Salaries Analysis Exercise (NPCA)

Daina Bouquin

```
library(ggplot2)
library(tidyverse)
library(scales)
library(psych)
library(ggpubr)
library(car)
library(knitr)
library(rmarkdown)
```

Five Questions

1. Which industries are the top 10% of earners associated with?
2. What are the characteristics of the top 10% of earners in each industry?
3. Are there statistically significant annual salary differences between genders in each industry? Overall?
4. Which industries pay the highest at entry level?
5. Which states have the highest salaries? What about just among nonprofits?

Data import and more cleanup

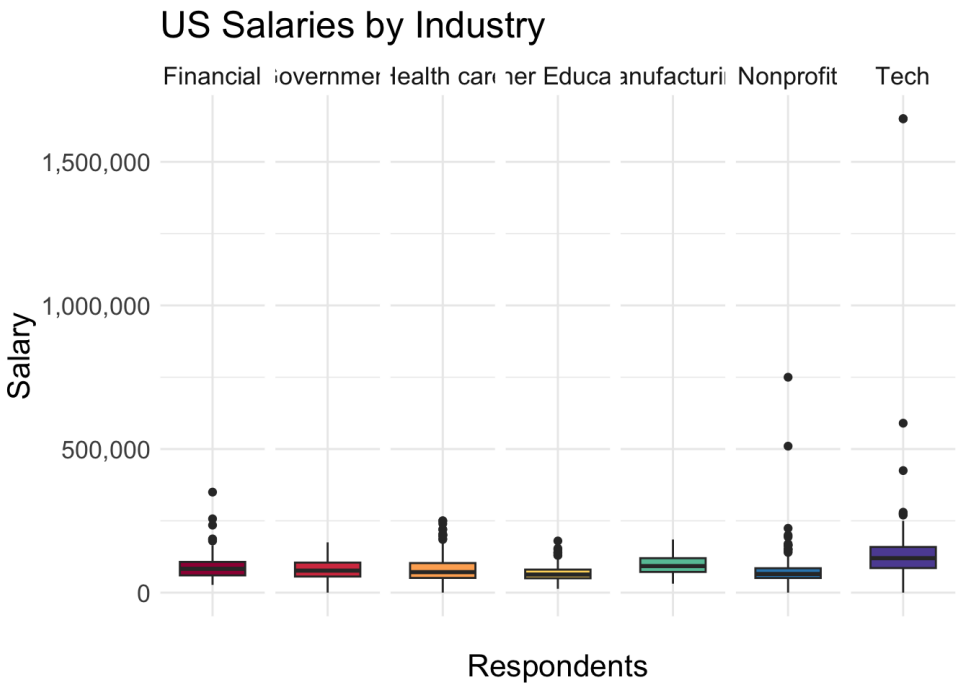
Subset to only keep data from the United States where the currency is USD. Convert the relevant columns to factors for plotting. I also create a new column to note whether or not a respondent is white or nonwhite and then subset the dataset to only include industries with at least 100 respondents.

```
salaries <- read.csv("clean_salaries.csv")

# remove "X" index column
salaries <- (subset(salaries, select = -c(X)))

# remove where country is not United States
salaries <- salaries[salaries$country == 'United States',]

# remove where currency is not USD
salaries <- salaries[salaries$currency == 'USD',]

# drop other currency column
salaries <- (subset(salaries, select = -c(other_currency)))

# convert categorical variables to factors
salaries$gender <- as.factor(salaries$gender)
salaries$race <- as.factor(salaries$race)
salaries$multiracial <- as.factor(salaries$multiracial)
salaries$state <- as.factor(salaries$state)
salaries$multistate <- as.factor(salaries$multistate)
salaries$industry <- as.factor(salaries$industry)
salaries$age_range <- as.factor(salaries$age_range)
salaries$all_experience <- as.factor(salaries$all_experience)
salaries$field_experience <- as.factor(salaries$field_experience)
salaries$job <- as.factor(salaries$job)
salaries$education.level <- as.factor(salaries$education.level) # column naming?

# Look at white vs. non-white - create categories
# Would keep the more granular data if I was doing a more granular analysis

# if white and not multiracial = white, otherwise = nonwhite
white_salaries <- subset(salaries, (multiracial== "No"))
white_salaries <- subset(white_salaries, !(race != "White"))

# create new column for white vs. nonwhite
salaries <- transform(
  salaries, whiteness = ifelse(id %in% white_salaries$id, "white", "nonwhite"))
salaries$whiteness <- as.factor(salaries$whiteness)

# drop the original race column
salaries <- (subset(salaries, select = -c(race)))

# drop records associated with more than one state (not looking into this aspect of the data and this will simpli
fy de-duplication)
salaries <- salaries[!(salaries$multistate=="Yes"),]
# drop column for multistate
salaries <- (subset(salaries, select = -c(multistate)))

# drop context columns
salaries <- (subset(salaries, select = -c(income_context)))
salaries <- (subset(salaries, select = -c(job_context)))

# Select industries where there are at least 100 people (meaningful sample of respondents)
# ** Here we create salaries2 **
industry_count <- table(salaries$industry)
salaries2 <- subset(salaries, industry %in% names(industry_count[industry_count > 100]))

# How many industries do we have now? (originally 70)
# length(unique(salaries2$industry)) # down to 7

# remove duplicates
salaries <- salaries[!duplicated(salaries$id), ]

# double check for duplicates
n_occur <- data.frame(table(salaries$id))
# n_occur[n_occur$Freq > 1,] # no duplicates left
```

