Louisville Government Employee Salary Analysis

Authors

Karthik S. Iyer

Vidit R. Sheth

School of Engineering and Applied Sciences, George Washington University

EMSE 6574 : Programming for Analytics

Prof. Michael Rossetti

December 19, 2023

TABLE OF CONTENTS

INTRODUCTION

In the domain of public administration, understanding government employee compensation holds significance in understanding the efficiency of the socio-economic and political standpoints of the government. We are looking at the real-time data available over a period of the last 5 years to understand the compensation structure for government employees of Louisville.

Starting with a data-driven investigation of government employee data in Louisville, this project combines advanced analytics with labour dynamics in the public sector. The dataset provides an extensive exploration of the complexities of pay structures and employment trends in government agencies. It is a veritable gold mine of data that includes Annual Rate, Regular Rate, Incentive Allowance, Overtime Rate, and Year-to-Date (YTD) Totals.

Beginning with a thorough Exploratory Data Analysis (EDA), the project uses statistical and visual methods to identify underlying trends, correlations, and possible outliers. This procedure, which makes use of Python and data visualization tools, reveals the subtleties of compensatory dynamics and lays the groundwork for more in-depth research.

The notion of intricate analysis is introduced by separating the dataset by departments, which uses Pandas dataframes to derive department-specific insights. This segmentation could be enhanced by multivariate statistical methods and regression techniques and decision trees are used in the path from descriptive statistics to predictive modelling. By incorporating factors like YTD Totals and departmental affiliations, machine learning libraries like scikit-learn play a crucial role in training models to estimate compensation changes or predict total profits.

The goal of this project is to go beyond data exploration by utilizing feature engineering techniques to extract new variables that improve prediction accuracy. The Python ecosystem includes scikit-learn, NumPy, and Pandas, serves as the toolkit for modelling, evaluating, and modifying the complex compensation landscape.

This report aims to shed light on the trends in the government compensation plans. Furthermore, the report will not only shed light on the spending capacity of the government for Louisville but will also help to understand the growth or decrease in compensation for employees over the 5 - year period from 2018 – 2023.

METHODOLOGY

The methodology used to analyse the government employee dataset in Louisville is based on a thorough and methodical approach to glean valuable insights from the abundance of available data.

Data Collection

The following steps are involved in this process:

- Defining our API URL where we are going to fetch data from. Upon going through documentation, we know that the API is open-source and doesn't have a separate API-Key for accessing the basic employee salary compensation data.
- We use data scrapping techniques to fetch the required data from the 'Louisville Metro Employee Salary Data[1]'
- We fetch the data in Json style and parse through it using the 'json' library.
- The data is stored into a data frame using the 'pandas' library for the next stage.

```
response = requests.get(url)
data = response.json()
#print(data)
```

Figure 1.1: Fetching data from API.

```
pd.json_normalize(data['features'])
```

Figure 1.2: Generating Data frame from API response.

We can fetch the data for the top 1000 entries from the API as seen in Figure 1.3. We replicate this process as shown in Figures 1.1 and 1.2 by packaging them into functions and looping thoroughly through the API data for the years 2018 through 2023. This helps us generate a dataset with 41030 entries as seen in Figure 1.4 for all the government employees from the year 2018 to 2023.

	attributes.CalYear	attributes.Employee_Name	attributes.Department	attributes.jobTitle	attributes.Annual_Rate	attributes.Regular_Rate
0	2018	Knop, Paul	Louisville Fire	Fire Prevention Inspector I	53930.24	38373.44
1	2018	Jones, Steven	Inspections, Permits & License	PR/B/M Inspector II	47278.40	46325.87
2	2018	Sheehan, Thomas	Louisville Fire	Fire Prevention Inspector II	61588.80	2368.80
3	2018	Martin, David	Library	Library Page L/U	27331.20	26936.96
4	2018	Bratcher, Elaine	Louisville Metro Police	Clerk Typist II-Police	34902.40	34532.80
...	...	...	...	...	...	...
995	2018	Kuchenbrod, Wayne	Parks & Recreation	Staff Assistant	10712.00	10061.11
996	2018	Watson, Lowell	Louisville Metro Police	Police Sergeant	70241.60	70234.56
997	2018	Bromback, Stephanie	Louisville Zoo	Events Coordinator	48318.40	47761.41
998	2018	Whitis, Jeffrey	Louisville Fire	Fire Apparatus Operator 56 Hr	56429.57	35316.49
999	2018	Thomas, Ali	Louisville Fire	Fire Company Commander 56 Hr	65581.57	40988.48

1000 rows × 11 columns

Figure 1.3: First API call data

	CalYear	Employee_Name	Department	jobTitle	Annual_Rate	Regular_Rate	Overtime_Rate	Incentive_Allowance	Other	YTD_Total	ObjectId
0	2018	Knop, Paul	Louisville Fire	Fire Prevention Inspector I	53930.24	38373.44	0.0	10807.55	None	49241.40	20288
1	2018	Jones, Steven	Inspections, Permits & License	PR/B/M Inspector II	47278.40	46325.87	0.0	0.00	None	47325.87	20292
2	2018	Sheehan, Thomas	Louisville Fire	Fire Prevention Inspector II	61588.80	2368.80	0.0	280.08	None	10999.49	20296
3	2018	Martin, David	Library	Library Page L/U	27331.20	26936.96	0.0	700.00	None	27636.96	20300
4	2018	Bratcher, Elaine	Louisville Metro Police	Clerk Typist II-Police	34902.40	34532.80	0.0	1563.12	None	36498.64	20304
...	...	...	...	...	...	...	...	...	...	...	...
41025	2023	Adams, Bradley M.	Louisville Metro Police Department	Police Recruit	46280.00	6230.00	0.0	0.00	None	6230.00	7196
41026	2023	Webb, Garland H	Louisville Metro Police Department	Police Recruit	46280.00	6230.00	0.0	0.00	None	6230.00	7197
41027	2023	Timpanaro, Keegan James	Louisville Metro Police Department	Police Recruit	46280.00	6230.00	0.0	0.00	None	6230.00	7198
41028	2023	Seago, Dylan	Louisville Metro Police Department	Police Recruit	46280.00	6230.00	0.0	0.00	None	6230.00	7199
41029	2023	Hutchins, Noah Michael	Louisville Metro Police Department	Police Recruit	46280.00	6230.00	0.0	0.00	None	6230.00	7200

41030 rows × 11 columns

Figure 1.4: Dataset for Louisville Government Employee Salary

The data consist of the following information:

- **CalYear**: Year when employee data was collected (Type: Long integer)
- **Employee_Name**: Name of the employee (Type: Character string)
- **Department**: Sector of the government where the employee was working (Type: Character string)
- **jobTitle**: The role/title of the employee (Type: Character string)
- **Annual_Rate**: Salary of the employee including bonuses, raises, benefits and base salary annually (Type: Double)
- **Regular_Rate**: Base salary of employee annually(Type: Double)
- **Overtime_Rate**: Additional hourly pay for employee for any overtime(Type: Double)
- **Incentive_Allowance**: Incentive amounts additional to base pay (Type: Double)
- **Other**: Any other bonuses, raises etc applicable to base salary (Type: Double)
- **YTD_Total**: Year to Date total amount paid to employee (Type: Double)
- **ObjectId**: Unique ID of all data points in dataset (Type: Long Integer)

Data Cleaning

Before we begin the process of analysing our data, we must first ensure that our data is clean. Data cleaning is an important step in the data analysis process that involves identifying errors and inconsistencies and inaccuracies within the dataset to ensure its usability. This meticulous process plays a crucial role in preparing data for meaningful analysis and interpretation. Here are key aspects and considerations in our data cleaning:

1. Handling Null and/or Missing Values

- The salaries dataset consists of numerous columns for different interpretations of salary compensation. We observe that the 'Other' column specifically, has huge number of 'Nan' values and will contribute nothing in our data analysis hence we take the decision to drop the entire column.

- We also observe that there remain few entries with no employee names. Since the number of such entries is minimal, we choose to omit these entries entirely from our data set.

2. Dealing with Duplicates

- Identifying and removing duplicate records ensures that each observation in the dataset is unique. This is particularly important when the dataset includes individual entities, such as employees, and duplicates could distort analyses.

- We hence look at our dataset and observe that although employee names repeat, these entries are not the same since there is a change in the Year and similarly different data is reflected in their compensations. Hence the are no duplicates.

3. Correcting Inconsistencies

- Addressing inconsistencies in data involves standardizing units of measurement. This ensures data consistency and accuracy. Normalization of data is a widely adopted method to do this.

- We normalize our calendar year column using the pd.datetime() function to ensure continuity and removal of inconsistencies.

Data cleaning is an ongoing process that requires a balance between thoroughness and practicality. It lays the foundation for robust and reliable analyses, enabling data scientists and analysts to draw meaningful insights and make informed decisions.

```
1  df.describe()
```

	CalYear	Annual_Rate	Regular_Rate	Overtime_Rate	Incentive_Allowance	Other	YTD_Total	ObjectId
count	40992.000000	40992.000000	40992.000000	40992.000000	40992.000000	0.0	40992.000000	40992.000000
mean	2020.474312	50491.170006	41324.779866	5812.400912	2370.439101	NaN	50608.547087	20496.500000
std	1.730951	21661.249854	24666.451928	10577.028665	4006.093649	NaN	31715.316283	11833.515454
min	2018.000000	0.000000	-320.640000	-282.600000	-2500.000000	NaN	-320.640000	1.000000
25%	2019.000000	37814.400000	24455.587500	0.000000	0.000000	NaN	27836.027500	10248.750000
50%	2020.000000	49504.000000	42238.300000	682.845000	0.000000	NaN	49850.135000	20496.500000
75%	2022.000000	61052.160000	55427.620000	6579.820000	3575.000000	NaN	70426.170000	30744.250000
max	2023.000000	238277.000000	235572.270000	129395.540000	43853.330000	NaN	262051.220000	40992.000000

Figure 2.1: Dataset Statistical Description

```
[7]  1  df.isnull().sum()

CalYear                0
Employee_Name          8
Department             0
jobTitle               0
Annual_Rate            0
Regular_Rate           0
Overtime_Rate          0
Incentive_Allowance    0
Other                  40992
YTD_Total              0
ObjectId               0
```

Figure 2.2 Sum of null values in our data

- Additionally, we introduce new columns into our dataset i.e. Employee ID which acts as a Unique Identifier for our analysis further down the road.

Exploratory Data Analysis

With this we are now able to perform our analysis on the Dataset.
We take a glance at the change in the Annual Rate, Regular Rate and the YTD totals for the employee in the 5-year period as shown in Figure 2.3.
We can observe that the annual rate is consistent with increase, but the YTD total fluctuates as per the Regular rate every year. This indicates that there may be a strong correlation between the Regular Rate and the YTD total for each employee.

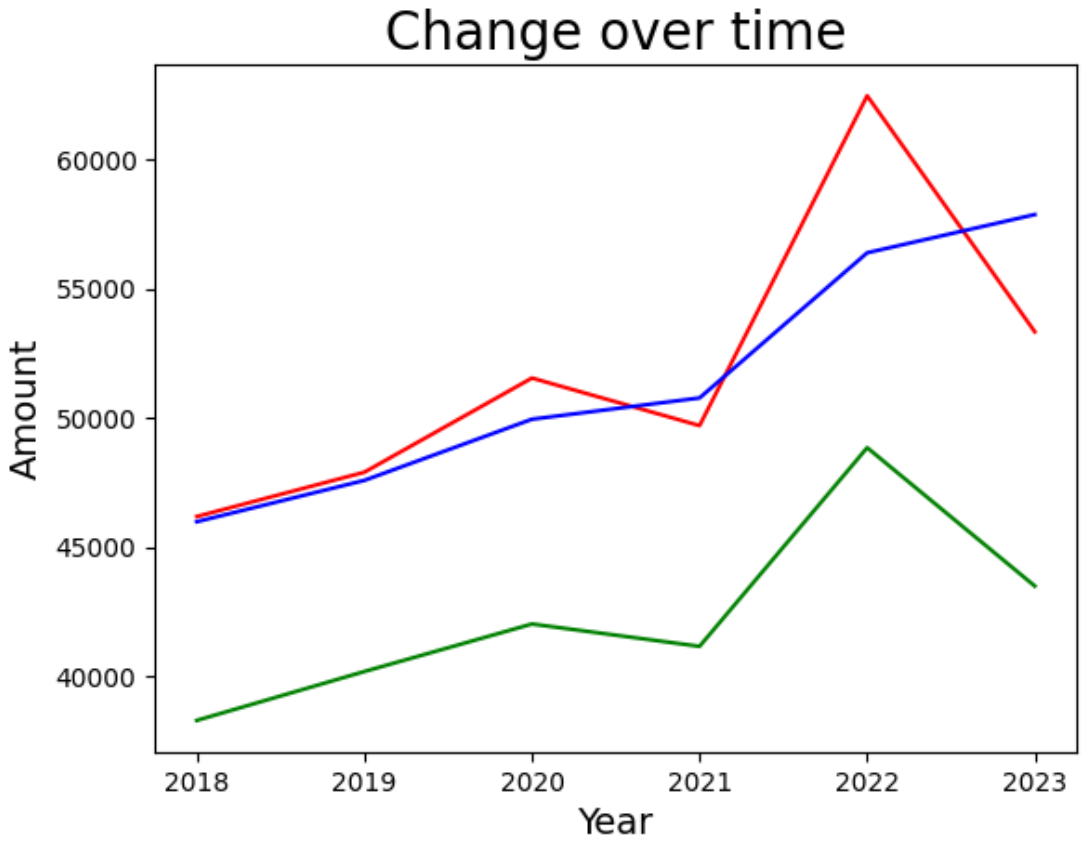

Figure 2.3 Trendline for Annual Rate(Blue), Regular Rate(Green), and YTD total(Red)

