Yelp Restaurant Reviews & the Weather

April Chung, Kendra Gedney, Yipin Lu, Ratnadeep Mitra

M.S. in Analytics

Introduction

Data Science Questions:

- Does the weather affect people's moods so much that they will rate restaurants lower or higher?

- Do people write more positive, five-star reviews on nice days because they are in a better mood?

- Does good weather lead to people have better experiences at restaurants, and lead to positive reviews?

Data Source:

- Restaurant Details and Reviews from Yelp
- Daily weather data from NOAA
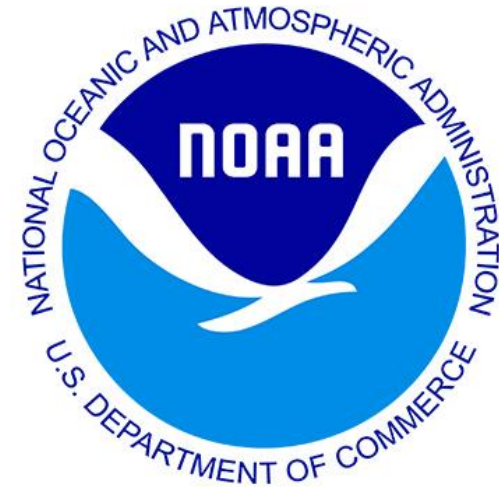
The Data

- Merging:
  - ➢Used *geopy* package to calculate the distances between restaurants and stations in a range of +/- 1 latitude and longitude (~69 miles)
  - ➢Selected the closest station's data that have observed values on the same day as each review
- Cleaning:
  - ➢Checked for outliers, missing values and duplicates
- Binning
  - ➢US Regions
- Final dataset:
  - ➢14,788 reviews from 5,263 restaurants with weather measurements
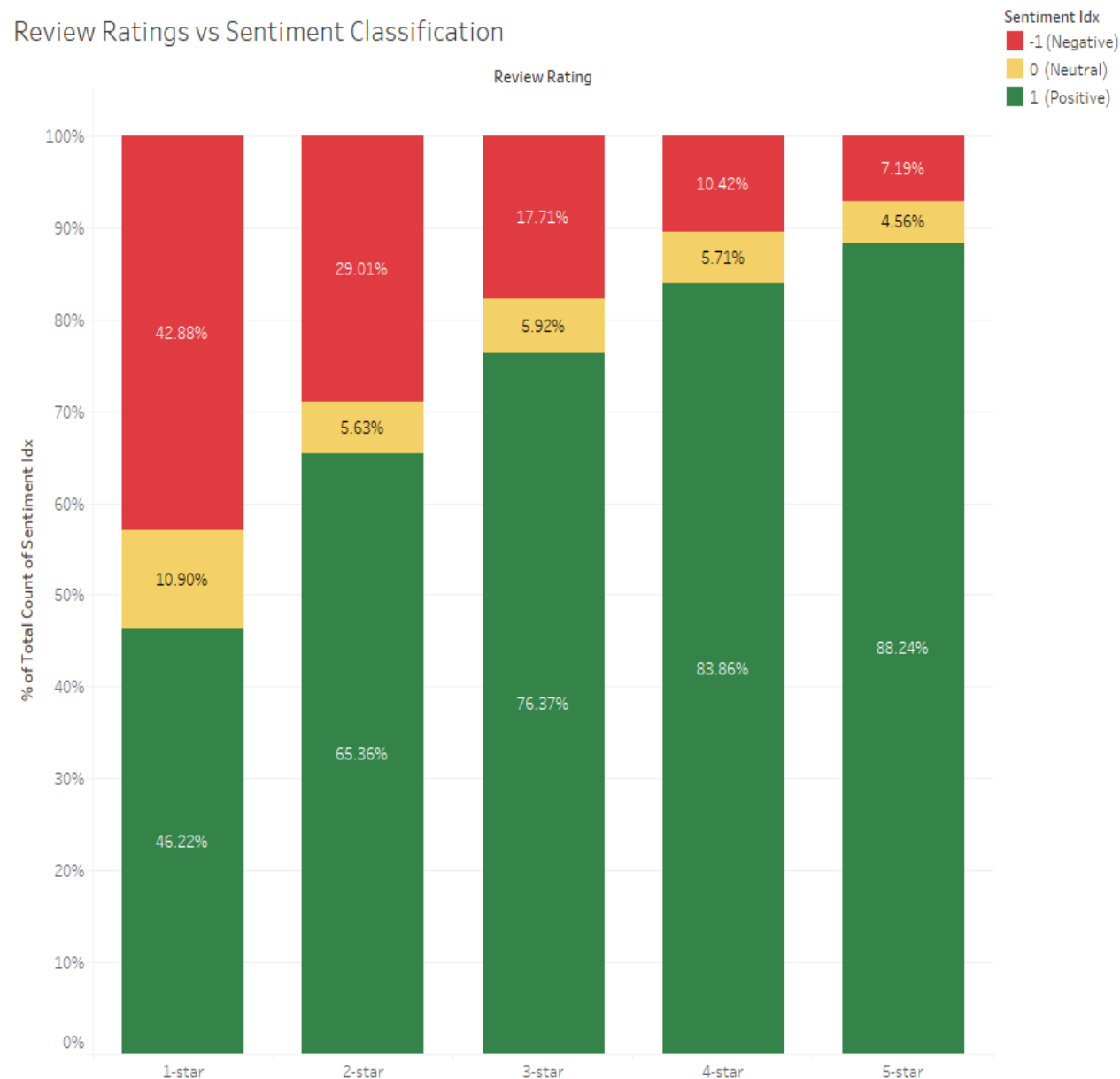
ANALYSIS

