Understanding Millennial Tastes: An Exploration Into Foursquare Venues

A. Introduction

A1. Background

Millenials are defined by Pew Research as the group of Americans born between the years 1981 and 1996, corresponding to an age range of 23 to 38 (Dimock, 2019). Each generation represents a cohesive demographic and psychological shift that is both palpable and measurable. Millenials on average are better educated, have more diversity in the workplace, and delay the formation of their own households until a later age. This means that they are more likely to be living at home with their parents, and are less likely to get married at a younger age (Bialik & Fry, 2019). This is only a minor selection of the various respects in which millennials differ from their predecessors. From political and religious views, to purchasing behaviors, millennial preferences are changing the American landscape.

The most tangible aspect of this effect can be seen in the composition of modern cities, and how they cater to millennial tastes. Two cities with the largest net migration of those between the age range of 23 and 38 are Portland, Oregon and Seattle, Washington (Geier, 2019). In Seattle, those between the ages of 30 and 39 make up 19% of the population, and those between 20 and 30 make up 20%, a combined total of 39%(U.S. Census Bureau, 2017).

A2. Problem

With such a prominent demographic starting to take over the mantle of society, it would be wise to understand their particular set of needs and desires. Whether this be from a business point of view, a sociological point of view, or merely from a cultural lens, it would help any party that interacts with millennials in any capacity to grasp how they are changing the culture. That is why for this project, I intend on using data to take a glance at the ways in which two of the most ‘booming’ millenial cities in the country are similar. By recognizing the commonalities, we may gain some insights as to what ways cities react to their growing pervasive millennial cultures.

A3. Interest

The outcome of this analysis would be well received by 3 groups of people, 2 in a general sense and 1 in a specific sense. Specifically, those who intend to either start up or pivot an existing business in the cities of Portland or Seattle would be appreciative to know what is trending and what potential competition might they come up against. For example, if hot dog stands were popular, but there were hundreds of them evenly distributed across the entire city, then it may not be sensible move to start up a hot dog stand. Similarly, if there were a suspiciously small number of gas stations, it could be because of some sociological reason such as the increasing environmental conscientiousness demands of the millennial population.

More broadly speaking, this analysis would be useful to sociologists and business owners across America who operate in areas with a growing proportion of people in their mid to late 20s. The

sociologists could use this to gain perspective on the demands faced by millennials. In a comparable regard, business owners would similarly like to gain perspective on the demands faced by millennials, but for a different reason. The growing field of behavioral design seeks to facilitate interactions between businesses and consumers by reducing the number of roadblocks that can potentially occur with any consumer behavior. One of the tenets of behavioral design is to segment your consumer-base into groups and appeal to their specific set of needs. Of these groups, millenials are emerging as a dominant force in the marketplace. Based on research by Accenture, millenial spending habits were examined and it was estimated that as a whole they spend over 600 billion dollars a year (Donnelly & Scaff, 2019). Not only that, but this number would grow by 2020 to 1.4 trillion.

B. Data

B1. Data Sources

The data for this project was collected from the following places:

- A table of neighborhoods in Seattle was scraped from here
- Latitude and Longitude coordinates were collected for each neighborhood using the Nominatim geocoder
- Foursquare API was used in conjunction with the neighborhood names and the coordinates to obtain a list of venues with their respective categories
- A CSV of Portland neighborhood names was obtained here

B2. Data Preprocessing

	Neighborhood Name	Within Larger District	Annexed	Locator Map	Street Map	Image	Notes
0	None	None	None	None	None	None	None
1	North Seattle	Seattle	Various				North of the Lake Washington Ship Canal[42]
2	Broadview	North Seattle[42]	1954[43]				[44]
3	Bitter Lake	North Seattle[42]	1954[43]				[45]
4	North Beach / Blue Ridge	North Seattle[42]	1940,[43] 1954[43]				[46]
5	Crown Hill	North Seattle[42]	1907,[47] 1952,[43] 1954[43]				[48]

The scraped wikipedia table

The initial dataframe of Seattle Neighborhoods contained many unwanted elements. The first step I did was drop the unnecessary columns, and drop rows that contained a Neighborhood Value of NA such as the very first row. This removed 1 row and brought down the total number of rows, and thus neighborhoods, to 127. Next, there were many artifacts left over from the wikipedia page, such as the numbers in brackets, as well as alternative names in parentheses. These would not yield coordinates when called with the geocoder, so I needed to adapt them into a useable form. After running a list of the

neighborhood names through several for loops to clean them, I had a dataframe free of artifacts. After mapping out the neighborhoods once, it was obvious that there were large gaps where no neighborhood was present, and these neighborhoods were not providing coordinates as output when ran through the geocoder.

At this point I isolated these neighborhoods and came to realize that some of them had alternative names that were more common and not listed in the wikipedia table. This was remedied when I manually changed these neighborhoods using the dataframe replace method. The result of this initial cleaning was a dataframe full of neighborhoods in Seattle that all yielded coordinates with the geocoder.

	Neighborhood
0	North Seattle
1	Broadview
2	Bitter Lake
3	North Beach
4	Crown Hill

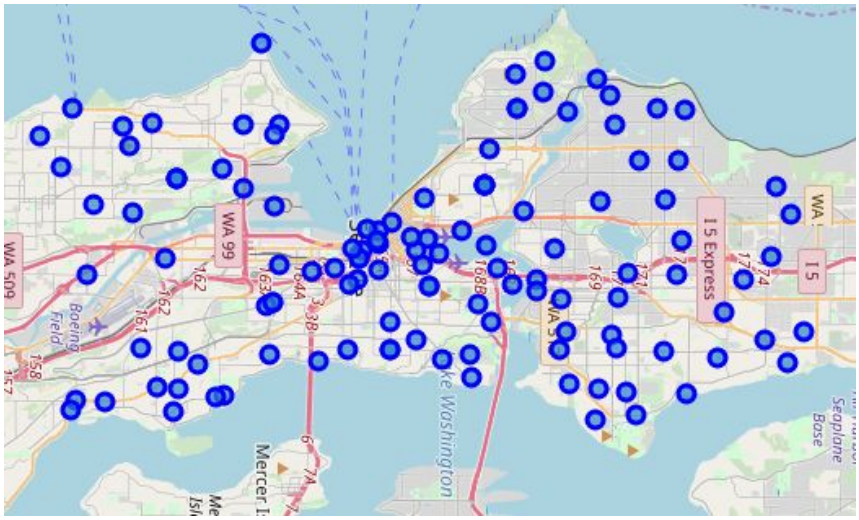
The dataframe of cleaned Seattle Neighborhoods

The coordinate data was then obtained by looping a slice of this dataframe containing all the neighborhoods with a for loop that geocoded each neighborhood, and appended that data to a list, and these data were added to the dataframe. Any dataframe that had a Latitude field of 0, indicating no coordinates were returned, was removed.

	Neighborhood	Latitude	Longitude
0	North Seattle	47.660773	-122.291497
1	Broadview	47.722320	-122.360407
2	Bitter Lake	47.726236	-122.348764
3	North Beach	47.696210	-122.392362
4	Crown Hill	47.694715	-122.371459

The Seattle dataframe with the coordinate data added

The neighborhoods and their respective coordinates were subsequently plotted using folium. The purpose of this was 2-fold. One was to visualize any glaring omissions of neighborhoods. In the early processing stage, some large gaps were found and manually handled as stated in the paragraph above. The second purpose was that using foursquare's api is more accurate with multiple segmented areas to search in a radius around.



Folium map of Seattle, rotated for space considerations

I then defined a function to call foursquare's api and return venue information including the venue's name and category. The resulting dataframe included the Neighborhood name, neighborhood coordinate data, the venue name, the specific venue coordinates, and the venue's category. Categories included labels such as 'Indian Restaurant' or 'Ice Skating Rink'. This dataframe had 11,094 rows indicating that 11,094 venues were located

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North Seattle	47.660773	-122.291497	Center for Urban Horticulture	47.657978	-122.290237	College Science Building
1	North Seattle	47.660773	-122.291497	University Village	47.662487	-122.298531	Shopping Plaza
2	North Seattle	47.660773	-122.291497	Din Tai Fung 鼎泰豐	47.661567	-122.299725	Dumpling Restaurant

The dataframe composed of the returned foursquare data

It became evident by looking at the folium map that because some of the neighborhoods were clustered so closely to each other, there would be an overlap of venues. That is, only some percentage of these 11,094 venues would be unique. To remedy this, I dropped values with duplicate values in the 'Venue', 'Venue Latitude', and 'Venue Longitude' columns. If I dropped based only on Venue name, it would drop things which were unique stores but belonged to a chain. For example, Dominos pizza or a 7-11 convenience store. This reduced the dataframe to a size of 2906 total unique venues.

A similar process was performed for the Portland, Oregon dataset. I imported the csv which contained 98 rows of neighborhoods.

	OBJECTID	NAME	COMMPLAN	SHARED	COALIT	HORZ_VERT	MAPLABEL	ID	Shape_Length	Shape_Area
0	1	CATHEDRAL PARK			NPNS	HORZ	Cathedral Park	31	11434.254777	5.424298e+06
1	2	UNIVERSITY PARK			NPNS	HORZ	University Park	88	11950.859827	6.981457e+06
2	3	PIEDMONT	ALBINA		NPNS	VERT	Piedmont	70	10849.327392	6.079530e+06
3	4	WOODLAWN	ALBINA		NECN	HORZ	Woodlawn	93	8078.360994	3.870554e+06
4	5	CULLY ASSOCIATION OF NEIGHBORS			CNN	HORZ	Cully Association of Neighbors	23	18179.392090	1.658062e+07

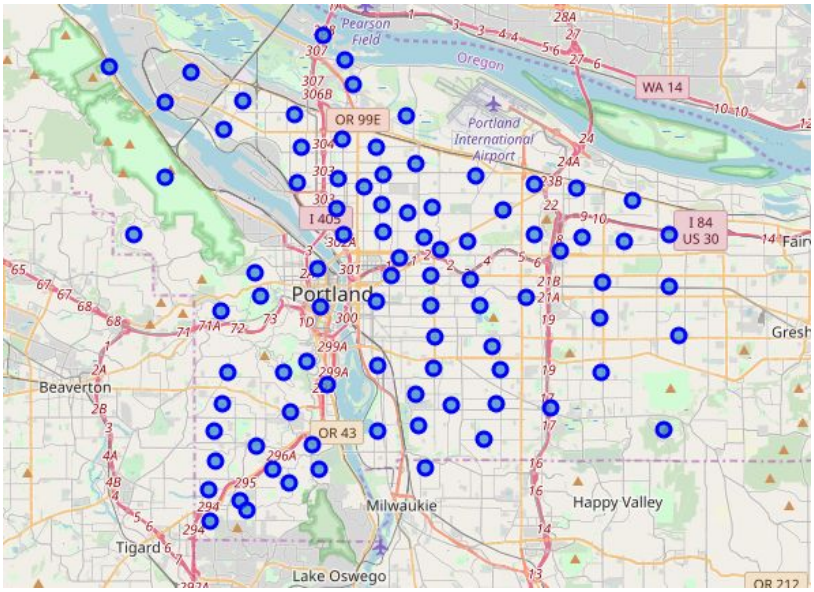
The imported Portland Neighborhoods

A simple 3 step process was used to clean this dataframe to a valid format. First, the superfluous columns were dropped. Next, MAPLABEL was renamed to Neighborhood. Finally, some neighborhoods that failed to yield coordinates from the geocoder were manually renamed.

	Neighborhood
0	Cathedral Park
1	University Park
2	Piedmont
3	Woodlawn

The cleaned Portland dataframe

The Portland dataframe was then treated in the same way as the Seattle dataframe. It was run through a for loop that geocoded all the coordinates, the coordinates were appended to the dataframe, and the data was then mapped out using folium. The Portland neighborhoods produced 7113 total venues, of which 2826 were unique.



A folium map of the Portland Neighborhoods

C. Methodology

Now with all of the data in one table, it had to be manipulated in a manner for it to make sense. I decided that the best representation of this data would be the ratio of each unique venue category divided by the total number of venues. This would standardize the Portland and Seattle data in such a way that it could be reasonably compared to each other.

Taking the dataframe of unique Seattle venues, I split it into 2 arrays. These arrays contained each unique venue category present, and the number of times it appears in the dataframe. I then divided each count by the total number of venues in the dataframe, and multiplied by 100, to obtain that categories percentage representation. A new dataframe was made of the Venues, their raw counts, and the percentage. This procedure was replicated for the Portland dataframe as well

	Venues	Raw_Counts	Ratio_Counts
0	Coffee Shop	204	7.019959
1	Park	120	4.129387
2	Pizza Place	101	3.475568
3	Bar	70	2.408809
4	Mexican Restaurant	64	2.202340

The top 5 represented categories in Seattle

	Venues	Raw_Counts	Ratio_Counts
0	Coffee Shop	186	6.581741
1	Pizza Place	109	3.857042
2	Park	106	3.750885
3	Bar	102	3.609342
4	Mexican Restaurant	75	2.653928

The top 5 represented categories in Portland

These tables are interesting and useful, but they aren't the most useful things we can derive from these datasets. The next step was to append both dataframes together. Appending the lists of both dataframe's venues together yields a prediction of 654 total venue categories,. but this will cause many of the venues to be duplicated as there will be overlap. The 2 dataframes were then joined using an outer join, which resulted in 398 rows indicating 398 unique venues between the 2 dataframes combined. Each NaN was changed to 0, so that numerical calculations could be performed.

	Venues	Raw_Counts_Seattle	Ratio_Counts_Seattle	Raw_Counts_Portland	Ratio_Counts_Portland
0	Coffee Shop	204.0	7.019959	186.0	6.581741
1	Park	120.0	4.129387	106.0	3.750885
2	Pizza Place	101.0	3.475568	109.0	3.857042
3	Bar	70.0	2.408809	102.0	3.609342
4	Mexican Restaurant	64.0	2.202340	75.0	2.653928

The merged dataframe of Portland and Seattle venue count data

I then defined a function for a Z-test for 2 proportions. This means for each row of data, if we assume that the true proportion of venues is actually equal, what is the chance that we observe the ratios we observed in the foursquare data. For example, if the underlying ratio for coffee shops in Portland and Seattle are actually equivalent, what is the chance that we observed the proportions of 6.58 and 7.02 that we did? Lists of each respective ratio was formed, and looped over using a for loop and the Zscore function. The Zscores were appended to a list, which was then added to the dataframe.

	Venues	Raw_Counts_Seattle	Ratio_Counts_Seattle	Raw_Counts_Portland	Ratio_Counts_Portland	Z_Score
12	Trail	38.0	1.307639	37.0	1.309271	-0.005473
305	Bed & Breakfast	1.0	0.034412	1.0	0.035386	-0.019880
282	Food Stand	1.0	0.034412	1.0	0.035386	-0.019880
277	Campground	1.0	0.034412	1.0	0.035386	-0.019880
270	Golf Driving Range	1.0	0.034412	1.0	0.035386	-0.019880

Each category, with the venue's ratio in each respective city, and the Z-score

A z-score with a large magnitude indicates that it is very unlikely that we would get those respective proportions if the actual underlying proportions were equal. A z-score with a small magnitude indicates that the actual proportions were not unlikely if the actual underlying proportions were equal. Setting an alpha value of 0.05 indicates a 5% risk of us making the error of concluding a difference when there is no actual underlying difference. This value of alpha corresponds to a critical value of 1.96. This means that we fail to reject the null hypothesis in cases where Z has a magnitude less than 1.96, and reject it when Z has a magnitude greater than 1.96.

I then created 2 new dataframes. One in which the magnitude of the Z-score is greater than 1.96 called differences, and one in which the magnitude of the Z-score is less than 1.96 called similarities.

	Venues	Raw_Counts_Seattle	Ratio_Counts_Seattle	Raw_Counts_Portland	Ratio_Counts_Portland	Z_Score
34	Hotel	20.0	0.688231	49.0	1.733900	-3.643776
14	Playground	37.0	1.273228	11.0	0.389243	3.711102
24	Convenience Store	27.0	0.929112	60.0	2.123142	-3.712318
21	Food Truck	28.0	0.963524	63.0	2.229299	-3.849129
27	Beach	25.0	0.860289	0.0	0.000000	5.010637

The 5 venue categories with the greatest magnitude of Z-score

	Venues	Raw_Counts_Seattle	Ratio_Counts_Seattle	Raw_Counts_Portland	Ratio_Counts_Portland	Z_Score
12	Trail	38.0	1.307639	37.0	1.309271	-0.005473
80	Indian Restaurant	9.0	0.309704	9.0	0.318471	-0.059724
75	Juice Bar	10.0	0.344116	10.0	0.353857	-0.062965
48	Taco Place	14.0	0.481762	14.0	0.495400	-0.074553
57	Salon / Barbershop	14.0	0.481762	14.0	0.495400	-0.074553

The 5 venue categories with the smallest magnitude of Z-score

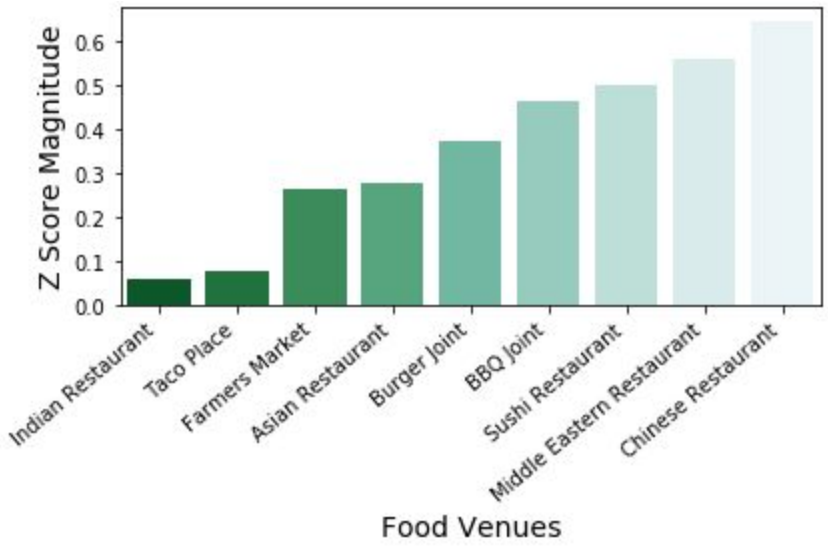
The differences dataframe contains information that could be used to make inferences on the differences between the two cities of Portland and Seattle, but that was not the purpose of this analysis. Rather, our goal was to find the venues which had a similar representation in each city indicated by a low magnitude of Z_score. For the dataframe of similarities, I filtered it so that the Raw Count for the venue in either Seattle or Portland had to be at least 8. This would eliminate the large amount of venues that had very few raw counts, and would be more difficult to draw broad practical assumptions from. Additionally I trimmed it down even further to only include those values with an especially small magnitude, under 0.7. This provided a final dataframe of 32 venues.

D. Results

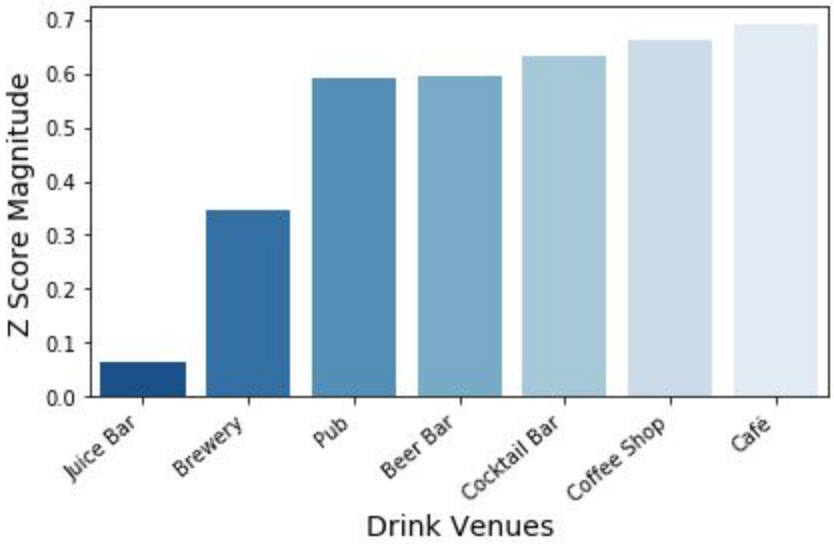
Using the Foursquare API we were returned 2906 unique venues from Seattle and 2826 unique venues from Portland. These 5732 unique venues were composed of 398 unique venue categories including specific labels such as “Spa” or “Chinese restaurant”. A Z-test for 2 proportion was performed across each of these 398 unique venue categories using an alpha of 0.05, and the data was further segregated into 2 dataframes. One in which the magnitude of the Z_score was over 1.96, and one in which it was less than 1.96, labeled differences and similarities respectively. The dataframe of particular interest, the one of similarities, possessed 361 of the 398 unique venue categories. After venues in which the number of venues in either the Seattle or Portland column was less than 8 were

removed, this left 73 venues. Finally a Z_Score cutoff magnitude of 0.7 was selected to represent those venues that were the most similar, leaving 32 venues of the original 398.

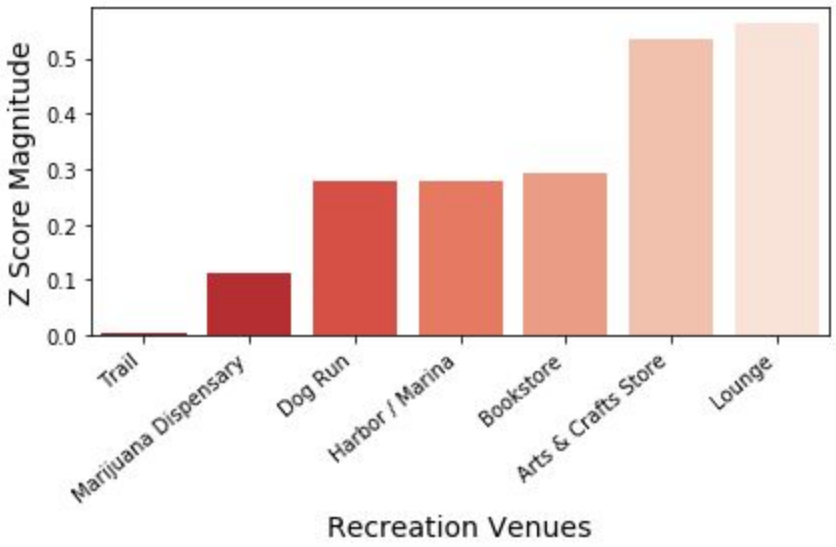
Those 32 venues were sorted into 5 different groupings. These groupings were Food, Drink, Recreation, Wellness, and City Essentials. Each group was plotted on a bar graph, save for City Essentials which includes venues that are common amongst almost any city. This includes venues such as gas stations or banks.



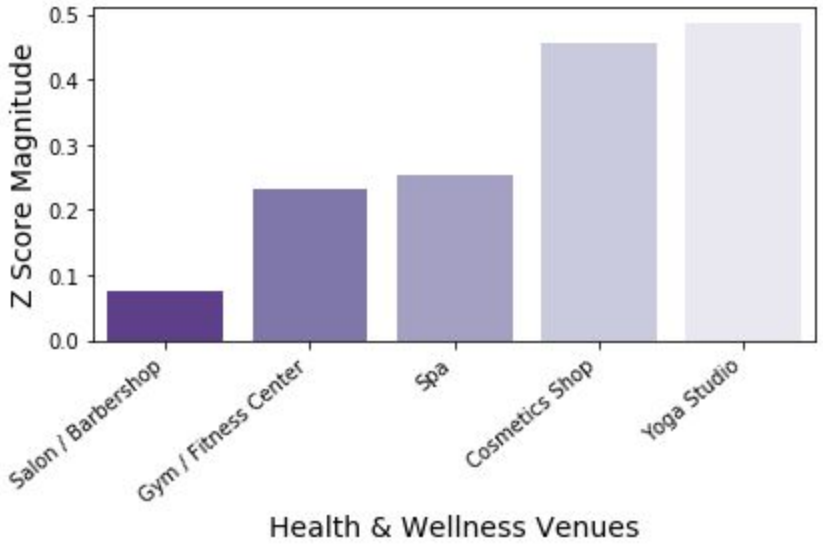
Graph of the common food venues which display the most similarity in representation



Graph of the common drink venues which display the most similarity in representation



Graph of the common recreation venues which display the most similarity in representation



Graph of the common recreation venues which display the most similarity in representation

E. Discussion

Based on these results, we can see a few possible trends that emerge from the data. These 2 millenial cities possess a very diverse international food scene. There is representation from multiple foreign cuisines, in addition to classic Americana such as Burgers and BBQ. Millennials appear to have a varied appetite and high degree of open mindedness.

There was a significant amount of alcohol and caffeine related establishments in each city, and the proportion of these was quite comparable. When viewed alongside of additional information, this can be solidified into a trend. After all, as of 2016 a Bloomberg study revealed that millenials consumed

44% of the coffee in the United States (Perez, 2016). This is despite making up only about roughly 25% of the population. In a similar vein, during the 7 year span of 2009 to 2016, millennials experienced the highest average increase in cirrhosis related mortalities (Tapper & Parikh, 2018).

From the table of recreation venues, we can see that millennials enjoy nature and indulging in mind-altering substances. In fact the data shows that nearly 1 in 4 people in the United States age 18-29 are regular users of cannabis (Statista Research Department, 2019).

Finally, despite their higher incidences of alcohol abuse, millennials possess health-conscious outlooks. They indulge in healthy activities such as yoga, or cardiovascular exercise, and are proponents of self-care rituals including spa days.

One final thing of note is the limitations of this analysis. Firstly, this analysis is limited to 2 cities, and not only that but cities which belong to the same geographic region. Although much of the data is corroborated by outside information, one should not understate the impact that this geographic isolation may have. To gain a more complete understanding, this process should be done on many more cities with a large millennial population, and compared against cities with a very small or shrinking millennial population. There are more factors involved that should be isolated, although this is a good starting point. The second limitation is in regards to the collection of the foursquare data. It should be analyzed to see if the data is actually representative of the total venues that actually exists within these cities, or if there is some bias at play. For example, younger people may be more likely to use foursquare, and this may influence the types of venues which have their data collected.

F. Conclusions

Based on this preliminary analysis of two booming millennial cities, we can conclude a few things. Although further research should be done into the topic, we can at least say with some degree of certainty that of the venues returned from foursquare's api in both the cities of Seattle and Portland, the venues that have the most similar representation are international restaurants, venues that serve physiologically altering substances(caffeine, alcohol, cannabis), nature trails, and wellness centers(spas, yoga studios, gyms). If we take this at face value, it appears that a certain picture is painted displaying the vague underlying values that are possessed by millenials, and are represented by their preference of venue.

G. References:

Bialik, K., & Fry, R. (2019, February 14). How Millennials compare with prior generations. Retrieved from https://www.pewsocialtrends.org/essay/millennial-life-how-young-adulthood-today-compares-with-prior-generations/

Dimock, M. (2019, January 17). Defining generations: Where Millennials end and Generation Z begins. Retrieved from https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/

Donnelly, C., & Scaff, R. (2019). Who are the Millennial shoppers? And what do they really want? Retrieved from https://www.accenture.com/us-en/insight-outlook-who-are-millennial-shoppers-what-do-they-really-want-retail

Geier, B. (2019, July 17). Where Millennials Are Moving 2019. Retrieved from https://smartasset.com/mortgage/where-millennials-are-moving-2019

Perez, M. G. (2016, October 30). Coffee-Loving Millennials Push Demand to a Record. Retrieved from https://www.bloomberg.com/news/articles/2016-10-30/millennial-hunt-for-caffeine-fix-propels-coffee-demand-to-record

Statista Research Department. (2019, August 9). Americans who smoke marijuana by age group 2018. Retrieved from https://www.statista.com/statistics/737849/share-americans-age-group-smokes-marijuana/

Tapper, E. B., & Parikh, N. D. (2018, July 18). Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: observational study. Retrieved from https://www.bmj.com/content/362/bmj.k2817

U.S. Census Bureau (2017). American Community Survey 1-year estimates. Retrieved from Census Reporter Profile page for Seattle, WA <http://censusreporter.org/profiles/16000US5363000-seattle-wa/>