DS311 Project Q2

Gabrielle Salamanca

March 17, 2023

## We-R-Finished

Before we can dive into the question, we must load in the necessary libraries and read the dataset into the markdown.

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.2.3

## Loading required package: lattice

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Warning: Coercing text to numeric in Y146963 / R146963C25: '45870'

## Warning: Coercing text to numeric in Y164631 / R164631C25: '76700'

Next, let’s rename the columns to make it easier for us to read.

Now, we must remove the sub job titles that have teaching jobs. The question specifically asks for data-related salaries in the job sub-categories, therefore we must remove: professor, attorney, assistant professor, and teacher.

## business analyst data analyst data scientist   
## 27811 3840 1227   
## management consultant software engineer   
## 770 99364

### II. What states (of those I am willing to move to) have the highest paying data-related salaries?

This is quite a broad range of states, but we will remove any US territories first and foremost for obvious reasons. This means Guam, Guamam, Palau, Northern Mariana Islands, Puerto Rico, and Virgin Islands will not be considered.

Now, that still leaves us with a large amount of states. When I tried running graphs to include work state and the job title sub variables, R had difficulty running them, because it was far too many to handle.

Now, the question asks *“What states (of those I am willing to move to)”*. Considering the group is staying in the Bay Area, we may be more picky about where we would want to live. California is much more friendly towards immigrants and quite open to people of all walks of life. We may be biased to stay within the West Coast, especially considering the natural disasters and how certain minorities are treated here either by law or general perception. This would greatly narrow the scope, so I have decided to use the top 10 work states as my pool, minus California. My reasoning is because we are already in California, and the state has a frequency of:

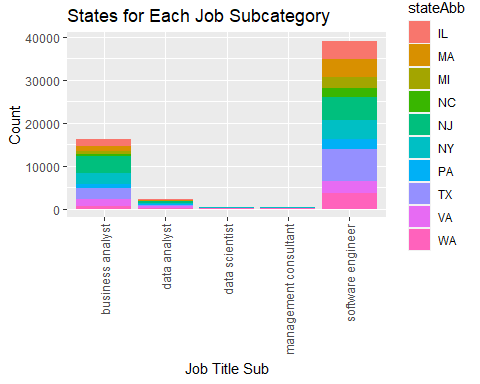
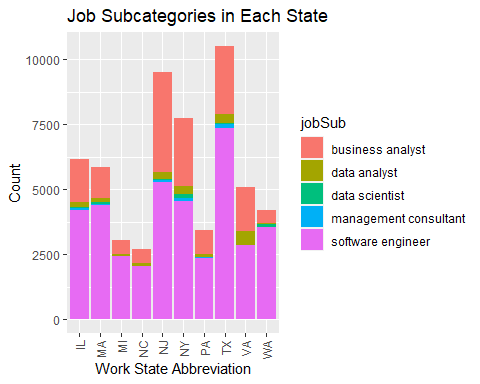
## California   
## 43833

This state completely dwarfs the others. See here:

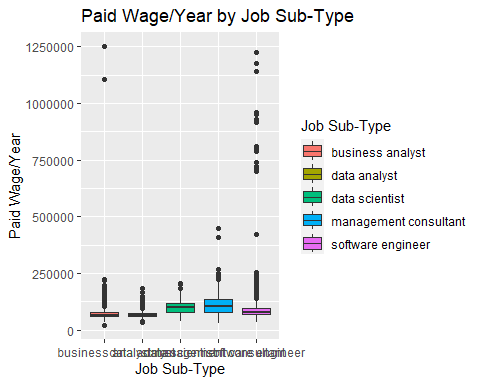
## Frequency  
## Texas 10526  
## New York 7753  
## New Jersey 9526  
## Illinois 6182  
## Massachusetts 5849  
## Virginia 5079  
## Pennsylvania 3447  
## Washington 4193  
## Michigan 3059  
## North Carolina 2702

California would skew the graphs if it were included, so we will only use these states for this question.