PART 1 – Sentiment Analysis

Sentiment Analysis

- Used *NLTK* and *Textblob*

- No strong correlation (0.32) between the computed sentiment and the given star rating

- Limitations –
  - Sarcasm
  - Metaphor
  - Following 1-star reviews were classified as positive –

  "Sitting here, I am absolutely astounded that previous reviews are, mostly, so overwhelmingly positive."

  "The one and only good thing about this place is they have a decent rose."

Review Ratings vs Sentiment Classification

Sentiment Analysis - Word Cloud

- Generated a word cloud based on the frequency of words in all the reviews

- The largest words (and thus most common) tend to be positive, for example, *"good"*, *"great"*, and *"love"*

- This is expected as the majority of reviews in our dataset were classified as positive (83%), and given 5-star reviews (55%)

ANALYSIS

PART 2 – Principal Analysis

Text Mining and Correlation Exploration

- Mined the reviews for weather-related terms. Only **0.7%** of reviews include weather-related terms.

  "What an afternoon of perfection. Great location, great food and awesome weather*! I know the Back Porch isn't responsible for the cool breeze, but the ample.."*

- We do not find strong correlations between temperature and the review ratings or the sentiment. Both correlation coefficients are near **zero**.

Correlation Coefficients

Region	TAVG & review_rating	TAVG & sentiment_polarity
Mountain	0.06	−0.06
SouthCentral	0.04	−0.01
MidAtlantic	0.03	0.02
Southeast	0.02	0.02
NorthCentral	0.01	0.02
Midwest	0.00	0.03
NewEngland	0.00	−0.03
Pacific	0.00	−0.01
Northwest	−0.06	0.06
All Regions	0.02	0.01

Clustering Analysis

- Used both k-means (with k=3) and DBScan (with eps=0.4) clustering analysis to explore the relationship among temperature, sentiment and the location of the restaurant

