GoDaddy - Microbusiness Density Forecasting

Forecast Next Month's Microbusiness Density

Dataset Description

Your challenge in this competition is to forecast microbusiness activity across the United States, as measured by the density of microbusinesses in US counties. Microbusinesses are often too small or too new to show up in traditional economic data sources, but microbusiness activity may be correlated with other economic indicators of general interest.

As historic economic data are widely available, this is a forecasting competition. The forecasting phase public leaderboard and final private leaderboard will be determined using data gathered after the submission period closes. You will make static forecasts that can only incorporate information available before the end of the submission period. This means that while we will rescore submissions during the forecasting period we will *not* rerun any notebooks.

Files

A great deal of data is publicly available about counties and we have not attempted to gather it all here. You are strongly encouraged to use external data sources for features.

train.csv

- `row_id` - An ID code for the row.
- `cfips` - A unique identifier for each county using the Federal Information Processing System. The first two digits correspond to the state FIPS code, while the following 3 represent the county.
- `county_name` - The written name of the county.
- `state_name` - The name of the state.
- `first_day_of_month` - The date of the first day of the month.
- `microbusiness_density` - Microbusinesses per 100 people over the age of 18 in the given county. This is the target variable. The population figures used to calculate the density are on a two-year lag due to the pace of update provided by the U.S. Census Bureau, which provides the underlying population data annually. 2021 density figures are calculated using 2019 population figures, etc.
- `active` - The raw count of microbusinesses in the county. Not provided for the test set.

sample_submission.csv A valid sample submission. This file will remain unchanged throughout the competition.

- `row_id` - An ID code for the row.
- `microbusiness_density` - The target variable.

test.csv Metadata for the submission rows. This file will remain unchanged throughout the competition.

- `row_id` - An ID code for the row.
- `cfips` - A unique identifier for each county using the Federal Information Processing System. The first two digits correspond to the state FIPS code, while the following 3 represent the county.
- `first_day_of_month` - The date of the first day of the month.

revealed_test.csv During the submission period, only the most recent month of data will be used for the public leaderboard. Any test set data older than that will be published in **revealed_test.csv**, closely following the usual data release cycle for the microbusiness report. We expect to publish one copy of **revealed_test.csv** in mid February. This file's schema will match **train.csv**.

census_starter.csv Examples of useful columns from the Census Bureau's American Community Survey (ACS) at data.census.gov. The percentage fields were derived from the raw counts provided by the ACS. All fields have a two year lag to match what information was avaiable at the time a given microbusiness data update was published.

- `pct_bb_[year]` - The percentage of households in the county with access to broadband of any type. Derived from ACS table B28002: PRESENCE AND TYPES OF INTERNET SUBSCRIPTIONS IN HOUSEHOLD.
- `cfips` - The CFIPS code.
- `pct_college_[year]` - The percent of the population in the county over age 25 with a 4-year college degree. Derived from ACS table S1501: EDUCATIONAL ATTAINMENT.
- `pct_foreign_born_[year]` - The percent of the population in the county born outside of the United States. Derived from ACS table DP02: SELECTED SOCIAL CHARACTERISTICS IN THE UNITED STATES.
- `pct_it_workers_[year]` - The percent of the workforce in the county employed in information related industries. Derived from ACS table S2405: INDUSTRY BY OCCUPATION FOR THE CIVILIAN EMPLOYED POPULATION 16 YEARS AND OVER.
- `median_hh_inc_[year]` - The median household income in the county. Derived from ACS table S1901: INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS).

In [1]:
```
# Importing Data Manupulation Library
import pandas as pd
import numpy as np

# Clean Notebook
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
train=pd.read_csv('train.csv')
train.head()
```

Out[2]:

	row_id	cfips	county	state	first_day_of_month	microbusiness_density	active
0	1001_2019-08-01	1001	Autauga County	Alabama	2019-08-01	3.007682	1249
1	1001_2019-09-01	1001	Autauga County	Alabama	2019-09-01	2.884870	1198
2	1001_2019-10-01	1001	Autauga County	Alabama	2019-10-01	3.055843	1269
3	1001_2019-11-01	1001	Autauga County	Alabama	2019-11-01	2.993233	1243
4	1001_2019-12-01	1001	Autauga County	Alabama	2019-12-01	2.993233	1243

