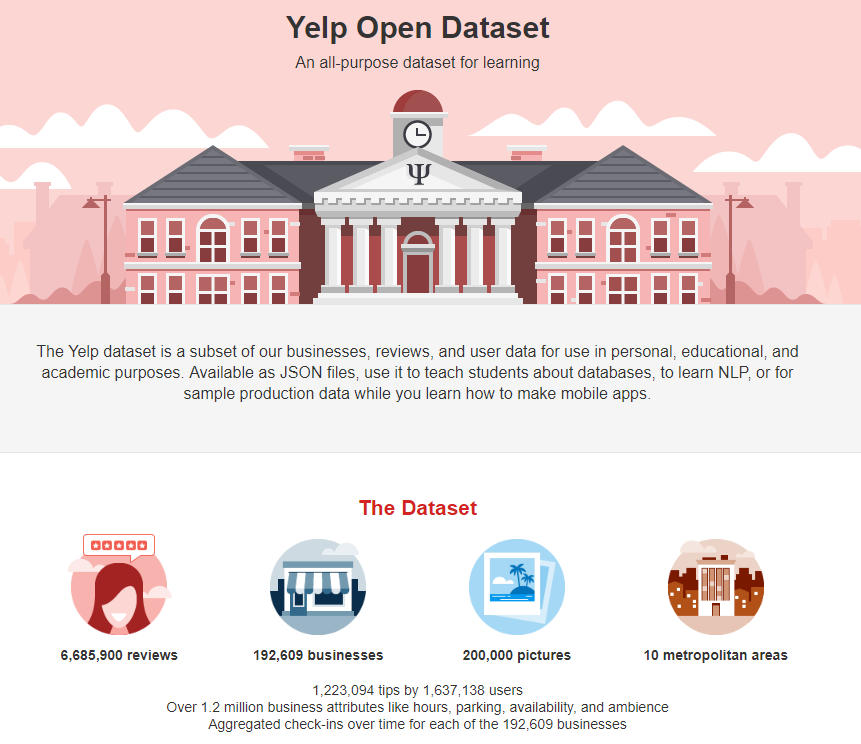
# DRT IST718 Final Project Checkpoint 1 (DRT = Debasis, Rich, TJ)

## Overview

The objective of this Project Checkpoint is for our team to describe the work that has been done on IST718 Final project, to date, and also to describe the work that we have planned.

This document comprises 5 sections, 4 are the requirements for this status report, and 1 is an appendix which describes Project Checkpoint 1.

The sections are as follows:

* Specification
* Observation
* Analysis
* Recommendation
* Appendix

## Specification

Online user reviews of products and services play an important part in the marketing, advertising and success of businesses today.

The goal of our project is to use the Yelp dataset from the Yelp Dataset Challenge to understand customer experience for a subset of businesses and see if we can predict ways that the user/purchaser experience can be improved.

Our objectives are:

1. Integrate social data and business data to the base Yelp data
2. Determine whether we can predict star ratings from the corresponding textual reviews
3. Determine profitability of a business based on the ratings
4. Geographically visualize how demanding customers are

