Final Report: Data Analysis and Modelling for the Louisville Metro Dataset

Submitted To:

Prof. Nasim Mobasheri

CS 418, UIC

Zhu Wang

CS 418

Prepared By:

Tarush Gupta

Pushti Dalal

Akhil Snehal Modi

December 06, 2020

Table of Contents

List of Figures

List of Tables

1. Introduction

This report summarizes all of the primary statistical modeling and analysis results associated with the Louisville Metro Government Department. The purpose of this report is to document both the implemented sampling design and all the corresponding data modelling as well as statistical analysis.

1a. Goal of the Project

This dataset will help us understand the inner workings of different government departments in the metro city of Louisville. The result from data analysis like this will be useful in determining and smoothing the functioning of different government departments. These results can also help in determining important aspects like funding, other government provisions, etc for these departments which ultimately work for the betterment of the people of Louisville. Also, it can also help in determining where the taxes paid by the people of Louisville are being used and are being put to better use or not.

2. Data Section

2a. Dataset

Our dataset contains the information about different government departments of the Louisville Metro Area. The dataset contains the features of CalendarYear, Annual Rate, Regular Rate, Overtime Pay, IncentivesAllowances, and so on for different government departments like Louisville Metro Police, Finance Department, County Attorney, Louisville Fire Department, etc. total of 93435 entries and 11 features.

2b. Features of Dataset

- Calendar Year: Represents the year the data was recorded in.
- Regular Rate: Regular is the salary that an employee is paid for the year.
- Overtime Pay: Overtime pay is the amount that the employee is being paid for any extra time he puts in.

- Incentive Allowance: Incentive Allowance is the extra money an employee makes like bonuses, etc.
- Annual Rate: Annual Rate is the total money an employee makes throughout the year. I.e., it is the summation of regular rate, overtime pay and the incentive allowance.

```
RangeIndex: 93435 entries, 0 to 93434
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   SalaryDataID       93435 non-null  int64
 1   CalendarYear       93435 non-null  int64
 2   EmployeeName       93433 non-null  object
 3   Department         93435 non-null  object
 4   JobTitle           93435 non-null  object
 5   AnnualRate         93435 non-null  float64
 6   RegularRate        93435 non-null  float64
 7   OvertimeRate       93435 non-null  float64
 8   IncentiveAllowance 93435 non-null  float64
 9   Other              93435 non-null  float64
 10  YearToDate         93435 non-null  float64
dtypes: float64(6), int64(2), object(3)
memory usage: 7.8+ MB
```

Figure 1: Dataset Description

3. Methods Section

3a. Source of Dataset

We found our data set from Data.gov which contains all the necessary data in the .csv formatted file that we require for the project.

3b. Reading and Cleaning the Dataset

As mentioned above, our dataset is stored in a .csv file format. So the first thing we do is read the .csv file using pandas dataframe library.

We cleaned the data set by checking all the nan values in the dataset and replacing them by string or int according to the column's dtype using *pandas.fillna() and by* dropping the attributes which were not needed to analyse the dataset such as, "other" using *pandas.drop().*

	SalaryDataID	CalendarYear	EmployeeName	Department	JobTitle	AnnualRate	RegularRate	OvertimeRate	IncentiveAllowance	YearToDate
0	1	2008	Robinson, Charles	Louisville Metro Police	Police Officer	46924.8	46872.02	707.79	5548.32	53412.13
1	2	2008	Schutte, Michael	Louisville Metro Police	Police Officer	46924.8	47257.92	17992.87	8139.04	78883.83
2	3	2008	Doyle, Lisa	Louisville Metro Police	Police Officer	46051.2	46344.32	3771.52	5748.80	57456.64
3	4	2008	Wilkins, Joseph	Louisville Metro Police	Police Officer	46051.2	46385.36	5422.09	6756.48	60771.93
4	5	2010	Locke, John	Louisville Metro Police	Police Officer	47902.4	47954.59	1792.55	6333.28	56780.42
...	...	...	...	...	...	...	...	...	...	...
93430	93431	2010	Bock, Asher	Metro Corrections	Corrections Officer	35089.6	31688.37	167.04	0.00	31855.41
93431	93432	2014	Hughes, Maranda	Metro Corrections	Corrections Officer	32156.8	6926.08	0.00	0.00	6926.08
93432	93433	2009	Wright, Jeremy	Metro Corrections	Corrections Officer	35089.6	32756.23	6672.88	0.00	39429.11
93433	93434	2009	Moore, Vince	Metro Corrections	Corrections Officer	43929.6	44328.32	8670.44	0.00	52998.76
93434	93435	2009	McCoomer, Damon	Metro Corrections	Corrections Officer	40435.2	40070.48	152.81	0.00	40223.29

93435 rows × 10 columns

Figure 2: Cleaned Sample Dataset

3c. Analysis we plan to perform

From the dataset, we would be evaluating the salaries of employees and would also be comparing their salaries depending on the job titles and the positions. From this data set we would be getting the insight of which position, jobs and department are more salarized. We can also analyse the Incentive Allowance depending on the job titles as well as Departments. This type of data analyzing will help us understand the workings of different departments in a city.

4. Data Analysis Section

To begin with, we print out the basic dataset description for all the features of the dataset that we would be using for further data analysis.

	SalaryDataID	CalendarYear	AnnualRate	RegularRate
count	93413.000000	93413.000000	93413.000000	93413.000000
mean	46715.814041	2013.869643	42404.892685	35110.284585
std	26970.552093	3.738136	18152.749683	20654.771773
min	2.000000	2008.000000	1300.000000	-542.100000
25%	23359.000000	2011.000000	32427.200000	21486.590000
50%	46715.000000	2014.000000	42577.600000	36507.230000
75%	70071.000000	2017.000000	52000.000000	48090.750000
max	93435.000000	2020.000000	216000.200000	184752.180000

	OvertimeRate	IncentiveAllowance	YearToDate
count	93413.000000	93413.000000	93413.000000
mean	3740.145868	1913.633608	41770.486301
std	7034.874840	3251.039162	25071.899694
min	-294.120000	-207.150000	-28699.390000
25%	0.000000	0.000000	24783.920000
50%	425.210000	0.000000	42407.000000
75%	3627.340000	1600.000000	58874.890000
max	129395.540000	33278.640000	218157.880000

Figure 3: Basic Dataset Description

4a. Comparing Overall Distribution of Feature

We got the insights of all numeric features like annual rate, overtime rate, regular rate and so on by plotting their overall distribution.

