Forecasting County-wise Micro-Business Density (MBD)

Mar.2023

Project Goal

- What is a micro-business?
  - Businesses with 10 or fewer employees
  - Has an online presence
- Current Models:
  - Leverages available internal and census data
  - Uses econometric approaches
- Motivation:
  - Potential to include additional data
  - Explore feature engineering techniques
  - Using more advanced approaches to improve predictions

Parties of Interest

Commercial Interests

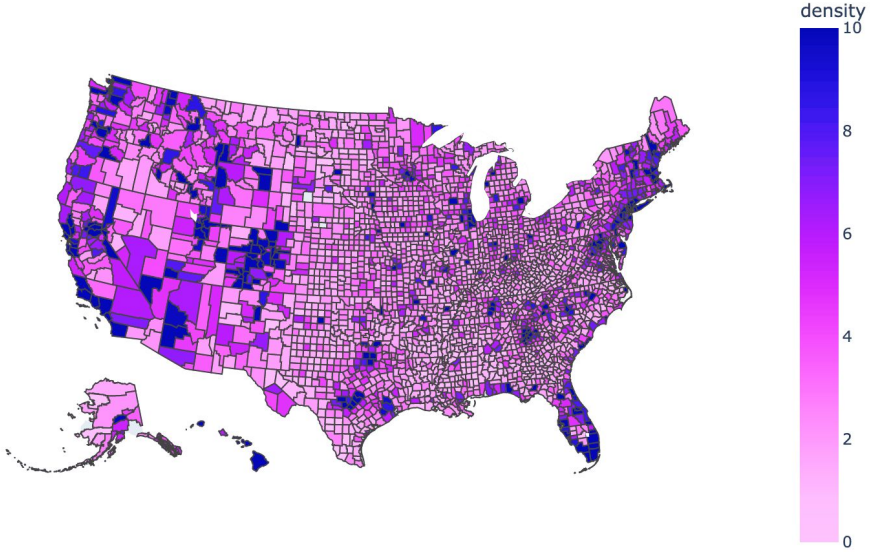
- ➔ Shopify
- ➔ Turbo Tax
- ➔ Office Depot
- ➔ MailChimp
- ➔ Home Depot

Academic Interests

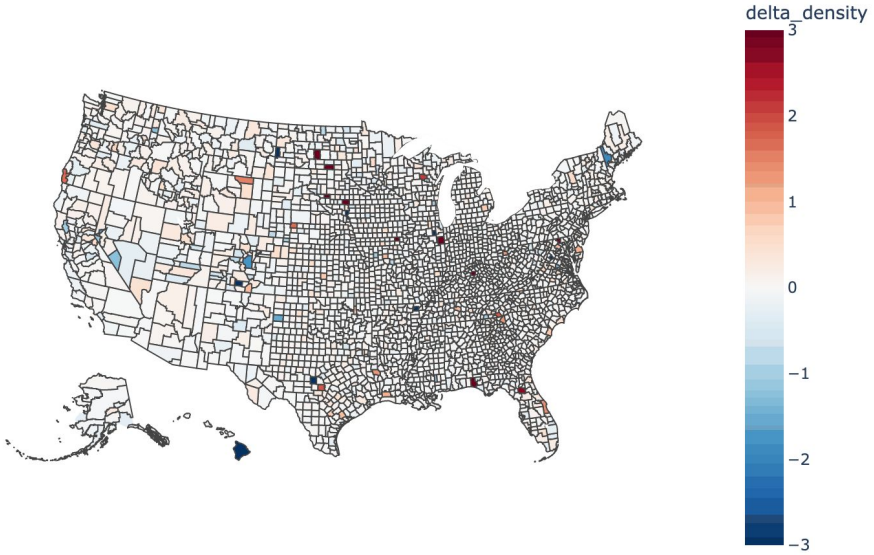
- ➔ Policymakers
- ➔ Economists
- ➔ Social Scientists

