Portland Gourmet

- **Programme:** MSc Smart Cities and Urban Analytics
- **Department:** The Bartlett Centre for Advanced Spatial Analysis
- **Module:** Spatial Data Capture, Storage and Analysis
- **Group 5:** World Capturer
- **Student Name:** Ying Chen, Zhengtao Yu, Hui Li, Yunqing Bai, Qingying Zhao
- **Word Count:** 5448
- **Website:** http://dev.spatialdatacapture.org/~ucfuyba/
- **API Documentation:** http://dev.spatialdatacapture.org:8881/
- **GitHub:**https://github.com/LouiseZhaozzz/Portland-Gourmet_Group5_World-Caputurer.git

1 Introduction

The modern restaurant industry is placing increasing emphasis on the role of dining ambience in the consumer experience (Liu, 2009) Ambience can be seen not only as a competitive advantage for restaurants but also as a determinant of customer satisfaction (Agoes and Pasaribu, 2015), with different atmospheres attracting different consumer groups.

In this project, we seek to explore restaurant atmospheres and their spatial distribution in Portland, USA. The NLP analysis was used to identify specific themes of high frequency and distinctive ambience. Hierarchical cluster analysis was then used to discover significant spatial differences between themes and to understand the distribution of local consumer groups. In addition, the analysis provides the best restaurants in each neighbourhood for that particular cluster and the results can be used to help restaurants better improve their service. A city perspective allows for improved urban planning and management. In addition, an interactive web page will visualise the results of our analysis.

2 Literature Review

People may have a prejudgment of restaurants' reputation and form their attitude to the product by reviews from the customers who have been to these restaurants (Lim and van der Heide, 2015). Reviews could be regarded as records of customers' impression of the restaurant. Therefore, these reviews could help to understand the atmospheres of the restaurants.

Sentiment analysis is one of the hot topics in the field of review research in recent years, mainly focusing on exploring positive and negative words (Tausczik and Pennebaker, 2010; Kang, Yoo and Han, 2012; Haddad et al., 2015), the factors that affect positive or negative emotions (Tian, Lu and McIntosh, 2021), and the sentiment analysis of a specific language like Japanese (Nakayama and Wan, 2019) and Cantonese (Zhang et al., 2011). In addition, green restaurants is another hot topic in this research area. For example, the impact of green restaurant attributes on customer satisfaction (Park et al., 2020) and their cognition of sustainable food (Park, Chae and Kwon, 2018).

Although many studies have been done on online reviews, most of these studies are still single-dimensional. For example, most sentiment analyses only divide restaurants into two groups (positive and negative) according to document-level overall sentiment (Bin et al., 2011). Relatively little attention has been paid to the composition of multidimensional topics in overall reviews (Vallurupalli and Bose, 2020). At present, multi-aspect work is mainly based on Latent Dirichlet Allocation(LDA)（Blei, 2012). For example, Özdağoğlu, Kapucugil-İkiz and Çelik (2018) used unsupervised LDA to divide restaurant reviews into 19 categories. In addition, some studies also used Supervised LDA to explore potential hot topics (Zhan, Loh and Liu, 2009; Mukherjee and Liu, 2012).

3 Research Question

The effects of restaurant ambience on customer experience were discussed in the literature review, and it can be hypothesised that restaurant ambience will significantly influence the overall effect of the consumer service experience and therefore customer satisfaction. Furthermore, overall satisfaction may be influenced by multiple independent attributes and it is important to measure their relative importance (Sulek and Hensley, 2004). The city of study for this paper is Portland, which is the largest and most populous city in Oregon (City of Portland, 2021) and also contains the highest number of parks per capita in the USA (Portland Tourism, 2019). Portland was also ranked the fourth-best city for food in the United States by the Washington Post (Melina, 2015).

The purpose of this analysis and website is to identify Portland's restaurant ambience themes by summarizing important customer evaluation attributes, to analyze the differences in restaurant characteristics and spatial distribution in different areas of Portland, and to characterize the corresponding consumer groups. Understanding the trends and important attributes of Portland restaurants in terms of spatial distribution within different environmental themes. Through differences in distribution, coupled with character and ambience themes, an analysis will be explored whether the spatial distribution can provide recommendations for visitor dining choices, restaurant operations, commercial companies and government strategies.

4 Research Area & Data

4.1 Research Area

The distribution and regional characteristics of Portland are shown in the figure below (USCB, 2019):

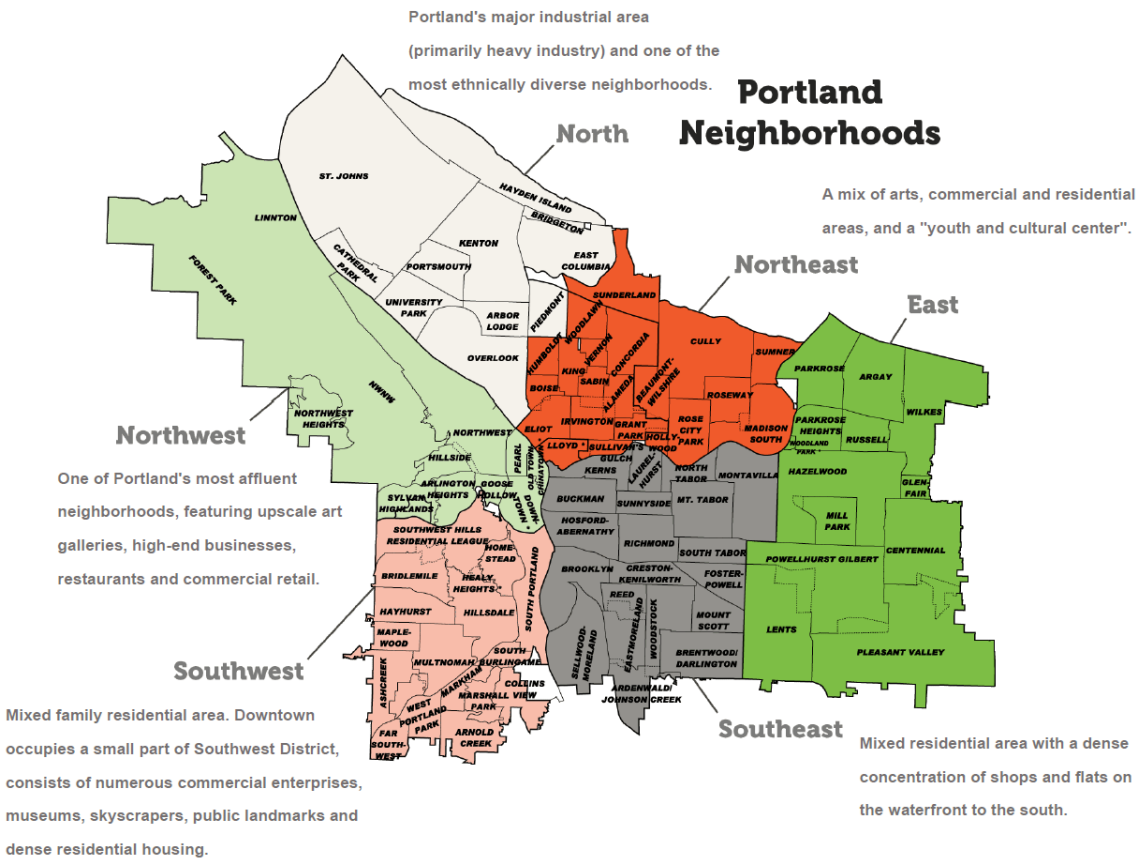

Figure 1. Portland Spatial Analysis

Schrock et al. (2015) indicate that Portland's foodservice industry exhibits a very clear tendency towards "centrality". The largest food services are concentrated in the core of the city, including the Downtown, Northwest, and Eastside areas, as well as in the corridors outside of the downtown business district, as the scale tends to be closer to population and employment centres.

In addition, the development of residential areas and consumer demand has significantly stimulated retail investment throughout the city; while the growth in tourism activity has led to growth in the number of restaurants offering full-service and specific dining needs, particularly in downtown Portland.

4.2 Datasets

Yelp is a popular online review community, which has been widely proved to have a great influence on consumers' decision-making (Ariyasriwatana and Quiroga, 2016). The Yelp dataset is a subset of businesses, reviews, and user data provided by the Yelp website (Yelp, n.d.), which involves 8 metropolitan areas, over 1.2 million business attributes, and over 8.6 million reviews. Exploring Portland's Restaurant data, there are **5644** restaurants in Portland with at least one yelp review, and **750395** reviews for those restaurants in total. However, due to computational limitations for text processing and

meaningless reviews for too long ago, this project only selected the recent 4 year's reviews, which leaves **297993** reviews in Portland since 2017/1/1.

