YELP DATA ANALYSIS

Stat 628 Module 2 [1/2]

Kai Wang, Yichen Shi, Yanran Wang

Group 4, Mon

TABLE OF CONTENTS

1. **Background and Overview**

2. **Further Analysis**

   — Time-space analysis

   — Text analysis

BACKGROUND AND OVERVIEW

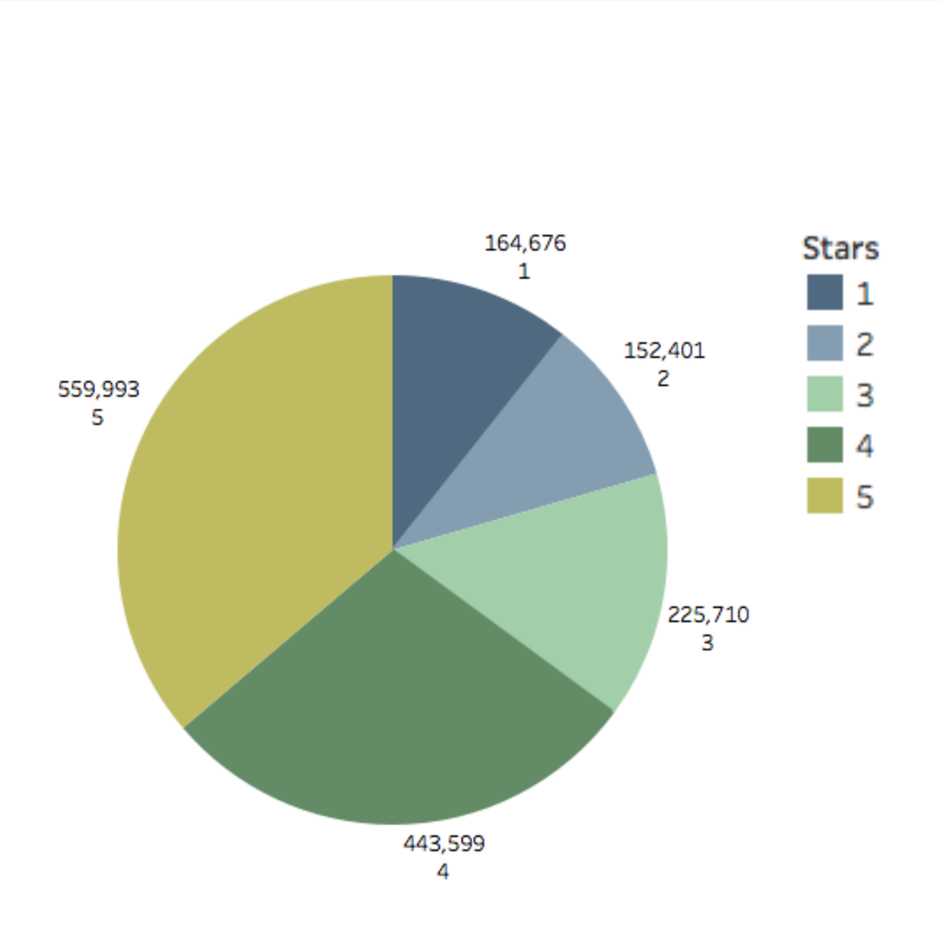
BACKGROUND AND OVERVIEW

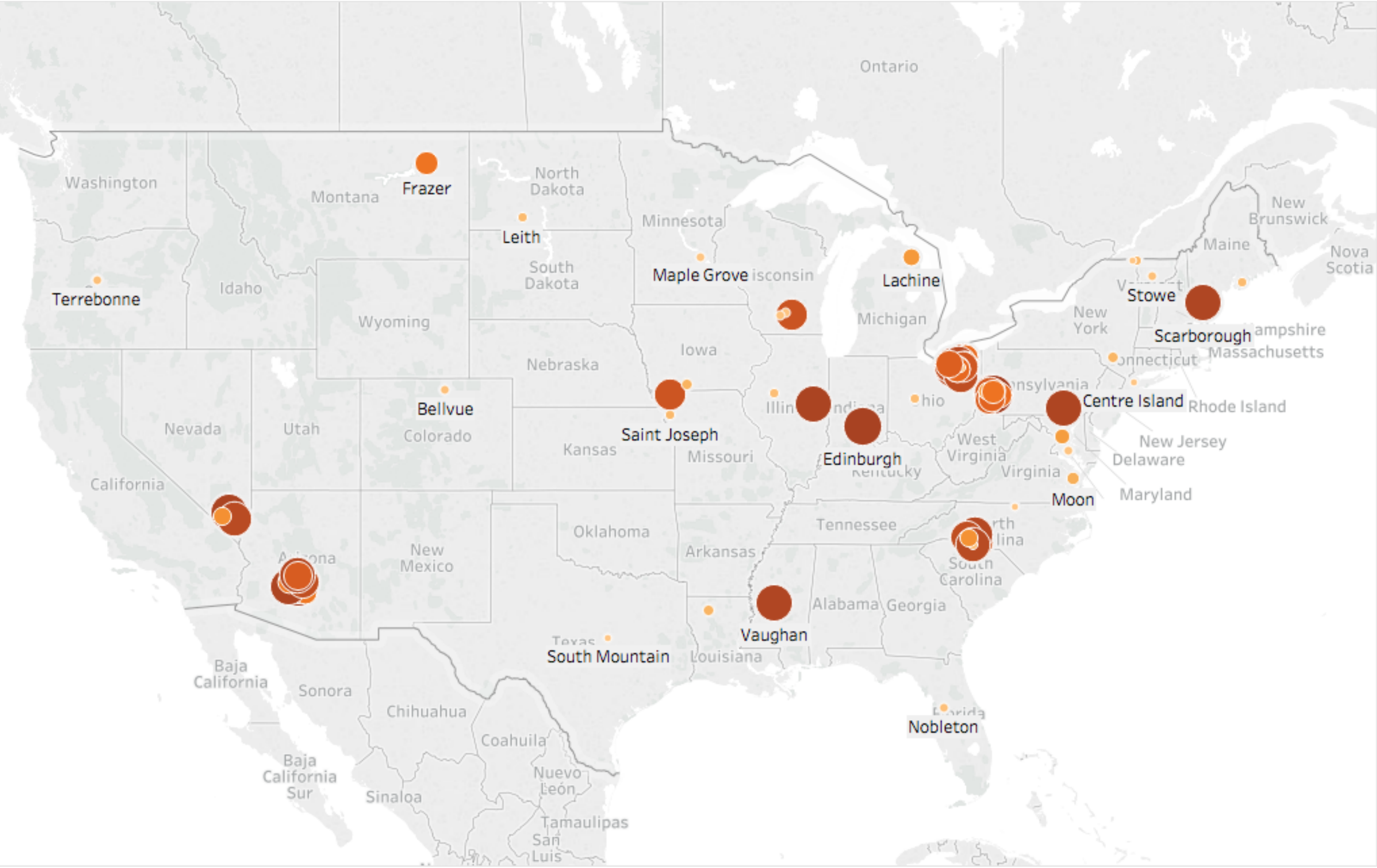
Goal:

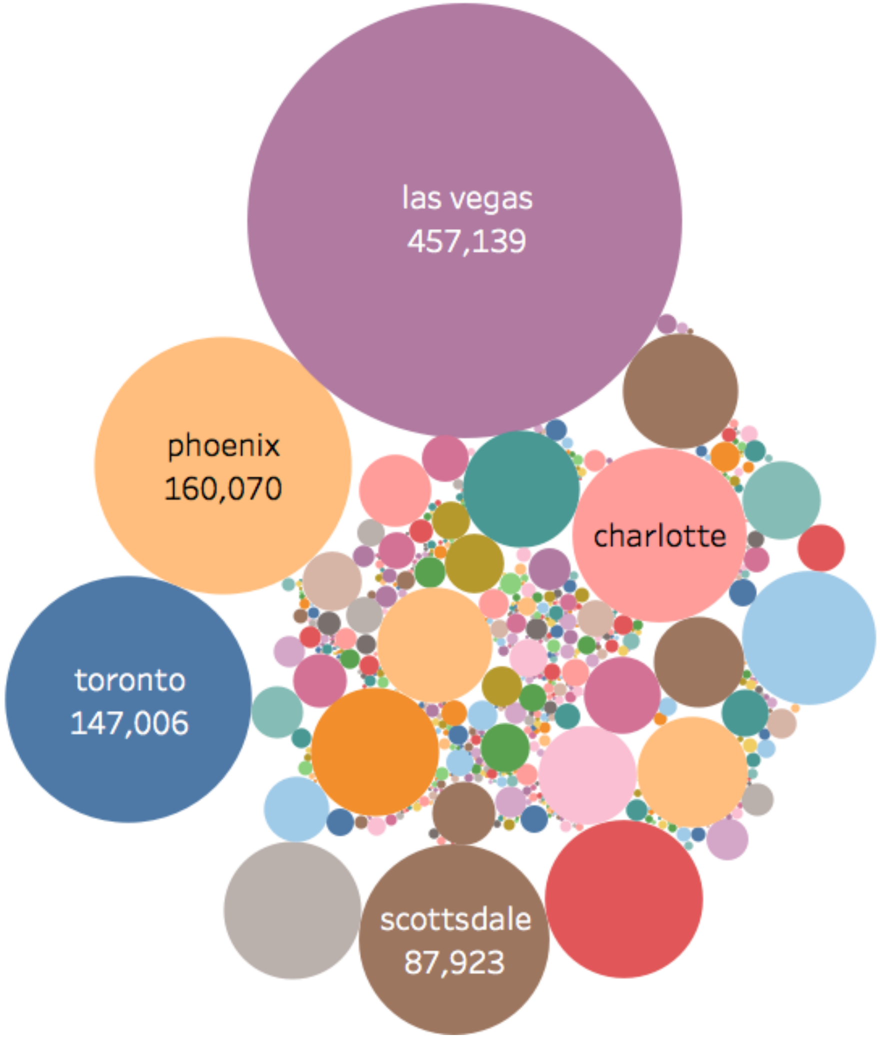
1. Find out WHAT makes a review positive or negative
2. Propose a model to PREDICT the ratings of reviews

