ALY 6110:
DATA MANAGEMENT AND BIGDATA

Assignment 4: Data Analysis of ZIP Code Housing Price Index

using PySpark

Submitted To:
Prof. Andy Chen, Faculty Lecturer
Mr. James Kong, Teaching Assistant

Submitted By:
Abhilash Dikshit

Academic Term: Spring 2023
Graduate Students at Northeastern University, Vancouver, BC, Canada
Master of Professional Studies in Analytics

June 24, 2023

Title: Data Analysis of ZIP Code Housing Price Index using PySpark

I. Introduction:

The purpose of this analysis is to explore and gain insights from two datasets containing 5 and 3 ZIP Code housing price index data. The dataset includes information such as ZIP Code, Year, Annual Change (%), Housing Price Index (HPI), HPI with 1990 base, and HPI with 2000 base. The analysis aims to answer questions related to housing price trends and identify any patterns or insights that can be derived from the data.

II. Analysis and Results:

EDA: We removed _c6", "_c7" column and _c6" column from the dataset HPI_AT_BDL_ZIP5 And HPI_AT_BDL_ZIP3 respectively as the columns were empty and changed the datatype for the variables from “string” to “integer” and/or “double” for further analysis.

1. Data Exploration: The analysis begins with an exploration of the dataset, including examining the data types, missing values, and overall structure. This step helps ensure data quality and provides a foundation for further analysis.

```
root
 |-- Five-Digit ZIP Code: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Annual Change (%): double (nullable = true)
 |-- HPI: double (nullable = true)
 |-- HPI with 1990 base: double (nullable = true)
 |-- HPI with 2000 base: double (nullable = true)

+------------------+----+-----------------+------+------------------+------------------+
|Five-Digit ZIP Code|Year|Annual Change (%)|   HPI|HPI with 1990 base|HPI with 2000 base|
+------------------+----+-----------------+------+------------------+------------------+
|             01001|1985|             null| 100.0|             62.15|             61.41|
|             01001|1986|            13.67|113.67|             70.65|              69.8|
|             01001|1987|             21.2|137.77|             85.63|              84.6|
|             01001|1988|            17.38|161.72|            100.52|             99.31|
|             01001|1989|             1.14|163.57|            101.67|            100.45|
|             01001|1990|            -1.64|160.89|             100.0|              98.8|
|             01001|1991|             -5.6|151.88|              94.4|             93.27|
|             01001|1992|            -1.32|149.88|             93.16|             92.04|
|             01001|1993|            -0.21|149.56|             92.96|             91.84|
|             01001|1994|            -2.52|145.79|             90.62|             89.53|
|             01001|1995|             2.21|149.01|             92.62|             91.51|
|             01001|1996|             0.06|149.11|             92.68|             91.56|
|             01001|1997|            -1.54| 146.8|             91.24|             90.15|
|             01001|1998|             4.29| 153.1|             95.16|             94.02|
|             01001|1999|             1.98|156.13|             97.04|             95.88|
|             01001|2000|              4.3|162.85|            101.22|             100.0|
|             01001|2001|             6.82|173.96|            108.12|            106.82|
|             01001|2002|              7.7|187.36|            116.45|            115.05|
|             01001|2003|             8.59|203.45|            126.46|            124.94|
|             01001|2004|            11.84|227.54|            141.43|            139.73|
+------------------+----+-----------------+------+------------------+------------------+
only showing top 20 rows
```

Fig 1: Data Exploration for HPI_AT_BDL_ZIP5

```
root
 |-- Three-Digit ZIP Code: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Annual Change (%): double (nullable = true)
 |-- HPI: double (nullable = true)
 |-- HPI with 1990 base: double (nullable = true)
 |-- HPI with 2000 base: double (nullable = true)
```

