

Paul Rees

COM 597: Community Data Science

Yelp API Final

June 12, 2015

Introduction

Yelp collects data on over 50 million businesses across 27 countries and generates terabytes of log data every single day. At the most basic level, this data is used to help users find businesses and share their experiences online. Beyond that, the potential uses for such a sizable dataset have yet to be realized. The people at Yelp know this, too; the company routinely holds contests that solicit technically creative examples from the data science community to uncover new ways to leverage their data.

In a world where the keepers of big data have been all but vilified, we are quick to interpret “leveraging data” as an inherently malicious activity. But that is not always the case, as is seen in the latest contest to come out of a partnership between [Yelp and the City of Boston](#) (and others). The goal of the contest is to use the Yelp reviews of restaurants around Boston to predict which businesses might be likely health code violators and base health inspections on that rather than current methods.

If Yelp can find ways to influence public health by mobilizing a community of data scientists, what else can we learn about a city through the social data of its businesses? Can we create a profile of a city or neighborhood by looking at the most common and most engaging businesses in that area? And what does that profile say about the people that live there?

The remainder of this paper will describe the process of answering the sub-questions that were required to inform the answers to the larger question. The first section will explain the steps taken to build the dataset, followed by a brief discussion of the analysis that took place to answer each sub-question and a concluding statement about how the results support the larger analysis. Finally, the report will conclude with a discussion about the limitations of the study as well as an evaluation of how successful the analysis was in learning about a community through the social data of its businesses.

Methods

Seattle is home to a diverse collection of businesses, particularly restaurants, which made it an attractive subject of an analysis such as this. In pursuit of the most common and most engaging businesses, we first need to define “common” and “engaging”:

- **Common:** a measure of the number of times that category appears in a defined location
- **Engaging:** the total number of reviews made for a particular business/businesses or category/categories

The next step was to build a sizeable dataset that would accommodate these definitions and provide enough information for a productive analysis.

Yelp’s Search API provides access to data on 20 businesses per API call. This meant that requesting all the businesses in Seattle from a single URL would not be possible. However, the API documentation includes a list of every neighborhood in Seattle (as well as the neighborhoods in other cities), and a file of all business categories as Yelp defines them.

These two files would prove very useful in building the dataset. The Python program that built the dataset used each neighborhood in the list to set the location parameter in the API call, and then it iterated through every category in the category file for each neighborhood. The result of the program was a file that included data for 728,304 records. However, the neighborhood tagging in the Yelp database is not 100% accurate. This resulted in a significant number of duplicate records that needed to be scrubbed out with a separate Python program. The cleaned up dataset listed 26,088 unique businesses. However, after further analysis of the data, it was discovered that a number of businesses were outside of the city limits and were marked as having a neighborhood of “none.” Since a primary goal of the analysis was to compare businesses in specific geographic areas within Seattle, any business with a neighborhood value of “none” was excluded from the analysis. This resulted in a final dataset of 17,118 businesses within the Seattle city limits. While this is likely not a complete list of Seattle businesses, it serves as a sample for analysis that we can then extrapolate trends to the larger business community.

The Yelp Search API returns many data points for each business in the database. On the surface, the response data seems to be geared towards building integrations with third party services that want to leverage Yelp to enhance the functionality of their application. This includes information like deals and offers, gift certificates, menu information, and reservation integration. However, the response data also includes values that would be useful to an analysis such as this. The most important data points for this study were the categories of a particular business, the location, and review count of that business.

The majority of the analysis pivots the data by most common and most engaging for each neighborhood, business, and category. This method allows us to zoom in and out on the data to get both a macro look at how Seattle Yelpers engage with businesses across the city, as well as a closer look at how those trends compare with businesses in specific areas.

Analysis

In order to discover what the most common and engaging businesses in Seattle say about the city, it was necessary to answer a series of sub-questions first and make conclusions about those findings.

Seattle's Most Common

The first question that needed to be answered was “what is the most common category in Seattle?” Yelp has defined over 1000 unique categories of businesses in their database. Seattle has businesses in 727 of them. Out of the 17,000 Seattle businesses in the dataset, the “Apartments” category was most common with 263 businesses occurring, followed by “Hair Salons” (241) and “Coffee & Tea” (222).

One doesn't need a database of business data to know that residential development is a booming industry in Seattle. The [2015 State of Downtown report](#) found that fifty-six new residential construction buildings are either planned, under construction, or recently completed in the downtown district alone. (The report defined downtown differently than Yelp, but the statistic still demonstrates the growth in this category.) The rise in residential development mirrors the population growth Seattle has experienced in the past five years. As new businesses

such as [Amazon](#) and [Expedia](#) expand their office space to accommodate more employees, it is expected that the number of apartment buildings will continue to grow.