In [3]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122265 entries, 0 to 122264
Data columns (total 7 columns):
 #   Column                  Non-Null Count   Dtype  
---  ------                  --------------   -----  
 0   row_id                  122265 non-null  object 
 1   cfips                   122265 non-null  int64  
 2   county                  122265 non-null  object 
 3   state                   122265 non-null  object 
 4   first_day_of_month      122265 non-null  object 
 5   microbusiness_density   122265 non-null  float64
 6   active                  122265 non-null  int64  
dtypes: float64(1), int64(2), object(4)
memory usage: 6.5+ MB
```

In [4]:
```
train.describe(include='all')
```

Out[4]:

	row_id	cfips	county	state	first_day_of_month	microbusiness_density	active
count	122265	122265.000000	122265	122265	122265	122265.000000	1.222650e+05
unique	122265	NaN	1871	51	39	NaN	NaN
top	1001_2019-08-01	NaN	Washington County	Texas	2019-08-01	NaN	NaN
freq	1	NaN	1170	9906	3135	NaN	NaN
mean	NaN	30376.037640	NaN	NaN	NaN	3.817671	6.442858e+03
std	NaN	15143.508721	NaN	NaN	NaN	4.991087	3.304001e+04
min	NaN	1001.000000	NaN	NaN	NaN	0.000000	0.000000e+00
25%	NaN	18177.000000	NaN	NaN	NaN	1.639344	1.450000e+02
50%	NaN	29173.000000	NaN	NaN	NaN	2.586543	4.880000e+02
75%	NaN	45077.000000	NaN	NaN	NaN	4.519231	2.124000e+03
max	NaN	56045.000000	NaN	NaN	NaN	284.340030	1.167744e+06

