Check Point: Crime Data and Housing Prices

Hsuan-Chi Chang, Wei-Ju Li
Department of Computer Science
Virginia Tech, Alexandria, VA, USA
hsuanchi@vt.edu, weijuli@vt.edu

***Abstract*—This project presents a scalable decision-support system that integrates housing market data and crime statistics to help homebuyers and investors make balanced decisions between growth and safety. We analyze the Zillow Home Value Index (ZHVI) and Washington, DC crime data to compute Month-over-Month (MoM) and Year-over-Year (YoY) growth rates, and construct a Housing–Crime Index (HCI) with user-defined weights. The system architecture combines a PostgreSQL–PostGIS database (Supabase) with a GCP backend and AWS Bedrock AI Agent for natural-language queries. To enhance robustness, we plan to integrate the HouseTS dataset, which provides multimodal features such as socioeconomic indicators and points of interest for over 6,000 ZIP codes nationwide. Normalization, weighting, and validation processes are explicitly defined to ensure transparency. Ethical safeguards are implemented to prevent data misinterpretation and stigmatization. The result is a cloud-based, interactive platform that visualizes neighborhood trade-offs and supports responsible urban analytics.**

***Index Terms*—Crime, Housing Prices, Spatial Analysis, Zillow, DC Crime**

I. Introduction

Crime data plays an essential role in homebuyers' decision-making. Research shows that when platforms such as Zillow remove crime information, buyers may perceive neighborhoods as safer than they truly are, leading to overpriced but unsafe purchases [1], [2].

Studies illustrate that higher crime rates—especially burglary—are linked to declines in housing prices, with stronger effects in areas distant from city centers [1]. Hedonic pricing models further show that including crime data corrects omitted variable bias, allowing for more accurate property valuations [3]. Even small differences in crime rates significantly affect buyers' willingness to pay [4].

If crime data is hidden, buyers rely on incomplete information, causing market distortions. Spatial econometric analyses confirm that detailed local crime data is critical for understanding neighborhood price differences and risk factors [5]. Overall, crime data is not peripheral, but a central component of housing valuation.

Building on these insights, this project develops a data-driven decision-support system that integrates housing market indicators and crime data at the ZIP-code level for the Washington, DC metropolitan area. We preprocess the Zillow Home Value Index (ZHVI) and DC crime datasets, derive Month-over-Month (MoM) and Year-over-Year (YoY) growth rates, and map all incidents to ZIP codes using the Nominatim reverse-geocoding API. The processed data are stored in a Supabase PostgreSQL–PostGIS database to support spatial queries. An Angular web interface connects to a cloud-based backend on Google Cloud Run, which computes a Housing–Crime Index (HCI) and communicates with an AWS Bedrock AI Agent for natural-language analysis and visualization.

To enrich the analysis, we incorporate the HouseTS dataset, which provides additional socioeconomic and point-of-interest (POI) features for the same ZIP codes. This integration enhances the robustness of our index and supports multimodal validation. In future work, the same architecture will be extended beyond Washington, DC to enable cross-city comparison and long-term spatiotemporal housing analysis.

II. Related Work

Multiple studies have combined house pricing data with crime map data to analyze the impact of crime on property values. For instance, Ceccato and Wilhelmsson (2020) integrate geocoded housing transaction data with spatial crime data by identifying crime hot spots using Getis-Ord statistics and hedonic price modeling. Their analysis in the Stockholm metropolitan area demonstrates that proximity to crime hot spots, particularly those with high incidences of vandalism, negatively affects house prices [6].

Similarly, research by McIlhatton and colleagues (2016) employs spatial econometric models to merge detailed housing price data with geographically referenced crime incident data in a UK city. Their approach uses spatial lag and spatial error models within a hedonic pricing framework to capture the localized effects of different crime types on house prices. They emphasize the role of spatial autocorrelation and local crime clusters in shaping property values, thereby exemplifying the integration of house pricing data and crime map data in their analytical models [7].

A third related study is proposed by Zhang, Adepeju, and Thomas (2022), who outline an experimental design to evaluate the effects of publicly available street-level crime maps on house prices. Their protocol leverages the natural variation introduced by geomasking in police.uk crime maps and combines these with detailed house pricing records from the UK Land Registry Price Paid dataset. By exploiting the differences between the published, geomasked crime data and the true underlying data, they aim to isolate the causal impact of crime map information on housing values [8].