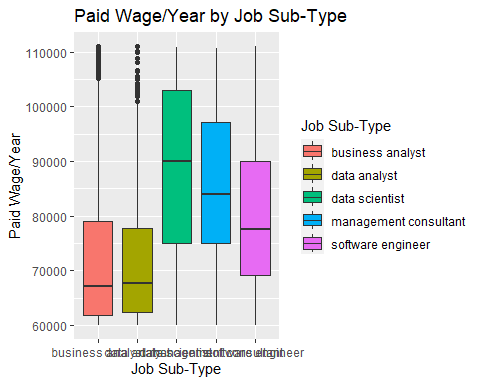
#### a. Differences between job sub-categories?



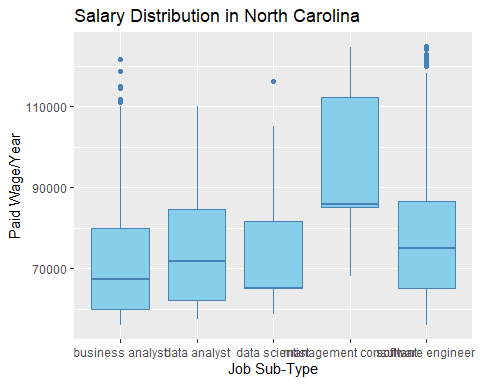
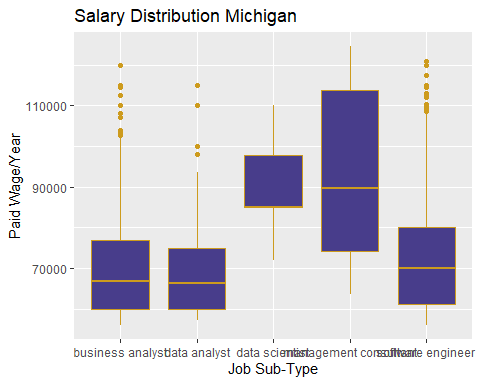
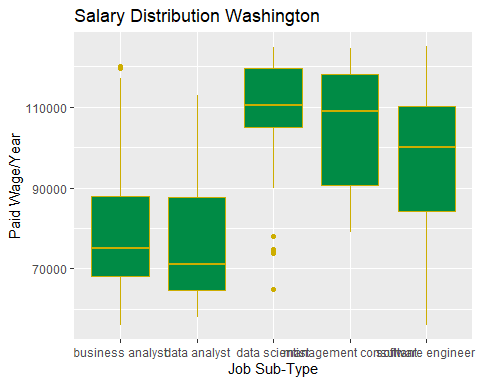
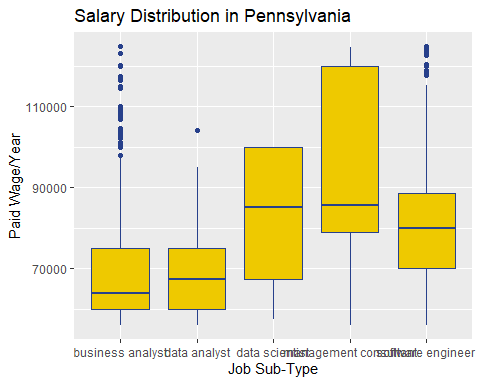
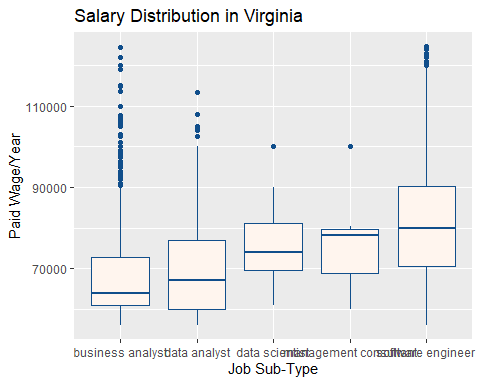
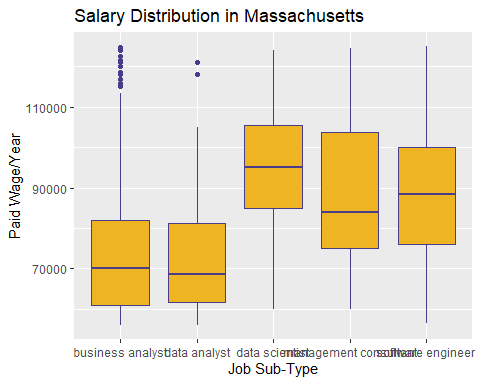
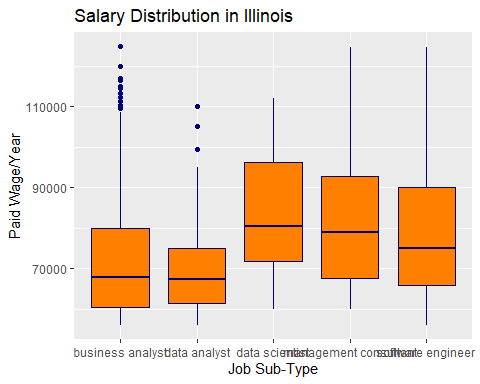
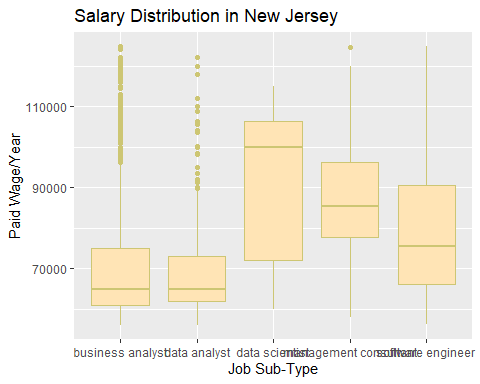
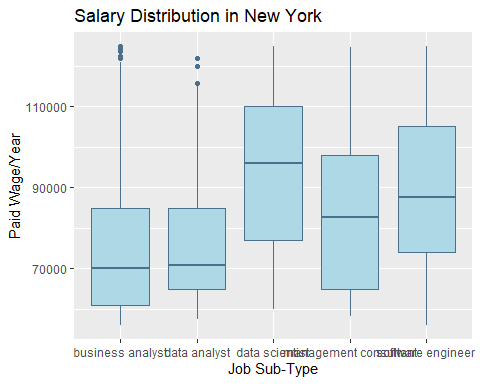
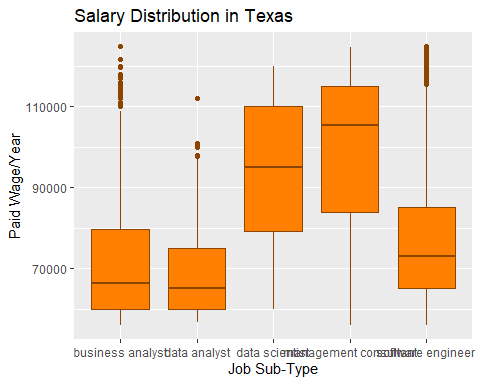
We see that software engineer and business analyst are the largest proportions compared to any other job sub-category in the first graph, especially in the second. Let’s see how each sub-type does in terms of salary.



## Warning: Removed 10534 rows containing non-finite values (`stat\_boxplot()`).



We see that we have many outliers in the first graph, especially in software engineer. But it seems management consultant and data scientist have larger salary ranges compared to the rest of the job sub-types. Now, if we did get rid of the data points outside of 10th and 90th percentile, we get pretty much a similar story. An analyst makes less than a data scientist, management consultant, and software engineer. These three jobs pay well, but it seems software engineer is the only one that’s being advertised to the most out of the them. Let’s see if it holds true for each state. We will only showcase the graphs without the outliers.



