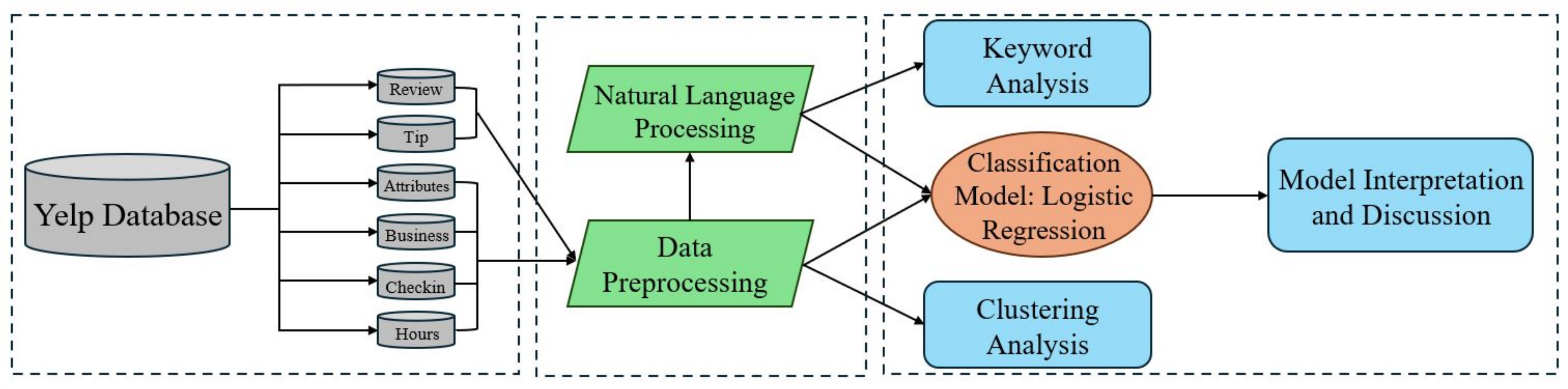
Yelp Star Rating Initiative

Ruo-Ying Qi: 261044984
Yudi Su: 261110250
Keane Dylan Yennoto: 261194825
Claire Zhao: 261194054
Kaibo Zhang: 261110409

Objective: examine the influence of various attributes of businesses and customer reviews on star ratings



01

Data and Preprocessing

YELP Dataset

01 Business.csv

Business profile

02 Attributes.csv

Business attributes information

03 Review&tip.csv

Reviews and tips written by users

04 Check-in.csv

Number of check-ins

NATURAL Classification

1,093 cities

67 states

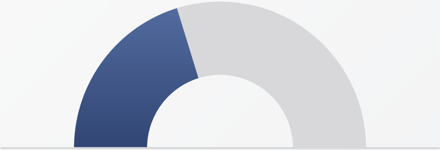
76,419 types

World Map with Latitudes and Longitudes - GIS Geography

	business_id	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open
83102	ZpK3cU9lIPddzFcaV_rygg	"XpresSpa"	NaN	"5757 Wayne Newton Blvd"	Las Vegas	NV	89119	-36.086009	-115.134643	4.0	17	1
90421	1yQUqh3_h1IOrXZmb4CBFw	"TriBeCa"	NaN	"88 Bruntsfield Place"	Edinburgh	EDH	EH10 4HG	89.999314	-142.466650	3.0	15	1

NLP Sentiment Analysis

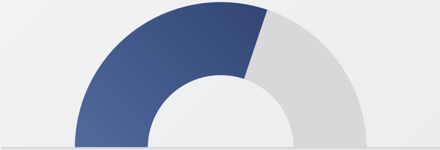
[-1.0, 1.0]



Polarity

A float where -1.0 is very negative and 1.0 is very positive

[0.0, 1.0]



Subjectivity

A float where 0.0 is very objective and 1.0 is very subjective

What to DROP ?

Numerical

KEEP!!!

Na rows

KEEP!!!

Accepts_Insurance

Unitary -> DROP!!!

What else???

[‘HairSpecializesIn_Coloring’,...,
‘ResturantsDelivery’,...,
‘DietaryRestrictions’,...]

