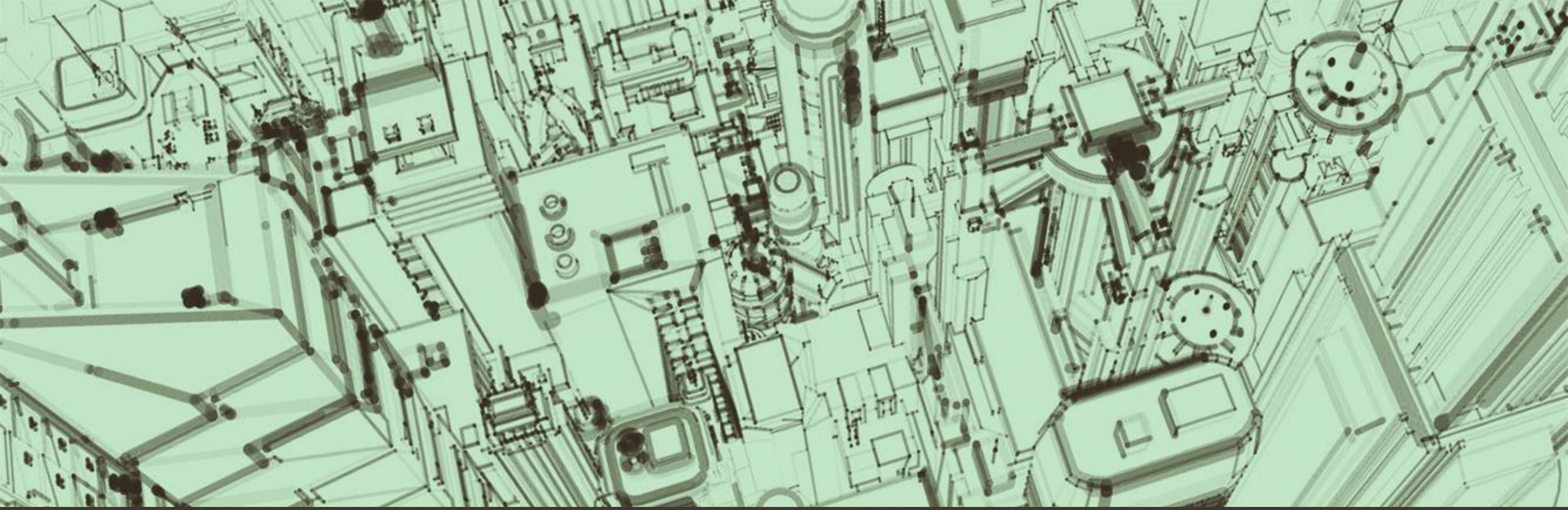


Data Scientist Salaries

Lessa Fleming

Contents

- Introduction
- Data
- Analysis
- Results
- Conclusion

Introduction

- This report is to review Data Scientist salaries from 2019-2022. This is being done as the salaries are important to those who are going into a Data science role as I am.
- Knowing salaries can help when the time comes to negotiate for future roles within the field based on job title or description.
- My goal is to be prepared to know past salaries and be able to estimate any increases year over year in the salaries.

Initial Analysis

- The mean salary between 2019-2022 $118225.6
- The most frequent salary was $120,000.00 at 435 occurrences
- The max salary was $1,250,000.00 and the min was $450
- Quartiles:

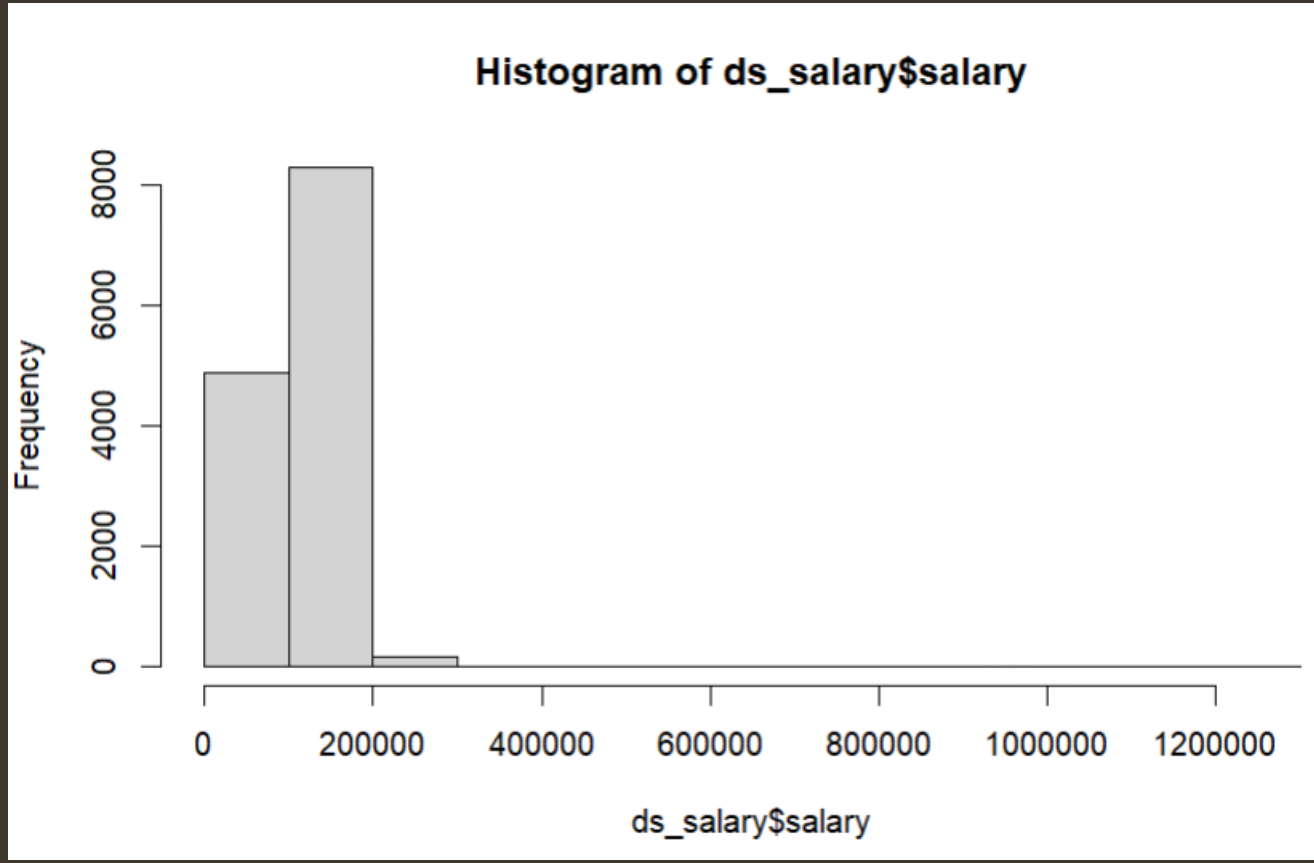
0%	25%	50%	75%	100%
450	91021	114000	142095	1250000

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

	startdate	mean_salary
1	2019	110556.2
2	2020	115905.6
3	2021	121754.8
4	2022	129671.9

Histogram 1- Frequency of Salary

- In this you can see that the most frequent salaries are between 100-200K
- The second highest frequency was between 0-100K
- There are minimal salaries over 200K with a significant drop



