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

Noman Arshad

Methods of Advanced Data Engineering

Friedrich-Alexander-Universität Erlangen-Nürnberg

Agenda

Introduction

Datasets

ETL Pipeline

Analysis

Outcomes

Conclusion

Introduction

- Investigating how startups influence household income in U.S. metropolitan areas
- Research Motivation: Cities aim to boost economic growth by fostering startups
- Key Question: Does an increase in startups lead to higher household income?
- Secondary Question: What other factors have positive impact on higher income housholds?

Datasets

1. Startup Data with Fundings and Locations

- **Metadata URL:** Kaggle Link
- **Data URL:** Dataset Link
- **Type:** CSV
- **Description:** Contains investment series at a company level with detailed locations.
- **License:** CC0 Public Domain

2. United States Household Income

- **Metadata URL:** Kaggle Link
- **Data URL:** Dataset Link
- **Type:** CSV
- **Description:** Provides average household income data by metropolitan area, including income distribution bins.
- **License:** CDLA-Sharing-1.0

Datasets

3. Unicorn Startups

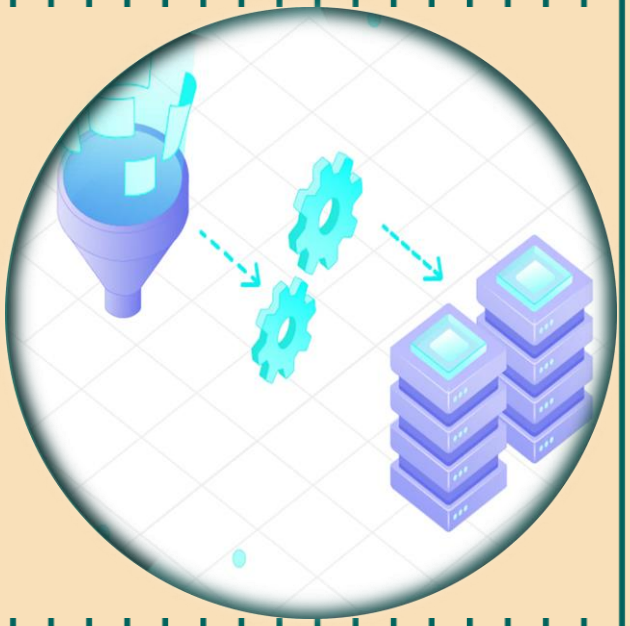
- **Metadata URL:** Kaggle Link
- **Data URL:** Dataset Link
- **Type:** CSV
- **Description:** Includes information about billion-dollar startups (unicorns), their funding rounds, and founding dates
- **License:** CC0 1.0

4. United States Cities Data

- **Metadata URL:** SimpleMaps Link
- **Data URL:** Dataset Link
- **Type:** CSV
- **Description:** Contains zip codes, population density, and geographic details for U.S. cities
- **License:** CC BY 4.0

5. U.S. Cities by Population

- **Metadata URL:** Kaggle Link
- **Data URL:** Dataset Link
- **Type:** CSV
- **Description:** Includes statistics like land area and population density, enabling higher-level regional analysis
- **License:** CC0 1.0

