STA 141B Final Project
House Price Web Scraping and EDA

Ayaka Okada (919554812),
Eric Kye (919455487),
Tianci Zhu (918613848)

Instructor: Peter Kramlinger

March 2024

Introduction:

According to The Wall Street Journal, the number of people owning stocks has been higher than it ever has, recording 58% of U.S households holding stocks, possibly due to financial instability and hope to retire early for some of the investors. However, investing in stocks is not the only way of building wealth. Furthermore, stocks are third place for long-term investment, ranking after real estate and gold (Gallop, 2023). While 18% believe that stock is the best long-term investment, 34% believe the real estate market is the way to go for American households (Gallop, 2023).

Real Estate has been a popular long-term investment due to the nature of stability, security, and the high return it gives investors. Even though real estate is not easily affected by short-term volatility, unlike the stock market, it can also be affected by market fluctuation. Since the pandemic in 2021, there has been a significant increase in home prices compared to before the pandemic. (Whitehouse, 2021) Just as stock investors would like to know how the stock trend of the business may be in the future, knowing the trend of real estate is as important for real estate investors. However, with more than 140 million houses in the U.S., it may be crucial to understand the trends and factors affecting each real estate price (Census, 2021).

The price of real estate can vary depending on where it is located. For example, the median house price in Iowa is roughly $200,000, whereas the median home price in New York was listed at around $450,000 in March of 2024, more than twice the price observed (Zillow, 2024). Real Estate prices reflect the market price of the city, and analyzing the components of the real state in major cities will provide a good comparison of why one place may be priced higher than the other location, as well as why salaries paid on similar roles vary depending on the location of work.

Not only do real estate prices affect investors, but also individuals looking for a new location to move and start a life in a new city. Especially for college students planning to pursue a career after graduation, it is important to know the role's market price and consider rent for the new house or apartment they are moving into.

This project uses web scraping techniques to gather real estate data from listing websites. This will be followed by exploratory data analysis to uncover the trends in real estate prices. This analysis aims to identify observable patterns and potentially formulate predictive models to

enhance decision-making in the real estate market based on the available data. The focus will be put on to derive a conclusion to the following research question:

"What combination of features, including those intrinsic to the property, temporal trends, and city-wide statistics, most accurately predict fluctuations in real estate prices?"

Five major cities will be focused on throughout this project due to the vast amount of information available regarding real estate. As there is an increasing demand in the tech industry and as this project is for a statistics course, five major cities with tech talent market and the workforce available in the tech industries (Conte, 2023). The following are the cities: San Francisco, New York City, Seattle, Los Angeles, and Dallas. Before the real estate market price is considered, the median income in these five cities will be used to consider if there is a difference in average income.

Methodology:

1. **Web scraping real estate listings from Realtor:**

The real estate data for the five major US cities listed above will be extracted from the collection of real estate listings from Realtor.com. Utilizing ScrapflyClient to manage web requests, the following information will be extracted regarding each house information regarding the five cities: price, number of bedrooms and bathrooms, square footage, and lot size. The web scraping is carefully limited to a specific number of pages (30) per city to avoid IP blocking. As there are endless listings available for each state, 240 house listings per city will be looked into. The extracted house information for each state is saved into a CSV file, 'real_estate_data.csv,' for further analysis.

Following the collection, the data in the CSV file is made into a single data frame using pandas. This data frame consists of information on 240 houses from each city regarding price, number of bedrooms and bathrooms, square footage, and lot size. Preliminary data cleaning tasks were conducted by removing non-numeric characters from the *Price* column and converting data types to numerical.

