Lessa Fleming

Project- Data Scientist Salaries

IST 707 Tuesdays

Introduction:

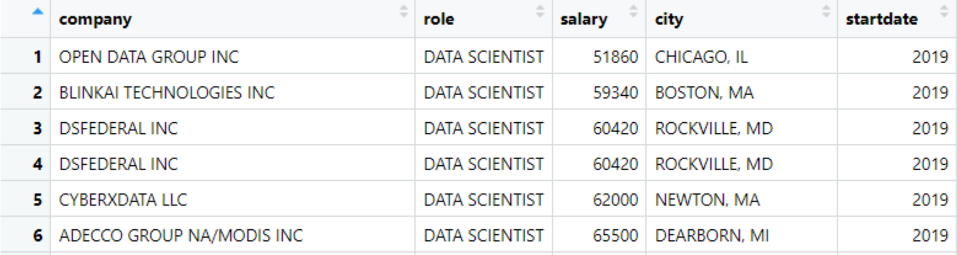
An analysis was completed on a data set from Kaggle of Data Science salaries offered from 2019-2022.This analysis will help to predict what salaries may be offered in 2023 and what a persona may expect when negotiating for a data scientist role. Multiple analysis methods will be used to help answer the question presented on the data. This is a replacement set of data as the first data set was unable to be cleaned in a timely manner to proceed with the analysis due to text issues. The new data set was precleaned and only required a small amount of cleaning to prepare the data for different methods.

Analysis and Models

About the data

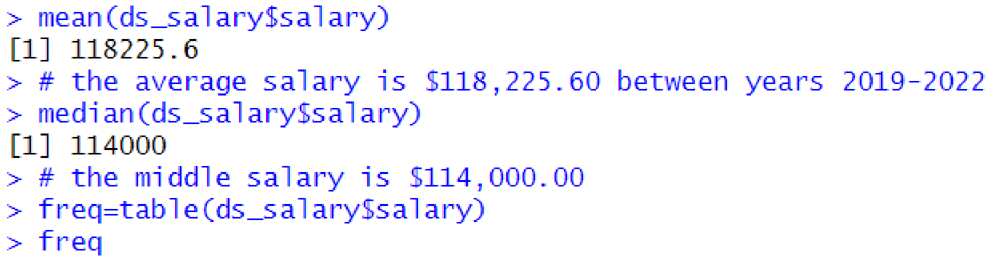
This data set includes five columns: Company, Role, Salary, City and Startdate and has 13,321 rows of data. This was already precleaned and had no empty or null data, and the salaries were free of and symbols and punctuation which all allowed for easier processing of the data.

To start the preparation of this data it was read into R and stored as a data set named ds_salary. The data was then viewed to see what was listed in the data set to do a double check to make sure there is no missing or incorrect data.

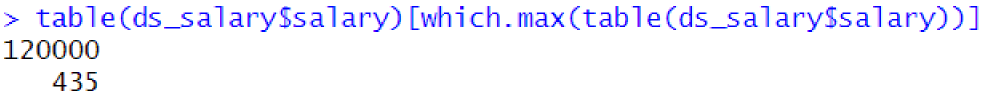


	company	role	salary	city	startdate
1	OPEN DATA GROUP INC	DATA SCIENTIST	51860	CHICAGO, IL	2019
2	BLINKAI TECHNOLOGIES INC	DATA SCIENTIST	59340	BOSTON, MA	2019
3	DSFEDERAL INC	DATA SCIENTIST	60420	ROCKVILLE, MD	2019
4	DSFEDERAL INC	DATA SCIENTIST	60420	ROCKVILLE, MD	2019
5	CYBERXDATA LLC	DATA SCIENTIST	62000	NEWTON, MA	2019
6	ADECCO GROUP NA/MODIS INC	DATA SCIENTIST	65500	DEARBORN, MI	2019

An initial review of the data was completed to see where most of the information was within the data. This was done by viewing the mean, median and Freq of the salary ranges.



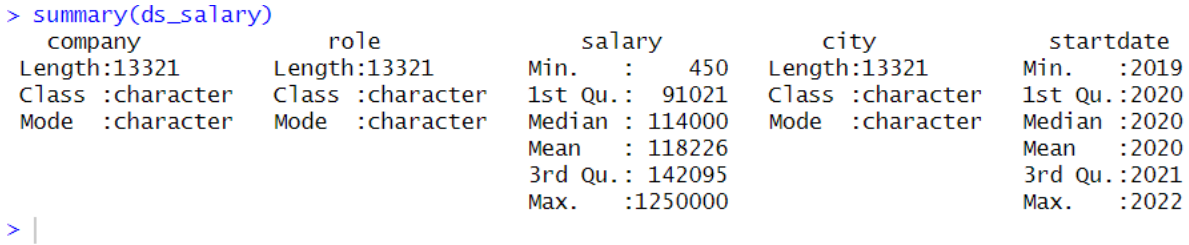
```
> mean(ds_salary$salary)
[1] 118225.6
> # the average salary is $118,225.60 between years 2019-2022
> median(ds_salary$salary)
[1] 114000
> # the middle salary is $114,000.00
> freq=table(ds_salary$salary)
> freq
```



```
> table(ds_salary$salary)[which.max(table(ds_salary$salary))]
120000
   435
```

