**Centre for Development of Advanced Computing (CDAC)**

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

Mumbai Centre – USM VITA

**Academic Year: Feb – Aug 2025**

**Project Synopsis**

NYC Housing Complaints & Affordable Living Index Prediction using Big Data & Machine Learning

**Guided By: Submitted By**

**Mr. Sarthak Dhargalkar Group 4**

## Title:

NYC Housing Complaints & Affordable Living Index Prediction using Big Data & Machine Learning

## Team:

* Team Leader : Sapedi Chetan Swaroop
* Team Member : Prateek Kulshrestha
* Team Member : Rushikesh Pabalkar
* Team Member : Sonu Lad
* Team Member : Abhishek Singh
* Team Member : Anannya Sarkar
* Team Member : Shakyadita Sonawane
* Team Member : Vedika Mule

## Problem Statement:

Housing-related complaints in New York City have been consistently increasing. However, identifying complaint-prone areas and evaluating affordable living conditions remains difficult due to large and unstructured datasets. There is a need for a system that can predict housing complaints and generate an Affordable Living Index to guide policy-making and tenant decisions.

## Real-Life Problem Being Solved:

- Helps tenants choose safer, well-maintained neighborhoods.  
- Assists NYC housing authorities in prioritizing inspection zones.  
- Supports real-estate agencies and urban planners in evaluating building livability.  
- Reduces public health and safety risks by early prediction of complaints.

## Introduction:

The rapid urbanization and population growth in New York City have led to increased pressure on housing infrastructure, often resulting in a spike in tenant complaints related to safety, maintenance, heating, plumbing, and other living conditions. NYC's 311 service receives millions of complaints annually, providing a massive but untapped source of data that can offer insights into the city’s housing quality trends. However, these datasets are vast, diverse, and often unstructured, making it difficult to extract actionable intelligence without sophisticated tools and techniques.  
  
This project aims to leverage open data from the NYC Open Data portal—including 311 Service Requests, HPD Housing Maintenance data, PLUTO property data, and Affordable Housing datasets—to analyze patterns in housing-related complaints and predict future occurrences using Machine Learning. Additionally, the project introduces a novel Affordable Living Index that quantifies the livability of residential buildings and neighborhoods based on various socio-economic and complaint-related features.  
  
To manage and process the 5–10 GB of data efficiently, we utilize a full-scale Big Data architecture on AWS, including Amazon S3 for data storage, AWS Glue for ETL operations, Amazon RDS for structured data warehousing, and Amazon EMR with PySpark for distributed machine learning. Insights from our ML models are presented in intuitive and dynamic dashboards using Power BI and Tableau, enabling end-users—such as city officials, residents, and real estate agencies—to make informed decisions.  
  
This project is a comprehensive application of the technologies and concepts covered in our PGDBDA course, including Data Engineering, Big Data, Cloud Computing, and Machine Learning. It not only builds technical competence but also addresses a socially significant problem—improving urban housing conditions and aiding residents in identifying safe and affordable housing options.

## Objectives:

- Predict types and frequency of housing complaints using ML.  
- Identify complaint-prone areas/buildings.  
- Design an Affordable Living Index for NYC.  
- Build a data warehouse integrating multiple housing datasets.  
- Visualize insights on a reporting dashboard for various stakeholders.

## Databases Used (from NYC Open Data):

|  |  |  |
| --- | --- | --- |
| Dataset Name | Source | Role |
| 311 Complaints | https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9 | Primary source of housing complaint data |
| HPD Maintenance Complaints | https://data.cityofnewyork.us/Housing-Development/Housing-Maintenance-Code-Complaints-and-Problems/ygpa-z7cr | Supplementary complaint data, more specific |
| PLUTO | https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-/64uk-42ks | Building characteristics |
| Affordable Housing Lottery Units | https://data.cityofnewyork.us/Housing-Development/Affordable-Housing-Lottery-Units/ | Locations of affordable units |
| Building Footprints | https://data.cityofnewyork.us/City-Government/BUILDING/5zhs-2jue | Spatial mapping of buildings |