We see in these graphs, generally the data scientist, management consultant, and software engineer are the higher paying jobs compared to the analyst jobs. Which one had the higher salary shifted in each state. But the most common patterns were, from highest to lowest: data scientist, management consultant, and software engineer; management consultant, data scientist, and software engineer. The most striking graph in these 10 was North Carolina, where data scientist was the lowest of the top 3. We even see that the analyst jobs even have a much larger quartile box, and that data scientist and management consultant have the lowest medians in the NC graph.

Ultimately, there is a difference in salaries between the job sub-types. The analyst jobs won’t be the highest paying in most states, so the other three will be the main ones I’d want to apply to. If I wanted to apply to the management consultant jobs, I would want to move to North Carolina, Michigan, Pennsylvania, and Texas. As for data scientist, I would want to move to New York, New Jersey, Illinois and Washington. As for software engineer, I would move to Virginia. However, considering I would most likely start in the analyst job first, North Carolina, Washington, and New York would be my choices.

#### b. Which companies have the highest salaries for those sub-types?

Get top 5 for each sub-type

## `summarise()` has grouped output by 'Sub'. You can override using the `.groups`  
## argument.

## # A tibble: 26 × 3  
## # Groups: Sub [5]  
## Sub `Employer Name` avg\_salary  
## <fct> <chr> <dbl>  
## 1 business analyst ELIE TAHARI LTD. 220000   
## 2 business analyst OFFICEMAX INCORPORATED 603712.  
## 3 business analyst PRECISION GLOBAL CONSULTING INC 179000   
## 4 business analyst SOROS FUND MANAGEMENT, LLC 175000   
## 5 business analyst STEVEN DOUGLAS ASSOCIATES, INC. 175000   
## 6 business analyst THE UNIVERSITY OF TEXAS SYSTEM ADMINISTRATION 677508   
## 7 data analyst AMICUS SYSTEMS INC 140575   
## 8 data analyst KNIGHT CAPITAL AMERICAS, L.P. 150000   
## 9 data analyst KNOWLEDGENT GROUP INC. 185000   
## 10 data analyst OPEN SYSTEMS TECHNOLOGIES, INC. 150000   
## # … with 16 more rows

## `summarise()` has grouped output by 'Sub'. You can override using the `.groups`  
## argument.

## # A tibble: 25 × 3  
## # Groups: Sub [5]  
## Sub `Employer Name` max\_salary  
## <fct> <chr> <dbl>  
## 1 business analyst ELIE TAHARI LTD. 220000   
## 2 business analyst HIREPOWER PERSONNEL, INC. 224640   
## 3 business analyst OFFICEMAX INCORPORATED 1103712.  
## 4 business analyst POPULUS GROUP 218400   
## 5 business analyst THE UNIVERSITY OF TEXAS SYSTEM ADMINISTRATION 1250784   
## 6 data analyst AMICUS SYSTEMS INC 166400   
## 7 data analyst KNIGHT CAPITAL AMERICAS, L.P. 150000   
## 8 data analyst KNOWLEDGENT GROUP INC. 185000   
## 9 data analyst OPEN SYSTEMS TECHNOLOGIES, INC. 150000   
## 10 data analyst WEATHERFORD INTERNATIONAL, INC. 148000   
## # … with 15 more rows

#### c. Will the answer change if I take standard of living into account?

knitr::opts\_chunk$set(echo = FALSE)  
# loading libraries  
library(readxl)  
library(caret)  
library(ggplot2)  
library(dplyr)  
  
