DECIPHERING MICROBUSINESS DENSITY GROWTH

Data Mining Platforms
Group 5
Mar 6, 2023

AGENDA

01. PROBLEM STATEMENT

- Case Introduction: Background and Purpose

02. DEFINITION OF VARIABLES & EDA

- Dataset Overview: Size, Variables, Missing Values & Outliers
- EDA: Data Distribution, Correlation, and Relationships

03. METHODOLOGY

- Data Impute
- Feature Engineering
- Model Build-Up

04. CONCLUSION & RECOMMENDATION

- Findings
- Business Solution & Recommendations

Problem Statement

- Microbusiness is a significant composition of both the local community and national economy.

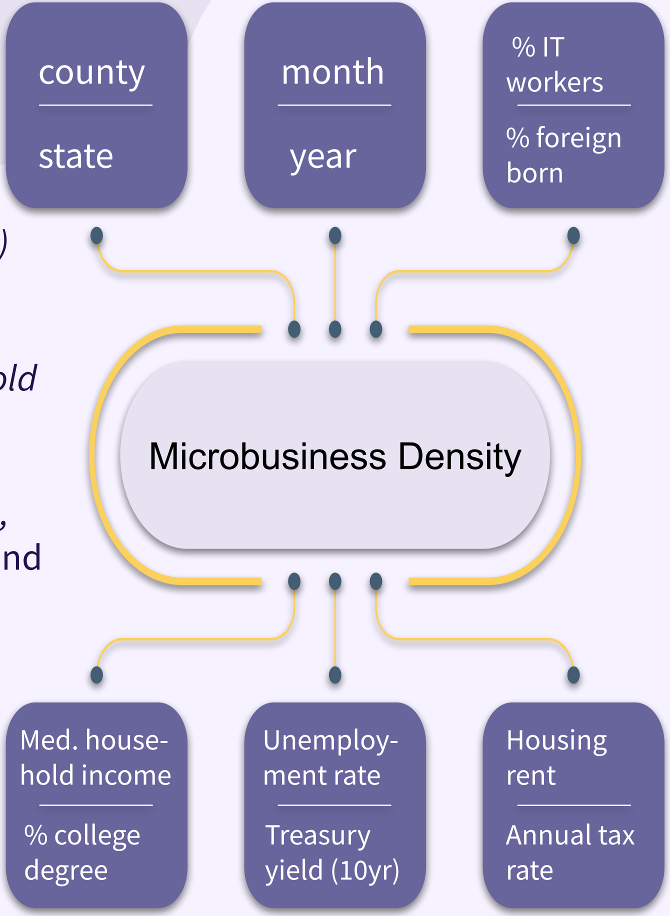
- However, due to the size limitations, it is difficult for stakeholders to track micro-business development, and provide further resources to boost growth.

- By combining data collected by GoDaddy, a domain registrar and web hosting company, and demographic information & macroeconomic indicators, we, as a business consulting team, are here to present our data mining findings and solutions, in order to:
  - ❖ Help policymakers to identify microbusiness pattern and improve resource allocation.
  - ❖ Facilitate investment decisions of institutional investors such as loan providers and venture capitalists.

DEFINITION OF VARIABLES & EDA

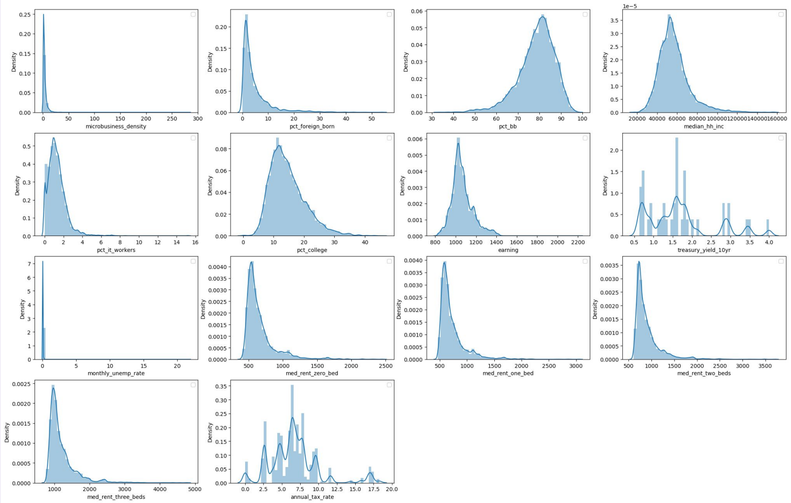
Dataset Overview:

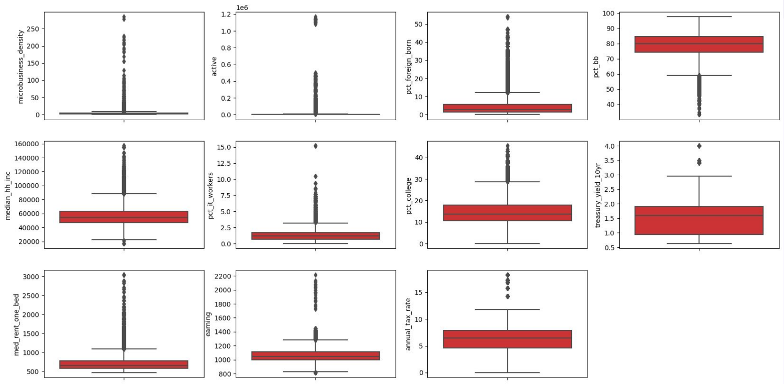
- 122,265 observations in total
- *Microbusiness Density (Microbusiness per 100 ppl over 18 years old)* as target variable
- Over 20 predictor variables, such as *county, date, median household income, housing rents, treasury yield*, etc.
- 90,916 missing values appear in county-level data of *% IT workers, % foreign born, median household income, % broadband access*, and *% college degree*
- Outliers are identified by checking the distribution and interquartile range of variables. Outliers in target variable are scaled, while other outliers are removed or transformed on a case-by-case basis

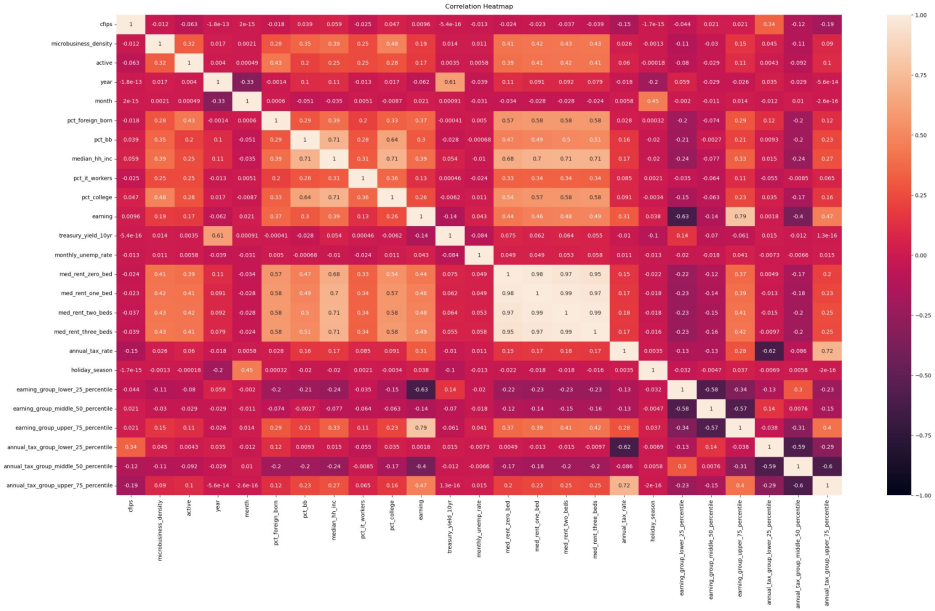


county
state
month
year
% IT workers
% foreign born
Microbusiness Density
Med. household income
% college degree
Unemployment rate
Treasury yield (10yr)
Housing rent
Annual tax rate

DEFINITION OF VARIABLES & EDA



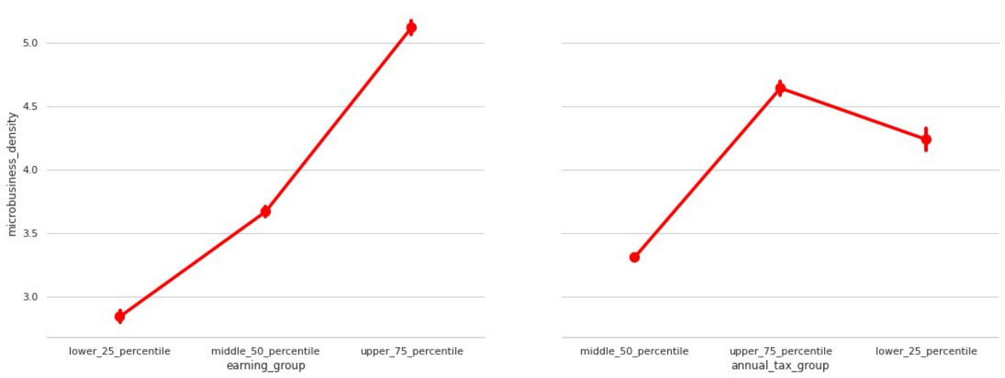



Correlation Heatmap

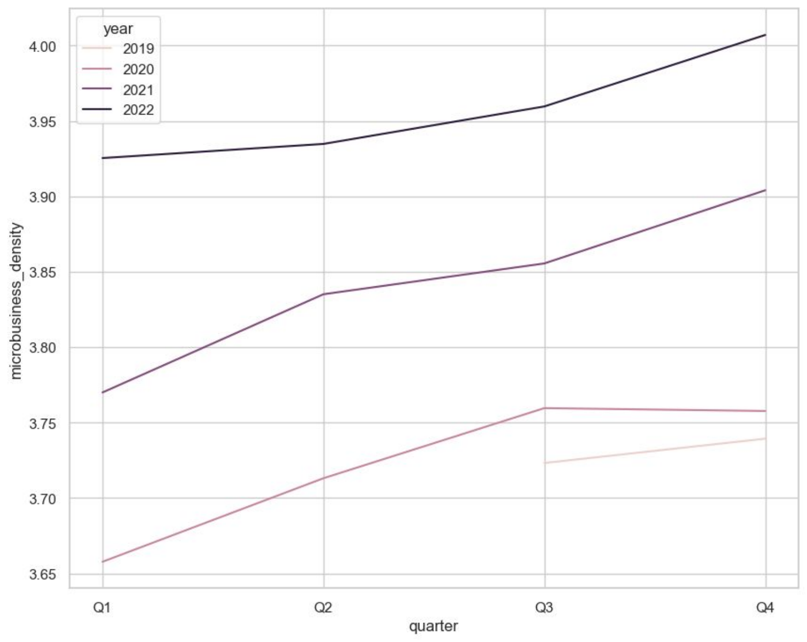
	cfips	microbusiness_density	active	year	month	pct_foreign_born	pct_bb	median_hh_inc	pct_it_workers	pct_college	earning	treasury_yield_10yr	monthly_unemp_rate	med_rent_zero_bed	med_rent_one_bed	med_rent_two_beds	med_rent_three_beds	annual_tax_rate	holiday_season	earning_group_lower_25_percentile	earning_group_middle_50_percentile	earning_group_upper_75_percentile	annual_tax_group_lower_25_percentile	annual_tax_group_middle_50_percentile	annual_tax_group_upper_75_percentile
cfips	1	-0.012	-0.063	-1.8e-13	2e-15	-0.018	0.039	0.059	-0.025	0.047	0.0096	-5.4e-16	-0.013	-0.024	-0.023	-0.037	-0.039	-0.15	-1.7e-15	-0.044	0.021	0.021	0.34	-0.12	-0.19
microbusiness_density	-0.012	1	0.32	0.017	0.0021	0.28	0.35	0.39	0.25	0.48	0.19	0.014	0.011	0.41	0.42	0.43	0.43	0.026	-0.0013	-0.11	-0.03	0.15	0.045	-0.11	0.09
active	-0.063	0.32	1	0.004	0.00049	0.43	0.2	0.25	0.25	0.28	0.17	0.0035	0.0058	0.39	0.41	0.42	0.41	0.06	-0.00018	-0.08	-0.029	0.11	0.0043	-0.092	0.1
year	-1.8e-13	0.017	0.004	1	-0.33	-0.0014	0.1	0.11	-0.013	0.017	-0.062	0.61	-0.039	0.11	0.091	0.092	0.079	-0.018	-0.2	0.059	-0.029	-0.026	0.035	-0.029	-5.6e-14
month	2e-15	0.0021	0.00049	-0.33	1	0.0006	-0.051	-0.035	0.0051	-0.0087	0.021	0.00091	-0.031	-0.034	-0.028	-0.028	-0.024	0.0058	0.45	-0.002	-0.011	0.014	-0.012	0.01	-2.6e-16
pct_foreign_born	-0.018	0.28	0.43	-0.0014	0.0006	1	0.29	0.39	0.2	0.33	0.37	-0.00041	0.005	0.57	0.58	0.58	0.58	0.028	0.00032	-0.2	-0.074	0.29	0.12	-0.2	0.12
pct_bb	0.039	0.35	0.2	0.1	-0.051	0.29	1	0.71	0.28	0.64	0.3	-0.028	-0.00068	0.47	0.49	0.5	0.51	0.16	-0.02	-0.21	-0.0027	0.21	0.0093	-0.2	0.23
median_hh_inc	0.059	0.39	0.25	0.11	-0.035	0.39	0.71	1	0.31	0.71	0.39	0.054	-0.01	0.68	0.7	0.71	0.71	0.17	-0.02	-0.24	-0.077	0.33	0.015	-0.24	0.27
pct_it_workers	-0.025	0.25	0.25	-0.013	0.0051	0.2	0.28	0.31	1	0.36	0.13	0.00046	-0.024	0.33	0.34	0.34	0.34	0.085	0.0021	-0.035	-0.064	0.11	-0.055	-0.0085	0.065
pct_college	0.047	0.48	0.28	0.017	-0.0087	0.33	0.64	0.71	0.36	1	0.26	-0.0062	0.011	0.54	0.57	0.58	0.58	0.091	-0.0034	-0.15	-0.063	0.23	0.035	-0.17	0.16
earning	0.0096	0.19	0.17	-0.062	0.021	0.37	0.3	0.39	0.13	0.26	1	-0.14	0.043	0.44	0.46	0.48	0.49	0.31	0.038	-0.63	-0.14	0.79	0.0018	-0.4	0.47
treasury_yield_10yr	-5.4e-16	0.014	0.0035	0.61	0.00091	-0.00041	-0.028	0.054	0.00046	-0.0062	-0.14	1	-0.084	0.075	0.062	0.064	0.055	-0.01	-0.1	0.14	-0.07	-0.061	0.015	-0.012	1.3e-16
monthly_unemp_rate	-0.013	0.011	0.0058	-0.039	-0.031	0.005	-0.00068	-0.01	-0.024	0.011	0.043	-0.084	1	0.049	0.049	0.053	0.058	0.011	-0.013	-0.02	-0.018	0.041	-0.0073	-0.0066	0.015
med_rent_zero_bed	-0.024	0.41	0.39	0.11	-0.034	0.57	0.47	0.68	0.33	0.54	0.44	0.075	0.049	1	0.98	0.97	0.95	0.15	-0.022	-0.22	-0.12	0.37	0.0049	-0.17	0.2
med_rent_one_bed	-0.023	0.42	0.41	0.091	-0.028	0.58	0.49	0.7	0.34	0.57	0.46	0.062	0.049	0.98	1	0.99	0.97	0.17	-0.018	-0.23	-0.14	0.39	-0.013	-0.18	0.23
med_rent_two_beds	-0.037	0.43	0.42	0.092	-0.028	0.58	0.5	0.71	0.34	0.58	0.48	0.064	0.053	0.97	0.99	1	0.99	0.18	-0.018	-0.23	-0.15	0.41	-0.015	-0.2	0.25
med_rent_three_beds	-0.039	0.43	0.41	0.079	-0.024	0.58	0.51	0.71	0.34	0.58	0.49	0.055	0.058	0.95	0.97	0.99	1	0.17	-0.016	-0.23	-0.16	0.42	-0.0097	-0.2	0.25
annual_tax_rate	-0.15	0.026	0.06	-0.018	0.0058	0.028	0.16	0.17	0.085	0.091	0.31	-0.01	0.011	0.15	0.17	0.18	0.17	1	0.0035	-0.13	-0.13	0.28	-0.62	-0.086	0.72
holiday_season	-1.7e-15	-0.0013	-0.00018	-0.2	0.45	0.00032	-0.02	-0.02	0.0021	-0.0034	0.038	-0.1	-0.013	-0.022	-0.018	-0.018	-0.016	0.0035	1	-0.032	-0.0047	0.037	-0.0069	0.0058	-2e-16
earning_group_lower_25_percentile	-0.044	-0.11	-0.08	0.059	-0.002	-0.2	-0.21	-0.24	-0.035	-0.15	-0.63	0.14	-0.02	-0.22	-0.23	-0.23	-0.23	-0.13	-0.032	1	-0.58	-0.34	-0.13	0.3	-0.23
earning_group_middle_50_percentile	0.021	-0.03	-0.029	-0.029	-0.011	-0.074	-0.0027	-0.077	-0.064	-0.063	-0.14	-0.07	-0.018	-0.12	-0.14	-0.15	-0.16	-0.13	-0.0047	-0.58	1	-0.57	0.14	0.0076	-0.15
earning_group_upper_75_percentile	0.021	0.15	0.11	-0.026	0.014	0.29	0.21	0.33	0.11	0.23	0.79	-0.061	0.041	0.37	0.39	0.41	0.42	0.28	0.037	-0.34	-0.57	1	-0.038	-0.31	0.4
annual_tax_group_lower_25_percentile	0.34	0.045	0.0043	0.035	-0.012	0.12	0.0093	0.015	-0.055	0.035	0.0018	0.015	-0.0073	0.0049	-0.013	-0.015	-0.0097	-0.62	-0.0069	-0.13	0.14	-0.038	1	-0.59	-0.29
annual_tax_group_middle_50_percentile	-0.12	-0.11	-0.092	-0.029	0.01	-0.2	-0.2	-0.24	-0.0085	-0.17	-0.4	-0.012	-0.0066	-0.17	-0.18	-0.2	-0.2	-0.086	0.0058	0.3	0.0076	-0.31	-0.59	1	-0.6
annual_tax_group_upper_75_percentile	-0.19	0.09	0.1	-5.6e-14	-2.6e-16	0.12	0.23	0.27	0.065	0.16	0.47	1.3e-16	0.015	0.2	0.23	0.25	0.25	0.72	-2e-16	-0.23	-0.15	0.4	-0.29	-0.6	1