Table 1. Dataset Information

Dataset	Source	Format	Attributes	Mainly Useful Information
Business	Yelp website	json	12	Restaurant ID，name，location，stars，review count **for each restaurant**
Review	Yelp website	json	9	Text content, stars，date, restaurant ID **for each review**
Portland Neighborhood Boundary	Open Data	shapelfile	11	Name, location and size of each community **for each neighborhood**

The Table shows the specific information of datasets. For example, business and review dataset, both from yelp. One is the information for each restaurant, which contains 12 attributes like restaurant ID, name, location, average stars, total review count. Another is the information for each review, including 9 attributes like restaurant ID, text content, stars, date. Besides, it also used the Portland neighbourhood boundary from Open Data (n.d.), which is a shapefile for spatial analysis.

5 Methodology

5.1 Review Data analysis method

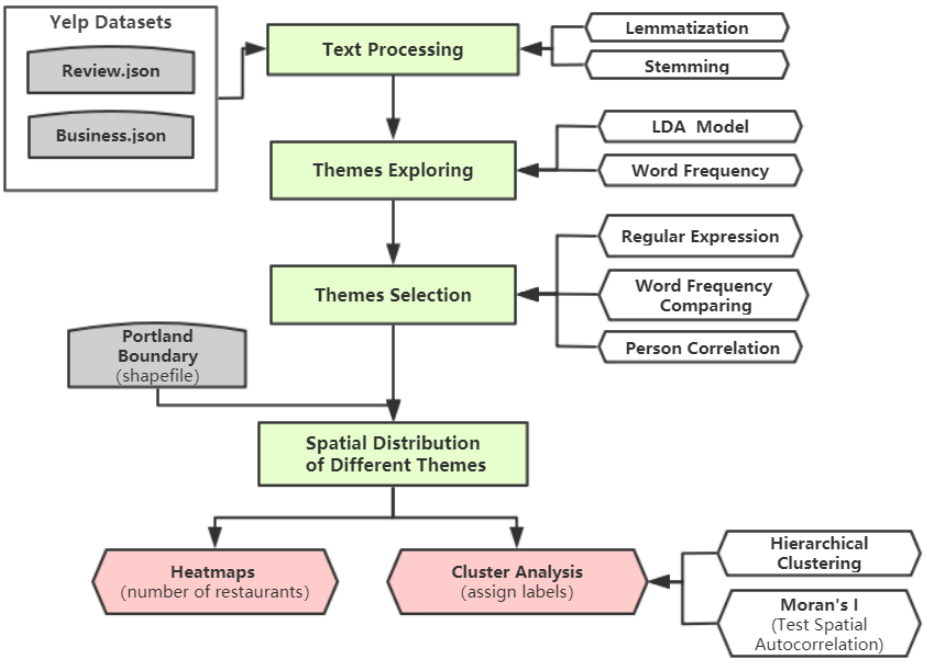

Figure 2. A flowchart of our methodology for data analysis in python

5.1.1 Text Processing

The goal of text processing is to remove unnecessary words and use the resulting clean texts in the LDA model to explore themes revealed in the reviews. To clean texts, we mainly used the nltk package and SpaCy library in python and the first step was to remove unwanted words, such as

characters, numbers, symbols and stop words. And checking spelling mistakes and converting all words to lowercase were also needed. Then, lemmatization and stemming were used to normalize the morphology. Korenius et al., (2004) and Balakrishnan and Lloyd-Yemoh (2014) have evaluated the overall precision performances between lemmatization and stemming and found lemmatization is a more precise language modelling technique than stemming but the differences are minor. The main difference is that stemming tries to remove and reduce affixes (not necessarily to express complete semantics), while lemmatization is to restore a word to its prototype (which can express complete semantics). There is a little cross between the two results and we will explore both methods and possibly choose one or more methods based on their actual performance.

5.1.2 Themes Exploring

To explore atmosphere-related keywords from reviews, our initial thought is to use the clean texts to build Latent Dirichlet Allocation (LDA) models, which is commonly used in thematic topic modelling. In short, the LDA model is a statistical model for discovering hidden topics in document sets (Blei, 2012). In LDA, words in a document are regarded as a "bag of words" model, and the distribution is inferred according to the word frequency in the document (Wang, Gu and Xu, 2018). And we computed LDA models by using the pyLDAvis package. We tried different topic numbers from 5 to 10 and different ranking stars in the LDA model, seeking the optimal classification results and comparing the differences between reviews for high-star restaurants and low-star restaurants. Besides, word frequency was also used to compare the importance of themes we have discovered.

5.1.3 Themes Selection

The results of LDA models are not ideal so we manually created a list of potential themes and keywords reflecting a restaurant's ambience. We calculated word frequency for different NLP techniques and themes and then chose to use Snowball Steamer to classify reviews by themes using the Regular Expressions. For example, the theme Family extracts reviews that contain either of the word "family", "kid", or "children", but because we applied Snowball Steamer to the texts, the word "family" becomes "famili". The figure below shows the python code in this step and it returns two new columns named "family_count" which shows the number of reviews containing family keywords for a restaurant and "family_rate" is normalized by dividing by the total available reviews regardless of themes. We repeated this step for all themes and the index of our final data frame is a restaurant's business id and some important columns are the count and the rate for all themes, and geometry of all restaurants. The final data frame prepares us for clustering Portland's neighbourhoods by our themes. Moreover, with the total nine themes we have created, to avoid themes with similar spatial representations, we computed the Pearson Correlation coefficients to only select themes that yield the best cluster.

```
family = df_review[df_review.snowball_reviews.str.contains(r'(?:kid|child|famili)',
                                                           regex=True, .groupby('business_id').
                                                           agg({'snowball_reviews':['count']}).reset_index()
family.columns =family.columns.get_level_values(1)
family.rename(columns={'':'business_id', 'count':'family_counts'}, inplace=True)

df_restaurant = pd.merge(df_restaurant,family, how= 'left',  left_on  = 'business_id',  right_on = 'business_id').filln
df_restaurant['family_rate']=df_restaurant['family_counts']/df_restaurant['number_of_reviews']
df_restaurant[5:7]
```

attributes	categories	number_of_reviews	liquor_counts	liquor_rate	vegan_counts	vegan_rate	romance_counts	romance_rate	family_counts	family_rate
Level': "u'loud'", hol': "u'none'"...	Restaurants, Mexican, Food Trucks, Food, Food ...	12.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.166667
itsPriceRange2': ', 'Alcohol': "u'...	Sushi Bars, Restaurants	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

Figure 3. Demonstration of classifying reviews by the Theme Family and adding back to our data frame.

5.1.4 Hierarchical cluster analysis

Spatial hierarchical clustering can define the level of spatial autocorrelation of different features as well as evaluating the statistical significance of clusters (Wang et al., 2020). The basic hierarchical clustering comes in two forms: agglomeration (beginning with each observation as the initial cluster and merging them to a single cluster) and division (starting with only one big cluster and dividing it into several clusters)(Wang et al., 2020). The similarity between objects needs to be quantified by distance measurement. Our analysis is based on geographic space and uses the two-dimensional Euclidean distance when clustering. Moreover, the distance between clusters needs to be measured through the linkage function. Among which, the Ward minimum variance method is defined as the minimum increase in the sum of intra-cluster variances due to the merging of two clusters (Govender and Sivakumar, 2020), which performs better than other clustering procedures in most cases and could give the best overall recovery (Saraçli, Doğan and Doğan, 2013). As for finding the optimal number of clusters, the hierarchical clustering algorithm generates a dendrogram, which illustrates a hierarchy of nodes and the sequence of nested clusters (Dubes and Jain, 1976; Fouedjio, 2016).

5.2 Visualization method

The website construction is mainly divided into two parts: web pages and maps (Figure 4). For web pages, HTML5 is the main programming language. The design and beautification of the webpage are done through Bootstrap. Bootstrap is the most popular HTML, CSS, and JS framework for the development of responsive layout, mobile device-first WEB projects. The website's responsive components and image layout are all done in Bootstrap.

For interactive maps, the map style API comes from Mapbox. Mapbox provides rich map styles and powerful data upload functions, but it cannot complete the embedding of multiple views and the correlation response between different views. Therefore, Tableau is used for interactive map production and data visualization. Tableau has a wealth of data visualization options. The biggest

advantage of this software is the ability to upload interactive maps and data to the Tableau Public server. In addition, Tableau can establish connections between different data files to ensure that only the required data is uploaded. This feature greatly enhances the stability of the interactive map and the speed of interactive feedback.

The final website is hosted on the UCL server, and all data and web page components are also stored under the same path.

