# Predicting the most profitable food truck locations

Chris Carney

March 20, 2021

1. Introduction

## 1.1. Background

With summer rapidly approaching I am looking for opportunities to make extra money over the summer to pay for school. I would like to open a food truck that specializes in frozen treats. With the failure rate for food trucks being as high as 60%, I want to make sure I complete a sound business analysis. After all, it's about location, location, location.

## 1.2. Problem

I am not sure what location would be best for my food truck. I would like to find a location that is close to neighborhood parks, hoping that people will stop for a treat while they are walking their dog or exercising. Unfortunately, I am not willing to travel far from home, so the location will need to either be in my hometown of Nashville, TN or where I go to school in Knoxville, TN. In order to make sure I am not wasting my time I need to understand weather patterns that could discourage interest in frozen treats, as well as any COVID hotspots.

## 1.3. Audience

This analysis could be used by a number of people, but is mostly focused on owner/operators of food trucks in the central/eastern portion of Tennessee. The goal is to find the most profitable locations. Owner/Operators outside of Tennessee may find this analysis interesting as well, but they would need to update the data for their location.

2. Data Acquisition and Usage

## 2.1. Data Sources

I will use the /v2/venues/explore API to access venue data through Foursquare's Places API. The venue data will alert me to area parks, recreation centers, gyms, and yoga studios.

- Foursquare APIs - https://developer.foursquare.com/docs/places-api/endpoints/

Weather data will be scraped from the climate sections on the city Wikipedia pages. I will be focused on average temperature by month and rainfall data.

- Nashville: https://en.wikipedia.org/wiki/Nashville,_Tennessee
- Knoxville: https://en.wikipedia.org/wiki/Knoxville,_Tennessee

COVID data will be extracted from Tennessee's Department of Health. I am interested in the total number of new cases and the number of vaccines being administrated on a daily basis.

- COVID data - https://www.tn.gov/health/cedep/ncov/covid-19-vaccine-information.html

2.2. Data Usage

The majority of data will come from the Foursquare APIs. This dataset will provide me with the location data about various parks, recreation centers, gyms, and yoga studios. I would also like to pull in historical weather data to understand what months are ideal for frozen treats. Lastly, I would like to try to avoid COVID hotspots, so I will pull in current COVID numbers by county. By analyzing these datasets together I hope to identify high traffic parks with warm temperatures that would encourage people to be outdoors.

2.3. Feature selection

All three data sets I selected contained a large amount of data. Only some of that data is pertinent to my analysis. The following table shows the breakdown of features I have selected for further analysis.

| Data Source | Feature(s) selected | Feature(s) dropped | Comments |
|---|---|---|---|
| Weather | Average high °F (°C),<br>Average precipitation inches (mm),<br>Average precipitation days (≥ 0.01 in) | Record high °F (°C),<br>Mean maximum °F (°C),<br>Average low °F (°C),<br>Mean minimum °F (°C),<br>Record low °F (°C),<br>Average snowfall inches (cm),<br>Average snowy days (≥ 0.1 in),<br>Average relative humidity (%),<br>Mean monthly sunshine hours,<br>Mean daily sunshine hours,<br>Mean daily daylight hours,<br>Percent possible sunshine,<br>Average ultraviolet index | My focus was on daily averages in regards to high temperatures and rainfall. Temperature records were not valuable since they are single time occurrences, and snowfall data wasn't relevant since I was targeting summer months. Mean daily sunshine hours and ultraviolet index did not exist in both datasets, so were not available for comparison. |
| COVID - Public- | DATE<br>COUNTY<br>NEW_CASES | TOTAL_CASES<br>TOTAL_CONFIRMED<br>NEW_CONFIRMED | My focus was on new cases by date in two counties; Davidson, |

| | | | |
|---|---|---|---|
| Dataset-County-New | | TOTAL_PROBABLE<br>NEW_PROBABLE<br>POS_TESTS<br>NEW_POS_TESTS<br>NEG_TESTS<br>NEW_NEG_TESTS<br>TOTAL_TESTS<br>NEW_TESTS<br>NEW_DEATHS<br>TOTAL_DEATHS<br>NEW_RECOVERED<br>TOTAL_RECOVERED<br>NEW_ACTIVE<br>TOTAL_ACTIVE<br>NEW_INACTIVE_RECOVERED<br>TOTAL_INACTIVE_RECOVERED<br>NEW_HOSPITALIZED<br>TOTAL_HOSPITALIZED<br>TOTAL_DEATHS_BY_DOD | and Knox. I wasn't interested in testing data, hospitalizations, or deaths. |
| COVID – COVID VACCINE COUNTY SUMMARY | DATE<br>COUNTY<br>NEW_VACCINE_COUNT | VACCINE_COUNT<br>RECIPIENT_COUNT<br>NEW_RECIPIENT_COUNT<br>RECIP_FULLY_VACC<br>NEW_RECIP_FULLY_VACC | My focus was on the number of vaccines administered each day in Davidson and Knox counties. |