Once that was completed the min, max and quartiles were viewed, and the summary of the data was displayed:

```
> sd(ds_salary$salary)  #      standard      deviation
[1] 36440.64
> max(ds_salary$salary)
[1] 1250000
> min(ds_salary$salary)
[1] 450
> range <- max(ds_salary$salary)       - min(ds_salary$salary)
> range
[1] 1249550
> qt    <- quantile(ds_salary$salary,   na.rm=TRUE)     #
alues
> qt
     0%     25%     50%     75%    100%
    450   91021  114000  142095 1250000
```



```
> summary(ds_salary)
   company               role               salary              city              startdate   
 Length:13321       Length:13321       Min.   :    450   Length:13321       Min.   :2019  
 Class :character   Class :character   1st Qu.:  91021   Class :character   1st Qu.:2020  
 Mode  :character   Mode  :character   Median : 114000   Mode  :character   Median :2020  
                                       Mean   : 118226                      Mean   :2020  
                                       3rd Qu.: 142095                      3rd Qu.:2021  
                                       Max.   :1250000                      Max.   :2022  
> |
```

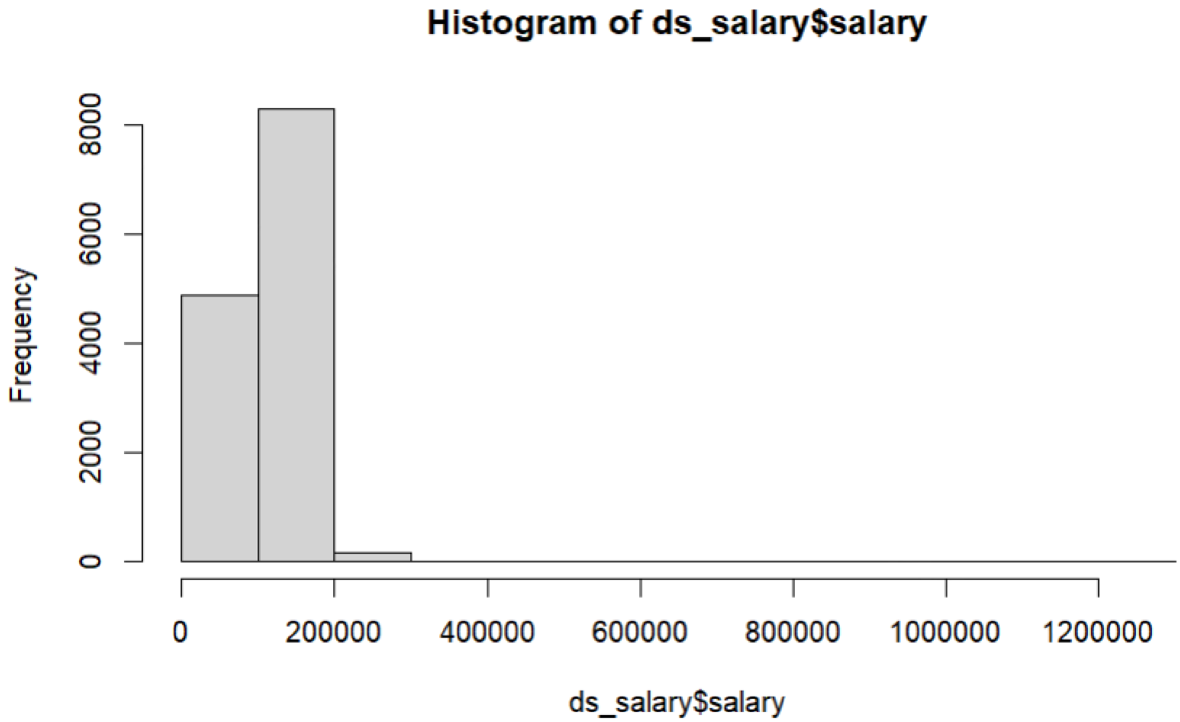
Initial Results before further Analysis

In the initial review the average salary for all the years combined was $118,225.00. The most frequent salary was $120,000, with 435 occurrences, and falls in the middle of the 50th and 75th percentile in the dataset. The maximum salary was $1,250,000.

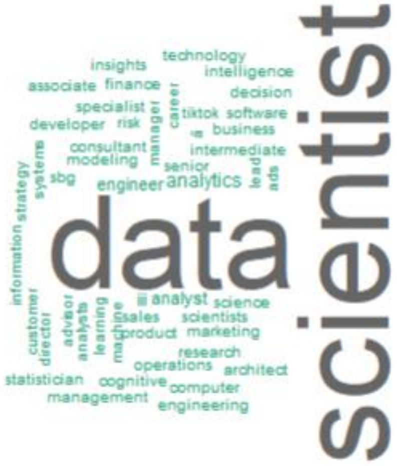
Models

The following models were created from the stored data sets to help understand the data and better analyze the data for results.

