Salary Data Set Analysis

Ana Paula Felix de Queiroz

## **Da**ta Cleaning

In this code chunk, I am performing data cleaning and data filtering tasks.

Firstly, I am using the **names** function to get the column names of the dataframe and the **head** function to get the first few rows of the dataframe. I am also using the **attach** function to attach the dataframe to the search path so that the column names can be used directly without specifying the dataframe name.

Next, I am checking for missing values in the dataframe by using the **is.na** function to create a logical matrix of missing values and then using the **colSums** function to count the number of missing values in each column.

After that, I am filtering the dataframe to keep only the data-related jobs. I define the job titles I want to keep as a vector and then filter the data using the **%in%** operator. I then check the number of rows in the cleaned data and the original dataframe using the **nrow** function.

Lastly, I am commenting out the code that removes rows with missing values and saves the cleaned data, as it is not being used in this analysis.

## CASE\_NUMBER CASE\_STATUS CASE\_RECEIVED\_DATE DECISION\_DATE   
## Length:167278 Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
## EMPLOYER\_NAME PREVAILING\_WAGE\_SUBMITTED PREVAILING\_WAGE\_SUBMITTED\_UNIT  
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## PAID\_WAGE\_SUBMITTED PAID\_WAGE\_SUBMITTED\_UNIT JOB\_TITLE   
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## WORK\_CITY EDUCATION\_LEVEL\_REQUIRED COLLEGE\_MAJOR\_REQUIRED  
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## EXPERIENCE\_REQUIRED\_Y\_N EXPERIENCE\_REQUIRED\_NUM\_MONTHS COUNTRY\_OF\_CITIZENSHIP  
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## PREVAILING\_WAGE\_SOC\_CODE PREVAILING\_WAGE\_SOC\_TITLE WORK\_STATE   
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## WORK\_STATE\_ABBREVIATION WORK\_POSTAL\_CODE FULL\_TIME\_POSITION\_Y\_N  
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## VISA\_CLASS PREVAILING\_WAGE\_PER\_YEAR PAID\_WAGE\_PER\_YEAR  
## Length:167278 Length:167278 Length:167278   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## JOB\_TITLE\_SUBGROUP order   
## Length:167278 Length:167278   
## Class :character Class :character   
## Mode :character Mode :character

## [1] "CASE\_NUMBER" "CASE\_STATUS"   
## [3] "CASE\_RECEIVED\_DATE" "DECISION\_DATE"   
## [5] "EMPLOYER\_NAME" "PREVAILING\_WAGE\_SUBMITTED"   
## [7] "PREVAILING\_WAGE\_SUBMITTED\_UNIT" "PAID\_WAGE\_SUBMITTED"   
## [9] "PAID\_WAGE\_SUBMITTED\_UNIT" "JOB\_TITLE"   
## [11] "WORK\_CITY" "EDUCATION\_LEVEL\_REQUIRED"   
## [13] "COLLEGE\_MAJOR\_REQUIRED" "EXPERIENCE\_REQUIRED\_Y\_N"   
## [15] "EXPERIENCE\_REQUIRED\_NUM\_MONTHS" "COUNTRY\_OF\_CITIZENSHIP"   
## [17] "PREVAILING\_WAGE\_SOC\_CODE" "PREVAILING\_WAGE\_SOC\_TITLE"   
## [19] "WORK\_STATE" "WORK\_STATE\_ABBREVIATION"   
## [21] "WORK\_POSTAL\_CODE" "FULL\_TIME\_POSITION\_Y\_N"   
## [23] "VISA\_CLASS" "PREVAILING\_WAGE\_PER\_YEAR"   
## [25] "PAID\_WAGE\_PER\_YEAR" "JOB\_TITLE\_SUBGROUP"   
## [27] "order"

