CS100 Final Project

Mariel Pacada

12/7/2019

Data Cleaning

```r
# Mis-typed number of employees
survey[survey$no_employees == "6/25/2019" ] <- as.factor(6-25)
survey[survey$no_employees == "1/5/2019" ] <- as.factor(1-5)
```

```r
unique_gender <- unique(survey$Gender)
levels(survey$Gender)[43] <- "Queer"

survey$Gender <- replace(survey$Gender, survey$Gender == "M", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "male", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "m", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Male-ish", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "maile","Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "something kinda male?", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Cis Male", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Mal", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Male (CIS)", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Make", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Guy (-ish) ^_^", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "male leaning androgynous", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Male ", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Man", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "msle", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Mail", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "cis male", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Malr", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "Cis Man", "Male")
survey$Gender <- replace(survey$Gender, survey$Gender == "ostensibly male, unsure what that really means

survey$Gender <- replace(survey$Gender, survey$Gender == "female", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "Trans-female", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "Cis Female", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "F", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "Woman", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "f", "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$Gender == "Female ", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "cis-female/femme", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "Trans woman", "Female")
survey$Gender <- replace(survey$Gender, survey$Gender == "Female (trans)", "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 == "queer/she/they", "Queer")
```

```r
survey$Gender <- replace(survey$Gender, survey$Gender == "non-binary", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Nah", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "All", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Enby", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "fluid", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Genderqueer", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Androgyne", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Agender", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "Neuter", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "A little about you", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == "p", "Queer")
survey$Gender <- replace(survey$Gender, survey$Gender == NA, "Queer")
```

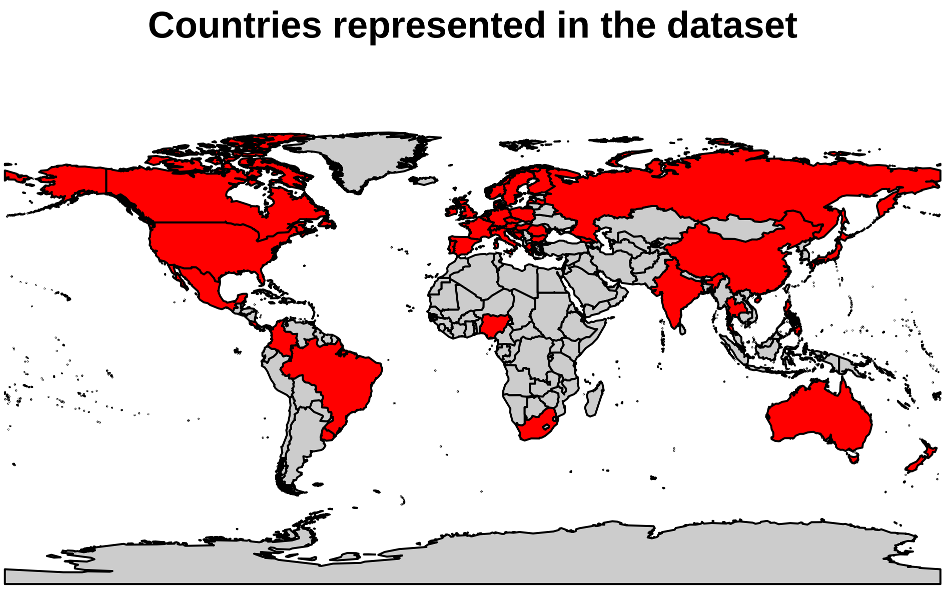
```r
survey <- survey %>%
            filter(Age > 16 & Age < 80)
```

```r
survey$Timestamp <- as.character(survey$Timestamp)

for (i in 1:nrow(survey)) {
  survey$Timestamp[i] <- substring(survey$Timestamp[i], 1, 4)
}

survey$Timestamp <- as.numeric(survey$Timestamp)
colnames(survey)[1] <- "Year"

survey <- survey %>%
            filter(Year == 2014 | Year == 2015)
```

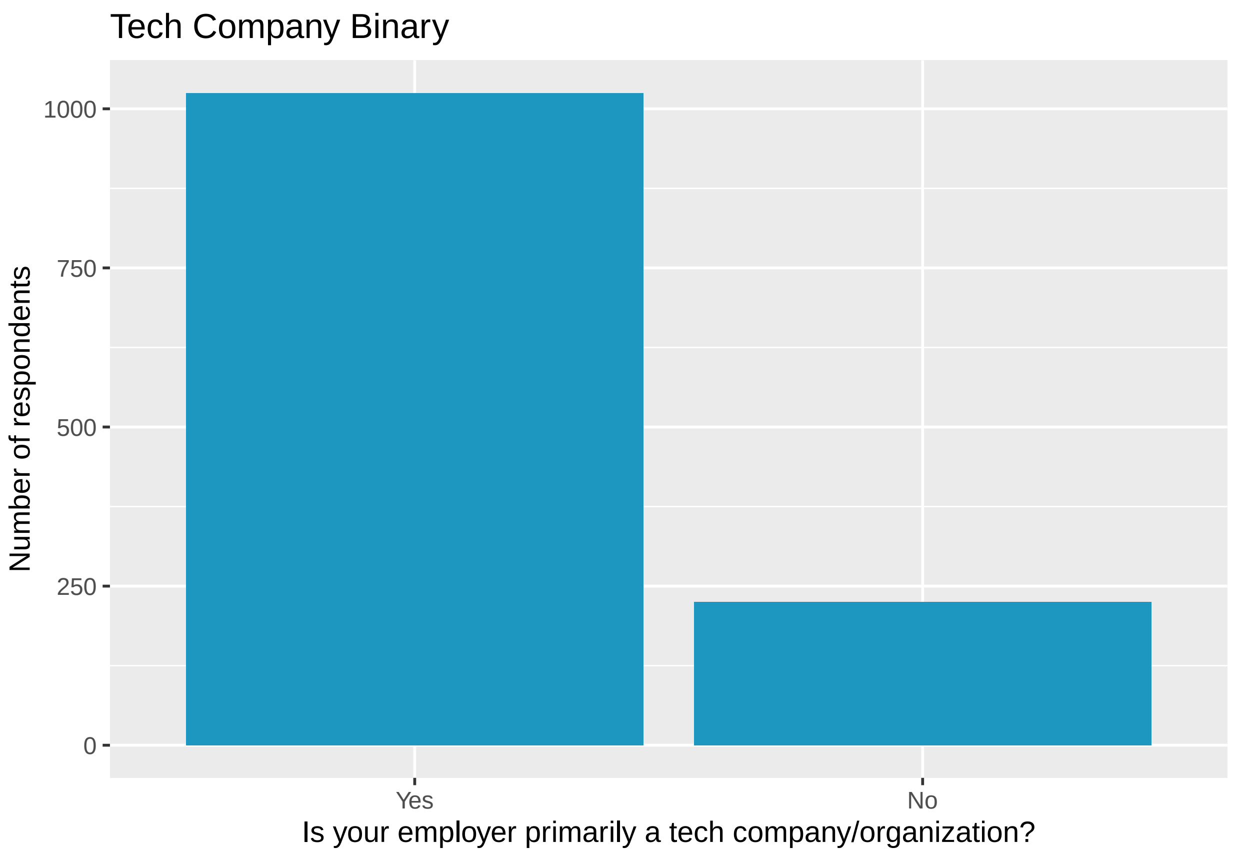
```r
colnames(countries)[5] <- "Country"
countries$Country[200] <- "Russia"

survey_countries <- unique(survey$Country)
survey_countries <- matrix(survey_countries, ncol = 1, byrow = TRUE)
survey_countries <- as.data.frame(survey_countries, stringsAsFactors = FALSE)
colnames(survey_countries) <- c("Country")

survey_countries <- merge(survey_countries, countries, by = "Country")


survey_countries <- survey_countries %>%
                      select(Country, Country.Code)
```