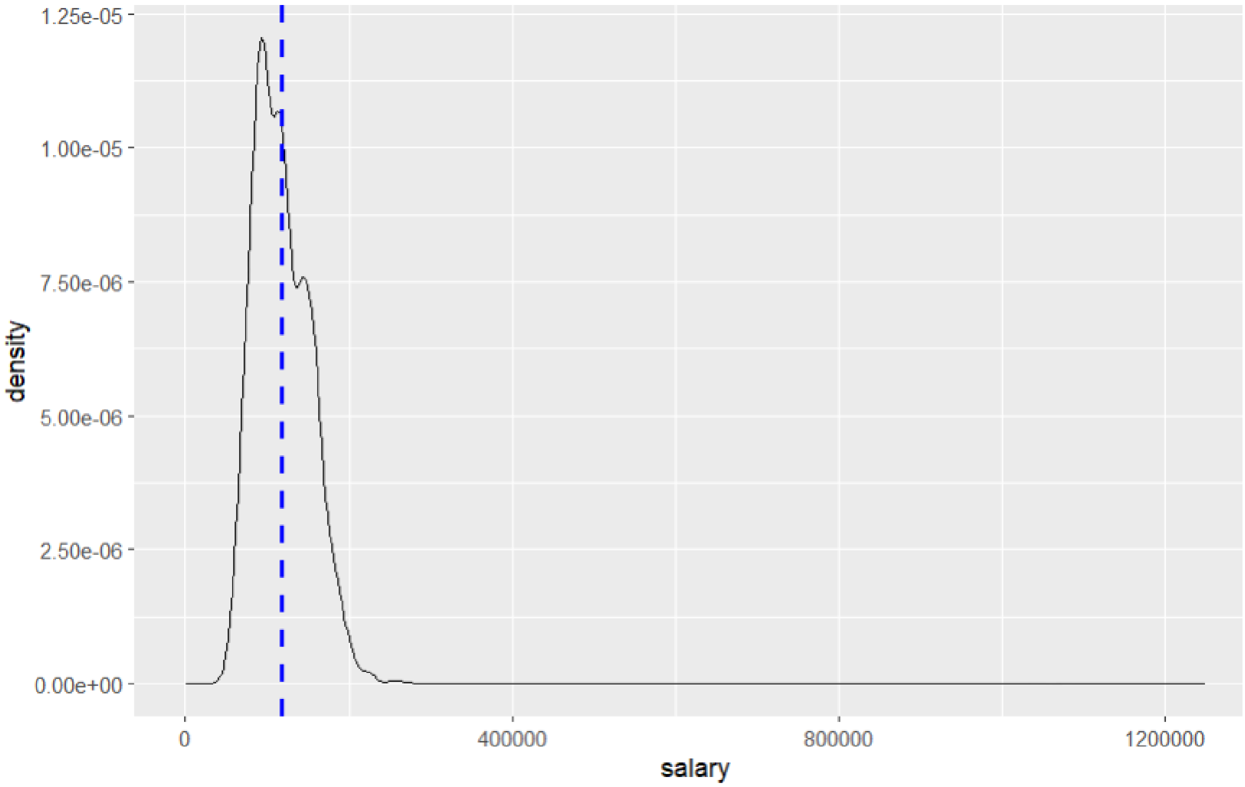
Histogram- Salary compared to frequency.

