Akshita Khajuria

Data & Business Analyst

EMAIL ME

About Me

Data & Research enthusiast with a **Masters in Quantitative Economics with specialization in Business & Data Science at UCLA.**
With 1+ year of experience in Data Analysis, Business Strategy, and Consulting- I bring passion for solving business puzzles across industries. Whether it's digging deep into qualitative insights or decoding patterns in complex datasets, I enjoy turning ambiguity into clarity and strategy. I'm here to connect the dots between business challenges and market success – one insight at a time.

LinkedIn

Portfolio

GitHub

Experience

- Worked in **Consulting, Data Science and Consumer Marketing Strategy** to solve business challenges ranging from Economic Development, Supply Chain Efficiency, Consumer Analytics and Marketing Strategy, and Data Mining and Management.

JUNE 2022–JAN 2023

Market Research Analyst

Independent Consultant to ITC & World Bank

JAN 2024 – MARCH 2024

Data Science Intern- Demand Planning & Operations

Harman International (Brand JBL)

OCT 2024 – DEC 2024

Project Manager- Data & Marketing Science

Harman International (Brand JBL)

JAN 2025 – MARCH 2025

Data Science Intern- Data Management

GordonMD® Global Investments LP

MAY 2025–PRESENT

Data Scientist- Business Operations

Los Angeles World Airports

GitHub

Project 1-
Shopper Signals: What Search Data Tells Us About Trust, Value, and Doubt
Decoding Shopper Psychology Through Search

Challenge- While online search data provides raw insights into what consumers are searching for, it often lacks the structured depth to inform strategic marketing and product decisions. Retailers struggle to decode why people search and what psychological intent drives that behavior?

Goal- Build a keyword-based behavioral segmentation model using Google Trends data to group shoppers based on psychological drivers (e.g., trust, price sensitivity, convenience). Then analyze interest trends over time, volatility, and seasonality

- Data Source: Google Trends API
- Tools: Python (Pandas, Seaborn, Matplotlib), Power BI
- Time period- Jan 2025- July 2025

Part 1- Segmentation Approach

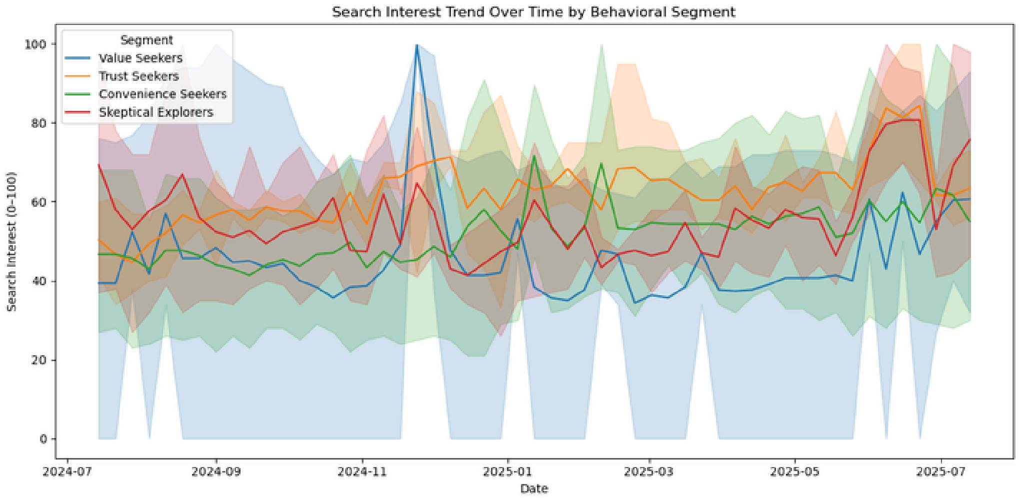
Grouped 12 shopping-related keywords into 4 behavioral clusters:

Value Seekers
- "promo codes fashion", "student discount", "free shipping"
- *Highly price-sensitive, deal-driven shoppers*

Trust Seekers
- "ethical brands", "return policy", "verified reviews"
- *Want assurance, reliability, and ethical transparency*

Convenience Seekers
- "same day delivery", "pickup near me", "try before you buy"
- *Prioritize speed and ease of shopping*

Skeptical Explorers
- "Temu vs Shein", "best product review", "Amazon fake reviews"
- *Research-heavy, cautious, questioning credibility*



