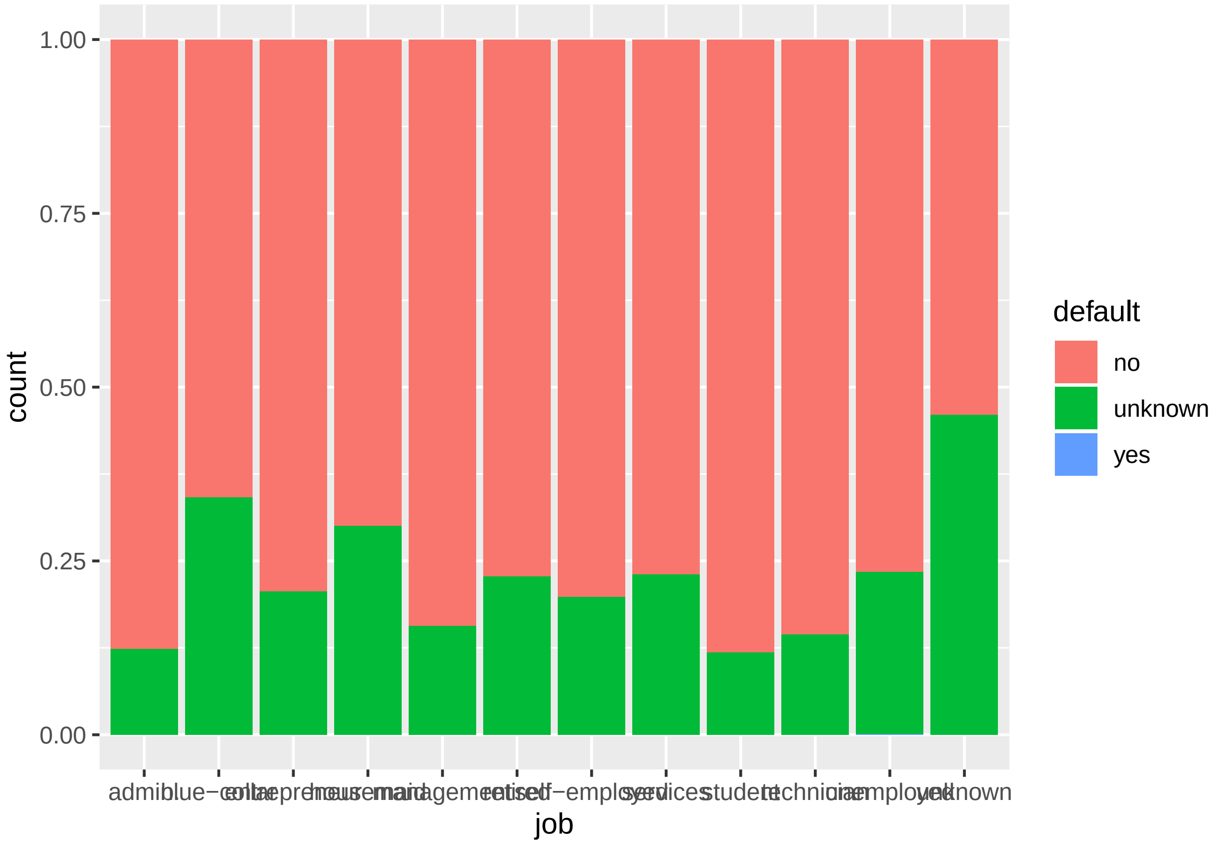
```
## Warning in ggcorr(cont_bank, label = TRUE, hjust = 1): data in column(s)
## 'poutcome' are not numeric and were ignored
```



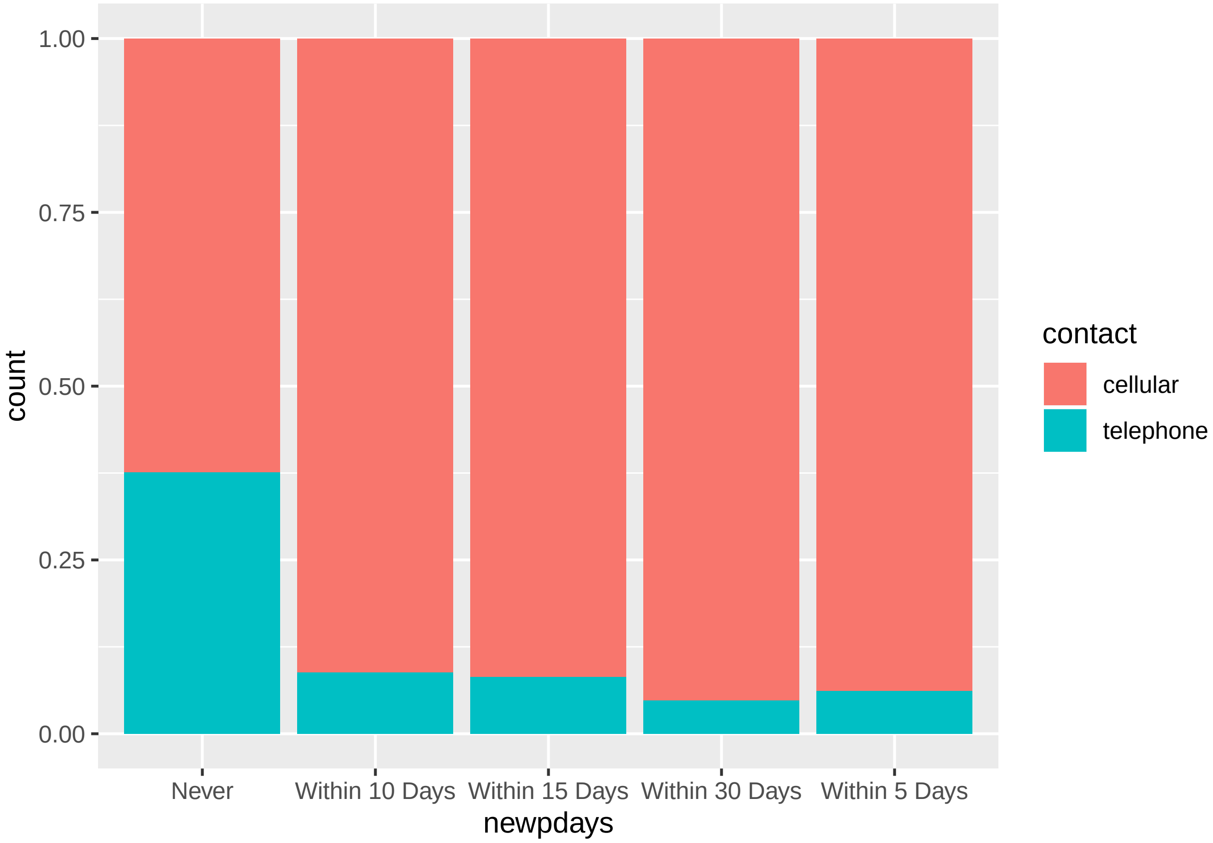
```
###
#Interactions on Categorical + Categorical  Variables

