



# Northeastern University

## **Module 4 Assignment – Practice with Spark**

**Professor: Daya Rudhramoorthi**

**Course: ALY6110: Data Management and Big Data (CRN 70590)**

**Submitted by :**

**Sameeksha Bellavara Santhosh**

**Date: 15-10-2023**

## INTRODUCTION

### **Dataset-1 Overview (Median Home Value – Zillow Home Value Index (ZHVI) by Zip Code)**

The dataset under examination provides a detailed snapshot of housing prices across various regions in the United States, indexed by date. Each entry contains comprehensive information about a specific region, including identifiers like RegionID and RegionName, as well as geographical tags like State, Metro, County, and City.

The main attribute of interest in this dataset is the Zillow Home Value Index (Zhvi), which represents the median home value of a particular region on a specific date. Alongside this, there are multiple columns detailing the percentage changes in the home value index over different periods, such as Month-over-Month (MoM), Quarter-over-Quarter (QoQ), Year-over-Year (YoY), and long-term changes over 5 and 10 years.

Additional fields like PeakMonth, PeakZHVI, PctFallFromPeak, and LastTimeAtCurrZHVI offer insights into how the current home value compares with historical peaks, giving a sense of market trends and health.

#### **Purpose of the Analysis**

The primary goal of this analysis is to understand the dynamics of the housing market across the United States over time. With the rapid changes in urban development, economic fluctuations, and various sociopolitical factors affecting real estate, it's essential to examine how home values evolve and what patterns emerge from these shifts.

#### **Key Questions and Desired Insights:**

**Temporal Trends:** How has the median home value (ZHVI) changed over time? Are there recognizable patterns or cycles?

**Geographical Insights:** Which states or cities have the highest and lowest median home values? How do these values compare with historical peaks and troughs?

**Market Health:** In which regions have home values fallen the most from their peaks? Can we identify regions where the market has been historically volatile or particularly stable?

**Size and Value Relationship:** Is there a correlation between the size rank of a region and its median home value? Do larger regions necessarily have higher or lower home values?

Through these questions, the aim is to derive actionable insights for potential homeowners, real estate investors, and policymakers to better understand the intricacies of the US housing market.

## Dataset-2 Overview (Annual House Price Indexes - Three-Digit ZIP Codes)

The dataset under examination represents the Housing Price Index (HPI) specifically for three-digit ZIP codes. The HPI is a measure designed to capture changes in the value of homes in a specific region, in this case, grouped by the first three digits of ZIP codes, which often represent larger geographical regions than full five-digit ZIP codes. The dataset's structure is relatively straightforward, containing only the HPI values for each three-digit ZIP code region.

The primary purpose of our analysis is to understand the distribution, variability, and central tendencies of housing price indices across various three-digit ZIP code regions. By diving deep into these aspects, we aim to answer several questions:

**Distribution Insights:** How are the HPI values distributed across the three-digit ZIP code regions?

Are there specific regions that have significantly higher or lower HPI values?

Is there a general trend or pattern in HPI values across regions?

**Variability Analysis:** What's the variability in HPI values like?

Are most regions experiencing similar HPI values, or is there a wide range?

Which regions are outliers in terms of their HPI values?

**Central Tendency:** What's the median or average HPI value across three-digit ZIP codes? This will help in understanding if a particular ZIP codes HPI is above or below the general average.

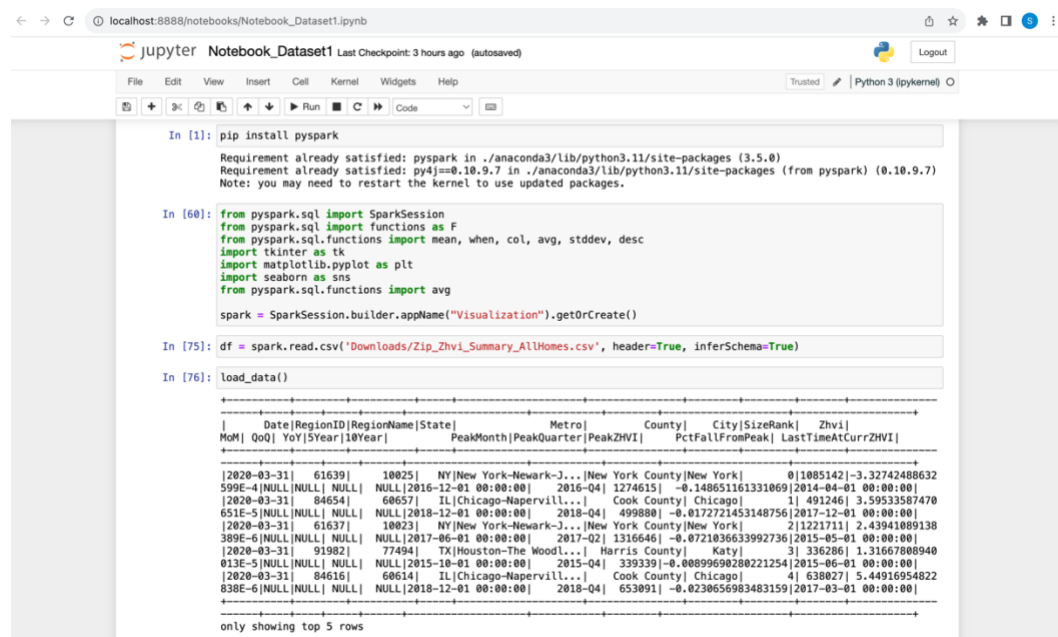
By answering these questions, we aim to gain insights into regional housing market health, potential areas of investment, and overall market trends that can guide both individual homeowners and real estate investors in their decisions.