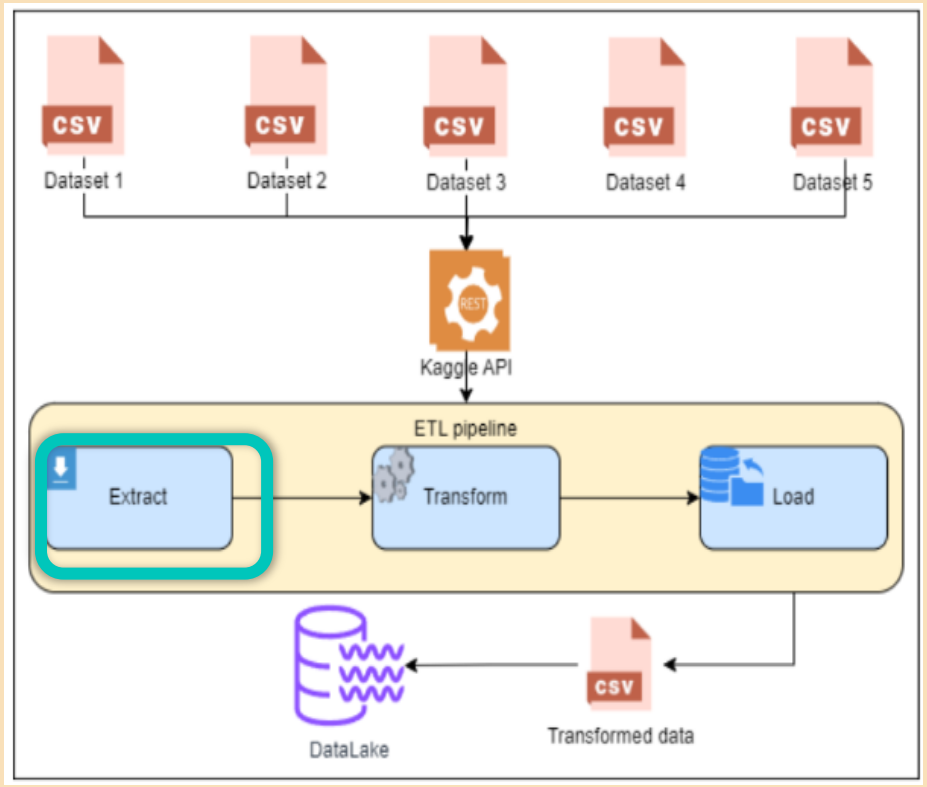


ETL Pipeline

Data Cleaning & Transformation

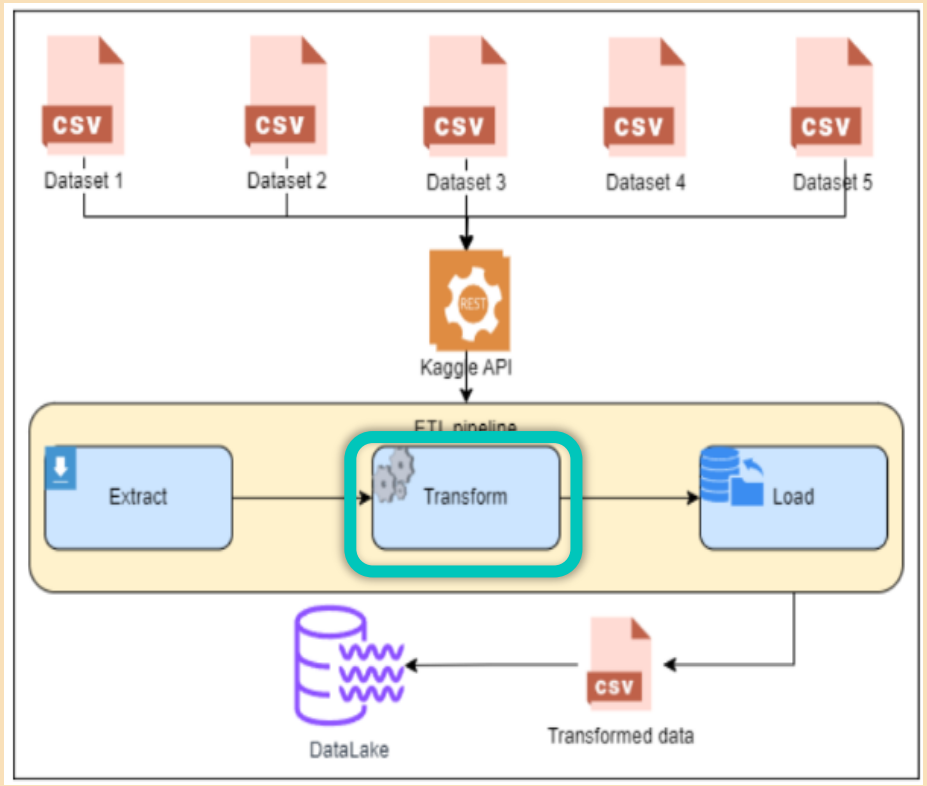
ETL Pipeline

- **Extraction:**
  - Data downloaded using Python, Kaggle API
  - Retry mechanism for fault tolerance
  - Auto unzipping
  - Logging for better visibility of process



