Data Analysis and Visualization

Group 21

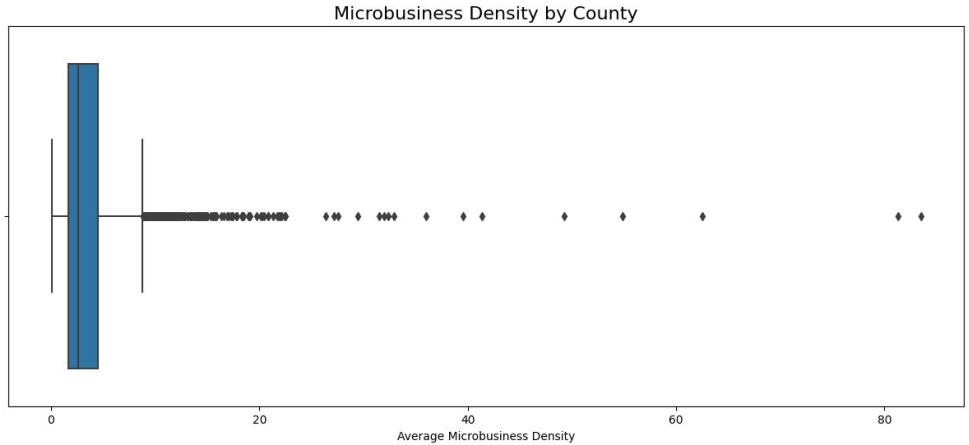
Daniel Clepper, Gaomingyu Fan, Darren Tung, Barry Yao, Julia Mengxuan Yu

Cleaning and sampling

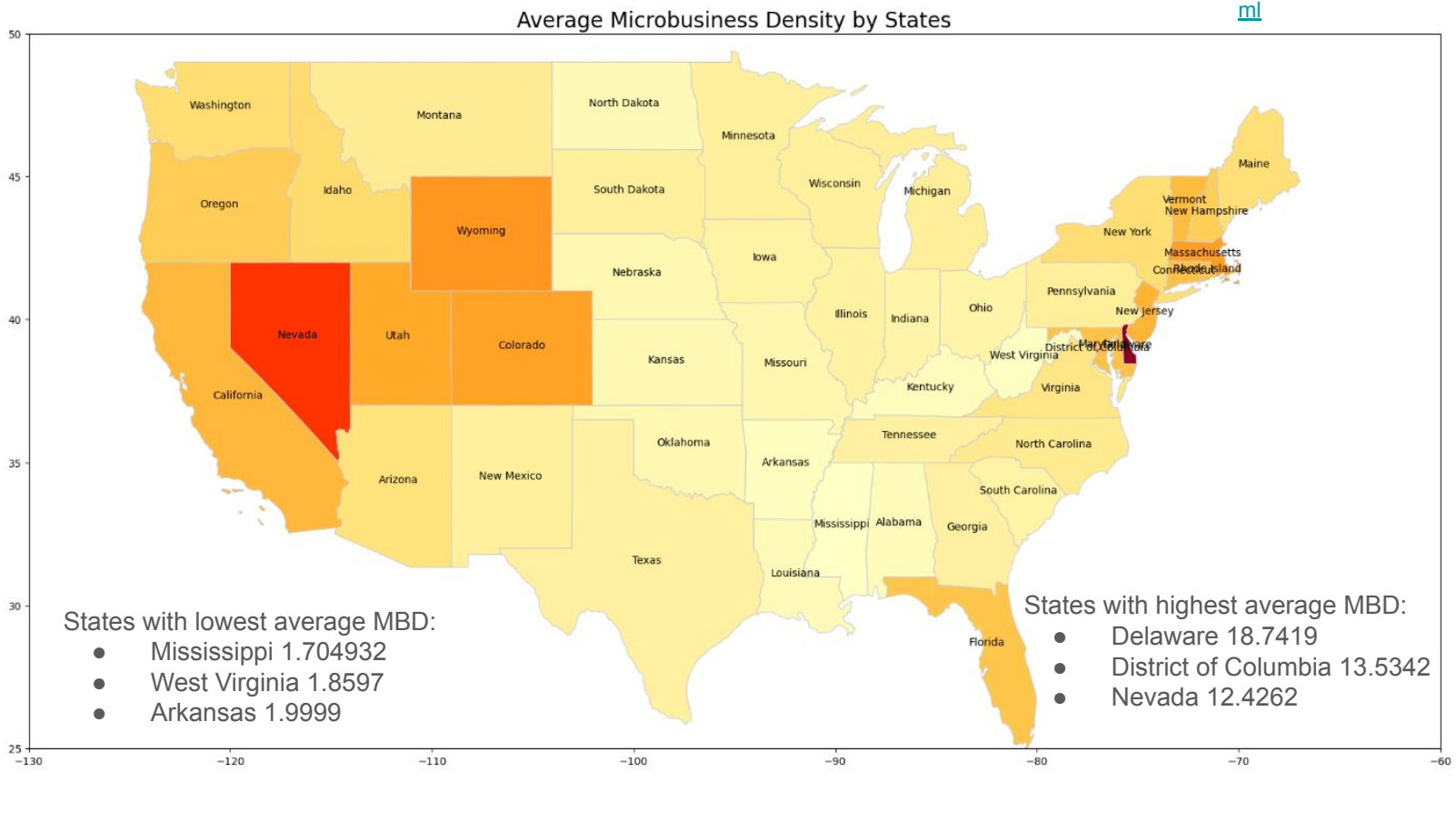
- Microbusiness density(MBD) shape: (128535, 7)
- Census data shape: (3142, 26)
- 0 missing values in MBD data, 14 missing values in Census data
  - Missing values imputed using mean
- 2 year lag
  - Census percentage values according to 2 years before MBD observation date saved as given observation's percentages (i.e. pct_college, pct_foreign_born)
  - Ex. MBD Observation Date 2019 -> corresponding county from pct_bb_2017
- Merged data shape: (128535, 12)
- Features: 'row_id', 'cfips', 'county', 'state', 'first_day_of_month', 'microbusiness_density', 'active', 'pct_bb', 'pct_college', 'pct_foreign_born', 'pct_it_workers', 'median_hh_inc'