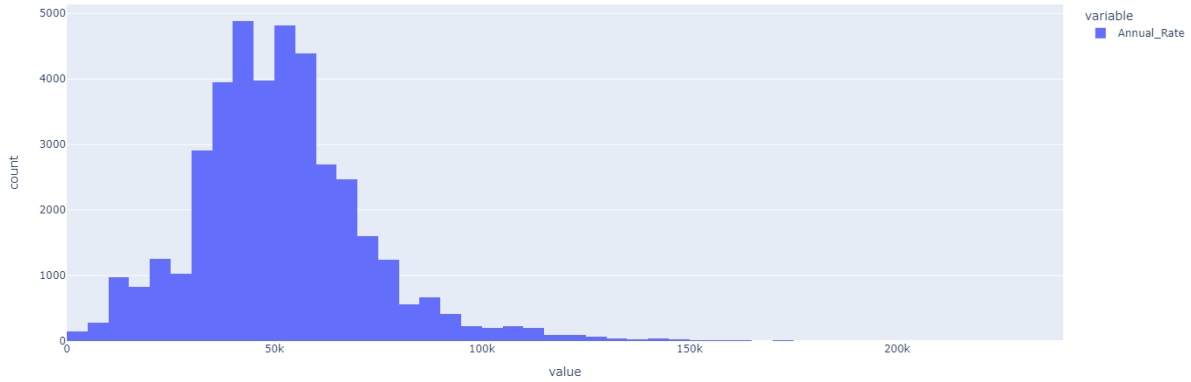

Figure 2.4: Histogram for Annual Rate

The above histogram for the Annual Rate of employees shows us how the employee compensation is normally distributed with a skew to the left. This can be interpreted that majority of employees, earn an annual salary of approx. 35k to 70k but there are few employees earning in the higher 150k. These employees can be considered as managers and higher government officials.

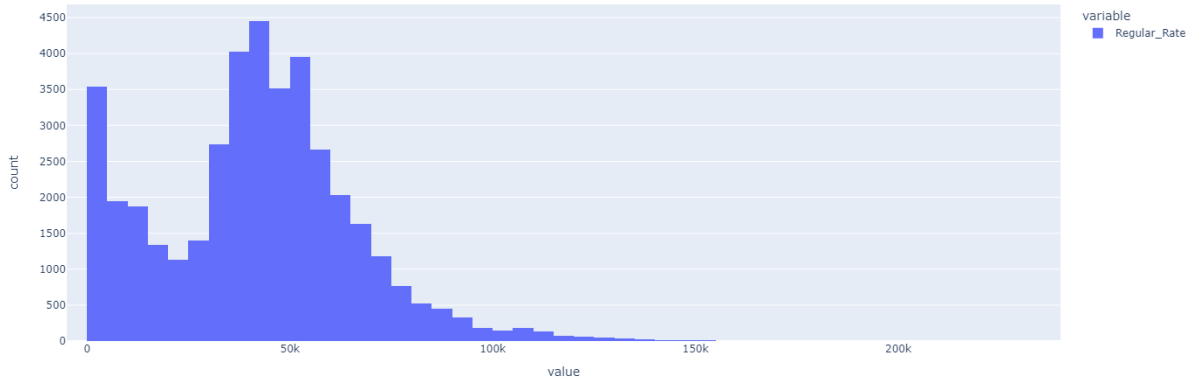

Figure 2.5: Histogram for Regular Rate

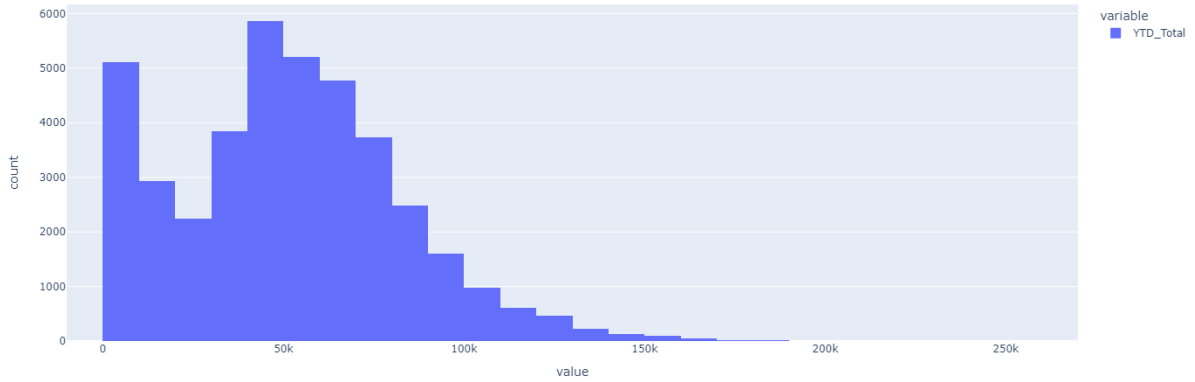

Figure 2.6: Histogram for YTD Total

Following the Histogram for annual rate, we take a glance at the histogram for regular and YTD total amounts. These histograms are almost identical to each other as seen in Figure 2.5 and 2.6. The mean for both is centralized to the left with minimal number of employees getting a higher amount of pay.

The above histograms are a good example to understand that the few employees earning over the 120k mark maybe considered as outliers. To now understand the fluctuation in compensation we calculate a new data column called Calculated Total.

