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 startup**s** 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.

3.1. Startup Investments Dataset

- **Source**: Startup Investments Crunch-base data via Kaggle
- **Description**: Detailed information on startup funding rounds, sectors, and locations across U.S. metropolitan regions.
- **Data URL**: https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase
- **License**: CC0 Public Domain / Open Data Commons Public Domain Dedication and License (PDDL).

3.2. U.S. Household Income by ZIP Code Dataset

- **Source**: U.S. Census Bureau via Kaggle
- **Description**: Household income distributions, including median and mean incomes, across U.S. ZIP codes with bins showing income distribution with bin size of 15000 USD.
- **Data URL**: https://www.kaggle.com/datasets/claygendron/us-household-income-by-zip-code-2021-2011
- **License**: Community Data License Agreement – Sharing – Version 1.0
    - Originally data is fetched from census bureau and is available with license: Attribution 4.0 International. License details can be found here

3.3. Unicorn Startups Dataset

- **Source**: Various sources via Kaggle

- **Description**: Lists of unicorn startups in the U.S., detailing their valuation, founding dates, and cities.
- **Data URL**: https://www.kaggle.com/datasets/ramjasmaurya/unicorn-startups
- **License**: CC0 1.0

3.4. U.S. Cities Database

- **Source**: Open datasets via Kaggle
- **Description**: Demographic information, population size, and city-level ZIP codes.
- **Data URL**: https://www.kaggle.com/datasets/sergejnuss/united-states-cities-database
- **License**: Creative Commons Attribution 4.0

3.5. U.S. Cities by Population Dataset

- **Source**: Open datasets via Kaggle
- **Description**: Population data for the largest U.S. cities, including geographic and state-level information.
- **Data URL**: https://www.kaggle.com/datasets/axeltorbenson/us-cities-by-population-top-330
- **License**: 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

4.1. Structure

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**: Includes over 20 columns detailing funding types, companies, and geographic locations.
- **Household Income Dataset**: Contains columns representing income brackets, ZIP codes, and median/mean household income.
- **Unicorn Dataset**: Flat dataset with fields such as city, valuation, and year founded.
- **Cities Database**: Hierarchical structure linking city and state data with demographic metrics.
- **Population Dataset**: Population statistics, state affiliations, and area data.

4.2. Quality

- **Startup Data**: Well-structured but includes missing funding round details in some records. Requires filtering for U.S.-specific startups.
- **Household Income Data**: High accuracy but includes percentage-based income brackets, necessitating conversion to absolute values.
- **Unicorn Startups**: Minimal missing data but requires standardization of city names.
- **Cities Database**: Consistent but requires normalization of ZIP codes and city/state names.
- **Population Data**: Comprehensive, though duplicates may need deduplication in 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:

- **Extraction**: Data files were downloaded using the Kaggle API.
- **Cleaning**: Missing values were addressed, duplicate rows were removed, and column names were standardized.
- **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).
- **Loading**: 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.

6. Challenges and Solutions:

Challenge: Handling large datasets during merging.

Solution: Optimized memory usage by selecting only relevant columns and using efficient merging techniques.

Meta-quality measures: The pipeline includes error handling during API calls and data validation checks to ensure integrity.

7. Result and Limitations

The output data from the pipeline includes columns for startup details, household income metrics, population density, and geographic attributes.

Structure and quality: The data is tabular, with consistent formatting and minimal missing values. It has been cleaned to remove irrelevant or redundant columns.

Output format: The final dataset is stored as a CSV file to ensure compatibility with various analytical tools. Data size is not huge and can be handled well with CSV files.

8. Limitations:

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.

9. 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.