

Data Report: The Impact of Hurricanes on U.S. Housing Markets

This report outlines the process of building an automated data pipeline to analyze the impact of hurricanes on the housing market in U.S. cities. The focus is on understanding how hurricanes influence housing market dynamics, including median sale prices and housing inventory.

Main Question

How do hurricanes impact housing market dynamics in affected U.S. cities, particularly in terms of sale prices and housing inventory?

This question explores the effects of natural disasters, specifically hurricanes, on the housing market, focusing on U.S. cities. By analyzing the changes in housing dynamics (such as sale prices and inventory levels) after a hurricane event, this study seeks to uncover trends and insights that could help policymakers and real estate professionals make informed decisions.

Data Sources and License Details

The datasets used in this project are sourced from Kaggle and are licensed under **CC0: Public Domain**, which allows for free use, modification, and distribution of the data without any restrictions.

U.S. Hurricanes and Landfalls (1851-2023)

- **Source:** [U.S. Hurricanes Dataset on Kaggle](#)
- **License:** CC0: Public Domain
- **Description:** This dataset contains records of hurricanes that have occurred since 1851, with detailed information about each storm, including its name, the date of landfall, the states affected, and the severity of the storm. It is crucial to understand the hurricane's impact on specific regions over time. The data will be used to track the hurricanes affecting U.S. cities.

U.S. Cities Housing Market Data

- **Source:** [U.S. Housing Market Data on Kaggle](#)
- **License:** CC0: Public Domain
- **Description:** This dataset contains monthly statistics of housing markets across U.S. cities, including median sale prices, housing inventory, and the average number of days on the market. These metrics will be essential in analyzing the housing market dynamics pre- and post-hurricane events. By merging this data with hurricane data, we can study how housing prices and inventory levels change when hurricanes affect cities.

Data Pipeline

The data pipeline for this project was developed to automate the extraction, transformation, and loading (ETL) of data from both the U.S. hurricanes and housing market datasets. The pipeline was implemented using Python, utilizing libraries such as pandas for data manipulation, sqlite3 for database management, and the Kaggle API for downloading the datasets.

Key Steps:

1. **Data Extraction:** The datasets are downloaded using the Kaggle API if they are not already present.
2. **Data Transformation:**

- The hurricane data is cleaned and structured, including splitting states affected by hurricanes, handling multiple states, and formatting dates.
 - The housing data is similarly cleaned and transformed, ensuring proper date alignment with the hurricane data and dealing with missing values.
3. **Data Loading:** The final dataset is merged and stored in an SQLite database for easy access and further analysis.

Technology Stack:

- **Python:** For the overall scripting and automation.
- **pandas:** For data cleaning, transformation, and manipulation.
- **sqlite3:** For storing the cleaned and merged dataset.
- **Kaggle API:** For downloading the datasets.

Data Cleaning and Transformation

Hurricane Data

- Removed duplicates and ensured proper formatting of dates.
- The "States Affected Names" column was split to handle cases where multiple states were affected by a single hurricane. Each state is placed in its own row for easy analysis.
- Extracted year and month from the hurricane event date to align it with housing market data.

Housing Data

- Removed duplicates and handled missing values in the housing market dataset.
- Converted the 'period_begin' column to a datetime format, ensuring proper alignment of dates for merging with hurricane data.

Merging Datasets

- Merged the datasets based on Year, Month, and State.
- Used left joins to ensure all housing data is preserved, even if no hurricane data is available for that period. To align the events. This allows for analysis of housing market metrics in affected cities during the same months that hurricanes occurred.

Problems Encountered

Multiple Affected States

- One of the challenges encountered during the data cleaning phase was dealing with cases where a hurricane affected multiple states. The pipeline splits the "States Affected Names" field into separate rows, ensuring that each state affected by a hurricane is treated individually. This step is important for accurately analyzing the impact on cities within those states.

Handling Missing Data

- Missing data was handled using the `coerce` parameter in the `pd.to_datetime()` function to avoid errors when parsing invalid dates. This ensures that any invalid date entries are handled gracefully without causing interruptions in the pipeline.

Result and Limitations

Output Data

The result of the pipeline is a cleaned and merged dataset that contains the following columns:

Column	Description
Year	The year of the hurricane event or the housing market data entry.
Month	The month of the hurricane event or the housing market data entry.
State	The U.S. state affected by the hurricane or the location of the housing market data.
City	The city affected by the hurricane or the location of the housing market data.
Median Sale Price	The median sale price of homes in a particular city during the specified month.
Inventory	The number of homes available in the market during the specified month.
SS HWS	The Saffir-Simpson hurricane wind scale rating (e.g., Category 1–5).
Max Winds (kt)	The maximum wind speed recorded for the hurricane.
Storm Names	The name of the hurricane.

The data is stored in an SQLite database, ensuring that it is easily accessible for future analysis.

Data Structure and Quality

- **Structure:** The data is structured with columns for each relevant metric, including housing price data, inventory data, and hurricane event data. The datasets are merged by year, month, and state to allow for meaningful comparisons of housing data and hurricane impact.
- **Quality:** The datasets were cleaned to remove duplicates, handle missing values, and ensure proper date formatting. However, there may still be gaps or inconsistencies in the data due to missing hurricane events or incomplete housing market data for certain months or cities.

Data Format

The final output is stored in SQLite format. This choice was made because SQLite is lightweight and well-suited for storing and querying structured data, making it easy to handle large datasets. Additionally, it allows for efficient querying when conducting analysis on the data.

Limitations and Future Work

Limitations:

- The dataset only includes hurricanes and housing market data until 2023, so any hurricanes or market trends that occur after this period won't be captured.
- Some cities may have incomplete data due to reporting inconsistencies or gaps in the housing market dataset.

Future Work:

- Expanding the dataset to include more granular data such as specific cities within each state and potentially adding more variables such as housing price trends over longer periods.
- Incorporating other factors that could affect the housing market