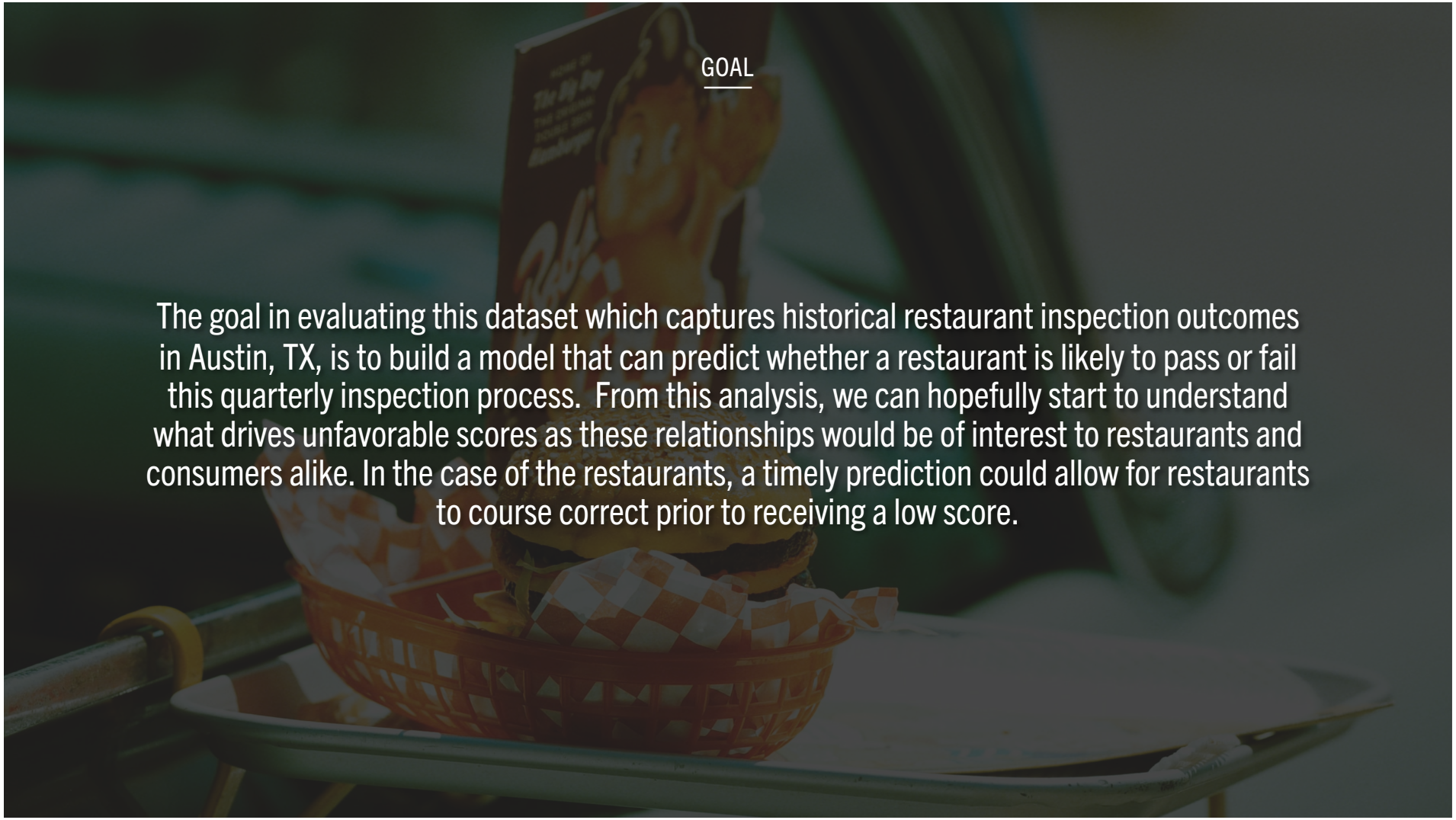
PREDICTING RESTAURANT INSPECTION OUTCOMES

CAPSTONE 2 - PRESENTATION



GOAL

The goal in evaluating this dataset which captures historical restaurant inspection outcomes in Austin, TX, is to build a model that can predict whether a restaurant is likely to pass or fail this quarterly inspection process. From this analysis, we can hopefully start to understand what drives unfavorable scores as these relationships would be of interest to restaurants and consumers alike. In the case of the restaurants, a timely prediction could allow for restaurants to course correct prior to receiving a low score.