Outlier detection and removal

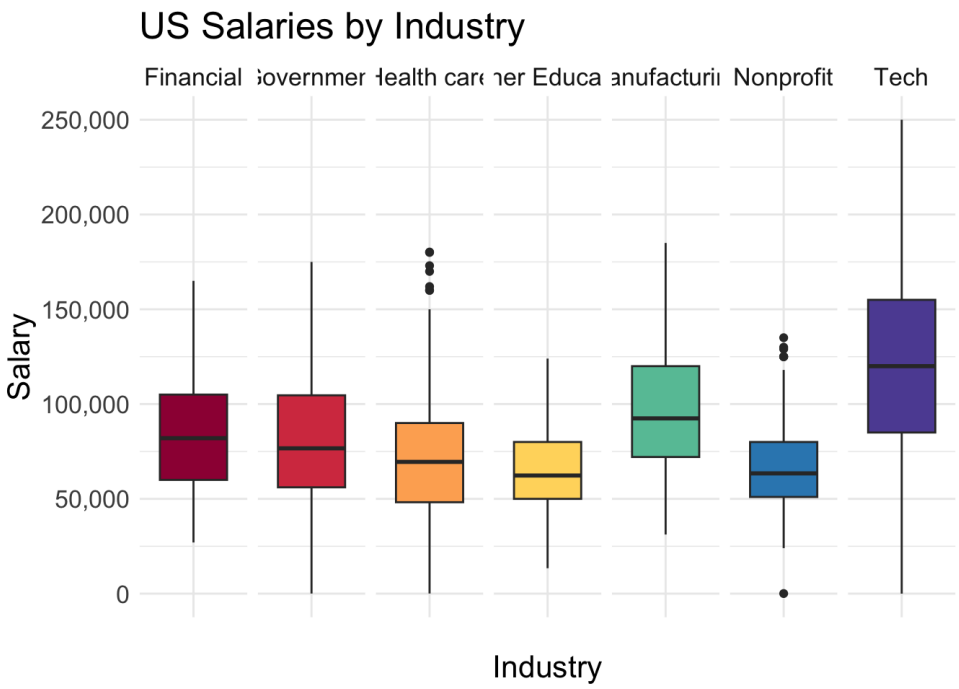
Outliers can actually be really useful so we'll identify them and store that information.

```
# Start with visual inspection

# build pretty palette:
ppretty <- c("#9e0142", "#d53e4f", "#fdae61", "#ffd86b", "#66c2a5", "#3288bd", "#5e4fa2")

# create boxplot
bp <- ggplot(salaries2, aes(x="", y=annual_salary, group=industry)) +
  geom_boxplot(aes(fill=industry)) + theme_minimal()
bp <- bp + scale_y_continuous(labels = label_comma())
bp <- bp + facet_grid(. ~ industry)
bp <- bp + scale_fill_manual(values=ppretty)
bp <- bp + theme(legend.position="none") # Remove legend
bp <- bp + theme(text = element_text(size=16), axis.title=element_text(size=16))
bp <- bp + labs(title = "US Salaries by Industry", x= "Respondents", y= "Salary")
bp
```



In ggplot2, an observation is defined as an outlier if it meets one of the following two requirements: * The observation is 1.5 times the interquartile range less than the first quartile (Q1) * The observation is 1.5 times the interquartile range greater than the third quartile (Q3).

```
# create a function to replicate outlier conditions
find_outlier <- function(x) {
  return(x < quantile(x, .25) - 1.5*IQR(x) | x > quantile(x, .75) + 1.5*IQR(x))
}

# create column to indicate suspected outliers
salaries2 <- salaries2 %>%
        group_by(industry) %>%
        mutate(outlier = ifelse(find_outlier(annual_salary), "Yes", "No"))

# create subset to work with where the outliers are removed
# ** Here we create salaries3 **
salaries3 <- salaries2[salaries2$outlier == 'No',]

# create a new boxplot using data without outliers
bp2 <- ggplot(salaries3, aes(x="", y=annual_salary, group=industry)) +
  geom_boxplot(aes(fill=industry)) + theme_minimal()
bp2 <- bp2 + scale_y_continuous(labels = label_comma())
bp2 <- bp2 + facet_grid(. ~ industry)
bp2 <- bp2 + scale_fill_manual(values=ppretty)
bp2 <- bp2 + theme(legend.position="none") # Remove legend
bp2 <- bp2 + theme(text = element_text(size=16), axis.title=element_text(size=16))
bp2 <- bp2 + labs(title = "US Salaries by Industry", x= "Industry", y= "Salary")
bp2
```



1. Which industries are the top 10% of earners associated with?

```
# top 10% overall (independent of industry)
n <- 10
top_10p <- salaries3[salaries3$annual_salary > quantile(salaries3$annual_salary, prob=1-n/100),]
```

```
kable(unique(top_10p$industry))
```

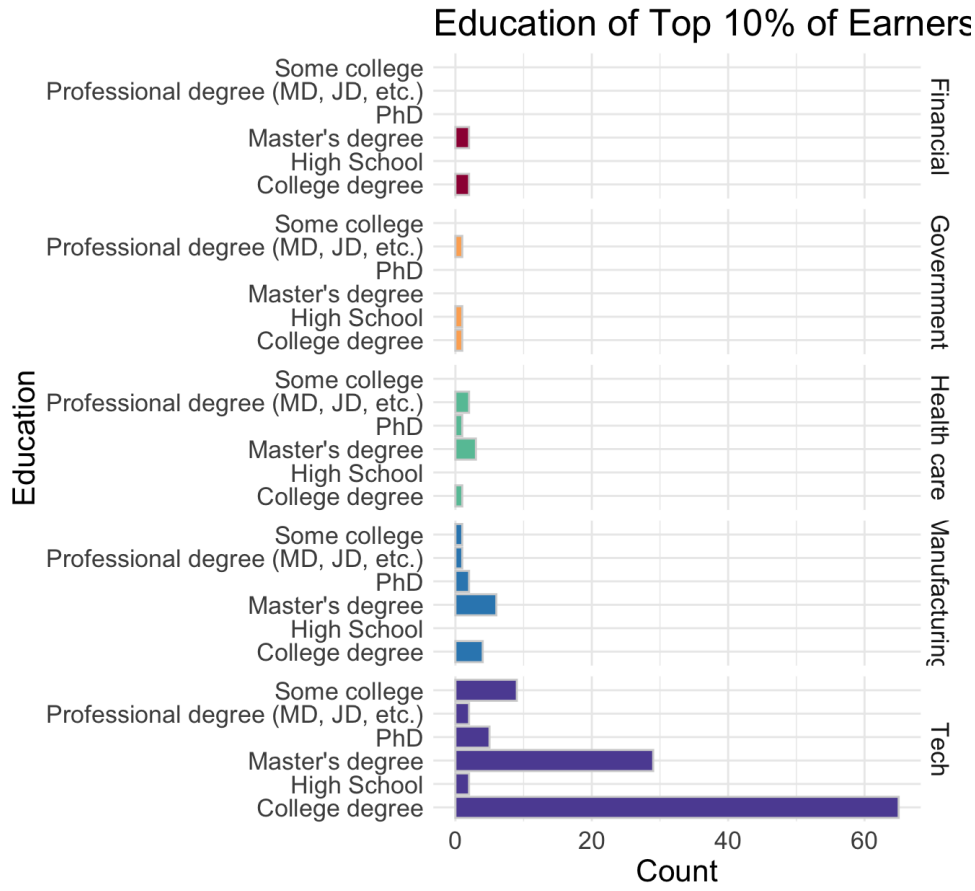
x
Manufacturing
Financial
Tech
Government
Health care

2. What are the characteristics of the top 10% of earners in each industry?

```
# find the top 10% of earners within each industry
by_industry <- salaries3 %>% group_by(industry)
top_10p_industry <- by_industry[by_industry$annual_salary > quantile(by_industry$annual_salary, prob=1-n/100),]
```

