Enterprise Relocation Analysis

Kevin Garvey

1. Introduction

An online automotive parts business is growing rapidly and is looking to expand. The business is currently headquartered in Las Vegas, Nevada, and has a large warehouse onsite. Since this company ships high end products with guaranteed overnight delivery throughout the United States, the business executives believe that it would be advantageous to locate the new warehouse in a strategic location. Several factors come into play. They would like to be located near their largest customer base to cut down on shipping costs. They would like to purchase an existing warehouse in an area that is affordable, to keep operational costs down. And finally, they would like an area that has an international airport nearby, so that they can speed up delivery times for international customers as well.

2. Data Sources

The data needed to help solve this problem will have to come from a few different sources. To target the company's largest customer pools, I will use an invoice roster to view shipping addresses of all previous sales. I can add up which zip codes have the most traffic, both with the number of sales and total items shipped. It is also important to factor in the price of the items sold, as higher ticket items are generally larger and heavier and more expensive to ship.

I can then merge this data with a dataset acquired from Zillow, that shows median home cost by zip code. This is an accessible, and applicable data file that will help to identify areas that may have business real estate prices that are too high. Once the data is narrowed down to a few strategic locations, I will use Foursquare location data to verify that International Airports are nearby. The company wants to be able to ship items to any location, as fast as possible.

Based on these analyses, I will be able to recommend some strategic locations for this company to expand to.

3. Methodology

3.1 Initial Customer Data

The company has provided a 12-month sales record that has been imported into a Jupyter Notebook for analysis. This is a fairly large dataset, with over 100,000 rows. To find and classify the high value customers, the data must be manipulated to count the frequency that a given zip code is being shipped to, as well the total cost

of the items being sent there. A third metric, recency, will be used to verify that customers that have bought recently are weighed more than customers who have not purchased recently.

	Date	InvoiceNumber	ShippingZip	QTY	Price
0	2019-10-24	1231066	98204	1	5250.00
1	2019-10-24	1231066	98204	1	5250.00
2	2019-10-24	1231081	91766	1	690.00
3	2019-10-24	1231073	75402	1	600.00
4	2019-10-24	1231070	67152	1	85.00
5	2019-10-24	1231070	67152	1	85.00
6	2019-10-24	1231070	67152	1	120.00
7	2019-10-24	1231070	67152	1	100.00
8	2019-10-24	1231079	33966	2	400.00
9	2019-10-24	1231078	33186	1	237.81

111,518 rows; 5 columns

3.2 RFM Segmentation

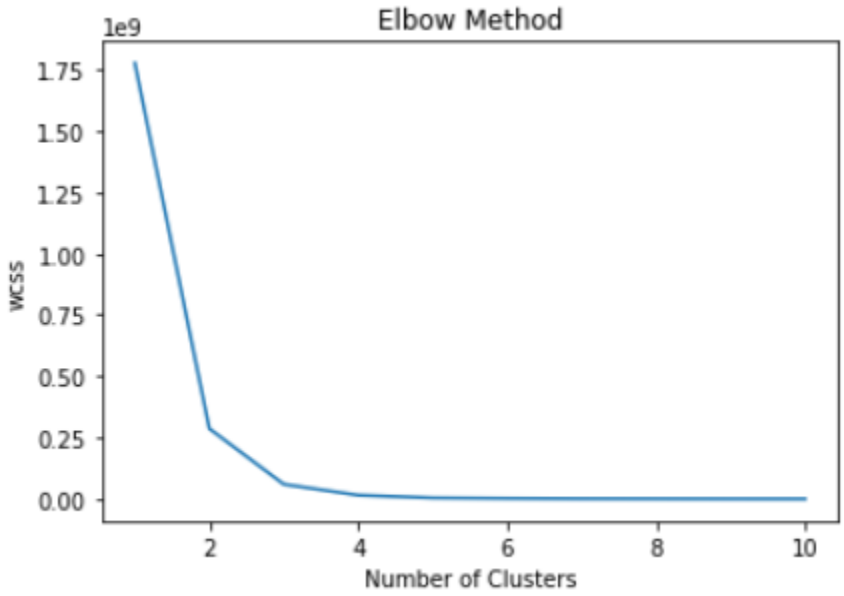
Below the dataset has been reduced to just one row for each zip code shipped to. The total amount cost of goods shipped is summed as the Monetary column. The number of orders shipped to each zip code are summed as the Frequency column, and the number of days since the last shipment to each zip code is shown as the Recency column. This has reduced to our dataset to just over 1000 unique rows.