Analysis-

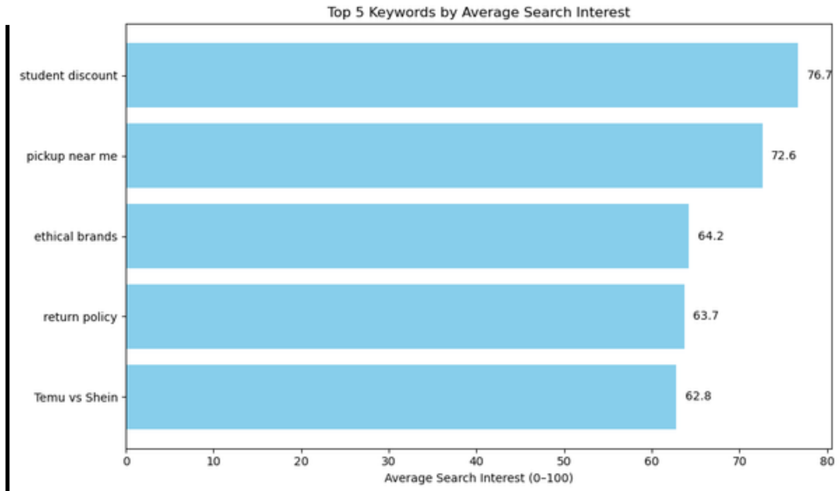
Value Seekers are highly volatile:
- Sharp spikes during sale seasons (November, May)
- Suggests this group is event-driven and opportunistic
- Great for short-term campaign planning

Trust Seekers lead in average interest

- Consistently high engagement over the past year
- Reflects a growing desire for brand transparency, ethical shopping, and secure return policies

Skeptical Explorers are rising
- Increasing interest in terms like “Temu vs Shein” and “Amazon fake reviews” in 2025
- Indicates growing doubt, digital noise, and buyer hesitation
- Suggests an opportunity to build confidence through verified content

Convenience Seekers remain stable
- Slight upward trend in interest for services like “pickup near me”
- Less volatile, but still important for optimizing user experience



