Placer.AI

Visitation Patterns at Stop & Shop

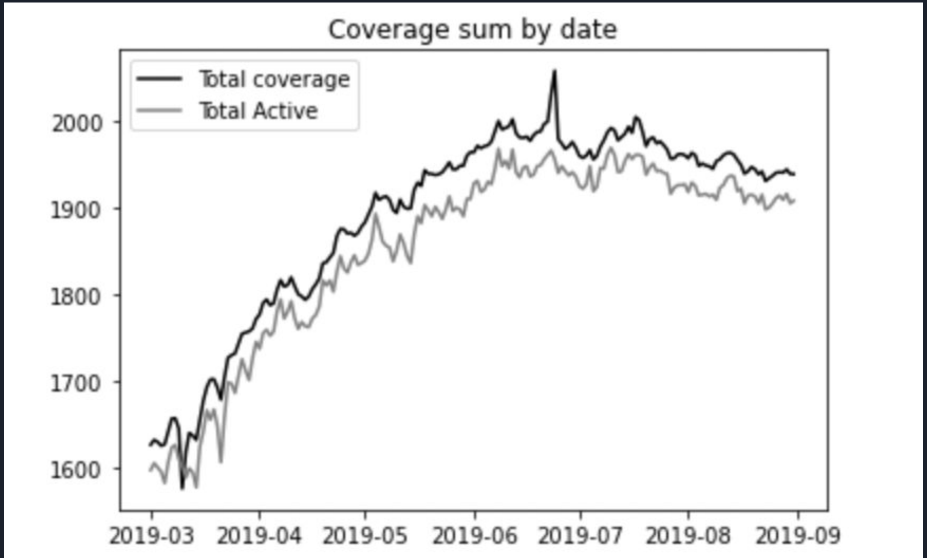
Ben Hendel

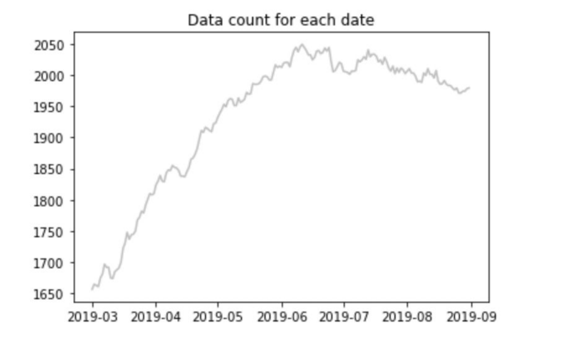
Background

Stop & Shop is a chain of 415 supermarkets located in the Northeastern United States. Over a time period of eight months user data was collected for devices and visitation. The goal of this presentation is to explain the data and assess what makes a customer loyal to Stop & Shop. Using this data, a comprehensive market analysis and model could be generated to potentially increase revenue. The primary tools used in this analysis were Python. Jupyter Notebooks, QGIS, and Kepler.gl

User Activity

A data value is "Coverage", for which a user is defined as active if coverage is at least 0.75. To see the overall coverage over time, these graphs sum grouped by date, which is misleading:

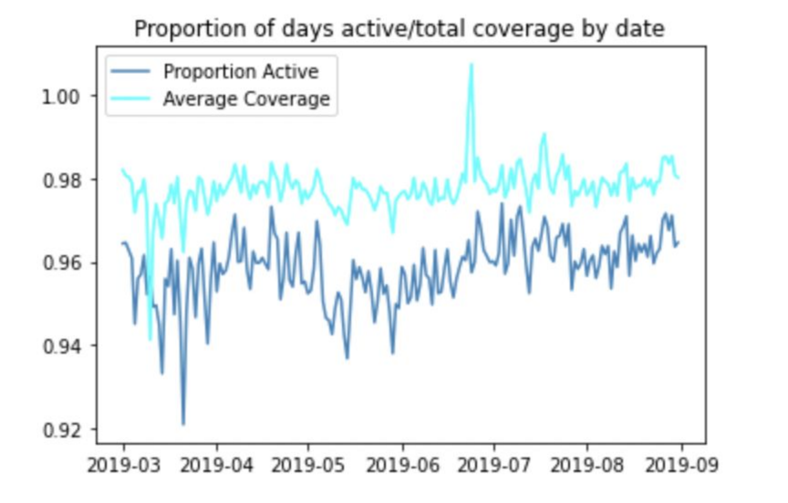




However, these graphs simply follow the total count of data collected in the sample, where more data was collected in later months

Though a trend is clear, it is simply the trend of the relative frequency of days. To fix this we will look at coverage and activity divided by how many times that date appears in the data.