When you consider the most common categories in each neighborhood individually, you get a more nuanced perspective of Seattle. The most common business in a neighborhood can offer insight into the commercial profile of the area and present additional questions for understanding why a certain type of business might be attracted to that area. For example, the most common category in the Downtown neighborhood is “Parking.” Anyone who has tried to park downtown recently might disagree with this finding, but just because parking is a challenge doesn’t mean that the lots and garages don’t exist. The Seattle Times [recently reported](#) that due to the community’s demands for quicker transit service and safer bike lanes, the city is outgrowing the 20th century traditions of on street parking. The focus on movability and livability means that there are less curb spaces for cars to go. This leaves room for dedicated lots and parking garages to step in.

Other neighborhoods offer similar data points for understanding its commercial profile. For example, from this analysis, we know to head to Pioneer Square for art galleries, the Industrial District for our beer, and Fremont for our cannabis (see Table 1).

Table 1: Most Common Categories			
Seattle	Apartments	Hair Salons	Coffee & Tea
	263	241	222
Downtown	Parking	Employment Agencies	Hotels
	23	20	20
Ballard	Auto Repair	Apartments	Coffee & Tea
	19	15	15
Capitol Hill	Apartments	Hair Salons	Real Estate
	25	15	15
Pioneer Square	Art Galleries	Parking	Antiques
	10	8	7
Queen Anne	Dog Walkers	Hair Salons	Parks
	7	7	6
Industrial District	Building Supplies	Breweries	Printing
	15	11	8
Fremont	Counseling	Massage Therapy	Cannabis Clinics
	9	9	7

Seattle's Most Reviewed

While category count was useful in identifying what businesses were most common in a particular area, review count can show how the community engages with those businesses. Seattle Yelpers have left a total of 637,816 reviews of local businesses. Splitting out the review totals by neighborhood, we can see that Downtown is by far the neighborhood with most business reviews, followed Capitol Hill and then Belltown and Ballard battle it out for the third and fourth spot. (see Table 2)

Table 2: Top 5 Neighborhoods With The Most Reviews	
Seattle	637816
Downtown	97402
Capitol Hill	63417
Belltown	42270
Ballard	41466
University District	36711

Downtown's commanding lead in number of reviews is likely due to the large number of tourists and visitors that cycle through the downtown district. The businesses in this area can draw from a larger pool of people who might make a review, where other neighborhoods in Seattle might see less traffic overall. Another explanation could be that Downtown simply has more businesses than the other neighborhoods in Seattle. The dataset had 1,890 businesses listed Downtown. The next highest neighborhood was Ballard with 1,029 businesses. More businesses in a neighborhood means more opportunities for that neighborhood to add to its review total.

Looking closer, we can see which businesses the Yelp community interacts with the most by comparing the specific businesses with the most reviews. The number one most reviewed business in all of Seattle is the popular Caribbean sandwich shop, Paseo, followed by Serious Pie and Pike Place Chowder. Paseo's first place spot is even more impressive since the reviews are for a single storefront. Serious Pie and Pike Place Chowder both have multiple

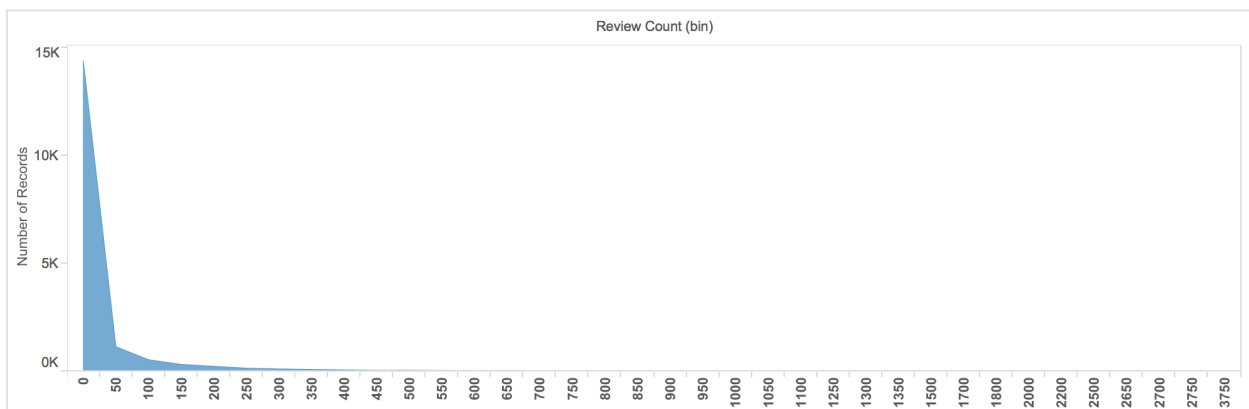
locations contributing to their second and third place positions. Paseo might be benefiting from the same phenomena that businesses

