Salary Data Set Analysis

Ana Paula Felix de Queiroz

Data 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 x 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.

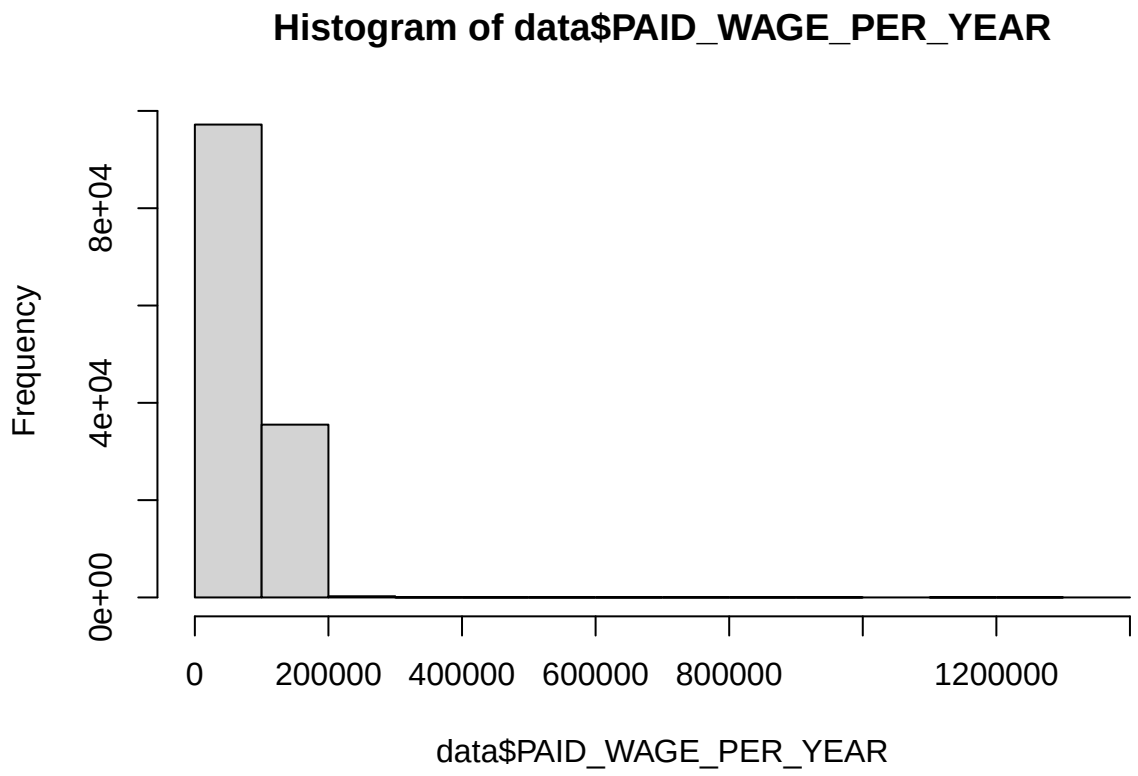
```
## The following objects are masked from data (pos = 3):
##
##     CASE_NUMBER, CASE_RECEIVED_DATE, CASE_STATUS,
##     COLLEGE_MAJOR_REQUIRED, COUNTRY_OF_CITIZENSHIP, DECISION_DATE,
##     EDUCATION_LEVEL_REQUIRED, EMPLOYER_NAME,
##     EXPERIENCE_REQUIRED_NUM_MONTHS, EXPERIENCE_REQUIRED_Y_N,
##     FULL_TIME_POSITION_Y_N, JOB_TITLE, JOB_TITLE_SUBGROUP, order,
```

