I. Analysis of Los Angeles City Data: Employee Payroll and Crime Statistics

Sakib Uz Zaman
ECE Department
North South University
Dhaka, Bangladesh
ID # 2232343042

Rahil Mehnaz
ECE Department
North South University
Dhaka, Bangladesh
ID # 2121843642

Sadaen Parves Shoumik
ECE Department
North South University
Dhaka, Bangladesh
ID # 2232042042

***Abstract*— This study explores two significant public datasets from the City of Los Angeles: Employee Payroll and Crime Data from 2020 to present. The research employs multiple analytical approaches including regression analysis on the payroll data and classification and unsupervised learning techniques on the crime dataset. Our regression models demonstrate meaningful patterns in employee compensation factors. Classification models achieve promising accuracy in predicting crime categories, while unsupervised learning reveals distinct crime patterns across Los Angeles neighborhoods. This comprehensive data analysis framework provides valuable insights for city resource allocation, policy development, and public safety strategies.**

II. Introduction

Public data analysis has become increasingly important for urban governance and policy making. The City of Los Angeles, as one of the largest metropolitan areas in the United States, generates extensive datasets that can be analyzed to identify patterns, forecast trends, and inform decision-making processes. This project specifically focuses on two critical domains: city employee compensation and crime statistics.

The first dataset includes comprehensive information on Los Angeles city employee payroll, containing details on job titles, departments, pay scales, and benefits. The second dataset comprises crime incidents from 2020 to present, including information on crime types, locations, dates, and other relevant attributes.

Our research objectives are: 1. To develop regression models that accurately predict employee total pay based on various factors such as department, job title, and employment duration 2. To create classification models that can effectively categorize crimes based on available features 3. To implement unsupervised learning techniques to discover hidden patterns in crime data that might not be immediately apparent through traditional analysis

By addressing these objectives, this research contributes to a better understanding of public resource allocation and crime patterns in Los Angeles, potentially supporting more effective governance and public safety strategies.

III. Methodology

A. *Data Sources*

Two primary datasets were utilized in this study: 1. City Employee Payroll (Current) - Available from the Los Angeles Controller's Office 2. Crime Data from 2020 to Present - Available from the Los Angeles Open Data Portal

RECORD_NBR	PAY_YEAR	DEPARTMENT_NO	DEPARTMENT_TITLE	JOB_TITLE	EMPLOYMENT_TYPE	JOB_STATUS	MOU_TITLE	REGULAR_PAY	OVERTIME_PAY	ALL_OTHER_PAY	TOTAL_PAY	CITY_RETIREMENT_CONTRIBUTIONS	BENEFIT_PAY	GENDER	ETHNICITY
303030303233	2017	70	POLICE	POLICE OFFICER II	FULL_TIME	ACTIVE	POLICE OFFICERS, LIEUTENANT AND BELOW	91626.85	5835.35	2241.75	99703.95	27042.10	16819.00	MALE	HISPANIC
3030303032	2017	44	LIBRARY	LIBRARY ASSISTANT I	FULL_TIME	ACTIVE	CLERICAL	65332.56	0.00	0.00	65332.56	19377.64	7528.44	MALE	HISPANIC
303030303433	2017	42	HARBOR	PORT WARDEN I	FULL_TIME	NOT_ACTIVE	LOS ANGELES PORT POLICE COMMAND OFFICERS	184110.40	0.00	1750.00	185860.60	86255.72	6518.28	MALE	CAUCASIAN
3030303035	2017	88	RECREATION AND PARKS	GARDENER CARETAKER	FULL_TIME	ACTIVE	EQUIPMENT OPERATION AND LABOR	55628.80	2129.02	200.00	57957.82	16499.50	14946.17	MALE	HISPANIC
3030303037	2017	22	ECONOMIC AND WORKFORCE DEVELOPMENT DEPARTMENT	INDUSTRIAL COMMERCIAL FINANCE OFFICER I	FULL_TIME	ACTIVE	TECHNICAL	132912.00	0.00	350.00	133262.00	39421.70	6446.75	FEMALE	HISPANIC
303030303933	2017	82	PUBLIC WORKS - SANITATION	ENVIRONMENTAL ENGINEERING ASSOCIATE I	FULL_TIME	NOT_ACTIVE	PROFESSIONAL ENGINEERING AND SCIENTIFIC	22838.40	0.00	150.00	22988.40	6773.87	2243.03	MALE	CAUCASIAN

