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

Noman Arshad
Friedrich-Alexander-Universität Erlangen-Nürnberg

1. Introduction

The relationship between entrepreneurial activity and regional economic development has become increasingly relevant as cities strive to become startup hubs. This analysis investigates whether an increase in the number of startups can lead to higher average household income in U.S. metropolitan areas. The question is particularly pertinent as cities invest resources in attracting and nurturing startups, often with the idea that this will boost local economic prosperity.

This research focuses on both direct and indirect relationships between startup activity and household income, examining these relationships at different geographical levels (city, county, and state) to provide a comprehensive understanding of the potential economic impact of startup ecosystems.

2. Data ETL Process & Transformed Data:

2.1. Data Preparation and Structure

The analysis is based on a comprehensive integrated dataset discussed in **"data-report"** that combines startup activity metrics with household income distributions across U.S. metropolitan areas. The datasets, primarily in CSV format, underwent extensive preprocessing to ensure consistency and accuracy. This involved standardizing ZIP codes, normalizing city names, converting income percentages to absolute values, and merging datasets on common fields (e.g., city, ZIP code). Missing values were filtered or imputed, and duplicate records were removed.

The final dataset (**startups_household_income_metropolitan_area.csv**) contained approximately **14,846** records with **75** columns, combining demographic, startup, and economic data with temporal data added as separate column with respective year, structured to capture both temporal and geographical dimensions of economic activity. Region column was also engineered to categorize data into west or east region based on longitude to analyze impact of startup activity on economic indicators in regions. It is difficult to show all **75** columns. **Table 1** shows the key features of transformed data.

Category	Variables	Description	Time Coverage
Geographic Data	ZIP code, city, county, state, lat/long, region (east U.S /west U.S.)	Location identifiers and coordinates	Static
Household Income	Income brackets ($10k to $200k+), median, mean	Distribution across 10 income levels	2018-2021
Startup Metrics	Company, valuations, industry, city	Startup activity and growth metrics	2017-2021
Demographics	Population, density, land area	Regional characteristics	Static

Table 1: Key Data Features

2.2. Data Sources and Attribution

This dataset integrates information from multiple sources, each used under appropriate open-data licenses and all data sources were used in compliance with their respective licenses and proper attribution:

1. **Startup Investment Data:** Sourced from Kaggle under CC0 Public Domain license
2. **U.S. Household Income:** Obtained from U.S. Census Bureau via Kaggle under CDLA-Sharing-1.0 license
3. **Unicorn Startups Data:** Kaggle dataset under CC0 1.0 license
4. **U.S. Cities Database:** Used under CC BY 4.0 license
5. **U.S. Cities Population Data:** Incorporated under CC0 1.0 license

3. Analysis

The analysis employed a comprehensive approach combining statistical methods, temporal analysis, and geographical comparisons. Multiple analytical techniques were used to ensure robust findings.

3.1. Statistical Methods:

3.1.1. Correlation Analysis

A correlation analysis was performed to determine the relationship between startup activity and household income. This analysis was conducted at three geographic levels; city, county, and state. The primary focus was on the correlation between:

- Number of startups and household mean income
- Number of startups and high-income household share
- Valuation of startups and high-income household share

The correlation matrices for each geographic level were visualized to highlight patterns and trends.