# dataset overview  
salary <- read\_excel("C:/Users/knigh/OneDrive/Desktop/Github/We-R-Finished/salary/salary\_data\_states.xlsx")  
# renaming  
colnames(salary)[1] = "Case Number"  
colnames(salary)[2] = "Case Status"  
colnames(salary)[3] = "Received Date"  
colnames(salary)[4] = "Decision Date"  
colnames(salary)[5] = "Employer Name"  
colnames(salary)[6] = "Submitted Prevail Wage"  
colnames(salary)[7] = "SPrW Unit"  
colnames(salary)[8] = "Submitted Paid Wage"  
colnames(salary)[9] = "SPaW Unit"  
colnames(salary)[10] = "Job Title"  
colnames(salary)[11] = "Work City"  
colnames(salary)[12] = "Required Edu"  
colnames(salary)[13] = "Required College Major"  
colnames(salary)[14] = "Exp Req"  
colnames(salary)[15] = "Exp Req (Months)"  
colnames(salary)[16] = "Citizenship"  
colnames(salary)[17] = "Prevail Wage SOC Code"  
colnames(salary)[18] = "PWSOC Title"  
colnames(salary)[19] = "Work State"  
colnames(salary)[20] = "WS Abb"  
colnames(salary)[21] = "WPostal Code"  
colnames(salary)[22] = "Full Time"  
colnames(salary)[23] = "Visa Class"  
colnames(salary)[24] = "Prevail Wage/Yr"  
colnames(salary)[25] = "Paid Wage/Yr"  
colnames(salary)[26] = "Job Title Sub"  
colnames(salary)[27] = "Order"  
# filter  
sal <- salary %>%  
 filter(!grepl("professor", `Job Title Sub`, ignore.case = TRUE) &   
 !grepl("attorney", `Job Title Sub`, ignore.case = TRUE) &  
 !grepl("assistant professor", `Job Title Sub`, ignore.case = TRUE) &   
 !grepl("teacher", `Job Title Sub`, ignore.case = TRUE))  
jobTitleSub <- as.factor(sal$`Job Title Sub`)  
summary(jobTitleSub)  
# filtering states  
datsal <- sal %>%  
 filter(!grepl("Guam", `Work State`, ignore.case = TRUE) &   
 !grepl("Guamam", `Work State`, ignore.case = TRUE) &  
 !grepl("Palau", `Work State`, ignore.case = TRUE) &   
 !grepl("Northern Mariana Islands", `Work State`, ignore.case = TRUE) &  
 !grepl("Puerto Rico", `Work State`, ignore.case = TRUE) &  
 !grepl("Virgin Islands", `Work State`, ignore.case = TRUE))  
CAfreq <- table(datsal$`Work State`)["California"]  
CAfreq  
TXfreq <- table(datsal$`Work State`)["Texas"]  
NYfreq <- table(datsal$`Work State`)["New York"]  
NJfreq <- table(datsal$`Work State`)["New Jersey"]  
ILfreq <- table(datsal$`Work State`)["Illinois"]  
MAfreq <- table(datsal$`Work State`)["Massachusetts"]  
VAfreq <- table(datsal$`Work State`)["Virginia"]  
PAfreq <- table(datsal$`Work State`)["Pennsylvania"]  
WAfreq <- table(datsal$`Work State`)["Washington"]  
MIfreq <- table(datsal$`Work State`)["Michigan"]  
NCfreq <- table(datsal$`Work State`)["North Carolina"]  
# dataframe  
Frequency <- c(TXfreq, NYfreq, NJfreq, ILfreq, MAfreq, VAfreq, PAfreq, WAfreq, MIfreq, NCfreq)  
stateFreq <- data.frame(Frequency)  
stateFreq  
topState <- c("Texas", "New York",  
 "New Jersey", "Illinois", "Massachusetts",  
 "Virginia", "Pennsylvania", "Washington",  
 "Michigan", "North Carolina")  
ssal <- datsal[datsal$`Work State` %in% topState, ]  
# variables  
stateAbb <- ssal$`WS Abb`  
jobSub <- ssal$`Job Title Sub`  
  