MOTIVATION AND CHALLENGES

MOTIVATION:

- GREAT OPPORTUNITY TO DO SOME ANOMALY DETECTION
- THE DATA IS HIGHLY CLASS IMBALANCED

CHALLENGES:

- THE DATA IS HIGHLY CLASS IMBALANCED ☺

(WITH ONLY XX% OF INSPECTIONS RESULTING IN A FAILING SCORE)

- MANY RESTAURANTS HAVE FEW OR NO REVIEWS

THE DATA

SOURCES:

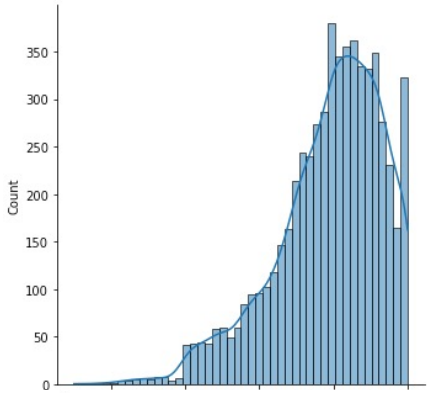
- AUSTIN RESTAURANT INSPECTIONS SCORES | AUSTIN PUBLIC HEALTH
- RESTAURANT REVIEWS | YELP

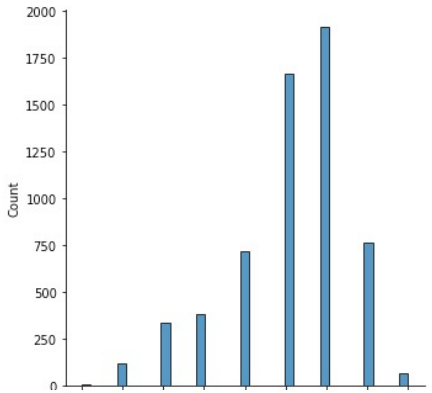
HIGH-LEVEL STATS:

- 3 YEARS OF INSPECTIONS HISTORY
- INSPECTIONS OCCUR QUARTERLY
- CONTAINS SCORES FOR XX UNIQUE RESTAURANTS
- XX TOTAL REVIEWS
- ALLOWING A 60-DAY LOOKBACK FOR REVIEWS-BASED FEATURES
- ASSSUMES THAT NON REVIEWS-BASED FEATURES ARE STATIC OVER THE 3 YEARS

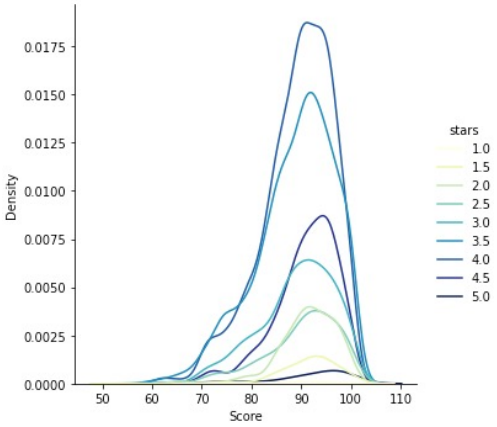
HOW IS THE DATA DISTRIBUTED?