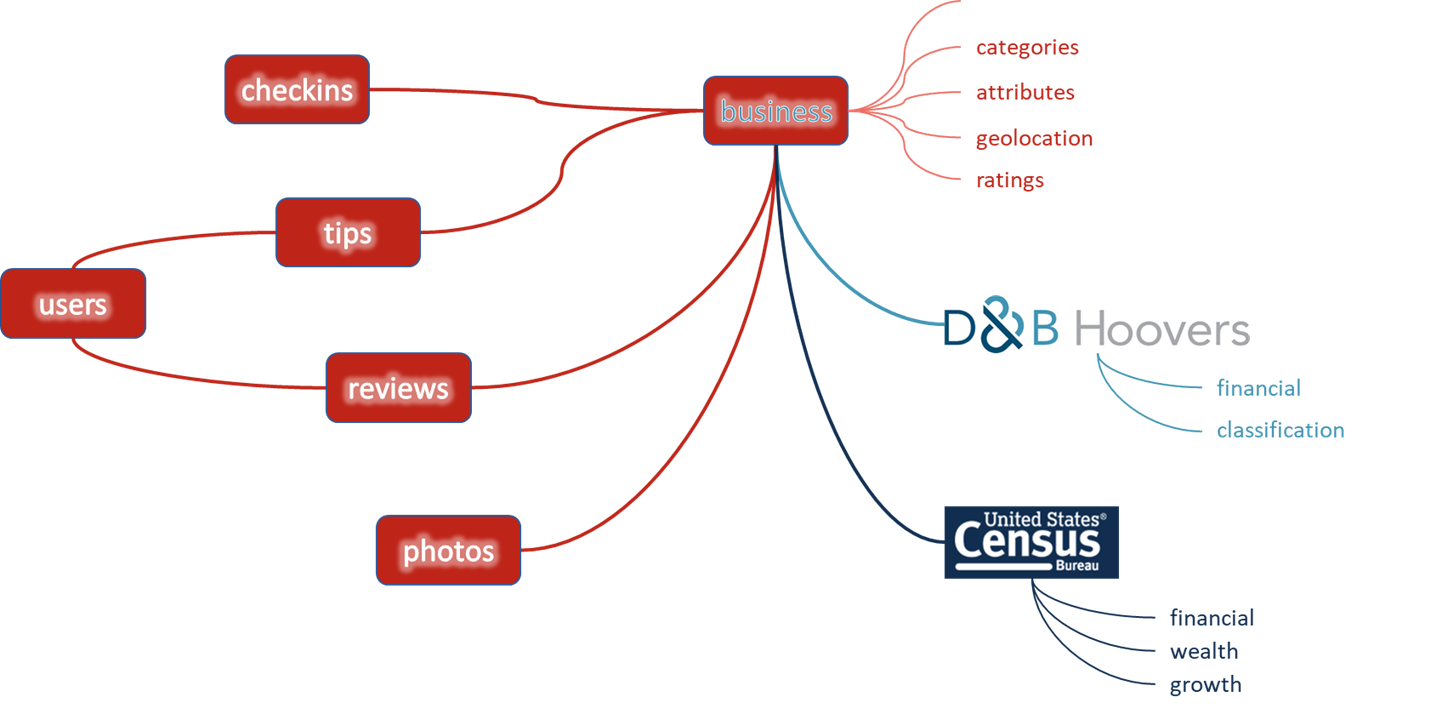
## Observation

The Yelp dataset comprises 6 datasets which we will use for our analysis.

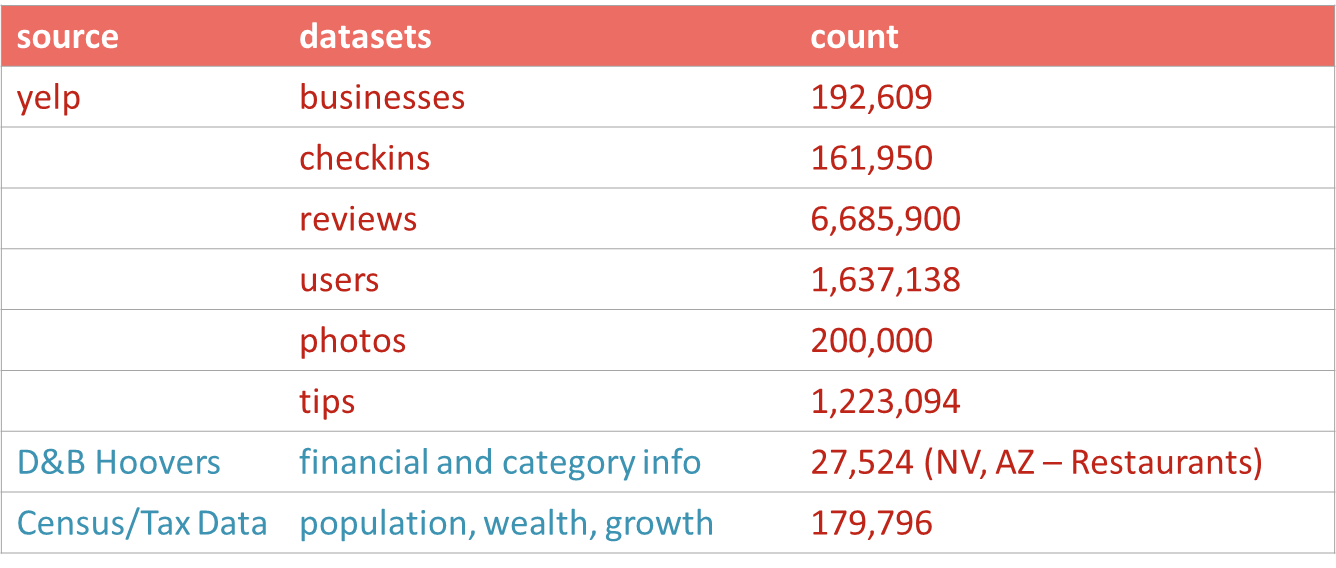
In addition, we will use D&B Hoovers business data to enrich our analysis

