Identifying Business Opportunities in Underserved Markets Using Yelp Data

A Data Driven Approach to Entrepreneurship

MIDS W-18-1 Project 2 Final Report December 6th 2016
Group 1: Mike Amodeo and Tom Seddon

1 Summary

A fundamental question for any entrepreneur interested in opening a business that serves customers from a physical location is being able to identify where the best opportunities exist in a given market.

Traditional approaches to this problem range from a reliance on the gut instinct of a good local real estate broker who 'knows the market' to complex modelling of local demographic, population and journey-time data. Each of these approaches have shortcomings. Reliance on a broker makes it difficult for an entrepreneur to be comfortable that all options have been considered, not just the ones the broker personally knows about or is incented to sell. Reliance on a complex demographic model can make it difficult for the entrepreneur to understand the specific logic of why each location is scored high or low, and also often involves complex tuning of the model's parameters by business category.

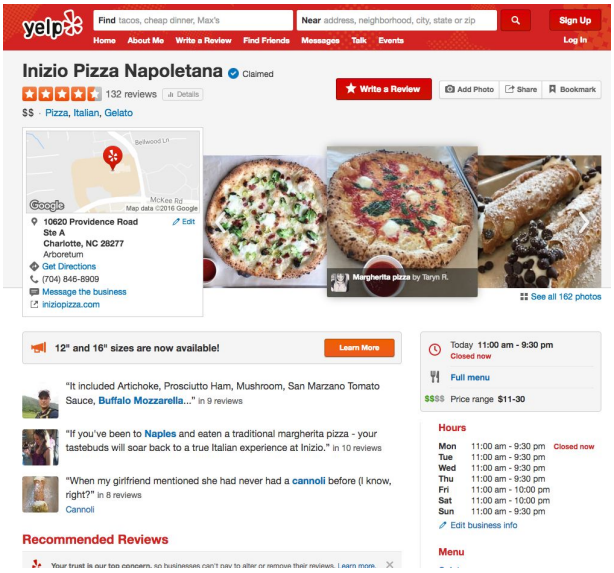
Our project created a way to identify promising business opportunities in a market that addresses the issues with traditional approaches by combining a solid data-driven foundation with transparent logic. This provides the entrepreneur with the assurance that they are looking across the whole market and that they understand very clearly why the selected areas have been identified.

We do this by using data made available from Yelp, a local business review company. We demonstrate how Yelp's data on business locations and customer feedback about those locations can be used both to identify areas which appear to have strong demand for a particular service and to identify areas that appear to have weak competition.

This approach demonstrates a service that does not seem to exist and that we believe might be valuable enough for Yelp to sell to entrepreneurs and commercial real estate brokers, either as a standalone service or bundled as additional value in an advertising package.

2 Background

Yelp is one of the world's leading online review sites, claiming to have over 95 million reviews of local businesses across more than 30 countries, as of Dec 31, 2015. Consumers write and submit their own reviews of businesses, including rating scores, text comments and photos. Consumers use Yelp to get recommendations about what businesses to use based on the feedback of other consumers. Yelp generates revenue primarily from the sale of advertising, although they have been seeking to grow revenue from other sources in the last three years.



An example of a Yelp business listing is shown to the left.

Yelp has made available some of their data in the past three years as part of the 'The Yelp Dataset Challenge'. The Challenge invites data scientists to demonstrate interesting techniques and mine insights from a subset of their data, largely for academic purposes. In the latest round (the eighth), Yelp provided data on users and businesses in 10 cities around the world. This rich dataset includes hundreds of thousands of reviews, ratings, check-ins, and other features.

Most of the published work on Yelp's data has concentrated on the text reviews as a way to demonstrate novel techniques for language and sentiment analysis. Our analysis instead utilizes the business rating and location data to assess the existing market for services.

3 How to identify an attractive area for a new business

Our approach is based on the idea that a good location for a new business has two critical characteristics. First, that there is a strong pool of demand for the service. Second, that there is weak existing competition for the service.

These are of course not the only considerations an entrepreneur would make in evaluating how attractive a location would be. For example high rent levels might make an otherwise attractive location unprofitable to serve. Or there may just be no suitable locations available in the area. But narrowing down a large metro area to a small number of candidate locations can help greatly in speeding and focusing a more detailed feasibility analysis.

It is possible to use the Yelp data to identify both areas that seem to have strong demand as well as areas that seem to have weak competition.

3.1 Identifying strong demand

On the face of it, the Yelp data might seem to contain no data about demand for services. However this overlooks the very important observation that market forces have already driven the existing pattern of supply. In theory, competition will have resulted in entrepreneurs already having optimized their business locations to reflect where demand is and is not. Even if this is imperfect in practice, which is why new opportunities may exist, the pattern of existing supply gives us some excellent information about the pattern of existing demand.

Looking at existing outlets as a way of assessing demand does have limitations as opposed to an approach that starts from population demographics; for example it ignores 'greenfield' new development opportunities. However it also holds advantages over that approach; for example, use of population data alone will not identify locations with high demand but low apparent residential population like a business or entertainment district. Additionally, the approach of viewing locations as a proxy for demand naturally accounts for complex differences in drive/walk time and existing neighborhood connectivity barriers, such as crossing roads/rivers etc.

So existing supply for a specific category of business can give us a very good proxy for demand where there is a lot of existing competition and the category is clearly defined. We take that approach for the 'Pizza' category in Phoenix.

We can also extend the usefulness of this approach by defining demand from outside only one specific business category. We may want to expand a category definition to capture a greater representation of demand. For example we may want to consider the competition to be not just other Pizza places. We take this approach when looking for pizza locations in Edinburgh for reasons explained later.

3.2 Identifying weak competition

Even if there is evidence to believe there is strong demand for a service in a location, it could be hard for a new business to get much share of that demand if the existing competition is strong. Ideally an entrepreneur would like to have reason to believe that customers could be induced to switch if an alternative was presented to them.

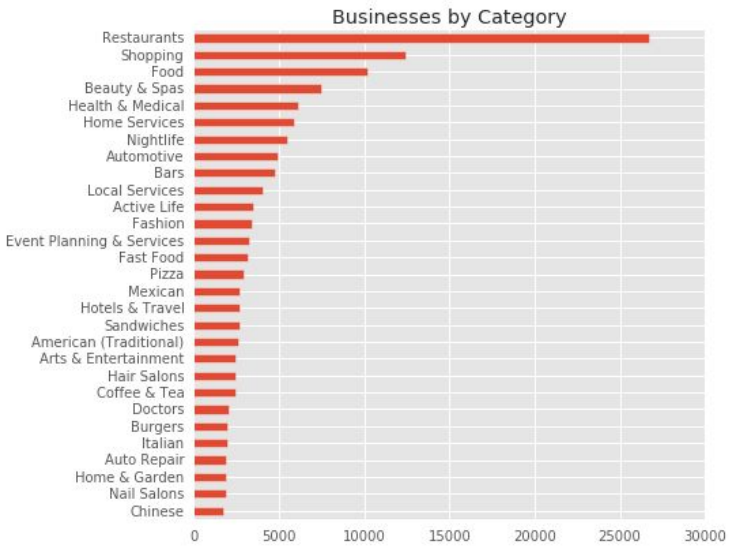
Here, Yelp's data on reviews can be directly helpful and provides a much richer source of information than would normally be available in a traditional demand modelling approach. From Yelp's point of view, the proprietary nature of this data also represents a reason why they may be able to turn the opportunity identification approach we describe into a for-profit business.

By looking at the review ratings of existing outlets, we can identify areas that may have many poorly-reviewed existing outlets of a particular category. These represent areas of hidden competitive weakness that are highly valuable for a potential business owner to identify.

When looking at the review ratings we may find variations in typical review scores by category or by city; if so, so we could define weakness as relative i.e. are the review scores weaker than the typical scores for that category in that city, rather than looking at absolute review scores. We also need to keep absolute levels of review counts in mind, as a low number of reviews could be a less reliable indicator of business quality.

This all assumes of course that review ratings are associated with business success i.e. higher rated businesses have customers who are more loyal and lower rated businesses have customers who are less loyal and so easier to attract with a new outlet. We do not attempt to prove this directly in our work, or consider how that may vary by business type.

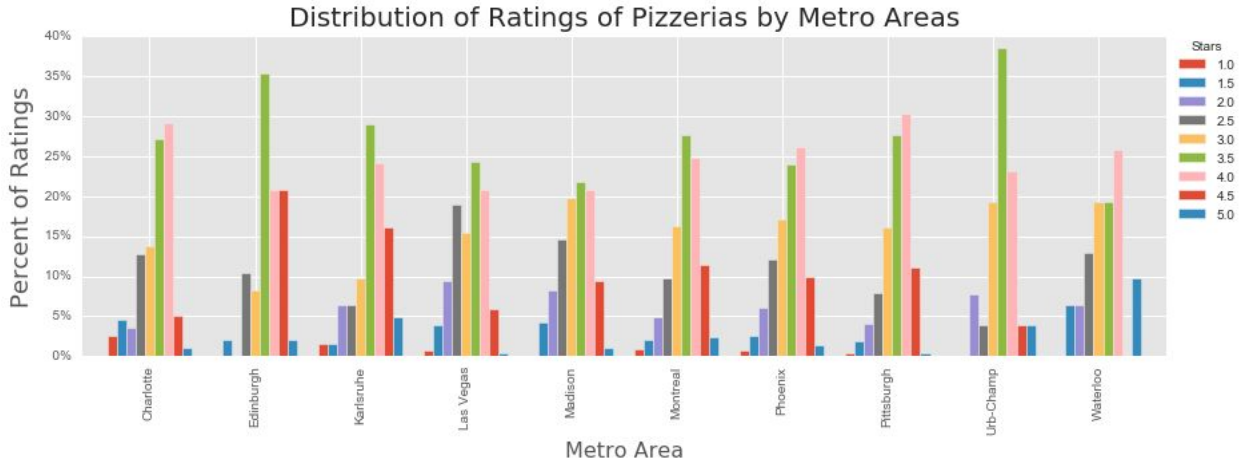
4 Categories considered and why



To demonstrate the approach, we wanted to initially focus on one category that had both representation across ten markets in the provided dataset and many locations in each market, indicating that it is a business that typically has a local trading area rather than pulling from across an entire metro. A larger representation and sample size would also give us a better illustration of the whether our methods gave results that seemed interesting and realistic. We profiled the number of business locations by metro area in each of the main categories. Restaurants were the most common business type, with pizza being the second most common Restaurant subcategory behind fast food. There is much overlap between the categories, as each business can be associated with any number of categories. There are no formal supercategories and subcategories, but there is much correlation. So, most pizza places are also classified as restaurants, although they don't have to be. Many of the pizza places may also be tagged as Italian or fast food.

We then checked to see if the Pizza outlets had a wide variety of ratings and if those ratings varied by metro area, which they did. This provides a basis to evaluate the metro areas differently. Charlotte has a large majority of ratings at 3.5 stars or above, whereas Las Vegas has a greater representation of ratings at 2.5 or below than other cities. We chose to analyze

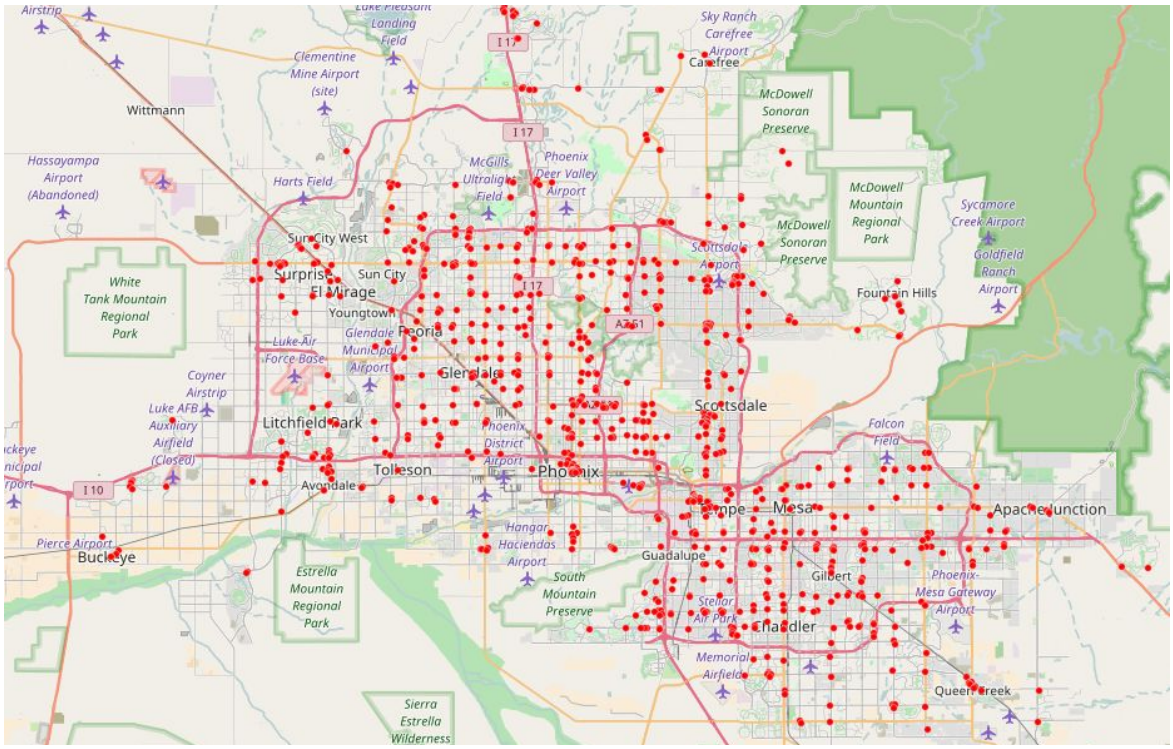
Phoenix in greater detail because it has a large number of pizza places and one of the more even distributions across rating categories.



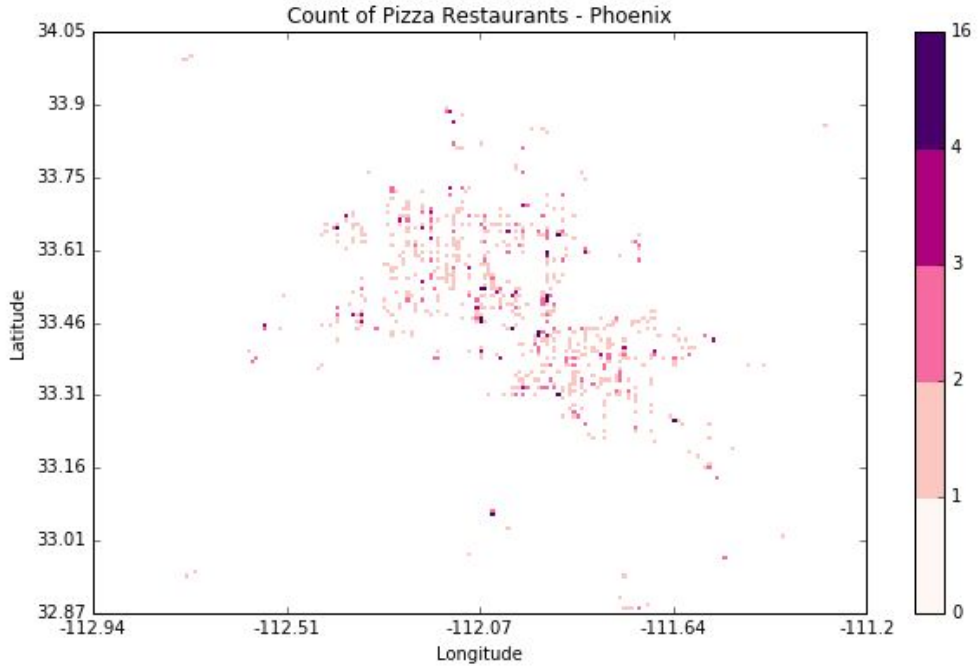
5 Where is a good place to open a Pizza restaurant in Phoenix?

5.1 Application to one category in one city

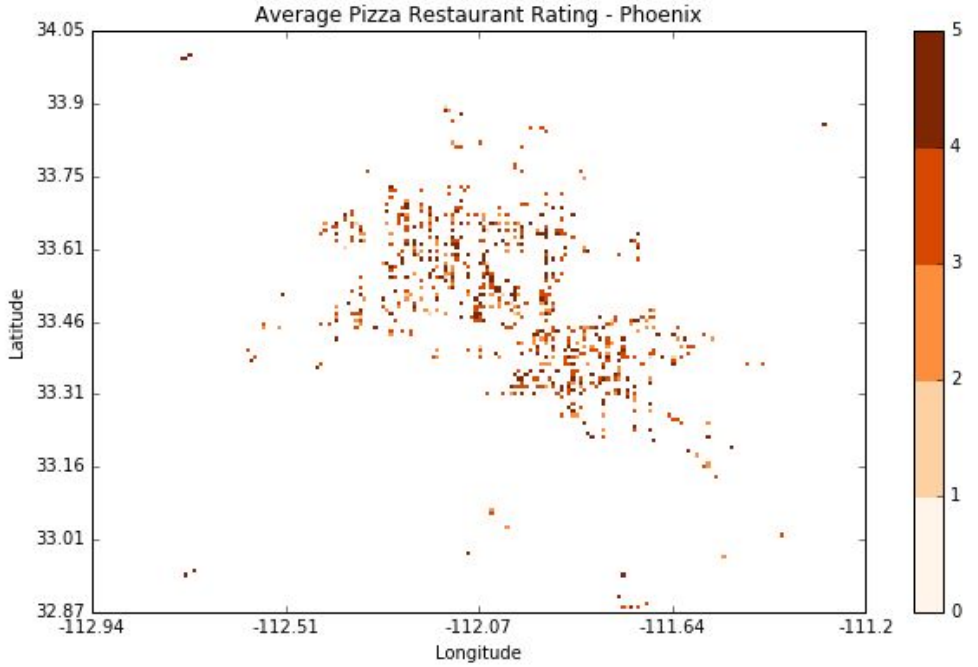
In order to begin applying our method, we narrowed down our search to Phoenix, the metro area with the most pizza places. The figure shows the all of the open pizza places in Phoenix.



We then group these spatially into 0.5 mile x 0.5 mile squares to show the variation in location density more directly. This is plotted to show a square as either empty or in one of five quintiles of density based on the cells that are not empty.



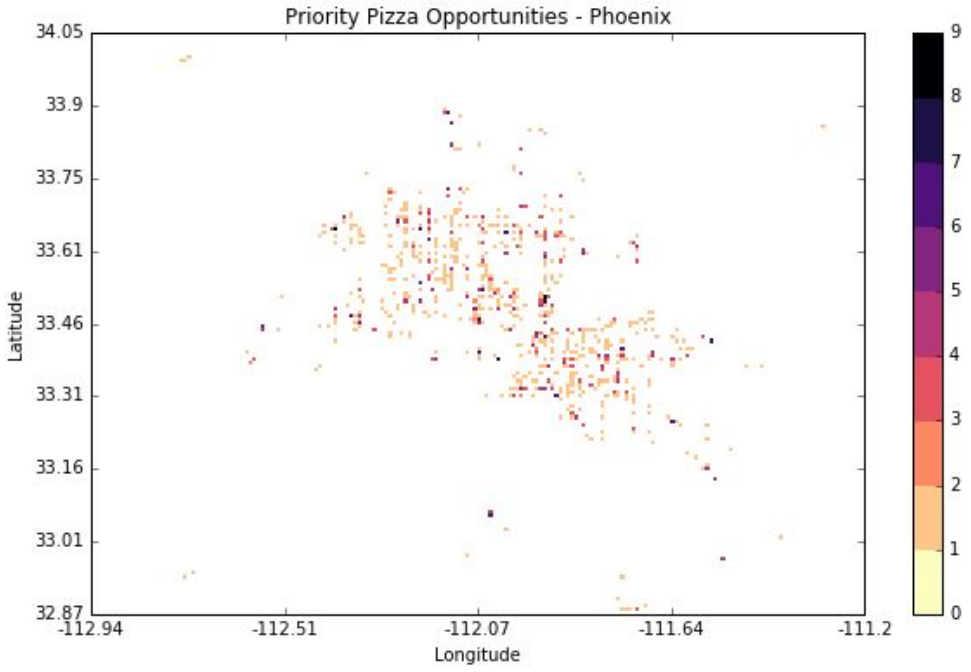
We now look at the average review of pizza restaurants in each of those cells. Again, this is plotted to show a cell as either empty or in one of five quintiles based on the cells that are not empty.



We can now look at how many cells fit into each of the quintiles of demand and review score. The coloration of the table shows the priorities we have assigned to each combination of average rating and number of restaurants in each grid. Better opportunities exist where there is high demand (more restaurants) but low quality (lower average rating).

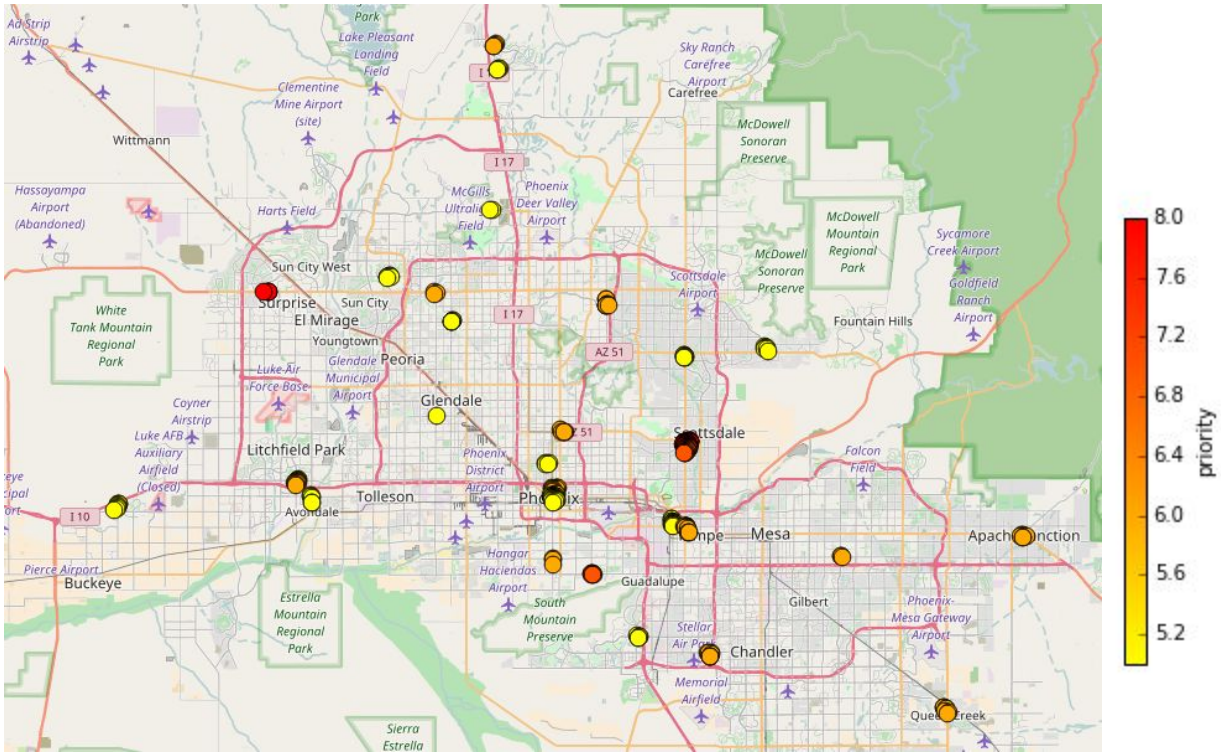
Count of cells		Average Rating of Pizza Places							
		0	**0.5 - 2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**All**
Counts of Pizza Places	**0**	32301	0	0	0	0	0	0	32301
	1	0	115	92	113	136	44	5	505
	2	0	11	25	32	36	8	1	113
	3	0	1	8	11	9	1	0	30
	4	0	0	0	5	3	1	0	9
	5 - 16	0	0	2	1	2	1	0	6

Finally we can combine these two elements as shown to create a final mesh that visualizes where the highest priority opportunities are.

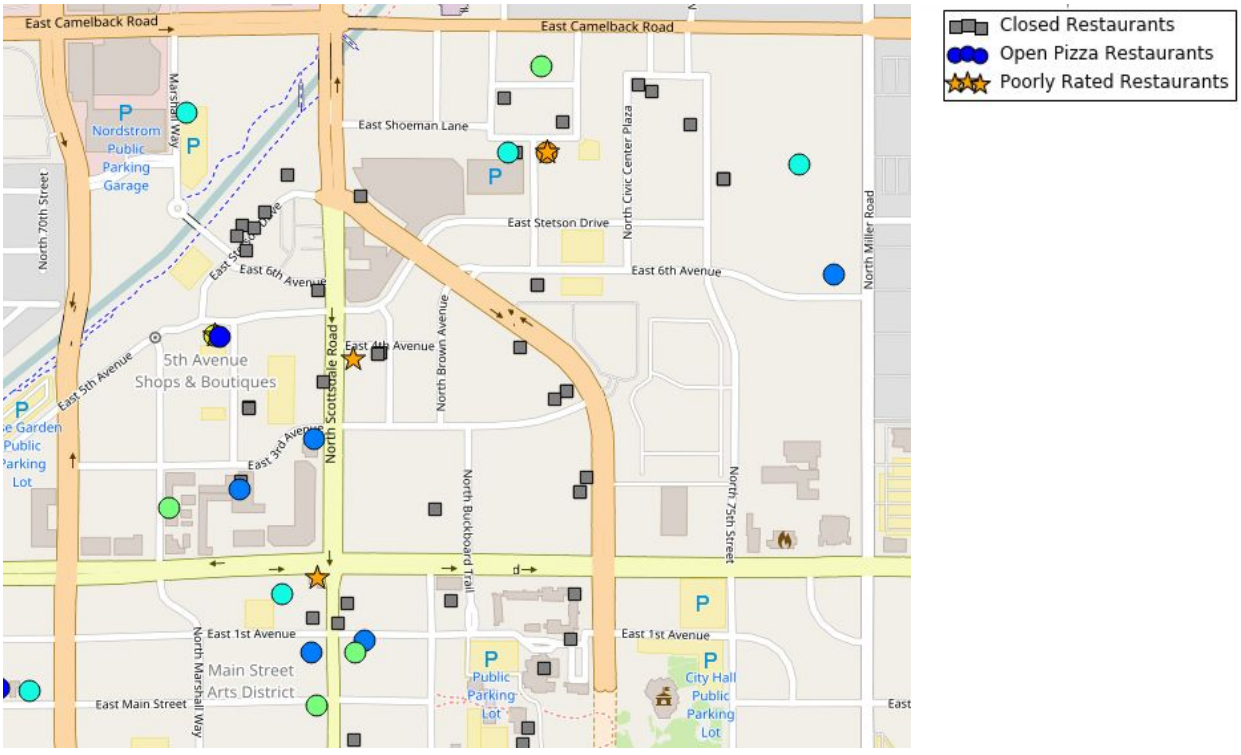


Beyond identifying priority areas, we can also utilize the Yelp dataset to target specific businesses that could be candidates for acquisition. By filtering businesses to show restaurant locations that are either very poorly rated or are currently closed and have not reopened as other businesses, we can even identify properties that may already contain the necessary

infrastructure for a restaurant, with the intent to potentially minimize the initial investments by a prospective entrepreneur.



Each of these priority opportunity areas would warrant a more detailed economic feasibility study. However, by zooming in on a map of one of these high priority locations, we can see simply some of the other features that would be analyzed, such as other attractions or competitors within the target area.



For instance, the commercial center of Scottsdale was one of the two highest priority locations we found, along with Surprise, AZ. This map shows the candidate locations as stars and competing pizza places as dots colored by their rating. There are many poorly rated and/or closed restaurants in this part of town, along with many local amenities, including a mall and some hotels. Presumably, there is a lot of commercial activity, including many restaurants. The northern part of the map (directly south of East Camelback Road) shows as a better opportunity site because there are several potential locations, and there are no highly rated restaurants in those blocks. The Main Street Arts District, however, has more highly rated restaurants and could be more competitive.

5.2 Comparison of finding Pizza opportunities in another city

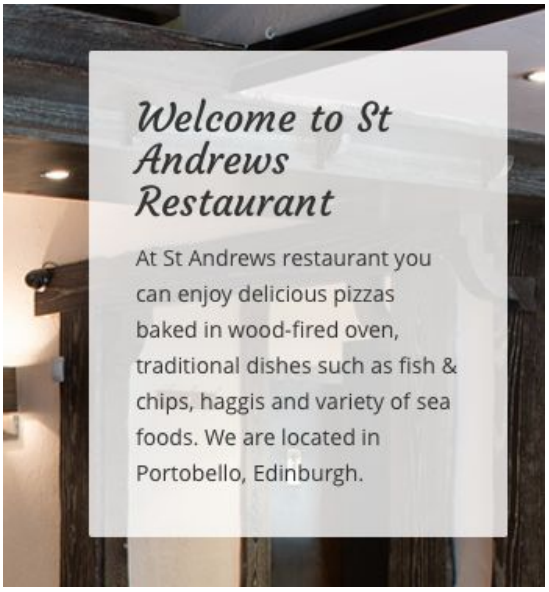
Applying the same approach to another city helps demonstrate the flexibility of the approach as well as some of the variables that can be applied in specific situations. We use Edinburgh, Scotland as our example as it is both another country, is less car-dependent than Phoenix and a much older city layout.

Our first observation is that the city is much more compact. We calculated the diagonal distance for all cities from the min lat/long pair to the max lat/long pair. The Phoenix metro area is 129 miles across while the Edinburgh metro area is 18 miles. In the Phoenix example we chose to use a 0.5 mile cell size, and in Edinburgh we chose to use 0.25 miles. This is not proportional to the two cities' difference in metro size, but we believe the most *informative* mesh size to use represents a tradeoff that we discuss more in section 5.3.

The choice of the most appropriate category is also illustrative. In Edinburgh there are only 54 businesses listed under "Pizza" while in Phoenix there are 1166. Since the Phoenix metro area is estimated to have a population 9x that of the Edinburgh metro area that would suggest there are far fewer pizza places per capita in Edinburgh.