Visualization of the Target:
Monthly Micro-Business Density and Density Change

Microbusiness density by county on 2019-08-01

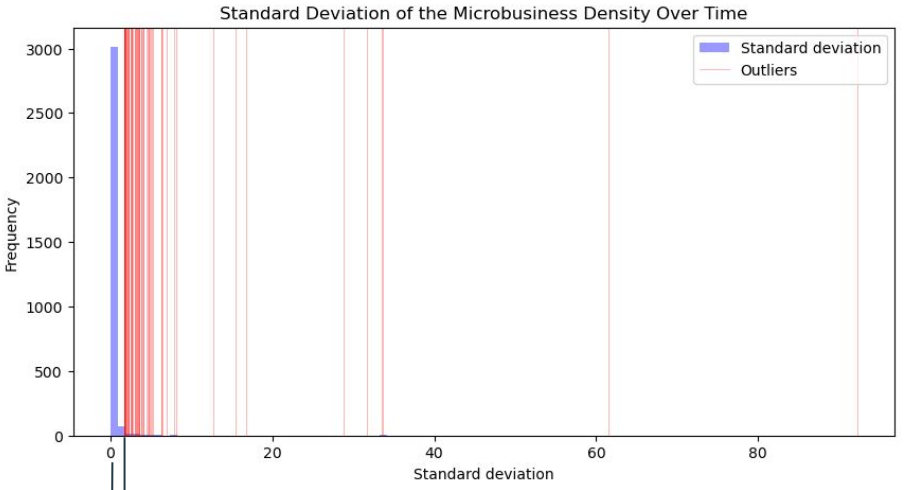


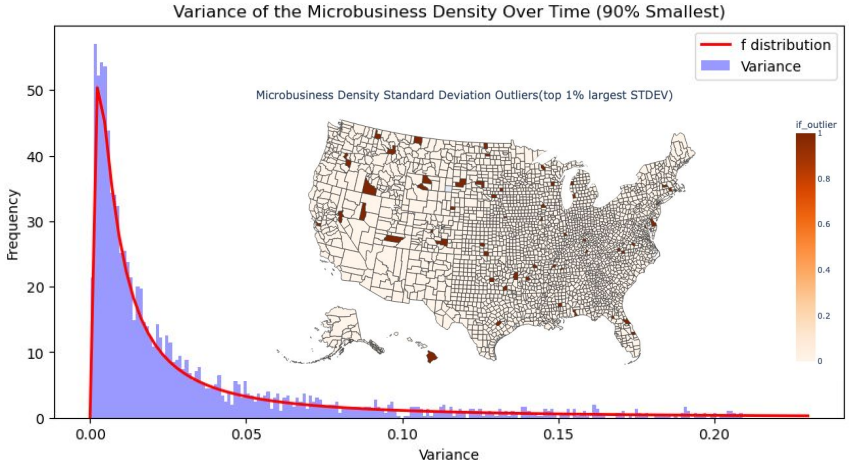
Monthly Change in Microbusiness density by county on 2019-09-01

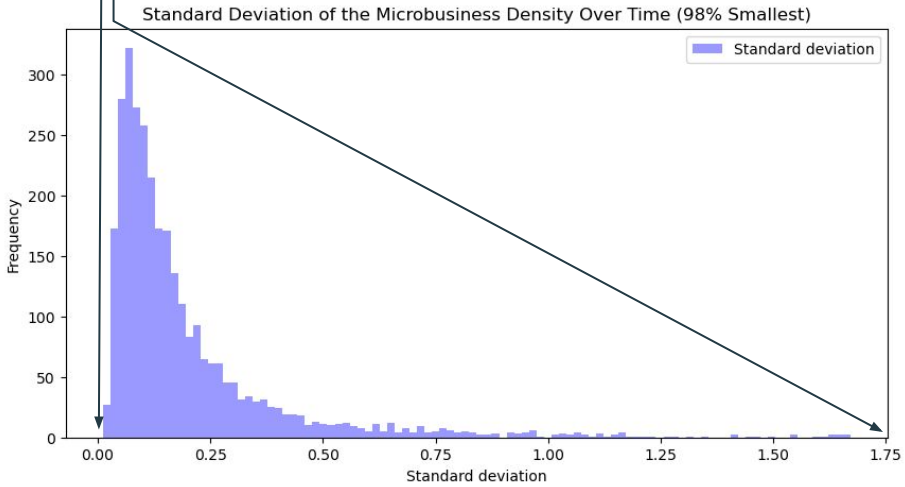


Distribution of the Targets

- The MBD's are not nice and normally distributed, it's heavy-tail-ish.
- Outliers, in red, with high leverage exists.
- Candidate Distributions: f, mielke, nct, betaprime, invgamma







Data Collection & Feature EDA

Types of Features

- Monthly
- Yearly
- Static

Important Features

- **
- Business Tax
-

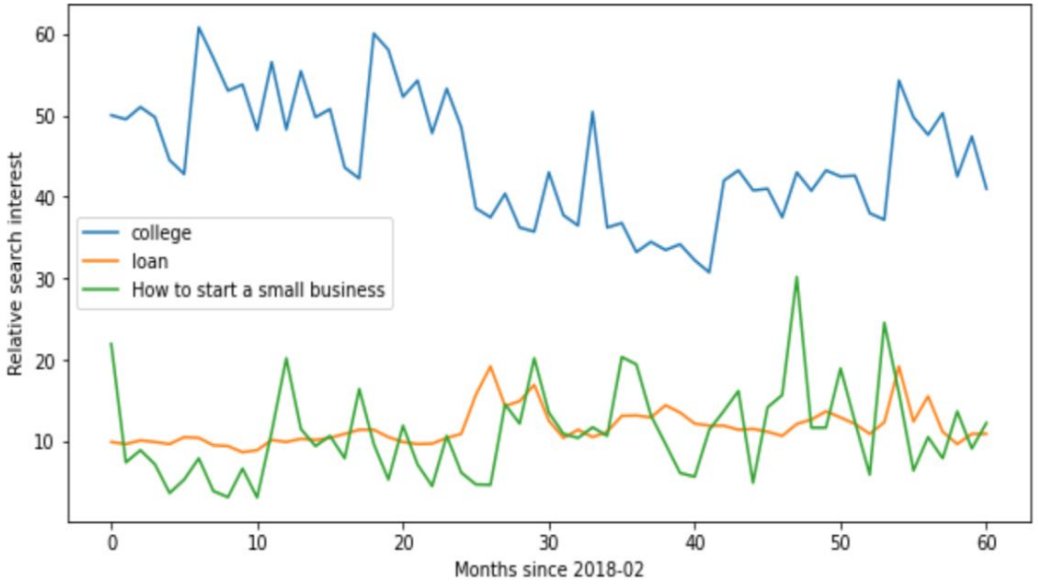
Google Trends Features

County-wise queries (7)

- “Alameda County Tax”

State-wise queries (17)

- “Tax”
- “How to start a small business”
- “Business loan”