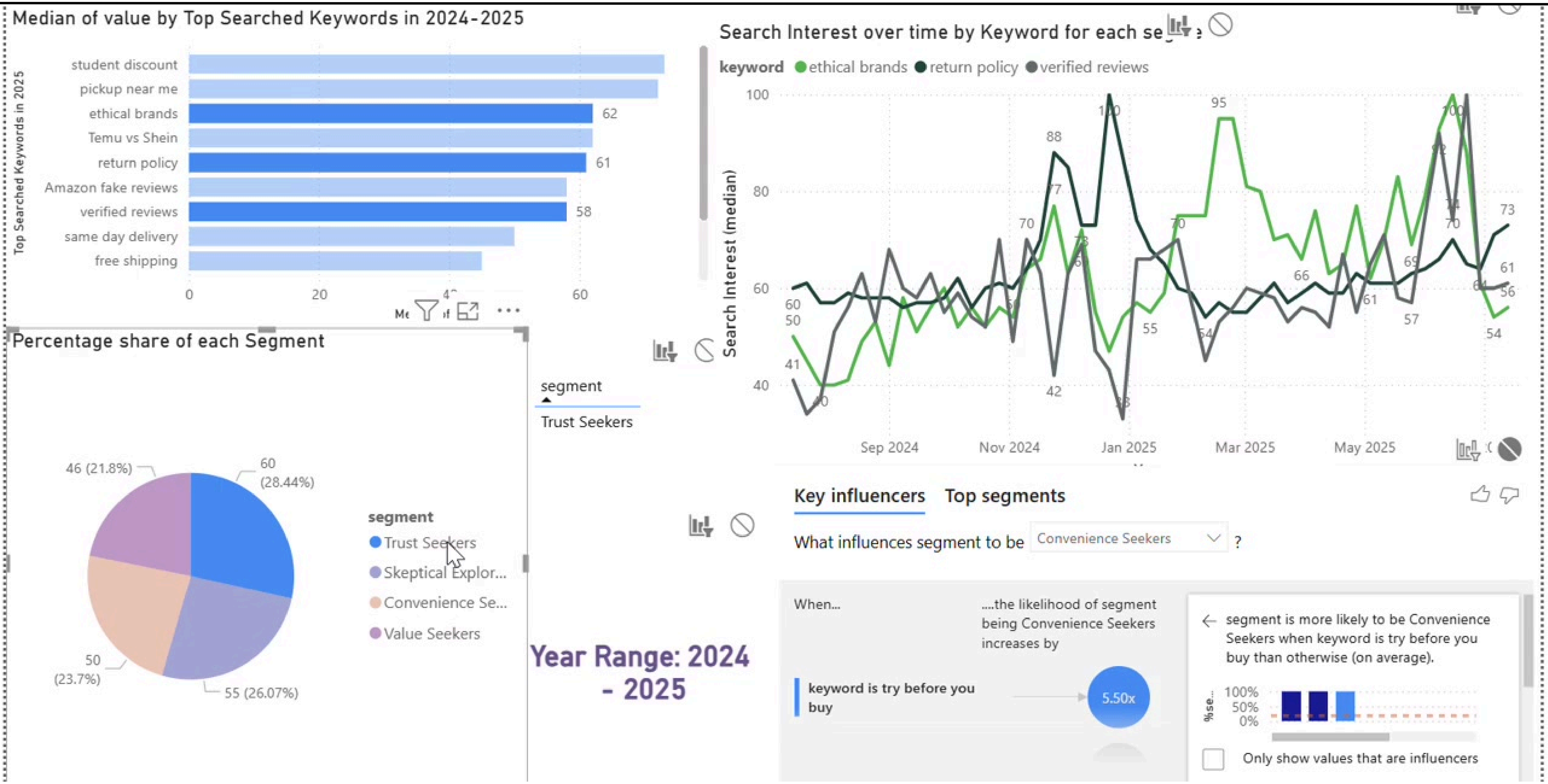
Key Insights-

- Trust and Ethics Are Driving Behavior: Keywords like “ethical brands” and “return policy” are among the top 5 most searched, showing that shoppers are actively looking for transparency, safety, and accountability in what they buy.

- Convenience and Discounts Still Matter: “Pickup near me” and “student discount” also rank high, highlighting that while trust is rising, convenience and savings remain core to consumer decision-making.

- Curiosity & Skepticism Are Climbing: “Temu vs Shein” making it to the top 5 signals a growing trend of consumers doing deeper comparisons — signaling more investigative, value-conscious buying behavior.

Project 1-
Shopper Signals: What Search Data Tells Us About Trust, Value, and Doubt
Decoding Shopper Psychology Through Search - POWERBI Dashboard



Project 2-
How can we improve operational efficiency in large-scale airport operations during peak congestion periods?

[GitHub](#)

Challenge- When flights get delayed, it causes a lot of chaos and confusion at the airport leading to several operational inefficiencies such as crowded security lines, limited security staff supply and high demand.

Goal- Using Historical flight and passenger-level data to forecast airport congestion, predict delays, and identify high-risk operational flight clusters.

Tools & Datasets Used

- Datasets:Air Traffic Passengers
- Flight Delay & Cancellation (2019–2023)

<u>Technologies: Python, Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, Seaborn, ARIMA, K-means Clustering</u>

Preparation of Dataset-

- Large Dataset Analysis- Cleaned and standardized over 3 million rows
- Engineered new features like congestion score, hour-of-day

Outcomes

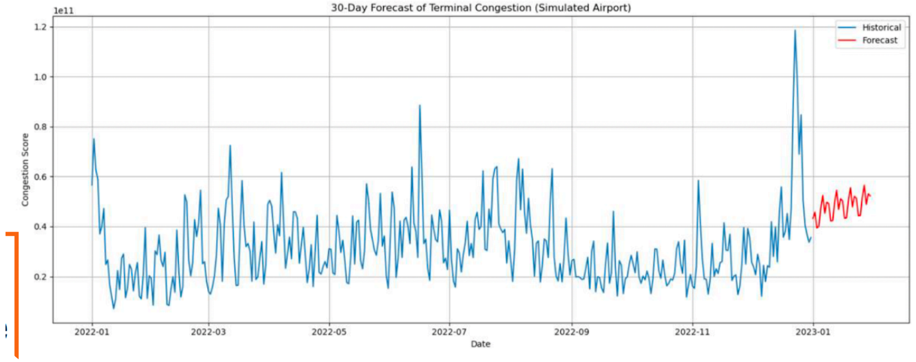
- ARIMA on the Congestion Score- Predicted future congestion rate with an 78% accuracy
- Seasons with Highest congestion (2019-2023)- June, July, August & December
- 1 PM – 3 PM is a critical period for airports like LAX — both volume pressure and delay risk peak.
- This time window demands:
  - Proactive staffing
  - Real-time flight monitoring
  - Contingency plans for runway/taxi delays & gate assignments