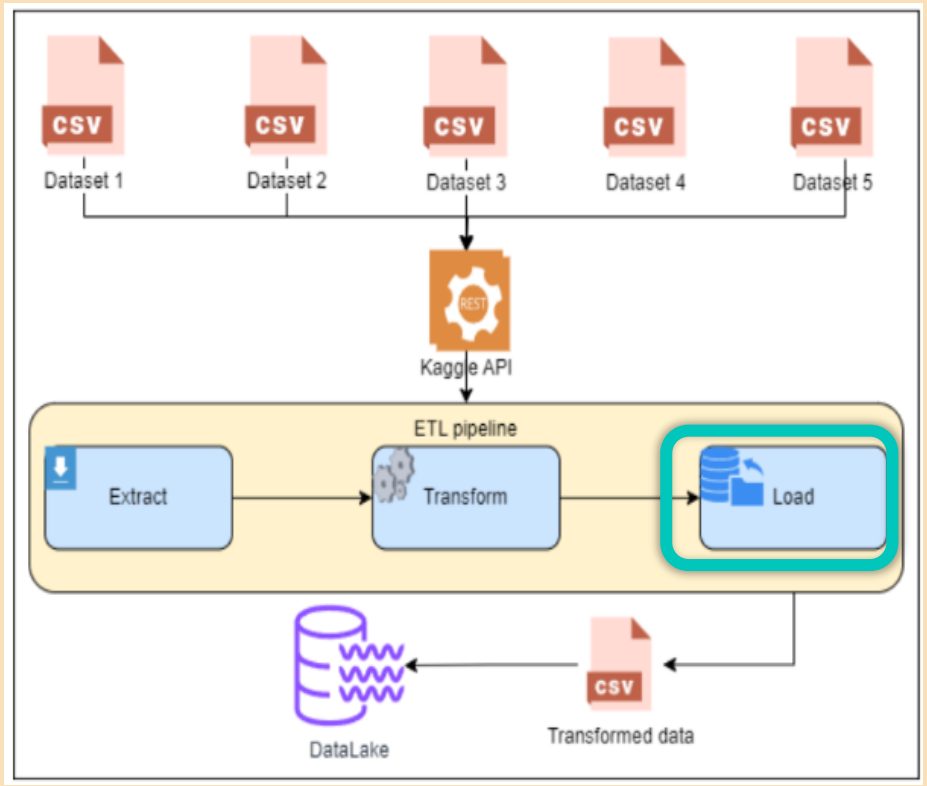
ETL Pipeline

- **Transformation:**
  - Removed duplicates
  - Standardized data types
  - Missing values were addressed
  - Column names were standardized and replaced spaces with "_" in city/state names for consistency
  - Merged all data frames for further analysis, using common columns like city and zip code



ETL Pipeline

- **Load**:
  - Transformed data stored in data folder in CSV file with14,846 records & 75 columns.
  - Features: Geographic identifiers, Income Brackets (2018-2021), Startup Metrics (2017-2021).