#Job and Marital interaction  - Potentially Useful
clean_bank_20 %>% ggplot(aes(x = job, fill = marital))  + geom_bar(position = "fill")
```



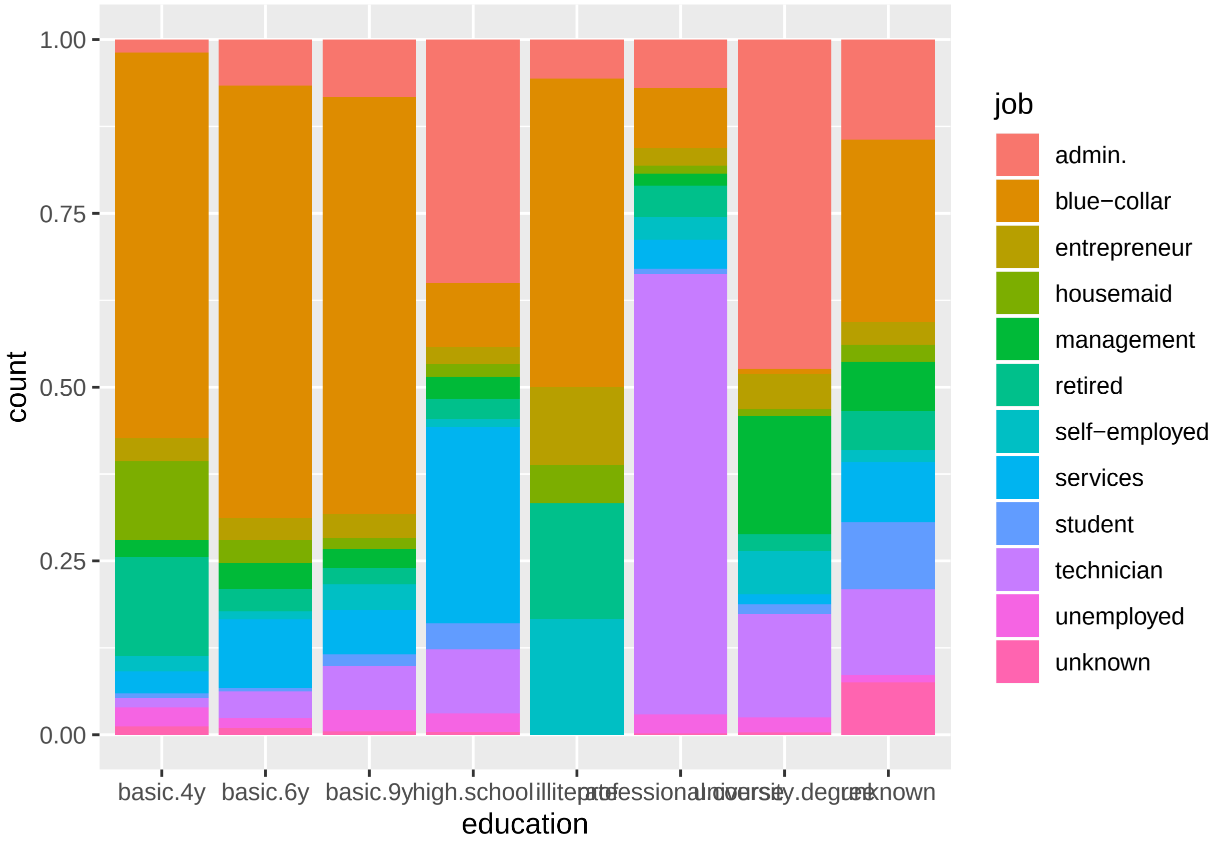
```
#Job and DEfault - Worth keeping, some interaction