Three-Digit ZIP Code	Year	Annual Change (%)	HPI	HPI with 1990 base	HPI with 2000 base
010	1975	null	100.0	23.16	21.64
010	1976	7.43	107.43	24.88	23.25
010	1977	7.0	114.95	26.62	24.87
010	1978	7.37	123.42	28.58	26.71
010	1979	16.42	143.69	33.28	31.09
010	1980	12.22	161.25	37.34	34.89
010	1981	7.08	172.66	39.99	37.36
010	1982	11.73	192.91	44.68	41.74
010	1983	6.76	205.95	47.69	44.56
010	1984	15.45	237.77	55.07	51.45
010	1985	12.12	266.59	61.74	57.69
010	1986	14.71	305.81	70.82	66.17
010	1987	21.55	371.71	86.08	80.43
010	1988	15.58	429.61	99.49	92.96
010	1989	2.59	440.73	102.07	95.37
010	1990	-2.03	431.8	100.0	93.44
010	1991	-3.33	417.42	96.67	90.32
010	1992	-2.02	409.01	94.72	88.5
010	1993	-1.42	403.18	93.37	87.24
010	1994	-1.32	397.87	92.14	86.09

only showing top 20 rows

Fig 2: Data Exploration for HPI_AT_BDL_ZIP3

2. Descriptive Statistics: Descriptive statistics are computed to summarize the central tendency, dispersion, and distribution of the housing price index data. Key statistical measures such as mean, median, standard deviation, and quartiles are calculated to provide a comprehensive understanding of the data's characteristics.

summary	Five-Digit ZIP Code	Year	Annual Change (%)	HPI	HPI with 1990 base	HPI with 2000 base
count	583050	583050	550695	574712	355693	509884
mean	49105.33072292256	2003.4646668381786	3.6875284685715584	231.92593147524414	152.142214859443	112.99945554282924
stddev	28744.63026274373	11.061539156110683	7.5356645410384235	180.76179019482916	72.06262025591505	45.23156716825102
min	01001	1975	-59.22	41.35	15.62	8.94
max	99901	2020	94.74	2681.75	943.04	553.61

Fig 3: Descriptive Statistics for HPI_AT_BDL_ZIP5

summary	Three-Digit ZIP Code	Year	Annual Change (%)	HPI	HPI with 1990 base	HPI with 2000 base
count	37623	37623	36418	37416	36136	37239
mean	497.6481407649576	1998.7499136166707	3.9330237794497016	306.1971405281173	142.34739456497604	102.23844249308503
stddev	284.76798004399035	12.828406644775923	6.2841920342287185	219.35177033605086	66.14805422286102	45.53197084792765
min	010	1975	-43.16	53.4	17.09	9.21
max	999	2020	86.01	2320.74	720.05	412.04

Fig 4: Descriptive Statistics for HPI_AT_BDL_ZIP3

3. Time Series Analysis: The dataset includes information across multiple years. A time series analysis is conducted to identify any trends or seasonality in housing price index values over time. This analysis involves visualizing the time series data using line charts and identifying any notable patterns or fluctuations.