2. **Web scraping city data from City-Data:**

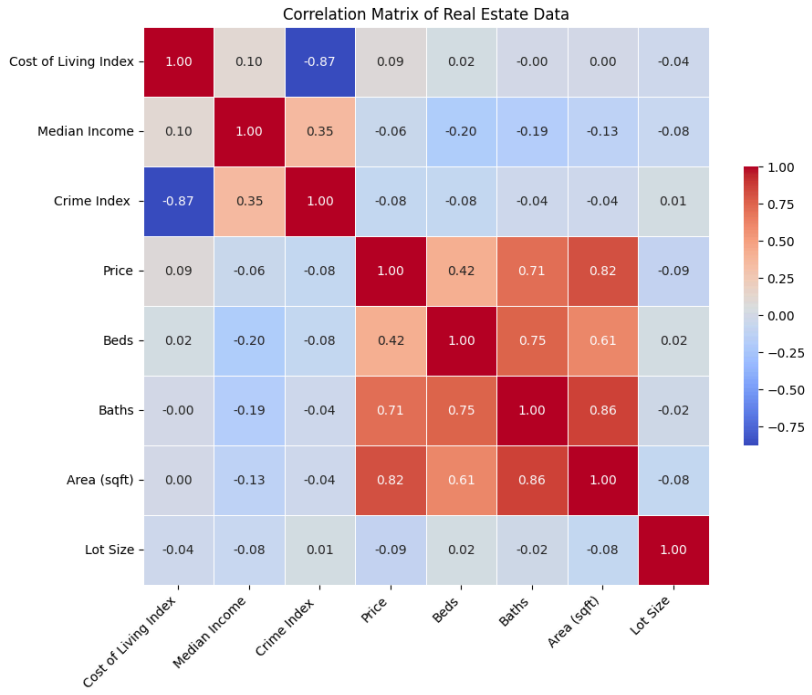
In addition to real estate listings, data from each city, such as the cost of living index, median income, and crime rates, were extracted by scraping through City-Data.com. Looking at the information regarding the city and the cost of living index may explain the trend of the housing price to determine why the real estate prices are different between the five locations. The information that was collected was merged into the CSV file to organize the information in one place.

Following is a summary of the cost of living index, median income, and crime rates of the five cities:

	City	State	Cost of Living Index	Median Income	Crime Index
0	Los-Angeles	California	145.1	70372	327.4
1	San-Francisco	California	141.1	121826	387.4
2	New-York	New York	160.2	67997	229.7
3	Seattle	Washington	118.5	110781	440.8
4	Dallas	Texas	96.1	57995	439.5

3. **Exploratory Data Analysis:**

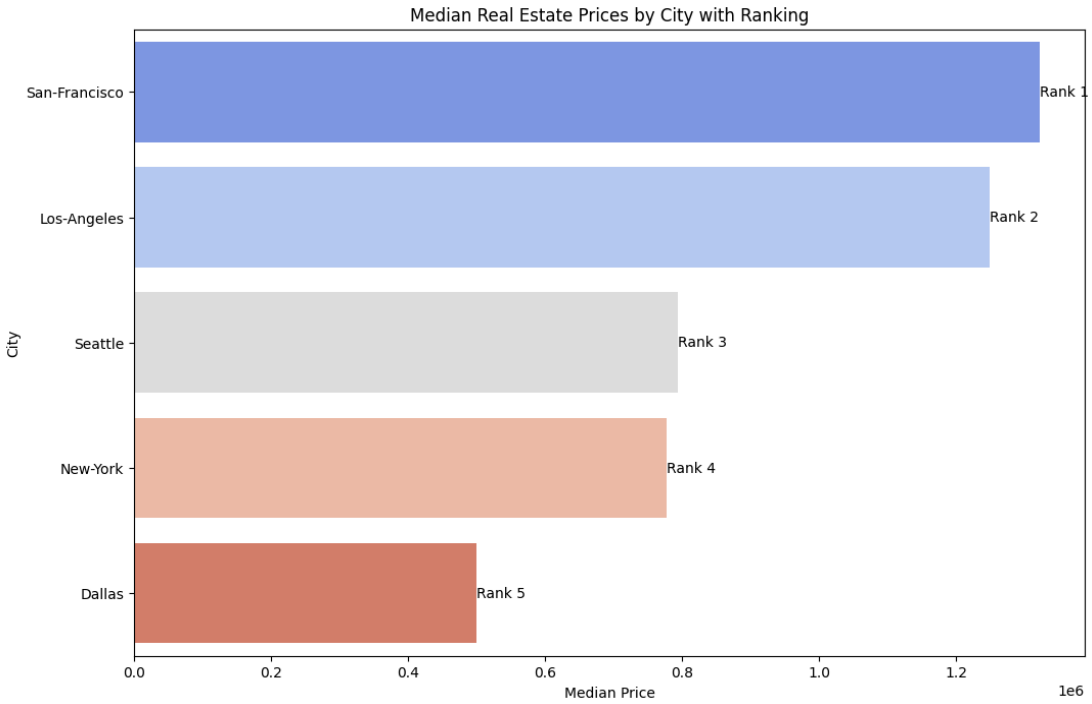
Once the dataset has been organized into one CSV file, exploratory data analysis can be conducted to uncover underlying possible patterns and relationships. First, a correlation matrix is generated to identify the variables that are most strongly associated with property prices. We used a heatmap to visualize the table. Heatmap allows efficient visualization of how each variable may be related to each other.