Selected Covid Indicators

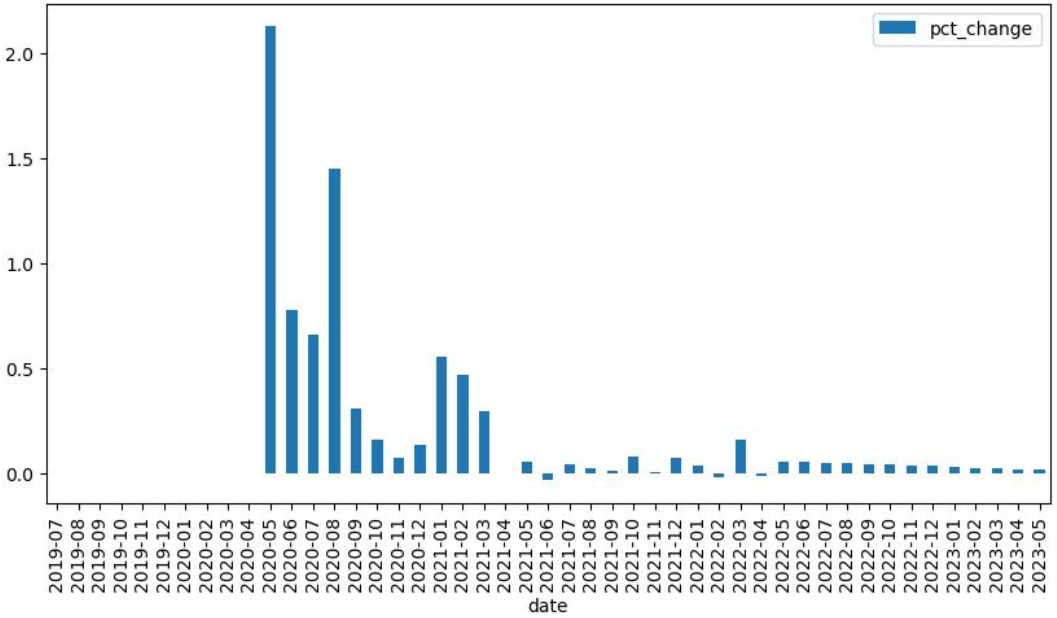
National Personal Income

- Hypothesis: average personal income should be a good proxy on the economic impact of covid.

Covid Death Data

- Hypothesis: Number of death and its' Diff contains signal of severity of covid at given time.
- Hypothesis: Pct Diff in Covid death could act as a proxy for how much people care about Covid. (plotted)

- A proxy for how much people care about Covid



ChatGPT into BERT 

County-wise queries to ChatGPT's API.

ZH

Give me bullet points for why I should or should not start a small business in Autauga County Alabama, 5 pros and 5 cons.

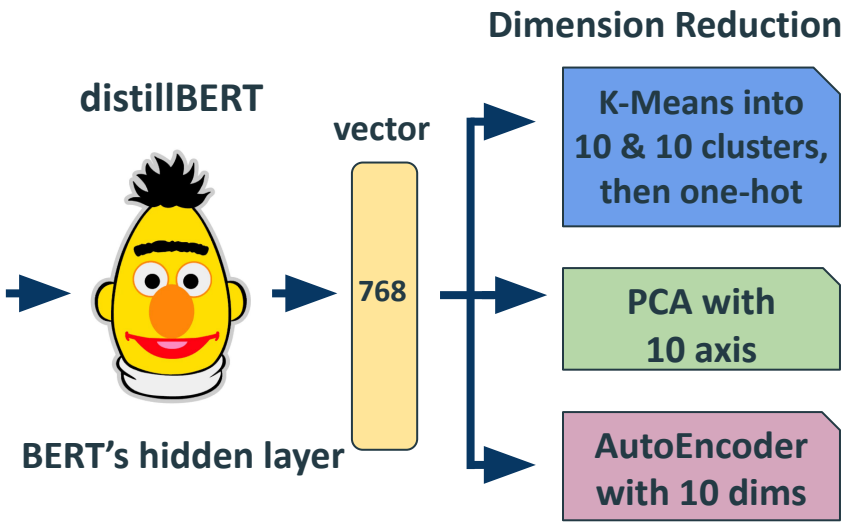
Pros of starting a small business in Autauga County, Alabama:

1. Low cost of living and doing business
2. Supportive business community
3. Favorable business climate
4. Resources available for small businesses, such as the Small Business Administration and local business incubators
5. Well-developed transportation infrastructure

Cons of starting a small business in Autauga County, Alabama:

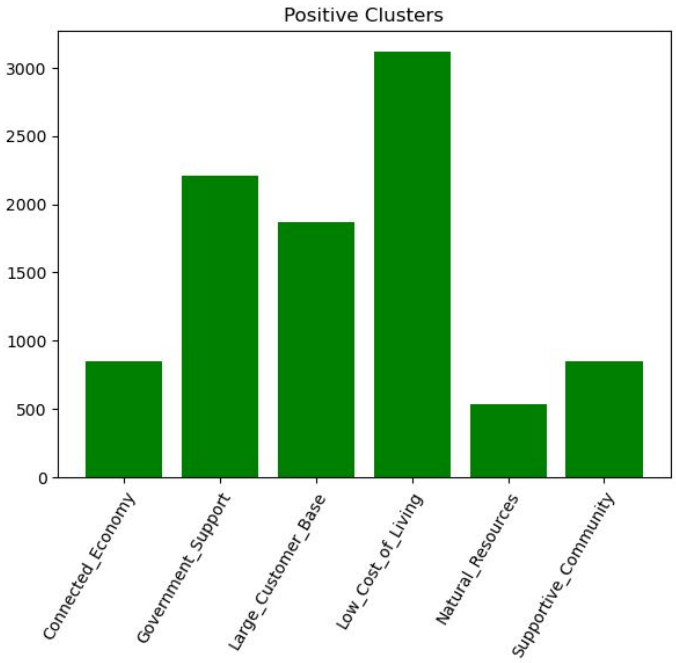
1. Limited access to investment capital
2. Small market size
3. Limited pool of skilled workers
4. Economy may be reliant on specific industries
5. High competition in some industries.