	ShippingZip	Recency	Frequency	Monetary
0	10577	137	1	0.00
1	11434	8	9	29282.81
2	14120	105	1	661.56
3	20008	80	1	0.00
4	20166	39	1	5.00
5	27410	3	105	50835.12
6	28110	312	1	900.00
7	30043	261	4	551.70
8	30354	3	235	328887.84
9	32583	360	1	500.00

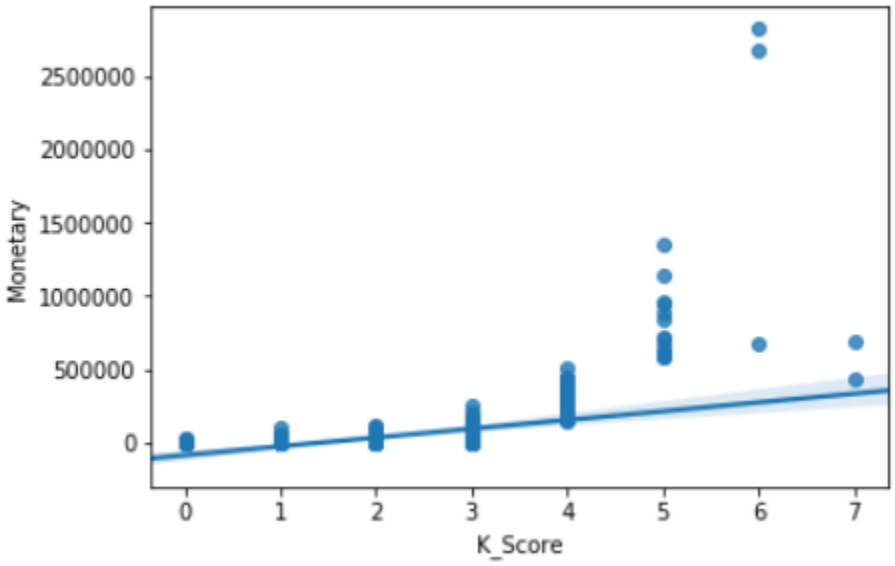
1,060 rows; 5 columns

3.3 K-Means Clustering

To segment the data to pinpoint the most valuable customers, K-Means Clustering is used to group similar data points together and to discover underlying patterns. A cluster score of 0-3 (4 clusters) is created for each of our three metrics. The scores for each metric are then summed to create a final K Score, so that the maximum possible K Score would be 9.



Below is a distribution of the K-Scores. In general, the score increases for zip codes that have higher dollars amounts shipped to them, although the highest score zip codes do not necessarily have the highest monetary amounts. These must be locations that are shipped to more frequently and/or more recently.



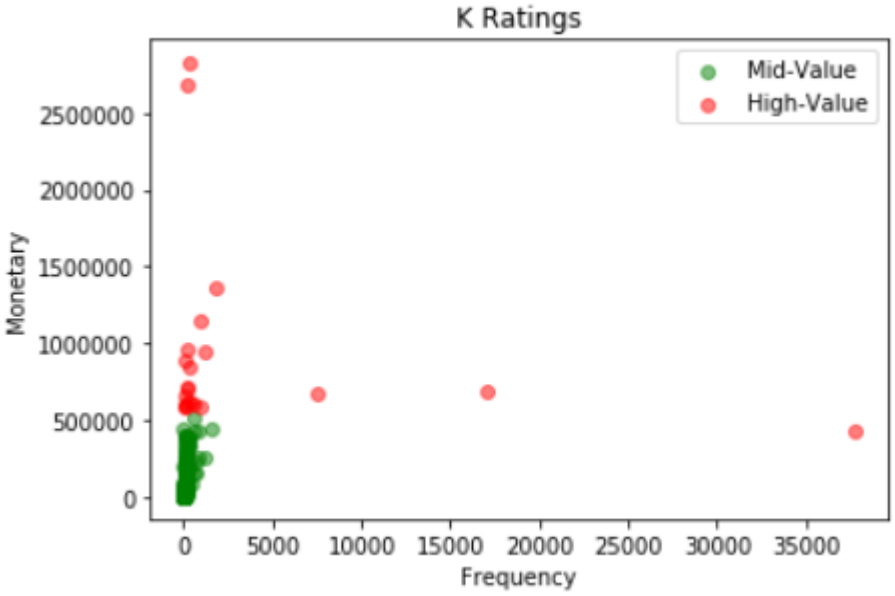
Below are the mean values for each of our K-Scores. As expected, the highest K-Scores do have the most frequent, most recent, and largest monetary values. Based on this distribution of scores, it is decided that customers having a score 5 or higher will be considered a 'High-Value' Customers and will be featured in future analyses to find the best strategic location.

