Xtern 2021 - FoodieX
Data Analysis Report

Brief Data Analysis Report

As a Data Science intern for Xtern team, I analyzed the Indianapolis restaurant data to gather valuable insights for the success of Xtern's new venture, FoodieX, a food delivery platform that is aimed to provide food delivery service to people residing in Indianapolis during this tough time of COVID-19. In this report, I present the key findings I feel are relevant for the smooth operation of FoodieX.

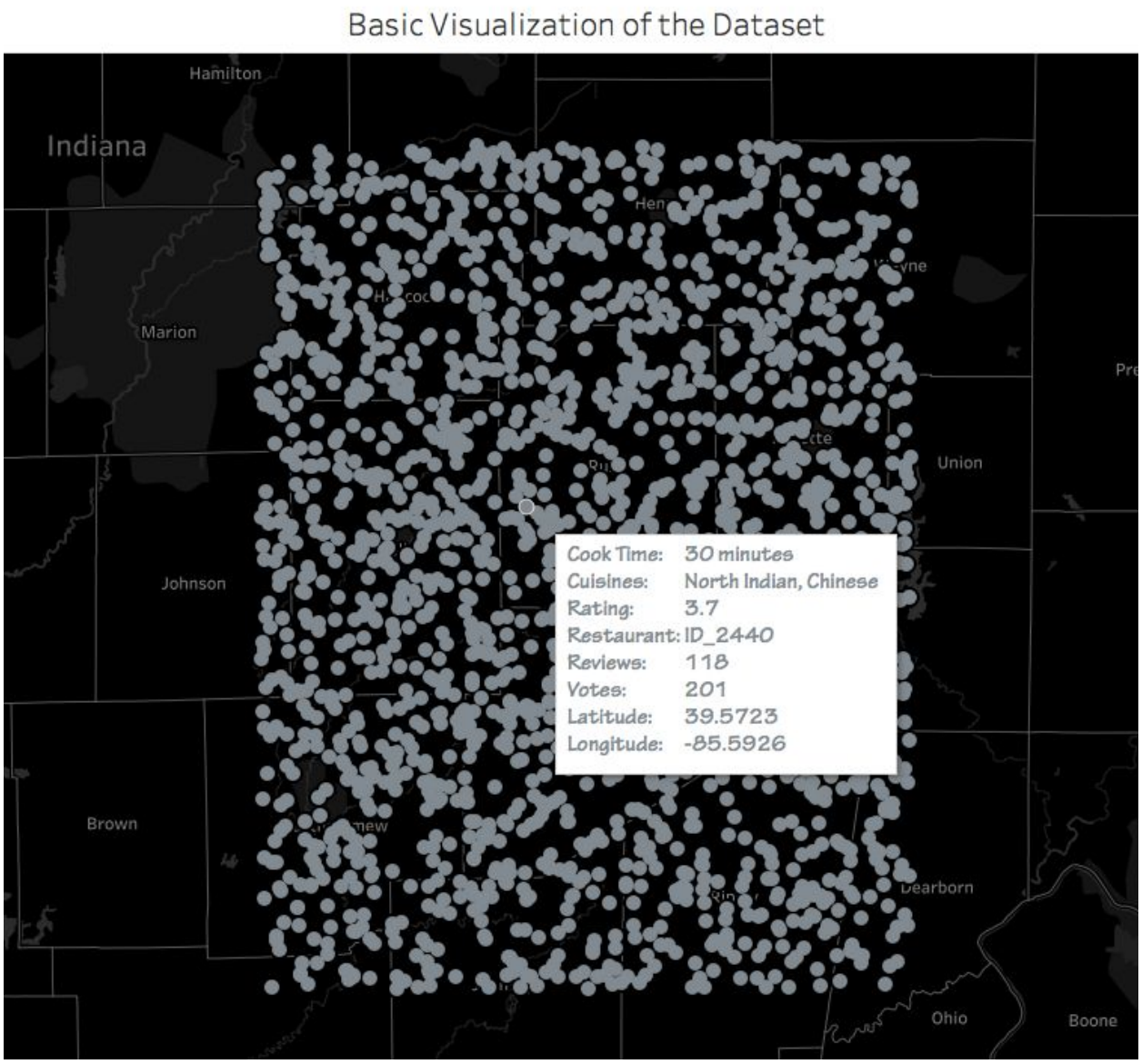
Here is the basic information about the dataset.

Number of Rows:
2019

Name of the columns:
'Restaurant',
'Latitude',
'Longitude',
'Cuisines',
'Average_Cost',
'Minimum_Order',
'Rating',
'Votes',
'Reviews',
'Cook_Time'

First Glance at the Dataset

As we can see, these rectangle-shaped location points are based in Indiana. Even though these location points are not exactly at Indianapolis, based on the description of the prompt, this analysis assumes that these locations are part of Indianapolis.