Microbusiness Density by County

- Mean Microbusiness_density by county:
  - Min = 0.06
  - Max = 83.48
  - Median = 2.61
  - Q1 = 1.65
  - Q3 = 4.5
- Carson City, Nevada: highest mean Microbusiness_density
- Issaquena County, Mississippi: lowest mean Microbusiness_density
- Many outliers greater than 4.5.
- Skewed Right



Source: https://www.census.gov/geographies/mapping-files/2018/geo/carto-boundary-file.html



States with lowest average MBD:

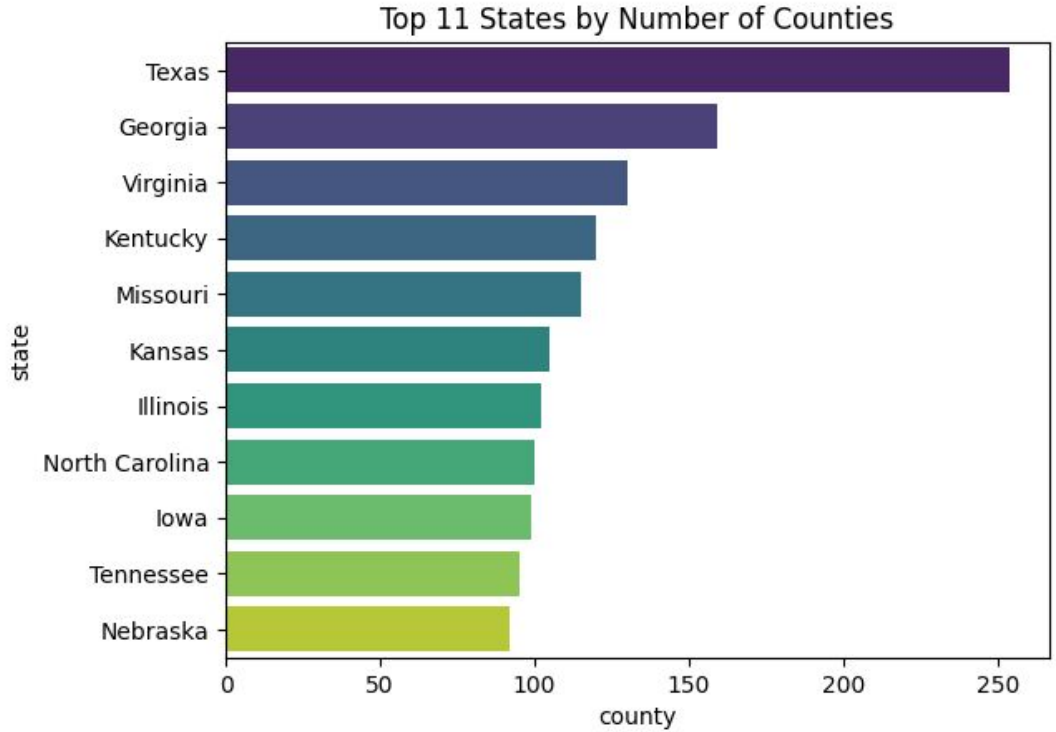
- Mississippi 1.704932
- West Virginia 1.8597
- Arkansas 1.9999

