Yelp Dataset

---by Project Team 1

Yili Wang	470106260
Zhuoyang Li	480164337
Kwon Nung Choi	309252806
Raja Sameer Gagguturu	480254199
Venkata Sai Harsha Mantena	480242978

THE UNIVERSITY OF SYDNEY

CONTENT

THE UNIVERSITY OF SYDNEY

01

Introduction

1.1 Introduction

YELP Dataset

- Yelp shares global crowdsourcing user data on various categories like restaurants, dentists across cities (such as Phoenix, Madison and Edinburgh)
- Yelp dataset - subset of businesses, reviews and user data, locations.
  - Used - Personal, Educational and Academic purposes.
  - It is available as JSON files and CSV files
    - Make mobile apps, teach databases, to learn NLP
- Yelp Dataset version *6* has total of six csv files, adding up to a size of *5.01 GB*
  - Focuses on Year *2004 to 2017*

THE UNIVERSITY OF SYDNEY

02

Data Preprocessing

2.1 Problems & Preprocessing

YELP Dataset

- **Date was a merged string of year, month and day**
  - had to be divided into year, month, day to allow monthly or yearly analysis
- **Original category column had multiple synonym category terms for each business**
  - e.g. **Brick Tavern + Tap** is classified as:
    "American (New); Nightlife; Bars; Sandwiches; American (Traditional) Burgers; Restaurants"
  - Hence, a primary category for each business needed to be created
- **Some valuable information was in string-format**
  - Hence needed a way to convert the string to numeric format

2.1 Problems & Preprocessing -cont.

YELP Dataset

➔ **Defining Dictionary for Primary Category (PC)**
- Tool: Jupyter Notebook
- Dataset: *yelp_business.csv*
- Reference:
  the blog of “The Complete Yelp Category List”
  (https://www.yelpblog.com/2018/01/yelp_category_list#section9)

➔ **Clustering Data for Business into Specific PC**

➔ **Redefine the Data Type of Business Attribute**
- Assign *Numeric values* to *Strings (one hot encoding)*
- Tool: Jupyter Notebook
- Dataset: *yelp_attribute.csv*
- Reason: Need to further analyze the *correlation* between business attributes and review level

Food

- Acai Bowls
- Bagels
- Bakeries
- Beer, Wine & Spirits
- Beverage Store
- Breweries

THE UNIVERSITY OF SYDNEY

03

Descriptive Findings

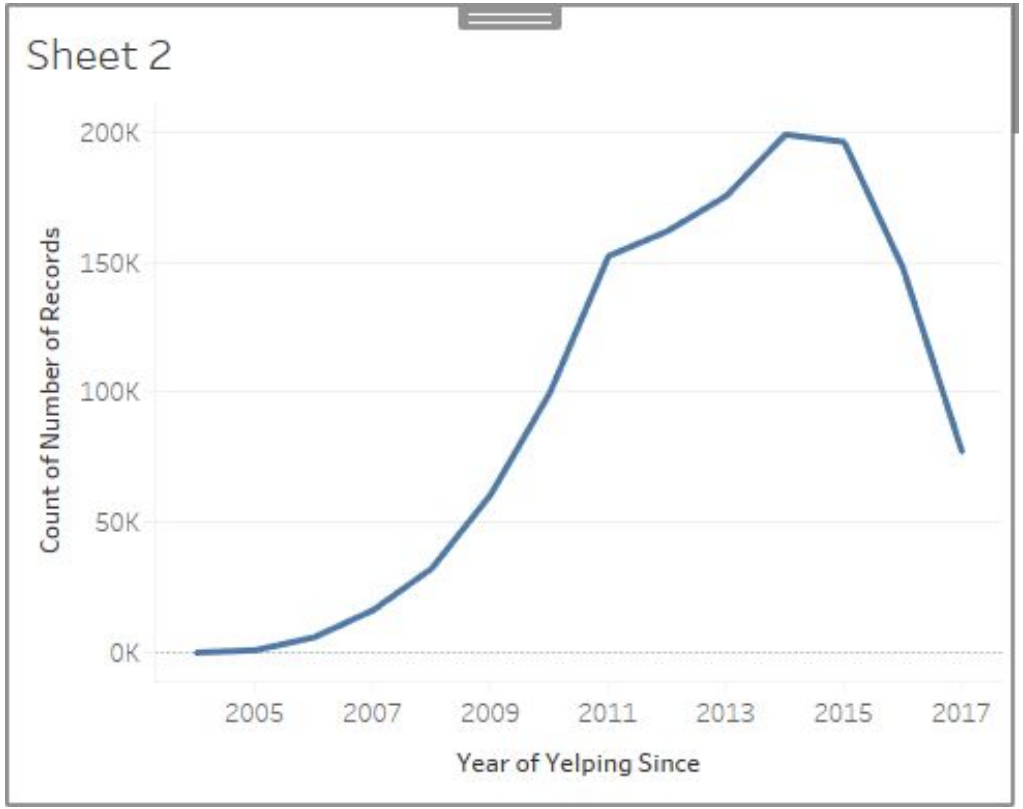
3.1 *Q1-* What was the trend in the popularity of Yelp over the years?

Founded in *2004*

Finding:
Number of user ***sign ups peaked*** in the year 2014(198,976)

Conclusion:
Later on due to the rising popularity of Google reviews and Amazon, yelp user base saw a steep downfall.

Currently ranked 2nd among review platforms preceded by google reviews and succeeded by Amazon