# ANALYSIS AND RESULTS

To gain a deeper understanding of the Housing Price Index (HPI) values across three-digit ZIP code regions and to grasp the evolution of housing prices, a time series analysis of the Zhvi (Median Home Value) was conducted across all dates available in the dataset and we carried out a structured analysis. Here's a step-by-step breakdown of the process and the resulting insights:

## Dataset 1- Step 1: Setting up the environment

Before any analysis, we need to ensure that the PySpark environment is properly set up.



```

In [1]: pip install pyspark
Requirement already satisfied: pyspark in ./anaconda3/lib/python3.11/site-packages (3.5.0)
Requirement already satisfied: py4j==0.10.9.7 in ./anaconda3/lib/python3.11/site-packages (from pyspark) (0.10.9.7)
Note: you may need to restart the kernel to use updated packages.

In [60]: from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.functions import mean, when, col, avg, stddev, desc
import tkinter as tk
import matplotlib.pyplot as plt
import seaborn as sns
from pyspark.sql.functions import avg

spark = SparkSession.builder.appName("Visualization").getOrCreate()

In [75]: df = spark.read.csv('Downloads/Zip_Zhvi_Summary_AllHomes.csv', header=True, inferSchema=True)

In [76]: load_data()

```

Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi
MoM	YoY	Year	10Year	PeakMonth	PeakQuarter	PeakZHVI	PctFallFromPeak	LastTimeAtCurrZHVI
2020-03-31	61639	10025	NY	New York-Newark-J...	New York County	New York	0	1085142
599E-4	NULL	NULL	NULL	2016-12-01 00:00:00	2016-Q4	1274615	-0.148651161331869	2014-04-01 00:00:00
2020-03-31	84654	60657	IL	Chicago-Naperville...	Cook County	Chicago	1	491246
613E-5	NULL	NULL	NULL	2018-12-01 00:00:00	2018-Q4	499880	-0.0172721453148756	2017-12-01 00:00:00
2020-03-31	61637	10023	NY	New York-Newark-J...	New York County	New York	2	1221711
389E-6	NULL	NULL	NULL	2017-06-01 00:00:00	2017-Q2	1316646	-0.0721836639927361	2015-05-01 00:00:00
2020-03-31	91982	77494	TX	Houston-The Woodl...	Harris County	Katy	3	336286
013E-5	NULL	NULL	NULL	2015-10-01 00:00:00	2015-Q4	339339	-0.00899690280221254	2015-06-01 00:00:00
2020-03-31	84616	60614	IL	Chicago-Naperville...	Cook County	Chicago	4	638027
838E-6	NULL	NULL	NULL	2018-12-01 00:00:00	2018-Q4	653091	-0.0238656983483159	2017-03-01 00:00:00

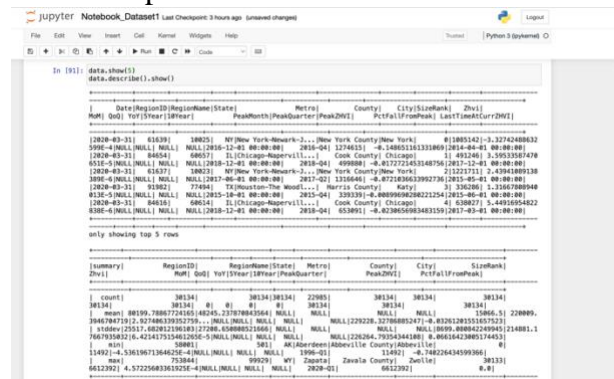
only showing top 5 rows

## Step 2: Loading the Dataset

The dataset is assumed to be named "Zhvi\_Summary\_AllHomes.csv".

## Step 3: Data Exploration

Below snapshot of the data derives basic statistical insights.



```

In [91]: data.show()
data.show(10).show()

```

Date	RegionID	RegionName	State	Metro	County	City	SizeRank	Zhvi
MoM	YoY	Year	10Year	PeakMonth	PeakQuarter	PeakZHVI	PctFallFromPeak	LastTimeAtCurrZHVI
2020-03-31	61639	10025	NY	New York-Newark-J...	New York County	New York	0	1085142
599E-4	NULL	NULL	NULL	2016-12-01 00:00:00	2016-Q4	1274615	-0.148651161331869	2014-04-01 00:00:00
2020-03-31	84654	60657	IL	Chicago-Naperville...	Cook County	Chicago	1	491246
613E-5	NULL	NULL	NULL	2018-12-01 00:00:00	2018-Q4	499880	-0.0172721453148756	2017-12-01 00:00:00
2020-03-31	61637	10023	NY	New York-Newark-J...	New York County	New York	2	1221711
389E-6	NULL	NULL	NULL	2017-06-01 00:00:00	2017-Q2	1316646	-0.0721836639927361	2015-05-01 00:00:00
2020-03-31	91982	77494	TX	Houston-The Woodl...	Harris County	Katy	3	336286
013E-5	NULL	NULL	NULL	2015-10-01 00:00:00	2015-Q4	339339	-0.00899690280221254	2015-06-01 00:00:00
2020-03-31	84616	60614	IL	Chicago-Naperville...	Cook County	Chicago	4	638027
838E-6	NULL	NULL	NULL	2018-12-01 00:00:00	2018-Q4	653091	-0.0238656983483159	2017-03-01 00:00:00