An instance of a data row can be viewed above, which highlights the data values for a restaurant with an id of 2440 which gives an idea of what constitutes the dataset. This visualization is created from Tableau. Just to show my full work, would really appreciate it if the Xtern team could once check out my Tableau visualization from this link. This interactive visualization is designed to show the geographic distribution of the restaurant based on various factors like rating, reviews, cook time, and votes.

<u>Questions that I was keen on answering from the dataset:</u>

Which are the trending and popular restaurants in Indianapolis?

Knowing the answer to this question is of particular significance to the FoodieX team since it is most likely that the maximum amount of orders are going to come from these places. Knowing the location of these places beforehand is crucial in optimizing our pick-up and delivery time.
Also, this piece of information might be useful to the operational team in deciding on the best route and setting up optimal pick-up zones.

I decided to write a simple algorithm to find out about the trendiest and popular restaurants in Indianapolis.

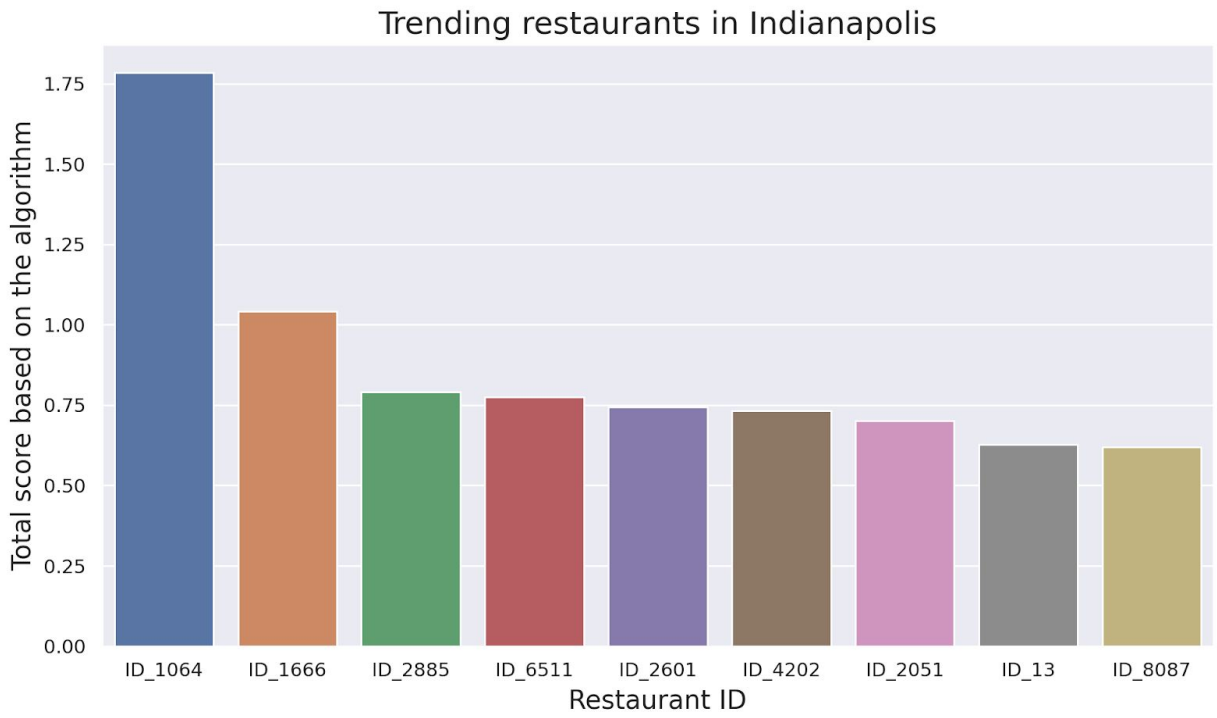
I figured that <***rating***, ***reviews***, *and* ***votes***> are the most significant variables from the dataset that could potentially answer this question.

I decided to create three additional columns **<*rating score***, ***reviews score***, and ***votes score***>
so that I could create another column **<total score>** which would be the sum of these three columns.

