cs100-fp-SL-1130

Silei Li and Mariel Pacada

11/30/2019

Data cleaning

```r
survey[survey$no_employees=="6/25/2019" ] <- as.factor(6-25)
survey[survey$no_employees=="1/5/2019" ] <- as.factor(1-5)
```

```r
diffgender <- unique(survey$Gender)
```

```r
#Male
survey$Gender<-replace(survey$Gender,survey$Gender=="M","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="m","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="male","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Male ","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="maile","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Mail","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Man","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Mal","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Malr","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="msle","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Make","Male")
#Female
survey$Gender<-replace(survey$Gender,survey$Gender=="F","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="f","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="female","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="Female ","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="Femake","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="Woman","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="woman","Female")
#Cis
#??partial match, do "fe" first for females, than "male" for males
#Cis female
survey$Gender<-replace(survey$Gender,survey$Gender=="Cis Female","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="cis-female/femme","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="Female (cis)","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="femail","Female")
survey$Gender<-replace(survey$Gender,survey$Gender=="","Female")
#Cis male
survey$Gender<-replace(survey$Gender,survey$Gender=="Cis Male","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="cis male","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Cis Man","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="Male (CIS)","Male")
survey$Gender<-replace(survey$Gender,survey$Gender=="","Male")
#Trans
```

Clean age

Will said: maybe store the rest as NA or other… but we can also just skip them, just filter Gender = Female and Male, and note that on the visualizations.

```r
print(length(unique(survey$Gender)))
```

```
## [1] 23
```

```r
diffgender <- unique(survey$Gender)
diffgender
```

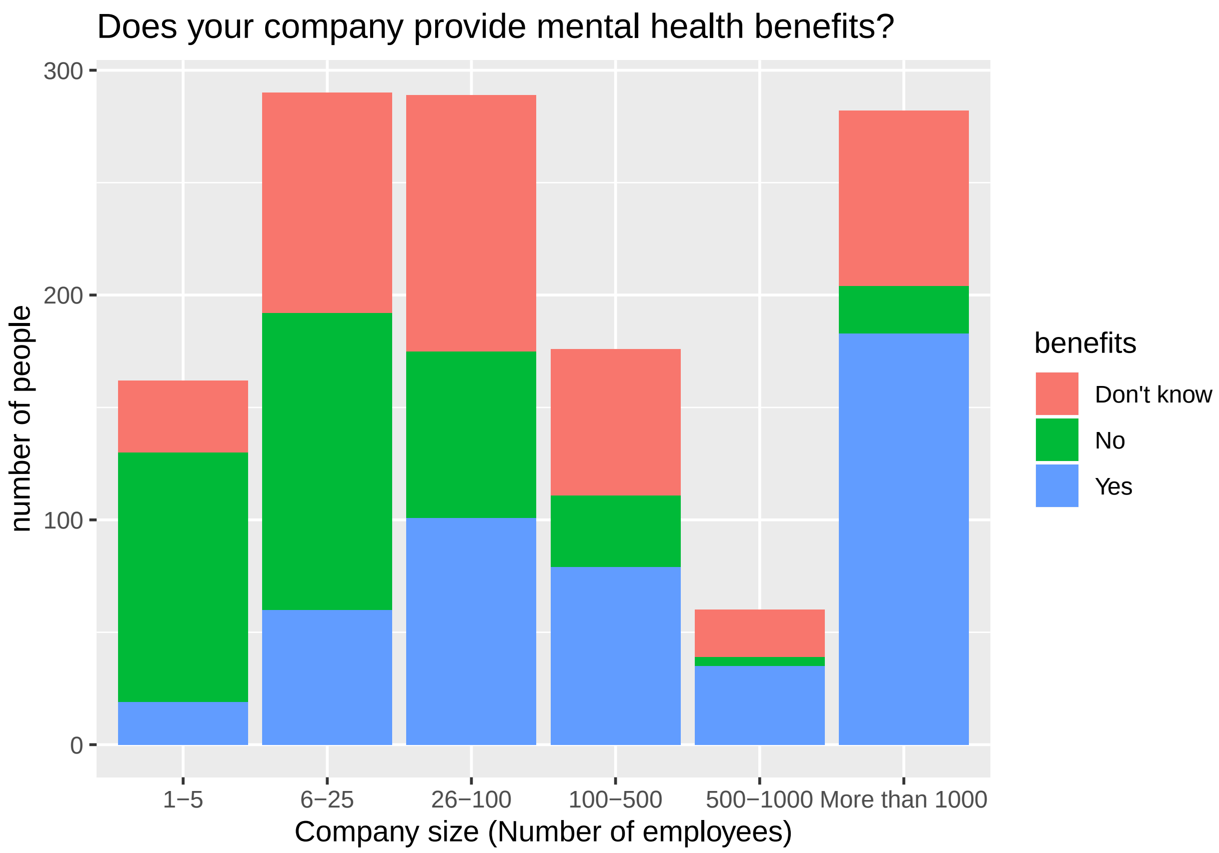
```
##  [1] Female
##  [2] Male
##  [3] Male-ish
##  [4] Trans-female
##  [5] something kinda male?
##  [6] queer/she/they
##  [7] non-binary
##  [8] Nah
##  [9] All
## [10] Enby
## [11] fluid
## [12] Genderqueer
## [13] Androgyne
## [14] Agender
## [15] Guy (-ish) ^_^
## [16] male leaning androgynous
## [17] Trans woman
## [18] Neuter
## [19] Female (trans)
## [20] queer
## [21] A little about you
## [22] p
## [23] ostensibly male, unsure what that really means
## 49 Levels: A little about you Agender All Androgyne ... Woman
```

```r
summary(survey)
```