However this highlights some of the differences to be aware of when choosing an appropriate category. This is shown when comparing the way that 'Italian' and 'Pizza' categories break down between the two cities. The table below shows the proportion of restaurants that are either Pizza or Italian and one or the other or both. In Phoenix we see only 20% being identified as Italian without being also in the Pizza category while in Edinburgh it's 66%. In Phoenix, more than two thirds of the pizza restaurants are not Italian

	Phoenix		Edinburgh	
	Count of businesses		Count of businesses	
Pizza or Italian	1453	100%	159	100%
Pizza, not Italian	773	53%	28	18%
Italian, not Pizza	287	20%	105	66%
Pizza and Italian	393	27%	26	16%



Does this mean that Italian restaurants in Edinburgh don't serve Pizza? Or is Pizza a different food category (like in Phoenix), though much less popular? Inspecting a few examples of the restaurants in Edinburgh that do fall into the 'Italian not Pizza' group makes clear that they serve pizza. The explanation for the difference seems to be that Yelp gets its category definitions from third party providers (e.g. phone companies) and business owners are then free to modify those categories (subject to the approval of Yelp moderators). So it seems that the definition of what is meant by each category may be influenced by local norms. The image on the left highlights that as it is one of the 'Italian, not Pizza' restaurants listed above. We would be surprised to find haggis, pizza and fish & chips being considered 'Italian' outside of Scotland.

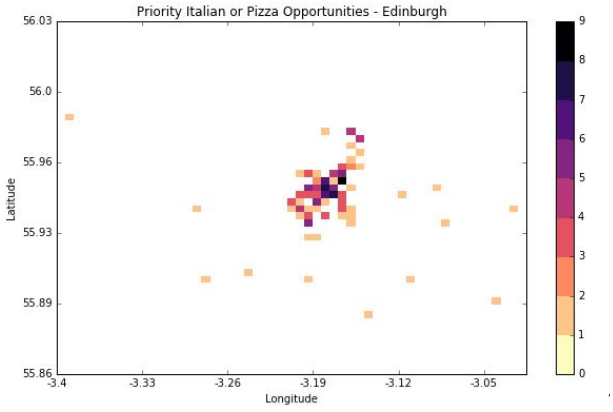
Looking in the other direction at the 'Italian, not Pizza' places in Phoenix, we find that after Pizza, the next most frequent category that an Italian restaurant is in is 'Chicken Wings'. Clearly there are different kinds of Italian restaurants.

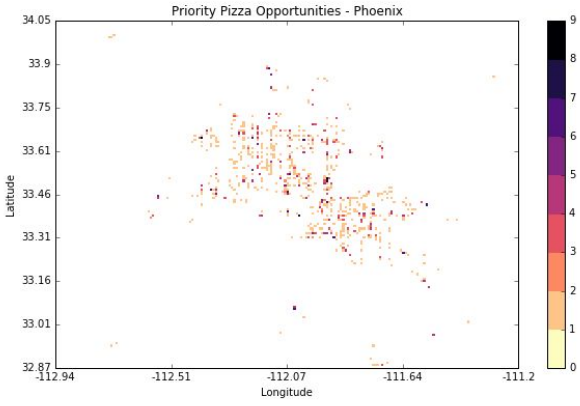
Phoenix restaurants categorized as 'Italian' other categories they are also in.
- count of restaurants, top 5 shown -

Restaurants	680
Pizza	393
Chicken Wings	87
Nightlife	72
Bars	70

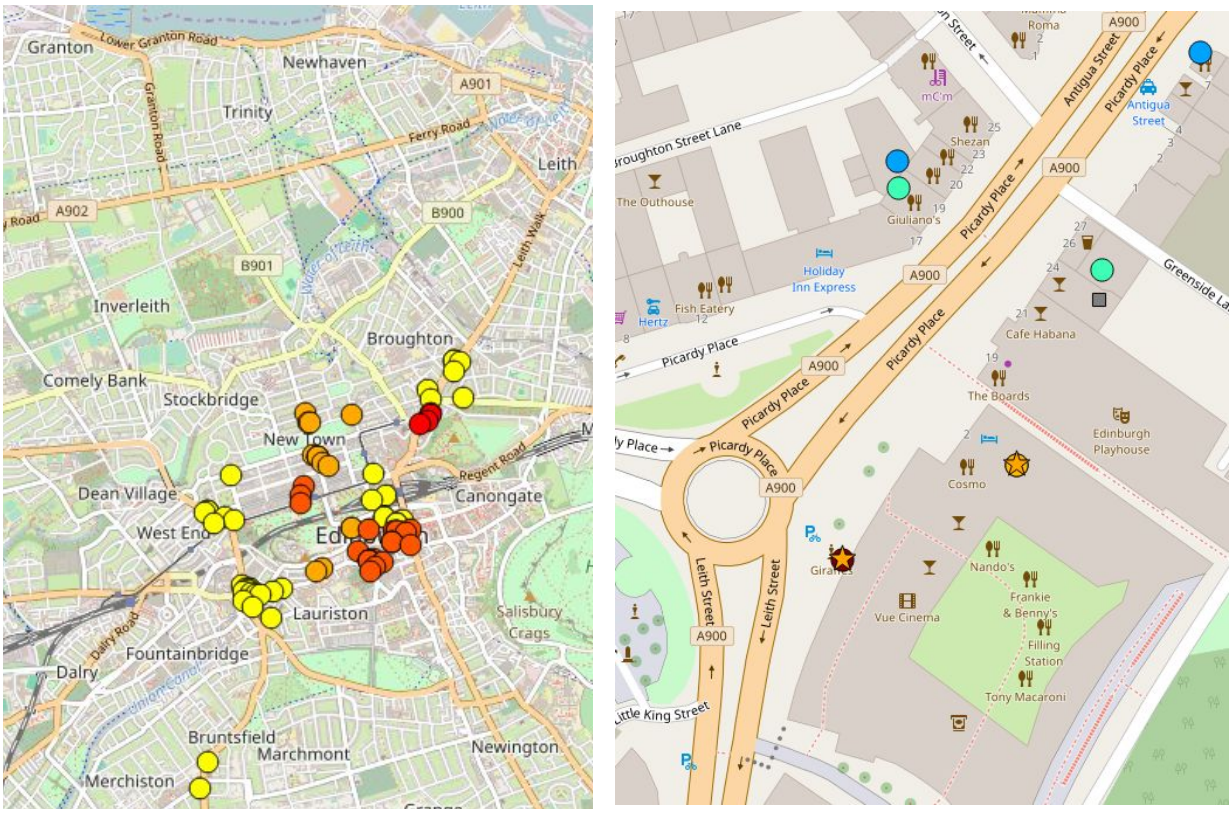
This shows the flexibility of our approach as it is easy to customize to create a category set that best matches what the entrepreneur thinks fits their concept the best. That might include 'Italian, Pizza but not Chicken Wings' for a more upscale concept.