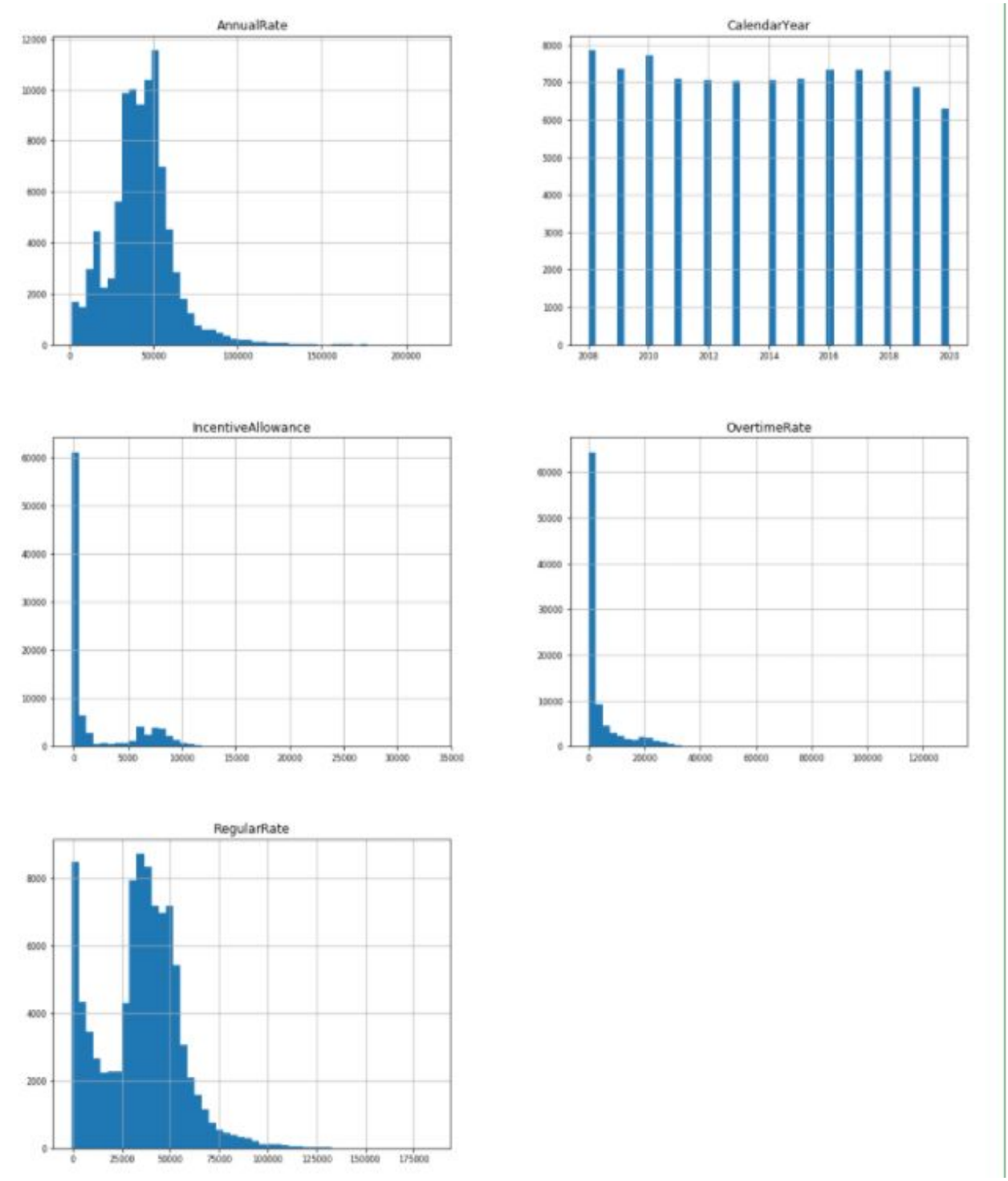

Figure 4: Overall Distribution of Features

After the overall distribution, we compare the features and plot them over the years.