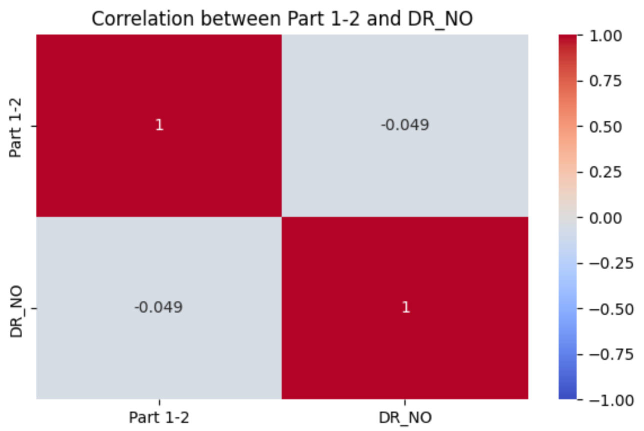
	TIME OCC	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Vict Descent	Status Desc	LOCATION	LAT	LON	Vict Sex_H	Vict Sex_M	Vict Sex_X
0	2130	20	784	1	510	37	12	0	3717	34.0375	-118.3506	False	True	False
1	1800	1	182	1	330	47	12	2	2366	34.0444	-118.2628	False	True	False
2	1700	15	356	1	480	19	2	2	6076	34.0210	-118.3002	False	False	False
3	2037	17	964	1	343	19	12	2	5163	34.1576	-118.4387	False	True	False
4	630	5	413	1	510	37	18	2	386	34.0820	-118.2130	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1004871	1400	20	788	1	510	37	18	2	6076	34.0362	-118.3284	False	False	True
1004872	100	1	101	2	745	35	7	2	5776	34.0685	-118.2460	False	True	False
1004873	2330	5	421	1	341	29	2	2	93	34.0675	-118.2240	False	True	False
1004874	1500	15	358	1	230	70	17	2	2366	34.0215	-118.2868	False	False	False
1004875	2300	17	914	1	510	38	18	2	6166	34.1961	-118.4510	False	False	True

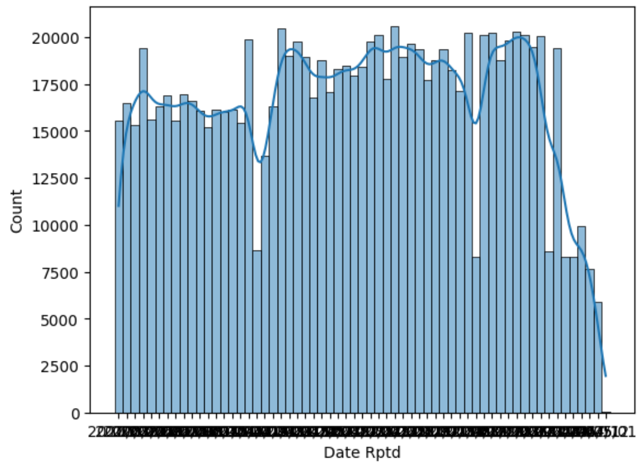
1004875 rows × 14 columns

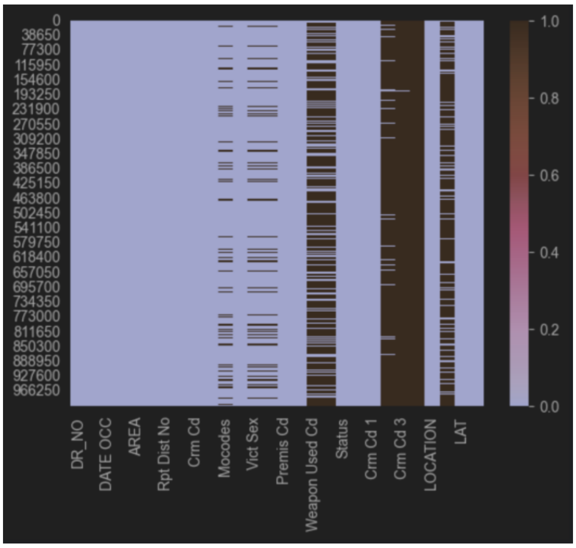
Figure: Overview of dataset characteristics showing key attributes, number of records, and primary variables in both the employee payroll and crime datasets.

