Information Visualization 2019-2020 – G29: Yelp Restaurant Reviews

Henrik Eriksson
Student (91961)
IST Lisboa
henrikseriksson@gmail.com

Joona Hietala
Student (95430)
IST Lisboa
joonah96@gmail.com

Teresa Steinebach
Student (95412)
IST Lisboa
teresa.steinebach@gmail.com

ABSTRACT
Nowadays, the generation of data per day is extraordinary, but human beings are not able to absorb and understand these huge volumes of data. Information visualization addresses the need for data processing and representation in a visual and meaningful way so that users can better understand and get insights from data. This project deals with restaurant review data by first exploring the problem domain and defining relevant user questions for which the visualization aims to deliver answers. Data preparation and processing serve as the base for the creation and design of interactive visualizations. The final idioms provide an overview on restaurant ratings as well as on behavior patterns of their clients. In addition, they reveal insights in restaurants across different cuisines and areas as well as across a time range of ten years which are relevant for restaurant owners and consumers.

Author Keywords
Data; Visualization; Business; User; Questions

INTRODUCTION
With the development of the second stage of the Internet, the web 2.0, the foundation for the triumph of user-generated content and a participatory culture was laid. Networking sites like social media, blogs and wikis evolved which enabled users to generate and share content, interact and collaborate with each other. The implementation of features, allowing them to evaluate content by rating or adding comments, led to a new way of opinion formation based on peer-to-peer reviews. This has a significant impact on the decision-making behavior of consumers nowadays. Instead of only listening to mouth-to-mouth recommendations, reviews from the online community provide a new source of information and evaluation. Feedback with similar experiences are easier to find and available in an enormous amount online. Platforms like Google, Amazon, TripAdvisor and Yelp offer unlimited possibilities to share purchasing and product experiences as well as to access created reviews.

User-generated reviews do not only provide useful insights to other users but also to the business owners about their products and services which are evaluated. Analyzing the decision-making and rating behavior of consumers may reveal trends and hidden potential for improvements. Both perspectives are of interest and a closer look into review data may be worth it.

For our visualization project, we selected a dataset by Yelp which is an international enterprise that publishes crowd-sourced online ratings and reviews about different kinds of businesses and services. Yelp was founded in the US in 2004 and grew continuously over the last 15 years reaching 141 million monthly unique visitors and 148 million reviews. Their review system is based on a 5-star rating, scaling from 1 to 5 with 5 as the best. Stars are given per review; the mean of all star ratings per business represents the overall rating of the business. Due to the huge size of the Yelp community, their ratings have an impact on users' decision-making behavior and therefore also on the trends and the development of rated businesses. Specifically, we decided to explore the domain of food and restaurant reviews, because it personally concerns us as customers of restaurants and lovers of food. Since our generation grew up with the peer-to-peer review culture and relies on the provided information, we are also users and interested to get more insights.

We defined questions which are – in our opinion – valuable to raise for business owners as well as users.

1. Does the average rating by users of businesses per cuisine differ by area (Q1)?
2. Do some states receive higher ratings compared to others (Q2)?
3. How are ratings distributed in certain areas (Q3)?
4. At what time of the year do people prefer to eat out/at restaurants (Q4)?
5. At what time of the year is it more popular to eat pizza (Q5)?
6. How does the number of times that people eat outside change during the days of the week (Q6)?

We try to answer these questions by looking at the ratings and reviews of restaurants in ten North-American states (two from Canada, eight from the US) which span over 39 different cuisines. With this, we intend to learn about the differences in user behavior and their preferences in terms of location, time and cuisines, in particular between states, cuisines and state/cuisine combinations as well as across a time range of ten years, specifically looking at the review behavior across the course of the year and on weekdays.

We aim to provide insights for users and business owners to learn on what time of the year and on which days of the week people tend to eat at restaurants and leave reviews more often

(Q4, Q6). This is useful for consumers to select times and days for eating out which are less busy or for businesses to focus on busy times and days to improve their services then. In addition, we provide information about the evaluation of cuisines which are more popular and rated highly vs. less popular and lower rated ones in general as well as at specific times (Q5) and in different locations / states (Q1, Q2). These information are supported by Q3 highlighting the distribution of ratings in order to avoid misleading overall ratings.