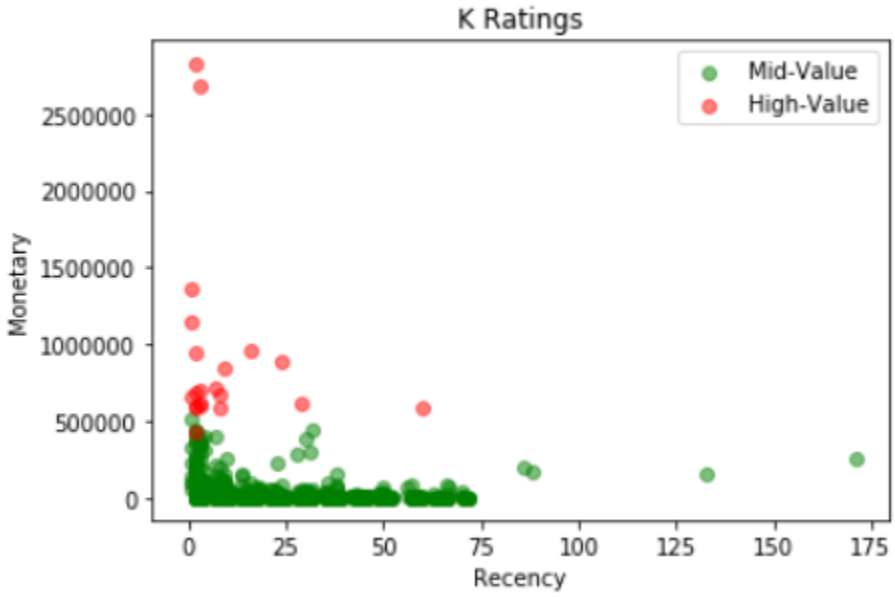
K_Score	Recency	Frequency	Monetary
0	318.570370	1.911111	1.814076e+03
1	225.662162	3.114865	4.950132e+03
2	121.693023	6.032558	7.177841e+03
3	26.070994	28.038540	2.322554e+04
4	6.666667	270.145833	2.795497e+05
5	10.687500	453.687500	7.756195e+05
6	4.333333	2700.000000	2.063107e+06
7	2.000000	27358.000000	5.574412e+05

3.4 Data Visualization

Looking at the distribution of the value assigned to our potential locations, only a small percentage is considered "High-Value." These are the locations that we want to focus on for further review.

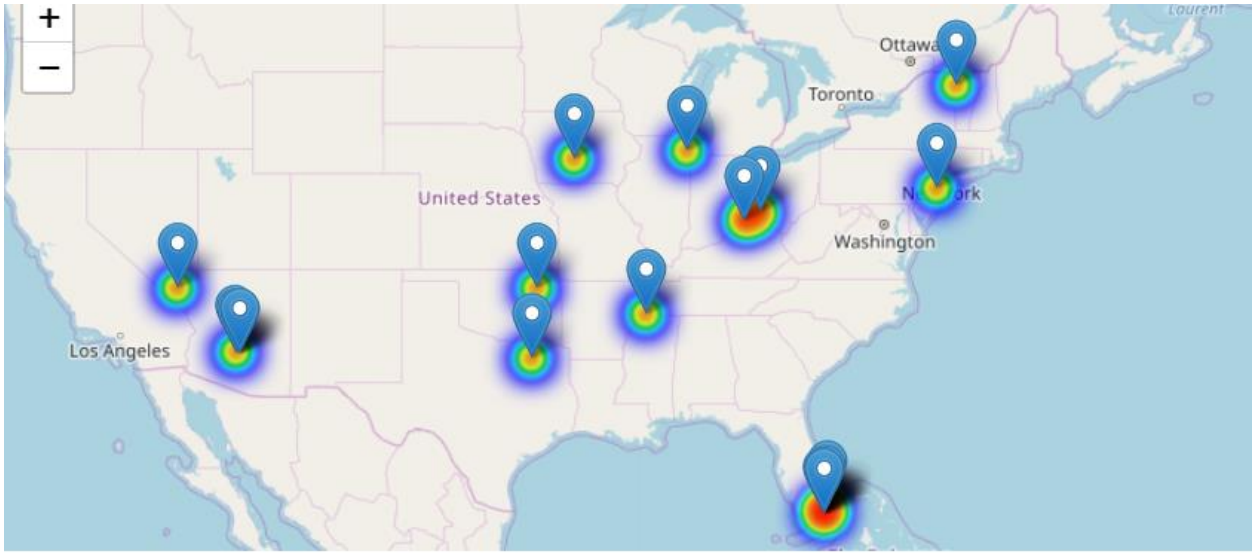
These scatter plots below verify that the 21 distinct High-Value locations often have multiple advantages over the Mid-Value locations. The High Value locations generally have a higher cost of goods shipped, tend to have more recent shipments, and tend to have more frequent shipment.