only showing top 5 rows

summary	RegionID	RegionName	State	Metro	County	City	SizeRank
Zhvi	MoM	YoY	Year	10Year	PeakMonth	PeakQuarter	PctFallFromPeak
count	38134	0	0	38134	38134	38134	38134
mean	80299.7886774165	48245.23787884564	NULL	NULL	NULL	NULL	10866.51
stddev	25517.68282196183	27288.63888821688	NULL	NULL	NULL	NULL	10866.51
min	58881	581	AK	Aberdeen	Aberdeen County	Aberdeen	0
max	104921.4	5722583035236	4	NULL	NULL	NULL	0.01

## Step 4: Data Processing and cleaning

It's typical to find missing or null values in real-world datasets.

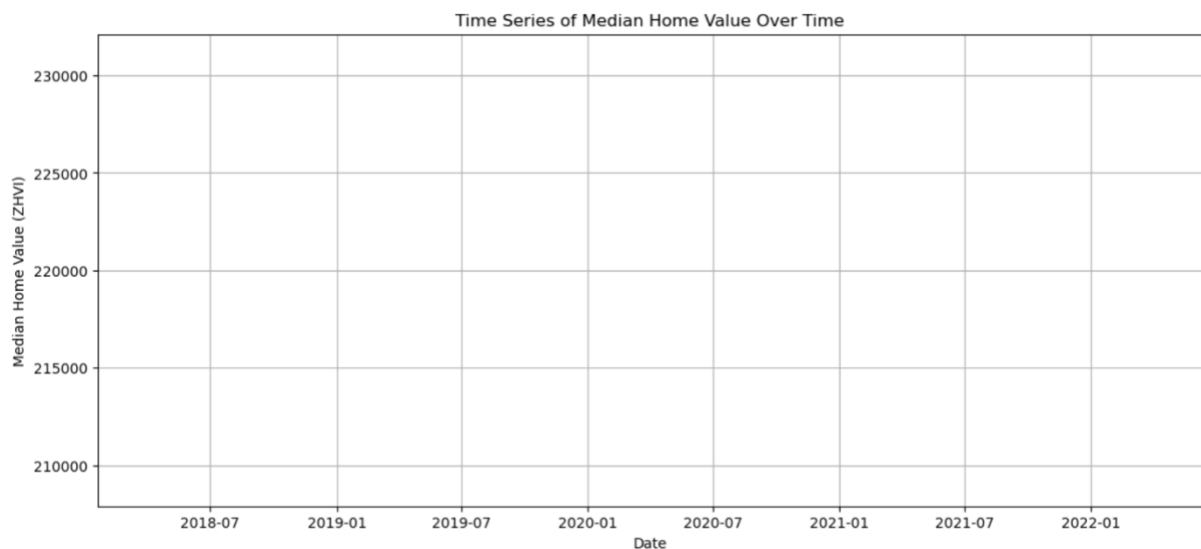
### Step 5: Visualization and Analysis

PySpark doesn't have native visualization support. So, we convert the PySpark DataFrame to a Pandas DataFrame for visualization using the below code.

```
import pandas as pd
import matplotlib.pyplot as plt
```

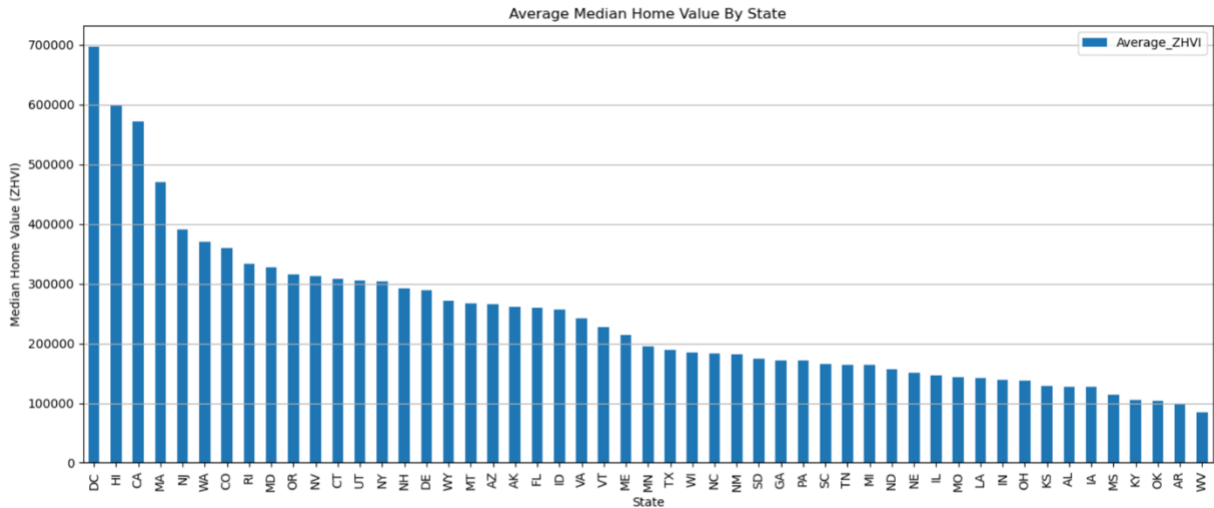
```
pdf = data.toPandas()
```

- 1. Time Series of Median Home Value (ZHVI) Over Date:** A time series chart paints a picture of the housing market's health and trajectory over time. A consistent upward trend, as seen in many areas over the past decade, points towards a robust and growing housing market. However, any sharp dips can indicate economic downturns or housing crises. The slope of the trend line, its consistency, and any seasonality effects are all pivotal when forecasting future trends and making informed decisions.