- Change Cfips as discrete value
- first day of the month datetime

In [5]:
```
# Change Data type
df1=train.copy()
df1.cfips=df1.cfips.astype(str)
```

In [6]:
```
# Data Type changed
df1.dtypes
```

Out[6]:
```
row_id                   object
cfips                    object
county                   object
state                    object
first_day_of_month       object
microbusiness_density    float64
active                     int64
dtype: object
```

In [7]:
```
# preview
df1.head(5)
```

Out[7]:

	row_id	cfips	county	state	first_day_of_month	microbusiness_density	active
0	1001_2019-08-01	1001	Autauga County	Alabama	2019-08-01	3.007682	1249
1	1001_2019-09-01	1001	Autauga County	Alabama	2019-09-01	2.884870	1198
2	1001_2019-10-01	1001	Autauga County	Alabama	2019-10-01	3.055843	1269
3	1001_2019-11-01	1001	Autauga County	Alabama	2019-11-01	2.993233	1243
4	1001_2019-12-01	1001	Autauga County	Alabama	2019-12-01	2.993233	1243

In [8]:
```
# Seperating active column not present in test
active=train.active
active
```

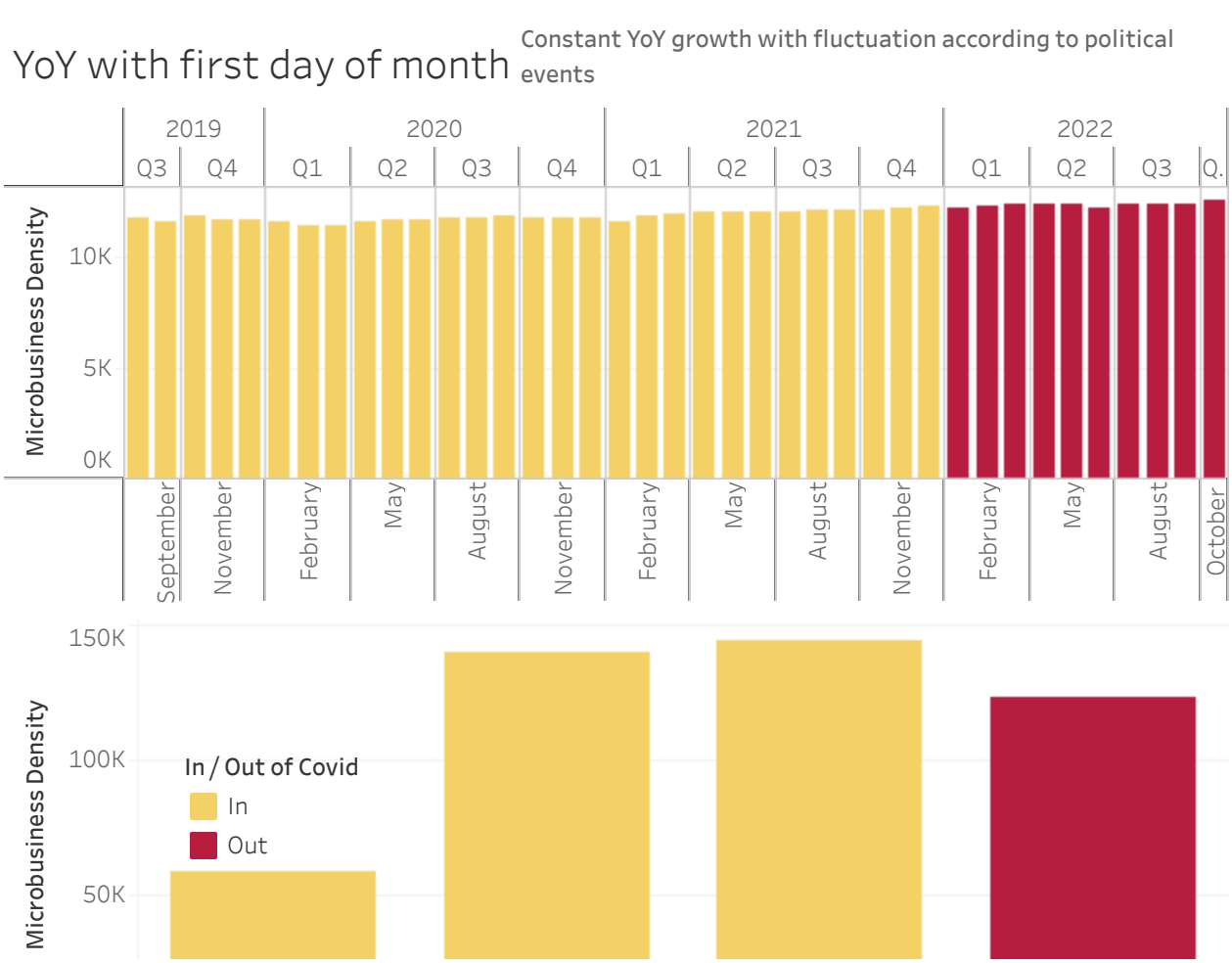
Out[8]:
```
0          1249
1          1198
2          1269
3          1243
4          1243
           ...
122260      101
122261      101
122262      100
122263      100
122264      100
Name: active, Length: 122265, dtype: int64
```

EDA

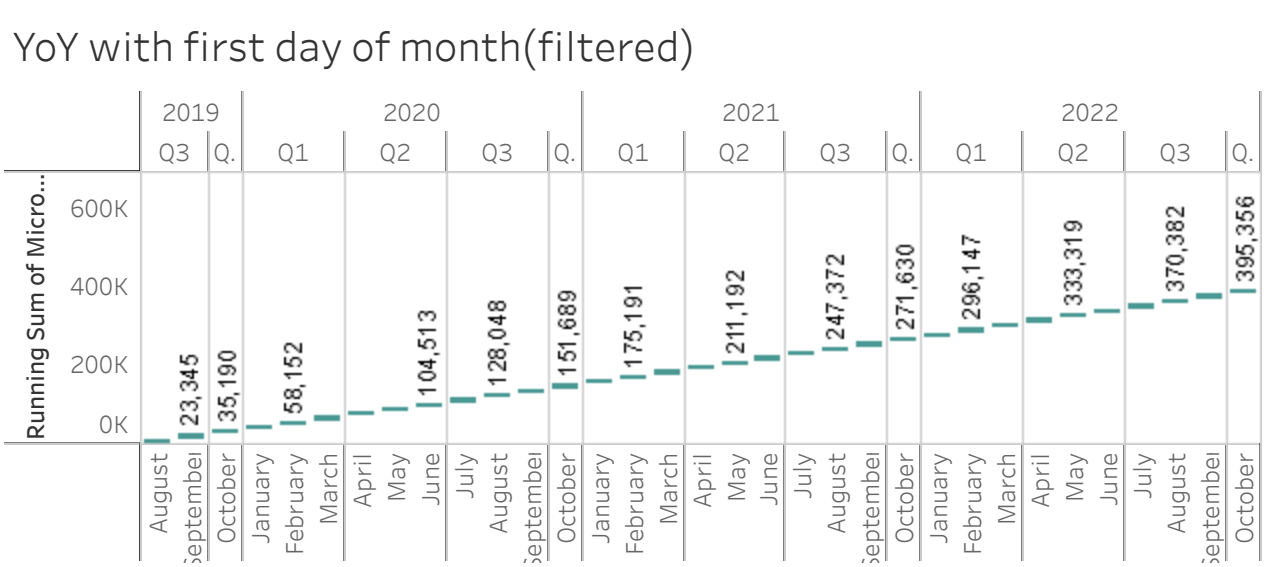
In [9]:
```
# Importing EDA libraries

