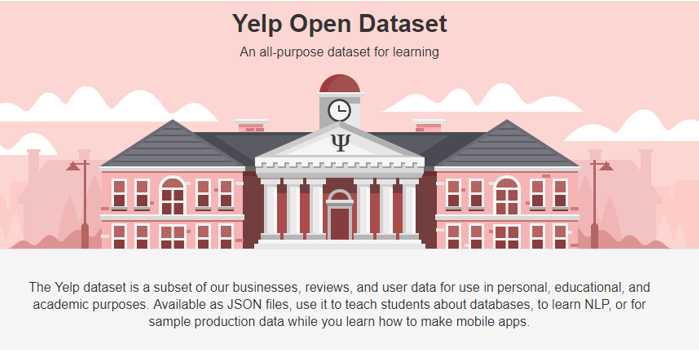
# Introduction

The objective of this report is to document the analysis and findings of a dataset containing data released by Yelp (the business listing and crowdsourcing review company).



This data is inclusive of reviews, business details that the reviews relate to, user who submitted the reviews, tips submitted buy users, checkins submitted by users and photos.

In addition, financial as well as demographic data would be used to enrich the analysis.

The main tasks that were worked on as input to this report are:

1. Data Exploration, Quality Management and Initial Transformation – to gain an understanding of the data attributes, to perform quality management checks and to conduct initial transformations on the source datasets.
2. Integration and Preparation – to integrate data from the source datasets into a larger combined dataset containing all required features, and then to prepare the integrated data for further analysis
3. Analysis and Geographic Visualization
4. Text Mining for sentiment analysis, authenticity analysis and topic modelling.
5. Forecasting to predict future counts, as well as the prediction validity.
6. Comparative analysis to determine how individual businesses are performing in relation to other similar businesses.
7. Results – describe what was found during the analysis.

# Analysis and Models

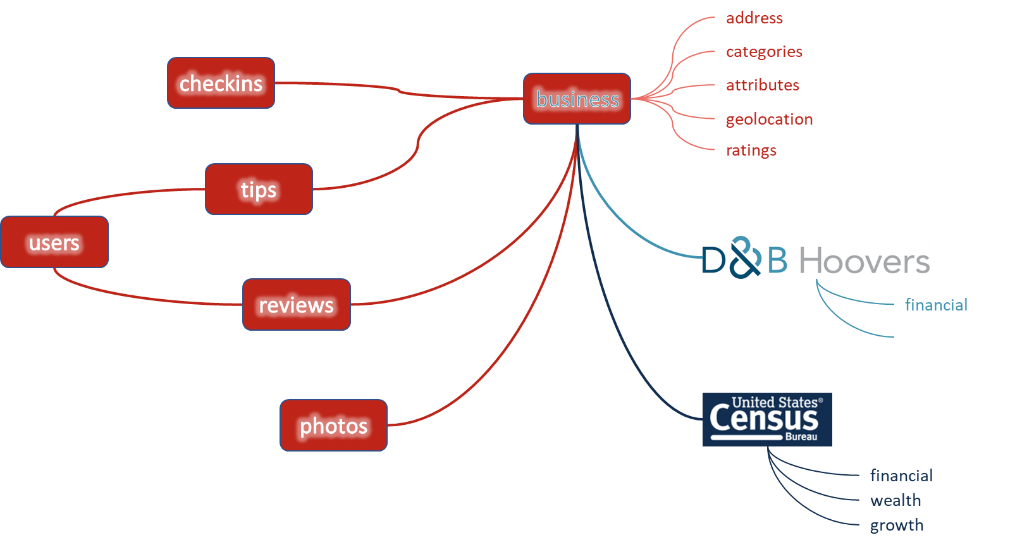
## About the Data

### Source Data

Data was sourced from three different places:

* Yelp Data as sourced from the Yelp Dataset challenge website,. This came in the form of a zipped file that contained 6 different JSON format datasets. The datasets contained the following data:
  + Reviews
  + Businesses
  + Tips
  + Checkins
  + Users
  + Photos
* Dunn and Bradstreet Hoovers business data; for restaurants in Arizona and Nevada
  + This data was selected and downloaded from the Dunn & Bradstreet Hoovers online business dataset
* US Census
  + This data was obtained from the US Census Bureau

This figure shows the conceptual data model that describes the data we obtained:



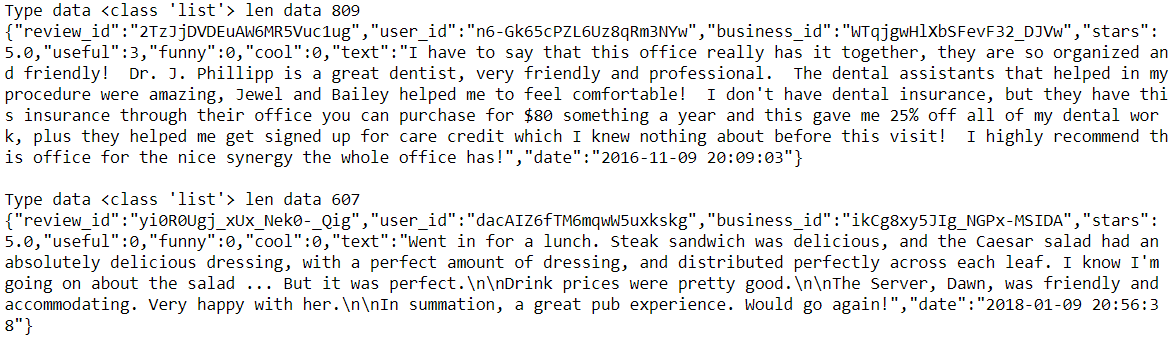
The eight sets of data are described below:

#### Yelp - Reviews

This is the dataset that most of our analysis and modelling revolved around. The dataset type is record type and comprises 6,685,900 observations across 6 variables.