RELATED WORK

The Yelp dataset has been published on their homepage for a couple of years now. The Yelp 'dataset challenge', with the intention to promote research and analysis, is currently in its 13th round, ending by December 2019. The availability of the dataset has led to numerous student projects and scientific papers utilizing the data for visualizations but especially for data science projects. In particular, the dataset has been used many times to explore rating and review data for research about recommender systems, acknowledging the possible influence on decision-making.

Bejarano et al. [1] explored the dataset for building and validating their mathematical model to measure user influence in recommender systems based on review and social attributes. Their work is claimed to be relevant for but not limited to marketing purposes. Also, Lei et al. [4] emphasize with their work the influence of reviews on decision-making behavior of users, especially regarding product selection. They leverage the dataset in their paper to perform a sentiment analysis on review data considering user and product attributes in order to propose a rating prediction for improving recommender systems. Their experiment results are illustrated in line and bar charts. Yu et al. [10] focus also on improving recommender systems and describe the distribution of Yelp feedback in a histogram. All of the aforementioned as well as [3], [5], [11] aim to learn about decision-making and influencing factors for users and to improve business performance and user experience, both more in favor of the business than the user perspective.

All of the papers show the huge potential of the used dataset, however they focus on delivering value for businesses when utilizing data. In contrast to this, Danone et al. [2] acknowledge the need for a visual summary of reviews for users who are overwhelmed by the number of product and service reviews. In their paper, they use a much smaller dataset of San Francisco only, conduct a comparative text analysis of reviews and focus their questions and effectively their visualization on five categories: responsiveness/ service, food quality/reliability, design and appearance, price and finally satisfaction. [2] provide a visualized comparison in form of bar chart, radar chart and table between specific restaurants and their reviews. Our analysis is based on star ratings not text feedback and total number of reviews instead of specific restaurants, providing an overall view on different cuisines and areas.

Mundada [6] explores the dataset based on the food business in Charlotte, NC, US only. Therefore his visualized work contains more details which are specific for this location. He creates a model by analyzing the sentiment of review texts to answer questions regarding the success or failure of businesses. Among others he illustrates the "Count plot for Open Restaurants" (bar chart), "Distribution of Star Ratings vs the status of the business" (violin plot), "Review Count by Neighborhood" (stacked bar chart), "Star rating distribution by neighborhood" (boxplot), "Neighborhood Popularity defined by the attributes" (heatmap) and more. The analysis and visualizations of Mundada [6] are extensive and contain more than 14 idioms covering multiple aspects of the initial Yelp dataset. Therefore they may serve as inspiration for further analysis.

In addition, Yu [9] focuses on the question what aspects are important to be a popular business on Yelp. She uses linear regression techniques to explore the dataset (restaurants in Las Vegas only) and focuses on similar but not the same attributes. For the visualization of results, she uses a word cloud to show the frequency of food categories / cuisines, bar chart for the several ratios (number of reviews, food over ambience photos) and a scatterplot for the comparison of actual vs. predicted reviews.

Another project [7] focused on features of restaurants and their influence on ratings. They summarize the data by visualizations: histograms, boxplots and smoothing (loess). Afterwards machine learning techniques are used to predict ratings based on restaurant features and demographic characteristics. For this, they combined the Yelp dataset with additional data like weather and demographic data.

Unlike the above mentioned sources, the student project [12] reports some interactivity between the idioms. A filter function is implemented in the treemap as well as in the price and rating scroll bar. Histogram and map adjust accordingly. Here, data is reduced to restaurants in Montreal only, with detailed information about each.

All of the aforementioned sources indicate the popularity of the Yelp dataset and the manifold opportunities to work with in order to learn more about the influence and meaning of review data for the decision-making behavior of consumers. Rather projects than scientific papers served as an inspiration for our visualizations. However, they still have a different approach and raise different questions than we do. With our questions, we focus on an overall point of view on user behavior across states, cuisines and time range.

THE DATA

The data, which has been used for our visualization, origins from a dataset that has been published by Yelp on their homepage as well as on general sources like www.kaggle.com. It has been made accessible without any barriers for public usage. No challenges have been faced during the process of accessing the dataset.