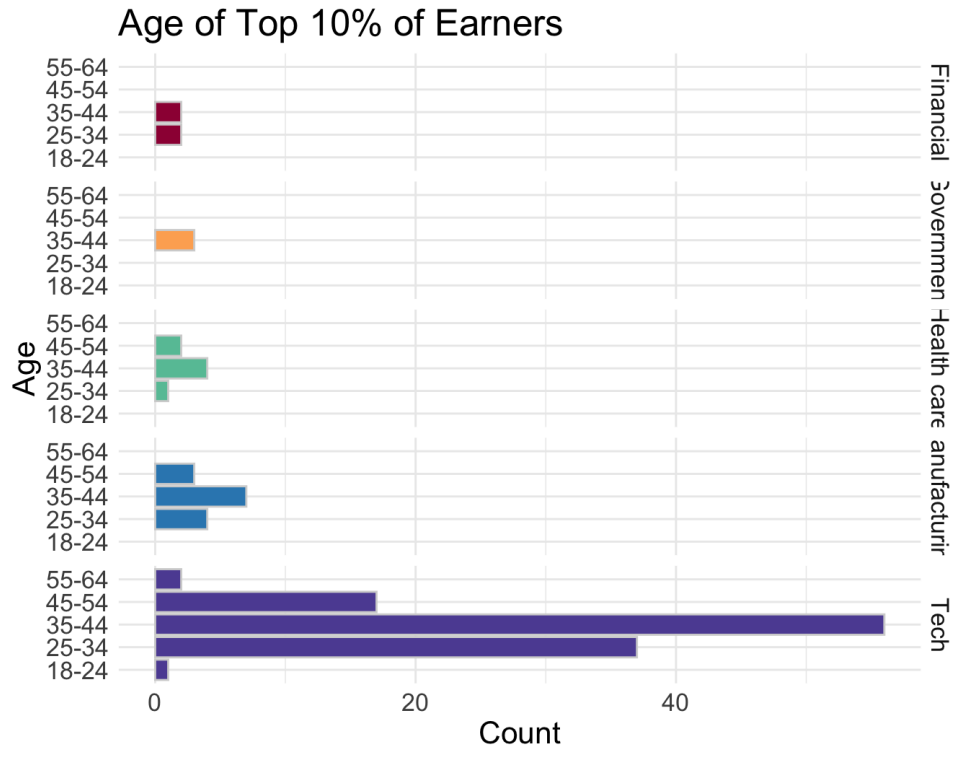
Education

```
# remove records without answers
top_10p_industry_plot1  <- top_10p_industry[!(top_10p_industry$education.level=="NaN"),]

# new pretty palette
ppretty2 <- c("#9e0142", "#fdae61", "#66c2a5", "#3288bd", "#5e4fa2")

h1 <- ggplot(top_10p_industry_plot1, aes(x=education.level, fill=industry)) + geom_histogram(stat="count", color=
"lightgrey")
h1 <- h1 + facet_grid(industry ~ .) + coord_flip() + theme_minimal()
h1 <- h1 + scale_fill_manual(values=ppretty2)
h1 <- h1 + theme(legend.position="none") # Remove legend
h1 <- h1 + theme(text = element_text(size=16), axis.title=element_text(size=16))
h1 <- h1 + labs(title = "Education of Top 10% of Earners", x= "Education", y= "Count")
h1
```



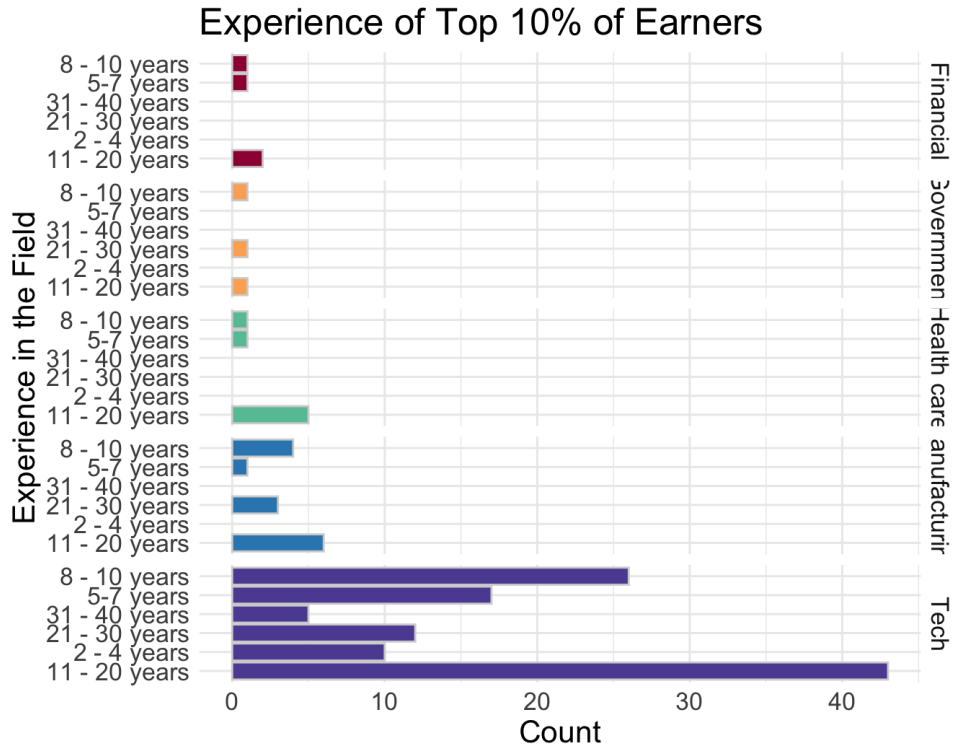
Age

```
h2 <- ggplot(top_10p_industry, aes(x=age_range, fill=industry)) + geom_histogram(stat="count", color="lightgrey")
h2 <- h2 + facet_grid(industry ~ .) + coord_flip() + theme_minimal()
h2 <- h2 + scale_fill_manual(values=ppretty2)
h2 <- h2 + theme(legend.position="none") # Remove legend
h2 <- h2 + theme(text = element_text(size=16), axis.title=element_text(size=16))
h2 <- h2 + labs(title = "Age of Top 10% of Earners", x= "Age", y= "Count")
h2
```