Part 1- ARIMA Forecasting

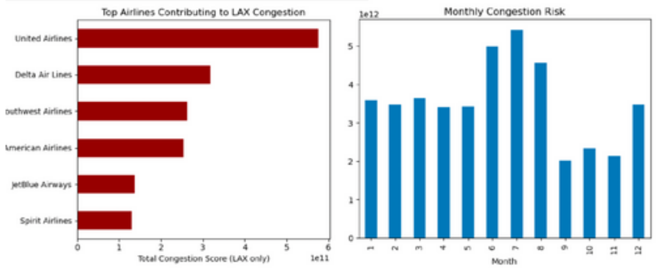
Question -Can we predict when airports are likely to face the most congestion?

Factors Considered:

- Historical adjusted passenger counts (enplaned + deplaned) per airline and terminal
- Date and time of flight (Monthly)
- Arrival delays
- Custom-engineered Congestion Score: Adjusted Passengers * Arrival Dela



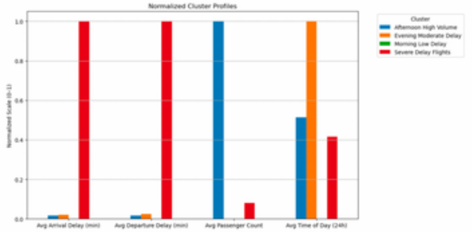
Predicted-



Part II – Regression & Unsupervised Machine Learning Model- Clustering

- **Challenge Identified**- For every 1-minute increase in average arrival delay, the odds of next-hour congestion increase by a factor of 2.57.
- **Solution**- Group flights based on operational similarity — so LAX can create tailored playbooks for handling different flight types.

	Cluster_Label	Avg Arrival Delay (min)	Avg Departure Delay (min)	Avg Passenger Count	Avg Time of Day (24h)
0	Afternoon High Volume	9.97	6.51	2573750.65	14.42
1	Evening Moderate Delay	11.52	8.67	939792.15	18.69
2	Morning Low Delay	5.88	1.97	942075.40	9.92
3	Severe Delay Flights	268.02	269.58	1071192.35	13.56



- Flights in the "**Afternoon High Volume**" (Blue) may not be delayed- But high Average Passengers (2:25 PM)
- **Morning Flights (Green)** Offer Operational Stability- ideal for handling maintenance, staffing resets, and preparing for peak congestion later (9:55 AM)
- **Severely Delayed Flights (Red)** Are Operational Red Flags- major disruption risks. – Need real-time monitoring and escalation protocols. Delay alerts, rerouting strategies, and contingency planning should be prioritized around this cluster. (1:34 PM)

Project 3-
How can we understand what drives Real Estate Prices to stay ahead in a competitive market ?

GitHub

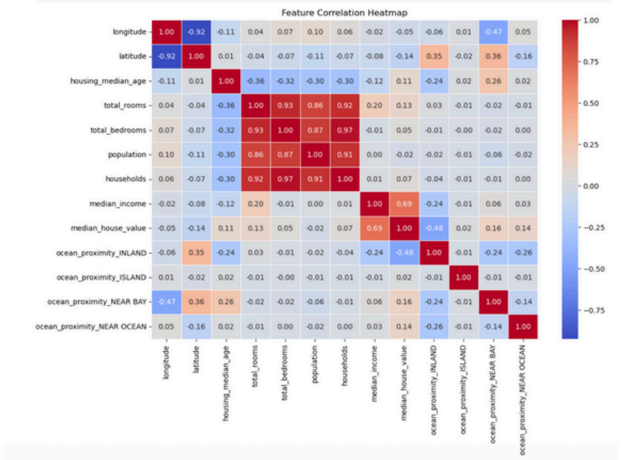
Challenge- Real Estate is one of the biggest revenue generating industry in the US economy, thus, how the real estate prices fluctuate over time and what are the major factors driving these prices make a huge impact on revenue and how the market performs.

Goal- Built predictive models (Linear Regression, Random Forest, XGBoost) to estimate house prices based on key features. Performed feature importance analysis, identifying income and proximity to the ocean as the most significant factors. Evaluated model accuracy using RMSE, R^2 scores, and visualized predictions to validate performance.