Part 1: Exploratory Data Analysis

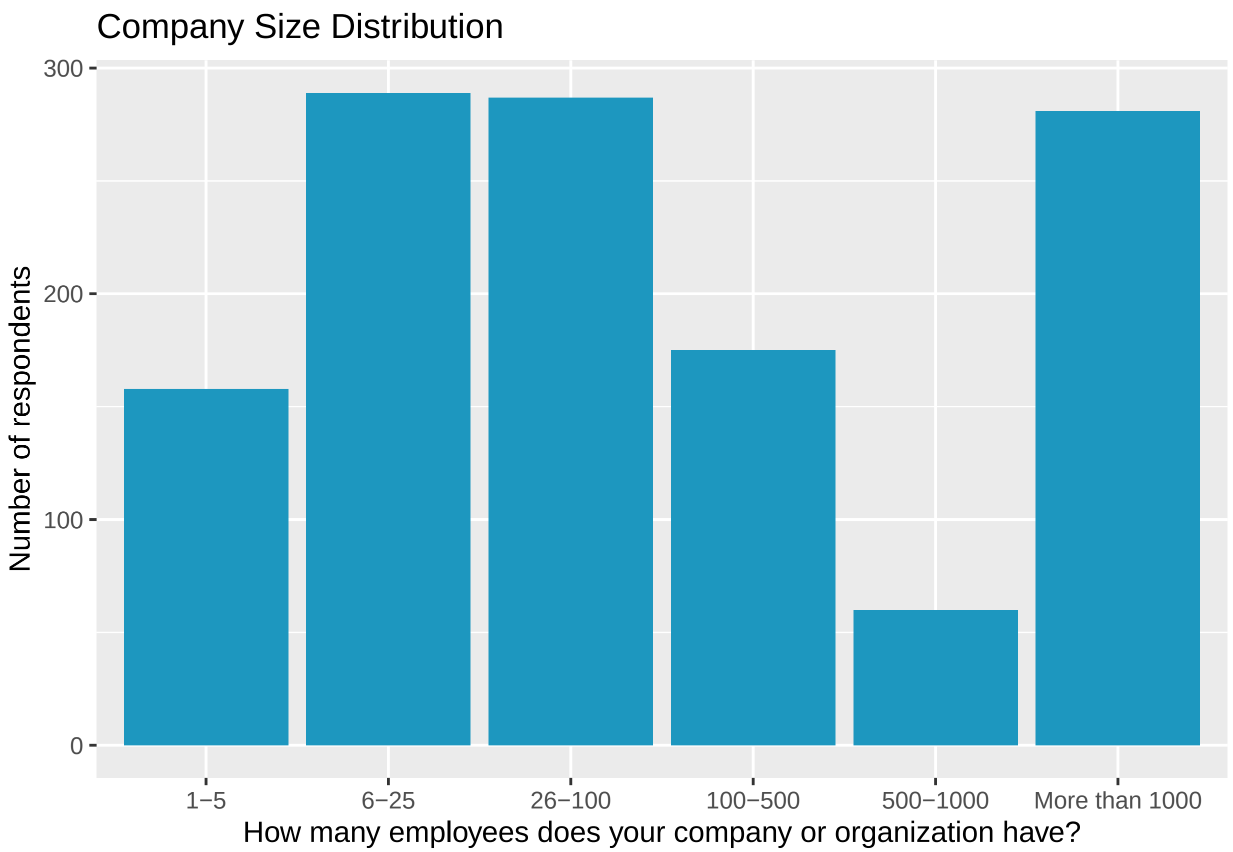
```r
data(wrld_simpl)
map_countries = wrld_simpl@data$NAME %in% survey_countries$Country

plot(wrld_simpl, col = c(gray(.80), "red")[map_countries+1], main = "Countries represented in the datas
```



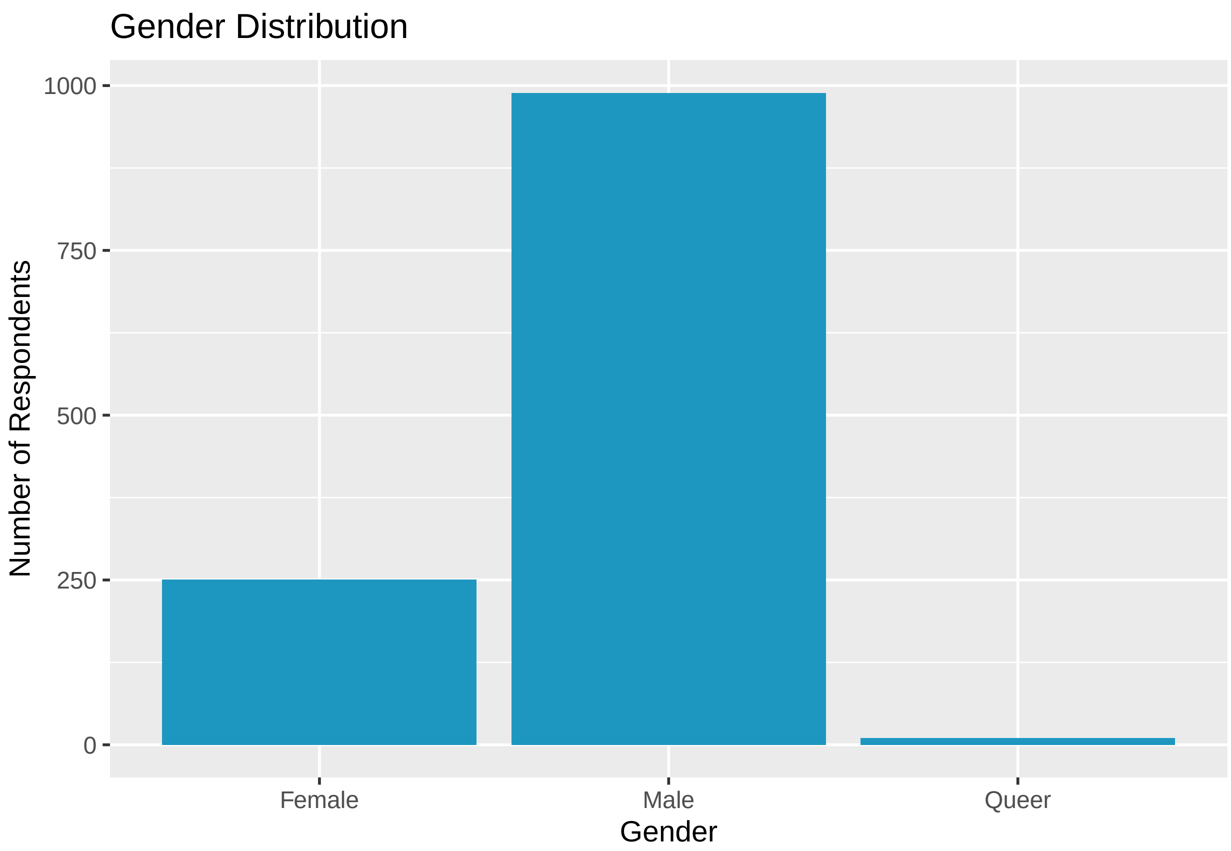
```
tech_binary <- survey %>%
                  filter(!is.na(tech_company)) %>%
                  select(tech_company) %>%
                  mutate(tech_company = factor(tech_company, levels = c("Yes", "No")))

ggplot(tech_binary, aes(x = tech_company)) + geom_bar(fill = "#1D97BF") + labs(x = "Is your employer pri
```