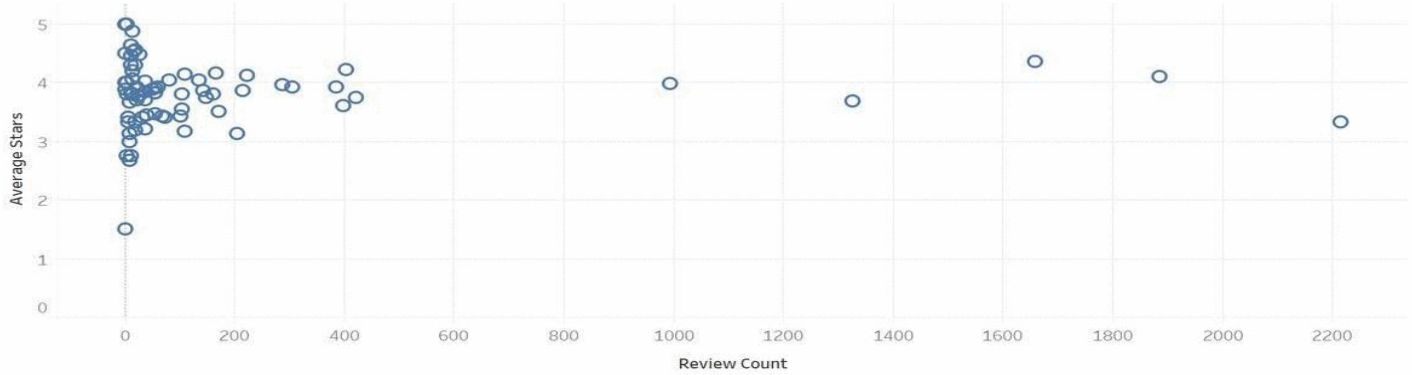


THE UNIVERSITY OF SYDNEY

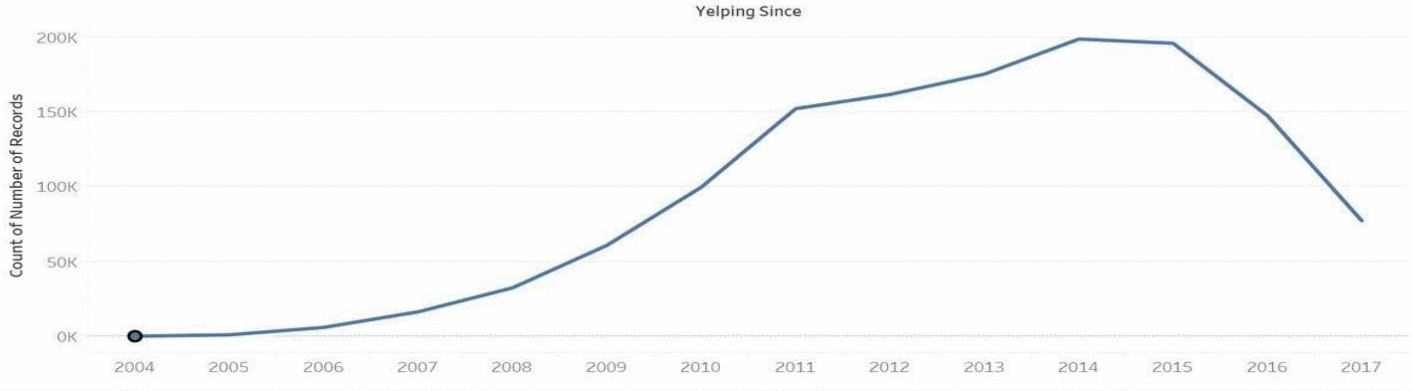
3.2 *Q2-* Does user rating pattern change over the years?

Finding & Conclusion

Sheet 1



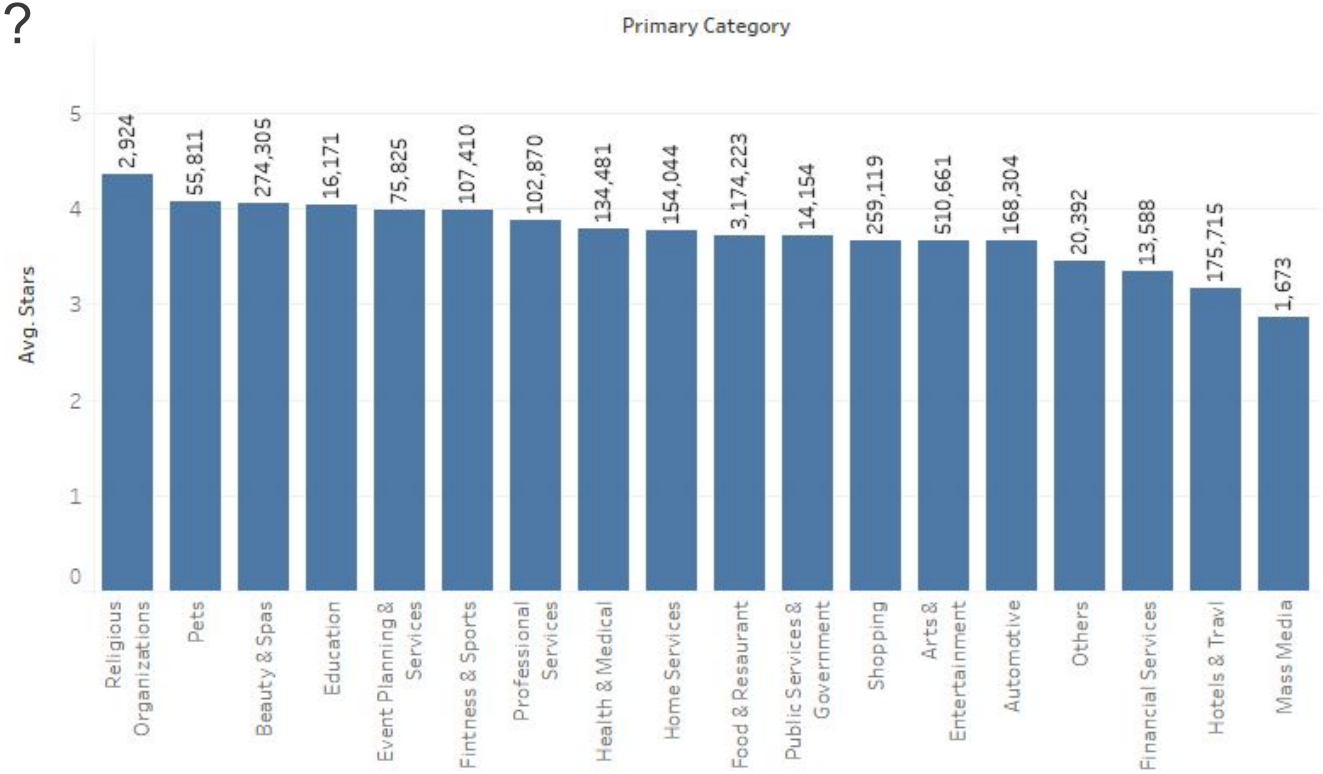
Sheet 2



THE UNIVERSITY OF SYDNEY

3.3 *Q3-* Which categories are users more likely to leave higher star ratings?