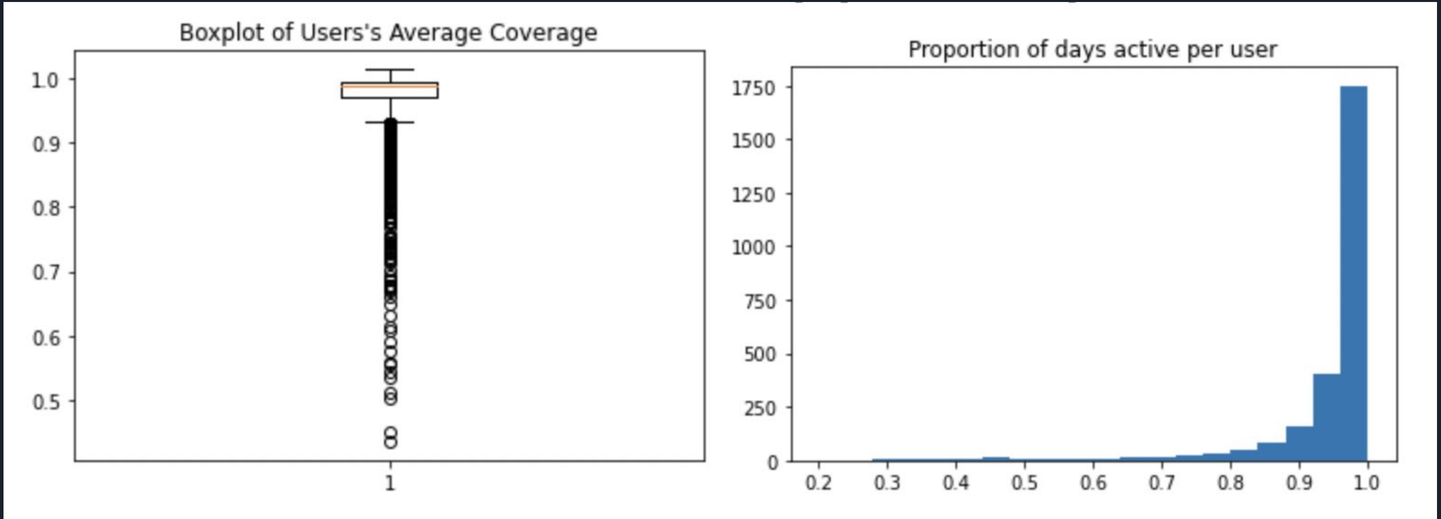
This is starting to look like more of a time series with a baseline and seasonality, for which a model could be built.



Note the anomalous peak with average daily coverage greater than 1: this is a due to rows that have a coverage value > 1, possibly an error, but the problem statement does not give enough information to make any assumptions

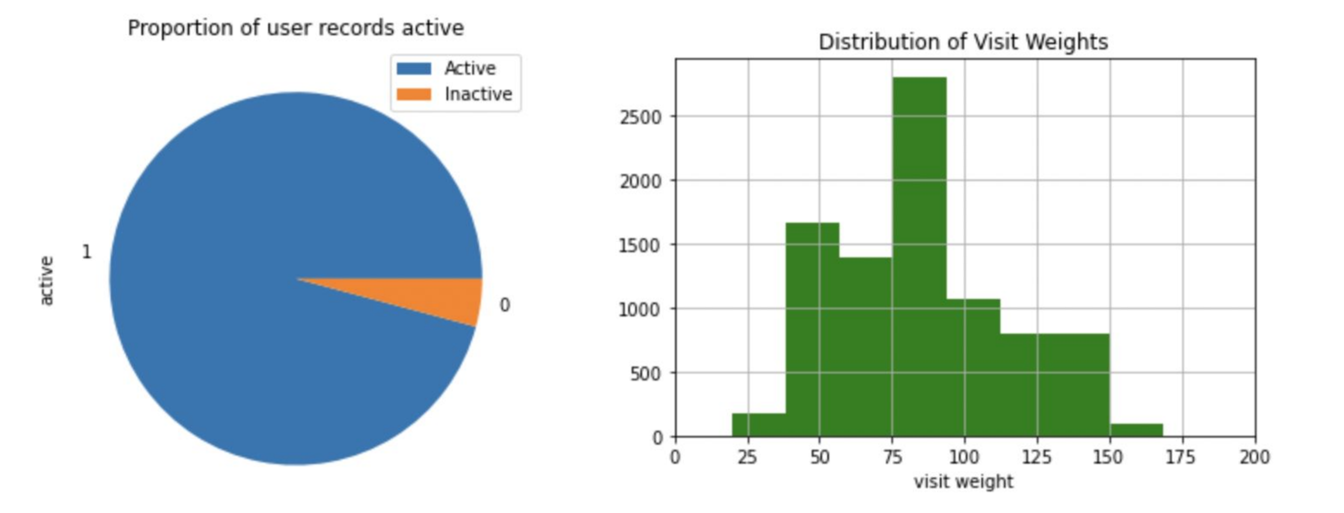
The user's average coverage distribution is skewed close to 1, with lower values being more rare.