The dataset has been provided in multiple CSV-files with each of them containing detailed data about either restaurant businesses, reviews, users, checkin, tips (short version of a review) or photos. The size of the initial dataset was five gigabytes consisting of 6,6 million reviews of 192 thousand restaurant businesses in eleven metropolitan areas. There was no need to add additional sources of data due to the size of the initial dataset. At the same time, the size required a cleanup and a combination of multiple files. This has been done by utilizing Pandas and NumPy to reduce the dataset to a subset of CSV-files of only restaurant businesses, their reviews and user data with a total of 44 attributes. Only attributes were selected which were identified to help answering the questions we raised in the beginning, when we explored the domain for the first time. The initial dataset also contained reviews of other businesses than restaurants; all of them have been removed in order to focus on the domain of food. In addition, we only included areas of businesses which are representative and geographically comparable, resulting in North-American states only (eight US states and two Canadian states). Random areas in Europe as well as non-representative American states (due to low number of reviews) have been removed. Food reviews were also given about categories which do not represent cuisines, like "nightlife" or "catering". Those have been excluded as well. We ended up with a reduced dataset containing a total of 2,14 million reviews and their ratings of 36,4 thousand restaurant businesses with 39 different cuisines in ten states.

Due to the remaining size of the dataset, we faced the challenge of performance and scalability issues during data processing as well as for the visualization. Regarding the latter, it was not possible to use the final dataset as it was and to rely on calculations of required values made by D3 only. Therefore, we decided to calculate the values needed for the visualizations upfront and store them in separate data files. The values in those smaller-sized files form the base of the visualization and ensure a smooth interactivity without loading times. The amount of time required per precedent calculation ranged from ten minutes up to five hours.

After checkpoint IV and the feedback of the first prototype, we decided to revise our data in order to add more details and provide more insights into our visualization. This resulted in new calculations which were in general the same but with smaller batches based on weekdays:

- Density plot: New averages for restaurant ratings were calculated to show detailed distributions on a continuous scale instead of previously discrete values which were rounded to 0.5.
- Heatmap: Based on the data for the density plot for restaurants and reviews, here we calculated the overall mean values with smaller batches based on weekdays.
- Radar chart: The data for this chart remained the same, because we had already counted reviews for weekdays. Specifically, we used the dates when reviews were created, defined weekdays for each review to make data available per state/cuisine.
- Line chart: Here, previous calculation were too limited and only showed the amount of ratings per day across the year. A visualization per weekday was not possible. Therefore we used a similar approach as for the radar chart: counting reviews for days of the year with the dates when reviews were made, then determined weekdays for them, divided days of a year in weeks, summed up all reviews per weekday per week of the year for all years included in the dataset (e.g. all Mondays of the first week of the year), for all weekdays across the total period of a year.

The result of the new calculations provide more insights in the date by adding weekdays.

VISUALIZATION

Overall Description

Our visualization consists of a navbar, placed on the top, and four idioms aligned below (figure 1).

The navbar contains the overall title of the project "Yelp Restaurant Reviews", a button saying 'Total" which initiates the display of total values in all idioms, and next to it the legend. The legend gives information about weekdays (all selected or one specific weekday) and offers to deselect one weekday in order to show all again. The button for deselection only appears right next to the label "Weekday" if one specific weekday has been selected. Further right, the color legend is placed. It labels the selected values which are visualized in the idioms by either green or orange color.

The body of the visualization consists of four idioms: heatmap, density plot, line and radar chart. In particular, the heatmap and radar chart offer a filter functionality in order to control the displayed values.

The central idiom is the heatmap which is located on the left-hand side and illustrates the average restaurant ratings per cuisine and state/area. The y-axis shows the 39 different cuisines, the ten states are shown on the x-axis with their official abbreviations (e.g. NV for Nevada, US or ON for Ontario, Canada). Average ratings are visualized per rectangle. Each of them represents a value for a state/cuisine combination. The more intense the saturation of the rectangle is, the higher the average rating. No values are available for white spots, because no reviews have been made for this combination of state and cuisine (e.g. for Thai food in South Carolina). Hovering over a rectangle shows a label with the exact value of the average restaurant rating. Furthermore, the heatmap serves as the main control panel for the other three idioms. Selections made in the heatmap change the values in the other idioms accordingly. Up to two selections can be made by clicking on either any axis label (x, y) or any rectangle or a combination of both. Latter ones, if selected, are highlighted with a colored border (green or orange). Selected axis label appear bold and in the respective color.