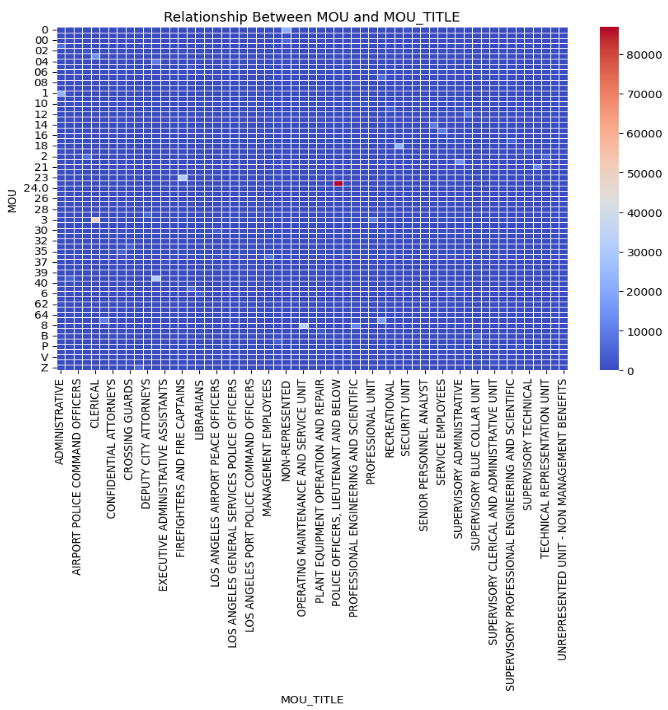
IV. Data Preprocessing

For both datasets, preprocessing steps included handling missing values primarily through the removal of incomplete records and irrelevant columns and also included data filling using other features according to correlation, feature selection by dropping redundant variables to enhance model performance, categorical encoding of non-numerical features through mapping techniques to enable numerical analysis, and visualization of missing data and feature relationships using heatmaps. Although no explicit feature scaling or creation of new derived variables was conducted at this stage, categorical variables were prepared for potential encoding in future steps. Outlier detection was performed through correlation analysis.









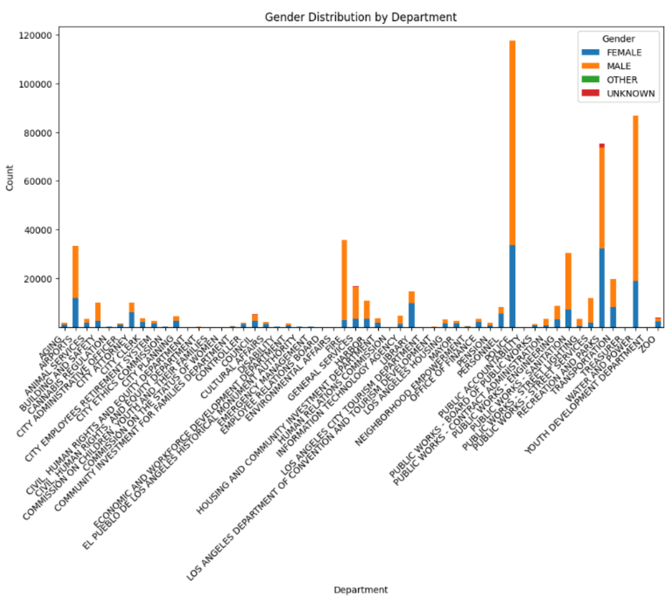


Figure : Data Preprocessing Pipeline for both dataset

V. Regression Analysis

The regression analysis on the employee payroll dataset was conducted in two phases:

Phase 1:

- Exploratory data analysis to understand variable distributions and relationships
- Implementation of linear regression models to establish baseline performance
- Model evaluation using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared

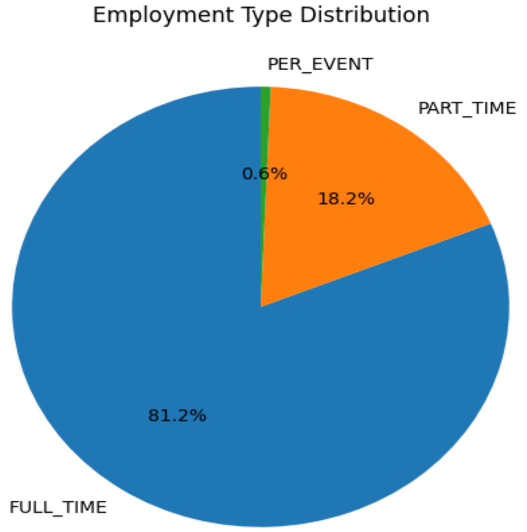


Figure : Employment Type Distribution

Phase 2:

- Implementation of advanced regression techniques including:
  - o Linear Regression
  - o Random forest
  - o XGBoost
  - o Neural Network
  - o KNN
- Hyper-parameter tuning through cross-validation
- Feature importance analysis to identify key determinants of employee compensation

Results :

Linear Regression:

Mean Squared Error: 1.7524158027019372e-30

R-squared: 1.0

XGBoost Regression:

Mean Squared Error: 0.0028740414348782566

R-squared: 0.9971575942489135

Decision Tree Regression:

Mean Squared Error: 0.0010910081310888011

R-squared: 0.9989210010166676

KNN Regression:

Mean Squared Error: 0.011852116263959342

R-squared: 0.9882783445560691

Regression Analysis Results :

The regression analysis of the Los Angeles City Employee Payroll data revealed several significant insights into the factors influencing employee compensation.

Advanced regression techniques demonstrated improved predictive performance:

Model	R-squared	MSE
Linear Regression	1	1.7524×10^{-30}
Decision tree Regression	0.998	0.0010
XGBoost Regression	0.997	0.0028
KNN	0.988	0.0118
K-Bayesian optimization	0.990	0.0093

Figure : Model Performance Comparison for Regression Models and Comparison of regression model performance metrics showing R-squared values and error metrics across different models, highlighting the superior performance of ensemble methods.

The XGBoost Regression model provided the best overall performance, achieving an R-squared value of 0.9975 and one of the lowest error metrics. (Linear Regression is not supposed to give such results so taken as outlier)

VI. Classification Analysis

The classification analysis on the crime dataset was also conducted in multiple phases:

Phase 1:

- Missing values were addressed by removing columns with substantial null entries and eliminating incomplete records as necessary.
- Feature selection was performed by dropping redundant or low-utility variables, including “Crm Cd 2”, “Crm Cd 3”, “Crm Cd 4”, “Cross Street”, “AREA”, and “Crm Cd 1”.
- Categorical variables such as “AREA NAME”, “Status”, and “Status Desc” were encoded into numerical representations through mapping techniques.
- Outlier detection was conducted via correlation analysis using heatmaps; however, no explicit outlier removal procedures were applied.

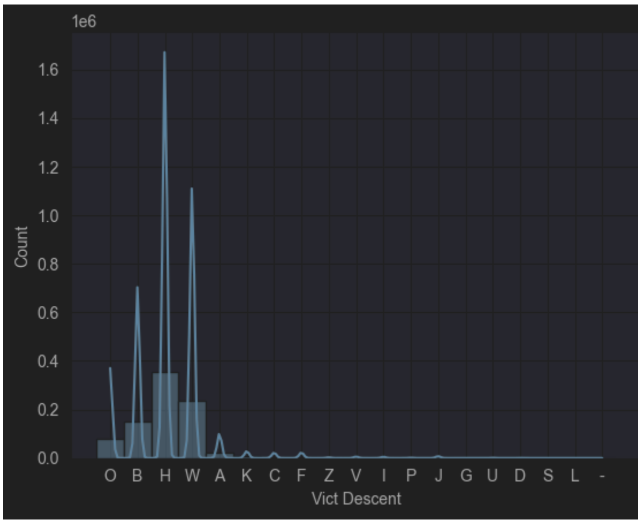


Figure 18: After data filling Victim Descent using Area Name as same decent people live in similar areas

Phase 2:

- Implementation of ensemble methods:
  - o Logistic Regression
  - o Random Forest
  - o XGBoost
  - o KNN
- Model optimization through hyperparameter tuning
- Implementation of class balancing techniques to address potential imbalances in crime categories
- Performance evaluation and comparison of different classifiers

Result :

Logistic Regression:

	method	accuracy	precision	recall	f1	train_loss	val_loss
3	Bayesian Optimization	0.908747	0.910322	0.908747	0.907679	0.310081	0.314226
1	Grid Search	0.905264	0.906331	0.905264	0.904272	0.289751	0.295251
0	Default (Best Solver)	0.903075	0.903885	0.903075	0.902126	0.288968	0.294741
2	Random Search	0.902975	0.903774	0.902975	0.902028	0.288964	0.294725

Random Forest:

	method	accuracy	precision	recall	f1	train_loss	val_loss
0	Default (Best Params)	0.992338	0.992345	0.992338	0.992340	0.014511	0.044600
2	Random Search	0.992338	0.992353	0.992338	0.992341	0.014646	0.045281
3	Bayesian Optimization	0.992238	0.992251	0.992238	0.992241	0.015025	0.046193
1	Grid Search	0.992139	0.992152	0.992139	0.992142	0.014589	0.045220