Similarly, the distribution of the proportion of days a user was active is skewed, with a mean of 94.7% chance of a user's day being active

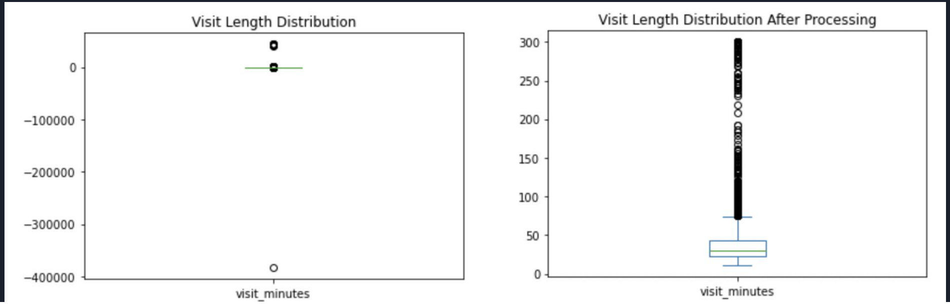


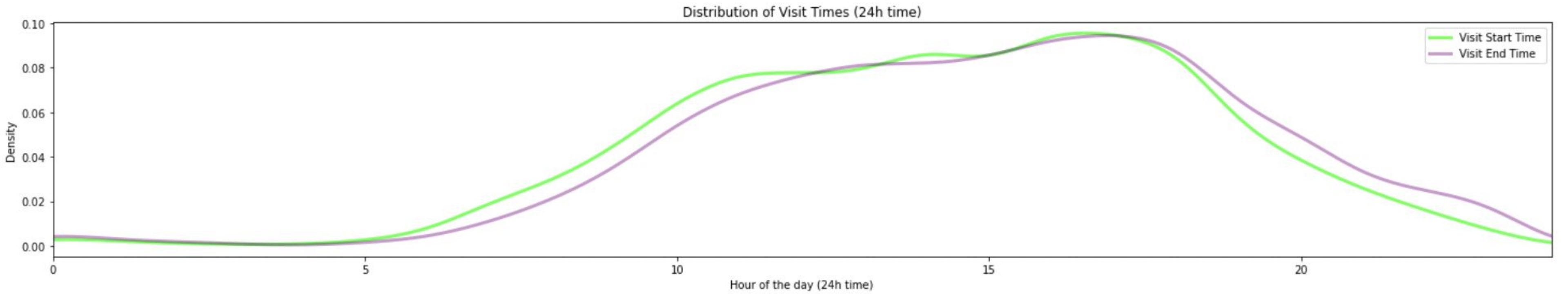
Visits

For visitations, the data is complete with no null or missing values. There are some issues with visit times we will get to later. Most records in the user data are “active”. Visits also have a weight field, which is distributed as follows. We will use it later for aggregation