In the case of Edinburgh we decided that the most appropriate category to identify whether there was demand for pizza was to look at anywhere that was listed as Pizza or Italian. This also represents a similar ratio of restaurants to population between the two metro areas.





This gave the results below, where the New Town area presented the highest priority opportunities.



The first observation is that the areas of opportunity are much more tightly clustered than in Phoenix. When looking at the map this seems to be because there is a strong concentration of businesses in the city center and then along the major transport routes in an out of the center.

The highest priority area identified, though, is outside of the center, along one of the main arteries. That suggests there is a hidden pocket of potential, where pizza/Italian exists, but the outlets are poor.

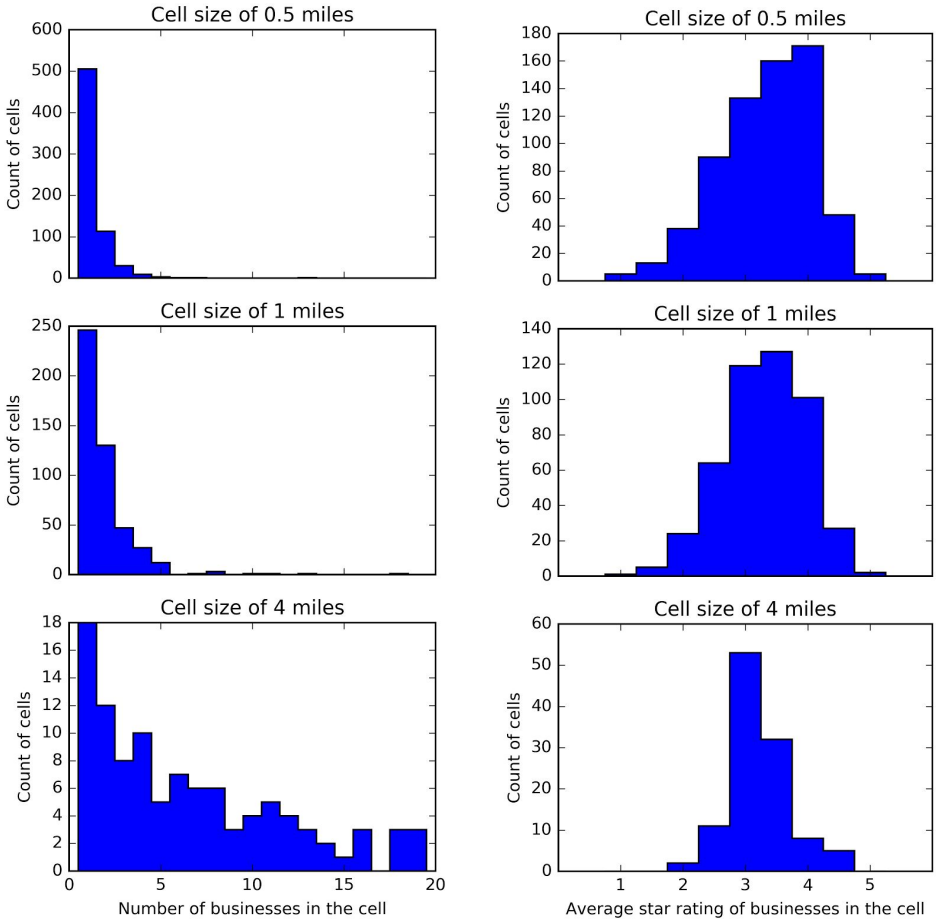
Note that the same approach was easy to apply despite the two cities exhibiting markedly different development patterns and walkability.

5.3 Impact of applying different cell size densities

As mentioned in the example of Edinburgh, choosing the most informative cell size to use for a particular category in a particular metro area represents a tradeoff. Using a large number of cells with small area sounds like it might give fine detail, but it means that many cells have only a single business in. Using a small number of cells with a large area each tends to spread out

the cell counts but mean that there is less variation in the review rating between cells. At the extremes, an extremely fine mesh would have only cells with one business while the coarsest mesh would be one cell for the whole metro area.

We illustrate this below with charts for Phoenix Pizza, showing how the spread of review counts and average star rating changes for three different choices of cell size. (Note that the y axis varies by chart because the total number of cells in the mesh changes with cell size).



We arrived at our chosen cell sizes for the two metro areas above by a process of trial and error. The choice is also impacted of course by the total number of outlets in the category, with a category with a larger number of outlets making it more possible to have a finer mesh if desired.

5.4 Application to another category

Although we have demonstrated this approach with the specific restaurant subcategory of Pizza places, this same methodology could be applied to other restaurant types or even other types of businesses.

Some parameters may change according to business type, such as the assumed willingness to travel and therefore the most useful cell size. While a customer may not want to travel farther than a mile for pizza, they may be willing to drive 30 minutes for a furniture store, a business type with fewer locations and a less frequent, less convenience-driven purchase pattern.

We could also choose to consider all Restaurants as competition and look for places where demand for restaurants was high, but rating for all restaurants, or just a set of subcategories, was poor. This approach would be useful when identifying locations for a very new concept with few or any existing businesses of that type in the market.

We could even use analogous categories if we wanted to explore the potential for a business that was not yet well distributed, but we had reason to believe that it shared demand patterns with another category. For example we might want to open a shoe shop and observe they are often found near clothing shops, so we use clothing shops as a proxy for demand.

6 Conclusions

6.1 Results so far

We believe this approach demonstrates the potential to use Yelp data to identify promising business opportunities in a market by combining a solid data-driven foundation with transparent logic. The great wealth of data associated with each business and the volume of businesses included in the dataset allows for an incredible amount of customization and specificity when determining what type of opportunities to pursue. Using just this simple example of pizza and Italian restaurants, we have given potential entrepreneurs a specific zone to look in and a short list of commercial spaces to analyze in greater detail.

In the Phoenix metropolitan area, we have been able to show that the commercial areas of Surprise and Scottsdale demonstrate high priority locations for the opening of restaurants serving pizza. Being the commercial centers of expansive residential communities, these locations are of course prime development opportunities. However, we have shown that these two in particular suffer a lack of quality restaurants, more so than the other commercial hubs of the metro area, including downtown Phoenix, Tempe, or Mesa. We have also shown specific sub-areas where the competition is weakest.

In Edinburgh, we were similarly able to show that a very specific area in New Town is currently supporting several Italian restaurants which were performing poorly based on their ratings relative to other Italian restaurants in the city. These locations are prime opportunities for improvement for an entrepreneur, as they know there is enough demand to keep even these poorly performing restaurants in business and there are some apparently convertable properties nearby.