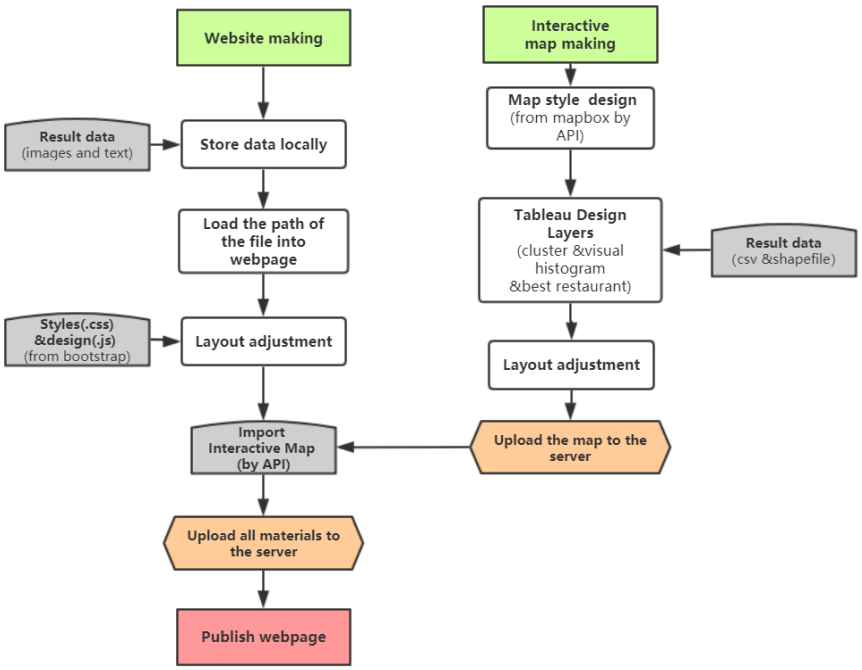

Figure 4. Flow Chart of Website Building

6 Results

6.1 Theme selection

6.1.1 Text processing result

Comparing different NLP's performances, we found that, although lemmatization performs better in keeping English words in the simplest form and still readable, it automatically deletes many repeated words within a review and reduces the total word counts for most themes. We think repeated words in a review should be kept because it reflects the importance of the word and the intensity of the theme is highlighted. For example, the word "family" occurs 13624 after stemming and 13493 after lemmatization, showing stemming performs better in keeping repeated words. We also compared two stemming techniques, Snowball Stemmer and Porter Stemmer from nltk stem packages. For example, the word "glorious" in the original text becomes "gloriou" in Porter Stemmer but keeps the same in Snowball Stemmer, showing Porter Stemmer is more likely to over-stemming. Snowball Stemmer is

an upgraded version of Porter Stemmer (Willett, 2006) and also known as Porter2 stemming (Heidenreich, 2018). Despite no differences in word counts using Porter or Snowball Stemmer, given Snowball Stemmer is updated, we chose Snowball Stemmer to classify reviews by themes in the next step. Figure 5 shows our review data frame after text processing using three NLP techniques.

	business_id	stars_r	snowball_reviews	porter_reviews	lemmatized_reviews
173627	xeFke1EuoYXOZaX1rTQYEw	5.0	wow best valu town spici chicken eggplant stea...	wow best valu town spici chicken eggplant stea...	good value town spicy eggplant steam rice egg ...
90508	4CxF8c3MB7VAdY8zFb2cZQ	4.0	want know voodoo hype person disappoint wait b...	We want know voodoo hype wa person disappoint ...	voodoo hype personally wait donut insane hour ...
226726	mRoX9rmDE-BM8Dg0eFasmw	5.0	expens onli thing save hungov af worth yessir	It expens onli thing save I hungov af worth ye...	expensive thing
92153	hXToK6UfN3BOXy7UScAmXw	1.0	iam peopl catch negat horribl servic rude peop...	iam peopl thi catch neg horribl servic rude pe...	people negative there horrible service rude pe...
210109	zWPEVx6Da-fixuYXDBMstQ	3.0	love locat stationari food truck littl lot sur...	love locat thi stationari food truck littl lot...	stationary food truck little lot food truck mi...

Figure 5. Demonstration of classifying reviews by the Theme Family and adding back to our data frame.

To illustrate the most commonly used 300 words in yelp reviews, we generated a word cloud (Figure 6) based on all lemmatized reviews except for some general unimportant words such as food, good, try, come, time, etc which were set as stopwords and excluded from the rest of texts. The word cloud illustrates words such as “atmosphere”, “experience”, “service” are also commonly used in reviews besides food items and adjectives. It also reveals people care about a restaurant’s atmosphere and their own experiences as much as they do for the food quality and price.

Figure 6. Word Cloud generated in python showing the most common 300 words in all reviews after lemmatization.

6.1.2 LDA model result

The results of LDA models did little contribution to identifying themes. Using all reviews for LDA models gives us no clusters that describe ambience. However, the reviews only for 5-star restaurants

formed a cluster of "vegan" and "vegetarian" (Table 2), giving us insights into creating a theme of Vagan. The reviews for 1-star restaurants (Table 3) revealed some reasons for bad reviews such as terrible service or attitude, cold delivery or other unsatisfied customer services. Overall, due to the unsupervised nature of LDA modelling, most clusters are grouped by food items or customer's experience which are not what we want to describe ambience. Therefore, borrowing some theme ideas from Rahimi, Andris and Liu (2017), we set our initial themes to nine categories and they are liquor, vegan, romance, family, classy, student, outdoor, music, and tourist.

Table 2. Terms with the highest beta values for 5-star restaurants

Topic cluster 1		Topic cluster 2		Topic cluster 3		Topic cluster 4		Topic cluster 5	
Term	Beta value	Term	Beta value	Term	Beta value	Term	Beta value	Term	Beta value
service	0.048	vegan	0.035	breakfast	0.02	fresh	0.022	chicken	0.057
experience	0.021	option	0.015	brunch	0.017	café	0.016	pizza	0.048
staff	0.021	sandwich	0.013	taco	0.017	coffe	0.015	cheese	0.023
friendly	0.016	vegetarian	0.011	egg	0.016	bowl	0.015	pie	0.013

Table 3. Terms with the highest beta values for 1-star restaurants

Topic cluster 1		Topic cluster 2		Topic cluster 3		Topic cluster 4		Topic cluster 5	
Term	Beta value	Term	Beta value	Term	Beta value	Term	Beta value	Term	Beta value
service	0.037	customer	0.023	location	0.037	pizza	0.118	chicken	0.031
terrible	0.012	business	0.022	service	0.03	cold	0.011	sauce	0.012
horrible	0.012	owner	0.015	customer	0.03	delivery	0.009	flavor	0.01

6.1.3 Word Frequency Comparison result

To check whether the nine themes we have created have enough word counts to be considered as important, Figure 7 shows word frequency for all of our potential themes. In the x-axis, the words in the same colour are different keywords under a theme. We sorted them based on the total number of word counts by themes. Liquor, showing in the red is the most frequent one and represented by different types of alcohol so this theme generally refers to bars. Romance, family, student, and tourist are mainly classified by target consumers using nouns such as "wife/husband", "girlfriend/boyfriend", "children", etc. By contrast, the theme Classy mainly used adjectives such as cosy, trendy, modern, etc. to highlight the atmosphere. Vegan provides a very specific cuisine so we put it alone. Outdoor and music focus on the built environment.

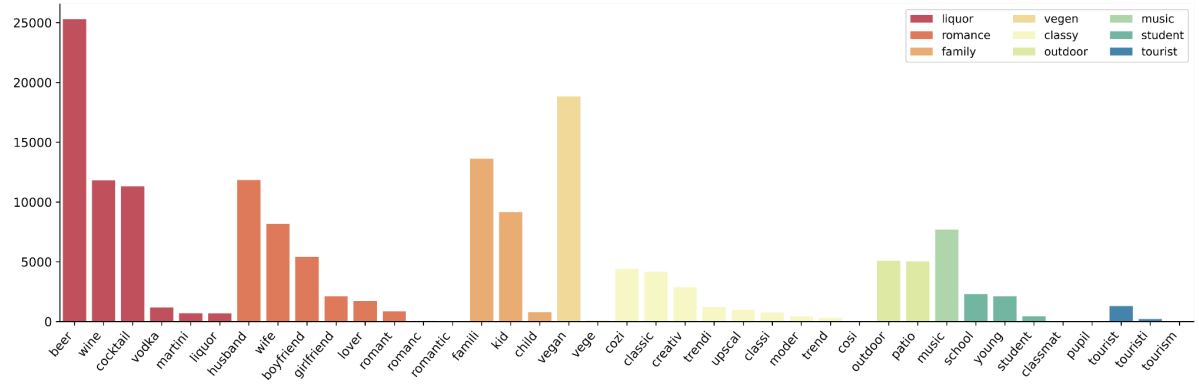

Figure 7. Bar chart created in python to show the word frequency for all of our potential themes and keywords in each theme.