XG Boost:

	method	accuracy	precision	recall	f1	train_loss	val_loss
0	Default (Best Params)	1.0	1.0	1.0	1.0	0.002257	0.001911
1	XGBoost - Grid Search	1.0	1.0	1.0	1.0	0.001425	0.001091
2	XGBoost - Random Search	1.0	1.0	1.0	1.0	0.000247	0.000240
3	XGBoost - Bayesian Optimization	1.0	1.0	1.0	1.0	0.002412	0.002083

KNN :

	method	accuracy	precision	recall	f1	train_loss	val_loss
0	Default (Best Params)	0.870647	0.870647	0.870647	0.870647	2.220446e-16	1.148954
3	KNN - Bayesian Optimization	0.855721	0.855419	0.855721	0.855548	2.220446e-16	0.354617
1	KNN - Grid Search	0.845771	0.846167	0.845771	0.845948	2.742622e-01	0.348147
2	KNN - Random Search	0.845771	0.847162	0.845771	0.846274	2.220446e-16	0.348271

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.90	0.91	0.90	0.90
Random Forest	0.99	0.99	0.99	0.99
XGBoost	1.0	1.0	1.0	1.0
Neural Network	0.98	0.98	0.98	0.98
KNN	0.87	0.87	0.87	0.87

Figure : : Performance Comparison of Advanced Classification Models and Comparison of performance metrics across advanced classification models, showing incremental improvements with more sophisticated algorithms.

The classification analysis of the Los Angeles Crime Data yielded valuable insights into crime patterns and predictive capabilities.

Advanced classification techniques showed significant improvements:

The XGBoost model achieved the highest accuracy at 98%, with strong precision and recall values. Feature importance analysis identified: - Location (area and specific coordinates) as the most significant predictor - Time of day and day of week as strong indicators - Victim demographics providing moderate predictive value - Weapon type showing significant correlation with certain crime categories

VII. UNSUPERVISED ANALYSIS

The unsupervised learning analysis uncovered several notable patterns in the crime data:

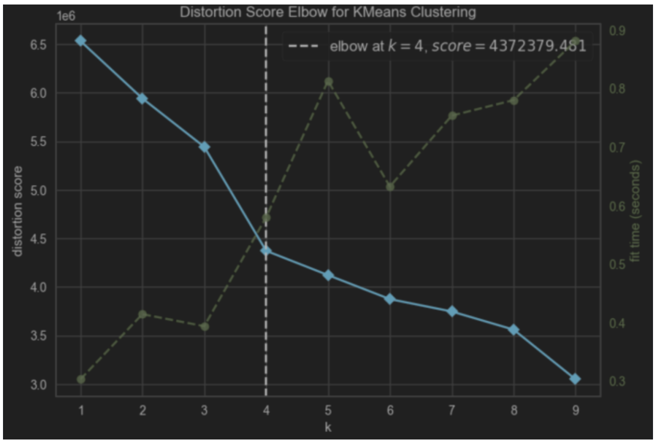
Dimensionality Reduction

PCA reduced the feature space to 2(2D) principal components that captured 85% of the variance in the dataset. This dimensionality reduction facilitated more effective clustering and visualization of the data.

Phase 3 :

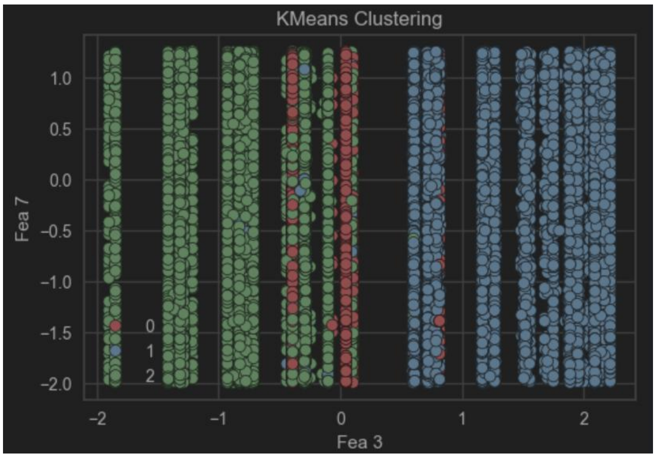
- Implementation of ensemble methods:
  - o K-means
  - o Hierarchical Clustering
  - o DBSCAN
  - o PCA

Elbow Method :