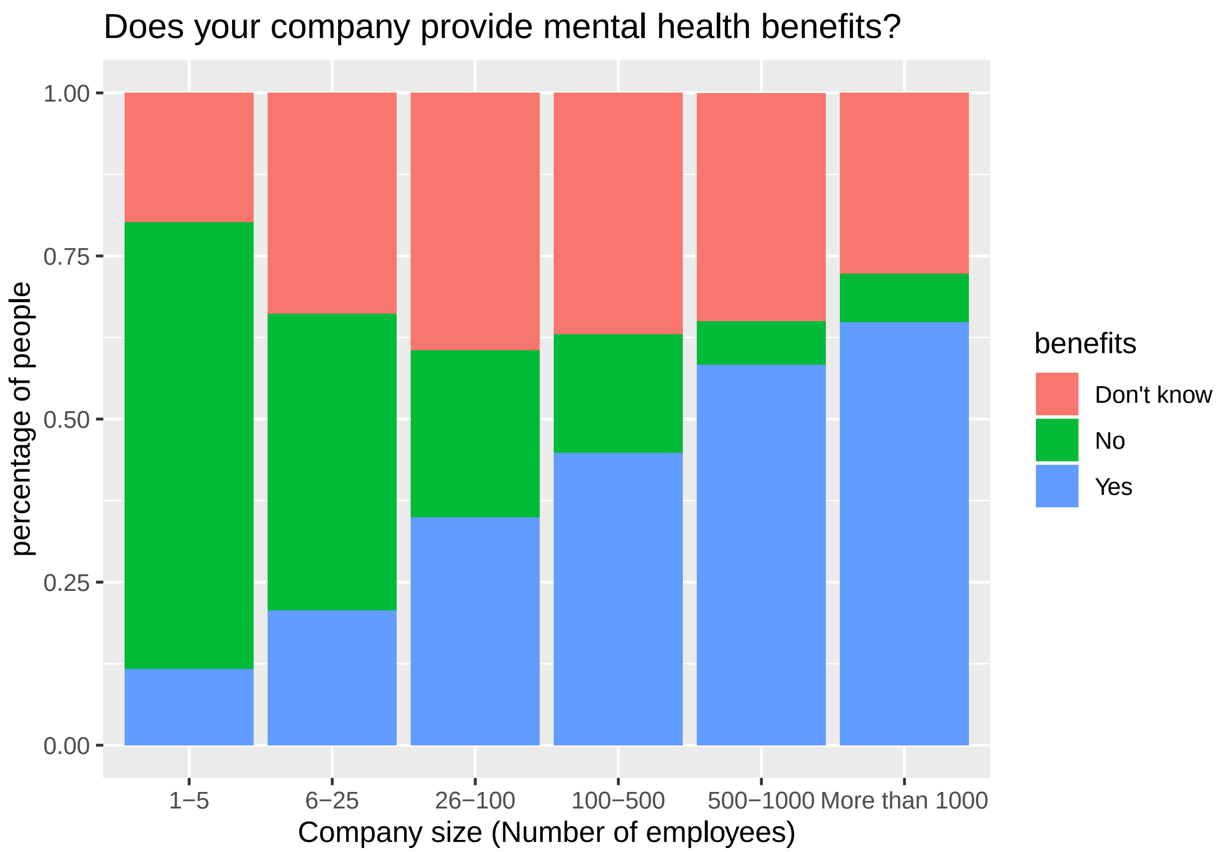
```
##                 Timestamp         Age                            Gender
##  2014-08-27 12:31:41:   2   Min.   :-1.726e+03   Male              :990
##  2014-08-27 12:37:50:   2   1st Qu.: 2.700e+01   Female            :247
##  2014-08-27 12:43:28:   2   Median : 3.100e+01   Female (trans)    :  2
##  2014-08-27 12:44:51:   2   Mean   : 7.943e+07   A little about you:  1
##  2014-08-27 12:54:11:   2   3rd Qu.: 3.600e+01   Agender           :  1
##  2014-08-27 14:22:43:   2   Max.   : 1.000e+11   All               :  1
##  (Other)            :1247                        (Other)           : 17
##           Country        state     self_employed family_history treatment
##  United States :751   CA     :138   No  :1095     No :767       No :622
##  United Kingdom:185   WA     : 70   Yes : 146     Yes:492       Yes:637
##  Canada        : 72   NY     : 57   NA's:  18
##  Germany       : 45   TN     : 45
##  Ireland       : 27   TX     : 44
##  Netherlands   : 27   (Other):390
##  (Other)       :152   NA's   :515
##    work_interfere         no_employees remote_work tech_company
##  Never    :213    1-5           :162   No :883     No : 228
##  Often    :144    100-500       :176   Yes:376     Yes:1031
##  Rarely   :173    26-100        :289
##  Sometimes:465    500-1000      : 60
##  NA's     :264    6-25          :290
##                   More than 1000:282
##
##        benefits     care_options   