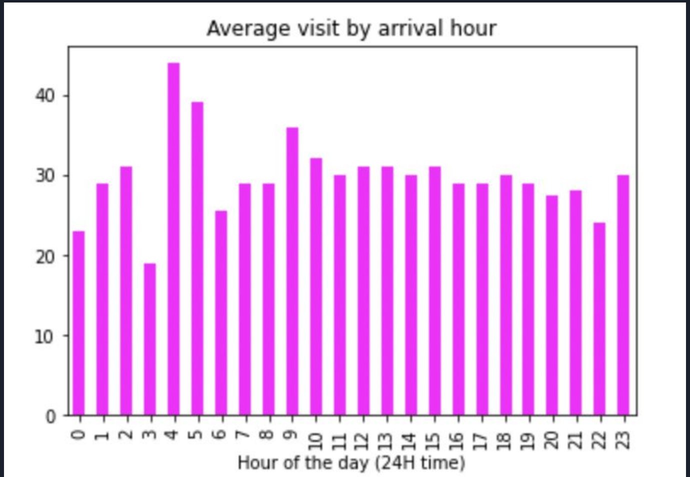


Predictably, end times are distributed like start times but shifted right. Visits to Stop & Shop are most popular at around 5pm and drop down from there, with noon to 5 being peak hours. One visit has the start time after the end time, and 60 visits are longer than 5 hours, often much longer, which is suspicious. Removing these rows could create a gap in the time series and is inadvisable. Instead, we will clip the visit length to 5 hours (an estimate of the most time anyone would reasonably visit a supermarket) and adjust the end time accordingly. For the row with negative visit length we will replace the visit length with the lowest visit length above 0.

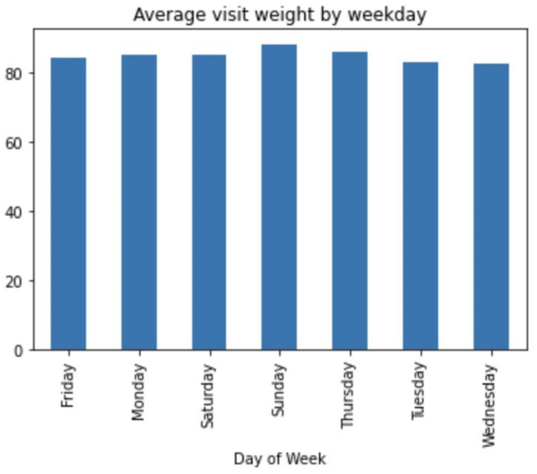


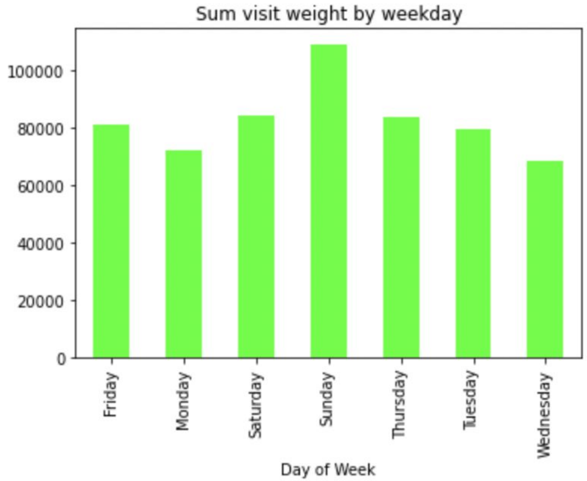


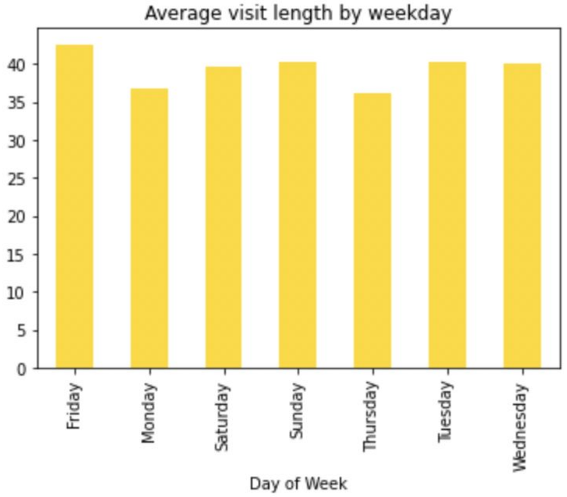
There is no glaring difference in average visit weight between days of the week.


