CIS 5200 Term Project Lab Tutorial

Authors: Christian Ahamba, Solange Ruiz, Damian Wilson

Instructor: Jongwook Woo, PhD

Date: 5/18/2025

Lab Tutorial

Christian Ahamba (cahamba@calstatela.edu)

Solange Ruiz (sruiz85@calstatela.edu)

Damian Wilson (dwilso33@calstatela.edu)

5/18/2025

Mapping U.S. Housing Trends with Big Data

Objective

The objective of this project is to:

- **Analyze** spatial and temporal trends in the U.S. housing market
- **Utilize** over **8 GB of Redfin housing market data**
- **Leverage big data tools** such as:
  - Hive (for querying)
  - Hadoop (for data aggregation and processing)
  - Tableau (for visualization)
- **Uncover**:
  - Patterns of price volatility
  - Long-term market shifts
  - Regional disparities in housing trends
- **Support decision-making** for:
  - Investors
  - Homebuyers
  - Policymakers

Platform Specification

- Component: 5 nodes (2 masters, 3 workers)
- Hadoop Cluster: 5 nodes (2 masters, 3 workers)

- Memory: 32 GB RAM per node
- CPU: 2.3 GHz per node
- Query Interface: Hive via Beeline
- SSH Access: 144.24.13.0

Step 1: Unzip the dataset

Download zip file from Kaggle.com
https://www.kaggle.com/datasets/thuynyle/redfin-housing-market-data

After zip file is downloaded to your local directory/local machine in your download folder, or any folder you will remember where to find the zip file there are five different datasets in TSV. Format. For smoother usage or preference, we converted the file from. TSV format to CSV.

1. Open Git Bash Terminal to unzip file:

```
$ unzip Redfin.zip -d redfin_data
```

This will extract the .tsv000 files into the redfin_data folder.

2. Convert .tsv000 to .csv format by changing to the (redfin_data) directory

```
$ cd redfin_data
```

Run the following loop to convert each .tsv000 file into a .csv file by replacing tabs with commas:

```
$ for file in *.tsv000; do
  cat "$file" | tr '\t' ',' > "${file% .tsv000}.csv"
```

3. Verify .csv files by listing the converted .csv files:

```
$ ls *.csv
```

You should see files listed like this:

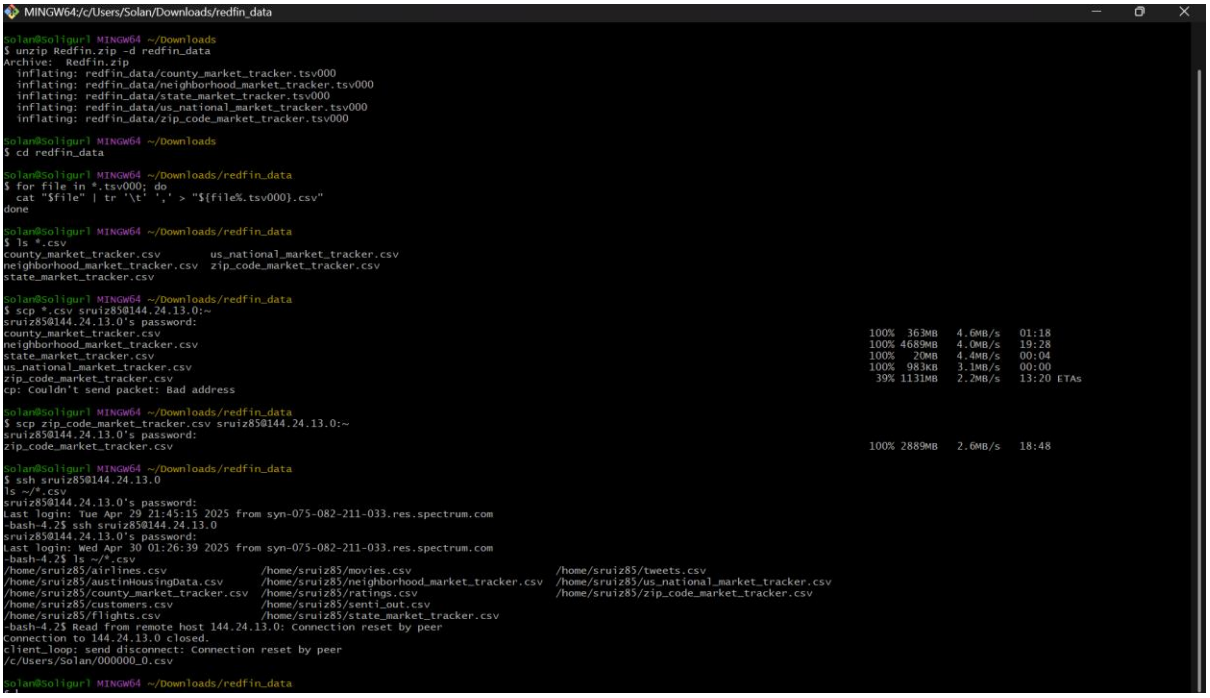
```
county_market_tracker.csv
us_national_market_tracker.csv
neighborhood_market_tracker.csv
zip_code_market_tracker.csv
state_market_tracker.csv
```