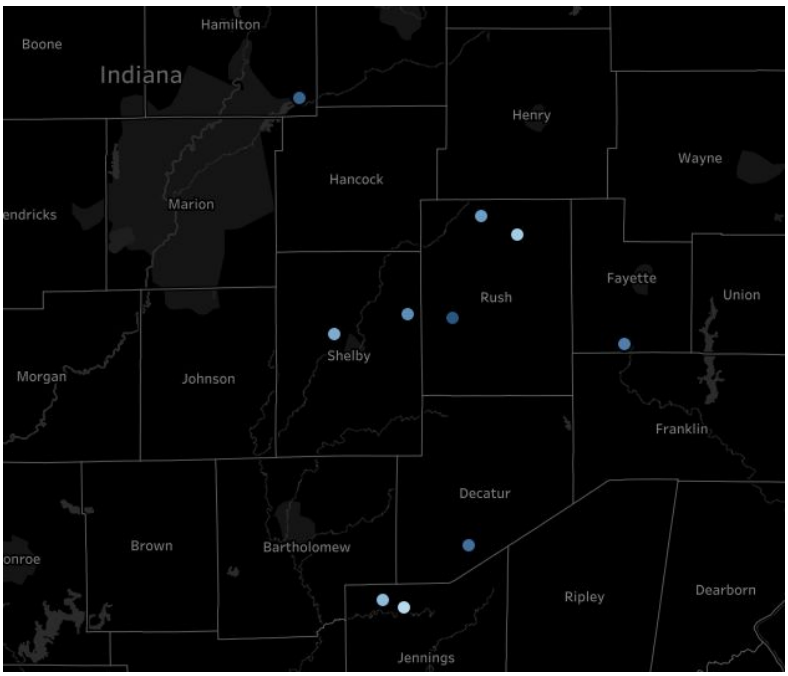
<rating score> is calculated as **<rating> / <sum rating> * 100 * 1/3**

I gave equal weight to each three of them i.e., 33.3333%

My analysis follows like this:



Tracking these 10 points on Tableau visualization shows that these 10 restaurants are scattered all around Indianapolis as shown below