And, we will also use US Census data to enrich our analysis.

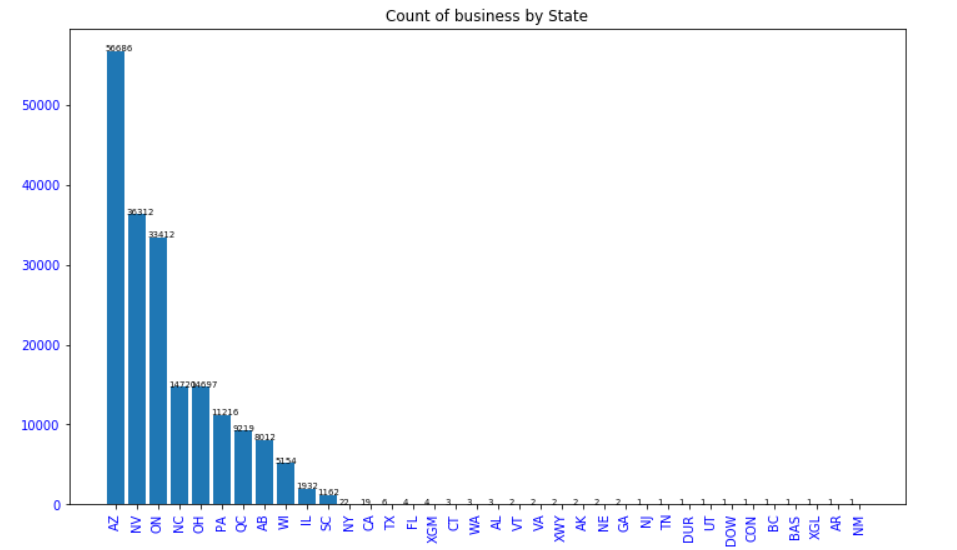
The diagram below provides a pictorial view of our conceptual data model:



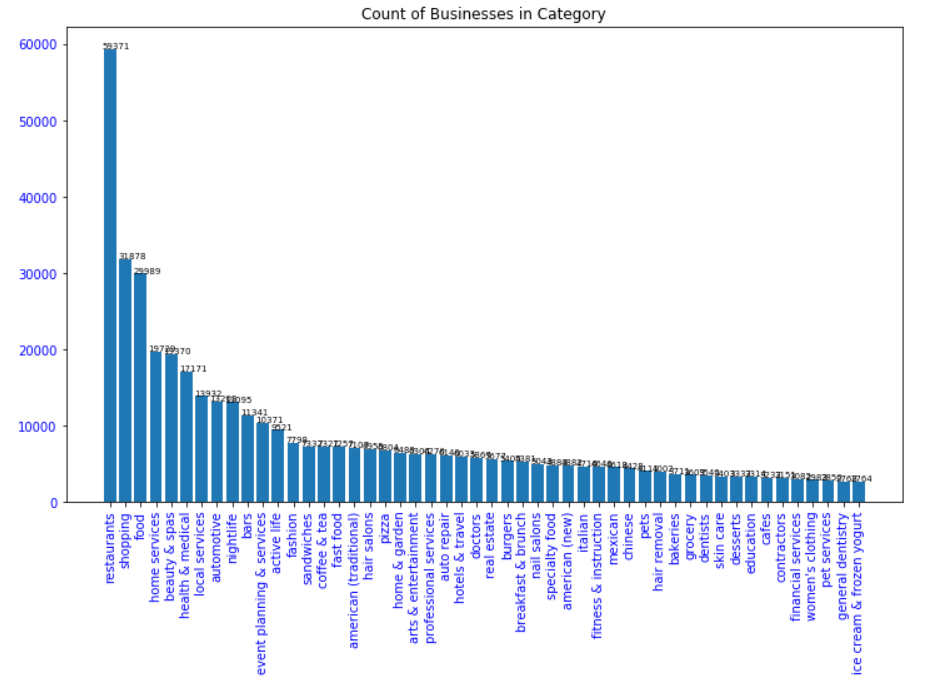
We have conducted initial analysis of the data and observe the following across the datasets of interest:



In addition to the above, we looked more closely at where the Yelp businesses are located, and see the following:



And, we examined what category the business is in, and see the following:



Based on the above, we have decided to conduct our analysis on businesses in NV and AZ only, and we have selected the top 5 categories as well.

In total, we have selected 19,122 business in NV and AZ, that falls into the following 5 categories:

category header IsRestaurants for category ^restaurants which is the top 1 count 5644

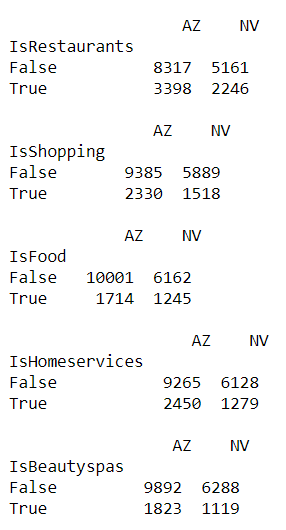
category header IsShopping for category ^shopping which is the top 2 count 3848

category header IsFood for category ^food which is the top 3 count 2959

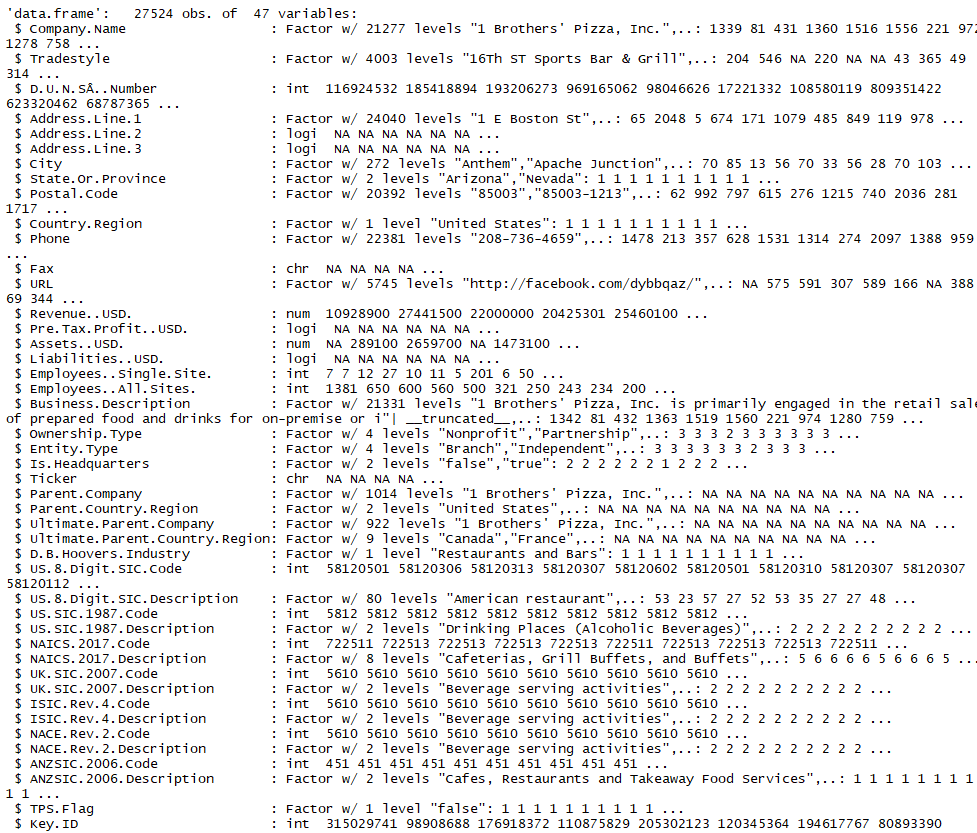
category header IsHomeservices for category ^home services which is the top 4 count 3729

category header IsBeautyspas for category ^beauty & spas which is the top 5 count 2942