ATTRIBUTES

Specific to business types

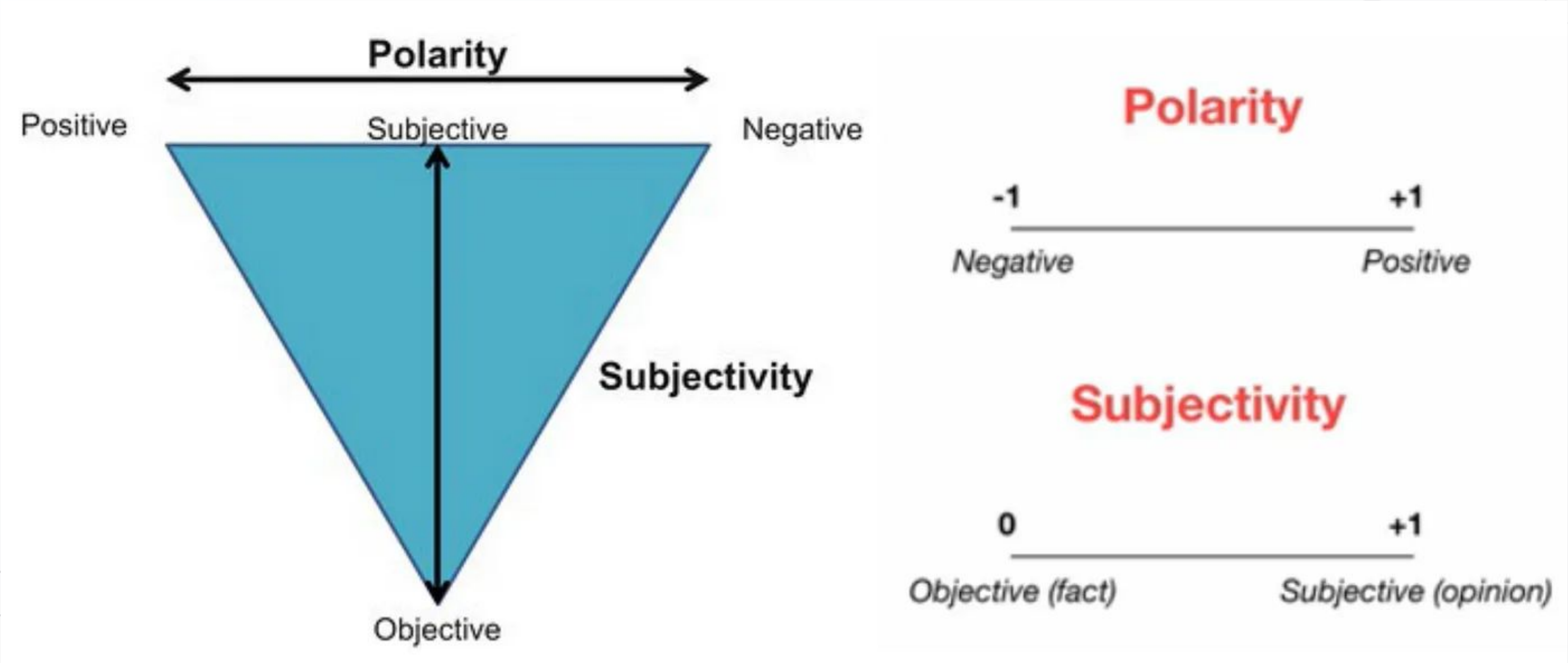
For the attributes file, columns that are specific to certain business types were **dropped**

NaN After table merging

0.5

0

	business_id	stars	review_count	label	sentiment_comment	subjectivity_comment	sentiment_tip	subjectivity_tip	checkins
0	FYWN1wneV18bWNgQjJ2GNg	4.0	22	NA	0.276481	0.562467	0.643083	0.692667	1.0
1	He-G7vWjzVUysIKrfNbPUQ	3.0	11	NA	0.277838	0.608054	0.650000	0.662500	1.0
2	KQPW8lFf1y5BT2MxiSZ3QA	1.5	18	NA	-0.044467	0.507554	NaN	NaN	1.0
3	8DShNS-LuFqpEWIp0HxijA	3.0	9	NA	0.184669	0.458150	0.223785	0.233333	1.0
4	PfOCPjBrlQAnz__NXj9h_w	3.5	116	NA	0.267249	0.596280	0.410907	0.555065	1.0
...	...	...	...	...	...	...	...	...	...
174561	ALV5R8NkZ1KGOZeuZI3u0A	4.0	4	NA	0.175780	0.446376	0.147500	0.450000	1.0
174562	gRGalHVu6BcaUDIAGVW_xQ	5.0	3	NA	0.348030	0.487755	NaN	NaN	NaN
174563	XXvZBIHoJBU5d6-a-oyMWQ	1.5	19	NA	-0.050504	0.517452	NaN	NaN	1.0
174564	INpPGgM96nPIYM1shxciHg	5.0	14	NA	0.360848	0.573782	0.850000	0.883333	1.0
174565	viKaP26BcHU6cLx8sf4gKg	5.0	4	NA	0.241796	0.447840	0.400000	0.375000	1.0