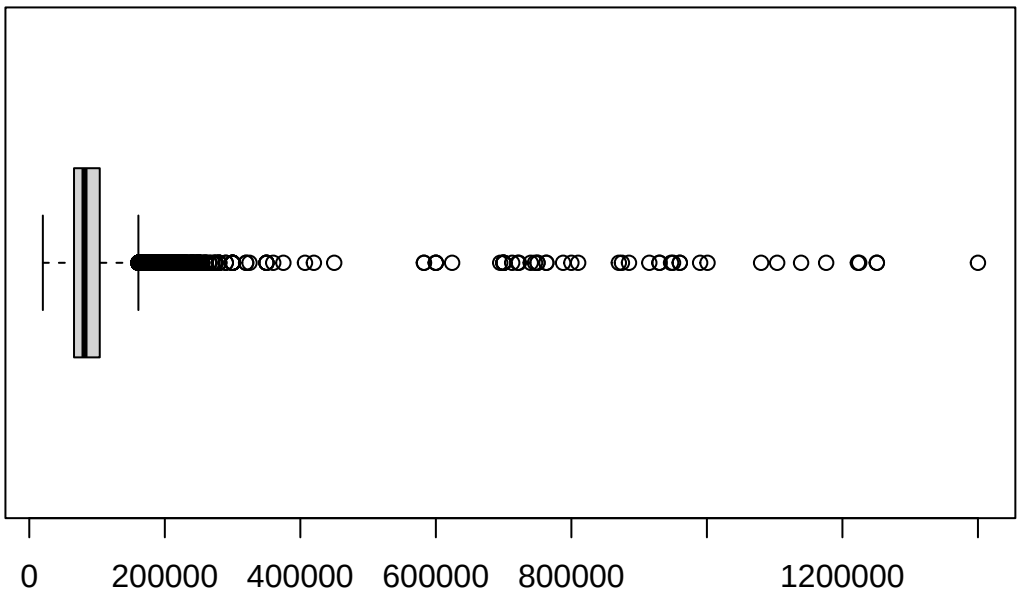
```
##     PAID_WAGE_PER_YEAR, PAID_WAGE_SUBMITTED, PAID_WAGE_SUBMITTED_UNIT,
##     PREVAILING_WAGE_PER_YEAR, PREVAILING_WAGE_SOC_CODE,
##     PREVAILING_WAGE_SOC_TITLE, PREVAILING_WAGE_SUBMITTED,
##     PREVAILING_WAGE_SUBMITTED_UNIT, VISA_CLASS, WORK_CITY,
##     WORK_POSTAL_CODE, WORK_STATE, WORK_STATE_ABBREVIATION

## 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
```