Specifically, Table 4 detailed listed word frequency for all themes and keywords we have created. Looking at total word counts for each theme, the table shows Liquor, Vegan, romance, Family, Classy, and Outdoor are common themes that contain more commonly used words in reviews.

Table 4. Total word counts for all potential themes and keywords

theme	keyword	word stem	frequency	total
liquor	beer	beer	25292	50991
	wine	wine	11816	
	cocktail	cocktail	11314	
	vodka	vodka	1187	
	martini	martini	695	
	liquor	liquor	687	
vegan	vegan	vegan	18826	18900
	vegetarian	vege	74	
romance	romantic	romant	850	30171
		romantic	6	
	romance	romanc	53	
	boyfriend	boyfriend	5417	
	girlfriend	girlfriend	2100	
	lover	lover	1737	
	husband	husband	11832	
	wife	wife	8176	
family	kid	kid	9139	23546
	child	child	783	
	family	famili	13624	
classy	classy	classi	762	15218
	upsacle	upscal	978	
	modern	moder	443	
	creative	creativ	2870	
	trend	trend	318	
	trendy	trendi	1218	
	classic	classic	4163	
	cozy	cozi	4401	
	cosy	cosi	65	
student	student	student	442	2758
	classmate	classmat	16	
	pupil	pupil	5	
	school	school	2295	
outdoor	outdoor	outdoor	5096	10133
	patio	patio	5037	
music	music	music	7705	7705
tourist	tourist	tourist	1296	1520
	tourism	tourism	19	
	touristy	touristi	205	

6.1.4 Pearson Correlation Matrix

With the total nine themes we have created, our goal is to select themes that have unique spatial representations. Recall we have calculated theme intensity named as "family_rate", "romance_rate", etc in our final data frame and based on all restaurant's theme intensity, we calculated the Pearson Correlation coefficients in python. If two themes always select the same number of reviews for every restaurant, they will be highly correlated and will show no difference in the clusters and we need to either remove one or combine them. The matrix (Figure 8) demonstrates the Liquor and Outdoor as well as the Student and Family are positively correlated (> 0.3), so we removed the Outdoor and the Student because they have fewer reviews (i.e. less important themes).

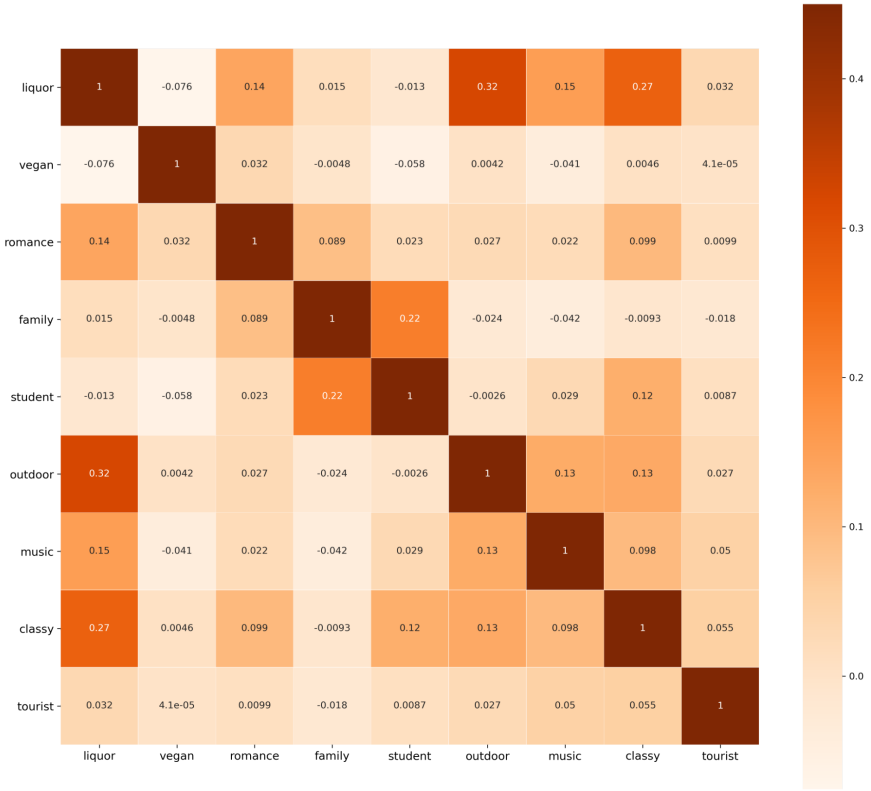

Figure 8. Pearson Correlation Matrix computed in python showing the correlation between theme intensity.

6.2 Spatial distribution of different themes

6.2.1 Heatmap

After checking the word frequency and the Pearson Correlation matrix, our final themes are Classy, Family, Liquor, Romance, and Vegan. To show how our themes are geographically related to restaurant's ranking scores, we also used the rating stars (1 to 5) from the Yelp Dataset. Figure 9 is composed of six screenshots of heatmaps made from Leaflet with an open street base map. It shows all themes sharing a common hotspot at Downtown Portland near the city river and this is also where most 5-star restaurants are located. Other than that, every theme has its sub-hotspot dispersed across the city.

Figure 9. Screenshots of heatmaps generated in python from Leaflet showing the geographical distribution of final themes compared with the restaurant's ranking stars.

6.2.2 Spatial Clustering

We employed agglomerative hierarchical clustering of the sklearn package in Python (Pedregosa et al, 2011), and we used 5 themes (Vegan, Classy, Family, Liquor, Romantic), number of reviews and star ratings to generate the clustering results. Next, we used a dendrogram to find the optimal number of clusters (Figure 10). Our analysis generates 5 clusters distributed across Portland by using linkage criterion "ward" and the set distance threshold to 5 without connectivity matrix.

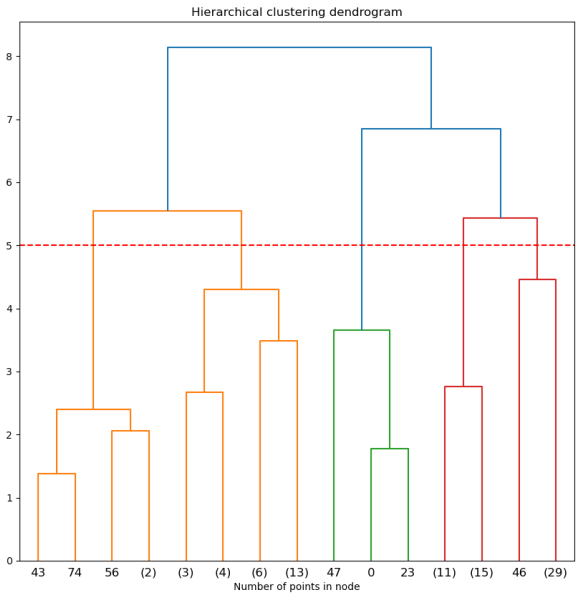

Figure 10. Dendrogram of the spatial hierarchical clustering result, when set distance threshold to 5, 5 clusters are defined.

Firstly, the obvious spatial differences and clusters among Portland by the restaurant ambience theme can be identified in Figure 11 below: Cluster 0 is dominant in the centre of Portland, with small parts distributed in outer Portland. Cluster 1 tends to group in the outer edges of Portland. While Cluster 2 and Cluster 3 majorly located in the western part of the city. And Cluster 4 randomly distributed across the city. Secondly, each cluster also varied in terms of numbers. Cluster 2 and Cluster 3 only contain 3 and 5 neighbourhoods, while 30 neighbourhoods are belonging to Cluster 0, and both Cluster 1 and Cluster 4 own 26 neighbourhoods respectively.