clean_bank_20 %>% ggplot(aes(x = job, fill = default))  + geom_bar(position = "fill")
```



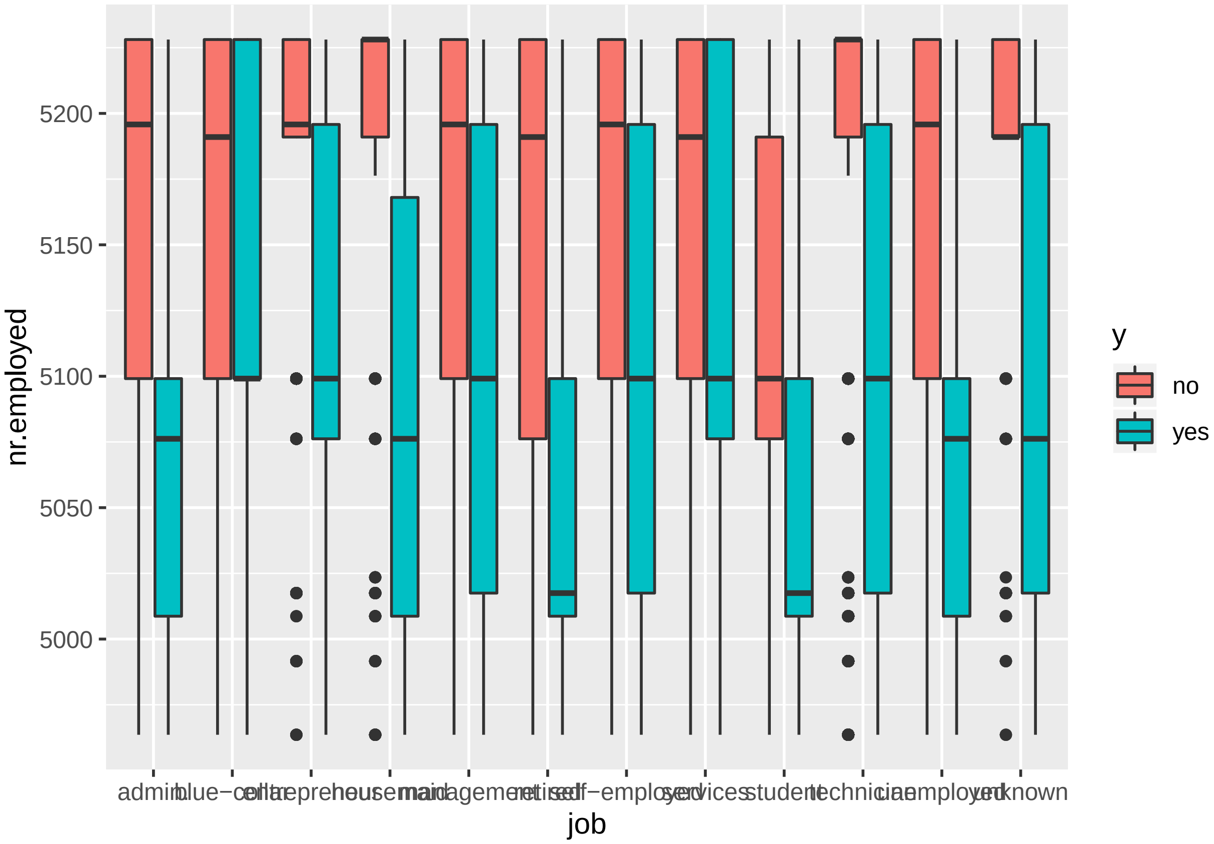
```
#Pdays sole categorical var and contact,   A little interaction,