Word Cloud

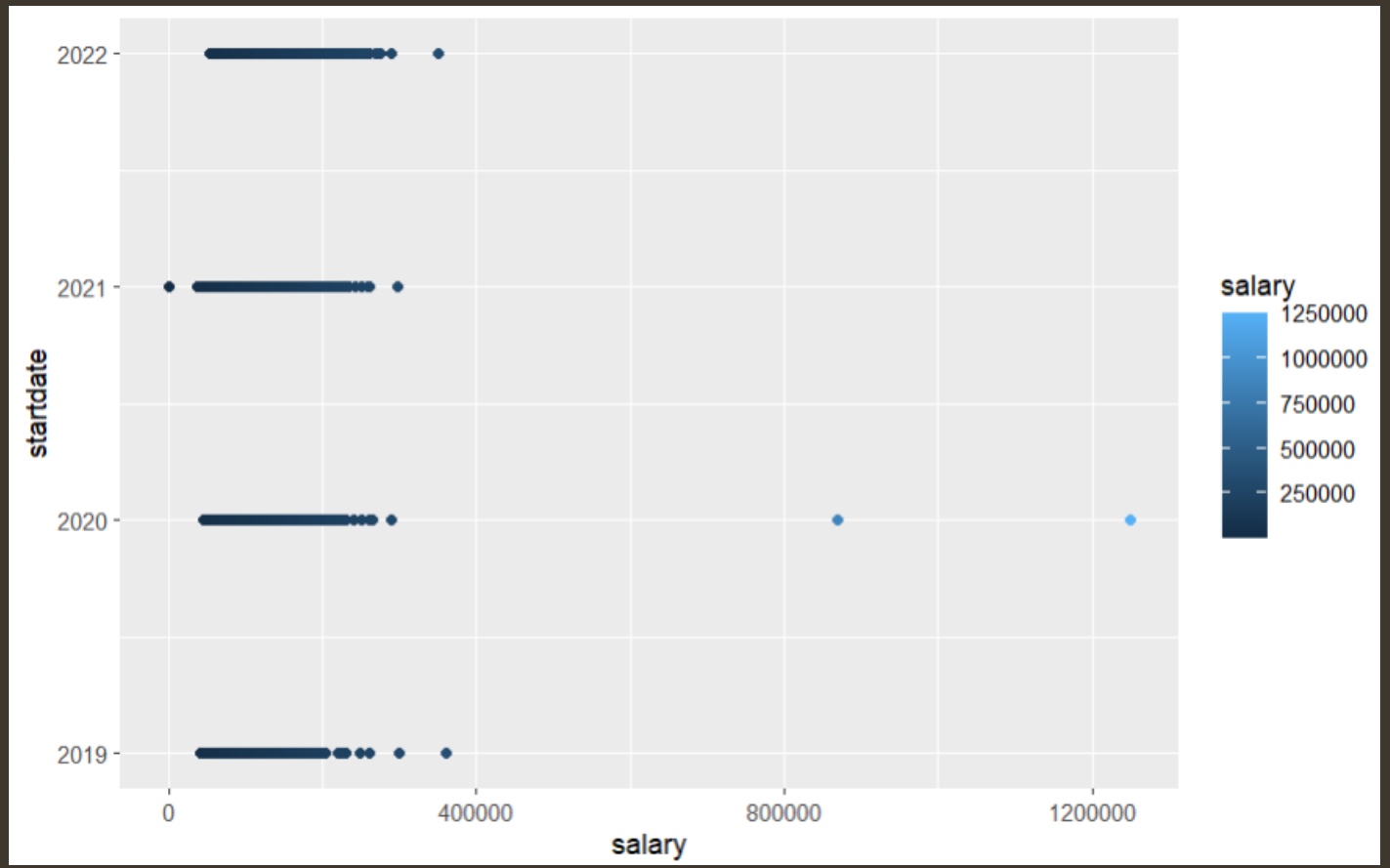
- A word cloud was created to see the most frequent word used in the role field
- As assumed Data Science was the top two words, after that analytics, intelligence, machine, risk and analyst were some of the additional top word that were surfaced
- This was minimized to pull 50 words as over that created a difficult word cloud