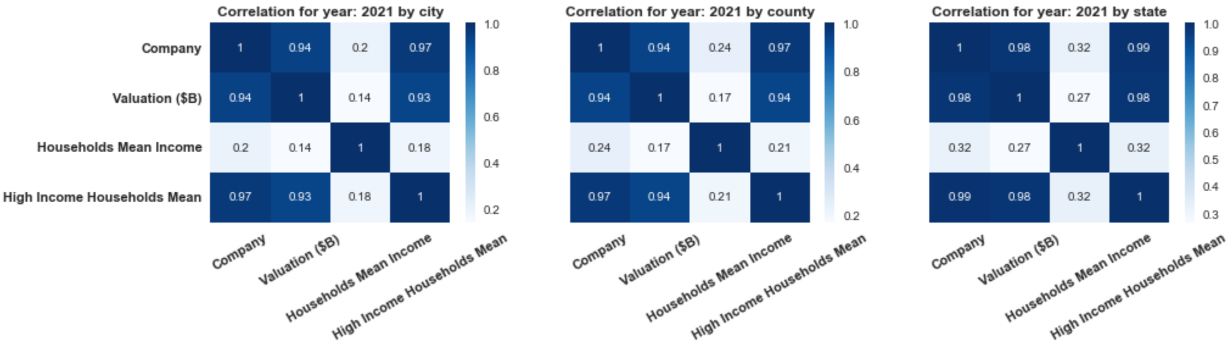

Figure 1: Correlation matrices for city, county, and state levels (2021)

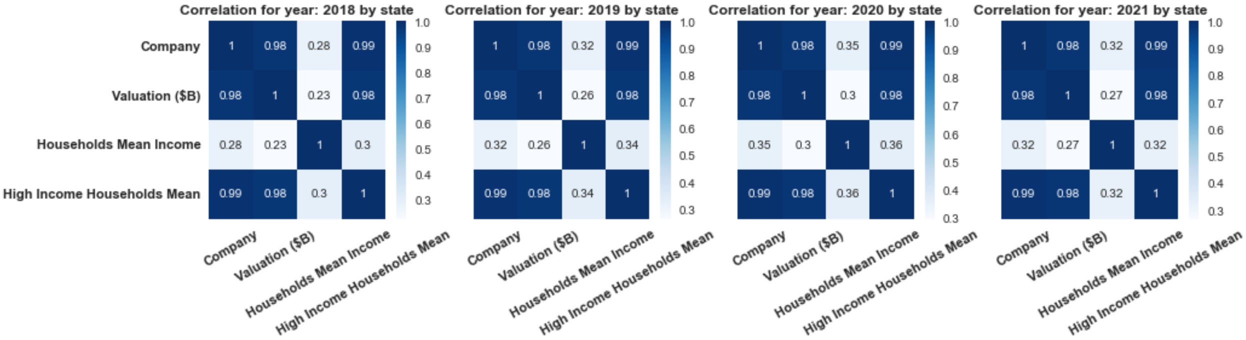

Figure 2: Year-over-year correlation trends (2018-2021) by State

Correlation coefficient **(r)** ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation. In this analysis, values around 0.2 indicate weak positive relationships, while values above 0.3 suggest moderate positive relationships and values above 0.9 shows strong positive relationships.

3.1.2. Regression Analysis

- Linear & exponential regression models were fitted to quantify relationships **(**Orange Line **in Figure 4,5 & 6)**
- R-squared values were calculated to measure explanatory power (achieved **R^2: 0.52**)
- The **R^2** value of **0.52** indicates that **52%** of the variation in household income can be explained by startup activity, suggesting other important factors influence income levels
- Residual analysis was performed to check for regional variations

3.2. Visual Analysis:

Generated various visualizations on aggregated data:

- Temporal evolution of startup count and income levels (**Figure 2**)
- Year over year changes in correlation strengths (**Figure 3**)
- East vs. West regional differences (**Figure 4, Figure 5 & Figure 6**) (sky blue **-> East,** dark blue **-> west**)
- City, County & State-level scatter plots with valuation-based sizing (**Figure 4, Figure 5 & Figure 6)**