Data Background:

1. About 1.5 million reviews with features on Yelp
2. Include stars, id, name, text, data, categories, city, etc.

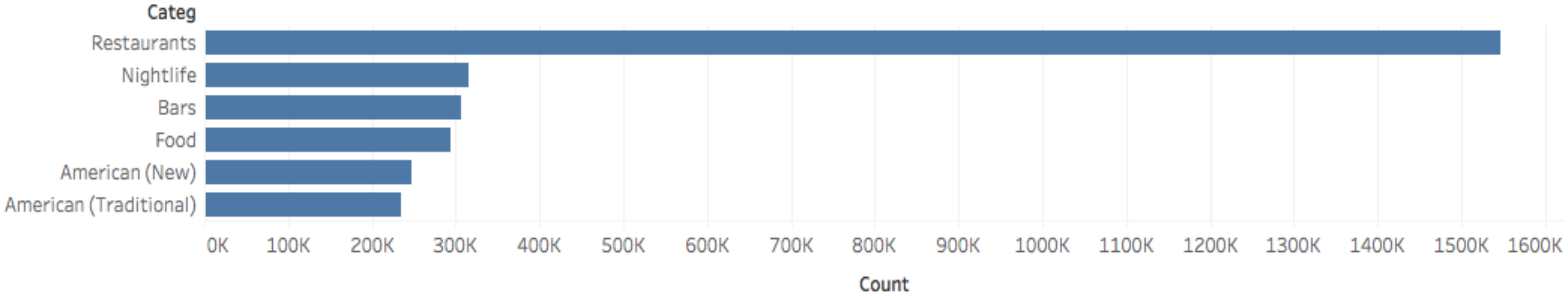


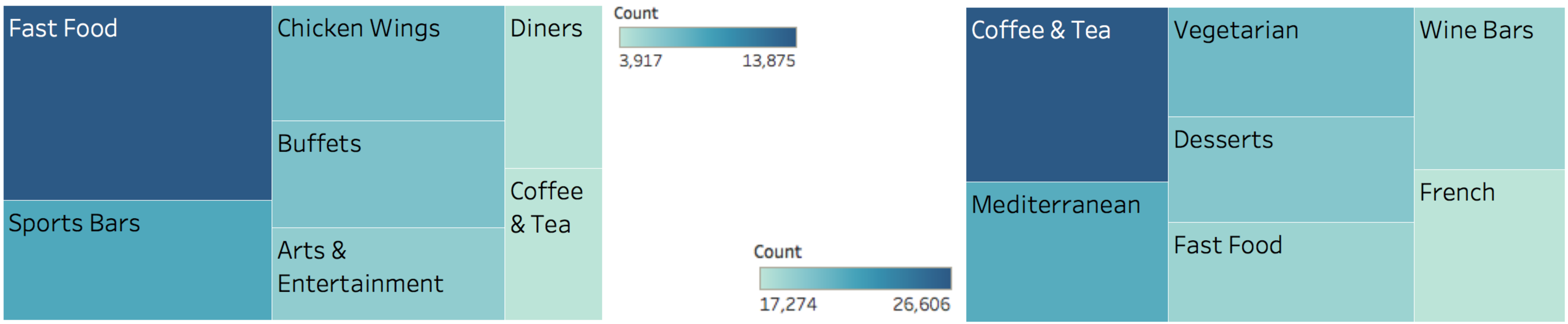




-- CITY --

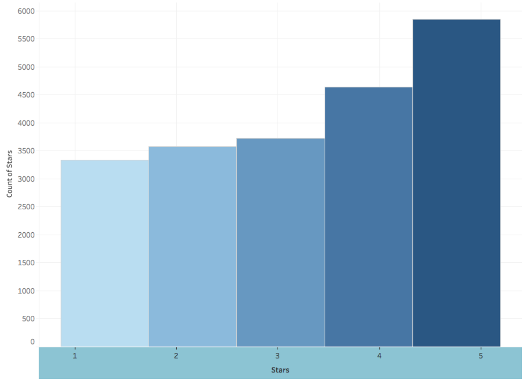
659 cities in total, Mostly ‘Las Vegas’



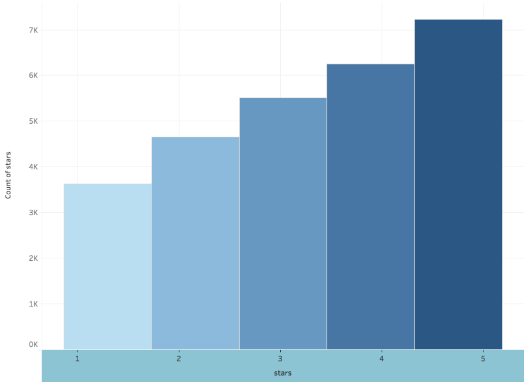


-- CATEGORIES --

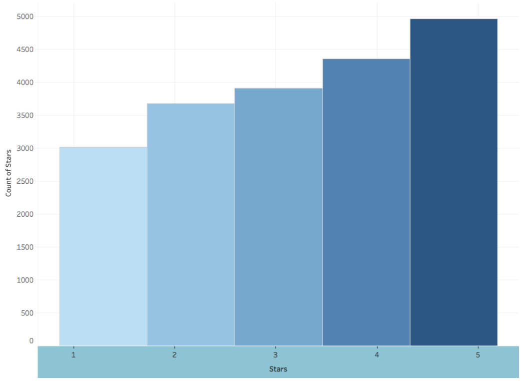
608 categories in total, Mostly 'Restaurants' [fig.1]
Top categories of star 1 and star 5 after removing duplicated part [fig. 2&3]