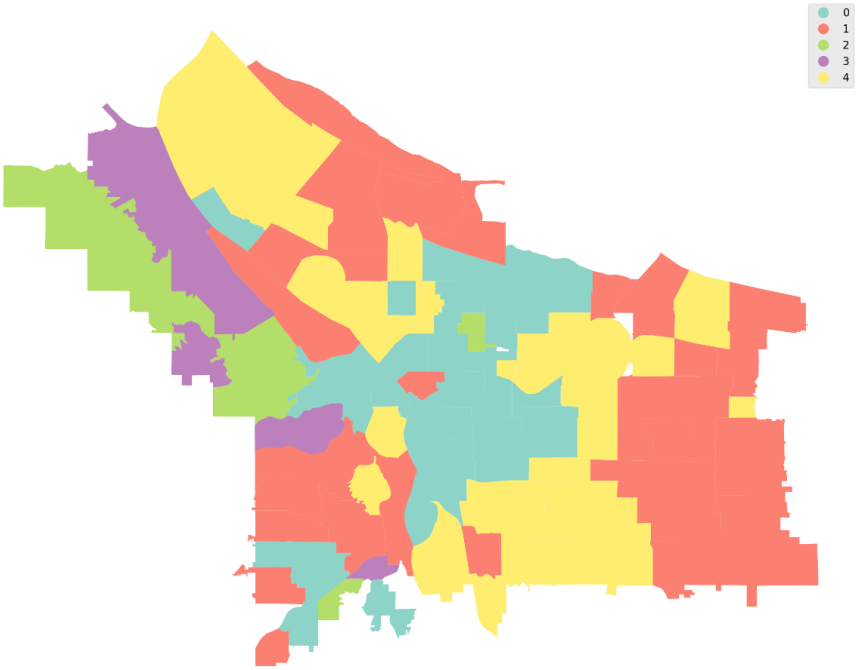

Figure 11. The spatial hierarchical clustering result

Based on the result of hierarchical clustering we assign each cluster with labels below:

Cluster 0 - Featured vegetarian food, high classy vibe with high stars

Cluster 1 - family gatherings, medium-low stars

Cluster 2 - prominent liquor

Cluster 3 - Highlighted romantic atmosphere

Cluster 4 - high stars

Figure 12 below visualises the performances of 5 clusters:

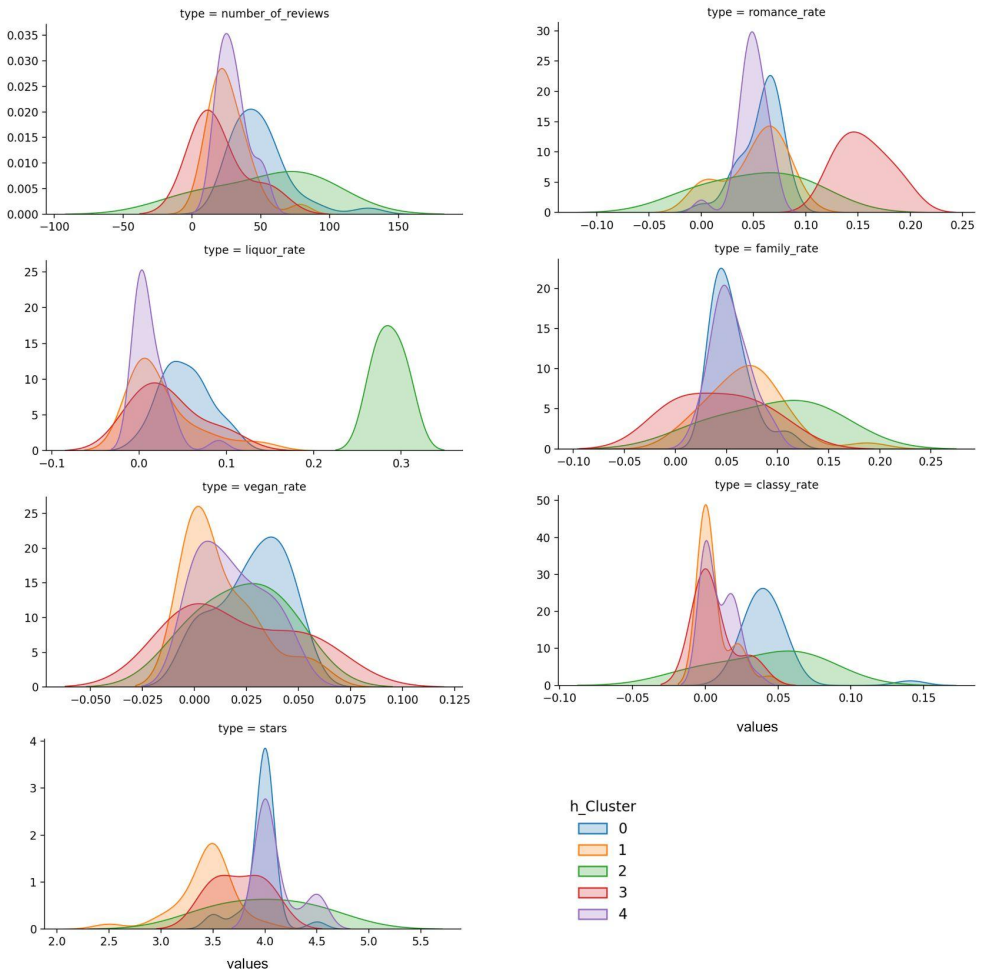

Figure 12. kernel density estimate (KDE) plot: performances of 5 clusters

Compared to the other clusters, Cluster 0 has a relatively high level of classy ambience with a high mean star rate of 4.0, meanwhile, this group owns many restaurants providing vegan food. The average star rate in Cluster 1 is slightly lower than others, but the vibe is suitable for family gatherings. Cluster 2 owns the widest review rates in general except for the beverages, which has the most outstanding performance in the 5 clusters. Cluster 3 owns the best romantic vibes and Cluster 4 has the highest average restaurant star rate.

The clustering result indicates that generally, restaurants with classy or modern vide tend to be gathered near city centres, which are some culturally diverse urban areas such as Old Town, Cully and Sunnyside, meanwhile, some restaurants serving excellent vegetarian food also gather in this area. While those restaurants welcomed by families are spread out across the cities with open space like Pleasant Valley and Mill Park in the east of the city, despite that these restaurants own an average medium-low star rate. As for dating couples, some are more likely to choose a restaurant near Forest Park or Northwest Heights, two of the more upscale parts of the Portland city (*Portland Neighborhood Guide*, no date) with outdoor seating and vast perspectives of the Tualatin Valley. Meanwhile, some restaurants providing high-quality liquor also gather around the Forest Park, a vast

woodland with only less than 10-minute commuting into downtown (*Portland Neighborhood Guide*, no date). Finally, restaurants with average high stars are distributed across the whole city, majorly in the blend of residential and commercial regions like Northeast Portland, East Portland, and North Portland. These outcomes help understand the diverse consumer groups in the neighbourhoods of Portland and might potentially attract different client groups (Rahimi, Andris and Liu, 2017).

6.3 Web Page Visualization

6.3.1 Introduction

The above analysis results will be displayed on the web page. The web page includes an introduction page, an analysis result page and an interactive map page. Users can interact and visualize on the map according to their needs. The purpose of this map is to provide dietary advice for visitors to Portland, so the main audience of this page is tourists who plan to travel to Portland. At the same time, the government can further explore the distribution of restaurants in Portland through the map, which will also help the planning of the tourism industry and the encouragement of the catering industry.

6.3.2 Web page Style Exploration

In the initial stage of web page creation, we used a lot of visual web page styles and interaction methods. Two of the examples give us great inspiration, one is the Interactive NBA Followers Map (Figure 13). This map inspired our initial interactive map style, and this visualization method of using different colours to represent different clusters and the interactive operation of displaying detailed charts and information after clicking became the basic method of our map in the end. The second is How Americans get to Work (Figure 14). This website made us realize that the user's accessibility to detailed information after selecting a specific area will affect their interactive experience. This makes the information we transmit must be short and concise, and at the same time avoid excessive text that causes viewers to feel irritated when reading.

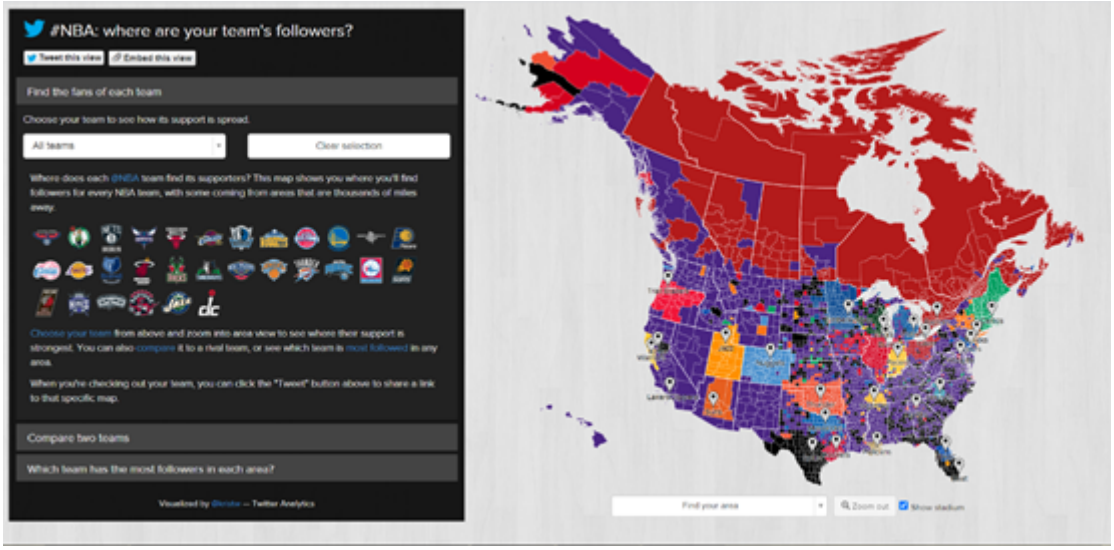

Figure 13. Example Website: Interactive NBA Followers Map
(Source: https://interactive.twitter.com/nba_followers)