# job sub/state plot  
ggplot(ssal, aes(x = stateAbb, fill = jobSub)) +   
 geom\_bar() +   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +  
 labs(x = "Work State Abbreviation", y = "Count", title = "Job Subcategories in Each State")  
  
# state/job sub  
ggplot(ssal, aes(x = jobSub, fill = stateAbb)) +   
 geom\_bar() +   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +  
 labs(x = "Job Title Sub", y = "Count", title = "States for Each Job Subcategory")  
# variable(s)  
paidWage <- ssal$`Paid Wage/Yr`  
  
# paid wage boxplot  
ggplot(ssal, aes(x = jobSub, y = paidWage, fill = jobSub)) +  
 geom\_boxplot() +  
 labs(title = "Paid Wage/Year by Job Sub-Type",  
 x = "Job Sub-Type",  
 y = "Paid Wage/Year",  
 fill = "Job Sub-Type")  
  
# no outliers  
ggplot(ssal, aes(x = jobSub, y = paidWage, fill = jobSub)) +  
 geom\_boxplot() +  
 labs(title = "Paid Wage/Year by Job Sub-Type",  
 x = "Job Sub-Type",  
 y = "Paid Wage/Year",  
 fill = "Job Sub-Type") +  
 scale\_y\_continuous(limits = quantile(paidWage, c(0.1, 0.9)))  
# variable  
ssal$Sub <- factor(ssal$`Job Title Sub`, levels = c("business analyst", "data analyst", "data scientist", "management consultant", "software engineer"))  
  
# Texas  
ssal %>%  
 filter(`Work State` == "Texas", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "darkorange1", color = "darkorange4") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in Texas")  
  
# New York  
ssal %>%  
 filter(`Work State` == "New York", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "lightblue", color = "skyblue4") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in New York")  
  
# New Jersey  
ssal %>%  
 filter(`Work State` == "New Jersey", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "moccasin", color = "khaki3") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in New Jersey")  
  
# Illinois  
ssal %>%  
 filter(`Work State` == "Illinois", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "darkorange1", color = "navyblue") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in Illinois")  
  
# Massachussetts  
ssal %>%  
 filter(`Work State` == "Massachusetts", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "goldenrod2", color = "darkslateblue") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in Massachusetts")  
  
# Virginia  
ssal %>%  
 filter(`Work State` == "Virginia", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "seashell", color = "dodgerblue4") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in Virginia")  
  
# Pennsylvania  
ssal %>%  
 filter(`Work State` == "Pennsylvania", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "gold2", color = "royalblue4") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in Pennsylvania")  
  
# Washington  
ssal %>%  
 filter(`Work State` == "Washington", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "springgreen4", color = "gold3") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution Washington")  
  
# Michigan  
ssal %>%  
 filter(`Work State` == "Michigan", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "darkslateblue", color = "goldenrod3") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution Michigan")  
  
# North Carolina  
ssal %>%  
 filter(`Work State` == "North Carolina", `Paid Wage/Yr` > quantile(`Paid Wage/Yr`, 0.05) & `Paid Wage/Yr` < quantile(`Paid Wage/Yr`, 0.95)) %>%  
 ggplot(aes(x = Sub, y = `Paid Wage/Yr`)) +  
 geom\_boxplot(fill = "skyblue", color = "steelblue") +  
 labs(x = "Job Sub-Type", y = "Paid Wage/Year", title = "Salary Distribution in North Carolina")  
ssal %>%  
 group\_by(Sub, `Employer Name`) %>%  
 summarise(avg\_salary = mean(`Paid Wage/Yr`)) %>%  
 top\_n(n = 5, wt = avg\_salary)  
#  
ssal %>%  
 group\_by(Sub, `Employer Name`) %>%  
 summarise(max\_salary = max(`Paid Wage/Yr`)) %>%  
 top\_n(n = 5, wt = max\_salary)