Business Closing Prediction based on Yelp Dataset

# Introduction

Over time, we observe a continuous evolvement in technological advancement. Researchers are more focus on the way to make people’s lives easier. The Internet carries a vast range of information resources and services that people use and depend on day-to-day. We saw in the meantime the proliferation of [website](https://en.wikipedia.org/wiki/Website) on which reviews can be posted about people, businesses, products, or services.

Yelp.com is known as one of the top review sites where consumers can share experiences about product quality and gather information about experience goods, which have quality that is observed only after consumption. Some articles revealed that online reviews have an impact on a business. Thus, businesses should care about them for preventing bad consequences or finding new business opportunities. On December 9, 2015, The New York Times titled in his Entrepreneurship section that: “A Bad Review Is Forever: How to Counter Online Complaints”.

However, can bad reviews lead to a business closure? Or what could be the main factors which impact closing business? The purpose of this project is to predict whether a business will remain open or going to close based on the Yelp dataset.

After importing the dataset from <<https://www.yelp.com/dataset>>, I will tidy and transform it to get only the data concerning businesses, reviews and users from Toronto. Then, I will generate some visualizations to raise patterns, show findings, and build models to answer the above questions.

# Literature Review

Do online consumer reviews affect restaurant demand [1]? Michael Luca from Harvard University investigated that question in 2016 by combining two datasets from Yelp.com and the Washington State Department of Revenue. It has been finding that variation in star rating can lead to an increase or a decrease in revenue. That is mostly visible by independent restaurants but not for restaurants with chain affiliation. Changes in quality retain more the attention of consumers and the users respond more when a review contains more information. Most of time, users rely more to rating when a restaurant got an important number of reviews compare to the size of reviewers.

Yelp reviews and ratings are important source of information to make informed decisions about a venue. After inspected hundreds of reviews, five dimensions have been finding about restaurant businesses. The research has been done by analyzing business reviews and building a classifier to describe different type of businesses in the restaurant industry. Those categories are “Food”, “Service”, “Ambience”, “Deals/Discounts” and “Worthiness” [2].

Using a deep learning method to analyze photos and reviews of restaurant posted on Yelp.com from December 2004 to December 2015, it has been found that the volume and the valence of photos are strong predictors of restaurant survival. When it comes to reviews only the valence matters [3], and a restaurant closure is strongly associated with a consumer sentiment.

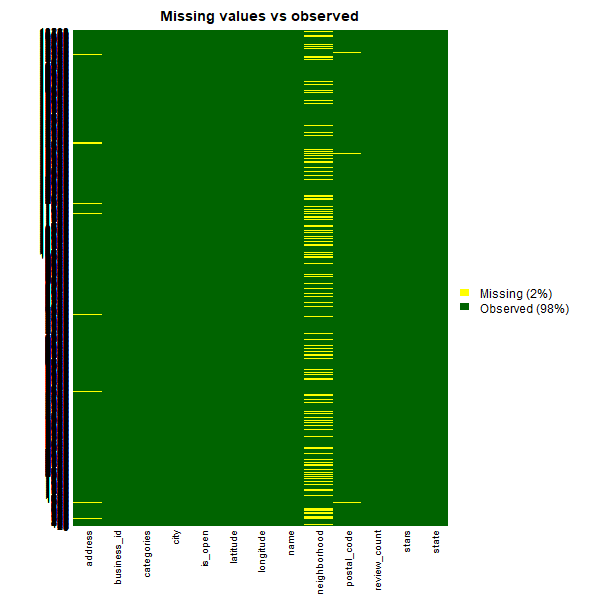
# Dataset

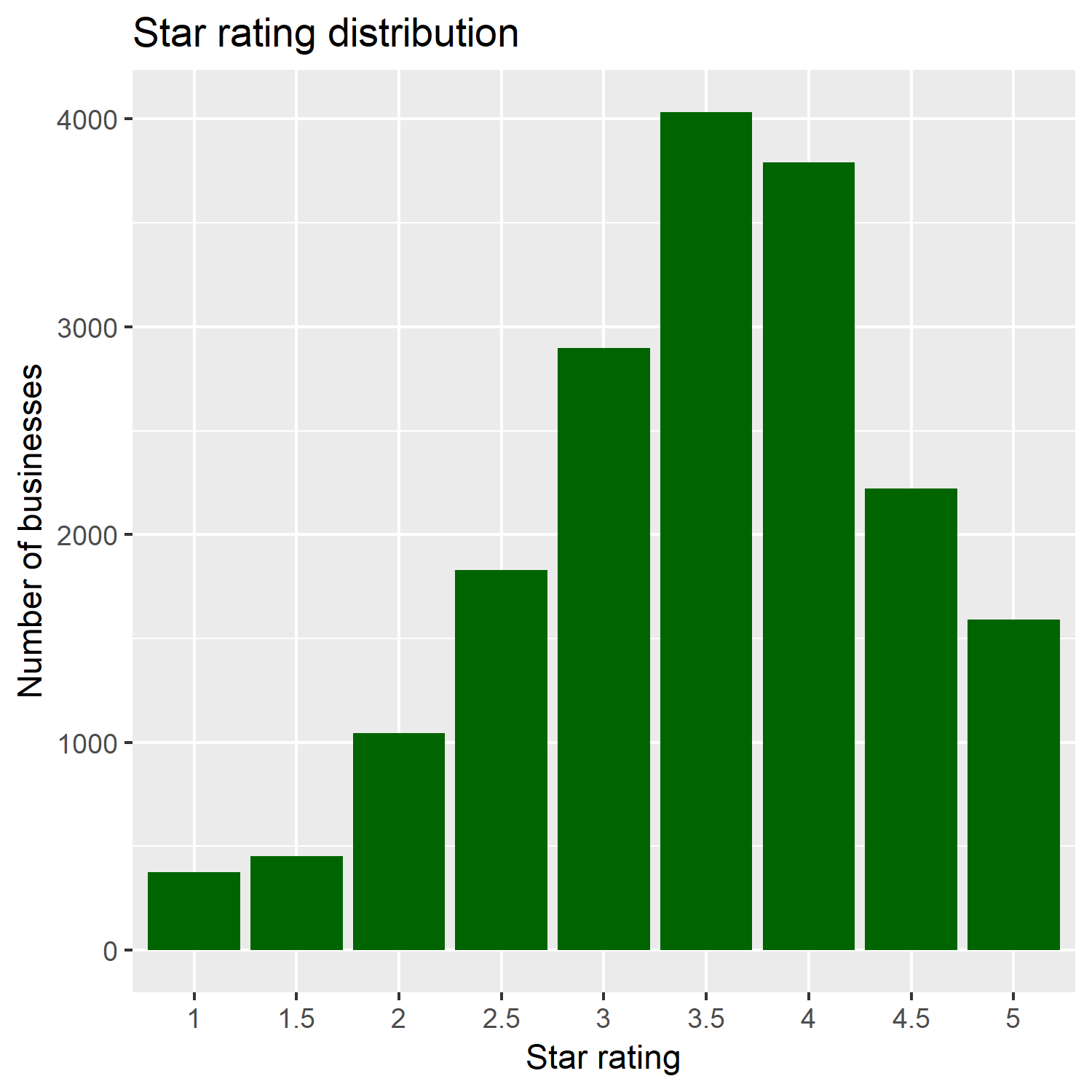
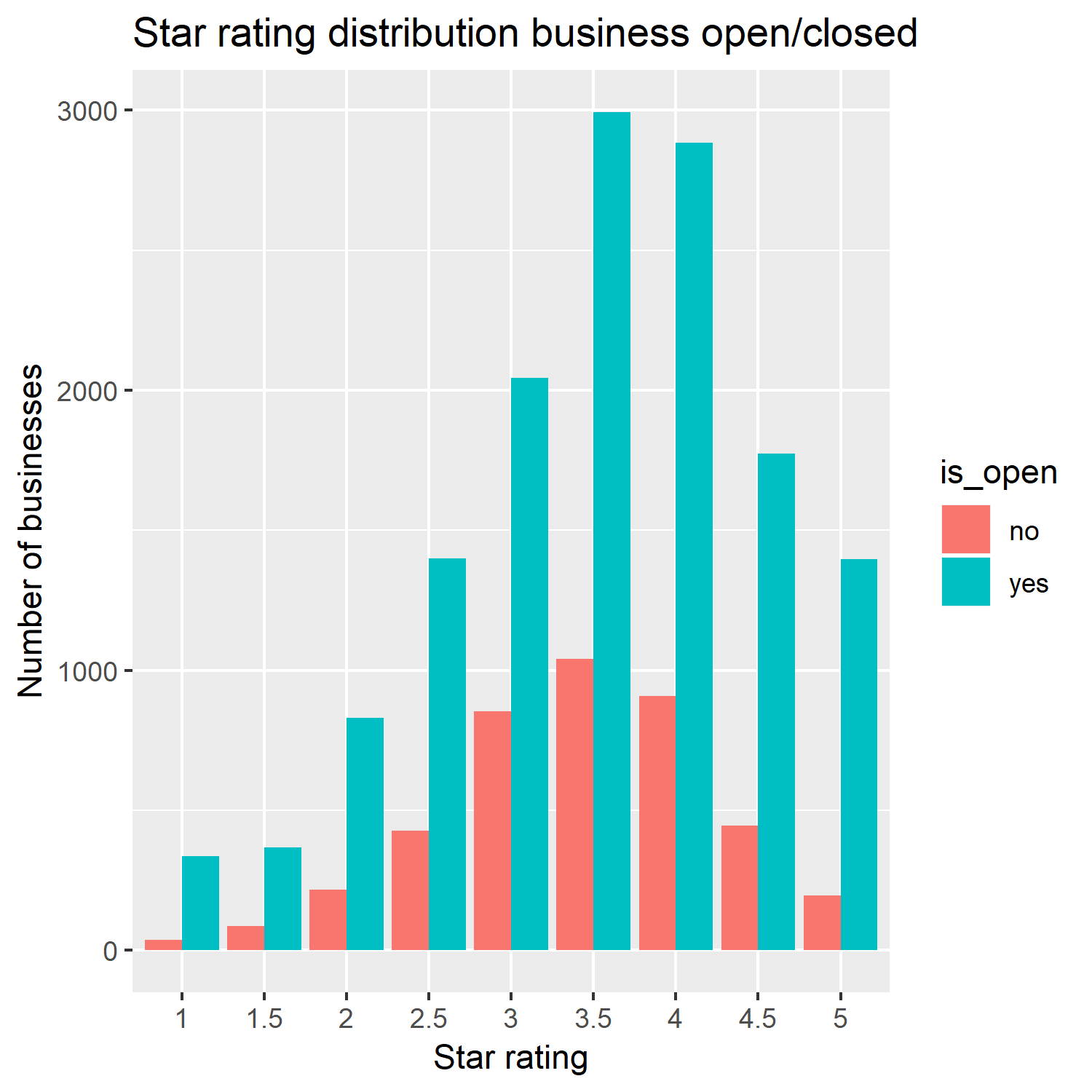
The dataset has been imported from <<https://www.yelp.com/dataset>>, it is JSON dataset which contains information for 188,593 businesses. This dataset has 5,996,996 reviews, 1,518,169 users, 280,992 photos, 157,075 check-ins, 1,185,348 tips for these businesses.

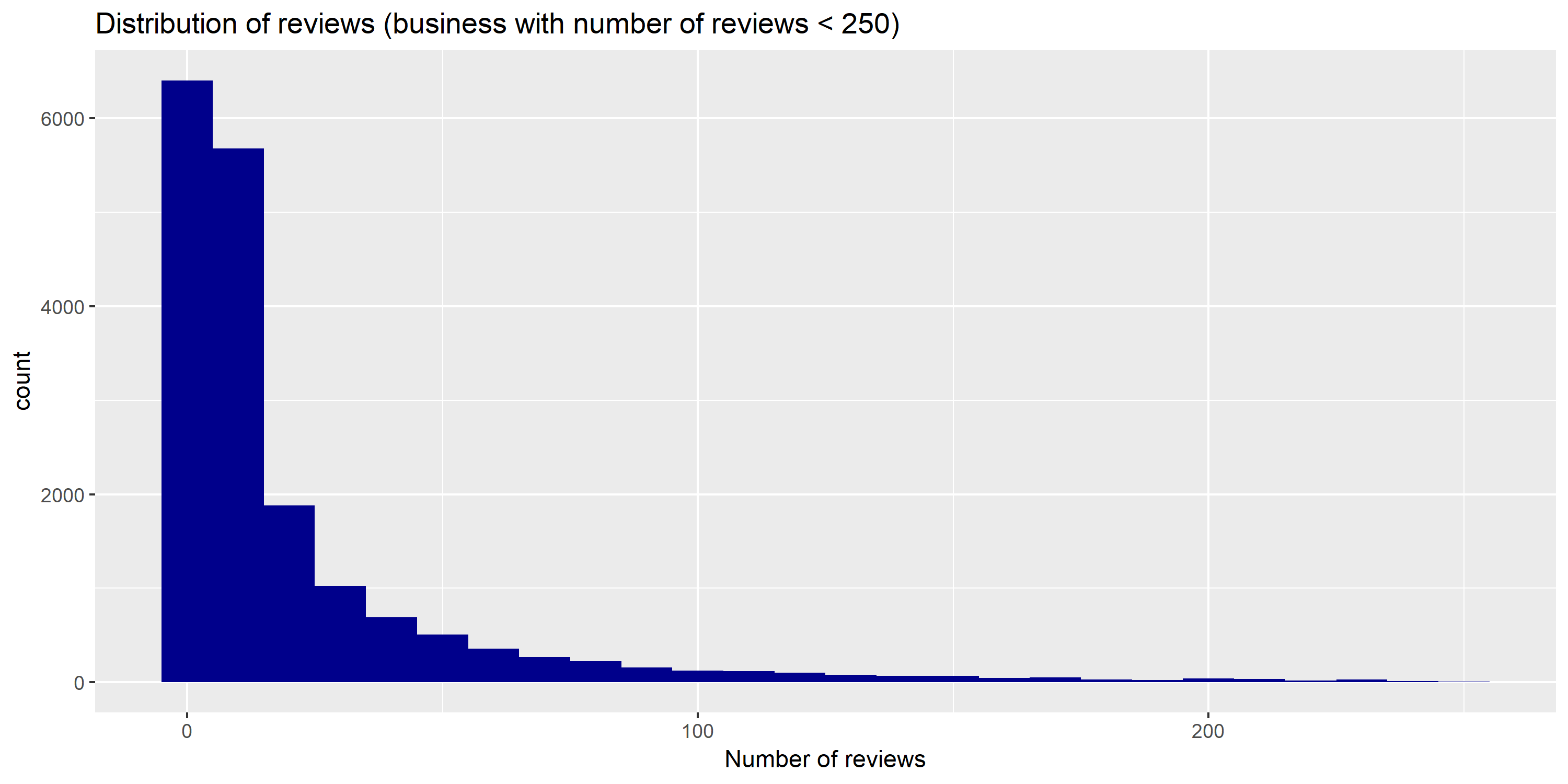
For the project, a subset of this dataset will be used by filtering the raw sets of data to get information related to businesses in the city of Toronto. This reduced the number of business to 18,233 with 474,803 reviews from 103,262 users. The goal will be to predict whether a business is close or open from the attribute “is\_open” in the business data.

The focus will be on the business data set which contains 18,233 observations and 13 variables. The variables are:

* address: character class with 14422 unique values and 283 missing values
* business\_id: character class with 18233 unique values and 0 missing values
* categories: character class with 10028 unique values and 33 missing values
* city: character class with 1 unique values and 0 missing values
* **is\_open (target variable)**: integer class with 2 unique values (0 or 1) and 0 missing values
* latitude: numeric class with 15366 unique values and 1 missing values
* longitude: numeric class with 15315 unique values and 1 missing values
* name: character class with 15292 unique values and 0 missing values
* neighborhood: character class with 80 unique values and 3435 missing values
* postal\_code: character class with 5261 unique values and 117 missing values
* review\_count: integer class with 380 unique values and 0 missing values
* stars: numeric class with 9 unique values and 0 missing values
* state: character class with 1 unique values and 0 missing values







From the graph above, it could be gathered that businesses "closed" received less reviews, and that does not relate to rating. “Closed” and “Open” businesses have the same distribution of rating.

Some additional information will be retrieve from the reviews data set to get more insights about each business. For that purpose, feature engineering will be applied to specific attributes like date of a review, the stars obtained by review and its text content.

# Approach

**Import the dataset**

**Tidy and transform**

**Visualize**

**Modelling**

**Conclusions**

## Step 1: Data loading

In this step, I first subscribed to the Yelp challenge to get the permission to download the dataset. After downloaded the data set and stored in a file, I imported into RStudio.

## Step 2: Data wrangling

This is the part where I tidied and transformed the data to get in a form that’s natural to work. Firstly, as it is a zip file containing 6 JSON files I ran a script to unzip it and then convert the JSON formats to R objects.

<<https://github.com/PatKakou/capstone-project/blob/master/Yelp_Dataset_JSON_to_RDS.Rmd>>

Secondly, I made some sub-setting to get the observations of Toronto city from the business data set and the review data set.

<<https://github.com/PatKakou/capstone-project/blob/master/Yelp_Toronto.Rmd>>

Finally, I will apply NLP techniques to extract keywords on reviews then label the reviews as positive and negative using a set of keywords. The purpose here is to gather some important insights from the review data set to create new variable and add them to the business data set.

## Step 3: Exploratory Data Analysis

A good visualization will show you things that you did not expect or raise new questions about the data [9]. In that third step, I will do some generate leads that I can later explore in more depth. I will iteratively generate questions about the data, search for answers by visualizing and transforming the data, use what I will learn to refine my questions and/or generate new questions.

My goal here is to develop a good understanding of that data set.

## Step 4: Model building

The purpose of this step is to create classification model to predict whether or not a business will survival. I will create different models, evaluate them make a comparison to choose the best one. After, I will tune that model to improve it.

## Step 5: Conclusions

Finally, the different inferences gather from the previous steps will be share here and some solutions will be discussed for the next steps.

**References**

[1] Luca, Michael, Reviews, Reputation, and Revenue: The Case of Yelp.Com (March 15, 2016). Harvard Business School NOM Unit Working Paper No. 12-016. Available at SSRN: <https://ssrn.com/abstract=1928601> or [http://dx.doi.org/10.2139/ssrn.1928601](https://dx.doi.org/10.2139/ssrn.1928601)

[2] <https://www.ics.uci.edu/~vpsaini/files/technical_report.pdf>

[3] Zhang, Mengxia and Luo, Lan, Can User Generated Content Predict Restaurant Survival: Deep Learning of Yelp Photos and Reviews (March 2018). Available at SSRN: <https://ssrn.com/abstract=3108288> or [http://dx.doi.org/10.2139/ssrn.3108288](https://dx.doi.org/10.2139/ssrn.3108288)

[4] Luca, Michael and Zervas, Georgios, Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud (May 1, 2015). Harvard Business School NOM Unit Working Paper No. 14-006. Available at SSRN: <https://ssrn.com/abstract=2293164> or [http://dx.doi.org/10.2139/ssrn.2293164](https://dx.doi.org/10.2139/ssrn.2293164)

[5] O'Hern, Matt S. and Rindfleisch, Aric and Schweidel, David A. and Antia, Kersi, The Impact of User-Generated Content On Product Innovation (March 3, 2011). Available at SSRN: <https://ssrn.com/abstract=1843250> or [http://dx.doi.org/10.2139/ssrn.1843250](https://dx.doi.org/10.2139/ssrn.1843250)

[6] McQuarrie, Edward F. and McIntyre, Shelby H. and Shanmugam, Ravi, What Motivates Consumers to Produce Online Reviews? Solidarity, Status, and the Soapbox Effect (February 1, 2013). Available at SSRN: <https://ssrn.com/abstract=2210707> or [http://dx.doi.org/10.2139/ssrn.2210707](https://dx.doi.org/10.2139/ssrn.2210707)

[7] [https://en.wikipedia.org/wiki/Review\_site#Impact](https://en.wikipedia.org/wiki/Review_site%23Impact)

[8] The New York times, A Bad Review Is Forever: How to Counter Online Complaints

<https://www.nytimes.com/2015/12/10/business/smallbusiness/small-business-counter-bad-reviews.html>

[9] Garrett, Grolemund and Hadley Wickham*,* R for Data Science (January 5, 2017). http://r4ds.had.co.nz/introduction.html