Question: How do frequent flash floods affect local housing prices and business revenues?

Data Sources:

1. [1] CyberFlood1104:
   - Information about the dataset and why it was chosen: This dataset contains information about various flood events from 1998 to 2008 out of which 203 are in the US, which is the focus of the project. The data contains the following attributes: **ID** (unique identifier), **Year**, **Month**, **Day**, **Duration** (how many hours the event lasted), **fatality** (number of individuals harmed), **Severity** (1/1.5/2 as per the Dartmouth Flood Observatory), **Cause** (such as cyclone, rain, snowmelt, etc), **Lat**, **Long**, **Country Code**, **Continent Code**.
     This dataset was chosen as it appears to contain the minimum information required to specify when and where a flood occurred. As our question aims to find a relation between flash floods, housing prices, and business revenues, knowing the location and date of the flood seems to be sufficient for our analysis.
   - Structure and quality: The data is structured and present as a CSV file which is tabular. The data is not entirely complete and consistent as many columns have missing values and some columns have values of multiple datatypes (string in one row and numeric in another)
   - License: Licensed under the CC BY 4.0 International license. The license is mentioned under 'Rights' in this link. There are no such obligations mentioned apart from citing the source.

2. [2] HPIData:
   - Information about the dataset and why it was chosen: The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since January 1975. The data covers dates from 1991 till the present date. This dataset was chosen as it is a very commonly used dataset when dealing with housing-prices-related data engineering problems. Also, there is an overlap of country and timeline of the housing prices in this dataset and flooding events mentioned in the previous dataset. This will help us find a relation, if any, between the two events.
   - Structure and quality: The data is structured and present as a CSV file which is tabular. The data is complete and consistent and does not require any preprocessing.
   - License: The FHFA House Price Index (HPI) dataset does not have a clearly specified license, but it is publicly available and free to use as part of the U.S. government's commitment to open data.

Data Pipeline:

High-Level Overview:

The data pipeline extracts datasets from given URLs, processes them for quality and consistency, and stores the cleaned data into local CSV files. The pipeline uses Python and the pandas library for data manipulation and transformation.

Data Transformation and Cleaning Steps:

1. CyberFlood1104:
   - Unnecessary Column Removal: Columns like Unnamed: 0, Unnamed: 12–Unnamed: 15, and datetime were dropped to eliminate irrelevant or redundant data.
   - Missing Value Handling: Critical columns like ID, Year, and Month had rows with missing values dropped. Non-critical columns (e.g., Day, Duration (hours), fatality, Severity, Cause, Lat, Long) were filled with placeholder values (-1 for numeric and 1000 for coordinates) to retain these rows for further inspection.
   - Outlier Removal: Row with ID = 834 was excluded due to mismatched data, improving dataset consistency.
   - Data Type Assignment: Explicit data types were assigned to columns to ensure uniformity during analysis.
   - Column Name Formatting: Column names were converted to lowercase and stripped of extra whitespace for consistent naming conventions.
   - Sorting and Resetting Index: Rows were sorted by ID and re-indexed to ensure a structured dataset.

2. HPI Dataset: No preprocessing was performed as the dataset was deemed clean.

Meta-Quality Measures and Error Handling:

1. Placeholder Values: Missing data was replaced with placeholder values (-1 or 1000), preserving the dataset for further analysis.
2. Outlier Handling: Outlier rows were removed based on prior inspection. Future iterations could automate this using statistical measures.
3. Dealing with errors and changing input data: While the pipeline processes the current dataset structure effectively, dynamic schema validation (e.g., checking for expected column names and types) could improve adaptability to evolving data formats. The pipeline assumes the URLs provide accessible CSV data. If a URL is invalid or the format changes, the pd.read_csv() function will raise an error. Adding a try-except block would improve robustness by handling such cases gracefully.
4. Problems Encountered and Solution: Most problems such as irregular column names, missing data, outliers, incorrect data-type assignment, etc. were solved respectively by pre-processing the data appropriately.

Result and Limitations:

- The output of the data pipeline are two CSV files with tabular data for each dataset respectively. The quality of the dataset seems good so far but can only be further understood if it can successfully answer the desired question.
- The data format chosen was CSV due to familiarity with the format and using it with pandas. Familiarity with MS Excel also led to choosing this format as results using code can be cross-checked by filtering in Excel.
- Limitations:
    - The flood events data might not be enough to convincingly answer the desired question. If that is the case, then a larger dataset can be used such as the United States Flood Database.
    - The question also mentions business revenues but currently no dataset for business revenues has been selected due to time constraints. More datasets can be added in the future, or the question can be revised to solely focus on flood events and housing prices.

References:

[1] Li, Z. (2020). United States Flood Database (v1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7545697

[2] Housing Price Index Data, Federal Housing Finance Agency https://www.fhfa.gov/hpi/download/monthly/hpi_master.csv