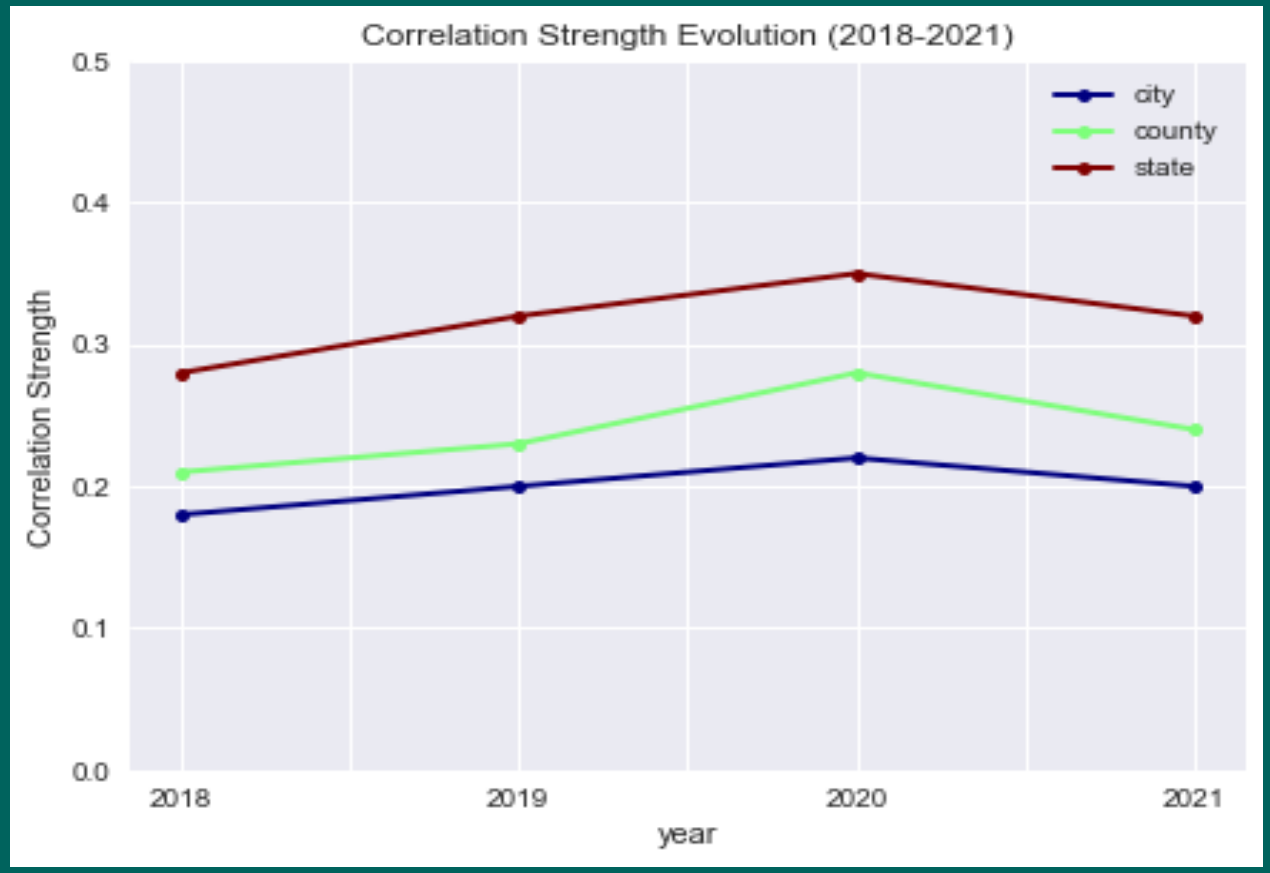


Analysis

Correlation Analysis

Identified temporal trends

1. Year over Year correlation between:
   - Household Income and Startup Count
2. A jump in year 2019-2020 reflects covid-19 effect
3. Broader region have stronger correlations



Correlation Analysis

Visualized geographic correlation trends for year 2021

- Correlation matrix at **City** level
- Between Startup count, Valuation, Mean Income and High-income households.