Figure 1: Overall Visualization for Yelp Restaurant Reviews

The applied color in the heatmap also represents the values in the other three idioms then and the color legend in the navbar provides additional orientation for the selection made.

The density plot, the second idiom, is located on the top right-hand side of the heatmap showing the distribution of the ratings on a scale from 1 to 5. It displays the values according to the two selections made in the heatmap and color coding is aligned as well. Hovering over each graph shows the particular selection, e.g. AZ, Mexican for Mexican cuisine in Arizona.

The third idiom is the line chart, positioned on the bottom right-hand side. It visualizes the total number of reviews across the year, with summed up values of a time range of ten years, with a filled line in orange or green. The x-axis represents the timeline in days of the year, the y-axis shows the number of reviews. The scale adjusts to the two selections. By brushing over the area of the line chart, zooming in for more details is enabled. The initial situation can be reached again by double-clicking on the chart.

The fourth idiom is the radar chart, located on the top right corner. It visualizes the number of reviews per weekday, which allows a comparison between different weekdays, based on the heatmap selection. Each weekday represents a variable which is provided with an axis that starts from the center. The seven axes are arranged radially with equal distances between each other. The scale of the axes adjust coherently according to the selected values. Grid lines from axis to axis support better readability. The values, corresponding to the selection in the heatmap, are plotted along the axes and connected together to form a polygon. Each of the two selections are visualized by a polygon. Hovering over the area of each polygon highlights the specific area. Hovering over the dots show the number of review for the specific weekday, where the dot belongs to.

Additional details have been added to the entire visualization by implementing the selection of weekdays. Therefore the radar chart allows the user to click on each label of the weekdays. The selected label will be highlighted in bold font. The legend in the navbar shows the weekday that was chosen. Next to it, a little button appears with a cross icon in order to deselect the weekday and to return back to values for all weekdays. The selection of one specific weekday in the radar chart initiates updates in the other three idioms: The heatmap represents then the values of the average ratings for the selected state/cuisine combination based on that particular weekday only; the density plot shows the distribution of ratings for the state/cuisine combination on that weekday only; the line chart visualizes the number of

reviews across the year for the selected state/cuisine combination on this weekday only. For the latter, the x-scale changes from days to weeks across the year, because each weekday occurs only once per week.

If the user aims to compare one selected state/cuisine combination to the total amount of reviews, the button 'Total' in the navbar needs to be clicked once. In case only total values are from interest, the user has to click twice on the button.

Rationale

We first listed out several potential visualization techniques for our data. We brainstormed which ones would make the most sense and could encode what we wanted to communicate with the data. In the data preparation phase we also tried out some basic versions of the idioms by making plots with python. This gave us some confidence that idioms we had chosen would work. We also frequently asked and got feedback from the faculty, which helped us determine final forms of each idiom.

In the heatmap we used color with different saturations in encoding the data. This is not the most understandable way to encode data, but we felt that it could communicate differences in average ratings fairly well and could pack a lot more information than other ways, while offering natural way of interacting with the dashboard for the user.

For the density plot and radar chart we used length to encode data, as it is one of the easiest to interpret by humans. This way was also an obvious choice for the data we wanted to show and questions we wanted to answer. "Height" of single point in both of these idioms encodes the number or frequency of observations having that value. In the case of density plot the distance from x-axis and in the case of radar chart distance from the center of the chart, both encode larger value, which is easy for humans to interpret.

In the line chart two different ways of encoding data are combined. Slope of the line allows the user to see the trend over time and between two points in time. Position of single time point on the other hand encodes value in each time point, higher distance from x-axis, corresponding to higher value in number of reviews. Both of these ways are very easy for humans to interpret, so they were obvious choices for encoding time-series data.

When we first started designing and choosing idioms, we wanted to include a choropleth map as one of the idioms, to communicate also the geographical aspects of our data in a visual way, not just state names. Soon we realized that it would not make much sense to use choropleth map as we had only two Canadian and eight US states with a significant amount of data and the map would end up looking very scarce.

Another idiom we ended up discarding was boxplot, that we eventually replaced with the density plot as it allowed us to communicate more granular details about the data than boxplot and would also allow comparison between combinations with different amount of observations.