Table 3: Most Reviewed Businesses			
Seattle	Paseo	Serious Pie	Pike Place Chowder
	3771	3148	3075
Downtown	Pike Place Chowder	Piroshky Piroshky	Pike Place Market
	3075	2734	2527
Capitol Hill	Molly Moon's	Poppy	Barrio Mexican Kitchen & Bar
	998	803	723
Belltown	Serious Pie	Umi Sake House	Lola
	2697	2243	2029

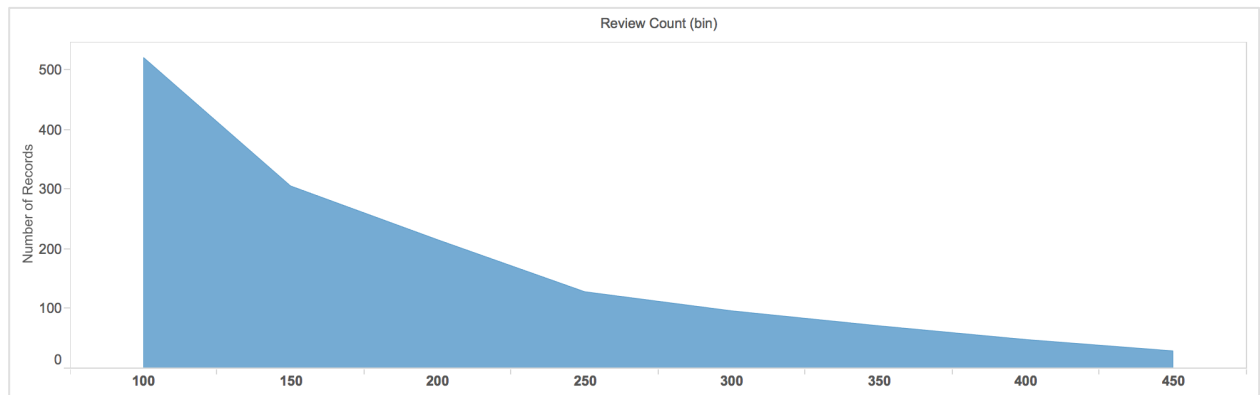
downtown benefit from, as it has received attention in the past from several food blogs touting it as “required eating” for any visit to Seattle. However, other businesses that have not received as much attention have still generated a substantial number of reviews. When you break down the most reviewed businesses by neighborhood, you can see the Yelp community’s favorite businesses to review (see Table 3). Note: Table 3 shows the most reviewed businesses in the top three most reviewed neighborhoods.

We should pause here to understand how review counts are distributed across the businesses in Seattle. Citywide, 91% of businesses have less than 100 reviews (see Histogram 1). Another 8% of businesses have between 100-499 reviews (see Histogram 2). Only 1% of businesses have more than 500 reviews (see Histogram 3). Although businesses like Paseo, Serious Pie, and Pike Place Chowder lead the pack with thousands of reviews, they are an anomaly in this dataset.

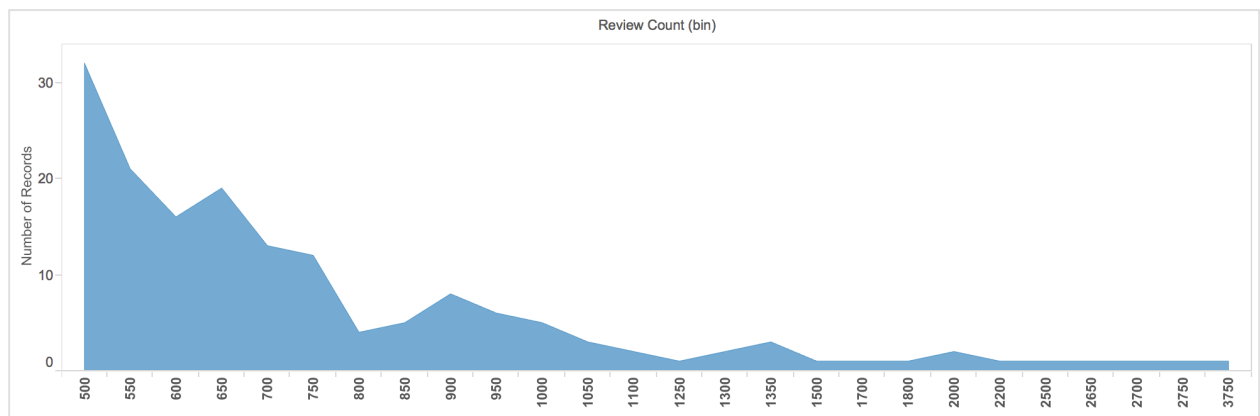
Histogram 1:



Histogram 2:



Histogram 3:



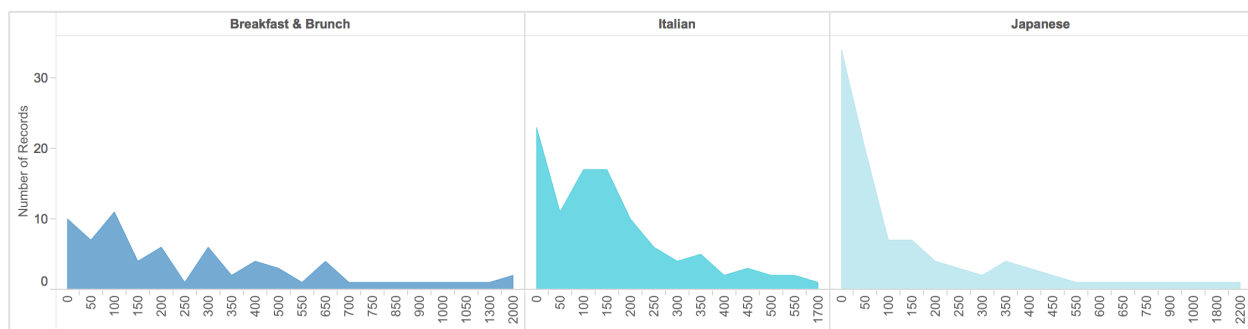
When we look at the distribution of review counts per category, we can see what factors contribute to that category being in the top three. Take Seattle's top three most reviewed categories: Breakfast & Brunch, Italian, and Japanese. Is it a handful businesses with many reviews or many businesses with a fair amount of reviews that give it a top status?

Table 4 shows that Breakfast & Brunch is the most reviewed category, with over 24,000 reviews, but there are only 68 businesses reviewed in total.

Table 4: Top Reviewed Categories			
Seattle	Breakfast & Brunch	Italian	Japanese
	24223	19542	19177
# of Businesses	68	103	94

Breakfast & Brunch has nearly 5,000 more reviews than the second most reviewed category, Italian, which has 103 businesses reviewed. When you compare the distribution of reviews for the top three most reviewed categories, you can get a closer look at the reviews of individual businesses that make up the category. The category histograms below follow a similar trend to the other review distributions, but we can see that although Breakfast & Brunch has the least amount of business reviewed out of the top three categories, there are more businesses with over 1,000 reviews that contribute to the total 24,000 reviews (see Histogram 4). This might suggest that there are less Breakfast & Brunch businesses worth reviewing in Seattle, but a handful of them are really good.

Histogram 4:



Claimed vs. Unclaimed

Yelp has the concept of a claimed and unclaimed business. A claimed business means that the business owner has verified their identity in order to gain extended access to the Yelp page for their business. Claiming a business allows business owners to add photos, business hours, and other information about the business, track visitors to their Yelp page, and respond to customer reviews. It is typically a sign that the business owner recognizes the value of the

Yelp platform and has taken an interest in leveraging it. It is also another key indicator in our pursuit to find the most engaging businesses in Seattle.

Seattle has 7,068 unclaimed businesses and 10,050 claimed businesses. Downtown has the most unclaimed businesses, with 823 businesses, followed by the University District and Ballard (see Table 5).

Table 5: Top Neighborhoods With Unclaimed Businesses	
Downtown	823
U-District	375
Ballard	365
Capitol Hill	356
First Hill	300

The claimed vs. unclaimed data is interesting because up until this point, the analysis has been measuring engagement from the perspective of the reviewer. Pulling in claimed data gives us the perspective of the business owner, and offers insight into how different types of businesses value social media.