6.2 Areas for further work

While these examples have worked quite well, there are a number of ways that this analysis could continue be improved. Here are a few areas for additional work:

Real world test	Work with entrepreneurs to apply to some real world searches and see if the areas surfaced do turn out to be attractive after additional feasibility work has been done.
Success	We make no claim that our approach is validated against actual business success. However, it would be powerful to see if business opening in locations identified by this technique performed well over time (assuming they could get good review scores).
Category variation	With more time it would be possible to test further which Categories the approach seemed to apply to better vs worse and identify patterns in that. Also, different prospective businesses could create comparisons in more types of categories, as shown in the Edinburgh example.
Lookalikes	Another extension to the approach for 'new to market' categories would be to look for locations that didn't have an existing outlet of a certain category, but that 'looked alike' to locations which did have an existing outlet of that kind, using an appropriate technique.
Increasing confidence in business rating	Additional scrutiny on the business rating data could create a more refined expectation of what businesses are performing well. For instance, analysis could show a refined rating based on more reliable reviewers, the number of total reviews, or the expectation of reviews based on some other business characteristic such as price (do expensive restaurants receive better or worse reviews on average than less expensive restaurants?)
Customization	As more examples are analyzed, more refinements can be made to the methodology and the prioritization criteria to better match a potential entrepreneur's desired business. Lists of comparables could be derived not only from the simple category, but also from deeper business data such as price point, ambience, etc.
Defining Business Trading Area	Willingness to travel varies per business type, and so the proximity criteria would change. As distance increases, adding a component for a drive time analysis would be useful

Target Location Improvement	Limitations with the addresses included in the dataset reduce confidence in the conversion opportunities. Being able to cross-reference other datasets such as commercial real estate listings or property databases would add value to the conversion opportunities.

7 Appendix - Data Validation

Before embarking on the analysis we spent considerable time validating the dataset. We ignored the files containing review text as that was not our area of interest, and focused on the 'business' dataset. This contains 85,901 rows x 14 columns, where each row represents information about one particular business in the cities that Yelp has made public. Several of the columns represent lists or dictionaries of multiple data items e.g. hours of operation by day.

	Illustrative contents (truncated)
business_id	--Y_2lDOtVDioX5bwF6GIw
attributes	{'Has TV': True, 'Happy Hour': True, 'Good for...
categories	[Bars, Comfort Food, Nightlife, Restaurants]
city	Madison
full_address	115 State St\nCapitol\nMadison, WI 53703
hours	{'Friday': {'open': '11:00', 'close': '02:00'}...
latitude	43.074739
longitude	-89.387112
city	Madison
name	Buck & Badger
neighborhoods	[Capitol]
open	TRUE
review_count	62
stars	2.5
state	WI
type	business

We verified that all business_ids were unique.

Our main areas of interest were the location data, the category data and the summary review data for the businesses.

The first focus was for us to group the businesses into the metro areas. This was not done in the raw data as even though there was address information, including 'city' and 'state' some of the metro areas were comprised of multiple cities and/or straddled multiple states. There were also anomalies in the addresses e.g. businesses listed as being in states that weren't in any of the metro areas (e.g. CA or AK).

Since the dataset also contained latitude and longitude information for each business we realized we could use that to cross-reference the addresses (plotting some sample maps verified the lat/longs looked correct). We first used 'state' information to do an initial grouping of businesses into rough metro areas, where the state:metro mapping was clear. This left 56 businesses which we then used lat/long information to calculate the distance to the average lat/long of the initial metro maps.

That helped clarify we had two types of issues we found we could correct:

- Small 'state' variations that were clearly mapped to one of the metros e.g. the state "RP":Rheinpfalz in Karlsruhe.
- Address errors in 'state' coding e,g, Gilbert plumbing listed as Mesa CA, but with a lat/long in Mesa, AZ. We checked all these via Yelp/Google and if a simple change made it a valid street address and Yelp's website showed the business at the corrected address then we made the change and kept the corrected record.

And two types of issue that we chose to discard the records entirely for:

- Some businesses that were hundreds of mile outside any of the metros or out of place e.g. Bocholt in Germany or a Chase bank with an address in Garland, TX but a lat/long for Las Vegas
- Ambiguous identification e.g. Texas de Brazil steakhouse has a lat/long for Las Vegas, where there is a Texas de Brazil restaurant, but contains an address for Dallas, TX, where there is also a Texas de Brazil. We don't know if the rest of the review information for the business refers to the Texas de Brazil in Dallas or the TdB in Las Vegas. Hard to know without asking Yelp to backcheck (although we could have checked the location of reviewers if we'd found a lot of these)

In the end we only needed to discard 8 of the businesses, giving us 85,983 in our working file.

We also did extensive exploration of the 'categories' data. There are over a 1000 unique categories and a business can be in multiple categories. There are 241 businesses that do not have any category listed, which we chose to keep. We explored the combinations of categories to understand, for example, if all 'Pizza' places are in 'Restaurants' (they are).

Customer satisfaction with a business is recorded by 'review_count' and 'stars'. We checked those to see how they varied and that they seemed reasonable. Review counts range from a minimum of 3 to a maximum of 6200, with a mean of 34. ('Mon Ami Gabi', a restaurant in the

‘Paris’ casino in Las Vegas has the 6200, which seems possible). ‘Stars’ ranges from 1.0 to 5.0 in 0.5 steps.

Finally, the dataset also has a column for whether the business is ‘open’ or ‘closed’. Of the 85,983 business records, there were 6,601 records for restaurants that are no longer in operation. Because of the potential cost savings of renovating an existing restaurant, closed restaurant sites became part of our conversion opportunity set. By referencing records for closed businesses against the open businesses based on their address data, we were able to remove 2,369 restaurants that had reopened as some other business at that same address. This left 4,232 records of closed restaurants across the ten metropolitan areas. We did some spot checks with online research to verify that this looked correct and was indeed what the data represented. It was valid in some locations, but there still seemed to be more records than there are in fact vacant commercial spaces. This is due to some businesses reopening with a slightly different address. Suite numbers at the reopened business came up as one reason why this would happen. Therefore, the closed business conversion opportunities is a layer with a lower degree of confidence and would require individual checks against other data.