Data Engineering Capstone

Data Pipeline Prototype

Data Cleaning and Transformations

Dataset: NOAA Historic Severe Storms

Data Pipeline Configuration (Classes):

- NOAABigQueryClient
- NOAADataFrameToCSV
- NOAADataRetrievalOrchestration
- NOAAExecutor

Data Pipeline Execution Sequence:

1. Initiate a Google BigQuery client
2. Execute a SQL query to retrieve and return all of the data, for a specified year, from each table as a DataFrame
3. Enable multithreading to concurrently execute step 4
4. Save each DataFrame as a CSV file

Data Cleaning and Transformations (Cont.)

Dataset: Homeowners & Renters Insurance by State

Data Pipeline Configuration (Classes):

- YearExtraction
- HomeInsuranceTableProcessing
- HomeInsuranceDataFrameCreation
- HomeInsuranceDataDisplayAndSave
- HomeInsurance Executor

Data Pipeline Execution Sequence:

1. Extract the year value from each of the table headers
2. Process each of the tables so that only 1 record exists per row
3. Compile all of the data from each of the tables into a single DataFrame
4. Display and save the dataframe as a CSV file

Data Cleaning and Transformations (Cont.)

Dataset: State to State Migration Flows

Data Pipeline Configuration (Classes):

- CensusDataDownloader
- CensusDataProcessor
- CensusDataMigration

Data Pipeline Execution Sequence:

1. Download the raw Excel files from Census.gov
2. Pre-process and clean the data from each of the Excel files into a DataFrame
3. Save each of the DataFrames into a cleaned Excel file

Exploratory Data Analysis (EDA)

Table of Content: EDA

Datasets:

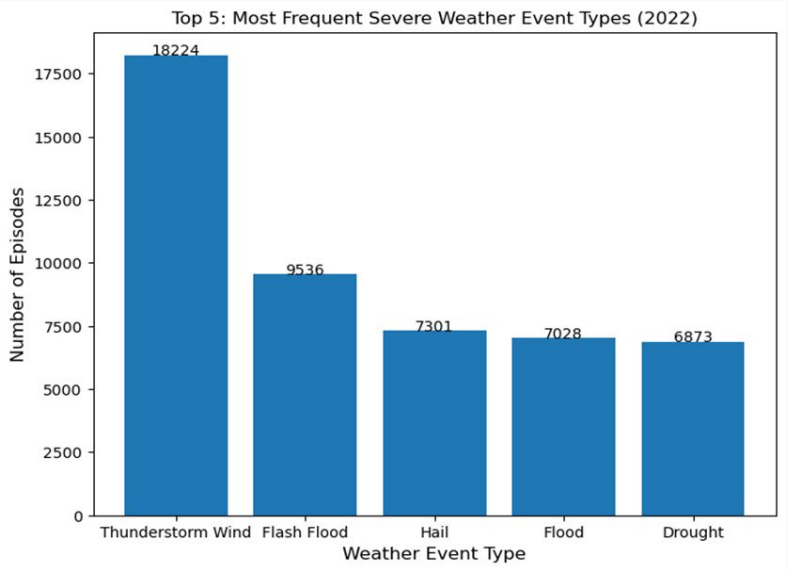
(*Based on the most recent year for which data is available*)

1. NOAA Historic Severe Storms (2022)
2. Homeowners & Renters Insurance by State (2007 - 2020)
3. State to State Migration Flows (2021)