clean_bank_20 %>% ggplot(aes(x = newpdays, fill = contact))  + geom_bar(position = "fill")
```



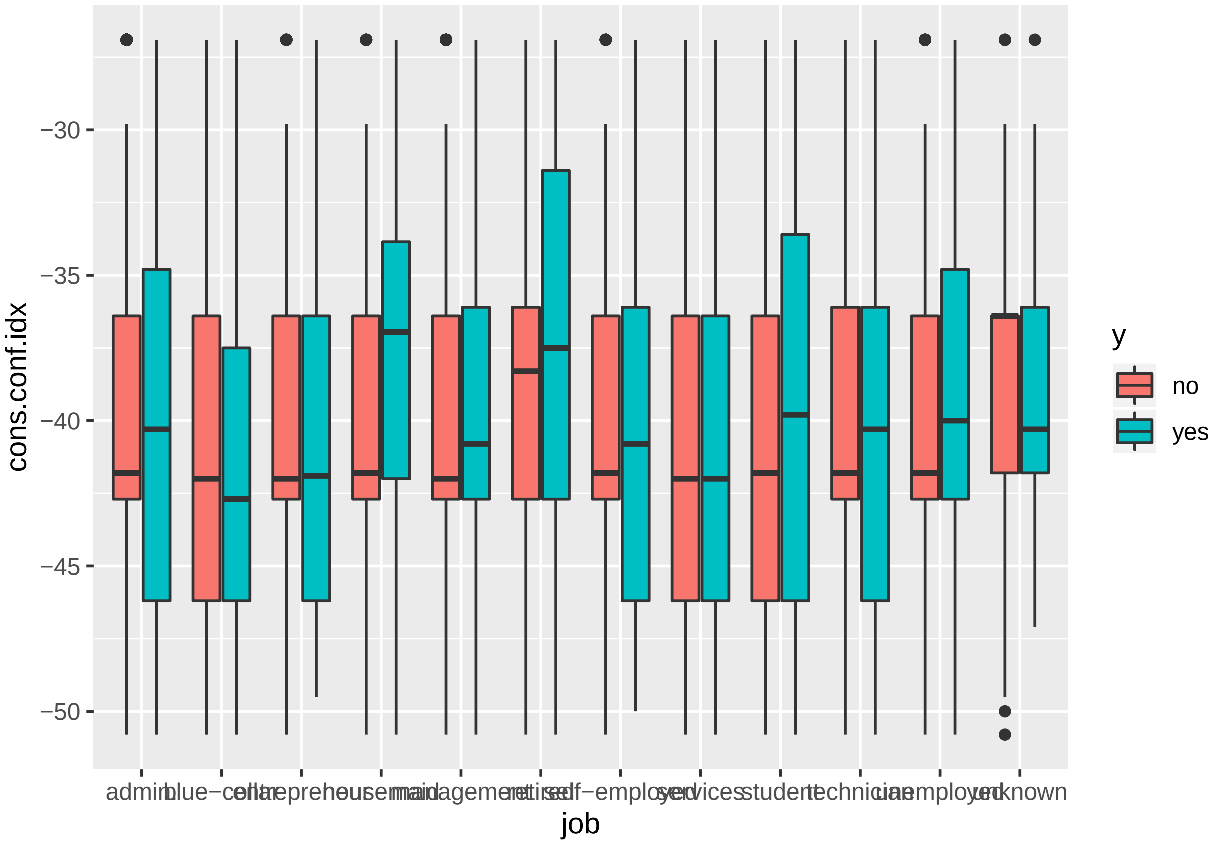
```
#education and job - Complex visually due to categories, but some interaction present for sure
clean_bank_20 %>% ggplot(aes(x = education, fill = job))  + geom_bar(position = "fill")
```



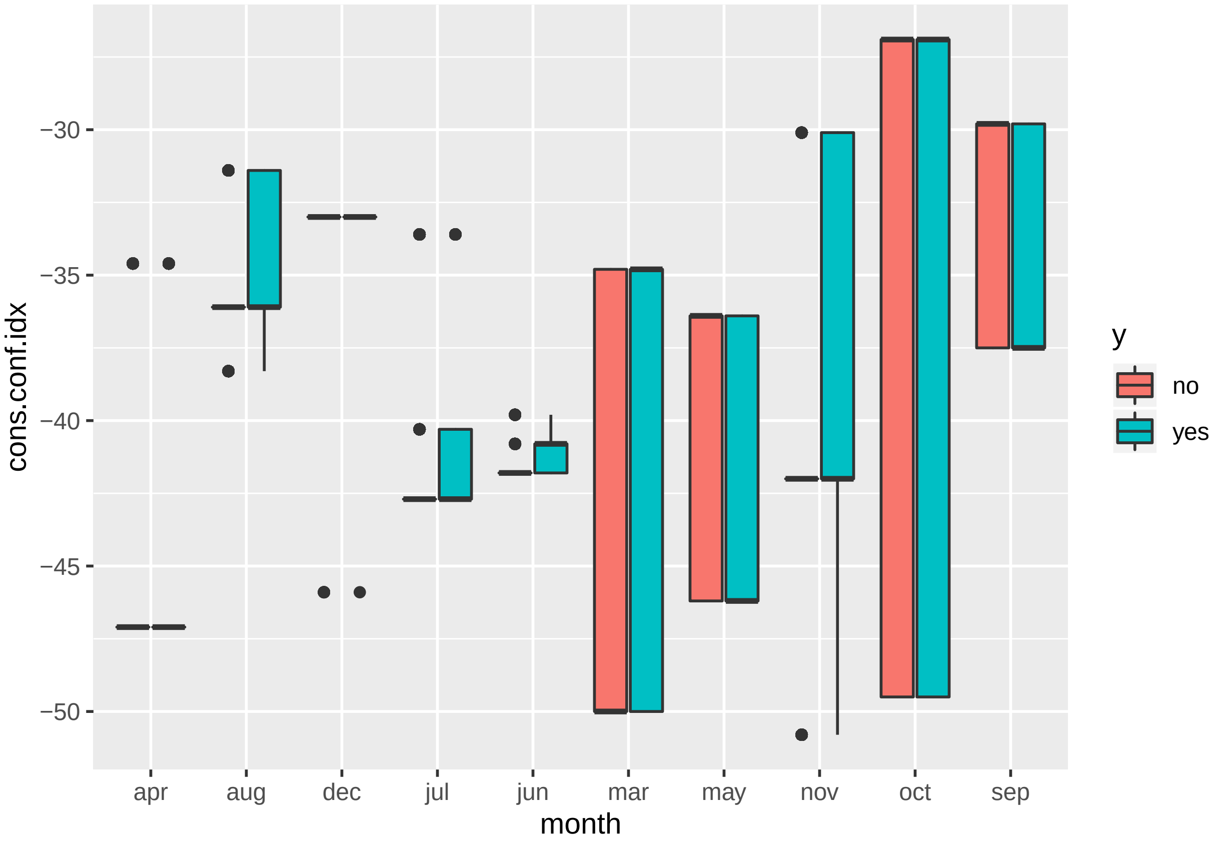
```
#


###
#Continuous + Categorical
#Job and nr. employed - Trend is confusing, but potentially some interaction
#Type of job related to employment in country?