Based on the reviewed research, at least three distinct studies have been conducted that focus on using house pricing data combined with crime map data for analysis. Each study utilizes rigorous methods—ranging from spatial statistical techniques to natural experiment designs—to examine how the geographical distribution of crime correlates with property market outcomes.

A. Integration of Multimodal and Spatiotemporal Datasets

Beyond traditional econometric models, recent advances in large-scale multimodal datasets have enabled richer and more reproducible analysis of urban housing dynamics. The *HouseTS* dataset introduced by Wang et al. (2025) provides monthly housing data for over 6,000 ZIP codes across 30 U.S. metropolitan areas, incorporating socioeconomic indicators, points of interest (POI), and high-resolution satellite imagery. It serves as a benchmark for multimodal, long-term housing analysis, facilitating consistent preprocessing and feature alignment across diverse regions. Compared with previous single-source datasets such as Zillow or Redfin, HouseTS offers unified data quality, spatiotemporal continuity, and cross-city scalability. In this project, we integrate the Washington, DC subset of HouseTS with local crime data to construct and validate a composite Housing–Crime Index (HCI) that connects spatial, economic, and safety dimensions in a transparent way.

III. Proposed Approaches

Our project focuses on combining housing value trends with crime trends to create a composite decision index at the ZIP code level in Washington, DC. The approach involves the following components:

1) **Data Preprocessing:** We obtain the Zillow Home Value Index (ZHVI) data at the ZIP code level and compute both Month-over-Month (MoM) and Year-over-Year (YoY) growth rates. For crime data, we aggregate incidents from the DC open-data portal, extracting fields such as `OFFENSE`, `WARD`, `NEIGHBORHOOD_CLUSTER`, and geolocation coordinates. Using the Nominatim reverse-geocoding API, each incident is mapped to a ZIP code. The processed datasets are stored in a Supabase PostgreSQL–PostGIS database with spatial indexing to support queries and choropleth visualization.
2) **Normalization:** To ensure comparability between variables with different scales, all numeric features—such as housing growth rates and crime frequencies—are normalized to a [0, 1] range using min–max normalization:

$$\widetilde{x}_z = \frac{x_z - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

where x_z denotes the raw value for ZIP code z. This transformation enhances interpretability by allowing 0 to represent the least favorable condition and 1 the most favorable. While the HouseTS dataset applies log-normalization for model training, we adopt min–max normalization to maintain transparency and consistency for user-facing visualization and index computation.
3) **Index Construction:** We construct a composite Housing–Crime Index (HCI) that merges housing appreciation and safety levels with user-defined weights:

$$\mathrm{HCI}_z = w_1 \widetilde{G}_z + w_2 (1 - \widetilde{C}_z), \tag{2}$$

where $\widetilde{G}_z$ represents the normalized growth indicator (combining MoM and YoY rates) and $\widetilde{C}_z$ represents the normalized crime measure (rate per 1,000 residents). Users can adjust the weights w_1 and w_2 in the web interface to emphasize investment potential or residential safety.
4) **Integration with HouseTS:** To enrich contextual understanding, we integrate features from the *HouseTS* dataset such as median rent, per capita income, and POI densities corresponding to each ZIP code. These features are aligned with our local DC datasets to support correlation analysis and validation of the HCI results.
5) **Validation:** We plan to validate the HCI by comparing it against socioeconomic indicators (e.g., income, rent) from HouseTS and checking spatial consistency. Areas known to have high-value yet high-crime dynamics (e.g., rapidly developing neighborhoods) will serve as case studies for evaluating the index's interpretability and robustness.

IV. System Design

A. System Overview

Figure 1 presents the conceptual workflow of the proposed system, integrating the Zillow ZHVI and DC Crime datasets to compute a composite Housing–Crime Index (HCI). The system follows a three-layer design that connects data ingestion, processing, and interactive visualization.