Polarity
Positive
Subjective
Negative
Subjectivity
Objective
Polarity
-1
+1
Negative
Positive
Subjectivity
0
+1
Objective (fact)
Subjective (opinion)

0 0 1

...	DogsAllowed_True	BusinessAcceptsBitcoin_False	BusinessAcceptsBitcoin_Na
...	0.0	0.0	1.0
...	0.0	0.0	1.0
...	NaN	NaN	NaN
...	0.0	0.0	1.0
...	0.0	0.0	1.0
...	...	...	...
...	NaN	NaN	NaN
...	0.0	0.0	1.0
...	0.0	0.0	1.0
...	0.0	0.0	1.0
...	0.0	0.0	1.0

02

Hypothesis Testing

Chi-square Test

H_0: the proportion of EU, NA, and SA labels in each cluster would be expected to be similar

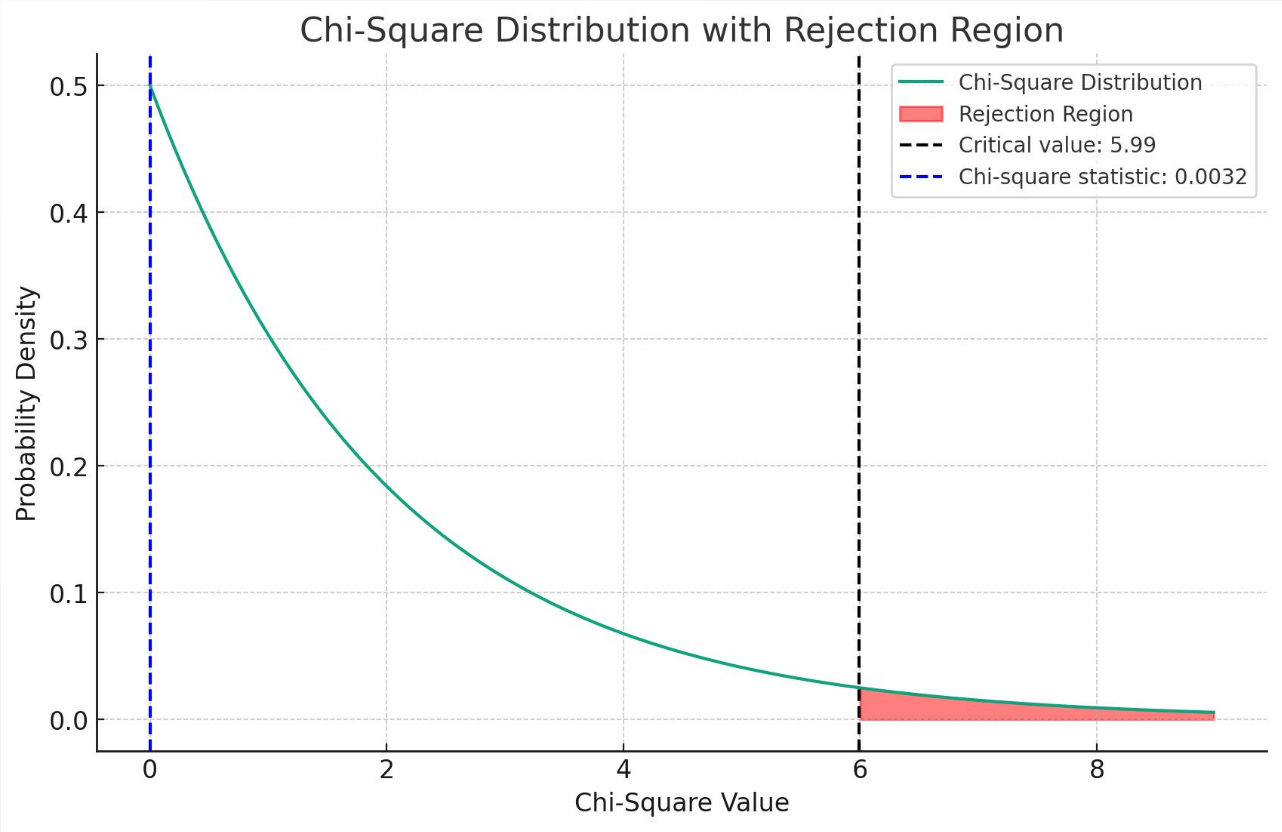
H_a: the proportion of EU, NA, and SA labels would differ between the clusters