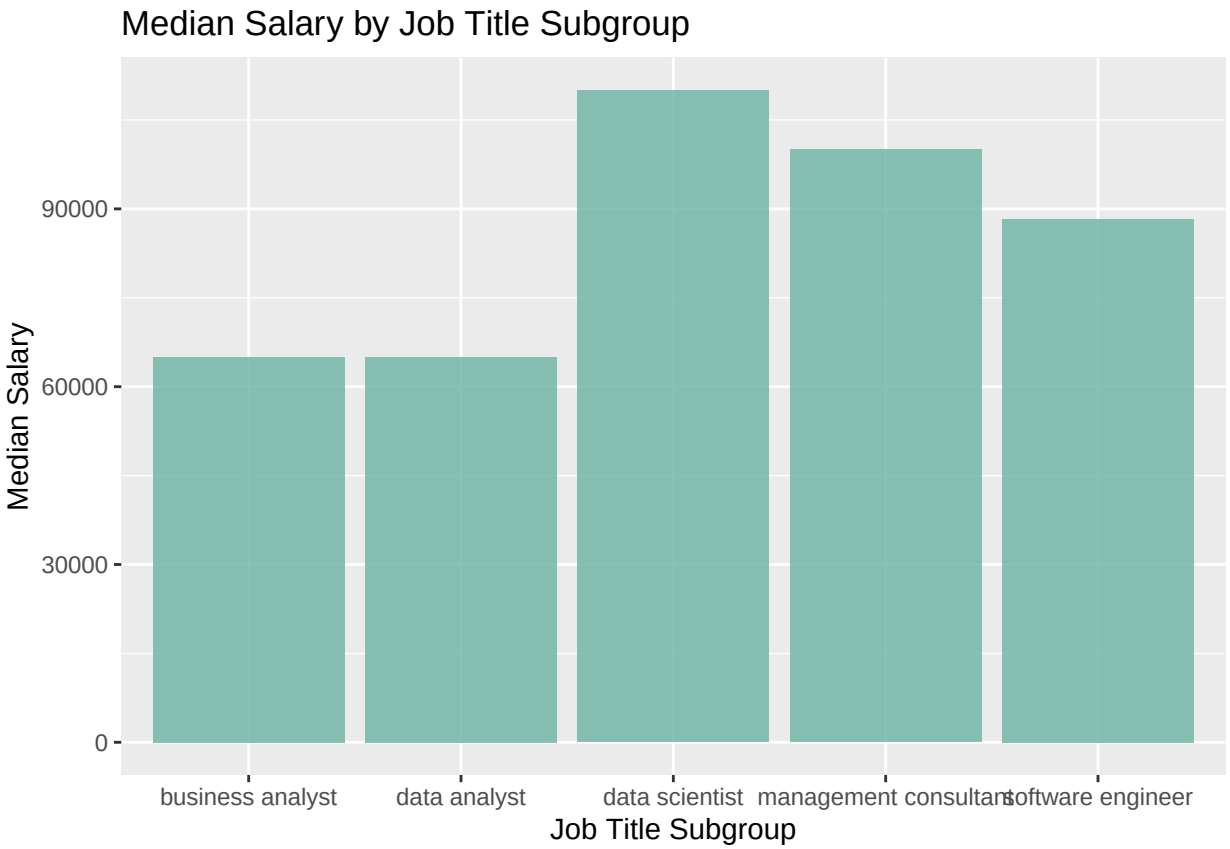
Descriptive Stats and Visualization by Job Title Subgroup

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.

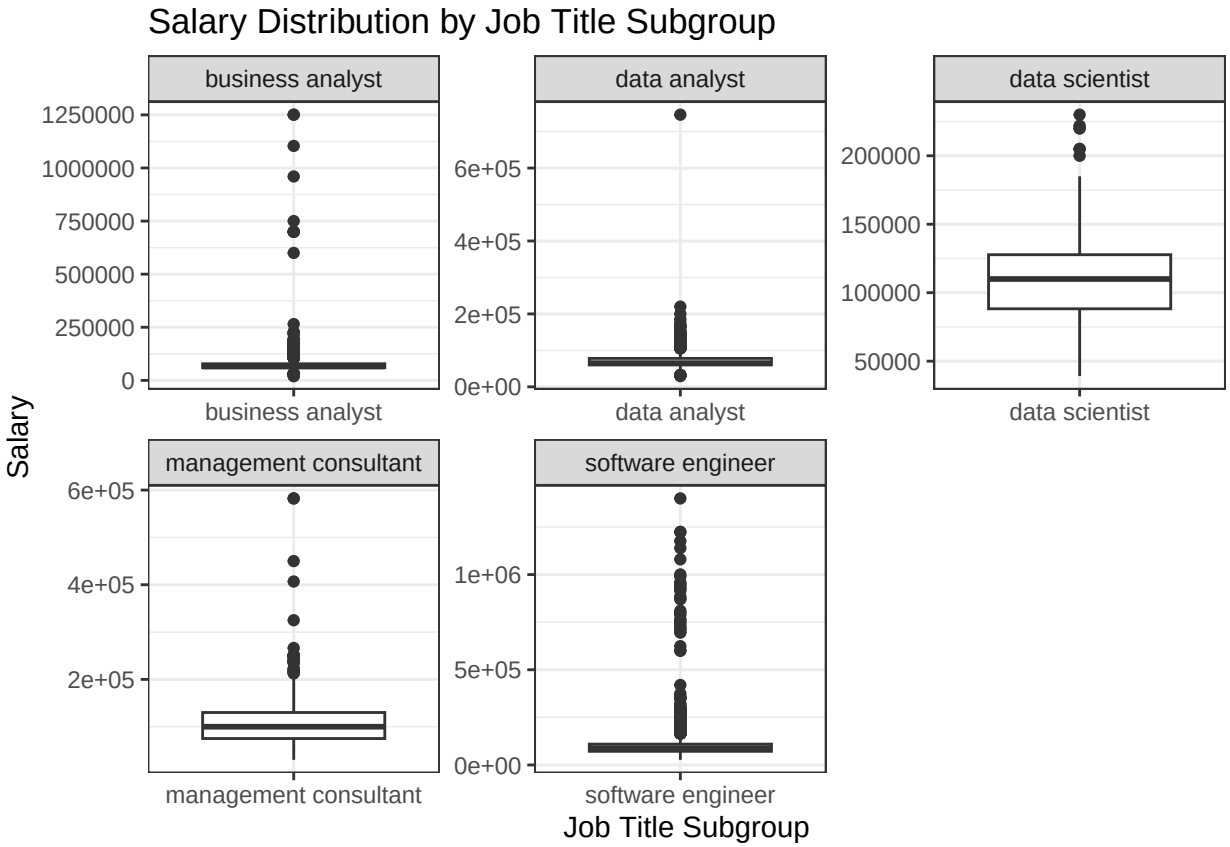
To accomplish this, we first group the data by job title subgroup and calculate the median salary for each subgroup. We then create box plots for each subgroup, which show the distribution of salaries and any potential outliers. The box plots are separated by subgroup to allow for easier comparison between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