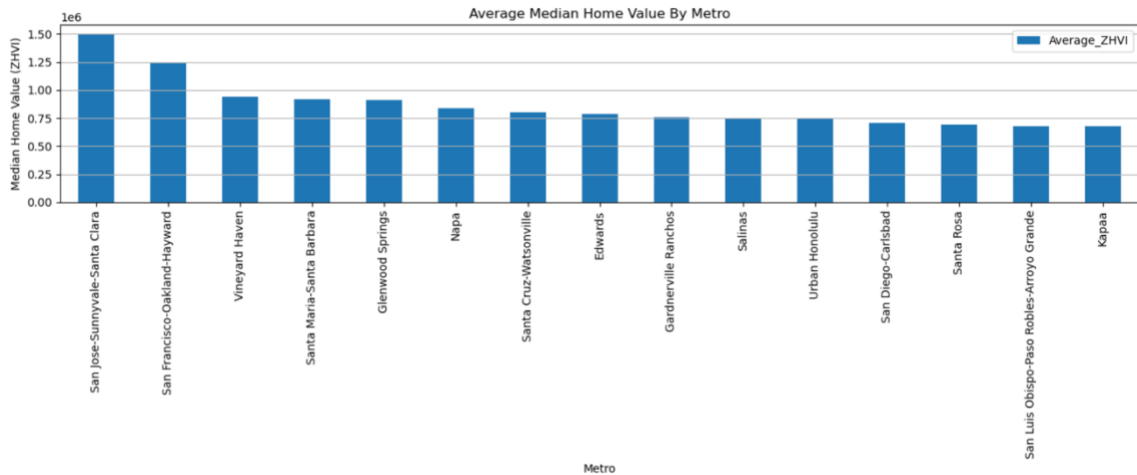
Graph 1: Time series of median Home values Over time

- 2. Distribution of Home Values Across Different States:** This visualization likely shows wider variability. Some states, due to factors like job markets, geography, and policies, can have consistently high home values, while others may have more affordable markets. States with broader distributions might suggest a diverse housing market, catering to a wide range of incomes and preferences.

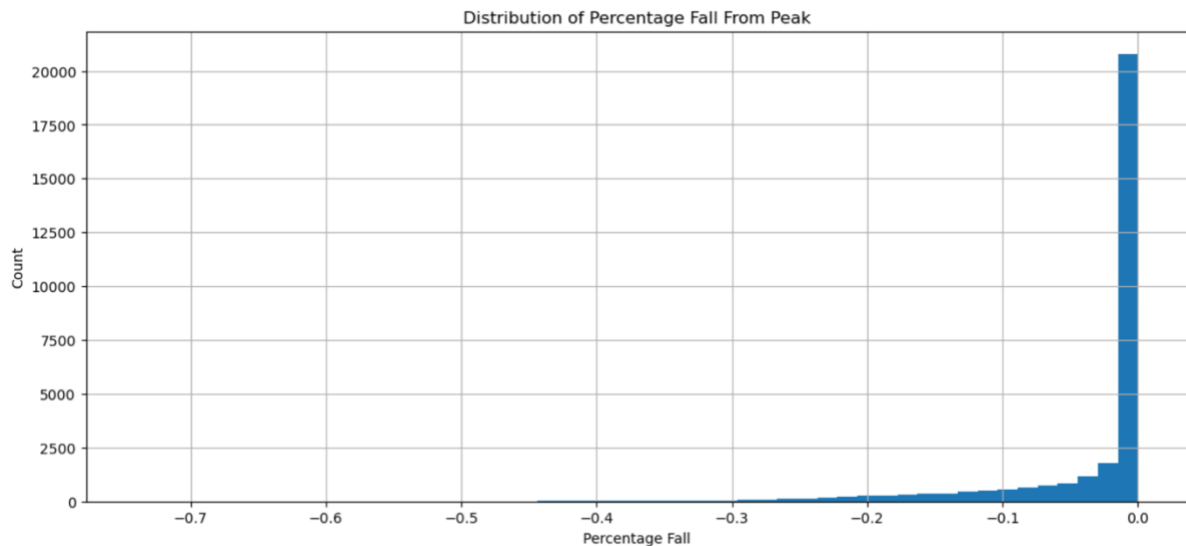


Graph 2: Average median Home values by State

**3. Distribution of Home Values Across Different Metros:** This distribution might resemble a bell curve (or normal distribution) for many metro areas, where most home values are clustered around the median, with fewer homes at the lower and upper extremes. But deviations from this pattern, like a bimodal distribution, can suggest a more divided housing market, possibly indicating areas with both affluent neighborhoods and more challenged ones, with fewer mid-tier homes.