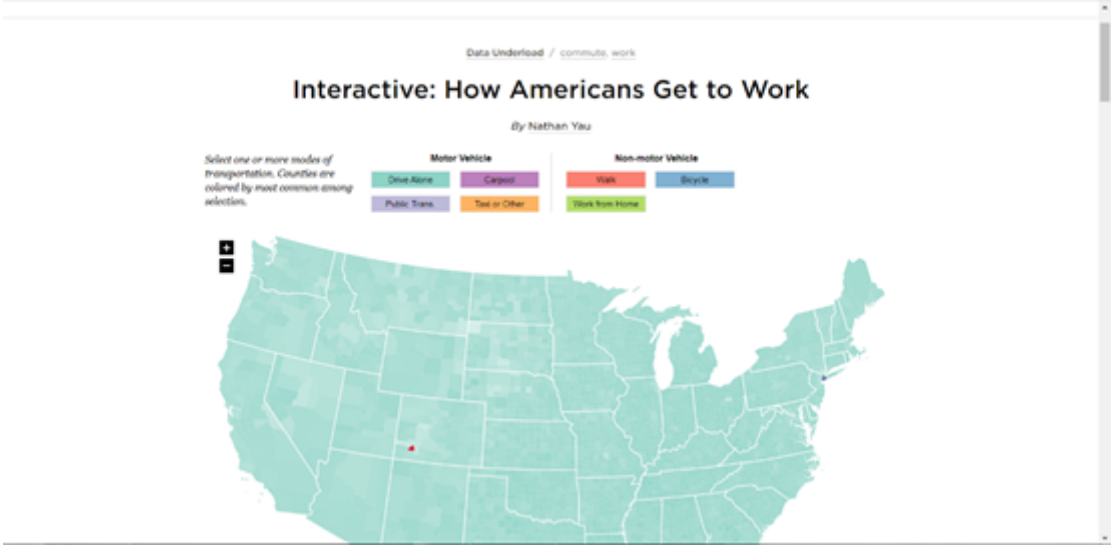

Figure 14. Example Website: How Americans get to Work
(Source: http://flowingdata.com/2015/01/20/how-americans-get-to-work)

6.3.3 Web page Design

The design of the website is based on HTML5 and CSS3. These two languages are currently the most popular web development methods. The components of the website are designed based on JavaScript. During the building process, Bootstrap was used as a development framework to improve the aesthetics and functionality of the web page.

6.3.4 External Libraries

For the map page, we imported Mapbox as an external library. Mapbox's Style API was imported into Tableau for further visualization processing. The reason for this operation is that we need two maps to implement different functions in the same view. At the same time, the two maps need to establish an

associated action instruction. When the user selects a specific area on the main map, the recommended restaurants need to be displayed on the mini-map at the same time. Tableau has a powerful dashboard function to achieve our purpose. At the same time, its rich components can allow bar graphs to be generated on the map to enhance the visualization effect. Its rich components and filter functions also help to establish a user-friendly interactive interface. Finally, Tableau allows designers to upload the map to their server (Tableau Public) and embed the map on a web page. This allows all the functions of our interactive map to be fully implemented in Tableau and publicly released. This feature makes the process of map easier (Figure 15).

JQuery, as another important tool for Web development, is also used in the design of web pages. In the analysis page, JQuery combines with an external Javascript library called Grayscale in Bootstrap to create a responsive indexing function. This feature is designed to allow visitors who are not familiar with data analysis to skip the analysis page and use the interactive map directly.

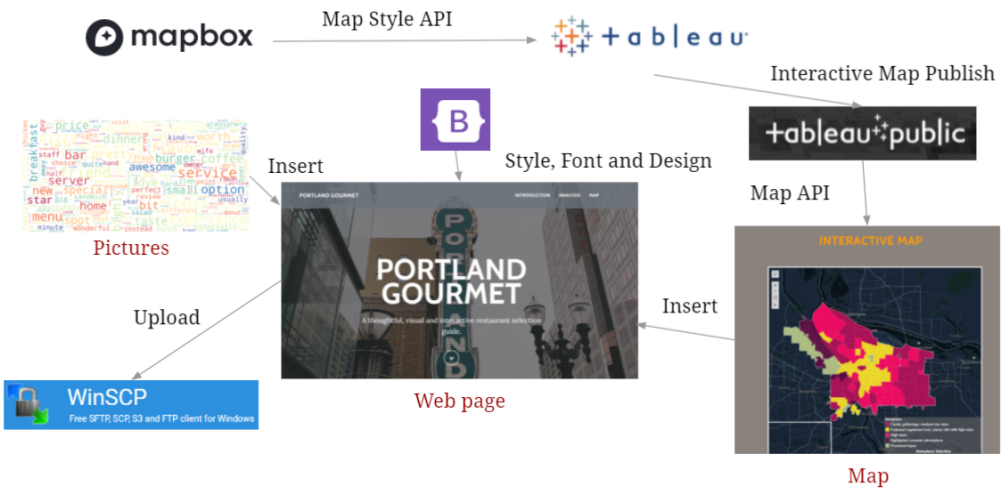

Figure 15. External Libraries Used in Website

6.3.5 User Interface Design and User Experience

Since the website's main audience is tourists who have little experience in data analysis and the ability to read and analyze charts, the interactive functions of the website need to be simplified and lightweight. It is unrealistic to give a complex histogram or box chart to the audience to read and understand. The purpose of the website is restaurant recommendation, so the final realistic result of the map should be specific restaurant information, including the name and address.

Based on this design principle, we optimized the interactive map. First, when the user opens the map(Figure 16), the interface will display a Portland map composed of different colour blocks. Each colour block represents a community, and the colour of the colour block represents the characteristics of the restaurant atmosphere in this area. Through our analysis of restaurant reviews, the five most popular styles will be displayed on the map page.

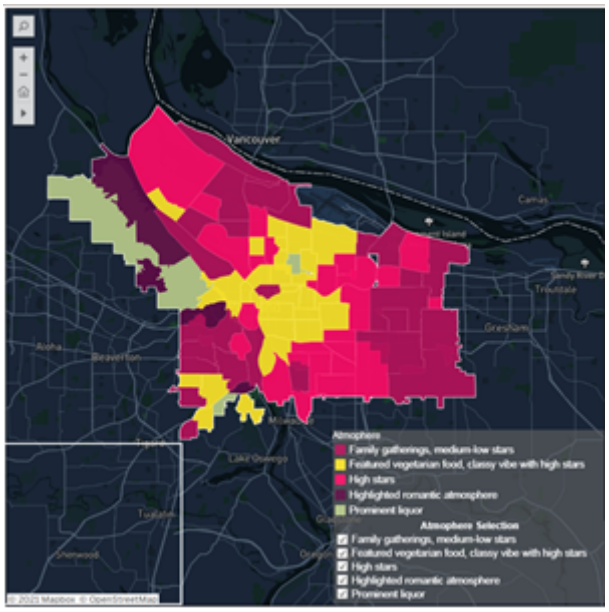

Figure 16. Initial View of the Interactive Map

Then, users can interact according to their individual needs. For example, if a user has booked a hotel in Madison South (a district of Portland), he or she can select Madison South's colour block on the map, and then he will see the most popular restaurant in this district at the bottom left In the small map of the corner (Figure 17). Another example of usage is that the user may not decide his or her destination but has certain requirements for the restaurant's service (such as providing good wine or being suitable for families). He or she can select the desired restaurant atmosphere in the filter in the lower right corner, and the map will show which restaurants in Portland perform well in this atmosphere. This can help users narrow their choices (Figure 17).