wellness_program      seek_help
##  Don't know:408   No      :501   Don't know:188     Don't know:363
##  No        :374   Not sure:314   No        :842     No        :646
##  Yes       :477   Yes     :444   Yes       :229     Yes       :250
##
##
##
##
##       anonymity                  leave     mental_health_consequence
##  Don't know:819   Don't know        :563   Maybe:477
##  No        : 65   Somewhat difficult:126   No   :490
##  Yes       :375   Somewhat easy     :266   Yes  :292
##                   Very difficult    : 98
##                   Very easy         :206
##
##
##  phys_health_consequence        coworkers          supervisor
##  Maybe:273              No          :260   No          :393
##  No   :925              Some of them:774   Some of them:350
##  Yes  : 61              Yes         :225   Yes         :516
##
##
##
##
##  mental_health_interview phys_health_interview  mental_vs_physical
##  Maybe: 207              Maybe:557             Don't know:576
##  No   :1008              No   :500             No        :340
##  Yes  :  44              Yes  :202             Yes       :343
##
##
##
##
##  obs_consequence
##  No :1075
##  Yes: 184
##
##
##
##
##
##
##  * Small family business - YMMV.
##
##  -
##  (yes but the situation was unusual and involved a change in leadership at a very high level in the c
##  A close family member of mine struggles with mental health so I try not to stigmatize it. My employe
##  (Other)
## NA's
```