3. Exploratory Data Analysis

### 3.1. Analysis of the weather data

The theory was that weather would impact someone's desire whether to purchase frozen treats. I used the data contained in the city Wikipedia pages to see if there was one location that was hotter and dryer between the two. Ultimately, since the cities are only 180 miles apart the data was largely the same. Knoxville is close to the Smokey Mountains which probably had some impact on the difference. Based on the data Nashville was slightly warmer and dryer than Knoxville. My analysis was focused on the months of June, July, and August. A snapshot of Nashville's data can be seen below.

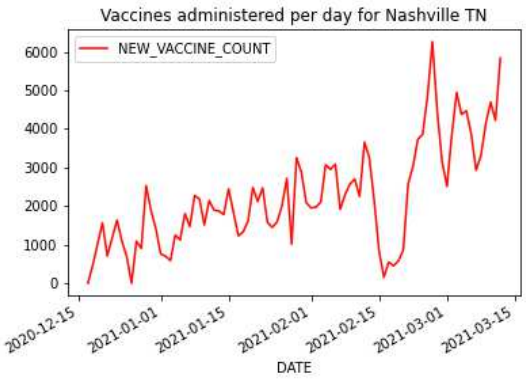
| Climate data for Nashville (Nashville Int'l), 1981–2010 normals,[b] extremes 1871–present[c] | | | | | | | | | | | | | [hide] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Year |
| Record high °F (°C) | 78 (26) | 84 (29) | 89 (32) | 91 (33) | 96 (36) | 109 (43) | 107 (42) | 106 (41) | 105 (41) | 99 (37) | 88 (31) | 79 (26) | 109 (43) |
| Mean maximum °F (°C) | 68.0 (20.0) | 73.0 (22.8) | 80.2 (26.8) | 85.1 (29.5) | 88.7 (31.5) | 94.2 (34.6) | 97.1 (36.2) | 96.5 (35.8) | 93.0 (33.9) | 85.5 (29.7) | 77.7 (25.4) | 68.5 (20.3) | 98.3 (36.8) |
| Average high °F (°C) | 46.9 (8.3) | 51.8 (11.0) | 61.0 (16.1) | 70.5 (21.4) | 78.2 (25.7) | 86.0 (30.0) | 89.3 (31.8) | 89.0 (31.7) | 82.4 (28.0) | 71.7 (22.1) | 60.3 (15.7) | 49.5 (9.7) | 69.8 (21.0) |
| Average low °F (°C) | 28.4 (−2.0) | 31.6 (−0.2) | 39.0 (3.9) | 47.5 (8.6) | 56.8 (13.8) | 65.4 (18.6) | 69.5 (20.8) | 68.4 (20.2) | 60.7 (15.9) | 48.9 (9.4) | 39.4 (4.1) | 31.3 (−0.4) | 49.0 (9.4) |
| Mean minimum °F (°C) | 8.8 (−12.9) | 13.5 (−10.3) | 22.5 (−5.3) | 31.2 (−0.4) | 42.7 (5.9) | 54.3 (12.4) | 61.6 (16.4) | 59.4 (15.2) | 45.4 (7.4) | 32.7 (0.4) | 23.8 (−4.6) | 13.8 (−10.1) | 4.7 (−15.2) |
| Record low °F (°C) | −17 (−27) | −13 (−25) | 2 (−17) | 23 (−5) | 34 (1) | 42 (6) | 51 (11) | 47 (8) | 36 (2) | 26 (−3) | −1 (−18) | −10 (−23) | −17 (−27) |
| Average precipitation inches (mm) | 3.75 (95) | 3.94 (100) | 4.11 (104) | 4.00 (102) | 5.50 (140) | 4.14 (105) | 3.64 (92) | 3.17 (81) | 3.41 (87) | 3.04 (77) | 4.31 (109) | 4.24 (108) | 47.25 (1,200) |
| Average snowfall inches (cm) | 2.6 (6.6) | 2.3 (5.8) | 0.9 (2.3) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | trace | 0 (0) | 0.5 (1.3) | 6.3 (16) |
| Average precipitation days (≥ 0.01 in) | 10.3 | 10.3 | 10.7 | 10.8 | 11.7 | 10.0 | 10.2 | 8.4 | 7.5 | 8.0 | 9.8 | 11.2 | 118.9 |
| Average snowy days (≥ 0.1 in) | 2.1 | 2.3 | 0.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | 1.0 | 6.2 |
| Average relative humidity (%) | 70.4 | 68.5 | 64.6 | 63.2 | 69.5 | 70.4 | 72.8 | 73.1 | 73.7 | 69.4 | 70.2 | 71.4 | 69.8 |
| Mean monthly sunshine hours | 139.6 | 145.2 | 191.3 | 231.5 | 261.8 | 277.7 | 279.0 | 262.1 | 226.4 | 216.8 | 148.1 | 130.6 | 2,510.1 |
| Mean daily sunshine hours | 4.5 | 5.2 | 6.2 | 7.7 | 8.4 | 9.3 | 9.0 | 8.5 | 7.5 | 7.0 | 4.9 | 4.2 | 6.9 |
| Mean daily daylight hours | 10 | 11 | 12 | 13 | 14 | 15 | 14 | 13 | 12 | 11 | 10 | 10 | 12 |
| Percent possible sunshine | 45 | 48 | 52 | 59 | 60 | 64 | 63 | 63 | 61 | 62 | 48 | 43 | 56 |
| Average ultraviolet index | 2 | 4 | 6 | 7 | 9 | 10 | 10 | 9 | 7 | 5 | 3 | 2 | 6 |
| Source 1: NOAA (relative humidity and sun 1961–1990)[81][89][90] | | | | | | | | | | | | | |
| Source 2: Weather Atlas (UV index)[91] | | | | | | | | | | | | | |