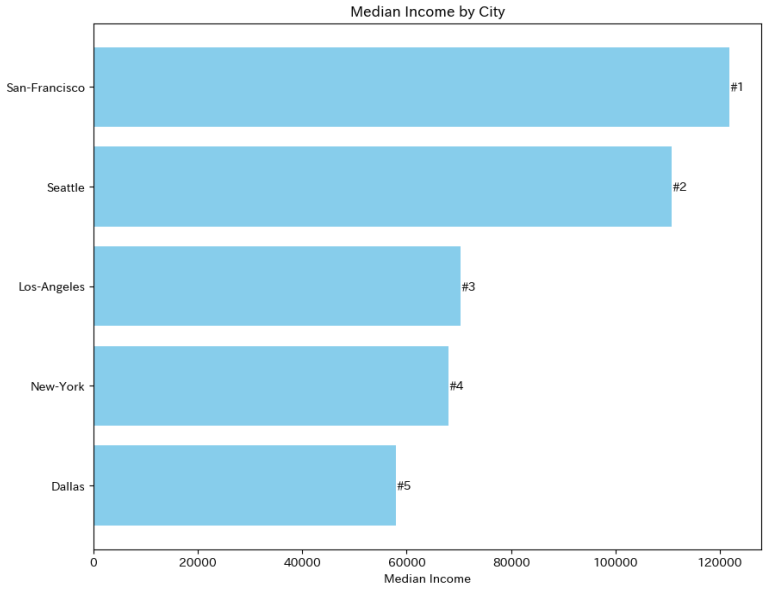
From the heatmap, it can be immediately interpreted that there are two blocks of square on the top left and middle to bottom left of the correlation matrix. When closely looking at the components of the correlation matrix, it can be found that these two are divided based on the combined dataset, one from real estate web scraping and one from city-data. This shows no strong correlations between the combined data from two sources.

The focus is on the price of real estate. In the fourth row, the highest correlation between the price, the area of the house, and the number of bathrooms and bedrooms is seen. However, with the other variables, especially the data extracted from the city, it seems to have a very small correlation, showing that there may be no association between the price and the individual's income in the city.

The bar graph compares median property prices across the five cities. This comparison purely highlights which cities have the highest real estate median price and the market dynamic of the city.



The median salary barplot of these cities is computed as well to see if a similar trend is observed.



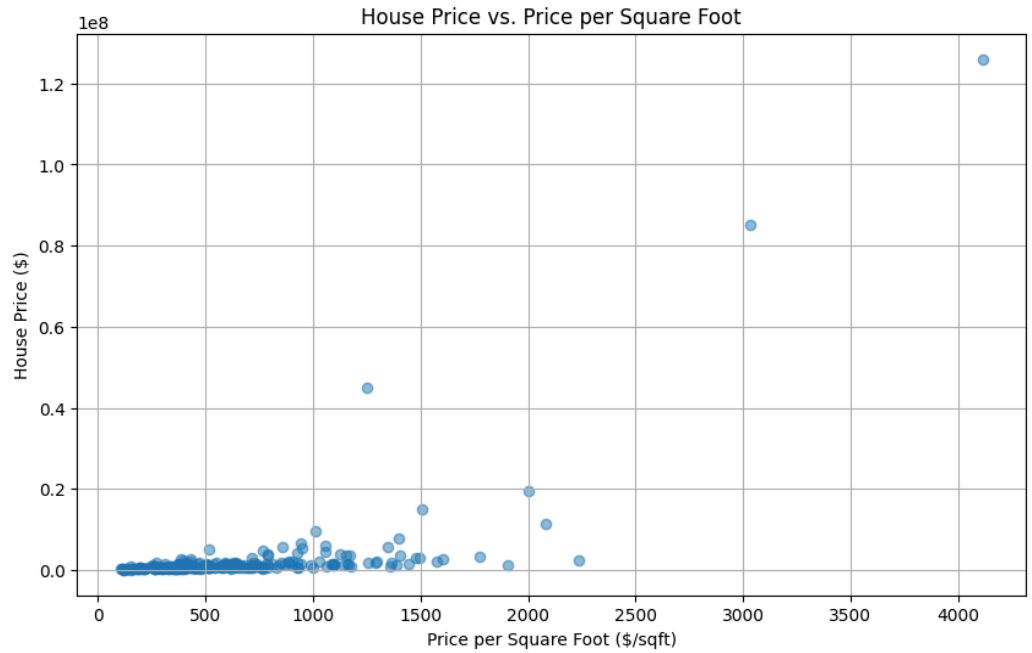
There is a similar trend between the price of the house and the median income other than the order of Seattle and Los Angeles, which is flipped for the median income compared to the house pricing.