When it comes to dealing with complexities of real data, that didn't produce issues that we weren't able to overcome. To use data in idioms we chose, it took a couple of iterations to get it into the right kind of shape, but wasn't that hard to implement, just required some additional work. Even though our initial dataset is quite big, by dividing it into smaller chunks before passing it to visualization components proved to produce wanted results in very short response time. D3 and JavaScript data filtering functions also proved to work fast enough for our data.

Demonstrate the Potential

The initial visualization shows total values in all four idioms in order to provide an overview of the data. This view allows to answer some of our questions raised in the beginning. If any selections have been made in the heatmap or radar chart, the user can return to the total values by clicking twice on the button 'Total' in the navbar.

With question one (Q1), we want to learn about differences by areas in average ratings of businesses per cuisine. The answer for this is provided by the heatmap (figure 2). The different saturation levels of the color in the rectangles illustrate that the average ratings per cuisine differs by areas. E.g. bakeries (4th row from the bottom) are higher rated in Quebec, CA and South Carolina, US (1st and last value on the x-axis from the left), than the other eight.

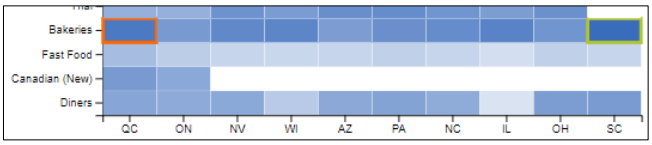

Figure 2: Heatmap - QC, Bakeries and SC, Bakeries

In Ontario, CA and Arizona, US, Bakeries received the lowest average rating. With this you can also learn, which cuisines are more popular in specific states: e.g. German food was rated on average with 5 stars in South Carolina, US, whereas Chicken Wings, Burgers and Fast Food received lows values (2.27-2.70) and twelve cuisines are not represented at all.

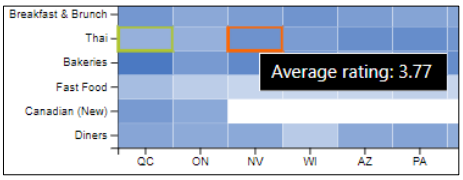

Figure 3: Heatmap - QC, Thai and NV, Thai

The overall view on the heatmap, without any selections, also addresses Q2 "Do some states receive higher ratings compared to others?" by following the rectangles and the corresponding values along the column of one state and others to be compared to: Consider the rectangles of Nevada, US and hover over them to see the values, then select another

state and hover of its rectangles. It is recommended to compare states in terms of the same cuisine (figure 3).

Q3 “How are ratings distributed in certain areas?” can be answered by the density plot. Therefore the user selects two values in the heatmap which should be compared: either the overall values of states by clicking on the state names respectively the labels on the x axis (like NV and AZ) or the specific ratings of a cuisine in a particular state (e.g. rectangle for Fast Food in NV). A combination of both is also possible, like looking at the distribution of NV overall and also at Fast Food in NV (figure 4 and 5)