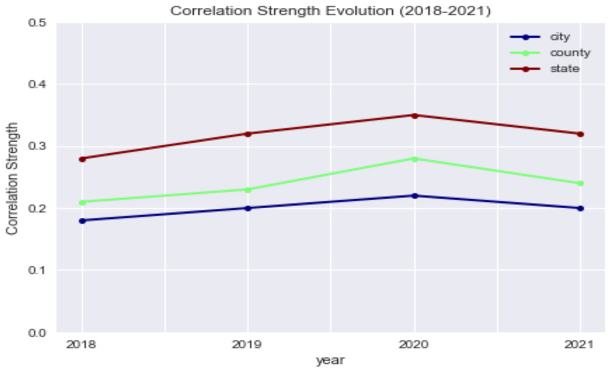

Figure 3: Year over Year Correlation Strength

4. Results & Interpretation

4.1. Startup Density and Average Income

The correlation analysis reveals a positive relationship between the startups and average household income across all geographic levels:

- **City level**: Moderate positive correlation (**r ≈ 0.20**)
- **County level**: Stronger positive correlation (**r ≈ 0.24**)
- **State level**: Strongest positive correlation (**r ≈ 0.32**)

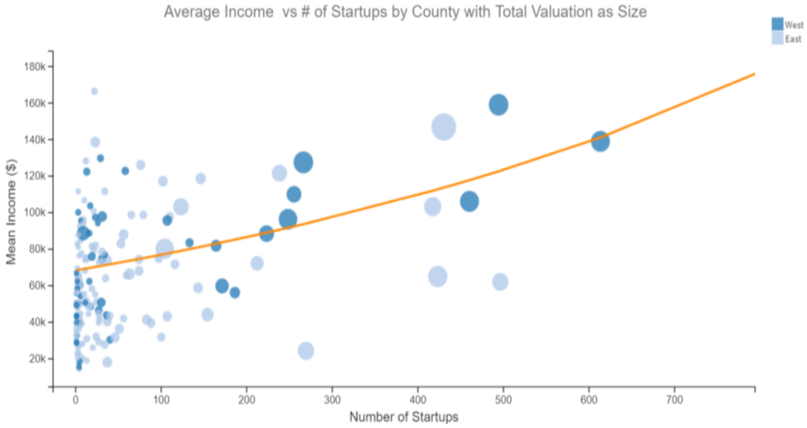

Figure 4: Mean Income vs. # of Startups by County with Valuation as size

The strengthening correlation at broader geographic scales suggests that startup impacts may have spillover effects beyond immediate city/county/state boundaries.

4.2. Startup Valuations and Income Levels

- Positive correlation between total startup valuations and mean household income (**r > 0.25**)
- The relationship is consistent across cities, counties, and states
- The correlation strengthened over the studied period (**2018-2021**)

This suggests that the quality and success of startups (measured by valuation) also have some influence on household income.

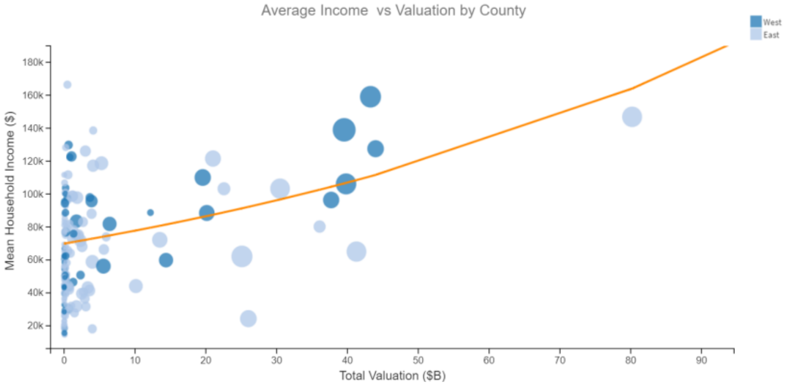

Figure 5: Mean Income vs. Valuation by County with regions

4.3. High-Income Household Distribution

The analysis of high-income households **(having annual income >$200,000**) reveals:

- Strong correlation with startup presence (**r > 0.95**)
- Higher concentration in areas with more valuable startup ecosystems