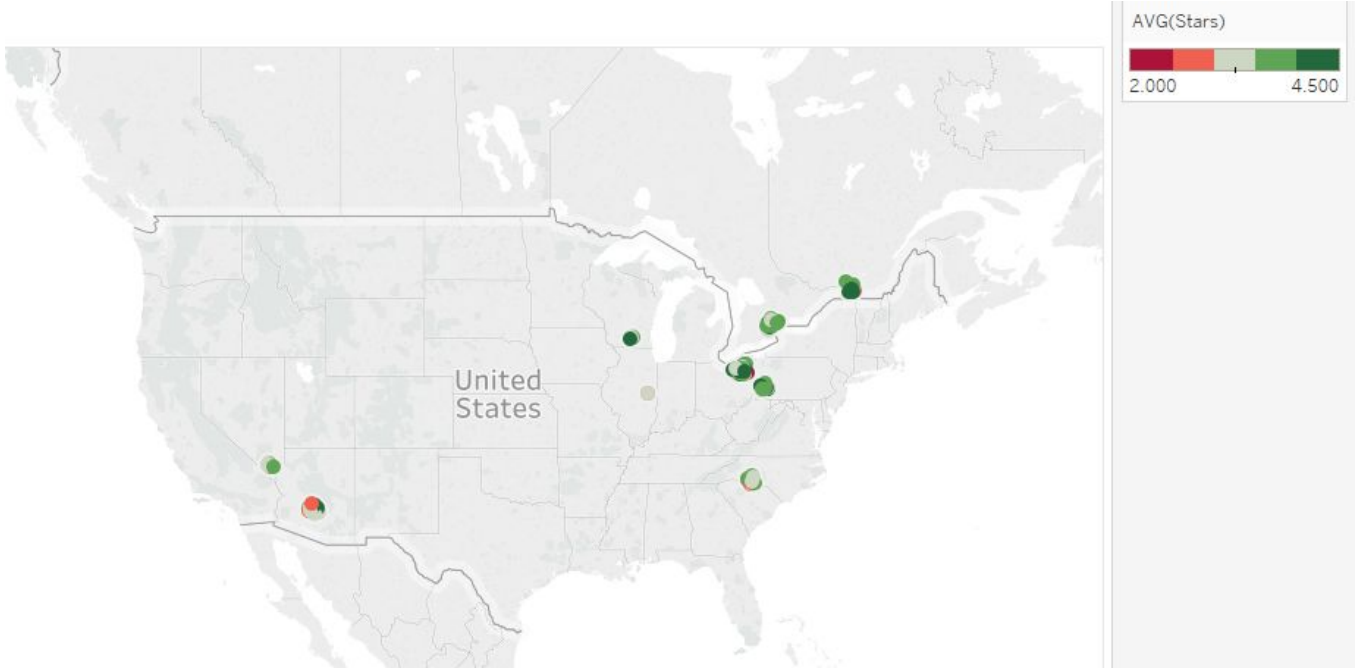
Finding & Conclusion:
People show more positive attitude towards religious org, pets, beauty & spas in comparison to mass media, hotels & travel



3.4 *Q4-* Compare various starbucks all over the US on the basis of reviews and find out which branches need improvement?

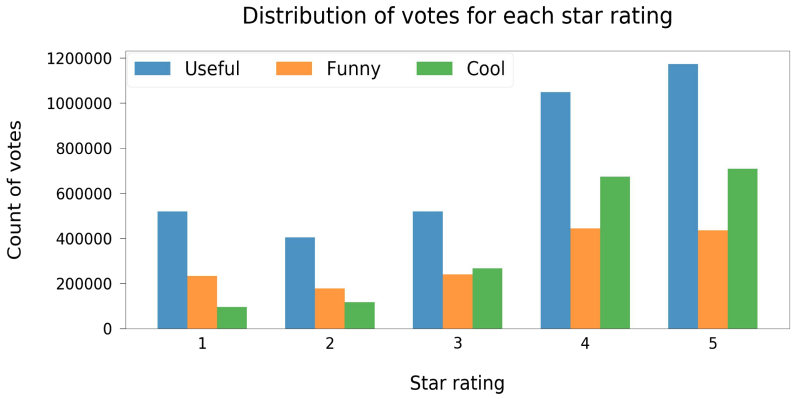
Finding:
The state of Arizona seems to be doing poorly as compared to other states in the US

Conclusion:
Further investigation is needed on why Starbucks in Arizona are not well received, as compared to other states



THE UNIVERSITY OF SYDNEY

3.5 *Q5-* For the comments for restaurant that have been upvoted (useful, funny, cool), is the distribution of the upvote different for each star rating?



Finding:

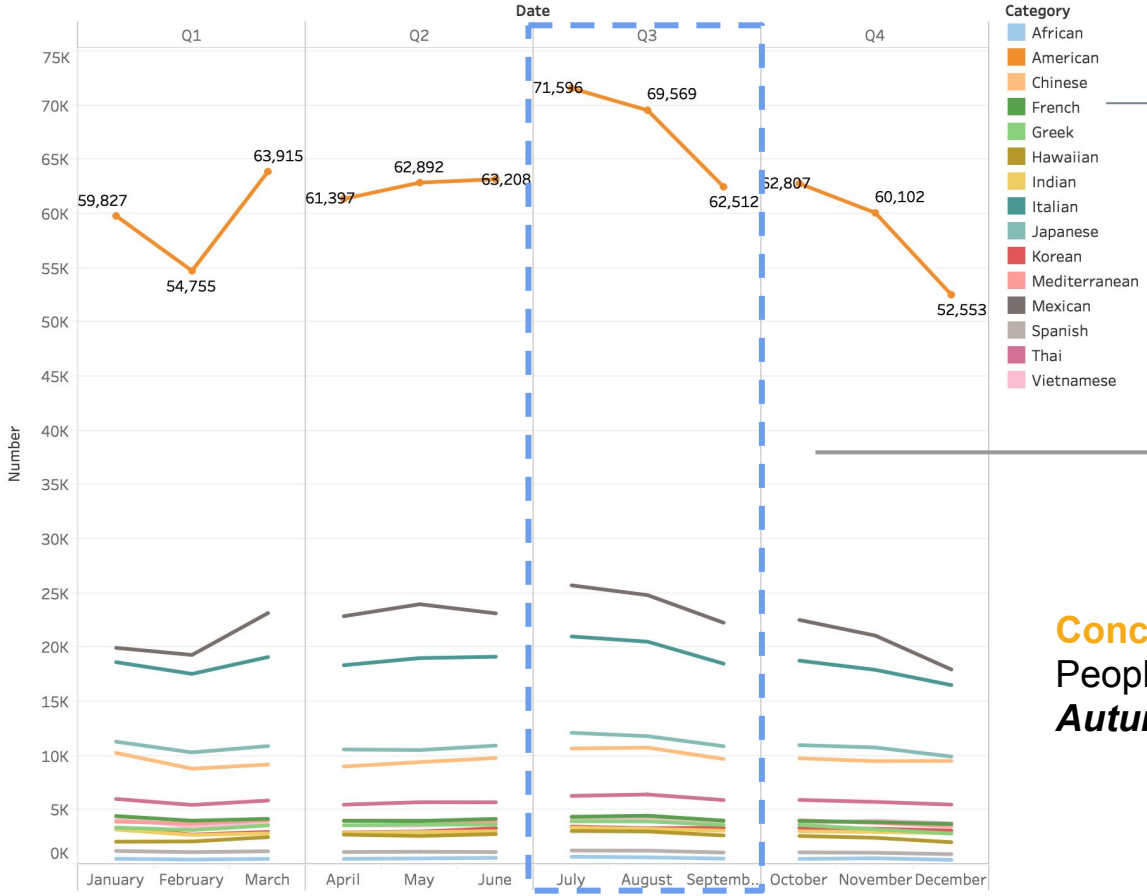
- Yes, distributions of 1 and 2 are different when you compare to distributions of 3,4,5 ratings.
- Increase of votes for cool with the increase of ratings.
- There are a lot of more 4 and 5 star ratings. We can say that people seem to review things they like.