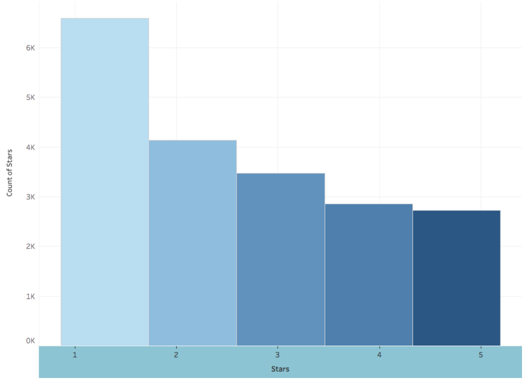
Mediterranean



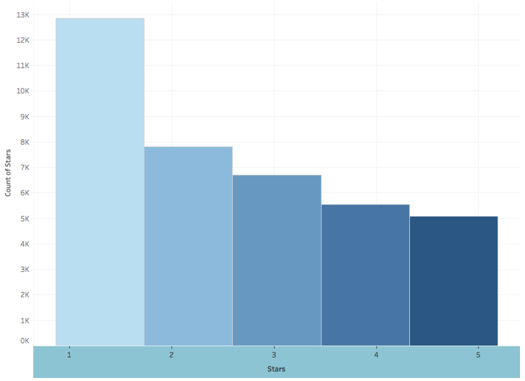
Coffee & teas



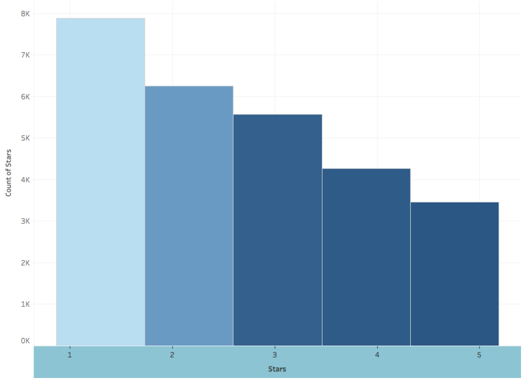
Wine bars



Chicken wings



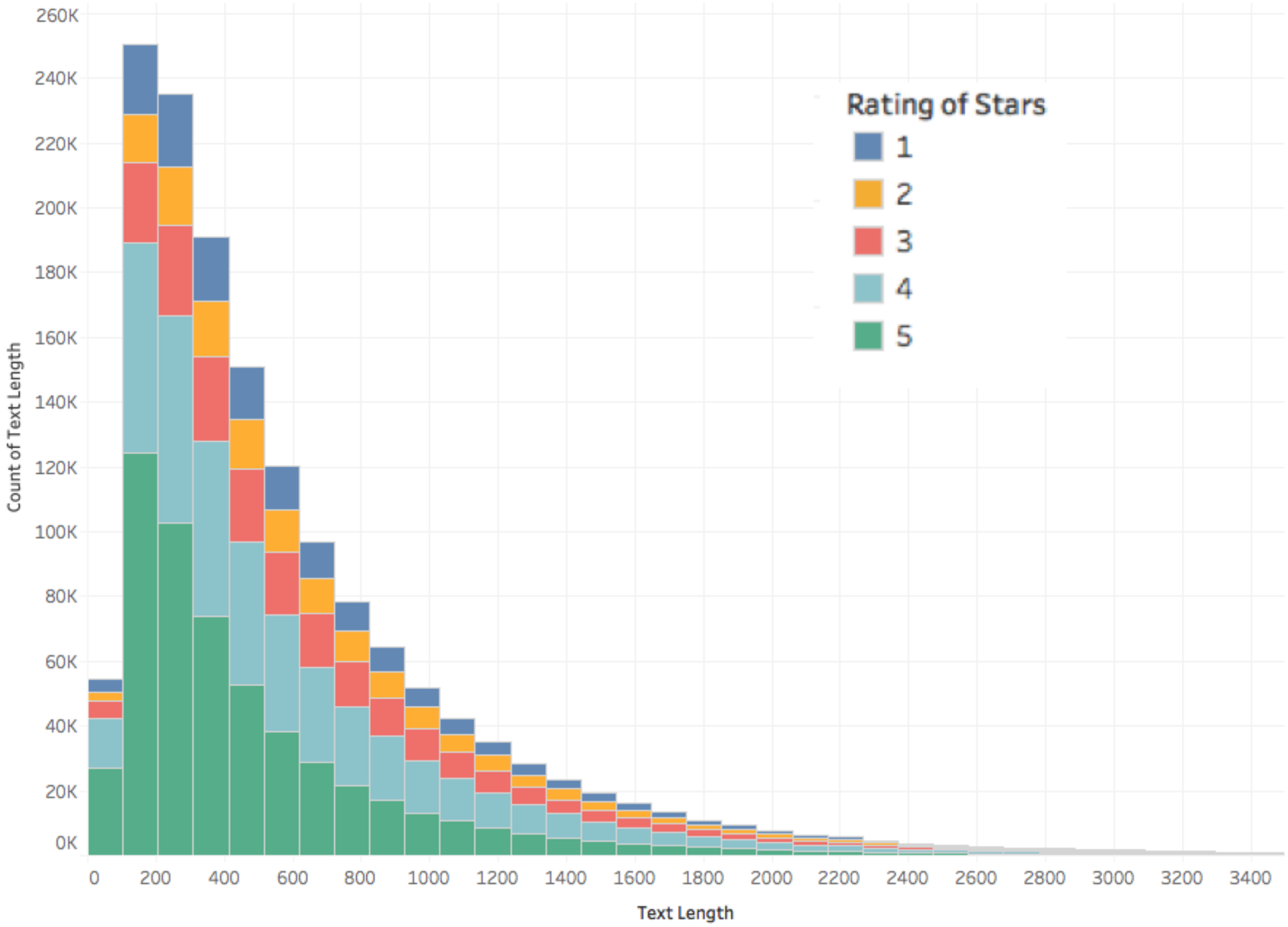
Fast food



Sports bars

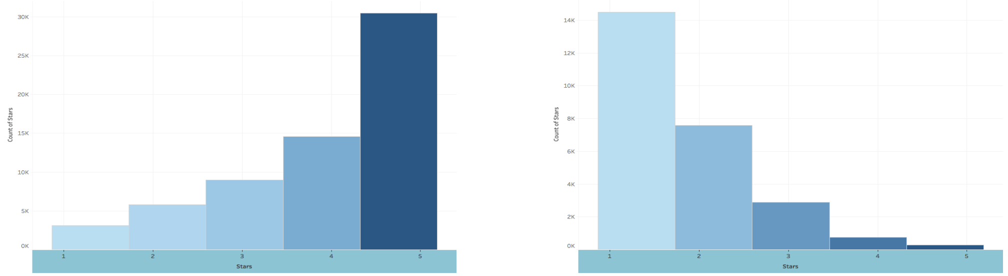
-- CATEGORIES –

Distributions of Rating of stars for certain category (after data balance)