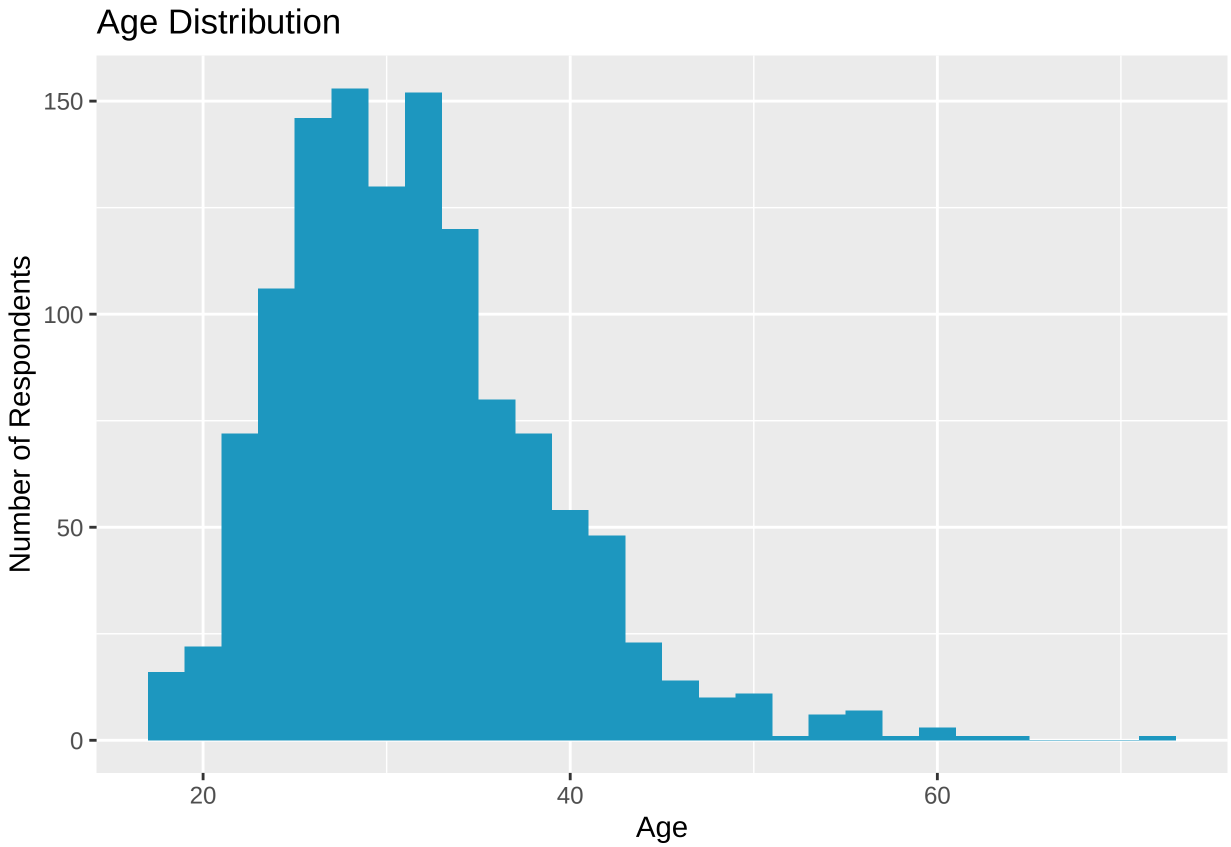
```
num_employee <- survey %>%
                  select(no_employees) %>%
                  mutate(no_employees = factor(no_employees, levels = c("1-5", "6-25", "26-100", "100-50
                                                                        "500-1000", "More than 1000")))

ggplot(num_employee, aes(x = no_employees)) + geom_bar(fill = "#1D97BF") + labs(x = "How many employees
```



```
ggplot(survey, aes(x = Gender)) + geom_bar(fill = "#1D97BF") + labs(y = "Number of Respondents", title
```