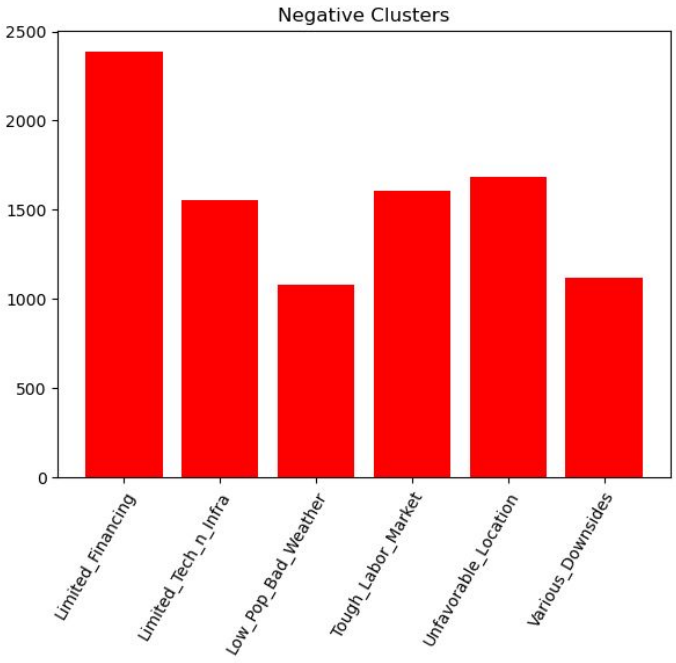
bullet point or paragraph-wise embedding into tensors



Cluster 9 - “Government_Support”:

- There are numerous resources available to help small businesses get started, such as the Small Business Development Center and the Baldwin County Economic Development Alliance.
- Access to resources such as the Blount County Chamber of Commerce, which provides support and resources to local businesses.
- Access to resources such as the Small Business Development Center and the Butler County Chamber of Commerce





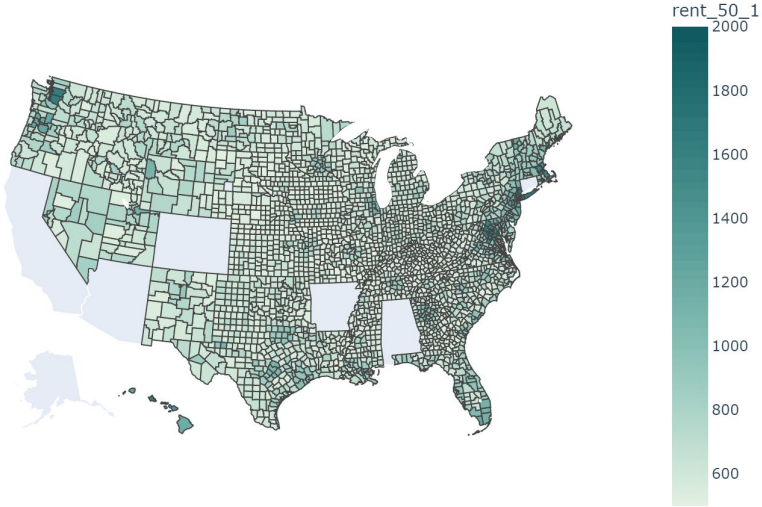
ChatGPT Signal Accuracy

- By visual inspection, rarely does ChatGPT make a wrong description about a county. (County checked: Autauga, Orange, Santa Barbara, Alameda, Los Angeles, Iron County Utah, a bunch of Washington counties and a bunch of Jefferson counties.)

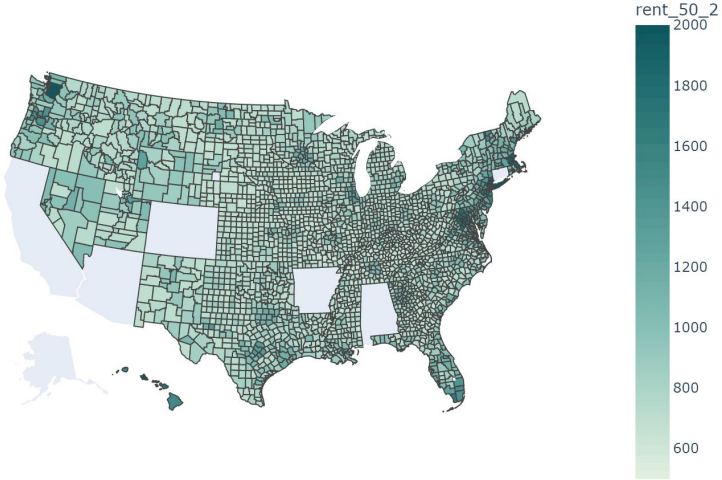
- By visual inspection, one in ten or 10% of the county wise description clustering is incorrect.