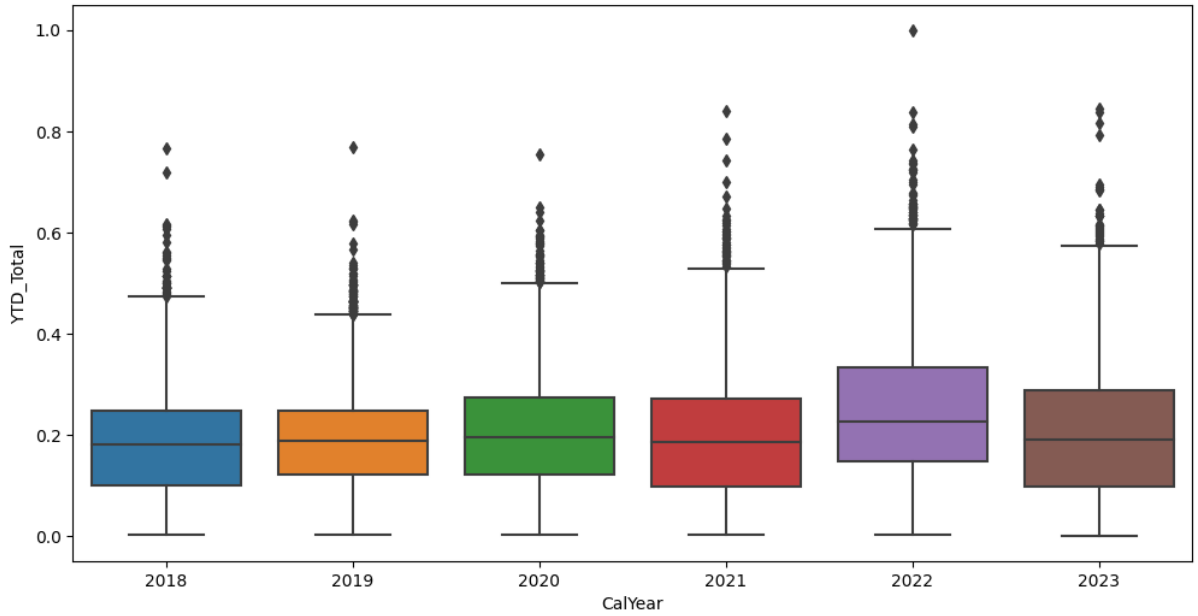

Figure 3.3: Boxplot for Year-To-Date from 2018 - 2023

The boxplot above represents the statistically significant values that have been spent by the government as part of the salary payment for all their employees. We can observe that in 2022, just after Covid lockdown impositions were slowly uplifted, there was a rise in the YTD-Total amounts paid towards employees.

The Annual Rate indicates the Annualized pay any employee should receive. And the YTD total indicates the amount received by the individual. The Calculated pay is a sum of the Regular Rate with the incentive pay and Overtime rate amount received by the individual. The values we find in this column are used to explain the difference in the actual amount the employee has received and the amount he was eligible to receive. We plot a line trend as seen in Figure 2.7 between the calculated pay and the annual rate for the different employees.

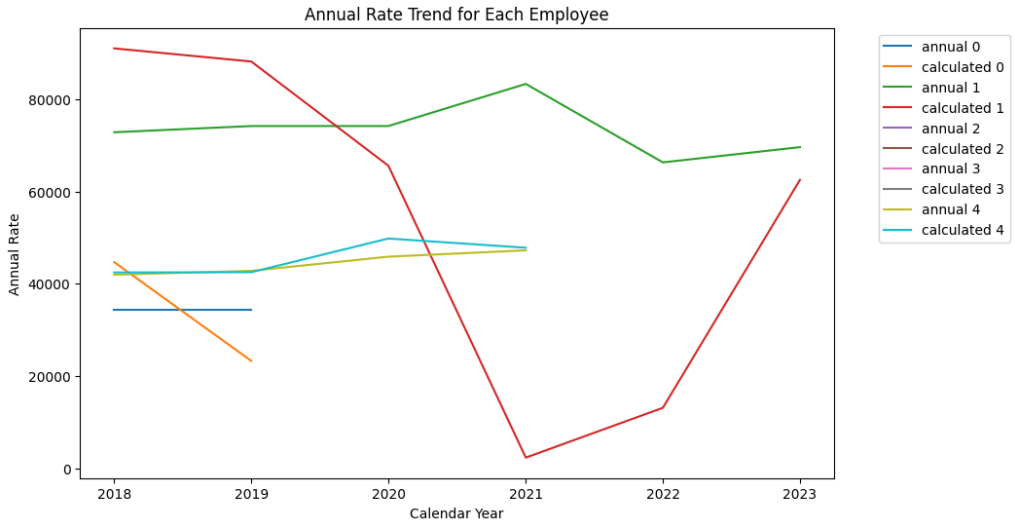

Figure 2.7: Line Plot for Calculated and Annual pay through the 5-year period.

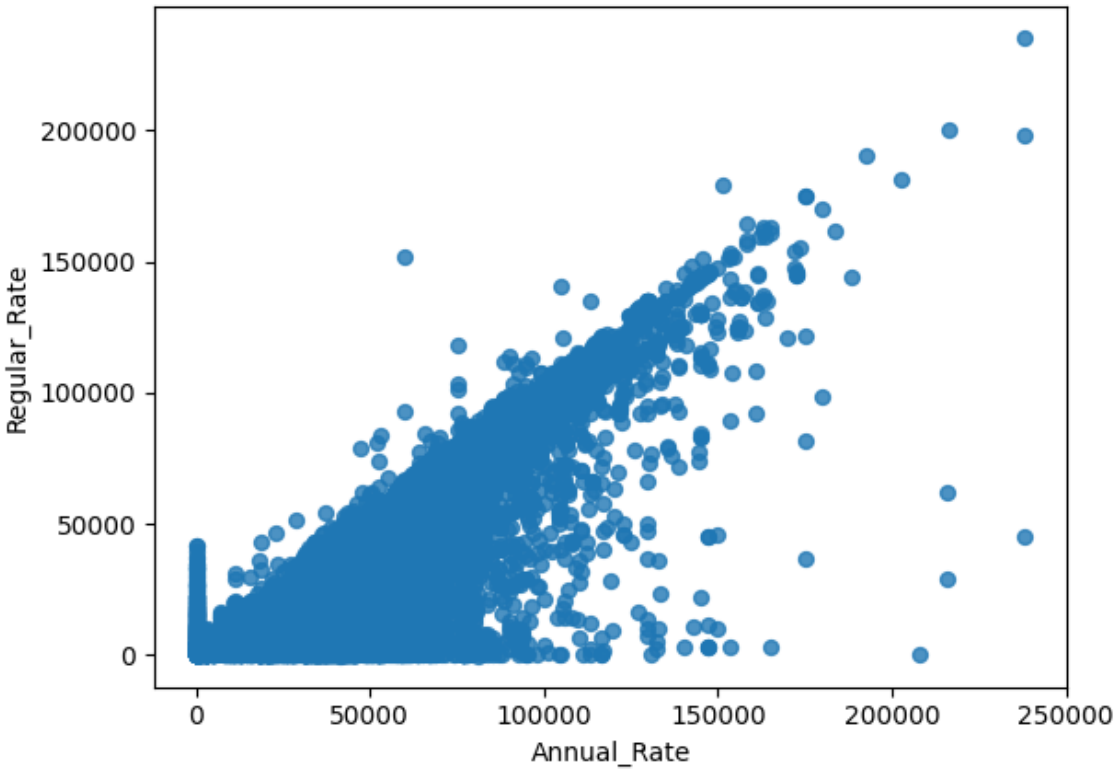

Figure 2.8: Annual Rate vs Regular Rate of employees

To understand the relationship and difference between the various parts of the salary data available, we plot the scatter plots for the different columns.
In this we come across the scatter plots for Annual Rate against Regular Rate.
We can see that there is high collinearity between the two, but there are a high number of data points spread across the lower half of the graph. This is representative of the fact that the Annual Rate for an employee may be higher even if the Regular rate is low. This is because there are numerous incentives and overtime rates which are applicable, leading to a rise in the annual rate.

