Group 2:
Benn Kulcsar, Mikayla Voorn,
Xinyi Wei, Xinyu Wu

IST 418

WHAT'S BREWING?

RECOMMENDATIONS FOR OPENING A NEW COFFEE SHOP

Project Overview

What is our project?

Our team has been hired to consult for a prospective coffee shop owner in Austin, Texas! We are tasked with giving insights into what customers like most and least when it comes to coffee shops!

Prediction, inference, and other goals

Identify the most and least important factors affecting ratings of coffee shops and determine their impact on customer satisfaction

Analyze the frequency of words appearing in negative and positive ratings to gain insights into the most common reasons for customer satisfaction or dissatisfaction

Find the differences between 5-star and 4-star reviews

Datasets

ratings and sentiments dataframe

coffee_shop_name	review_text	rating	num_rating	cat_rating
The Factory - Caf...	11/25/2016 1 che...	5.0 star rating	5	HIGH
The Factory - Caf...	12/2/2016 Listed...	4.0 star rating	4	HIGH
The Factory - Caf...	11/30/2016 1 che...	4.0 star rating	4	HIGH
The Factory - Caf...	11/25/2016 Very ...	2.0 star rating	2	LOW
The Factory - Caf...	12/3/2016 1 chec...	4.0 star rating	4	HIGH
The Factory - Caf...	11/20/2016 1 che...	4.0 star rating	4	HIGH
The Factory - Caf...	" 10/27/2016 2 ch...	Anderson Lane is...	the pro vs con f...	really. 3. They ... | i
The Factory - Caf...	" 11/2/2016 2 che...	it's not child f...	make sure your c...	the sex images i... | t
The Factory - Caf...	" 10/25/2016 1 ch...	come here. Pop+a...	3.0 star rating	3
The Factory - Caf...	11/10/2016 3 che...	5.0 star rating	5	HIGH
The Factory - Caf...	" 10/22/2016 1 ch...	$$ Price tag of ...	don't come here ...	food & food
The Factory - Caf...	11/20/2016 The s...	3.0 star rating	3	LOW
The Factory - Caf...	11/17/2016 1 che...	3.0 star rating	3	LOW
The Factory - Caf...	12/5/2016 This i...	5.0 star rating	5	HIGH
The Factory - Caf...	11/13/2016 Beaut...	5.0 star rating	5	HIGH
The Factory - Caf...	11/9/2016 1 chec...	5.0 star rating	5	HIGH
The Factory - Caf...	11/6/2016 Really...	5.0 star rating	5	HIGH
The Factory - Caf...	10/25/2016 1 che...	4.0 star rating	4	HIGH
The Factory - Caf...	10/15/2016 1 che...	4.0 star rating	4	HIGH
The Factory - Caf...	12/1/2016 So muc...	4.0 star rating	4	HIGH

raw yelp review dataframe

coffee_shop_name	full_review_text	star_rating
The Factory - Caf...	11/25/2016 1 che...	5.0 star rating
The Factory - Caf...	12/2/2016 Listed...	4.0 star rating
The Factory - Caf...	11/30/2016 1 che...	4.0 star rating
The Factory - Caf...	11/25/2016 Very ...	2.0 star rating
The Factory - Caf...	12/3/2016 1 chec...	4.0 star rating
The Factory - Caf...	11/20/2016 1 che...	4.0 star rating
The Factory - Caf...	" 10/27/2016 2 ch...	Anderson Lane is...
The Factory - Caf...	" 11/2/2016 2 che...	it's not child f...
The Factory - Caf...	" 10/25/2016 1 ch...	come here. Pop+a...
The Factory - Caf...	11/10/2016 3 che...	5.0 star rating
The Factory - Caf...	" 10/22/2016 1 ch...	$$ Price tag of ...
The Factory - Caf...	11/20/2016 The s...	3.0 star rating
The Factory - Caf...	11/17/2016 1 che...	3.0 star rating
The Factory - Caf...	12/5/2016 This i...	5.0 star rating
The Factory - Caf...	11/13/2016 Beaut...	5.0 star rating
The Factory - Caf...	11/9/2016 1 chec...	5.0 star rating
The Factory - Caf...	11/6/2016 Really...	5.0 star rating
The Factory - Caf...	10/25/2016 1 che...	4.0 star rating
The Factory - Caf...	10/15/2016 1 che...	4.0 star rating
The Factory - Caf...	12/1/2016 So muc...	4.0 star rating

