

Part 6

Project Presentation

Present your project to the class.

Your presentation should include:

- Project title
- Team members
- Question(s) sought to answer
- Data preparation work
- Tools used
- Classification/clustering/etc applied,
- Knowledge gained
- How that knowledge can be applied.

Presentation tips:

- Use 20-point font minimum
- *Pictures say a thousand words* much easier to convey info than lots of words on each slide
- Make sure we can read your content.
- Everyone must speak during presentation.
- You have only 6 minutes to present.

You will need to do the following:

- Submit a video to Github labeled Group#_Project_Title_Part6_Video.[extension] discussing the topics listed above.
- Submit your slide deck to Github labeled Group#_ProjectTitle_Part6.PDF

Yelp Data Mining

Chris Powell, Jacob Bostick, Donald Boles



What is the Yelp Dataset Challenge?

Yelp Dataset Challenge

Discover what insights lie hidden in our data.



Question(s) sought to answer

- Where should a person open a business based on these reviews in Yelp?
 - Determine the type, location, name, and reviewers

Tools used

- **Programming Language:** Python
- **Libraries:** pandas, plot.ly, matplotlib, geopandas, shapely, folium, numpy, seaborn, StandardScaler, KMeans, PCA, DecisionTreeClassifier, train_test_split, export_graphviz, Image
- **Repository:** github
- **IDE:** Jupyter Notebook

Datasets

- Within the Yelp dataset, we plan on using the reviewer, business, checkin, and user json files.
- The Yelp dataset can be found at <https://www.yelp.com/dataset/challenge>.
- Local download on all team members machines.

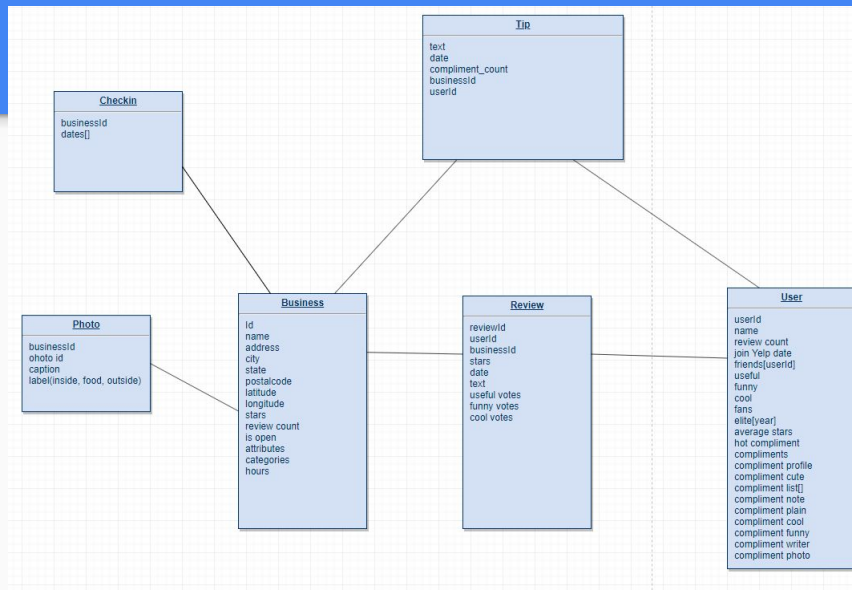
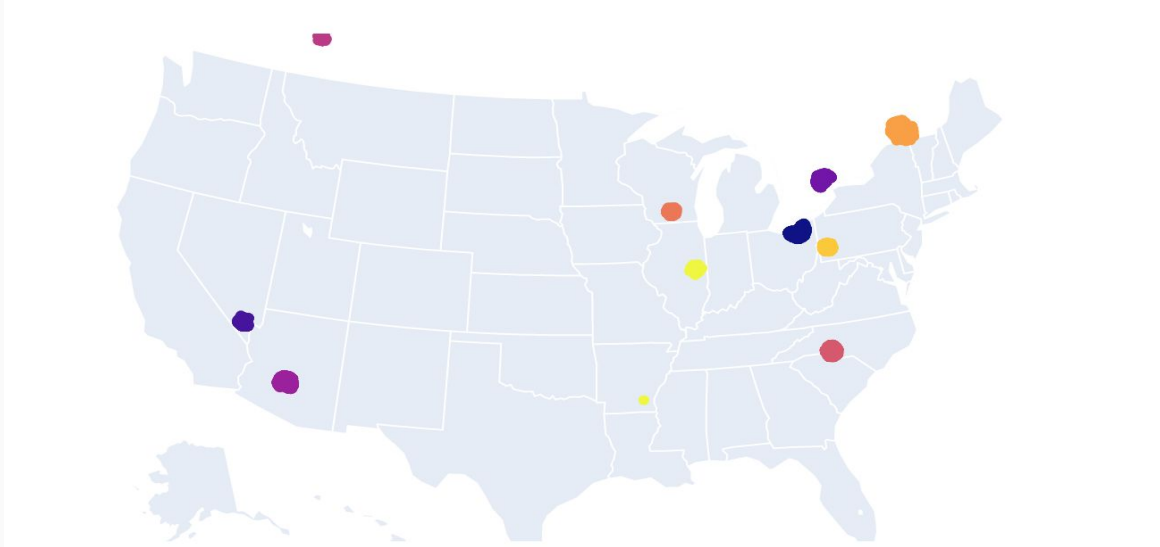