Correlation Analysis

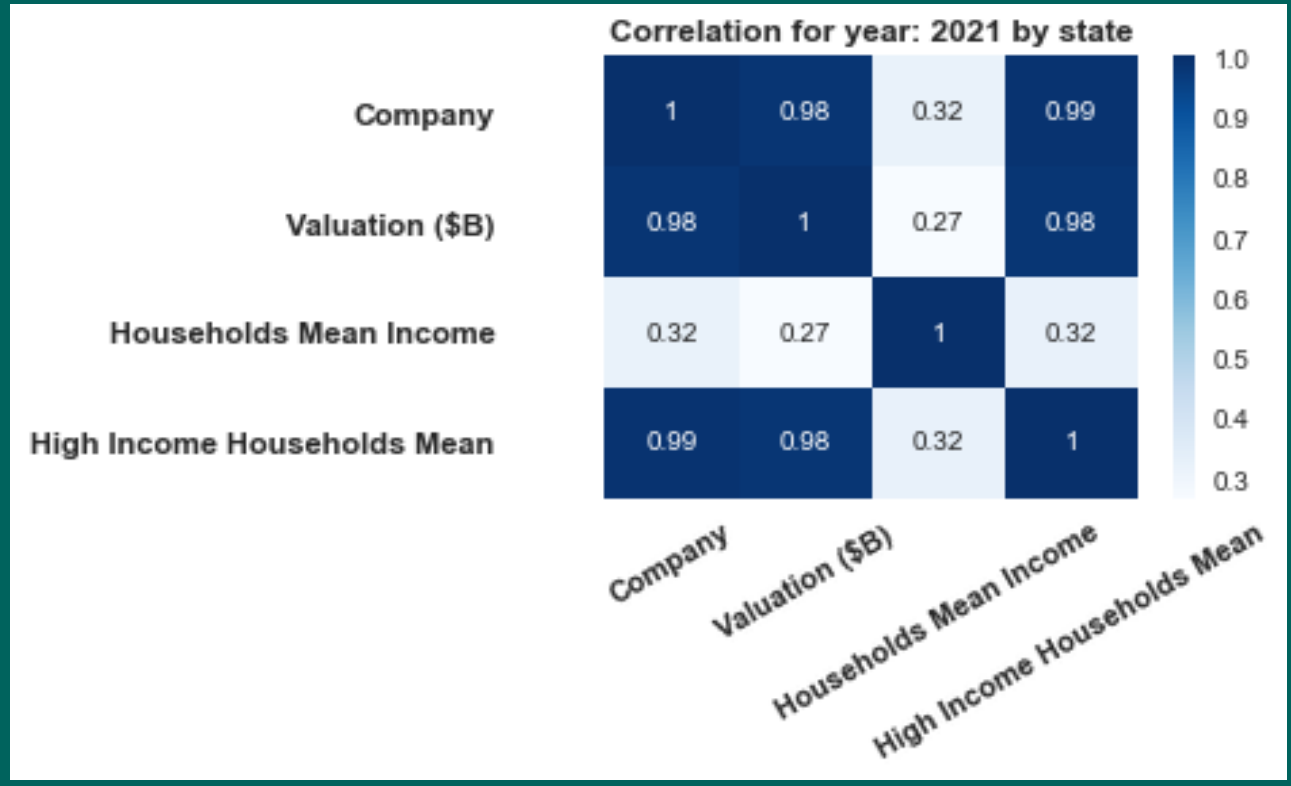
Visualized geographic correlation trends for year 2021

- Correlation matrix at **County** level
- Between Startup count, Valuation, Mean Income and High-income households.