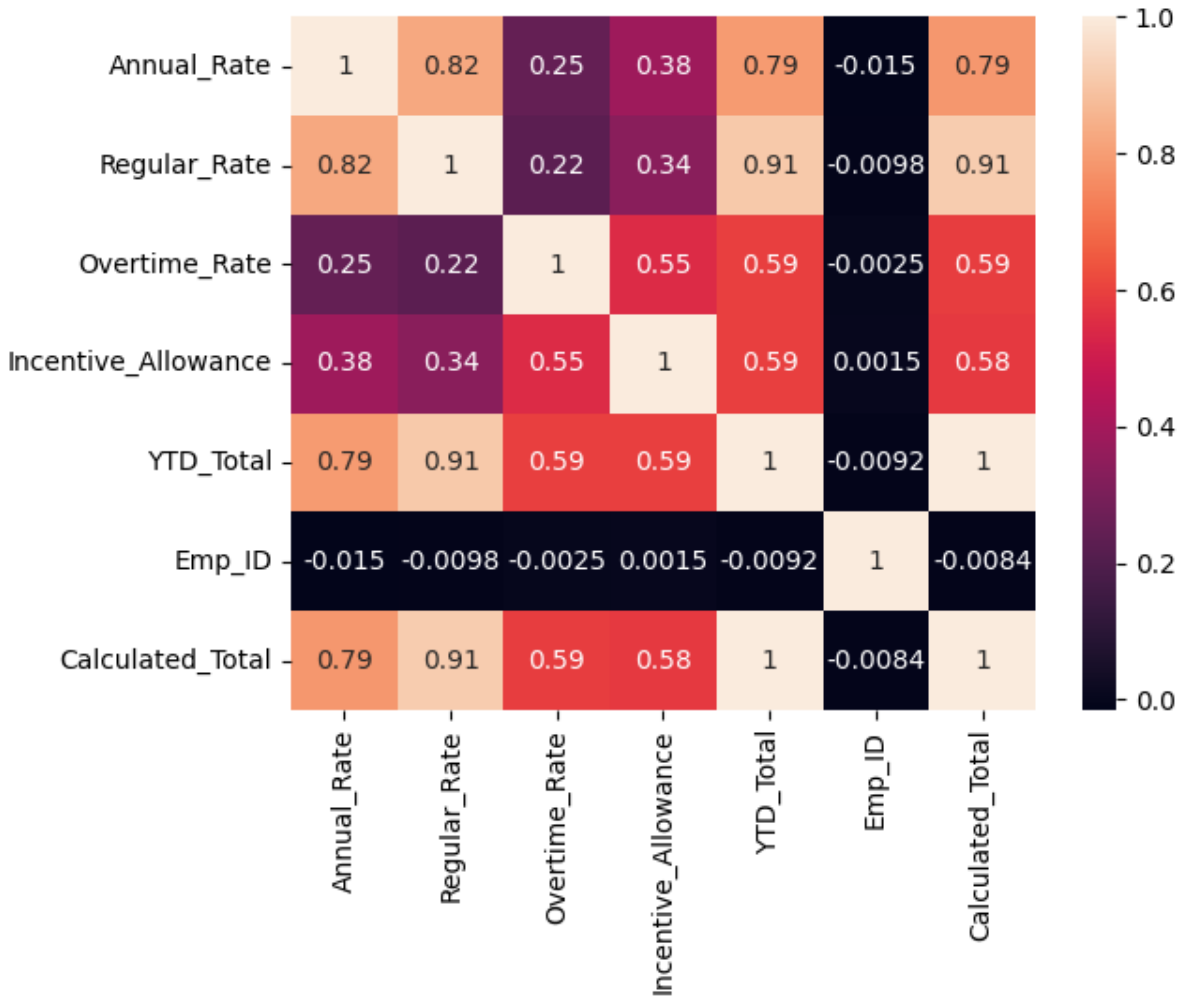

Figure 2.9 Correlation Matrix for Salaries

We can come to similar conclusion by observing the correlation matrix for our data. Within the correlation matrix, we can observe that the YTD_total column is highly correlated with the Annual and Regular Rates. The Annual and Regular Rate is correlated with each other with a factor of 0.82.

We can hence conclude that the YTD total can be easily interpreted based on annual and regular rates. The YTD total column also has a 0.59 correlation with Overtime Rate and Incentive Allowance.

With this we now take a glance at the salary distribution across the different departments of the government.

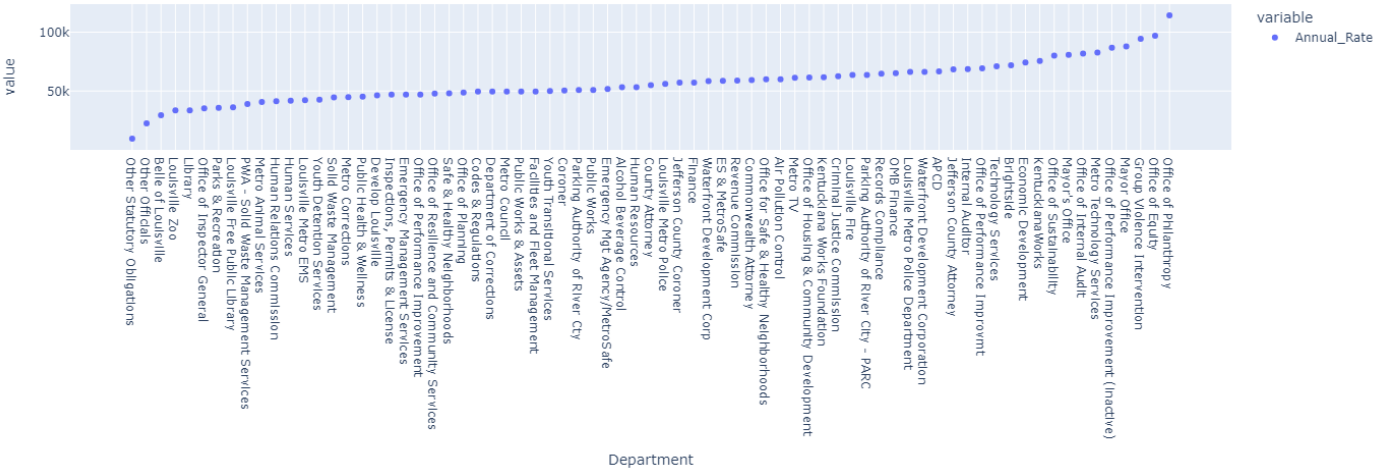

Figure 3.0: Mean salary for each department

From the Figure 3.0 we can clearly state that departments such as Parks & Recreation, Solid Waste management, Louisville Zoo and other officials are at the lower spectrum of the salary compensation plan. On the other hand, the offices of Philanthropy and Equity, the mayor's office, Office of Internal Audit etc have a higher salary compensation plan.