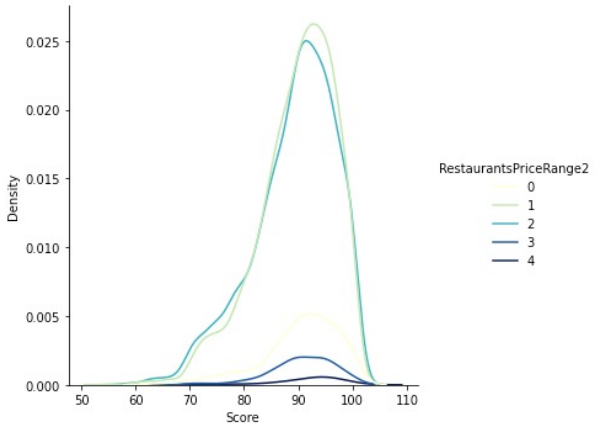
- INSPECTION SCORES AND RATINGS ALIKE ARE LEFT-SKEWED AND TEND TO BE HIGH





- INTERESINGLY AND ENCOURAGINGLY INSPECTIONS SCORES FOLLOW A SIMILAR DISTRIBUTION ACROSS RATINGS AND PRICE RANGE





THE SCORING IS HIGHLY IMBALANCED

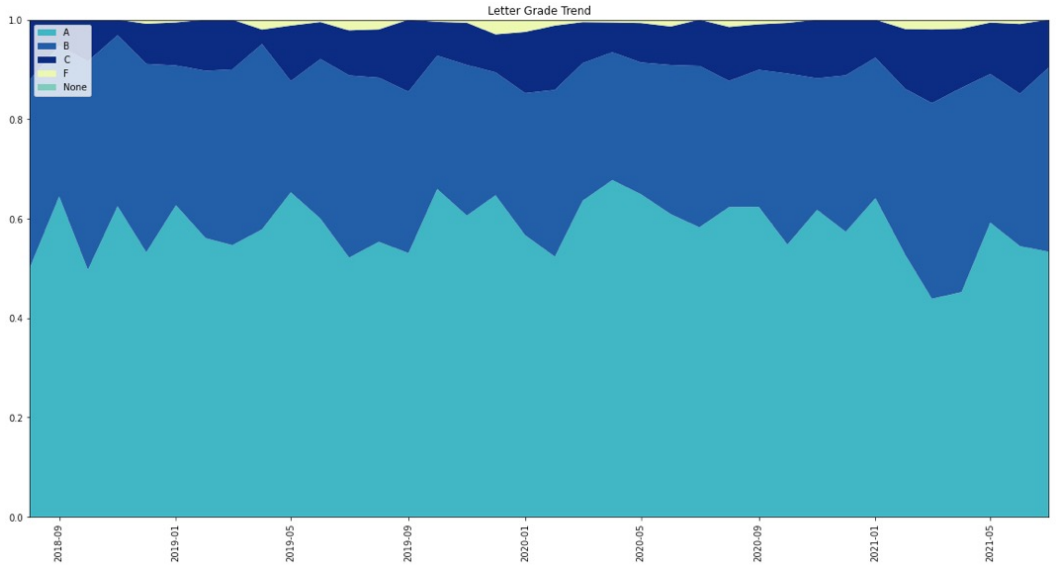
A: 58%

B: 32%

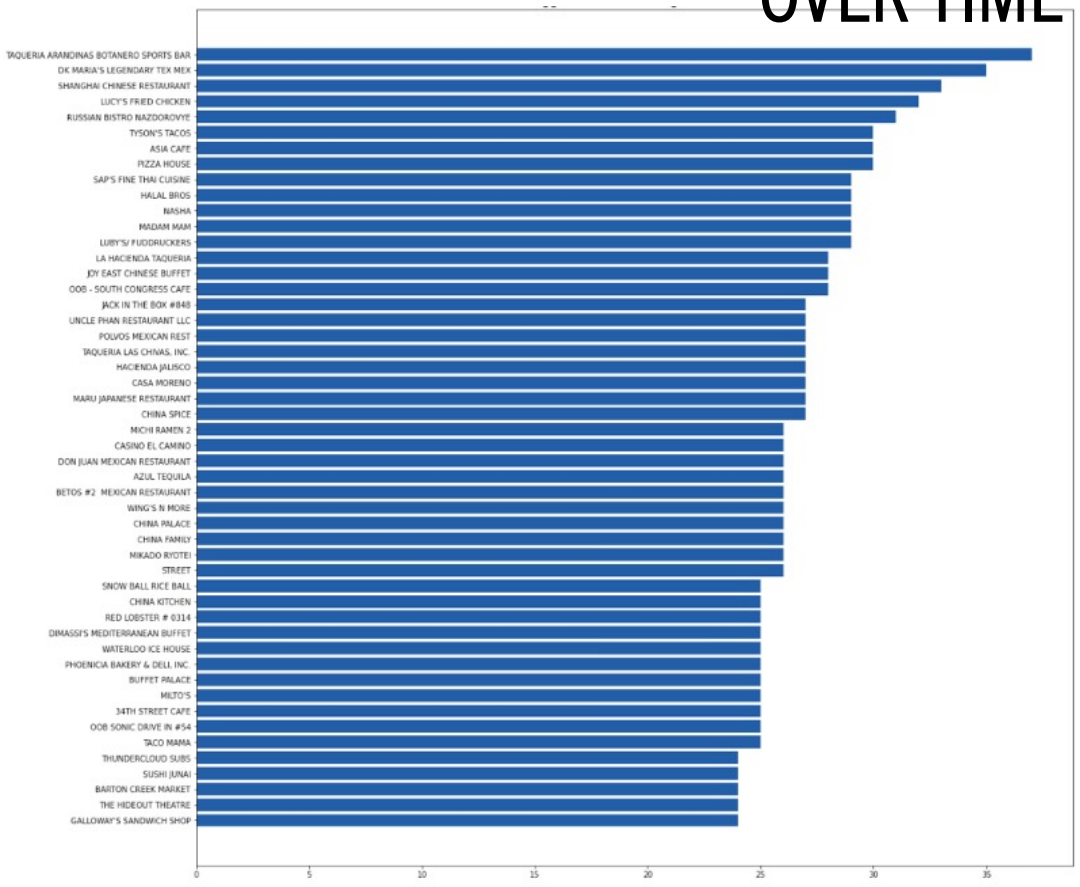
C: 10%

F*: 0.9%

* 70 OR BELOW



BUT AT THE RESTAURANT LEVEL WE DO SEE VARIATION OVER TIME

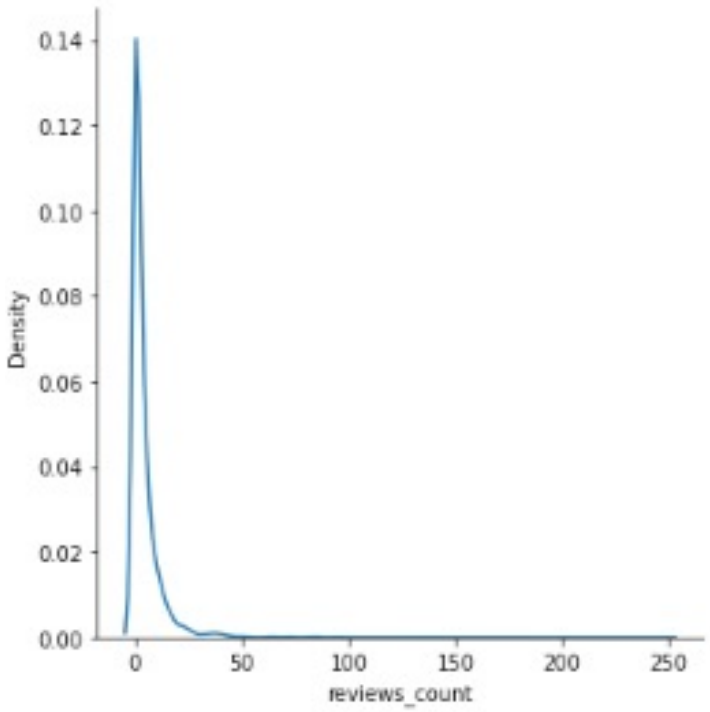


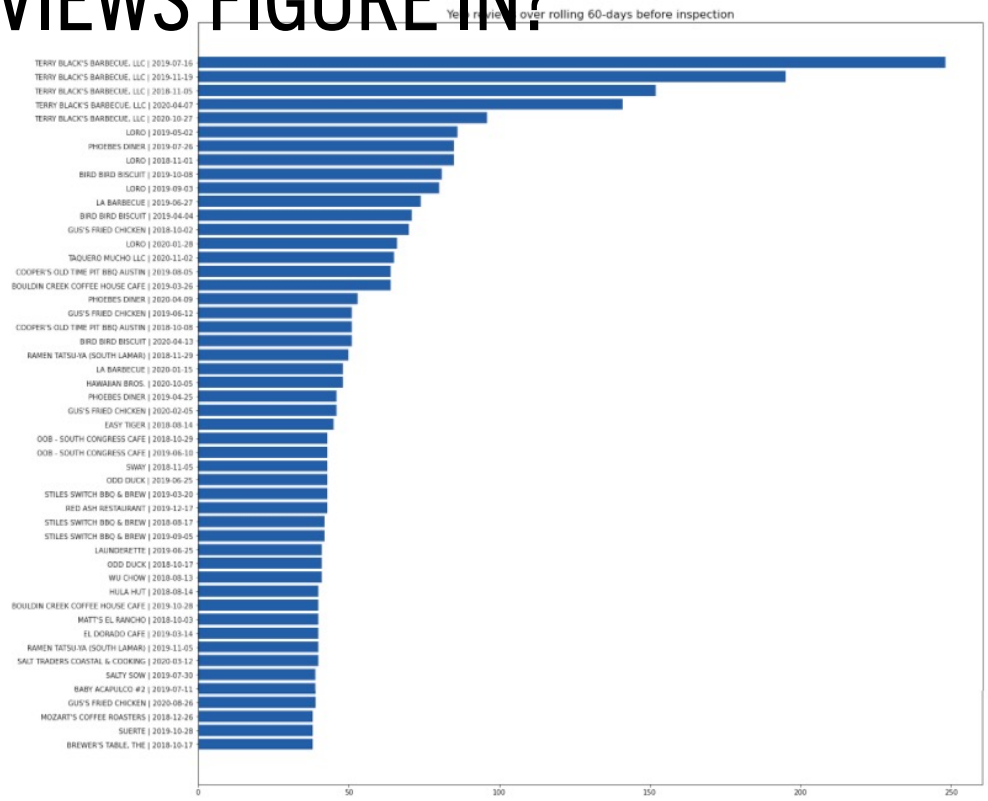
- IT'S A MIXED BAG IN TERMS OF RESTAURANTS SCORE OVER TIME