- Majority of continuous variables are right-skewed
- Box plots also show both our target variable and features have many outliers
- Microbusiness_density has relatively high correlation with broadband internet access percentage, household income median, college rate and housing rents.

DEFINITION OF VARIABLES & EDA

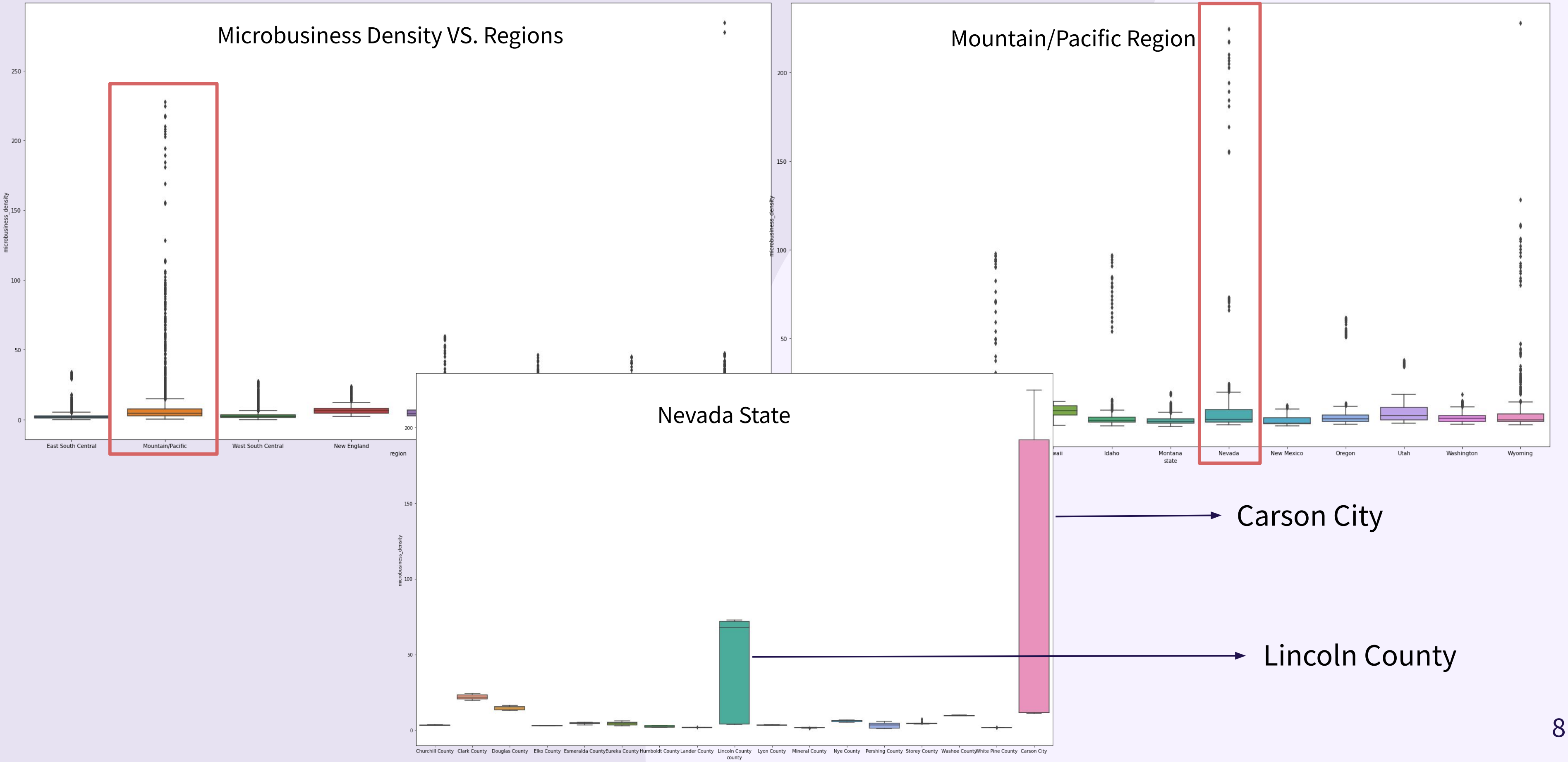


- Microbusiness density increases over the years
- Q4 usually has the highest density



- Microbusiness density is positively related to the average household income.
- High microbusiness density only occurs in region with tax that is either higher than 75 percentile or lower than 25 percentile.

DEFINITION OF VARIABLES & EDA


Microbusiness Density VS. Regions

Mountain/Pacific Region

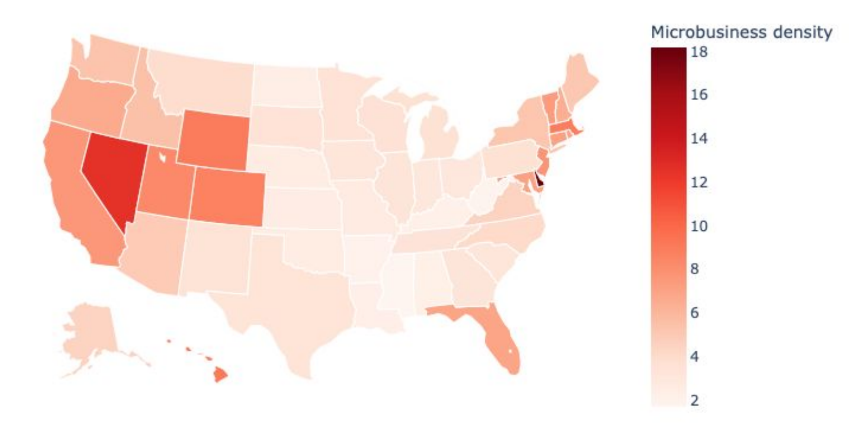
Nevada State

Carson City

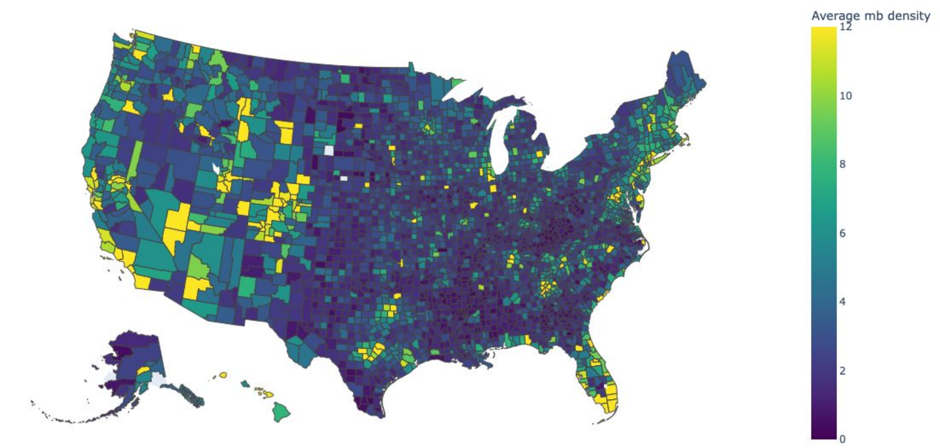
Lincoln County

EDA - GIS VISUALIZATION

- **State-wise**, we see that the East Coast and the North-West Coast of the U.S. have higher micro-business densities than others
- **Top three counties** with highest average microbusiness density are: Carson City(NV), Sheridan County(WY), Rio Grande County(CO). However, their trends of micro-business density differ (e.g. Carson City's short peak)
- **Bottom three counties** with lowest average microbusiness density are: Issaquena County(MS), Echols County(GA), and Greensville County(VA)
- With basic census information, we see counties with highest micro-business density has an above-average level of access to the internet, percentage of foreign born people, and median household income; the bottom 1 county has the opposite



Average Micro-Business Density by State



Average Micro-Business Density by County

PREPROCESSING & FEATURE ENGINEERING

1

CENSUS DATA

Interpolated monthly data— percentage foreign-born, percentage broadband access, percentage college graduates, using yearly-updated values

2

IMPUTATION

Missing values were thoroughly investigated and imputed appropriately given the data type, descriptive statistics, and considerations of downstream analytical needs. All imputations and rationales are summarized in the appendix.