Diagram of json files and attributes

Data preparation work

- Files ranged from .5 gigabyte to 6 gigabytes and contained millions of records and tens of millions of overall data points.
- Merging files using the IDs into pandas dataframes.
- Building a dictionary of all words in business names
- Building a dictionary of all attributes.
- Creating a boolean for the 5 star rated businesses as a column in the dataframe.
- Creating a dataframe with a record for each attribute for the decision tree to learn off of.

K-means Clustering plotly visualization



Choosing a cluster: Nevada?

Open rate of the businesses

- 0.17241500916913532

Average number of reviews

- 35.327955282832555

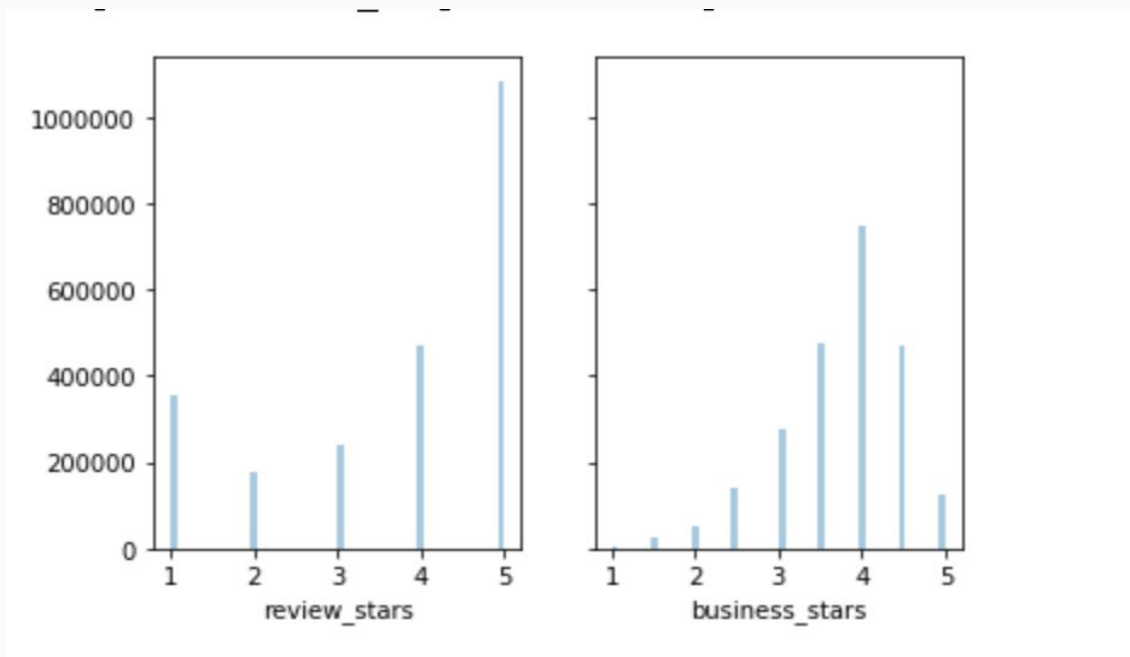
Average rating of businesses

- 3.7071783749471012

Classification/clustering/etc applied

- Naive Bayes classification
 - (used to find the probability of getting 5 stars)
- K-means clustering
 - (used in clumping locations based on geographical region)
- Information Gain using Decision Tree
 - (used to find the highest entropy for all attributes - also used to take a closer look at categories and the business names)