Data Layer: This layer collects and stores the two main datasets. The Zillow Home Value Index (ZHVI) provides ZIP-code–level monthly home value indices, while the DC Crime dataset aggregates crime incidents geocoded by ZIP code. Both datasets are stored in a PostgreSQL–PostGIS database that supports spatial queries and geometry indexing.

Processing Layer: Data preprocessing aligns ZHVI and crime data by ZIP code, removes missing values, and normalizes the attributes. Two types of housing growth rates are computed:

- **Month-over-Month (MoM):** captures short-term fluctuations.
- **Year-over-Year (YoY):** captures long-term appreciation trends.

Crime incidents are aggregated by ZIP code and normalized by population (per 1,000 residents). The composite HCI index is constructed by combining the normalized housing growth ($\widetilde{G}_z$) and crime rate ($\widetilde{C}_z$) with user-defined weights:

$$\mathrm{HCI}_z = w_1 \widetilde{G}_z + w_2 (1 - \widetilde{C}_z), \tag{3}$$

where w_1 and w_2 represent the weights assigned to housing growth and safety respectively, satisfying $w_1 + w_2 = 1$. The subcomponents G_z and C_z are further defined as:

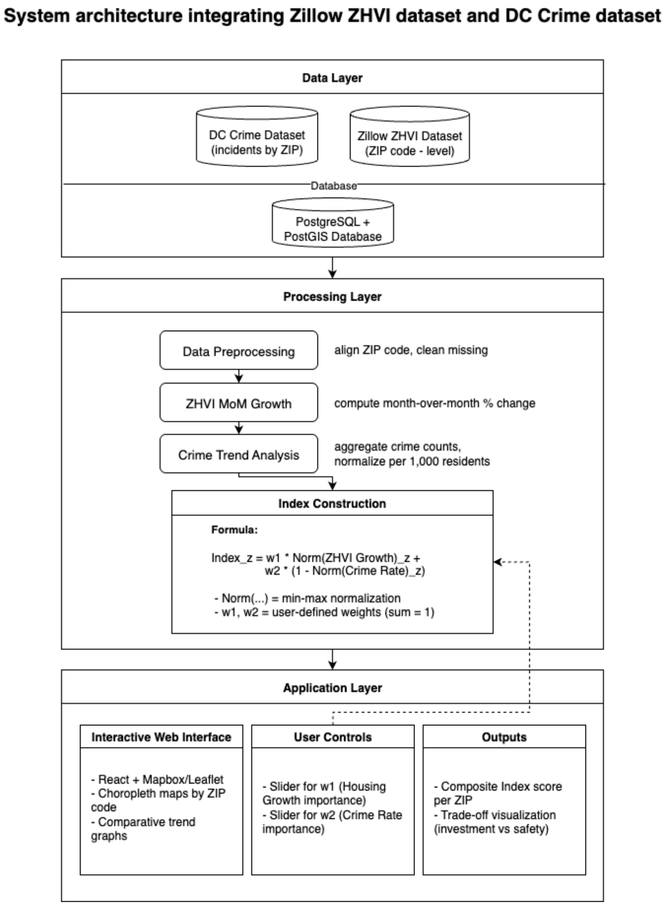

System architecture integrating Zillow ZHVI dataset and DC Crime dataset
Data Layer
DC Crime Dataset (incidents by ZIP)
Zillow ZHVI Dataset (ZIP code - level)
Database
PostgreSQL + PostGIS Database
Processing Layer
Data Preprocessing
align ZIP code, clean missing
ZHVI MoM Growth
compute month-over-month % change
Crime Trend Analysis
aggregate crime counts, normalize per 1,000 residents
Index Construction
Formula:
Index_z = w1 * Norm(ZHVI Growth)_z + w2 * (1 - Norm(Crime Rate)_z)
- Norm(...) = min-max normalization
- w1, w2 = user-defined weights (sum = 1)
Application Layer
Interactive Web Interface
- React + Mapbox/Leaflet
- Choropleth maps by ZIP code
- Comparative trend graphs
User Controls
- Slider for w1 (Housing Growth importance)
- Slider for w2 (Crime Rate importance)
Outputs
- Composite Index score per ZIP
- Trade-off visualization (investment vs safety)

Fig. 1: Conceptual system architecture integrating Zillow ZHVI dataset and DC Crime dataset.