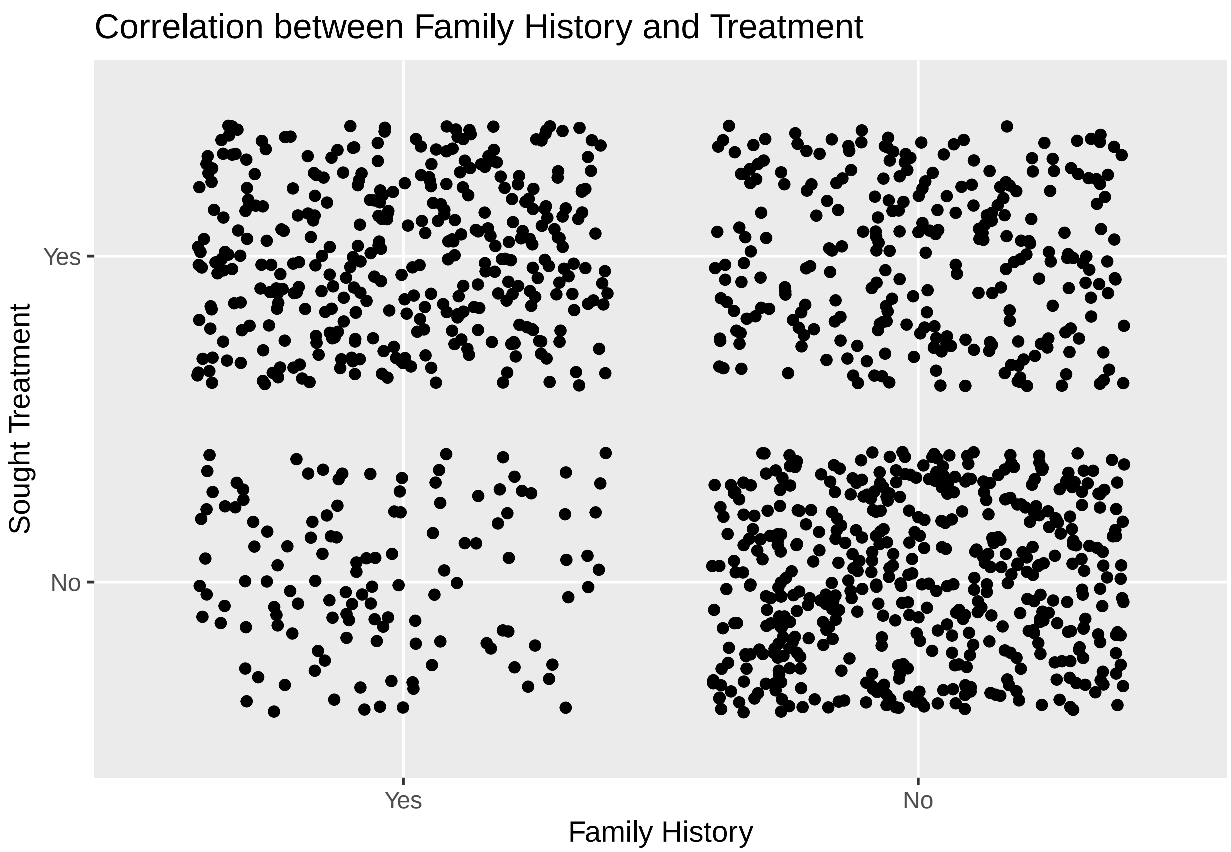


```
ggplot(survey, aes(x = Age) )+ geom_histogram(binwidth = 2, fill = "#1D97BF") + labs(y = "Number of Resp
```