Review Stars vs Business Stars



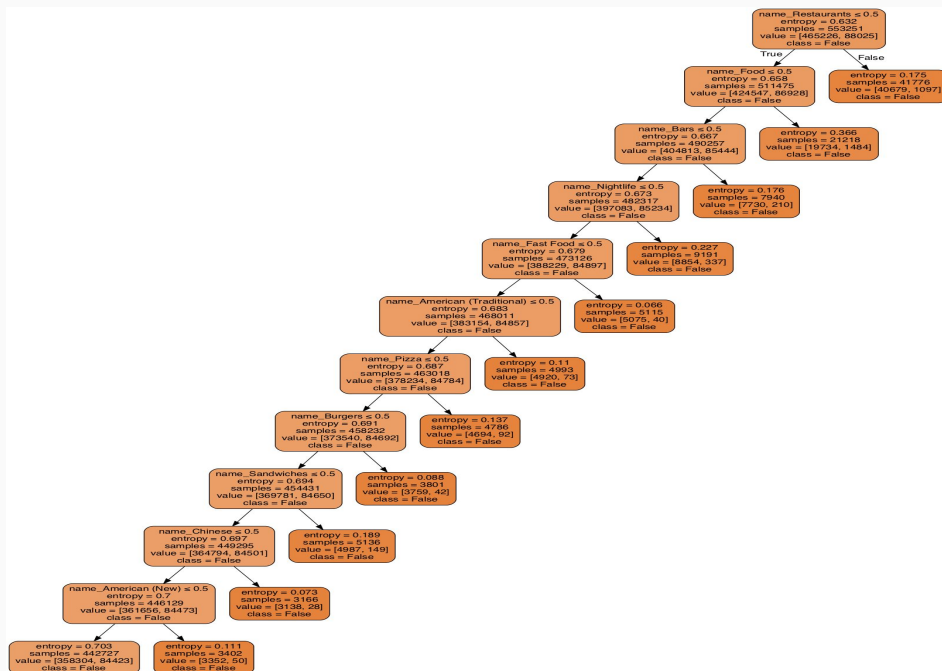
Knowledge gained

- The most likely name of a business with a 5-star rating is 'DDS'. This is a shorthand for doctor of dentistry.
- Restaurants, Food, and Bars were the top attributes of a business with a 5-star rating.

```

graph TD
    A["name_McDonald's ≤ 0.5  
entropy = 0.604  
samples = 136951  
value = [116711, 20240]  
class = False"]
    B["name_Starbucks ≤ 0.5  
entropy = 0.606  
samples = 136379  
value = [116139, 20240]  
class = False"]
    C["name_DDS ≤ 0.5  
entropy = 0.608  
samples = 135638  
value = [115406, 20232]  
class = False"]
    D["name_Subway ≤ 0.5  
entropy = 0.606  
samples = 135362  
value = [115258, 20104]  
class = False"]
    E["name_Pizza Hut ≤ 0.5  
entropy = 0.608  
samples = 134830  
value = [114730, 20100]  
class = False"]
    F["entropy = 0.608  
samples = 134593  
value = [114493, 20100]  
class = False"]
    G["entropy = 0.0  
samples = 237  
value = [237, 0]  
class = False"]
    H["entropy = 0.064  
samples = 532  
value = [528, 4]  
class = False"]
    I["entropy = 0.996  
samples = 277  
value = [148, 128]  
class = False"]
    J["entropy = 0.086  
samples = 741  
value = [733, 8]  
class = False"]
    K["entropy = 0.0  
samples = 572  
value = [572, 0]  
class = False"]

    A -- True --> K
    A -- False --> B
    B -- True --> C
    B -- False --> J
    C -- True --> D
    C -- False --> I
    D -- True --> E
    D -- False --> G
    E -- True --> F
    E -- False --> H
  
```



How that knowledge can be applied

- Knowledge of which business markets rate higher
 - Starting a business
- Knowledge of which businesses need help
 - Which businesses to consult to
- Being a dentist you should:
 - Put your name in the business name. 'Premiere Dentistry with Jacob Bostick DDS'.
 - Sell food at your dentist office.