Conclusion:

In general, people like to write an review for a ***positive experience*** than a ***negative one***.

THE UNIVERSITY OF SYDNEY

3.6 *Q6-* Peak Incidence of The Restaurant



Preprocess:

We classified the restaurant by different ***Cuisine*** based on the category list posted on Yelp Official Blog

Finding:

From July to September, ***the count of reviews*** of restaurants reached the ***peak***, especially the ***American*** restaurant.

Conclusion:

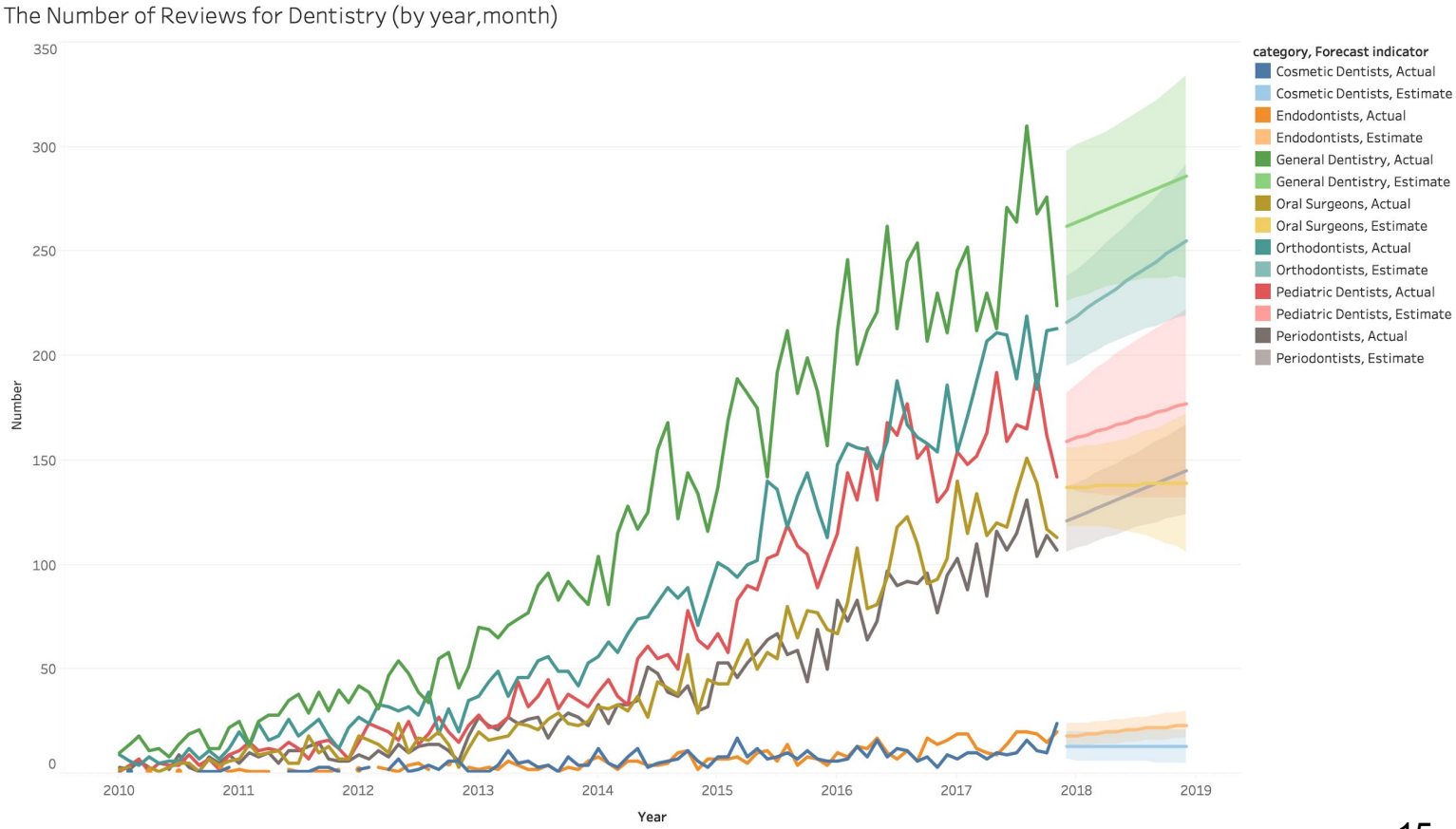
People are ***more likely go to restaurants in Autumn*** more than in other seasons

THE UNIVERSITY OF SYDNEY

3.7 *Q7-* Trend Description Analysis of Dental Reviews

Finding:
In general, the number of dental reviews were ***rising*** year by year, even, it may ***continuously rise*** in the near future.

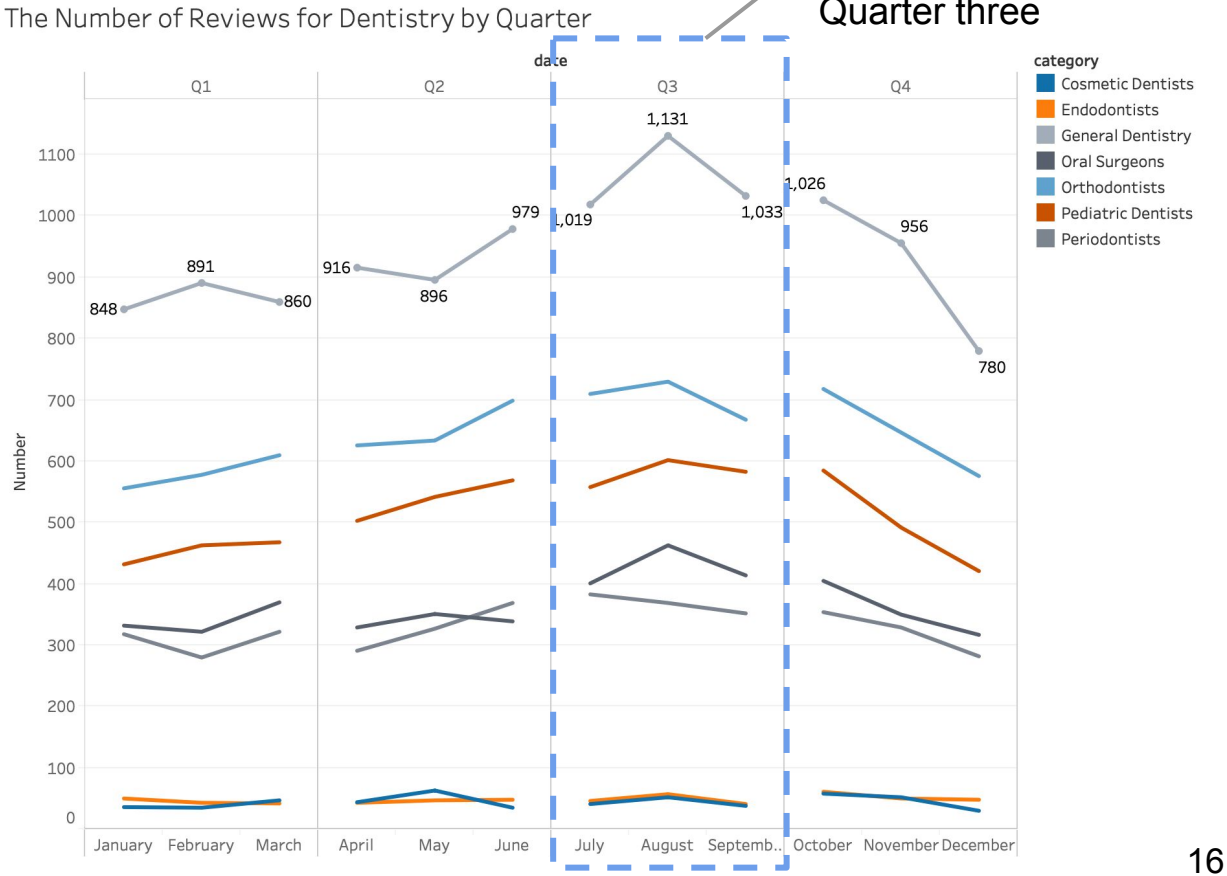
Conclusion:
Oral health problems are getting more and more ***people's concerns***