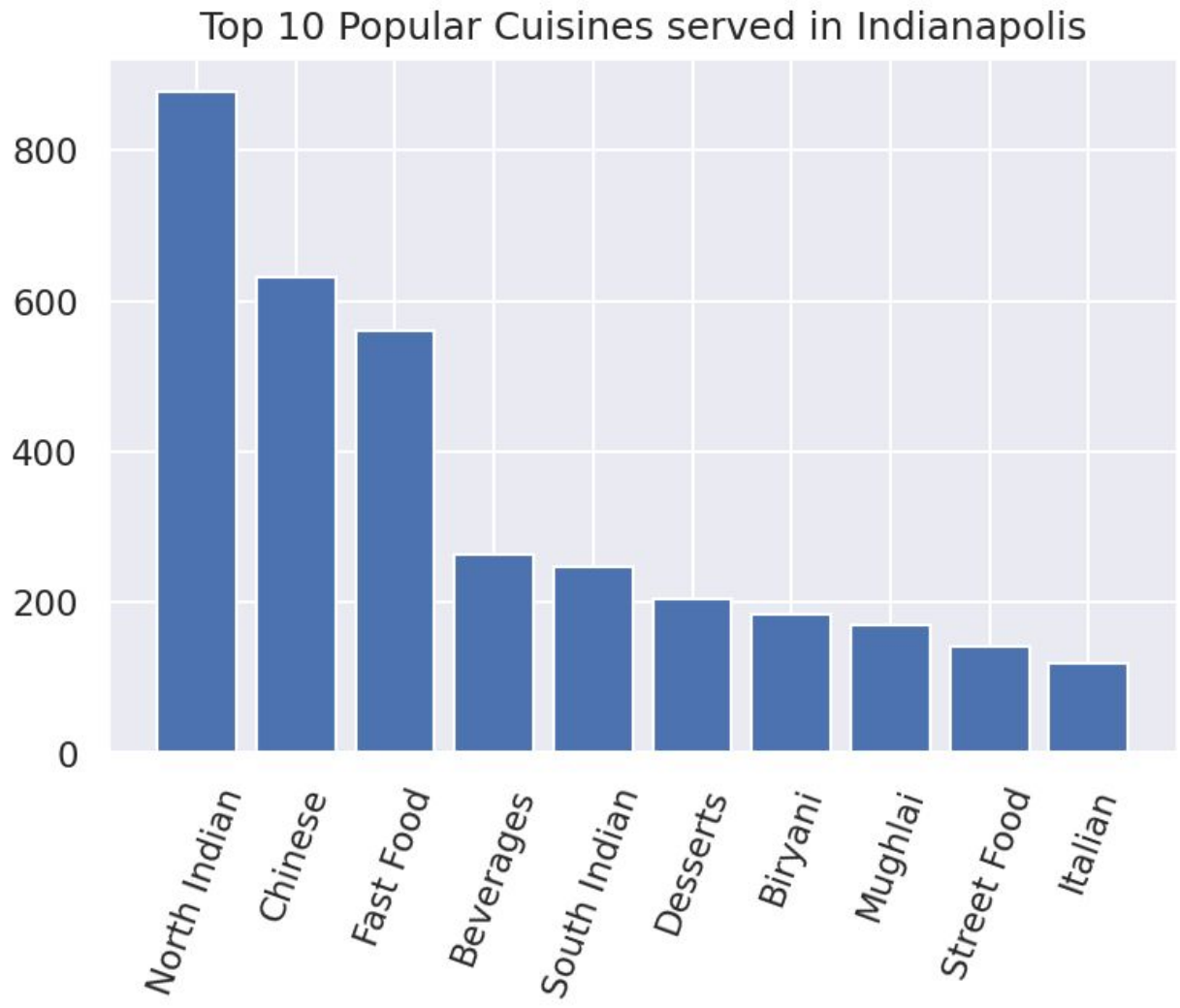


This concludes that it is very vital to have multiple pickup zones around the town that could need the consumer with as little delay as possible.

What are the most popular and most available cuisines in town?

Knowing the answer to this question could give FoodieX team a slight edge in the initial phase. Are these food at risk of spilling while driving? Are the market dominated by one particular type of food?

The result of the data goes this way.



It is interesting to see that sub-continental food gaining popularity in Indianapolis. On average, 'North Indian' and 'Chinese' food take relatively longer cook-time. For the hybrid restaurants with multiple cuisines, the expected cook-time for these foods might be underestimated. Therefore, it is important for

the FoodieX team to keep a note of this and constantly get themselves updated on popular and most demanded cuisines.

What would be the estimated cook-time for a particular cuisine?

This question is directly concerned with efficiency. It is not common in the food delivery business to have a time mismatch where one party has to spend idle waiting for some time before entering into a transaction with the other party. The sequences of these delays could result in a big business loss for both parties, i.e., restaurants and food delivery service.

I decided to test 3 machine algorithms for this task of potentially estimating the correct time to reduce the business loss.
The key variable for this estimation is the type of cuisines which is my predictive variable. Here, the target variable would be cooking time.

The result of my test follows like this:

Machine Learning Algorithm	Accuracy Score
Decision Tree	60%
Random Forest	63%
Logistic Regression	65%

These results are definitely not good enough to be used as a predictive model. However, these numbers do make sense since the dataset is pretty small with just 2019 data points. More data points are needed before we could deploy a model which gives a much higher accuracy rate.

However, it is key for the operational team to constantly keep track of the cooking time and keep a record so that it could be referenced in the future to get a relativity better estimate or even the data could be used to build a predictive model in the future.

Which and how many locations could act as pickup zones for FoodieX?

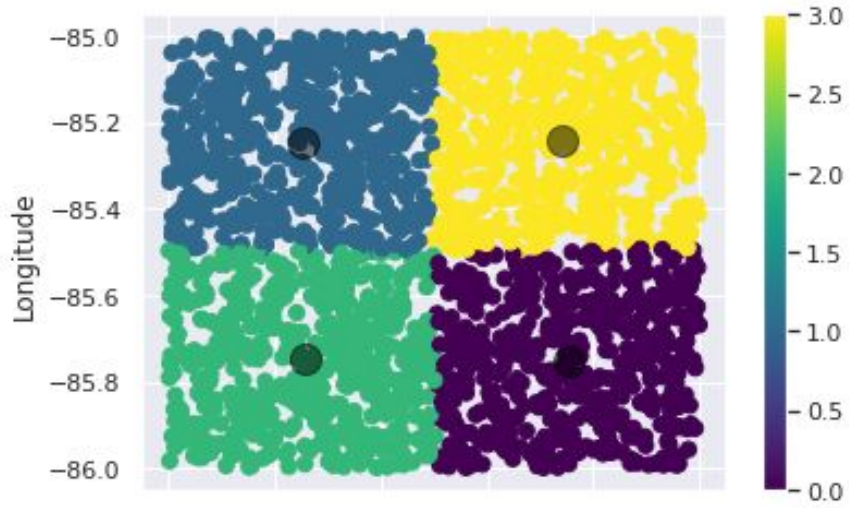
Another question which is important from both efficiency and customer service point of view.

Based on the elbow method, a mathematical way of determining the best number of clusters based on the dataset, for K-means Clustering, it appeared from my analysis that 4 would be the right amount of clusters for this dataset.

The 4 location points that are best as pickup zones as derived from the K-means clustering algorithm are as follows:

[[39.25413726 -85.75092214]
 [39.75333502 -85.74907981]
 [39.25492652 -85.24365223]
 [39.74026978 -85.24679707]]

where *x* is latitude and *y* is longitude as shown in the diagram below:



I hope this analysis was insightful to the Xtern team.