**ALY6110 70591 Data Management & Big Data SEC 08 Fall 2023 CPS**

**Module 4 Assignment — Practice with Spark REPORT**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Hema Seshadri

**Date**

12/02/2023

**Analysis of House Price Indexes and Home Values using Spark**

**Introduction:**

This report presents an analysis of two government-sourced datasets, "Annual House Price Indexes - Five-Digit ZIP Codes" and "Annual House Price Indexes - Three-Digit ZIP Codes," alongside a commercially sourced dataset, "Median Home Value – Zillow Home Value Index (ZHVI) by Zip Code." The objective is to utilize Spark and R to process these large datasets effectively and extract meaningful insights, particularly focusing on trends, variations, and correlations in house price indexes and home values.

**Analysis and Results:**

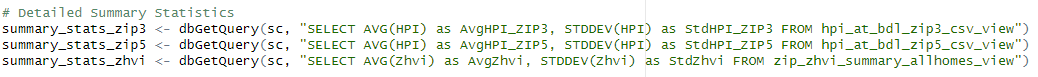
**Analysis and Results:**

The datasets were loaded into Spark using the SparklyR library, and temporary SQL views were created for each dataset for ease of processing and analysis.

A computer code on a white background

Description automatically generated

**Summary Statistics:**

The analysis began with computing detailed summary statistics for each dataset:  


* For the three-digit ZIP code dataset, the average House Price Index (HPI) was found to be approximately 306.2 with a standard deviation of 219.35.  
  
* The five-digit ZIP code dataset showed an average HPI of around 231.93 with a standard deviation of 180.76.  
  
* The Zillow Home Value Index (ZHVI) for all homes averaged at 220,009.4 with a substantial standard deviation of 214,881.2, indicating significant variability in home values.  
  

**Trend Analysis:**

The trend analysis revealed a steady increase in the average HPI for both three-digit and five-digit ZIP codes from 1975 to 2020. This trend was visualized in a line graph, which clearly depicted the upward trajectory of HPI over the years.  
A screenshot of a computer

Description automatically generated

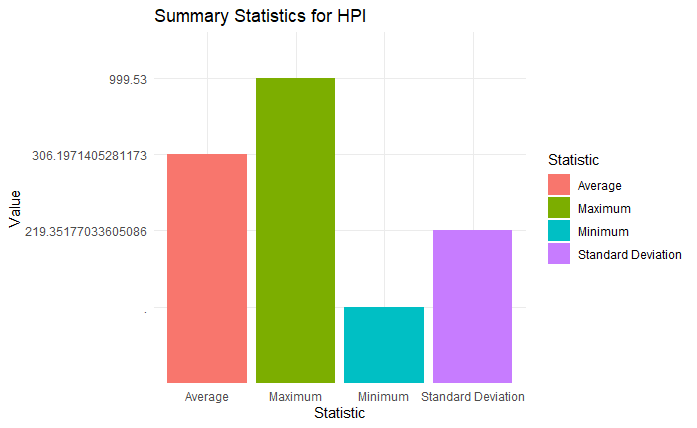
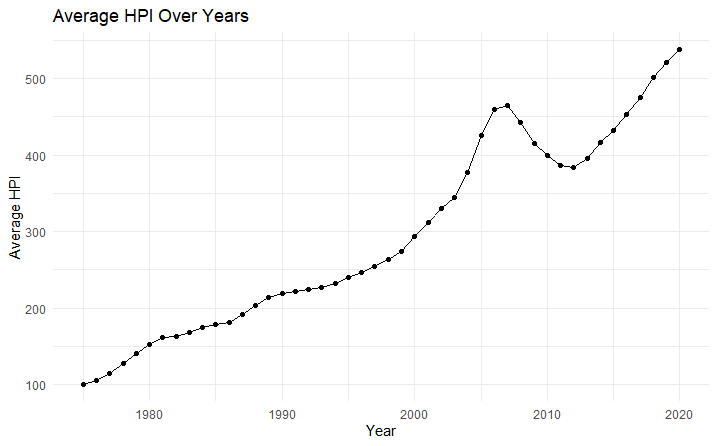
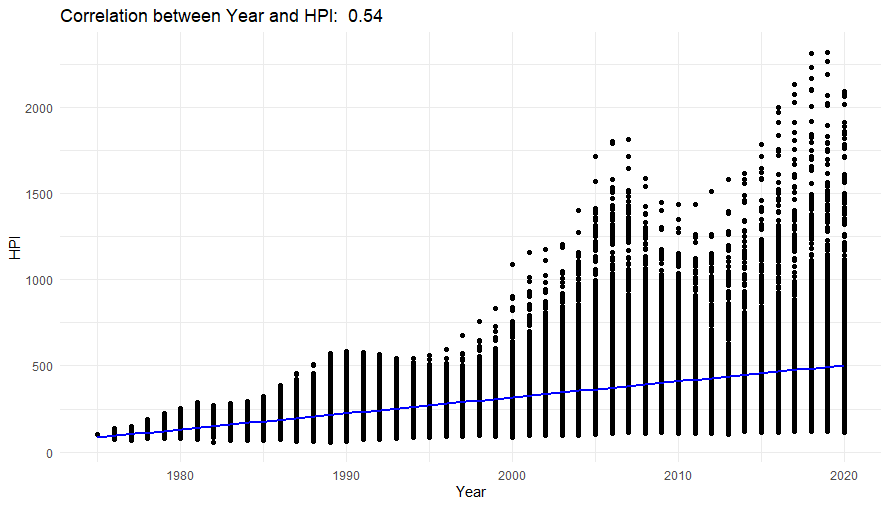
**Comparison Between ZIP Code Datasets:**

