NESTOR

The Nest Advisor

Sidratul Muntaha Antara

Table of Contents

1. Abstract

Nestor, a data-driven real estate recommendation engine developed to assist prospective homebuyers in identifying optimal housing markets based on their individual preferences and priorities. Nestor leverages publicly available datasets encompassing home prices, school quality indices, crime statistics, employment metrics, and healthcare access to evaluate and rank U.S. states by overall livability and affordability.

Unlike traditional real estate platforms that focus on static listings, Nestor applies a customizable scoring framework — the Desirability Score — to deliver tailored housing recommendations. This composite score integrates multiple factors that influence housing desirability, allowing users to filter and prioritize regions using criteria such as budget constraints, bedroom count, safety preferences, and school ratings.

The core functionality of Nestor is implemented as a web-based interactive application using Streamlit, enabling real-time exploration of housing recommendations. Users are provided with top-ranked regions, visual insights through charts, downloadable filtered results, and detailed region-level snapshots.

Key Outcomes and Insights:

- States such as West Virginia and Oklahoma consistently rank highly due to a favorable combination of affordability and school quality.
- The Desirability Score framework allows users to understand trade-offs between affordability, safety, and educational access.
- The platform successfully demonstrates the feasibility of an integrated housing recommendation tool that aligns with both economic data and user-specific lifestyle criteria.

Nestor offers a scalable, modular approach that can be further enhanced with real-time APIs, machine learning models, or expanded to metro-level housing data in future iterations.

2. Background and Problem Statement

The homebuying process represents one of the most significant financial and lifestyle decisions an individual or family can undertake. While numerous online platforms such as Zillow and Redfin offer access to extensive real estate listings, they largely focus on superficial or single-dimensional filters — primarily price, location, and property features. These tools lack the analytical depth required to evaluate broader socioeconomic factors that heavily influence quality of life, such as employment opportunities, public safety, school performance, and healthcare accessibility.

In today's data-rich environment, homebuyers are often overwhelmed by fragmented and unstructured information. The absence of an integrated system that consolidates multiple regional indicators into a single, actionable score leads to decision fatigue, inefficient research, and potentially suboptimal choices.

Problem Definition

Despite the availability of data, there is no accessible solution that synthesizes critical variables such as housing affordability, crime rates, job market strength, and educational quality into a unified framework tailored to individual preferences. Consequently, homebuyers lack the ability to objectively evaluate and compare locations in a way that balances personal needs with quantitative indicators.

Project Motivation

This project introduces Nestor, a real estate recommendation engine designed to bridge that gap. Nestor aggregates and analyzes key housing, economic, and social indicators at the state level and generates ranked recommendations based on user-defined preferences. The platform aims to simplify the decision-making process for prospective homebuyers by providing a transparent, interactive, and data-driven tool that aligns housing options with personalized lifestyle criteria.

3. Objectives and Scope

3.1 Project Objectives

This capstone project aims to develop Nestor, a data-driven recommendation system that assists homebuyers in identifying U.S. states that best fit their lifestyle and financial priorities. Unlike traditional property search platforms, Nestor provides an interactive decision-support tool that evaluates regions based on multiple socioeconomic factors.

Key objectives include:

- Integrating diverse datasets on housing affordability, employment, education, safety, and healthcare.
- Standardizing and normalizing data for consistent cross-region comparisons.
- Designing a Desirability Score that ranks states based on user-defined preferences.
- Delivering an interactive Streamlit dashboard that enables users to:
    - ➢ Filter regions by budget, bedrooms, school rating, crime rate, and healthcare access.
    - ➢ View ranked recommendations and detailed region summaries.
    - ➢ Export customized results for further analysis.

3.2 Project Scope

The project focuses on state-level housing recommendations across the United States. It evaluates five key dimensions:

Metric Category	Key Variables
Housing Affordability	Median Home Price
Employment Factors	Job Growth Rate, Unemployment Rate
Education Quality	School Performance Index (HUD Data)
Public Safety	Combined Violent Crime Rate (Murder, Assault, Rape)
Healthcare Access	Simulated Accessibility Score (1–10)

Currently, the scope excludes individual property-level analysis, live API data, and advanced machine learning models. These areas are identified as future opportunities for scaling and refinement.

4. Data Collection and Sources

The Nestor recommendation engine integrates publicly available datasets representing key socioeconomic factors that influence residential decision-making. Data was collected from trusted government and industry sources to evaluate affordability, employment, education, safety, and healthcare access across U.S. states.

4.1 Overview of Data Sources

Category	Data Source	Key Variables
Housing Affordability	Zillow Research / Kaggle	Median Home Price (HomePrice)
Employment Indicators	Federal Reserve Economic Data (FRED)	Unemployment Rate, Job Growth Rate
Education Quality	U.S. Department of Housing and Urban Development (HUD)	School Proficiency Index (SchoolRating)
Crime and Public Safety	FBI Crime Reports / Kaggle datasets	Murder, Assault, Rape rates (CrimeRate)
Healthcare Access	Simulated Data (Prototype placeholder)	Accessibility Score (HealthcareAccess)

4.2 Data Integration and Structure

All datasets were processed and merged into a unified structure where each row represents a state-month combination, with associated metrics. The merged dataset contains approximately 3,000 observations and was saved as combined_data_final.csv for use in the Streamlit application.

Note: The healthcare access metric was simulated for this prototype due to the lack of consistent state-level public data. Future versions aim to integrate real healthcare access data from verified sources.

5. Data Preparation and Cleaning

Following data collection, multiple preprocessing steps were applied to ensure consistency, accuracy, and comparability across regions. These steps included merging datasets, handling missing values, feature transformation, and the construction of a composite scoring model.