import seaborn as sns
import plotly.express as ex
import matplotlib.pyplot as plt
```

YoY Increase in Density



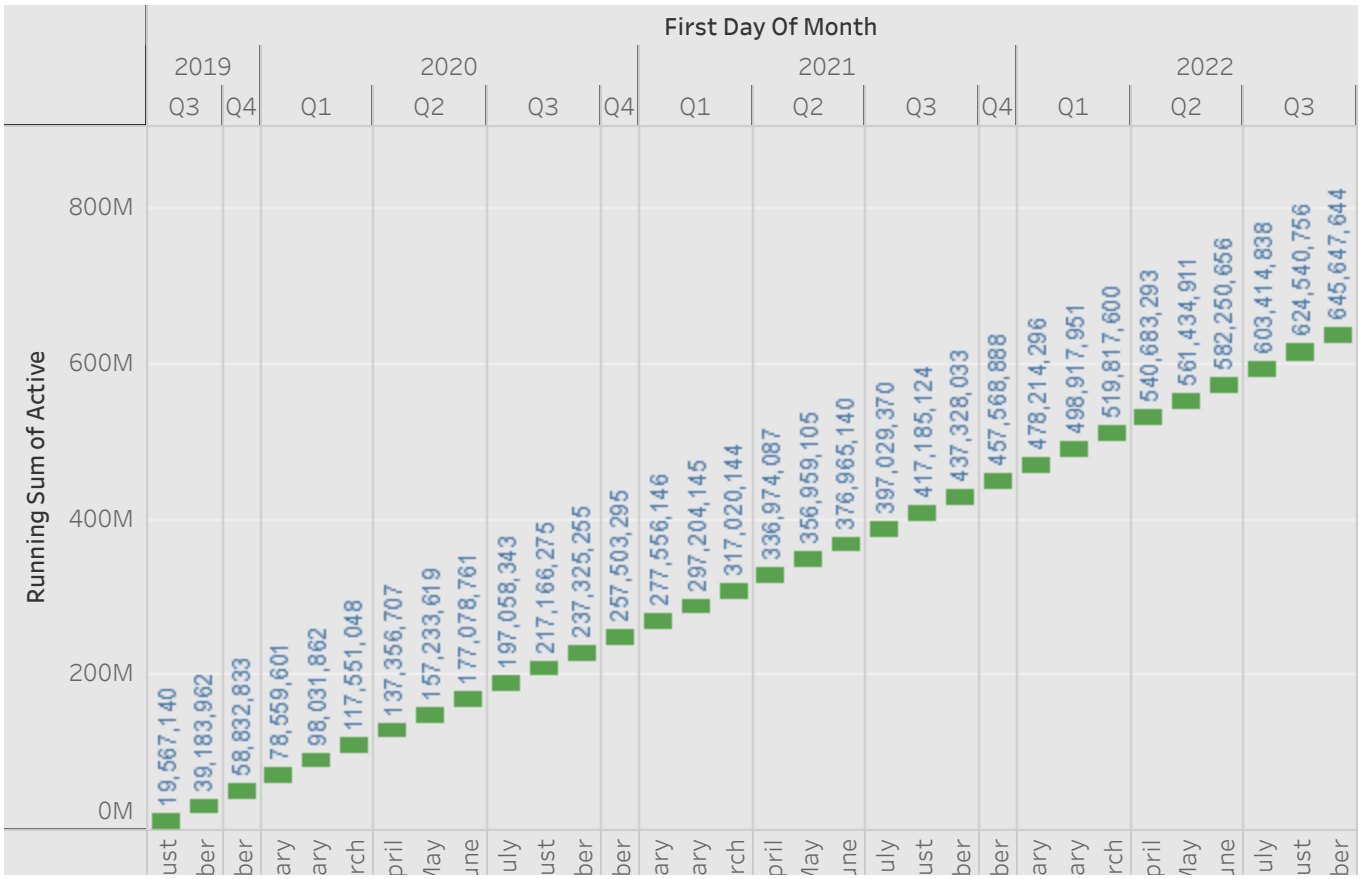
Data shows downfall for 2022 due to inusfficient data. Rest of the years have Nov and Dec Data, Where as 2022 doesnt



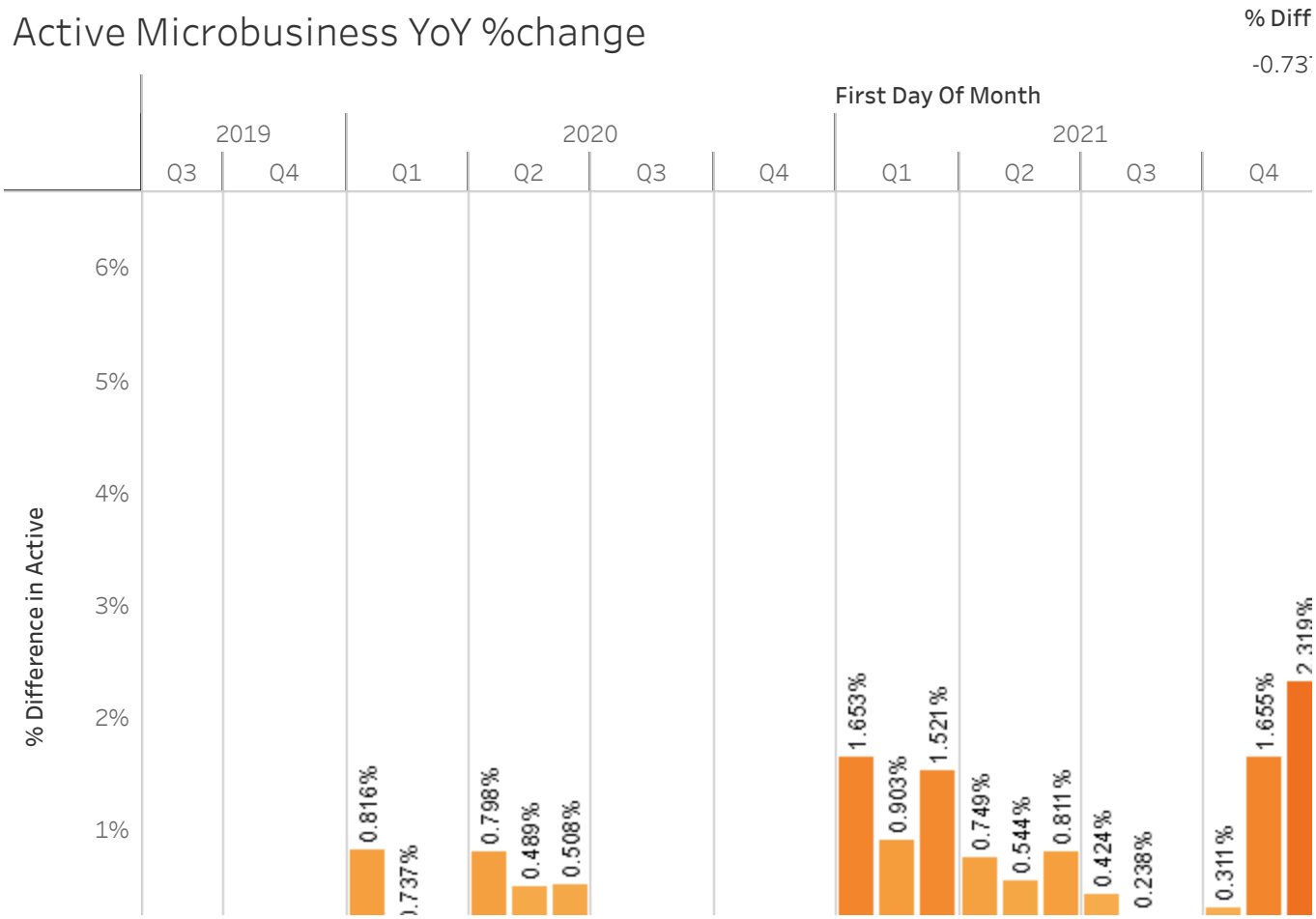
Filtered YoY with Density with constant months

YoY active Micro Business's