Datasets from Kaggle

sentiments by shop dataframe

coffee_shop_name	num_reviews	rating	coffee	tea	vibe	internet	food
Manana Coffee & J...	33	4.848484848	0.6666667	0.03030303	0.515151515	0	0.212121212
Brian's Brew	45	4.844444444	0.8888889	0	0.044444444	0	0.111111111
Flitch Coffee	28	4.821428571	0.5714286	0.071428571	0.464285714	0	0
Third Coast Coffe...	56	4.821428571	0.75	0	0.160714286	0	0.035714286
Kowabunga Coffee	16	4.8125	0.0625	0.125	0	0	0
Legend Coffee	28	4.714285714	0.8928571	0.035714286	0	0	0.178571429
Fleet Coffee	57	4.701754386	1	0	0.315789474	0.035087719	0.087719298
Holy Grounds	30	4.633333333	0.5333333	0	0.566666667	0	0.433333333
Anderson's Coffee...	100	4.62	0.63	0.13	0.12	0	0
Apanas Coffee & B...	136	4.580882353	0.8088235	0.014705882	0.551470588	0.029411765	0.198529412
Flat Track Coffee	63	4.571428571	0.7936508	0	0.222222222	0.015873016	0.126984127
Friends & Neighbors	29	4.551724138	0.3793103	0	0	0	0.310344828
Lola Savannah Coffee	104	4.55	0.68	0.265	0.22	0.015	0.2
Corona Coffee	100	4.53	0.76	0.08	0.28	0.05	0.35
Figure 8 Coffee P...	100	4.5	0.88	0	0.47	0.03	0.11
Sister Coffee	17	4.470588235	0.8235294	0.058823529	0	0	0.117647059
Lucky Lab Coffee	25	4.44	0.6	-0.04	0.2	0	0.36
Cafe Ruckus	68	4.426470588	0.8676471	0.117647059	0	0.029411765	0.205882353
Cafe Creme	100	4.37	0.54	0.02	0.47	0.04	0.52
Fat Cats Organic ...	94	4.361702128	0.7978723	0.031914894	0.180851064	0.021276596	0.404255319

Ratings & Sentiment Columns: coffee_shop_name, review_text, rating, num_rating, cat_rating, bool_high, overall_sent, and sentiment for individual words such as: service, seating, and location

Raw Yelp Review Columns: coffee_shop_name, full_review_text, star_rating

Sentiments by Shop Columns: coffee_shop_name, num_reviews, rating, coffee, tea, vibe, internet, food, alcohol, seating, service, parking, location, local, price, hours