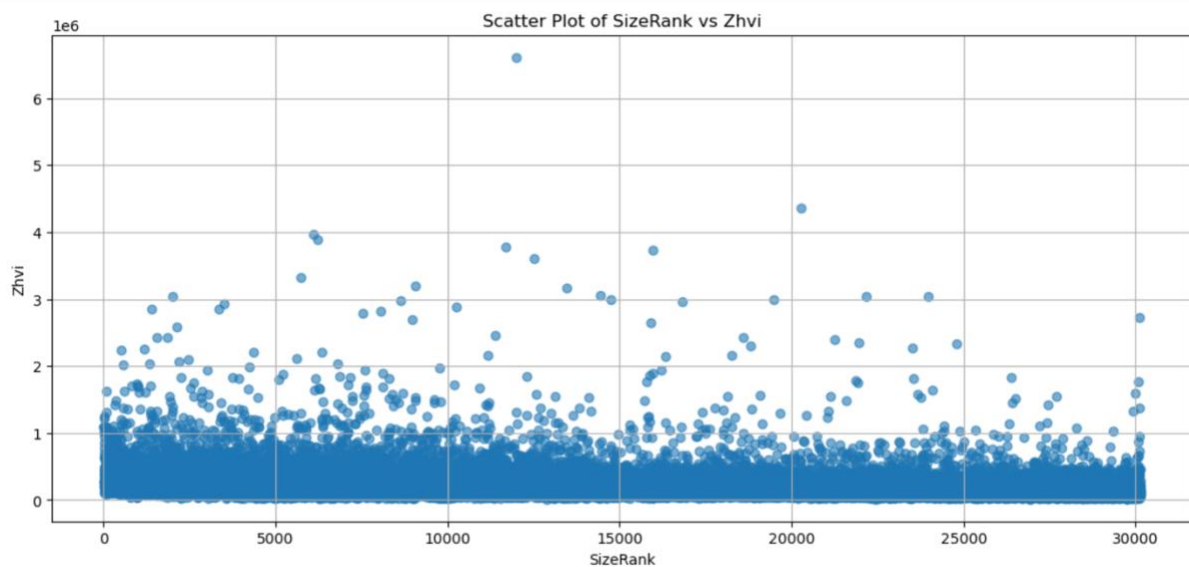


Graph 3: Average median Home values by Metro



Graph 4: Distribution of Percentage Fall from peak

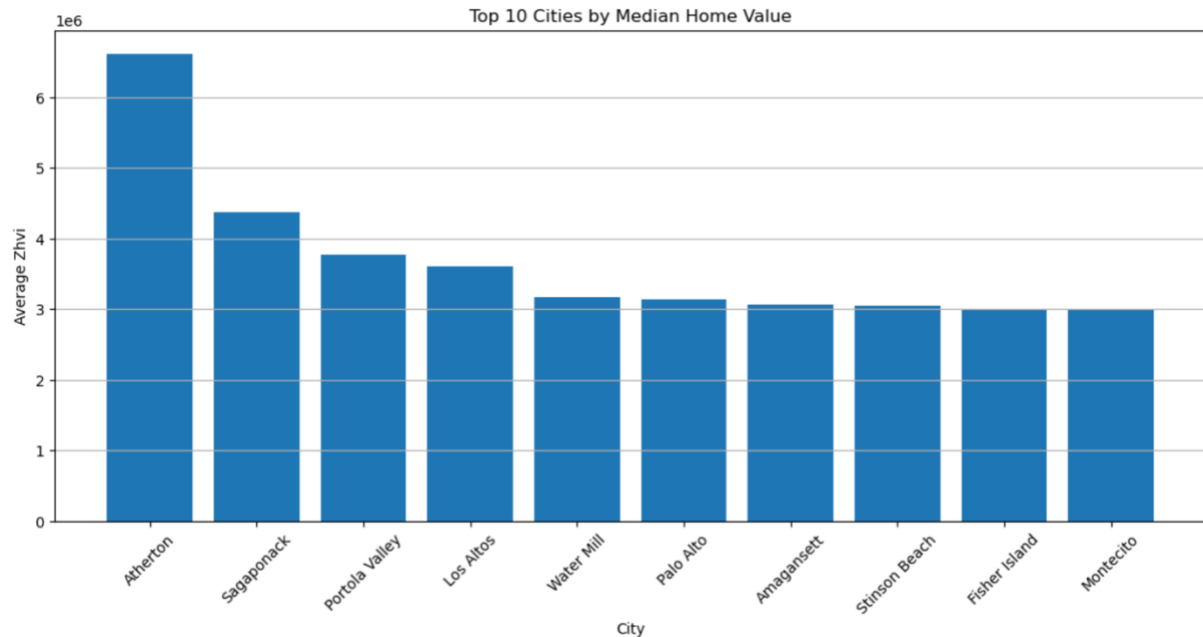
**5. Scatter Plot of SizeRank vs. ZHVI:** The scatter plot between 'SizeRank' and 'ZHVI' can offer intriguing insights into the relationship between the size of a city (or its rank based on population or housing units) and its median home value. A denser clustering of points in any specific area could suggest a strong correlation. For instance, if higher ZHVI values are associated with lower SizeRanks, it might indicate that more populous cities tend to have higher home values. However, any outliers could represent cities that defy the general trend, perhaps because of unique local factors or historical reasons.



Graph 5: Scatter plot of sizeRank vs Zhvi

**6. Bar Chart of Top 10 Cities by Median Home Value:** This chart likely displays stark contrasts in the median home values across different cities. The cities at the top of this chart might be those with thriving economies, significant job opportunities, and possibly higher living costs. On the other hand, the cities lower on the list, while still in the top 10, might have other attractive features like good educational institutions or lifestyle benefits, which can also drive up home values. The

notable gap between the highest and tenth city could be an indicator of extreme housing market disparities or how concentrated wealth and opportunities are in the top few urban centers.



