Capstone Project - Week 5

S. Hennessy

Table of Contents

Introduction

Business Problem

A client is seeking to enter the food business in the State of Maine by establishing a restaurant operation. The client is a seasoned cook/chef with little exposure to business operations. Nevertheless, he wants to launch his dream of operating his own restaurant.

The client shared two primary goals: He wants to establish his business in an optimum location for the peak tourism season. He is already fired up for the 2019 season and plans to launch his business in the spring although he has not selected a region or city to pursue. He wants to promote the local fishing industry by offering a selection of appealing seafood items, yet, he is unsure whether he can create enough revenue solely from a seafood-based menu. From his culinary experience, he is confident that seafood is the optimum food item to market.

Background

The State of Maine is the most northeastern location in the USA.

At one time, Maine's primary source of revenue was based upon the harvesting of raw wood material through a significant Pulp and Paper industry. With the global collapse of the wood industry in the last decade or so and with the surge in local entrepreneurs, Maine diversified its work force into other streams and has recovered, somewhat, economically.

Despite the downturn in the forestry industry, two sources of revenue have remained consistent: Tourism and Aquaculture. By tapping into both of these revenue streams, the client is seeking to optimize the volume of hungry customers by having his new restaurant and mouthwatering seafood menu join the culinary landscape of dining options in Maine.

Maine is a State that embraces all four seasons. Its climate responds well to year-round tourism-related activities (*source:* visitmaine.com). For the first year of operations, the client has decided to concentrate on the peak tourism season, but also wants consideration to off season revenue, if possible.

Summer vacation traffic is almost entirely driven by visitors interested in viewing, touring and staying along Maine's Atlantic Ocean/Gulf of Maine coast line. This coast line is serviced by Maine's Route 1 highway and is the basis for determining the ideal location in this Capstone project.

Data

The source data will be built upon a collection of identified northeastern towns along Maine's Route 1, dotted by tourist attractions and fishing communities from Kittery to Calais, a distance of 316 miles along the southeastern coastline of Maine.

The starting point for this project is the initial list of town Route 1 exits, comprising 46 towns across 8 counties.

The data will be converted from files to data frames that are built, augmented, trimmed, and grouped. Those data frames will be used throughout this project in geo mapping, generating FourSquare API calls, building fresh data frames, analysis, top 10 venue lists, clustering with unsupervised machine learning (k-means) and final analysis augmented from files with government stats.

Methodology

The methodology used during this Capstone project is based upon course material covered in this course (Applied Data Science Capstone) and three prior courses (Python for Data Science, Data Analysis with Python, Data Visualization with Python), beginning with the use of Python as the language to facilitate Data Science activities. The depth of supporting libraries are major contributors in support of Python for Data Science. During this project, the following libraries were used for the listed reasons:

- Pandas: To enable data analysis
- numpy: To handle data in a vectorized manner
- matplotlib: To enable the plotting of graphs
- requests: To support requests (i.e. Get request for a JSON file)
- csv: To enable the import of csv files into Python
- folium: To enable map rendering within Python
- json: To handle JSON files
- sklearn: To support k-means clustering

Using Python and the imported libraries noted above, the following key activities were performed in the noted code sections to deliver:

Exploratory data analysis: Analysis of the initial mapping of towns along the route indicated the need to tighten geographic area (Code sections 2.1, 2.2 & 2.3). This led to focusing on the geo center of the first county (York) and mapping from that point. Analysis of that data revealed results (Code section 3) that skewed in favor of a town (Portland) outside of the county and an alternate approach was taken (Code 4) with the analysis of each town along Maine's Coastal Route 1. Using the first collection of towns, analyses of unique venues was performed (Code Sections 4.10 & 4.11).

Inferential statistical testing: Following the use of One Hot Encoding (OHE), mean was used to determine the frequency of venue occurrences with town groupings (Code Section 4.8 & 5.4).

Machine learning: OHE was used during the analysis in understanding of unique venues across the initial collection of towns (Code Section 4.6) and all towns (Code Section 5.3) to allow for statistical testing. Clustering (via k-means) was then used to perform unsupervised machine learning and segment the towns (Code Section 7) This allowed the elimination of the majority of towns and reduced further analysis down to seven outlier towns (Code Section 8).