Stacked barplot

```
survey%>%
  mutate(no_employees = factor(no_employees, levels = c("1-5","6-25","26-100","100-500","500-1000","Mor
  ggplot(aes(fill=benefits,x=no_employees,y=1))+geom_bar(position="stack", stat="identity") +
  labs(x="Company size (Number of employees)", y="number of people",title="Does your company provide me
```

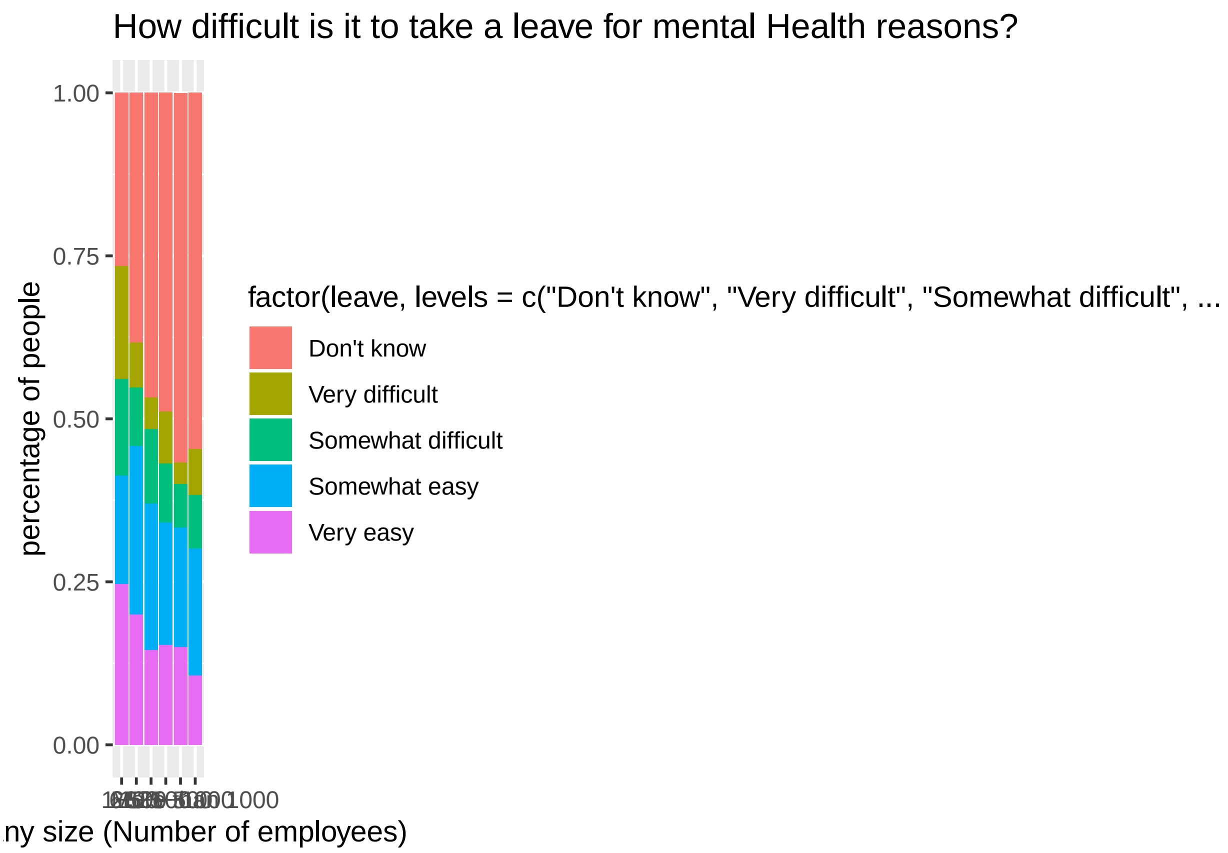


```
survey%>%
  mutate(no_employees = factor(no_employees, levels = c("1-5","6-25","26-100","100-500","500-1000","Mor
  ggplot(aes(fill=benefits,x=no_employees,y=1))+geom_bar(position="fill", stat="identity") +
  labs(x="Company size (Number of employees)", y="percentage of people",title="Does your company provid
```