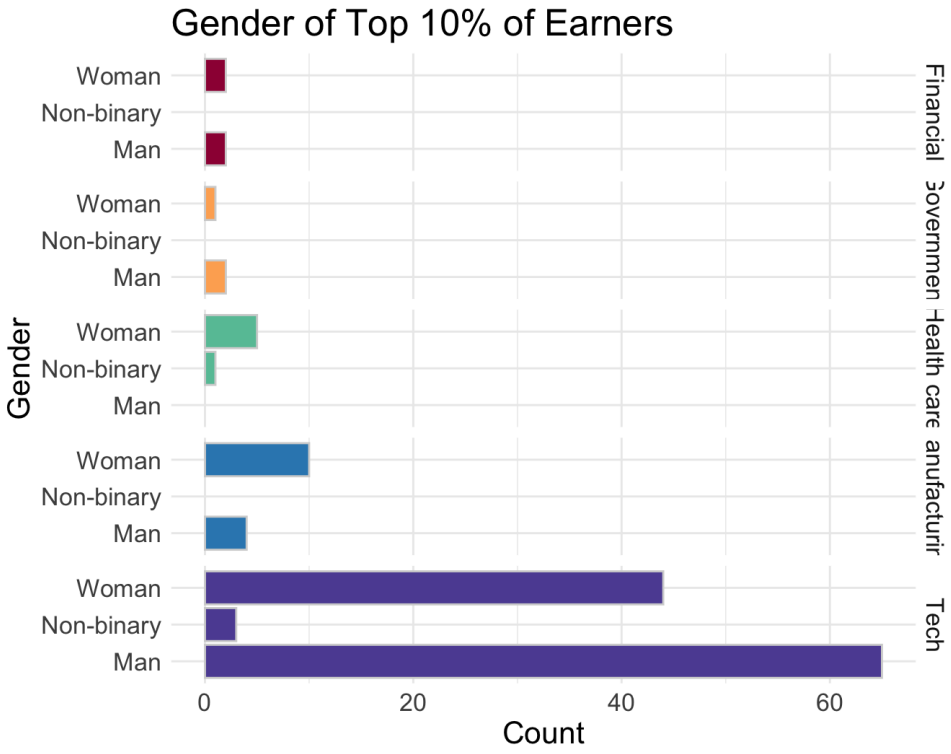
Experience

```
h3 <- ggplot(top_10p_industry, aes(x=field_experience, fill=industry)) + geom_histogram(stat="count", color="ligh
tgrey")
h3 <- h3 + facet_grid(industry ~ .) + coord_flip() + theme_minimal()
h3 <- h3 + scale_fill_manual(values=ppretty2)
h3 <- h3 + theme(legend.position="none") # Remove legend
h3 <- h3 + theme(text = element_text(size=16), axis.title=element_text(size=16))
h3 <- h3 + labs(title = "Experience of Top 10% of Earners", x= "Experience in the Field", y= "Count")
h3
```



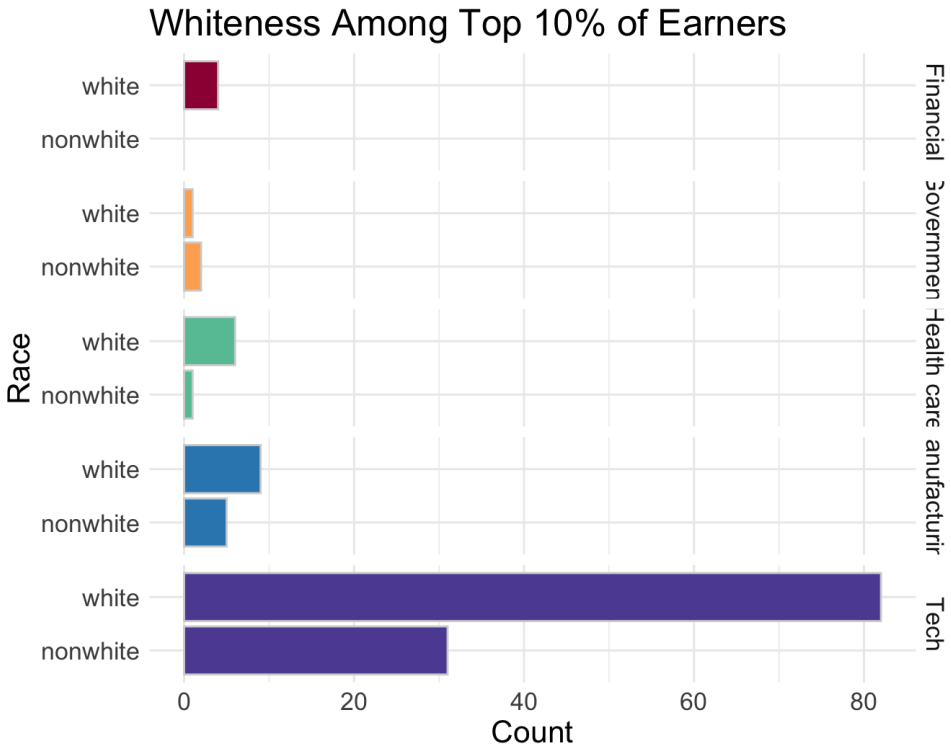
Gender

```
# remove records without answers
top_10p_industry_plot2  <- top_10p_industry[!(top_10p_industry$gender=="NaN" | top_10p_industry$gender=="Other or
prefer not to answer"),]

h4 <- ggplot(top_10p_industry_plot2, aes(x=gender, fill=industry)) + geom_histogram(stat="count", color="lightgre
y")
h4 <- h4 + facet_grid(industry ~ .) + coord_flip() + theme_minimal()
h4 <- h4 + scale_fill_manual(values=ppretty2)
h4 <- h4 + theme(legend.position="none") # Remove legend
h4 <- h4 + theme(text = element_text(size=16), axis.title=element_text(size=16))
h4 <- h4 + labs(title = "Gender of Top 10% of Earners", x= "Gender", y= "Count")
h4
```



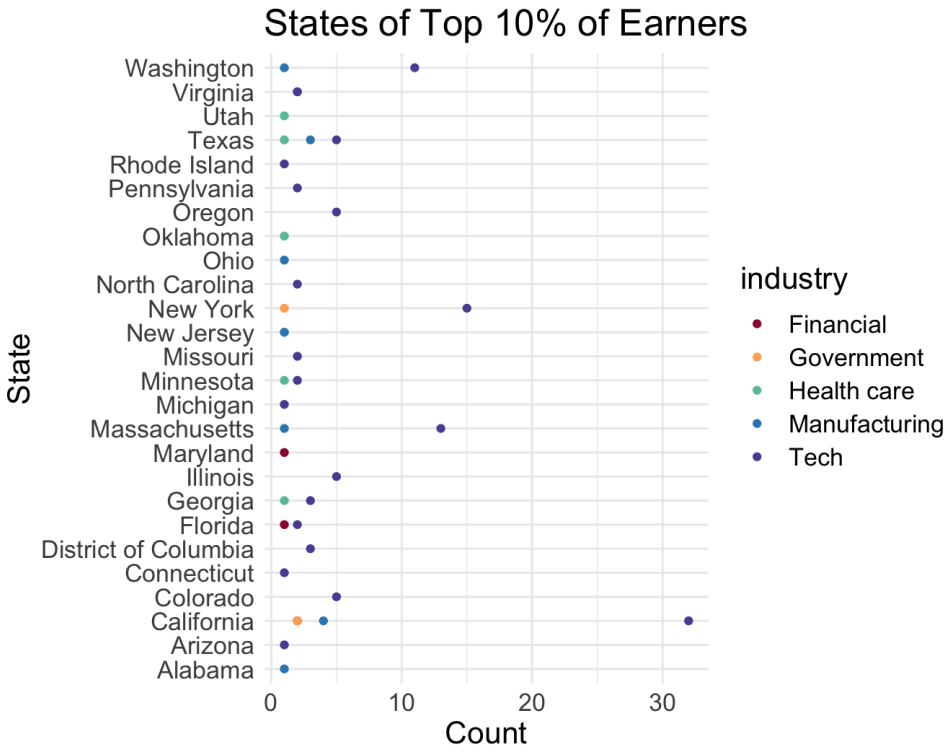
Whiteness

```
h5 <- ggplot(top_10p_industry, aes(x=whiteness, fill=industry)) + geom_histogram(stat="count", color="lightgrey")
h5 <- h5 + facet_grid(industry ~ .) + coord_flip() + theme_minimal()
h5 <- h5 + scale_fill_manual(values=ppretty2)
h5 <- h5 + theme(legend.position="none") # Remove legend
h5 <- h5 + theme(text = element_text(size=16), axis.title=element_text(size=16))
h5 <- h5 + labs(title = "Whiteness Among Top 10% of Earners", x= "Race", y= "Count")
h5
```