Yearly Median Rent Features

median rent (1 bed/s) by county for year 2019

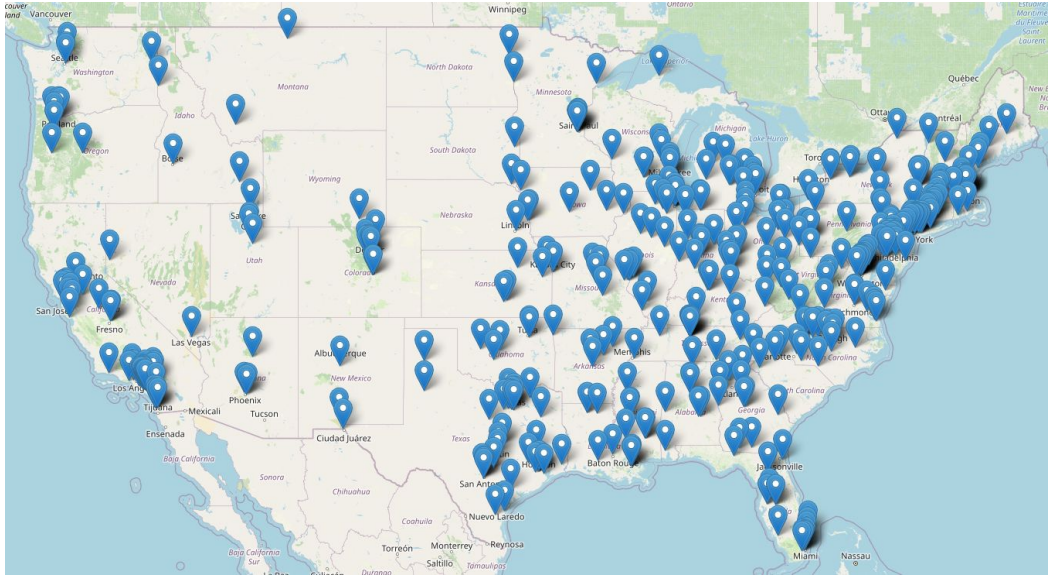


median rent (2 bed/s) by county for year 2019



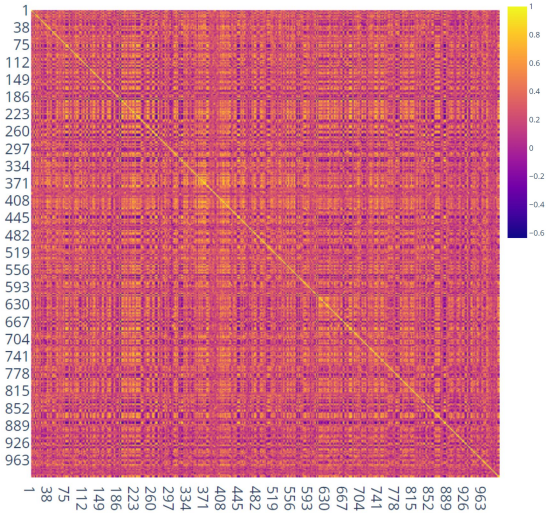
Nearest National University Features

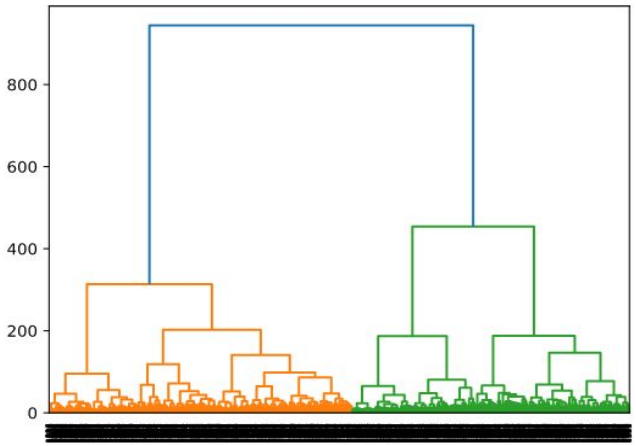
- Tuition
- Enrollment
- Acceptance Rate
- SAT/ACT Scores
- Graduation Rate
- US News Score
- Average Debt at Graduation

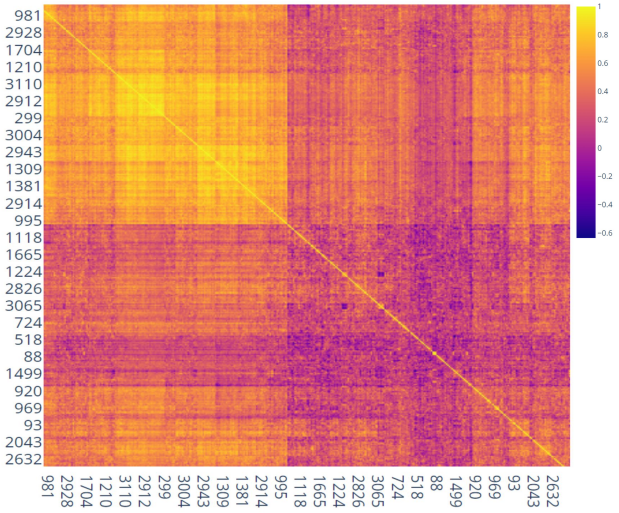


Clustering Target Density Correlation

- Cluster the correlations between the relative change in density between all counties







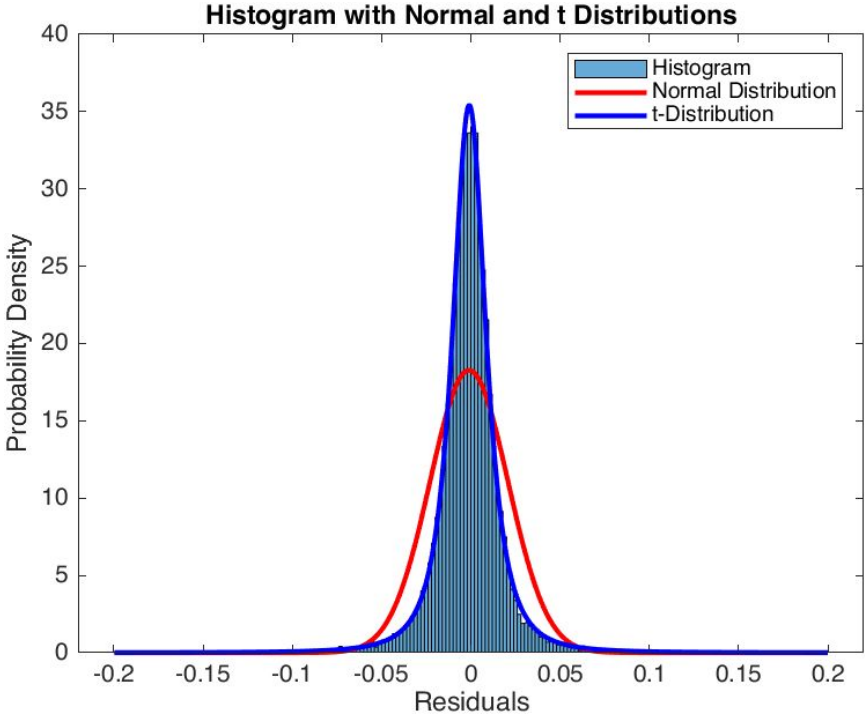
Modeling the data

- Used various regression models
  - Linear Regression
    - OLS
    - Lasso
    - Ridge
    - Elastic Net
  - Decision Trees (regression trees)
  - Ensemble of Trees
    - Random Forest
    - Boosting - Adaboost, LightGBM
  - Neural Networks (MLP)
- PCA on Feature Space