This drives us to the conclusion that the white-collar job offices have a higher salary than the blue-collar jobs. But in terms of departmental employees let us look at the distribution of how many employees does each department have over the years.

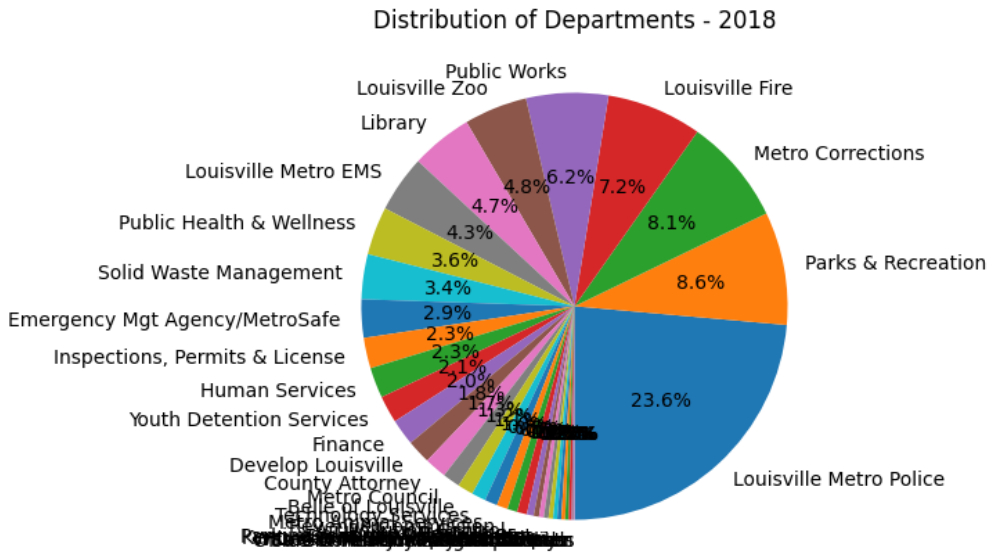

Figure 3.1: Distribution of Departments – 2018

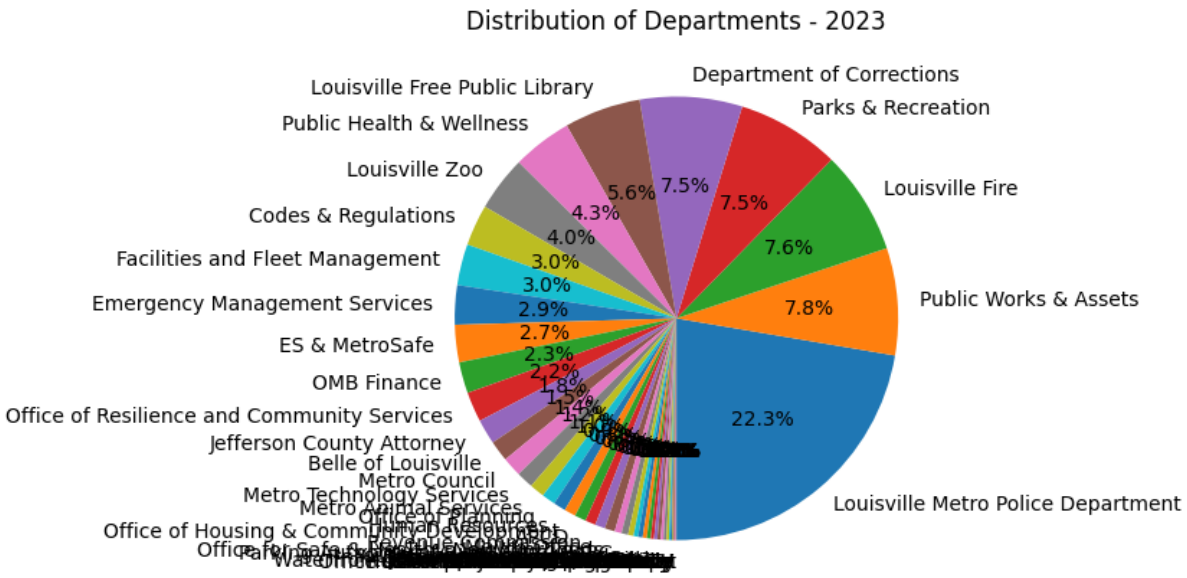

Figure 3.2: Distribution of Departments – 2023

We can observe that the highest number of employees are in the Louisville Metro Police Department and second highest number is Public Works & Assets. Putting together the information from Figure 3.0, 3.1, 3.2 we can state that although these departments have the highest number of employees, they pay scales lies somewhere between the Annual Rates of all government employees.
With the above explanatory data analysis, we can observe that there is a high linear correlation between the Calculated total and the YTD Total.