- Used PCA to visualize results. Although DBScan better identified the outliers, **no clear clusters formed**. This suggests that there is **no strong relationship** among temperature and Yelp reviews.
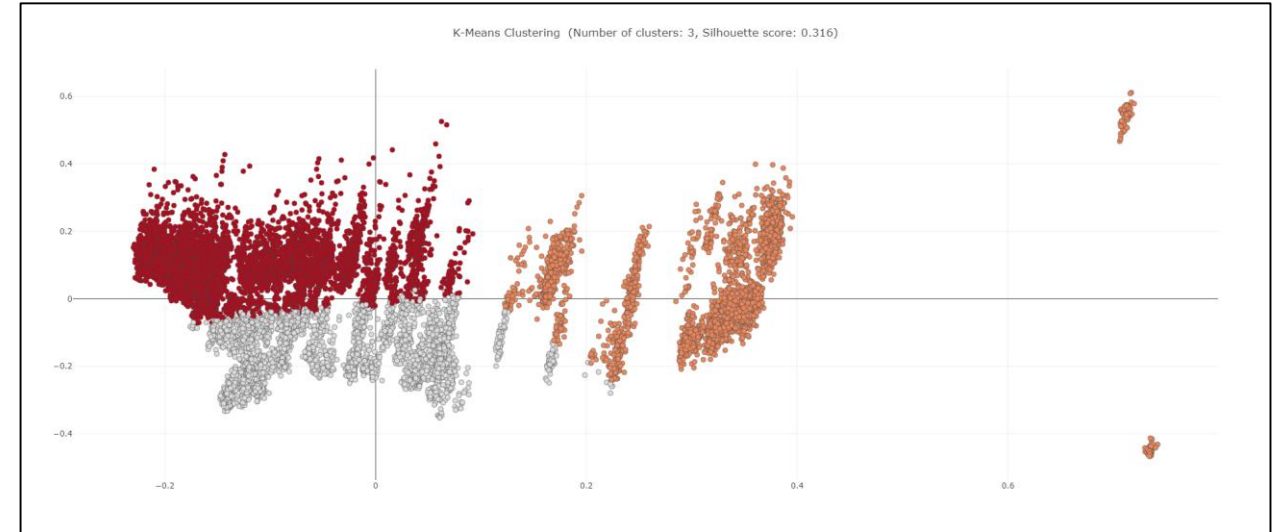

K-means Clustering
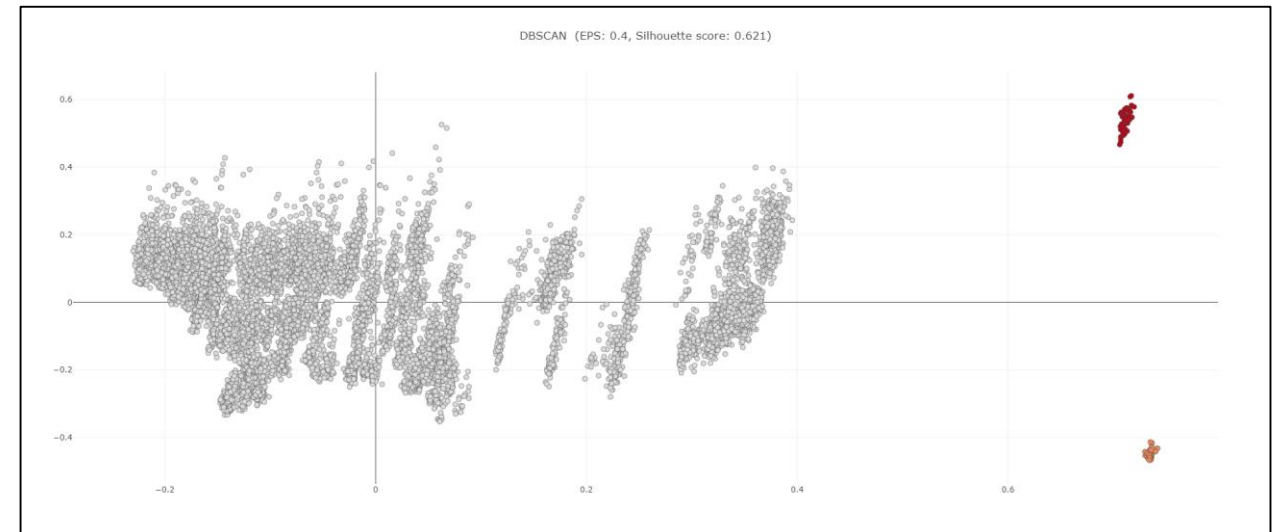
DBScan Clustering

Multivariate Linear Regression

- Conducted multivariate linear regression analysis to explore whether rainy climate or level of temperature would influence customer's sentiment.