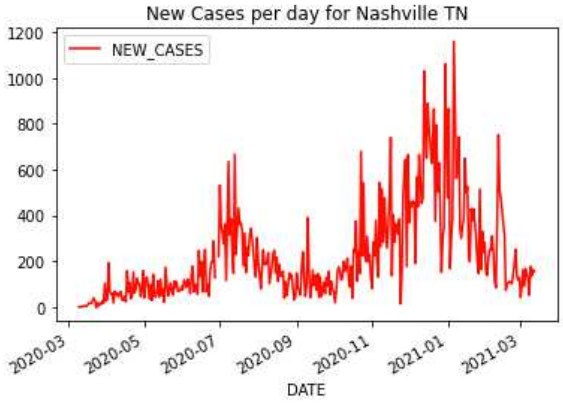
**Climate data pulled from [https://en.wikipedia.org/wiki/Nashville,_Tennessee#Climate](https://en.wikipedia.org/wiki/Nashville,_Tennessee#Climate)*

3.2. Analysis of the COVID datasets

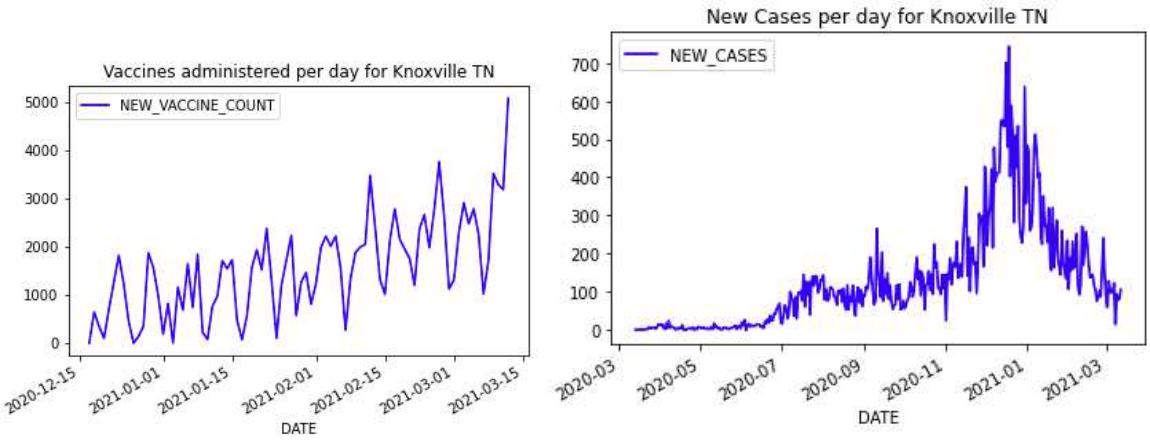
In general, the hypothesis was that cities seeing a drop in the new case count for COVID would be safer to do business in. As I reviewed the data I realized they had added the vaccine numbers, so I pulled that data and charted it for both cities using matplotlib.pyplot. The charts for each city are below. Ultimately, both cities were doing a good job administering the vaccine and controlling new cases of COVID. I selected Knoxville as the better city due to a significantly less count of new cases, but that could also be related to the difference in population size as well. Nashville's population is nearly 4 times greater than Knoxville.