Results

The results of this capstone project reflect the conversion, consumption, augmentation, cleaning, grouping, geo tagging and mapping, augmenting with FourSquare API JSON data, unsupervised machine learning (k-means) and final analysis augmented with files with government stats. For this Results section, the highlights will focus on illustrations of actions performed that led to this report's conclusion.

Figure 1 shows the results of building the Route 1 geo tagging dataframe.

Figure 2 illustrates the geo mapping of the towns being analyzed along Maine's coastal Route 1.

Figure 3 shows the geo mapping of the identified Route 1 towns and geo center of York.

Out[5]:

	Town	County	Latitude	Longitude
0	Kittery	York	43.088448	-70.736847
1	York	York	43.165944	-70.635096
2	Wells	York	43.322181	-70.580978
3	Kennebunk	York	43.384092	-70.545273
4	Biddeford	York	43.492584	-70.453384
5	Saco	York	43.500918	-70.442829
6	Scarborough	Cumberland	43.596226	-70.330056
7	South Portland	Cumberland	43.641472	-70.240881
8	Portland	Cumberland	43.661028	-70.254860
9	Falmouth	Cumberland	43.729525	-70.241993
10	Cumberland	Cumberland	43.655499	-70.259263

Figure 1 - Generation of Route 1 towns with Geo Tagging

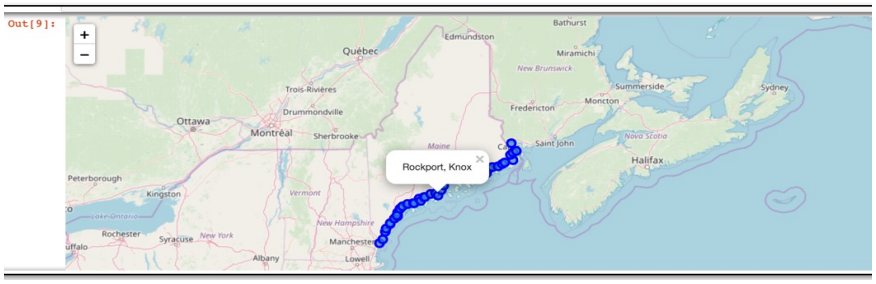

Figure 2 - Geo plotting of towns along coastal Route 1

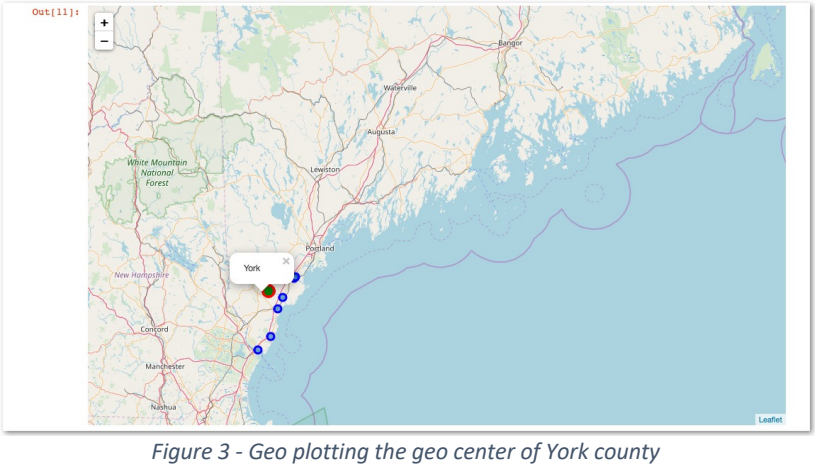

Figure 3 - Geo plotting the geo center of York county

Figure 4 is based upon the geo center of York county – queried FourSquare API data, converted JSON data and displayed in the dataframe.

Figure 5 displays results in a change in analysis toward all identified towns along Route 1 in York county.

ut[16]:

	name	categories	city	lat	lng
0	Village Tavern	American Restaurant	West Kennebunk	43.406402	-70.573385
1	Congdon's Doughnuts	Donut Shop	Wells	43.306432	-70.585085
2	Rachel Carson National Wildlife Refuge	Nature Preserve	Wells	43.347446	-70.548384
3	Wells Beach	Beach	Wells	43.302541	-70.566702
4	Rococo Artisan Ice Cream	Ice Cream Shop	Kennebunkport	43.362043	-70.476594
5	The Maine Diner	Diner	Wells	43.341794	-70.563127
6	Ogunquit Beach	Beach	Ogunquit	43.250342	-70.593937
7	Marginal Way Walk	Trail	Ogunquit	43.243314	-70.590440
8	Drakes Island Beach	Beach	Wells	43.321669	-70.553495
9	Drakes Island	Beach	Wells	43.325598	-70.551920
10	Palace Diner	Diner	Biddeford	43.492589	-70.454716
11	Goose Rocks Beach	Beach	Kennebunkport	43.399693	-70.410285
12	Elements: Books Coffee Beer	Café	Biddeford	43.494060	-70.458017
13	Perkins Cove	Harbor / Marina	Ogunquit	43.237309	-70.590848
14	Train Tavern	Cocktail Bar	Lebanon	43.415895	-70.864733
15	Mabel's Lobster Claw	Seafood Restaurant	Kennebunkport	43.349169	-70.472348
16	The Ramp Bar & Grill	American Restaurant	Kennebunkport	43.366133	-70.431636
17	Barreled Souls Brewing	Brewery	Saco	43.523646	-70.427539

Figure 4 - York county dataframe

4.4 - Display the resulting dataframe

In [19]:
```
print(local_county_venues.shape)
local_county_venues
```

```
(600, 7)
```

Out[19]:

	Town	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kittery	43.088448	-70.736847	Tributary Brewing Company	43.090371	-70.735942	Brewery
1	Kittery	43.088448	-70.736847	The Black Birch	43.085805	-70.744154	Gastropub
2	Kittery	43.088448	-70.736847	Lil's	43.085915	-70.743462	Coffee Shop
3	Kittery	43.088448	-70.736847	Anju Noodle Bar	43.085812	-70.743627	Asian Restaurant
4	Kittery	43.088448	-70.736847	Tulsi	43.086387	-70.745017	Indian Restaurant
5	Kittery	43.088448	-70.736847	Beach Pea Baking Company	43.090030	-70.749025	Bakery
6	Kittery	43.088448	-70.736847	Prescott Park	43.077172	-70.752168	Park
7	Kittery	43.088448	-70.736847	Loco Coco's Tacos	43.087291	-70.747909	Mexican Restaurant
8	Kittery	43.088448	-70.736847	Moxy American Tapas Restaurant	43.077792	-70.756757	Tapas Restaurant
9	Kittery	43.088448	-70.736847	Ceres Bakery	43.077058	-70.756209	Bakery

Figure 5 - All towns in York county

Figure 6 built on the analysis from Figure 5 which then allowed for determining the frequency of venues, per town, in York county.

4.8 - Now determine the frequency of occurrence of each category when grouped by towns in the county

In [23]:
```
county_grouped = county_onehot.groupby('Town').mean().reset_index()
county_grouped
```

Out[23]:

	Town	Accessories Store	American Restaurant	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bar	Baseball Field	Beach	Bed & Breakfast	Board Shop	Bookstore	Bowling Alley	Breakfast Spot
0	Biddeford	0.00	0.04	0.00	0.01	0.01	0.00	0.02	0.02	0.01	0.09	0.00	0.00	0.00	0.01	0.02
1	Kennebunk	0.00	0.10	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.08	0.04	0.00	0.00	0.00	0.02
2	Kittery	0.00	0.03	0.00	0.01	0.02	0.01	0.04	0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.01
3	Saco	0.00	0.04	0.00	0.01	0.01	0.00	0.02	0.01	0.01	0.10	0.00	0.00	0.00	0.01	0.02
4	Wells	0.00	0.09	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.12	0.00	0.01	0.00	0.00	0.03
5	York	0.01	0.15	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.05	0.02	0.00	0.00	0.00	0.02

Figure 6 - Venue frequency by town in York county

Figure 7 shows the next result as the Top 10 listing by towns in York county.

Figure 8 highlights, from the dataframe, the Top 10 list displayed for each town in York county.