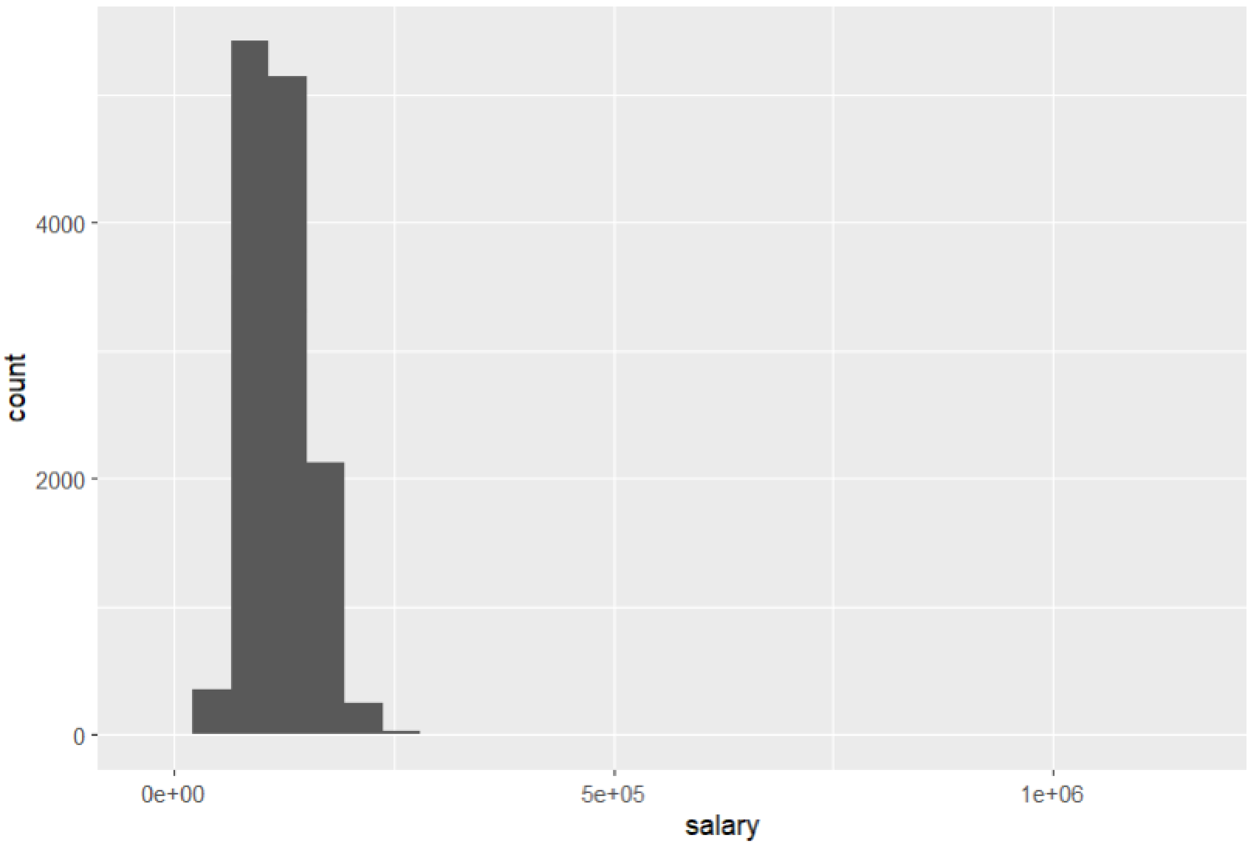


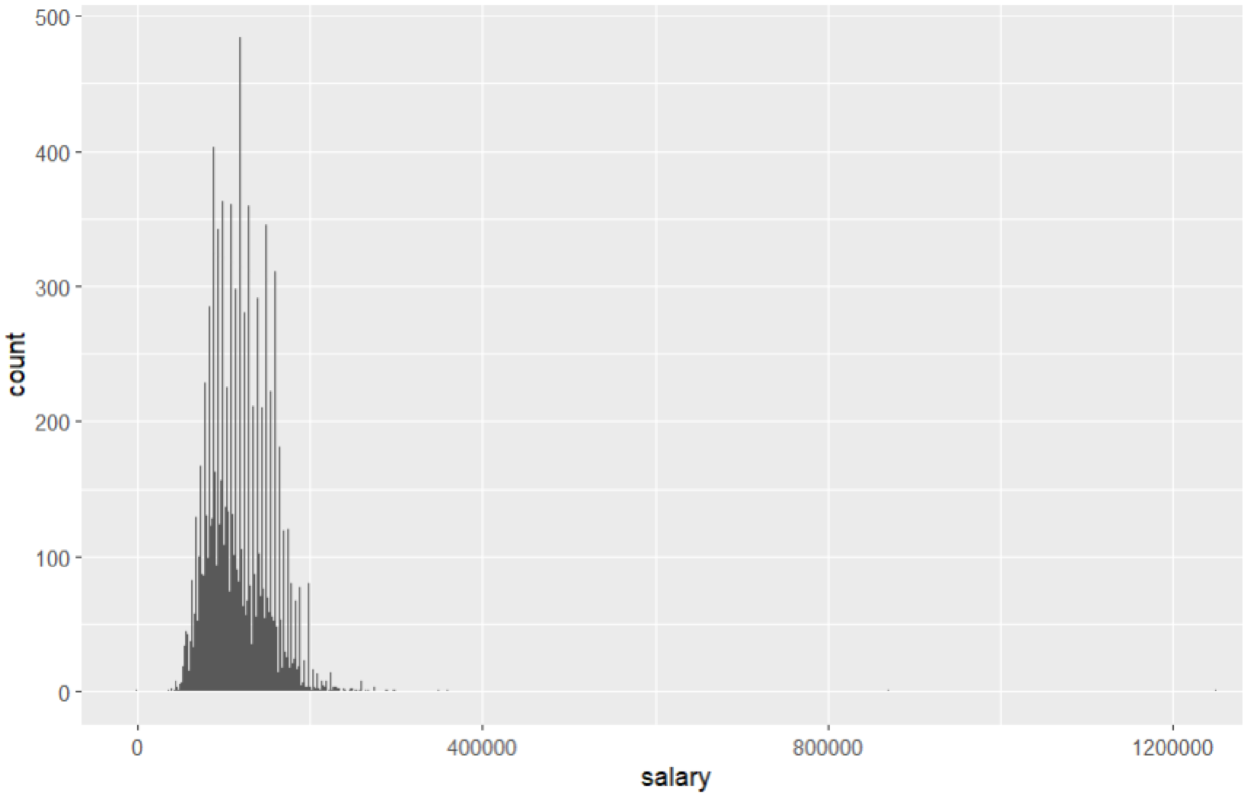
Word Cloud- Views the most frequent words (limited to 50)