4. Securely copy data to remote Hadoop node
Use scp to transfer files to the Hadoop cluster (IP: 144.24.13.0). Example:

```
scp county_market_tracker.csv sruiz85@144.24.13.0:~
```

Note: Repeat the command for each file as needed.
If the IP or DNS resolution fails, make sure the address is correct, and the server is accessible.

Example:



Step 2: Train NLP

Access Hive from HDFS via SSH and Beeline

1. SSH into your Hadoop environment (open your terminal and connect to the Hadoop cluster)

```
$ sruiz85@144.24.13.0
```

2. Once logged in, start Hive with Beeline

```
$ Beeline
```

You should see the beeline > prompt, ready for SQL commands.
3. Check available data (Optional)

```
hdfs dfs -ls /user/Group5-Folder/
```

4. Filter raw data to U.S state-level

```
CREATE TABLE my_state_data AS
SELECT *
FROM Group5-Folder
WHERE region_type = 'state';
```

5. Create Hive External Table from HDFS Path

```
CREATE EXTERNAL TABLE sruiz85_redfin_state (

)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/sruiz85/my_state_data'
TBLPROPERTIES ('skip.header.line.count' = '1');
```

6. Run Hive SQL Queries for Tableau (First line Chart)

```
SELECT state, YEAR(TO_DATE(period_begin)) AS year,
       AVG(median_list_price) AS avg_list_price
FROM sruiz85_redfin_state
WHERE median_list_price IS NOT NULL
GROUP BY state, YEAR(TO_DATE(period_begin));
```

7. Map: Avg Median List Price in 2021

```
SELECT state, AVG(median_list_price) AS avg_list_price
FROM sruiz85_redfin_state
WHERE YEAR(TO_DATE(period_begin)) = 2021
  AND median_list_price IS NOT NULL
GROUP BY state;
```

8. Heatmap: Price Volatility (Std Dev) by State-Year

```
SELECT state, YEAR(TO_DATE(period_begin)) AS year,
       STDDEV(median_list_price) AS volatility
FROM sruiz85_redfin_state
WHERE median_list_price IS NOT NULL
GROUP BY state, YEAR(TO_DATE(period_begin))
ORDER BY state, year;
```

9. Bar Chart: Top 10 States by Price Growth (2012-2021)

```
SELECT
  state,
  MAX(CASE WHEN YEAR(TO_DATE(period_begin)) = 2012 THEN median_list_price
END) AS start_price,
  MAX(CASE WHEN YEAR(TO_DATE(period_begin)) = 2021 THEN median_list_price
END) AS end_price,
  ROUND(100 * (
    MAX(CASE WHEN YEAR(TO_DATE(period_begin)) = 2021 THEN median_list_price
END) -
    MAX(CASE WHEN YEAR(TO_DATE(period_begin)) = 2012 THEN median_list_price
END)
  ) /
    MAX(CASE WHEN YEAR(TO_DATE(period_begin)) = 2012 THEN median_list_price
END), 2) AS percent_growth
FROM sruiz85_redfin_state
WHERE median_list_price IS NOT NULL
```

```
GROUP BY state
ORDER BY percent_growth DESC
LIMIT 10;
```

10. Export Hive Query Result for Tableau, save output to server local path.

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/sruiz85/output/chart1'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','