Average visit by arrival hour

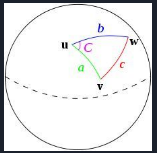
Based on total visit weight and total visits, it seems that Sundays are busiest. Customers also spend more time at Stop & Shop during Friday visits, and less time when arriving at 10pm. Late night arrivals are higher variance and less trustworthy since the sample size is so much smaller


Average visit weight by weekday


Sum visit weight by weekday

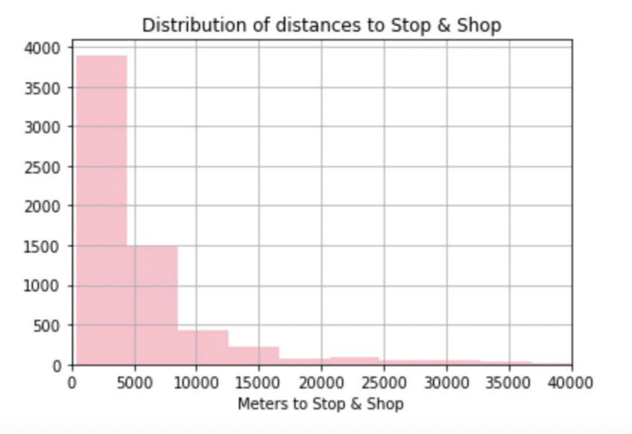

Average visit length by weekday

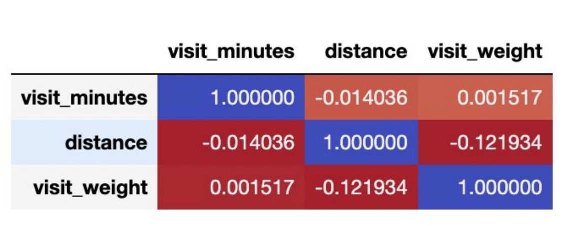
Distances and Correlations



We could get the distances by simply doing the distance formula, but this cartesian line is inaccurate due to the curvature of the earth's surface. The Haversine formula is a much more accurate approximation.

There are a huge range of varied values for distances, as we will see later. There are even points on the West Coast. Despite the quality of Stop & Shop groceries, it's unlikely that someone would fly from Los Angeles to shop there!



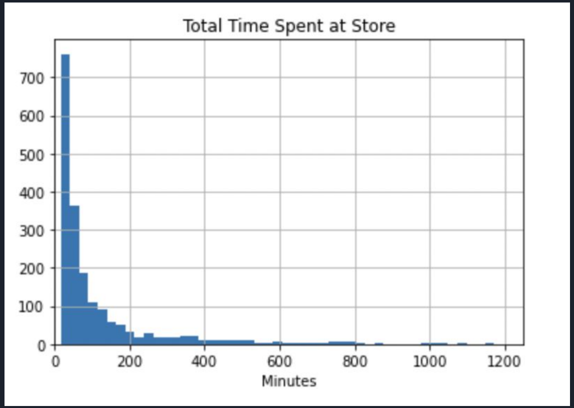


	visit_minutes	distance	visit_weight
visit_minutes	1.000000	-0.014036	0.001517
distance	-0.014036	1.000000	-0.121934
visit_weight	0.001517	-0.121934	1.000000

Visit weight and distance are negatively correlated, which may mean distance alone isn't the best way to determine loyal customers

Loyal Customers Part One: Time at Store

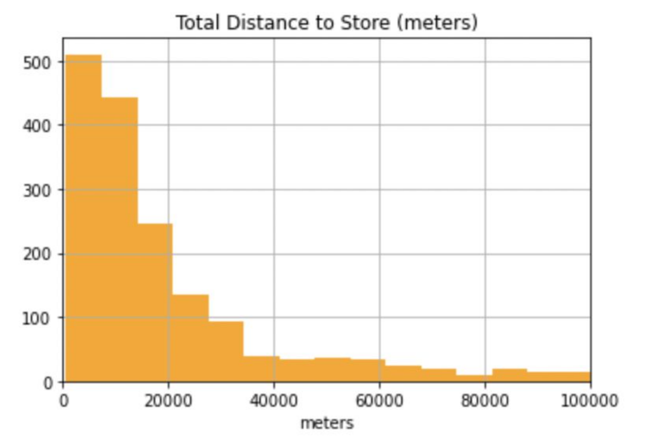
In order to determine what makes a loyal customer, let's first take a look at customers that spend a lot of time at the store. Customers that spend a lot of total time in the store are usually the ones purchasing the most.