3.5 Analysis from Customer Dataset

Using Folium, I have generated a map of the 21 High-Value locations, which are scattered across the country with a heavier bias towards the Midwest and East coast. A heatmap of the potential locations, shows multiple targets in Ohio and Florida. Below is also the final dataset for the initial analysis.



	ShippingZip	Recency	Frequency	Monetary	R_Cluster	F_Cluster	M_Cluster	K_Score	K_Rating
0	33126	2	37718	427524.94	3	3	1	7	High-Value
1	45177	2	565	597780.50	3	0	2	5	High-Value
2	74116	3	389	616513.63	3	0	2	5	High-Value
3	89118	60	36	584000.00	3	0	2	5	High-Value
4	05404	29	240	620400.00	3	0	2	5	High-Value
5	07751	24	30	886650.00	3	0	2	5	High-Value
6	33069	8	124	587793.53	3	0	2	5	High-Value
7	33122	1	959	1141977.18	3	0	2	5	High-Value
8	33166	1	1810	1359175.83	3	0	2	5	High-Value
9	33172	2	1127	943721.66	3	0	2	5	High-Value
10	33178	2	960	592635.14	3	0	2	5	High-Value
11	41018	16	159	961284.49	3	0	2	5	High-Value
12	50266	3	65	596749.13	3	0	2	5	High-Value
13	60622	9	288	844000.00	3	0	2	5	High-Value
14	75402	1	90	655672.99	3	0	2	5	High-Value
15	85226	3	178	704930.02	3	0	2	5	High-Value
16	85286	7	239	716628.65	3	0	2	5	High-Value
17	33025	2	16998	687357.42	3	2	2	7	High-Value
18	85225	8	7570	678866.75	3	1	2	6	High-Value
19	38118	3	239	2681122.23	3	0	3	6	High-Value
20	85034	2	291	2829331.90	3	0	3	6	High-Value

3.6 Zillow Home Value Index

Zillow is an online real estate database company with price estimates on more than 110 million homes across the United States. The company has created the *Zillow Home Value Index*, which is the median Zestimate valuation for a given geographic area on a given day. This data can be used to pinpoint locations where rent or labor costs may be too high. The top ZHVI values, show zip codes in San Francisco, New York, and Los Angeles, which are areas that this company would like to avoid because of the associated costs.

	ShippingZip	State	Metro	County	City	Zhvi
0	94027	CA	San Francisco-Oakland-Hayward	San Mateo County	Atherton	5902400
1	90210	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	Beverly Hills	4777200
2	90402	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	Santa Monica	3942400
3	94301	CA	San Jose-Sunnyvale-Santa Clara	Santa Clara County	Palo Alto	3731600
4	94022	CA	San Jose-Sunnyvale-Santa Clara	Santa Clara County	Los Altos	3571400
5	94028	CA	San Francisco-Oakland-Hayward	San Mateo County	Portola Valley	3514400
6	11976	NY	New York-Newark-Jersey City	Suffolk County	Water Mill	3399200
7	11930	NY	New York-Newark-Jersey City	Suffolk County	Amagansett	3162500
8	94024	CA	San Jose-Sunnyvale-Santa Clara	Santa Clara County	Los Altos	3060000
9	90272	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	Los Angeles	3010200

The Zillow data is merged with our list of potential locations. The zip codes are sorted by K-Score and then by the Zillow Home Value Index. Some of our potential locations have well above average home prices. Zip code 60622 near downtown Chicago, and zip code 05404 in Vermont near Lake Champlain have home prices well into the $400,000 range and are no longer considered viable locations. The median Zillow Home Value Index across the country is $231,000. Since we are looking to avoid areas with higher than average costs, initially all locations with a Home Value Index over $300,000 will be removed. We can also remove locations close to the current warehouse in Las Vegas, so zip codes in Nevada and Arizona are also removed.