Data Exploration

66

Different Coffee Shops

102

Average reviews per store

7,616

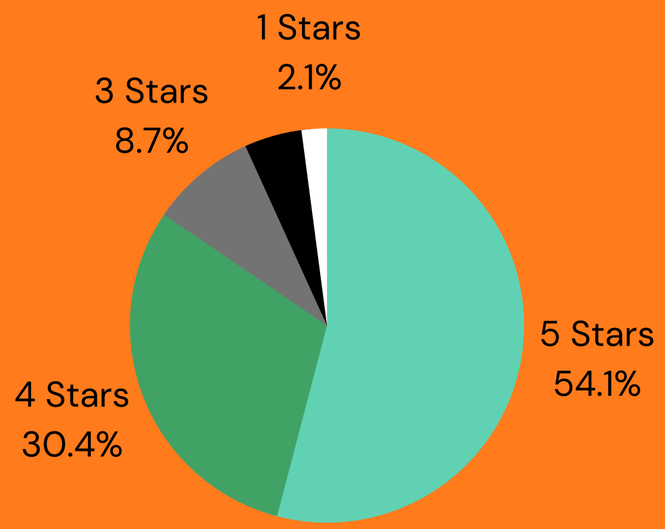
Reviews

400

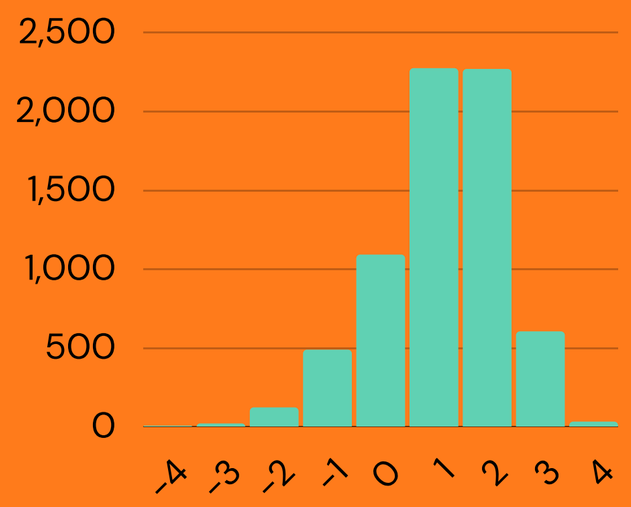
The maximum review a store

4.3★

Average rating amongst coffee shops



Distribution of Review Ratings



Distribution of Review Sentiments

Data Transformations

- Encoded the overall sentiment to be a binary value: 1 for positive and 0 for negative
- Selected only coffee shops with 5 star reviews to carry out machine learning tasks
- Removed columns pertaining to individual word sentiments (we were unsure how these numbers were calculated)
- Removed columns with string data in either the overall sentiment or star rating column

overall_sent	sentiment
4.0	1
3.0	1
2.0	1
1.0	1
2.0	1
0.0	0
3.0	1
0.0	0
0.0	0
1.0	1
1.0	1
1.0	1
3.0	1
2.0	1
3.0	1
2.0	1
2.0	1
2.0	1
1.0	1
-1.0	0

Methods Used

Sentiment Analysis

PySpark Logistical Regression:

- Tokenize the words, separating the text into words
- Stop words remover
- TF-IDF transformer
- Logistical Regression, to predict a binary value
- Check the word weights

Unsupervised Learning

Clustering:

- Transform coffee shop review into vectors
- Standardize data before applying unsupervised learning algorithms
- Define a PCA transformer
- Normalizes each row of data
- Defines a Spark KMeans clustering object

Interesting Results

What does it really mean to be 5 stars?

After performing logistic regression to predict if a review earns 4 or 5 stars, we got an accuracy of 67%

This proves just how subjective stars are since they are based on user input without clear guidelines

```
tokenizer = feature.RegexTokenizer(minTokenLength=2)\
  .setGaps(False)\
  .setPattern("\\p{L}+")\
  .setInputCol("review_text")\
  .setOutputCol("words")

tok = Pipeline(stages=[tokenizer])

reviews_1 = tok.fit(df_4_5).transform(df_4_5)

train, validate, test = reviews_1.randomSplit([0.6, 0.3, 0.1], seed = 123)

stop_words = requests.get('http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words').text.split()

sw_filter = feature.StopWordsRemover()\
  .setStopWords(stop_words)\
  .setCaseSensitive(False)\
  .setInputCol("words")\
  .setOutputCol("filtered")

cv = feature.CountVectorizer(inputCol='filtered', outputCol='tf')
idf = IDF().\
    setInputCol('tf').\
    setOutputCol('tfidf')

pipe_features = Pipeline(stages=[sw_filter, cv, idf])

feat_df = pipe_features.fit(train).transform(train)

#feat_df.show()

##

lr = LogisticRegression().\
    setLabelCol('num_rating').\
    setFeaturesCol('tfidf').\
    setRegParam(0.0).\
    setMaxIter(100).\
    setElasticNetParam(0.)

lr_pipe = Pipeline(stages=[pipe_features, lr]).fit(train)

log_df = lr_pipe.transform(validate)
```