Nearly every category has both claimed and unclaimed business listings. In order to understand which categories might be more likely to have unclaimed businesses, we can use an “unclaimed ratio.” The unclaimed ratio is the number of unclaimed businesses divided by the number of claimed businesses in a particular category.

There were 177 categories that had an unclaimed ratio greater than 1, meaning that there were more unclaimed businesses than claimed businesses in that category. The businesses with the highest unclaimed ratio tended to be municipal organizations (such as parks, public transportation, and libraries), or medical facilities (such as hospitals and pediatricians). See Appendix 1 for a list of the ten categories with the highest unclaimed ratios.

One assumption when looking at the list of businesses with the highest unclaimed ratios is that businesses like these have little need or interest in a social platform like Yelp. Or in the case of the Local Flavor category (events and landmarks such as the Fremont Solstice Parade

and the UW Cherry Blossoms), there might not be anyone responsible for social media who would claim the businesses listing.

But Yelpers will review a business whether it is claimed or not. Although ratings and review counts tend to be lower for unclaimed businesses versus claimed,

Table 6: Claimed vs. Unclaimed		
	Claimed	Unclaimed
Avg. Rating	4.16	3.81
Total Reviews	611802	207154

there are over 200,000 reviews of unclaimed businesses on Yelp (see Table 6). Not an insignificant number.

Looking at the businesses with the lowest unclaimed ratio, we can see which categories of businesses are most likely to be claimed. Cosmetic Density tops the list with a 67 out of 72 businesses claimed. Looking at the other categories with the lowest unclaimed ratio, such as Piercings, Gastropubs, and Steakhouses, you might say that these businesses are more dependent on positive reviews and might value the opinion of their customer more than other types of businesses (see Appendix 2).

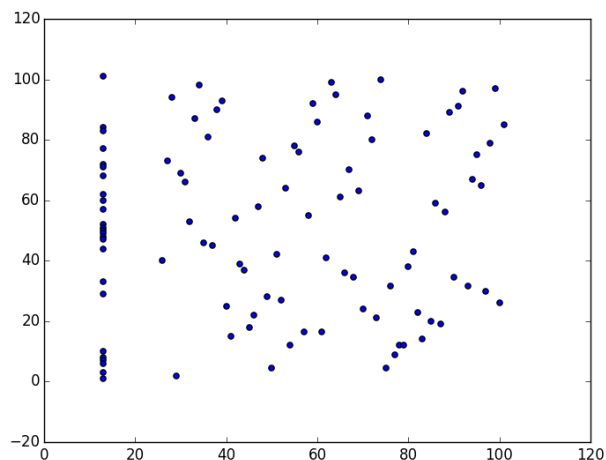
Followers vs. Reviews

The number of reviews a business has on Yelp is a strong indicator of how a community responds to a particular business, but does that engagement carry over to other social media platforms as well?

To answer this question, I took the top 100 most reviewed businesses in Seattle and wrote a Python program to get the number of Twitter followers from Twitter's API. Out of the top reviewed businesses, Starbucks had significantly more Twitter followers than the other businesses (8.1 million). Safeway and Pike Place Market had more modest follower counts

(66,580 and 44,053, respectively) but were still outliers compared to the other top reviewed businesses. See Appendix 3 for a list of the 100 businesses and their review and follower counts.

I was hoping that by comparing review counts and follower counts for the top 100 most reviewed businesses, I could find a correlation between the two social platforms. I wrote a Python program that assigned each business a rank in order of most Twitter followers and another rank in

