CS506 Report

Small Businesses (Team 5)
Simon Cen
Reshab Chhabra
Dave Godfrey
Xiaojie Li

TABLE OF CONTENTS

Preliminary Information

Abstract

The purpose of this project is to explore the small businesses in District 4. With more context into the small business landscape, the current gaps that exist can be identified and analyzed. The goal is to see what will attract more small businesses to the area while understanding the current issues small businesses face.

Checklist

- ☐ Examine Main Streets Business Data, Foot Traffic Data, Spending Data, and District Shape files to inspire ideas to depict what a small business in District 4 can improve on. We will also be touching on a few proposed basic questions.

- ☐ Answer two fundamental questions:
    1. What businesses are over-represented?
    2. What businesses exist for each district?

- ☐ Answer a key question that inspires our extension project: Which districts have the best spending rate?

- ☐ Expand from requirements to create an extension project. We will cover the following:
    - Spending rate in district 4 versus other districts
    - Foot traffic in district 4 versus other districts
    - A conclusion statement, covering what we believe small businesses in District 4 should have, and what types of businesses District 4 would need to prosper more.
    - Future proposals, where we propose next steps for this project and acknowledge ideas we've considered but did not pursue due to data limitations.

Preliminary Data Analysis

Main Streets Business Data

First, we can examine the covariates and their relation to examine any patterns. We briefly investigated the following data columns (in order of analysis)

NAICS_2017_2digit_desc, NAICS_2017_6digit_desc, employment_buckets, street_name, latitude & longitude, mainstreet.

Looking into Business Type Popularity

We decided to investigate what business types are the most popular among the small businesses in the Boston area

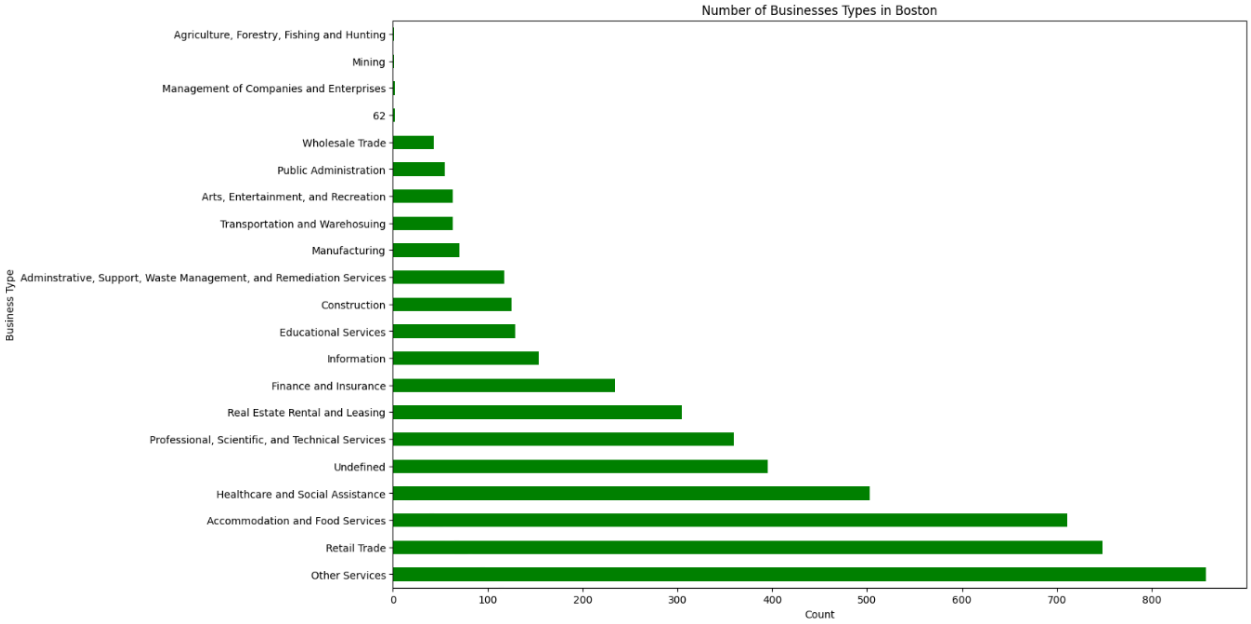


Figure 1: Number of business types in Boston for each business type

We can see here that Retail Trade, Food services and Healthcare appear to be the most popular. When it comes to promoting small businesses to relocate to this city, we can focus on these top business groups, as these appear to be more popular and larger in number. Hence, by attracting these business types, we can proportionally attract more businesses.

In contrast, we can also focus on the smaller ones and tell these small businesses that they could play a huge role in the Boston area if they decide to move here. For example, if there is no plumbing in the city, we should definitely have a plumber since that would help a ton.

We felt these the names Retail Trade was a bit too ambiguous, so we went forward with investigating the larger descriptions, NAICS_2017_6digit_desc. As these descriptions are larger, plotting them would result in too many elements. Hence, we will only display the top 10.

```
NAICS_2017_6digit_desc
Full-Service Restaurants                                        608
                                                                395
Offices of Real Estate Agents and Brokers                       209
Beauty Salons                                                   208
Religious Organizations                                         109
Other Social Advocacy Organizations                             109
Offices of Lawyers                                              102
Barber Shops                                                     86
Offices of Physicians (except Mental Health Specialists)         78
Offices of Dentists                                              73
Name: NAICS_2017_6digit_desc, dtype: int64
```

Figure 2: Top 10 business types in the Main Streets Business Data

We believe it could be insightful to view businesses' employment numbers by location, but first we will look at the number of employees, or the employment buckets in order to examine which employee count is the most often occuring.

We noticed that full-service restaurants, real estate, and beauty salons are quite prevalent, so we will later investigate how densely close they are in terms of location.

Looking at Employment Numbers

We believe it could be insightful to view businesses' employment numbers by location, but first we will look at the number of employees, or the employment_buckets.

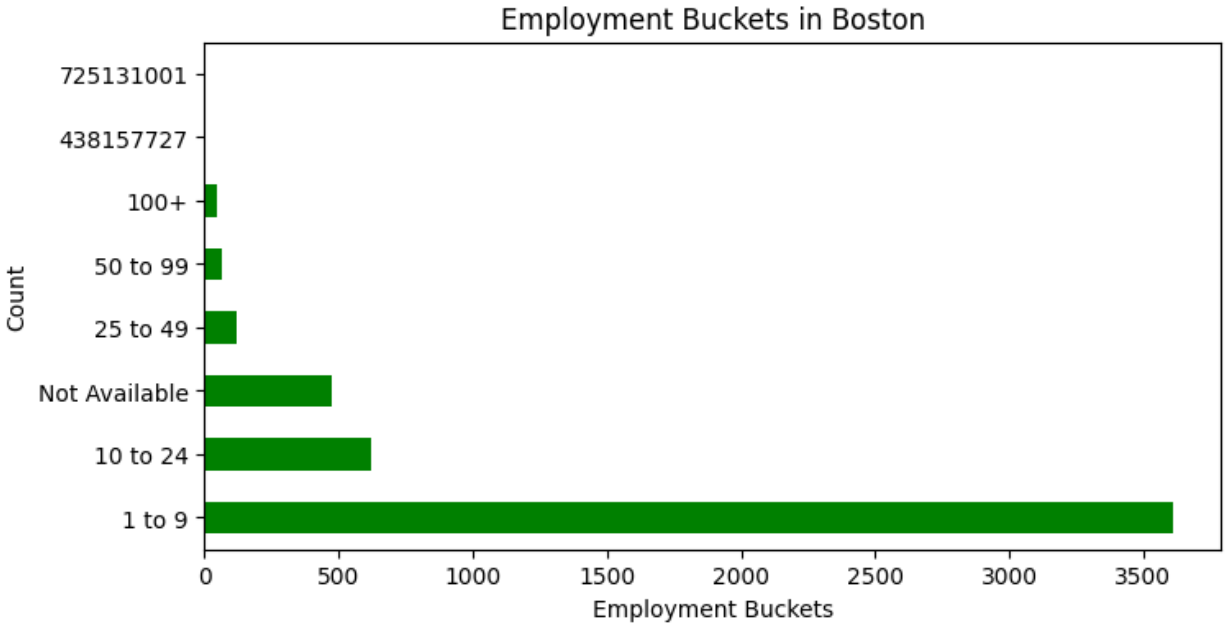


Figure 3: Distribution of the number of employees for small business in the Main Street Business Data

We notice that the dataset has a lot of businesses centered around 1-9 employees. We can potentially focus on this `1-9` bucket and view it by location.

Looking at Street Name

Our team believes looking at the Street Name (`street_address`) by the business type (`NAICS_2017_6digit_desc`) could be insightful because small businesses could potentially gain attraction and popularity by relocating to a specific street.

First, we will look at how dense a street name is in terms of number of small businesses

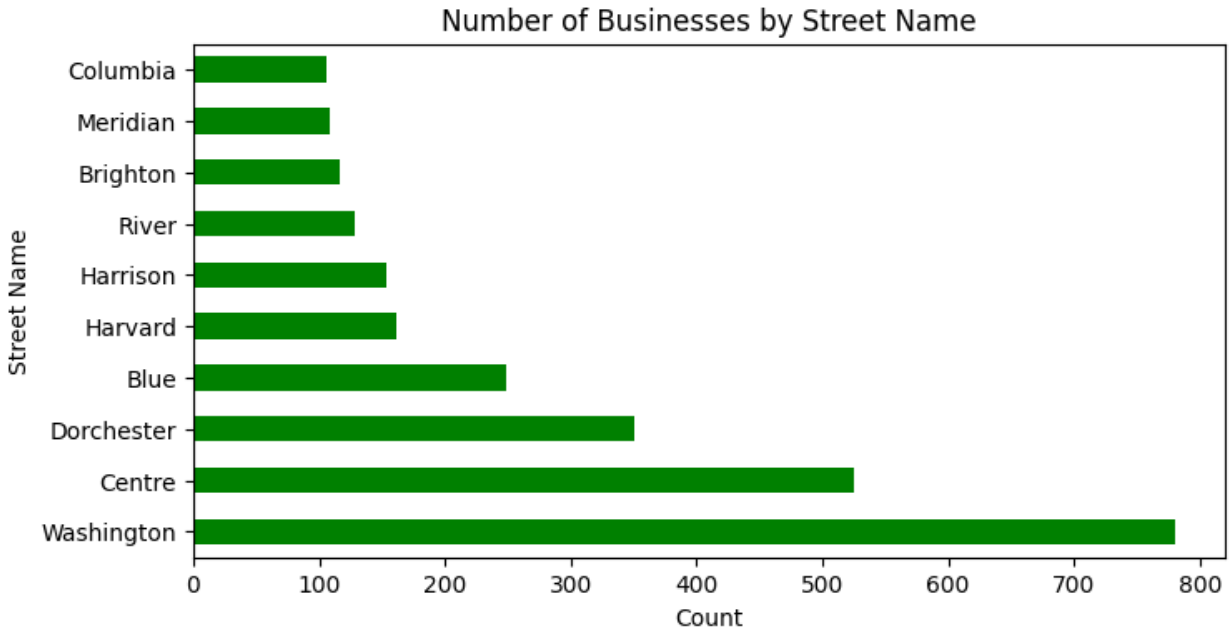


Figure 4: Count of Businesses by Street Name

We found it very peculiar that a single street like `Washington` can have up to about 800 businesses on it. We later googled the street to find that it is about 37 miles – it consists of 5 different streets – insinuating that street name may not be a great indicator.

It would have been interesting if we computed the density of how many businesses per mile, but we decided that looking at the `latitude` and `longitude` can be more insightful.

Looking at Latitude and Longitude

We believe this is the key indicator to view what small businesses are overrepresented. To begin, for our preliminary, we will simply graph the latitude and longitude of each small business in Boston. We will also have the districts colored for better visualization.

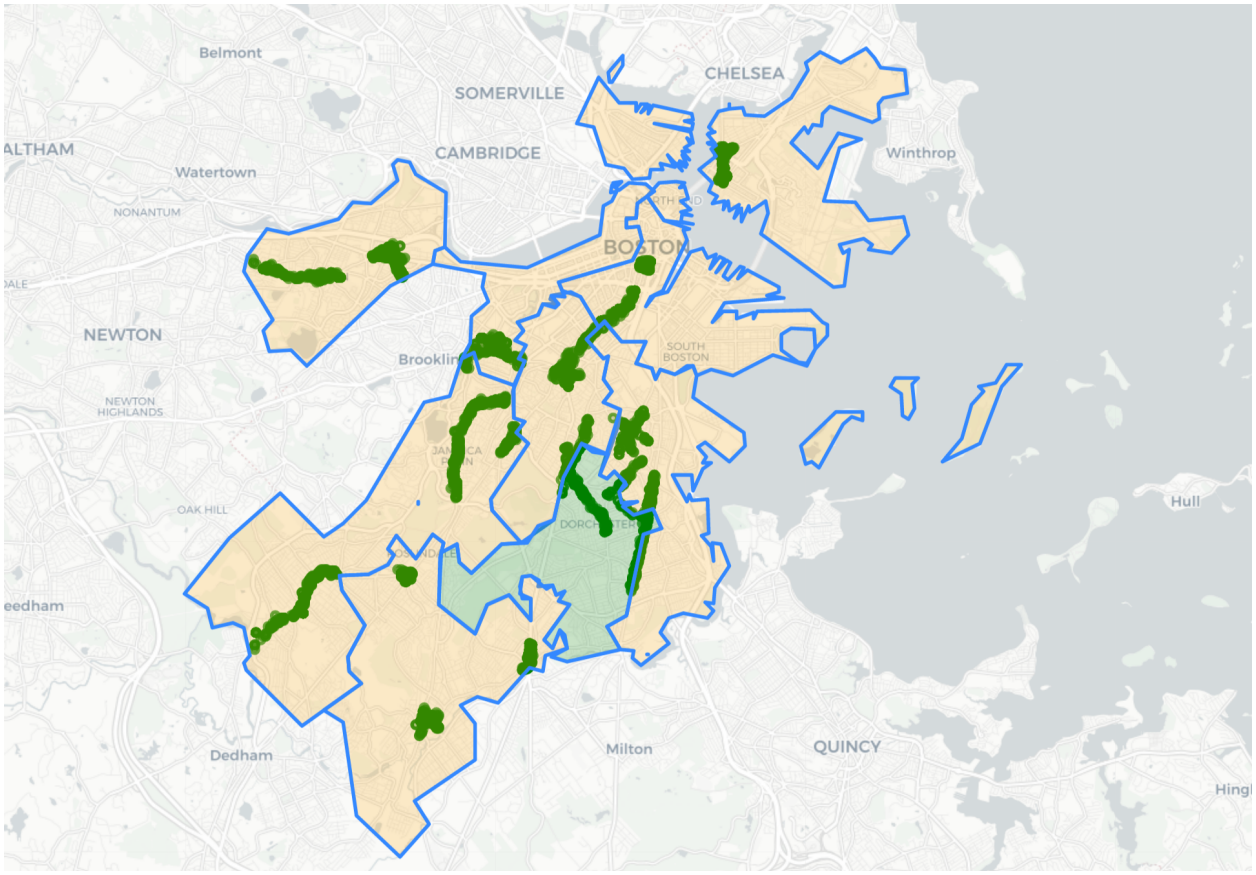


Figure 5: Small business locations plot with each districts sectioned by color

We have District 4 area plotted in *green* and the others plotted in *orange*. Here, we noticed that in District 4, small businesses are primarily located on 2 streets. We hope to further investigate which businesses they are and which ones are over or underrepresented. We also hope to narrow down the datasets by whether they are in District 4, as we can abstract what types of businesses are in each district.

Looking at Mainstreet

Our group thought, similar to street names, that the mainstreet could be important to small businesses, as maybe they can relocate their business here based on traffic. To begin, we will plot mainstreet by count. *We hope to narrow this down to only District 4 later on and look into business types*.

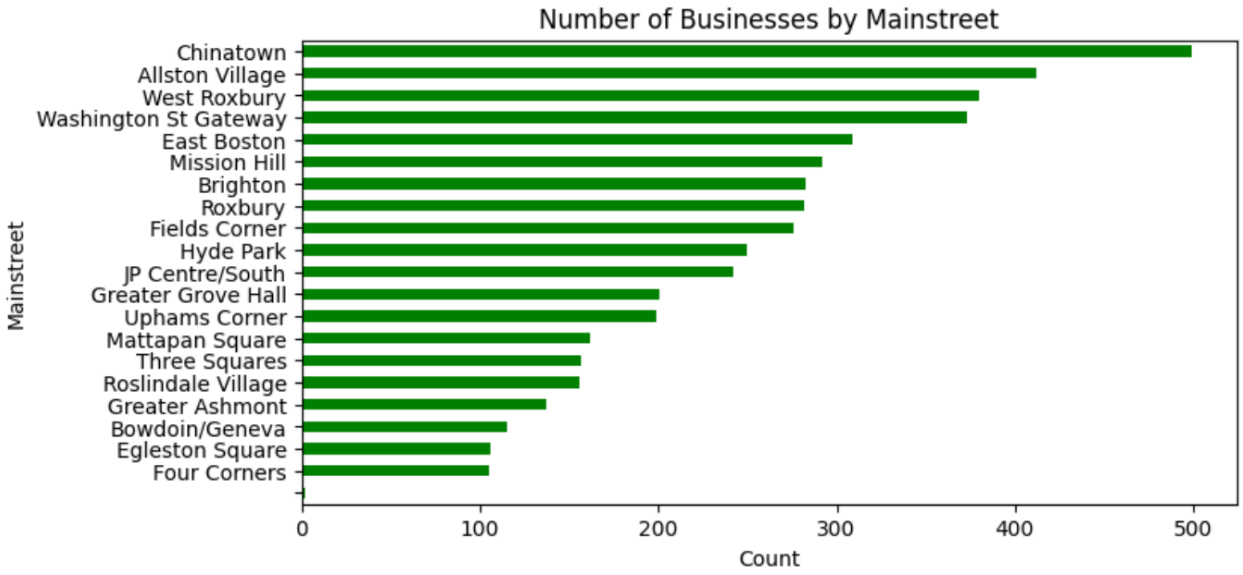


Figure 6: Number of businesses by each mainstreet

Chinatown appeared to be on top, but this is not district 4. The goal here though was to see if large numbers occur, which there seems to be. Once again, we hope to narrow our data down to district 4 to further investigate.

Foot Traffic Data

As the only content in this data set is foot traffic time series analysis for each mainstreet, we believe that we can model foot traffic by mainstreet to help view which areas are the most popular.

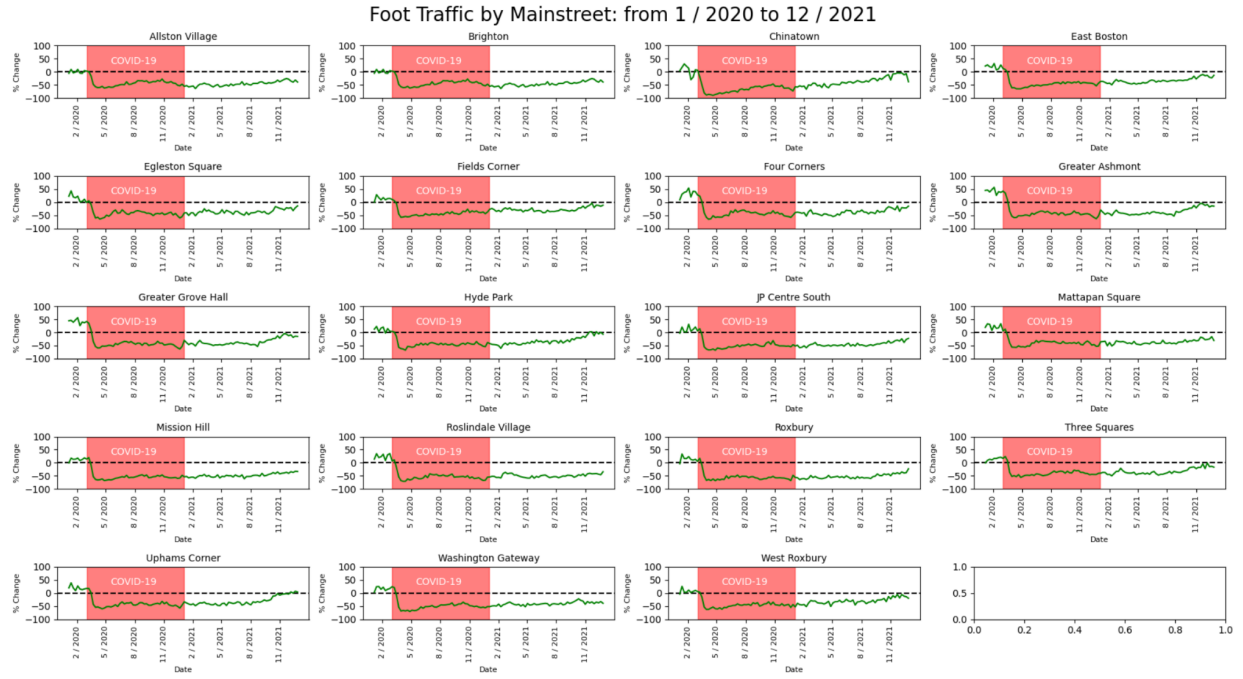


Figure 7: Foot Traffic by mainstreet from January 2020 to December 2021

We decided to label COVID-19 in the graphs because we noticed the huge decrease around March 2020. We defined the interval to be from March 2020, to when Vaccines began to be distributed -- Jan 2021. We noticed after we got vaccines, the overall foot traffic in Boston for each of these mainstreets began to increase.

We would have liked to see the initial foot traffic values to help quantify how popular Boston is, so we decided to look it up and plot the actual foot traffic values. This can be useful in the future when providing metrics on how busy Boston really is.

In terms of comparing each mainstreet to another, we see a similar pattern: a drop during covid, and a slight increasing trend in foot traffic -- yes it is still decreasing, but the percent change is slowly approaching back to a positive number.

Spending Data

As the spending data has the same content as the foot traffic data – mainstreet and a time series analysis, we've decided to create a similar graph.

Figure 8: Spending rate for each mainstreet from January 2020 to March 2022

We found, in contrast to foot traffic, that the main streets' overall spending rate had higher variability and wider differences among each main street. For example, we can see the areas East Boston, Fields Corner, Hyde Park, Greater Grove Park, and Mission Hill have a high spending rate after COVID-19's impact, but many other areas such as Four Corners are still going through the crisis.

What we would like to investigate more is *which main street area corresponds to what district, and to compare which district has higher spending rates than others*. This would allow us to see what district 4 can improve on to ultimately attract more small businesses.

Obtaining District from Longitude and Latitude

As we are currently focused on District 4, we've decided to investigate the district map data to label a longitude and latitude point by its associated District number. This will be key when we need to compare businesses given its longitude and latitude, and will be key when pairing main street with a district number. Here's an example of the dataframe showing some businesses that have had their districts added.

NAICS_2017_2digit_desc	estimated_employment	employment_buckets	mainstreet	district
Retail Trade	1	1 to 9	Brighton	9
Undefined	0	Not Available	Brighton	9
Healthcare and Social Assistance	2	1 to 9	Brighton	9
Other Services	3	1 to 9	Brighton	9
Information	13	10 to 24	Brighton	9
...	...	...	...	...
Retail Trade	9	1 to 9	Mission Hill	8
Undefined	0	Not Available	Mission Hill	8
Undefined	0	Not Available	Mission Hill	6

Figure 9: Added district number column to the original Mainstreet Business data set

Key Questions

We will look into the following questions in this section:

- What businesses exist? E.g. restaurants, barbershop
- What businesses are over-represented (e.g. within one mile of Codman square there are 30 barber shops)

What Businesses are over-represented?

To investigate this, we will take our business plotted by location, and now add a color to the marker to indicate what business type it is.

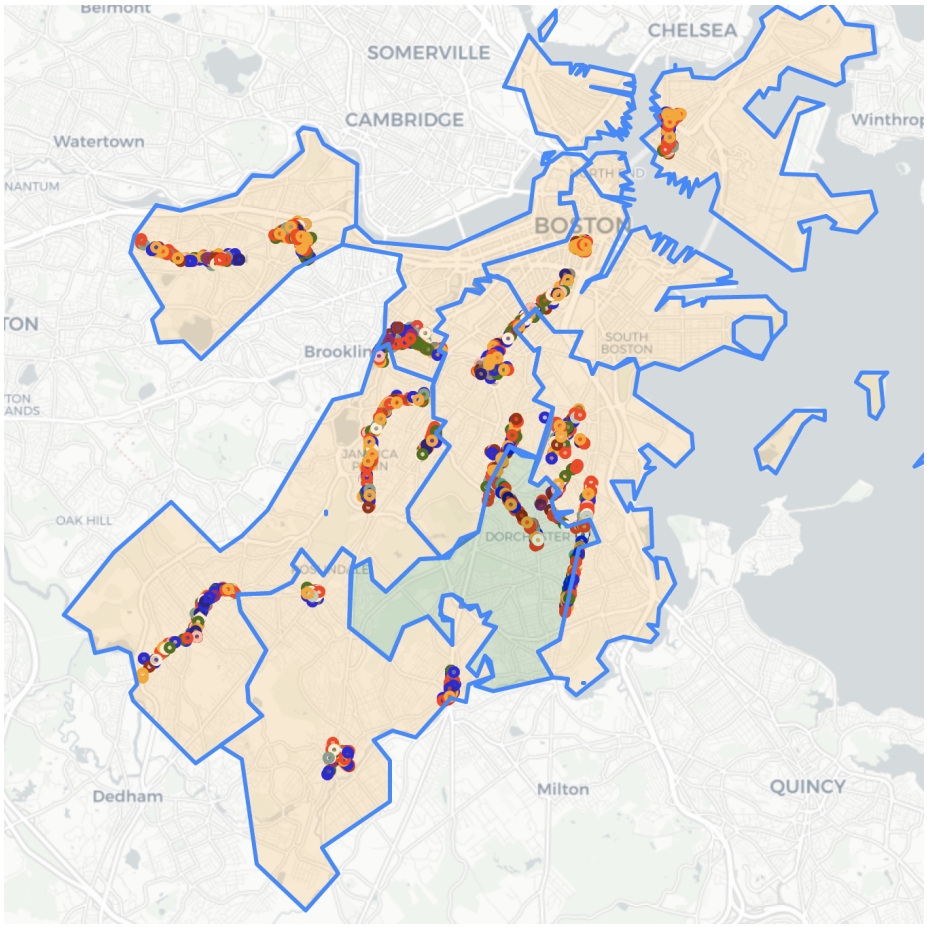


Figure 10: Colored business types plotted by location, where each district is sectioned.

We noticed that there are too many colors, so instead of looking at all businesses, we will rather look at the top 5 most prevalent ones to gather more information on overrepresentation. Moreover, the bottom ones are more likely to be less represented.

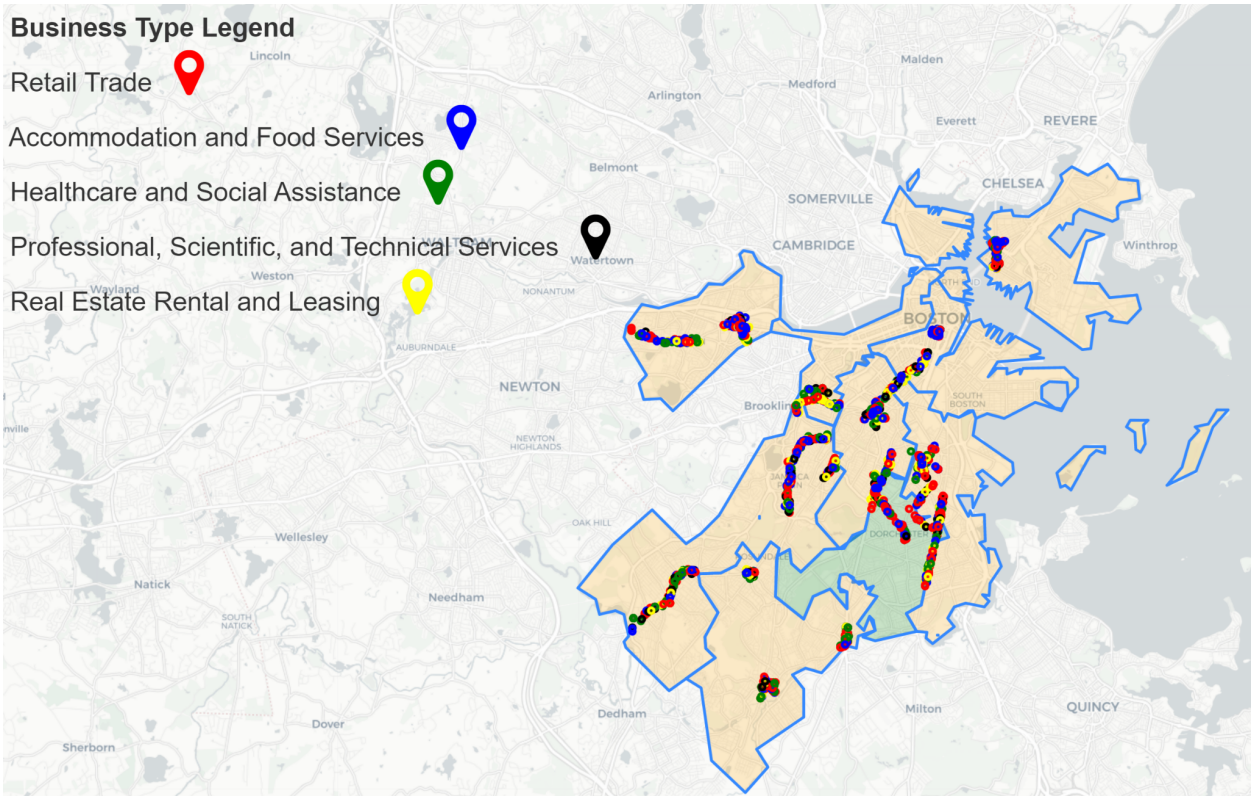


Figure 11: Top 5 business types plotted by location, where each district is sectioned

With more prominent colors, we can see the following that are **closest** to each other:

- District 4 seems to have a lot of reds, or Retail Trade that are close to each other and a severe lack of blue, Accomodation and Food Services, compared to other areas like North End and Jamaica Plain. As these are food services, we can see a lack of restaurants in District 4.
- We can also see a lack of green density, Healthcare and Social Assistance, in District 4. Having more of these can promote more Bostonians living in this area and the potential of having more small businesses come in. This is because the business owners can rest well, assured they have health care nearby.
- The color black, Professional, Scientific, and Technical Services, appear to be equally dense among all districts. The same followed for Real Estate Rental and Leasing.

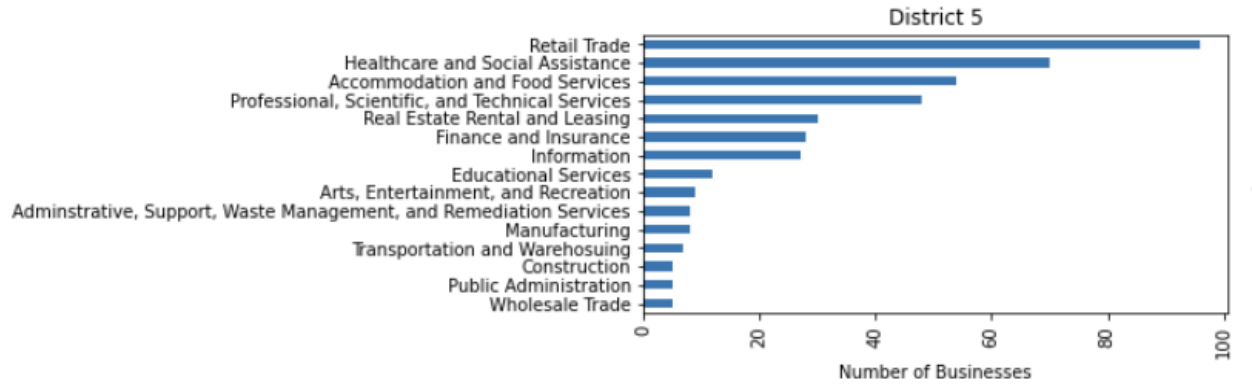
Next Steps

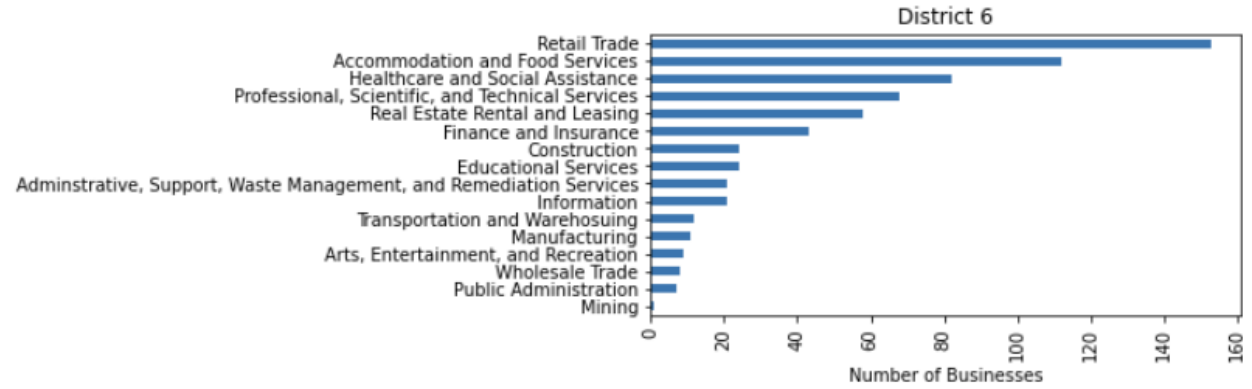
To investigate what more small businesses in District 4 should really have, we want to graph Spending Rate by District to see which districts have the highest spending rate. We can then learn from the higher spending rates area and say what District 4 can have. The only limitation would be socioeconomic factors, which we do not fortunately have.

What Businesses Exist?

We decided to look at what business types exist for each district and compare them to the required District 4 in addition to our Preliminary Analysis.


District 4
Retail Trade
Healthcare and Social Assistance
Accommodation and Food Services
Professional, Scientific, and Technical Services
Information
Finance and Insurance
Real Estate Rental and Leasing
Construction
Educational Services
Transportation and Warehosuing
Manufacturing
Adminstrative, Support, Waste Management, and Remediation Services
Wholesale Trade
Arts, Entertainment, and Recreation
Public Administration
0
10
20
30
40
50
60
70
80
Number of Businesses


District 5
Retail Trade
Healthcare and Social Assistance
Accommodation and Food Services
Professional, Scientific, and Technical Services
Real Estate Rental and Leasing
Finance and Insurance
Information
Educational Services
Arts, Entertainment, and Recreation
Adminstrative, Support, Waste Management, and Remediation Services
Manufacturing
Transportation and Warehosuing
Construction
Public Administration
Wholesale Trade
0
20
40
60
80
100
Number of Businesses


District 6
Retail Trade
Accommodation and Food Services
Healthcare and Social Assistance
Professional, Scientific, and Technical Services
Real Estate Rental and Leasing
Finance and Insurance
Construction
Educational Services
Adminstrative, Support, Waste Management, and Remediation Services
Information
Transportation and Warehosuing
Manufacturing
Arts, Entertainment, and Recreation
Wholesale Trade
Public Administration
Mining
0
20
40
60
80
100
120
140
160
Number of Businesses


District 7
Retail Trade
Accommodation and Food Services
Healthcare and Social Assistance
Professional, Scientific, and Technical Services
Finance and Insurance
Real Estate Rental and Leasing
Adminstrative, Support, Waste Management, and Remediation Services
Educational Services
Information
Construction
Public Administration
Transportation and Warehosuing
Manufacturing
Arts, Entertainment, and Recreation
Wholesale Trade
0
10
20
30
40
50
60
70
Number of Businesses

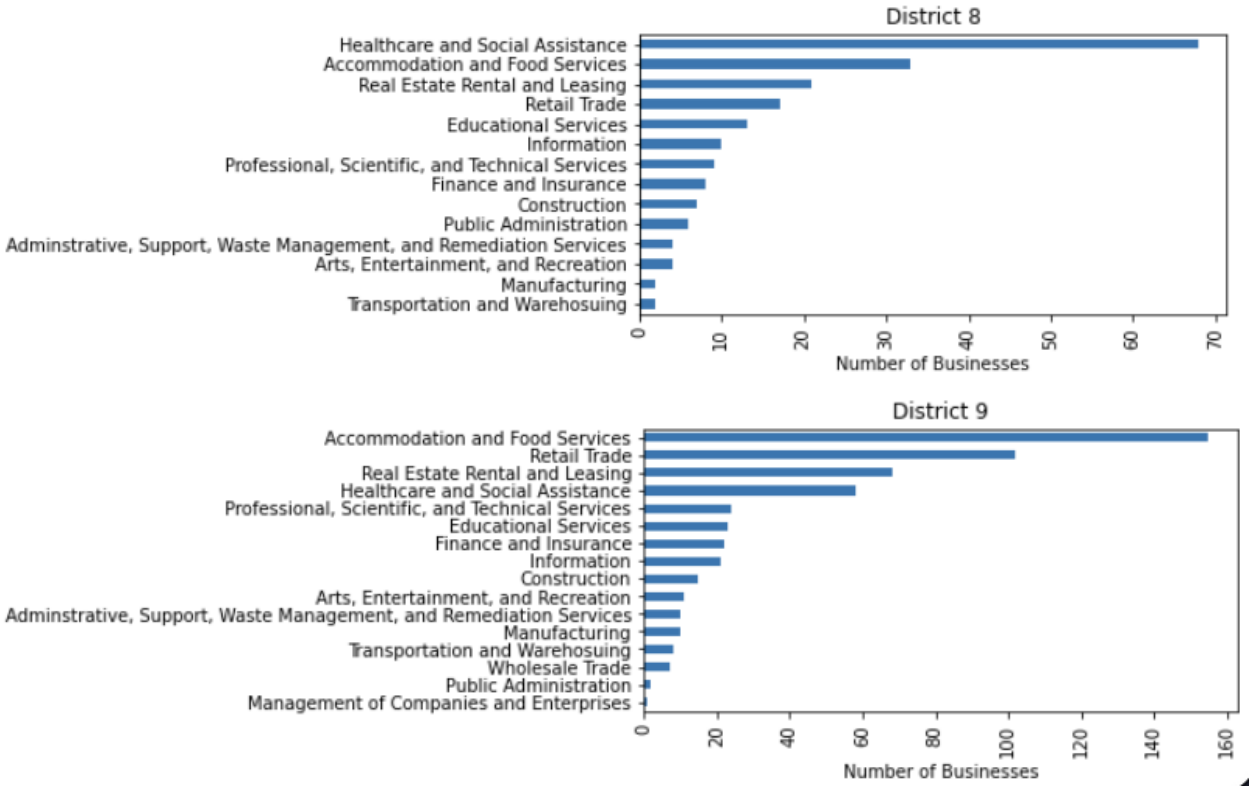


Figure 12: A histogram for the most popular business types for each district

District 4 appears to have a lot of Retail Trade in terms of number and a decent number of Healthcare and Social Assistance. What we can see lacking slightly though against other districts is Accommodation and Food Services, which District 2 and 9 have a lot of. We believe this can help enhance District 4 more.

We also noticed a difference in Educational Services. For example, District 8 has slightly more educational services than District 4.

Lastly, a huge thing we noted is that District 4 has one of the least number of Finance and Insurance compared to the other Districts. This could be vital in telling District 4's economic state and the socioeconomic culture of that area. Perhaps we need more small businesses to come to promote financial assurance, which would definitely bring in profit.

Additional Key Question: Which Districts Have the Best Spending Rate?

We believe checking what districts have the best spending rate will add significance. For example, if a district has a very high spending rate, we can say that a lot of money is spent there. Then, we can see what business types or what fields exist.

Creating Association Between Mainstreet and District Number

First, we noticed that the spending data graph has main streets as the columns. Hence, to help extrapolate district numbers, we will check what mainstreets are associated with which districts.

districts	mainstreet
[9]	Allston Village
[3, 4]	Bowdoin/Geneva
[9]	Brighton
[2]	Chinatown
[1]	East Boston

Figure 13: A subsection of the created dataset: each mainstreet is listed by on each district

Now we will combine data points. Consider the mainstreet `Uphams Corner`. We see that district 3 and district 7 are associated with it, so we will add all of those data points from this mainstreet to both districts 3 and 7 plot.