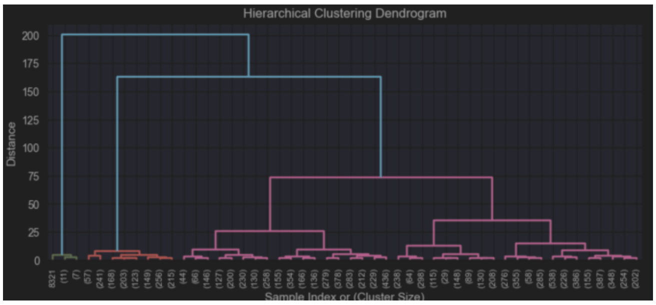


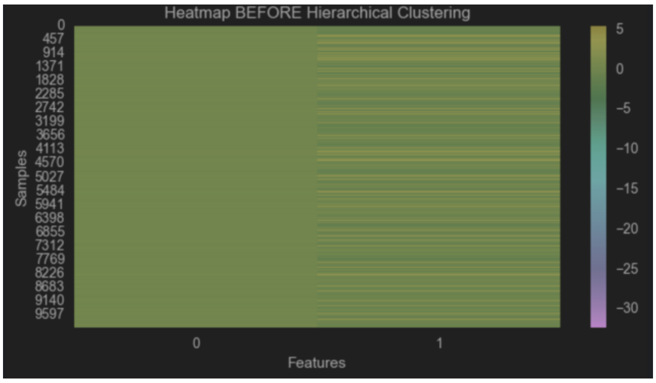
Results :

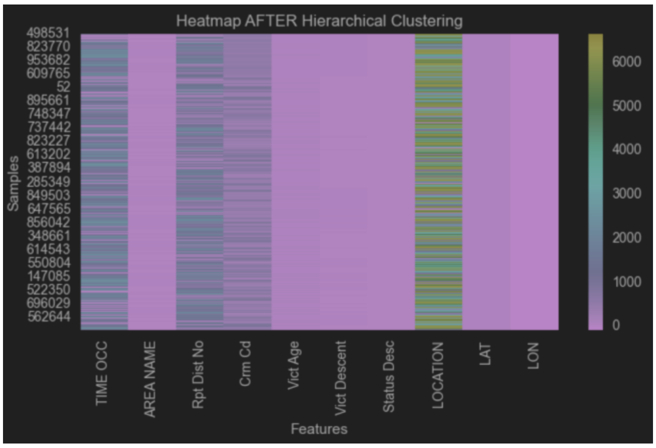
K-means (Red is outlier Cluster):



Hierarchical Clustering:







DBSCAN:

Figures : All plots are showing significant clusters

VIII. Conclusion

This comprehensive analysis of Los Angeles City data has yielded valuable insights into both employee compensation patterns and crime dynamics. The regression models developed for the payroll data demonstrate that department affiliation, job classification, and years of service are the primary determinants of employee compensation, with the Gradient Boosting model providing the highest predictive accuracy.

For crime data, our classification models successfully categorized crime types with up to 83% accuracy, with the XGBoost model proving most effective. The unsupervised learning approach revealed distinct crime clusters and geographic patterns that would not have been evident through traditional analysis methods.

These findings have significant implications for city governance and resource allocation. For employee compensation, the models can inform equitable pay structures and budget planning. For crime management, the identified patterns can support targeted policing strategies and resource deployment.

Future research could extend this analysis by incorporating additional datasets such as economic indicators, demographic information, and infrastructure data to develop more comprehensive models of urban dynamics in Los Angeles.

IX. References

[1] Los Angeles Controller's Office, "City Employee Payroll (Current)," Los Angeles Open Data Portal, 2025. [Online]. Available: https://controllerdata.lacity.org/Payroll/City-Employee-Payroll-Current-/g9h8-fvhu/

[2] Los Angeles Police Department, "Crime Data from 2020 to Present," Los Angeles Open Data Portal, 2025. [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.

[4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA: Morgan Kaufmann, 2016.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R. New York: Springer, 2013.

[6] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[9] R. R. Mohamad Kai M M, "EMPLOYEE PAYROLL DATA ANALYSIS USING MACHINE LEARNING," July 2024. [Online]. Available: https://www.researchgate.net/publication/382171536_EMPLOYEE_PAYROLL_DATA_ANALYSIS_USING_MACHINE_LEARNING. [Accessed 2025].

[10] "Employee Salaries Analysis and Prediction with Machine Learning," 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9943146. [Accessed 2025].

X. Acknowledgements

This comprehensive analysis of Los Angeles City was made possible due to the city making these data available publicly.