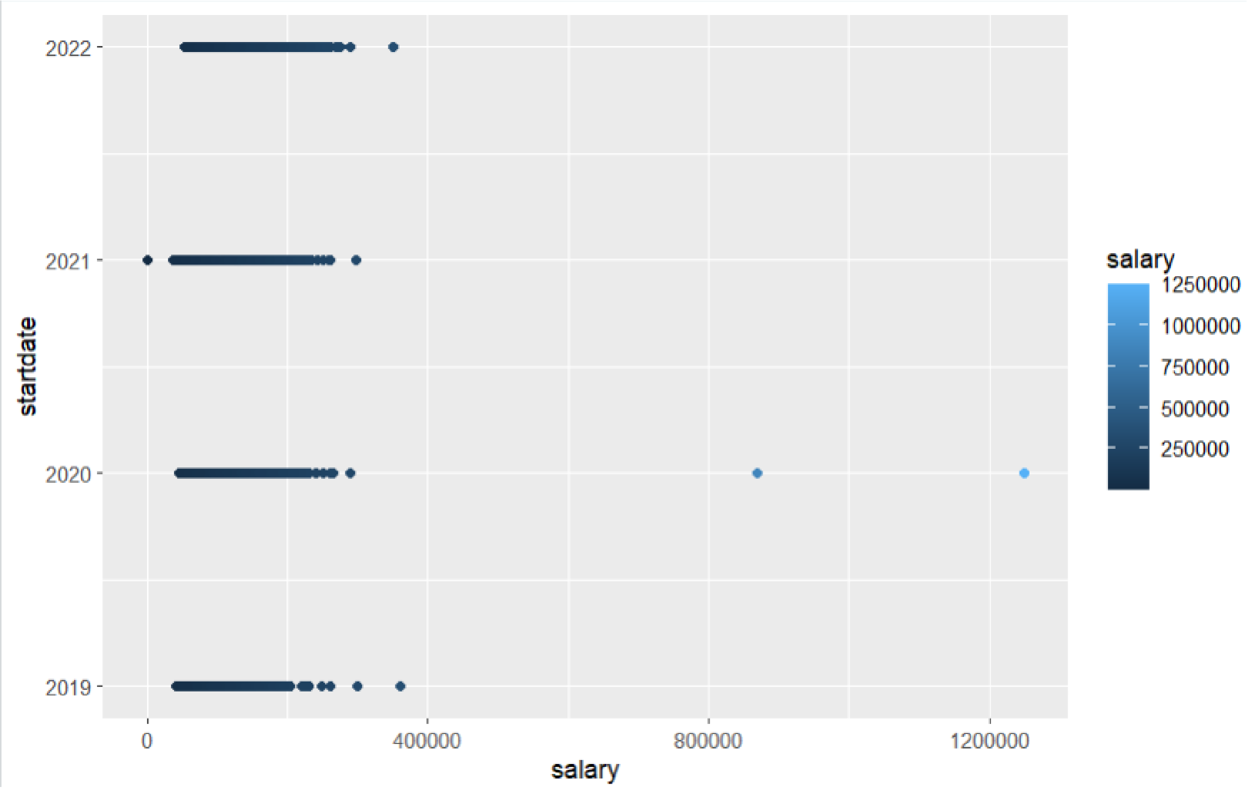


GGPLOTS-density with mean line, bar chart showing the count of the salaries from the entire dataset. Salary broken down by year, the coloring is the salary range.

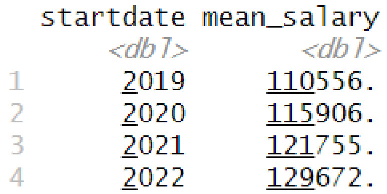

1.25e-05
1.00e-05
7.50e-06
5.00e-06
2.50e-06
0.00e+00
density
0
400000
800000
1200000
salary


count
4000
2000
0
0e+00
5e+05
1e+06
salary

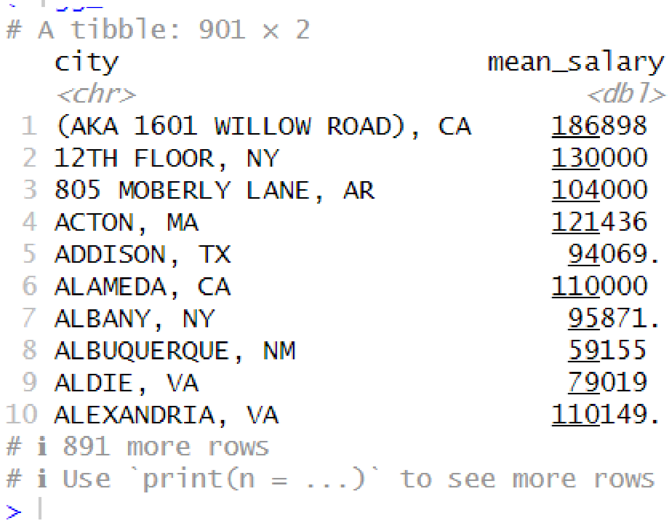

500
400
300
200
100
0
count
0
400000
800000
1200000
salary



Additional Models- Mean salaries per year, Mean salaries broken down by city.