5.1 Dataset Consolidation and Integration

All datasets were merged based on state and monthly timeframes.
Key preprocessing tasks included:

- Standardizing state names and date formats.
- Removing duplicates and correcting inconsistencies.
- Aligning variables across all records for consistency.

5.2 Handling Missing and Incomplete Data

Some variables, such as healthcare access and school quality, had gaps or incomplete records. These were addressed as follows:

- Healthcare Access: Simulated scores were used as placeholders due to the lack of consistent state-level public data.
- School Quality: Missing values were filled using the median score across states.
- Crime Data: Incomplete records were excluded from scoring to avoid bias.

5.3 Feature Normalization

To ensure fair comparison, all key variables were normalized on a scale of 1 to 10. This included affordability, education, and safety metrics. Normalization ensured that no single factor dominated the scoring process due to scale differences.

5.4 Construction of the Desirability Score

To synthesize insights across key housing criteria, a **Desirability Score** was formulated. This score reflects a weighted combination of affordability, education, and safety indicators, calibrated to reflect their perceived importance in home selection.

- **Affordability**: 40% weight
- **Education**: 30% weight
- **Safety**: 30% weight

This score enables users to compare states holistically and serves as the primary basis for recommendations within the Nestor platform.

5.5 Final Prepared Dataset

The final cleaned dataset included essential columns such as:

- Regional identifiers (State, RegionName, Date)
- Core metrics (HomePrice, UnemploymentRate, JobGrowthRate, SchoolRating, CrimeRate, HealthcareAccess)
- Engineered outputs (Bedrooms, DesirabilityScore)

This processed file, titled combined_data_final.csv, was used as the input source for the Nestor recommendation engine and visual dashboard.

6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the underlying distributions, trends, and relationships among the variables used in the recommendation engine. This stage was instrumental in validating assumptions, identifying outliers, and uncovering patterns that guided the final scoring model and dashboard design.

6.1 Descriptive Statistics

Initial descriptive analysis revealed significant variation across U.S. states in terms of home prices, educational quality, crime rates, and job growth:

- Home Prices ranged from under $110,000 in economically modest states to nearly $1 million in more affluent regions such as California and Massachusetts.
- Unemployment Rates spanned a wide range, with some states maintaining levels below 3%, while others experienced temporary spikes above 10% during recessionary periods.
- School Ratings exhibited noticeable clustering, with several southern states averaging below the national median, while northeastern and midwestern states tended to score higher.
- Crime Rates were highly variable, with urbanized states showing elevated levels of violent crimes.

These differences underscore the importance of normalization to facilitate equitable comparisons across all metrics.

6.2 Geographic Patterns

The analysis revealed that:

- High-scoring states like West Virginia, Kentucky, and Oklahoma balanced affordability, safety, and education.

- Lower-scoring states such as California and New York were penalized due to high housing costs and crime rates, despite strong job markets and school systems.