4.10 - Let's examine each town along with the top 10 most common venues

In [25]:
```
num_top_venues = 10

for Town in county_grouped['Town']:
    print("----"+Town+"----")
    temp = county_grouped[county_grouped['Town'] == Town].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----Biddeford----
                venue  freq
0               Beach  0.09
1      Sandwich Place  0.05
2      Ice Cream Shop  0.05
3          Donut Shop  0.04
4   Seafood Restaurant  0.04
5         Coffee Shop  0.04
6          Campground  0.04
7         Pizza Place  0.04
8  American Restaurant  0.04
9             Brewery  0.03
```

Figure 7- Top 10 listings for York county

Next create a dataframe to display the top 10 for each town in the county

In [27]:
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Town']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
towns_venues_sorted = pd.DataFrame(columns=columns)
towns_venues_sorted['Town'] = county_grouped['Town']

for ind in np.arange(county_grouped.shape[0]):
    towns_venues_sorted.iloc[ind, 1:] = return_most_common_venues(county_grouped.iloc[ind, :], num_top_venues)

towns_venues_sorted
```

Out[27]:

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Biddeford	Beach	Sandwich Place	Ice Cream Shop	American Restaurant	Campground	Donut Shop	Pizza Place	Coffee Shop	Seafood Restaurant	Café
1	Kennebunk	Seafood Restaurant	American Restaurant	Beach	Hotel	Bed & Breakfast	Pizza Place	Deli / Bodega	Brewery	Coffee Shop	Restaurant

Figure 8 - Top 10 from York county dataframe

Figure 9 is based on the analysis, at this point, that led to the change in approach for gathering the Top 10 list for all identified towns along the Maine coast of Route 1 (2353 venues).

5.2 - Display the resulting dataframe for all identified towns along the Maine coast on Route 1

In [31]:
```
print(Route1_venues.shape)
Route1_venues
```

```
(2354, 7)
```

Out[31]:

	Town	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kittery	43.088448	-70.736847	Tributary Brewing Company	43.090371	-70.735942	Brewery
1	Kittery	43.088448	-70.736847	The Black Birch	43.085805	-70.744154	Gastropub
2	Kittery	43.088448	-70.736847	Lil's	43.085915	-70.743462	Coffee Shop
3	Kittery	43.088448	-70.736847	Anju Noodle Bar	43.085812	-70.743627	Asian Restaurant
4	Kittery	43.088448	-70.736847	Tulsi	43.086387	-70.745017	Indian Restaurant
5	Kittery	43.088448	-70.736847	Beach Pea Baking Company	43.090030	-70.749025	Bakery
6	Kittery	43.088448	-70.736847	Prescott Park	43.077172	-70.752168	Park
7	Kittery	43.088448	-70.736847	Loco Coco's Tacos	43.087291	-70.747909	Mexican Restaurant
8	Kittery	43.088448	-70.736847	Moxy American Tapas Restaurant	43.077792	-70.756757	Tapas Restaurant

Figure 9 - Top 10 list for all towns on Coastal Route 1

Figure 10 builds on the change in analysis that facilitated the identification of the frequency of venues of all towns along the Maine coast of Route 1.

Figure 11 shows the result that underscored the client's certainty that 'seafood restaurant' was the leader for venues along the coast of Maine.

5.4 - Now determine the mean of the frequency of occurrence of each category when grouped by towns along Route 1

```
In [34]: route1_grouped = route1_onehot.groupby('Town').mean().reset_index()
         route1_grouped
```

Out[34]:

	Town	Accessories Store	Airport	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Auto Workshop	Automotive Shop	BBQ Joint
0	Bath	0.000000	0.000000	0.000000	0.047619	0.00	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.015873
1	Belfast	0.000000	0.000000	0.000000	0.037037	0.00	0.018519	0.000	0.000000	0.000000	0.000000	0.000000	0.018519	0.000000
2	Biddeford	0.000000	0.000000	0.000000	0.040000	0.00	0.000000	0.000	0.000000	0.010000	0.010000	0.000000	0.000000	0.000000
3	Brunswick	0.000000	0.000000	0.000000	0.040816	0.00	0.000000	0.000	0.000000	0.000000	0.010204	0.010204	0.000000	0.000000
4	Bucksport	0.000000	0.000000	0.000000	0.076923	0.00	0.038462	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Calais	0.000000	0.000000	0.000000	0.050000	0.00	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Cherryfield	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Columbia Falls	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 10 - Frequency of venues across all towns

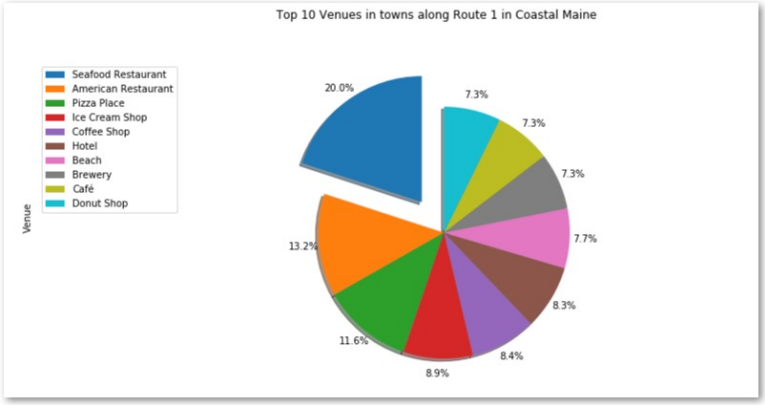

Figure 11 - Top 10 venues along Maine's Coastal Route 1

The next significant project result was the merging and creation of the clustering dataframe.

7.2 - Merge dataframes into a new dataframe for clustering