YoY with active MB

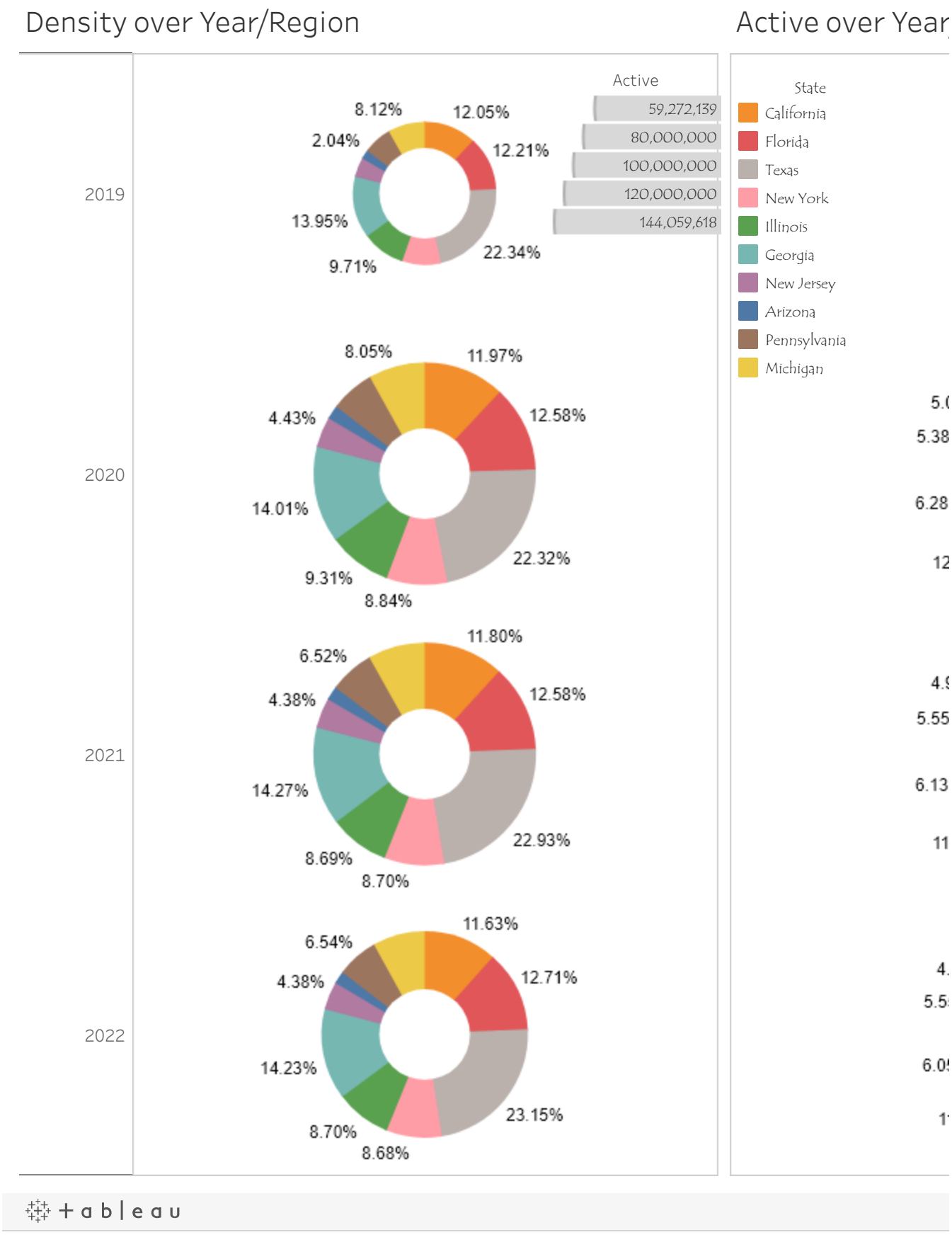


YoY Data shows linear growth for both Density and Active

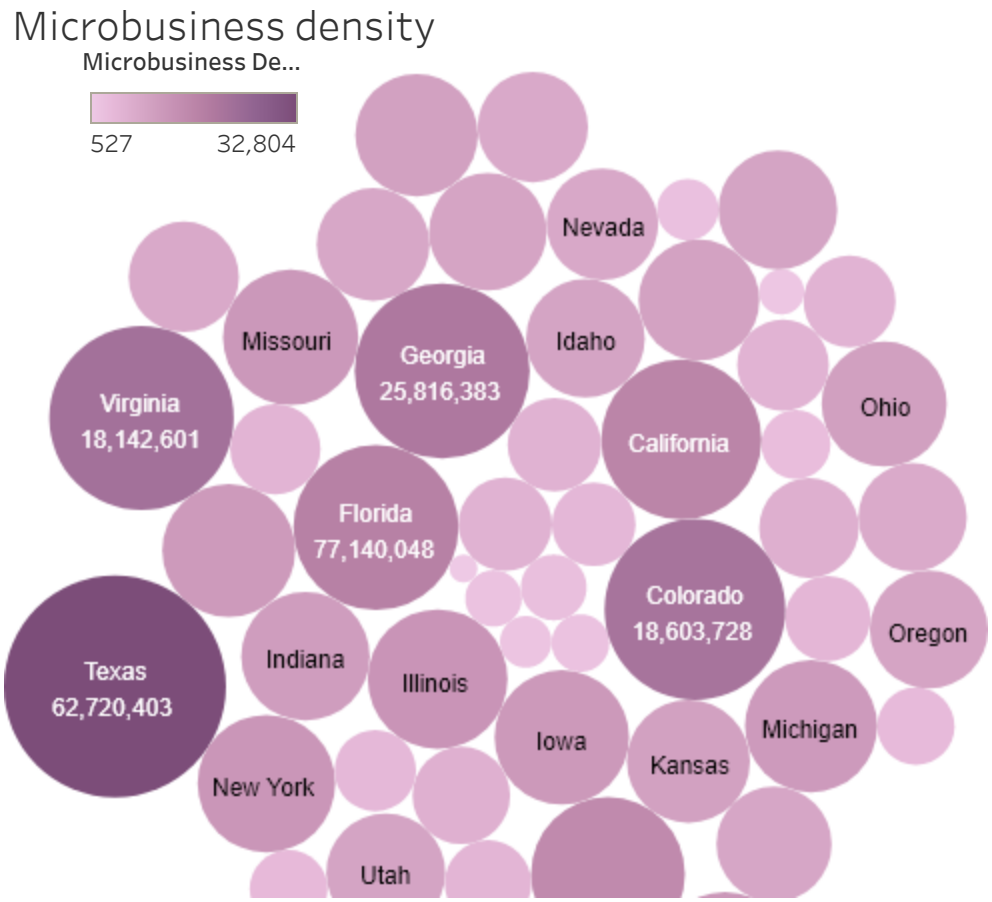


PCT Change shows huge drop in active business during Lockdown.

Region Divided by Years %



Microbusiness Density by Region



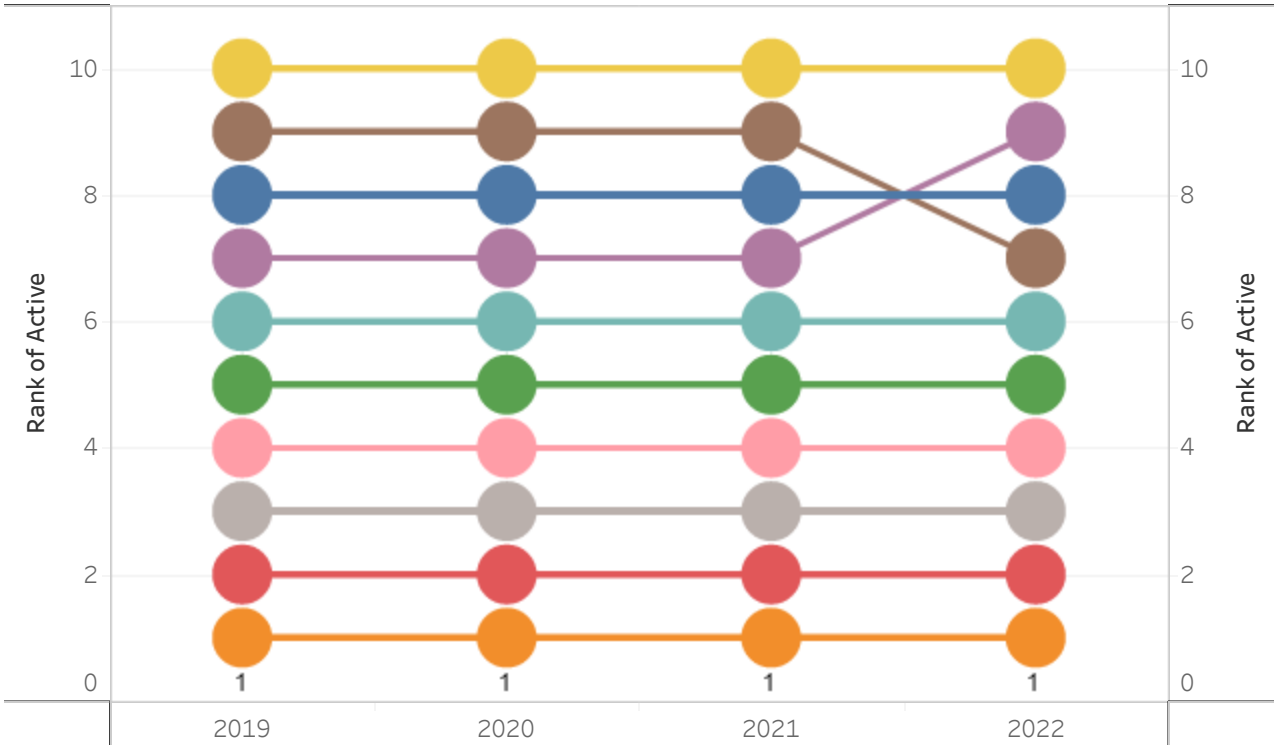
Active YoY View

Active YoY table By states

State	2019	2020	2021	2022	Grand Total