Head and Tail

	company	role	salary	city	startdate
	<chr>	<chr>	<dbl>	<chr>	<dbl>
1	FIDDLER LABS INC	DATA SCIENTIST	450	BOSTON, MA	2021
2	MACHINE INTELLIGENCE TECHNOLOGIES LLC	DATA SCIENTIST	37149	MIAMI, FL	2021
3	RANGE DIGITAL MARKETING LLC	DATA SCIENTIST	40000	MINNEAPOLIS, MN	2019
4	THE REINALT-THOMAS CORPORATION	DATA SCIENTIST I	40706	SCOTTSDALE, AZ	2019
5	MACHINE INTELLIGENCE TECHNOLOGIES LLC	DATA SCIENTIST	41184	LAUDERDALE, FL	2021
6	INNOVATIVE ADVOCATE GROUP	DATA SCIENTIST	44117	RED BANK, NJ	2019

	company	role	salary	city	startdate
	<chr>	<chr>	<dbl>	<chr>	<dbl>
1	BYTEDANCE INC	DATA SCIENTIST - TIKTOK ADS	297000	BELLEVUE, WA	2021
2	CITADEL AMERICAS LLC	DATA SCIENTIST	300000	NEW YORK, NY	2019
3	GOODWATER CAPITAL LLC	DATA SCIENTIST	350000	BURLINGAME, CA	2022
4	NETFLIX INC	DATA SCIENTIST	360000	LOS GATOS, CA	2019
5	TAGUP INC	DATA SCIENTIST	870000	SOMERVILLE, MA	2020
6	FACEBOOK INC	DATA SCIENTIST	1250000	MENLO PARK, CA	2020

- The data was placed into a data frame and then ordered by Salary to see if over the years the salary changes
- You can see from the data that the salary range is not a clear progressive increase, the salary ranges vary through the years based on the sort
- The highest salary was at Facebook in 2020



GGPLOT

This breaks the salary down by year showing the count of the salaries by year with the color based on the salary amount. In this you can see the there are outliers in the 2020 data.