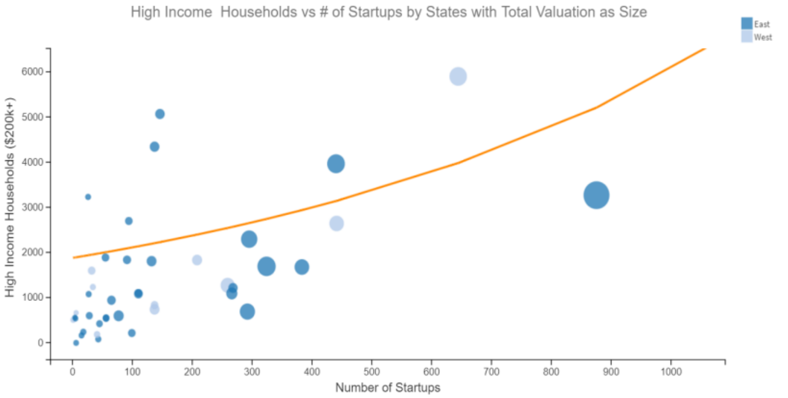

Figure 6: High Income Households vs. Startup Density

- More pronounced effect in metropolitan areas with established startup clusters
- Overall **Increasing effect (shown by regression lines)** in High Income Households **($200k+)** with **increase in startup density**.

As shown in Figure 1 and , the analysis reveals a positive but modest correlation between startup activity and household income, with patterns strengthening consistently from city to state level across the study period **(2018-2021**). Correlations are stronger at the state level **(r ≈ 0.32)** compared to the city level (r ≈ 0.20), and a similar geographic scaling exists between startup valuations and income measures **(r = 0.14 at city level to r = 0.27 at state level)**. While startups contribute to economic growth, other factors such as local policies, infrastructure, and education play significant roles. Notably, there is a high correlation between startup growth and the number of high-income households (**r > 0.96**), suggesting that areas with high startup activity tend to see an increase in households **earning over $200k**.

5. Conclusion

The analysis provides evidence supporting a **positive relationship** between **startup activity** and **household income** in metropolitan areas. However, the relationship is **modest**, and the effect **strengthens** at larger **geographic levels** (e.g., counties, states). It appears to be part of a broader ecosystem of economic factors that collectively contribute to increased metropolitan prosperity. The **stronger correlations** with **startup valuations** indicate that the quality and success of startups are somewhat more important than quantity in driving economic benefits.

5.1. Key Findings:

1. Metropolitan Areas with high startup activity tend to see an increase in **high income households ($200k and more)**
2. The relationship is stronger when considering startup valuations rather than just the number of startups
3. The impact appears to be regional rather than purely local
4. The effect is more pronounced in areas with established startup ecosystems

5.2. Limitations

The question was partially answered by demonstrating a positive correlation between startup growth and household income. However, several limitations and uncertainties exist:

- **Data Limitations**: Aggregated data may mask intra-regional variations. Startups also have offices in multiple cities, due to this overall impact have distributed effect, not showing true impact of number of startups per region
- **Causality**: The analysis establishes correlation and it does not imply causation. Other factors like educational attainment, living cost or pre-existing economic conditions may drive both startup growth and income levels
- **Time Lag**: Economic benefits from startups may take time to materialize, which is not fully captured by this analysis
- **Covid-19:** Time range 2018-2021 does include Covid-19 period and it have impact on **overall Income Mean in 2019 & 2020**, as shown in **Figure 3**

These findings suggest that cities might benefit more from policies focused on supporting startup quality and growth with good infrastructure rather than just increasing number of startups.

6. Future Research Directions

1. Analysis based on educational attainment data
2. Investigation of specific industry sectors' impacts
3. Analysis of startup failure rates and their economic impacts
4. Examination of other economic indicators beyond household income like cost of living etc.

7. References:

- OpenAI. (2024). *ChatGPT (January 2025 version)*. Retrieved from https://openai.com/chatgpt. Used as a tool to improve document conciseness and shorten its length