THE UNIVERSITY OF SYDNEY

3.8 *Q8-* Peak Incidence of The Dentistry

Similar to the trend with the restaurant in Quarter three



Finding:

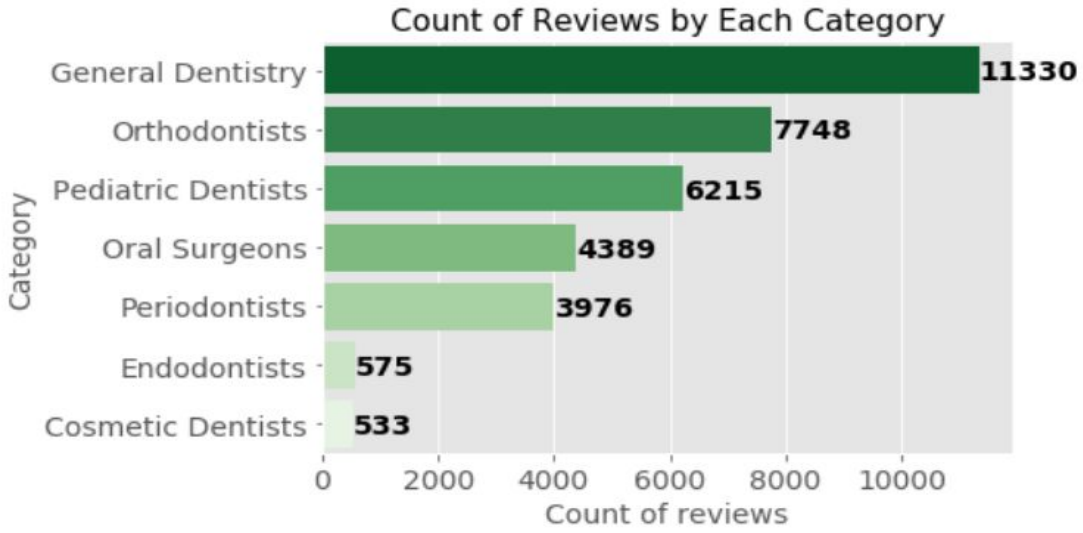
In general, the number of dental reviews were ***rising*** year by year, even, it may ***continuously rise*** in the near future.

Conclusion:

Oral health problems are getting more and more ***people's concerns***

THE UNIVERSITY OF SYDNEY

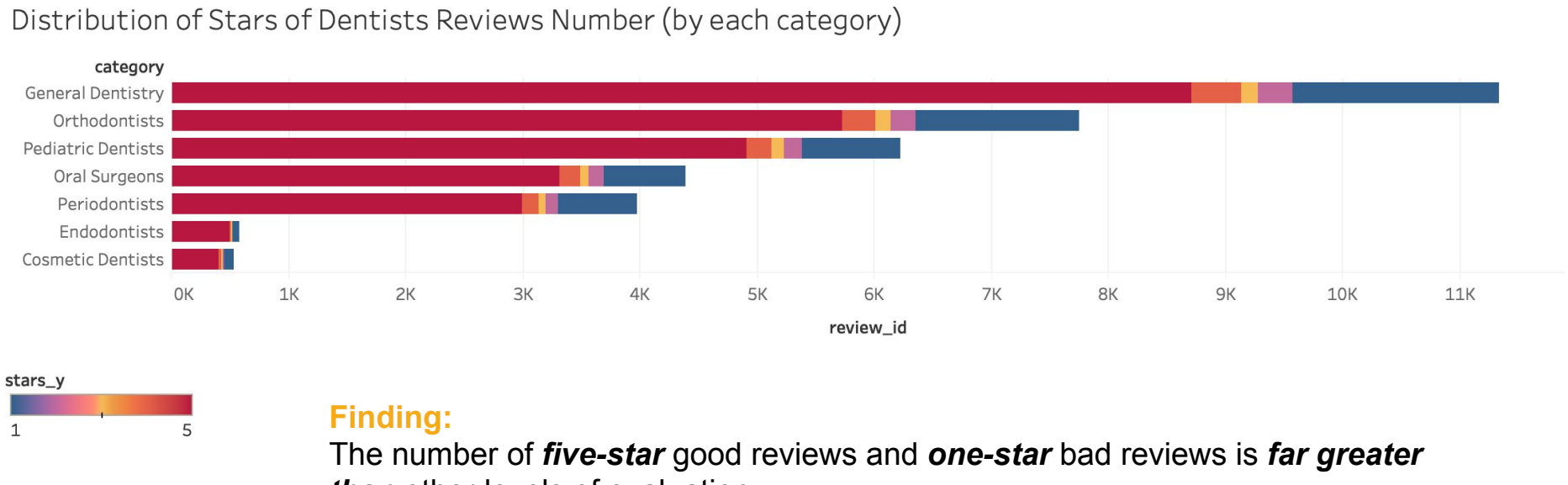
3.9 *Q9-* Distribution of Dentistry Reviews



Preprocess: Getting the count of reviews for each category in dentistry by using Python3

THE UNIVERSITY OF SYDNEY

3.9 *Q9-* Stars Distribution of Dentistry Reviews *cont.*



Finding:

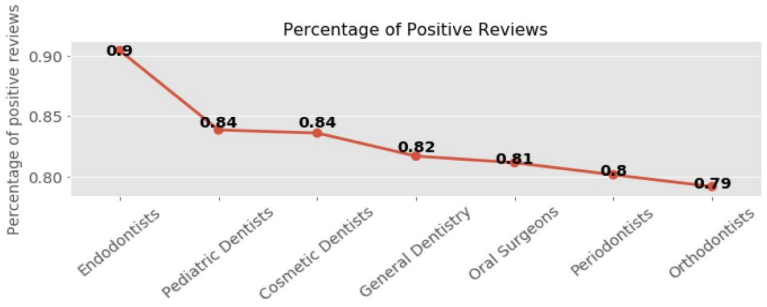
The number of ***five-star*** good reviews and ***one-star*** bad reviews is ***far greater than*** other levels of evaluation.