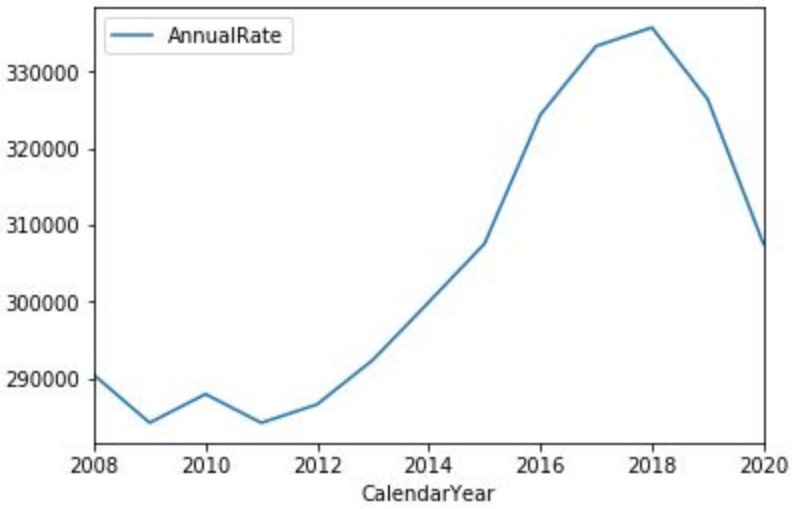

Figure 5: Overall Distribution of Annual Rate over the years

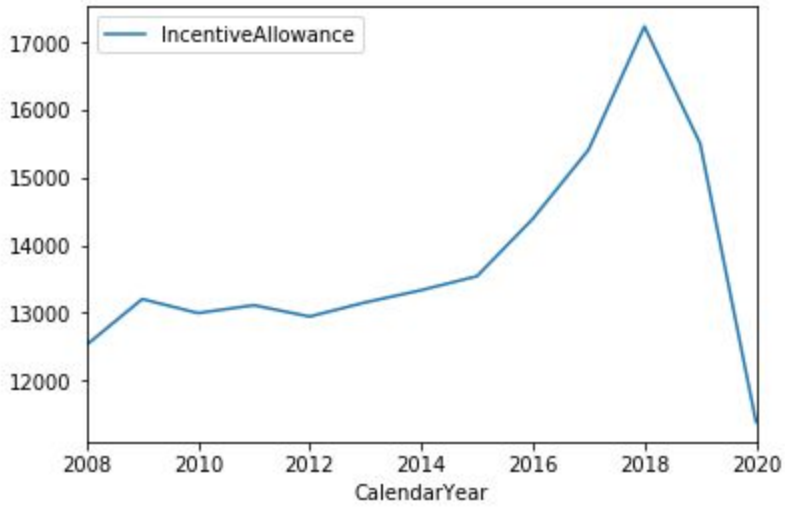

Figure 6: Overall Distribution of Incentive Allowance over the years

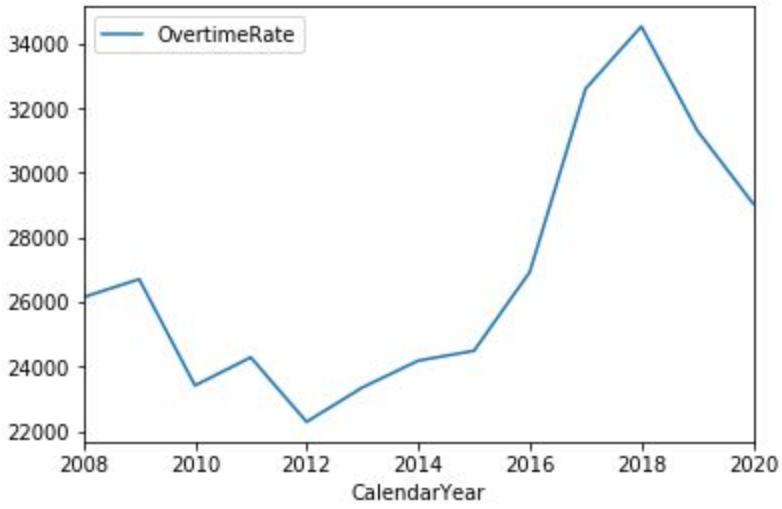

Figure 7: Overall Distribution of Overtime Rate over the years

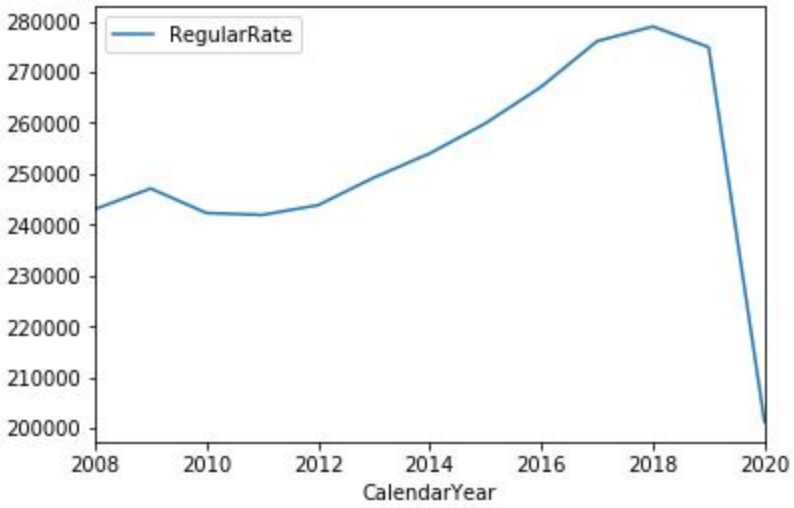

Figure 8: Overall Distribution of Regular Rate over the years

Through the graphs, we were able to gain the following information:

a. There was a steady increase in the Annual Rate, the Regular Rate, and the Incentive Allowance until the year 2018. But then there was a sudden drop in those features in the years 2019 and 2020.

b. For the Overtime Rate, we notice that it initially decreased between the years of 2008 and 2012, but then, like the rest of features, we see a steady growth in the Overtime Rate over the next 6 years i.e., from 2012 to 2018 and sudden drop in the years of 2019 and 2020.

4b. Questions to be answered

1. Which department is the most salarized?

- Looking at Figure 9, we notice that the Louisville Metro Police Department consists of 26.90% of total salary distributed by the Louisville government, hence making it the most salarized.

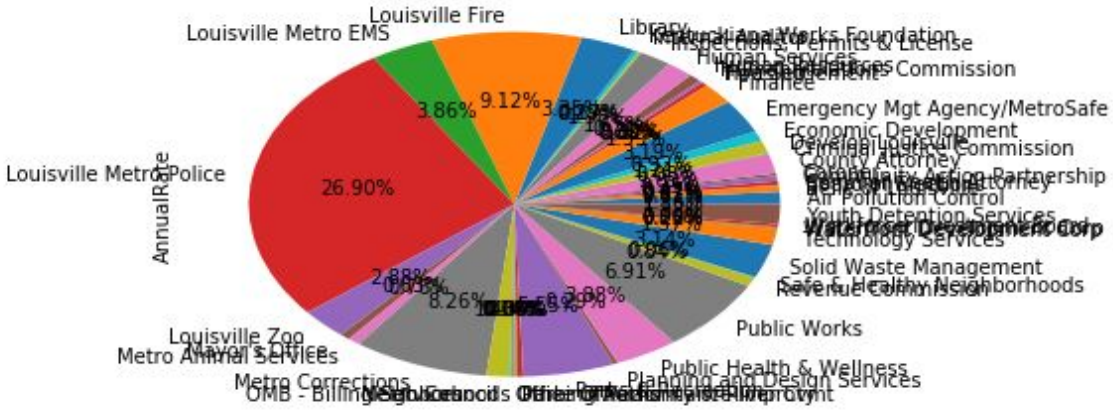

Figure 9: Highest Salaried Department

2. Which Department gets more benefit from other departments?

- Looking at Figure 10, we notice that the Louisville Metro Police Department gets more benefits from other departments. About 66.55% of the Incentives are allocated to the Louisville Metro Police Department.

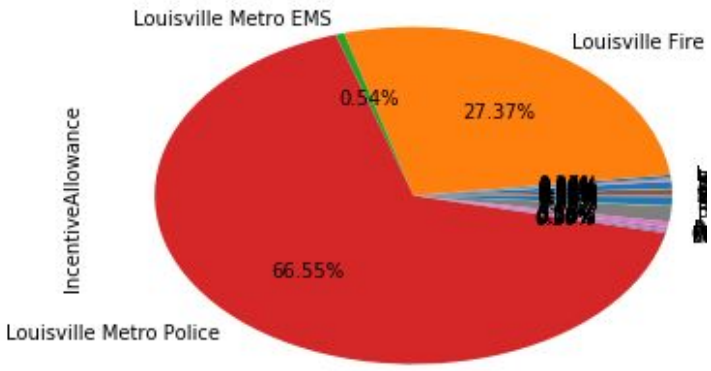

Figure 10: Department with the most incentives

3. Which department is being preferred by most of the employees and why it is being preferred?

- Out of all the government employees in the Louisville Metro Area, 22.93% employees prefer working in the Louisville Metro Police Department because from the previous analysis, we saw that the Police Department is the most salarized as well as getting the maximum benefits than any other departments.

