

Connecticut Single-Family Home Market Analysis: Data Workflow Project

Author: Shotaro Oyama
Course: AI Mastery Capstone - Module 1
Date: February 2026

1. Overview

This project builds a comprehensive data workflow analyzing over 400,000 single-family home sales transactions in Connecticut from 2001 to 2023. Using the Real Estate Sales 2001-2023 GL dataset from the Connecticut Office of Policy and Management (available at <https://catalog.data.gov/dataset/real-estate-sales-2001-2018>), the analysis examines long-term price trends, regional variations, assessment accuracy, and seasonal patterns. This workflow establishes a clean, reproducible foundation for future machine learning applications in real estate price prediction.

2. Dataset Description

The Real Estate Sales 2001-2023 GL dataset contains 1,141,722 real estate transactions recorded by Connecticut municipalities between October 2001 and September 2023. The dataset includes 14 variables: Serial Number, List Year, Date Recorded, Town, Address, Assessed Value, Sale Amount, Sales Ratio, Property Type, Residential Type, Non Use Code, Assessor Remarks, OPM remarks, and Location coordinates.

This analysis focuses on single-family homes, which represent 401,612 transactions (35.2% of the dataset). Key variables analyzed include Sale Amount (transaction price), Assessed Value (municipal property assessment), Date Recorded (transaction date), and Town (municipality). The temporal span of 23 years enables long-term trend analysis, while geographic coverage across 170 Connecticut towns allows regional comparison. Missing values are present in Property Type (33.5%), Residential Type (35.3%), and remarks fields (>70%), requiring careful data quality management.

3. Workflow Description

3.1 Data Ingestion

The workflow begins by loading the 130MB CSV file using Pandas (McKinney, 2010) and performing initial quality assessments. Basic statistics reveal data types, missing value patterns, and variable distributions. This inspection phase identifies data quality issues requiring attention in subsequent cleaning steps.

3.2 Data Cleaning

Data cleaning employs multiple strategies to ensure analysis validity:

1. **Property Type Filtering:** Subset to single-family homes to create a homogeneous comparison group
2. **Outlier Removal:** Implement town-year specific IQR (Interquartile Range) method with hierarchical fallback, accounting for regional and temporal price variations (Osborne & Overbay, 2004)
3. **Date Conversion:** Transform text dates to datetime objects and extract temporal features (Year, Month, Quarter)
4. **Missing Value Treatment:** Remove rows with missing critical variables (Sale Amount, Date Recorded, Assessed Value)

The town-year IQR approach is superior to global percentile thresholds because it adapts to local market conditions. For example, a \$2M sale is typical in Greenwich but represents an outlier in smaller towns. This context-aware method removed 3.93% of transactions (15,796 records) while preserving legitimate high-value sales up to \$7.05M.

3.3 Exploratory Data Analysis

EDA examines five key research questions through statistical summaries and visualizations:

1. **Temporal Trends:** Time series analysis of median prices from 2001-2023
2. **Price Distribution:** Histogram revealing right-skewed distribution typical of real estate markets
3. **Regional Variation:** Box plots comparing prices across top 10 towns by volume
4. **Assessment Accuracy:** Scatter plot analyzing assessed value vs. actual sale price
5. **Seasonality:** Heatmap showing transaction volume by month and year

Each analysis includes both quantitative metrics (medians, correlations, growth rates) and qualitative interpretation of market dynamics.

3.4 Data Visualization

Five primary visualizations communicate findings (Hunter, 2007; Waskom, 2021):

- **Figure 1 (Time Series):** Line plot with economic event markers (2008 Financial Crisis, COVID-19 Pandemic)
- **Figure 2 (Distribution):** Histogram of sale prices showing market concentration
- **Figure 3 (Regional):** Box plots comparing price ranges across towns
- **Figure 4 (Assessment):** Scatter plot with regression line and perfect assessment diagonal
- **Figure 5 (Seasonality):** Heatmap revealing summer peaks and winter troughs

All visualizations follow best practices with clear titles, labeled axes, and appropriate color schemes for accessibility.

3.5 Summary and Interpretation

The final section synthesizes findings, acknowledges limitations (missing property characteristics, nominal dollars, scope restrictions), and discusses implications for future AI/ML work including feature engineering strategies, model selection considerations, and ethical considerations for algorithmic fairness.