## Key Columns and Roles:

|  |  |  |
| --- | --- | --- |
| Dataset | Columns | Role |
| 311 Complaints | Complaint Type, Created Date, Borough, Status, Location | Train ML model to predict complaint trends |
| HPD Complaints | Complaint Type, Received Date, Building ID, Location | Cross-validation and deeper issue detection |
| PLUTO | BBL, Land Use, Year Built, Building Area, Residential Units | Used for feature engineering in livability scoring |
| Affordable Units | Address, Borough, Project ID, Location | Compare livability between affordable vs. market housing |
| Building Footprints | Building ID, Area, GeoShape | Accurate geolocation for dashboards and mapping |

## Use Cases:

- Predict which buildings are likely to receive mold/heating complaints in the next month.  
- Score each neighborhood/building by its 'Affordable Living Index'.  
- Visual dashboard for public users to check livability by zip code.  
- Report generation for government agencies.

## Tech Stack:

|  |  |
| --- | --- |
| Layer | Tools/Technologies |
| Cloud | AWS (S3, EC2, EMR, Glue, RDS, Lambda, VPC) |
| Data Processing | PySpark (on EMR), Pandas |
| ETL | AWS Glue, Lambda |
| ML Models | PySpark MLlib, Scikit-learn (for local testing) |
| Visualization | Tableau, Power BI |
| Storage | AWS S3 (raw data lake), MySQL on AWS RDS (Data Warehouse) |
| Orchestration | AWS Lambda & Glue triggers |
| Dashboard Deployment | Power BI Server or AWS QuickSight |

## Project Flow:

|  |  |  |
| --- | --- | --- |
| Step | Description | Tools Used |
| 1 | Download and store raw datasets in S3 | AWS S3 |
| 2 | Perform ETL and transformation | AWS Glue |
| 3 | Process large data using Spark | AWS EMR (PySpark) |
| 4 | Load transformed data into RDS warehouse | AWS RDS |
| 5 | Train ML models to predict complaints | PySpark MLlib |
| 6 | Generate Affordable Living Index | PySpark (weighted scoring) |
| 7 | Visualize insights on dashboard | Tableau/Power BI |
| 8 | Automate pipeline with triggers | AWS Lambda + Glue |
| 9 | Secure the infra & enable monitoring | AWS VPC, CloudWatch |

## Expected Outcomes:

- A functioning ML model to predict housing complaints.  
- Livability scoring index for buildings and areas.  
- Centralized Data Warehouse built using cloud tools.  
- Interactive visual dashboards for decision-makers and citizens.  
- Automated cloud-native ETL + ML pipeline.

## Conclusion:

This project brings together the power of Big Data, Machine Learning, and Cloud Computing to solve a real-world urban issue—improving the quality of housing in New York City. By analyzing large volumes of complaint data from the NYC Open Data platform, we can identify patterns in housing issues like heating failures, maintenance delays, and structural safety concerns.

The predictive machine learning models help forecast future complaints, enabling city officials to act proactively rather than reactively. The creation of the **Affordable Living Index** adds even more value, as it provides a clear, data-driven score to evaluate how livable and affordable a building or neighborhood is for residents.

Using tools like AWS S3, Glue, EMR, and RDS, we built an end-to-end cloud-based data pipeline that can handle large-scale data ingestion, transformation, storage, and analysis. The insights are presented through interactive dashboards using Tableau or Power BI, making it easy for decision-makers, tenants, and housing planners to access and use the results effectively.

In summary, this project not only applies technical skills learned during the course—such as data integration, modeling, visualization, and cloud orchestration—but also addresses a socially important problem. It offers practical value by helping the city become smarter in managing housing issues and guiding residents to safer and more affordable living options.

## Future Scope:

- Expand to include crime, noise, and health inspection data.  
- Predict housing complaint trends citywide using time-series forecasting.  
- Build a mobile app or public-facing dashboard.  
- Offer complaint alerts to tenants and authorities in real time.  
- Deploy ML models as REST APIs via AWS SageMaker for reuse in other cities.