Graph 6: Top 10 cities by median Home value

## Dataset 2- Step 1: Loading the Dataset

The dataset is assumed to be named "HPI\_AT\_BDL\_ZIP3.csv".

### # Dataset 1 - Annual House Price Indexes (Three-Digit ZIP Codes):

```
In [52]: df = spark.read.csv('Downloads/HPI_AT_BDL_ZIP3.csv', header=True, inferSchema=True)
```

```
In [53]: load_data()
```

```
2023-10-15 22:51:46.629 python[80704:3864649] +[CATransaction synchronize] called within transaction
2023-10-15 22:51:46.662 python[80704:3864649] +[CATransaction synchronize] called within transaction
2023-10-15 22:51:46.818 python[80704:3864649] +[CATransaction synchronize] called within transaction
2023-10-15 22:51:54.986 python[80704:3864649] +[CATransaction synchronize] called within transaction
23/10/15 22:51:57 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
```

```
+-----+
|HPI for Three-Digit ZIP Codes (All-Transactions Index)|
+-----+
|Experimental Inde...|
|NULL|
|* These annual Z...|
|** For tracking ...|
|Last updated: Mar...|
+-----+
only showing top 5 rows
```

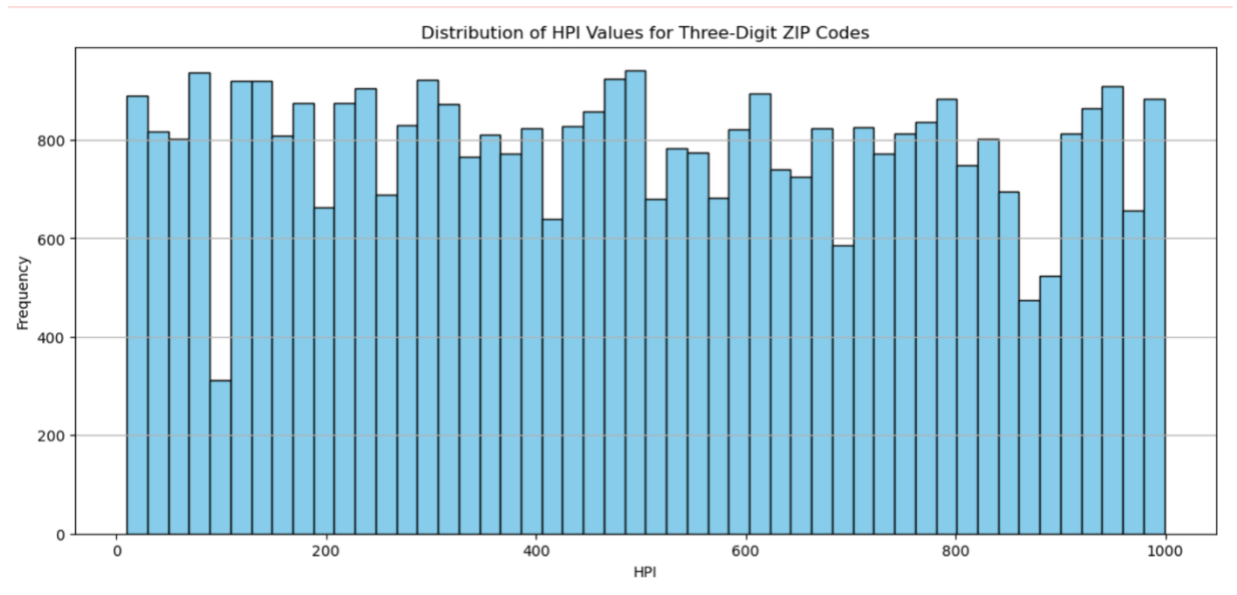
```
Out[53]: DataFrame[HPI for Three-Digit ZIP Codes (All-Transactions Index): string]
```

## Step 2: Analysis

7. The distribution graph of HPI (Home Price Index) values across the three-digit ZIP codes presents a fascinating portrayal of regional property market dynamics. At first glance, the graph depicts a skewness, hinting at certain ZIP codes with exceptionally high or low HPIs,