Ordinary Least Squares (already not too shabby)

```
Root Mean Squared Error: 0.0716
R-squared: 0.0176,  Adjusted R-Squared: 0.012
F-statistic vs. constant model: 3.14, p-value = 3.14e-36
```

	tStat	pValue
Remaining_Tax_Burden_per_1_000OfPersonalIncome_	6.7726	1.2901e-11
XXCountyRent	4.905	9.3964e-07
XXCountyLoan	4.8919	1.004e-06
National_PI	4.5592	5.156e-06
(Intercept)	-4.1756	2.9816e-05
XXCountyInsurance	4.0506	5.1226e-05
XXCountyApartments	3.77	0.00016357
XXCountyBusiness	-3.7004	0.00021567
Public_Employees_Per_10_000_of_Population__full_timeEquivalent_	-3.5612	0.00036974
XXCountyLoan_pct	-3.5437	0.00039521
Sales_Tax_Burden_per_1_000OfPersonalIncome_	3.5007	0.00046485
XXCountyTax_pct	3.4646	0.00053178
office_pct	-3.4595	0.00054193
GPT_PCA_dim6	-3.0109	0.0026071
GPT_OH_Tough_Labor_Market	-2.8551	0.004305
XXCountyTax	-2.8238	0.0047493
prev_year_excessive_drinking	2.7745	0.0055322
business	2.6869	0.0072167
prev_year_high_school_completion	-2.6743	0.007492
GPT_OH_Connected_Economy	2.6361	0.0083919



Baseline RMSE (Validation) = 0.1599

Baseline RMSE (Test) = 0.0728

Model Number	Model Type	Preset	PCA	RMSE (Validation)	RMSE (Test)
4.3	Linear Regression	Robust Linear	25 numeric components kept	0.072015	0.027674
8	Neural Network	Optimizable Neural Network	25 numeric components kept	0.07199	0.027681
4.1	Linear Regression	Linear	25 numeric components kept	0.071946	0.027741
4.20	Neural Network	Narrow Neural Network	25 numeric components kept	0.071843	0.028105
4.21	Neural Network	Medium Neural Network	25 numeric components kept	0.040067	0.028227
2.3	Linear Regression	Robust Linear	Disabled	0.07188	0.028257
3.3	Linear Regression	Robust Linear	Disabled	0.07188	0.028257
4.25	Kernel	SVM Kernel	25 numeric components kept	0.071918	0.028376
3.20	Neural Network	Narrow Neural Network	Disabled	0.071479	0.028995
4.2	Linear Regression	Interactions Linear	25 numeric components kept	0.07175	0.029124
4.26	Kernel	Least Squares Regression …	25 numeric components kept	0.07122	0.029674
2.1	Linear Regression	Linear	Disabled	0.071359	0.029862
3.1	Linear Regression	Linear	Disabled	0.071359	0.029862
3.26	Kernel	Least Squares Regression …	Disabled	0.070284	0.030133
3.21	Neural Network	Medium Neural Network	Disabled	0.039694	0.031723
4.22	Neural Network	Wide Neural Network	25 numeric components kept	0.06923	0.038694
3.23	Neural Network	Bilayered Neural Network	Disabled	0.040187	0.041377
3.22	Neural Network	Wide Neural Network	Disabled	0.066478	0.047606
3.24	Neural Network	Trilayered Neural Network	Disabled	0.040166	0.051431
3.15	Ensemble	Bagged Trees	Disabled	0.10896	0.061872
3.7	Tree	Coarse Tree	Disabled	0.1107	0.06252
4.15	Ensemble	Bagged Trees	25 numeric components kept	0.10896	0.062625
3.14	Ensemble	Boosted Trees	Disabled	0.10806	0.062882
4.7	Tree	Coarse Tree	25 numeric components kept	0.11066	0.063084
4.6	Tree	Medium Tree	25 numeric components kept	0.10998	0.063152
3.6	Tree	Medium Tree	Disabled	0.10968	0.063708
4.14	Ensemble	Boosted Trees	25 numeric components kept	0.10798	0.064262
1	Tree	Fine Tree	Disabled	0.10923	0.065197
3.5	Tree	Fine Tree	Disabled	0.10923	0.065197
7	Ensemble	Optimizable Ensemble	25 numeric components kept	0.10153	0.076547
4.5	Tree	Fine Tree	25 numeric components kept	0.10834	0.079892
4.24	Neural Network	Trilayered Neural Network	25 numeric components kept	0.035562	0.10459
5	Tree	Custom Tree	25 numeric components kept	0.10251	0.18817
6	Tree	Optimizable Tree	25 numeric components kept	0.10251	0.18817