States with highest average MBD:

- Delaware 18.7419
- District of Columbia 13.5342
- Nevada 12.4262

Number of Counties by State

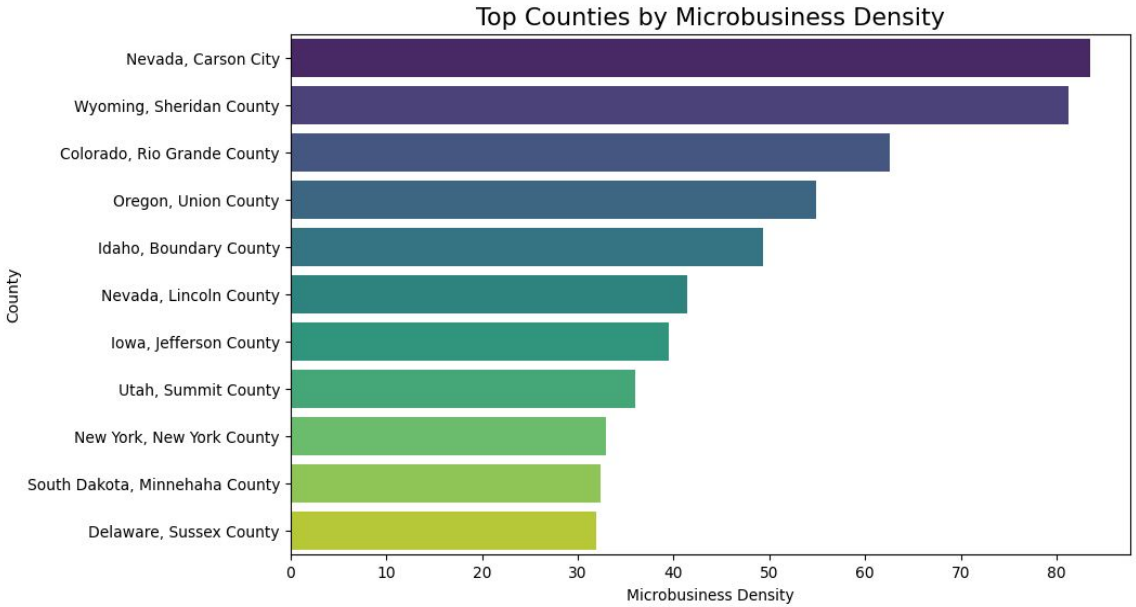
- Texas has most counties (254)
- District of Columbia has fewest counties (1)
- As number of counties increases, microbusiness density decreases
    - Bottom 3 Average MBD States are not inside Top 11