```
In [43]: route1_merged = df_RT1_Maine

         # add clustering labels
         route1_merged['Cluster Labels'] = kmeans.labels_

         # merge route1_grouped with route1_data to add latitude/longitude for each neighborhood
         route1_merged = route1_merged.join(route1_venues_sorted.set_index('Town'), on='Town')

         route1_merged.head() # check the last columns!
```

Out[43]:

	Town	County	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Kittery	York	43.088448	-70.736847	1	Seafood Restaurant	Pizza Place	Bakery	Brewery	Café	American Restaurant	Italian Restaurant
1	York	York	43.165944	-70.635096	1	Seafood Restaurant	American Restaurant	Pizza Place	Beach	Hotel	Candy Store	Ice Cream Shop
2	Wells	York	43.322181	-70.580978	1	Seafood Restaurant	Beach	American Restaurant	Pizza Place	Hotel	Café	Ice Cream Shop
3	Kennebunk	York	43.384092	-70.545273	1	Seafood Restaurant	American Restaurant	Beach	Hotel	Bed & Breakfast	Pizza Place	Brewery

Figure 12 - Clustering dataframe

Use of the clustering dataframe resulted in the generation and mapping of k-means clustering as shown in Figure 13.

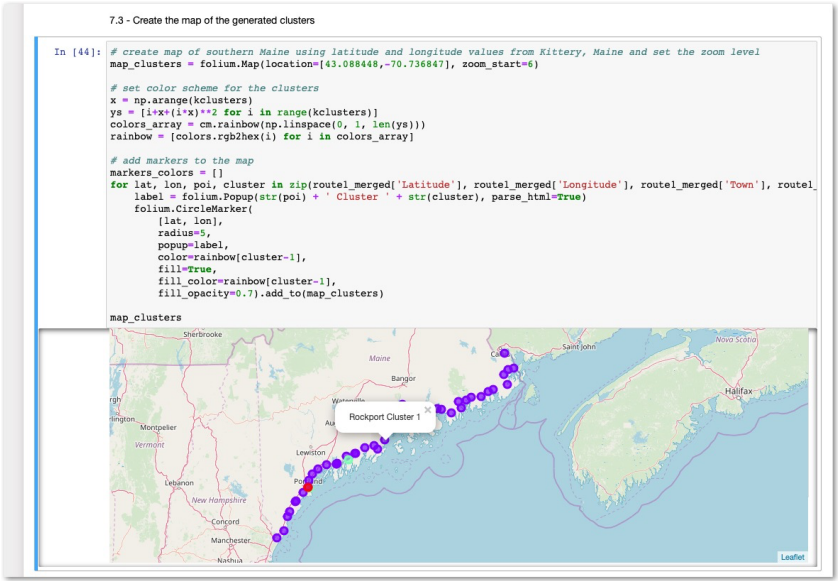

Figure 13 - Clustering (k-means) and Mapping

Cluster 1 was excluded from the final analysis based upon various iterations and quantity of Clustering performed. The seven remaining outlier towns are from clusters 0, 2, 3, and 4.

Discussion