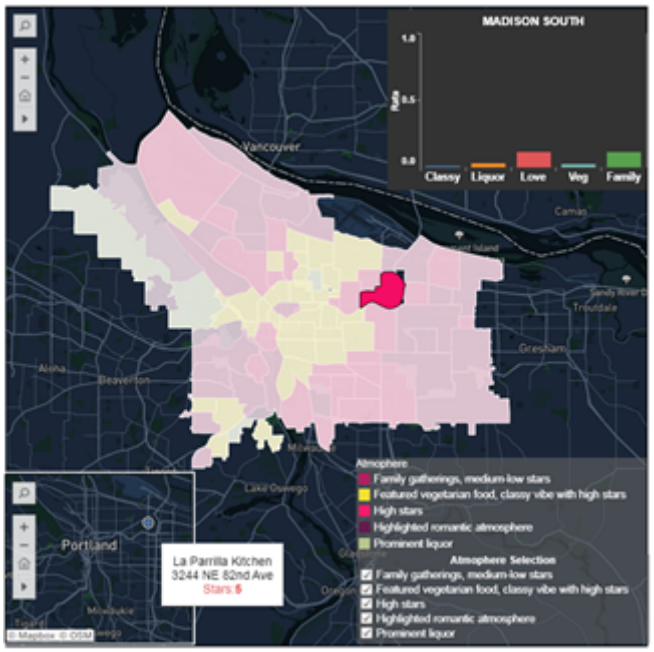

Figure 17. Detailed Information View of Restaurant Recommendation

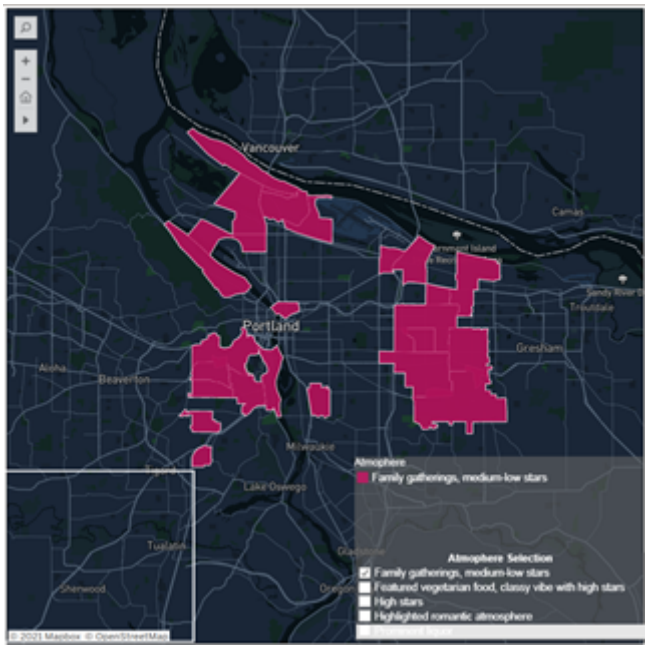

Figure 18. Filtered View of Restaurant Atmosphere

The other major user group of the website is the government, especially data analysts in charge of tourism and catering. We hope that our data exploration of the restaurant atmosphere can help the Portland government optimize the distribution of local restaurants. Therefore, for experienced data analysts, we expect them to be inspired by our analysis process, so we have inserted many analysis charts and provided some suggestions based on the analysis results. In the visual map, the government can see the distribution characteristics of the restaurant atmosphere through the distribution of different colour blocks. For example, restaurants in the central area of Portland are highly rated by the vegetarian group (Figure 19). Taking these factors into account, the functionality of interactive maps will depend on the needs of users. This makes this map useful in providing recommendations or data exploration.

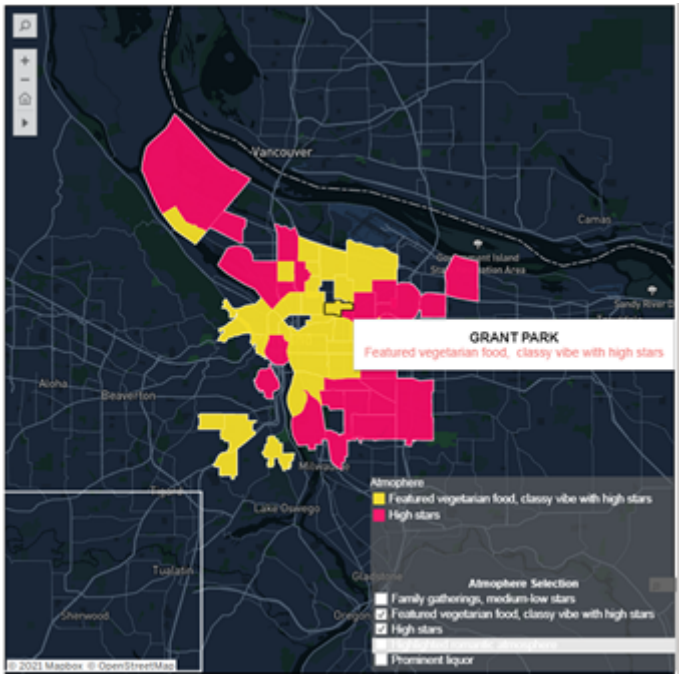

Figure 19. Cluster Analysis View of Restaurant Atmosphere

7 Discussion

7.1 Further Suggestions

7.1.1 Suggestions for Tourists

The analysis results can give visitors a general reference of the city restaurant atmosphere. For those neighbourhoods with romantic themes, there are attractions and commercial facilities that are suitable for lovers to play with. For example, Forest Park, South Burlingame and Sylvan-Highlands are all tourist attractions or near tourist attractions. So, a couple of travellers could give priority to these areas. And this project's website also recommends the best restaurants for each neighbourhood, which can also be a reference.

7.1.2 Suggestions for Restaurants

For restaurant operators, the results would help to redefine the atmosphere more suitable for the specific neighbourhoods and improve the service quality of the restaurant. For example, the northeast area of Downtown has good commercial and public facilities, such as schools, fashion culture centres, convention centres. In our result, this area is a cluster of family-low stars. Thus, for restaurants in this area, more consideration could be given to setting the family atmosphere. Meanwhile, they could focus more on the quality of dishes and service, which may be distinguished from other restaurants and obtain more customers.

7.1.3 Suggestions for City Government

Cluster analysis will also be beneficial to urban development planning and management. For planning commercial facilities, if the government and companies want to build a theme park about romance, the neighbourhood with high romance could be firstly considered. Because the proportions of couples in the consumer groups are relatively higher in these neighbourhoods, which could be more likely to promote customers' consumption of the theme park. For urban management, one of the examples is that Forest Park, Alameda and Markham are the only three neighbourhoods with high liquor. Because of the danger of alcohol, there might be more safety and noise issues during the night. Thus, it would be better to increase police patrols at night and improve safety supervision in these areas.

7.2 Limitations of the project

7.2.1 Limitations of Topic Selection

Although the unsupervised LDA model was used, the direct classification results were not ideal. Because most topics still focus on common words, such as service and food. In essence, we mainly chose the topics that we think are meaningful from the high-frequency word list and LDA models manually. In addition, to highlight the spatial distribution and find different customer groups, Pearson correlation was used to delete one of the highly relevant topics. Highly relevant topics were not

merged because they might not fit together. However, this did make some topics disappear, which limits the number of topics. In further research, more methods are needed to improve the classification, such as the Supervised LDA model, Structural Topic Model (STM).

7.2.2 Limitations of Spatial Analysis

Table 5. Global Moran'I of Each Theme

Theme	Moran's I	P-value
vegan	0.1636	0.0060
stars	0.1574	0.0090
classy	0.1305	0.0220
liquor	0.1237	0.0360
family	0.0679	0.1020
romance	0.0587	0.1600
number of reviews	0.0035	0.3880

In cluster analysis, hierarchical clustering was used without considering the connectivity matrix. Some researchers employed spatial constraint to achieve spatial connectedness (Pawitan and Huang, 2003). The dissimilarities of this method can be non-Euclidean distances while the weights of the observations can be non-uniform(Chavent et al., 2018).

The table above shows the results of the global Moran's I for each theme. Most topics' P-value was not significant if the significance level α was 0.01. Meanwhile, there was significant spatial autocorrelation with themes vegan (0.164) and stars (0.157). It could try to impose spatial constraints on vegan and stars for further cluster analysis. Although it might generate poor outcomes due to dissimilar units being gathered primarily (Anselin, 2020). In addition, a more urban comparison may be needed, which will also be the key point of further spatial research.

7.2.3 Limitations of Web Page

For map visualization, the map views opened on different devices may not match. This is because the map format of the inserted page cannot exactly match the screens of different devices. Besides, The interactive design and function of the interactive map still need to be improved. For example, if the colour of each cluster could be embedded in the button, there may be no need to display the legend repeatedly. The web's user experience can be improved for tourists. The interactive map can only give the user one restaurant in the selected area, which will limit the user's choice. It still needs to collect more user feedback to improve the interactivity of the web page.

8 Conclusion

This project identified different themes for Portland's restaurants and examined the differences in the spatial distribution of restaurants under different themes. Recommendations were also provided to visitors, restaurants and the government based on the analysis results. In addition, our website provides an interactive platform to help users participate in the analysis and help them filter the most

attractive atmospheres and neighbourhoods. If we have more time to improve our project in the future, we will focus on improving the analysis and the website in the limitation part and enhancing the user experience.

