Get Yelp

A Comprehensive Analysis of Review Trends

Michael Gigiolio, Dante Pasionek and Dominic Deckys

Question

What lies in our data?

What relationships lie within the different components of Yelp reviews? What patterns exist within presumably unrelated data such as stars, location and hours of business? How does a user’s profile and activity affect the ratings they give?

The Dataset

Reviews

- Review ID
- User ID
- Business ID
- Stars
- Date

Companies

- Business ID
- Name
- Address
- Latitude
- Longitude
- Stars
- Review Count
- Categories

Users

- User ID
- Name
- Review Count
- Average Stars

Process

Exploration

Understanding Data

Determining the content of the datasets and considering how they could be applied. Identifying areas where there may be trends to uncover.

Cleaning

Preprocessing

Making the dataset easy to work with.

- Removing useless columns
- Handling missing values
- Chunking data

Mining

Trend Location

Processing the data and comparing one field to another. Plotting the data types to provide visualization. Narrowing down where trends are located for further investigation.

Tools

Python

- Base code for project
- Easily accessible notebooks
- Easy to import modules

SciPy

- Useful for complicated math and statistics
- K-means clustering

Seaborn

- Logistic Regression
- Graph display

Tools

Folium

- Facilitated mapping
- Easy to plot points and clusters
- Provided intuitive visualization

Matplotlib

- Easy graphing
- Customizable labels
- Allowed different displays of data
- Allowed for advanced visualization

Pandas

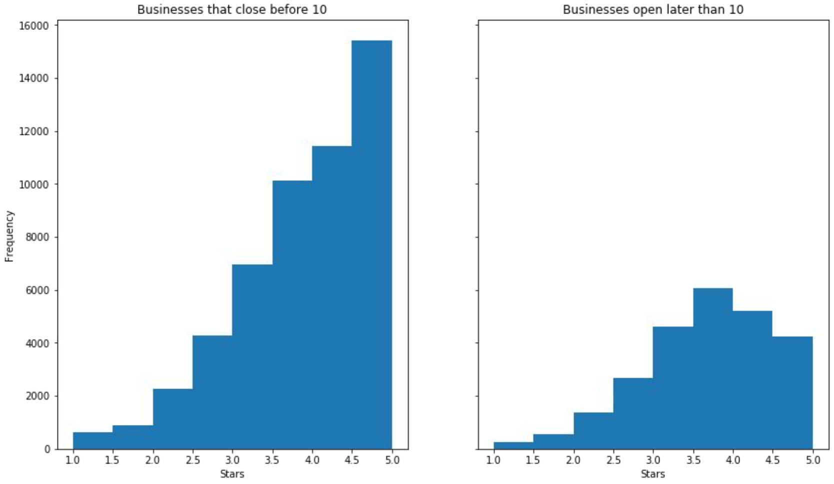
- Allowed easy import of data
- Facilitated data cleaning
- Provided some easy calculation techniques

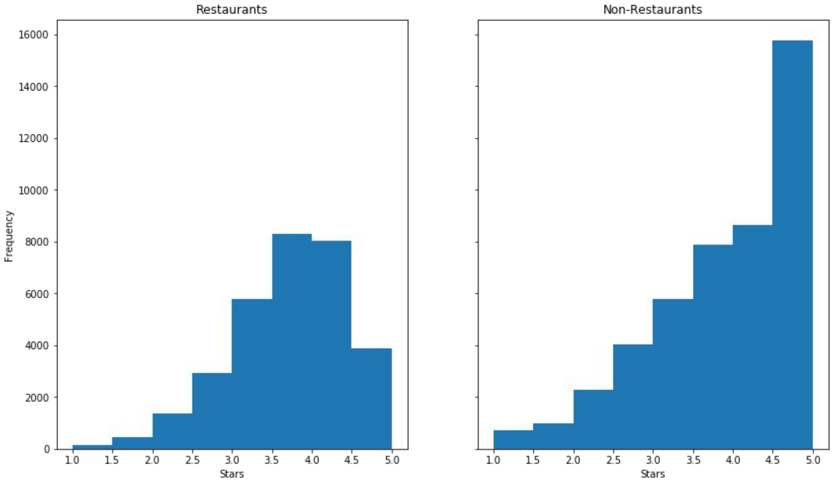
Findings

Rating Frequency

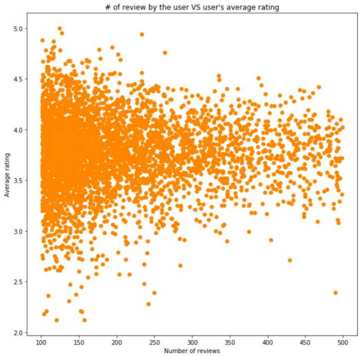
- Businesses that close early fair better
- Most businesses close early
- Businesses open later have a mean closer to the sample average