States

```
# remove records without answers
top_10p_industry_plot3  <- top_10p_industry[!(top_10p_industry$state=="NaN"),]

s1 <- ggplot(top_10p_industry_plot3, aes(x=state, fill=industry, color=industry)) + geom_point(stat="count") + co
ord_flip()
s1 <- s1 + scale_color_manual(values=ppretty2) + theme_minimal()
s1 <- s1 + theme(text = element_text(size=16), axis.title=element_text(size=16))
s1 <- s1 + labs(title = "States of Top 10% of Earners", x= "State", y= "Count")
s1
```

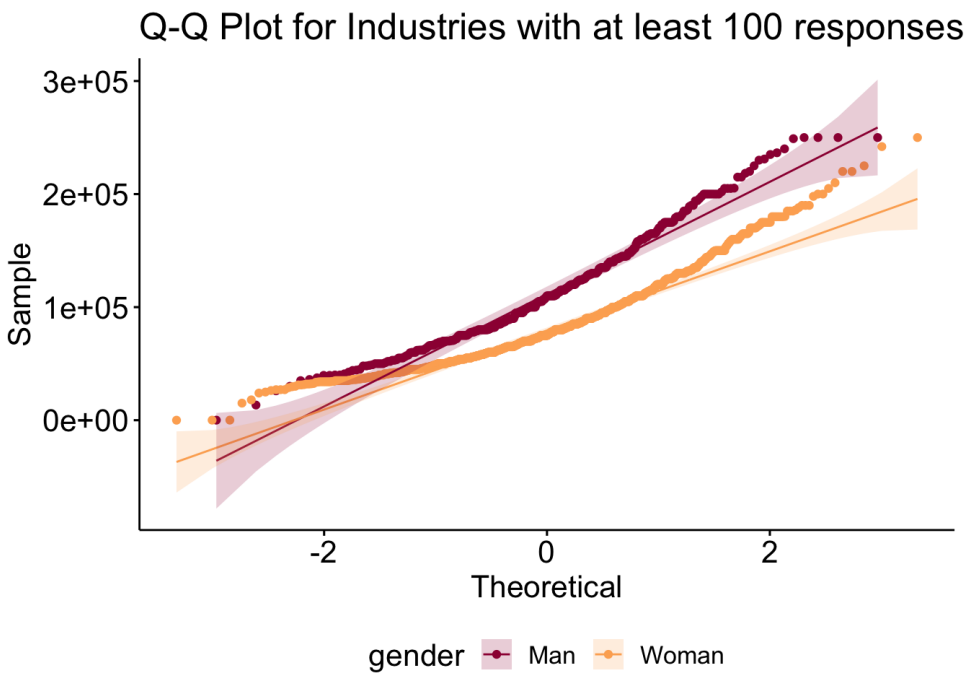


3. Are there statistically significant annual salary differences between genders in each industry? Overall?

```
# Go back to using salaries3 (industries with at least 100 records + outliers removed)
# remove records without answers

salaries3_gender  <- salaries3[!(salaries3$gender=="NaN" | salaries3$gender=="Other or prefer not to answer" | sa
laries3$gender=="Non-binary"),] # drop non-binary for two sample t-test
```

Can we do a between groups t-test? The data needs to be normally distributed with *no* significant difference in variance

```
# check for normal distribution visually

qq1 <- ggqqplot(salaries3_gender, x = "annual_salary",
   color = "gender", palette = c("#9e0142", "#fdae61")) # not normally distributed
qq1 <- qq1 + theme(legend.position="bottom")
qq1 <- qq1 + theme(text = element_text(size=16), axis.title=element_text(size=16))
qq1 <- qq1 + labs(title = "Q-Q Plot for Industries with at least 100 responses")
qq1
```



The plot shows that the data are not normally distributed. Run Shapiro-Wilk test to confirm.

```
shapiro.test(salaries3_gender$annual_salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salaries3_gender$annual_salary
## W = 0.93185, p-value < 2.2e-16
```

From the output, the p-value is < 0.05 implying that the distribution of the data are significantly different from a normal distribution. We therefore cannot assume the normality.

We'll run an F-test to check the variance and further confirm that we can't do a t-test.

```
# use F-test to see if variances are equal between groups
res.ftest <- var.test(annual_salary ~ gender, data = salaries3_gender)
res.ftest # way less than the significance level 0.05, therefore there is a significant difference between the tw
o variances.
```

```
##
##  F test to compare two variances
##
## data:  annual_salary by gender
## F = 1.845, num df = 332, denom df = 1129, p-value = 2.625e-13
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.557904 2.203197
## sample estimates:
## ratio of variances
##           1.844997
```