Correlation Analysis

Visualized geographic correlation trends for year 2021

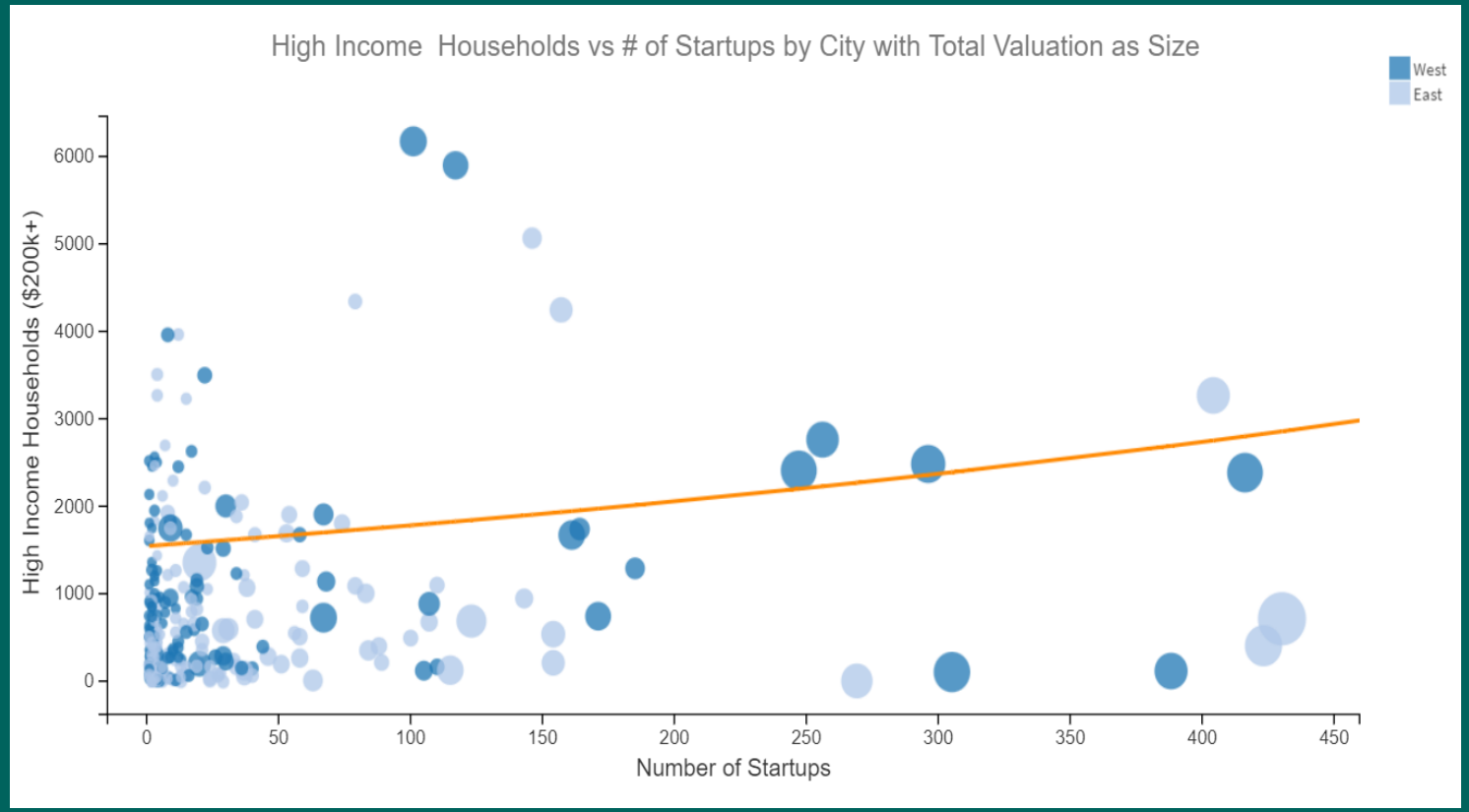
- Correlation matrix at **State** level
- Between Startup count, Valuation, Mean Income and High-income households.



Regression Analysis ($R^2 = 0.48$)

Regression Analysis: Quantified relationships ($R^2 = 0.48$)