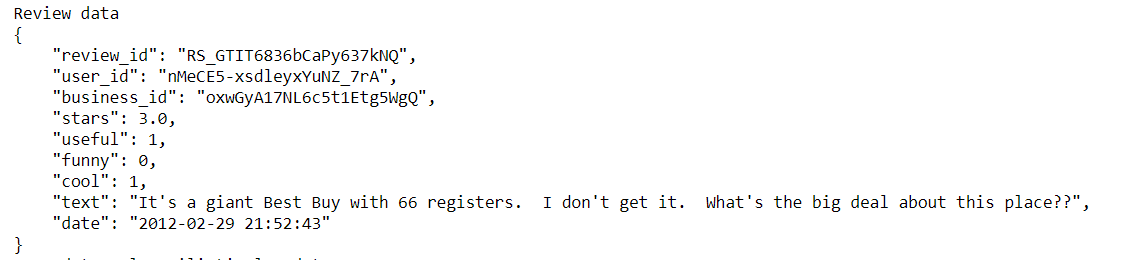
The attribute type of three of the 5 columns are nominal; and the other 1 attributes was numeric.

Data was in JSON form, sample of two rows as followed:



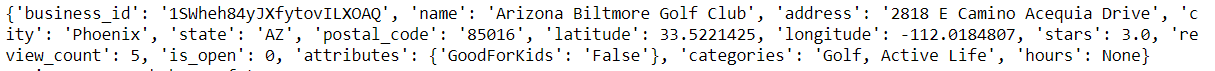
Due to the size of this dataset, the data had to be read and processed on an observation by observation basis. It was NOT possible to use standard JSON handlers to parse the dataset in one go.

A sample of 1 observation of the data parsed out is given below:

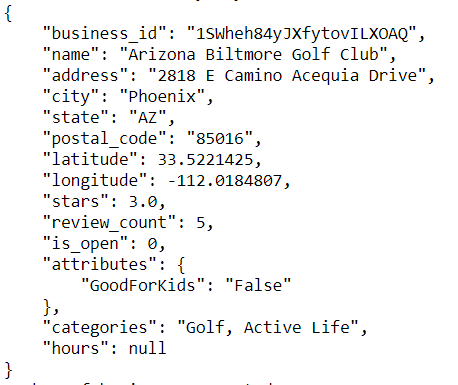


#### Yelp Businesses dataset

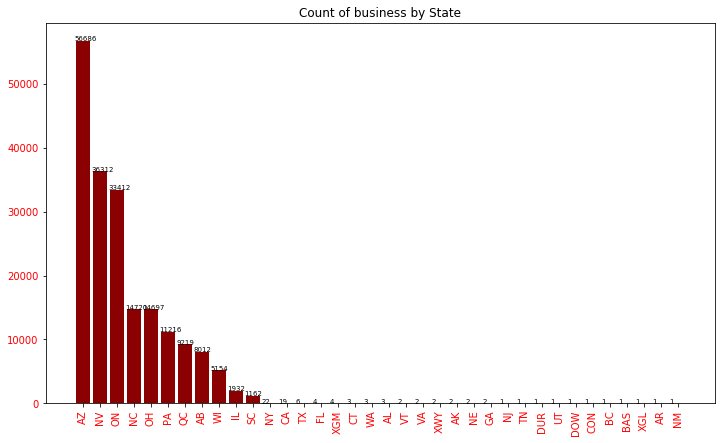
This is the dataset that most of our analysis and modelling revolved around. The dataset type is record type and comprises 19,122 businesses across x variables.

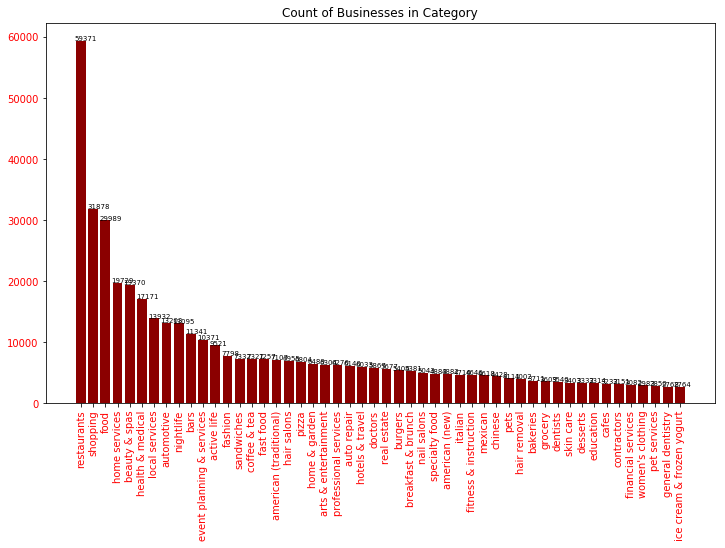
A sample of raw data for 1 business is given below:  


And this translates into a more meaningful form as follows:



From an analysis of count of business by state, the following bar chart is produced:



In addition to looking at business by state, businesses were also examined by category of business. The chart below shows how many businesses there are in each category. 