## # A tibble: 6 × 27  
## CASE\_NUMBER CASE\_…¹ CASE\_…² DECIS…³ EMPLO…⁴ PREVA…⁵ PREVA…⁶ PAID\_…⁷ PAID\_…⁸  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 I-200-14073-2… denied 3/14/2… 3/21/2… ADVANC… 6217100 year 62171 year   
## 2 A-15061-55212 denied 3/19/2… 3/19/2… SAN FR… 5067600 year 91440 year   
## 3 I-200-13256-0… denied 9/13/2… 9/23/2… CAROUS… 4947000 year 49470 year   
## 4 I-200-14087-3… denied 3/28/2… 4/7/20… HARLIN… 251052… month 43800 year   
## 5 I-203-14259-1… denied 9/16/2… 9/23/2… SIGNAL… 84573.… bi-wee… 170000 year   
## 6 I-200-14092-4… denied 4/2/20… 4/9/20… CAPGEM… 113610 month 114421 year   
## # … with 18 more variables: JOB\_TITLE <chr>, WORK\_CITY <chr>,  
## # EDUCATION\_LEVEL\_REQUIRED <chr>, COLLEGE\_MAJOR\_REQUIRED <chr>,  
## # EXPERIENCE\_REQUIRED\_Y\_N <chr>, EXPERIENCE\_REQUIRED\_NUM\_MONTHS <chr>,  
## # COUNTRY\_OF\_CITIZENSHIP <chr>, PREVAILING\_WAGE\_SOC\_CODE <chr>,  
## # PREVAILING\_WAGE\_SOC\_TITLE <chr>, WORK\_STATE <chr>,  
## # WORK\_STATE\_ABBREVIATION <chr>, WORK\_POSTAL\_CODE <chr>,  
## # FULL\_TIME\_POSITION\_Y\_N <chr>, VISA\_CLASS <chr>, …

## CASE\_NUMBER CASE\_STATUS   
## 0 0   
## CASE\_RECEIVED\_DATE DECISION\_DATE   
## 0 0   
## EMPLOYER\_NAME PREVAILING\_WAGE\_SUBMITTED   
## 0 0   
## PREVAILING\_WAGE\_SUBMITTED\_UNIT PAID\_WAGE\_SUBMITTED   
## 0 0   
## PAID\_WAGE\_SUBMITTED\_UNIT JOB\_TITLE   
## 0 0   
## WORK\_CITY EDUCATION\_LEVEL\_REQUIRED   
## 0 0   
## COLLEGE\_MAJOR\_REQUIRED EXPERIENCE\_REQUIRED\_Y\_N   
## 42 0   
## EXPERIENCE\_REQUIRED\_NUM\_MONTHS COUNTRY\_OF\_CITIZENSHIP   
## 6128 0   
## PREVAILING\_WAGE\_SOC\_CODE PREVAILING\_WAGE\_SOC\_TITLE   
## 0 0   
## WORK\_STATE WORK\_STATE\_ABBREVIATION   
## 0 0   
## WORK\_POSTAL\_CODE FULL\_TIME\_POSITION\_Y\_N   
## 0 0   
## VISA\_CLASS PREVAILING\_WAGE\_PER\_YEAR   
## 0 0   
## PAID\_WAGE\_PER\_YEAR JOB\_TITLE\_SUBGROUP   
## 0 0   
## order   
## 0

## [1] "software engineer" "assistant professor" "teacher"   
## [4] "business analyst" "management consultant" "data analyst"   
## [7] "attorney" "data scientist"

## [1] 133012

## [1] 167278

## The following objects are masked from df:  
##   
## CASE\_NUMBER, CASE\_RECEIVED\_DATE, CASE\_STATUS,  
## COLLEGE\_MAJOR\_REQUIRED, COUNTRY\_OF\_CITIZENSHIP, DECISION\_DATE,  
## EDUCATION\_LEVEL\_REQUIRED, EMPLOYER\_NAME,  
## EXPERIENCE\_REQUIRED\_NUM\_MONTHS, EXPERIENCE\_REQUIRED\_Y\_N,  
## FULL\_TIME\_POSITION\_Y\_N, JOB\_TITLE, JOB\_TITLE\_SUBGROUP, order,  
## PAID\_WAGE\_PER\_YEAR, PAID\_WAGE\_SUBMITTED, PAID\_WAGE\_SUBMITTED\_UNIT,  
## PREVAILING\_WAGE\_PER\_YEAR, PREVAILING\_WAGE\_SOC\_CODE,  
## PREVAILING\_WAGE\_SOC\_TITLE, PREVAILING\_WAGE\_SUBMITTED,  
## PREVAILING\_WAGE\_SUBMITTED\_UNIT, VISA\_CLASS, WORK\_CITY,  
## WORK\_POSTAL\_CODE, WORK\_STATE, WORK\_STATE\_ABBREVIATION