A comparative analysis between the three-digit and five-digit ZIP code datasets indicated consistently higher average HPI values for the three-digit ZIP codes across the analyzed period.

A graph showing the number of numbers

Description automatically generated

**Visualizations:**

1. **Summary Statistics Visualization**: A bar graph illustrated the average, minimum, maximum, and standard deviation of the HPI, providing a clear visual representation of the spread and central tendency of the data.  
   
2. **Trend Analysis Graph**: The line graph of HPI trends over the years for both datasets offered a visual narrative of the rising house price indexes, making the growth pattern easily comprehensible.  
   
3. **Correlation Graph**: The scatter plot, along with a fitted line, demonstrated a moderate positive correlation (approximately 0.54) between the Year and HPI, suggesting that house prices have generally increased over time.  
   

**Insights:**

1. **Spatial Variation in House Price Appreciation**

The analysis of the datasets aligns with findings by Bogin, Doerner, and Larson (2019) in "Local House Price Dynamics: New Indices and Stylized Facts." They observe that house price appreciation rates decrease with distance from the central business district (CBD) in large cities, suggesting an increasing desirability of housing units closer to CBD locations. Our trend analysis, which shows a general increase in house prices over the years, might be capturing this central city appreciation effect, especially in the three-digit ZIP code areas that likely include city centers.

2. **Non-Stationarity and Supply Elasticity**

Consistent with the observations in the same study, non-stationarity in real house prices near CBDs aligns with our finding of high price volatility, particularly in three-digit ZIP code areas. This volatility could indicate lower supply elasticity in central city locations, where housing expansion is challenged by limited space and higher demand.

3. **Price Dynamics Post-Acceleration**

Their research also discusses price dynamics post-acceleration. The sustained real price increases and high price volatility near city centers, as observed in our data, might reflect underlying economic fundamentals in these areas. This could explain the more pronounced trends in the three-digit ZIP code areas, which likely include major city centers.

4. **Implications for Mortgage Credit Modeling**

The findings emphasize the importance of using localized house price indices for mortgage credit modeling. Our analysis, which provides detailed insights into local HPI trends, supports this notion. The more granular data from five-digit ZIP codes could be particularly valuable for predicting property values and assessing mortgage performance, as it captures submarket variations more effectively.

5. **Broader Economic Implications**

The trend of increasing house prices, especially in areas closer to city centers, has broader implications for urban development, housing affordability, and economic policy. It highlights the need for careful urban planning and policy interventions to ensure affordable housing availability, especially in rapidly appreciating central city areas.

**References:**

Alexander N. Bogin, William M. Doerner, William D. Larson. 2019. "*Local House Price Dynamics: New Indices and Stylized Facts.*" Real Estate Economics, 47(2), 365-398. <https://onlinelibrary.wiley.com/doi/full/10.1111/1540-6229.12233>

Alexander N. Bogin, William M. Doerner, William D. Larson. 2019. "*Local House Price Paths: Accelerations, Declines, and Recoveries*." Journal of Real Estate Finance and Economics, 58(2), 201-222. <https://link.springer.com/article/10.1007/s11146-017-9643-y>

Alexander N. Bogin, William M. Doerner, William D. Larson. 2019. "*Missing the Mark: Mortgage Valuation Accuracy and Credit Modeling*." Financial Analysts Journal, 75(1), 32-47. <https://www.tandfonline.com/doi/abs/10.1080/0015198X.2018.1547051>

Javier Luraschi, K. K. (n.d.). *Mastering spark with R*. Chapter 2 Getting Started. https://therinspark.com/starting.html#starting-prerequisites