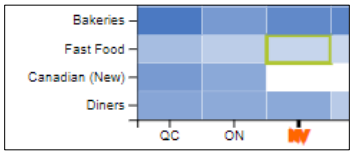

Figure 4: Heatmap - NV and NV, Fast Food

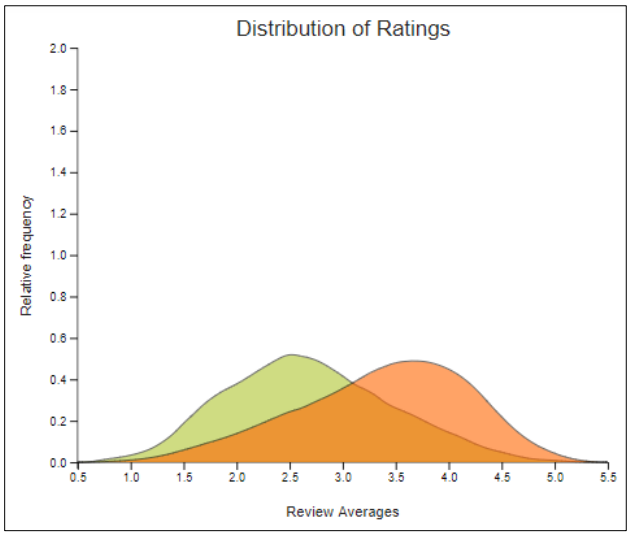

Figure 5: Density Plot - NV and NV, Fast Food

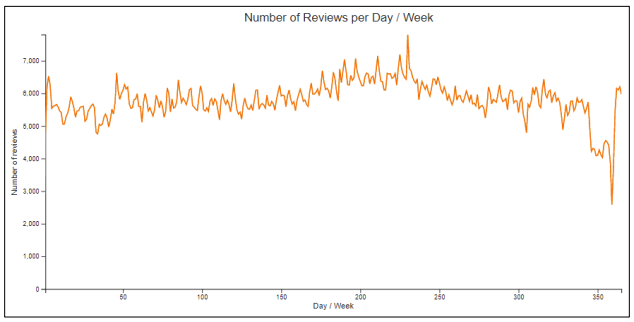

Figure 6: Line Chart - Total

The line chart helps answering Q4 and Q5. For Q4 “At what time of the year do people prefer to eat out/at restaurants?” it is necessary to look at total values without any selections (figure 6). Therefore the user should click the button ‘Total’ in the navbar twice in order to show total values only. Now, the line chart visualizes the total number of reviews given per day of the year. The run of the curve shows that the highest number of reviews per day are created during the middle of the year which reflects summer time.

In order to explore more details, the user can zoom by brushing the particular area, e.g. with the highest peak. By doing this, it is possible to see that the peak value happens at the end of August on day 230 of the year which is August 18 (figure 7).

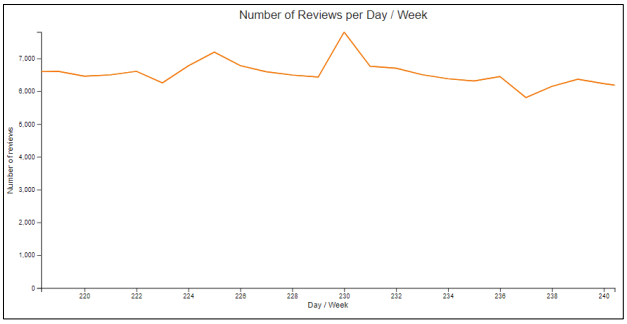

Figure 7: Line Chart - Brushing for Peal Value

In addition, it is observable that the number of reviews created at the end of the year drops to its lowest value of about 3.000 reviews compared to the peak day with more than 7.000 reviews. Zooming in reveals that the day hitting the bottom is day 359 which is December 25, Christmas day (figure 8). It is interesting to learn that people tend to eat out less often during Christmas time but much more often during the rest of the year, especially in summer.

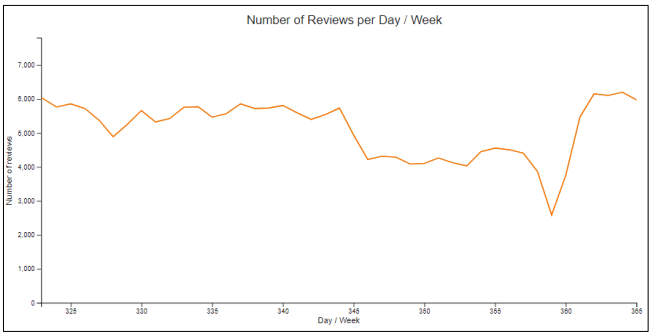

Figure 8: Line Chart - Brushing for Bottom Value

It is also possible with the line chart to learn more about the consumer behavior in terms of states and cuisines: “At what time of the year is it more popular to eat pizza?” (Q5). First, the user clicks on the label of the y-axis of the heatmap, which says ‘Pizza”. This selection causes an update of data in all idioms. Now total values (orange) and values for Pizza across all ten states (green) are visualized. In the line chart, the green line appears for Pizza which enables the comparison to the total (figure 9). It can be observed that Pizza is popular across the whole year without any major peaks or drops, except for one: Christmas again. This analysis can be done for any other cuisine, state or state/cuisine combination.