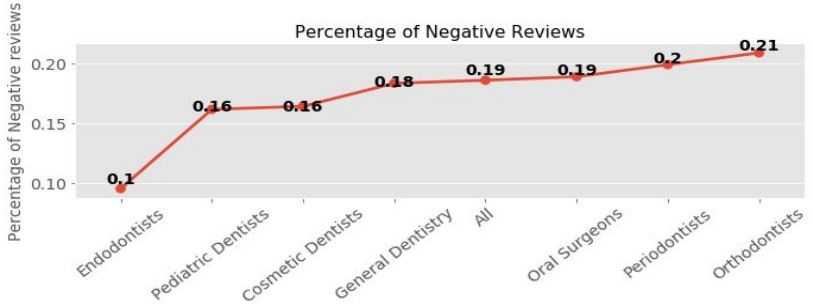
Conclusion:

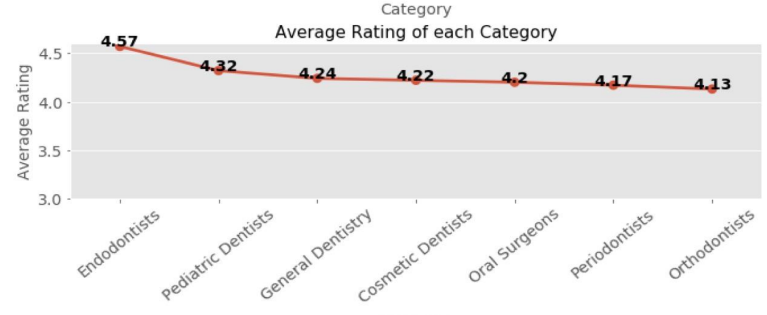
Customers are more willing to give ***the most positive and the most negative feedback***.

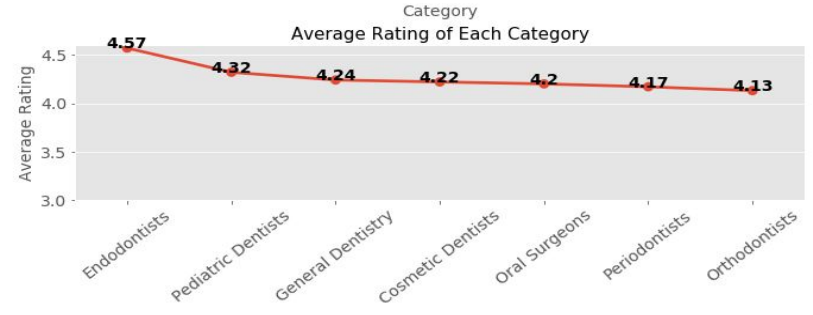
THE UNIVERSITY OF SYDNEY

3.10 *Q10-* Comparison of Stars of Reviews for Each Dental Category









Finding: **Endodontists got the highest star rating** in all dentistry, and **Orthodontists is the lowest**.

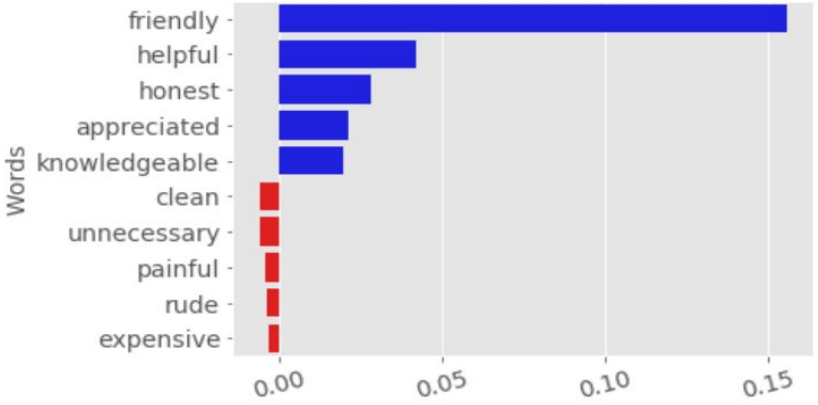
Conclusion: Endodontists is the category of dentistry that easier to get praise.