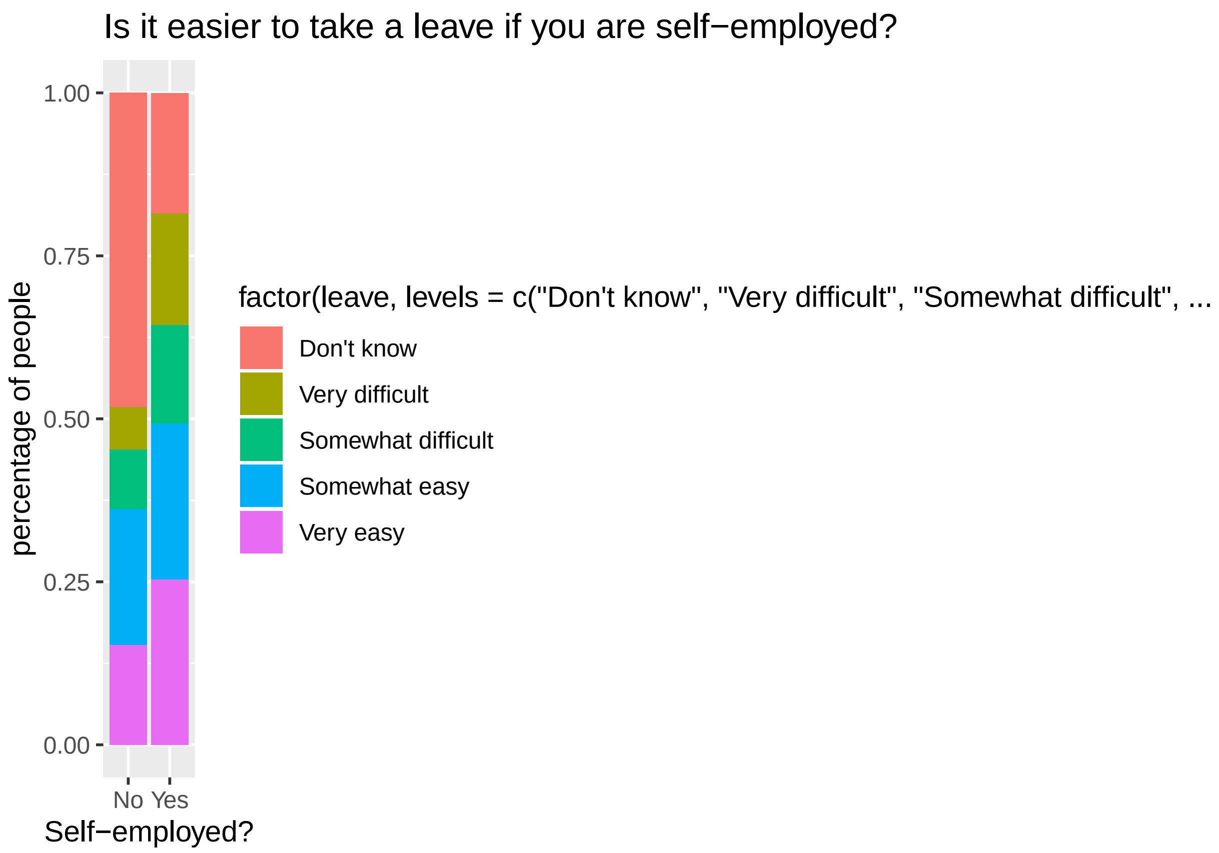
Need to reorder tha bars aes(z, x, fill=factor(y, levels=c(“blue”,“white”)))) + geom_bar(stat = “identity”)

```
survey%>%
  mutate(no_employees = factor(no_employees, levels = c("1-5","6-25","26-100","100-500","500-1000","More
  ggplot(aes(fill=factor(leave, levels=c("Don't know", "Very difficult", "Somewhat difficult", "Somewha
  labs(x="Company size (Number of employees)", y="percentage of people",title="How difficult is it to ta
```



Does self-employment affect how easy it is to take a leave? (Excluding "NA") Should we also exclude "Don't know"? maybe not? It seems that self-employed people face a similar level of difficulty to ask for a leave for mental health reasons - maybe because they are the boss they need to stay even more.

```
survey%>%
  filter(self_employed != "NA")%>%
  ggplot(aes(fill=factor(leave, levels=c("Don't know", "Very difficult", "Somewhat difficult", "Somewha
  labs(x="Self-employed?", y="percentage of people",title="Is it easier to take a leave if you are self-
```