num_rating	prediction
0	1.0
1	0.0
1	1.0
1	0.0
1	1.0
1	1.0
1	0.0
1	1.0
1	1.0
1	1.0
0	1.0
1	1.0
1	1.0
1	1.0
1	1.0
1	0.0
1	1.0
1	1.0
1	1.0
1	0.0

```
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator(rawPredictionCol = 'rawPrediction', labelCol='num_rating', metricName = 'areaUnderROC')

auc = evaluator.evaluate(log_df)

print(f"AUC: {auc}")
```

```
AUC: 0.6685899032027139
```

Interesting Results

word	score
growing	-6.419837
uninviting	-4.932099
pot	-4.826837
jar	-4.739572
style	-4.692448
alright	-4.284726
bfast	-4.208210
topochico	-4.120738
picky	-3.963064
charger	-3.783096
october	-3.694816
disappointing	-3.608396
die	-3.602675
nighters	-3.565938
hangover	-3.565938
created	-3.528071
swinging	-3.490718
solo	-3.490718
ummmm	-3.242681
cranberries	-3.242681

word	score
crepes	3.996679
resisted	2.991568
great	2.519338
avoid	2.359395
digged	2.347262
limeades	2.198689
patient	2.142910
checked	1.937350
connectivity	1.903888
tuesday	1.730738
freshly	1.729827
preparation	1.650829
moreso	1.634072
slicker	1.580678
southside	1.529011
continued	1.529011
matter	1.506122
anytime	1.488364
storefront	1.436782
mmmmmm	1.427484

Looking at every review, what words indicate a positive coffee shop experience? What about negative?

Users like:

- crepes
- limeades
- patience (workers)
- connectivity (either workers or WIFI)
- freshly + prepared (food)
- southside (location)

Users dislike:

- style (this one is vague and subjective)
- topochico (type of drink)
- uninviting (spaces)
- alright (things that are average)
- cranberries

All reviews

word	score
growing	-6.419837
uninviting	-4.932099
pot	-4.826837
jar	-4.739572
style	-4.692448
alright	-4.284726
bfast	-4.208210
topochico	-4.120738
picky	-3.963064
charger	-3.783096
october	-3.694816
disappointing	-3.608396
die	-3.602675
nighters	-3.565938
hangover	-3.565938
created	-3.528071
swinging	-3.490718
solo	-3.490718
ummmm	-3.242681
cranberries	-3.242681

word	score
crepes	3.996679
resisted	2.991568
great	2.519338
avoid	2.359395
digged	2.347262
limeades	2.198689
patient	2.142910
checked	1.937350
connectivity	1.903888
tuesday	1.730738
freshly	1.729827
preparation	1.650829
moreso	1.634072
slicker	1.580678
southside	1.529011
continued	1.529011
matter	1.506122
anytime	1.488364
storefront	1.436782
mmmmmm	1.427484

4 star ratings

word (4 stars)	score (4 stars)
peeps	0.908046
stumbled	0.856763
brag	0.849955
years	0.821812
wife	0.797027
somewhat	0.746249
meat	0.722419
pricy	0.692157
struggle	0.672241
owned	0.668760

word (4 stars)	score (4 stars)
pairing	-4.509169
hyped	-4.509169
cherry	-4.258818
goat	-4.258818
enchanting	-4.258818
acidic	-2.853716
smoothest	-2.853716
liquid	-2.853716
plugs	-2.805956
tons	-2.805956

5 star ratings

word (5 stars)	score (5 stars)
latta	2.312575
antonio	1.687476
return	1.401851
kudos	1.401851
heaven	1.346816
superb	1.343509
tables	1.268758
mascarpone	1.265766
coffees	1.257388
offerings	1.257388

word (5 stars)	score (5 stars)
burnet	-9.454737
dammit	-8.562844
ridiculously	-8.390448
combination	-8.115990
sad	-7.220616
nailed	-6.575661
righteous	-6.498804
brazil	-6.498804
shape	-5.092633
created	-4.458723

Interesting Results

What is the difference between 4 and 5 star ratings?

