**HOUSE DATA ANALYSIS**

**Introduction**:

This report presents the analysis of two datasets, the Annual House Price Index (HPI) and the Zip\_Zhvi\_Summary\_AllHomes. The analysis aims to gain insights into the housing market by examining the trends and changes in housing prices over time. By utilizing the capabilities of Spark, we performed various data processing and visualization techniques to extract valuable information from the datasets.

**Analysis 1: Annual House Price Index (HPI)**

The Annual House Price Index dataset provides information about the HPI values for different years. The analysis steps undertaken for this dataset include:

1. Data Loading and Exploration: The HPI dataset was loaded into Spark, and we explored the structure of the dataset, examining the columns and their data types.

A picture containing text, screenshot

Description automatically generated

A screenshot of a computer screen

Description automatically generated with medium confidence

1. Calculating Average HPI by Year: Using Spark SQL, we grouped the data by year and calculated the average HPI for each year. This aggregation allowed us to understand the overall trend of housing prices over time.

A screen shot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated

The results of the analysis showcased the average HPI values for different years, providing insights into the housing market's performance. By observing the changes in average HPI, we could identify periods of significant growth or decline in housing prices.A picture containing screenshot, plot, diagram, line

Description automatically generated

Insights:

* The average HPI serves as an indicator of the housing market's performance.
* Years with higher average HPI indicate periods of growth and increased housing prices.
* Analyzing the average HPI helps in identifying the overall trend of the housing market and understanding periods of market fluctuations.

**Analysis 2: Zip\_Zhvi\_Summary\_AllHomes Data**

The Zip\_Zhvi\_Summary\_AllHomes dataset contains detailed information about housing price indices for different ZIP codes. The analysis steps undertaken for this dataset include:

1. Data Loading and Exploration: The ZIP code housing price dataset was loaded into Spark, and we examined the structure of the dataset, including column names and data types.

A screen shot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated

1. Annual Change in HPI for a Specific ZIP Code: Focusing on a specific ZIP code (ZIP code 1001), we analyzed the annual changes in the HPI. By comparing HPI values between consecutive years, we calculated the percentage change in the HPI.

A picture containing screenshot, plot, diagram, line

Description automatically generated

Insights:

* The annual change in HPI helps identify the fluctuations in housing prices for a specific ZIP code.
* Positive percentage changes indicate an increase in housing prices, while negative changes represent a decrease.
* Analyzing the annual change in HPI can help identify the housing market's volatility and understand the trends specific to a ZIP code.

*References:*

Chen, T., & Guestrin, C. (2016). *XGBoost*. <https://doi.org/10.1145/2939672.2939785>

*S Spark: Cluster Computing with Working Sets*. (2019). <https://www.usenix.org/legacy/events/hotcloud10/tech/full_papers/Zaharia.pdf>