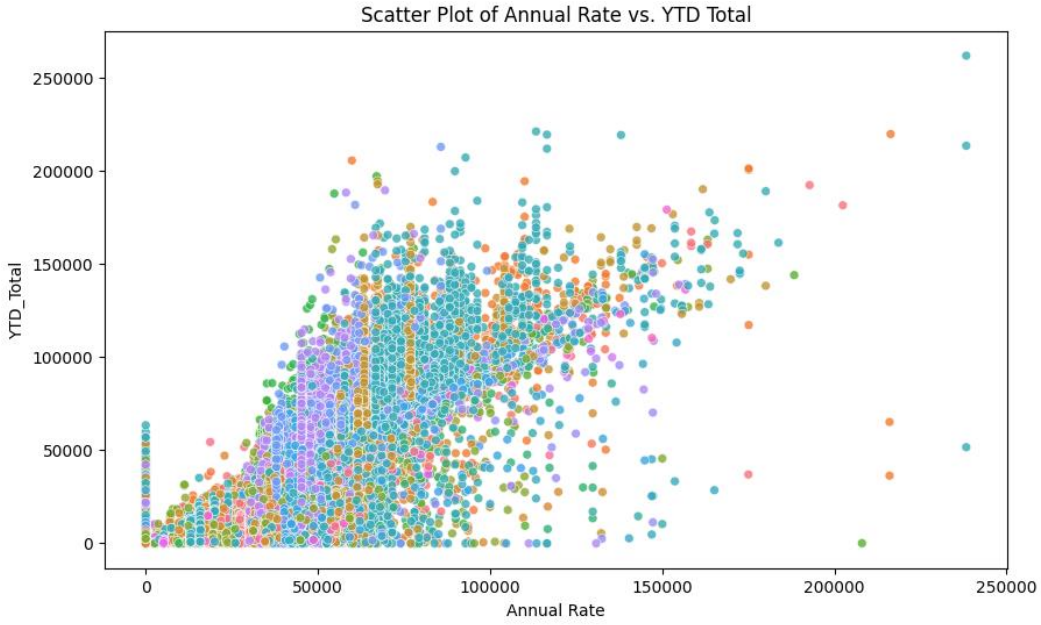

Figure 3.3 Scatter plot for Annual Rate vs YTD Total

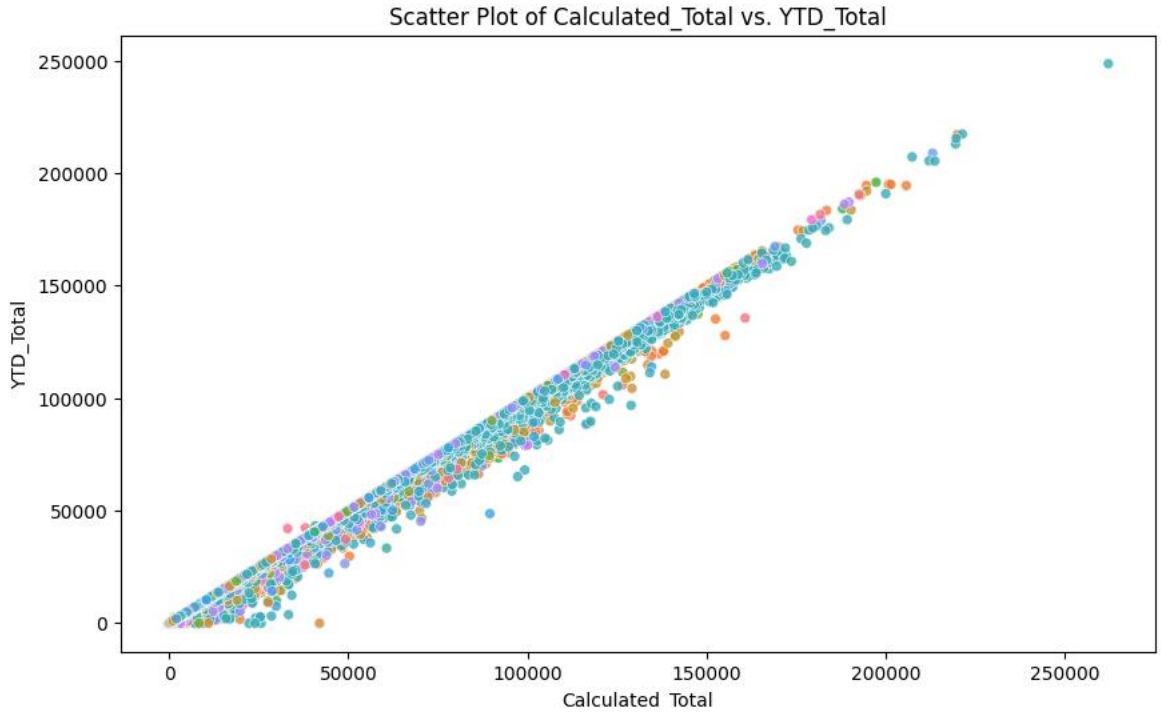

Figure 3.4 Scatter plot for Calculated Total vs YTD Total

From graph 3.3 and 3.4 we can note that the Annual Rate although in definition should be close to the YTD payment for employees is scatter over a wide range with the different colours representing the different departments. In Figure 3.4 we can observe that the Calculated total is a better variable to use in training a linear regression model. The data provided to us can be used to predict the amount of money any employee may earn at the end of the year, based on their annual rate, overtime rate and incentive allowance for the employee.

RESULTS

With the help of 'dash' library in python we build a dashboard for the smooth analysis of the data. The dashboard helps us to easily move the data around to find numerous insights. Dashboards leverage various data visualization elements such as charts, graphs, tables, and maps to represent complex data in a visually accessible manner. Common visualization types include line charts, bar charts, pie charts, heatmaps, and scatter plots. An effective analytical dashboard allows users to interact with the data dynamically. Users can filter, drill down, or zoom into specific data points to gain deeper insights. Interactive elements enhance the user experience and enable more nuanced analysis.

Dashboards can integrate data from various sources, allowing users to view a comprehensive picture of their business or operations. Integration capabilities enable a holistic analysis by combining data from different departments or systems. Users should have the flexibility to customize dashboards based on their

specific needs and preferences. This may include choosing which metrics to display, adjusting time frames, or selecting specific data subsets.

The following images are snapshots from the Dashboard developed.