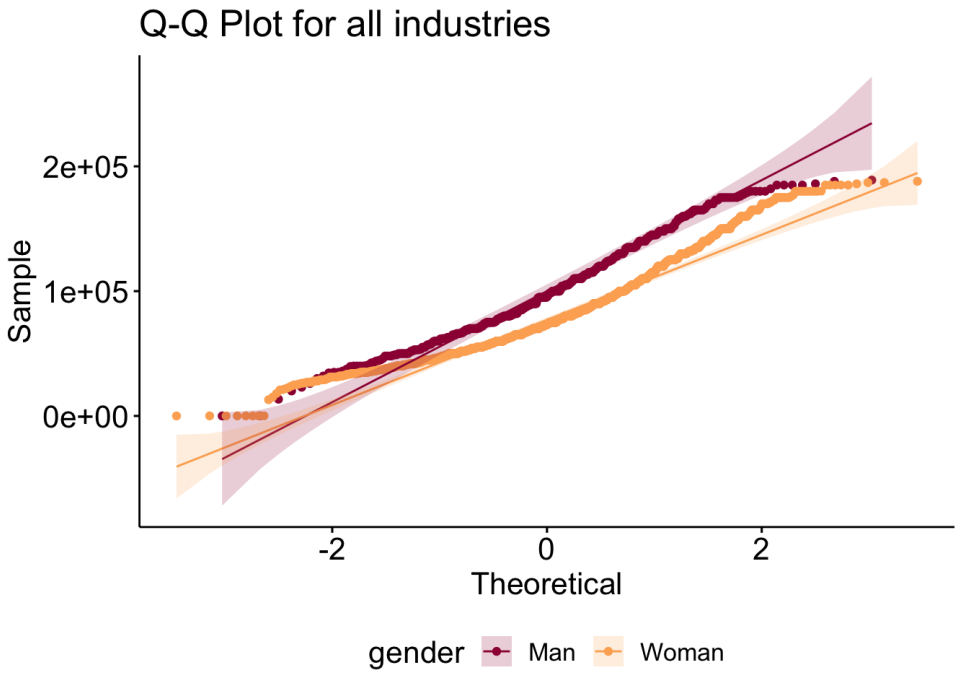
The p-value of F-test is much less than the significance level 0.05, so there is a significant difference between the two variances.

We can't use the t-test for the subgroup of industries with at least 100 responses, but can we do it for the full dataset?

```
# Go back to using salaries (contains all industries and still contains outliers)
salaries_gender  <- salaries[!(salaries$gender=="NaN" | salaries$gender=="Other or prefer not to answer" | salari
es$gender=="Non-binary"),] # drop non-binary for two sample t-test

# Need to remove outliers
# use previously defined outlier function
# create column to indicate suspected outliers
salaries_outliers_removed <- salaries_gender %>%
        mutate(outlier = ifelse(find_outlier(annual_salary), "Yes", "No"))

# create subset to work with where the outliers are removed
salaries_outliers_removed <- salaries_outliers_removed[salaries_outliers_removed$outlier == 'No',]
```

Check for normal distribution visually

```
qq2 <- ggqqplot(salaries_outliers_removed, x = "annual_salary",
   color = "gender", palette = c("#9e0142", "#fdae61")) # not normally distributed
qq2 <- qq2 + theme(legend.position="bottom")
qq2 <- qq2 + theme(text = element_text(size=16), axis.title=element_text(size=16))
qq2 <- qq2 + labs(title = "Q-Q Plot for all industries")
qq2
```



Distriburion is definitely not normal, but run another Shapiro-Wilk to confirm and a F-test to assess variance for the full dataset.

```
shapiro.test(salaries_outliers_removed$annual_salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salaries_outliers_removed$annual_salary
## W = 0.9548, p-value < 2.2e-16
```

The distribution of the data are significantly different from a normal distribution.

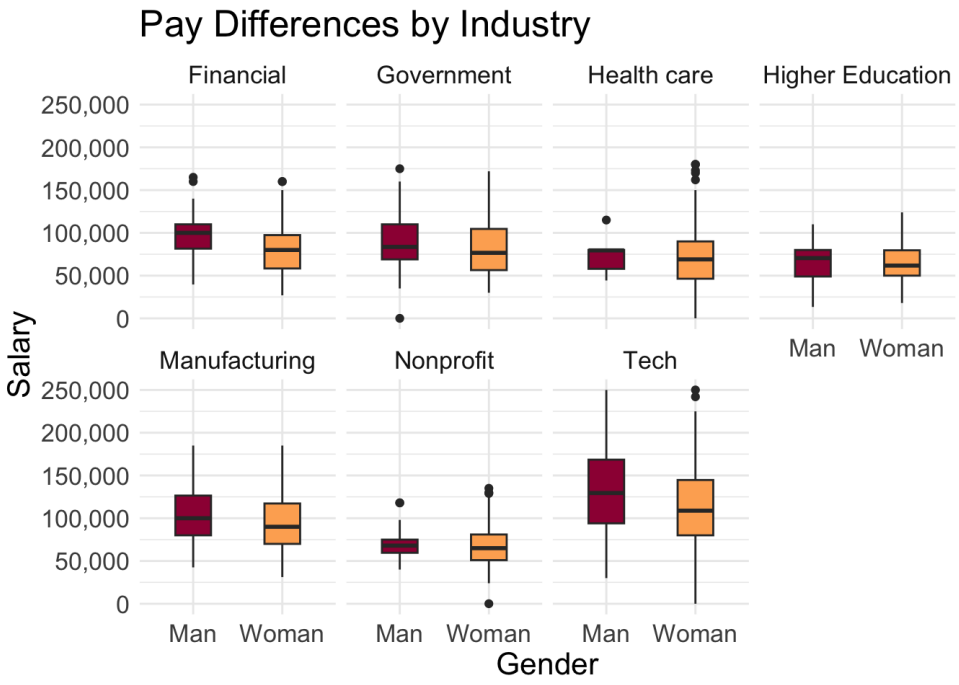
```
# F-test
res.ftest2 <- var.test(annual_salary ~ gender, data = salaries_outliers_removed)
res.ftest2 # variance differs still, so we can't do a statistical examination
```

```
##
##  F test to compare two variances
##
## data:  annual_salary by gender
## F = 1.32, num df = 402, denom df = 1786, p-value = 0.000238
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.136774 1.544294
## sample estimates:
## ratio of variances
##           1.319987
```

F-test confirms that there is a significant difference between the two variances in the full dataset.

We can still take a look to see if there are differences in pay by industry, but we are limited in our ability to attribute any observable differences to gender from a statistical standpoint. If we had been able to do a t-test or more in-depth analysis we would want to further control for years of experience and education. Unfortunately the data significantly limits our ability to do this because we only have year ranges (which are categorical) and it's not clear that those ranges are representative of the thresholds at which you would be able to detect a significant difference using regression techniques. For instance, does pay start to be impacted at 13 years of experience or 17? We only have a range "11-20 years" so we can't test this. The same is true with the age ranges. These limitations would also be relevant in a granular examination of race or education within this dataset.

```
# comparing just men and women within industries with at least 100 respondents
bp3 <- ggplot(salaries3_gender, aes(x=gender, y=annual_salary, fill=gender)) +
    geom_boxplot(width=0.4) + theme_minimal()
bp3 <- bp3 + theme(text = element_text(size=16), axis.title=element_text(size=16))
bp3 <- bp3 + scale_y_continuous(labels = label_comma())
bp3 <- bp3 + facet_wrap(industry ~ ., ncol=4)
bp3 <- bp3 + scale_fill_manual(values=ppretty2)
bp3 <- bp3 + labs(title = "Pay Differences by Industry", x= "Gender", y= "Salary")
bp3 <- bp3 + theme(legend.position="none")
bp3
```