Based on the top 10 venue lists from clusters 0, 2, 3, and 4, one town emerges that does not list Seafood Restaurant. The town is Scarborough, which, however, is mitigated by the fact that its location is immediately adjacent to Portland, a major city in Maine. With Cumberland, Scarborough & South Portland in a relatively densely populated area (within 10 miles of the City of Portland and having a significant number of Seafood Restaurants), these communities are removed from consideration.

The elimination of Scarborough reduced the final towns to four for review and selection of the optimum location. The towns are Belfast, Edgecomb, Rockport and Stockton Springs.

Since tourism is the major contributor to revenue for a venue of this type, sites of interest play a key role. (Order of towns with most to least sites of interest are: Rockport-8; Belfast-5; Stockton Springs-2; Edgecomb-0. *source*: Wikipedia.)

Additional details on town makeup were needed to determine the viability of year round operations in addition to tourism revenue.

To obtain data as up to date as possible, the Maine Department of Administrative and Financial Services for Economic & Demographics on the Government of Maine web portal was accessed (http://econ.maine.gov/index/build). From this page, details were extracted into CSV files.

Three areas were selected in determining the year-round viability for running a seafood restaurant. The belief is that a higher level of household income, retail sales and higher educational attainment are key determining factors. Specifically, the factors retrieved for each town are:

- Income: 2016 Median Household