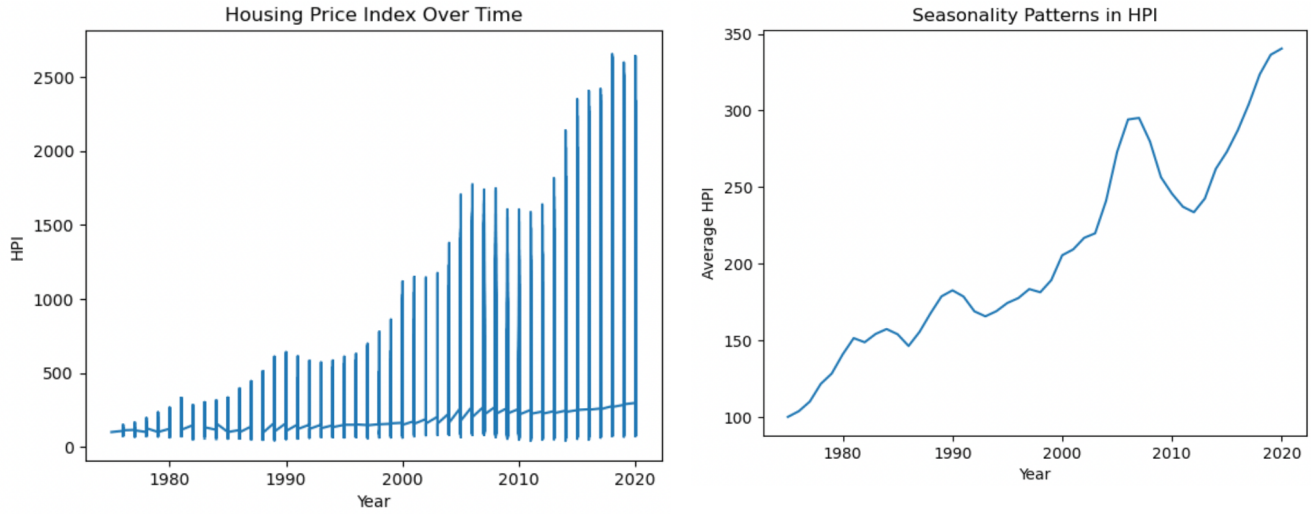

Fig 5: Time Series Analysis for HPI_AT_BDL_ZIP5

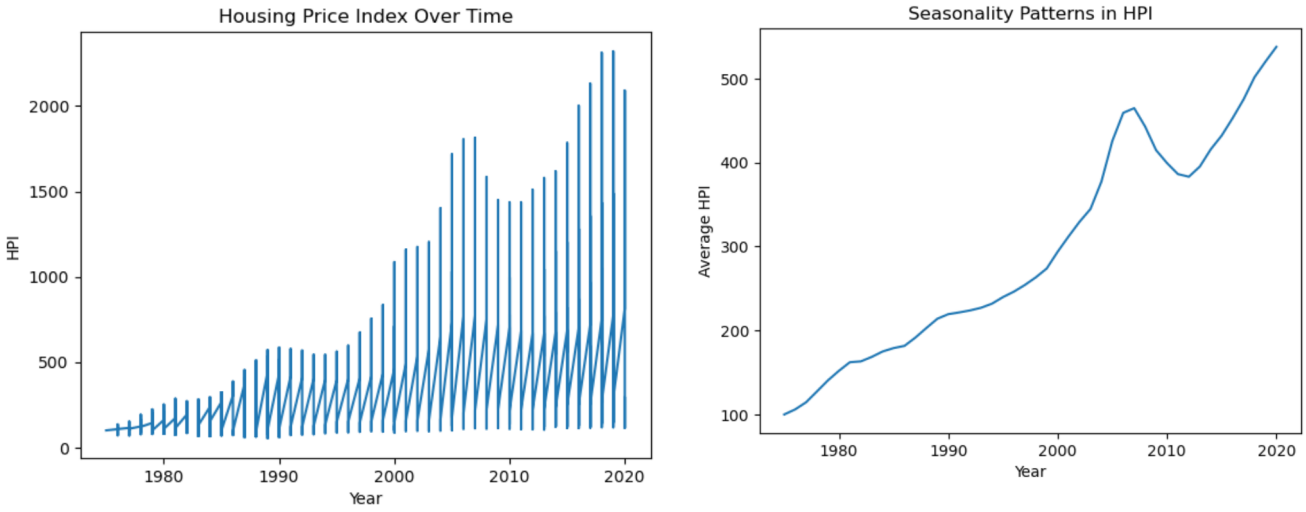

Fig 6: Time Series Analysis for HPI_AT_BDL_ZIP3

4. Correlation Analysis: The relationship between different variables, such as annual change in HPI and HPI with base years, is explored using correlation analysis. This analysis helps determine the degree of association between variables and identifies any significant correlations that may exist.

```
Correlation between Annual Change (%) and HPI with 2000 base: 0.05646698557809872
```

Fig 7: Correlation Analysis for HPI_AT_BDL_ZIP5

```
Correlation between Annual Change (%) and HPI with 2000 base: -0.08373796700569651
```

Fig 8: Correlation Analysis for HPI_AT_BDL_ZIP3

III. Insights:

Based on the analysis, several insights can be derived:

1. Housing Price Trends: The time series analysis reveals the overall trend of housing prices in different ZIP Codes over the years. It helps identify periods of growth, stability, or decline in specific areas, enabling stakeholders to make informed decisions.