4. Key Decisions and Assumptions

4.1 Cleaning Choices

Decision 1: Town-Year IQR Method for Outliers

Rather than using a simple global percentile threshold (e.g., removing top 1%), we implemented a sophisticated hierarchical approach. This decision recognizes that Connecticut's housing market is heterogeneous, with luxury coastal towns having legitimately higher prices than inland areas. The IQR method calculates $Q3 + 1.5 \times IQR$ separately for each Town×Year group (with fallbacks for small samples), ensuring outlier detection adapts to local context (Tukey, 1977).

Justification: This approach is consistent with robust statistical practices that account for data structure (Osborne & Overbay, 2004). A global threshold would either over-filter expensive areas or under-filter cheap areas, introducing systematic bias.

Decision 2: Minimum Price Threshold (\$2,000)

We retained the dataset documentation's \$2,000 minimum to filter non-market transactions (family transfers, errors). While this is somewhat arbitrary, it aligns with domain knowledge that legitimate home sales rarely occur below this threshold.

Trade-off: This removes approximately 500 transactions that could be legitimate distressed sales or mobile homes, but preserves data quality for the majority of market-rate transactions.

Decision 3: Single-Family Focus

Filtering to single-family homes creates internal validity (consistent property type) but reduces external validity (findings may not generalize to condos, multi-family properties).

Justification: Single-family homes represent the largest market segment (401,612 transactions) and have more consistent characteristics for meaningful trend analysis.

4.2 EDA Focus Areas

The five research questions were selected to provide complementary perspectives on market dynamics:

- **Temporal analysis** reveals long-term growth and crisis impacts
- **Distribution analysis** shows market concentration and typical price ranges
- **Regional analysis** identifies geographic heterogeneity
- **Assessment analysis** evaluates municipal valuation accuracy
- **Seasonal analysis** uncovers cyclical patterns

This multi-faceted approach is consistent with recommended exploratory analysis frameworks (Tukey, 1977; Peng & Matsui, 2015).

4.3 Visualization Design Rationale

Time Series Plot: Line plot chosen over bar chart to emphasize continuous temporal progression. Economic event markers (vertical lines) provide historical context for interpreting price changes.

Price Distribution: Histogram with 50 bins balances detail with readability. Right-skewed distribution is expected for housing data (Case & Shiller, 1989).

Regional Comparison: Box plots effectively communicate median, quartiles, and outliers simultaneously across multiple groups. Ordering by transaction volume prioritizes data-rich comparisons.

Assessment Scatter Plot: Includes perfect assessment diagonal ($y=x$) as reference line and regression line to quantify relationship strength. Log scale considered but rejected to maintain interpretability.

Seasonality Heatmap: Color intensity encodes transaction volume, making temporal patterns immediately visible. Year-over-year comparison reveals both seasonal consistency and crisis impacts.

These design choices follow visualization best practices emphasizing clarity, appropriate chart types for data structure, and meaningful context (Few, 2012; Tufte, 2001).

5. Results and Interpretation

5.1 Long-Term Price Growth with Crisis Volatility

Finding: Connecticut single-family home prices increased approximately 40% from 2001 (\$220K median) to 2023 (\$310K median), representing ~1.5% annual growth when accounting for inflation. However, this growth occurred in distinct phases rather than linear progression.

Figure 1 shows three clear periods:

- 2001-2007: Steady growth to \$270K peak
- 2008-2012: Decline to \$235K trough (13% drop)
- 2013-2023: Recovery and acceleration to \$310K

Interpretation: The modest long-term growth rate reflects Connecticut's mature housing market. When adjusted for inflation (approximately 2% annually over this

period), real price growth was minimal, suggesting housing tracked general economic conditions rather than experiencing speculative bubbles outside the 2008 crisis period.

5.2 Divergent Economic Crisis Impacts

Finding: The 2008 Financial Crisis and COVID-19 Pandemic had opposite effects on housing prices despite both representing major economic disruptions.

- **2008 Crisis:** Prices declined 13% over 5 years, with recovery taking until 2017
- **COVID-19:** Prices increased 19% from 2019-2023, the steepest growth period in the dataset

Interpretation: This divergence reflects fundamentally different economic conditions. The 2008 crisis was housing-driven with credit constraints and foreclosures directly suppressing demand (Mian & Sufi, 2014). COVID-19 combined low interest rates, remote work flexibility increasing housing demand, and constrained supply due to construction shutdowns, creating upward price pressure (Mondragon & Wieland, 2022). This finding demonstrates that economic disruptions affect housing markets through diverse mechanisms.

5.3 Significant Regional Market Segmentation

Finding: *Figure 3* reveals substantial price variation across Connecticut towns, with median prices differing by 2-3x between locations even after controlling for property type (single-family homes only).