$$G_z = \alpha\,\widetilde{\mathrm{YoY}}_z + (1-\alpha)\,\widetilde{\mathrm{MoM}}_z, \tag{4}$$

$$C_z = \beta\,\widetilde{\mathrm{CrimeRate}}_z + (1-\beta)\,\widetilde{\mathrm{CrimeTrend}}_z, \tag{5}$$

and each variable is normalized into a [0, 1] interval using min–max scaling:

$$\widetilde{x}_z = \frac{x_z - \min(x)}{\max(x) - \min(x)}. \tag{6}$$

This normalization scheme provides interpretability by ensuring that 0 represents the least favorable condition and 1 represents the most favorable one.

Application Layer: The user interacts with the system through a web-based interface built with React and Mapbox/Leaflet. A pair of sliders allows users to adjust w_1 and w_2 dynamically to reflect personal preferences (e.g., prioritizing investment potential or safety). Outputs include ZIP-code–level choropleth maps, comparative trend graphs, and scenario-based recommendations.

—

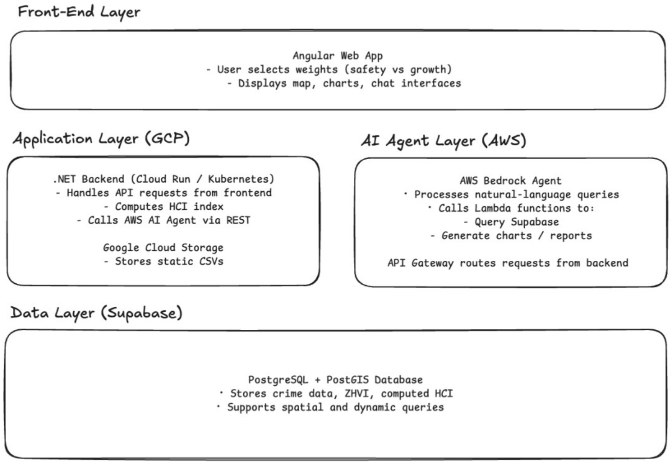
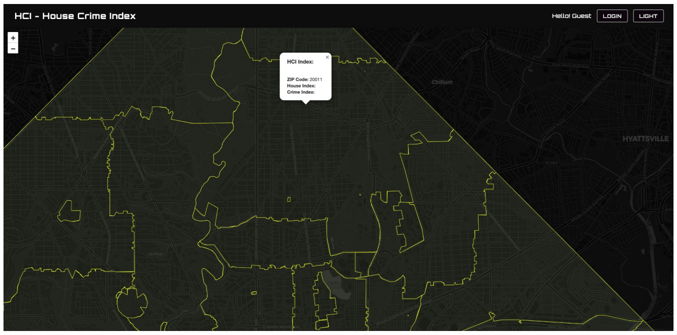

Front-End Layer
Angular Web App
- User selects weights (safety vs growth)
- Displays map, charts, chat interfaces
Application Layer (GCP)
.NET Backend (Cloud Run / Kubernetes)
- Handles API requests from frontend
- Computes HCI index
- Calls AWS AI Agent via REST
Google Cloud Storage
- Stores static CSVs
AI Agent Layer (AWS)
AWS Bedrock Agent
· Processes natural-language queries
· Calls Lambda functions to:
- Query Supabase
- Generate charts / reports
API Gateway routes requests from backend
Data Layer (Supabase)
PostgreSQL + PostGIS Database
· Stores crime data, ZHVI, computed HCI
· Supports spatial and dynamic queries

Fig. 2: Four-layer implementation architecture using Supabase, GCP, and AWS.



HCI Index:

ZIP Code: 20011
House Index:
Crime Index:

Fig. 3: Front-end user interface of the Housing–Crime Index (HCI) web application. The Angular + Mapbox interface visualizes Washington, DC ZIP code boundaries, allowing users to explore housing and crime trends, adjust weight preferences, and toggle between light/dark themes. Neutral color tones and contextual tooltips are used to promote ethical visualization.

B. Implementation and Technologies

To support scalability and AI integration, the system is deployed as a four-layer cloud-based architecture across Supabase, Google Cloud Platform (GCP), and Amazon Web Services (AWS), as shown in Figure 2. Each layer is modular, containerized, and independently deployable, ensuring flexibility and maintainability.