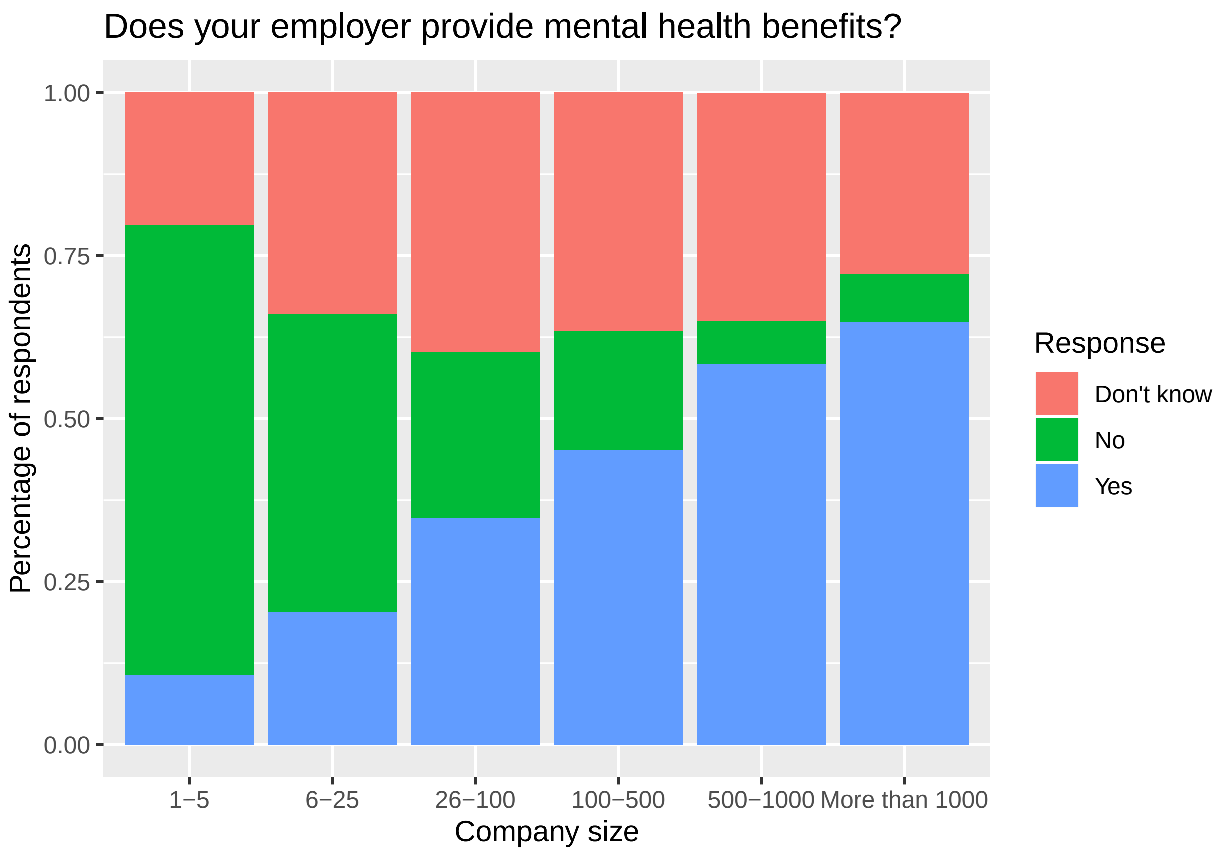


Part 2: Categorical Data Analysis

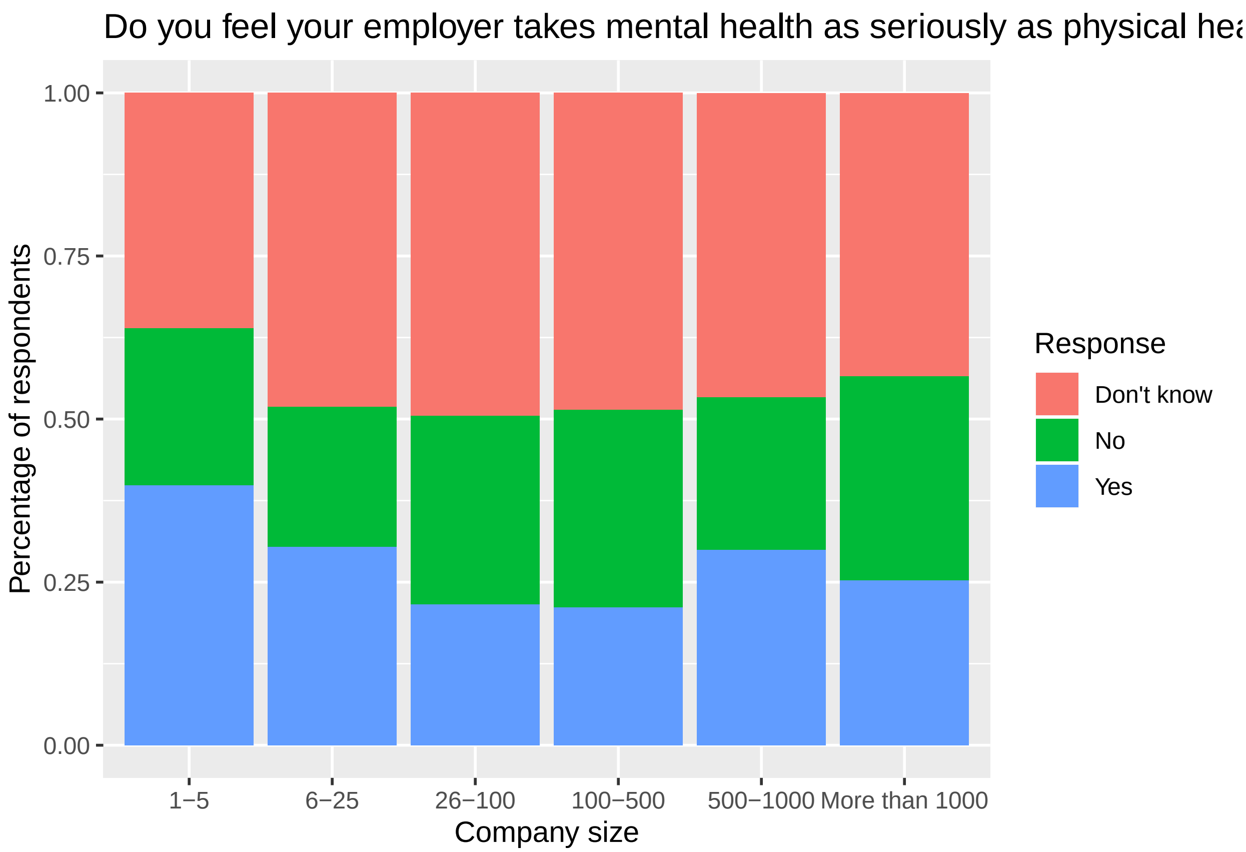
```
fam_treatment <- survey %>%
                 select(family_history, treatment) %>%
                 mutate(family_history = factor(family_history, levels = c("Yes", "No")))

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



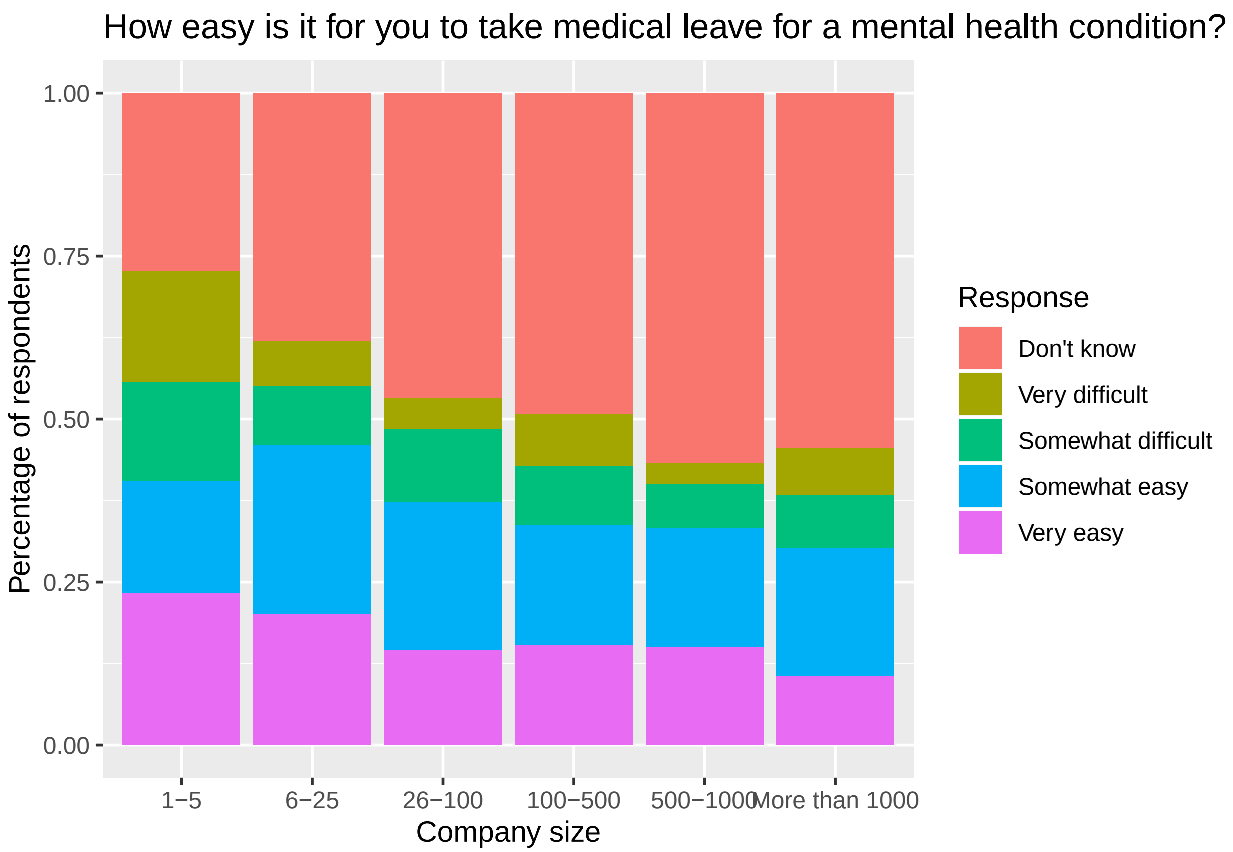
```
emp_benefits <- survey %>%
                  select(no_employees, benefits) %>%
                  mutate(no_employees = factor(no_employees, levels = c("1-5", "6-25", "26-100", "100-5
                                                                         "500-1000", "More than 1000")))