clean_bank_20 %>% ggplot(aes(x = job, y = nr.employed, fill = y)) + geom_boxplot()
```



```
#Cons confidence and job - not quite significant
clean_bank_20 %>% ggplot(aes(x = job, y = cons.conf.idx, fill = y)) + geom_boxplot()
```

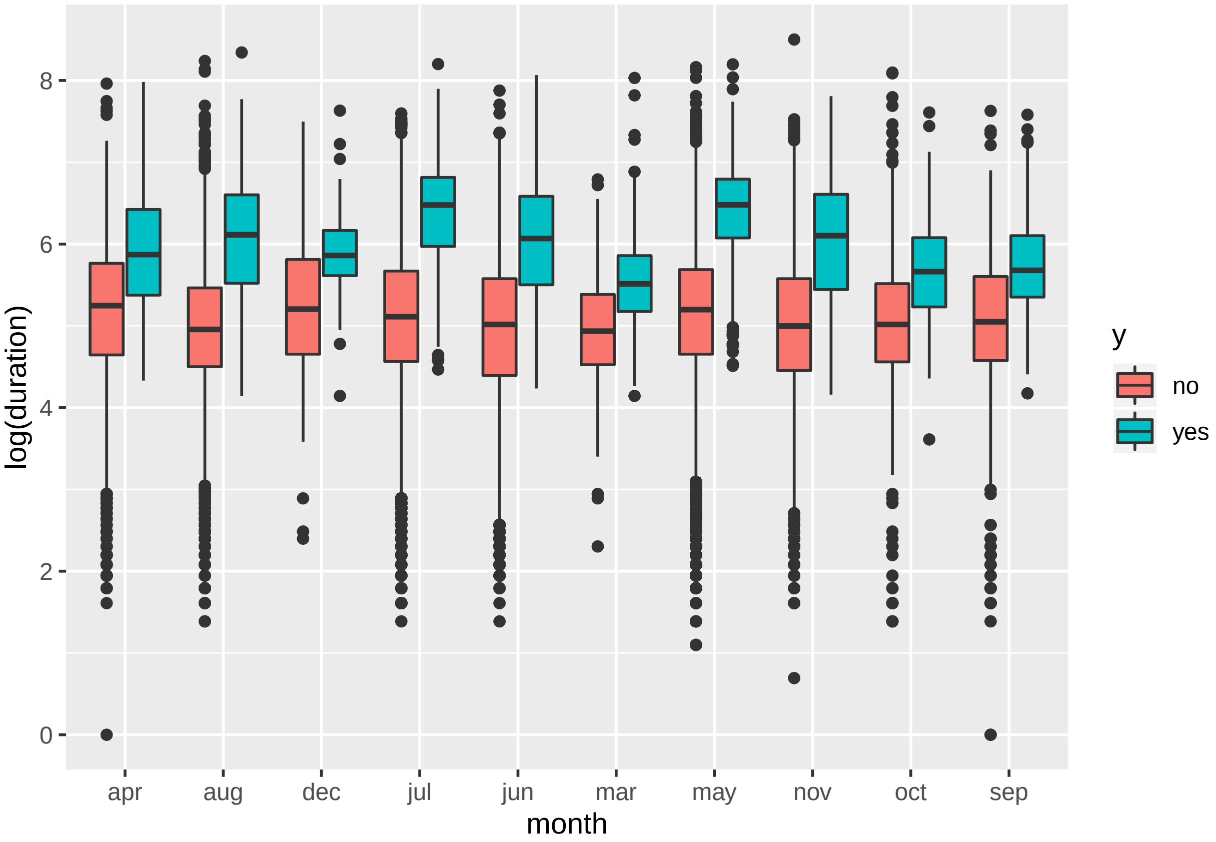


```r
#Month with cons conf and duration, month seems useful, but interactions are confusing, seem odd.