## Descriptive statistics

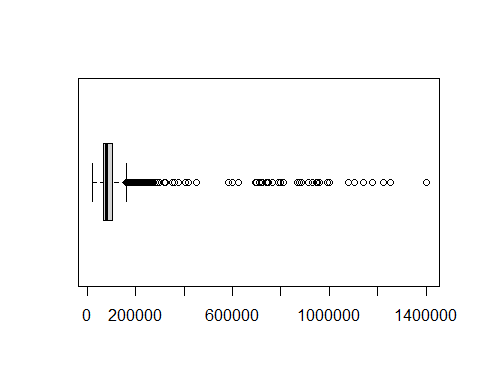
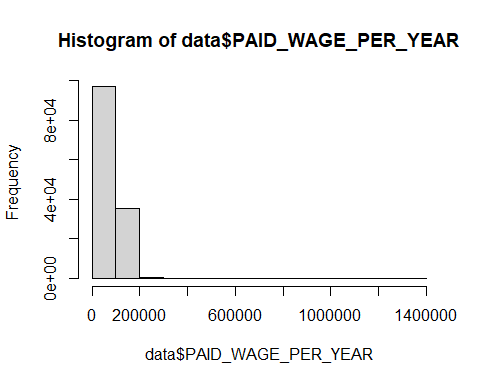
In this section, I will be exploring the distribution of salaries in the dataset. I will be calculating summary statistics for the salary distribution, including mean and median, and creating visualizations such as histograms and boxplots to better understand the distribution of salaries.

## The following objects are masked from data (pos = 3):  
##   
## CASE\_NUMBER, CASE\_RECEIVED\_DATE, CASE\_STATUS,  
## COLLEGE\_MAJOR\_REQUIRED, COUNTRY\_OF\_CITIZENSHIP, DECISION\_DATE,  
## EDUCATION\_LEVEL\_REQUIRED, EMPLOYER\_NAME,  
## EXPERIENCE\_REQUIRED\_NUM\_MONTHS, EXPERIENCE\_REQUIRED\_Y\_N,  
## FULL\_TIME\_POSITION\_Y\_N, JOB\_TITLE, JOB\_TITLE\_SUBGROUP, order,  
## PAID\_WAGE\_PER\_YEAR, PAID\_WAGE\_SUBMITTED, PAID\_WAGE\_SUBMITTED\_UNIT,  
## PREVAILING\_WAGE\_PER\_YEAR, PREVAILING\_WAGE\_SOC\_CODE,  
## PREVAILING\_WAGE\_SOC\_TITLE, PREVAILING\_WAGE\_SUBMITTED,  
## PREVAILING\_WAGE\_SUBMITTED\_UNIT, VISA\_CLASS, WORK\_CITY,  
## WORK\_POSTAL\_CODE, WORK\_STATE, WORK\_STATE\_ABBREVIATION