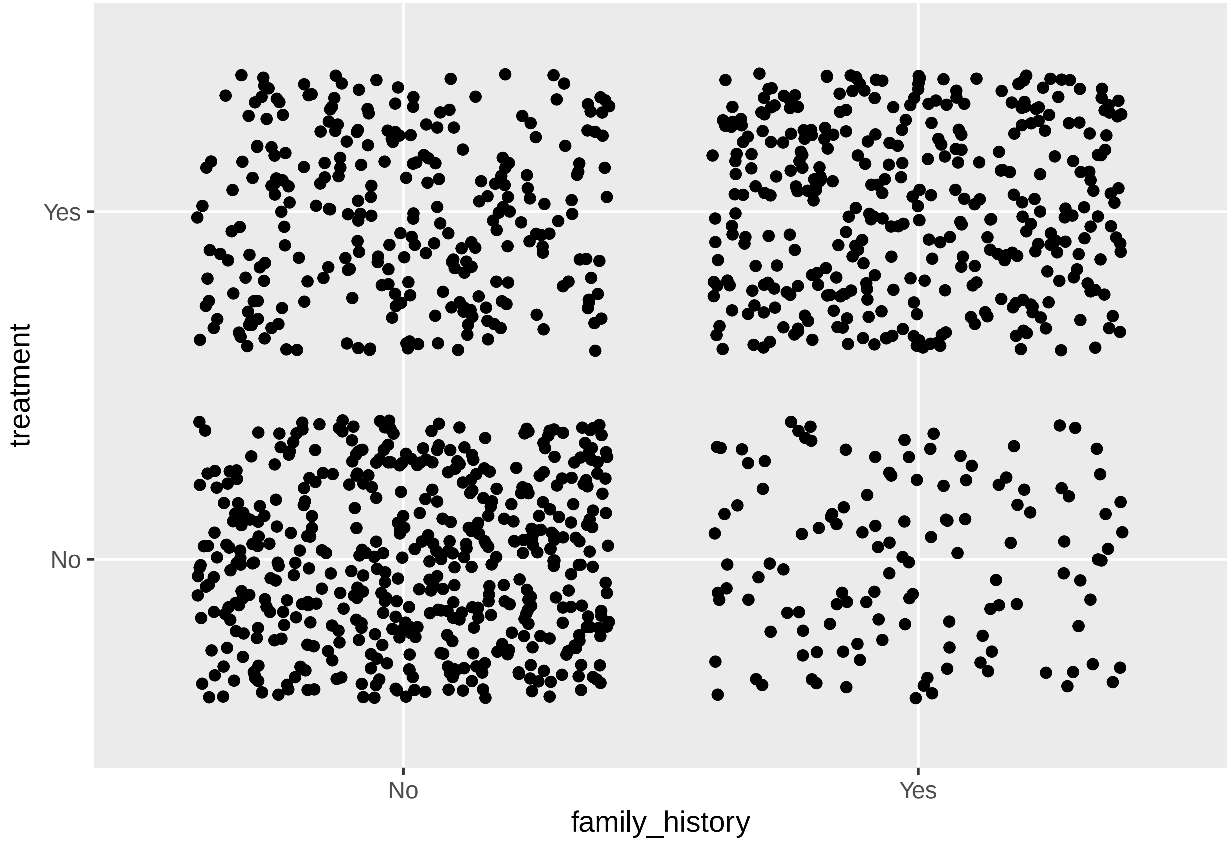


decision tree predicting what factors make people want more benefit?

(Take into account family history)

There is some weak correlation between family history of mental health problem and whether the individual has seeked treatment. However this is not a significant relationship, so most of the difference can be attributed to the other factors.

```
ggplot(survey, aes(x=family_history, y=treatment))+geom_point(position="jitter")
```



Word cloud of comments? It seems our comments need to be read in context, not just key words... not sure if it's a good idea to do a word cloud

```
#comments <- survey$comments
#comments %>% with(wordcloud(comments, n, max.words = 25))
#comments %>% with(wordcloud(comments, n, max.words = 100, random.order = FALSE, colors = brewer.pal(8,
```