- Taxable Retail Sales Quarterly (2018 Q1, Q2 & Q3) by County
- Educational Attainment

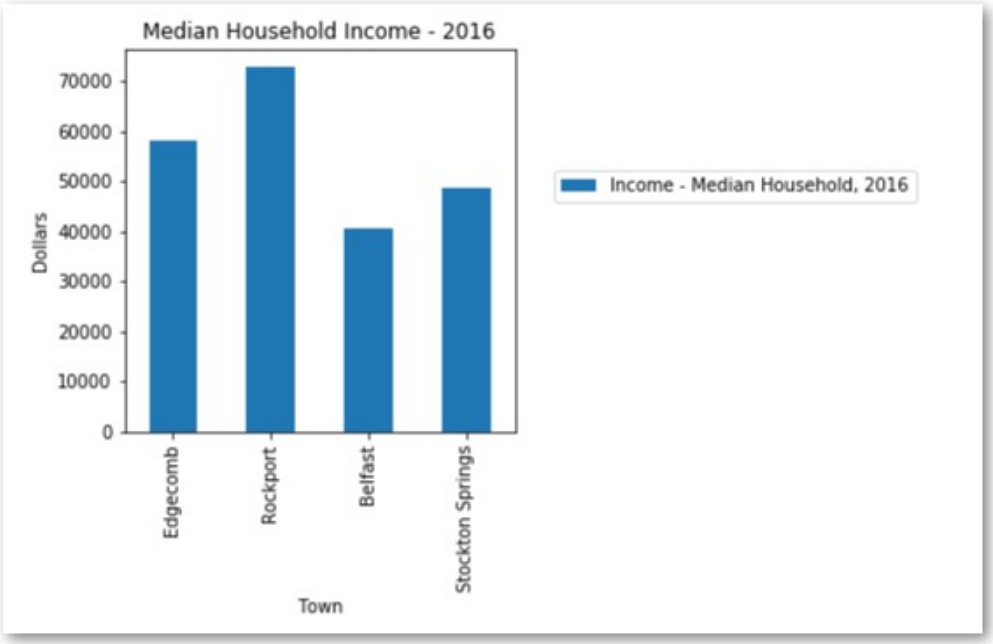

Figure 14 - Median Household Income (2016)

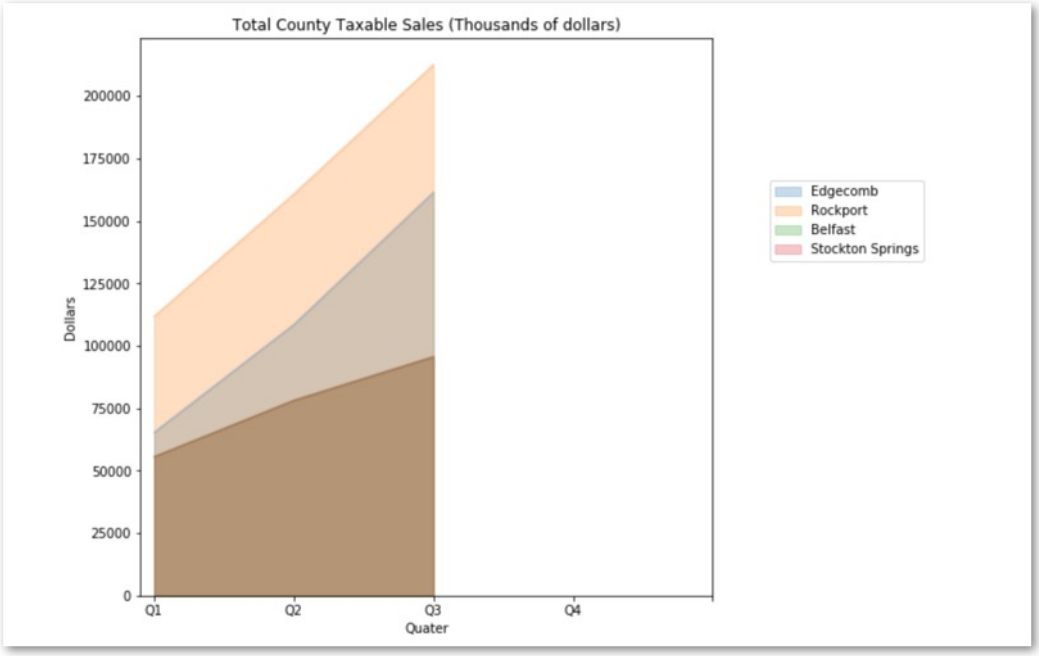

Figure 15 - Total County Taxable Sales (2018 -Q1, Q2 &Q3)

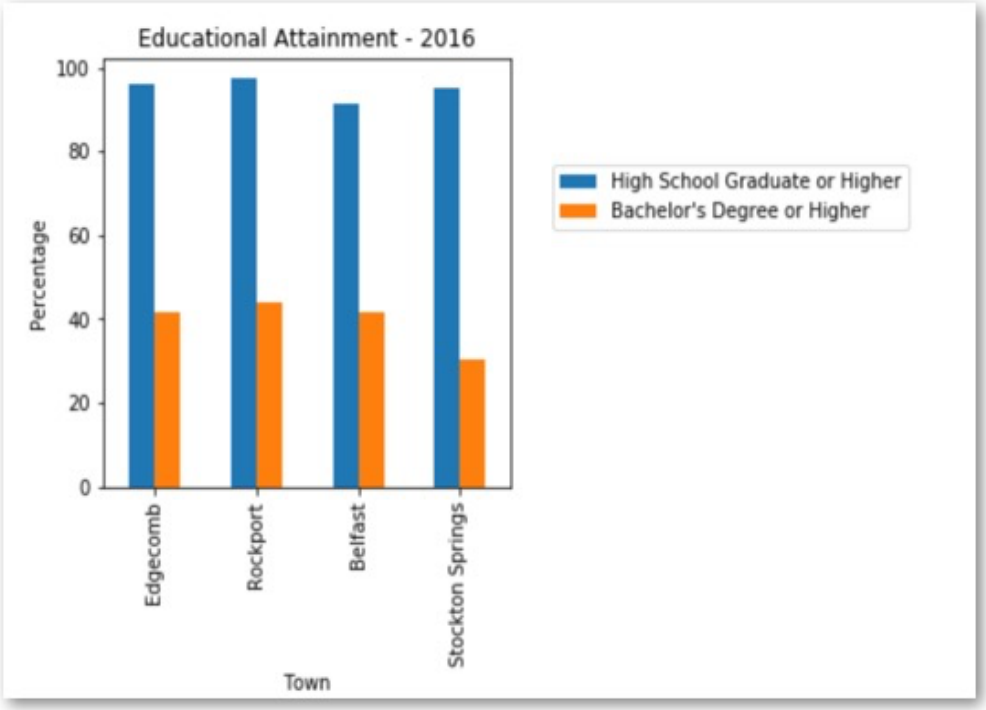

Figure 16 - Education Attainment (2016)

Conclusion

Based upon results, it is important to again review the client's primary goal: to establish a restaurant in an optimum location for the peak tourism season and, if possible, to consider off season revenue.

With supporting data from the Government of Maine website, converting that data to CSV files, as well as pulling relevant FourSquare data into various data frames, the clear choice for town selection is Rockport, Maine. The town has the following factors in its favor:

- Greatest number of points of interest for tourism
- Generated largest sales volume for the first three quarters of 2018
- Largest median income
- Highest level of educational attainment

The quantity of points of interest leads to greater likelihood of capturing tourism traffic in the local area during the summer months. During off season from the tourism revenue stream, we would want to have a stable economic base with likelihood of increased disposable income for dining at a restaurant. Again, Rockport leads against the four finalists.

In closing, Rockport is selected and recommended as the preferred client location for a Seafood Restaurant!