<u>Technologies: Python, Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, Seaborn, Machine learning models (Linear Regression, Random Forest, and XGBoost), we identify key factors influencing prices and develop a predictive model.</u>

Preparation of Dataset-
- Large Dataset Analysis- Cleaned and standardized over 20k rows and 10 columns

Conducted Exploratory Data Analysis (EDA)
- Visualize the distribution of house prices.
- Analyze correlations between features.
- Plot geographical price variations.



Key Business Takeaways for a Real Estate Companies:
Income is the most influential factor in home prices, making it crucial for pricing models.
Coastal and near-bay properties command higher prices, highlighting the importance of location-based marketing.
Further inland homes tend to be cheaper, which could guide investment strategies for affordability-focused developments.
Household size and total rooms are highly correlated, suggesting that pricing strategies should consider family-oriented housing demand.

Part 1- Machine Learning Model (Baseline)

Goal: Establish a simple Linear Regression model to predict house prices and evaluate performance.

Outcome- Linear Regression RMSE: 70060.52
Linear Regression R^2 Score: 0.63

- This means the model's predictions deviate by ~70K USD on average from the actual house prices.
- The model explains 63% of the variance in house prices, which is decent but leaves room for improvement.

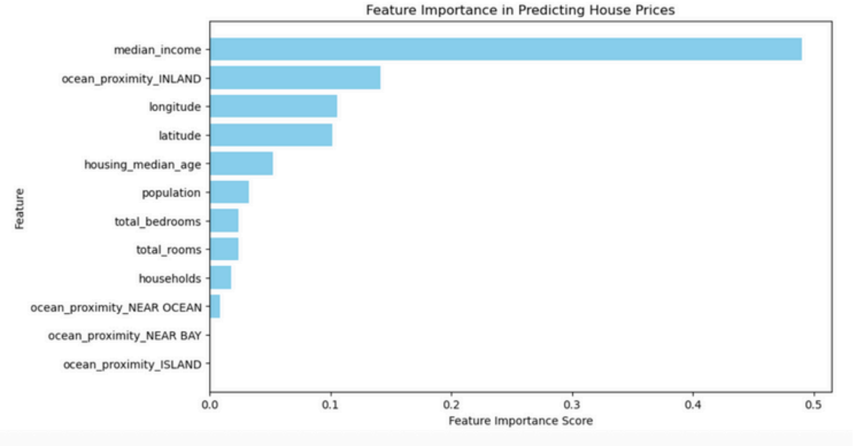
Part II- Improve Model Performance with Random Forest

Goal- Use a more advanced model (Random Forest Regressor) for better accuracy

Outcome- RMSE (Root Mean Squared Error) = $49,008.79
- The average prediction error is ~$49K, which is a significant improvement over Linear Regression (70KUSD).

R^2 Score = 0.82
- The model now explains 82% of the variance, meaning it's capturing more of the housing price patterns which is decent but leaves room for improvement.

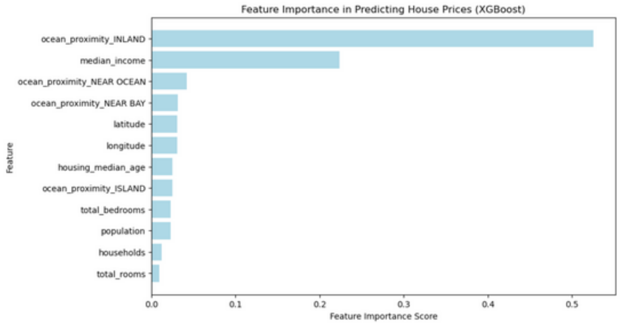
Hyperparameter Tuning for Random Forest
Goal - Improve accuracy by optimizing model parameters.
Found- Income and proximity to the ocean are driving the Real estate prices the most.



Part III – XGBoost for Further Improvement

Interpretation of Results:
- RMSE (Root Mean Squared Error) = $47,215.29- The average prediction error is ~47K USD, lower than Random Forest (48K USD), meaning better accuracy
- R^2 Score = 0.83- explains 83% of the variance, slightly improving over Random Forest (82%).