- word importance differences
- a lot stronger weighted words for the 5 star reviews (so a good reason to only use 5 start reviews when looking for recommendations)

4 star ratings

word (4 stars)	score (4 stars)
peeps	0.908046
stumbled	0.856763
brag	0.849955
years	0.821812
wife	0.797027
somewhat	0.746249
meat	0.722419
pricy	0.692157
struggle	0.672241
owned	0.668760

word (4 stars)	score (4 stars)
pairing	-4.509169
hyped	-4.509169
cherry	-4.258818
goat	-4.258818
enchanting	-4.258818
acidic	-2.853716
smoothest	-2.853716
liquid	-2.853716
plugs	-2.805956
tons	-2.805956

5 star ratings

word (5 stars)	score (5 stars)
latta	2.312575
antonio	1.687476
return	1.401851
kudos	1.401851
heaven	1.346816
superb	1.343509
tables	1.268758
mascarpone	1.265766
coffees	1.257388
offerings	1.257388

word (5 stars)	score (5 stars)
burnet	-9.454737
dammit	-8.562844
ridiculously	-8.390448
combination	-8.115990
sad	-7.220616
nailed	-6.575661
righteous	-6.498804
brazil	-6.498804
shape	-5.092633
created	-4.458723

Data Transformations

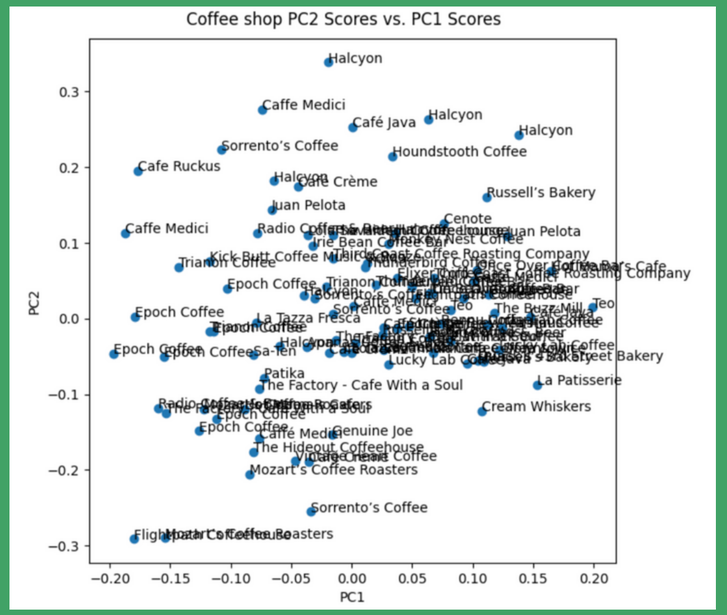
Unsupervised Learning

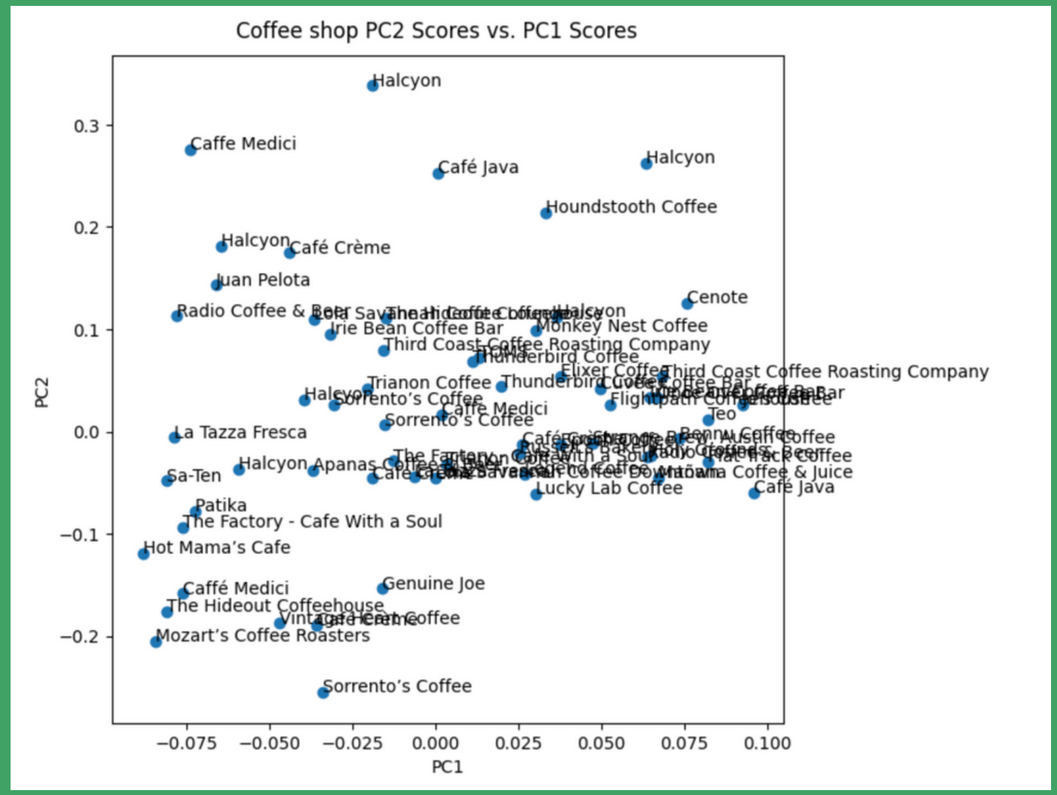
- Kept only the "coffee_shop_name", "review_text", and "rating" columns of the rat_sent dataframe
- Created a new column called "num_rating" which is the numeric rating derived from the "rating" column
- Took 100 random samples of the 5 stars rating
- Created a pipeline with tokenizer, stop words filter, CountVectorizer, and idf to fit the dataframe

coffee_shop_name	review_text	rating	num_rating
The Factory - Caf...	11/25/2016 1 che...	5.0 star rating	5.0
The Factory - Caf...	11/10/2016 3 che...	5.0 star rating	5.0
The Factory - Caf...	12/5/2016 This i...	5.0 star rating	5.0
The Factory - Caf...	11/13/2016 Beaut...	5.0 star rating	5.0
The Factory - Caf...	11/9/2016 1 chec...	5.0 star rating	5.0