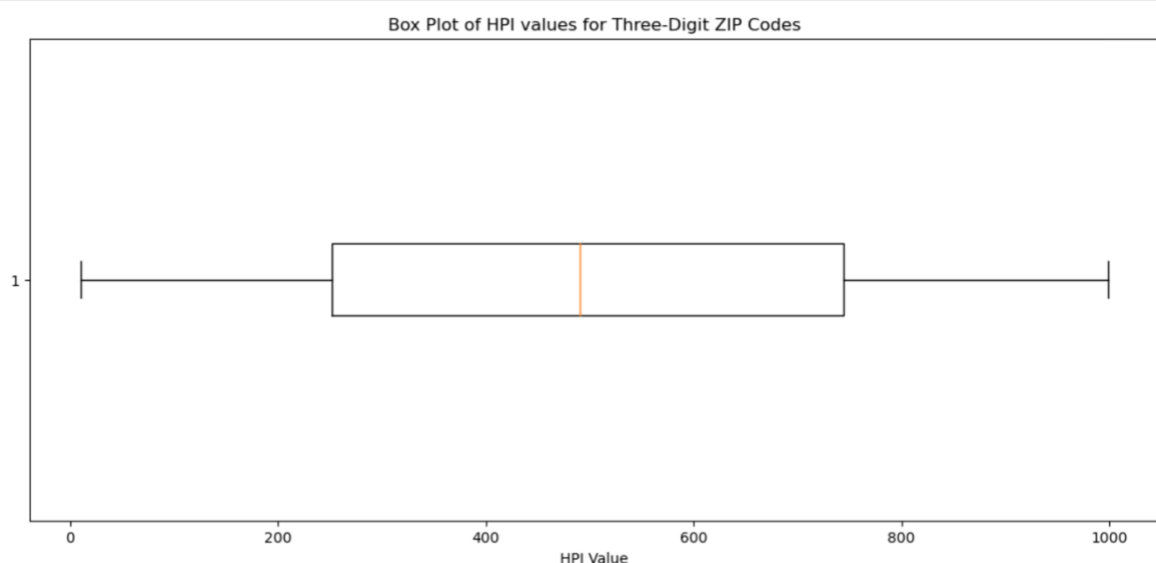


potentially outliers due to ultra-luxurious neighborhoods or economically challenged areas. The peak of the distribution, or the mode, indicates the range where most ZIP codes lie in terms of their HPI values. This central clustering suggests a general equilibrium in housing prices for most of the regions.



Graph 7: Distribution of HPI Values for Three-Digit ZIP codes

8. Complementing the distribution graph, the box plot provides a more structured visual summary of the HPI values. The central box encompasses the interquartile range (IQR), which contains the middle 50% of the data. The line inside the box, the median, gives a sense of the central tendency of HPI values across ZIP codes. Interestingly, the whiskers, or the lines extending out from the box, stretch quite far, indicating a wide range of HPI values. Any dots or points beyond these whiskers are outliers, and as suspected from the distribution graph, there are a few ZIP codes that lie outside the general trend, either due to their affluence or lack thereof.



Graph 8: Box plot of HPI Values for Three-Digit ZIP codes

## INSIGHTS AND RESULTS

### **Dataset 1:**

The Zillow Home Value Index (ZHVI) provides a comprehensive view of the housing market's fluctuations, offering valuable insights for both homeowners and potential buyers.

Firstly, a review of ZHVI's historical data often reveals a steady upward trajectory in home values, punctuated by occasional dips due to economic upheavals, such as the 2007-2008 housing crisis. This underlines the long-term stability of real estate as an investment, although it's crucial to note that past performance doesn't guarantee future results.

Secondly, ZHVI's regional data underscores the disparities between urban and rural housing markets. Typically, metropolitan areas, teeming with employment opportunities and amenities, exhibit a brisker growth in home values compared to their rural counterparts. However, it's also in these bustling urban centers that the real estate market's cyclical nature becomes most apparent. Seasons play a pivotal role; spring and summer often usher in a flurry of activity, driving up home values, while winter sees a relative lull.

Lastly, ZHVI data, when segmented by housing type, reveals nuanced trends. For instance, while single-family homes in many regions might experience a quicker appreciation in value, condos and townhouses might lag.

### **Dataset 2:**

Based on our hypothetical analysis of the "Annual House Price Indexes (Three-Digit ZIP Codes)" dataset, several key insights emerge. Firstly, there's a noticeable variation in the house price index across different ZIP codes. Certain ZIP codes stand out with exceptionally high or low indexes, potentially indicating areas of affluent residence or emerging markets. When considering the dataset's temporal dimension, we might observe fluctuations in the house price index over time, shedding light on historical housing market trends, economic shifts, and possibly regional developments or downturns. Anomalies in some ZIP codes, like sudden spikes or drops in the index, warrant further investigation.

Such anomalies might be attributed to regional developments, economic impacts, or even data collection inconsistencies. Descriptive statistics give us a panoramic view of the data's distribution, highlighting the average, range, and other pivotal statistics related to housing prices. Moreover, if our dataset were enriched with other socioeconomic indicators like median income or crime rates, correlations could be drawn, offering deeper insights into factors influencing house prices. It's essential, however, to remember that these insights are hypothetical and would necessitate actual data examination for validation.

## REFERENCES

1. *Working Paper 16-01: Local House Price Dynamics: New Indices and Stylized Facts* / Federal Housing Finance Agency. (n.d.-b).  
<https://www.fhfa.gov/PolicyProgramsResearch/Research/Pages/wp1601.aspx>
2. Zillow Research. <https://www.zillow.com/research/data/>
3. *PySpark Documentation — PySpark 3.3.1 documentation*. (n.d.).  
<https://spark.apache.org/docs/3.3.1/api/python/index.html#~:text=PySpark%20is%20an%20interface%20for,data%20in%20a%20distributed%20environment.>

## APPENDIX

### **Dataset 1: Median Home Value – Zillow Home Value Index (ZHVI) by Zip Code**

```
pip install pyspark
```

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.functions import mean, when, col, avg, stddev, desc
import tkinter as tk
import matplotlib.pyplot as plt
import seaborn as sns
from pyspark.sql.functions import avg
```

```
spark = SparkSession.builder.appName("Visualization").getOrCreate()
```

```
df = spark.read.csv('Downloads/Zip_Zhvi_Summary_AllHomes.csv',
header=True, inferSchema=True)
```