Cluster label	0	1
label		
EU	0.044976	0.0625
NA	0.954812	0.9375
SA	0.000212	0.0000

Results

- Chi-squared statistics: 0.003278
- P-value: **0.998387**
- Degree of Freedom: 2

- Given the high p-value, **we fail to reject the null hypothesis.**
- No evidence showing that geographical location systematically leads to higher star ratings.



03

Logistic Regression Modelling

Why Logistic Regression?

Inference

Inference as our main objective

Simple Model

Simple model over complex to prioritize interpretability

Binary Class Handling

Logistic Regression's capability in predicting binary classes (positive & negative businesses

Step by Step

stars >= 4

Regard businesses with stars >= 4 as positive

Drop “Problematic” Variables

Assume {attribute}_Na: as false -> keep only {attribute}_True

Convergence Issues -> dropped binary variables with minority class less than 0.05

Retrain Model

All except one of the retrained model were statistically significant at 0.05 level of significance

Retrain Part 2

Drop “WheelchairAccessible_True”

10-fold CV Accuracy: 80.96%

Findings

sentiment_comment

significantly influences a business’s likelihood to be “positive”

checkins

doesn’t impact a restaurants performance. Odds near to 1

valet parking

Actually decreases the odds of a business being favorable

Final Logistic Model Summary

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 fstars   No. Observations:               174564
Model:                          Logit   Df Residuals:                   174554
Method:                           MLE   Df Model:                            9
Date:                Mon, 25 Mar 2024   Pseudo R-squ.:                  0.3822
Time:                        21:27:55   Log-Likelihood:                -74738.
converged:                       True   LL-Null:                   -1.2097e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
===================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                             -0.2014      0.007    -30.423      0.000      -0.214      -0.188
sentiment_comment                  2.5612      0.013    195.827      0.000       2.536       2.587
subjectivity_comment              -0.4024      0.008    -52.605      0.000      -0.417      -0.387
sentiment_tip                      0.0415      0.007      5.649      0.000       0.027       0.056
subjectivity_tip                   0.0189      0.007      2.709      0.007       0.005       0.033
checkins                          -0.0258      0.008     -3.399      0.001      -0.041      -0.011
BusinessAcceptsCreditCards_True    0.1546      0.007     22.185      0.000       0.141       0.168
BusinessParking_garage_True        0.1280      0.007     18.788      0.000       0.115       0.141
BusinessParking_valet_True        -0.1104      0.006    -17.873      0.000      -0.123      -0.098
BikeParking_True                   0.0340      0.006      5.240      0.000       0.021       0.047
===================================================================================================
```

04

Proposed Initiative & Potential Benefits

Text Analysis

STEP 1

Calculate sentiment scores

STEP 2

Identify reviews in top & lower 25% sentiment scores

STEP 3

Extract only the “interested” keywords

STEP 4

Aggregate filtered words & count frequencies

Review Summaries

Figure 1: Recurring Keywords in Reviews with **Top 25%** Sentiment Score

Figure 2: Recurring Keywords in Reviews with **Bottom 25%** Sentiment Score

Advisory Service to enhance ratings

Increase ratings

Guide on recurring positive aspects to improve business performance

Reduce chances to receive lower ratings

Suggest to avoid actions associated with common negative keywords

Potential Benefits

Encourage more businesses to register

Provide insights to assist businesses in comprehending and forecasting their probability of achieving top ratings

More & higher quality reviews

Users know their feedback is valued

Yelp gains increased trust among businesses and users -->
win-win-win situation

05

Post-Implementation Strategy

Maintenance and Feedback Loop

01

Update data source

Focus on collecting data on current attributes for future analysis

02

Analyze star-rating improvements

If tangible improvements in star ratings occurs as business owners integrate customer- friendly features

Regional Analysis

01

Zoom in

Analyze the business characters of different regions or cities

THANKS!