- The **random points** in the space and the **low R^2 score (-0.0012)** showed there is no relationship between X (weather attributes) and Y (sentiment_polarity)
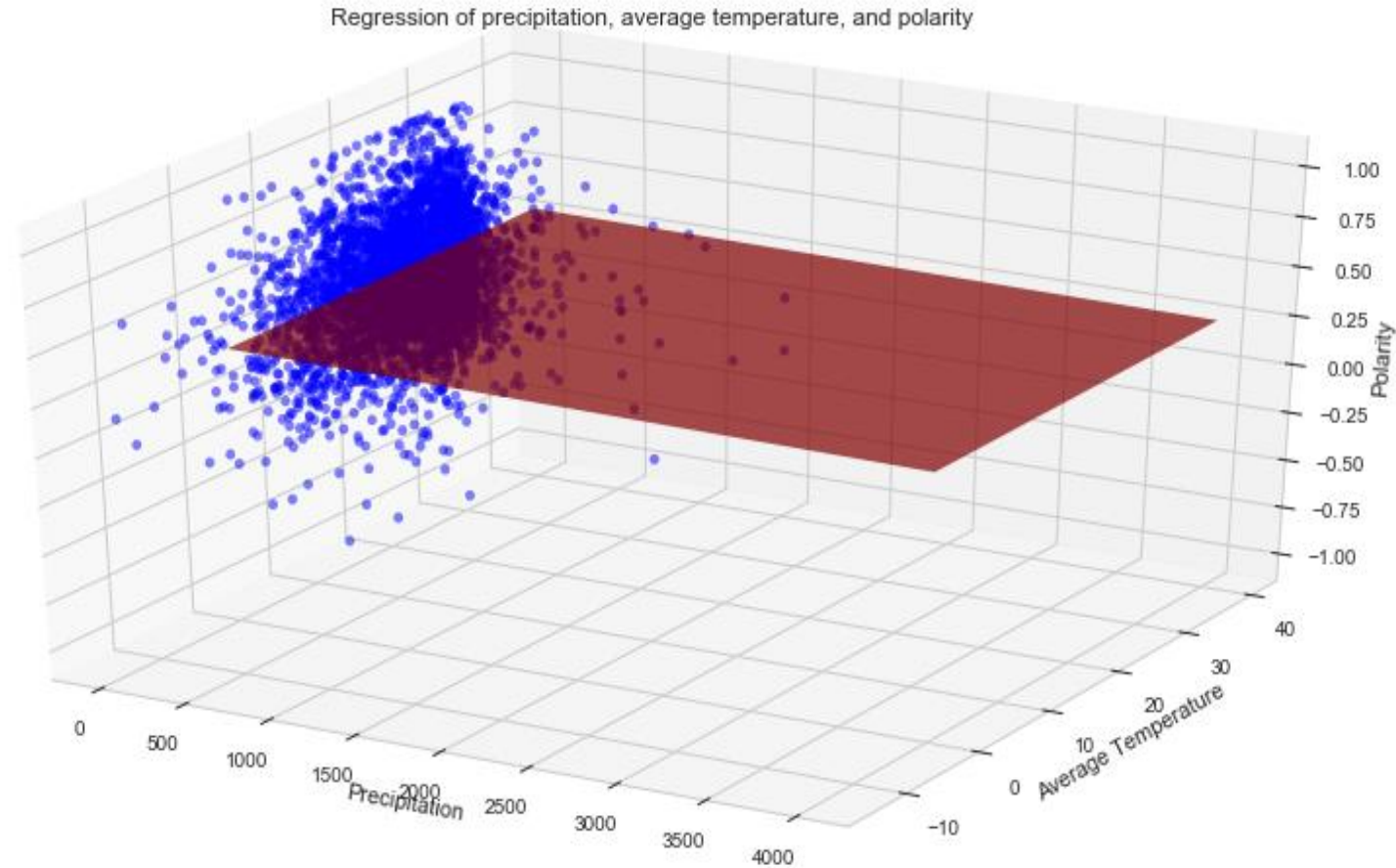

Multivariate Linear Regression

Classifying Reviews based on Weather and Location

- Used **five machine learning** algorithms to test whether we could predict the rating based on location and weather alone:
  - ➢ Grouped the rating into two classes: High (4-5 stars) and Low (0-3 stars)
  - ➢ Applied each machine learning algorithm
- Naïve Bayes and SVM performed the best in training dataset, with **~80%** accuracy