Time total roughly follows an exponential distribution, which is commonly used in statistics to measure the amount of time until something happens. In our example, that is the time until the customer leaves the store. Notice the "tail end" of customers who don't behave normally, these are worth investigating

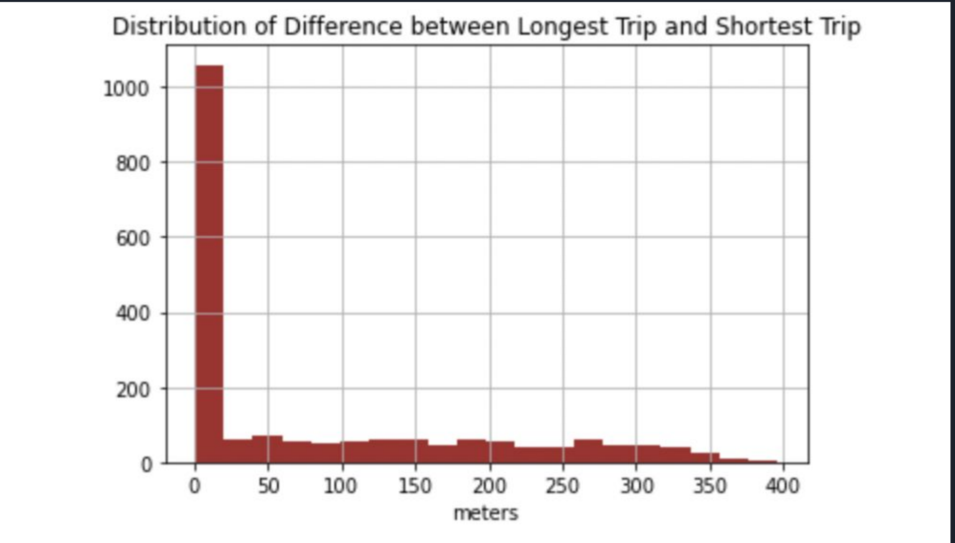
Loyal Customers Part Two: Distance to Store

Next let's look at customers that have traveled the most total distance to the store. It logically follows that if a customer is willing and has traveled a large total distance to the store, they are loyal. Individual trips can vary greatly and data can be less than perfectly reliable,, but a large total distance traveled is indicative. Notice a "long tail" of customers that have traveled greater than ~40km to get to Stop & Shop. The patterns of these customers are worth looking into.



Loyal Customers Part Three: Range of Trips

Something interesting to note: The home coordinates for each device are not the same every time and vary considerably. This feature may represent something interesting that hasn't yet been explained in the presentation- this could be the user making multiple trips from multiple different locations. In that case, a good way to assess customer loyalty is to ask "Is the customer making trips to the store from many different locations, far and close?"



After all, distant trips are low weight and high variance, and could mean one-time necessities; it is unknown if a customer coming from far has another closer, preferred store in mind. But if customer chooses to visit Stop&Shop from very different starting points, it means that they likely prefer shopping there no matter where they are. This was not entirely clear in the description.

To the left are customers divided by the range of trip distances. Notice the two groups, the low values and the tail. The phenomena of the customers in the tail are worth looking into.

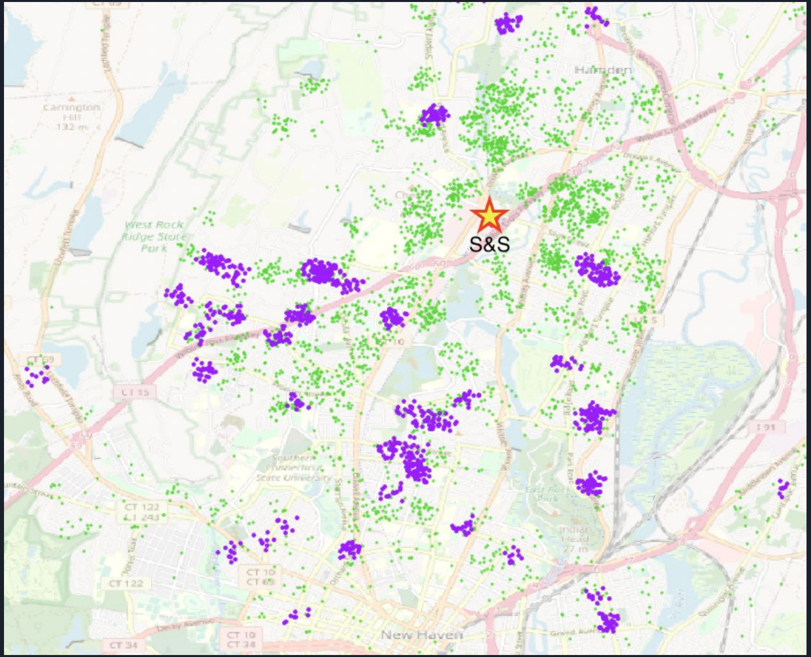
Loyal Customers: Putting it all together