SELECT ...
```

Replace **SELECT** ... with any of the chart queries above.

11. From local machine, run:

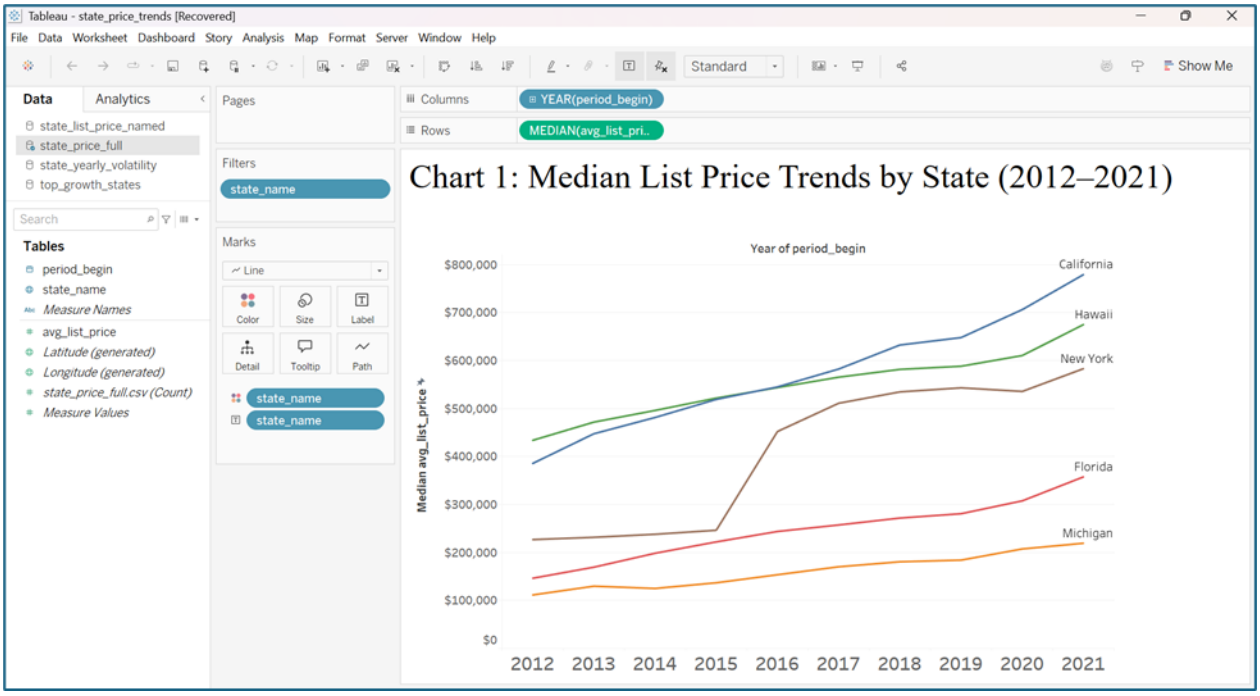
```
scp sruiz85@144.24.13.00:/home/sruiz85/output/chart1/* ~/Downloads/chart1.csv
```

Step 3: Visualization

Import saved CSV files into Tableau **(Text File Data Source)**

1. Launch Tableau Desktop or **Tableau public** -> click **"Text File"** under connect -> navigate your download folder and select the **CSV file** (ex. Chart 1), Tableau will load the file into the **Data Source** tab.

2. **Chart 1: Line Chart – Median List Price Trends**



From: state_price_full.csv

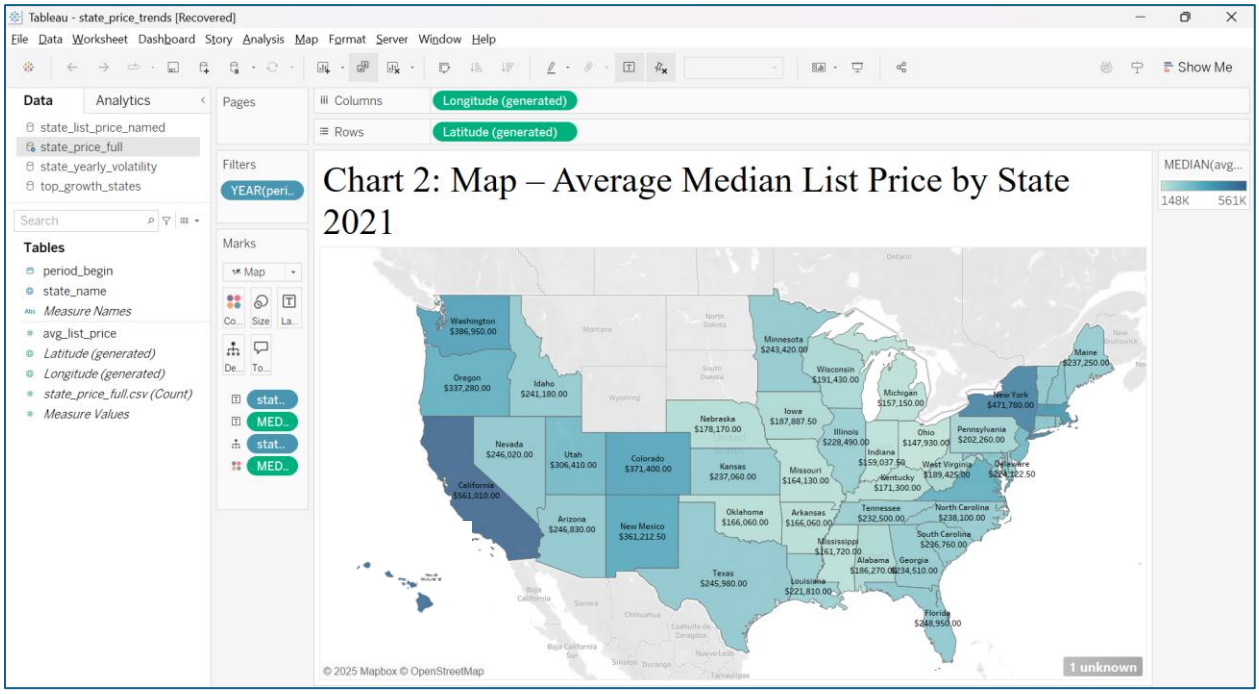
Fields Needed:

- state → Dimensions
- year → Dimensions
- avg_list_price → Measure

Tableau Setup:

1. Drag year to **Columns**
2. Drag avg_list_price to **Rows**