Training Dataset Results		
Method	mean	std
NB:	0.80	-0.01
SVC:	0.80	-0.01
KNN:	0.77	-0.01
RF:	0.76	-0.01
CART:	0.75	-0.01
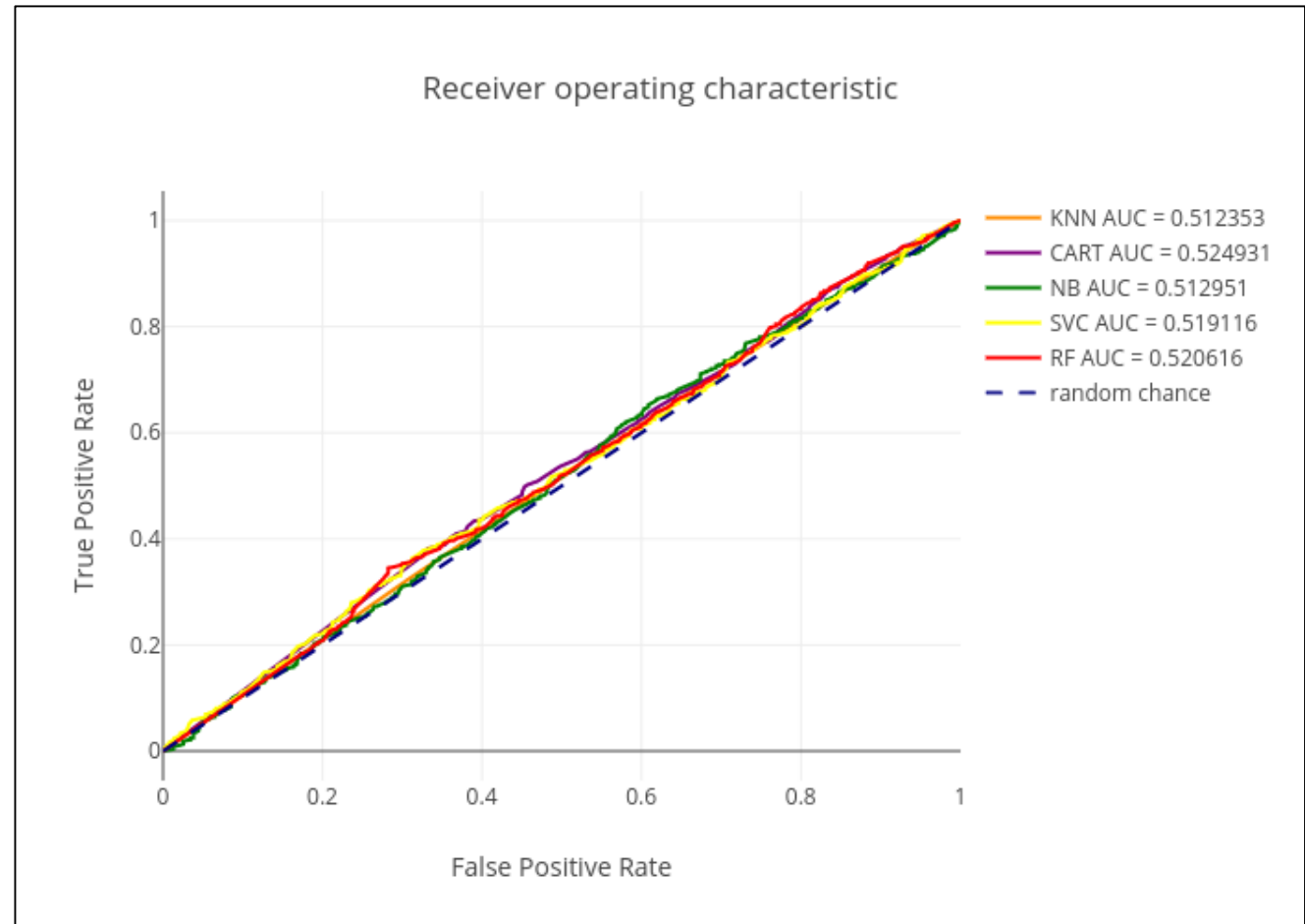
The machine learning algorithms