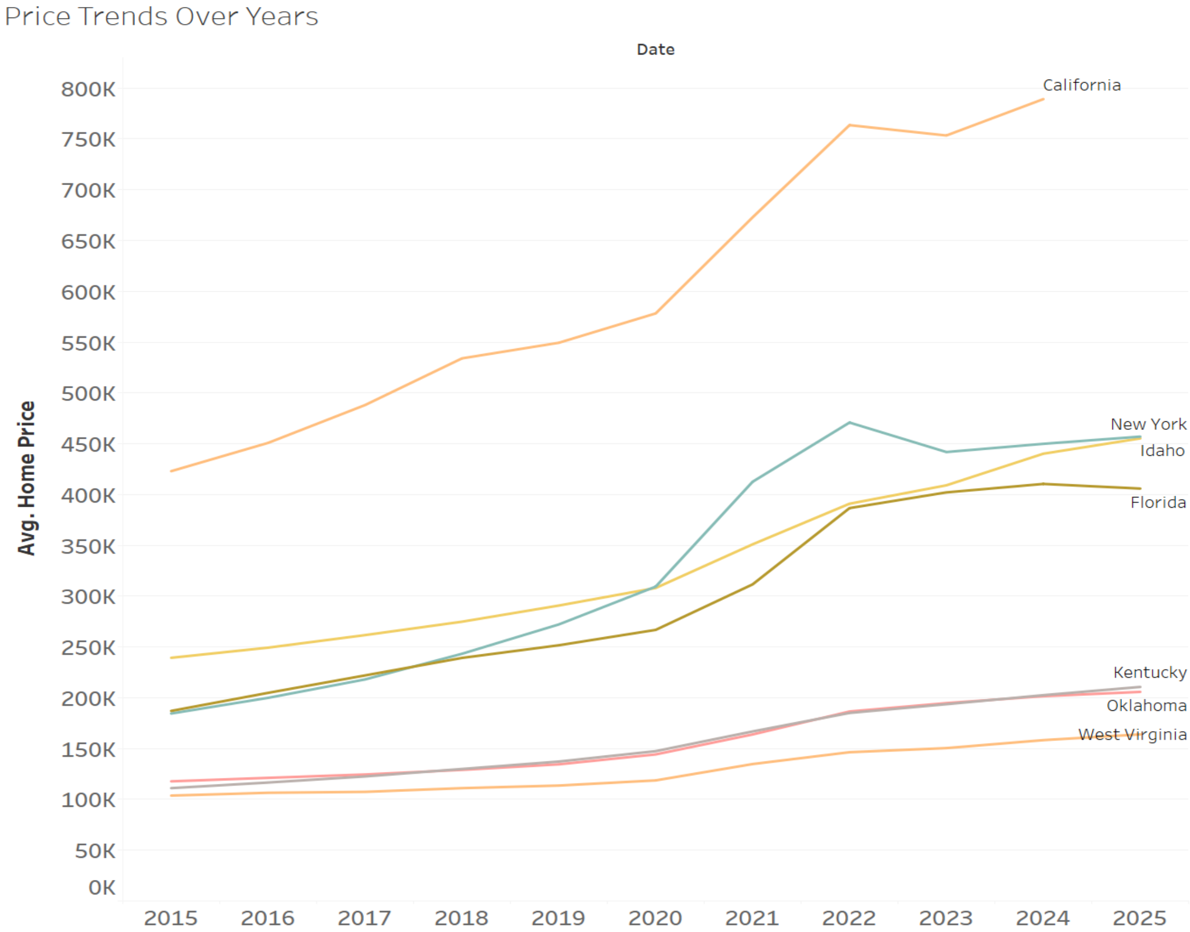

Fig 6.2: Price Trends Over Years in New York, California (Average Higher Price), Kentucky, Oklahoma, West Virginia (Average Lower Price)

6.3 Correlation Insights

Correlations were identified among key variables:

- Negative correlation between home prices and desirability, confirming affordability as a major driver.
- Positive correlation between school quality and desirability, particularly in safer regions.
- Negative impact of high crime rates on overall rankings.

These insights inform the final weight assignments in the scoring model.

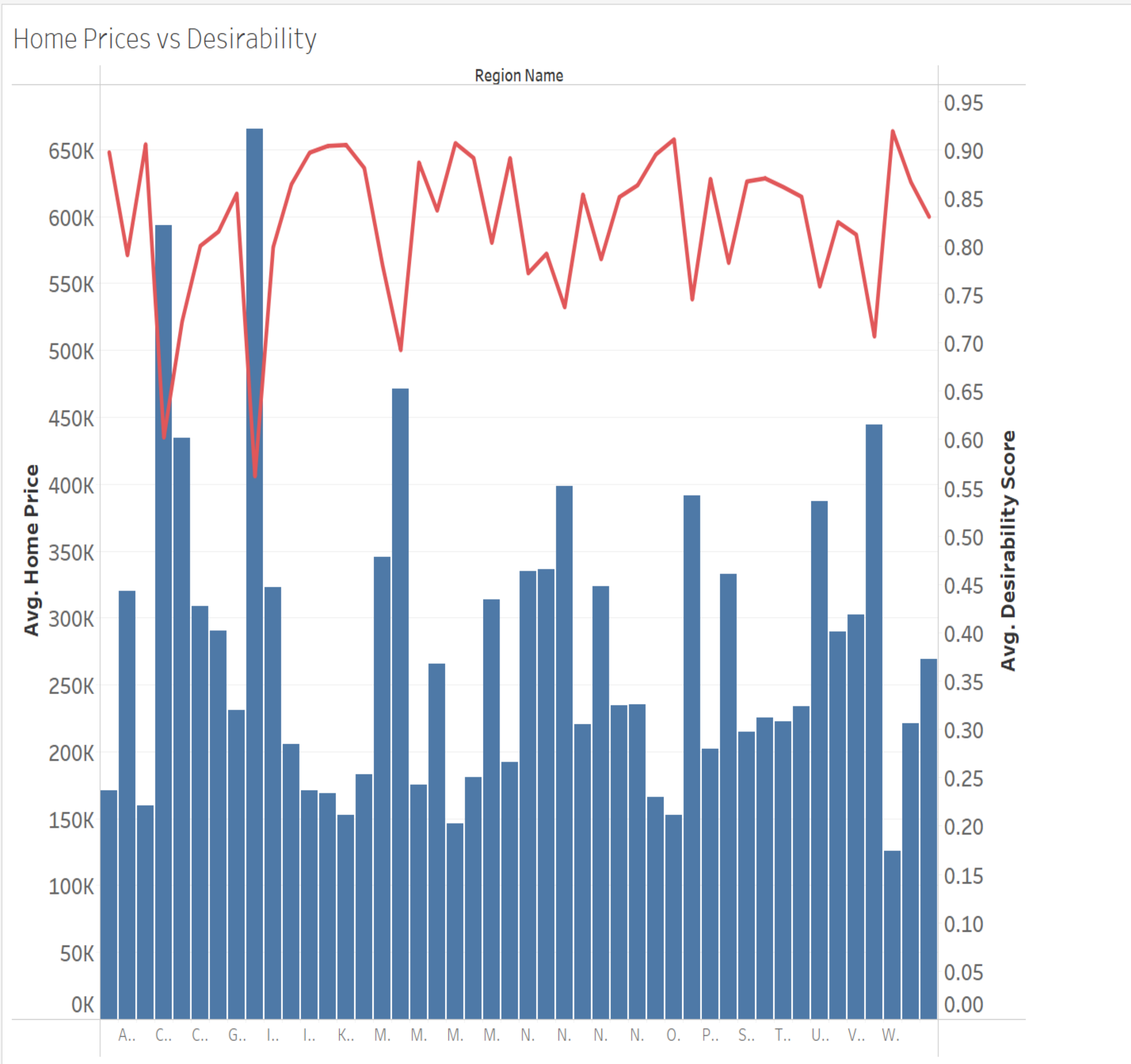

Fig 6.3.1 : Dual Axis Plot showing how prices affect the desirability score

The chart shows an inverse relationship between average home prices and the desirability score across regions. States with lower housing costs tend to achieve higher desirability scores, reinforcing affordability as a key driver in housing decisions.

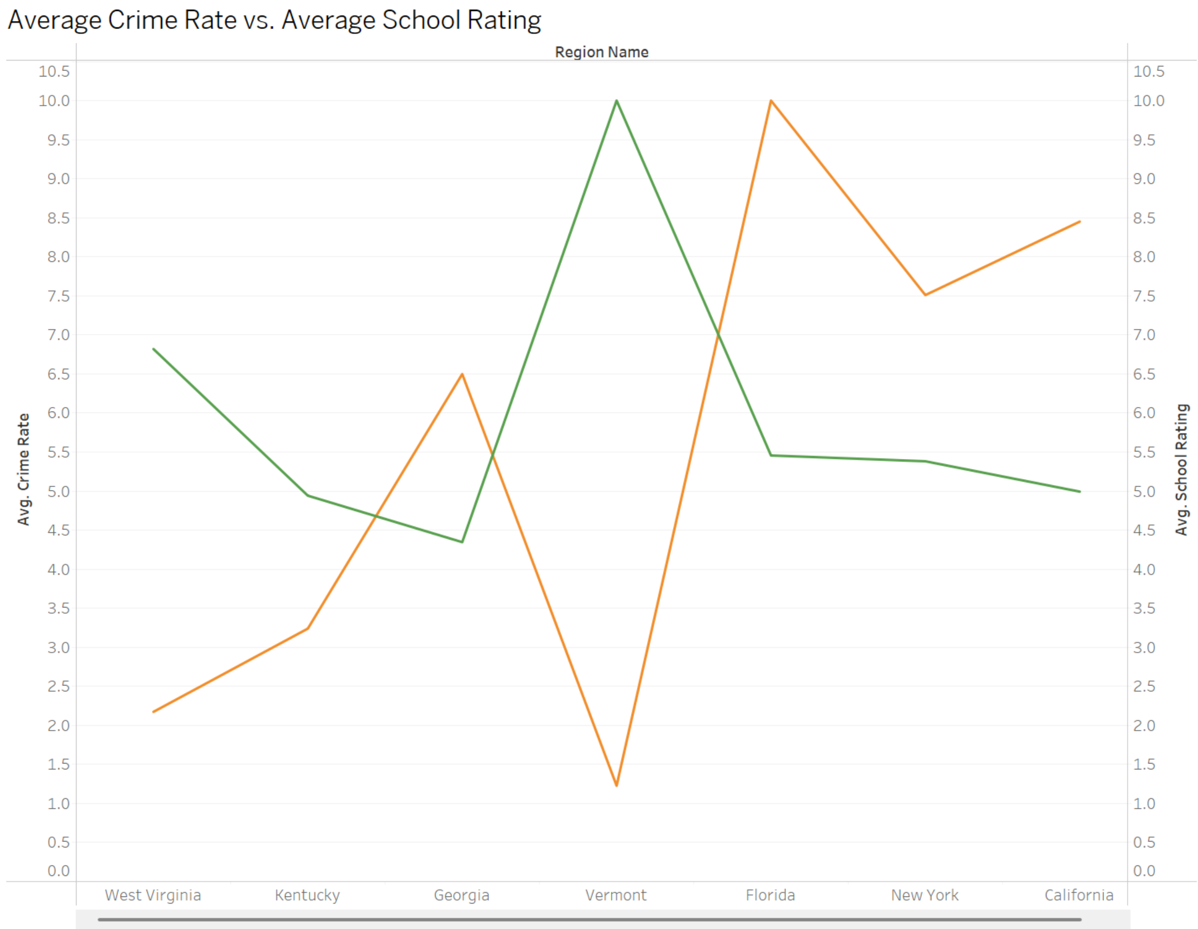

Fig 6.3.2: Comparison of Average Crime Rate vs. Average School Rating across Selected States. This chart highlights the trade-offs between safety and education quality in West Virginia, Kentucky etc.

6.4 Distribution of Desirability Scores

The distribution of the final DesirabilityScore was right-skewed, with most states falling in the 0.80–0.90 range. A few exceptional states achieved scores above 0.93, driven by well-balanced profiles across affordability, education, and safety.

(We can see it in Fig 6.3.1)

This distribution also reinforced the importance of ranking, as even small differences in the score could meaningfully shift a state's position in the recommendation output.

6.5 Visual Highlights from the Dashboard

The Streamlit dashboard includes several key visualizations derived from this analysis:

- Bar Chart: Displays the top regions based on user-defined criteria and updated in real-time.
- Scorecard: Highlights the top recommended state along with its Desirability Score.
- Region Snapshot: Presents a breakdown of affordability, safety, and education metrics for a selected region.

- Downloadable Table: Provides a ranked list of all matching states for user review and export.

These components collectively offer an intuitive, data-informed interface for housing decision-making.

7. Recommendation Engine Architecture

At the heart of Nestor is a multi-criteria recommendation engine designed to evaluate and rank U.S. states based on user-defined lifestyle and financial preferences. Unlike traditional platforms that offer basic filtering, Nestor delivers personalized, ranked recommendations using a transparent scoring framework.

7.1 Design Philosophy

The recommendation engine was built on three guiding principles:

1. **Personalization**: Enable users to customize key criteria such as budget, school quality, crime tolerance, and healthcare access.
2. **Interpretability**: Ensure the scoring system is transparent and easy to understand, with clearly defined component scores.
3. **Responsiveness**: Deliver real-time recommendations via an interactive dashboard interface, with updated rankings and visualizations upon each input change.

7.2 Input Parameters

Users can interact with the app via sidebar filters to define their preferences:

- Budget Range: Minimum and maximum home price
- Preferred Number of Bedrooms
- Minimum School Rating
- Maximum Acceptable Crime Rate
- Minimum Healthcare Access Score

These inputs are applied as dynamic filters to reduce the dataset to only those states meeting user-defined criteria.

7.3 Scoring Framework

The recommendation engine uses a composite metric called the Desirability Score to rank matching states. This score integrates three major dimensions:

Dimension	Description	Weight
Affordability	Inverse of median home price (normalized)	40%
Education Quality	Scaled school proficiency index from HUD	30%
Public Safety	Inverse of combined violent crime rate (normalized)	30%

Each dimension is scaled from 1 to 10 to ensure proportional weighting. The score is recalculated dynamically as users change their filters, allowing for highly personalized rankings.

7.4 Output and Ranking

Once filtered and scored, states are sorted in descending order by their Desirability Score. The top result is displayed as a "Top Pick", followed by a ranked table of all matching states. Additional outputs include:

- A horizontal bar chart of the top five states.
- A summary card for the selected region, including key metrics.
- A downloadable CSV containing filtered results for offline use.

This structured output allows users to quickly interpret, compare, and act on housing recommendations.

7.5 Scalability and Future Enhancements

While the current implementation uses a weighted scoring algorithm, the architecture is modular and can accommodate future enhancements such as:

- Machine learning–based personalization (e.g., regression or ranking models)
- User profile integration for long-term tracking
- Granular location scoring (city or ZIP code level)
- API integration with live housing or census data

These improvements would further increase the sophistication and real-world applicability of the platform.

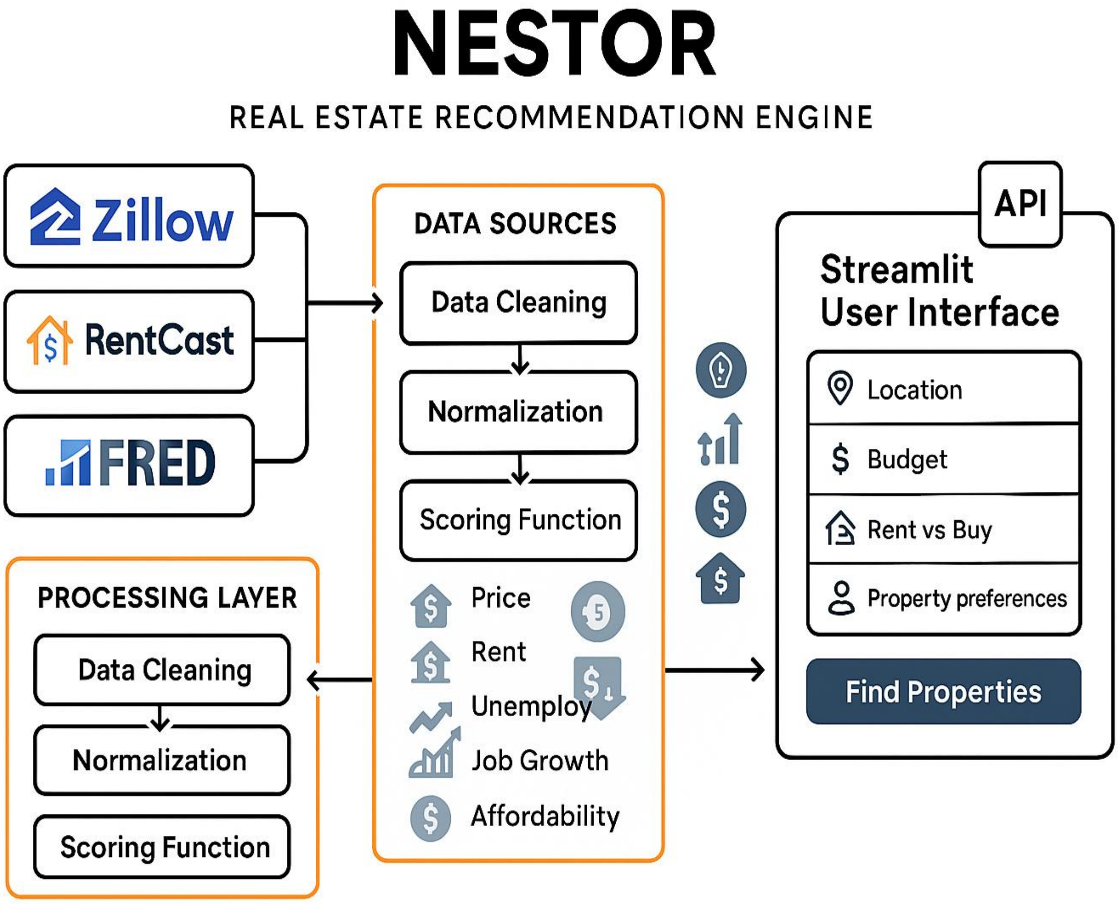

Figure 7.5 : Nestor's system architecture illustrating how data is sourced, processed, and delivered through the user-facing Streamlit web application.

8. App Implementation

To provide an intuitive and interactive user experience, the Nestor recommendation engine was deployed as a web-based application using Streamlit, an open-source Python framework for building custom data dashboards. The platform supports real-time filtering, scoring, and visualization, making it accessible to users with no technical background.

8.1 Technology Stack

Component	**Technology Used**
Web Interface	Streamlit (Python)
Data Processing	Pandas, NumPy
Scoring & Normalization	Scikit-learn
Visualization	Matplotlib, Streamlit Charts
Data Format	Pre-processed CSV File

Streamlit was chosen for its simplicity, responsiveness, and tight integration with the Python data science ecosystem. It enables rapid deployment without the need for front-end development expertise.

8.2 User Interface Design

The dashboard consists of:

• **Sidebar Filters :** Users can define preferences such as budget range, number of bedrooms, school rating, crime tolerance, and healthcare access.

• **Main Display Panel :** Displays real-time results including:

➢ Top Pick Summary

➢ Ranked Table of Matching States

➢ Bar Chart of Top 5 Regions

➢ Region Summary Cards

• **Export Functionality**

Users can download their personalized recommendations in CSV format for offline review.

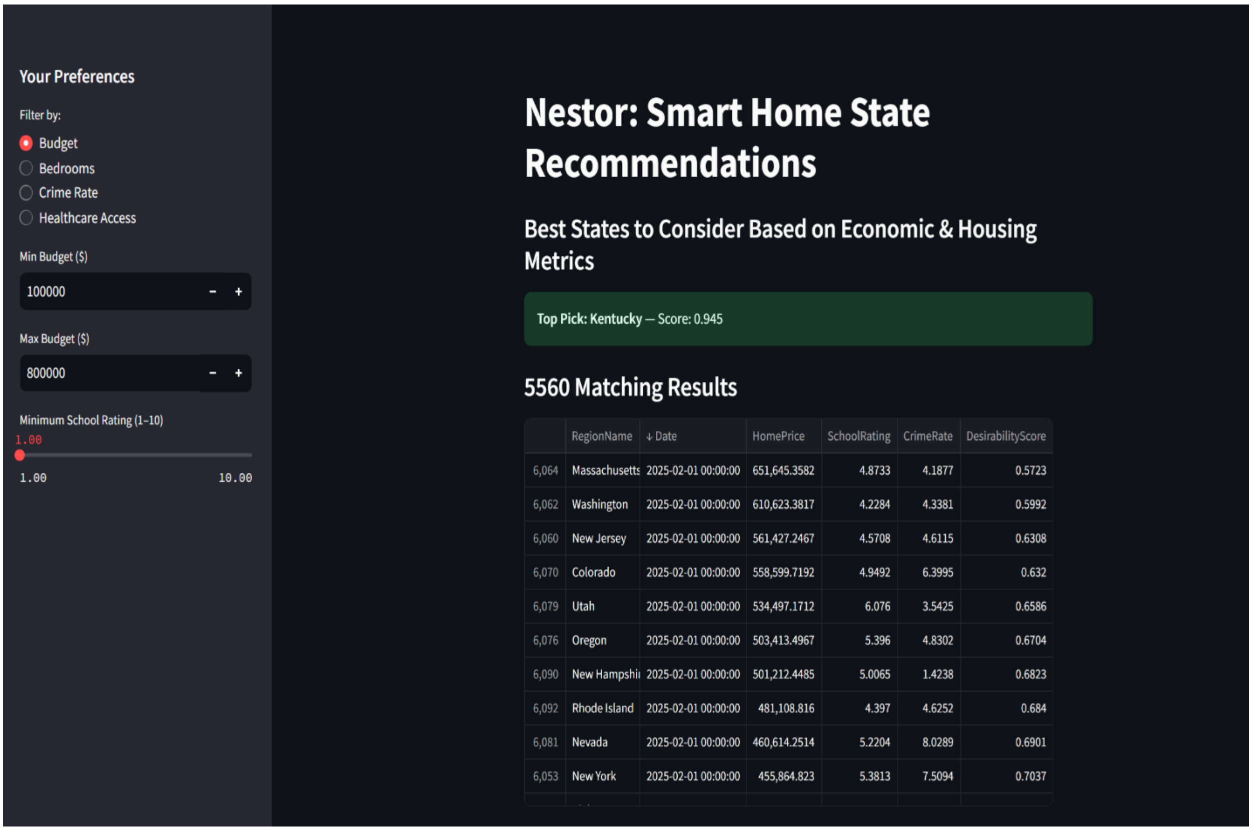

Fig 8.2: Nestor's interactive Streamlit dashboard showing user-defined budget and school rating filters on the sidebar, with real-time ranked results displayed in the main panel. The top-ranked state, Kentucky, is highlighted based on the Desirability Score.

8.3 Interactivity and Responsiveness

One of the core strengths of the Nestor dashboard is its ability to instantly reflect user input changes. Filters update the dataset in real time, and the resulting visualizations and tables are immediately refreshed. This responsiveness enhances user engagement and facilitates rapid scenario testing.

8.4 Visual Design and Accessibility

The dashboard features:

- Clear visual summaries using scorecards and bar charts.
- Interactive elements that respond to user input.
- Downloadable results for ease of reporting or consultation.

The overall design prioritizes usability, ensuring that insights are accessible and actionable without requiring advanced data literacy.

8.5 Price Trends Visualization

To provide users with deeper market insights, Nestor includes a Price Trends Visualization that displays historical home price trends for the top-ranked regions based on user-selected criteria.

This feature uses a line chart to compare price changes over time across the top five states identified by the Desirability Score. It helps users evaluate whether a region's affordability is stable, rising, or volatile, offering additional context to support long-term investment decisions.

By integrating this time-series analysis, Nestor not only ranks states based on current data but also allows users to visualize market trends, enhancing their ability to make forward-looking housing decisions.

Fig 8.5.1: Streamlit Dashboard Showing Bar Chart of Top Regions by Desirability Score and Price Trends Over Time for those Top Regions

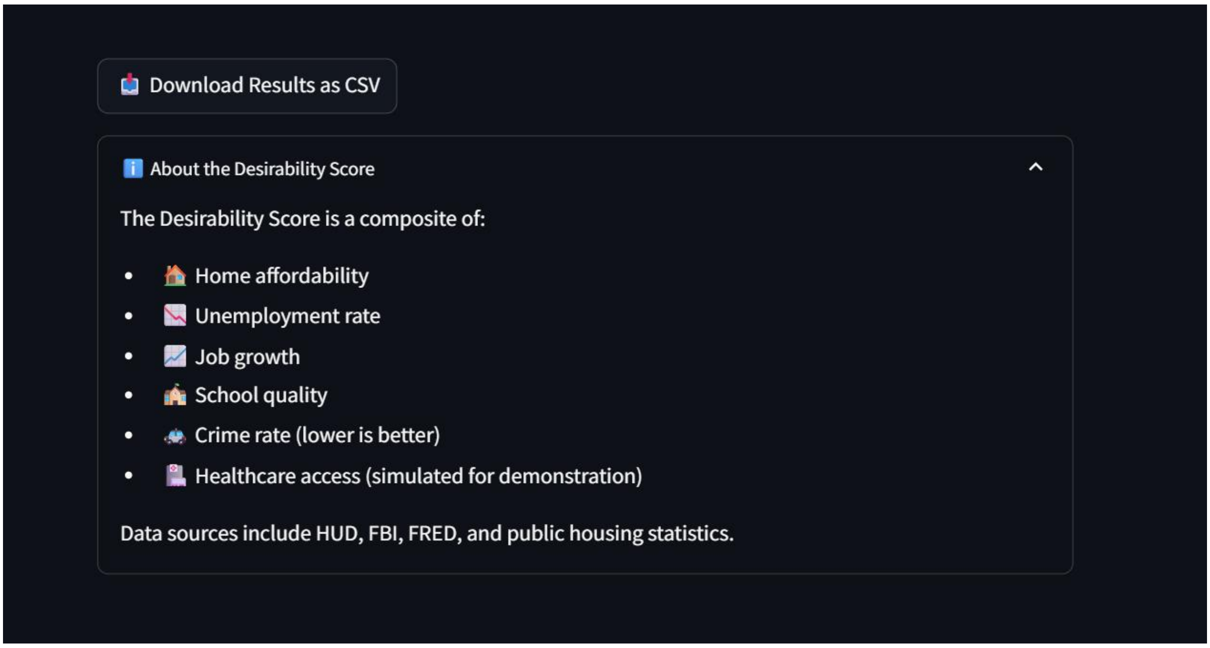

Fig 8.5.2: Streamlit Dashboard Showing CSV download option and desirability score composition

Please note that the healthcare access filter shown here uses simulated placeholder data and is intended for demonstration purposes only.

9. Results and Insights

The Nestor recommendation engine successfully delivered personalized, data-driven housing rankings across U.S. states. Through user interaction and dynamic scoring, several key trends and observations emerged.

9.1 Regional Rankings and Patterns

Nestor identified clear regional patterns:

- Top-Ranked States: West Virginia, Kentucky, and Oklahoma led the rankings by balancing affordability, safety, and education.
- Trade-off States: California, New York, and Massachusetts scored lower due to high housing costs and elevated crime, despite strong job markets and schools.
- Underserved Regions: Some highly affordable states ranked lower overall due to weak education or higher crime, showing that price alone is not enough.

9.2 Personalization Impact

Nestor's interactive filters allowed users to adjust their priorities, revealing how rankings change based on:

- Higher school rating requirements, favouring northern states.
- Lower crime tolerance, highlighting safer southern regions.

- Larger budgets, surfacing higher cost but high-quality states like Colorado.

This adaptability reinforces Nestor's value as a flexible, user-driven tool.