## The following objects are masked from df:  
##   
## CASE\_NUMBER, CASE\_RECEIVED\_DATE, CASE\_STATUS,  
## COLLEGE\_MAJOR\_REQUIRED, COUNTRY\_OF\_CITIZENSHIP, DECISION\_DATE,  
## EDUCATION\_LEVEL\_REQUIRED, EMPLOYER\_NAME,  
## EXPERIENCE\_REQUIRED\_NUM\_MONTHS, EXPERIENCE\_REQUIRED\_Y\_N,  
## FULL\_TIME\_POSITION\_Y\_N, JOB\_TITLE, JOB\_TITLE\_SUBGROUP, order,  
## PAID\_WAGE\_PER\_YEAR, PAID\_WAGE\_SUBMITTED, PAID\_WAGE\_SUBMITTED\_UNIT,  
## PREVAILING\_WAGE\_PER\_YEAR, PREVAILING\_WAGE\_SOC\_CODE,  
## PREVAILING\_WAGE\_SOC\_TITLE, PREVAILING\_WAGE\_SUBMITTED,  
## PREVAILING\_WAGE\_SUBMITTED\_UNIT, VISA\_CLASS, WORK\_CITY,  
## WORK\_POSTAL\_CODE, WORK\_STATE, WORK\_STATE\_ABBREVIATION

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20000 66000 81500 87657 104000 1400000

## [1] 81500



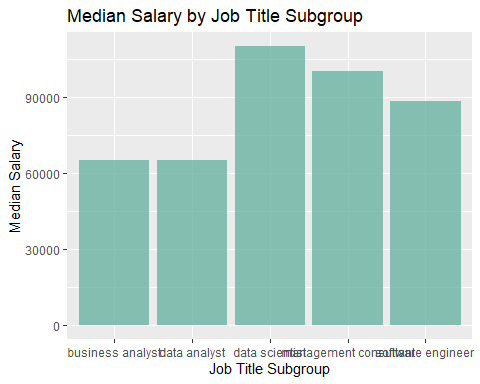
## Descriptive Stats and Visualization by Job Title Subgroup

Here, I’m looking at how much people in different job groups get paid. I’m using the median and mean salaries for each job group to get a sense of what’s typical. This helps me compare salaries across different job titles and see if some job groups pay more than others.

The bar graph helps visualizing the median salaries for each job group. This way we can see which job groups pay the most and which ones pay the least.

## JOB\_TITLE\_SUBGROUP PAID\_WAGE\_PER\_YEAR  
## 1 business analyst 65000.0  
## 2 data analyst 65000.0  
## 3 data scientist 110000.0  
## 4 management consultant 100000.0  
## 5 software engineer 88275.2

## JOB\_TITLE\_SUBGROUP PAID\_WAGE\_PER\_YEAR  
## 1 business analyst 71300.08  
## 2 data analyst 70030.08  
## 3 data scientist 108021.04  
## 4 management consultant 108251.33  
## 5 software engineer 92505.30



## **Outlier Analysis**

In this code, we are examining the distribution of salaries across different job title subgroups to identify any potential outliers. Outliers are values that are significantly different from the majority of the data and can have a disproportionate impact on summary statistics and model accuracy.

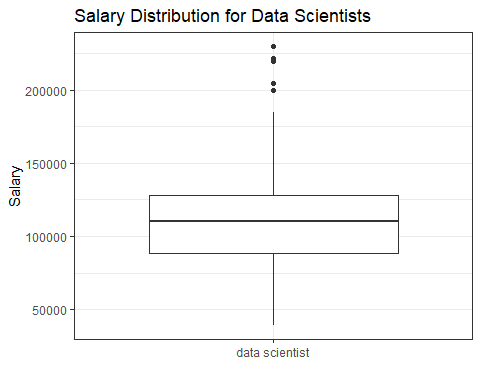
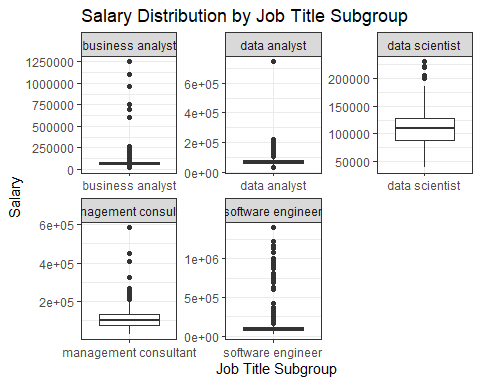
To accomplish this, we first group the data by job title subgroup and calculate the median salary for each subgroup. We then create box plots for each subgroup, which show the distribution of salaries and any potential outliers. The box plots are separated by subgroup to allow for easier comparison between subgroups.