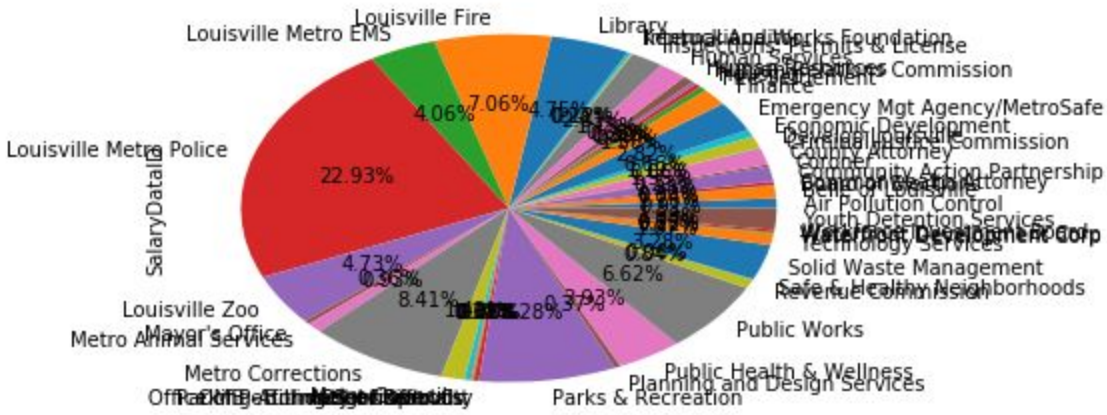

Figure 11: Most preferred Department

4c. Some more data analysis

1. Analyzing Regular Rate over the Calendar Years for each Department

- By plotting the graph of Regular rate over calendar years for each department

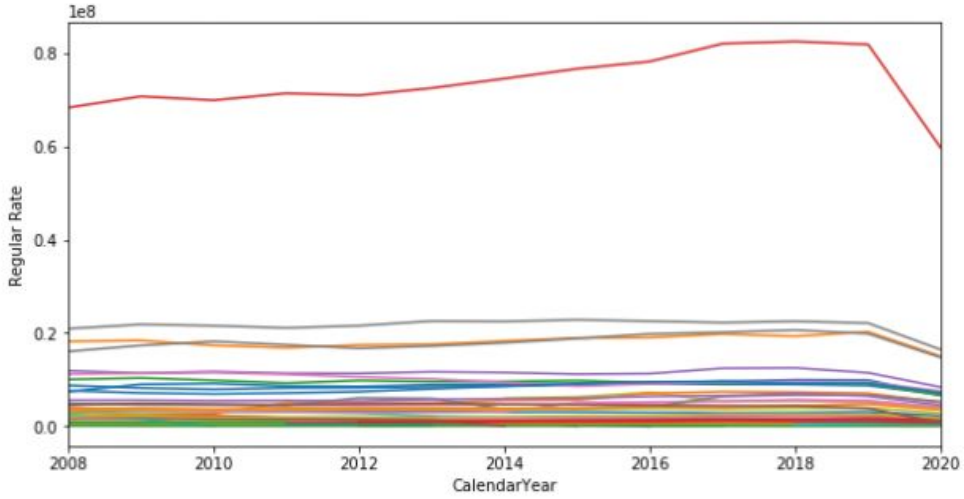

Figure 12: Regular Rate over Calendar Years for each Dept.

We notice that the regular rate is maximum in the department of Police and has been increasing over the years especially we saw a gradual increase in overtime rate in the years from 2016 to 2018 while overtime rate in other departments is decreasing over the years.

2. Analyzing Overtime Rate over the Calendar Years for each Department

- By plotting the graph of Regular rate over calendar years for each department

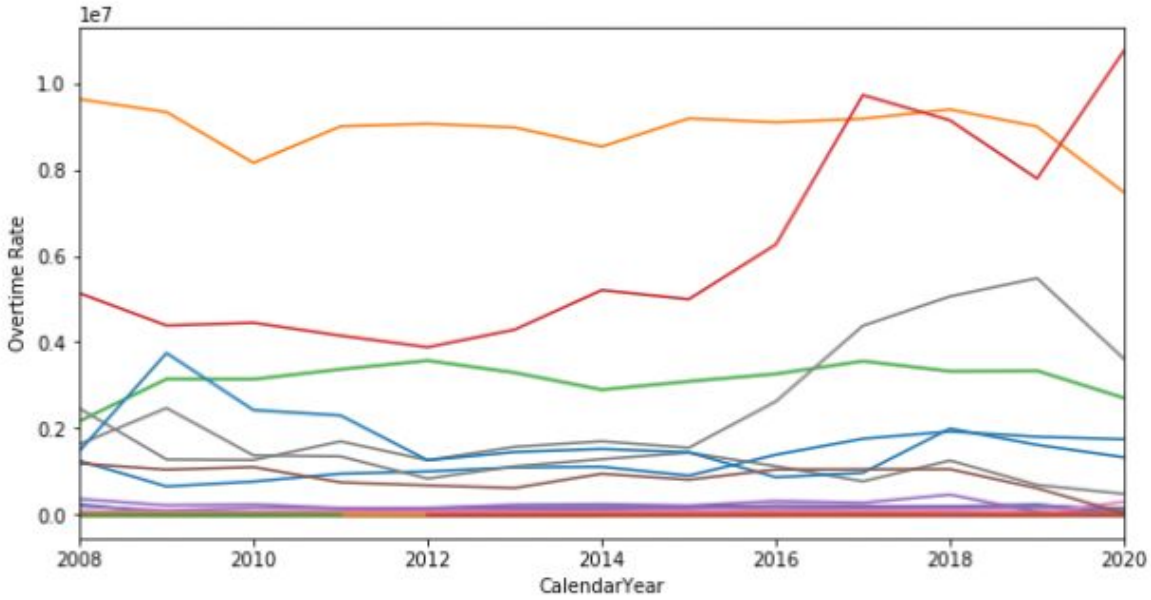

Figure 13: Overtime Rate over Calendar Years for each Dept.

We notice that the regular rate is maximum in the department of Police and has been increasing over the years especially we saw a gradual increase in overtime rate in the years from 2016 to 2018 while overtime rate in other departments is decreasing over the years.

3. Establishing a Correlation between the features

We established a correlation of all the other features with the Annual rate of each department. Doing this, gave us a deep insight into how the regular rate, overtime rate and incentive allowance are affected by the annual rate of each department.

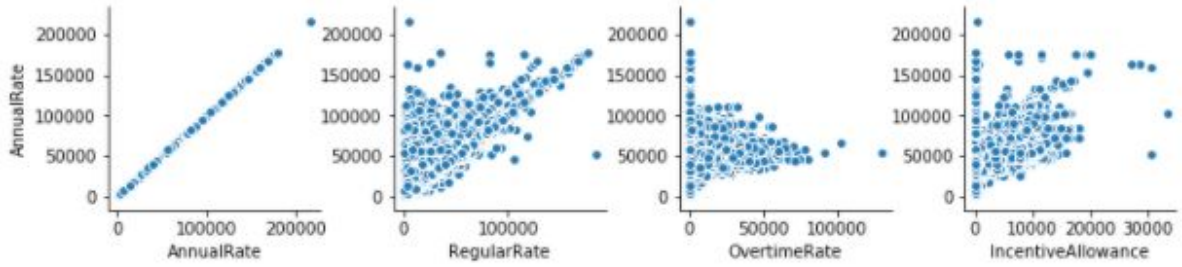

Figure 14: Correlation of other Features with Annual Rate

4d. Linear Regression

1. Means Squared Error

- We print Mean squared errors for each feature in the dataset to get an insight about features that are best to be considered in the prediction model.

```
SalaryDataID
Mean squared error: 674251275.57
CalendarYear
Mean squared error: 12.60
AnnualRate
Mean squared error: 87784596.40
RegularRate
Mean squared error: 17003590.90
OvertimeRate
Mean squared error: 11777643.06
IncentiveAllowance
Mean squared error: 4373968.99
YearToDate
Mean squared error: 18391977.83
[679407768.98, 12.54, 86047191.39, 17079127.86, 11583304.63, 4467842.73, 176274
31.23, 18391977.833168447]
```

Figure 15: Means Squared Errors

2. Predictions

2a. Predicting Annual Rate

- Taking all the numeric features, we fitted a linear regression model to predict annual rate for the upcoming 10 years which is from year of 2021 to 2030 and perceive that the annual rate is increasing steadily.

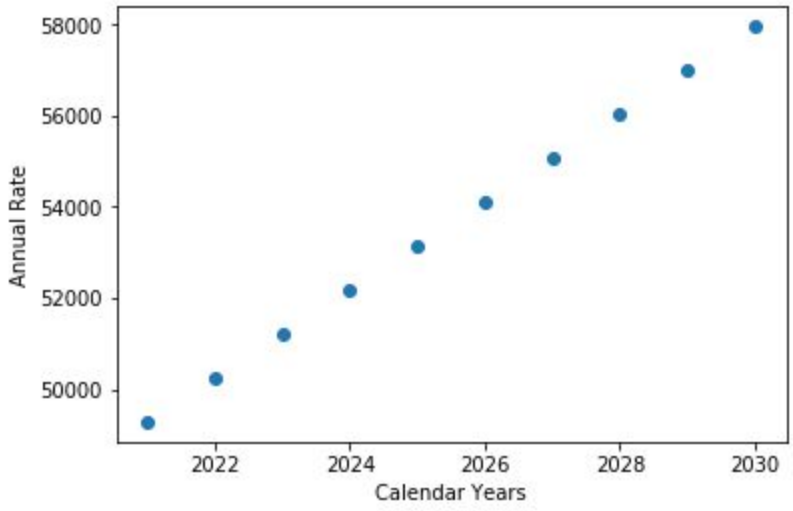

Figure 16: Annual Rate over the next 10 years

- The purpose of doing this was to differentiate how the predicted annual rate differs from the actual annual rate and to get an idea about how the pandemic has affected the annual rate of different government departments.

2b. Predicting Overtime Rate

- We took 5 different annual rates that we predicted in the previous graph, and fitted a linear regression model to predict overtime rate that can be expected over those annual rates

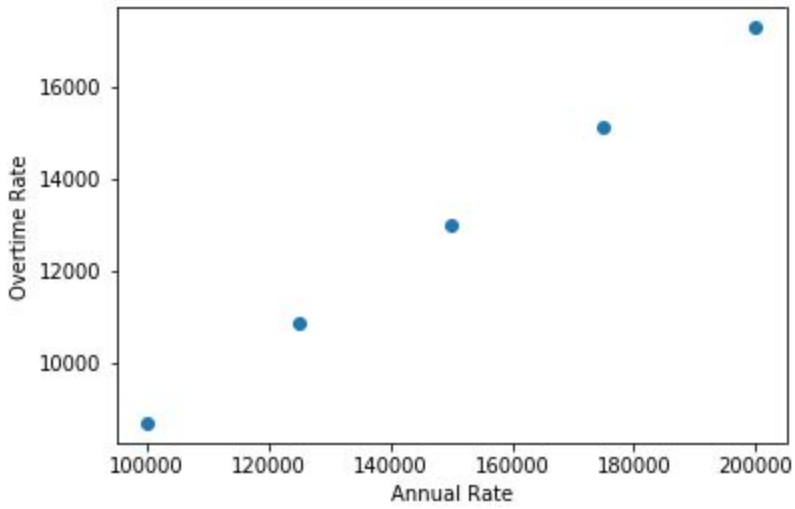

Figure 17: Overtime Rate over different Annual Rates

2c. Predicting Incentive Allowances

- We took 5 different annual rates that we predicted in the previous graph, and fitted a linear regression model to predict Incentives that can be expected over those annual rates

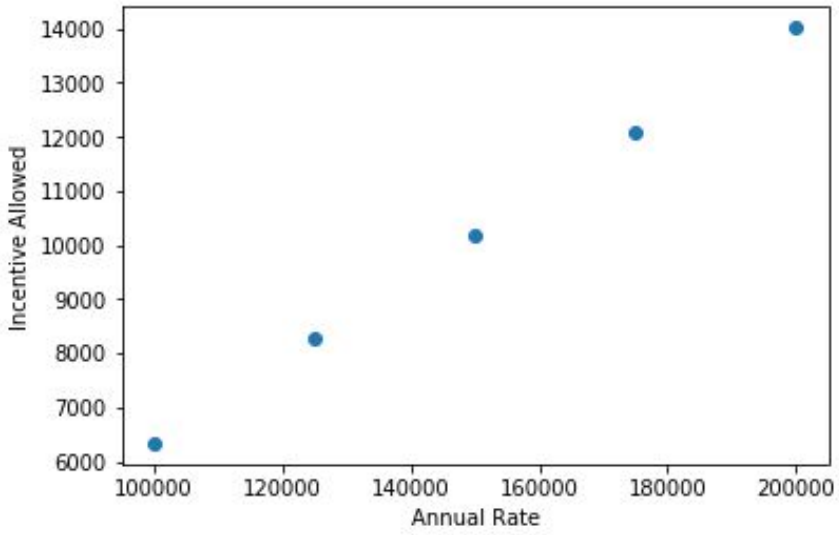

Figure 18: Incentives Allowed over different Annual Rates

3. Data Modelling

3a. Labels

- We made a new column and named it “predict” and copied all the data from the annual rate column into the new column. Then, we classified the data into three different categories i.e., Annual Rate ranging below 35,000, second range, Annual rate ranging between 35,000 - 45,000 and the third ranges beyond 45,000. Then we used label encoding over the predict column which classified it in class 0, class 1 and class 2.
- The reason to classify the dataset into these three specific classes was to get an even distribution of data points in all the three classes which was determined using the mean and standard deviation of the annual rates

3b. Data Scaling and Splitting

We make a dataframe using columns from the original dataset which are either float 64 data type for int 64 data type and then we split it into 80% training data and 20% testing data.

- We used standard scaler to eliminate the outliners and to make the dataset normalized
- We fit the different models to predict the ranges of Annual Rates using the features of Regular Rate and Overtime Rate, as these features had the least mean squared errors.

3c. Models

We used 5 different models and compared the precision and accuracy rates between them.

- **Model 1: KNeighborsClassifier**

	precision	recall	f1-score	support
0	0.76	0.76	0.76	17
1	0.92	0.95	0.94	152
2	0.85	0.71	0.77	31
avg / total	0.90	0.90	0.90	200

```
Accuracy 0.9
[[ 13   3   1]
 [  4 145   3]
 [  0   9  22]]
```

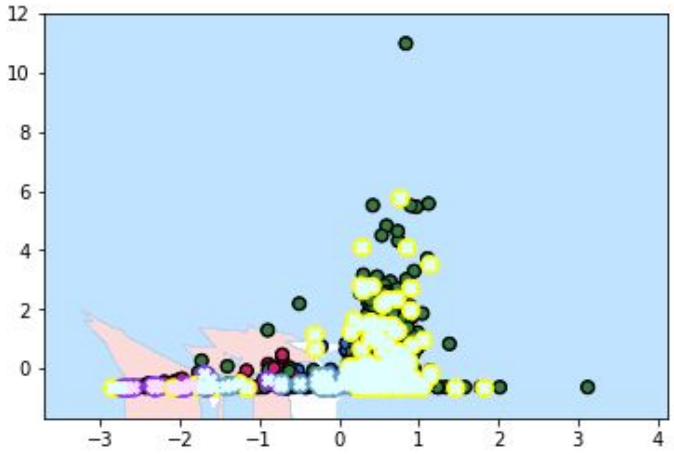

Figure 19: Decision boundary using KneighborsClassifier Model

- **Model 2: Linear SVM Model**

	precision	recall	f1-score	support
0	0.48	0.88	0.62	17
1	0.85	0.94	0.89	152
2	0.00	0.00	0.00	31
avg / total	0.68	0.79	0.73	200

```
Accuracy 0.79
[[ 15   2   0]
 [  9 143   0]
 [  7  24   0]]
```

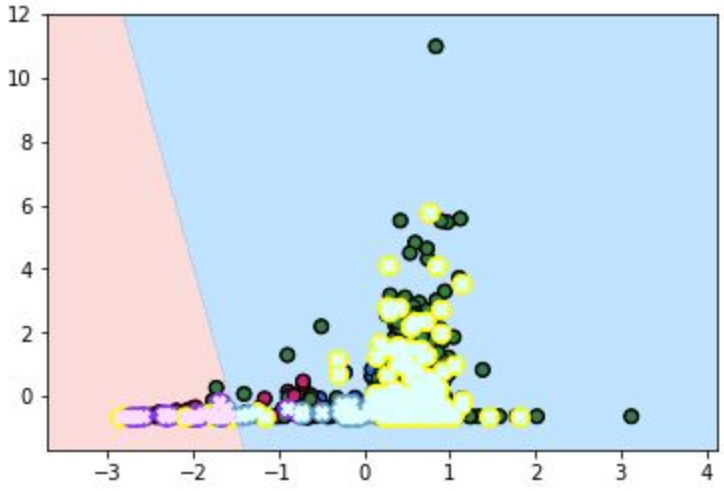

Figure 20: Decision boundary using Linear SVM model

- **Model 3: Non-Linear SVM Model**

	precision	recall	f1-score	support
0	0.48	0.88	0.62	17
1	0.95	0.93	0.94	152
2	0.84	0.52	0.64	31
avg / total	0.89	0.86	0.87	200

```
Accuracy 0.865
[[ 15   0   2]
 [  9 142   1]
 [  7   8  16]]
```

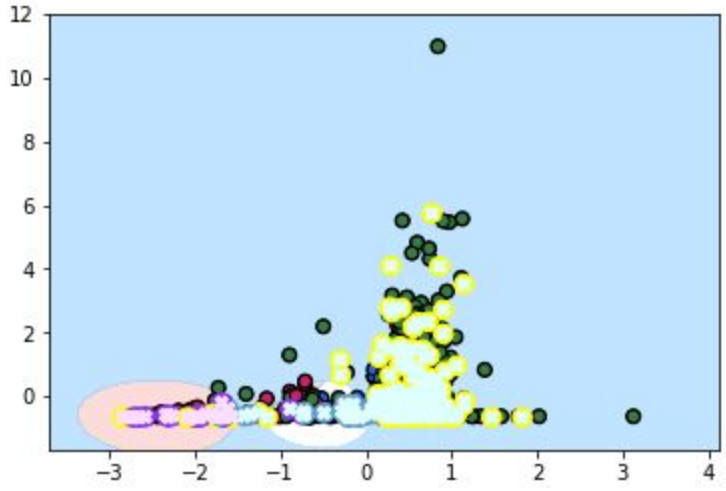

Figure 21: Decision boundary using Non-linear SVM Model

- **Model 4: Decision Tree Model**

	precision	recall	f1-score	support
0	0.69	0.65	0.67	17
1	0.90	0.95	0.92	152
2	0.83	0.65	0.73	31
avg / total	0.87	0.88	0.87	200

```
Accuracy 0.875
[[ 11   6   0]
 [  4 144   4]
 [  1  10  20]]
```

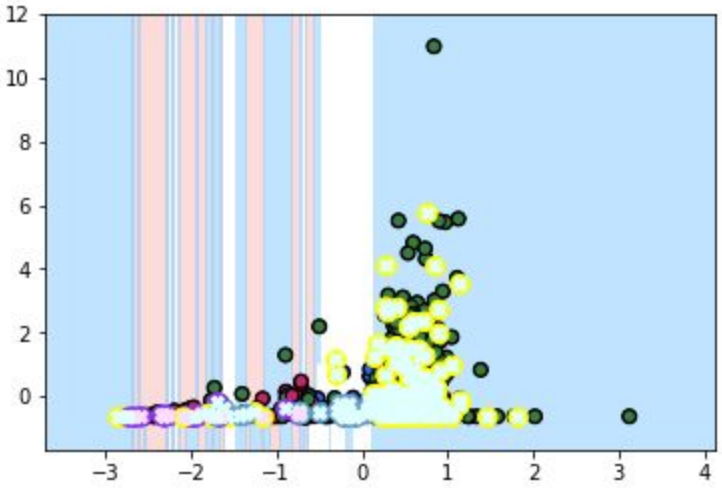

Figure 22: Decision boundary using Decision Tree Model

- **Model 5: Naive-Bayes Classifier**

	precision	recall	f1-score	support
0	0.45	0.88	0.60	17
1	0.89	0.93	0.91	152
2	0.56	0.16	0.25	31
avg / total	0.80	0.81	0.78	200

```
Accuracy 0.805
[[ 15   0   2]
 [  9 141   2]
 [  9  17   5]]
```

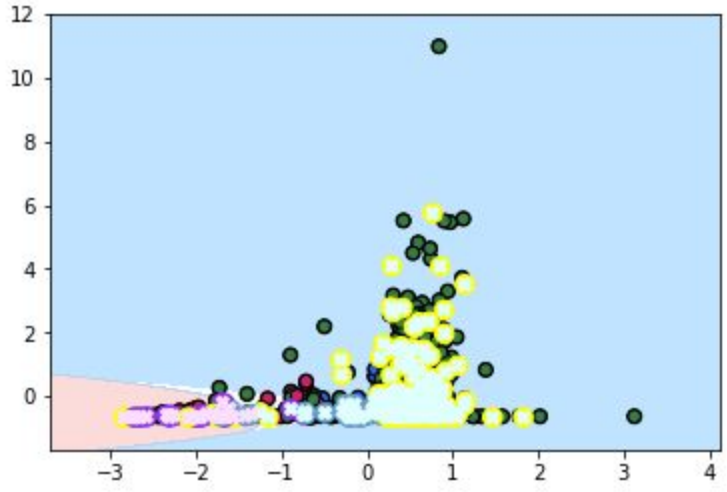

Figure 23: Decision boundary using Naive-Bayes Classifier

- **Comparing Different Models**

Models	Precision			Accuracy
	Class 0	Class 1	Class 2	
KNN	0.76	0.92	0.85	0.900
Linear SVM	0.48	0.85	0.00	0.790
Non-Linear SVM	0.48	0.95	0.84	0.865
Decision Tree	0.69	0.90	0.83	0.875
Naive Bayes	0.45	0.89	0.56	0.805

Table 1: Comparing different models

3d. Further Data Modelling

- According to the statistics of precision and accuracy score, we saw KNN is the most accurate model that we can use for our further predictions.
- Hence, we use KNN to make further decision boundaries using different features like Regular Rate v/s Overtime Rate, Regular Rate v/s Incentive Allowance and Overtime Rate v/s Incentive Allowance.

- Regular Rate v/s Overtime Rate

	precision	recall	f1-score	support
0	0.76	0.76	0.76	17
1	0.92	0.95	0.94	152
2	0.85	0.71	0.77	31
avg / total	0.90	0.90	0.90	200

```
Accuracy 0.9
[[ 13   3   1]
 [  4 145   3]
 [  0   9  22]]
```

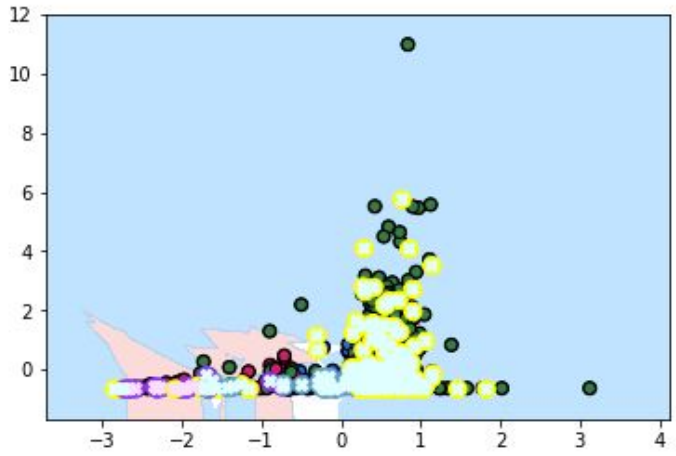

Figure 24: Regular Rate v/s Overtime Rate Decision Boundary

- **Regular Rate v/s Incentive Allowance**

	precision	recall	f1-score	support
0	0.94	0.94	0.94	17
1	0.97	0.97	0.97	152
2	0.90	0.87	0.89	31
avg / total	0.95	0.95	0.95	200

```
Accuracy 0.955
[[ 16   1   0]
 [  1 148   3]
 [  0   4  27]]
```

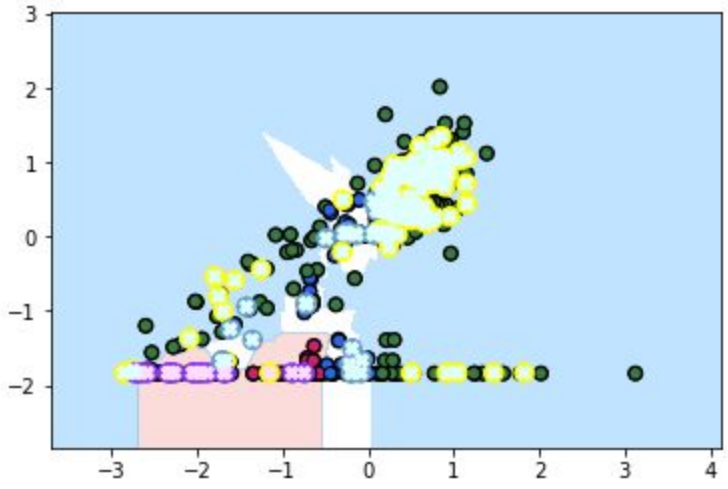

Figure 25: Regular Rate v/s Incentive Allowance Decision Boundary

- **Overtime Rate v/s Incentive Allowance**

	precision	recall	f1-score	support
0	1.00	0.41	0.58	17
1	0.87	0.95	0.91	152
2	0.70	0.61	0.66	31
avg / total	0.85	0.85	0.84	200

```
Accuracy 0.85
[[  7  10   0]
 [  0 144   8]
 [  0  12  19]]
```

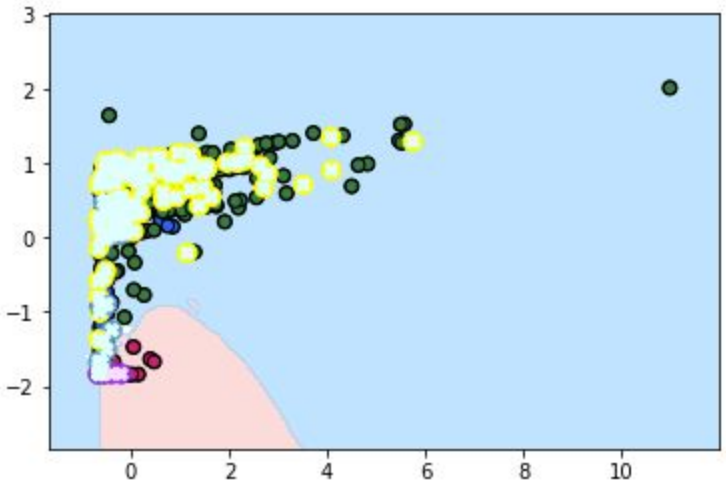

Figure 26: Incentive Allowance v/s Overtime Rate Decision Boundary

	Class 0	Class 1	Class 2	Accuracy
RegularRate and OvertimeRate	0.76	0.92	0.85	0.90
RegularRate and IncentiveAllowance	0.94	0.97	0.90	0.95
OvertimeRate and IncentiveAllowance	1.00	0.87	0. 70	0.85

Table 2: Modelling different features using KNN

3d. Clustering

We used clustering as we had a big dataset which was not well structured. By implementing clusters we were able to narrow down our conclusions.

- We clustered our Ranges of Annual rate over Annual rate and regular rate and we got the homogeneity score and completeness score to be of 0.4939 and 0.4923 respectively for the Kmean model with 3 clusters and 2 features.

```
Homogeneity Score: 0.493972189853
Completeness Score: 0.492320020614
```

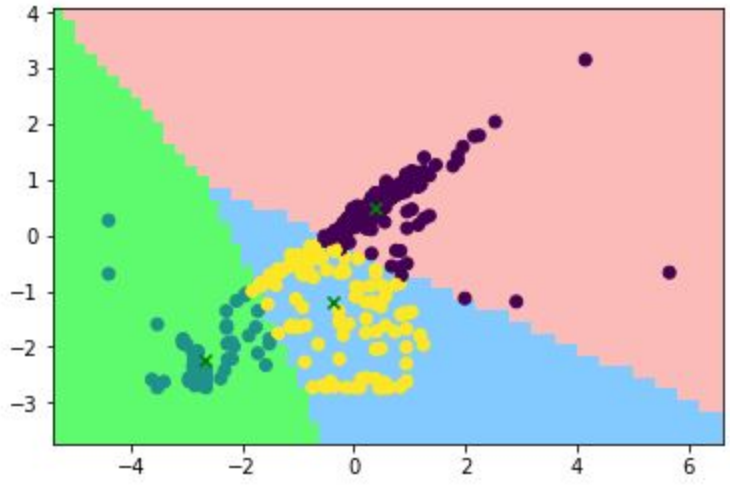

Figure 27: Ranges of Annual Rate over Annual Rate and Regular Rate

- We clustered our ranges of Annual rate over the calendar year and Regular rate and got the homogeneity score and completeness score of 0.4235 and 0.3872 respectively.
- Here, the employee salary would be the most clustered in the range of 35k and 45k in the years 2008, 2012 and 2016.

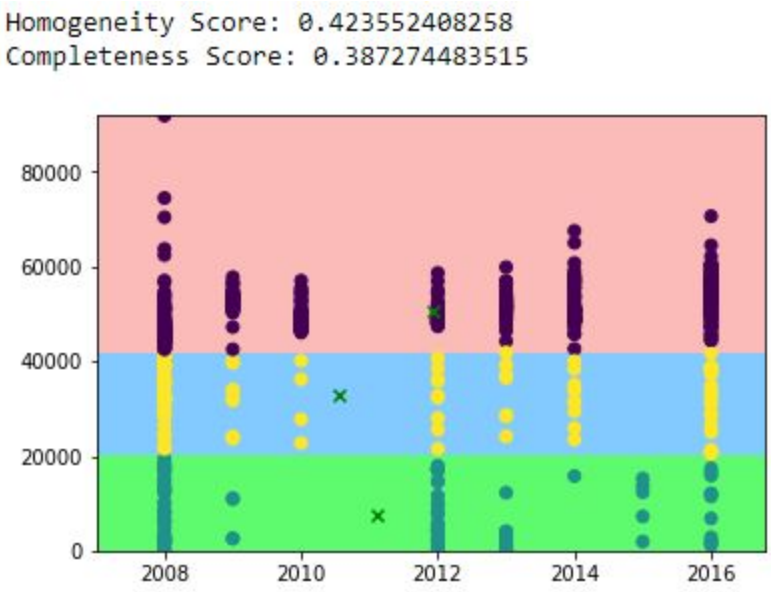

Figure 28: Ranges of Annual Rate over Regular Rate and Calendar Years

5. Challenges

The major challenge we faced was to plot a clean and readable chart because we had a large number of attributes to comprehend in the charts. Moreover, we had a large data-set comprising about 90,000 entries and hence, we used about 10,000 entries for our analysis.

6. Results

To summarize our analysis, we were able to evaluate the salaries of employees by comparing their salaries depending on the job titles and the positions. We found out that the AnnualRate was highest around the year 2018. It started to increase by the year 2013 and then it dipped down again in 2020 major cause of this would surely be the ongoing pandemic. Similarly incentive allowance was also highest around the year 2018 and which is currently dipping down due to loss of job opportunities in the year 2020. Overtime rate and regular rate followed a similar trend, as the business and economy

coming back to normal this downward trend could possibly see an upward curve in the coming years.
We were also able to reach a consensus that the Louisville Metro Police Department is the most salarized department covering around 26.90% of the total.
Through our analysis we were able to conclude that the Louisville Metro Police Department received around 67% of all allocated incentives.
We came to a conclusion that the Louisville Metro Area, 22.93% employees prefer working in the Louisville Metro Police Department because the Police Department is the most salarized as well as getting the maximum benefits than any other departments.
In our analysis we were able to differentiate between the regular rate and the overtime rate. We found out that the regular rate is the maximum in the department of Police during the years 2016 to 2018 whereas the overtime rate in the other departments decreased over the years.
We were able to predict how the pandemic has affected the annual rate of different government departments and how it would look in the course of upcoming years. By using 5 different models we were able to compare the precision and accuracy rates for different departments.

7. Conclusion

To conclude, we were successfully able to implement the sampling design and were able to perform the corresponding data modelling as well as the statistical analysis which helped us to come across various insightful details and answers to various integral questions which we planned to get through this statistical analysis. Our main goal was to understand the inner workings of different government departments in the metro city of Louisville through statistical modelling. We were able to register some key facts and information which could be used in future to a greater extent and could be further used to determine the smooth functioning of different government departments.
Now, through our graphical analysis we can get a better understanding of the annual rate, regular rate and the incentive allowance in the coming years depending on the trend from the past. If you are looking for a government job in or around the city of Louisville then the Louisville Metro Police Department is one of the most salarized departments to be considered. Over time rate is some of the factors that both an employee and an employer keeps an eye on and surely Louisville Metro Police Department serves as one of the best when it comes to Over time rate.
KNN classifier gave us the best precision and accuracy score to enhance our data modeling analysis further. Homogeneity scores and completeness scores were the key

points for this aspect of the modeling which further enhanced our overall analysis for the Louisville Metro Department.