3. Drag state to **Color** (or **Filters** if you want to limit to a few states)
4. Choose **Line** as the mark type

3. **Chart 2: Map – Avg List Price by State (2021)**



From: state_price_full.csv

Fields Needed:

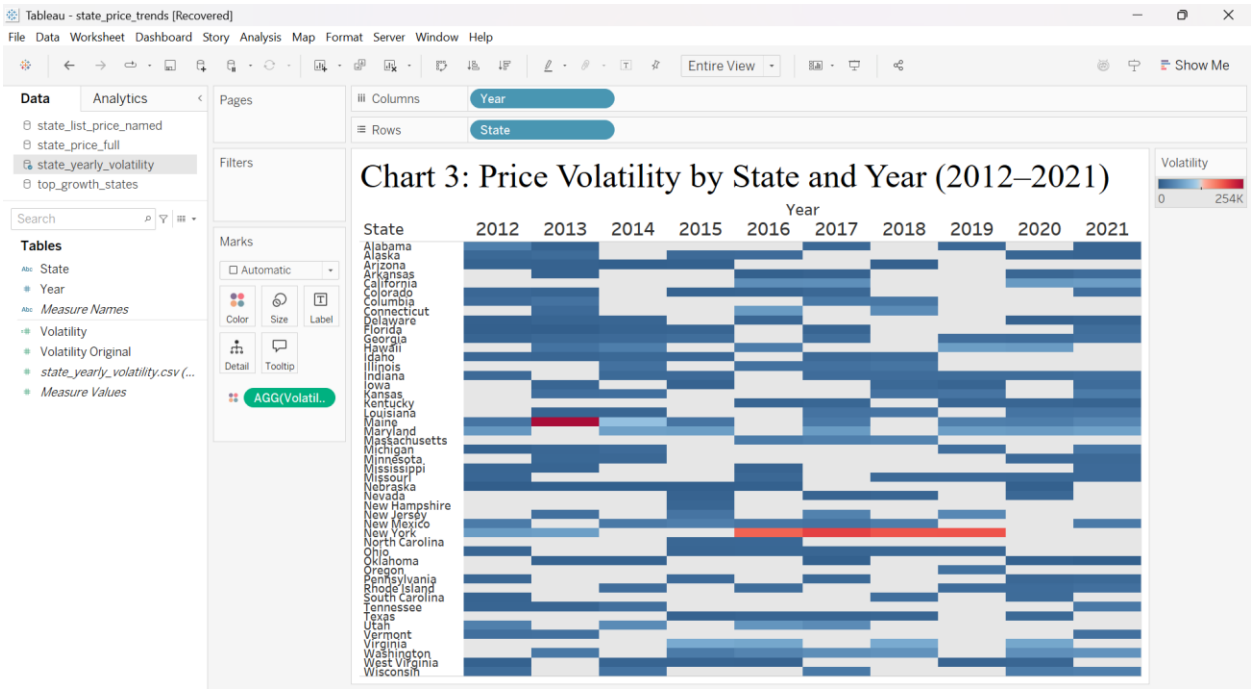
- state → Geographic Dimension
- avg_list_price → Measure

Tableau Setup:

1. Make sure state is set to **Geographic Role > State/Province**

2. Double-click state to generate a map
3. Drag avg_list_price to **Color** and **Label**

4. Chart 3: Heatmap – Price Volatility by State and Year



From: state_yearly_volatility.csv

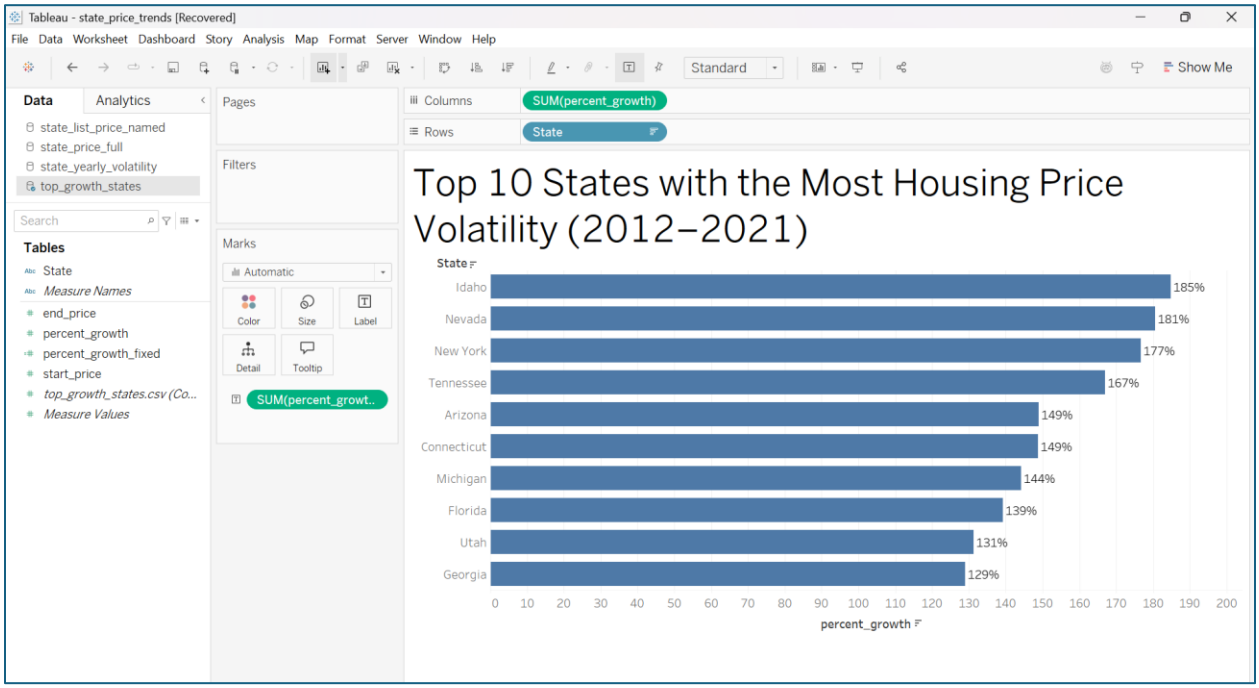
Fields Needed:

- state, year → Dimensions
- volatility → Measure

Tableau Setup:

1. Drag year to **Columns**
2. Drag state to **Rows**
3. Drag volatility to **Color**
4. Set mark type to **Square** or **Heatmap**
5. (Optional) Add volatility to **Label**

5. **Chart 4: Bar Chart – Top 10 States by Price Growth**



From: top_growth_states.csv

Fields Needed:

- state → Dimension
- percent_growth → Measure

Tableau Setup:

1. Drag state to **Rows**
2. Drag percent_growth to **Columns**
3. Use **Bar** mark type
4. Sort descending
5. Drag percent_growth to **Label**

References

Data Source:
Redfin Housing Market Data – Kaggle

GitHub Repository (Project Code):
https://github.com/Soligrl/CIS-5200

Tableau Mapping Guide:
Tableau Help: Create Simple Maps

FIPS Code and Alternative Map Projections:
Flerlage Twins: Alternative Map Projections in Tableau

Hive Subqueries Tutorial:
BigDataNSQL: Sub-Queries in Apache Hive

Professor's Lab Tutorial and Course Lectures:
CIS 5200 – Big Data Analytics, Spring 2025.
Lab materials and lecture notes provided by Prof. **Jongwook,Woo, PhD**, California State University, Los Angeles.