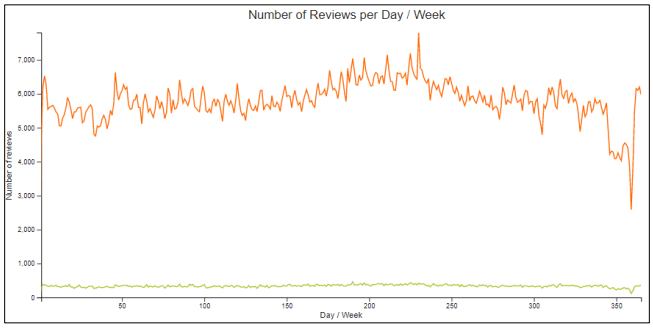

Figure 9: Line Chart - Total vs. Pizza

The last question (Q6) "How does the number of times that people eat outside change during the days of the week?" asks for more details on the time scale. It is interesting for business owners as well as restaurant clients to know which days are more popular and busier for eating out. Therefore we implemented the radar chart (figure 10). The user can either look at the total values or select more specific ones: E.g. selecting Fast Food in Ontario, CA (green) and Diners in Ontario, CA (orange) reveals that Sunday is the most popular day to eat out in Diners, followed by Monday and Tuesday, but overall all weekdays seem balanced; the picture looks different for Fast Food. Here, Sunday also reaches the highest number but the difference to the days during the week is much bigger and on Tuesday to Friday, it is less popular to eat out. Exact numbers of reviews can be seen by hovering over the dots of the variables (weekdays).

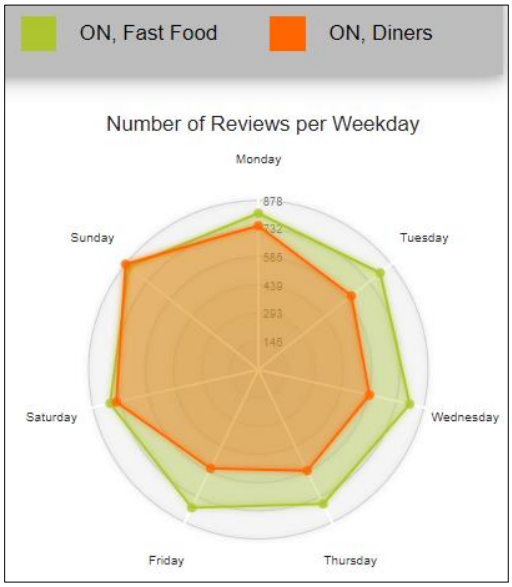

Figure 10: Radar Chart - ON, Fast Food and ON, Diners

In addition, we learned some specific insights as well: Canadian (New) food is only represented in the two Canadian states but not at all in the eight US states. It might be interesting for entrepreneurs to explore whether US states have potential for Canadian restaurants. Also, in South Carolina twelve out of 39 cuisines are not available as per Yelp data, maybe because citizens there are less open to try different food or they are and the potential for some unusual cuisines might be high but unexplored. Interestingly, Fast Food has been overall rated very poorly in all ten states. Except for Quebec (3.08), the values were between 2.54 and 2.82. This is surprising since the US are known for their Fast Food culture. Same applies to the average rating of Chicken Wings. In contrast, consumers seem to be more satisfied with restaurants serving German food, with five states rated over 4.00 stars, especially in South Carolina (5.00). Furthermore, the radar chart confirmed our expectation that Sunday is the most busiest day for eating at restaurants, but surprisingly in terms of Greek cuisine, Wednesday is super popular too.

IMPLEMENTATION DETAILS

We had a slow start with D3 as it was quite different from what we had done before. Every member of the group had at least some experience working with web technologies, mainly HTML, CSS, and one or several JavaScript frameworks.

The different idioms were implemented separately and after determining how to control the data in the different idioms we made functions for the different filtering methods, by cuisine/state and by weekday in order to control the data for all idioms. This let us keep the idioms separate with only connecting them functionally by a set of main functions. There had to be extra data handling functions for the different idioms. For example, if only a state was selected, the number of reviews had to be summed up over all cuisines for that state for the line chart and the radar chart.

We used D3's csv import function to import all data from all csv files in the same JavaScript promise to be sure to have everything ready from the beginning.

When integrating all four idioms, all code had to be made sure to be globally unique so there would be no conflicts or overwritten variables. Integration was somewhat difficult in the beginning as everyone had followed different tutorial while developing individual idioms. However we were able to overcome problems by coming up with one single way to update each idiom that is used while modifying data that each idiom displays.

The code for the radar chart was mainly taken from [12]. The main challenges were related to controlling the data and the colors. Some changes had to be made to be able to click on different weekdays and make sure the selected weekday stays highlighted even after the radar chart is updated with new data.