Averages for men are higher in all industries.

```
# table to list averages by industry
mean_salaries <- salaries3_gender %>% group_by(industry,gender) %>%
  summarise(mean_salary = mean(annual_salary),
            .groups = 'drop') %>%
  as.data.frame()
kable(mean_salaries)
```

industry	gender	mean_salary
Financial	Man	99421.07

industry	gender	mean_salary
Financial	Woman	81311.93
Government	Man	87377.27
Government	Woman	82309.44
Health care	Man	74880.83
Health care	Woman	73954.14
Higher Education	Man	67268.00
Higher Education	Woman	65438.85
Manufacturing	Man	105322.86
Manufacturing	Woman	94814.13
Nonprofit	Man	72176.47
Nonprofit	Woman	67674.13
Tech	Man	133547.42
Tech	Woman	113505.27

```
# create a table to describe the pay gap by industry
pay_gap <- mean_salaries %>% group_by(industry) %>% mutate(pay_gap = mean_salary-lag(mean_salary,default=first(me
an_salary)))
pay_gap <- (subset(pay_gap, select = c(industry, pay_gap)))
pay_gap <- pay_gap[!(pay_gap$pay_gap==0),]
kable(pay_gap)
```

industry	pay_gap
Financial	-18109.1438
Government	-5067.8317
Health care	-926.6914
Higher Education	-1829.1467
Manufacturing	-10508.7256
Nonprofit	-4502.3435
Tech	-20042.1452

The largest pay gap between men and women is seen in the Tech industry.

```
### What about salaries across the whole dataset?
stat <- salaries_outliers_removed %>% group_by(gender) %>% summarise(mean_annual_salary = mean(annual_salary))

bp4 <- ggplot(salaries_outliers_removed, aes(x=gender, y=annual_salary, fill=gender)) +
    geom_boxplot(width=0.15) +
    stat_summary(fun.y=mean, geom="point", color="white", fill="white") +
    geom_text(data = stat,
            aes(label = round(mean_annual_salary, 2), x = gender, y = mean_annual_salary - 0.1),
            vjust = 0, nudge_x = 0.22) +
    theme_minimal()
bp4 <- bp4 + theme(text = element_text(size=16), axis.title=element_text(size=16))
bp4 <- bp4 + scale_y_continuous(labels = label_comma())
bp4 <- bp4 + scale_fill_manual(values=ppretty2)
bp4 <- bp4 + labs(title = "US Salaries by Gender", x= "Gender", y= "Salary")
bp4 <- bp4 + theme(legend.position="none")
bp4
```

What's the difference in average pay between men and women across all industries?

```
pay_gap_all <- stat %>% mutate(pay_gap_all = mean_annual_salary-lag(mean_annual_salary,default=first(mean_annual_
salary)))
pay_gap_all <- (subset(pay_gap_all, select = pay_gap_all))
pay_gap_all <- pay_gap_all[!(pay_gap_all$pay_gap_all==0),]
kable(pay_gap_all)
```

pay_gap_all
-20693.8

On average women make $20,693.80 less than men (20.58% less).

4. Which industries pay the highest at entry level?

```
# back to salaries2 (not subset by industries with at least 100 respondents)
# remove outliers
salaries2_outliers_removed <- salaries2 %>%
        mutate(outlier = ifelse(find_outlier(annual_salary), "Yes", "No"))
salaries2_outliers_removed <- salaries2_outliers_removed[salaries2_outliers_removed$outlier == 'No',]

# subset to only entry-level salaries
entry_salaries <- salaries2_outliers_removed[!(salaries2_outliers_removed$field_experience != "1 year or less"),]

# top 10% of entry salaries
n <- 10
entry_top_10p <- entry_salaries[entry_salaries$annual_salary > quantile(entry_salaries$annual_salary, prob=1-n/10
0),]

kable(unique(entry_top_10p$industry))
```

x
Tech
Financial
Health care

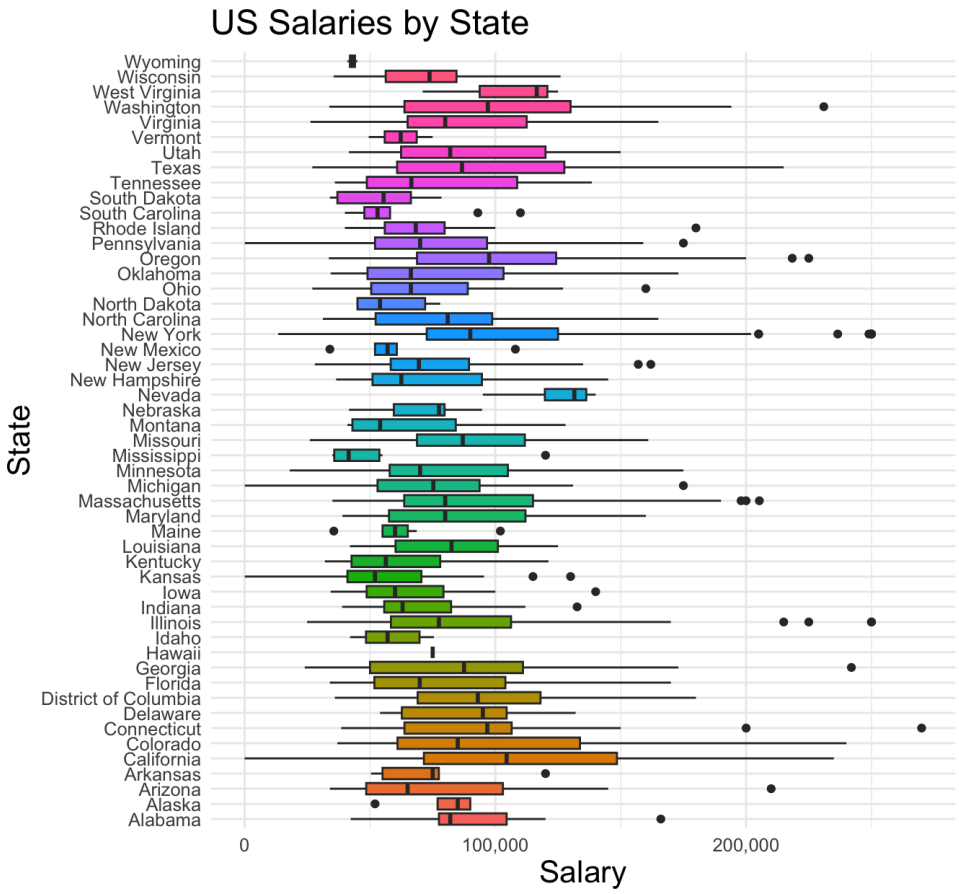
Who are these high earning entry-level employees?

```
entry_top_10p <- (subset(entry_top_10p, select = -c(timestamp, country, currency, multiracial, job, outlier, add_
compensation)))
kable(entry_top_10p)
```