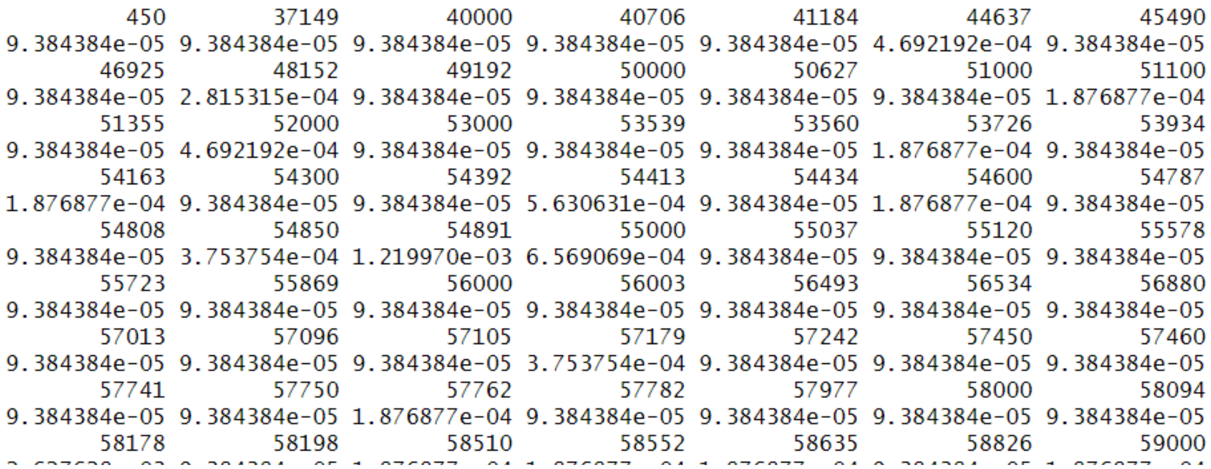
```
  startdate mean_salary
      <dbl>       <dbl>
1      2019     110556.
2      2020     115906.
3      2021     121755.
4      2022     129672.
```



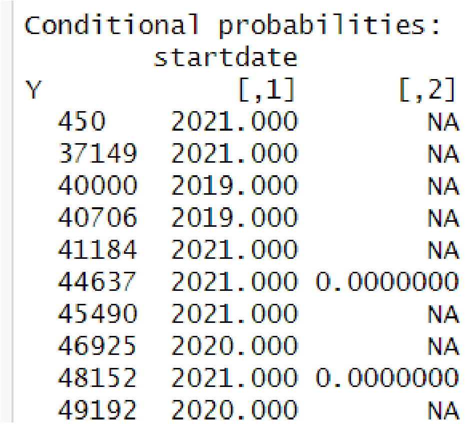
```
# A tibble: 901 × 2
   city                          mean_salary
   <chr>                               <dbl>
 1 (AKA 1601 WILLOW ROAD), CA        186898
 2 12TH FLOOR, NY                    130000
 3 805 MOBERLY LANE, AR              104000
 4 ACTON, MA                         121436
 5 ADDISON, TX                        94069.
 6 ALAMEDA, CA                       110000
 7 ALBANY, NY                         95871.
 8 ALBUQUERQUE, NM                    59155
 9 ALDIE, VA                          79019
10 ALEXANDRIA, VA                    110149.
# ℹ 891 more rows
# ℹ Use `print(n = ...)` to see more rows
> |
```

Data Mining Models-Probability table, Naïve Bayes Model, Predicted Salary with Naïve bayes test data, Decision tree, and randomForest.