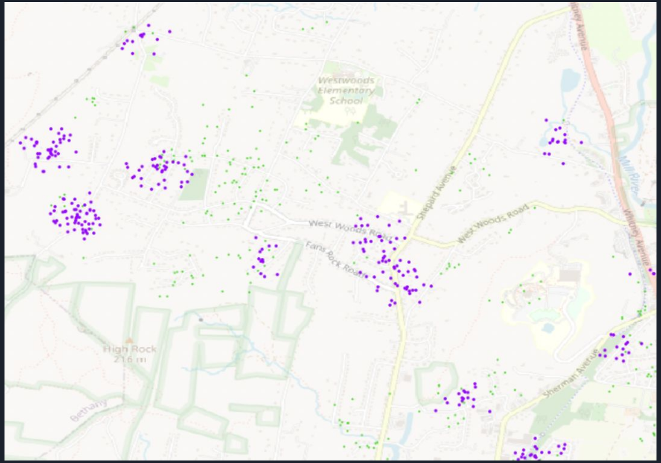
We have created three clusters based on these criteria: Time spent at store, time spent traveling to store, and variety in distances traveled to store.

Predictably, these often tend to be the same customers, so once we take the customers in common between groups we get 91 of the most "Loyal" customers (out of 1928, the top 5%). Together, they account for 1803 out of the 6793 active visits (27%).

Thus, our definition of a loyal customer is contingent on this unique combination of criteria. Let’s visualize what these look like on a map

Mapping


S&S
Hamden
Carrington Hill 132 m
West Rock Ridge State Park
CT 15
CT 122
CT 10
I 91
Southern Connecticut State University
New Haven