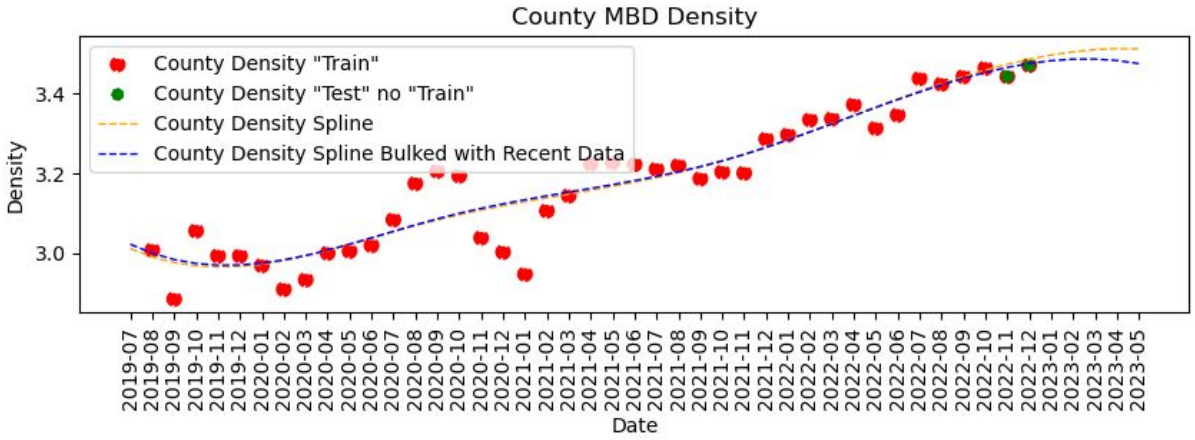
Preliminary Modeling Conclusions

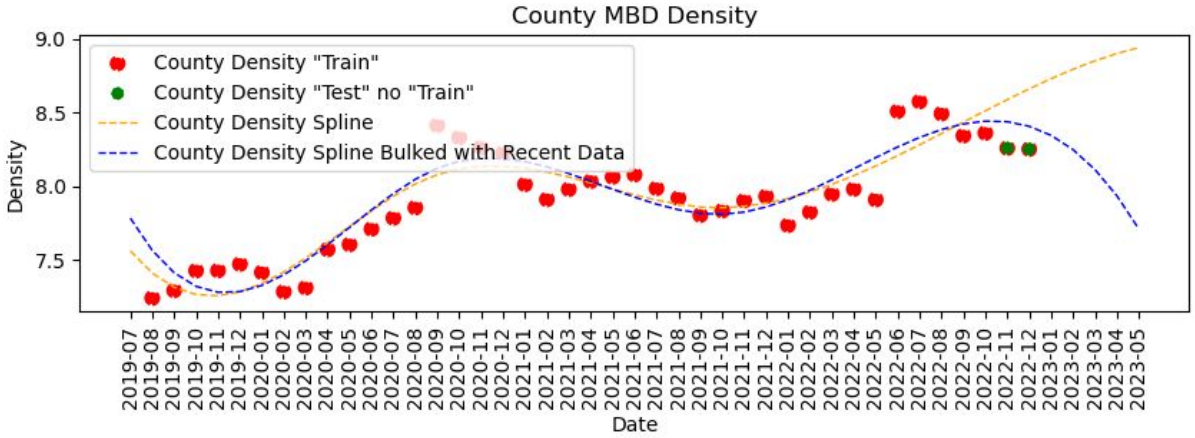
- Linear Models Works Pretty Well
  - Lasso on PCA data with 30 principal axis had the best cross validated performance
  - Lasso selected 8 PCA features
- More Feature Engineering to be Done
  - Tree based model were expected to perform better but not in reality.
  - Lasso and Elastic net could be better after better timewise stationarilization.
- Modeling on outliers the time axis
  - Current model excludes only spatial outliers
  - Time wise anomalies exist, shown on the next slides

Base Models:

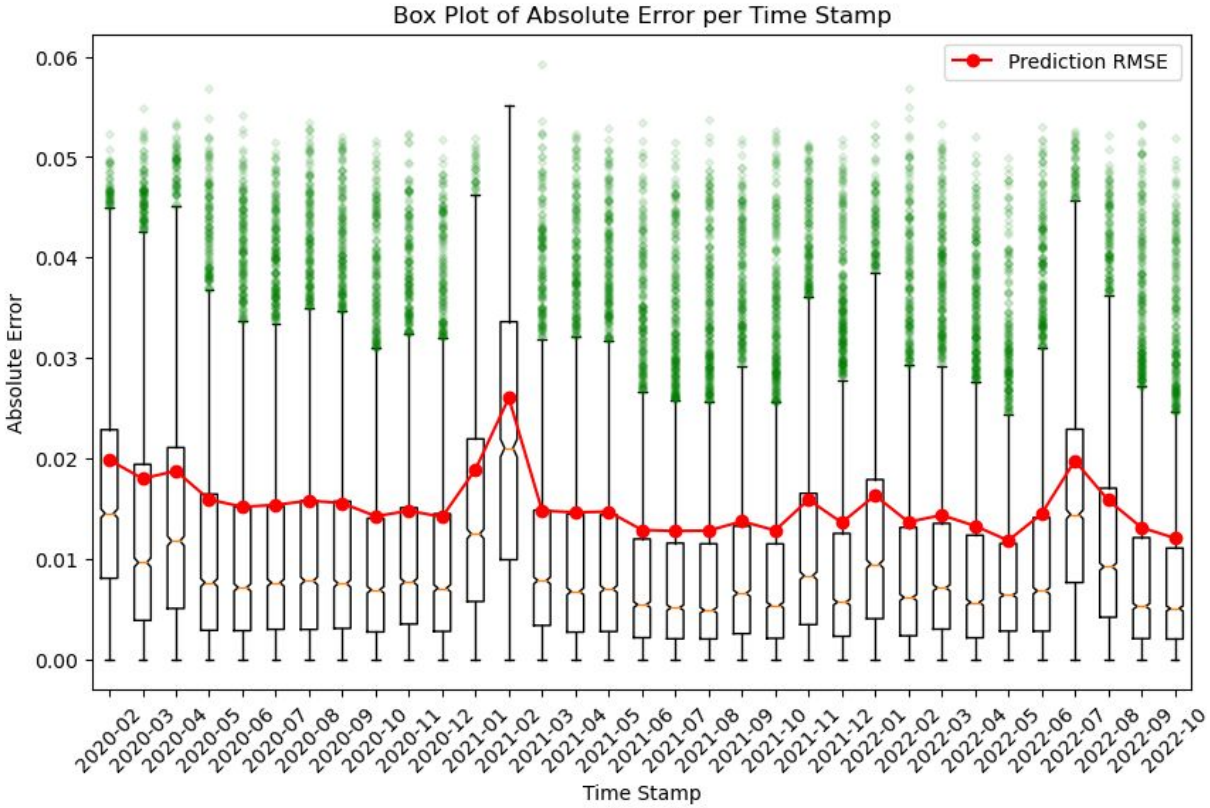
County-wise Spline:

- Good Performance on recent extrapolations, not as much far into the future.