Front-End Layer (Angular Web App): Implements an interactive user interface using Angular and Mapbox, as illustrated in Figure 3. Users can adjust the index weights (safety vs. growth), visualize ZIP-level scores, and access a chat interface powered by the AI Agent for natural-language insights.

Application Layer (GCP Backend): A .NET backend deployed on Google Cloud Run handles RESTful API requests from the frontend. It computes the HCI values, manages user sessions, and communicates with the AWS AI Agent through secured API calls. Google Cloud Storage hosts processed

TABLE I: Technology Stack Summary.

Layer	Technologies / Services	Purpose
Front-End	Angular, TypeScript, Mapbox	Interactive visualization, weight adjustment, and chat interface
Application (GCP)	.NET Core, Cloud Run, Google Cloud Storage	Compute HCI, manage API requests, and cache processed datasets
AI Agent (AWS)	AWS Bedrock, Lambda, API Gateway	Handle natural-language queries, generate charts and summaries
Data Layer	Supabase (PostgreSQL + PostGIS)	Store spatial data, support geospatial queries, and maintain normalized metrics

CSVs for caching and efficient retrieval. **AI Agent Layer (AWS Bedrock):** The AWS Bedrock Agent processes natural-language queries, calls AWS Lambda functions to query Supabase, and generates visual summaries or recommendations. API Gateway manages secure communication between the backend and Bedrock services. **Data Layer (Supabase):** Supabase hosts a PostgreSQL–PostGIS database that stores housing and crime data, computed HCI scores, and geographic geometries. Spatial indexing (GIST) and materialized views improve query efficiency, while SQL triggers maintain data consistency.

This hybrid GCP–AWS architecture decouples computation, data management, and AI interaction. The GCP backend focuses on orchestrating index computation and frontend requests, while AWS provides elastic AI processing and visualization generation. Supabase serves as the unified data layer for persistent storage and spatial queries, ensuring transparency, reproducibility, and scalability for future integration with multimodal datasets such as HouseTS.

V. Database Design and Schema Optimization

A. *Database Design Overview*

The project utilizes a relational database architecture hosted on Supabase (PostgreSQL + PostGIS) to store, normalize, and query housing and crime data at the ZIP-code level. The database schema is designed for modularity and scalability, supporting spatial joins and analytical queries across multiple datasets such as ZHVI, DC Crime, and HouseTS.

Figure **??** illustrates the entity–relationship (ER) design. Each ZIP code serves as a primary linkage key between datasets, enabling efficient spatial aggregation and comparative analysis. The design supports both temporal and spatial queries through indexing and caching mechanisms described in Section V-B.

B. *Schema Details*

The database consists of three major tables: `crime_data`, `zhvi_data`, and `composite_index`. Each table is normalized to reduce redundancy and optimized for spatial queries.

1) Crime Data Schema:

2) Housing Data Schema:

3) Composite Index Schema:

TABLE II: DC Crime Data Schema (2025).

Field Name	Type	Description
X, Y	FLOAT	Projected coordinates (State Plane)
CCN	VARCHAR(20)	Case Control Number (unique police report ID)
REPORT_DAT	TIMESTAMP	Date and time when the incident was reported
SHIFT	VARCHAR(20)	Police shift (DAY, EVENING, MIDNIGHT)
OFFENSE	VARCHAR(100)	Type of crime (e.g., THEFT/OTHER, BURGLARY)
BLOCK	VARCHAR(100)	Street block of the incident
WARD, DISTRICT	INTEGER	DC political and police subdivisions
LATITUDE, LONGITUDE	FLOAT	Geographic coordinates
ZIPCODE	VARCHAR(10)	ZIP code derived using Nominatim API
START_DATE, END_DATE	TIMESTAMP	Start and end time of incident

TABLE III: Zillow ZHVI Data Schema.

Field Name	Type	Description
ZIPCODE	VARCHAR(10)	ZIP code identifier
REGIONNAME, COUNTYNAME	VARCHAR(50)	Regional and county labels
MOM	FLOAT	Month-over-month growth rate (%)
YOY	FLOAT	Year-over-year growth rate (%)
CURRENTPRICE	FLOAT	Latest ZHVI value (USD)

TABLE IV: Composite Housing–Crime Index Schema.