ggplot(emp_benefits, aes(x = no_employees, y = 1, fill = benefits)) + geom_bar(position = "fill", stat =
```



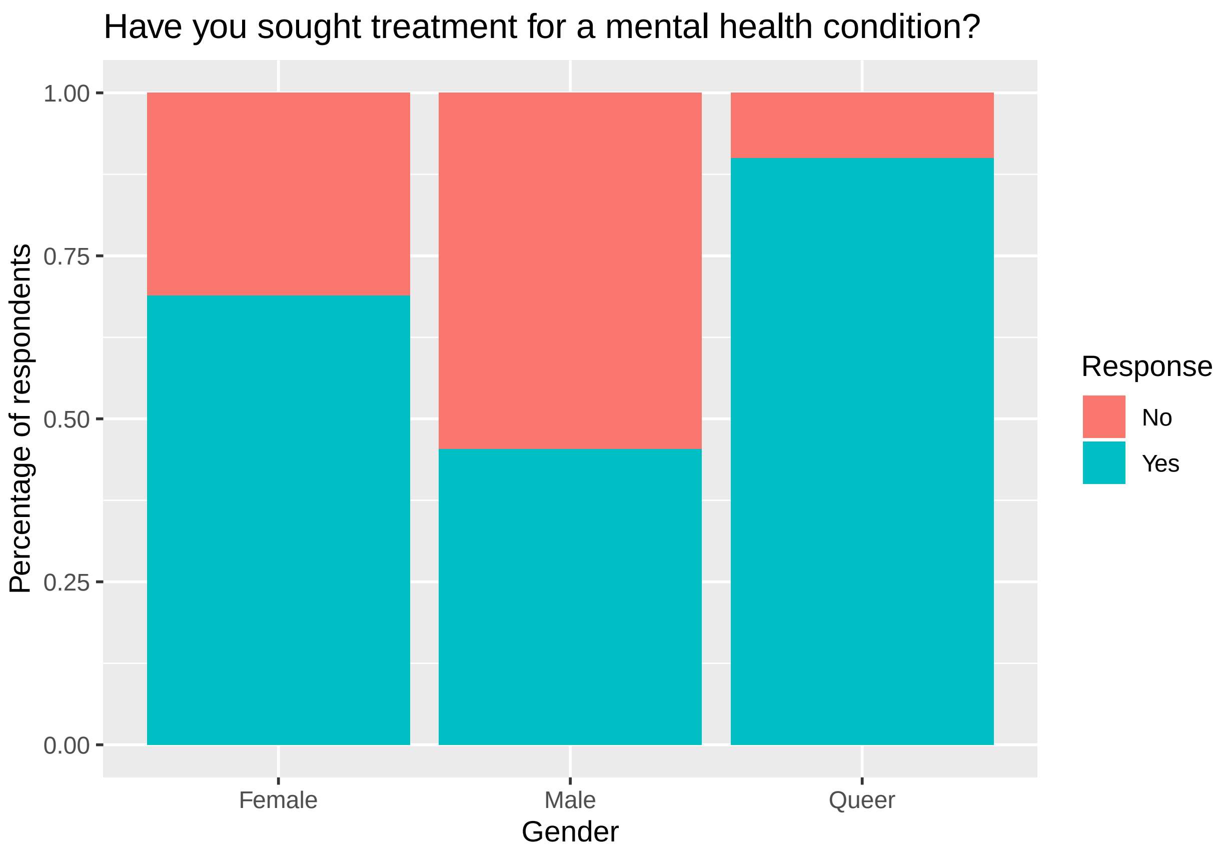
```
ment_vs_phys <- survey %>%
                  select(no_employees, mental_vs_physical) %>%
                  mutate(no_employees = factor(no_employees, levels = c("1-5", "6-25", "26-100", "100-5
                                                                        "500-1000", "More than 1000")))

ggplot(ment_vs_phys, aes(x = no_employees, y = 1, fill = mental_vs_physical)) + geom_bar(position = "fi
```



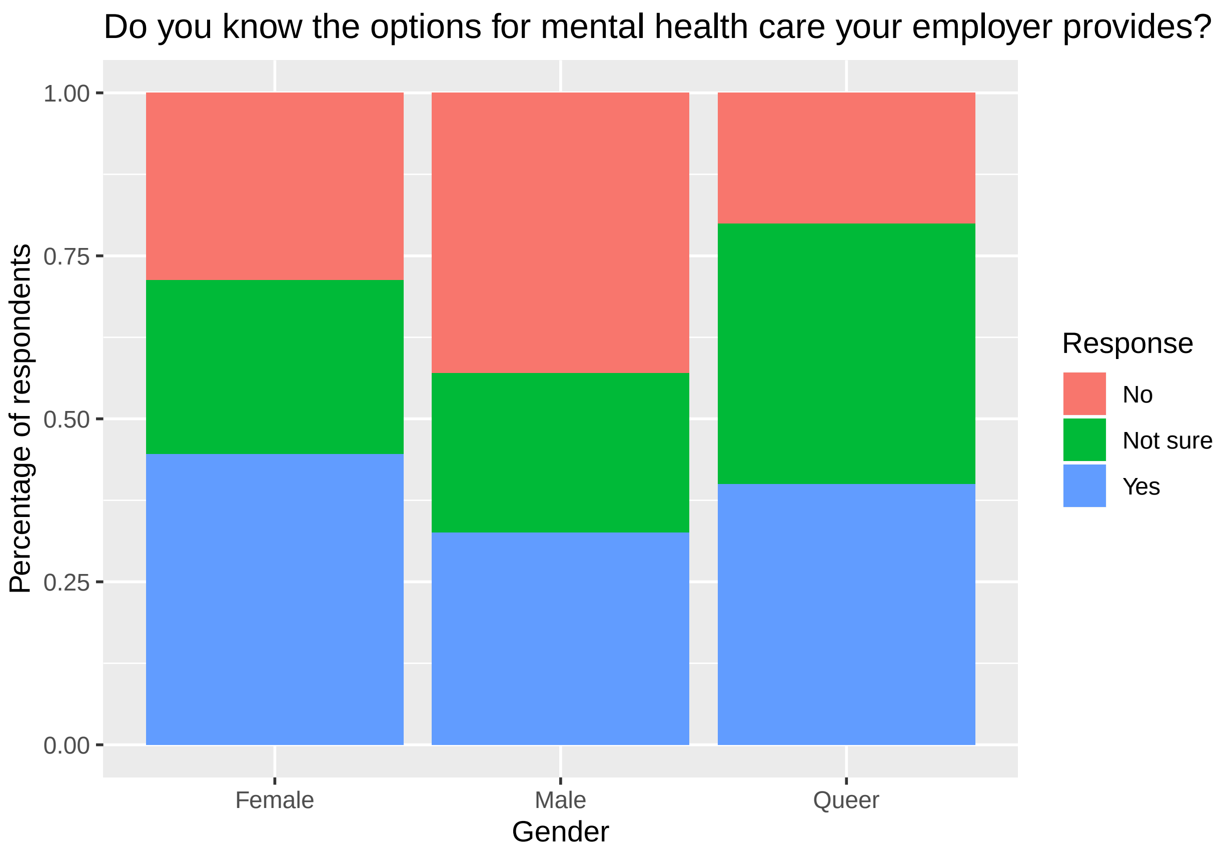
```
emp_leave <- survey %>%
               select(no_employees, leave) %>%
               mutate(no_employees = factor(no_employees, levels = c("1-5", "6-25", "26-100", "100-500"
                                                                     "500-1000", "More than 1000")))