Code for heatmap [13], density plot [14] and line chart [15] are mainly taken from the same source of tutorials. Main challenges while integrating these were building the data filtering according to user selection. Also updating idioms and how those worked originally demanded tweaking, removing old svg elements for example. As mentioned before, special cases such as selecting total, only state or cuisine also needed handling and some custom summation functions needed to be built for the line chart and histogram.

Views are linked by an update function that controls the state of all idioms. Heatmap functions as a "control panel" for manipulating other idioms. By clicking a cell, state or cuisine in the heatmap, information is passed to update function that updates the state of the visualization according to user selection. Every controlled idiom is using essentially the same state (state + cuisine), so updating several idioms works the same way in terms of every idiom, data is filtered by state and cuisine information. State of the visualization is saved in a dictionary with key value pairs that are used to filter data in each idiom. In each idiom, these key-value pairs are compared to attributes of the data-array passed to the function and filtering of data is carried out accordingly.

CONCLUSION & FUTURE WORK

With this visualization project, we have learned several things, starting with theoretical knowledge about information visualization and techniques to process data with Pentaho as well as coding own visualization with the D3 library. In terms of our dataset, we explored the food / restaurant industry in North America and learned about the diversity of available cuisines, which of them are rated better across the states and which weekdays are the busiest. In addition, we came to know that people prefer to eat out during summer time and rather spend Christmas at home (or at least less time in restaurants).

If we were to start all over again, we would probably explore the content of reviews by conducting a text analysis. The results could be visualized in a word cloud, for example.

Another option to enrich our project would be to focus on trends by looking more into data over the years: Which cuisines have developed positively or negatively? Does it change over time when users eat out (days or time of the year). In addition option, we would consider features of restaurants as well as users to explore their influence on reviews and ratings.

All in all, the dataset offers unlimited opportunities for data visualization and analysis. As elaborated in the 2nd chapter "Related Work", this can range from simple visualizations to complex machine learning projects. Our intention was to explore the data and provide an overview which is useful for restaurant owners as well as their (potential) clients.

REFERENCES

[1] Bejarano, A., Jindal, A., & Bhargava, B. (2017). Measuring user's influence in the Yelp recommender system. PSU Research Review, 1(2), 91-104.

[2] Danone, Y., Kuflik, T., & Mokryn, O. (2018, March). Visualizing Reviews Summaries As a Tool for Restaurants Recommendation. In 23rd International Conference on Intelligent User Interfaces (pp. 607-616). ACM.

[3] Kim, H., & Arguello, J. (2017). Evaluation of features to predict the usefulness of online reviews. Proceedings of the Association for Information Science and Technology, 54(1), 213-221.

[4] Lei, X., Qian, X., & Zhao, G. (2016). Rating prediction based on social sentiment from textual reviews. IEEE Transactions on Multimedia, 18(9), 1910-1921.

[5] McAuley, J., & Leskovec, J. (2013, October). Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM conference on Recommender systems (pp. 165-172). ACM.

[6] Mundada, R. (2017),

https://raunakm90.github.io/posts/yelp-visualization/

[7] Shin, H., Yuan, Q., Wang, Y.,

https://sites.google.com/site/bio260final/overview

[8] Viry, M. (2017),

http://bl.ocks.org/mthh/7e17b680b35b83b49f1c22a3613bd89f

[9] Yu, P. (2019),

https://towardsdatascience.com/the-restaurant-guide-how-to-be-popular-on-yelp-a77591a13c8c

[10] Yu, X., Ren, X., Sun, Y., Sturt, B., Khandelwal, U., Gu, Q., ... & Han, J. (2013, October). Recommendation in heterogeneous information networks with implicit user feedback. In Proceedings of the 7th ACM conference on Recommender systems (pp. 347-350). ACM.

[11] Zhao, G., Qian, X., & Xie, X. (2016). User-service rating prediction by exploring social users' rating behaviors. IEEE transactions on multimedia, 18(3), 496-506.

[12] N.N. (2016),

https://yihaozhou.github.io/Data-Visualization-Final-Project/about.html

[13] https://www.d3-graph-gallery.com/heatmap

[14] https://www.d3-graph-gallery.com/density.html

[15] https://www.d3-graph-gallery.com/line.html