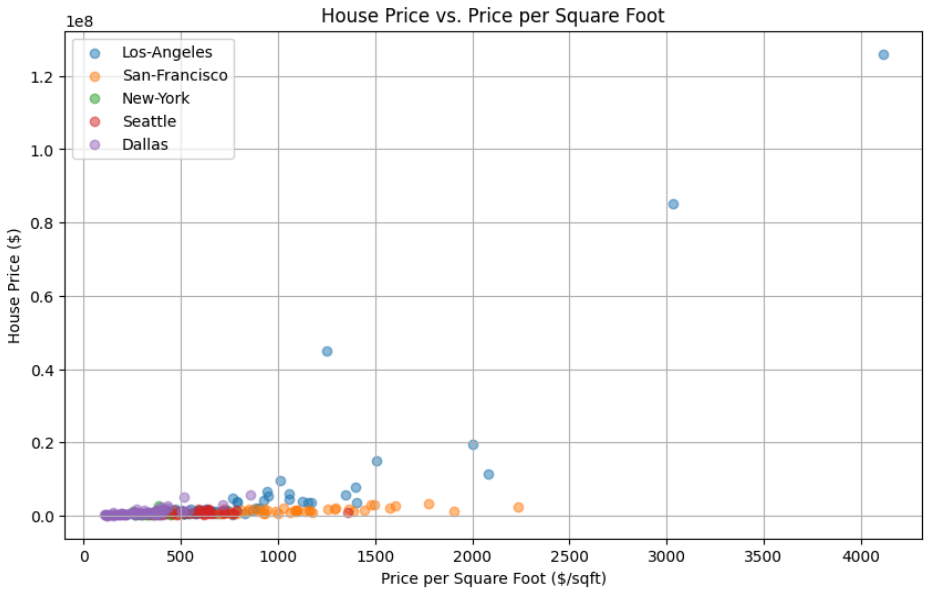
4. **Use API to get historical house prices:**

Attom API was used to find and gather property information and historical assessment data. First, property IDs were retrieved based on specified criteria and then fetched detailed historical assessment information for each property. The data was organized into a data frame and saved as CSV files. Due to the limitations of the API itself, the exact address was not extracted, so only approximate latitude and longitude were called.

5. **Combine information to make interpretations of data**

First, data was cleaned to conduct interpretations. The currency symbols and commas were removed from prices. Then the columns were converted to appropriate data types, and irrelevant columns were removed. Missing values are dropped, and a new column for price per square foot is calculated and added. Below is the scatter plot for the house price vs price per square foot.



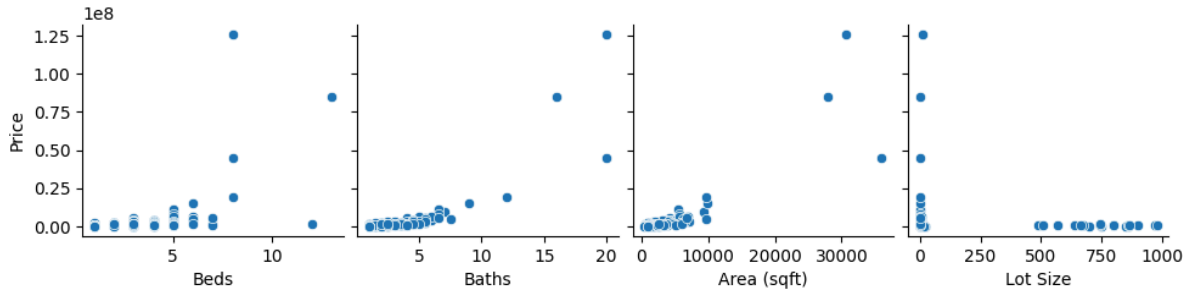


Since the above scatter plot includes all housing information, the cities have been categorized by each color to see how the data differ between cities.

The linear regression model was fitted along with train-test split functionality for model evaluation. Initially, relevant features were selected such as the number of bedrooms, bathrooms, the area in square feet, and lot size, as predictors, while the price serves as the target variable. Subsequently, data was split into training and testing sets. Two linear regression models were

trained: a simple one and a multiple regression model. Both models are evaluated using root mean squared error (RMSE) on both training and testing data. Additionally, relationships between predictors and price were visualized using a pair plot from Seaborn, providing insights into their correlations.