3

ECONOMIC INDICATORS

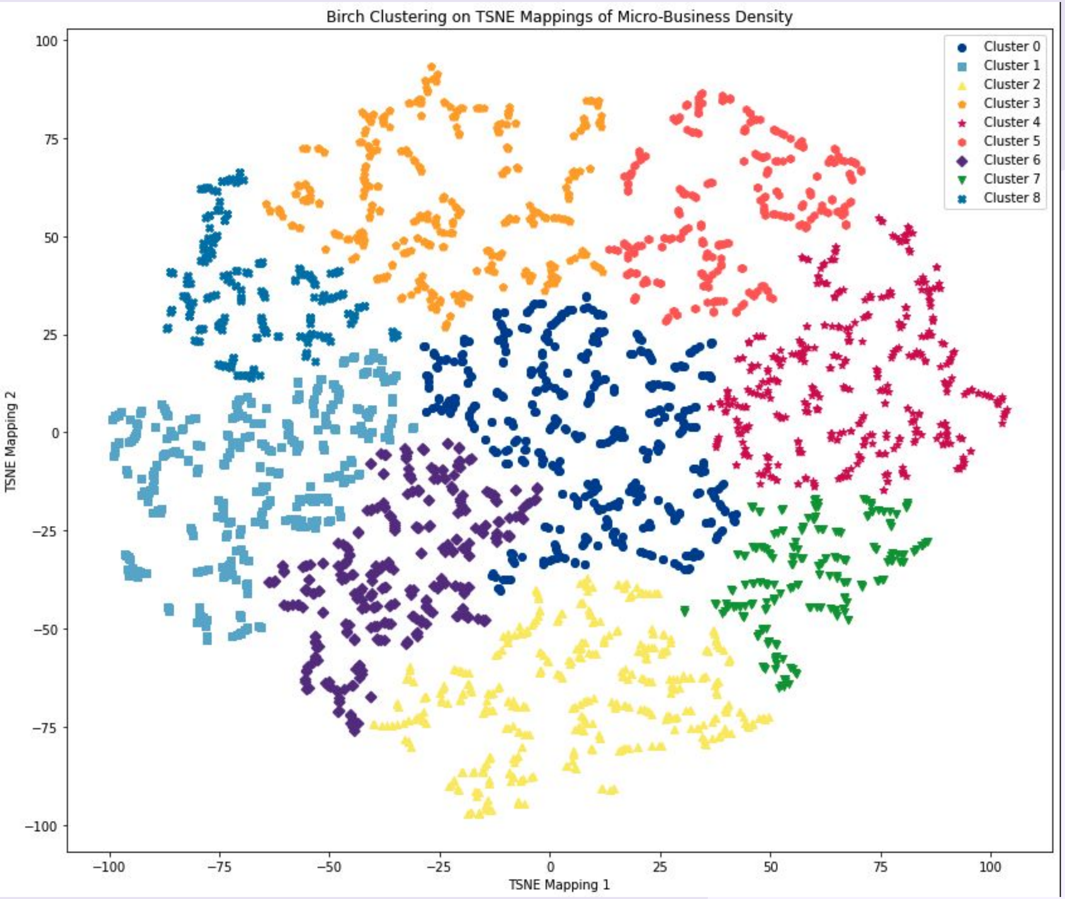
Included new variables like earnings, rent, U.S. treasury bill rates, unemployment rate, tax rates, etc. from external sources

4

CATEGORICAL VARS

Engineered categorical variables for regionality, fiscal quarters, seasonality, covid, and binned continuous features

CLUSTERING WITH BIRCH

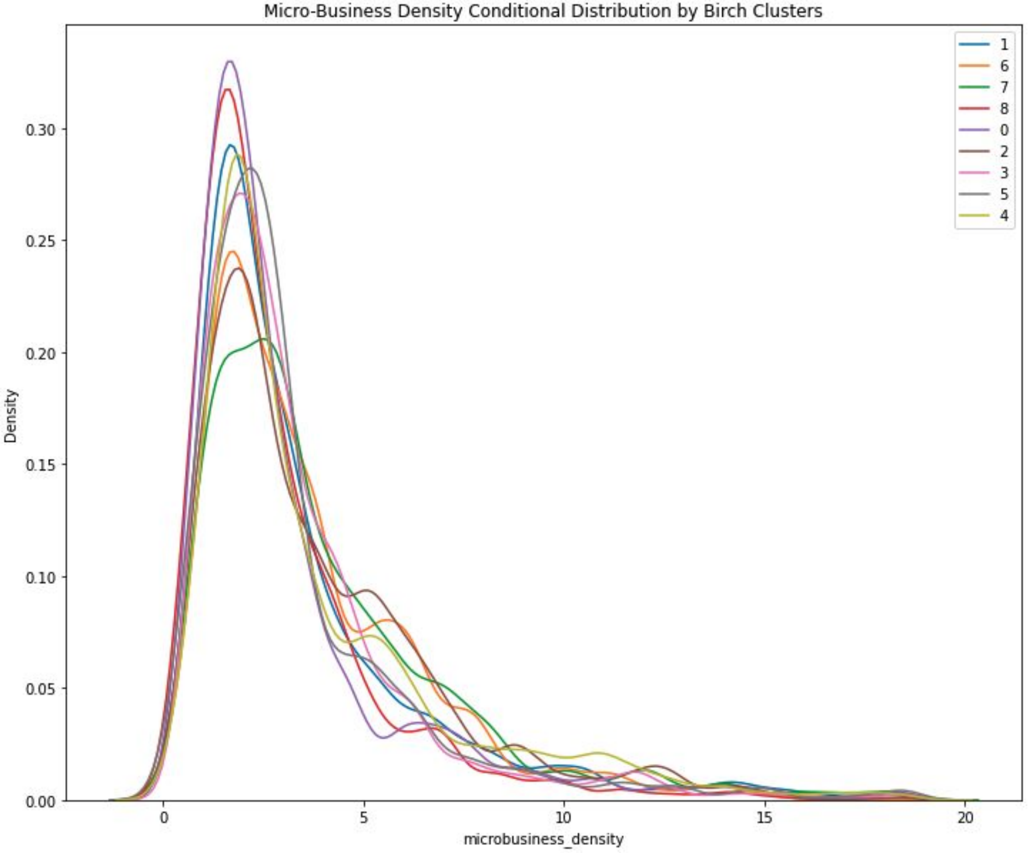


- Number of clusters selected based on distortion score and Calinski-Harabasz Index (higher = better defined clusters) recommended by the literature.

- Visualization uses TSNE with tuned perplexity and learning rate, and initialized using PCA.

- This clustering model is compared with Optics, HDBSCAN, and direct Agglomerative Clustering; the most sensible model is selected based on business interpretation needs that follow— i.e., trade-off between too many clusters and too few.

CLUSTERING EVALUATION

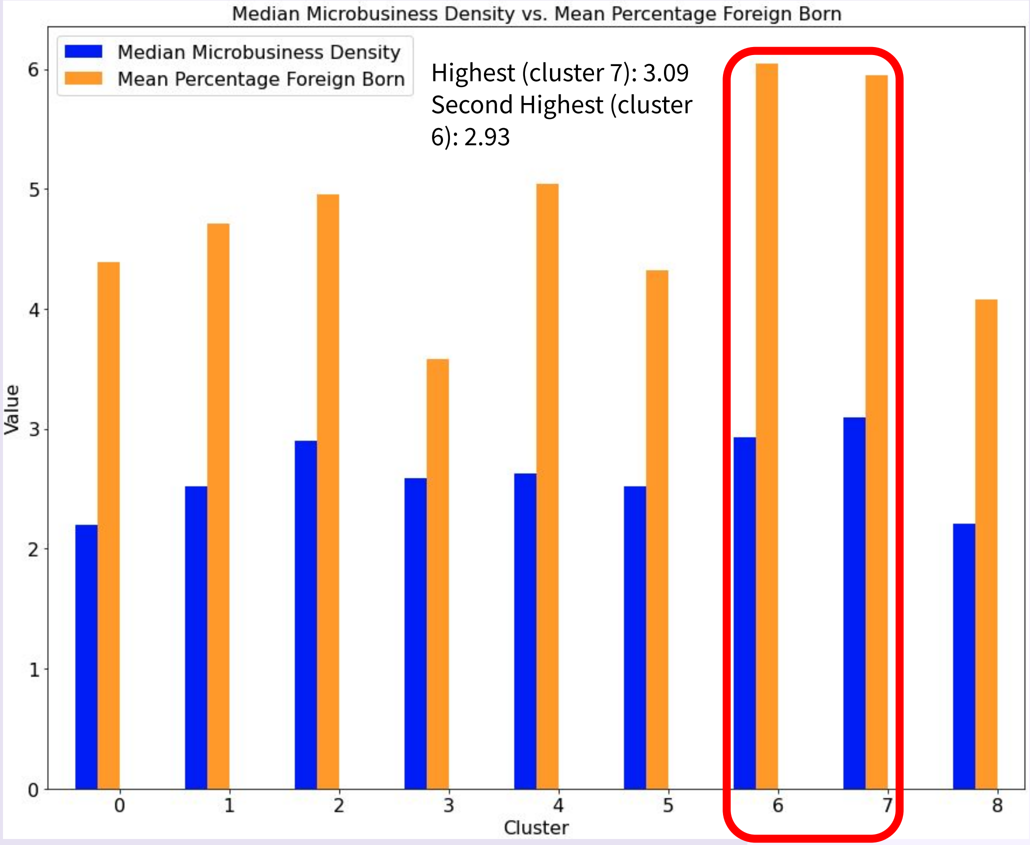


- **Business Question**: Is the chosen clustering model useful for segmenting and capturing the differences in the distributions of micro-business densities to aid business interpretations?

- **Approach**: We use KDE to estimate the conditional densities and conduct pairwise Kolmogorov-Smirnov test on simulated samples. All conditional distributions are found to have distributions that are **not** identical.

Test results are included in the notebooks.

CLUSTER CHARACTERISTICS

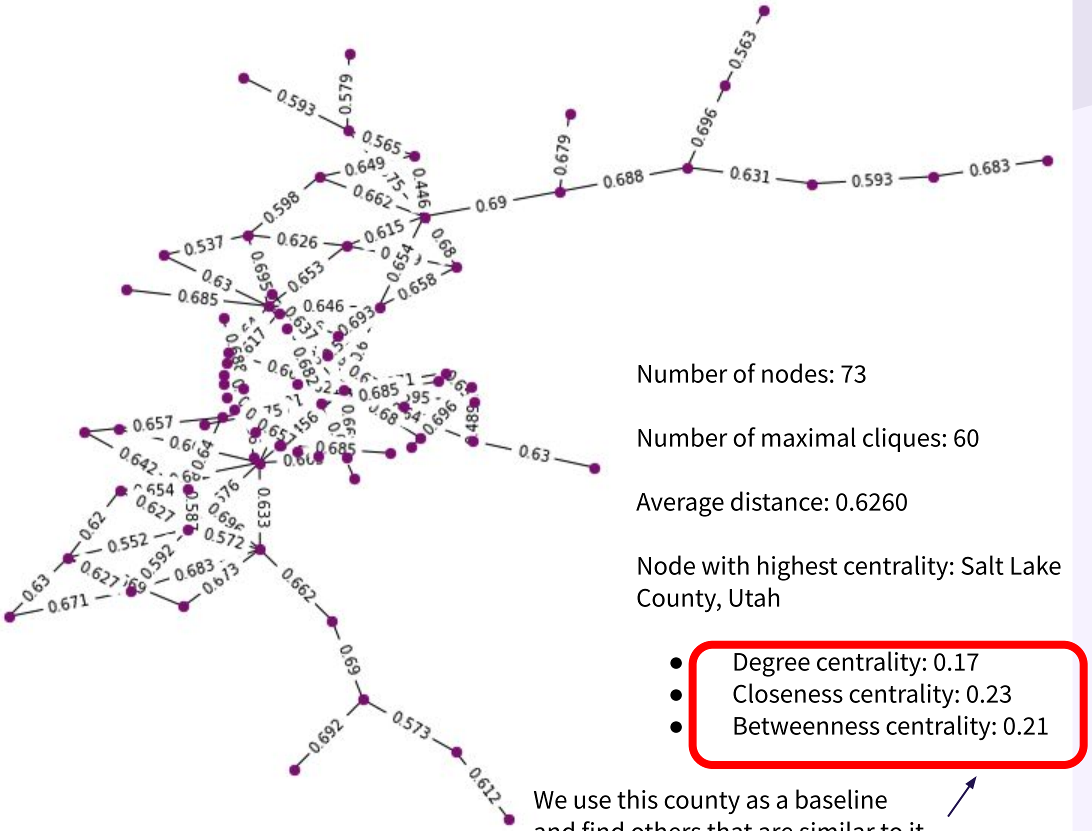


Highest (cluster 7): 3.09
Second Highest (cluster 6): 2.93

Key Characteristics of High Micro-business Density Clusters

- The top two clusters with the **highest median densities** are also **top two in percentage foreign born** ~ 6 % and 5.9% respectively.
- The top three clusters with the highest densities are also **top three in median household income** ~ 57,806 - 61,848.
- The cluster with the highest median density is the cluster that has the **highest percentage of college graduates** at 16.35 %.

CLUSTER ANALYSIS (GRAPH NETWORK)



Number of nodes: 73

Number of maximal cliques: 60

Average distance: 0.6260

Node with highest centrality: Salt Lake County, Utah

- Degree centrality: 0.17
- Closeness centrality: 0.23
- Betweenness centrality: 0.21

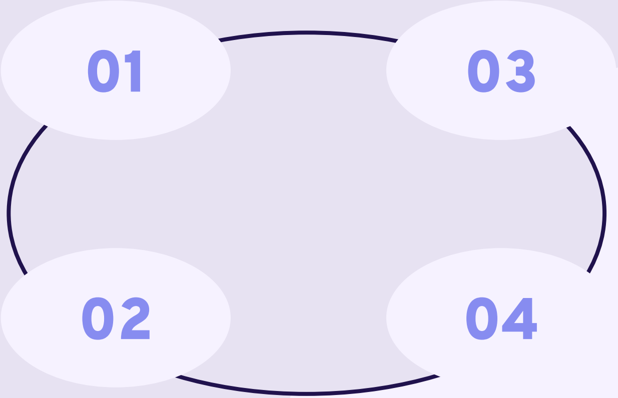
We use this county as a baseline and find others that are similar to it

- **Business Question**: What do high micro-business density counties look like? What attributes do they share in common?

- **Approach**:

  For cluster 7 (highest median micro-business density), construct a network.
  - Each node is a county
  - The edge/link attributes are the euclidean distances between pairs of counties
  - A link between a pair of counties exists if their euclidean distance is below a threshold (0.7)

CLUSTER ANALYSIS (NEAREST NEIGHBOR)



01

02

03

04

Regionality

8 of 11 (72 %) of these counties are in mountain/pacific coastal states.

Broadband Access

All 11 counties have > 85% (6 of them have > 90%) of population with broadband access compared to 78% average for out-of-cluster counties.

Favorable Tax Policy

The average annual marginal corporate tax rate is 5.4 % compared to 6.6 % for out-of-cluster counties.

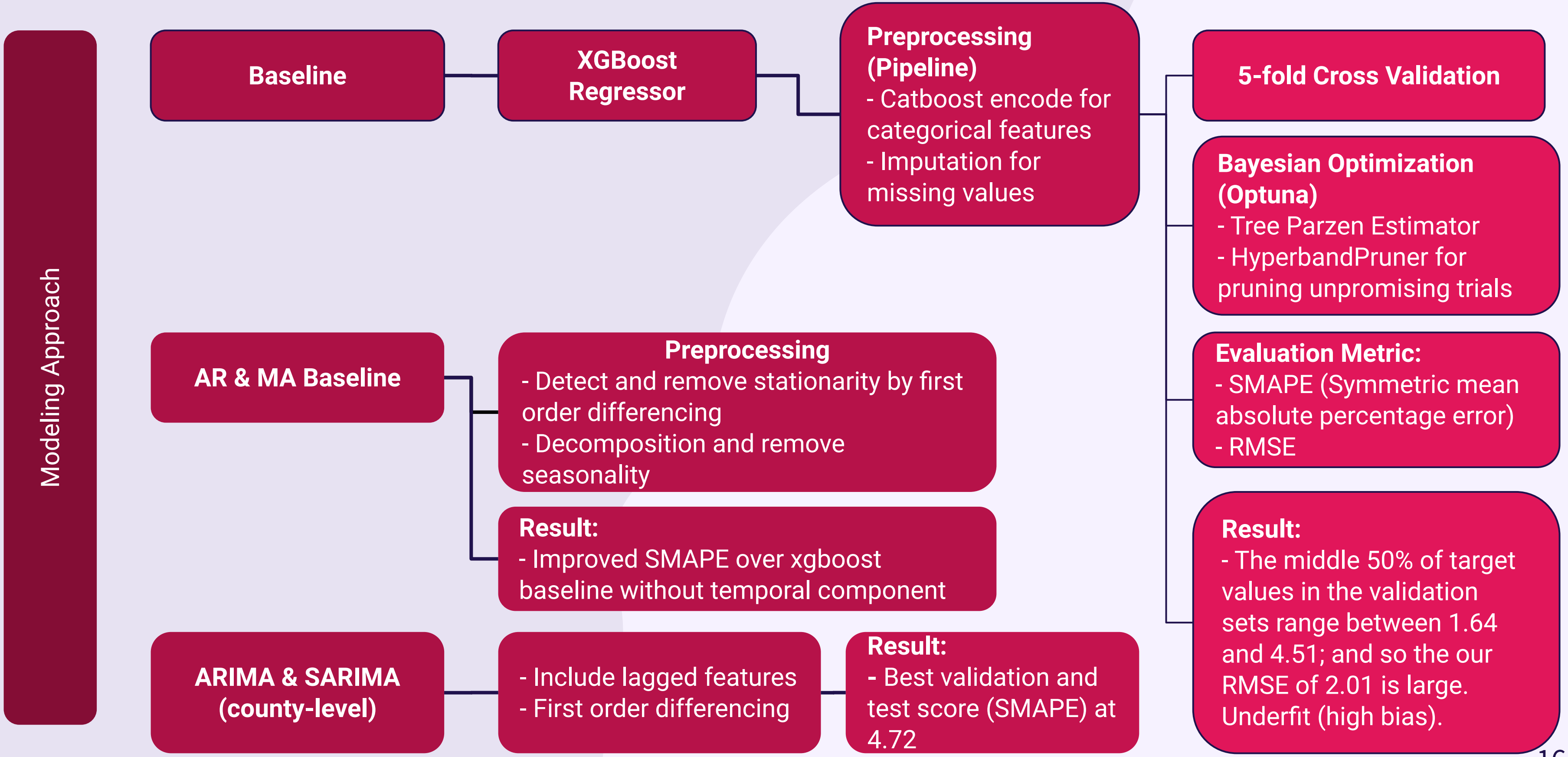
Socio-Economic

- **Unemployment**: 4% v.s. 6% (out-of-cluster)
- **Education**: 26% v.s. 14% (out-of-cluster)
- **Foreign-born**: 8% v.s. 3% (out-of-cluster)

10 Nearest Neighbors of Salt Lake County

County	Marin, CA	Placer, CA	El Paso, CO	Pitikin, CO	San Miguel, CO	Canyon, ID	Douglas, KS	Davis, UT	Fauquier, VA	Winnebago, WI
Closest City	SF	Sacramento	CO Springs	Aspen	Telluride	Boise	Lawrence	SLC	Warrenton	Oshkosh

PREDICTIVE MODELING FRAMEWORK



Modeling Approach

Baseline

XGBoost Regressor

Preprocessing (Pipeline)
- Catboost encode for categorical features
- Imputation for missing values

5-fold Cross Validation

Bayesian Optimization (Optuna)
- Tree Parzen Estimator
- HyperbandPruner for pruning unpromising trials

Evaluation Metric:
- SMAPE (Symmetric mean absolute percentage error)
- RMSE

Result:
- The middle 50% of target values in the validation sets range between 1.64 and 4.51; and so the our RMSE of 2.01 is large. Underfit (high bias).

AR & MA Baseline

Preprocessing
- Detect and remove stationarity by first order differencing
- Decomposition and remove seasonality

Result:
- Improved SMAPE over xgboost baseline without temporal component

ARIMA & SARIMA (county-level)

- Include lagged features
- First order differencing

Result:
- Best validation and test score (SMAPE) at 4.72

CONCLUSION

East, West Coast and Rockies

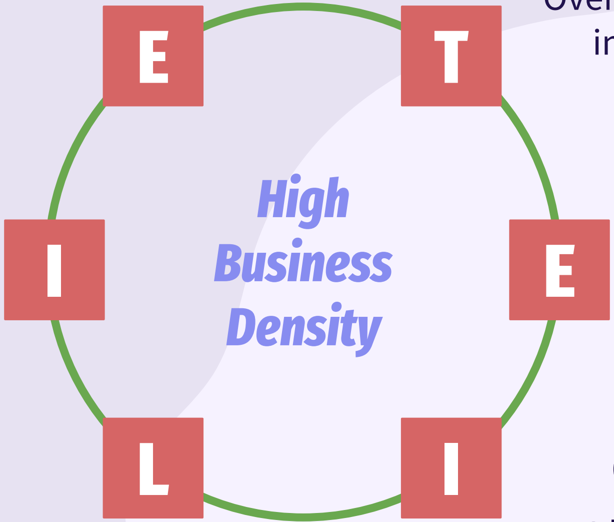
Generally, these counties have higher microbusiness density than others

Infrastructure

Access to Internet are crucial for micro-business growth

Labor Force

High foreign-born workers and low unemployment rate helps boost business growth



E

T

I

High Business Density

E

L

I

Trend

Overall, micro-businesses in the U.S. are growing

Education

College education might foster entrepreneurship and business acumen

Geoeconomic Characteristics

Counties with tourism resources or suburban areas bring more opportunities

RECOMMENDATIONS

1 **Government Entities which wish to boost local microbusiness**

Improving broadband access, investing in workforce education of IT & digital literacy, providing favorable tax policy and labor regulations

2 **Private Investors & financial institutions**

Increasing access to venture capital and financing for targeted areas, such as providing micro-lending programs, community development financial institutions (CDFIs), or other financing options

3 **Services Providers of microbusiness**

Investing more resources on areas with high probability of micro-business growth, such as discount program for company domain sales, marketing tools (see actionable examples in appendix)

4 **Non-profit Organizations**

Facilitating local marginal groups through trainings (e.g. language, digital skills) to secure employment in micro-business and gig-economy

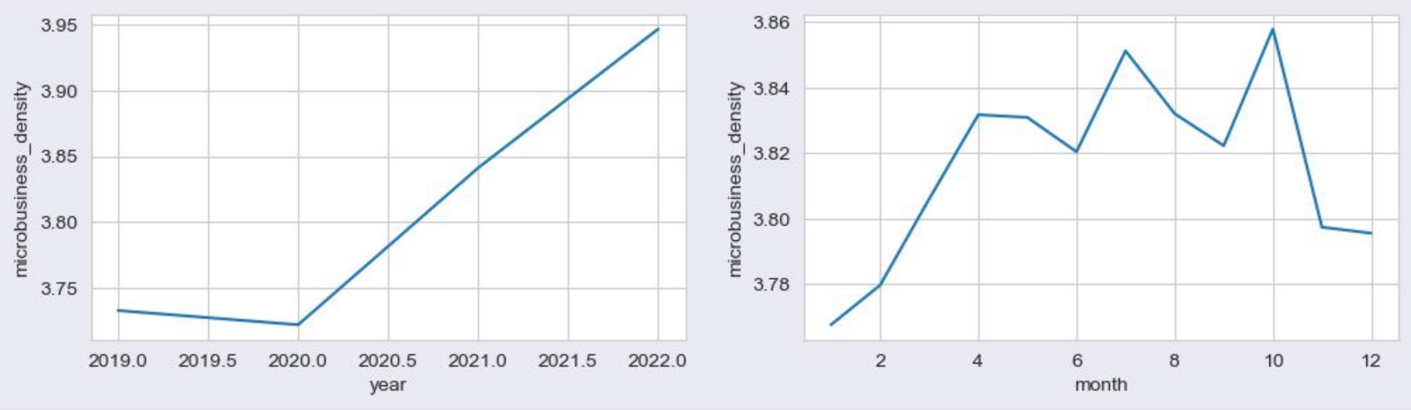
Appendix - Imputation

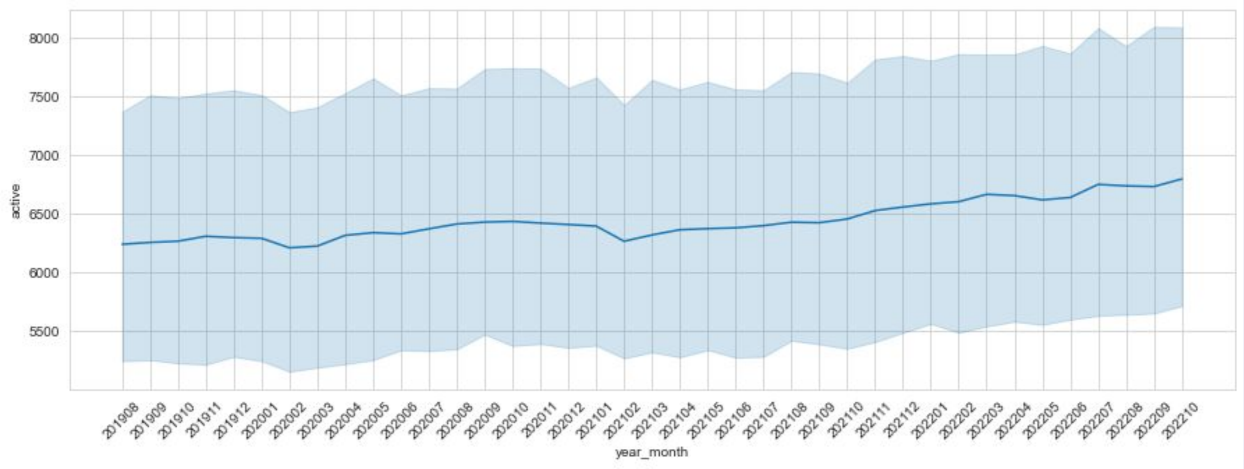
- Missing values in some census data features are due to the fact that there are years covered in the training data when these features are not available in the census data— e.g., 2022. For these features, our imputation strategy can be summarized as follows:
  - For each unique value of ‘cfips’ (county ID), we use the median of the previous three years of data to impute the values for year 2022 *if the year-over-year changes fluctuate a great deal.*
  - For each unique value of ‘cfips’, we use last year’s data (2021) to impute the values for year 2022 *if the year-over-year changes are not significantly different*.

- There is one county in the training set in which unemployment data is not available— Kalawao County, Hawaii. For these missing values, we impute using the state-level monthly unemployment rate as proxies of the county-level unemployment rates.

All above steps are detailed and explained in a single Jupyter notebook.

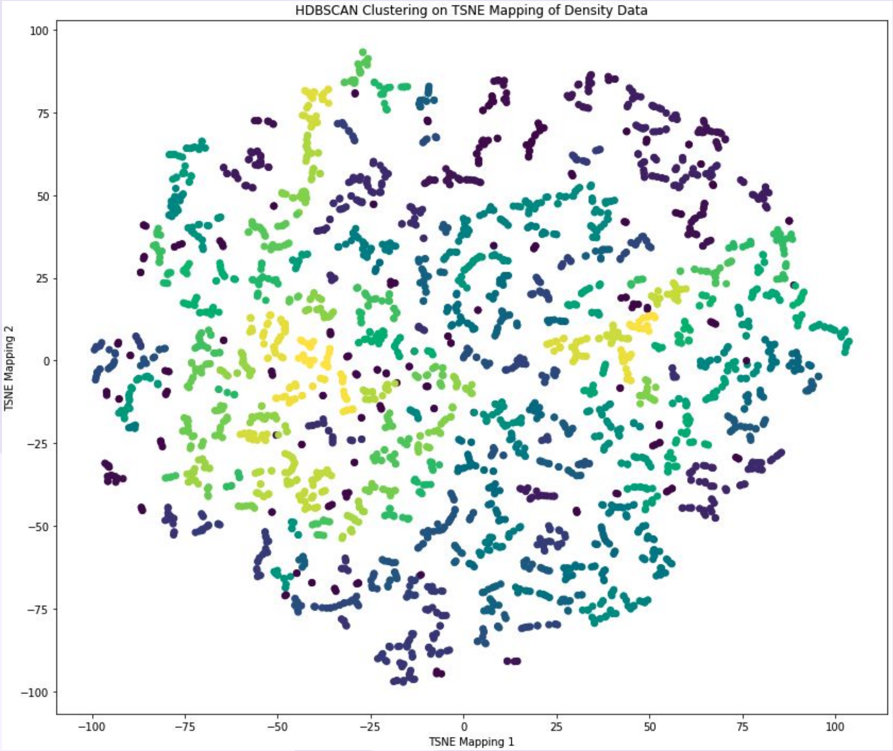
Appendix - EDA

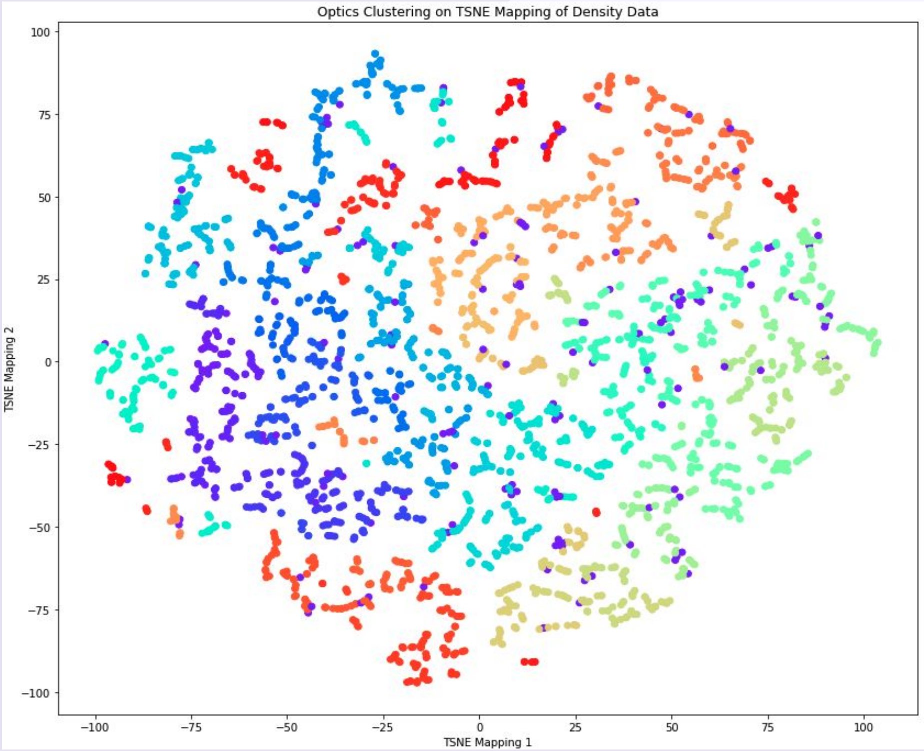




- Microbusiness_density increases over the years except a very small drop in 2022
- May have some seasonality patterns, further explore that by adding new features ‘quarter’ and ‘holiday_season’

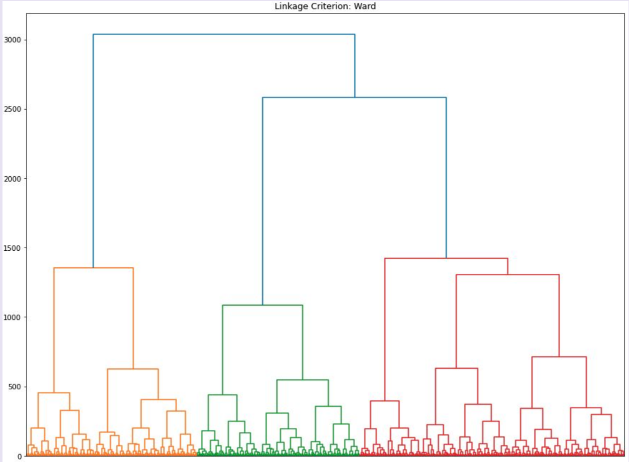
Appendix - Clustering Models

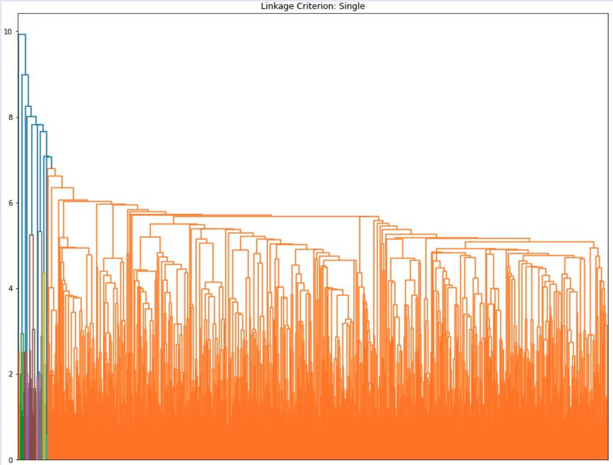


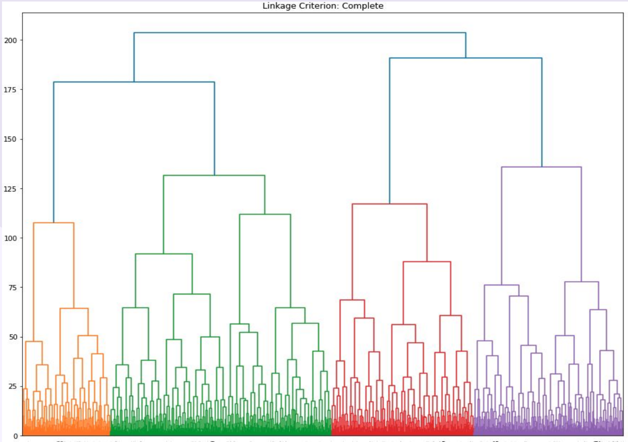


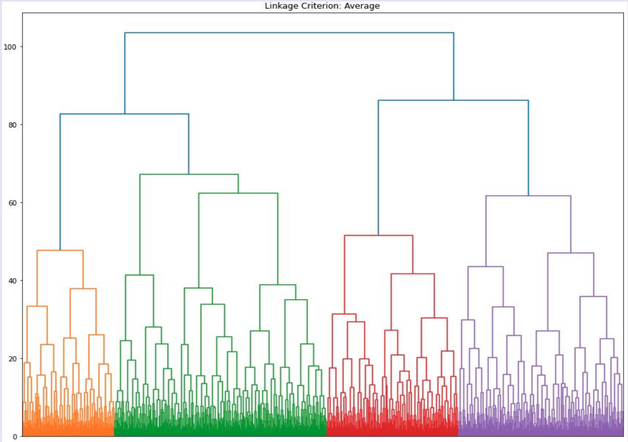
HDSCAN and Optics resulted in many clusters (> 500 and > 1000), which may not be conducive to business interpretations..

Appendix - Clustering Models


Linkage Criterion: Ward


Linkage Criterion: Single

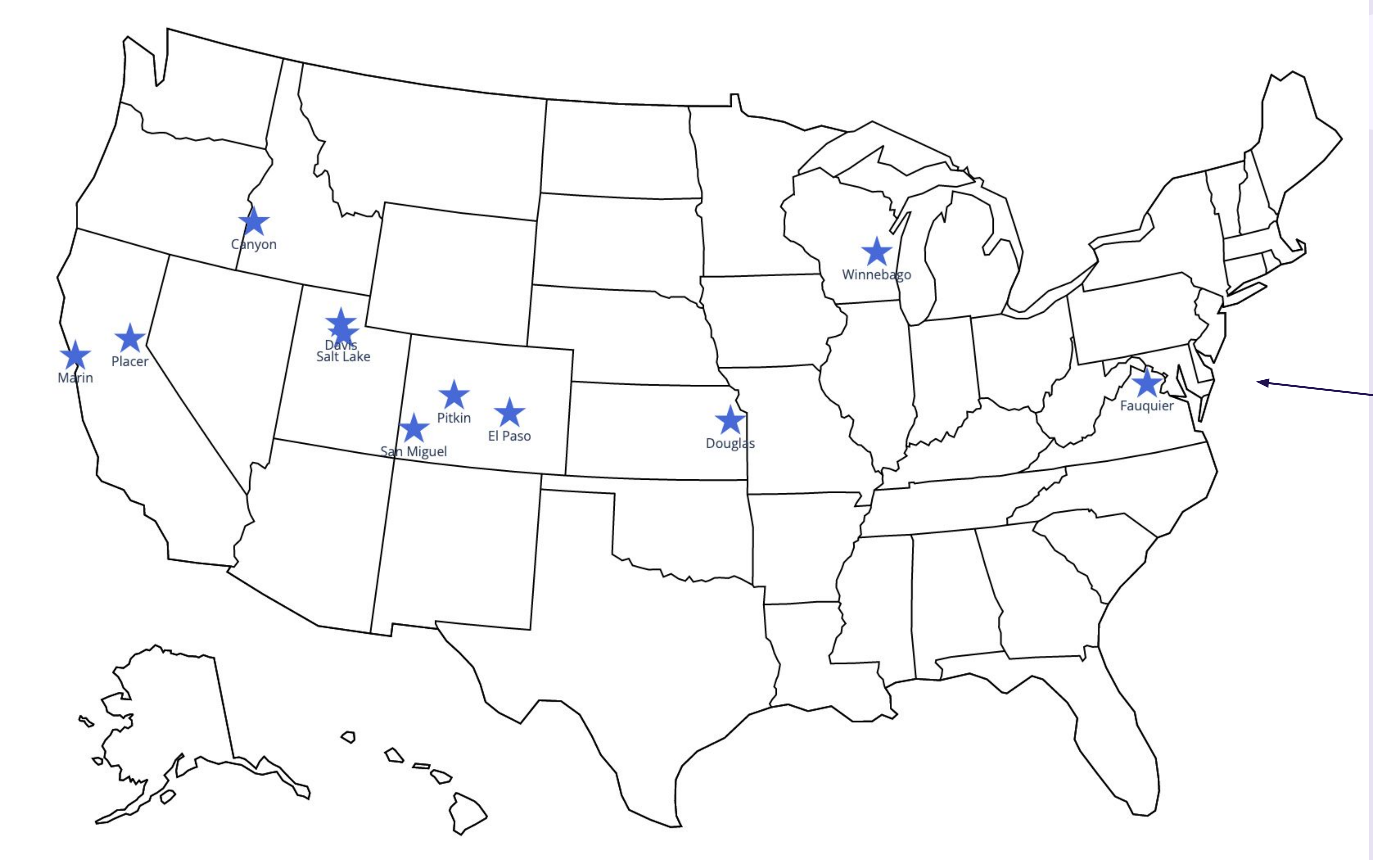

Linkage Criterion: Complete


Linkage Criterion: Average

Agglomerative clustering resulting in the fewest number of clusters among all clustering models.

While the number of clusters are manageable, it may note be useful for segmenting and capturing the differences between micro-business density

Appendix - KNN MAP

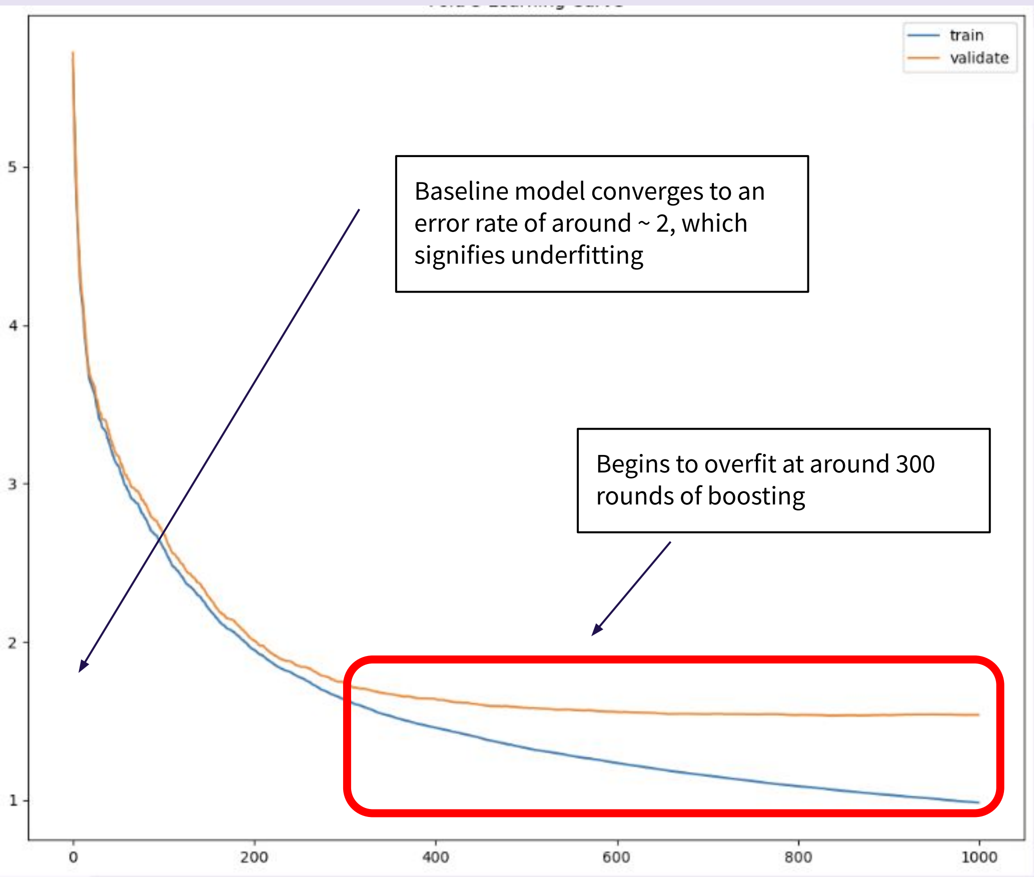


These are similar (neighbor) counties belonging to cluster 7, which has the highest median micro-business density.

Appendix - Micro-Business Programs & Services

- Place County Micro Biz Grants
- Davis County Community & Economic Development— Davis Fund
- Marin County Micro-Business Grant
- Pitikin County Small Business Resource Center
- San Miguel County Small Business Resource Center
- Douglas County E-Community Loan Program, Youth Entrepreneurship Challenge, E-Accelerator Program
- Fauqueir County Micro-Loan Program
- Winnebago County Small Business Development Program

Appendix - Baseline Modeling Learning Curve

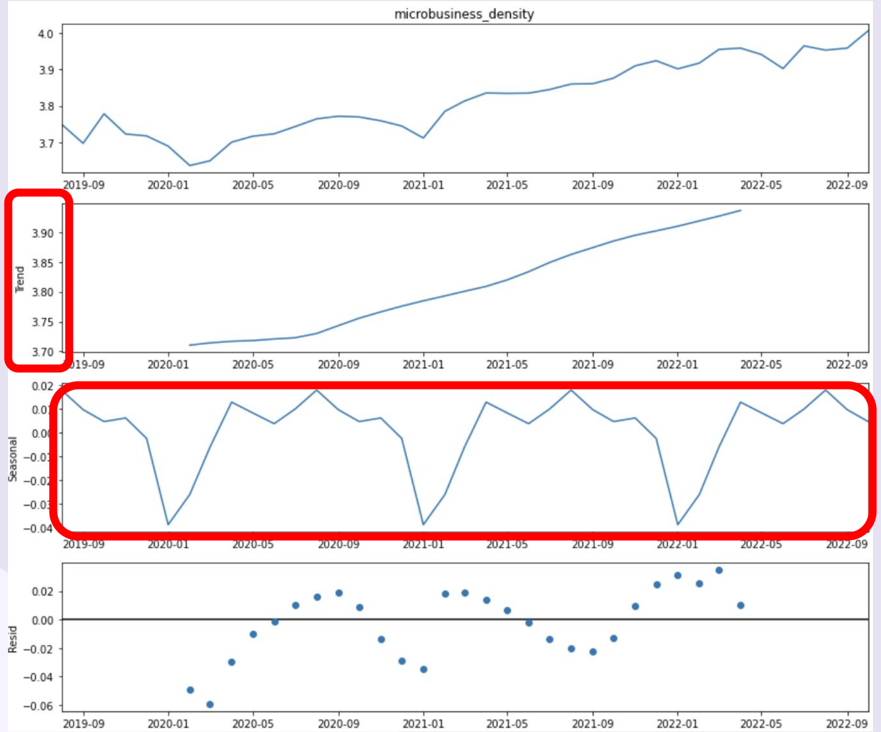


Baseline model converges to an error rate of around ~ 2, which signifies underfitting

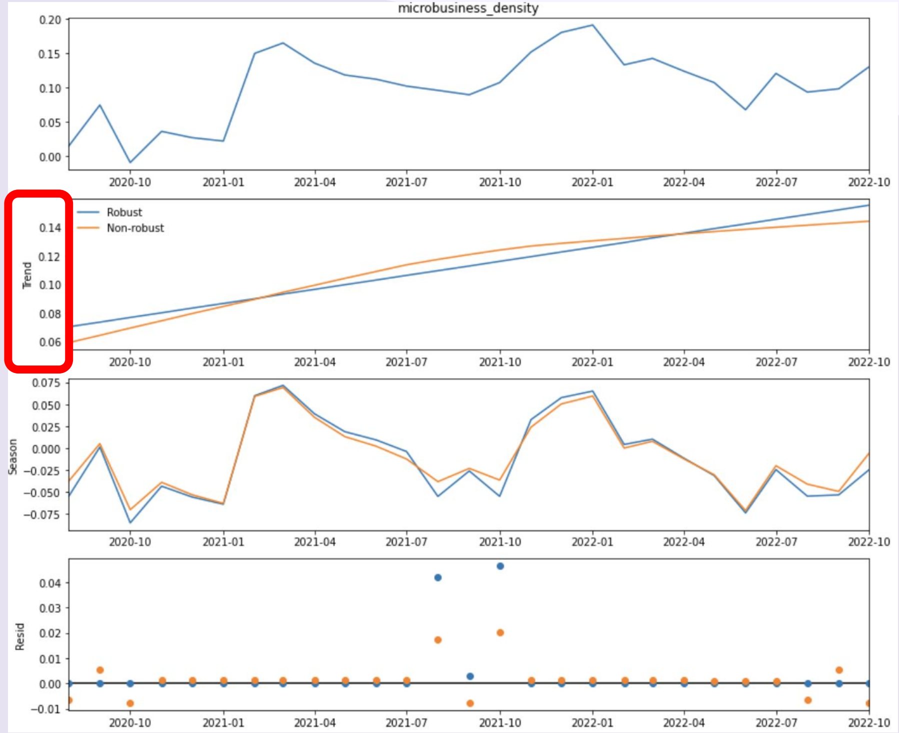
Begins to overfit at around 300 rounds of boosting

Appendix - Time Series Decomposition

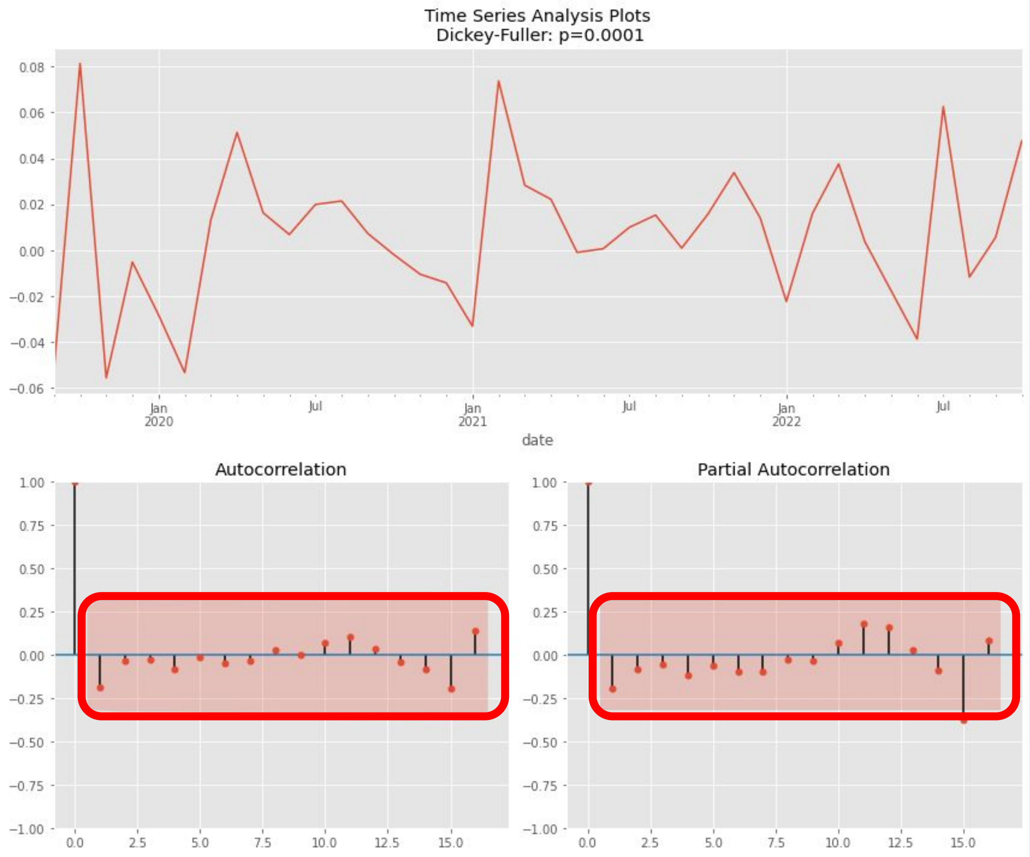
Before Differencing



After Differencing (Reduced Seasonality and Trend)



Appendix - Dickey-Fuller Test for Stationarity

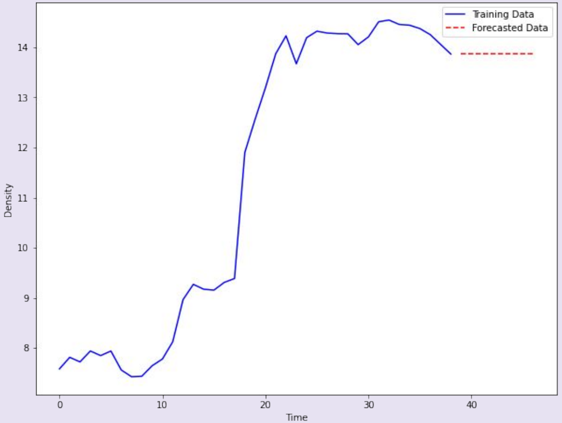


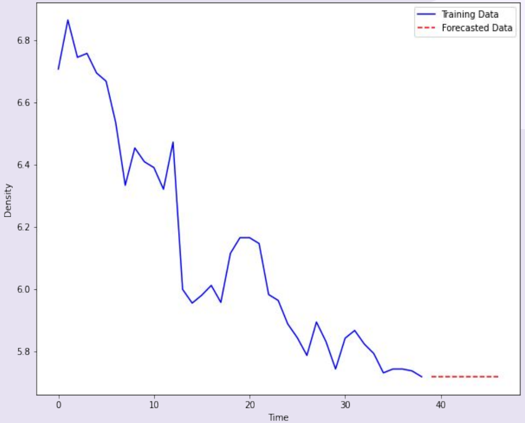
- The x-axis is the number of lagged periods k

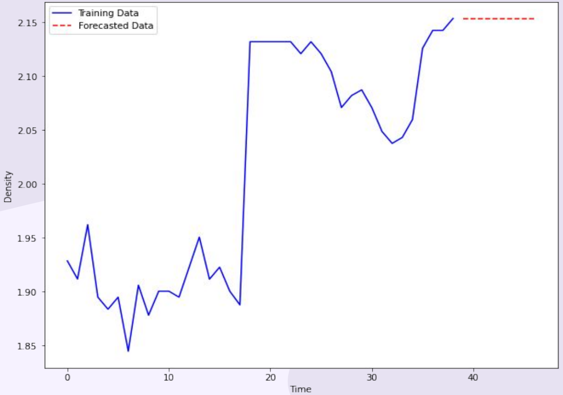
- The y-axis is the autocorrelation coefficient between between -1 and 1 with a confidence interval.

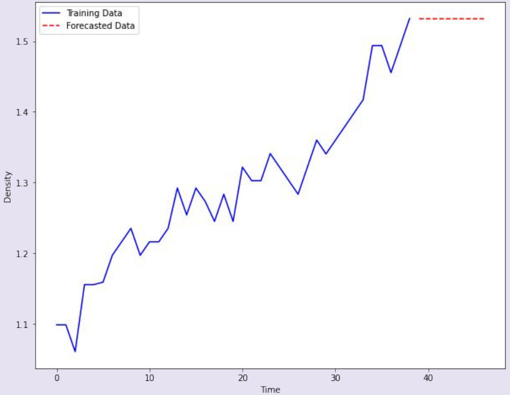
- The autocorrelation measures the linear relationship between lagged values of a time series.