Nashville's charts:




Knoxville's charts:




3.3. Analysis using K-means clustering with Foursquare Places API data

The goal with this analysis was to identify the neighborhoods that had the highest density of parks, recreation centers, gyms, and yoga studios. These venues are where I believe I will find my target customers. In order to do the analysis, I had to identify the neighborhoods in each city and track down their lat longs. Using the lat longs of the neighborhood data I called the Places API to return any venues within a radius of 500 meters.

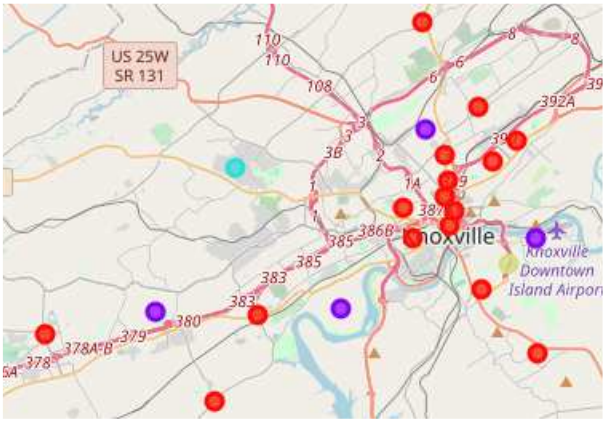
Once I had the data on venues for all the neighborhoods, I used one hot encoding to match the venue category with the neighborhood using 1 for present or 0 for not. With that data frame I could generate a mean score for each category by neighborhood. See table below.

| | Neighborhood | ATM | Accessories Store | American Restaurant | Antique Shop | Art Gallery | Arts & Crafts Store | Asian Restaurant | Athletics & Sports |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Antioch | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | Belle Meade | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | Bellevue | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.023256 | 0.023256 | 0.000000 |
| 3 | Bordeaux | 0.000000 | 0.000000 | 0.071429 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | Donelson | 0.055556 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 5 | East Nashville | 0.000000 | 0.000000 | 0.037037 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.037037 |
| 6 | Germantown | 0.000000 | 0.000000 | 0.043478 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 7 | Green Hills | 0.000000 | 0.000000 | 0.046875 | 0.000000 | 0.000000 | 0.000000 | 0.015625 | 0.000000 |
| 8 | Hermitage | 0.000000 | 0.000000 | 0.068966 | 0.000000 | 0.000000 | 0.000000 | 0.034483 | 0.000000 |

Once this was complete I was able to identify the top venues for each neighborhood as shown below.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Antioch | Mattress Store | Convenience Store | Comfort Food Restaurant | Mobile Phone Shop | Trail |
| 1 | Belle Meade | Gym | Sports Bar | Moving Target | Pool | Golf Course |
| 2 | Bellevue | Fast Food Restaurant | Pizza Place | Ice Cream Shop | Gas Station | Mexican Restaurant |
| 3 | Bordeaux | Fast Food Restaurant | River | Boutique | Discount Store | Shoe Store |
| 4 | Donelson | Pizza Place | ATM | Men's Store | Donut Shop | Coffee Shop |

I then used clustering to group similar neighborhoods allowing me to comb through the data easier. Once the neighborhoods were clustered I was able to easily identify a few targets neighborhoods for each city. Nashville revealed Woodbine, Belle Meade & Old Hickory as good targets, while Knoxville neighborhoods were Island Home Park, West Hills & Oakwood-Lincoln Park. Below is a map of the Clusters for each city.


4. Results

Based on my analysis I identified six target neighborhoods for my frozen treats food truck. These neighborhoods were selected based on their proximity to outdoor activities, weather patterns, and current COVID data. The preferred neighborhoods are listed below.

| City | Neighborhood |
|---|---|
| Nashville | Woodbine |
| Nashville | Belle Meade |
| Nashville | Old Hickory |
| Knoxville | Island Home Park |

| | |
|---|---|
| Knoxville | West Hills |
| Knoxville | Oakwood-Lincoln Park |

5. Discussion

This was an interesting experiment. As I analyzed the data, I realized that this could have benefited from additional datasets. Demographic data and crime statistics would be useful for one. Island Home Park in Knoxville sits right on top of the airport. That seems like a strange location to choose for a food truck, but I can't be sure without further analysis.

Since the summer is still 3 months away, it could have been useful to build a predictive model to forecast what COVID numbers might look like during the summer months.

6. Conclusion

In this study, I analyzed a number of factors (weather, COVID, venue categories) that could influence location for a frozen treats food truck. Using k-means clustering I was able to segment the neighborhoods into similar groups and plot them on a map. This analysis would be valuable to food truck operators trying to determine location, and with a few minor tweaks could be used to help support locations for winter months, or even rainy days based on different target venues. With the additional items mentioned in the discussion section I feel this could be a very accurate tool for prediction.