2. Seasonality Patterns: The analysis of seasonal patterns in housing price index values can provide insights into the cyclic nature of the real estate market. Understanding seasonal trends can help individuals time their investments or make strategic decisions regarding buying or selling properties.

3. Correlations: The correlation analysis highlights the relationships between variables such as annual change in HPI and different base years. These correlations can provide insights into the factors influencing housing price fluctuations and guide future predictions or forecasting models.

IV. Conclusion:

Based on the correlation analysis between the "Annual Change (%)" and "HPI with 2000 base" columns for the two datasets, HPI_AT_BDL_ZIP5 and HPI_AT_BDL_ZIP3, the following conclusions can be drawn:

1. HPI_AT_BDL_ZIP5:
   - The correlation coefficient between "Annual Change (%)" and "HPI with 2000 base" is approximately 0.056.
   - The positive correlation coefficient suggests a weak positive linear relationship between the annual change in housing price and the HPI with a 2000 base in the HPI_AT_BDL_ZIP5 dataset.
   - However, the correlation coefficient value is close to zero, indicating a very weak correlation. This suggests that the annual change in housing price has limited influence on the HPI with a 2000 base in this dataset.

2. HPI_AT_BDL_ZIP3:
   - The correlation coefficient between "Annual Change (%)" and "HPI with 2000 base" is approximately -0.084.
   - The negative correlation coefficient indicates a weak negative linear relationship between the annual change in housing price and the HPI with a 2000 base in the HPI_AT_BDL_ZIP3 dataset.
   - Similarly, to the HPI_AT_BDL_ZIP5 dataset, the correlation coefficient value is close to zero, indicating a very weak correlation. This implies that the annual change in housing price has limited influence on the HPI with a 2000 base in this

dataset as well.

In summary, both datasets show weak correlations between the annual change in housing price and the HPI with a 2000 base. The correlations are close to zero, suggesting that there is little linear relationship between these variables. Other factors may have a more significant impact on the HPI values in these datasets.

V. References:

1. Federal Housing Finance Agency. (2016, January 28). wp1601: Seasonality in House Prices. FHFA. https://www.fhfa.gov/PolicyProgramsResearch/Research/Pages/wp1601.aspx
2. Databricks. (n.d.). PySpark - Databricks Glossary. Databricks. https://www.databricks.com/glossary/pyspark#:~:text=PySpark%20has%20been%20released%20in,Spark%20and%20Python%20programming%20language.