Appendix - Example Forecast

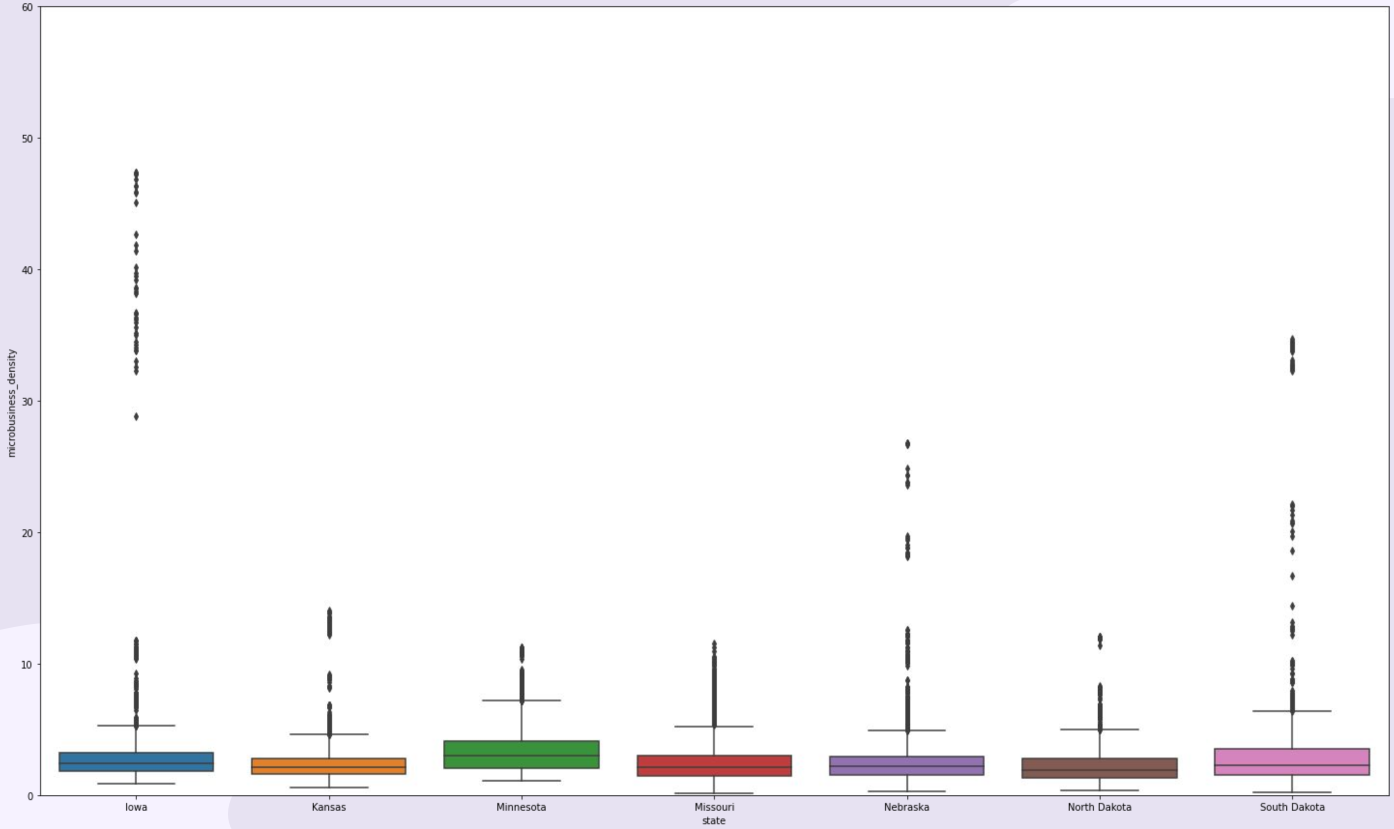




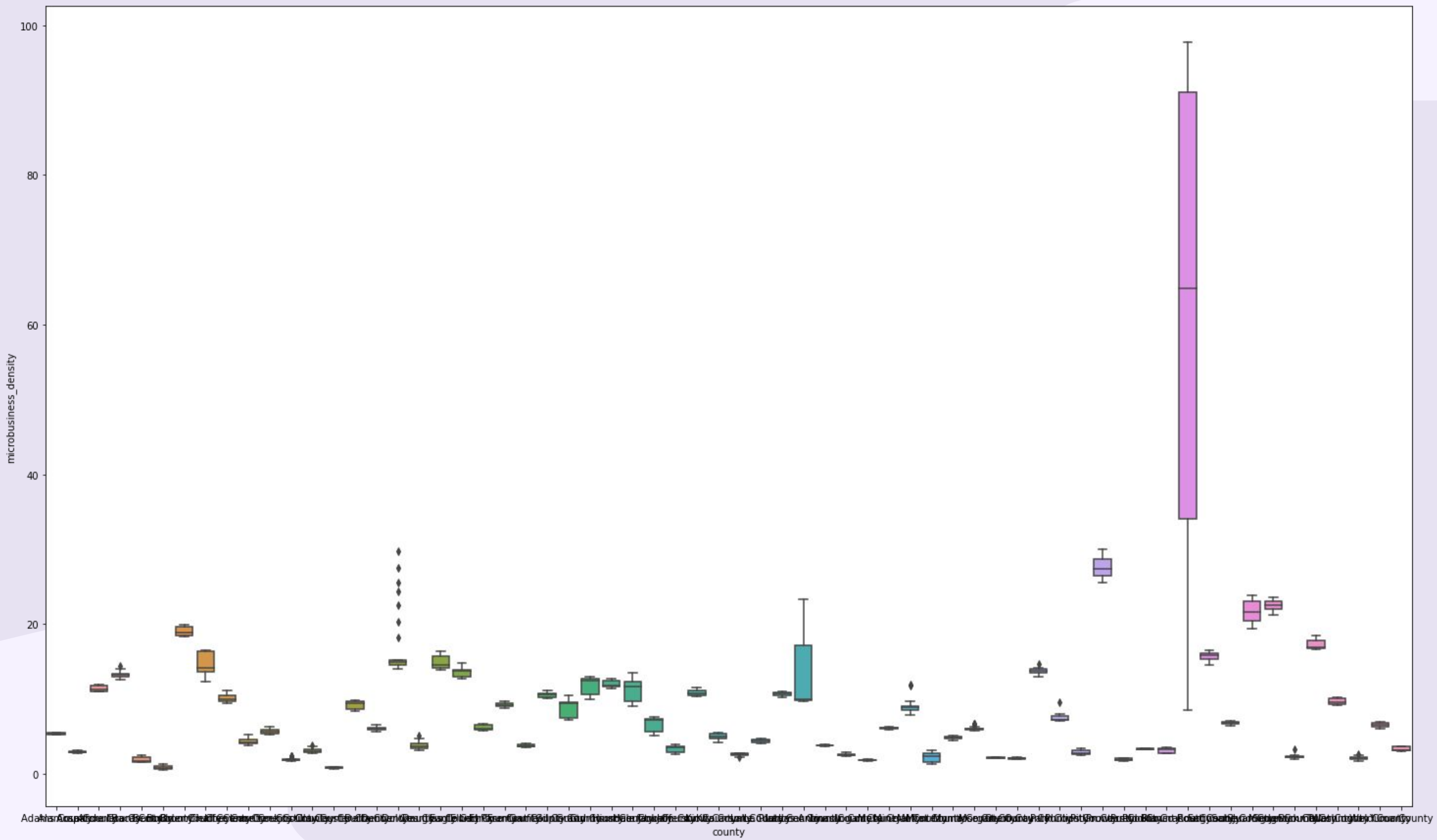




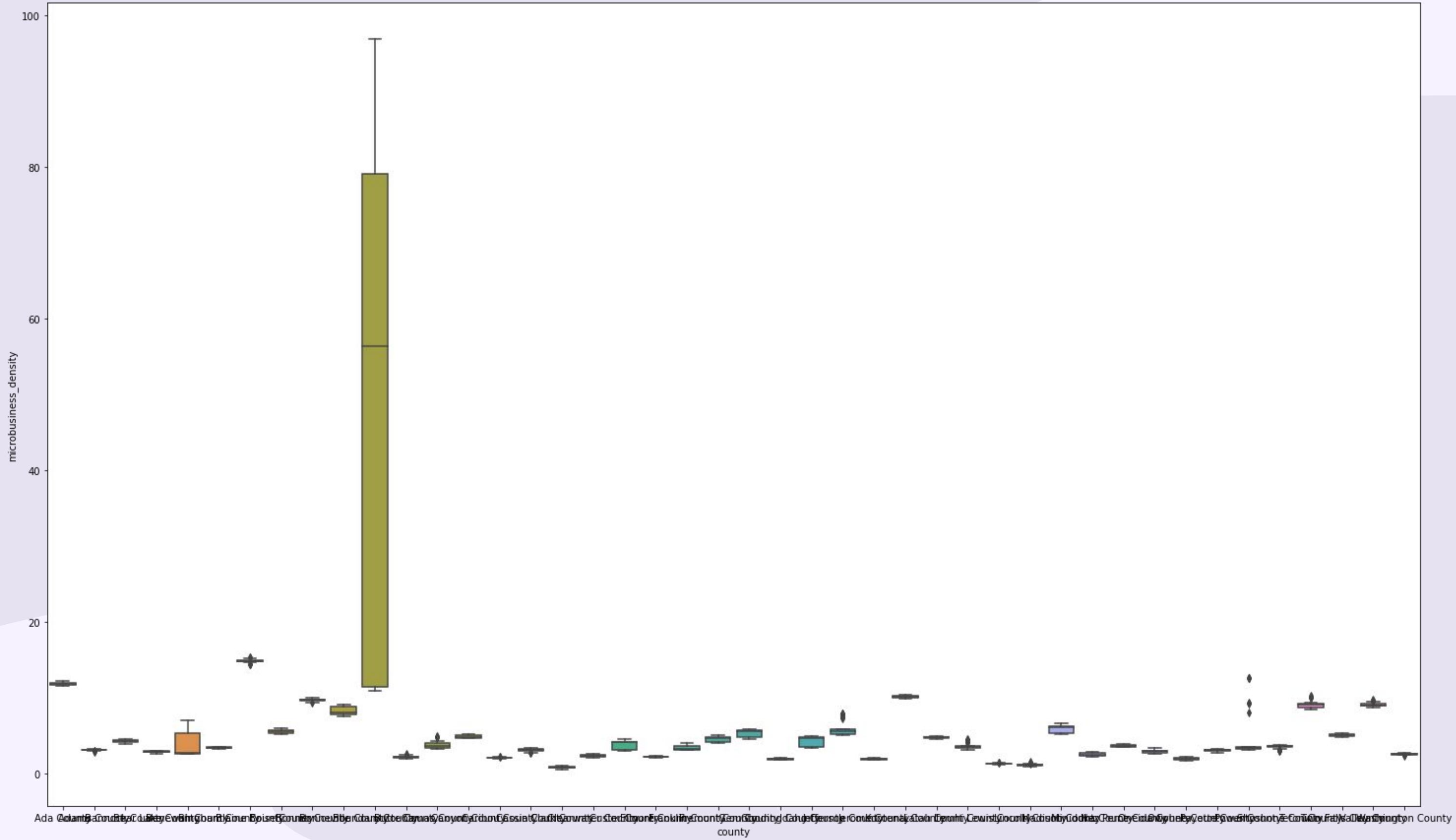
Appendix - Region x Density (West North Central)



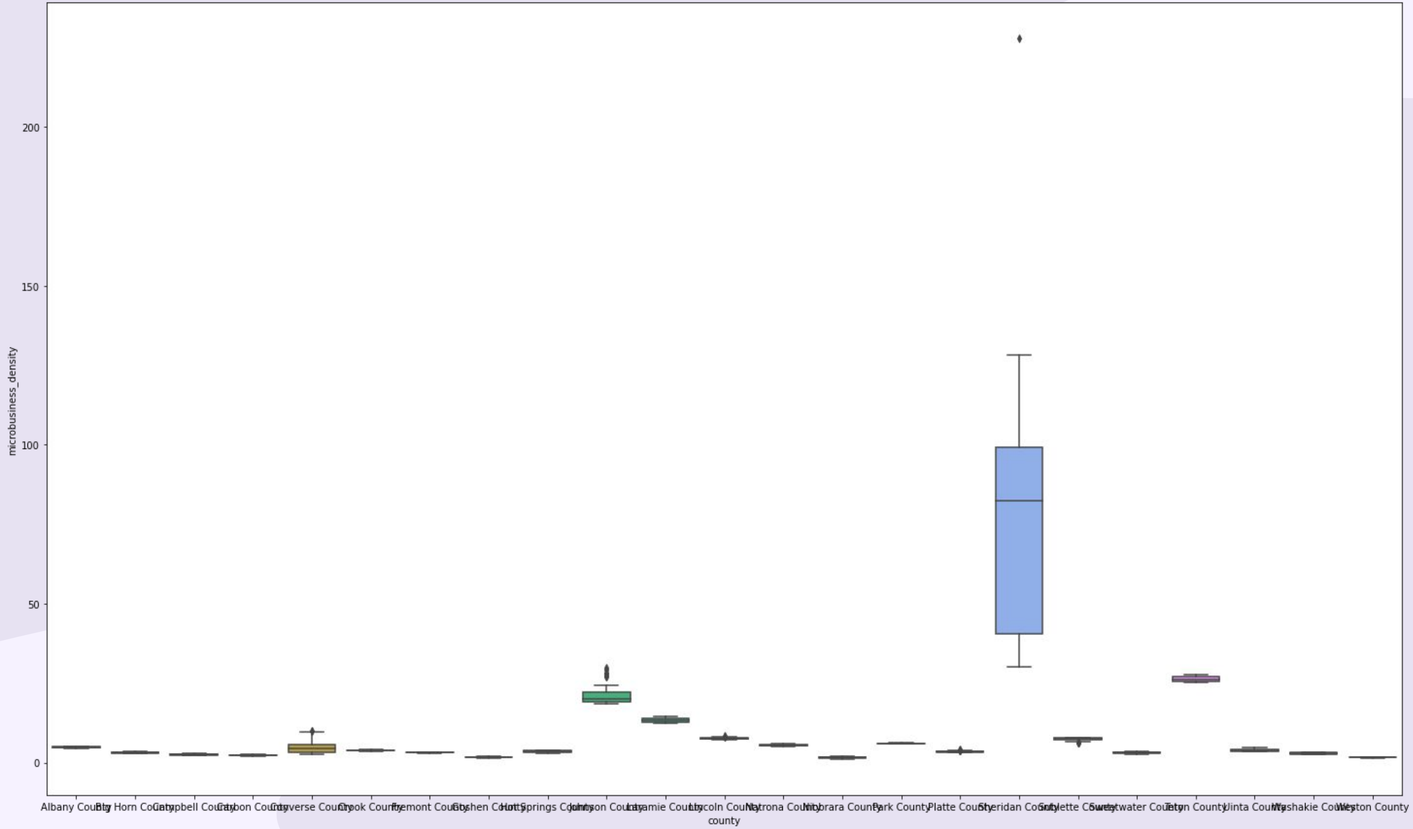
Appendix - Region x Density (Colorado)



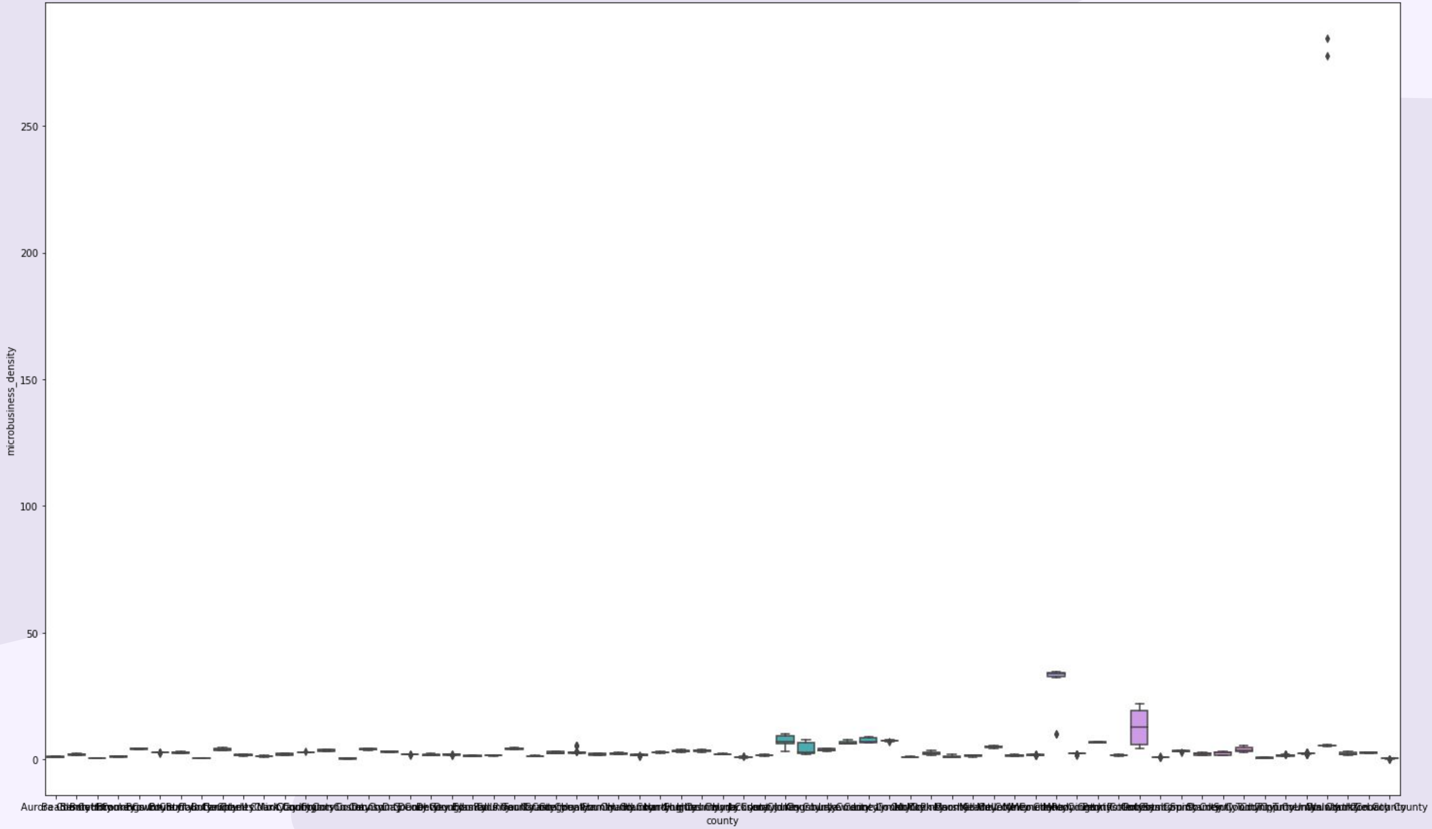
Appendix - Region x Density (Idaho)