NOAA Historic Severe Storms - 2022

The leading severe weather event across the United States was Thunderstorm Wind

Thunderstorm Wind events occurred almost 2x as frequently as Flash Flood

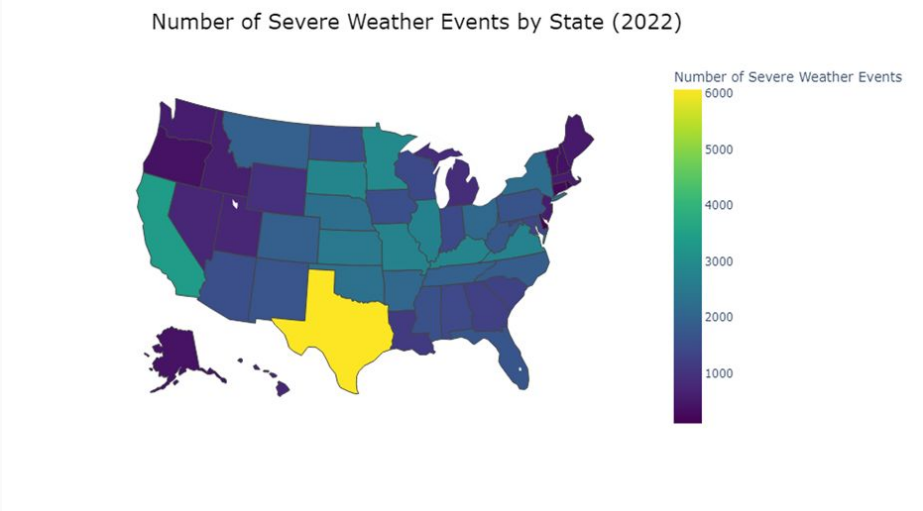


NOAA Historic Severe Storms - 2022 (Cont.)

Texas experienced the greatest number of severe weather events (totaling 6,067) - Nearly 2x as many as California

Second to Texas, California experienced a total of 3,358 severe weather events

In 2022, Delaware and Rhode Island experienced the least amount of severe weather events (totaling 103 and 119 respectively)

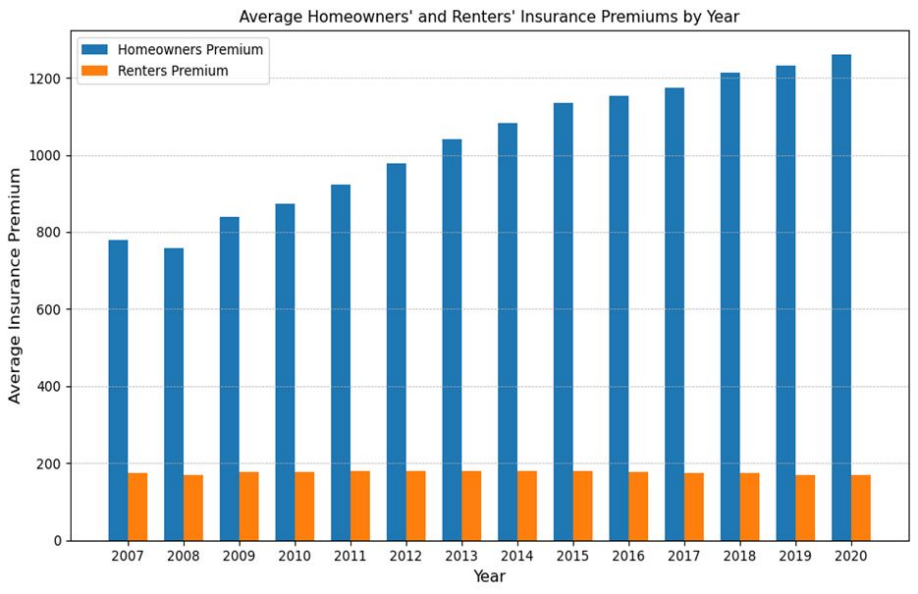


Homeowners & Renters Insurance

Renters’ insurance does not fluctuate materially over the years

Homeowners’ insurance, on the other hand, has risen from:

- Under 800 USD (in 2007)
- To over 1,200 USD (in 2020)



State to State Migration Flows - 2021

Florida and Texas experienced the greatest inflows

Whereas, California and New York suffered the largest outflows

Of the respondents who moved to Florida, most emigrated from New York (totaling 91, 758)

Of the respondents who moved to Texas, the majority emigrated from California (totaling 107,546)