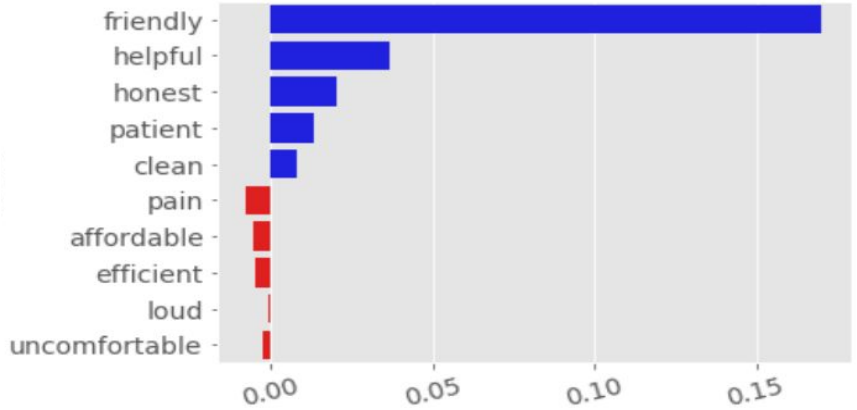
THE UNIVERSITY OF SYDNEY

04

Predictive Findings

4.1 *Q1- What factors can impact the Dentists' review?*



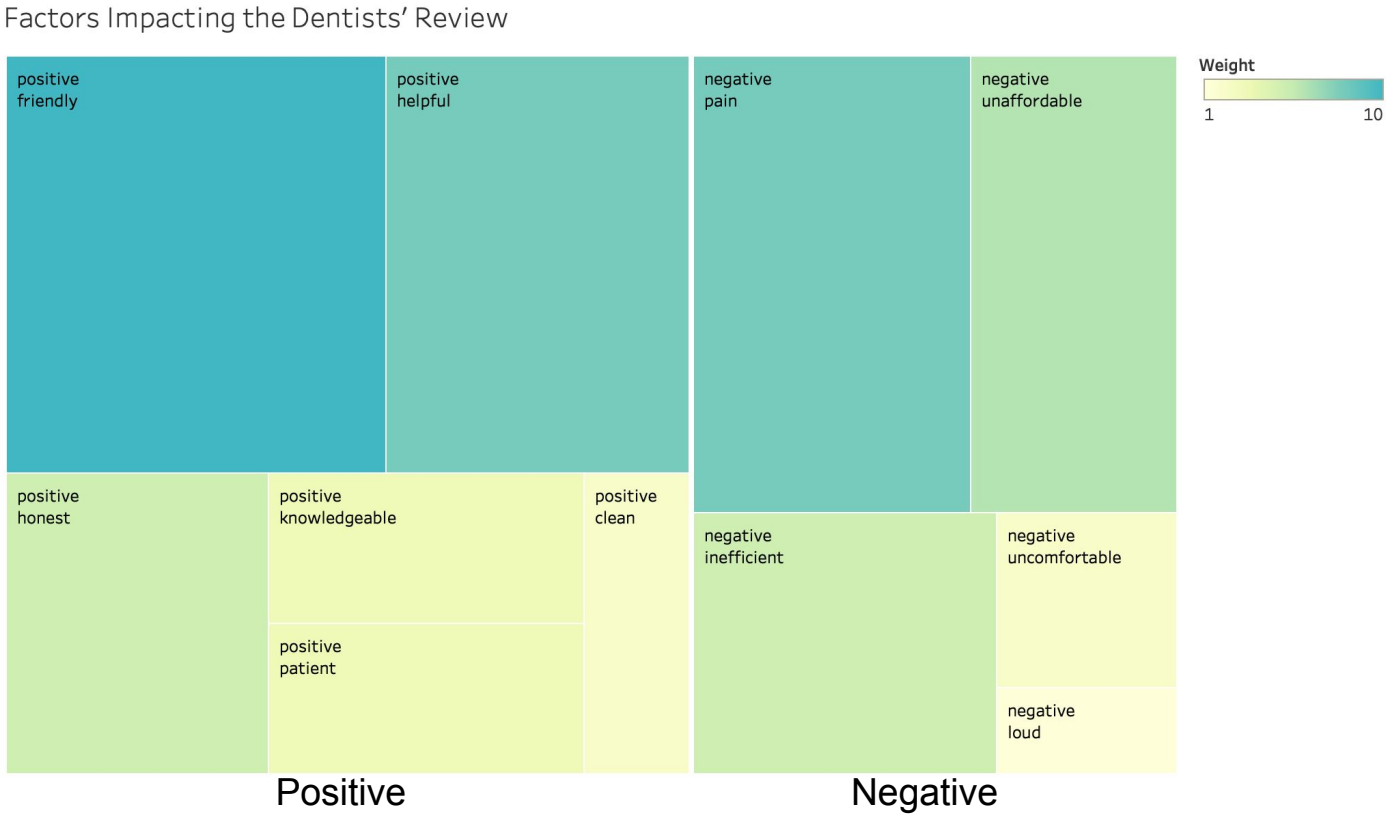


Preprocess: Import two file that contains common meaningless positive words such as 'like' 'pretty' and negative words such as 'bad' respectively, to delete those useless words from the words of reviews. By the way, we can keep more positive and negative words to apply Support Vector Machine model for extracting reviews that contain those words. After got the score of each word and the total count of extracted reviews of each category by SVM model, we calculate the frequency by each word, then multiplied the frequency with the word score and normalized the results by the total count of extracted reviews, so that we can get the polarity score --- the value that show how essential the positive/negative word among the reviews is.

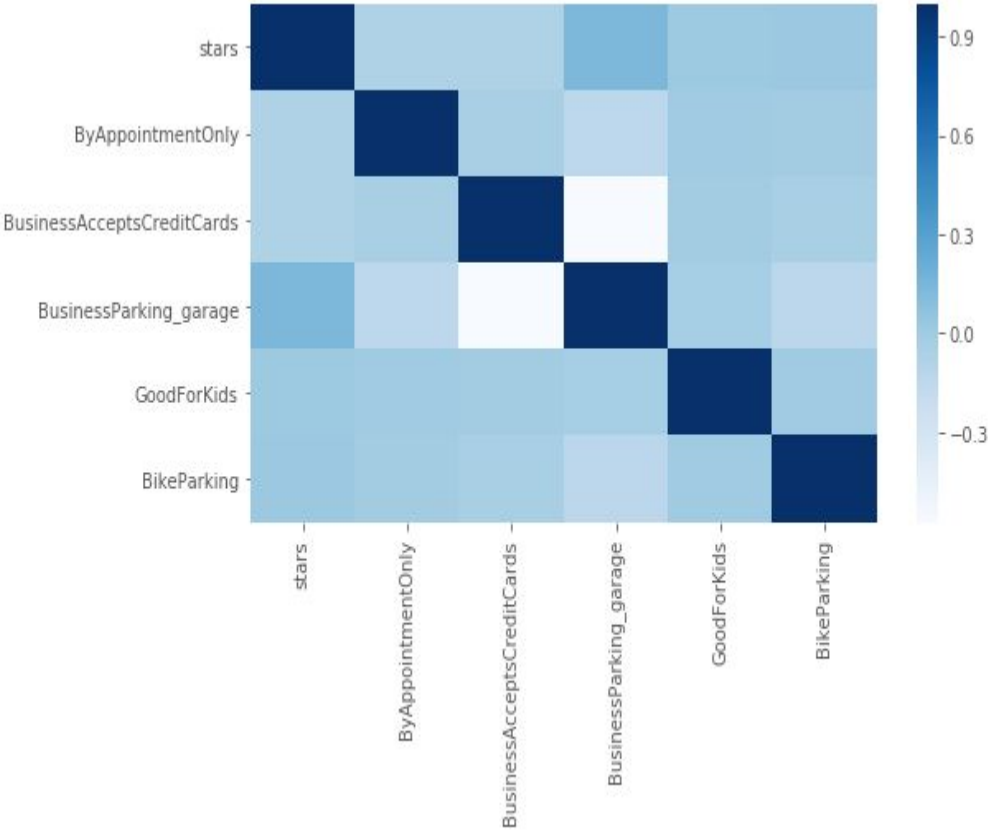
4.1 *Q1- What factors can impact the Dentists' review? -cont.*

Finding: Customers would like to share their attitude and feeling by reviews.

Conclusion:
The personality of dentist, environment of clinic, feeling experience of customers could impact the level (the star level) of reviews.



