**1. Introduction**

**1.1 Background:**

Opening a restaurant, similar to any other business, involves as extensive decicision-making structure. Some decisions that are made can be reversed given time, such as a menu, a theme, and even a business strategy regarding the type of customers being catered to. However, one decision that can reasonably be considered permanent is the choice of location for a restaurant.

**1.2 The Problem:**

The decision to open a restaurant in a neighborhood is not a decision that should be taken lightly. Once picked, a bad location can be disastrous, necessitating a need to get everything not just alright, but perfect. Needless to say, picking a bad location has been the doom of many a restaurant, and is even touted as one of the major reasons restaurants fail[[1]](#endnote-1). Therefore, it stands to be questions: is it possible to use data to determine which locations are ideal for opening a restaurant?

**1.3 Intended Beneficiaries:**

This project is intended to benefit anyone that is looking to start operating a restaurant in a neighborhood in New York by providing them insights regarding those neighborhoods in New York where restaurants tend to be more successful.

**2. Data Description:**

**2.1 Data Sources:**

The data for this project has been sourced from the FourSquare API, particularly from the neighborhoods from around New York City. The data used herein has been sourced only from FourSquare, although the data pre-processing techniques used herein are easily applicable to a far more diverse range of data sources.

**2.2 Data Description:**

The dataset schema looks as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| venue.id | venue.name | venue.location.lat | venue.location.lng | Likes |
| String | String | Float | Float | Int |

**2.3 Data cleaning:**

Most of the data provided by the FourSquare API was in String format. For instance, the number of likes that each restaurant got was preceded by a string sequence. The first n characters were removed from all rows corresponding to the “likes” column in order to obtain numeric data. The latitude and longitude columns are float types and did not require any further cleaning. There was no missing data. Each venue had a valid name, id, co-ordinates, and likes. In total, there were 65 total venues on the FourSquare API database in NYC.

1. <https://www.thebalancesmb.com/ten-reasons-restaurants-fail-2888628> [↑](#endnote-ref-1)