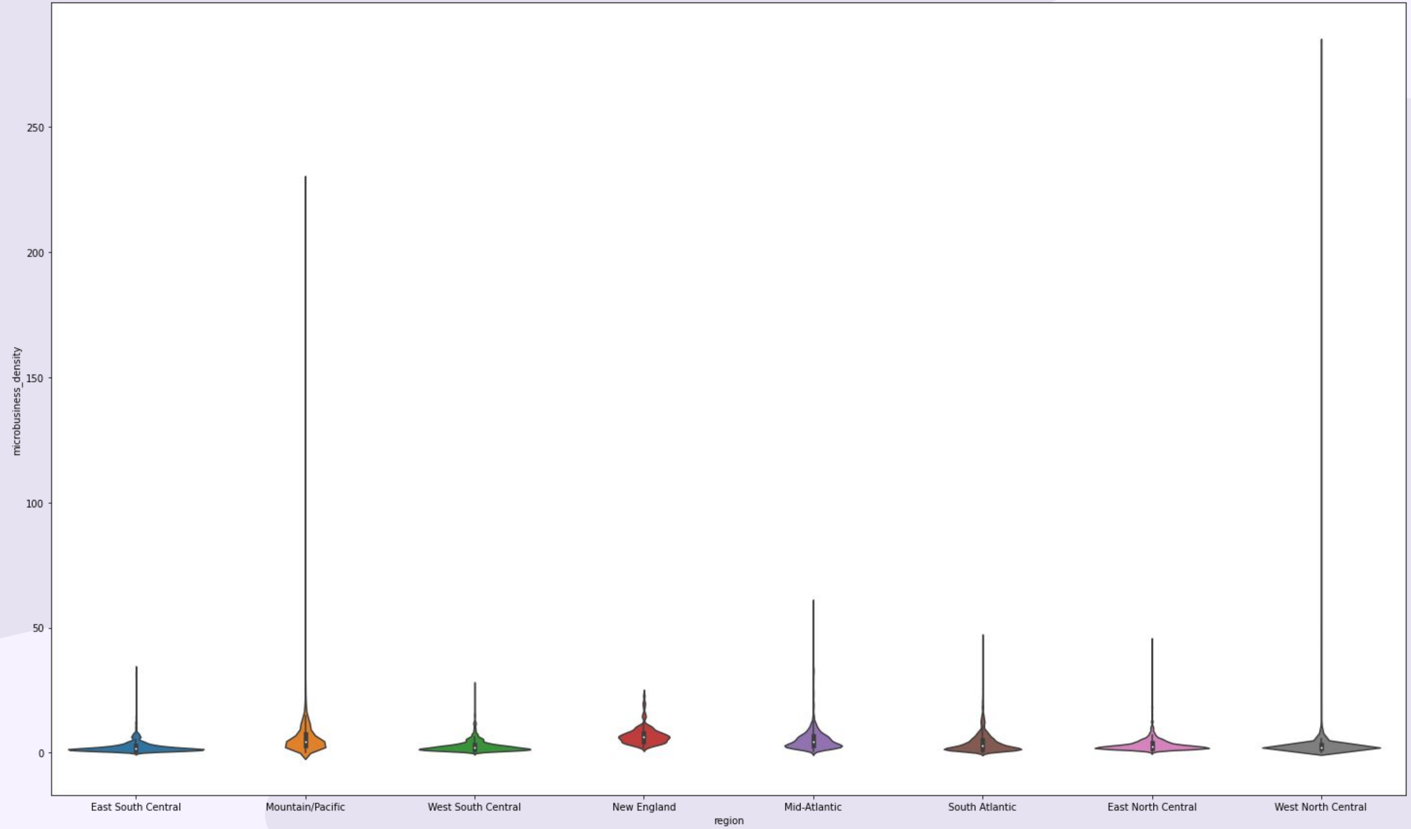
Appendix - Region x Density (Wyoming)



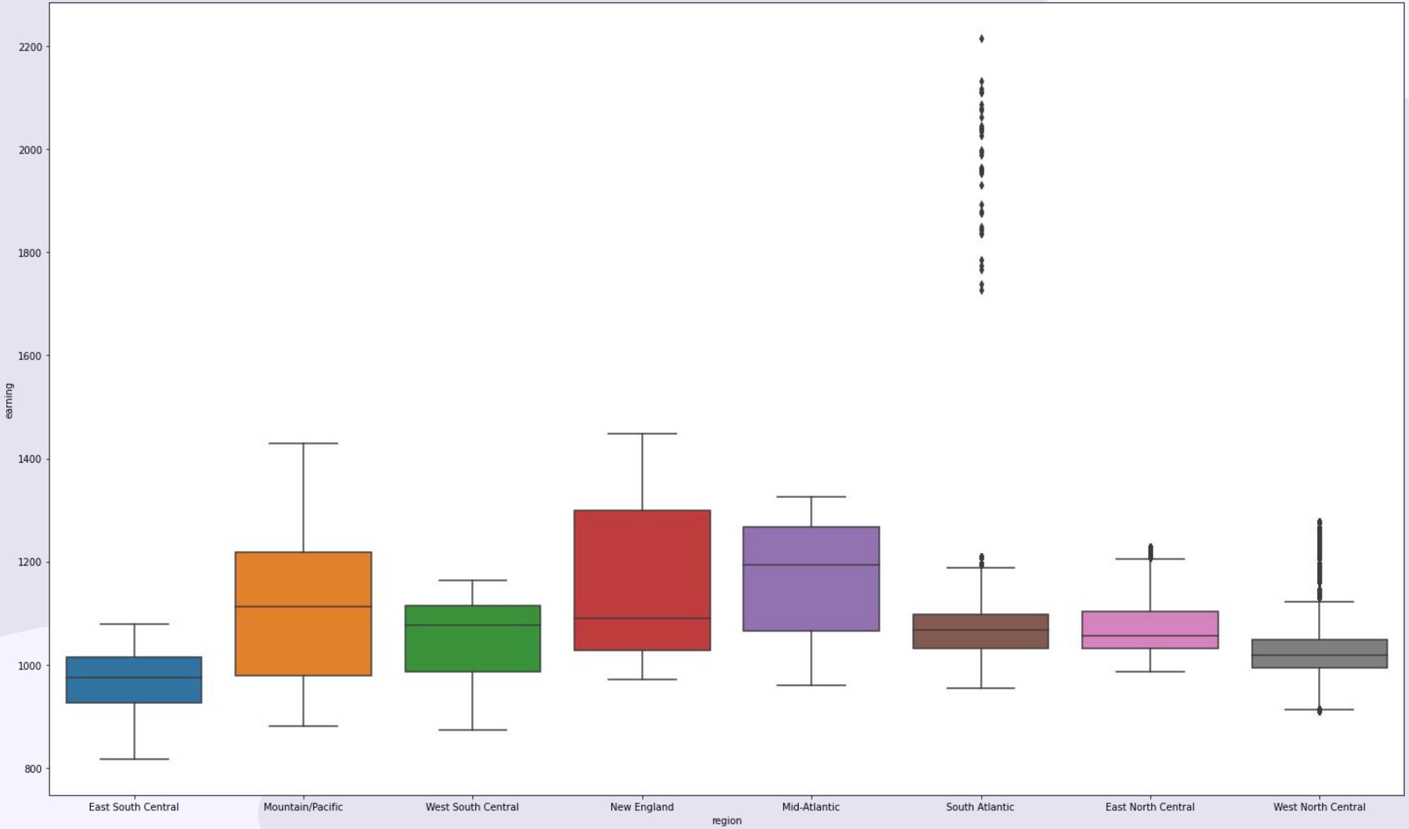
Appendix - Region x Density (South Dakota)



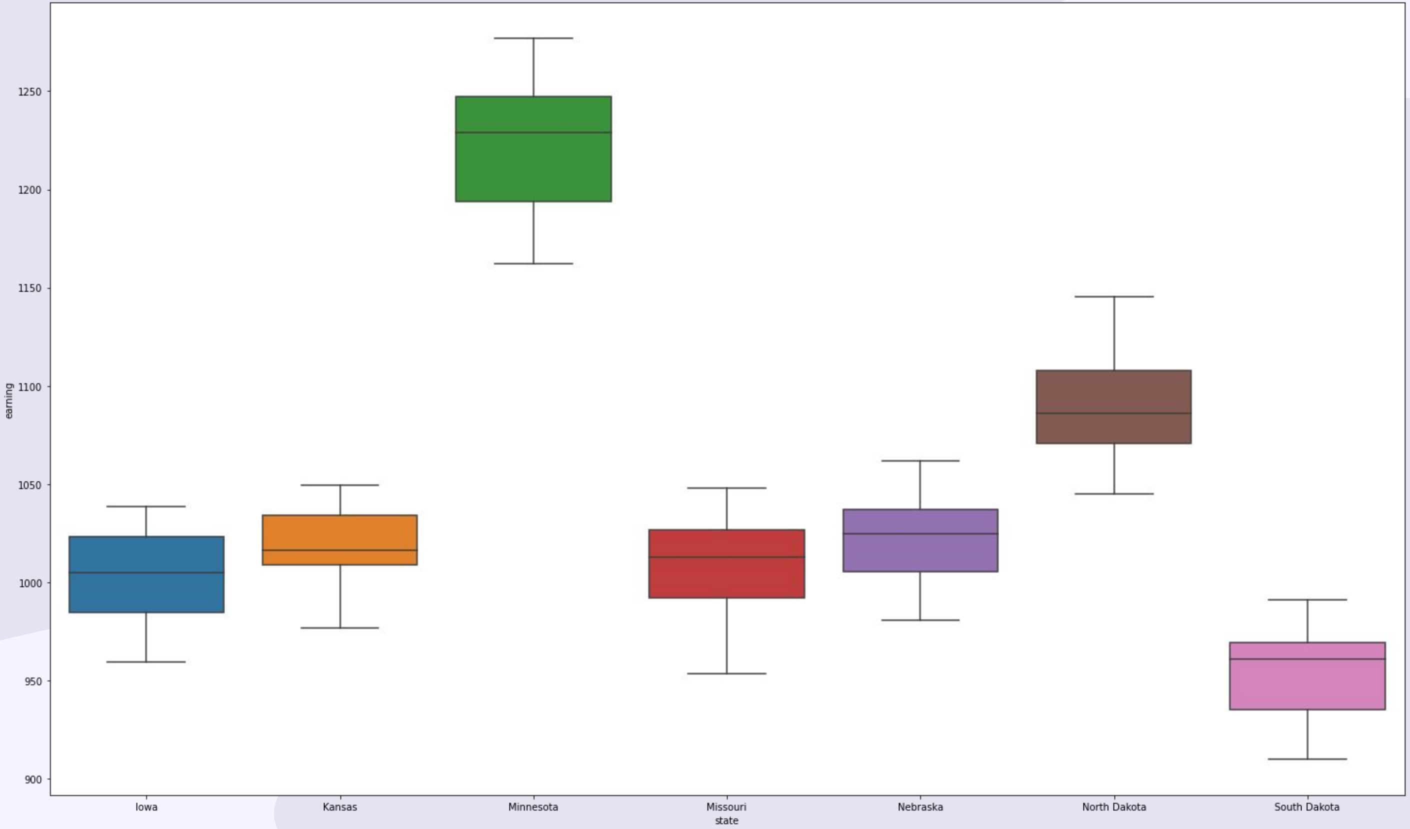
Appendix - Region x Density (Violin)



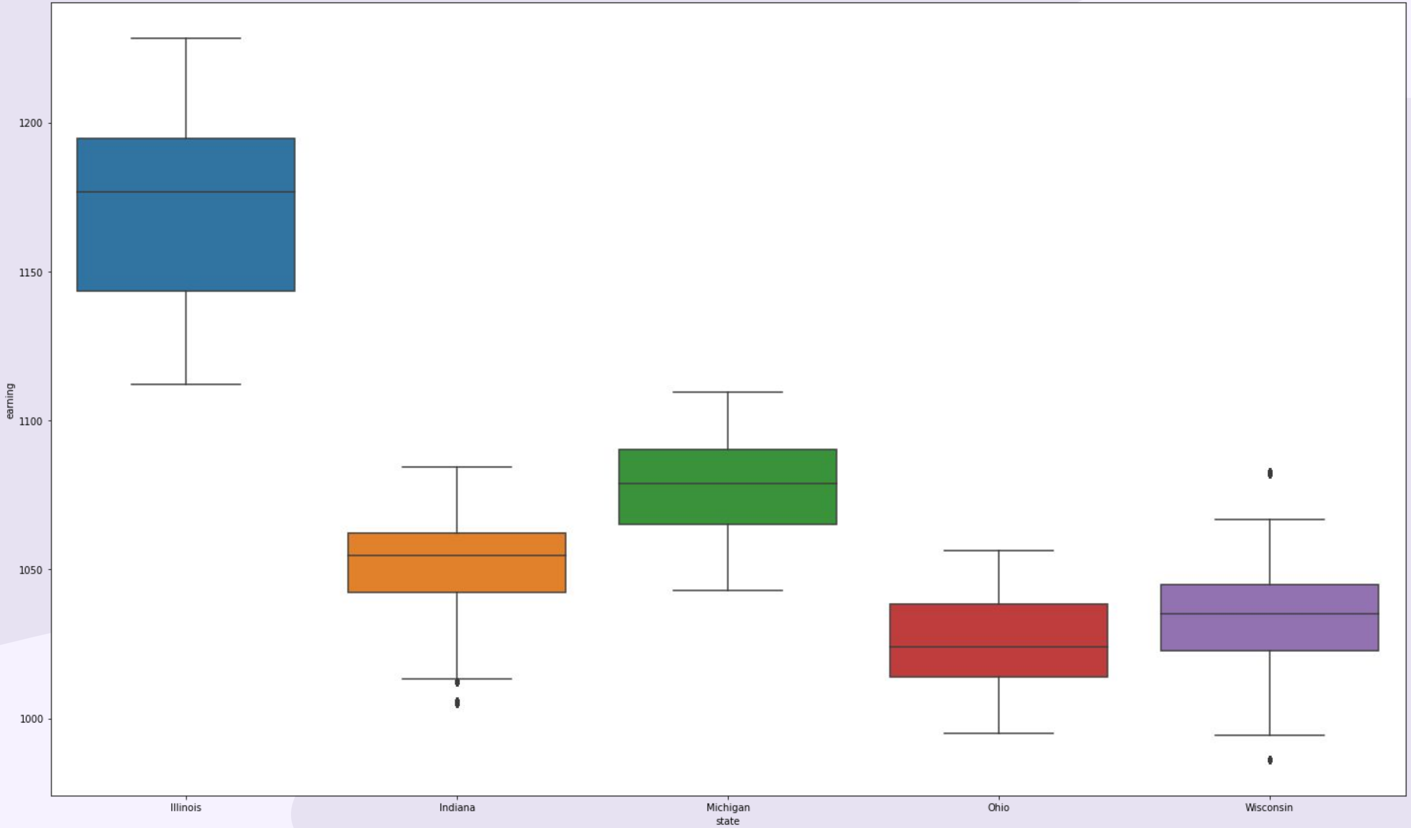
Appendix - Region x Earning



Appendix - Region x Earning (West North Central)



Appendix - Region x Earning (East North Central)


earning
Illinois
Indiana
Michigan
Ohio
Wisconsin
state