	ShippingZip	Recency	Frequency	Monetary	K_Score	K_Rating	Latitude	Longitude	ZHV
0	33126	2	37718	427524.94	7	High-Value	25.78	-80.29	195400
1	33025	2	16998	687357.42	7	High-Value	25.99	-80.28	269900
2	38118	3	239	2681122.23	6	High-Value	35.03	-89.93	77700
3	41018	16	159	961284.49	5	High-Value	39.01	-84.60	139200
4	45177	2	565	597780.50	5	High-Value	39.45	-83.80	155200
5	33122	1	959	1141977.18	5	High-Value	25.80	-80.32	163900
6	75402	1	90	655672.99	5	High-Value	33.09	-96.09	165400
7	33069	8	124	587793.53	5	High-Value	26.23	-80.16	168300
8	33172	2	1127	943721.66	5	High-Value	25.79	-80.36	189800
9	74116	3	389	616513.63	5	High-Value	36.19	-95.84	211200
10	50266	3	65	596749.13	5	High-Value	41.57	-93.81	236000
11	07751	24	30	886650.00	5	High-Value	40.37	-74.25	254800

3.7 Foursquare API

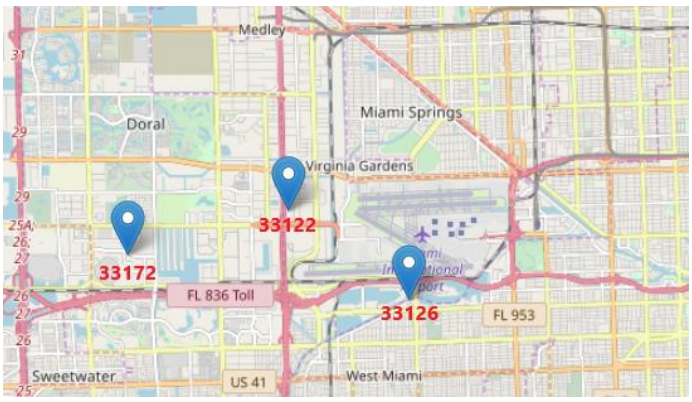
We look to continue to reduce the size of potential locations for the next warehouse. The final requirement is that the new area be located next to an International Airport. This company frequently ships next day air domestically, and also has a number of international customers. Being close to an International Airport is a top priority. For this we use Foursquare's API, to make a loop of our latitude and longitude coordinates for each potential destination and search for airports within 8000 meters, or approximately 5 miles. Our locations are given a pass/fail binary score, and the locations without international airports are removed. The result is our final dataframe, with 5 potential locations.

	ShippingZip	Recency	Frequency	Monetary	K_Score	K_Rating	Latitude	Longitude	ZHV	Intl Airport
0	33126	2	37718	427524.94	7	High-Value	25.78	-80.29	195400	1
1	38118	3	239	2681122.23	6	High-Value	35.03	-89.93	77700	1
2	41018	16	159	961284.49	5	High-Value	39.01	-84.60	139200	1
3	33122	1	959	1141977.18	5	High-Value	25.80	-80.32	163900	1
4	33172	2	1127	943721.66	5	High-Value	25.79	-80.36	189800	1

4. Results and Discussion

It is possible zip code 38118 could be considered because the potential building and labor costs could be much cheaper. The location is in Memphis, Tennessee and the home costs are half that of the other potential locations. 38118 also had the largest single Monetary value, meaning that high dollar items are being shipped there.

However, three of the final five potential locations are actually neighboring zip codes in Miami, Florida, located right next to Miami International Airport. These were noticed right away based on the initial heatmap. Even before the Foursquare analysis, 5 of the final 12 locations were in Florida. This is an obvious hotspot for this company's customers and being located in Florida could drastically reduce costs as overnight deliveries would not require air transport. The immediate recommendations would be to further explore warehouse locations in zip codes 33126, 33122, and 33172.



5. Conclusion

The need for expanding to a second location is backed up by the data. After the initial RFM segmentation and K-Means Clustering based on the initial company provided sales data, 18 of the 21 possible locations were in the Midwest or Eastern United States. After merging the sales data with Zillow's home value data, 5 of the top 12 possible locations were in Florida. And finally, after using Foursquare's API to verify proximity to International Aiports, 3 of the final 5 possible locations were also in Florida. In fact, these final three zip codes all share borders with each other. They are effectively one shipping destination. So, moving to this cluster would drastically cut costs. And it's proven that these are affordable and viable markets, so that facility and labor costs would remain low.