

# Capstone Project - Week 5

S. Hennessy



# Agenda

- Executive Summary
- Data
- Methodology
- Results
- Discussion
- Conclusion



# Executive Summary

## Background:

The State of Maine is the most northeastern location in the USA. Despite recent economic instability, two sources of revenue have remained consistent: Tourism and Aquaculture.

## Business Case:

A client is seeking to enter the food business in the State of Maine by establishing a Restaurant operation. The client believes a seafood restaurant would be the most popular, is seeking to identify the ideal town for tourism season and would also support off season operations.

## Data:

Based upon the desire to maximize tourism opportunities for restaurant operations, 46 towns across 8 counties along Maine's Coastal Route 1 were selected for this project. (a distance of 316 miles along the southeastern coastline of Maine)

# Data – 1 of 2

The source data was built upon a collection of identified northeastern Towns along Maine's Route 1, from Kittery to Calais, a distance of 316 miles along the southeastern coastline of Maine.

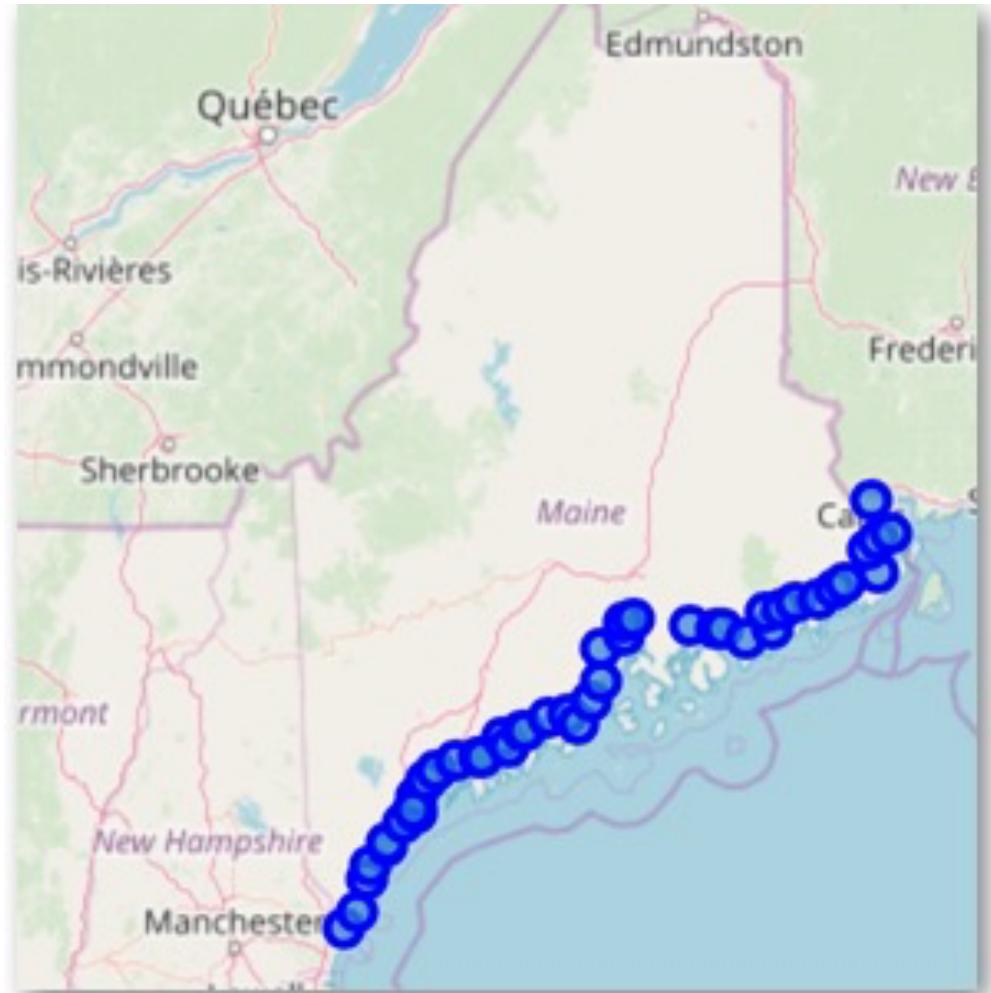
The initial list of town exits from Coastal Maine Route 1 with 46 towns across 8 counties is the starting point for this project.