-- TEXT --

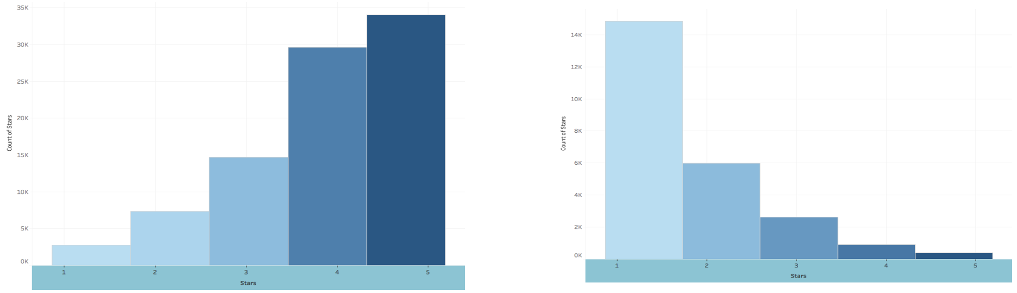
Text length vs. Number of reviews



Amazing

Terrible

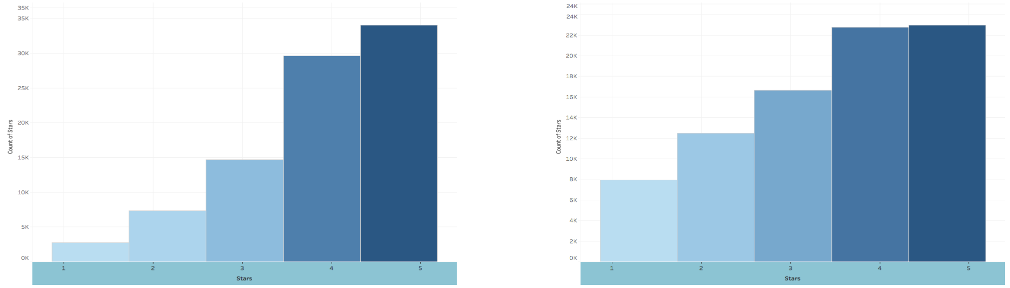
- Feeling -



Friendly

Rude

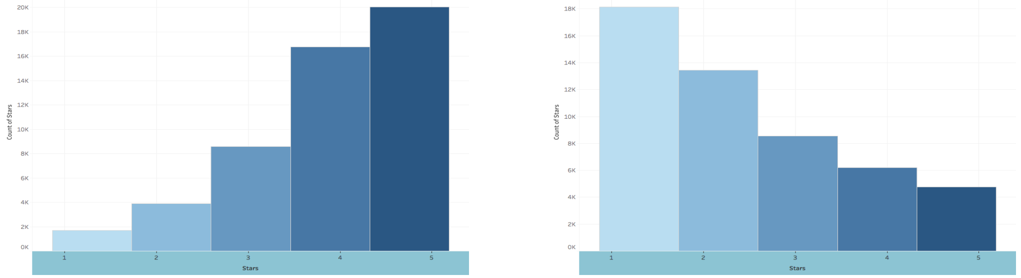