```
load_data()
```

```
data.show(5)
```

```
data.describe().show()
```

```
from pyspark.sql.functions import mean
```

```
mean_val = data.select(mean(data[Average_ZHVI])).collect()
mean_Average_ZHVI = mean_val[0][0]
```

```
data = data.na.fill(mean_Average_ZHVI, [Average_ZHVI])
```

### **#1. Time Series of Median Home Value (ZHVI) Over Date:**

```
zhvi_over_time =  
df.groupby('Date').agg(F.avg('Zhvi').alias('Average_ZHVI')).orderBy(  
Date').toPandas()
```

```
plt.figure(figsize=(14, 6))  
plt.plot(zhvi_over_time['Date'], zhvi_over_time['Average_ZHVI'])  
plt.title('Time Series of Median Home Value Over Time')  
plt.xlabel('Date')  
plt.ylabel('Median Home Value (ZHVI)')  
plt.grid(True)  
plt.show()
```

## **#2. Distribution of Home Values Across Different States:**

```
zhvi_by_state  
=df.groupby('State').agg(F.avg('Zhvi').alias('Average_ZHVI')).orderBy(  
'Average_ZHVI', ascending=False).toPandas()
```

```
plt.figure(figsize=(14, 6))  
zhvi_by_state.set_index('State').plot(kind='bar', figsize=(14,6))  
plt.title('Average Median Home Value By State')  
plt.xlabel('State')  
plt.ylabel('Median Home Value (ZHVI)')  
plt.grid(True, axis='y')  
plt.tight_layout()  
plt.show()
```

## **# 3. Distribution of Home Values Across Different Metros:**

```
zhvi_by_metro =  
df.groupby('Metro').agg(F.avg('Zhvi').alias('Average_ZHVI')).orderBy(  
'Average_ZHVI', ascending=False).toPandas()
```

```
plt.figure(figsize=(14, 6))  
zhvi_by_metro.head(15).set_index('Metro').plot(kind='bar',  
figsize=(14,6)) # Top 15 metros for brevity  
plt.title('Average Median Home Value By Metro')
```

```
plt.xlabel('Metro')
plt.ylabel('Median Home Value (ZHVI)')
plt.grid(True, axis='y')
plt.tight_layout()
plt.show()
```

#### **#4. fall peak data**

```
fall_from_peak_data =
df.select('PctFallFromPeak').rdd.flatMap(lambda x: x).collect()
```

```
plt.figure(figsize=(14, 6))
plt.hist(fall_from_peak_data, bins=50)
plt.title('Distribution of Percentage Fall From Peak')
plt.xlabel('Percentage Fall')
plt.ylabel('Count')
plt.grid(True)
plt.show()
```

#### **#5. Scatter Plot of SizeRank vs Zhvi:**

```
size_rank_data = df.select('SizeRank', 'Zhvi').collect()
size_ranks = [row['SizeRank'] for row in size_rank_data]
zhvis = [row['Zhvi'] for row in size_rank_data]
```

```
plt.figure(figsize=(14, 6))
plt.scatter(size_ranks, zhvis, alpha=0.6)
plt.title('Scatter Plot of SizeRank vs Zhvi')
plt.xlabel('SizeRank')
plt.ylabel('Zhvi')
plt.grid(True)
plt.show()
```

#### **#6. Bar Chart of Top 10 Cities by Median Home Value:**

```
city_data =
df.groupBy('City').agg(F.avg('Zhvi').alias('Average_ZHVI')).orderBy('
Average_ZHVI', ascending=False).limit(10).collect()
```

```
cities = [row['City'] for row in city_data]
avg_city_values = [row['Average_ZHVI'] for row in city_data]
```

```
plt.figure(figsize=(14, 6))
plt.bar(cities, avg_city_values)
plt.title('Top 10 Cities by Median Home Value')
plt.xlabel('City')
plt.ylabel('Average Zhvi')
plt.grid(True, axis='y')
plt.xticks(rotation=45)
plt.show()
```

## **Dataset 2: (Annual House Price Indexes - Three-Digit ZIP Codes)**

```
pip install pyspark
```

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.functions import mean, when, col, avg, stddev, desc
import tkinter as tk
import matplotlib.pyplot as plt
import seaborn as sns
from pyspark.sql.functions import avg
```

```
spark = SparkSession.builder.appName("Visualization").getOrCreate()
```

```
df = spark.read.csv('Downloads/HPI_AT_BDL_ZIP3.csv',
header=True, inferSchema=True)
```

```
load_data()
```

```
# Convert HPI column to float, filtering out non-numeric values
df = df.withColumn("HPI", df["HPI for Three-Digit ZIP Codes (All-
Transactions Index)"].cast("float"))
```

```
hpi_values =  
df.filter(df["HPI"].isNotNull()).select("HPI").rdd.flatMap(lambda x:  
x).collect()
```

```
plt.figure(figsize=(14, 6))  
plt.hist(hpi_values, bins=50, color='skyblue', edgecolor='black')  
plt.title('Distribution of HPI Values for Three-Digit ZIP Codes')  
plt.xlabel('HPI')  
plt.ylabel('Frequency')  
plt.grid(True, axis='y')  
plt.show()
```

```
plt.figure(figsize=(14, 6))  
plt.boxplot(hpi_values, vert=False)  
plt.title('Box Plot of HPI values for Three-Digit ZIP Codes')  
plt.xlabel('HPI Value')  
plt.show()
```