Interpretation: This heterogeneity suggests Connecticut comprises distinct sub-markets rather than a homogeneous housing market. Variation likely reflects differences in school quality (Nguyen-Hoang & Yinger, 2011), employment accessibility, coastal proximity, and local amenities. For predictive modeling, this finding indicates geographic features will be critical predictors, potentially requiring location-specific models or sophisticated feature engineering.

5.4 Systematic Assessment Gap

Finding: *Figure 4* shows properties consistently sell above assessed values, with median ratio of 1.41x. The scatter plot reveals most points lie above the perfect assessment line ($y=x$), indicating systematic under-assessment.

Interpretation: Municipal assessments lag market reality, likely due to infrequent revaluation cycles (Connecticut municipalities typically reassess every 3-5 years). This creates predictable inefficiency: homeowners pay property taxes based on outdated values, benefiting from lower tax burdens but potentially affecting municipal revenue stability. The relationship ($R^2=0.74$) suggests assessed value is a strong but imperfect proxy for market value.

5.5 Consistent Seasonal Patterns

Finding: *Figure 5* demonstrates strong seasonality with June-August consistently showing highest transaction volumes across all years, while January-February show lowest activity.

Interpretation: This pattern reflects practical constraints (weather, school calendars) and behavioral factors in housing decisions. The consistency across 23 years, even during crisis periods, suggests these seasonal forces are deeply embedded in market dynamics. For time series forecasting, models must incorporate seasonal adjustments to avoid mistaking cyclical patterns for trends.

6. Responsible Practice: Bias and Data Quality

6.1 Potential Bias Sources

Selection Bias from Filtering: Our filtering decisions (single-family only, outlier removal) could introduce bias. Removing 3.93% of transactions as outliers may disproportionately affect rapidly appreciating neighborhoods or unique luxury properties. If these areas differ systematically in demographics or other characteristics, our analysis may underrepresent certain market segments.

Geographic Representation Bias: Towns with fewer transactions receive less precise treatment in our hierarchical IQR method. Small towns ($n<20$ per year) use broader thresholds, potentially missing local market nuances. This creates an analytical advantage for data-rich municipalities, potentially distorting comparative conclusions.

Temporal Sampling Bias: If municipalities have inconsistent reporting during economic crises, our crisis impact analysis could be skewed. For example, if foreclosed properties are underreported in certain years, we may underestimate the 2008 crisis severity.

Assessment-Based Bias: Our finding that properties sell for 1.5-2x assessed value could perpetuate systemic inequities if assessment practices differ across wealthy vs. lower-income areas. If reassessments are more frequent in affluent areas, our analysis could amplify existing assessment disparities.

6.2 Mitigation Strategies

To reduce bias risk, we implemented several safeguards:

1. **Documented Decision Thresholds:** All filtering criteria (\$2,000 minimum, IQR multiplier) are explicitly documented with rationale
2. **Sensitivity Analysis Preparation:** Code structure enables easy threshold adjustment to test result robustness
3. **Heterogeneity Acknowledgment:** Regional and temporal analyses explicitly examine whether patterns differ across segments
4. **Limitation Transparency:** Results section acknowledges constraints and interprets findings cautiously

For future ML applications, additional steps would include:

- **Stratified Validation:** Ensure model performance is equitable across price ranges, towns, and time periods
- **Fairness Auditing:** Test for disparate impact if demographic data becomes available
- **Continuous Monitoring:** Track whether predictions differ systematically across groups
- **External Validation:** Compare findings to alternative data sources (Zillow, Realtor.com)

6.3 Data Quality Trade-offs

Missing Data: Rather than imputing missing values (which could introduce artifacts), we removed rows with missing critical variables. This reduces sample size but preserves data integrity. Future work could explore multiple imputation methods if retaining more data becomes critical (Rubin, 1987).

Nominal vs. Real Dollars: Our analysis uses nominal dollars rather than inflation-adjusted values. While this simplifies interpretation, it potentially overstates real price growth. Future iterations should incorporate Consumer Price Index adjustments for more accurate economic analysis.

Non-Market Transactions: Despite our \$2,000 threshold and ratio filters, some family transfers and distressed sales likely remain. These represent real transactions but may not reflect typical market conditions. More sophisticated classification (if additional transaction type data were available) could improve market representation.