- Service -



Delicious

Fresh

- Food -



Recommend

Left

- Else -

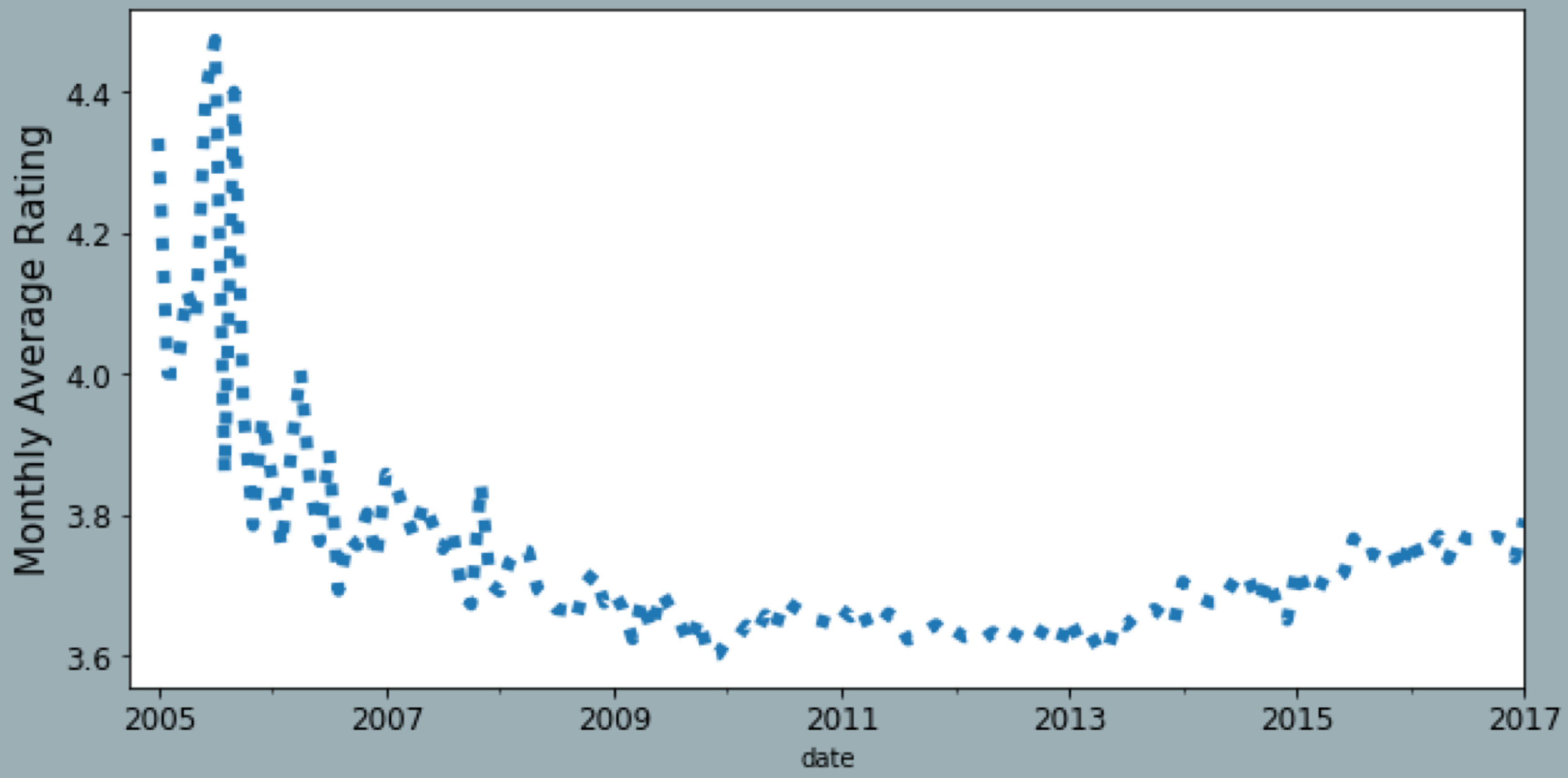
-- TEXT --

Distributions of Ratings of stars for certain Text (after data balance)

