1. Data Sources and Description

The data used for the analysis was sourced through www.simplemaps.com. This website provided updated data (Jan 2021) on latitude and longitude for over 108,000 cities and towns in the US for all 50 states. The geodata was used in conjunction with the folium library to visualize the geographic locations of the counties. The data was read into the IDE via a CSV file.

Another website used was https://data.census.gov/cedsci/. This is the official website of the US Census Bureau which provided the population, population density and average income per household for the counties in Florida. The data was used to analyze each county using their attributes to identify the most suitable county for a startup. The data was read into IDE via a CSV file.

Foursquare is a social media mobile application used to discover various types of businesses in a geographic location. The application provided the trending venues by category for the cluster analysis model.

All of the datasets used in this project were checked for null/NAN values and duplicates. Columns with data not relevant to the scope of this project were removed (State Id, time zone, zip code, ranking, military, etc) The Foursquare category data was grouped by county and one hot encoded to convert the categorical data. Table 3.1 is a data dictionary for all of the datasets used in this project. Table 4.1 is the table created from www.simplemaps.com and https://data.census.gov/cedsci/

Table 3.1 – Data Dictionary

Term	Data Type	Description
County _name	Object	Name of the county
lat	Float64	Latitude of the object
lng	Float64	Longitude of the object
average income	Int64	Average income for each household in the county
population	Int64	Population of each county
Population density	Float64	Population density of each county
County	Object	Name of the county
County Latitude	Float64	Latitude of the county
County Longitude	Float64	Longitude of the county
Venue	String	Name of the venue
Venue latitude	Float64	Latitude of the venue
Venue Longitude	Float64	Longitude of the venue
Venue Category	String	Classification of the venue
Most Common Venue	String	10 most trending venues
Cluster Labels	Int64	The number of cluster in each county

Table 4.1 – County Data

	county_name	lat	lng	average income	population	population densty
0	Broward	26.1412	-80.1464	59547	1951866	2271.75
1	Collier	29.9807	-81.8108	69653	140250	1143.53
2	Duval	29.4499	-83.2818	55832	1227744	1183.75
3	Hillsborough	27.5908	-81.5081	48452	3855386	1814.40
4	Lee	28.9681	-81.6482	57832	1528309	807.05