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

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

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



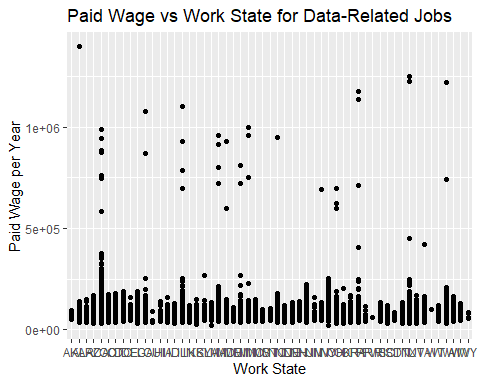
**Wage payed across different US states:**

This code is used to analyze and visualize the paid wage for data-related jobs across different states in the US. I created a scatter plot of paid wage versus work state, with the x-axis representing the states’ abbreviations and the y-axis representing the median paid wage per year.

I also Identified the top and bottom five states with the highest and lowest median paid wages per year. These states are grouped and arranged by their median wage values in descending and ascending order, respectively.

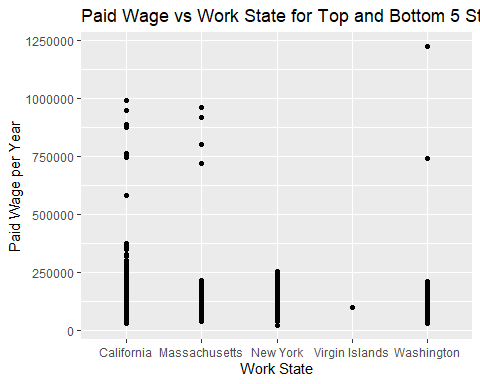
In the third block of code, a scatter plot is created to show the paid wage versus work state, but only for the top and bottom five states identified in the second block.

The fourth block of code calculates and prints the top ten states with the highest median paid wages.



## # A tibble: 5 × 2  
## WORK\_STATE median\_wage  
## <chr> <dbl>  
## 1 California 105000  
## 2 Washington 102000  
## 3 Virgin Islands 99788  
## 4 Massachusetts 85000  
## 5 New York 85000

## # A tibble: 5 × 2  
## WORK\_STATE median\_wage  
## <chr> <dbl>  
## 1 Montana 60000  
## 2 Wyoming 58205  
## 3 West Virginia 55000  
## 4 Puerto Rico 53000  
## 5 Guam 44699



## WORK\_STATE PAID\_WAGE\_PER\_YEAR  
## 5 California 105000  
## 52 Washington 102000  
## 50 Virgin Islands 99788  
## 23 Massachusetts 85000  
## 34 New York 85000  
## 39 Oregon 82846  
## 49 Vermont 80250  
## 48 Utah 79726  
## 30 Nevada 79563  
## 22 Maryland 77000

## **State, Job Title Subgroup and Salary**

Here, i am exploring the relationship between median salary, job title subgroup, and state.

I am using the data on wage payed, state, and job title to gain insights into median salaries across different states and job title subgroups.

I first group the data by work state and job title subgroup and calculate the median salary for each combination of these two variables. This allows me to see how median salaries vary for different job titles in different states.

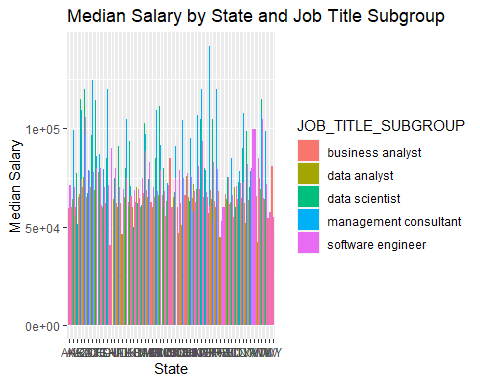
Then, I create a grouped bar plot to visualize the median salaries for each job title subgroup in each state. This plot provides an easily interpretable visual representation of the differences in median salaries across job title subgroups and states.