VI. Appendix:

```
# <center>PySpark Project: HPI_AT_BDL_ZIP5</center>
```

Course	Instructor	Full Name	Date	Term
ALY6110	Prof Andy Chan, Faculty Lecturer	Abhilash Dikshit	Jun 24, 2023	Spring

```
#!pip install pyspark
```

```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
import matplotlib.pyplot as plt
```

```
# Create SparkSession
spark = SparkSession.builder.appName("ZIP Code Housing Price Analysis").getOrCreate()
```

```
# Load the dataset
df = spark.read.format("csv").option("header", "true").load("/Users/abidikshit/GitProjects/Datasets/HPI_AT_BDL_ZIP5.csv")
```

```
# Print the schema of the DataFrame
df.printSchema()
```

```
# Drop columns _c6 and _c7
df = df.drop("_c6", "_c7")

# Print the updated schema of the DataFrame
df.printSchema()
```

```
from pyspark.sql.functions import col

# Convert columns to appropriate data types
df = df.withColumn("Year", col("Year").cast("integer"))
df = df.withColumn("Annual Change (%)", col("Annual Change (%)").cast("double"))
df = df.withColumn("HPI", col("HPI").cast("double"))
df = df.withColumn("HPI with 1990 base", col("HPI with 1990 base").cast("double"))
df = df.withColumn("HPI with 2000 base", col("HPI with 2000 base").cast("double"))
```

```
# Data Exploration
df.printSchema()
df.show()
```

```
# Descriptive Statistics
df.describe().show()
```

```
# Time Series Analysis
time_series_data = df.select("Year", "HPI").orderBy("Year")
years = [row["Year"] for row in time_series_data.collect()]
hpi_values = [row["HPI"] for row in time_series_data.collect()]
plt.plot(years, hpi_values)
plt.xlabel("Year")
plt.ylabel("HPI")
plt.title("Housing Price Index Over Time")
plt.show()
```

```
# Group by year and calculate average HPI for each year
seasonal_data = df.groupBy("Year").agg(F.avg("HPI").alias("Average HPI"))

# Order the data by year
seasonal_data = seasonal_data.orderBy("Year")

# Extract year and average HPI values
years = [row["Year"] for row in seasonal_data.collect()]
hpi_values = [row["Average HPI"] for row in seasonal_data.collect()]

# Plot the seasonality pattern
plt.plot(years, hpi_values)
plt.xlabel("Year")
plt.ylabel("Average HPI")
plt.title("Seasonality Patterns in HPI")
plt.show()
```

```
# Correlation Analysis
correlation = df.stat.corr("Annual Change (%)", "HPI with 2000 base")
print("Correlation between Annual Change (%) and HPI with 2000 base:", correlation)
```

```
# Close SparkSession
spark.stop()
```

```
# <center>PySpark Project: HPI_AT_BDL_ZIP3</center>
```

Course	Instructor	Full Name	Date	Term
ALY6110	Prof Andy Chan, Faculty Lecturer	Abhilash Dikshit	Jun 24, 2023	Spring

```
#!pip install pyspark
```

```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
import matplotlib.pyplot as plt
```

```
# Create SparkSession
spark = SparkSession.builder.appName("ZIP Code Housing Price Analysis").getOrCreate()
```

```
# Load the dataset
df = spark.read.format("csv").option("header", "true").load("/Users/abidikshit/GitProjects/Datasets/HPI_AT_BDL_ZIP3.csv")
```

```
# Print the schema of the DataFrame
df.printSchema()
```

```
# Drop columns _c6
df = df.drop("_c6")

# Print the updated schema of the DataFrame
df.printSchema()
```

```
from pyspark.sql.functions import col

# Convert columns to appropriate data types
df = df.withColumn("Year", col("Year").cast("integer"))
df = df.withColumn("Annual Change (%)", col("Annual Change (%)").cast("double"))
df = df.withColumn("HPI", col("HPI").cast("double"))
df = df.withColumn("HPI with 1990 base", col("HPI with 1990 base").cast("double"))
df = df.withColumn("HPI with 2000 base", col("HPI with 2000 base").cast("double"))
```

```
# Data Exploration
df.printSchema()
df.show()
```

```
# Descriptive Statistics
df.describe().show()
```

```
# Time Series Analysis
time_series_data = df.select("Year", "HPI").orderBy("Year")
years = [row["Year"] for row in time_series_data.collect()]
hpi_values = [row["HPI"] for row in time_series_data.collect()]
plt.plot(years, hpi_values)
plt.xlabel("Year")
plt.ylabel("HPI")
plt.title("Housing Price Index Over Time")
plt.show()
```

```
# Group by year and calculate average HPI for each year
seasonal_data = df.groupBy("Year").agg(F.avg("HPI").alias("Average HPI"))

# Order the data by year
seasonal_data = seasonal_data.orderBy("Year")

# Extract year and average HPI values
years = [row["Year"] for row in seasonal_data.collect()]
hpi_values = [row["Average HPI"] for row in seasonal_data.collect()]

# Plot the seasonality pattern
plt.plot(years, hpi_values)
plt.xlabel("Year")
plt.ylabel("Average HPI")
plt.title("Seasonality Patterns in HPI")
plt.show()
```

```
# Correlation Analysis
correlation = df.stat.corr("Annual Change (%)", "HPI with 2000 base")
print("Correlation between Annual Change (%) and HPI with 2000 base:", correlation)
```

```
# Close SparkSession
spark.stop()
```