Interesting Results

Randomly pick one hundred reviews from different coffee shops





Interesting Results

Different Coffee Shops grouped by k-means group

kmeans_feat	desc
1	[Caffé Medici , Lucky Lab Coffee , Sorrento's Coffee , Once Over Coffee Bar , Genuine Joe , Cuvée Coffee Bar , Radio Coffee & Beer , La Tazza Fresca , Strange Brew, Austin Coffee , Quack's 43rd Street Bakery , Lola Savannah Coffee Downtown , Café Java , Hot Mama's Cafe]
3	[Mozart's Coffee Roasters , Flat Track Coffee , Third Coast Coffee Roasting Company , Bennu Coffee , Halcyon , Cream Whiskers , Halcyon , Tea Haus , Once Over Coffee Bar , Trianon Coffee , Teo , Teo , Teo , La Patisserie , Russell's Bakery , Hot Mama's Cafe]
4	[Café Crème , Sorrento's Coffee]
2	[Cenote , The Buzz Mill]
0	[Vintage Heart Coffee , Mozart's Coffee Roasters , Mozart's Coffee Roasters , Monkey Nest Coffee , Café Crème , Café Crème , Café Crème , Legend Coffee , Epoch Coffee , Epoch Coffee , Epoch Coffee , Epoch Coffee , The Factory – Cafe With a Soul , Caffe Medici , Caffe Medici , Caffe Medici , Holy Grounds , Mañana Coffee & Juice , Epoch Coffee , Epoch Coffee , Epoch Coffee , Epoch Coffee , Epoch Coffee , The Factory – Cafe With a Soul , The Factory – Cafe With a Soul , Houndstooth Coffee , Caffé Medici , Flightpath Coffeehouse , Flightpath Coffeehouse , Third Coast Coffee Roasting Company , Thunderbird Coffee , Thunderbird Coffee , Apanas Coffee & Beer , Halcyon , Cherrywood Coffeehouse , Jo's Coffee , Patika , Cenote , The Hideout Coffeehouse , The Hideout Coffeehouse , Lucky Lab Coffee , Lucky Lab Coffee , Lola Savannah Coffee Lounge , Halcyon , Irie Bean Coffee Bar , Irie Bean Coffee Bar , Sa–Ten , Sorrento's Coffee , Sorrento's Coffee , Elixer Coffee , Cafe Ruckus , Halcyon , Halcyon , Halcyon , TOMS , Genuine Joe , Radio Coffee & Beer , Radio Coffee & Beer , Trianon Coffee , Trianon Coffee , Trianon Coffee , Kick Butt Coffee Music & Booze , La Tazza Fresca , Juan Pelota , Juan Pelota , Juan Pelota , My Sweet Austin , My Sweet Austin , My Sweet Austin , Café Java , Café Java , Russell's Bakery , Russell's Bakery , Hot Mama's Cafe]

Findings & Recommendations

What makes a coffee shop successful?

Based on the sentiment of reviews overall, here are the three most important factors:

- **Unique food items; such as crepes and limeades**
- **Quality of staff; the people you hire should be patient and connect with the customers**
- **Spaces should be inviting and styled in a way that appeals to people; you should check out coffee shops similar to what style you're interested in**

Based on the sentiment of only 5-star reviews, here are the four most important factors:

- **Coffee quality is pivotal!**
- **Create the best possible experience for customers; aim to make it superb**
- **Tables should be clean and usable, that is where a lot of customers will spend their time!**
- **There should be enough diversity in product offerings to have something for everything**

Recommendation for the coffee shop

If you love coffee shops such as If you enjoy coffee shops like Mozart's Coffee Roasters, Flat Track Coffee, or Third Coast Coffee, you should definitely give these places a try:

- **Bennu Coffee**
- **Halcyon**
- **Cream Whiskers**
- **Halcyon**
- **Tea Haus**
- **Once Over Coffee Bar**
- **Trianon Coffee**

If you love coffee shops such as Caffé Medici, Lucky Lab Coffee, or Sorrento's Coffee, you should definitely give these places a try:

- **Genuine Joe**
- **Cuvée Coffee Bar**
- **Radio Coffee & Beer**
- **La Tazza Fresca**
- **Strange Brew**
- **Austin Coffee**

COFFEE PLEASE

Problems Encountered

- Encoding the overall sentiment to be a binary value: 1 for positive and 0 for negative
- Randomly pick 100 coffee shop reviews in the dataset to do the unsupervised learning due to the dataset size
- To delve deeper into the results and findings, we further narrowed down the selection by only choosing coffee shops with 5-star reviews for clustering.
- The stop words remover was noth enough to filter out unnecessary words, to improve the model we would have to manually remove words not associated with coffee shops

Summary

Through utilizing different methodologies we were able to capture what leads to high ratings and goes into making a successful coffee shop.

The keywords that led to a higher rating were not necessarily as helpful as we had anticipated, but still gave us interesting insights into what costumers like to see in their favorite coffee shops.

Through clustering we are able to recommend coffee shops to check out to gain inspiration for creating a new cafe. However, because of the large amount of shops included in our dataset, we had to limit the ones we used to best visualize the clusters.

THANK YOU!

ANY QUESTIONS?