Figure 3.6: Dashboard snapshot 1

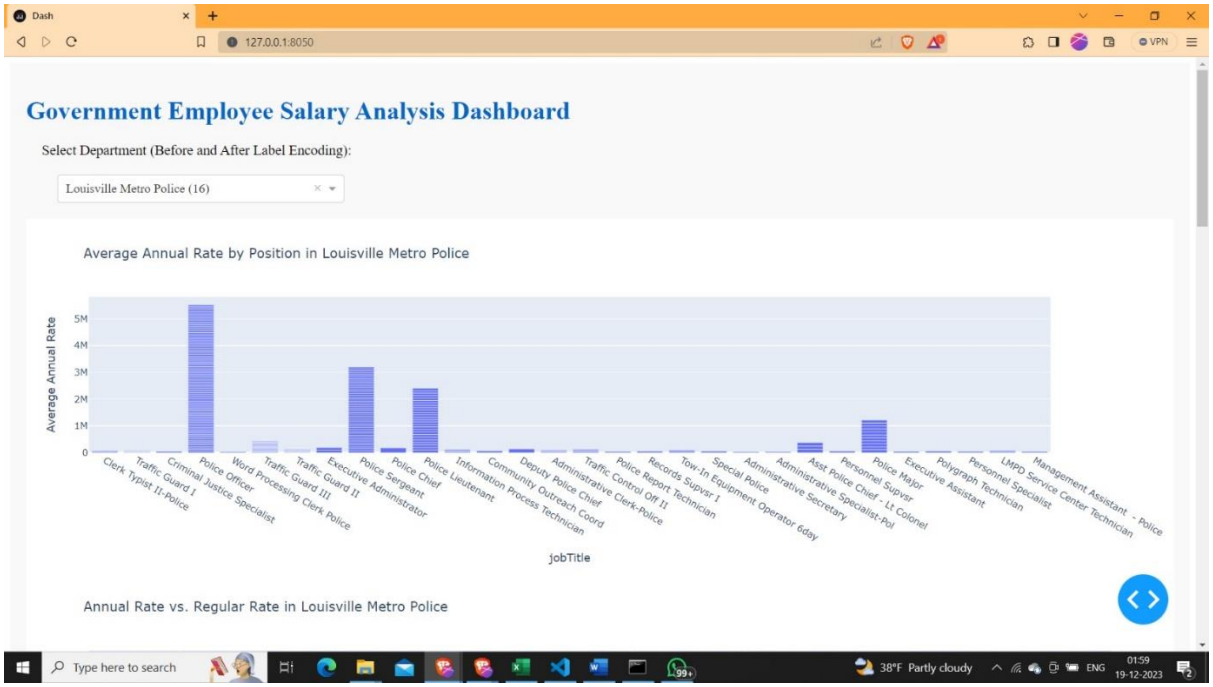

Figure 3.7: Dashboard snapshot 2

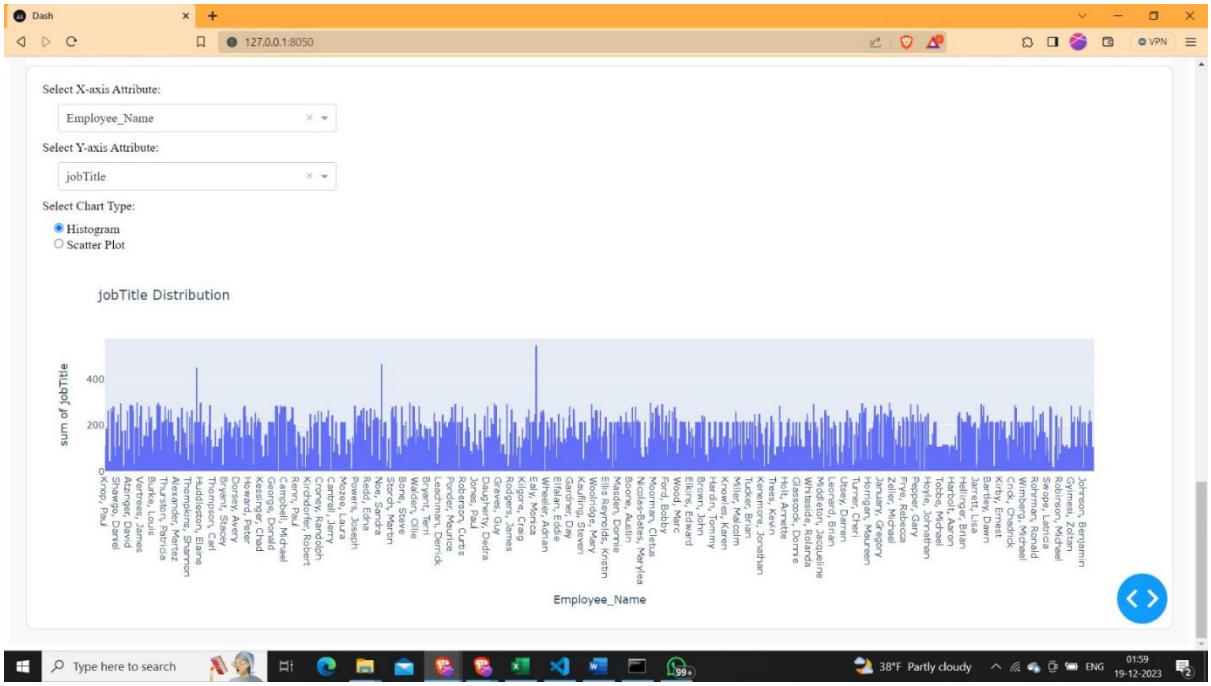

Figure 3.8: Dashboard snapshot 3

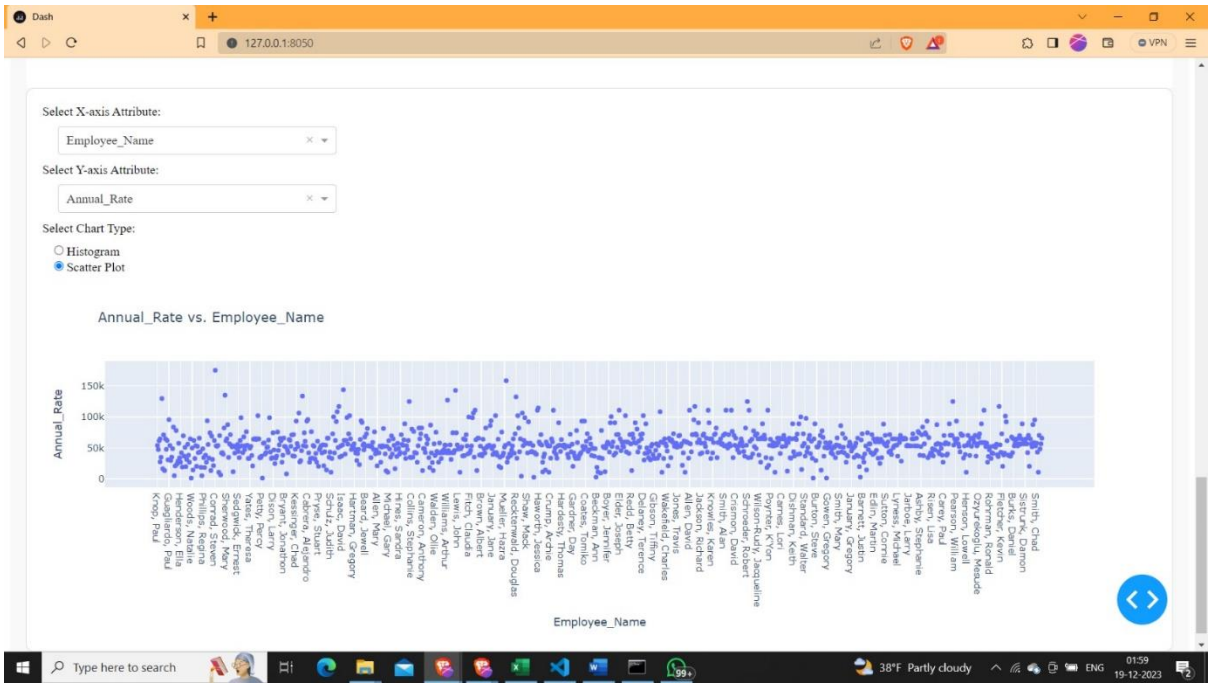

Figure 3.9: Dashboard snapshot 4

CONCLUSION

To summarize, the Government Employee Salary Analysis Dashboard is helpful resource for understanding the nuances of the public sector compensation. This data-driven platform helps local authorities make strategic decisions and increases transparency. Through the integration of different types of information, the dashboard enables effectiveness and responsibility.