just barely outperforms the dotted line. This indicates that we cannot predict the rating by weather and location alone better than random chance
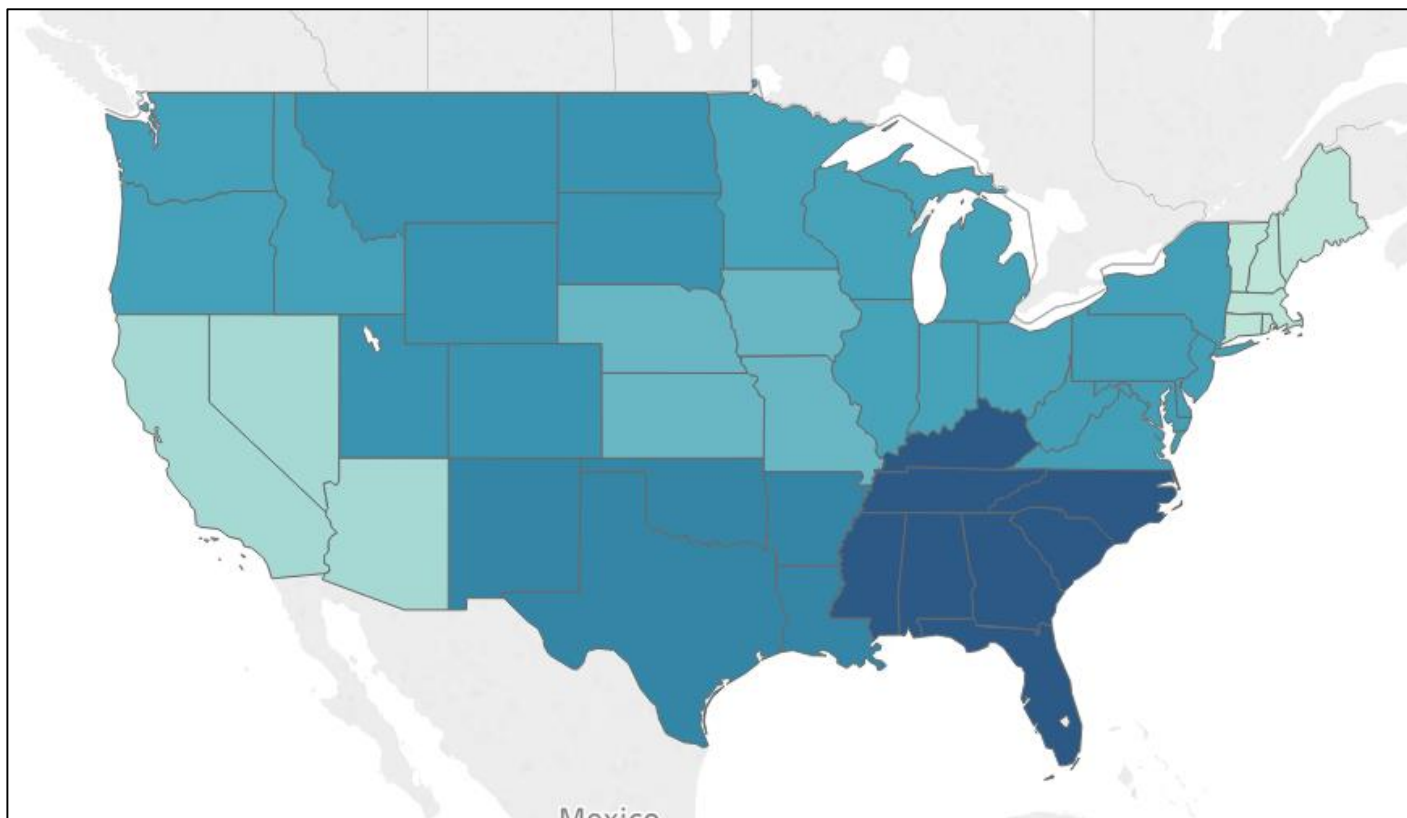
10