9 Reference

Agoes, A. & Pasaribu, N. F., 2015. Kajian Pustaka Mengenai Restaurant Atmosphere. *Binus business review*, 6(1), pp.110–116.

Ariyasriwatana, Weranuj & Quiroga, Luz Marina, 2016. A thousand ways to say 'Delicious!'—Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp. *Appetite*, 104, pp.18–32.

Balakrishnan, V. and Lloyd-Yemoh, E. 2014. 'Stemming and lemmatization: A comparison of retrieval performances'. in. *Proceedings of SCEI Seoul Conferences*, Seoul, Korea, pp. 174–179.

Bin Lu et al., 2011. Multi-aspect Sentiment Analysis with Topic Models. *2011 IEEE 11th International Conference on Data Mining Workshops*, pp.81–88.

Blei, D M. 2012. Probabilistic topic models, *Communications of the ACM*, 55(4), pp.7784.

City of Portland, 2021. *May 2021 – Bureau of Planning and Sustainability (BPS) Updates*, Available at: https://www.portland.gov/bps/news/2021/5/4/may-2021-bureau-planning-and-sustainability-bps-updates (Accessed: 15 May 2021).

Chavent, M. et al. 2018. ClustGeo: an R package for hierarchical clustering with spatial constraints, *Computational Statistics,* 33(4), pp. 1799–1822.

Cesario E., Vinci A., Zhu X. 2020. Hierarchical Clustering of Spatial Urban Data. In: Sergeyev Y., Kvasov D. (eds) Numerical Computations: Theory and Algorithms. NUMTA 2019. *Lecture Notes in Computer Science*, 11973.

Dubes, R., Jain, A.K., 1976. Clustering techniques: the user's dilemma. *Pattern Recognit*. 8, p.247–260.

Fouedjio, F. 2016. A hierarchical clustering method for multivariate geostatistical data, *Spatial Statistics*, 18, p. 333–351.

Govender, P. and Sivakumar, V, 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), *Atmospheric Pollution Research*, 11(1), p. 40–56.

Haddad, Mark et al., 2015. The development and psychometric properties of a measure of clinicians' attitudes to depression: the revised Depression Attitude Questionnaire (R-DAQ). *BMC psychiatry*, 15(1), p.7.

Heidenreich, H. 2018. *Stemming? Lemmatization? What? Medium.* Available at: https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8 (Accessed: 18 May 2021).

Heung, V. C. S. and GU, T. 2012. Influence of restaurant atmospherics on patron satisfaction and behavioral intentions. *International Journal of Hospitality Management*, 31(4), pp.1167-1177.

Huang, J., Rogers, S. and Joo, E. 2014. *Improving Restaurants by Extracting Subtopics from Yelp Reviews*. Available at: https://www.ideals.illinois.edu/handle/2142/48832 (Accessed: 18 May 2021).

Philip Wilkinson, n. d. *Introduction to hierarchical clustering (Part 3 — Spatial clustering)*. Available at: https://towardsdatascience.com/introduction-to-hierarchical-clustering-part-3-spatial-clustering-1f8cbd451173 (Accessed: 15 May 2021).

Jia, S., Zhang, X., Wang, X. and Liu, Y. 2018. Fake reviews detection based on LDA. in *2018 4th International Conference on Information Management (ICIM)*. *2018 4th International Conference on Information Management (ICIM)*, pp. 280–283.

Kang, Hanhoon, Yoo, Seong Joon & Han, Dongil, 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert systems with applications*, 39(5), pp.6000–6010.

Korenius, T., Laurikkala, J., Järvelin, K. and Juhola, M. 2004. Stemming and lemmatization in the clustering of

finnish text documents. in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. New York, NY, USA: Association for Computing Machinery (CIKM '04), pp. 625–633.

Lim, Young-shin & Van Der Heide, Brandon, 2015. Evaluating the Wisdom of Strangers: The Perceived Credibility of Online Consumer Reviews on Yelp. *Journal of computer-mediated communication*, 20(1), pp.67–82.

Liu, Y. 2009. The effects of dining atmospherics: an extended Mehrabian–Russell model. *International Journal of Hospitality Management*, 28 (4), pp. 494-503.

Tom Sietsema, 2015. The search for America's best food cities: New York, *The Washington Post (Online)*, 2015–01-01.

Mukherjee, A., & Liu, B., 2012. Aspect extraction through semi-supervised modeling. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*,1, pp.339-348.

Nakayama, Makoto & Wan, Yun, 2019. The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. *Information & management*, 56(2), pp.271–279.

Open data, n.d. *Neighborhood Boundaries*. Available at:https://gis-pdx.opendata.arcgis.com/datasets/1ef75e34b8504ab9b14bef0c26cade2c_3/data?geometry=-122.965%2C45.478%2C-122.223%2C45.646 (Accessed: 24 April 2021).

Özdağoğlu, Güzin, Kapucugil-İkiz, Aysun & Çelik, Ayhan Fuat, 2018. Topic modelling-based decision framework for analysing digital voice of the customer. *Total quality management & business excellence*, 29(13-14), pp.1545–1562.

Pawitan, Y. and Huang, J., 2003. Constrained clustering of irregularly sampled spatial data', J*ournal of Statistical Computation and Simulation*, 73(12), pp. 853–865.

Portland Neighborhood, n.d.. *Portland Neighborhood Guide.* Available at: https://portlandneighborhood.com/portland-neighborhoods (Accessed: 15 May 2021).

Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python, *JMLR 12*, pp. 2825-2830.

Park, Eunhye, Chae, Bongsug & Kwon, Junehee, 2018. The structural topic model for online review analysis. *Journal of hospitality and tourism technology*, 11(1), pp.1–17.

Park, E., 2020. The Effects of Green Restaurant Attributes on Customer Satisfaction Using the Structural Topic Model on Online Customer Reviews. *Sustainability (Basel, Switzerland)*, 12(7), p.2843.

Rahimi, S., Andris, C. and Liu, X., 2017. Using yelp to find romance in the city: A case of restaurants in four cities, Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, *UrbanGIS* 2017-January.

Geo data, n.d. *Spatial Clustering*. Available at: https://geodacenter.github.io/workbook/9c_spatial3/lab9c.html (Accessed: 15 May 2021).

Saraçli, S., Doğan, N. and Doğan, I. 2013. 'Comparison of hierarchical cluster analysis methods by cophenetic correlation', *Journal of Inequalities and Applications*, 2013, pp. 1–8. doi: 10.1186/1029-242X-2013-203.

Schrock, G., Green, J., and Liu, J. 2015. Portland's Food Economy: T conomy: Trends and Contributions ends and Contributions, *Institute for Sustainable Solutions Publications and Presentations*, pp.88.

Sulek, J.M. and Hensley, R.L. 2004. The relative importance of food, *atmosphere, and fairness of wait: the case of a full-service restaurant. Cornell Hotel & Restaurant Administration Quarterly*, 45(3).

Travel Portland. 2021. *Explore diverse shopping, dining and cultural offerings in downtown Portland.*,

Available at: https://www.travelportland.com/neighborhoods/downtown/ (Accessed: 15 May 2021).

Tausczik, Yla R & Pennebaker, James W, 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1), pp.24–54.

Tian, Guang, Lu, Liang & McIntosh, Christopher, 2021. What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data. *Food quality and preference*, 88, p.104060.

USCB, 2019. Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More, *Population Division*. Retrieved May 15, 2021

Vallurupalli, Vamsi & Bose, Indranil, 2020. Exploring thematic composition of online reviews: A topic modeling approach. *Electronic markets*, 30(4), pp.791.

Wang, Zhuo, Gu, Songmin & Xu, Xiaowei, 2018. GSLDA: LDA-based group spamming detection in product reviews. *Applied intelligence (Dordrecht, Netherlands)*, 48(9), pp.3094–3107.

Wang, S. et al. (2020) 'Spatial disparity and hierarchical cluster analysis of final energy consumption in China', *Energy*, 197, pp. 117195.

Willett, P. 2006. The Porter stemming algorithm: then and now. *Program*. Emerald Group Publishing Limited, 40 (3), pp. 219–223.

Yelp, n.d. *Yelp Open Dataset*. Available at:https://www.yelp.com/dataset (Accessed: 24 April 2021).

Yu, W. 2019. A mathematical morphology based method for hierarchical clustering analysis of spatial points on street networks, *Applied Soft Computing Journal*, 85, p. 105785.

Zhan, Jiaming, Loh, Han Tong & Liu, Ying, 2009. Gather customer concerns from online product reviews – A text summarization approach. *Expert systems with applications*, 36(2), pp.2107–2115.

Zhang, Ziqiong et al., 2011. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert systems with applications*, 38(6), pp.7674–7682.