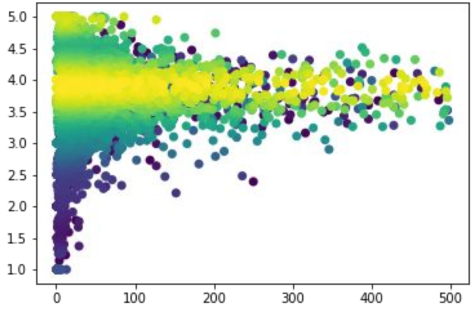
- Restaurants fair worse than non-restaurants
- Very unlikely for restaurants to get 5 stars
- It is likely for non-restaurants to get 5 stars

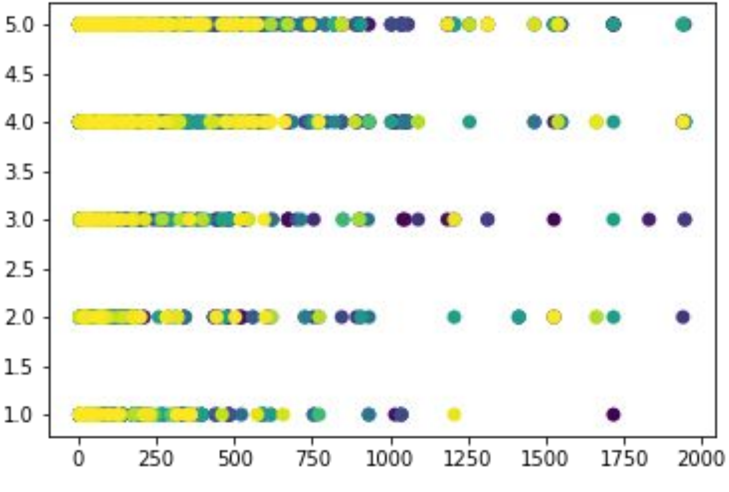




Correlations







Monomodal, negatively skewed data

Global mean of 3.7277

Notable outliers

High density as count grows large around the mean

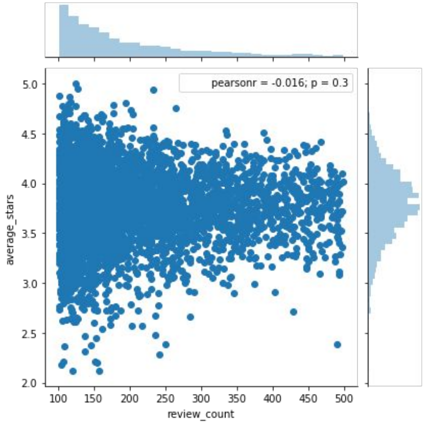
High density of 5s

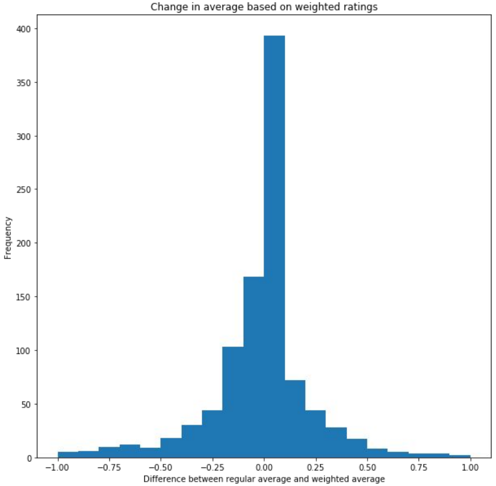
Restaurants given rounded scores

Similar trends to user graphs

Interesting high density outliers

Normalization





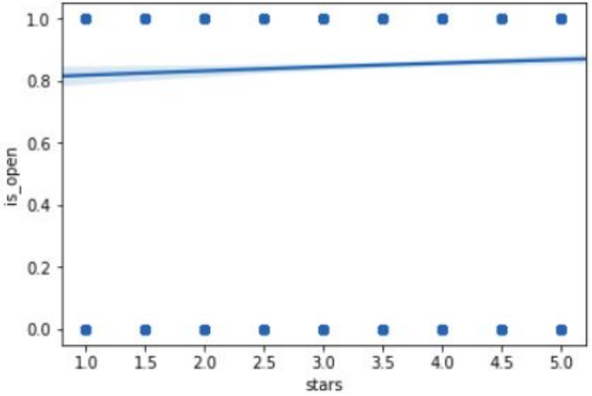
Problems with reviews:

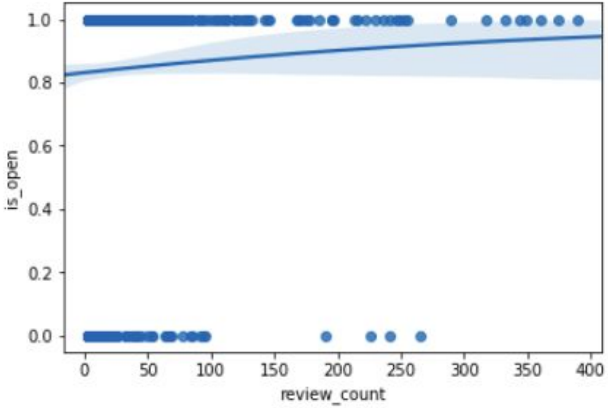
- User bias
- User reliability

Normalization based on:

- User review average
- Number of user reviews

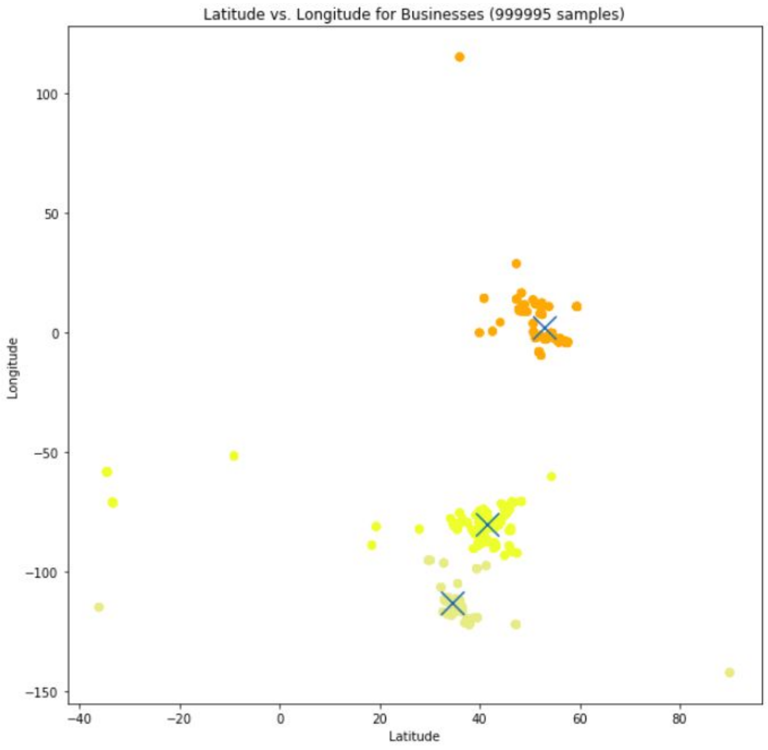
Regression





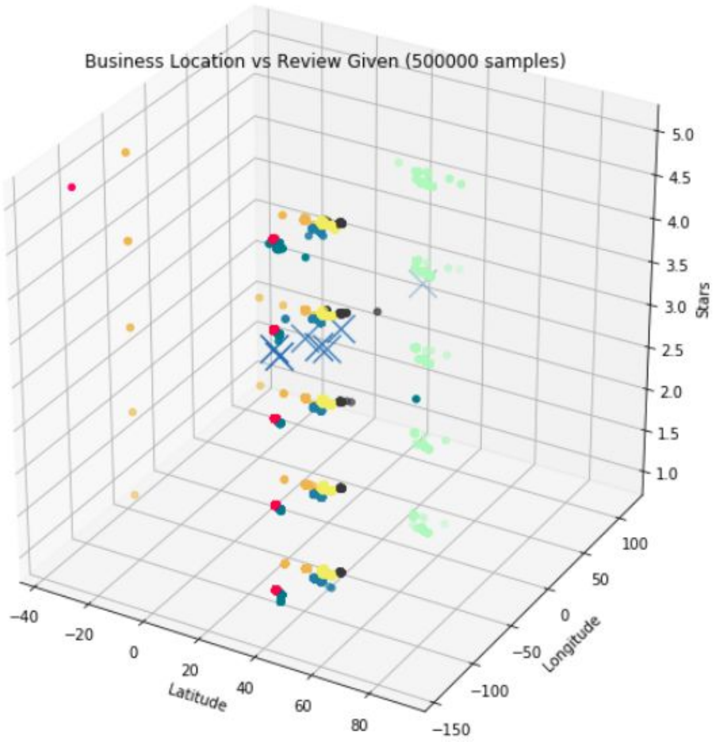
- Logistic regression to determine if Star rating had effect on if a business is currently open
- No effect!

KMeans



- 2D/3D clustering for business location and star reviews
- 3 large clusters across two continents
- Relatively even distribution across location

KMeans cont.



- 2D/3D clustering for business location and star reviews
- 3 large clusters across two continents
- Relatively even distribution across location

Knowledge & Applications

- Restaurants perform worse than other businesses
- Businesses open later do not rate as well
- All businesses receive bad reviews
- Location isn't extremely important

- Creating a more comprehensive review system
- Remove location factor from real estate analyses