clean_bank_20 %>% ggplot(aes(x = month, y = cons.conf.idx, fill = y)) + geom_boxplot()
```

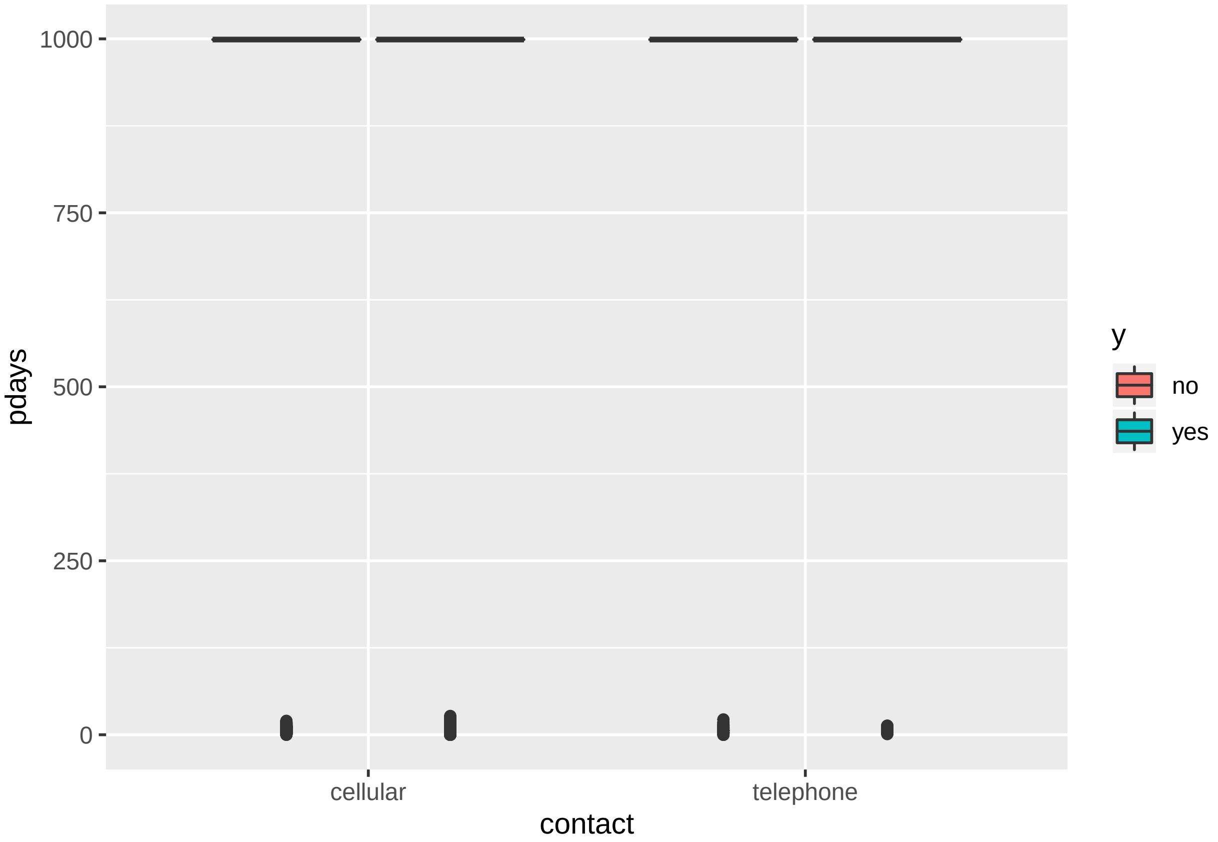


```
clean_bank_20 %>% ggplot(aes(x = month, y = log(duration), fill = y)) + geom_boxplot()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```

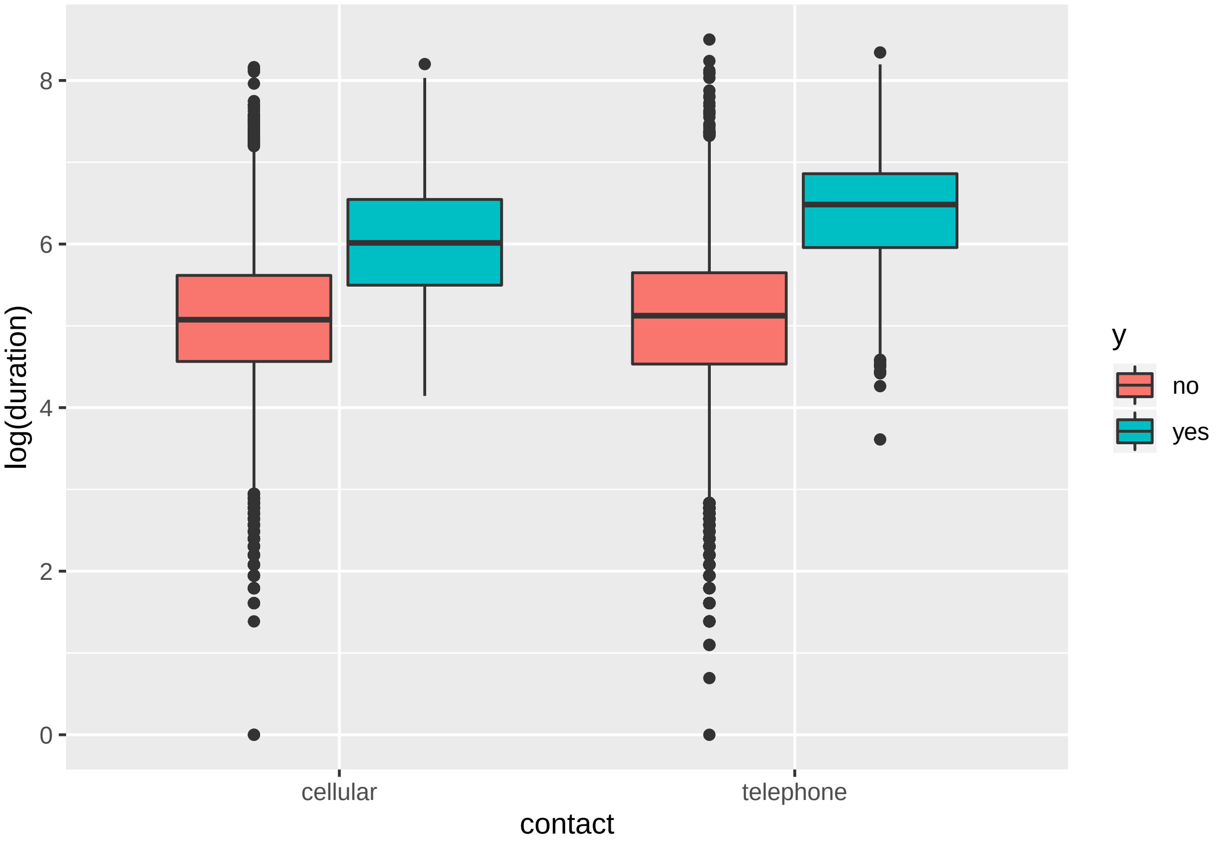


```
#Original Pdays and contact - VV Hard to truly see interaction here dunno how to turn on/off for plot o
#Lets limit pdays interactions then..
clean_bank_20 %>% ggplot(aes(x = contact, y = pdays, fill = y)) + geom_boxplot()
```



```
#Contact Type and Duration - I Don't think this is significant
clean_bank_20 %>% ggplot(aes(x = contact, y = log(duration), fill = y)) + geom_boxplot()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



```
###
#Continuous + Continous

#Use the ggcorr plot from before...
#Emp. var rate, cons price index,  and euribor3m are multicolinear - include 1 or none




#job*default + contact*duration + pdays*contact + pdays*duratio
```

###OBJECTIVE ONE

```
###Forward Selection Model Creation

#Forward selected model returns this set of variables
logr_Forward <-glm(y ~ duration + job + contact + day_of_week + default + previous+ pdays, family = bin



###Backward Selection Model Creation
logr<- glm(y ~ job + education + default + contact +duration + previous + pdays + campaign, family = bi



###Stepwise Selection Model Creation
logr_Stepwise <- glm(y ~ job + default + contact + month + duration + campaign + pdays +poutcome, famil
###Table of accuracies, etc


###ROC Curve Building



###ROC Curve Printing
```