The attribute type of eight of the 14 columns are nominal; and the other 6 attributes are numeric.

Source data is in JSON form,

#### Yelp Users dataset

This is the dataset that most of our analysis and modelling revolved around. The dataset type is record type and comprises 1,637,138 observations across 6 variables.

The attribute type of three of the 5 columns are nominal; and the other 1 attributes was numeric.

Source data was in JSON form

A sample of raw data for 1 user is show here:



This parses out into the data shown below:





Data attributes on the Yelp Users dataset, in alphabetical order, are:

|  |
| --- |
| average\_stars |
| compliment\_cool |
| compliment\_cute |
| compliment\_funny |
| compliment\_hot |
| compliment\_list |
| compliment\_more |
| compliment\_note |
| compliment\_photos |
| compliment\_plain |
| compliment\_profile |
| compliment\_writer |
| cool |
| elite |
| fans |
| friends |
| funny |
| name |
| review\_count |
| useful |
| User\_id |
| yelping\_since |

#### Yelp Tips dataset

This is the dataset that most of our analysis and modelling revolved around. The dataset type is record type and comprises 1,223,094 observations across 6 variables.

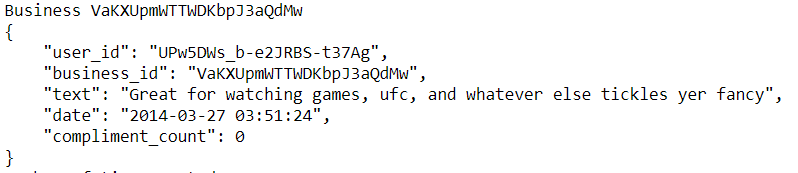
The attribute type of three of the 5 columns are nominal; and the other 1 attributes was numeric.

Source data was in JSON form.

A sample of 1 observation in raw form is:



This translates into a more readable form as follows:



The data attributes on the tips dataset are:

|  |
| --- |
| user\_id |
| business\_id |
| text |
| date |
| compliment\_count |

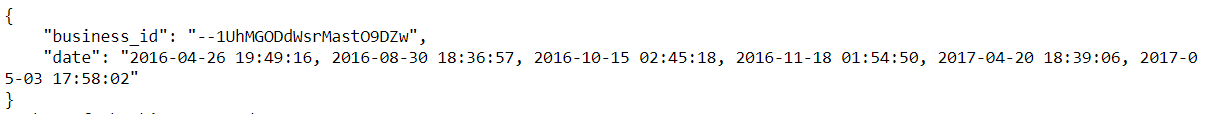
#### Yelp Checkins dataset

This is the dataset that most of our analysis and modelling revolved around. The dataset type is record type and comprises 161,950 observations across 2 variables.

A sample of raw data for 1 user is given as follows:



And is transformed into a more meaningful set like this:



The attribute type of one of the 2 columns is nominal; and the other 1 attributes is numeric.

Source data is in JSON form.

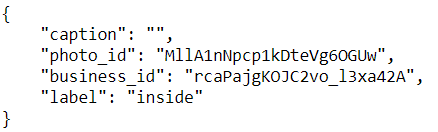
#### Yelp Photos dataset

This is the dataset that our photo analysis was used for. In reality, the JSON dataset is an “index” dataset, that is of record type and comprises 200,000 observations. What this does is provide access to a set of photos, each of which is referenced via a photo\_id that exists in the index file.

A sample of the raw data for 1 photo record is shown here:



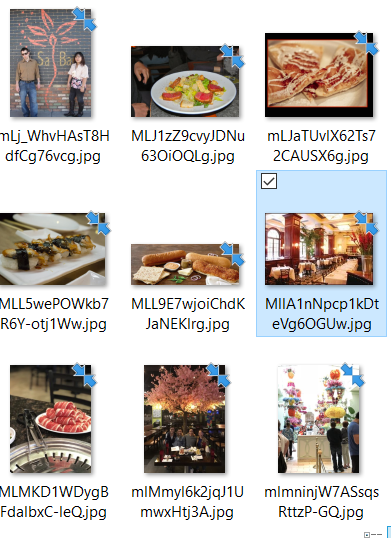
And it expands as follows:



The attribute type for all 4 photo index columns are nominal.

This photo index acts as a pointer to photo “MllA1nNpcp1kDteVg6OGUw.jpg” in the image dataset.

A sample of nine images in the image dataset are shown below:



And the image for the photo corresponding to the data shown above is shown here:

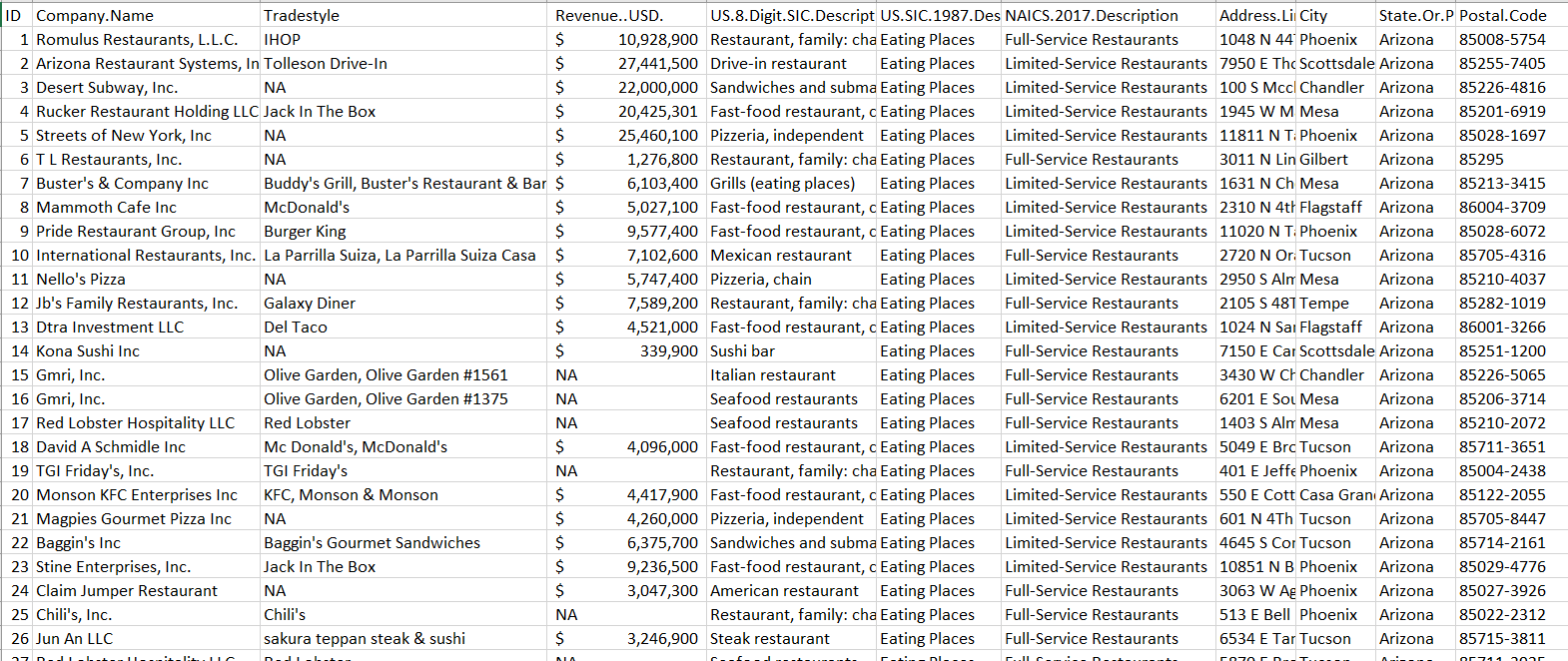


Source data of the photo index file is in JSON form; format of the photo is in .jpg form.

#### D&B Hoovers financial dataset

The D&B Hoovers Financial dataset was obtained from the online catalog, where all business in Nevada and Arizona that are classified as restaurants were extracted. The dataset type is record type and comprises 27,524 observations with multiple financial and business variables.

A sample of the data is given below:



Source data is in excel form.

#### 2010 Census dataset

The intention was to use the census data to enrich our review dataset. The dataset type is record type and comprises 1,622,831 observations across 6 variables.

A sample of the data is given as follows:

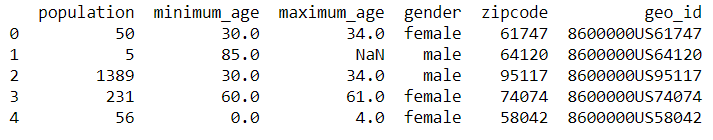


Figure :Sample Census 2010 Population data by zip

This was summarized by zipcode into 33,119 observations with 2 attributes; zipcode and population count per zipcode – with descriptive statistics below.

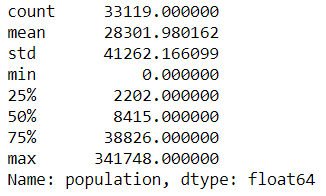


Figure : Descriptive stats for Population 2010 by zipcode data

#### Missing Data

Two observations must be considered here:

1. Obvious data that is missing, based on the project brief is “Annual Donations” data.
2. Regression showed a 90% adjusted R^2 meaning; equivalent to 90% confidence in the factors we use to predict our results. The implication is that there are more attributes that we should include in our regression – but we did not include them in this study.

### Quality

After reading in all 5 datasets, the underlying quality of the data was reviewed.

### Transformation and Filtering

Due to the number of observations of the review data, as well as the difficulty workin with 6MM observations, it was decided to concentrate on reviews for business in Arizona and Nevada that are in the top 5 business categories. Categories selected were (1) Restaurants, (2) Shopping, (3) Food, (4) Home Services and (5) Beauty Spas.

The count of business selected, by category, is given in the table below:



Subsequent to selection of the relevant businesses, all other subsets were filtered to represent the businesses selected. This brought the observation counts for the Yelp data down, as shown below:

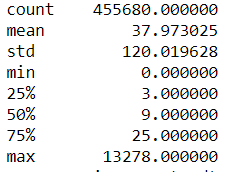


The selection and transformation per dataset is further described in the sections below:

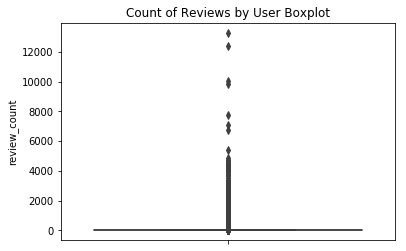
#### Yelp Users Filtering

Users were selected and analyzed based on reviews they gave for the businesses of interest.