ggplot(emp_leave, aes(x = no_employees, y = 1, fill = factor(leave, levels = c("Don't know", "Very diff
```



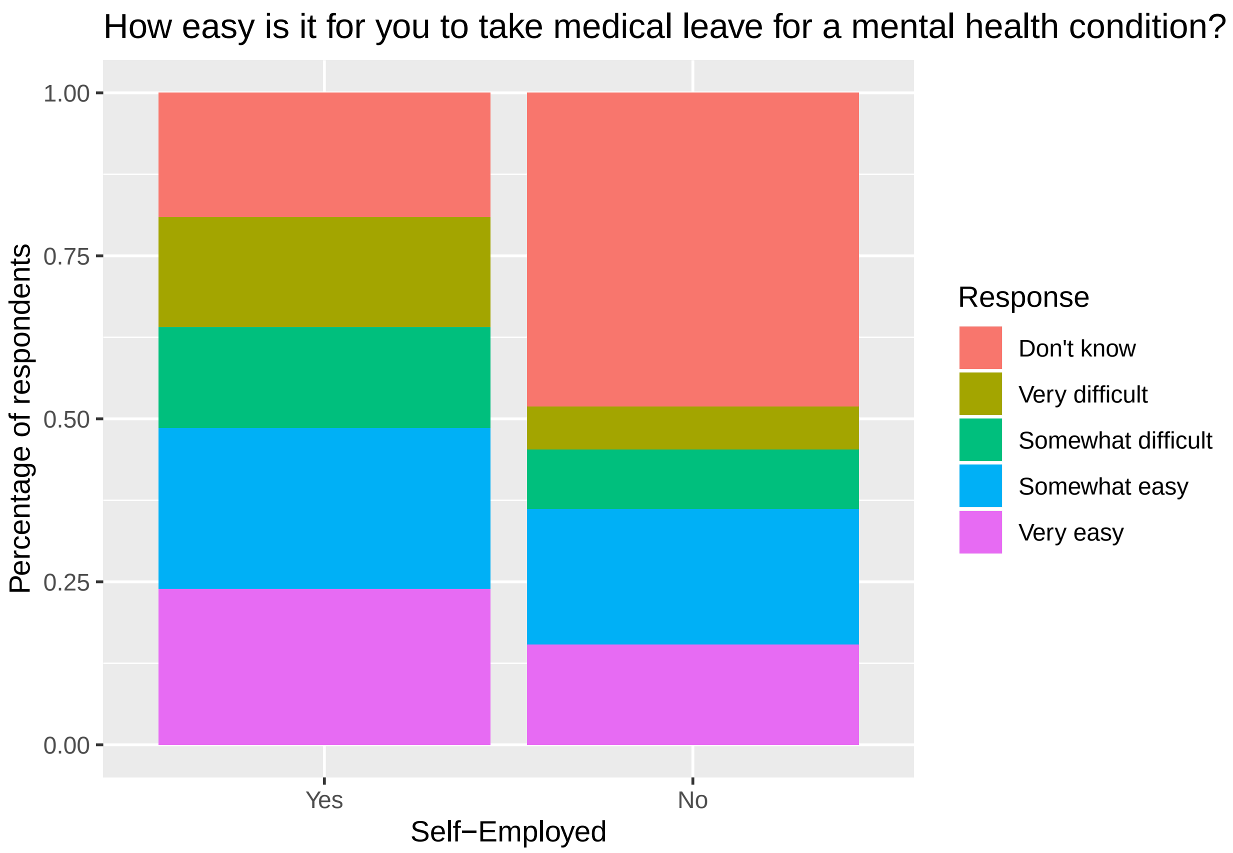
```
ggplot(survey, aes(x = Gender, y = 1, fill = treatment)) + geom_bar(position = "fill", stat = "identity
```



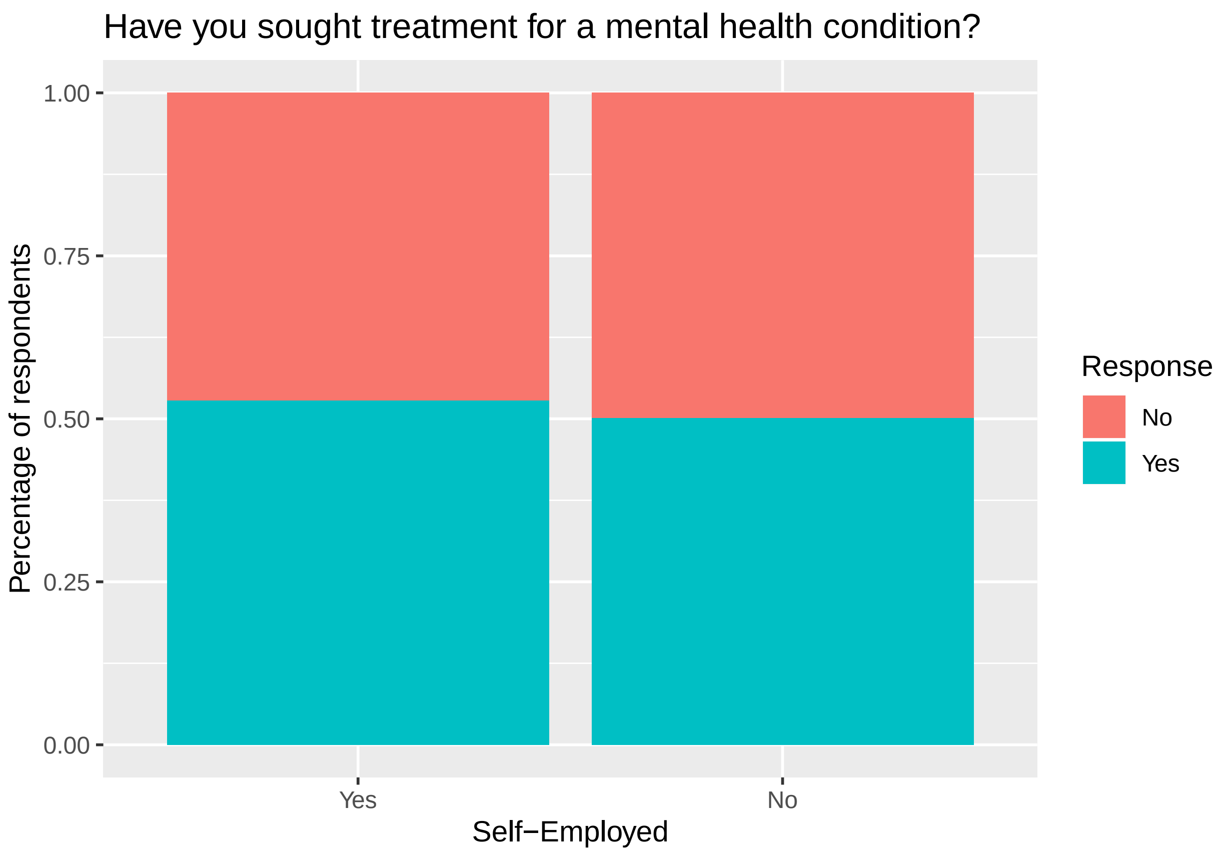
```
ggplot(survey, aes(x = Gender, y = 1, fill = care_options)) + geom_bar(position = "fill", stat = "ident
```



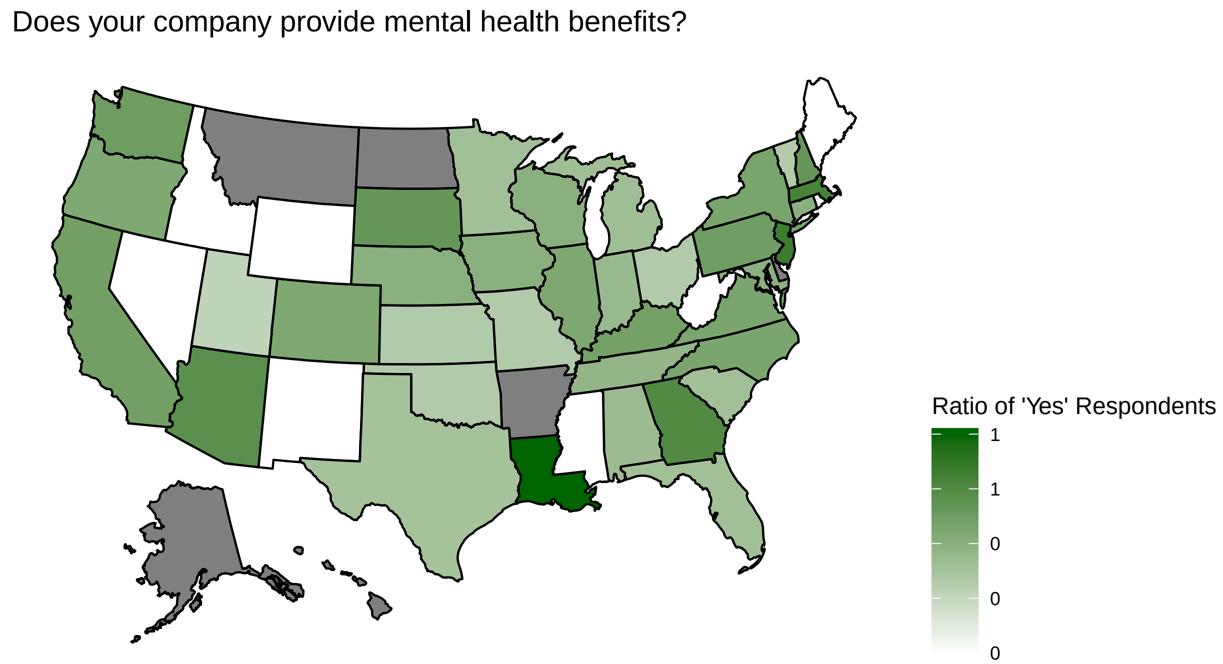
```
self_leave <- survey %>%
              filter(!is.na(self_employed)) %>%
              select(self_employed, leave) %>%
              mutate(self_employed = factor(self_employed, levels = c("Yes", "No")))