Naïve Bayes

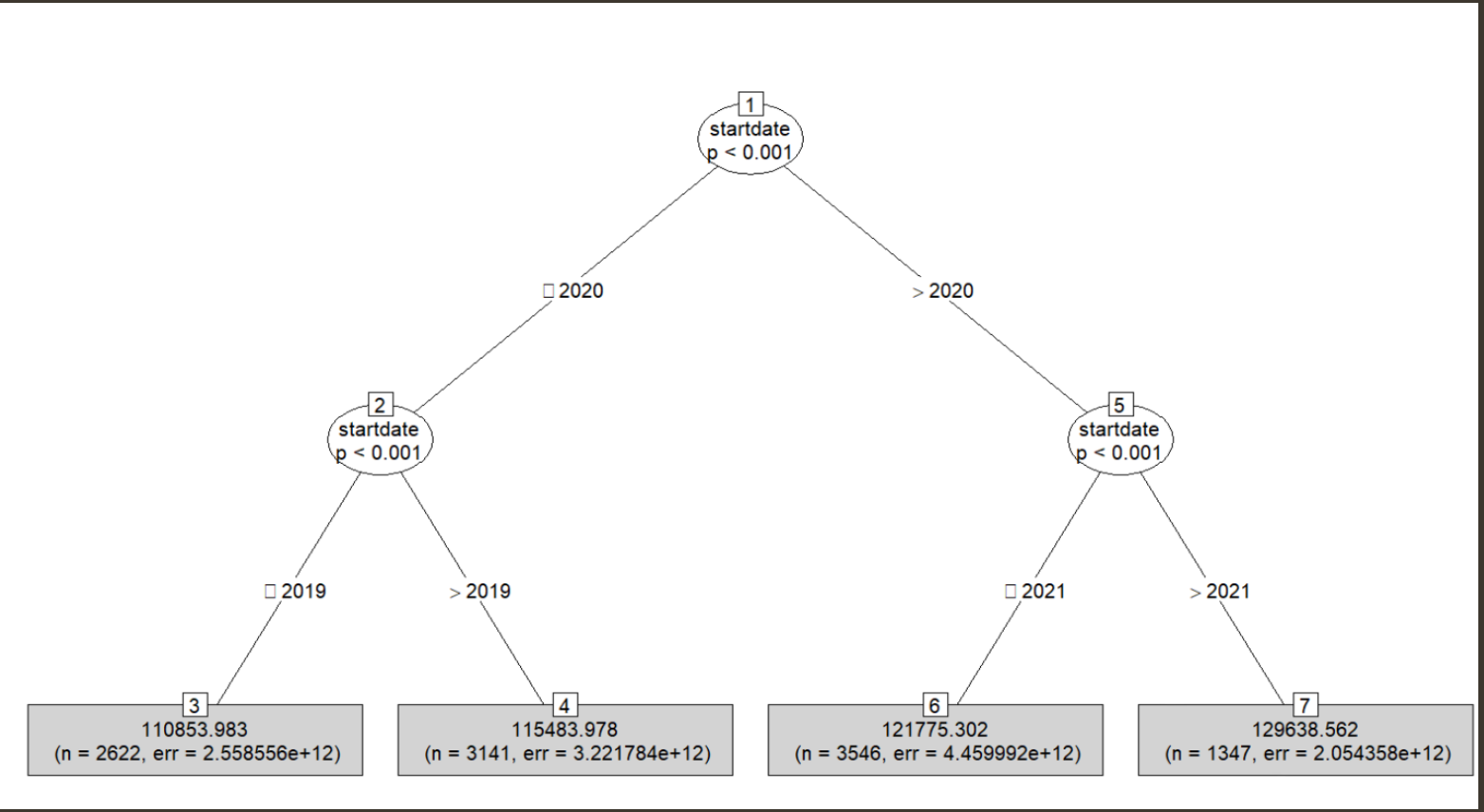
In the results below you can see the bayes predictions of the salaries. Based of the sampling of the data. To the side you can see a sampling of the probabilities for salaries in the sample set as well for the test data after being trained on the training data.

```
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
  50000  2019.000        NA
  50627  2020.000        NA
  51000  2019.000        NA
  51100  2021.000 0.0000000
  51355  2019.000        NA
  52000  2019.200 0.4472136
  53000  2020.000        NA
```

```
  69618  2021.000 0.0000000
  69659  2021.000 0.0000000
  69680  2019.000 0.0000000
  69826  2021.000 0.0000000
  69840  2021.000        NA
  69950  2021.000 0.0000000
  70000  2019.915 0.8961245
  70034  2019.750 0.5000000
  70096  2021.000        NA
  70100  2020.500 0.7071068
  70200  2019.333 0.5773503
  70242  2019.000 0.0000000
  70340  2020.000 0.0000000
  70400  2020.000        NA
  70450  2019.000        NA
  70491  2021.000        NA
  70512  2021.500 0.7071068
  70720  2019.000        NA
  70736  2022.000        NA
  70803  2021.500 0.7071068
  70845  2019.000        NA
  70880  2020.000 0.0000000
```

```
> df_test$pred <- predict(model_naive, newdata = df_test, type = "class")
> df_test$pred_up <- predict(model_naive, newdata = df_test, type = "class")
> head(df_test)
# A tibble: 6 × 7
  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



This decision tree shows 2020 and under and then greater than 2021 and the probability of the pay based on the factor of the year hired. In the end you can see that the highest salary was in the node for over 2020 and then over 2021. with a salary of $129,638.562.

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

- After analysis salaries for Data Scientist progressively increase as the year increase. Initial results show that there is on average a mean increase of around $5500.00 year to year.
- Majority of salaries with the exceptions of a few outliers were within the 100-200K area in all the years combined.
- In predicting salaries the Naive Bayes predicted the salaries higher across all of the sample data set.

Conclusion

- When reviewing what range of salaries to ask for, looking for key words that describe the skills needed can help for negotiation. The salaries offered do range greatly but on average for 2023 using the average increase from 5000-7000 dollars will give you a predicted range of $124,671 - $134,671 using the 2022 mean salary for the starting salary.
- The prediction for the salaries using the Naïve Bayes predicted a range that increases as well as decreases the salaries in the test data. This used the salary, start date and city.
- The randomForest produced a results that the highest change for a higher salary was a position over the year 2021 with a salary mean of $129638.