California	17,026,164	40,658,483	39,787,258	34,270,973	32,935,720
Florida	9,409,287	22,930,792	23,865,484	20,934,485	19,285,012
Texas	7,733,849	18,891,486	19,240,705	16,854,363	15,680,101
New York	7,130,895	17,314,683	17,197,204	14,843,324	14,121,527
Illinois	3,768,132	9,001,354	8,824,321	7,580,476	7,293,571
Georgia	3,154,362	7,706,530	8,002,439	6,953,052	6,454,096
New Jersey	2,938,662	7,158,514	7,077,564	6,118,053	5,823,198
Arizona	2,845,716	6,856,825	6,963,898	6,137,148	5,700,897
Pennsylvania	2,740,467	6,663,111	6,917,961	6,252,963	5,643,626
Michigan	2,524,605	6,101,654	6,182,784	5,279,863	5,022,227
North Carolina	2,427,068	5,945,528	6,057,550	5,266,977	4,924,281
Washington	2,412,753	5,819,976	5,896,536	5,083,874	4,803,285
Colorado	2,295,264	5,830,372	5,583,871	4,894,221	4,650,932
Virginia	2,293,071	5,574,849	5,525,478	4,749,203	4,535,650
Nevada	2,101,073	5,258,987	5,628,437	4,628,231	4,404,182
Massachusetts	2,235,385	5,396,861	5,240,328	4,513,970	4,346,636
Ohio	2,172,647	5,247,016	5,257,695	4,575,673	4,313,258

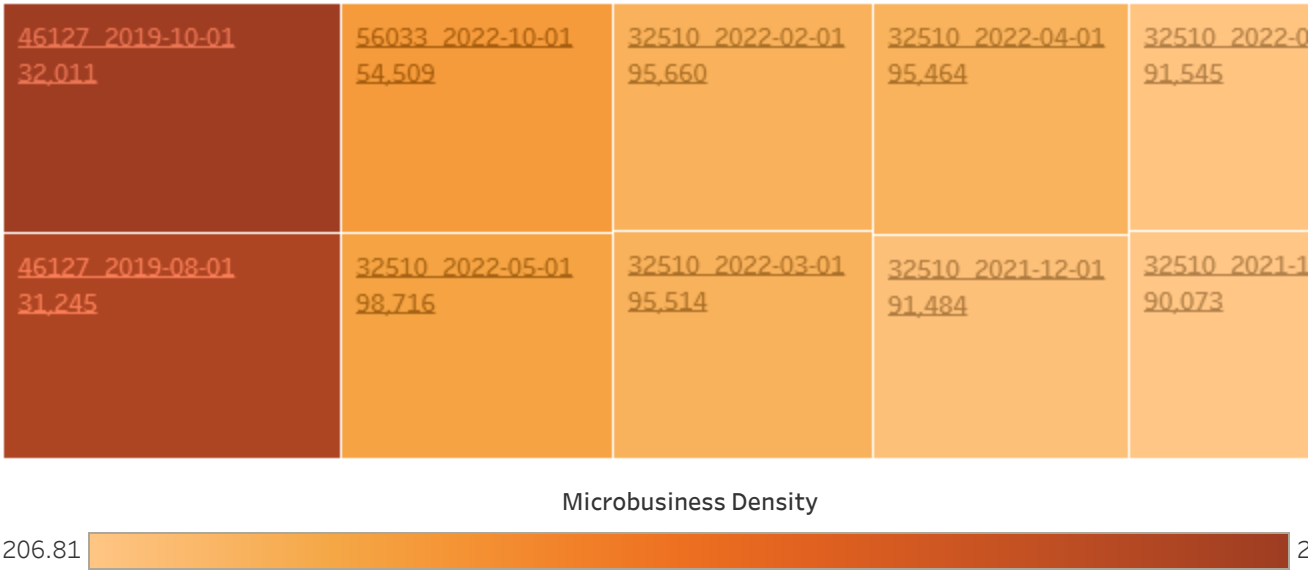
Active
113,7
Active

Active YoY Rank By states



Row ID with Density

Row ID with density



In [ ]: