Mini Project 3

Vijay Teja Gottipati, 110622613

Introduction

The goal of this project is to do exploratory analysis on the Yelp academic dataset 1 and build some interesting model. The dataset consists of 5 JSON files, one each for: Users, Reviews, Tips, Check-ins and Businesses.

Data

The dataset consists of around 1.6M reviews, 500K tips, 366K users, 61K businesses and aggregated check-in information for 45K businesses.

The relevant features of each of the 5 entities are:
Business: city, review_count, neighborhoods, full_address, hours, state, longitude, stars, latitude, attributes, open, categories
Users: yelping_since, votes, elite, compliments, fans, average_stars, review_count, friends
Checkins: checkin_info, business_id,
Tips: user_id, text, business_id, likes, date
Reviews: votes, user_id, review_id, text, business_id, stars, date

The dataset is overall very clean and did not require any preprocessing for initial analysis.

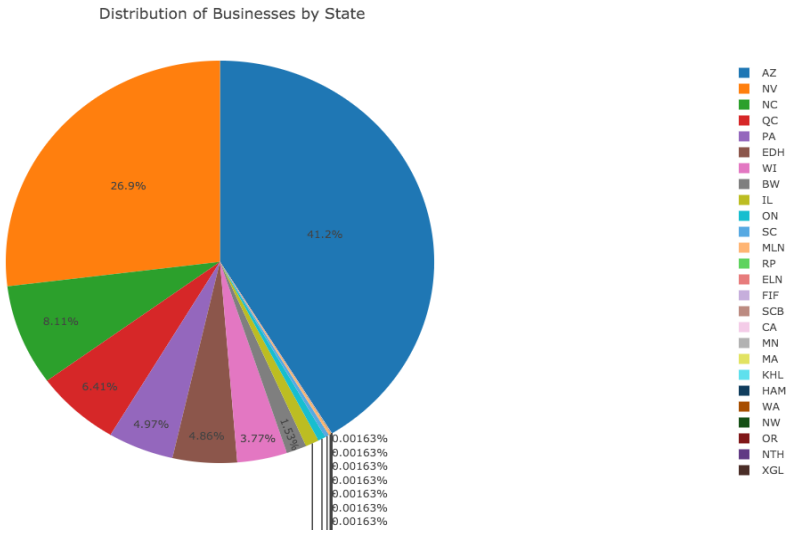
Problem Statement

The primary goals of the project are:

1. Exploratory analysis.
2. Model to predict if a user is an elite yelper.

Exploratory Analysis

Let us first see the spread of business by state.

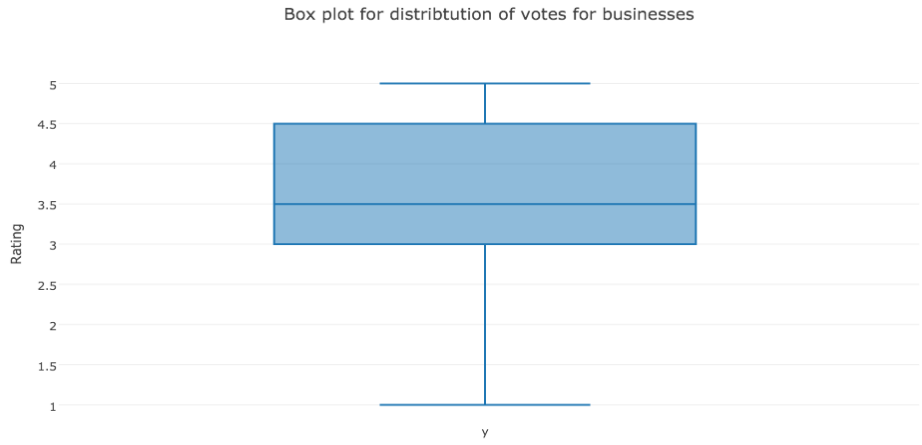


Businesses from Arizona and Nevada constitute the vast majority of the businesses.

City wise, the top 5 cities with most places are:

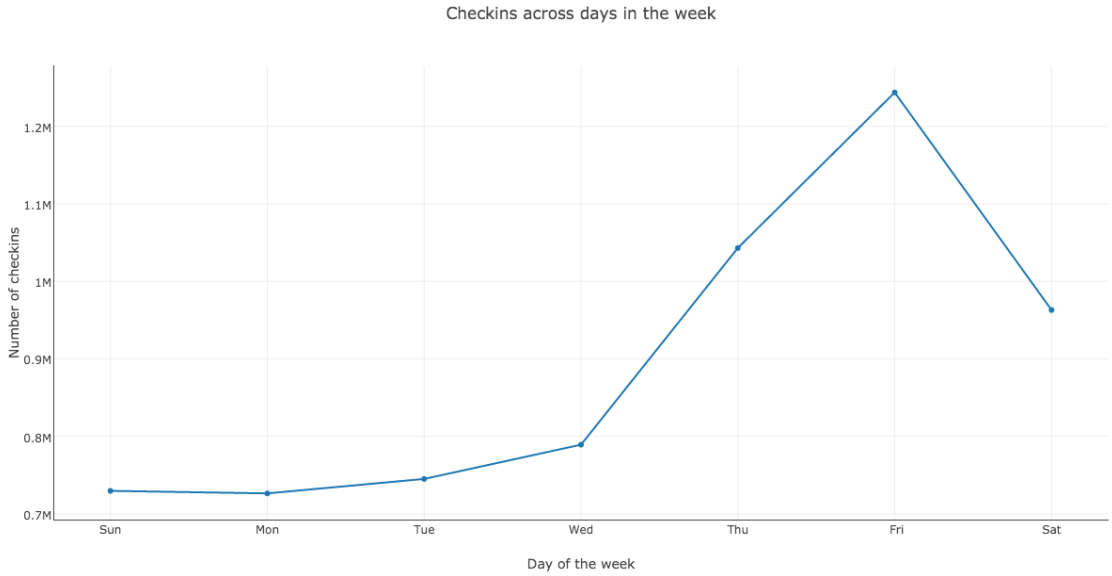
City	Percentage of places	Number of businesses
Las Vegas	22.229668	13601
Phoenix	13.745424	8410
Charlotte	6.903766	4224
Scottsdale	6.601399	4039
Edinburgh	4.953910	3031

Next let's look at the distributions of average stars that all the business get.

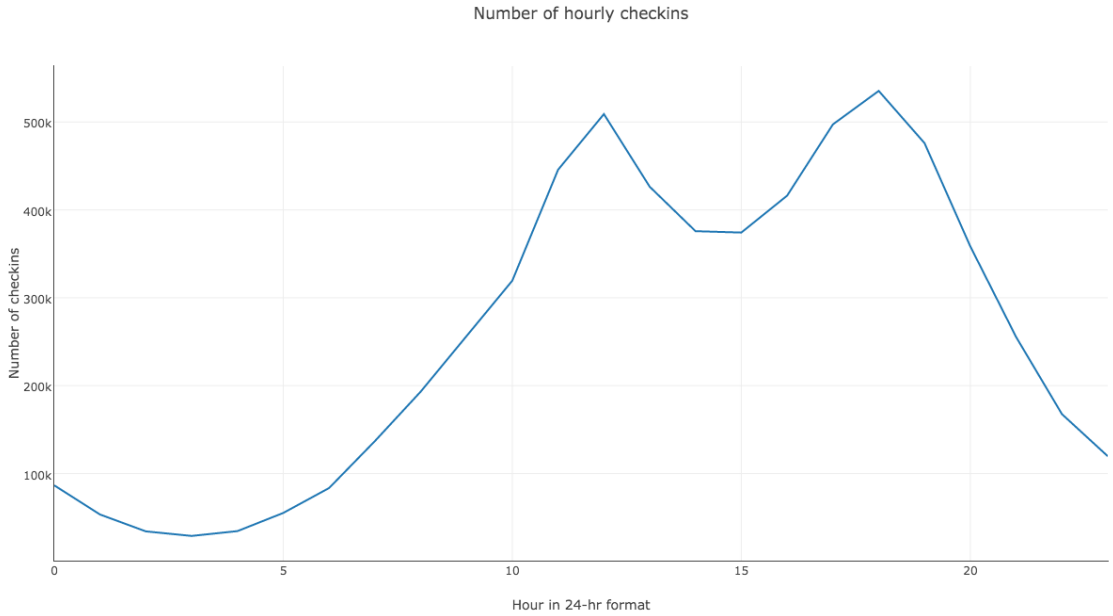


From the box plot it appears that the distribution is skewed towards higher ratings. Ratings of below 3 constitute of just 25% of the businesses.

Lets next look at the aggregate check-ins across time of the day and day of the week.

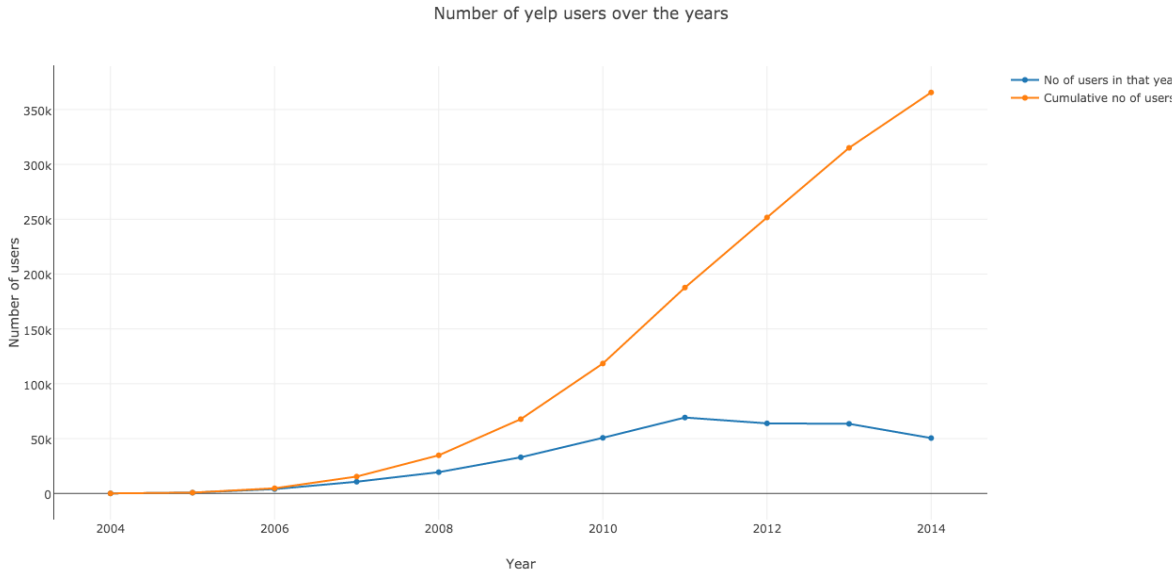


Looking at the day distribution, it is intuitive that the check-ins are peaking on Friday. It is a little surprising that Sundays get less check-ins than Wednesday, the middle of the week. Business can use this distribution to provision appropriate amount of resources in terms of staff and inventory management based on the day, or use promotions to attract customers on the low days.



The early distribution also looks intuitive, with bi-modal peaks at 12PM and 6PM; typical lunch and dinner times.

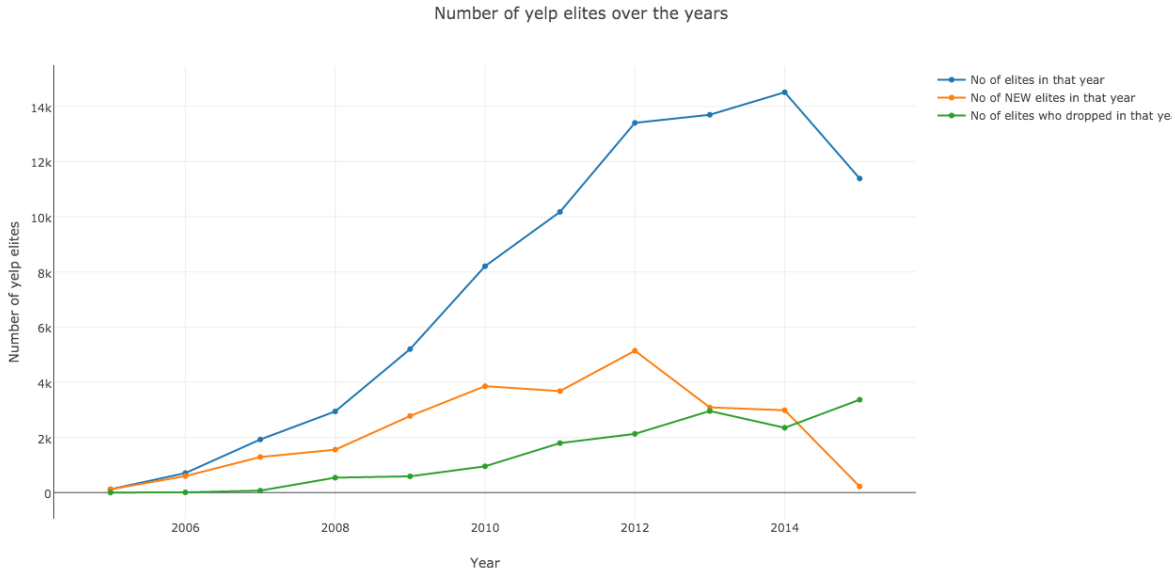
Let us now look at the user numbers and growth:



The aggregate number of users seem to be rising exponentially. But the number of new users added in a year has peaked at 2011 and has been marginally decreasing. This suggests that the market in these areas have saturated and there is less scope for growth in terms of users.

Elite yelper classification

Yelp has a concept of elite squad where they select a small number of users and give them elite status. Elite status is given for a year. At the end of a year, the status is again re-evaluated. Looking at the numbers, as of 2014 they are 14512 elites in that year, or just ~4% of the users. Let us look at the growth and attrition of elite users:



Note: The data points for 2015 are still incomplete since the year is not yet over, but I am still including it for reference.

Looking at the graph, we can see that the elite squad is increasing over the years, exponentially at first, but mellowed from 2012 onwards. This peak also coincides with the peak of number of new users who get the elite status. This number is decreasing from 2012 onwards. A possible reason could be the saturation that we saw in total number of new users in the previous graph. This number was calculated by incrementing the count only for the 1st year that a user was an elite member. Repeat elite members were not included.
We can also see that the number of elites who did not retain their status from previous year are also increasing. This number was calculated by looking at all the years that a user had elite status and incrementing the drop count for every year that the user did not retain the elite status.

Yelp describes that they look for Authenticity, Contribution, and Connection in a user to qualify them for an elite status. On the surface these qualities look subjective and hard to quantify. Let us see if we can build a model to predict if a given user is an elite or not based on the features from the user data.

Feature selection: Let us look at the various attributes of a user and see which of them could be a helpful feature. The attributes are:
yelping_since, user_id, name, votes, elite, compliments, fans, average_stars, review_count, friends
From here we can remove the user_id and name since they do not give any useful information about the user. All the other attributes look like they can provide something helpful about the user that the other attributes do not.

Feature cleaning: These features have to be tweaked to make them suitable for a model.

- I have transformed the elite feature from list of years a user was elite to a binary variable that is True if the user was elite in any year. This would be the class variable that we want to predict.

- Having a list of friends is not suitable for a model. I replaced it with the number of friends that the user has.
- Transformed the yelping_since attribute to age attribute that just records how long the user has been on yelp.
- The votes attribute has vote counts for 3 attributes: funny, cool and helpful. I split the attribute into 3 attributes: funny_votes, useful_votes, and cool_votes.

Model selection: Gaussian based Naïve-Bayes classifier seems like an apt model for this binary classification task.

Training: The 366K users were split into training and test data in 66.66:33.33 ratio. The model was evaluated with a 10-fold stratified cross validation. The metric chosen to evaluate the model was F-1 score since it is important to be both accurate in the prediction (less false positives) and still not miss out on classifying the correct instances (less false negatives). On an average, the model has an F-1 score of 0.71 in predicting the elite class during the cross valuation.

Evaluation: Running the model on the 33.33% of the test data, the model performed with an F-1 score of 0.7, with a precision and recall of 0.74 and 0.67 respectively.
The confusion matrix is as follows:

	Predicted Non-elite	Predicted Elite
Actual Non-elite	37437	560
Actual Elite	772	1570

Conclusion

Most of the business in this dataset were from Arizona and Nevada with Las-Vegas and Phoenix being the major business hubs on this Yelp data. Looking at the check-in information it shows that the weekends bring most of the business and the food time of 12 and 6PM are the most popular times for businesses. This day distribution in particular should be helpful to business owners by making informed decisions on resource and inventory management. Even though Yelp describes their selection for elite squad in terms of subjective criteria like Authenticity, Contribution, and Connection, by building a Gaussian based Naïve-Bayes classifier model with an accuracy of ~74% we can see that their selection is not so subjective and is relatively simple since despite making the Gaussian and independence assumptions, we still get a decent. This is further validated by the fact that the model is able to distinguish an elite user even though the data was disproportionately skewed with elite users being just ~6% of all the users. It possible that by including the reviews that a user has given as a feature, we could build a better model and capture the ‘authenticity’ aspect of the selection.