ANALYSIS

PART 3 – Regional Analysis

Regional Differences

- Significant difference in **ratings** (p-value: 0.0003)

- Slightly significant difference in **sentiment** (p-value: 0.055)

Average Rating by Region	
Region	Mean *review_rating*
Southeast	4.27
SouthCentral	4.22
Mountain	4.20
MidAtlantic	4.19
Northwest	4.19
NorthCentral	4.19
Midwest	4.16
Pacific	4.12
NewEngland	4.10

Regional Differences

Hierarchical Clustering Dendrogram of Reviews by Region (sihouette score = 0.33)

cosine similarity distance
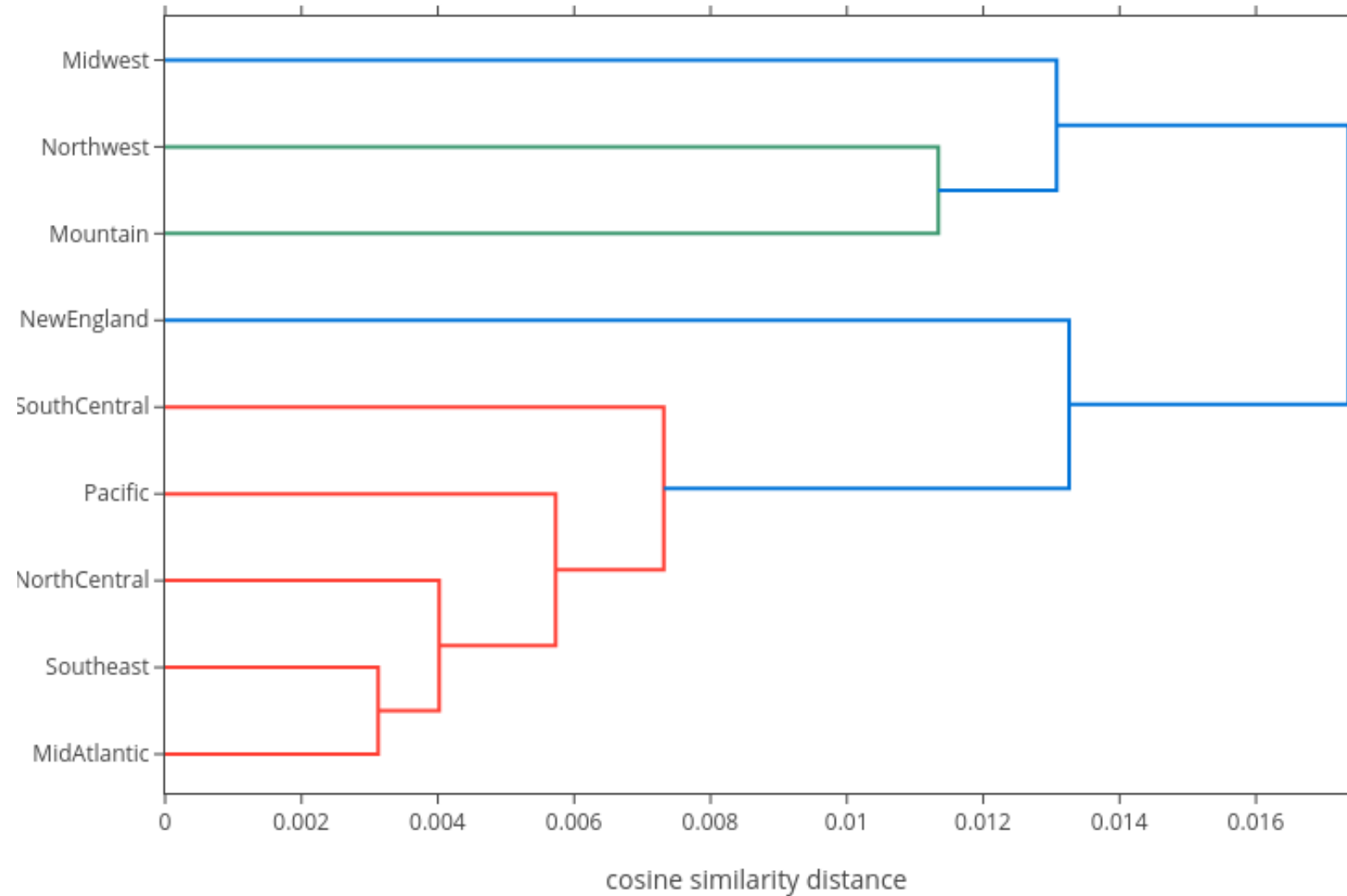
ANALYSIS

PART 4 – Text Analysis

Network Analysis

- Looking into our review data, we wanted to see if there were other attributes such as food items that affect ratings.