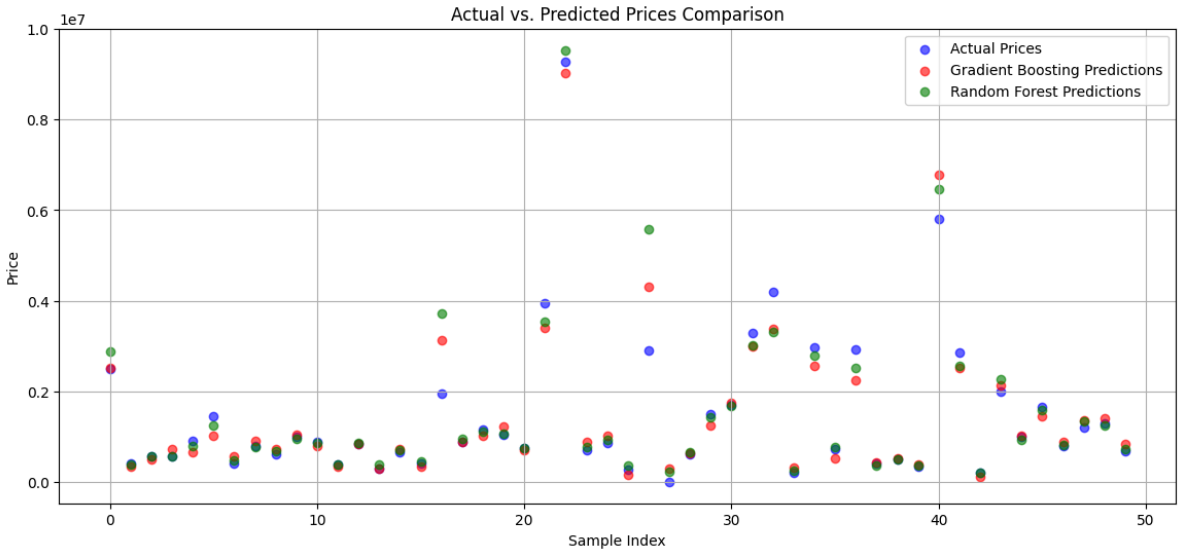
```
Linear Regression Train RMSE: 5647266.412078352
Linear Regression Test RMSE: 1667388.0664332015
Multiple Linear Regression Train RMSE: 5647266.412078352
Multiple Linear Regression Test RMSE: 1667388.0664332015
```



A predictive price model was created using gradient boosting and random forest models. The housing dataset was processed by imputing missing values, excluding features with high missingness, and encoding categorical variables.

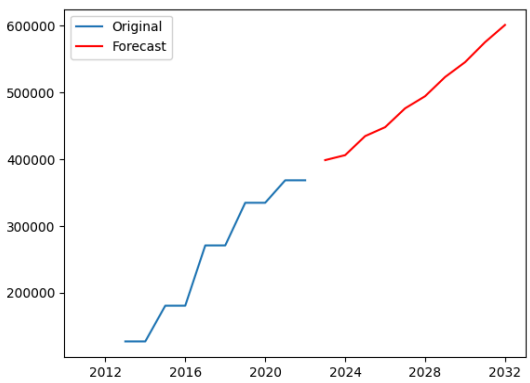
Two predictive models were built and evaluated, a gradient-boosting regressor and a random forest regressor. The models were assessed based on Root Mean Square Error (RMSE) and R-squared metrics, revealing that the Random Forest model underperformed the Gradient-Boosting model in predicting house prices. A visualization comparing actual vs. predicted prices qualitatively assessed the models' performance, indicating both models' effectiveness.

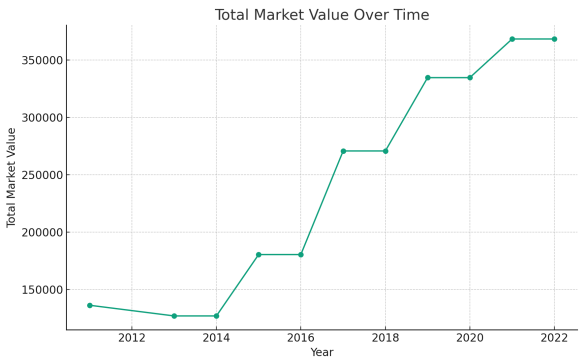
```
Gradient Boosting Regressor RMSE: 596821.0123913316
Random Forest Regressor RMSE: 6863261.678899245


Gradient Boosting Regressor R-squared: 0.916629882541568
Random Forest Regressor R-squared: 0.8602937386705457
```