	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

# Data – 2 of 2

The data was converted from files to data frames that were built, augmented, trimmed, and grouped.

The data frames were used in geo mapping, generating API calls, analysis, top 10 venue lists, clustering with unsupervised machine learning (k-means) and final analysis augmented with government stats.



# Methodology – 1 of 2

## Course Content:

The methodology used during this Capstone project is based upon course material covered in this course (Applied Data Science Capstone) and the three prior courses (Python for Data Science, Data Analysis with Python, Data Visualization with Python).

## Python & Libraries:

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

# Methodology – 2 of 2

## **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 Route1. 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 for the elimination of the majority of towns and reduced further analysis to seven outlier towns (Code Section 8).

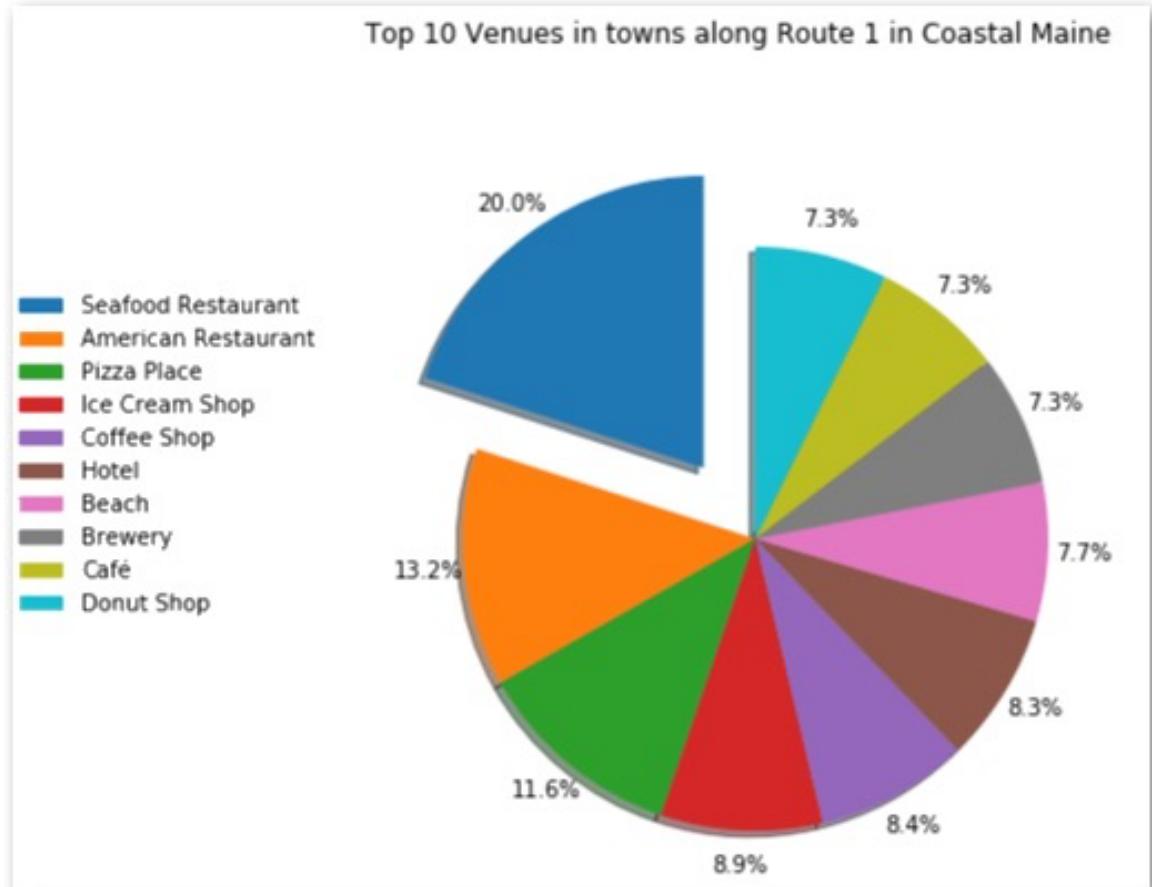
# Results – 1 of 2

## Validation of assumption

The first major result was from analysis in gathering the Top 10 list for all 2353 venues in the identified towns along the Maine coast of Route 1.