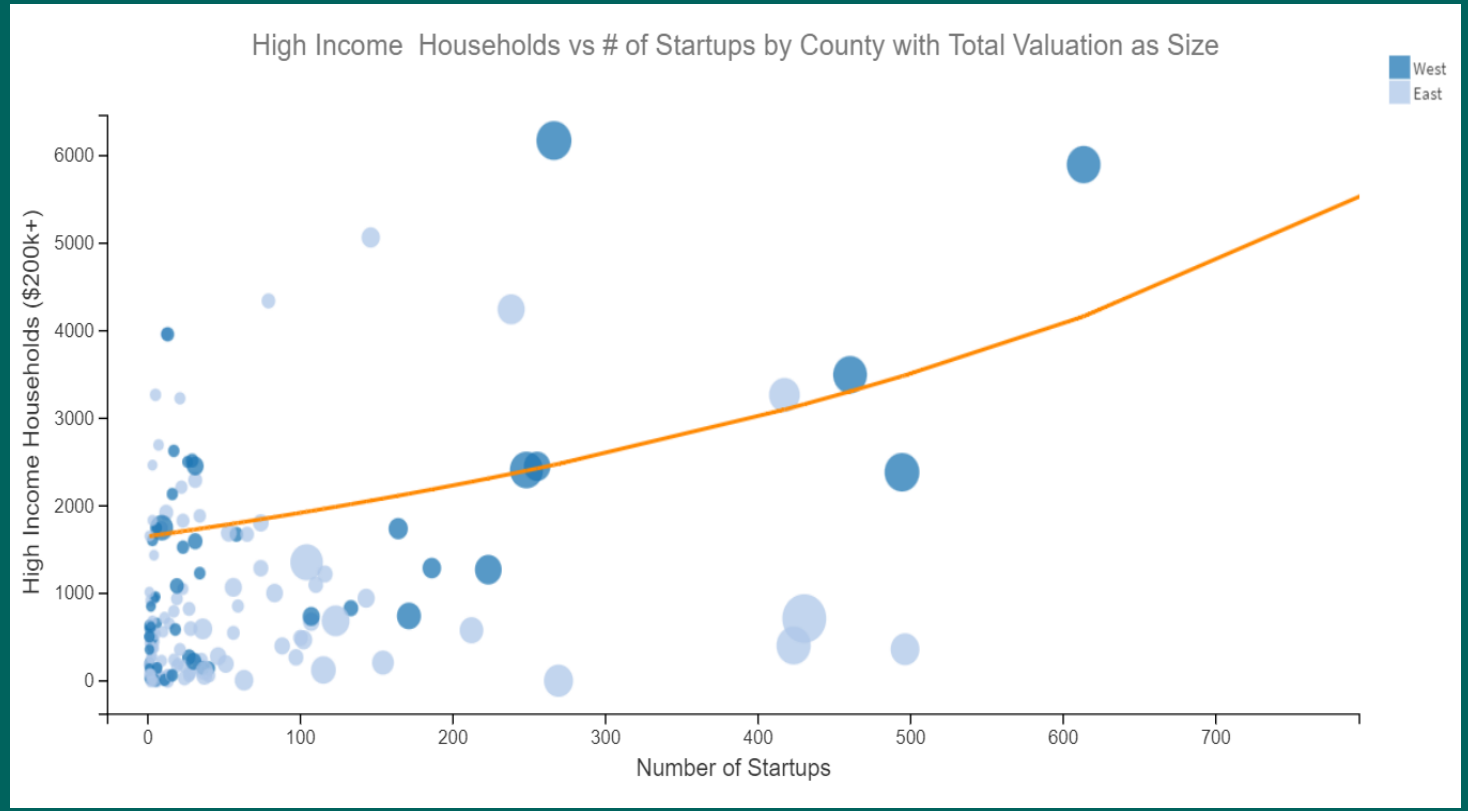
- Number of Startups vs High Income Households (>200k$)
- Regression line in Orange shows overall increasing relation, but with $R^2 = 0.48$
- Size of bubble shows total valuation by **City**
- High valuation shows more prominent effect on income



Regression Analysis ($R^2 = 0.52$)

Regression Analysis: Quantified relationships ($R^2 = 0.52$)