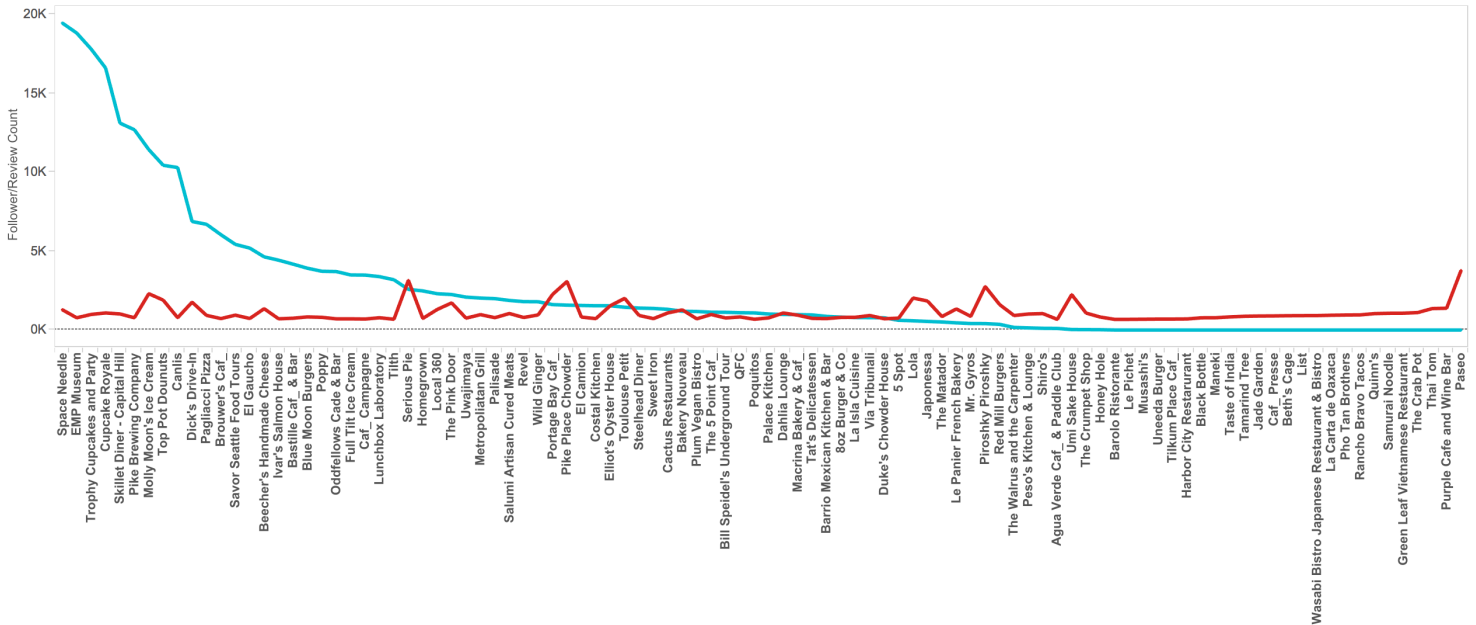


order of most Yelp reviews using Spearman's rank correlation coefficient, and then graphed the results in a scatterplot. Scatterplot 1 shows each business, with the Twitter rank along the x-axis and the Yelp rank along the y-axis. If the business didn't have a Twitter account, their follower count was set to 0.

Unfortunately, there was no obvious correlation between the two platforms in terms of rank. Because of the way the Spearman calculation handles ties, all of the businesses without Twitter accounts are assigned the same rank and are grouped together on the left side of the graph. The Yelp ranks for these businesses span the entire range, showing further evidence that Twitter followers and Yelp reviews are not correlated.

Looking at the absolute follower and review counts, however, there are some interesting trends. By sorting the top 100 businesses by Twitter followers and graphing the number of followers against the number of reviews for each business, we can see how the two

social metrics compare for these businesses (see Line Graph 1). Since the outliers mentioned earlier (Starbucks, Safeway, and Pike Place Market) have such high follower counts, they are excluded from this visualization.



For the majority of businesses, Twitter followers outnumber Yelp reviews. However, there are a few instances where review counts spike above the Twitter followers, especially towards the tail end of the chart. These instances are indicators that the business might be missing out on significant social activity and customer loyalty activities that could happen on Twitter.

Remember that the businesses in this list are the top 100 most reviewed in Seattle. Even the least reviewed business in this list has over 600 reviews. These are businesses that have very active profiles on Yelp, but a portion of them are missing an opportunity to engage an already active audience on a new social platform.

Discussion

So what can we learn about a city based on the most common and most engaging businesses in it? We can tell that Seattle has a growing population and a booming residential development industry. We know that certain types of businesses are attracted to certain areas. These businesses sometimes correlate with civic trends, such as the increase in parking-related businesses in a downtown area with decreasing curbside spots. With more contextual information about the individual neighborhoods and their history, we could offer additional reasons for why businesses are attracted one area over the other. What makes breweries so popular in the Industrial District, for instance? Or what do art galleries get in Pioneer Square that they don't get in other neighborhoods?

Overall, the vast majority of business listings in Seattle have mediocre review counts; 91% of businesses have less than 100 reviews. Without additional data about when these reviews were made, it is difficult to tell if the low review counts are due to new businesses that are in the early stages of developing their Yelp listing, or if Seattle Yelpers are generally infrequent reviewers.

For the most engaging organizations, however, something about them catapults their review counts into the thousands. The catalyst for such significant review counts is likely due to offline factors—a fantastic Cuban sandwich, a strong tourist season, or out of this world pizza. Whatever the reason for the reviews, people are actively seeking a place to share their experiences with that organization.

We see people engage not only with major landmarks around the city, like Pike Place Market and the Space Needle, but also with local favorites like Molly Moon's Ice Cream, Serious

Pie, and Lola. Seattleites love their breakfast and ethnic food spots. This suggests that the Yelp community is made up of a healthy mix of locals and tourists alike. If the business is truly exceptional, an army of reviewers will get behind it to express their enthusiasm online. But that social engagement doesn't always carry over from one social network to another.

Although Seattle is a city known for its forward approach to technology and digital trends, a number of its businesses seem to be slow to adopt social media, even when the community has decided to engage anyway. The unclaimed municipal businesses and parks sometimes have hundreds of reviews (as is the case with Gas Works Park, 352 reviews), but seemingly no one on the other end is listening. And in the case of Twitter, businesses with hundreds of reviews sometimes don't have any Twitter profile whatsoever (Paseo attracts thousands of reviews but doesn't maintain a Twitter account). These businesses are potentially passing up the opportunity to engage with an already enthusiastic community of fans by not interacting with them on Twitter. It's good practice for any organization to be aware of the social engagement they are generating on one platform and look for ways to leverage that community on a new platform.

Limitations

There were several limitations to the study that would have to be addressed should one continue this work. First of all, the Yelp API itself presented challenges in accessing all the data that was needed. The 20-business limit meant that businesses were unavoidably left out of the dataset. This issue was further compounded by the errors in how the businesses were tagged with neighborhood data. With better neighborhood tagging, more unique businesses could

have been returned with each API call, which ultimately would have made better use of the 20 business limit.

Secondly, the API excluded all user and review content data entirely. This would have made for a much more dynamic study. Having access to user profiles and specific review content could have offered the opportunity to see how often certain Yelpers make reviews, what types of businesses they review most often, and look for trends that might indicate if external factors influence the frequency or sentiment of reviews.

In terms of finding the most engaging businesses, another metric that might be used to measure cross-platform engagement is to monitor the number of tweets about a certain business. This would offer an additional dimension to the follower count metric included in this study. I explored this possibility, but I used the Twitter Search API and was only able to return a week's worth of data. This was not sufficient to draw any conclusions. One potential workaround for this, with more time, one could write a program using Twitter's streaming API and keep it running over an extended period of time to accumulate a sizeable dataset. Once compiled, additional analysis could be conducted on the content of the tweets.

Conclusion

In the end, there are many interesting insights that can be uncovered by analyzing the most common and most engaging businesses in a city. The city's historical, societal, and economic influences manifest in the profile of its businesses community. The interests and passions of the community come through as well. Although some people question the value social media can offer a business, the ultimate value of social media will be how the platform leverages the data it collects to improve our understanding of the community itself.

Appendix 1

Category	Unclaimed	Claimed	Unclaimed Ratio
Parks	143	1	143
Local Flavor	33	1	33
Playgrounds	31	1	31
Post Offices	28	1	28
Public Transportation	21	1	21
Libraries	34	2	17
Public Services & Government	34	2	17
Food Court	15	1	15
Beaches	14	1	14
Ear Nose & Throat	14	1	14

Appendix 2

Category	Unclaimed	Claimed	Unclaimed Ratio
Cosmetic Dentists	5	67	0.07
Gastropubs	1	12	0.08
Piercing	1	12	0.08
Steakhouses	1	13	0.08
Career Counseling	1	11	0.09

Appendix 3

name	handle	follower_count	review_count
Starbucks	Starbucks	8122887	1538
Safeway	Safeway	66580	758
Pike Place Market	pike_place	44053	2527
Space Needle	space_needle	19456	1287
EMP Museum	EMPmuseum	18821	775
Trophy Cupcakes and Party	trophycupcakes	17806	987
Cupcake Royale	CupcakeRoyale	16609	1082
Skillet Diner - Capital Hill	skilletstfood	13111	1018
Pike Brewing Company	pikebrewing	12693	777
Molly Moon's Ice Cream	mollymoon	11432	2302
Top Pot Dounuts	Toppot	10439	1893
Canlis	canlis	10298	781
Dick's Drive-In	DicksDriveIns	6877	1764
Pagliacci Pizza	pagliaccipizza	6700	929
Brouwer's Caf_	brouwerscafe	6040	724
Savor Seattle Food Tours	savorseattle	5427	943
El Gaucho	ElGauchoSteak	5192	730
Beecher's Handmade Cheese	BeechersSeattle	4636	1351
Ivar's Salmon House	IvarsClam	4426	711
Bastille Caf_ & Bar	bastilleseattle	4176	747
Blue Moon Burgers	BlueMoonBurgers	3915	830
Poppy	poppyseattle	3723	803
Oddfellows Cade & Bar	oddfellowscafe	3704	708
Full Tilt Ice Cream	FTicecream	3490	708
Caf_ Campagne	Cafe_Campagne	3481	696
Lunchbox Laboratory	Lunchbox_Lab	3382	777
Tilth	seattletilth	3184	683
Serious Pie	SeriousPiePike	2573	3148
Homegrown	homegrownian	2480	734
Local 360	Local360Seattle	2299	1302
The Pink Door	PinkDoorSeattle	2250	1720
Uwajimaya	Uwajimaya	2084	751
Metropoliatan Grill	MetGrill	2025	976
Palisade	PalisadeSea	1990	781
Salumi Artisan Cured Meats	SalumiSeattle	1873	1046
Revel	revelseattle	1800	794
Wild Ginger	WildGingerRes	1792	958
Portage Bay Caf_	portagebaycafe	1611	2257

Pike Place Chowder	PikePLChowder	1577	3075
El Camion	elcamionseattle	1557	819
Costal Kitchen	coastalkitchen	1539	722
Elliot's Oyster House	ElliottsSeattle	1538	1541
Toulouse Petit	ToulousePetit	1442	2005
Steelhead Diner	steelheaddiner	1391	923
Sweet Iron	SweetIronWaffle	1368	722
Cactus Restaurants	eatatCACTUS	1312	1087
Bakery Nouveau	bakerynouveau	1188	1276
Plum Vegan Bistro	PlumRestaurants	1173	708
The 5 Point Caf_	The5PointCafe	1129	984
Bill Speidel's Underground Tour	TourUnderground	1123	762
QFC	QFCGrocery	1098	827
Poquitos	vivapoquitos	1084	683
Palace Kitchen	palace_kitchen	1013	770
Dahlia Lounge	dahlialounge	986	1079
Macrina Bakery & Caf_	MacrinaBakery	976	941
Tat's Delicatessen	tatsdeli	956	737
Barrio Mexican Kitchen & Bar	BarrioChino_	864	723
8oz Burger & Co	8ozburger	818	799
La Isla Cuisine	laislacuisine	789	807
Via Tribunali	ViaTribunali	782	921
Duke's Chowder House	DukesChowder	774	713
5 Spot	5spotseattle	616	753
Lola	LOLASeattle	584	2029
Japonessa	JaponessaSushi	544	1837
The Matador	MatadorRest	509	853
Le Panier French Bakery	LePanierBakery	455	1336
Mr. Gyros	mrgyroSeattle	411	864
Piroshky Piroshky	PiroshkyBakery	408	2757
Red Mill Burgers	RedMillBurgers	358	1615
The Walrus and the Carpenter	thewalrusbar	164	919
Peso's Kitchen & Lounge	pesoskitchen	135	1013
Shiro's	ShirosSushi	108	1041
Agua Verde Caf_ & Paddle Club	aguaverdecafe	101	674
Umi Sake House	umi_sake_house	31	2243
The Crumpet Shop	thecrumpetshop_	27	1072
Honey Hole	HoneyholeSandwi	20	818
Paseo	no account	0	3771
Purple Cafe and Wine Bar	no account	0	1381

Thai Tom	no account	0	1359
The Crab Pot	no account	0	1107
Green Leaf Vietnamese Restaurant	no account	0	1066
Samurai Noodle	no account	0	1061
Quinn's	no account	0	1039
Rancho Bravo Tacos	no account	0	959
Pho Tan Brothers	no account	0	951
La Carta de Oaxaca	no account	0	937
Wasabi Bistro Japanese Restaurant & Bistro	no account	0	915
List	no account	0	913
Beth's Cage	no account	0	907
Caf_ Presse	no account	0	894
Jade Garden	no account	0	887
Tamarind Tree	no account	0	872
Taste of India	no account	0	832
Maneki	no account	0	778
Black Bottle	no account	0	774
Harbor City Restarurant	no account	0	702
Tilkum Place Caf_	no account	0	694
Uneeda Burger	no account	0	693
Musashi's	no account	0	685
Le Pichet	no account	0	677
Barolo Ristorante	no account	0	673