id	age_range	industry	annual_salary	state	all_experience	field_experience	education.level	gender	whiteness
385	25-34	Tech	105000	Washington	2 - 4 years	1 year or less	Master's degree	Woman	nonwhite
1416	25-34	Financial	135000	Colorado	5-7 years	1 year or less	College degree	Woman	white
1532	18-24	Tech	142000	California	1 year or less	1 year or less	College degree	Woman	nonwhite
2038	35-44	Tech	131000	California	21 - 30 years	1 year or less	Some college	Woman	nonwhite
2241	25-34	Health care	125000	Massachusetts	5-7 years	1 year or less	Master's degree	Woman	white
2416	25-34	Tech	110000	New York	2 - 4 years	1 year or less	College degree	Woman	nonwhite
2767	18-24	Tech	116000	Colorado	1 year or less	1 year or less	College degree	Man	white

5. Which states have the highest salaries?

```
salaries2 <- salaries2[!(salaries2$state=="NaN"),]

# remove outliers
salaries2_outliers_removed <- salaries2 %>%
        mutate(outlier = ifelse(find_outlier(annual_salary), "Yes", "No"))
salaries2_outliers_removed <- salaries2_outliers_removed[salaries2_outliers_removed$outlier == 'No',]

# plot the results
bp5 <- ggplot(salaries2_outliers_removed, aes(x=state, y=annual_salary, fill=state)) +
  geom_boxplot() + theme_minimal() + coord_flip()
bp5 <- bp5 + scale_y_continuous(labels = label_comma())
bp5 <- bp5 + theme(legend.position="none") # Remove legend
bp5 <- bp5 + theme(text = element_text(size=12), axis.title=element_text(size=16))
bp5 <- bp5 + labs(title = "US Salaries by State", x= "State", y= "Salary")
bp5 <- bp5 + theme(plot.title = element_text(size=18))
bp5
```

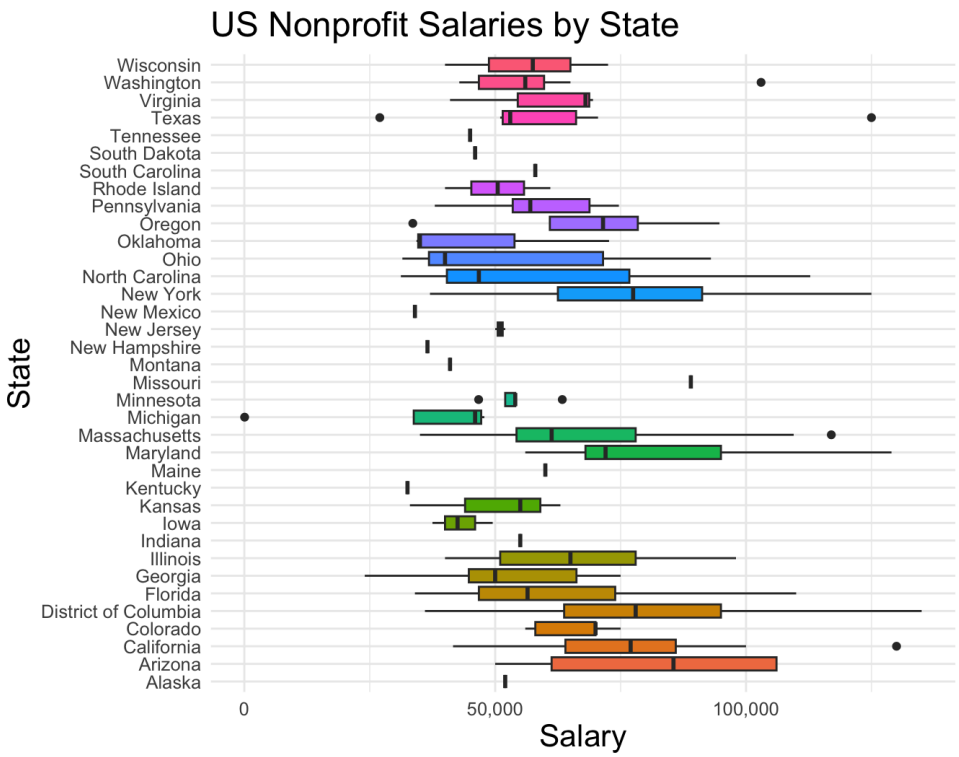


```
high_paying_states <- head(arrange(salaries2_outliers_removed,desc(annual_salary)), n = 50) #top 50 highest payin
g jobs
high_paying_states_list <- as.data.frame(unique(high_paying_states$state))
colnames(high_paying_states_list) <- c("High Paying States")
kable(high_paying_states_list)
```

High Paying States
Connecticut
New York
Illinois
Georgia
Colorado
California
Washington
Oregon
Texas
Arizona
Massachusetts

What about pay within just the nonprofit industry?

```
nonprofit_data <- filter(salaries2, industry == "Nonprofit")

nonprofit_data_outliers_removed <- nonprofit_data %>%
        mutate(outlier = ifelse(find_outlier(annual_salary), "Yes", "No"))

# create subset to work with where the outliers are removed
nonprofit_data_outliers_removed <- nonprofit_data_outliers_removed[nonprofit_data_outliers_removed$outlier == 'N
o',]

# plot the results
bp6 <- ggplot(nonprofit_data_outliers_removed, aes(x=state, y=annual_salary, fill=state)) +
  geom_boxplot() + theme_minimal() + coord_flip()
bp6 <- bp6 + scale_y_continuous(labels = label_comma())
bp6 <- bp6 + theme(legend.position="none") # Remove legend
bp6 <- bp6 + theme(text = element_text(size=12), axis.title=element_text(size=16))
bp6 <- bp6 + labs(title = "US Nonprofit Salaries by State", x= "State", y= "Salary")
bp6 <- bp6 + theme(plot.title = element_text(size=18))
bp6
```




```
high_paying_states_nonprofit <- head(arrange(nonprofit_data_outliers_removed,desc(annual_salary)), n = 50) #top 50 highest paying jobs
high_paying_states_nonprofit <- as.data.frame(unique(high_paying_states_nonprofit$state))
colnames(high_paying_states_nonprofit) <- c("High Paying States (Non-Profits)")
kable(high_paying_states_nonprofit)
```

High Paying States (Non-Profits)
District of Columbia
California
Maryland
New York
Texas
Massachusetts
North Carolina
Florida
Arizona
Washington
Illinois
Oregon
Ohio
Missouri