7. Reproducibility

7.1 Technical Reproducibility

This project is designed for complete reproducibility through several mechanisms:

Requirements Management: The `requirements.txt` file specifies exact package versions used:

```
pandas==2.1.3
numpy==1.26.2
matplotlib==3.10.8
seaborn>=0.12.0
scipy>=1.11.4
```

Anyone can recreate the exact environment using `pip install -r requirements.txt`. Pinning major package versions (pandas, numpy, matplotlib) while allowing minor updates for others (seaborn, scipy) balances reproducibility with security updates.

Code Organization: All analysis code resides in a single Jupyter Notebook (`data_workflow.ipynb`) that can be executed top-to-bottom without modification. Each section is clearly marked with markdown headers, and code cells include comments explaining logic.

Functional Abstractions: Key operations (filtering, outlier detection, date conversion) are encapsulated in documented functions with clear inputs/outputs. This promotes:

- Code reuse
- Testing capability
- Clear documentation via docstrings
- Modification transparency (users can examine function logic)

Data Accessibility: While the 130MB dataset cannot be included in the Git repository, the README provides explicit download instructions from the public Data.gov source. Future users can obtain identical data.

7.2 Version Control Practices

The project employs professional Git practices demonstrating workflow transparency:

Branching Strategy:

- `main` branch: Final, stable version
- `data-cleaning` branch: Development of cleaning functions and outlier handling
- `eda` branch: Exploratory analysis and visualization development

This structure shows incremental development and allows reviewers to understand evolution of analytical decisions.

Commit History: Multiple commits (10+) document specific changes:

- "Add data cleaning functions with docstrings"
- "Implement town-year IQR outlier detection"
- "Complete EDA: assessment analysis and filtering"

Commit messages explain *what* changed and often *why*, creating an audit trail of analytical decisions.

GitHub Repository: Public repository (<https://github.com/yourusername/ct-housing-analysis>) provides:

- Complete source code
- Version history
- Branch visualization showing workflow
- Issue tracking for potential improvements

7.3 Documentation Standards

README.md: Concise setup instructions enable users to run the analysis within 10 minutes, assuming Python installation.

This Report (module_summary.pdf): Detailed methodology documentation explains rationale for each decision, providing context beyond code comments.

Notebook Markdown Cells: Each section begins with explanatory text describing:

- What analysis is being performed
- Why this approach was chosen
- How to interpret results

Function Docstrings: All custom functions include NumPy-style docstrings with:

- Parameter descriptions and types
- Return value specifications
- Usage examples
- Methodological notes

7.4 Reproducibility Verification

To verify reproducibility, an independent user would:

1. Clone repository
2. Download dataset from Data.gov link
3. Install dependencies from requirements.txt
4. Execute notebook top-to-bottom

Expected execution time: 3-5 minutes on standard laptop. All outputs should match those committed to repository (figures, statistics, printed summaries).

Potential Reproducibility Challenges:

- **Data updates:** If Connecticut updates the dataset on Data.gov, results may change. We specify the version accessed in February 2026.
- **Random sampling:** Some visualizations use `.sample()` for performance. Setting `random_state=42` ensures consistent sampling across runs.
- **System dependencies:** Matplotlib rendering may vary slightly across operating systems, though numerical results remain identical.

8. Sources and Citations

Data Source

State of Connecticut Office of Policy and Management. (2023). *Real Estate Sales 2001-2023 GL* [Data set]. Data.gov. <https://catalog.data.gov/dataset/real-estate-sales-2001-2018>

Methodological References

- Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. *American Economic Review*, 79(1), 125-137.
- Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Analytics Press.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. <https://doi.org/10.1109/MCSE.2007.55>
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61. <https://doi.org/10.25080/Majora-92bf1922-00a>
- Mian, A., & Sufi, A. (2014). What explains the 2007–2009 drop in employment? *Econometrica*, 82(6), 2197-2223.
- Mondragon, J., & Wieland, J. (2022). Housing demand and remote work. *NBER Working Paper* 30041. National Bureau of Economic Research.
- Nguyen-Hoang, P., & Yinger, J. (2011). The capitalization of school quality into house values: A review. *Journal of Housing Economics*, 20(1), 30-48.
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(6). <https://doi.org/10.7275/qf69-7k43>
- Peng, R. D., & Matsui, E. (2015). *The art of data science*. Leanpub.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>

Documentation References

- Pandas Development Team. (2023). *pandas documentation*. <https://pandas.pydata.org/docs/>
- Python Software Foundation. (2023). *Python 3 documentation*. <https://docs.python.org/3/>
- Git Documentation. (2023). *Git user manual*. <https://git-scm.com/docs>
-

End of Report