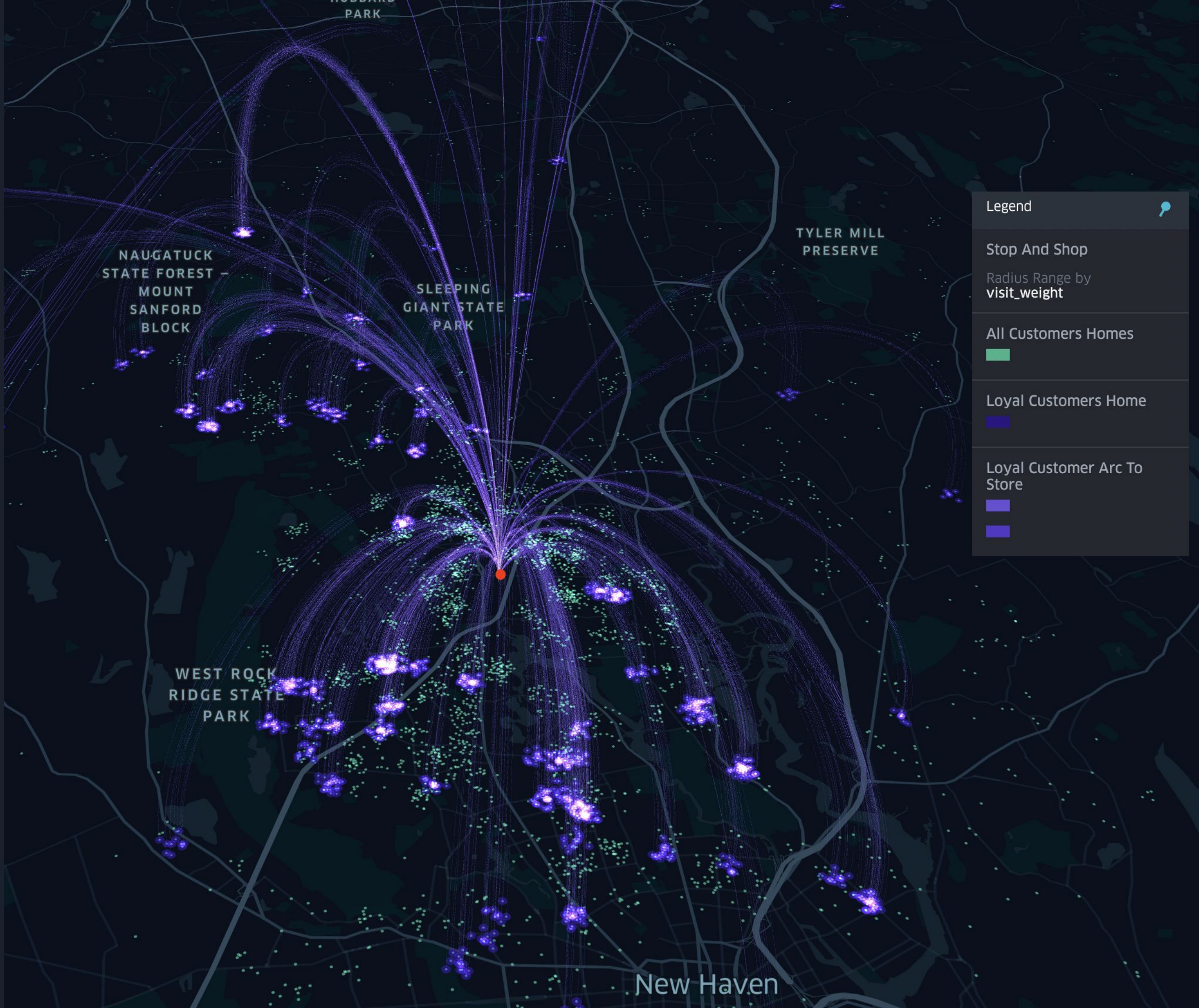


A zoomed in look at the clusters of loyal customers (purple) vs all customers (green)

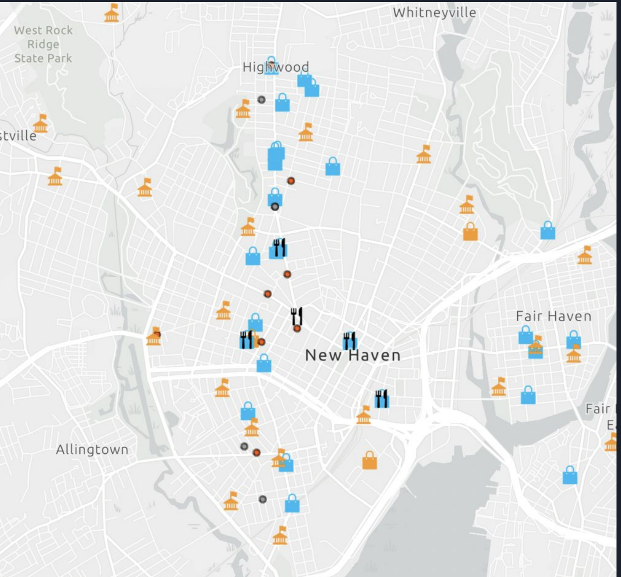
By uploading the data to GIS software, we get a clear visual picture of the segmentation of loyal customers. Interestingly, the loyal customers (purple points) tend to cluster together fairly well in small neighborhoods and subdivisions. Analysing multiple attributes had the positive effect of removing high variance/long distance visits, such as the few odd ones coming all the way from Los Angeles.

Although I am most comfortable with GIS software for mapping, let's create the same kind of map using Kepler since that is what was stated in the prompt

For this kind of project, Kepler proves to be useful and aesthetically more pleasing to look at, though I was not able to get the package and widgets installed to make it visible natively on Jupyter Notebook. The 3D map shows clusters of loyal customers arcing to the location of Stop & Shop


PARK
NAUGATUCK STATE FOREST – MOUNT SANFORD BLOCK
SLEEPING GIANT STATE PARK
TYLER MILL PRESERVE
WEST ROCK RIDGE STATE PARK
New Haven
Legend
Stop And Shop
Radius Range by visit_weight
All Customers Homes 🟩
Loyal Customers Home 🟪
Loyal Customer Arc To Store 🟪 🟪

Further Additions



A further demographic and analysis would be valuable to Stop & Shop from a business perspective. If given more time and a salaried role, I would download layers of US Census data, land use shapefiles from the New Haven county website, building records from Zillow, and nearby POI to get a sense of access to competitors. I have done this workflow before as part of a suitability analysis for a new construction site.

Thank you for your consideration!

- Ben Hendel