4.1 *Q1- What factors can impact the Dentists' review? -cont.*

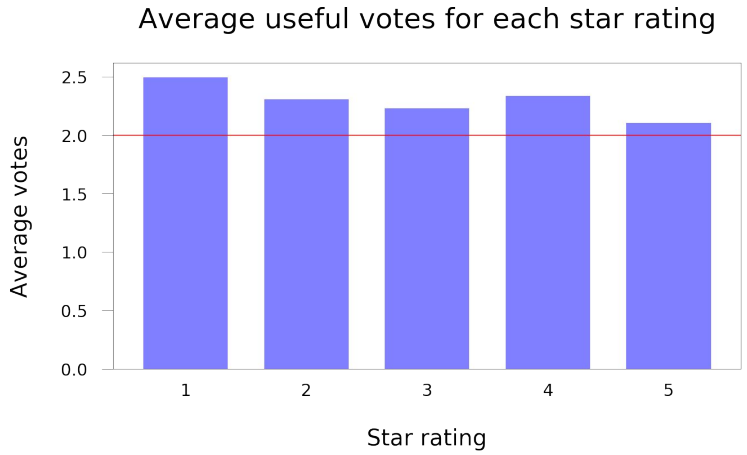


Like we mentioned before, we re-defined the data type of business attributes to calculate the correlations as we want to see whether these attributes such as 'ByAppointmentOnly' and 'AcceptsCreditCards' are impact the Dentists' review.

Finding: We calculate the average value of the whole business attributes, then we found that only 5 of them have non-zero value. Then we use heatmap to show the correlations of the 5 attributes with star rating.

Conclusion: BusinessParking_garage is the business attribute that affect the rating more than other attributes of Dental clinic.

4.3 *Q4- For restaurant and foods, can we predict star ratings based on user comments?*



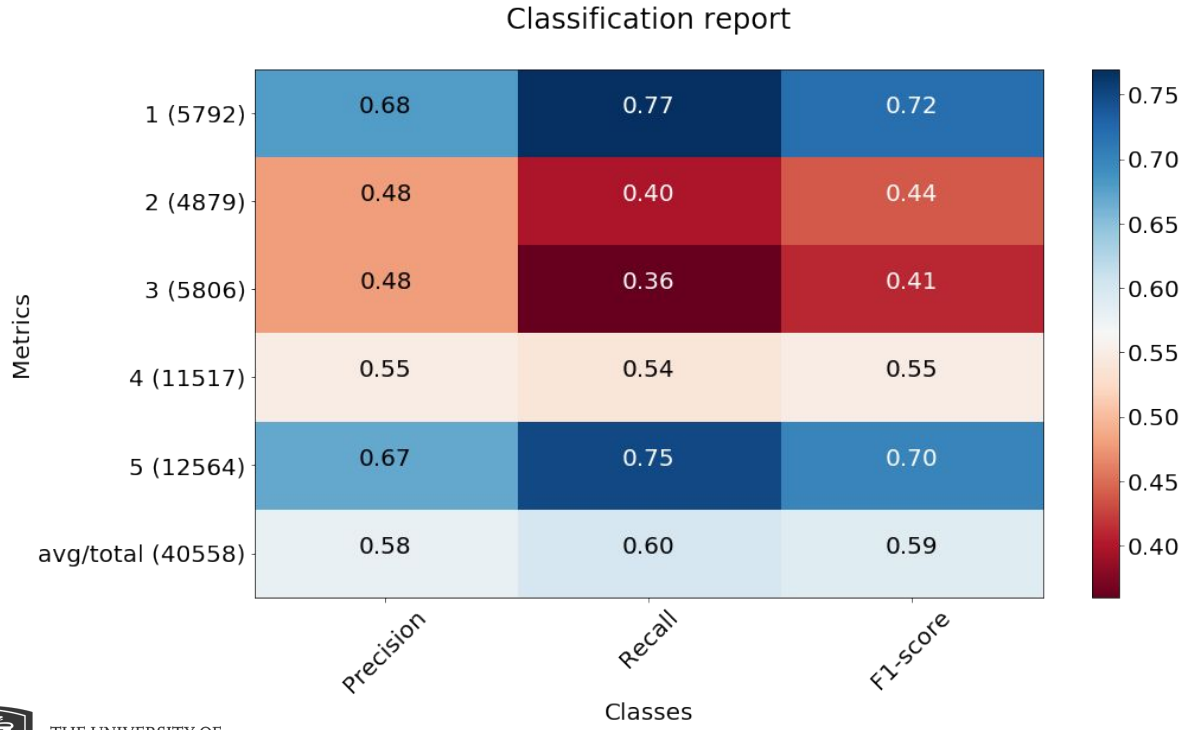
- Focus on user comments that have been voted as ‘useful’ more than twice (average cut-off for all ratings)
- Cleaning user comment
  ➔ Remove regex and stopwords, convert to lowercase, lemmatize words
- Unigram and bigram model was used to extract single and pair of words that were highly associated with each star rating

THE UNIVERSITY OF SYDNEY

4.3 *Q4- For restaurant and foods, can we predict star ratings based on user comments?*

Top 10 Keywords per category					
Star Ratings					
	5	4	3	2	1
1	delicious	perfect	definitely return	bland	mediocre
2	excellent	excellent	back	mediocre	terrible
3	amaze	perfectly	gem	unfortunately	horrible
4	awesome	amaze	thanks	rude	bland
5	perfect	fantastic	deserve	lack	disappointment
6	tasty	awesome	so good	terrible	meh
7	favourite	yummy	sick	horrible	overprice
8	fantastic	reasonable	heaven	overprice	rude
9	friendly	perfection	incredible	disappointment	disappointing
10	amazing	amazing	perfection	meh	sad

4.3 *Q4- For restaurant and foods, can we predict star ratings based on user comments?*



- Accuracy = 61.5%

THE UNIVERSITY OF SYDNEY

References

Carroll, J. (2018). *The Complete Yelp Category List*. Retrieved from https://www.yelpblog.com/2018/01/yelp_category_list#section9

Datacollaboratives.org. (2019). *Yelp Dataset Challenge*. [online] Available at: http://datacollaboratives.org/cases/yelp-dataset-challenge.html [Accessed 21 Feb. 2019].

Glaeser, Edward, Kim, Hyunjin, and Luca, Michael. “Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity.” NBER Working Paper Series (2017): n. pag. Web.

Hegde, Sindhu, Satyappanavar, Supriya, and Setty, Shankar. “Restaurant Setup Business Analysis Using Yelp Dataset.” 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Vol. 2017-. IEEE, 2017. 2342–2348. Web.

Takashima, Yuri, and Aono, Masaki. “Predicting the Usefulness of Cosmetic Reviews.” 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). IEEE, 2017. 1–6. Web.

Te, Yiea-Funk et al. “PREDICTING THE GROWTH OF RESTAURANTS USING WEB DATA.” Economic and Social Development: Book of Proceedings. Varazdin: Varazdin Development and Entrepreneurship Agency (VADEA), 2018. 237–256. Web.

THANKS

THE UNIVERSITY OF SYDNEY