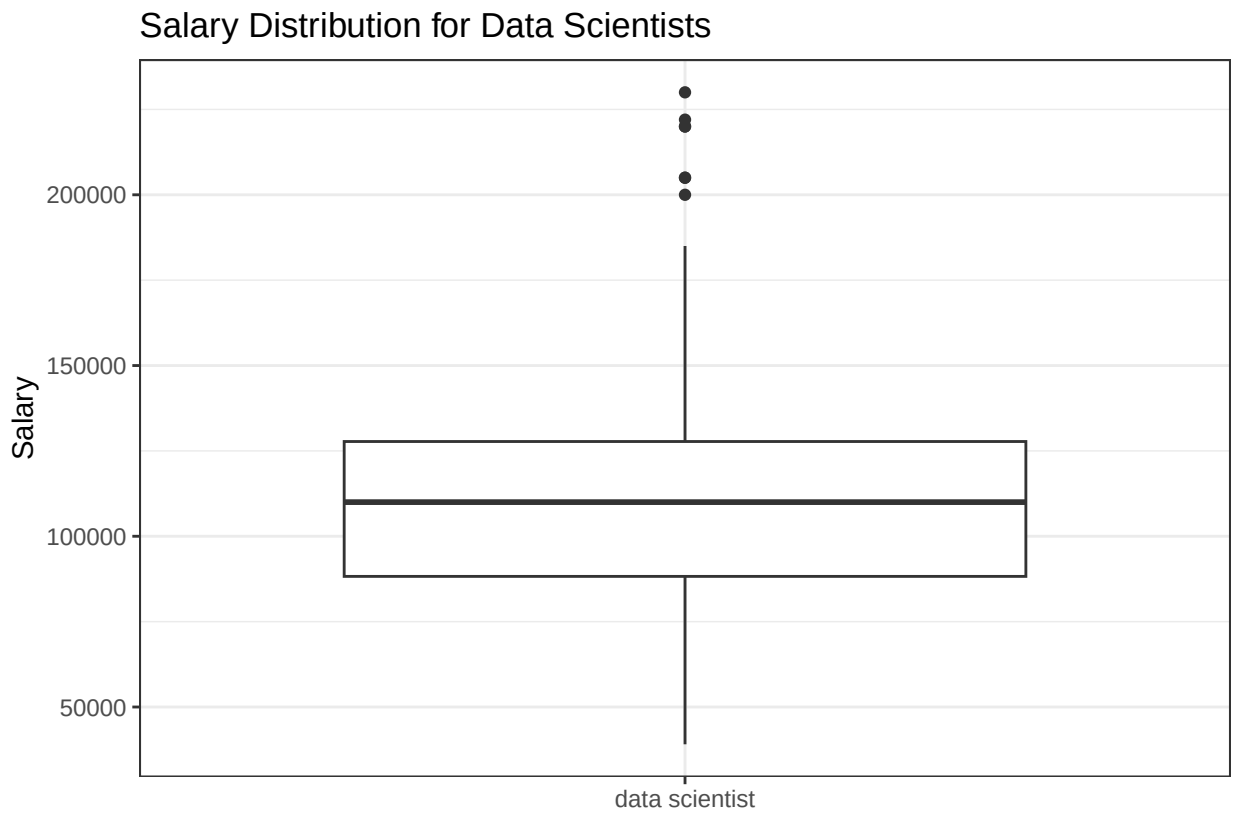
```
##
## Attaching package: 'dplyr'

## 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
```


Salary Distribution by Job Title Subgroup
business analyst
data analyst
data scientist
management consultant
software engineer
Salary
1250000
1000000
750000
500000
250000
0
6e+05
4e+05
2e+05
0e+00
200000
150000
100000
50000
1e+06
5e+05
business analyst
data analyst
data scientist
management consultant
software engineer
Job Title Subgroup



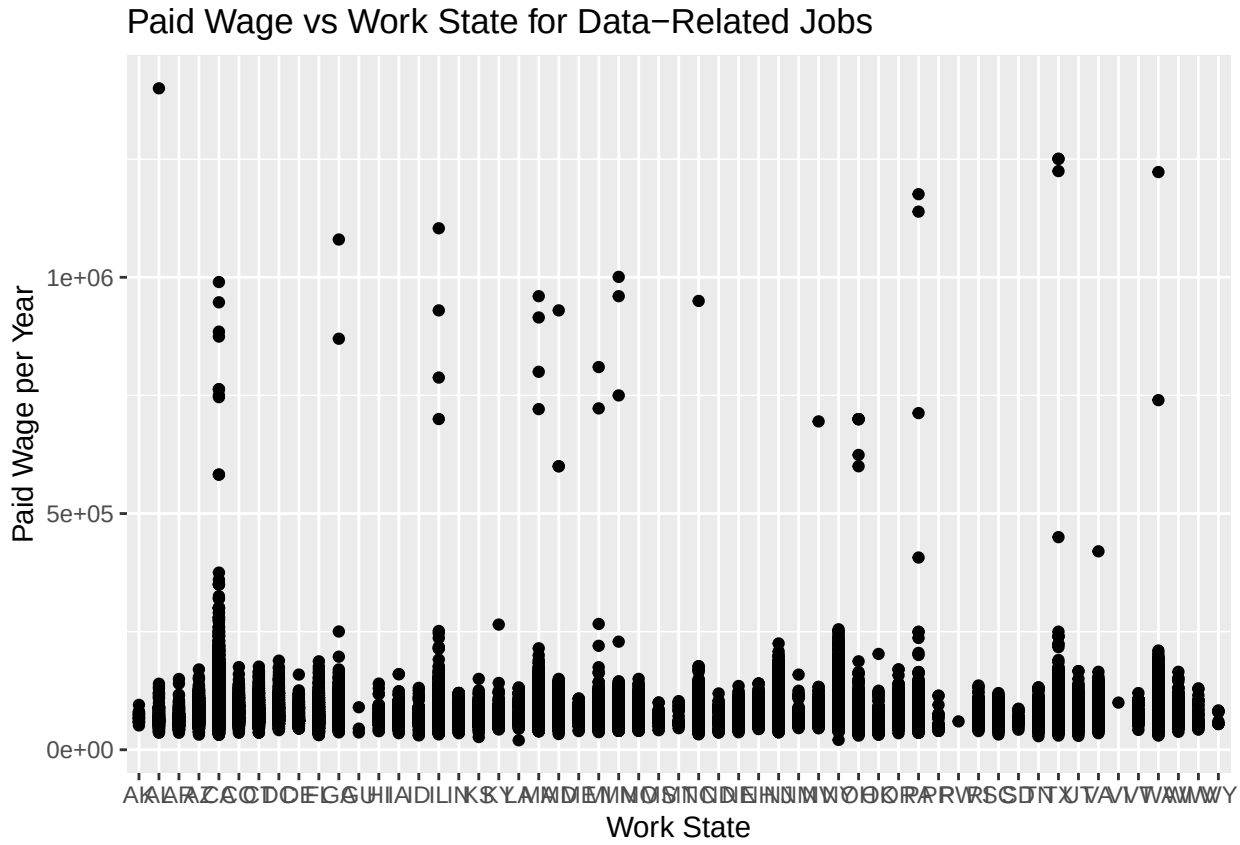
Wage payed across different US states:

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.

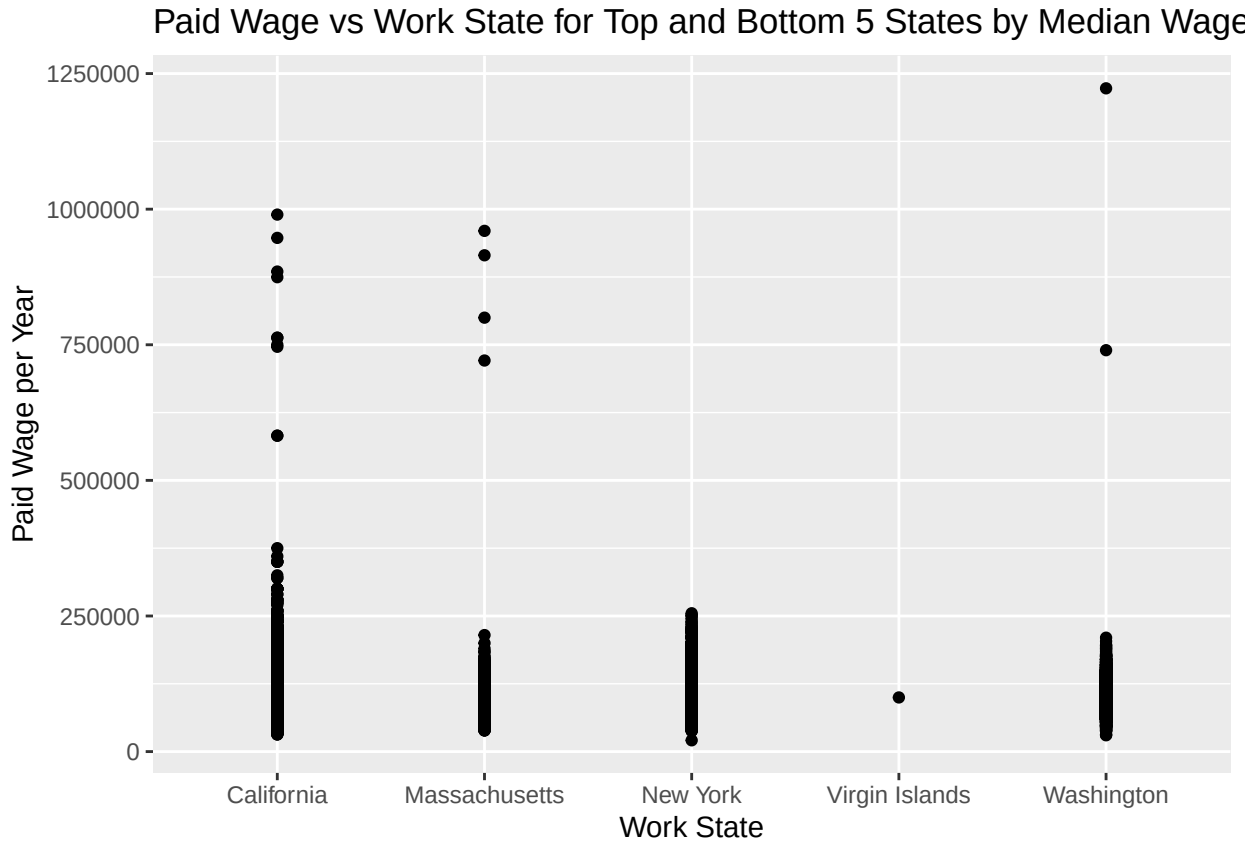
In the third block of code, a scatter plot is created to show the paid wage versus work state, but only for the top and bottom five states identified in the second block.

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




```
## # A tibble: 5 x 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 x 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.

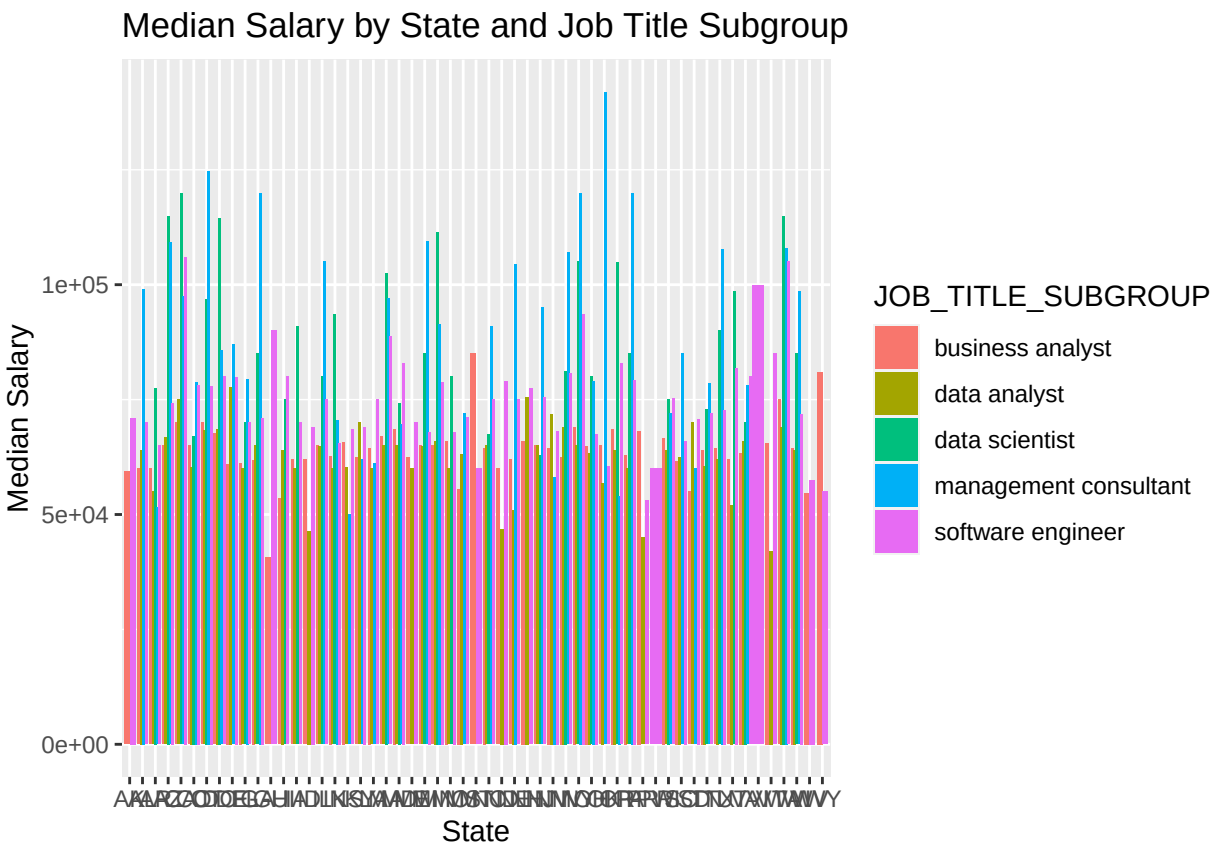
Then, I create a grouped bar plot to visualize the median salaries for each job title subgroup in each state. This plot provides an easily interpretable visual representation of the differences in median salaries across job title subgroups and states.

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 x 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.
```



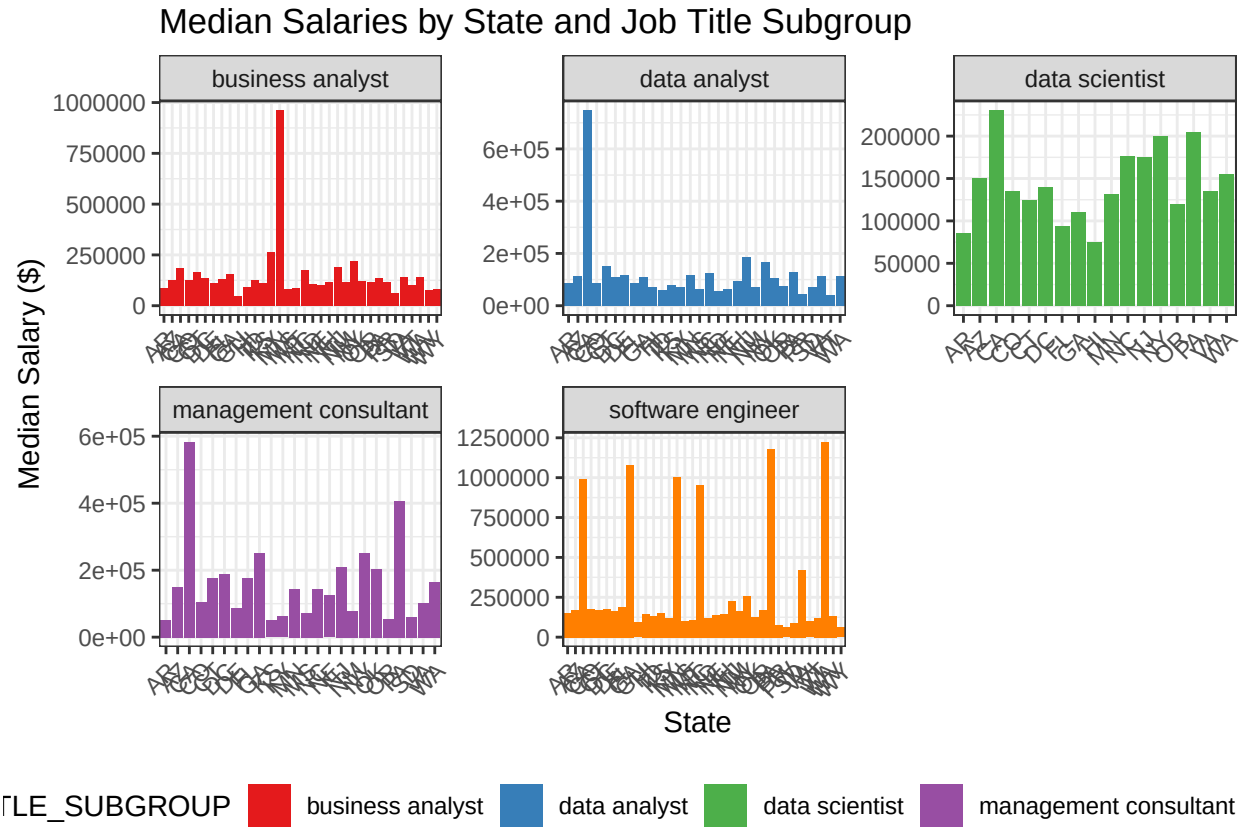
In this code block, the user is analyzing data related to salaries across different states and job title subgroups. They are using the dplyr and ggplot2 libraries to group and filter the data, and to create visualizations that help them identify patterns and trends in the data.

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 x 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
```