To further deepen the analysis of the prediction, another predictive model of house prices was modeled using time series. First, historical data from a CSV file was loaded into a pandas data frame. The 'year' column in the DataFrame is converted to DateTime format and set as the index for time series analysis. ARIMA model was then defined with specified parameters, such as order. This model is fitted to the historical data to learn patterns and relationships. based on the fitted values, a forecast of future values was computed. Future timestamps are generated to facilitate plotting. The original data was plotted along with the forecasted values to visualize the ARIMA model. The first graph shows the given data up to now, while the second graph shows the predictive model showing the trend over time.

As there is a positive slope, the price of housing has been trending upward over the last decade. Based on this positive trend, the forecast, drawn by the red line, seems to indicate that the housing price will likely continue to increase in the next few years.





Discussion:

Our study aims to illustrate the complex relationship between real estate prices and various factors in five major U.S. cities with technology industries. This observation is consistent with the economic principle that housing affordability is closely related to local income levels. Although the overall trend suggests higher median incomes are associated with higher real estate prices, however, this trend is not consistent in all cities, highlighting the nuances of the real estate market, where local factors such as the availability of amenities, historic property values, and city policies related to homeownership can significantly impact prices.

In addition, our exploratory data analysis, such as using heat maps and correlation matrices, allows us to visually quantify the relationship between various factors and real estate prices. For example, depending on the city, we observed varying correlations between property characteristics such as the number of bedrooms and bathrooms, square footage, parcel size, and price. This variation suggests market dynamics and homebuyer preferences may differ significantly across urban areas. Our forecasting model uses techniques such as gradient boosting, random forests, and time series to predict future real estate prices based on historical data.

In summary, our report provides a general overview of the current state of real estate, highlighting the impact of economic factors on the real estate market. These findings not only contribute to academic discussions, but also provide practical implications for investors, policymakers, and individuals dealing with real estate investments.

Limitations:

Even though we intended to use it for educational purposes and did not infringe on legal rights or copyrights, we encountered limitations in conducting Realtor web scraping mainly due to commercial protection. The main challenge came from anti-scraping methods that websites implemented to prevent commercial exploitation, which made our scraping work potentially hostile in their view. We used the ScrapFly API server to bypass these measures, but this solution has limitations. Due to the concern of being blocked, there is the limitation that only 240 data points per script were collected per city, which is less than required to perform a more robust

analysis. Additionally, the inability to use automation tools such as Selenium limited our access to detailed information about each property's detail page. Thus, we could only analyze correlations with a few key factors.

Also, while our forecasting models help to make predictions based on historical data, they may not fully accommodate future market changes triggered by unforeseen economic, political, and other situational changes. The volatility of the real estate market means that our findings may become outdated.

References:

Ališauskas, Bernardas. “How to Scrape Real Estate Property Data Using Python.” *ScrapFly Blog*, ScrapFly Blog, 7 Aug. 2023, scrapfly.io/blog/how-to-scrape-real-estate-property-data-using-python/.

Attom Developer Platform, api.developer.attomdata.com/home. Accessed 20 Mar. 2024.

“Housing Prices and Inflation.” *The White House*, The United States Government, 30 Nov. 2021, www.whitehouse.gov/cea/written-materials/2021/09/09/housing-prices-and-inflation/.

Niccolo Conte. “Mapping the Biggest Tech Talent Hubs in the U.S. and Canada.” *Visual Capitalist*, 26 Dec. 2023, www.visualcapitalist.com/biggest-tech-talent-hubs-in-us-canada/#:~:text=California’s%20Bay%20Area%2C%20which%20includes,compared%20to%20378%2C870%20in%202021.&text=Washington%20D.C.&text=Toronto%20remains%20the%20third%20tech,Bay%20Area%20and%20New%20York.

Saad, Lydia. “Real Estate’s Lead as Best Investment Shrinks; Gold Rises.” *Gallup.Com*, Gallup, 7 Feb. 2024, news.gallup.com/poll/505592/real-estate-lead-best-investment-shrinks-gold-rises.aspx.

U.S. Census Bureau Quickfacts: United States, www.census.gov/quickfacts/fact/table/US/VET605222. Accessed 21 Mar. 2024.

Code:

Juptyer Notebook

Github Repo