Analyzing the Impact of Startup Growth on Household Income in U.S. Metropolitan Areas

1. Motivation

The rapid growth of startups has been advertised as a driver of regional economic development. However, little empirical evidence exists on whether this increase in entrepreneurial activity significantly affects household income. This study seeks to explore the potential relationship between startup density and average household income in U.S. metropolitan areas, leveraging extensive data integration through an automated ETL pipeline.

2. Research Question

2.1. Primary Question:

- Can an increase in number of startups cause higher average household income in a metropolitan area?

2.2. Supporting Questions:

- How does startup density vary across metropolitan regions?
- What role do high-growth startups (unicorns) play in influencing household income?

3. Data Sources

The following datasets were chosen to provide comprehensive coverage of financial, geographical, and demographic information relevant to analyzing the relationship between startup density and household income.

ID	Dataset	Description	Source	URL	License
1	Startup Investments	Detailed information on startup funding rounds, sectors, and locations across U.S. metropolitan regions	Kaggle	Link	CC0 Public Domain / PDDL
2	U.S. Household Income	Household income distributions, including median and mean incomes, across U.S. ZIP codes with bins showing income distribution with bin size of 15000 USD	U.S. Census Bureau via Kaggle	Link	CDLA-Sharing-1.0 Census bureau data is available with license: Attribution 4.0 International. Details can be found here
3	Unicorn Startups	Lists of unicorn startups in the U.S., detailing their valuation, founding dates, and cities	Kaggle	Link	CC0 1.0
4	U.S. Cities Database	Demographic information, population size, and city-level ZIP codes	Kaggle	Link	CC BY 4.0.
5	U.S. Cities by Population	Population data for the largest U.S. cities, including geographic and state-level information	Kaggle	Link	CC0 1.0
Each dataset is under a standard open-data license. Detailed license information can be found at their respective Kaggle dataset pages.					

4. Structure and Quality of Data

The datasets used in this project are in CSV format, containing rows for cities, ZIP codes, or startups and columns for various attributes.

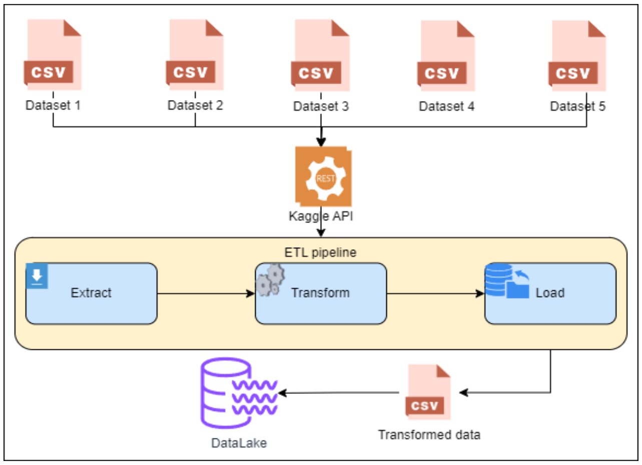
- **Startup Investments Dataset:** Contains over 20 columns detailing funding types, companies, and geographic locations. The data is well-structured but includes missing funding round details in some records, requiring filtering for U.S.-specific startups.
- **Household Income Dataset:** Contains columns representing income brackets, ZIP codes, and median/mean household income. The data is highly accurate but uses percentage-based income brackets, necessitating conversion to absolute values.
- **Unicorn Dataset:** A flat dataset with fields such as city, valuation, and year founded. Minimal missing data but requires standardization of city names.
- **Cities Database:** A hierarchical structure linking city and state data with demographic metrics. It is consistent but requires normalization of ZIP codes and city/state names.
- **Population Dataset:** Contains population statistics, state affiliations, and area data. While comprehensive, it may include duplicates that need deduplication during merging stages.

5. Data Pipeline

The data pipeline was implemented using Python, with Pandas for data manipulation and the Kaggle API for data extraction. Data pipeline steps are described below:

5.1. **Extraction:** Data files were downloaded using the Kaggle API with *retry* mechanism for fault tolerance. Downloaded dataset with automatic unzipping of data

5.2. **Transformation:** Unnecessary columns were removed, data types were standardized, and datasets were merged (with inner join on common columns e.g. city, state and zip_code). Missing values were addressed, duplicate rows were removed, column names were standardized and replaced spaces with "_" in city/state names.



- **Crunchbase Data**:
  - Filtered U.S. only startups and converted funding fields to float.
  - Added a new column Valuation ($B) by summing funding rounds and converting to billions.
- **Household Income Data**:
  - Dropped error and families columns
  - Converted income percentages to absolute counts using total households
  - Standardized ZIP codes to 5-digit format
- **Unicorn Data**:
  - Filtered U.S. unicorns and normalized city names for consistency in merging
  - Convert founded date type from string to date using pd.to_datetime() function
- **Cities Database**:
  - Exploded ZIP code lists and handled duplicate cities by retaining the most populated entries
- **Population Data**:
  - Normalized city and state names and retained only relevant columns for merging

5.3. **Merging (part of transformations):**

- Combined all transformed datasets using pd.merge() and pd.concat() function
- Concatenated Crunchbase _startup with unicorn data at axis=0 (appended rows).
- Merged Crunchbase _startup and unicorn data with cities, ZIP codes, and income data.
- Final dataset created by linking all components based on city and zip_code columns.

5.4. **Load:** Lastly, transformed and merged dataframe were stored as CSV file in data directory of project and can be used for further analysis to answer research question. The transformed dataset is saved to the ***"data/"*** directory in project main directory with file name "**startups_household_income_metropolitan_area**".

The table below summarizes the data quality analysis across various datasets used in this project.

Aspect	**Strengths**	**Potential Issues**	**Solutions**
Accuracy	Data from reliable sources (Census Bureau, Crunchbase).	Missing or incorrect funding details in some startup entries.	Standardized and converted data formats, filtered out invalid data.
Completeness	Covers most necessary information for the analysis.	Some missing early-stage startup data, error fields in income.	Dropped incomplete records; kept most relevant fields.
Consistency	Uniform data format achieved after cleaning (ZIP codes, funding).	Initial inconsistencies in city names and funding fields.	Standardized names and ensured numeric conversions.
Timeliness	Focus on recent data (2021 income, active startups).	Older startup records may introduce bias.	Filter out older inactive startups for modern economic trends.
Relevancy	Data directly supports research question (startup density & income impact).	Outdated unicorn data could affect results.	Remove closed startups data.

6. Challenges and Solutions:

- Missing information from dataset: In first 2 dataset I could not able to find all the data needed so I had to combine five different datasets, which could potentially cause inconsistencies in data.
- Meta-quality measures: The pipeline includes error handling during API calls and data validation checks to ensure integrity. Raising errors on columns that are required.
- Inconsistent City column name across various datasets: Standardized column names for all dataset.

7. Result and Limitations

The output data is tabular, with consistent formatting and minimal missing values. It has been cleaned to remove irrelevant or redundant columns. The final dataset is stored as a CSV file to ensure compatibility with various analytical tools and it is ready to be aggregated for further analysis. Data size is not huge and can be handled well with CSV files.

The dataset may not account for all startups or household income fluctuations over time. Aggregated data could mask regional variations or outliers. These limitations will be considered in the final analysis. Aggregated data might also have duplicate rows due to multiple branches of company in different cities that might impact future analysis.

8. Conclusion

The ETL pipeline efficiently integrates startup, geographic, and economic datasets, providing a robust foundation for answering the research question: **"Can an increase in startups cause higher average household income in metropolitan areas?"** Future steps involve statistical modeling and correlation analysis to validate this hypothesis.