	GREATER GROVE HALL	MISSION HILL	ROXBURY	UPHAMS CORNER	WASHINGTON GATEWAY
0	11.0	54.0	5.0	7.0	8.0
1	6.0	32.0	5.0	-4.0	0.0
2	46.0	77.0	18.0	14.0	18.0
3	21.0	20.0	24.0	3.0	12.0
4	16.0	30.0	18.0	8.0	12.0

Figure 14: the spending rate values from the Spending Data for each mainstreet belonging to District 7

We notice that a single district can be associated with multiple mainstreets. Hence, the challenge is posed: How do we combine them? Well, since they are all part of district 4, we will average each value. The limitation here is that when we average, we assume that there are equally the same number of samples for each week taken for each location. This is an unfortunate limitation we face, but it is worth acknowledging. On the brighter end, we believe the data for each week follows the law of large numbers, so we believe averaging them will be a sufficient indicator.

	district_1_avg	district_2_avg	district_3_avg	district_4_avg	district_5_avg	district_(
0	17.0	8.0	5.0	15.2	7.333333	
1	12.0	0.0	3.0	13.8	1.666667	
2	27.0	18.0	15.8	34.2	20.333333	
3	14.0	12.0	-1.0	10.0	8.666667	
4	16.0	12.0	1.8	13.6	9.333333	

Figure 15: District spending data subset, averaged by all involved mainstreets for each district

These averages will be used to draw the line, but to show the multiple mainstreets involved for a single district, we will plot them as points in the graph.

Putting the Dataset Together

Lastly, we will group all these districts to create our desired dataset

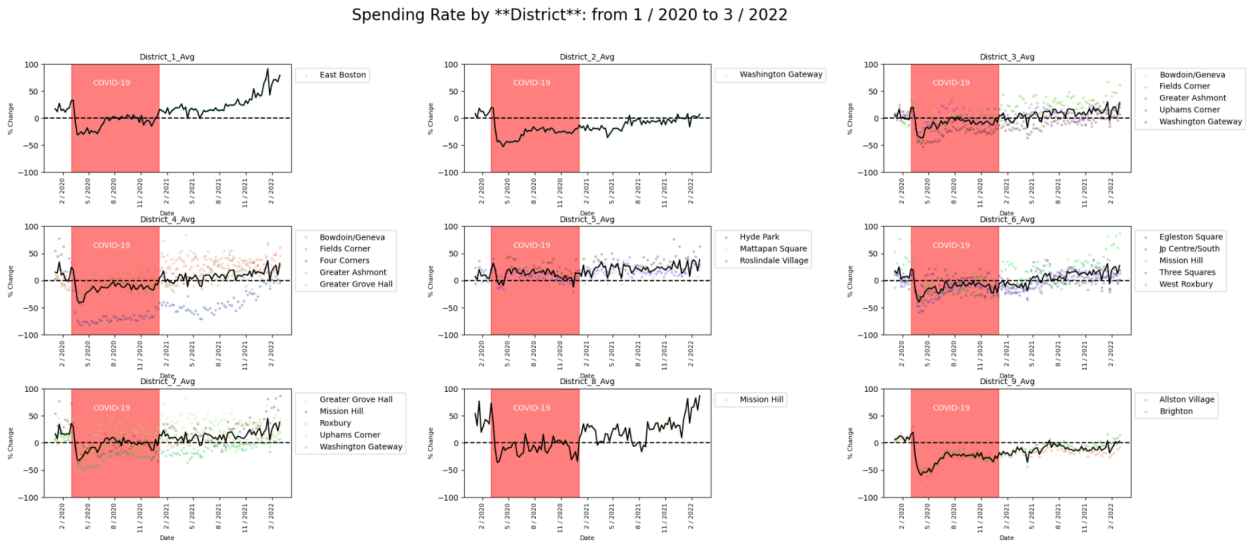


Figure 16: Spending Rate by District from January 2020 to March 2022

District 1: We can see District 1 has strongly recovered from the pandemic and is booming with a large spending rate increase. We believe our interest, District 4, can learn from what District 1 does. To look into this, we will use the previously founded business types in District 1 and will compare this into our extension project.

District 2: Nothing significant to note in spending rate besides its slight recovery from the pandemic.

District 3: We can see a larger spread in values, but each of the mainstreets associated have a pattern: they all share a very slight increase in spending rate, so there isn't anything significant to note here.

District 4: District 4, on average, is similar to District 3 in behavior, but what is noticeably different among the other districts is the large spread of values. We have mainstreets like Four Corners that faced a massive loss in spending rate, but are fortunately climbing back up. We can also maintain a positive and the classic slight increase in spending rate.

District 5: What was most notable was how District 5 maintained a consistent behavior despite COVID-19 taking place. We can look into how District 5 maintains its consistency by using the provided data, but note we are limited in data sourcing. This consistency we hypothesize is important to small businesses --- having consistent and long time customers are vital.

District 6: same comments as district 3

District 7: same comments as districts 3 and 6, except we can also see a large spread here. It is worth noting too that Mission Hill is included in District 7, but it is only part of District 8. The reason this is of significance is because Mission Hill also is a high climber in spending rate: it is almost at 100% like District 1 is and exponentially increased post-pandemic.

District 8: Same comments as District 1

District 9: District 9 seems to be the district with, on average, the largest overall decrease during the pandemic. District 9 additionally seems to be recovering slowly, but it is worth noting what type of businesses are in district 9 to ultimately state what district 4 should have maybe less of.

District 4 Improvements

We can definitely see that on average, district 4 maintains a healthy spending rate level, but it is definitely not on par with districts like 1 and 8. We can look into our other created graphs to see what District 1 and District 8 small businesses entail in contrast to District 4 (part of extension project). We can also look into how District 9 has a slow recovery rate compared to the others, and why this is the case. This will allow us to extrapolate what district 4 should have less of and what small business types will prosper the most in District 4.

Extension Project

Topic

Our topic is focusing on analyzing the popularity of small businesses in District 4 compared to the other districts

Points of Interest

We are interested in the following:

- spending in district 4 vs. other districts
- foot traffic in district 4 vs. other districts
- makeup of the businesses within district 4 compared to other districts

Spending In District 4 Versus Other Districts:

We did this for the Additional Key Question

Foot Traffic in District 4 Versus Other Districts

Similar to what we did with the spending rate, we want to see how foot traffic varies among other districts (we will also average the values similar to the spending rate). As we do not have this exact data, we will again extrapolate using mainstreet names from two datasets: the foot traffic one and mainstreet businesses.

