PROJECT TITLE: Eastern US Housing Market: What Drives It?

TEAM MEMBERS:
ELENA GORGEVSKA, AINA TENG, DYLAN D'COSTA, ISAAC OBODO

DESCRIPTION:
In this data analysis project, we explore the dynamic real estate market to gain valuable insights for investors and sellers. Our project aims to answer fundamental questions about property prices, the correlation between key variables, and factors affecting property costs. We analyze the influence of bedrooms, size, acres, and zip code on overall property costs. The project emphasizes data-driven decision-making and provides stakeholders with valuable insights to navigate the complex real estate market effectively.

DATASETS USED:
To address these research questions, we gathered and analysed a comprehensive dataset covering relevant economic indicators and housing prices for the selected states within the US, using a dataset from Kaggle. Ahmed Shahriar Sakib uploaded Real Estate listings in the US broken by State and zip code less than a month ago.

ROUGH BREAKDOWN OF TASKS:
This project will be conducted as a collaborative effort, with each team member assigned specific roles and responsibilities to ensure a cohesive and efficient workflow. By adopting a pseudo-director approach, accountability will be evenly distributed among team members, fostering a sense of shared ownership and teamwork. The methodology will encompass the following aspects:

Dylan: Lead for coding and data visualization using Matplotlib. Collaborating with Elena on the slide deck and ensuring high-quality data visualizations that meet project requirements.

Aina: Responsible for managing GitHub uploads, ensuring regular commits, and maintaining a professional-quality final repository. She will also contribute to collecting dataset and data cleaning, Jupyter Notebook and written analysis.

Isaac: In charge of the written analysis, tracking progress, and coordinating inputs from other team members. Additionally, he will assist Aina in managing the Jupyter Notebook.

Miloh: Responsible for data management using pandas, ensuring the datasets are relevant and clean.

Data Collection and Preprocessing:
The team will collectively gather the dataset. Data preprocessing will involve cleaning, normalisation, and standardisation to ensure consistency and accuracy. Aina and Isaac will maintain the Jupyter Notebook, keeping track of updates and data integrity.

Data Visualization:
Dylan will lead the team in using Matplotlib, and along with Elena, they will create insightful data visualizations that enhance the understanding of the dataset. Regular meetings will be held to cross-compare progress and ensure the data visualizations align with the project's goals.

Regression Analysis:
The team will collaboratively perform regression analysis for the Eastern US Housing Market. We will conduct the analysis in Jupyter Notebooks, where Aina and Isaac will ensure the datasets used are up-to-date and accurate.

Written Analysis and Report:
Isaac will compile the written analysis based on inputs from all team members. The report will summarise the research questions, methodology, data analysis, and findings. Each team member will contribute their insights to ensure a well-rounded and comprehensive report.

Team Meetings and Collaboration:
The team will meet frequently to discuss progress, share insights, and offer support to each other. Collaborative discussions will help address challenges and ensure everyone is on the same page. The final presentation will be rehearsed together to ensure a seamless and cohesive delivery.

This collaborative approach, with designated pseudo-directors for each component, will empower the team to tackle the project’s objectives. By leveraging the individual strengths of each team member and fostering a cohesive machine-like workflow, we aim to produce valuable insights into what drives housing prices in the US. Together, we will deliver a high-quality analysis and presentation that contributes meaningfully to the field of housing market research and successfully meet the project requirements.

QUESTIONS & ANSWERS

Q1: How does the average property price vary across different regions?

A study analysing the average price of a property in different regions across the country can provide valuable insights for real estate investors and sellers interested in the market.

The initial dataset had over 600,000 rows and was cleaned down to a little over 6300 rows. In this clean dataset, the number of real estate listings varies across different regions. New York has the highest number with 1993 listings, followed by New Jersey with 1543 listings, and Connecticut with 1035 listings. Puerto Rico has 454 listings, while Maine has 428. Pennsylvania has 330 listings, Delaware has 251, and Rhode Island has 199. Massachusetts has 62 listings, and New Hampshire has only 4 listings. Vermont has 3 listings, and the Virgin Islands have just 1 listing. These varying numbers of listings provide insights into the real estate market activity in each region.

We conducted an analysis of average property prices across different regions using a bar plot. Based on the 'PRICE' column in the clean_data DataFrame, the code calculates the average property price per state (or region). The data is grouped by 'STATE' and then selected by 'PRICE' using the groupby function. As a result of the mean() function, the average property prices for each state are calculated. The sort_values(ascending=False) function arranges the states in descending order. As we can see from the bar plot, property prices vary across regions.

To create our bar plot, we used Seaborn, a data visualization library. With a width of 12 inches and a height of 6 inches, the plot is both clear and visually appealing. Plots show the x-axis representing the states (or regions) based on average_price_by_zip series, while y-axis represents average property prices for each state.

We applied a 'coolwarm' color palette to the plot to make it more readable. Moreover, we aligned the tick labels on the x-axis to the right and rotated them 45 degrees. By making this adjustment, state names do not overlap and all labels are legible.

Moreover, we labelled the x and y axes appropriately, indicating state names and average property prices. The title contextualises plot content: 'Region'. Finally, plt.tight_layout() optimises the layout, avoiding overlapping elements within the plot.

Stakeholders in the real estate market can use this visual representation to determine whether a region's average property price is higher or lower, thus aiding in investment or pricing decisions.

Q2(A): What is the correlation, if any between household features and property prices?

As a starting point, we selected the relevant columns from clean_data DataFrame to focus on the variables of interest: 'BED', 'HOUSE SIZE', 'ACRE LOT', and 'PRICE'. The numbers in these columns show the number of bedrooms, house size, and lot size, as well as the price of the property.

To find out how these columns relate, we calculated the correlation matrix. With a heatmap, we could quickly assess the strength and direction of the correlation matrix.

To analyse the correlation between the number of bedrooms and property prices specifically, we utilised a scatter plot. Each point on the scatter plot represents a property with its corresponding number of bedrooms and property price. By observing the scatter plot, we can visually discern any trends or patterns that suggest a correlation between the two variables. The plot provides insights into whether an increase in bedrooms correlates with higher property prices or vice versa.

We encountered properties with five bedrooms that exhibited interesting property price outliers. To further analyse this aspect, we computed relevant statistical metrics for properties with five bedrooms and compared them with properties having more than five bedrooms. By calculating summary statistics, such as mean, median, standard deviation, and quartiles for both groups, we gained deeper insights into property prices' variation and central tendency.

Additionally, we performed statistical significance tests to ascertain whether the differences in property prices between these two groups are statistically significant or merely due to chance.

Overall, the data-driven analysis provides us with a more comprehensive understanding of the impact of different variables on overall property costs. These insights are invaluable for making informed decisions in the dynamic real estate market.

Q2(3): What does the data tell us about which features are deemed more important when deciding listing price? How does scope of data impact the analysis?

The analysis indicates that there are several factors that determine the overall price of a property. Among the variables examined, location (zip code) appears to have the biggest impact on property prices, while acres have the smallest influence. In our analysis, acreage and price are negatively correlated, suggesting that location matters much more when determining property value. Furthermore, the size of the house is also an important factor in determining the price, while bedrooms appear to have a smaller impact.

Q(4): Drilling into one state, Massachusetts, how does lat & lon impact listing prices?

A need to drill down further into our dataset and find out if there were other factors such as zip codes that played a factor in property prices. We created a scatter plot of latitude vs longitude and to make the visualization effective the plots were colored by the price of the listing. We found that generally the more expensive listings were located at the south of the state. We worked with this additional nugget of data we gleaned from the state and plotted the same listing on a map. The map clearly makes the analysis come alive by helping us realize that the expensive properties were by the waterfront and it was a subtle message to us data analysts that location matters..

Bonus: What are the top ten most expensive listings in this dataset?

The top 10 most expensive listings in this Eastern region are primarily located in New York New York and Greenwich Connecticut. We were able to drill down the whole dataset by sorting the values by price, then creating a new dataset of a slice of these top 10 listings and mapping them using hvplot. The conclusion we come to is that while bedrooms and house size are important the strongest driver of house prices is similar to the famous saying that for real estate all that matters is “Location! Locations ! Location!

Limitations:

While we can observe the correlation between property prices and factors such as size, acres, and zip code, understanding the specific elements that make a location appealing is challenging to substantiate with this dataset. The missing aspects here are the lack of historical pricing and sentiment analysis by scanning review sections of listings or comments. This limits our ability to perform a more comprehensive analysis of the underlying factors that drive property value in different locations. As a result, we cannot fully explore the intricate nuances that contribute to a location's attractiveness in the real estate market using this particular dataset. The dataset used was also limited to 17 states in the Eastern Coast of the US and thus cannot paint a comprehensive image of the US as whole.