Field Name	Type	Description
ZIPCODE	VARCHAR(10)	ZIP code identifier (primary key)
HCI_SCORE	FLOAT	Computed housing–crime composite index
WEIGHT_GROWTH	FLOAT	Weight assigned to housing growth
WEIGHT_SAFETY	FLOAT	Weight assigned to safety/crime rate
UPDATED_AT	TIMESTAMP	Last computation timestamp

C. *Optimization Strategies*

To ensure scalability and fast query performance, multiple optimization techniques were implemented:

- **Spatial Indexing:** Each dataset containing geographic data (latitude/longitude or ZIP polygons) employs a GIST index on its geometry column, significantly improving query performance for spatial joins and map rendering:

```
CREATE INDEX crime_geom_idx
ON crime_data USING GIST (geom);
```

Listing 1: Creating a GIST index for spatial queries

- **Materialized Views:** Frequently used aggregated results, such as monthly crime counts and average HCI per ZIP code, are precomputed and stored as materialized views to minimize redundant computation during API requests.
- **Caching Strategy:** Static CSV files generated from preprocessing (ZHVI and crime summaries) are stored in Google Cloud Storage and fetched by the GCP backend when needed, reducing latency and database load during repeated visualizations.
- **Data Consistency and Triggers:** PostgreSQL triggers automatically update derived metrics (e.g., recomputed HCI scores) when source tables are modified. This guarantees data consistency between raw datasets and computed indices.

This design enables fast data retrieval for interactive map rendering and AI-assisted querying, while maintaining normalization, consistency, and scalability for future integration of multimodal datasets such as HouseTS.

VI. Ethical and Data Reliability Considerations

A. Ethical Presentation of Crime Data

While combining crime and housing data provides valuable insights for homebuyers and investors, it also raises important ethical challenges. Crime statistics are often unevenly reported across neighborhoods, and their visualization can unintentionally reinforce stereotypes or stigmatize communities. To mitigate this risk, our system avoids labeling any ZIP code as "safe" or "unsafe." Instead, it presents continuous normalized indicators, emphasizing comparative trends rather than absolute rankings.

In addition, choropleth maps are designed using neutral color palettes (e.g., blue–gray scales) instead of alarming red tones to avoid misleading visual emphasis. Descriptive tooltips clarify that crime data reflect reported incidents, not verified convictions, and should be interpreted as relative measures of risk rather than definitive safety indicators.

B. Fairness and Contextualization

The project explicitly integrates contextual socioeconomic variables to promote fairness and reduce data misinterpretation. By referencing supplementary datasets such as HouseTS—containing information on population density, income levels, and points of interest—the visualizations help users interpret housing and crime trends within a broader social context. This approach ensures that the platform encourages informed decision-making rather than simplistic judgments about neighborhood safety.

C. Data Reliability and Limitations

Both the ZHVI and DC Crime datasets have inherent reliability constraints. Zillow data may underrepresent rapid price fluctuations in smaller markets, while crime data are subject to underreporting and inconsistent classification. For instance, police records may vary in completeness across different wards, and temporal lags can occur between the incident date and official publication. To account for these issues, our system performs outlier detection, temporal smoothing, and missing-value interpolation during preprocessing. All derived metrics are accompanied by timestamps and data-source metadata to ensure transparency.

D. Transparency and Responsible Use

The web application includes a dedicated "Data Disclaimer" section that outlines the purpose and limitations of the analysis. Users are reminded that the Housing–Crime Index (HCI) is a relative indicator intended for exploratory comparison, not as an absolute measure of neighborhood desirability. All data sources are publicly available and licensed under open data terms, and the processing code will be released for academic reproducibility. This transparency ensures that the project aligns with responsible data science practices and upholds ethical standards in public data visualization.

TABLE V: Team Progress Summary.