FURTHER ANALYSIS

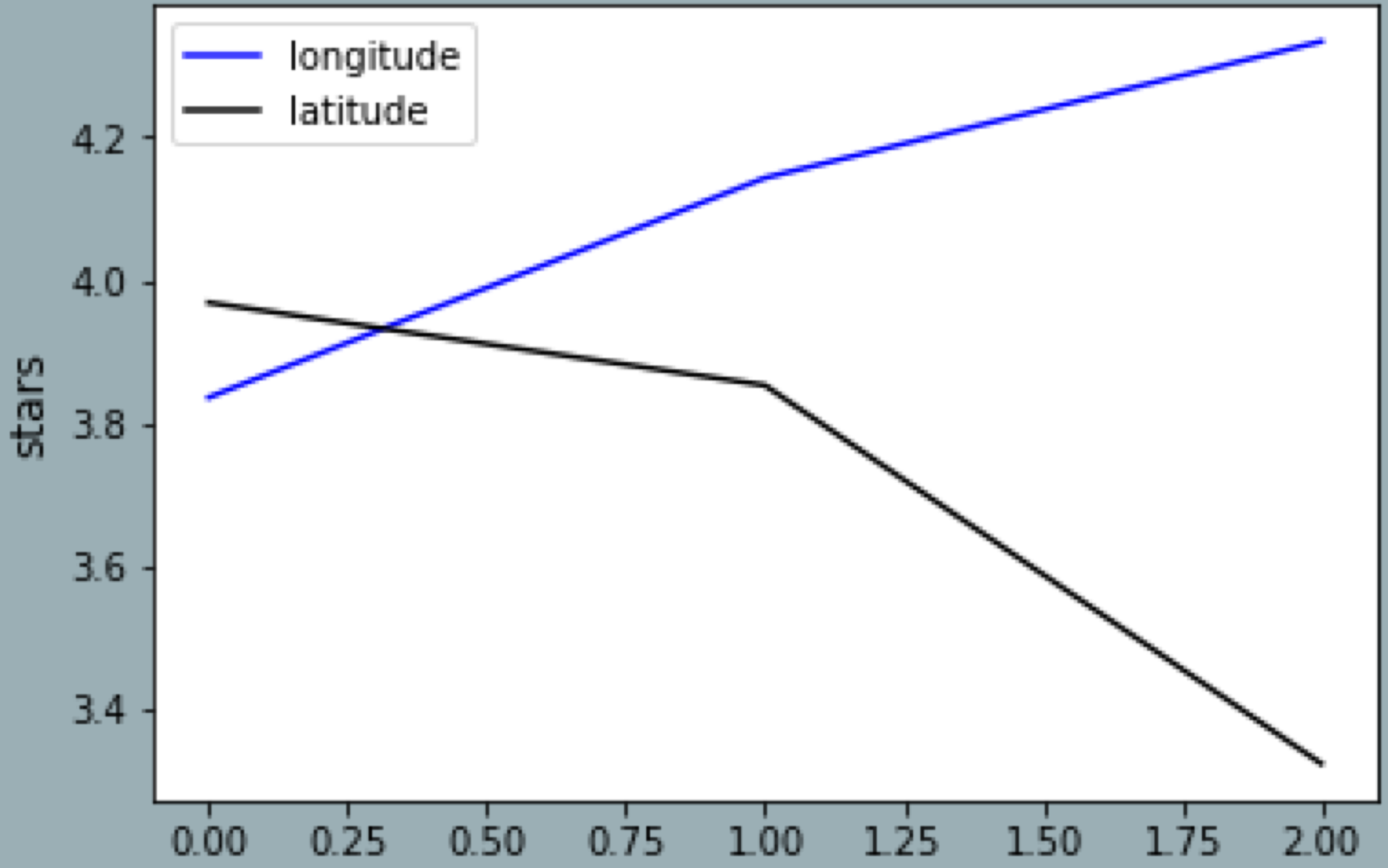
TIME-SPACE ANALYSIS

TIME EFFECT

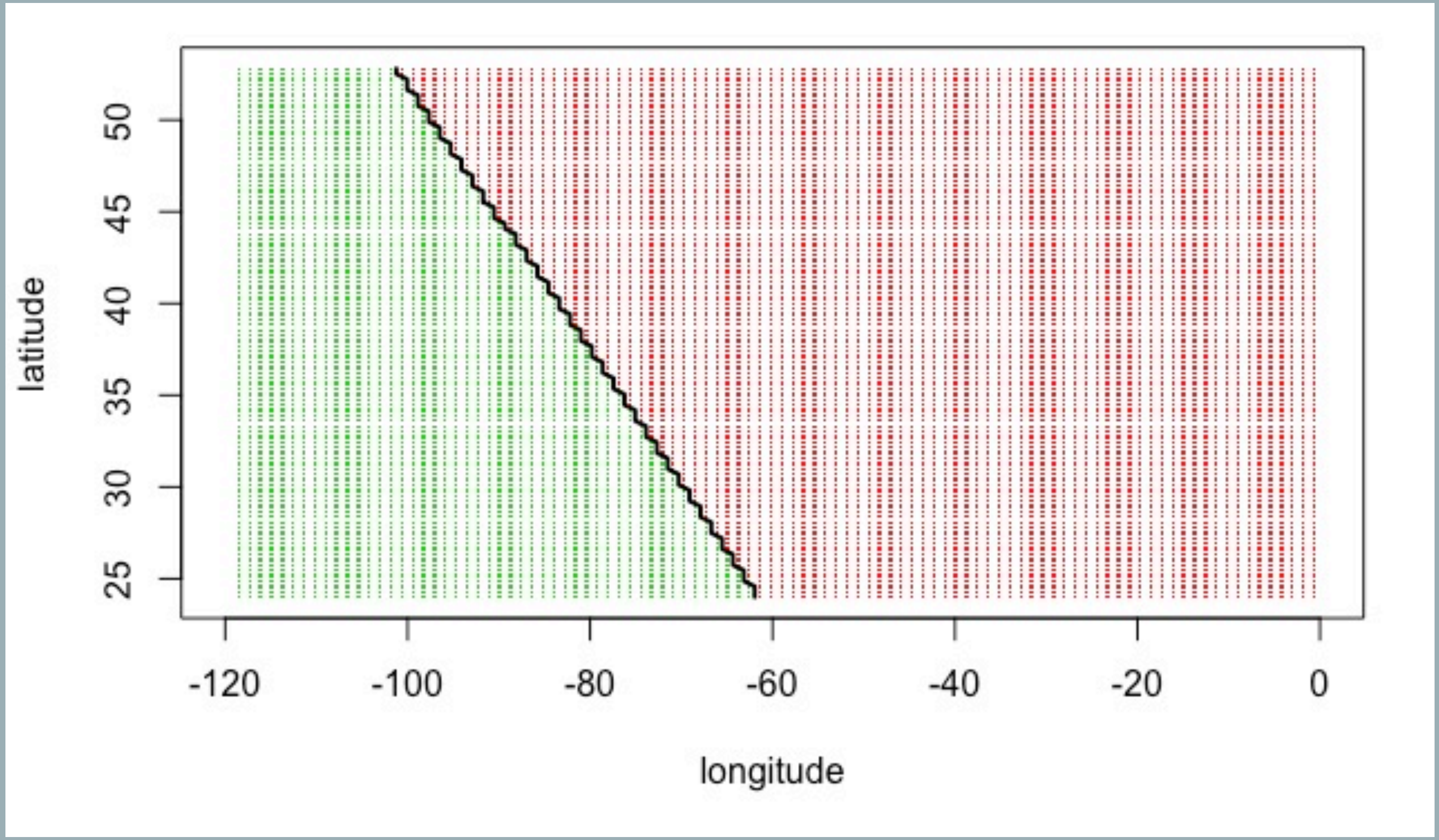


Monthly Average Rating
4.4
4.2
4.0
3.8
3.6
2005
2007
2009
2011
2013
2015
2017
date

GEOGRAPHIC EFFECT



LINEAR DISCRIMINANT ANALYSIS ON LONGITUDE AND LATITUDE

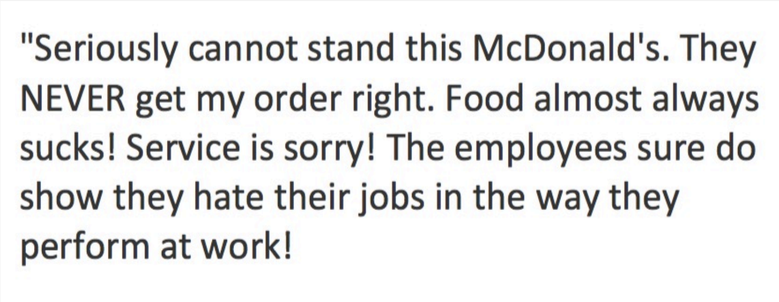


FURTHER ANALYSIS

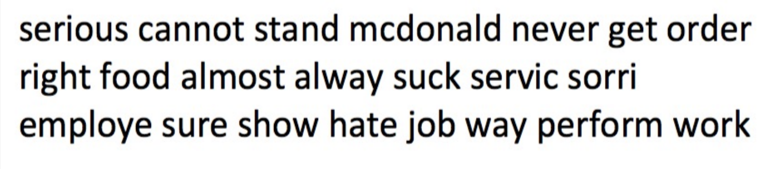
TEXT ANALYSIS

TOKENIZATION

1. Remove punctuation and extra whitespace
2. Remove stop words
3. Lexicon Normalization: Lemmatization



"Seriously cannot stand this McDonald's. They NEVER get my order right. Food almost always sucks! Service is sorry! The employees sure do show they hate their jobs in the way they perform at work!



serious cannot stand mcdonald never get order right food almost alway suck servic sorri employe sure show hate job way perform work

TEXT TRANSFORMATION

1. Generate corpus
2. Vectorize all texts
3. dimension reduction (Plan)
   - PCA (Principal Component Analysis)
   - SVD (Singular Value Decomposition)
   - LDA (Latent Dirichlet Allocation)

LDA

Topic 1: review, use, star, thing, old, look, tast, say, mean, actual

Topic 2: servic, friend, price, time, menu, delici, come, definit, amaz, littl

Topic 3: order, time, wait, tabl, ask, servic, server, minut, first, took

Topic 4: chicken, rice, soup, dish, spici, roll, noodl, thai, sauc, beef

Topic 5: fri, burger, sandwich, chicken, chees, meat, sauc, bbq, side, onion

Topic 6: breakfast, buffet, egg, coffe, brunch, ice, cream, toast, pancak, waffl

Topic 7: bar, drink, beer, night, hour, room, fun, happi, cool, patio

⋮ ⋮

REGRESSION AND CLASSIFICATION (PLAN)

- Feature selection
- Value Prediction
- Methods: (Logistic regression, SVM, Random Forest, GBM…)