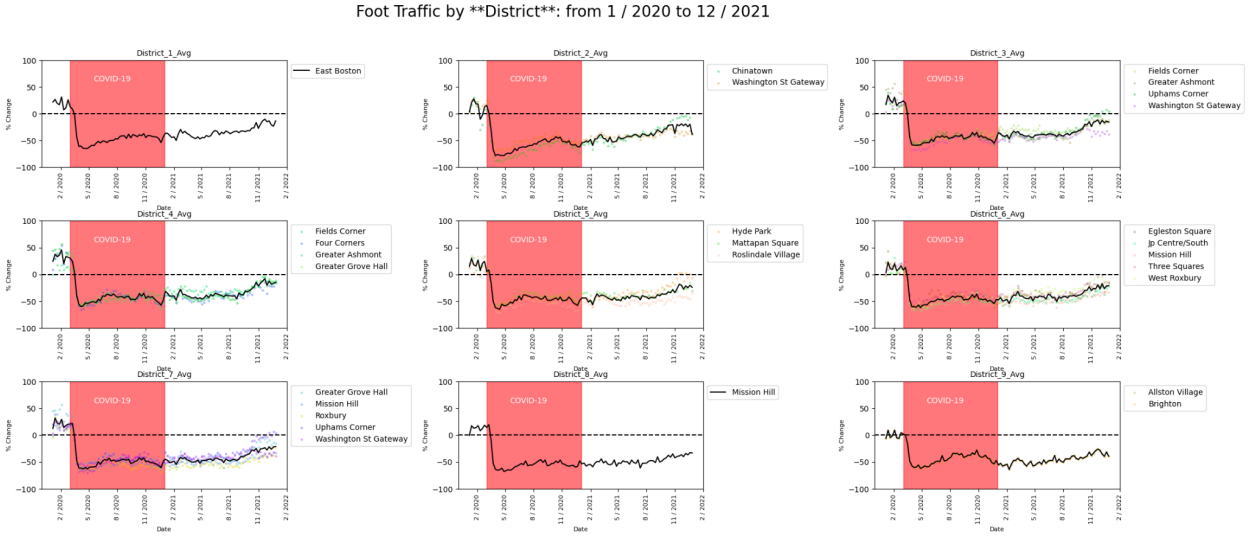


Figure 17: Foot Traffic by District from January 2020 to March 2022

All the districts follow a similar trend: there is a large dip in foot traffic at the start of the pandemic, then a slight recovery after the first vaccine date. Hence, there is unfortunately no distinctive key features to abstract from this data when it comes to figuring out what small businesses in District 4 need.

Business Types in District 4 Versus Other Districts

We did this when answering a fundamental question since we expanded this question beyond what was originally asked: what businesses exist? We expanded it by checking for each district to reach a larger conclusion.

Conclusion

District 4's Spending Rate

Considering the data given, a small business that is looking to grow in both size and revenue in District 4 would need to follow spending trends found in successful, high spending rate districts such as District 1 due to the fact that East Boston appears to have the largest and quickest spending rate increase. District 1 consists mainly of the area around Logan Airport, which means that there could be a significant number of people that come in and out of that area. However, District 4 lacks a hotspot similar to Logan Airport, which we hence believe that District 4 may not have the same advantage.

However, to help replicate this successful rate, District 4 can potentially have more attractions. For example, district 4 has a zoo (see https://bostonmainstreets.org/districts/four-corners-main-street/) Having more attractions like this can invoke more foot traffic and, by definition, would yield higher spending rate or revenue.

District 4's Foot Traffic

As the foot traffic rate for District 4 shows a parallel relationship among other districts, it still remains inconclusive mainly as the graph shows percentage changes. Hence, we believe that as the *rates are similar, this may not be the same for the actual number* for foot traffic. For example, we believe District 1 has a significantly higher foot traffic than District 4. But due to data limitations, we do not truly know whether District 4 has less foot traffic than other districts. Hence, more data would help conclude what District 4 can do to either increase or decrease their foot traffic when it comes to attracting more small businesses.

District 4's Small Business Location Density

District 4 has two concentrated streets densely packed with small businesses, whereas more successful districts – ones with higher spending rates – such as District 1 have a small cluster. Perhaps, it could be better for small businesses in district 4 to be concentrated too in terms of location.

District 4's Small Business Types

Businesses relating to Retail Trade, Healthcare, Social and Food Services serve as the top types for small businesses within District 4. They each have more than 40 units each within the District, and are fairly spread out as seen on the District map. However, when other Districts are looked at for their top types, such as District 8 and District 9, it becomes apparent that District 4 is lacking in types that seem to bring more traffic to 8 and 9, with Real Estate Rental and Leasing being one of them (this is simply a hypothesis, as we cannot guarantee higher traffic given only rates and not number).

How District 4 Can Improve Profit

Based on the data which highlighted the over representation businesses, we're aligning the most popular business within District 4 with the highest grossing businesses. With that being said, comparing District 4 with other districts, District 4 should attempt to replicate the same business model, such as having more Social and Food Services. However, an issue with an assumption arises that is explained in more detail in the next section.

What Businesses District 4 Should Have More Of

This is a bit difficult to answer with the amount of information we have currently. For example, an economically successful district might have many restaurants, but that doesn't necessarily mean District 4 should just have more restaurants. If there aren't enough customers or foot traffic to support that, then it would be meaningless. We can explore what types of businesses are most successful within District 4 and then compare it with other Districts to see if this specific type of business would actually improve District 4's economy.

Data Limitations

Spending Rate

We use spending rate as a metric, but it is not clear whether this spending rate is only for small businesses, or if it is for the whole mainstreet itself. We worry that this data also considers franchises which made us question its relevance. However, the rates we believe do vary by location overall, which is why we kept it. Moreover, their spending rate is measured by percent change, which is not a good indicator of how much was actually spent. Rather it gives us a relative model of how much each district continues to spend relative to how much they spent the previous week.

Foot Traffic

We cannot base our assumption solely on the foot traffic data for the District, and that's because it does not account for population density. Foot Traffic is also measured in percent change by week, similar to spending rate, which poses similar issues. It is difficult to make a conclusion whether or not a certain district has more foot traffic than another. Ideally, data collected in the future would have a raw estimate of the amount of people walking in each District, which would ultimately allow us to make more conclusive comparisons between the districts.

Updated District Shapes

For the scope of this project, we are still utilizing old District 4 lines, which could impact the decisions we are making or predicting with our project. A change in district shape would change the scope and range of the businesses we are currently analyzing, and naturally it would change the predictions we are making right now as well.

Future Proposals

Machine Learning Model

We believe a useful feature that can be made when predicting what a small business needs to prosper in District 4 is a machine learning model that compares the district's businesses with other districts that may be doing better economically. We could collect certain features such as business types, latitude and longitude, foot traffic, and cell phone traffic to compare the different districts and see what features help businesses succeed and use that to improve District 4's economic decisions.

However, one of the main problems with the scope of the project currently is that there is just not enough data to solve this broad question, and even if we do make a prediction it is likely to be inaccurate and not helpful, which is why we propose it for the future next steps. Another downside is that machine learning models require large computation power, which the desired company may not have. Thus, it is also worth noting another option: attempting to model small businesses' success with a differential equation.

More Datasets – Small Businesses' Values

As non-small businesses owners, it is difficult for us to perceive what small businesses truly value if they were to move to Boston, and what they want in terms of desires. Perhaps questioning small business owners via a poll or an online credible dataset can prove to be beneficial when analyzing datasets, as having this would allow us to know what we need to focus more on rather than having an ambiguous idea. This would also allow us to have more insight on the features that help small businesses succeed. For example, we assume spending rate is associated with profit which can be something a small business values. We also thought small businesses value culture and religion, so having datasets about these would be insightful.