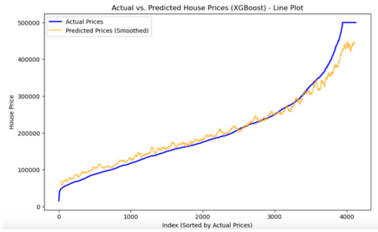
Key Takeaways:
- XGBoost performed better than Random Forest, indicating that boosting techniques improved predictions.
- Lower RMSE means better generalization, making XGBoost a stronger choice for real estate price modeling.
- Both models perform well, but XGBoost is slightly more accurate, making it ideal for final deployment.



- Ocean Proximity (INLAND) is now the most important factor, replacing median income, which was previously the strongest predictor. Proximity to water (NEAR OCEAN, NEAR BAY) significantly impacts prices, highlighting location's importance. Median income remains a key factor but has less influence compared to location. With improved model accuracy, the feature importance ranking shifted, revealing deeper real estate pricing trends. Traditional factors like number of rooms and population contribute less than geographic and economic factors.

Summary - With improved model accuracy, ocean proximity (INLAND vs. NEAR OCEAN) replaced median income as the strongest predictor of house prices, emphasizing location over affordability. Traditional factors like room count and population have minimal impact compared to geographic and economic influences.

Part IV - Predicting future Real Estate prices using XG Boost

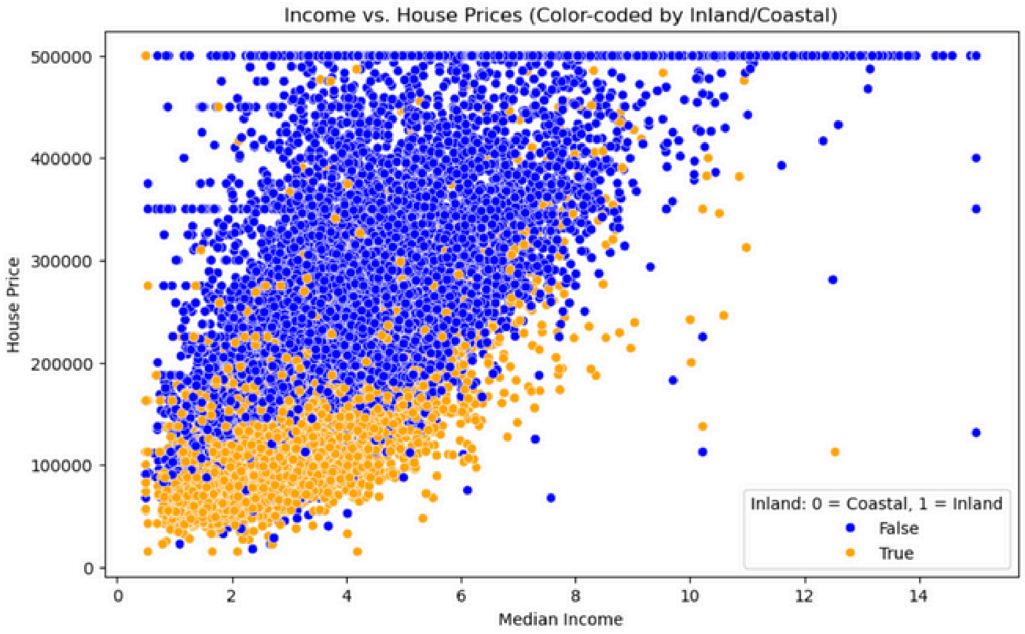


- The blue line (actual prices) and the orange line (predicted prices) show that the model captures market trends well, making it useful for pricing strategies.
- Predictions align closely with actual values in the mid-range housing market, meaning the model can reliably estimate prices for average properties.
- However, in the higher price segment, the model underestimates values, suggesting that luxury housing pricing may require additional variables (e.g., amenities, proximity to premium locations).
- . With improved model accuracy, the feature importance ranking shifted, revealing deeper re

GitHub

Project 3-
How can we understand what drives Real Estate Prices to stay ahead in a competitive market ?

Visualizing Price Trends Across Key Variables



To combine both major factors, we can visualize how income levels change in different regions and affect housing prices.

- This visualization highlights the relationship between median income and house prices, with inland (orange) and coastal (blue) properties distinguished:
- Coastal Properties (Blue) Are More Expensive
- Most high-income regions correspond to higher house prices, especially for coastal properties. Even at lower income levels, coastal properties tend to have higher prices than inland properties.
- Inland Properties (Orange) Are More Affordable
- Houses in inland areas remain relatively cheap even at higher income levels. This suggests inland regions may be more suited for affordable housing projects or investment opportunities for appreciation.

Outcomes

- Luxury Real Estate: Focus on coastal areas where higher-income buyers are willing to pay premium prices.
- Affordable Housing: Inland areas offer opportunities for budget-conscious buyers or rental investments.
- Investment Potential: Inland properties could provide growth potential as incomes rise.

Use case- **A real estate company can use this to adjust pricing strategies, target marketing efforts, and identify high-return investment areas based on income levels and proximity to the ocean.**

Major Takeaways
Income and location (inland vs. coastal) significantly impact house prices. Coastal properties command higher prices across all income levels, making them prime targets for luxury real estate. Inland properties remain more affordable, making them better for budget-conscious buyers or long-term investment.

XGBoost performed best, achieving RMSE ≈ $47,215 and $R^2 = 0.83$, making it the most reliable model.

How This Can Be Used Further
- Price Prediction Tool: A real estate company can integrate this model into a pricing tool for property valuation.
- Investment Insights: Use feature importance analysis to identify high-growth areas based on income shifts and location demand.
- Urban Planning & Development: Policymakers can leverage this to plan affordable housing projects in inland areas and infrastructure improvements in high-value regions.

GitHub

Project 4- Optimizing a Diversified Stock Portfolio Using Monte Carlo Simulation for Maximum Sharpe Ratio and Efficient Frontier Visualization

Challenge- In Market Uncertainties especially in 2025, the stock market has been extremely volatile and thus, there is a high risk of losing money while investing large sum of money in stock market.

Goal- This project demonstrates the process of constructing an optimized investment portfolio using Modern Portfolio Theory (MPT). The focus is on selecting a diversified set of 10 stocks, calculating the maximum Sharpe ratio, and determining the optimal asset allocation.

Technologies: Python, Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, Seaborn, Monte Carlo Simulation

Preparation of Dataset-

- Data Collection: Using yfinance to gather daily adjusted closing prices for the selected stocks.

Monte Carlo simulation to explore a large number of potential portfolio configurations (6000 in this case), evaluates their risk and return characteristics, and identifies the portfolio that maximizes the Sharpe ratio. This is used for portfolio optimization to find the optimal asset allocation.

Part 1- Correlation Analysis

This correlation plot provides insights into the relationships between the different stocks you have chosen.

Here's what it indicates:

1. Negative Correlation:

DXCM and SERV (-0.86): There is a strong negative correlation between Dexcom (DXCM) and ServiceMaster (SERV), meaning that when DXCM's price tends to go up, SERV's price tends to go down, and vice versa. SERV and LULU (-0.72): ServiceMaster (SERV) and Lululemon (LULU) also show a strong negative correlation, indicating that these stocks move in opposite directions.

2.Positive Correlation:

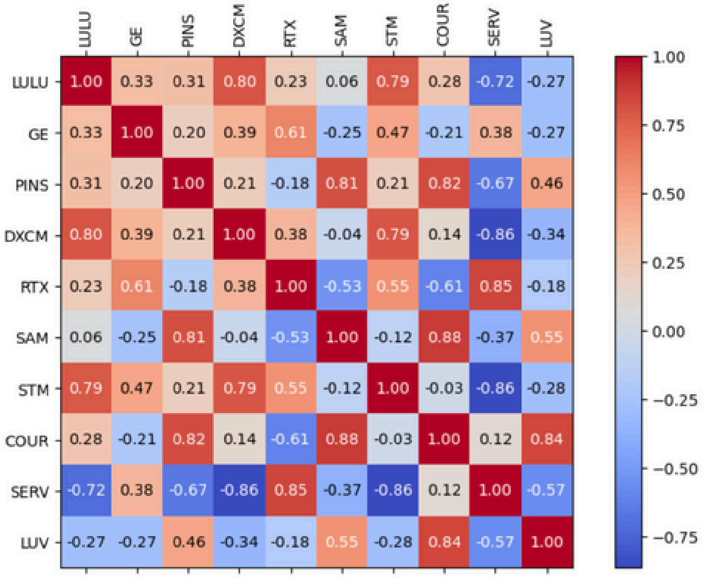
SERV and RTX (0.85): ServiceMaster (SERV) and Raytheon Technologies (RTX) have a high positive correlation, suggesting that their stock prices tend to move in the same direction. COUR and SERV (0.84): Coursera (COUR) and ServiceMaster (SERV) also exhibit a strong positive correlation, meaning they generally rise and fall together.

3.Moderate Correlation:

PINS and SAM (0.81): Pinterest (PINS) and Boston Beer Company (SAM) show a moderate positive correlation, indicating some co-movement between their prices. LULU and STM (0.79): Lululemon (LULU) and STMicroelectronics (STM) have a moderate positive correlation, suggesting that these stocks move somewhat together.

4.Low to No Correlation:

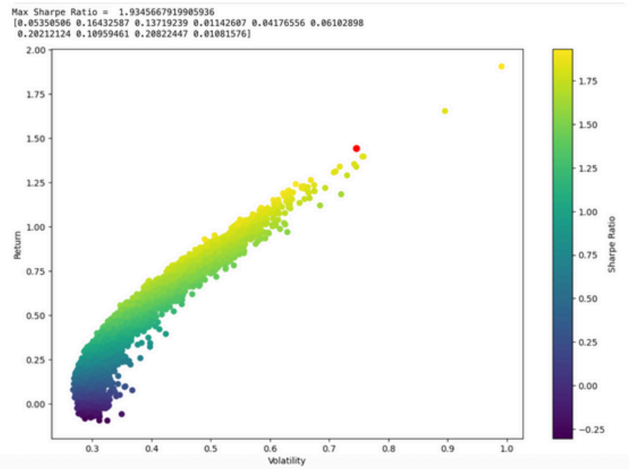
LULU and SAM (0.06): Lululemon (LULU) and Boston Beer Company (SAM) show very little correlation, indicating that their price movements are relatively independent of each other. GE and PINS (0.20): General Electric (GE) and Pinterest (PINS) exhibit a low correlation, meaning that their stock prices do not necessarily move together.



Implications for the Portfolio:
Diversification: The mix of both negatively and positively correlated stocks indicates a diversified portfolio. Having negatively correlated stocks can help reduce overall portfolio risk since they can offset each other's price movements. However, stocks with high positive correlations (like SERV and RTX) may increase the portfolio's risk, as they could move in the same direction during market events.
Risk Management: By analyzing these correlations, we can make informed decisions on how to allocate our investments to balance risk and return according to Modern Portfolio Theory. This diversified selection could be beneficial for balancing the portfolio's volatility and potential returns.

Part II – Portfolio Simulation

Implementing a Monte Carlo simulation to explore various asset weight combinations, aiming to maximize the portfolio's Sharpe ratio.



- Max Sharpe Ratio: The maximum Sharpe ratio achieved by this portfolio is approximately 1.93. The Sharpe ratio measures the risk-adjusted return, with higher values indicating better performance relative to the risk taken. A Sharpe ratio above 1 is generally considered good, while anything above 2 is considered excellent. A ratio of 1.93 indicates that the portfolio is well-balanced, providing strong returns for the level of risk involved.
- Efficient Frontier Analysis: Efficient Frontier Curve: The efficient frontier curve shows the relationship between expected return and portfolio volatility (risk). Each point represents a different combination of asset weights in the portfolio, illustrating the trade-offs between risk and return. Red Dot (Optimal Portfolio): The red dot marks the portfolio with the highest Sharpe ratio. This portfolio represents the best trade-off between return and risk, maximizing returns for each unit of risk taken.

Outcomes

- Optimal Risk-Return Balance: **The portfolio with the highest Sharpe ratio (1.93) suggests a well-optimized selection of assets that offers a favorable balance between risk and return.** The portfolio is efficiently diversified, minimizing unnecessary risk while enhancing potential returns.
- Volatility Consideration: **The optimal portfolio is positioned with moderate volatility, indicating a balance between conservative and risky investments.** This suggests that the portfolio is designed to achieve higher returns without exposing the investor to excessive risk.
- Portfolio Effectiveness: **The combination of stocks chosen for this portfolio has been effective in creating a diversified mix that reduces risk while maintaining strong returns, as reflected by the efficient frontier's shape and the high Sharpe ratio.**

Overall Assessment:

- The portfolio is well-diversified and optimized for risk-adjusted returns. It strikes a good balance between growth and stability, making it suitable for an investor seeking both security and potential for higher returns. **The efficient construction of this portfolio is evident in its ability to offer strong returns without taking on excessive risk.**