This analysis is important because it allows me to identify patterns in median salaries across different states and job title subgroups, which can provide insights into the factors that influence salaries in different fields and locations. This information can be valuable for making decisions related to career planning, job searching, and negotiating salaries.

## `summarise()` has grouped output by 'WORK\_STATE\_ABBREVIATION'. You can override  
## using the `.groups` argument.

## # A tibble: 225 × 3  
## # Groups: WORK\_STATE\_ABBREVIATION [55]  
## WORK\_STATE\_ABBREVIATION JOB\_TITLE\_SUBGROUP median\_salary  
## <chr> <chr> <dbl>  
## 1 OK management consultant 141848.  
## 2 CT management consultant 124786.  
## 3 GA management consultant 119995.  
## 4 NY management consultant 119995.  
## 5 PA management consultant 119995.  
## 6 CA data scientist 119800   
## 7 AZ data scientist 115000   
## 8 WA data scientist 115000   
## 9 DC data scientist 114500   
## 10 MN data scientist 111457   
## # … with 215 more rows

## `summarise()` has grouped output by 'JOB\_TITLE\_SUBGROUP'. You can override  
## using the `.groups` argument.



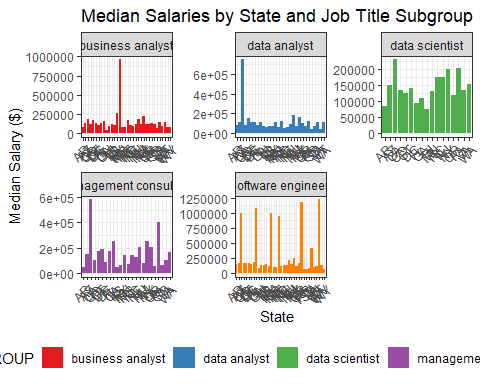
In this code block, the user is analyzing data related to salaries across different states and job title subgroups. They are using the dplyr and ggplot2 libraries to group and filter the data, and to create visualizations that help them identify patterns and trends in the data.

First, they calculate the median salary by state and job title subgroup using the group\_by and summarise functions. They then identify the top 5 and bottom 5 states based on median salary for each job title subgroup using the top\_n and arrange functions.

Next, they filter the data to only include the top and bottom states using the filter function. Finally, they create a grouped bar plot using ggplot2 that shows median salaries by state and job title subgroup, with each subgroup shown in a separate facet.

This approach is useful because it allows the user to easily compare salaries across different states and job title subgroups, and to identify which states have the highest and lowest median salaries for each subgroup. The visualization makes it easy to see patterns and trends in the data, and can help the user identify areas where further investigation or action may be needed.

## `summarise()` has grouped output by 'JOB\_TITLE\_SUBGROUP'. You can override  
## using the `.groups` argument.



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

## # A tibble: 50 × 3  
## # Groups: JOB\_TITLE\_SUBGROUP [5]  
## EMPLOYER\_NAME JOB\_TITLE\_SUBGROUP mean\_sa…¹  
## <chr> <chr> <dbl>  
## 1 THE UNIVERSITY OF TEXAS SYSTEM ADMINISTRATION business analyst 677508   
## 2 OFFICEMAX INCORPORATED business analyst 603712.  
## 3 SIGMATEK SYSTEMS, LLC software engineer 600000   
## 4 CO-CREATION PARTNERS, INC. management consultant 582400   
## 5 ALIASWIRE, INC. software engineer 528000   
## 6 LOAD DYNAMIX, INC. software engineer 486650   
## 7 INSIDE, INC. software engineer 474896.  
## 8 INTUIT data analyst 433162.  
## 9 KEY software engineer 412500   
## 10 LANDIS GYR TECHNOLOGY, INC software engineer 401912   
## # … with 40 more rows, and abbreviated variable name ¹​mean\_salary