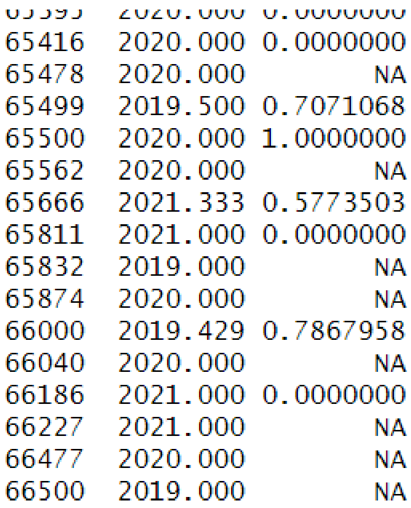
Probability-



```
         450        37149        40000        40706        41184        44637        45490 
9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 4.692192e-04 9.384384e-05 
       46925        48152        49192        50000        50627        51000        51100 
9.384384e-05 2.815315e-04 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 1.876877e-04 
       51355        52000        53000        53539        53560        53726        53934 
9.384384e-05 4.692192e-04 9.384384e-05 9.384384e-05 9.384384e-05 1.876877e-04 9.384384e-05 
       54163        54300        54392        54413        54434        54600        54787 
1.876877e-04 9.384384e-05 9.384384e-05 5.630631e-04 9.384384e-05 1.876877e-04 9.384384e-05 
       54808        54850        54891        55000        55037        55120        55578 
9.384384e-05 3.753754e-04 1.219970e-03 6.569069e-04 9.384384e-05 9.384384e-05 9.384384e-05 
       55723        55869        56000        56003        56493        56534        56880 
9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 
       57013        57096        57105        57179        57242        57450        57460 
9.384384e-05 9.384384e-05 9.384384e-05 3.753754e-04 9.384384e-05 9.384384e-05 9.384384e-05 
       57741        57750        57762        57782        57977        58000        58094 
9.384384e-05 9.384384e-05 1.876877e-04 9.384384e-05 9.384384e-05 9.384384e-05 9.384384e-05 
       58178        58198        58510        58552        58635        58826        59000 
```

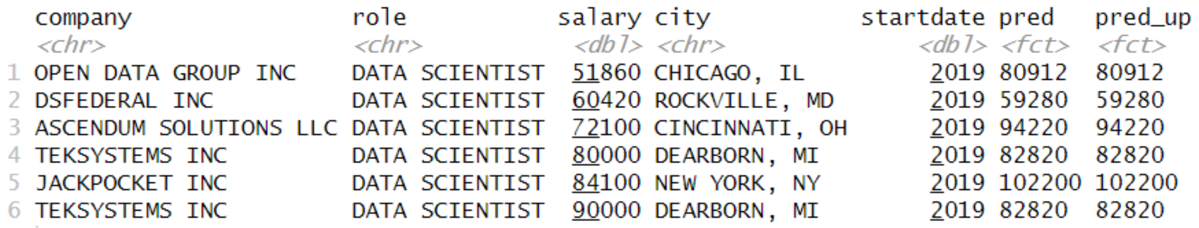


```
Conditional probabilities:
       startdate
Y            [,1]      [,2]
  450    2021.000        NA
  37149  2021.000        NA
  40000  2019.000        NA
  40706  2019.000        NA
  41184  2021.000        NA
  44637  2021.000 0.0000000
  45490  2021.000        NA
  46925  2020.000        NA
  48152  2021.000 0.0000000
  49192  2020.000        NA
```



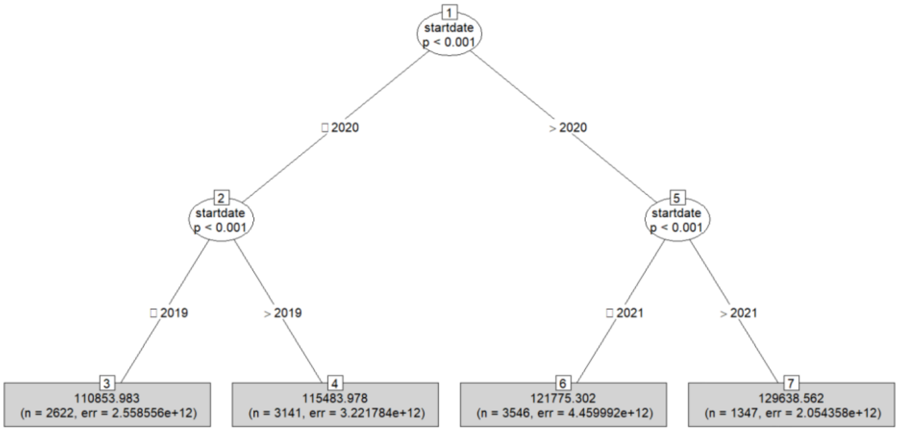
```
  65416  2020.000 0.0000000
  65478  2020.000        NA
  65499  2019.500 0.7071068
  65500  2020.000 1.0000000
  65562  2020.000        NA
  65666  2021.333 0.5773503
  65811  2021.000 0.0000000
  65832  2019.000        NA
  65874  2020.000        NA
  66000  2019.429 0.7867958
  66040  2020.000        NA
  66186  2021.000 0.0000000
  66227  2021.000        NA
  66477  2020.000        NA
  66500  2019.000        NA
```