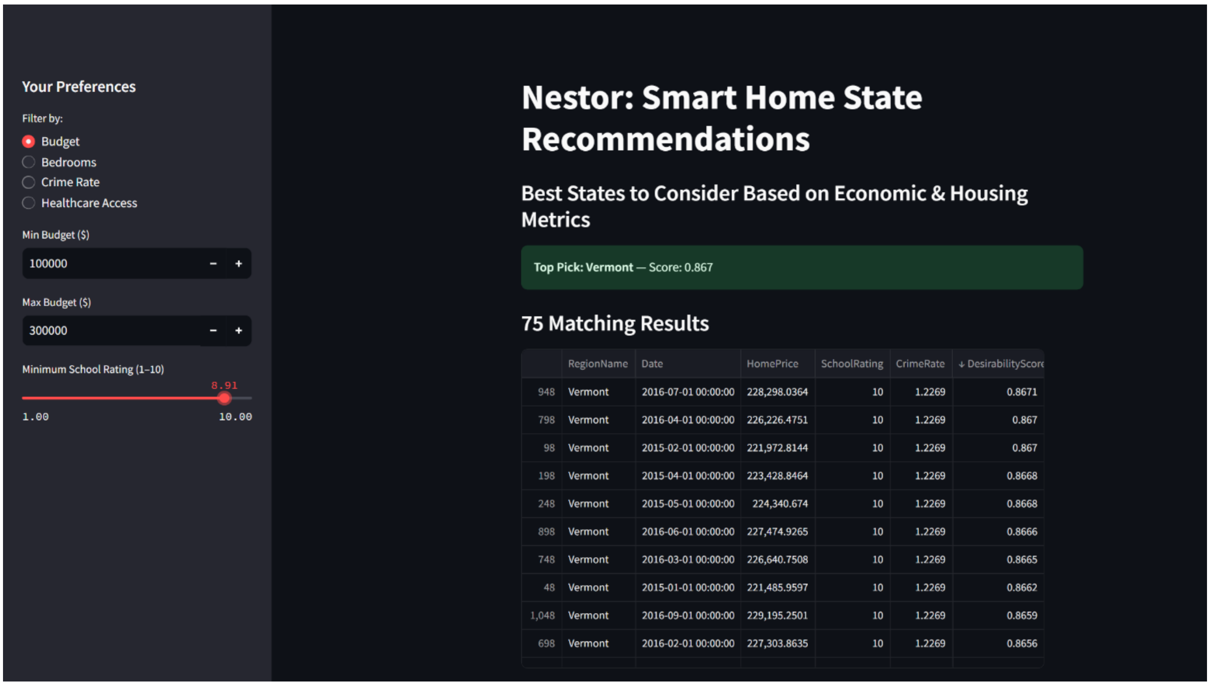

Fig 9.2 : compared to Fig 8.2, the max budget has been reduced and the school rating preference got higher and the top pick gets changed which shows the personalization impact in preferences

9.3 Score Distribution and Interpretation

Most states clustered between 0.80 and 0.90 on the Desirability Score scale.
A few exceeded 0.93, standing out for their balance across affordability, safety, and education. Subtle score differences between states helped users compare trade-offs that aren't obvious in traditional platforms.

9.4 Key Insights and Observations

- Balanced Decision-Making: Nestor emphasizes that no single factor—like price or school quality—should be viewed in isolation.
- Objective Rankings: The scoring model treats all regions equally, reducing bias toward popular markets.
- User-Friendly Insights: Clear charts, scorecards, and rankings make complex data easy to understand.

9.5 Limitations and Data Considerations While the prototype delivers actionable insights, some limitations were acknowledged:

- The healthcare access metric was simulated due to lack of consistent public data.
- The app currently operates at the state level; metro or ZIP-level data would offer more granular recommendations.

- The scoring system uses fixed weights; future enhancements could include user-weighted inputs or machine-learned ranking models.

9.6 Competitive Comparison with Existing Platforms

While platforms like Zillow, Redfin, and Realtor.com list properties, they lack tools to balance lifestyle and economic factors. Nestor stands out by offering real-time, multi-factor rankings tailored to user preferences, helping buyers make informed decisions.

Feature / Capability	Nestor	Zillow	Redfin	Realtor.com
Core Focus	Regional ranking based on livability and buyer preferences	Property listings and Zestimate valuations	Property listings with agent services	Property listings with market insights
Recommendation Approach	Multi-factor scoring (affordability, safety, education, healthcare)	Price and location filters	Price and location filters with agent suggestions	Price and location filters with local market data
User-Defined Trade-Off Analysis	Yes, balances multiple factors	No, users compare manually	No, users compare manually	No, users compare manually
Real-Time Personalized Ranking	Yes, updates instantly based on preferences	No	No	No
Data Transparency	Clear scoring with explained weights	Proprietary Zestimate	Proprietary agent rankings	Popularity-based rankings
Visual Insights (Charts & Rankings)	Bar charts, scorecards, snapshots	Basic overviews	Limited visuals	Market trends only
Interactive Decision Support	Yes, live dashboard with export option	No	No	No
Regional (State/City) Scoring	Yes, with future metro-level expansion	No, property-level only	No, property-level only	No, property-level only
Export / Download Option	Yes, CSV export	No	No	No
Expansion Potential	API, city-level data, user-weighted scoring	Listing-focused with ads	Agent-focused platform	Listing-focused with limited insights

10. Conclusion and Recommendations

Nestor demonstrates how public data and user-driven scoring can transform the homebuying process. By moving beyond traditional filters like price and location, Nestor provides a multi-criteria ranking system that helps buyers make more informed, balanced decisions.

10.1 Key Achievements

- Developed a transparent scoring framework based on affordability, education, and safety.
- Integrated publicly available data from multiple sources.
- Built a real-time interactive dashboard with dynamic filtering and ranked outputs.
- Highlighted regional trade-offs that traditional platforms overlook.

10.2 Why Nestor Adds Value

Unlike typical listing sites, Nestor allows users to:

- Compare regions holistically, not just by price.
- Adjust preferences and see rankings update instantly.
- Download ranked results for further review.

This makes Nestor a valuable decision-support tool for homebuyers seeking more than just listings.

10.3 Future Enhancements

To expand Nestor's impact, the following improvements are recommended:

- City or ZIP-Level Expansion for more detailed recommendations.
- Live API Integration with real-time housing and economic data.
- User-Weighted Scoring to reflect individual priorities.
- Machine Learning Models for smarter, data-driven personalization.
- Cloud Deployment to make Nestor accessible to a wider audience.

10.4 Final Reflection

By combining data transparency, user interactivity, and actionable insights, Nestor offers a scalable solution to improve homebuying decisions. With further development, it has the potential to set a new standard in real estate technology.

11. References

The following data sources, tools, and libraries were used in the development of the Nestor recommendation engine:

- **Zillow Research**
  https://www.zillow.com/research/data/
- **FRED (Federal Reserve Economic Data)**
  https://fred.stlouisfed.org/
- **HUD School Proficiency Index**
  https://hudgis-hud.opendata.arcgis.com/
- **FBI Crime Data Explorer / Kaggle**
  https://cde.ucr.cjis.gov/
  https://www.kaggle.com/datasets/mathchi/violent-crime-rates-by-us-state
- **Simulated Data for Healthcare Access**
  Synthetic data generated to represent regional access to healthcare (placeholder for future integration).
- **U.S. Census Bureau** (optional for future extensions)
  https://data.census.gov/