ggplot(self_leave, aes(x = self_employed, y = 1, fill = factor(leave, levels = c("Don't know", "Very di
```



```
self_treatment <- survey %>%
                    filter(!is.na(self_employed)) %>%
                    select(self_employed, treatment) %>%
                    mutate(self_employed = factor(self_employed, levels = c("Yes", "No")))

ggplot(self_treatment, aes(x = self_employed, y = 1, fill = treatment)) + geom_bar(position = "fill", s
```



```
state_benefits <- survey %>% select(state, benefits) %>% na.omit()
states <- unique(state_benefits$state)

benefit_ratio <- c()
for (i in 1:length(states)){
  total_states <- nrow(state_benefits %>% filter(state == states[i]))
  total_yes <- nrow(state_benefits %>% filter(state == states[i] & benefits =="Yes"))
  new_ratio <- total_yes/total_states
  benefit_ratio <- append(benefit_ratio, new_ratio)
}

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

plot_usmap(data = states, values = "benefit_ratio") +
  scale_fill_continuous(name = "Ratio of 'Yes' Respondents", label = scales::comma, low = "white", high
                        "darkgreen") + theme(legend.position = "right") +
                        labs(title = "Does your company provide mental health benefits?")
```


Does your company provide mental health benefits?
Ratio of 'Yes' Respondents
1
1
0
0
0

Part 3: Classification

Part 4: Hypothesis Testing

```r
gender_vs_treatment <- survey %>%
                       select(Gender, treatment) %>%
                       filter(Gender != "Queer")

gender_table <- with(gender_vs_treatment, table(Gender, treatment))
gender_table <- gender_table[c(16, 30), 1:2]

gender_t_test <- fisher.test(gender_table)
gender_t_test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  gender_table
## p-value = 2.114e-11
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2753241 0.5080557
## sample estimates:
## odds ratio
##   0.375182
```

```
gender_chi_test <- chisq.test(gender_table)
gender_chi_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_table
## X-squared = 43.381, df = 1, p-value = 4.506e-11
```

```
country_table <- with(survey, table(Country, benefits))
country_table <- addmargins(country_table, FUN = list(Total = sum), quiet = TRUE)
country_table <- country_table[, 2:3]

north_america <- country_table[8,] + country_table[46,]
not_north_america <- country_table[49,] - north_america

country_table <- matrix(c(north_america, not_north_america), ncol = 2, byrow = TRUE)
colnames(country_table) <- c("No", "Yes")
rownames(country_table) <- c("North America", "Other")
country_table <- as.table(country_table)

country_t_test <- fisher.test(country_table)
country_t_test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  country_table
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.04623651 0.09883596
## sample estimates:
## odds ratio
## 0.06809495
```

```
country_chi_test <- chisq.test(country_table)
country_chi_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  country_table
## X-squared = 256.25, df = 1, p-value < 2.2e-16
```