Prediction-

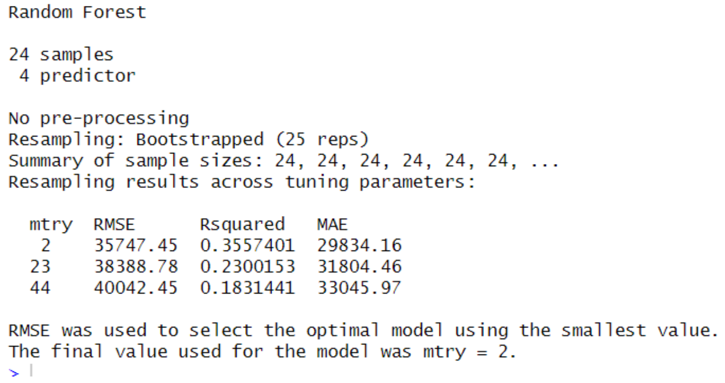


```
  company               role           salary city            startdate pred   pred_up
  <chr>                 <chr>           <dbl> <chr>               <dbl> <fct>  <fct>  
1 OPEN DATA GROUP INC   DATA SCIENTIST  51860 CHICAGO, IL          2019 80912  80912  
2 DSFEDERAL INC         DATA SCIENTIST  60420 ROCKVILLE, MD        2019 59280  59280  
3 ASCENDUM SOLUTIONS LLC DATA SCIENTIST 72100 CINCINNATI, OH       2019 94220  94220  
4 TEKSYSTEMS INC        DATA SCIENTIST  80000 DEARBORN, MI         2019 82820  82820  
5 JACKPOCKET INC        DATA SCIENTIST  84100 NEW YORK, NY         2019 102200 102200 
6 TEKSYSTEMS INC        DATA SCIENTIST  90000 DEARBORN, MI         2019 82820  82820  
```

Decision tree-



Random Forest-



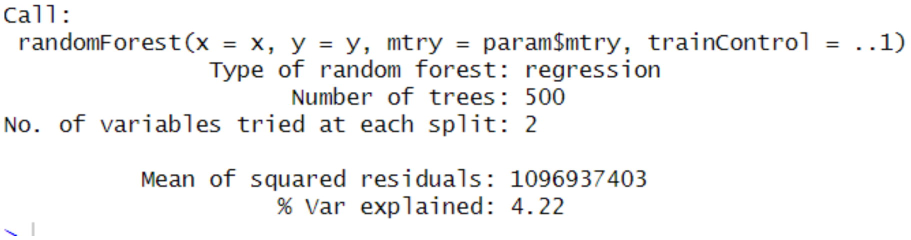
```
Random Forest

24 samples
 4 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 24, 24, 24, 24, 24, 24, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    35747.45  0.3557401  29834.16
  23    38388.78  0.2300153  31804.46
  44    40042.45  0.1831441  33045.97

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
>
```



```
Call:
 randomForest(x = x, y = y, mtry = param$mtry, trainControl = ..1)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 1096937403
                    % Var explained: 4.22
>
```

Results

The analysis of the data set produced many different answers to the posed question of what salary in 2023 would be best to negotiate for with the skills and knowledge of an education for Data Science. The initial results showed an increase year over year in the mean salary offered from the different job listings. The mean salary was $118,000.00 for all the years combined and the mean salary for 2022 was $129,000 which is close to the mean for all the years. T can also be seen that there was around a $5200.00 increase each year to the mean salary for the first three years and the las year showed an increase of around $7000 to the 2022 mean salary.

The first model is a histogram of salary and frequence, this showed that the most frequent salaries were between $100-200K with under $100,000 being the second. Over $200,000 made a significant drop with the least amount of reportted salaries. The density charts, and additional bar charts showing the frequencey and desity reinfoced this with showinng the same infomration on where the most frequent salaries were displayed.

A word cloud was created to display the most frequenct key words used in the decription of the job listing. These key words were limited to the top 50 words for easier review and display. These showed words like Data Scientist which would be expected as the top as this is listed in every description. After those, word like machine, analytics, insight, analyst, and engeinner were displayed.

The models then move into the data mining portion of the analysis. The first is the probability table, in this all the salaries are listed and their probabilities. The most frequent probability was the 9.384384… and 1.876877… this was run on the training data. Once that was viewed the Naïve Bayes model was created to use for the prediction model. Looking at the head of the tst dataset after prediction was applied the predictions show a range of different salaries for the top 6 chosen.

A decision tree was created uing the salary and start date to determine which lead to the best salary choice. There are two branches on this tree, 2020 and under, and greater than 2020. Different tuning was done to try and create the best model and this set up created the most understandable model. The results show that the greatest salary is over the year 2021.

RandomForest was the used for the final models in the data. The data was sampled and a training and test set created. The probability table used the training data that contained the sample set. All prbabilities with the exception of one were 0.04166667. This was then applied to the training data and a model created. There were 24 samples and 4 predictors. Mtry of 2 was used as the final value as this was the lowest and determined to be the optimal model.

Conclusions

In a final review of the data, year after year the salaries for a Data Scientist increase. There was a steady increase of an average of $5000.00 for the first three years than an increase of $7000.00 from 2021 to 2022's salary averages. It can be seen in the average salaries per city that there are great differences in the salaries.

While there are differences in the average salaries per city, the average salary overall for all years was $188,000 and the most frequent salary offered overall was $120,00.00. After performing prediction analysis on the data, it can be seen in the top 6 results around a $10,000.00 difference in what was offered and what was predicted.

When applying for a Data Science role in 2023 based on all analysis and results seen, a negotiated or expected amount for a salary shows to be between $124,671 - $134,671.