- THE LARGEST VARIATION (IN ABSOLUTE TERMS) OVER THE 3 YR PERIOD WAS 35 POINTS

WHAT CAN WE LEARN FROM THE YELP DATASET??

HOW DO REVIEWS FIGURE IN?





MOST RESTAURANTS DO NOT HAVE VERY MANY IF ANY REVIEWS DURING THE 60-DAY LOOKBACK WINDOW

THE COVERAGE IS NOT GREAT, BUT IT IS A RICH DATASET SO LET'S DIG A BIT DEEPER

SENTIMENT ANALYSIS


```
positive['text'][1:2].values
```

```
array(['Definitely one of my favorite places for vietnamese grub! Very authentic and although they
h, the people here are really nice. Will always recommend others to try it.'],
```

```
negative['text'][3:4].values
```

```
array(["I have been to this place a couple of times and each time, I never ceased to be disappointed.
par at best and the place is dirty.  One time a friend of mine went there and found a lizard (yes, that
ard) in his vermicelli bowl.  The service staff is not very helpful either.  They are hard to understan
em to be in a bad mood.  So if you enjoy paying too much for a can (yes, CAN) of soda with a cup of ice
ards in your food, I would recommend Pho Van."],
```

THIS ANALYSIS PRODUCED SOME INTEREST LEARNINGS BUT IT WOULD NEED TO REVISITED A BIT MORE DEEPLY FOR BETTER FEATURE EXTRACTION AS WELL AS TO GAUGE THE POTENTIAL FOR INFORMATION GAIN IN A SUBSEQUENT MODEL