Team Member	Work Completed	Next Steps
Hsuan-Chi Chang	Implemented data preprocessing pipeline for ZHVI and DC Crime datasets. Integrated Nominatim API to map coordinates to ZIP codes and imported cleaned data into Supabase. Set up PostgreSQL–PostGIS schema and tested spatial queries.	Integrate additional socioeconomic variables from HouseTS (median rent, income, POI density). Conduct validation and correlation analysis between HCI and HouseTS metrics.
Wei-Ju Li	Developed the Angular frontend, including choropleth map visualization and weight-adjustable sliders. Implemented responsive user interface and connected frontend with backend API endpoints.	Integrate AWS Bedrock AI Agent for natural-language queries and improve visualization ethics (disclaimer panel, color scheme). Finalize user interaction testing.
Both Members	Collaboratively defined system architecture and HCI computation methodology. Wrote initial proposal and extended sections on normalization, ethics, and database optimization.	Perform system integration between GCP backend and AWS AI layer. Prepare final presentation and documentation for submission.

VII. Progress Report and Next Steps

A. Progress Report

Table V summarizes the progress achieved by each team member as of the current checkpoint. Both members have contributed to system design, data integration, and validation, following an iterative and collaborative workflow.

B. Next Steps

Over the next development phase, the team will focus on three main tasks:

1) **Validation and Testing:** Conduct cross-validation of HCI results with socioeconomic indicators from HouseTS and ensure accuracy of spatial joins and normalization.
2) **Frontend Integration:** Complete full linkage between the Angular interface, GCP backend, and AWS AI Agent, enabling interactive and conversational analytics.
3) **Performance Optimization:** Implement caching and materialized views to reduce query latency and prepare the system for final demonstration and report submission.

The team maintains a collaborative development workflow through version-controlled repositories and weekly progress meetings, ensuring consistent progress and clear task ownership toward project completion.

References

[1] V. Ceccato and M. Wilhelmsson, "The impact of crime on apartment prices: Evidence from stockholm, sweden," *Urban Studies*, vol. 41, no. 13, pp. 2873–2894, 2004.

[2] E. G. Goetz, "Information transparency in housing markets: Zillow, crime data, and the limits of online platforms," *Housing Policy Debate*, vol. 29, no. 3, pp. 473–488, 2019.

[3] R. Thaler, "A note on the value of crime control: Evidence from the property market," *Journal of Urban Economics*, vol. 5, no. 1, pp. 137–145, 1978.

[4] S. Gibbons, “The value of good neighbours: A study of the impact of crime on house prices,” *Urban Studies*, vol. 41, no. 1, pp. 105–123, 2004.

[5] G. E. Tita, T. L. Petras, and R. T. Greenbaum, “Crime and residential choice: A neighborhood level analysis of the impact of crime on housing prices,” *Journal of Quantitative Criminology*, vol. 22, no. 4, pp. 299–317, 2006.

[6] V. Ceccato and M. Wilhelmsson, “Do crime hot spots affect housing prices?” *Nordic Journal of Criminology*, 2020.

[7] D. McIlhatton, W. McGreal, P. T. de la Paz, and A. Adair, “Impact of crime on spatial analysis of house prices: evidence from a uk city,” *International Journal of Housing Markets and Analysis*, vol. 9, no. 4, pp. 627–647, 2016.

[8] M. L. Zhang, M. Adepeju, and R. Thomas, “Estimating the effects of crime maps on house prices using an (un)natural experiment: A study protocol,” *PLOS ONE*, vol. 17, no. 12, p. e0278954, 2022.

Appendix

Appendix A: Task Assignment

- **Hsuan Chi Chang:** Responsible for ZHVI data preprocessing, YoY growth calculations, and database setup. Focuses on DC Crime dataset aggregation, trend analysis, and normalization by ZIP code.
- **Wei Ju Li:** Develops the frontend interface, including map visualization and index weighting controls.
- **All Members:** Collaborate on system integration, testing, and report writing.

Appendix B: Schedule

- **Week 1:** Literature review and dataset exploration.
- **Week 2–3:** Data preprocessing, alignment of ZHVI and crime datasets.
- **Week 4:** Calculation of YoY growth and crime trends by ZIP code.
- **Week 5:** Index formula design and initial testing.
- **Week 6:** Frontend map visualization and user interface development.
- **Week 7:** System integration, validation, and final presentation/report preparation.