We then examined them by state:



From the total population of 6,685,900 reviews, we selected 1,042,847 reviews corresponding to the 19,122 businesses we are interested in.

As for the D&B Hoovers business dataset, we downloaded this data for Restaurant business in AZ and NV. The following figure shows the dictionary and structure of this data:  


For our Census and Tax datasets, we will use data that was acquired as part of our IST718 Lab2 project.

The Data Dictionary for the IRS 2016 Adjusted Gross Income (aka tax) dataset is attached here.



The work that is still required for data understanding and preparation is:

1. Continued reduction (to limit the stats and businesses of interest)
2. Descriptive stats on full sets as well as reduced sets
3. Transformation
4. Data quality
5. Integration
6. Visualization

## Analysis

The hypotheses we are attempting to validate are:

1. Predict sentiment of each review
   1. Model: SVM
2. Verify authenticity of reviews
   1. Model: Naïve Bayes
3. Verify quality and consistency of reviewer
   1. Model: tbd
4. Predict quality of photos
   1. Model: tbd
5. Predict star ratings from reviews
   1. Model: tbd
6. Determine new business opportunities
   1. Model: tbd
7. Segment customers by “persona”
   1. Model: kmeans; decision tree
8. What type of customer will frequent a restaurant
   1. Model: Linear Model
9. Association Rules for Business Attributes
   1. Association Rule analysis
10. Forecast Number of Reviews (in total or by business)
    1. Prophet; Arima

## Recommendation

Based on what we know thus far, which is:

1. We have the data and can access it
2. We know what data preparation work we must do
3. We know the hypotheses we are looking to prove

We anticipate that the outcome of our data analysis and modelling will drive our recommendations.

## Appendix

### DRT IST718 Final Project Checkpoint 1

Online user reviews of products and services play an important part in the marketing, advertising and success of businesses today.

The goal of our project is to use the Yelp dataset from the Yelp Dataset Challenge to understand customer experience for a subset of businesses (tbd) and see if we can predict ways that the user/purchaser experience can be improved.

Our objectives are:

1. Integrate social data and business data to the base Yelp data
2. Determine whether we can predict star ratings from the corresponding textual reviews
3. Determine profitability of a business based on the ratings
4. Geographically visualize how demanding customers are

The data we will use is derived from various sources:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SubjectArea** | **SourcDataset Source** | **Records** | **Attributes** | **Obtainable** | **Notes** |
| Business | Yelp Challenge Dataset | 192609 | 13+ | Y | attribute counts varies per business |
| checkin | Yelp Challenge Dataset | 161950 | 2 | Y | Date attribute has multiple checkins for a business |
| photoix | Yelp Challenge Dataset | 200000 | 4 | Y |  |
| tip | Yelp Challenge Dataset | 1223094 | 5 | Y |  |
| review | Yelp Challenge Dataset | tbd | 8 | Y |  |
| user | Yelp Challenge Dataset | tbd |  | Y |  |
|  | Instagram | ? | ? | Maybe | integrate instagram reviews per business |
| business | DBHoovers | ? | ? | Maybe | integrate business information |
| tweets | Twitter | ? | ? | Maybe | integrate tweets per business |

Technically, we anticipate using a combination of Python and R programming for this work. Python for the heavier data processing and integration processes (likely from within Watson), and R for the geographic visualization (the feature set appears richer, as well as easier to use).

Our project team members (and roles) are:

* Debasis Chatterjee Geo-architect and visualizer
* Richard Paterson Storyboard and Developer
* Tajudeen Abdulazeez Architect and integrator

This project is important to our team as it will give us a deeper understanding of an online customer review process, and it will allow us to determine whether we can identify predictor variables that drive review responses.