This result facilitated the identification of the frequency of venues of all towns along the Maine coast of Route 1.

Subsequently, that resulted in validating the client's belief that 'seafood restaurant' was the leader for venues.



# Results – 2 of 2

## Clustering

Utilizing the top 10 analysis, the next significant project result was the merging and creation of the clustering dataframe.

This clustering dataframe was then leveraged in the unsupervised machine learning and identification of five distinct clusters for further analysis.

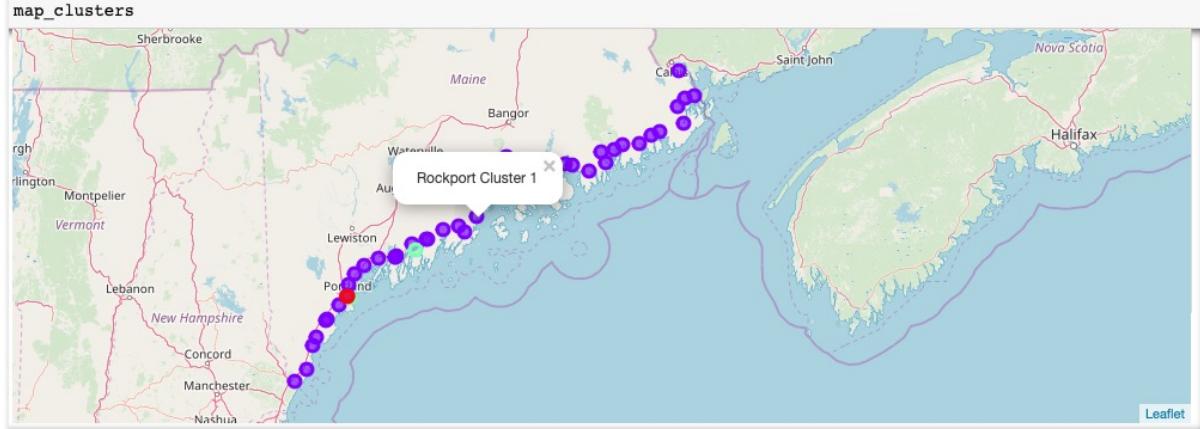
### 7.3 - Create the map of the generated clusters

```
# create map of southern Maine using latitude and longitude values from Kittery, Maine and set the zoom level
map_clusters = folium.Map(location=[43.088448,-70.736847], zoom_start=6)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i*x+(i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(route1_merged['Latitude'], route1_merged['Longitude'], route1_merged['Town'], route1_merged['Cluster']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



# Discussion – 1 of 2

## Outliers

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

Based upon the top 10 venue lists from clusters 0, 2, 3, & 4, Cumberland, Scarborough & South Portland are in relatively densely populated area within 10 miles of Portland and removed from consideration for selection.