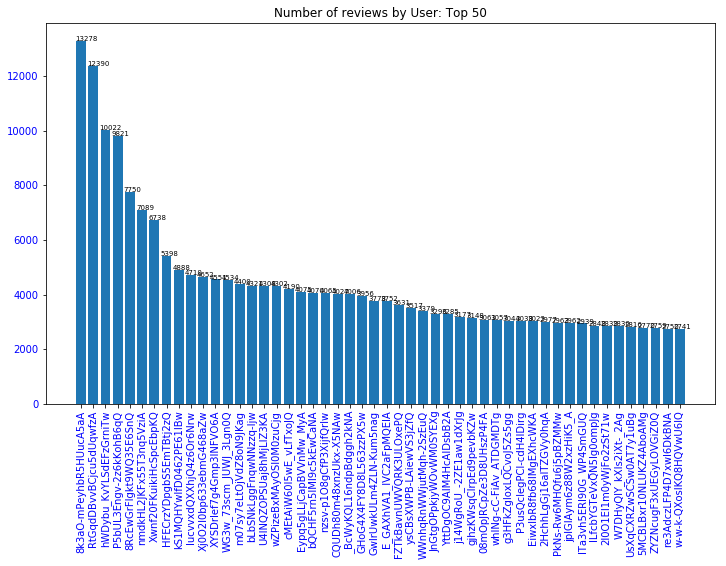
Given that there are 19,122 businesses of interest, with 1.042,847 reviews given by 455,680 users; the descriptive stats of count of reviews per user is shown as follows:



With a corresponding boxplot:



And the Top 50 reviewers are shown here:



#### Yelp Tips dataset

Based on the data attributes on the Tips dataset, filtering for the purposes of this analysis can be done in one of two different ways.

Way 1: Filter by business\_id. When tips are filtered by business, the count of tips selected was 219,946 tips. This is a relatively simple selection process because the business are first filtered by state and category, and the then tips are filtered based on the subset of businesses.

Way 2: Filter by user\_id. When tips are filtered by user\_id, the count of tips selected was 721,253. This is a more convoluted filtering process which accounts for the difference in tips counted. The path is (1) our selected business -> (2) our selected reviews -> (3) our selected users -> (4) our selected tips.

For subsequent analysis including tips, the tips by business subset was used.

### Integration

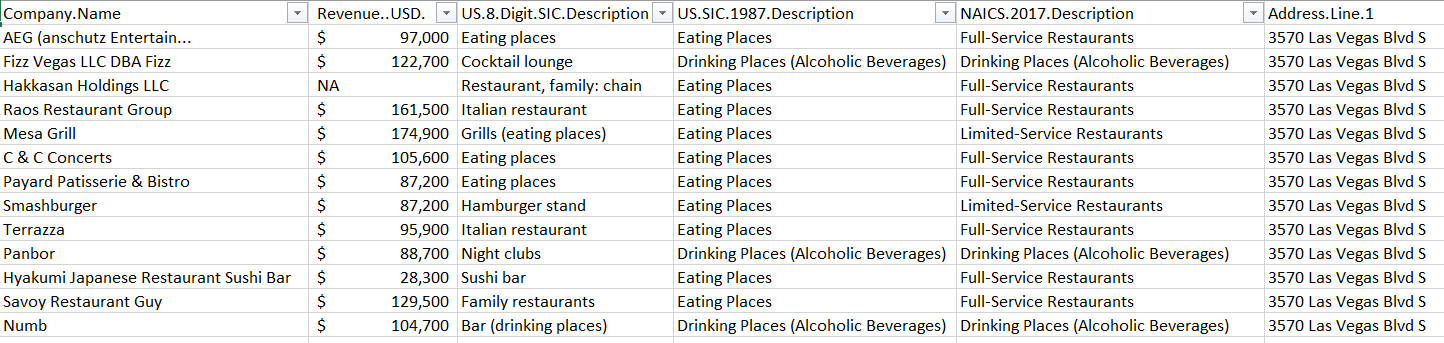
The approach that was taken to integration was to rely on the Yelp identifiers to integrate the Yelp datasets that are of interest:

* Our Yelp Businesses will become the master dataset.
* Our Yelp Reviews links to Our Yelp Businesses via business\_id.
* Our Yelp Photos links to Our Yelp Businesses via business\_id.
* Our Yelp Users integrates to Our Yelp Reviews via user\_id.
* Our Yelp Checkins integrates to Our Yelp Users via user\_id.
* Our Yelp Tips integrates to Our Yelp Users via user\_id.

#### D&B Hoovers Financial dataset integration

It became apparent that integrating businesses on the Financial dataset with those on the Yelp business dataset was not practical, as the criteria for what was considered a Restaurant in the D&B Hoovers dataset was completely different to that on the Yelp business dataset.

As an example, the Bacchanal Buffet (one of the most reviewed restaurants in Las Vegas was listed on the Yelp data as being at 3570 S Las Vegas Boulevard, Las Vegas, 89109; whereas there is NO Bacchanal Buffet on the D&B Hoovers list. The Restaurants that are listed at this address on D&B Hoovers are:



It should be note that the Bacchanal Buffet is a restaurant in the Caesars Palace resort.