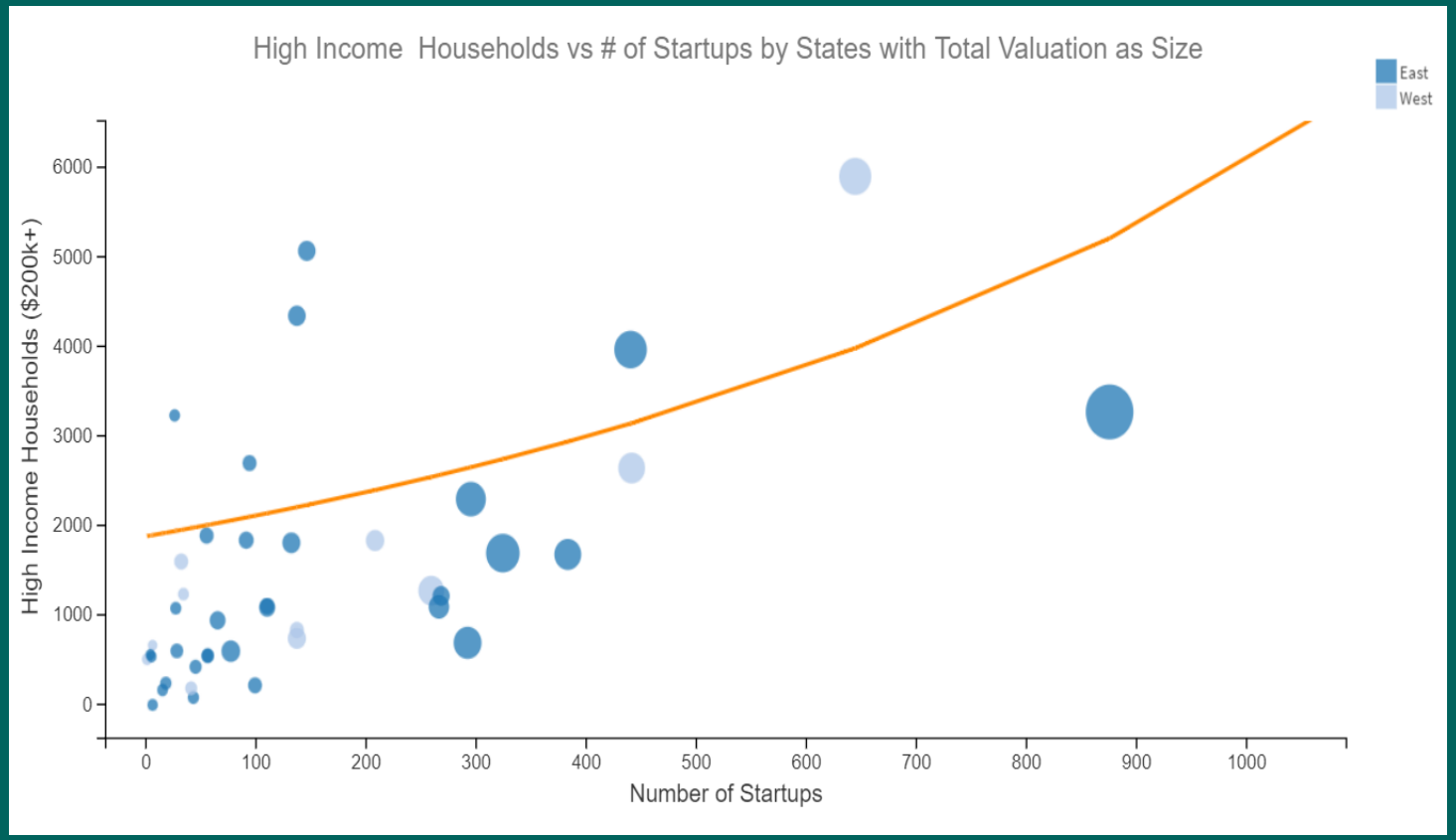
- Number of Startups vs High Income Households (>200k$)
- Regression line in Orange shows overall increasing relation
- Size of bubble shows total valuation by **County**
- High valuation shows more prominent effect on income



Regression Analysis ($R^2 = 0.52$)

Regression Analysis: Quantified relationships ($R^2 = 0.52$)

- Number of Startups vs High Income Households (>200k$)
- Regression line in Orange shows overall increasing relation
- Size of bubble shows total valuation by **state**
- High valuation shows more prominent effect on income



Outcomes

Outcomes

1. **Startup density vs High household income (>200K$)**

   Strong positive correlations, with increasing correlations at broader geographic levels **(r ≈ 0.97 to 0.99)**

2. **Startup density positively correlates with household income**

   Weak positive correlations, with increasing correlations at broader geographic levels **(r ≈ 0.20 to 0.32)**

3. **Startup valuations vs Household Income**

   Weak positive correlation shows quality of startup have influence on high-income households **(r ≈ 0.15 to 0.27)**

4. **Impact is regional rather than purely local**

   Regional spill over

Conclusion

Conclusion

Key Findings:

1. Evidence supporting a **moderate positive relationship** between **startup activity** and **household income** in metropolitan areas
   - The effect strengthens at larger geographic levels (e.g., counties, states)
2. The **weak positive correlations** between **Income & startup valuations,** indicates more successful startups drives higher income in region
3. Metropolitan Areas with high startup activity have **high correlation** with **high income households ($200k and more)**, shows Positive impact
4. The effect is more **pronounced** in areas with established startup ecosystems like San Francisco and New York

Limitations:

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

Thank You