11 States with most counties

Top Counties by Microbusiness Density

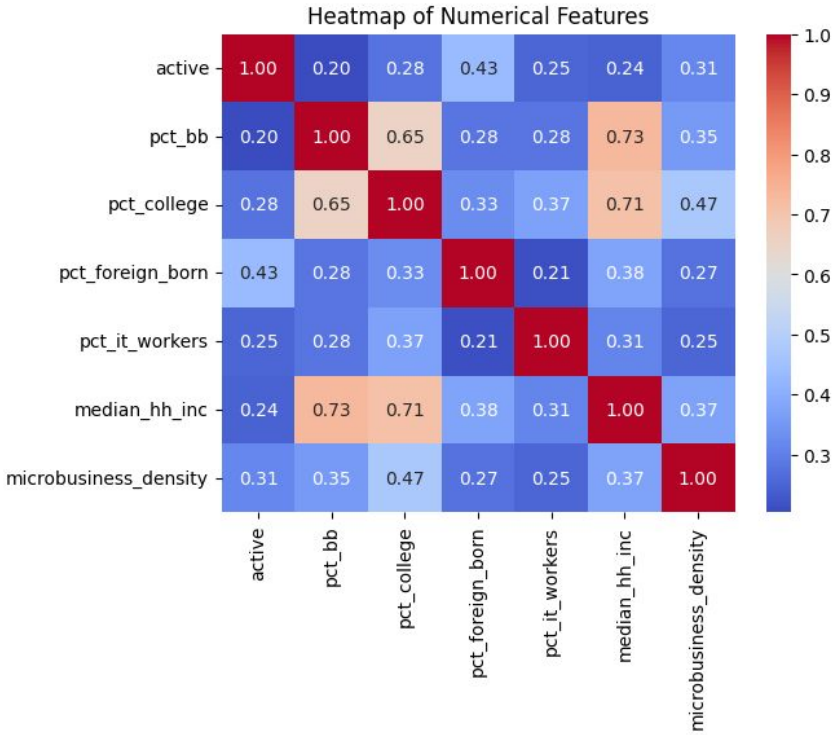
- Extreme MBD values in Carson city and Lincoln country (Nevada) may explain the high ranking of Nevada among states
- Delaware has no county in Top 10 despite being state with highest MBD
    - Sussex County, Delaware: 11th-highest



Top 11 Counties with Highest MBD

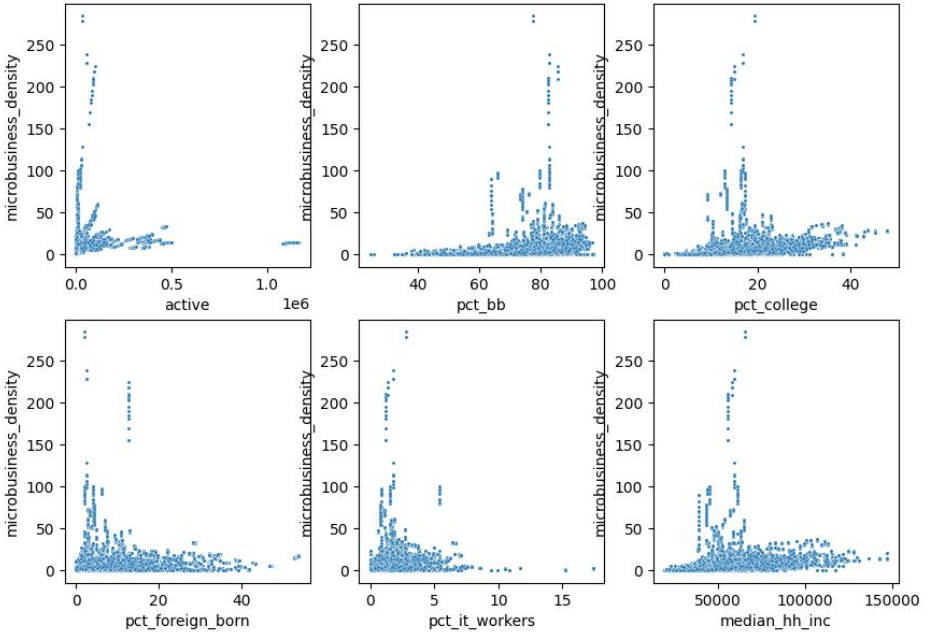
Correlation Heatmap

- (pct_bb, pct_college): r=0.65
- (pct_bb, median_hin_inc) r=0.73
- (pct_college, median_hin_inc) r=0.71
- All variables have positive correlation
- Features are not highly correlated