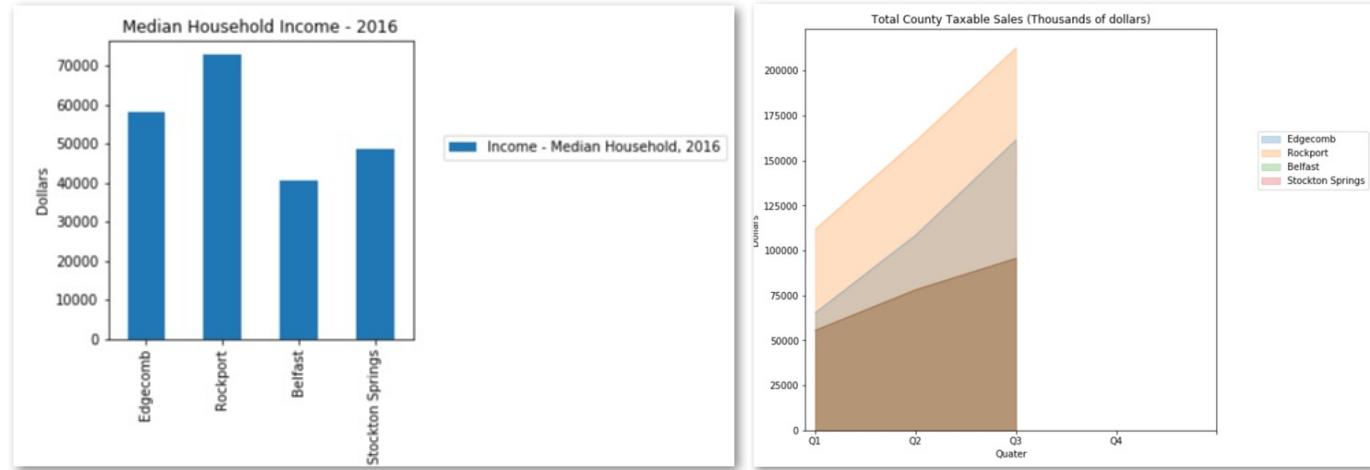
The final towns were reduced to four for review and selection of the optimum location of the seafood restaurant: Belfast, Rockport, Edgecomb and Stockton Springs.

	Town	County	Latitude	Longitude	Cluster Labels
6	Scarborough	Cumberland	43.596226	-70.330056	1
7	South Portland	Cumberland	43.641472	-70.240881	3
10	Cumberland	Cumberland	43.655499	-70.259263	0
18	Edgecomb	Lincoln	43.958413	-69.630602	3
24	Rockport	Knox	44.184524	-69.076149	1
26	Belfast	Waldo	44.426119	-69.006736	4
27	Stockton Springs	Waldo	44.489520	-68.856976	2

# Discussion – 2 of 2

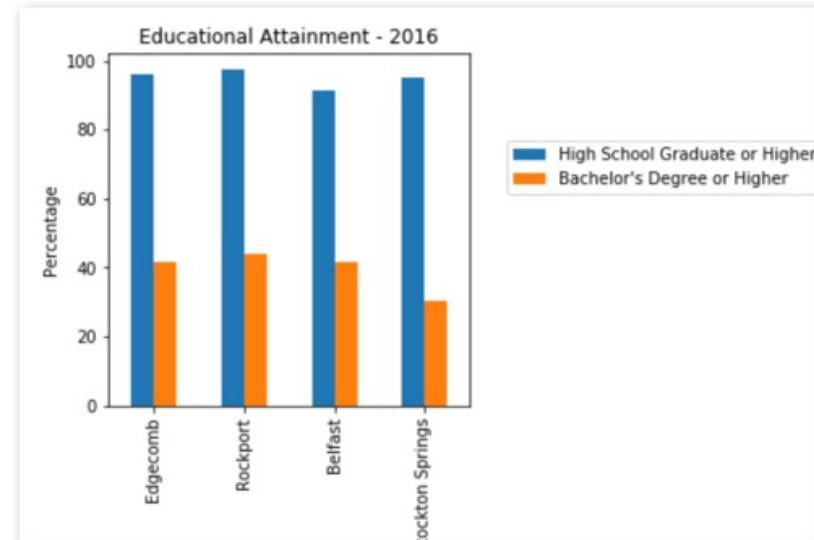
## Government Data

The belief is that a higher level of household income, retail sales and higher educational attainment are key determining factors in off season revenue generation for a restaurant.



Those factors were obtained from the Government of Maine portal\* and retrieved for each town

- Income: Median Household
- Taxable Retail Sales Quarterly
- Educational Attainment



\*Source: <http://econ.maine.gov/index/build>

# 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.

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.

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 clear choice for town selection is Rockport, Maine.