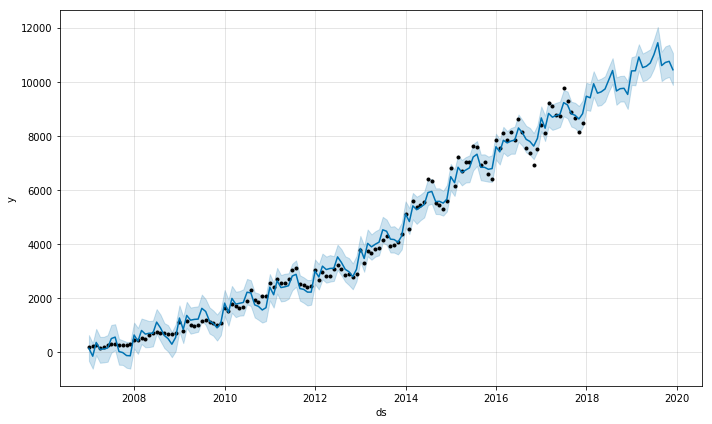
As a second example, the Gordon Ramsay BurGR restaurant was also examined, with the same results. No ability to match across datasets.

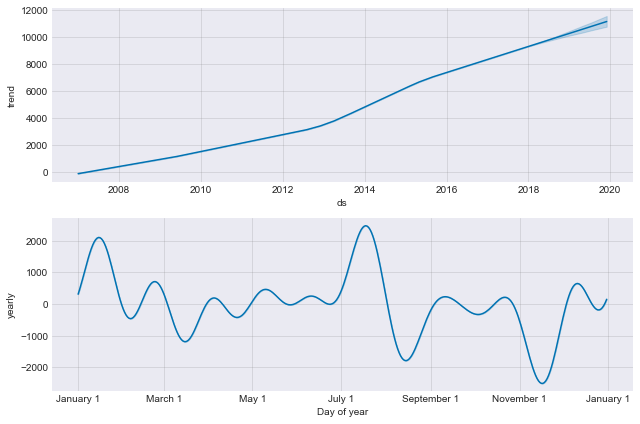
As a result of this, we did not use the Financial dataset for subsequent analysis.

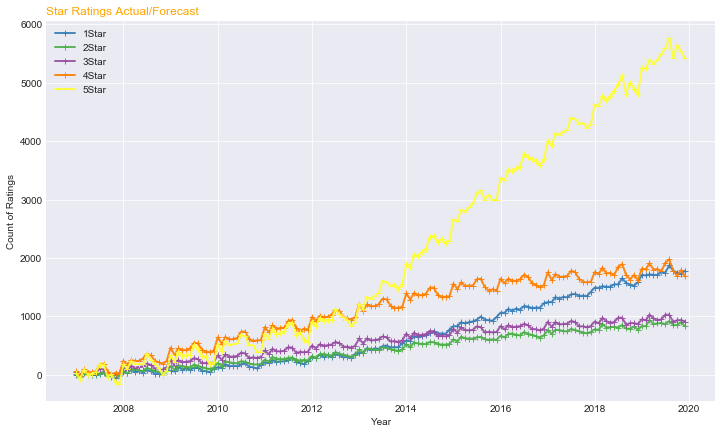
## Model1: Forecasting Reviews

@rich tbd

The process for forecasting reviews.



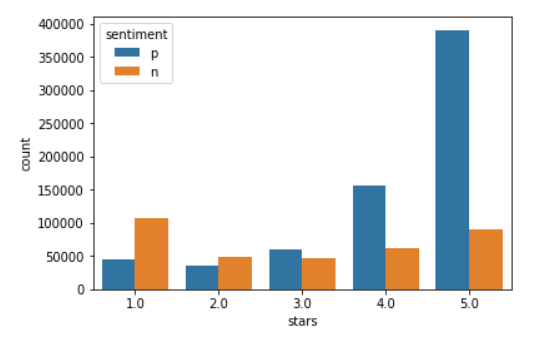




## Model2: Reviews Sentiment Analysis

@tj tbd

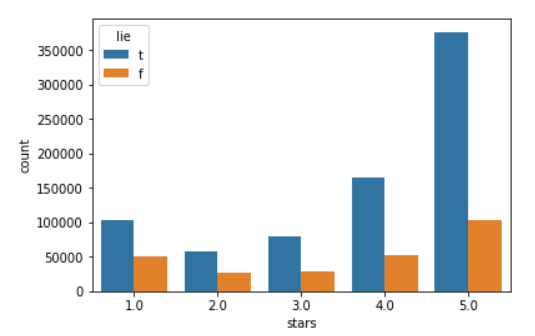
The process that was followed for Review Sentiment Analysis was



## Model3: Reviews Authenticity Analysis

@tj tbd

The process that was followed for Review Authenticity analysis was.