Scatterplots

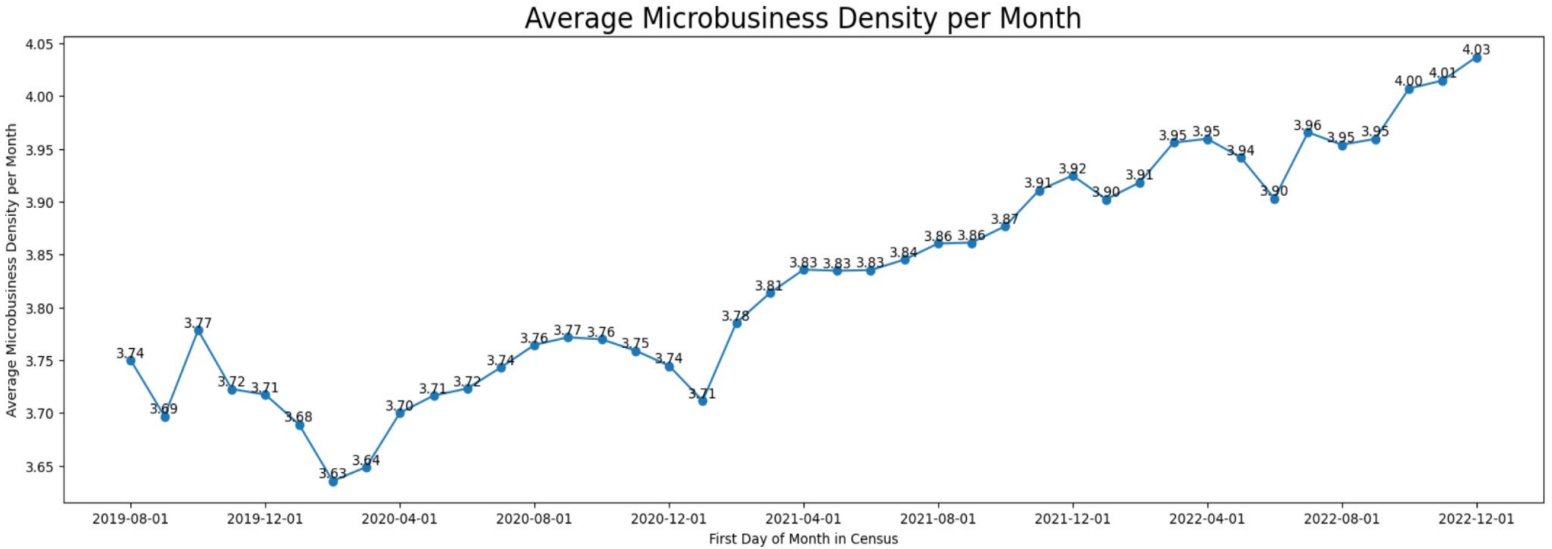
- Heteroscedasticity present in “active”
  - May drop feature before training ML models
- Considerations for outliers
  - For all features, there exists outliers that spikes up significantly
  - Partitioning the counties in ways such as by size or population may be needed



Average Microbusiness Density by Month

The trend of average Microbusiness Density per month:

- From 2019 to 2022, average Microbusiness Density per month increased over time
  - Occasional drops, but not seasonal
- Feb 2020 had the lowest average Microbusiness Density 3.63



Machine Learning techniques proposed to be implemented

- Baseline Model:
- Linear Regression
  - Comparing L1/L2/Elastic-Net
- Gradient Boosting
  - XGBoost
  - Hyperparameter Tuning
- Time Series
  - AR/MA/ARMA/ARIMA
- Evaluation Metric: MAPE
  - Mean Absolute Percentage Error
  - Used over Mean Squared Error because we care about small errors
- Other Techniques
  - Early Stopping
  - Structured Splitting