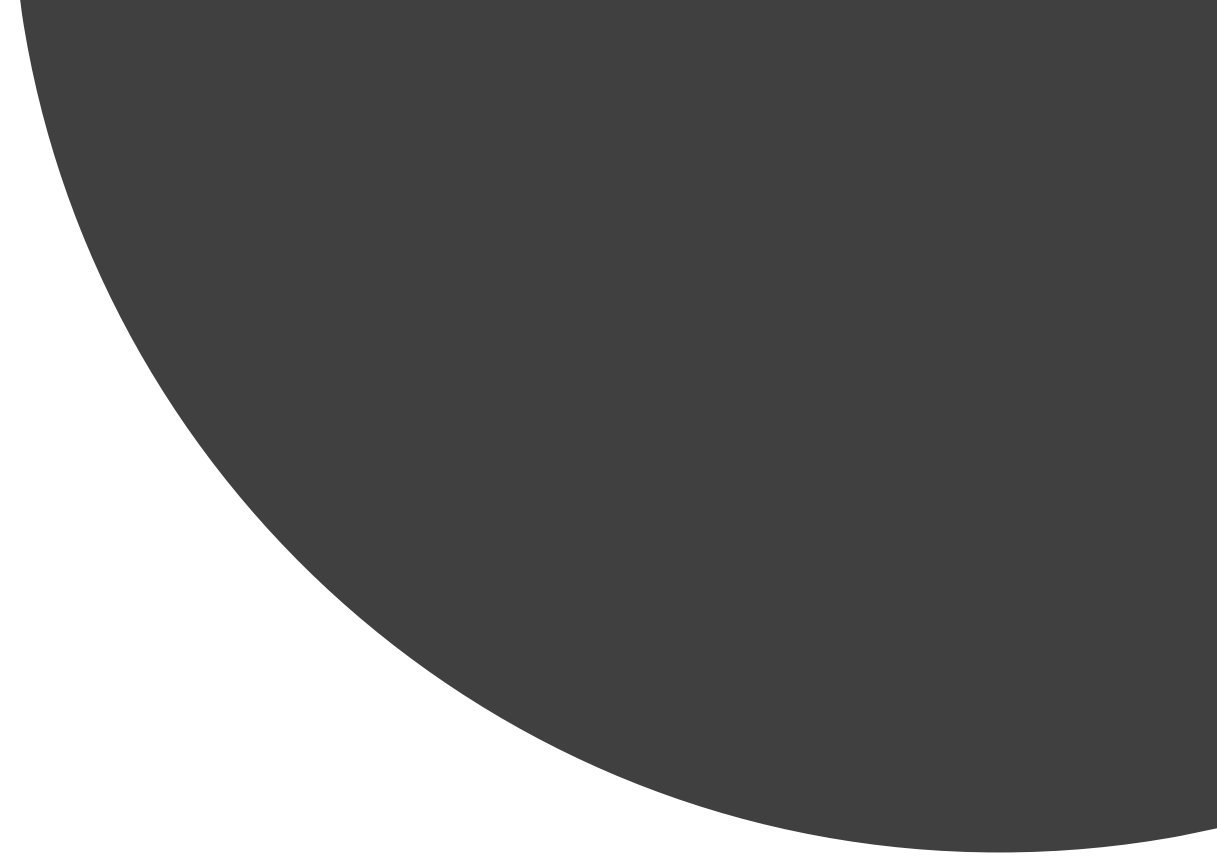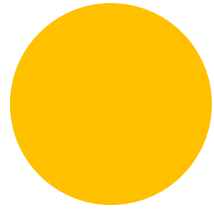- Top popular food items overall were shared across the nation with "chicken", "cheese", and "fries"

Truncated View of Associations on High Rating Reviews			Truncated View of Associations on Low Rating Reviews		
Food 1	Food 2	Co-occurrence Count	Food 1	Food 2	Co-occurrence Count
chicken	steak	26	chicken	fries	11
chicken	pork	22	chicken	cheese	7
pork	brisket	16	fries	fish	6
chicken	cheese	16	chicken	plate	5
cheese	fries	15	meat	plate	5
cheese	steak	15	cheese	meat	5
fish	seafood	15	chicken	pork	5
chicken	fries	11	chicken	side	5
chicken	beef	10	chicken	shrimp	5
shrimp	crab	10	chicken	fish	4
chicken	shrimp	10	cheese	fries	4
			cheese	bread	4
			pork	plate	4

Topic Modeling

- Used Latent Dirichlet Allocation (LDA) to create top 10 topics from Yelp reviews
- **Topic 5** is mostly composed of specific food items: *chicken, cheese, salad, burger*
- **Topic 7** has many atmosphere-related words without any food words: *atmosphere, cute, seating, inside, bar*
- **Topic 8** has the most service-related words: *service, friendly, staff, attentive*

Limitations

- Well-known challenges faced by sentiment analysis
  - ➢ Sarcasm
  - ➢ Metaphor

- Restrictions of the Yelp API
  - ➢ Only 3 reviews per restaurants
  - ➢ Returns only excerpts of reviews

Thank you! | Questions?