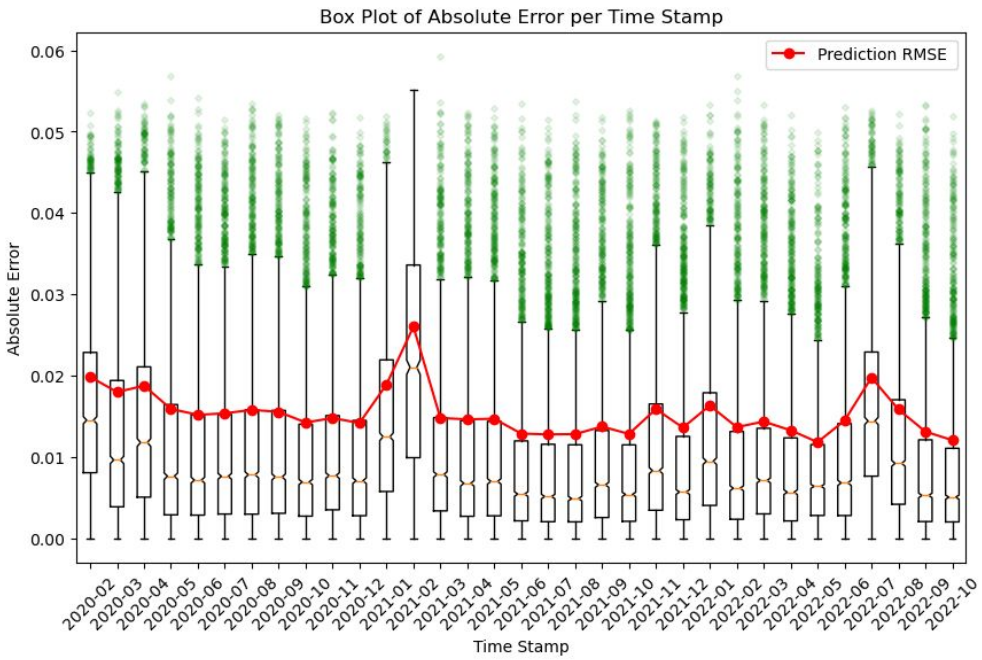




Lasso on PCA performance across time



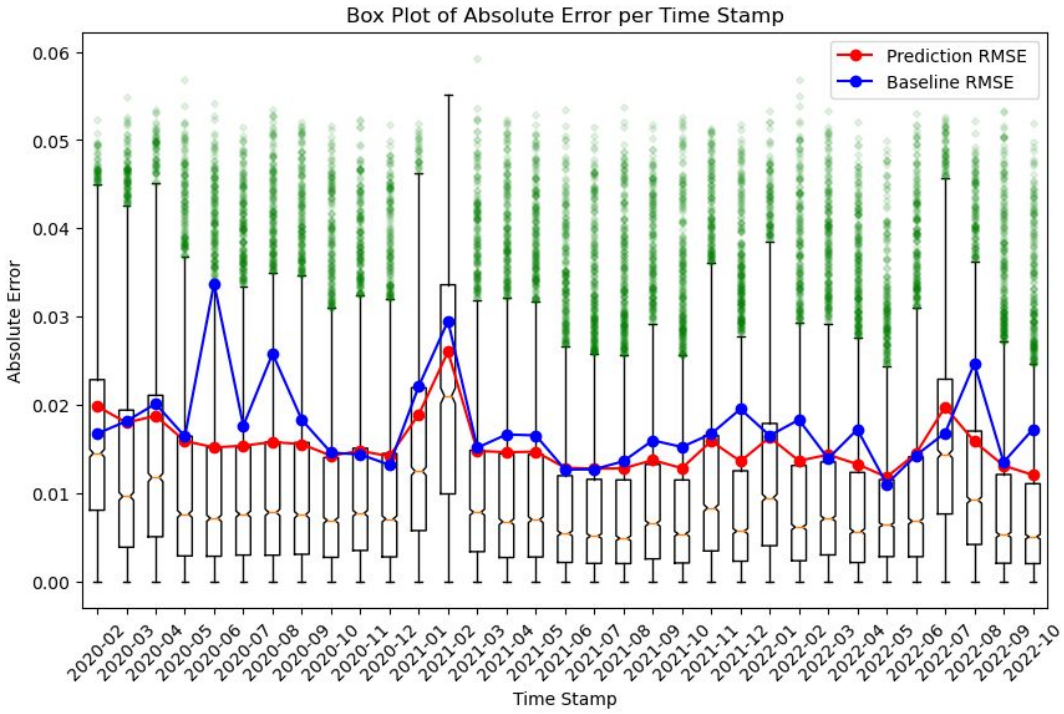
How well does the model deal with Covid



- Simple Hypothesis 1:
  - Model takes into account of covid from the features' signals such as personal income, covid death, and pct growth in death rate.
- Simple Hypothesis 2:
  - Covid had a lagged effect on microbusinesses, causing a turbulence from 2020-11 to 2020-03 (notable cross-sectional variance from our estimator).

Compared to best baseline (outlier excluded mean)

- On most time stamps, we outperforms the best case baseline estimator.



What’s next?

- Include additional features (geospatial)
- Drop smallest x% of counties (currently using 93%)
- Different kinds of clustering (currently only using PCA)
- Better feature engineering (log? inverse?)
- Create multiple models for different clusters/county sizes
- Inspect residuals more closely