Maybe we can make a US state map

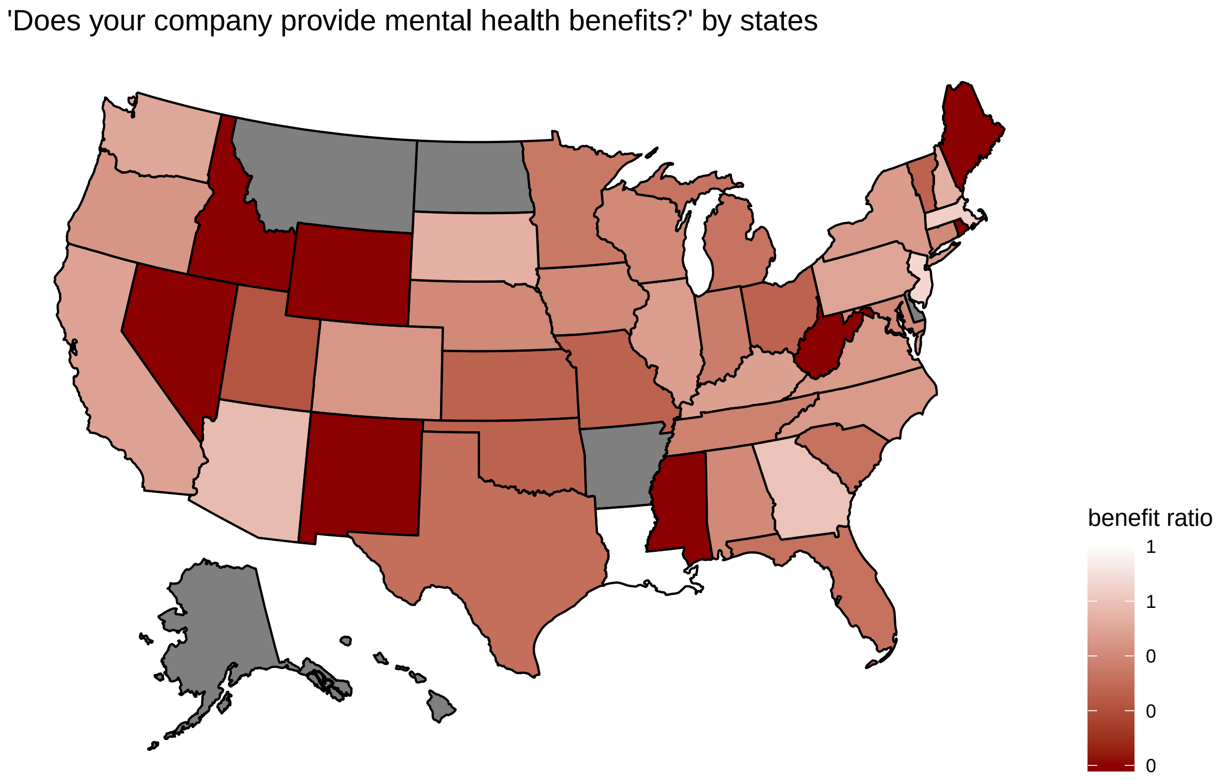
```
library(maps)
library(usmap)
#plot_usmap(regions = "state")
#plot_usmap(regions = "state",label_color = "grey" )
#plot_usmap(include = na.omit(survey$state))
statebenefits <- survey %>% select(state, benefits) %>% na.omit()
states <- unique (statebenefits$state)
```

```
benefitratio <- c()
for (i in 1:length(states)){
  totalstates <- nrow(statebenefits %>% filter(state==states[i]))
  totalyes <- nrow(statebenefits %>% filter(state==states[i] & benefits =="Yes"))
  newratio <- totalyes/totalstates
  benefitratio <- append(benefitratio, newratio)
}
states <- data.frame(states)
```

```
benefitratio <- data.frame(benefitratio)
states <- cbind(states, benefitratio)
states <- states %>%
  rename(state=states)
```

Benefit ratio: “Does your company provide mental health benefits?”

```
plot_usmap(data = states, values = "benefitratio") +
  scale_fill_continuous(name = "benefit ratio", label = scales::comma, low="darkred", high="white") +
  theme(legend.position = "right") +
  labs(title="'Does your company provide mental health benefits?' by states")
```


'Does your company provide mental health benefits?' by states
benefit ratio
1
1
0
0
0