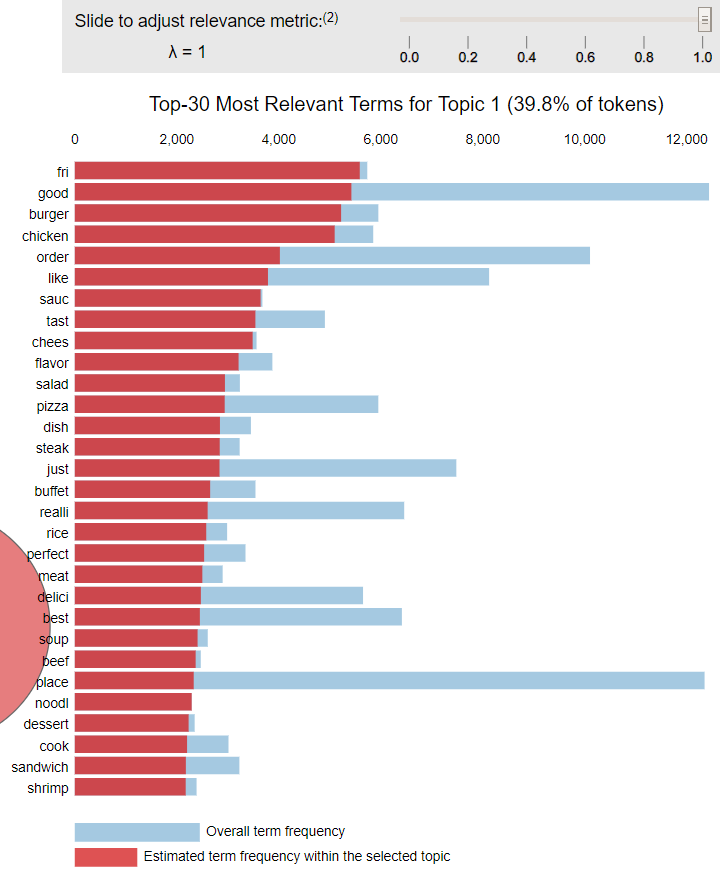
## Model4: Review Topic Modelling

@tj tbd

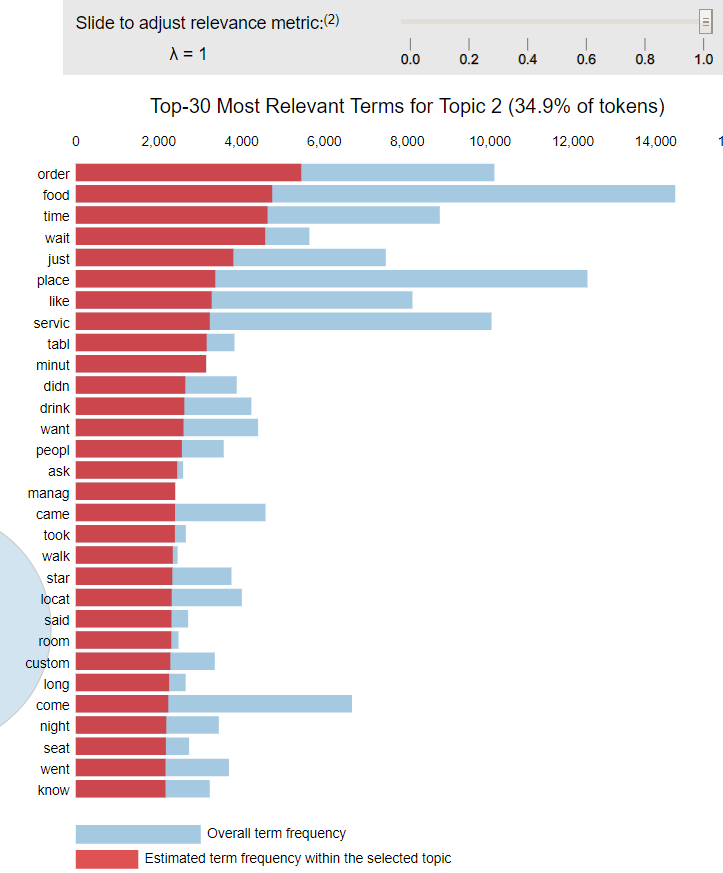
The process that was followed for Review Topic Modelling was…

Topic modelling segmented the reviews into three topics, Food, Service and Experience.

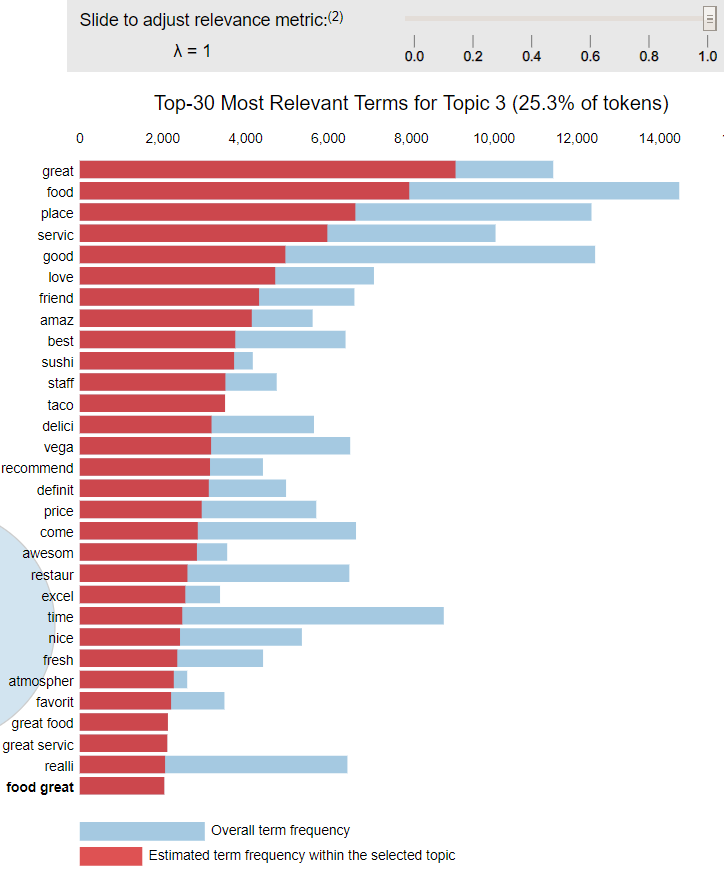
The top 30 most relevant terms for the Food topic are shown in this chart:



The top 30 most relevant terms for the Service topic are shown in this chart:



e top 30 most relevant terms for the Experience topic are shown below:



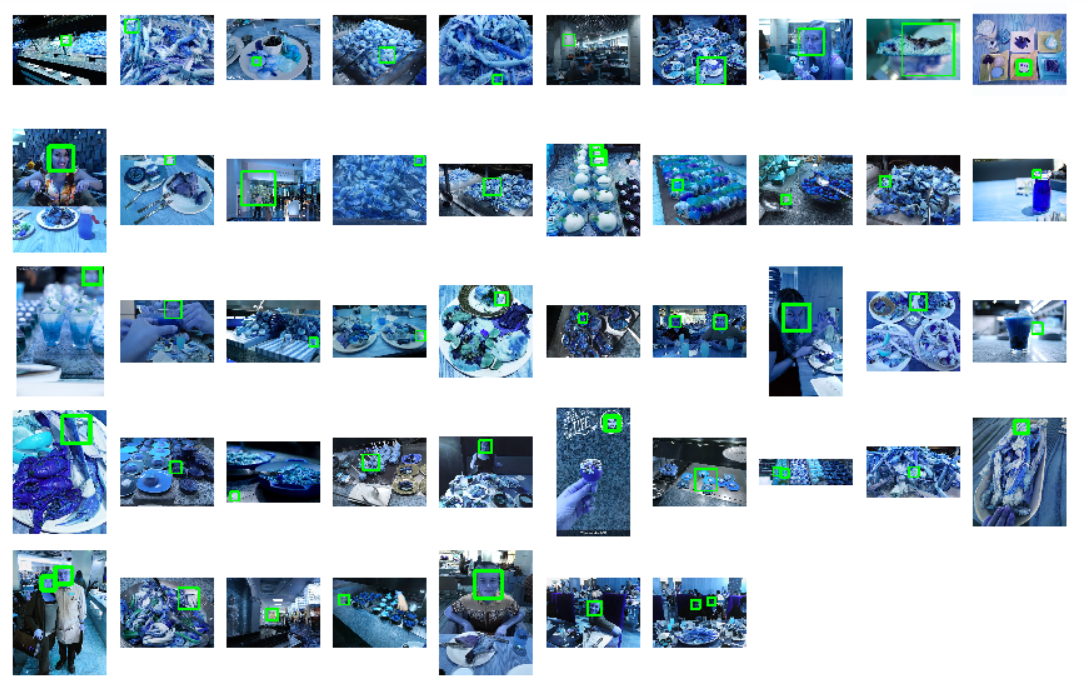
## Model5: Tip Sentiment Analysis

@rich tbd The approach that was followed for Tip Sentiment Analysis

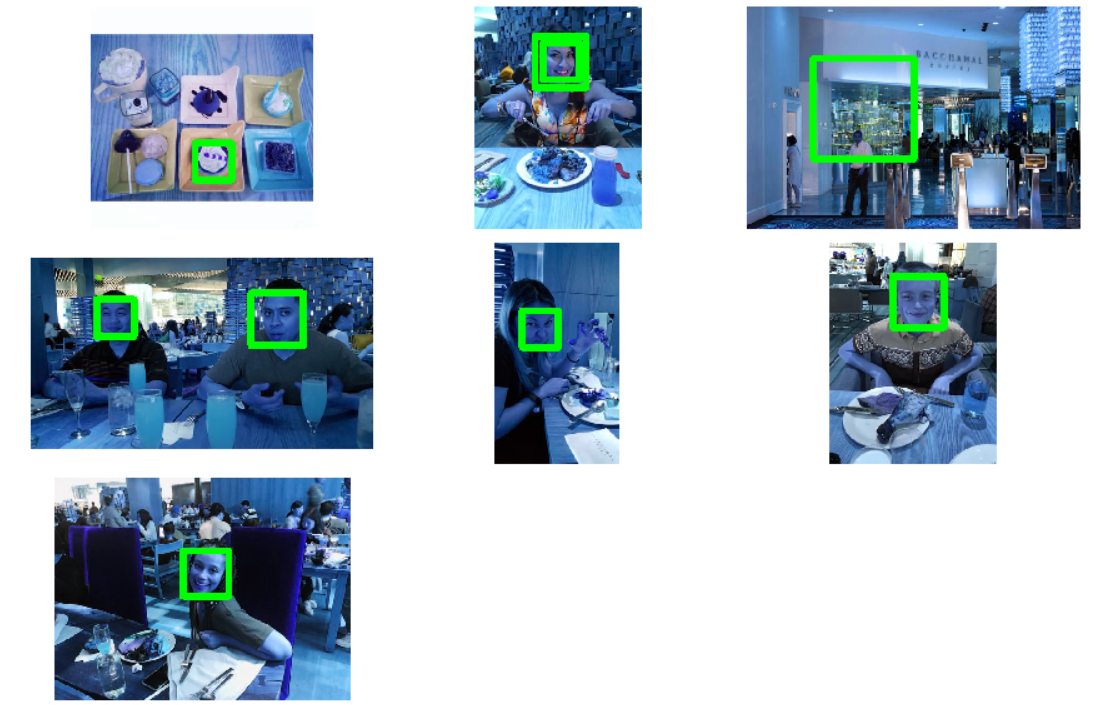


## Model6: Photos with People

@rich tbd



Iteration 2



# Results

## Observation: Review Forecasting shows Seasonality

Four predictions were run to determine what the salary (TotalPay) of the Syracuse

## Observation: Review Sentiment does not match star rating

Tbd

## Observation: Review Authenticity does not match start rating

tbd

# Conclusions

## General

tbd

# Data Dictionary

The main features of the xxx dataset are described in this table

@rich tbd