Applied Data Science Capstone – Destination Classifier

Introduction/Business Problem

Suppose someone wants to go on a vacation somewhere in the U.S, but does not know where he/she wants to go.


Photos found on pexels.com

My program attempts to cluster several destinations based either on the estimated gas prices at the destination, estimated max/min temperature yesterday at the destination, or venue type frequencies (according to Foursquare) to help the user decide on a place to travel to. Travel companies such as Momondo.com may be interested in it because it would allow an adventurous customer an additional screening feature for finding flights. A traveler who plans to rent a car at the destination may be interested in visiting a place with gas prices similar to their city of departure. Furthermore, travelers may want to choose a destination with a similar climate/temperature to another destination they know about. If a customer likes these clustering features, the customer may re-use the service, and service provider (e.g. Momondo.com) may make more money.

Data

I obtain data from 5 sources: (1) Bureau of Transportation Statistics, (2) Wikipedia, (3) Gas Buddy, (4) Weather Underground, and (4) Foursquare.

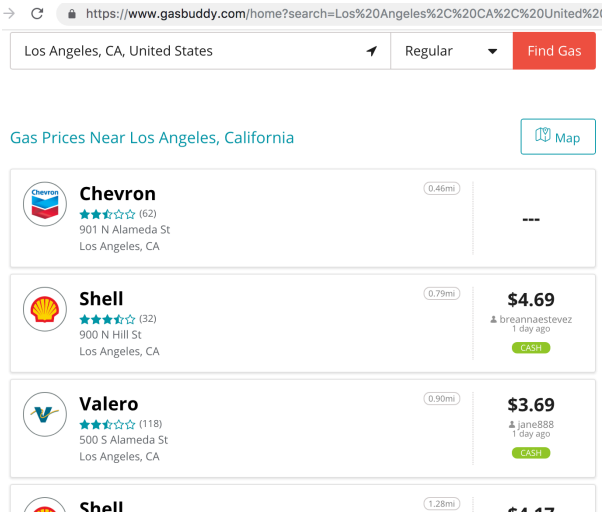
I obtain a CSV file from the BTS containing airports around the world, their city and state if applicable, and their coordinates.

I use a Wikipedia article, 'List of the busiest airports in the United States' to shorten the table of airports found in the CSV file. I consider ~30 large hubs (e.g. SFO) from the first table of the article. A sample of the table is shown below:

://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

Rank (2017)	Airports (large hubs)	IATA Code	Major city served	State	2017[3]	2016[4]	2
1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	GA	50,251,962	50,501,858	4
2	Los Angeles International Airport	LAX	Los Angeles	CA	41,232,416	39,636,042	3

I use Gas Buddy to estimate the gas price at the destination. If I manually search for 'Los Angeles, CA, United States' on Gas Buddy, I see at most 10 gas prices under a section called, 'Gas Prices Near Los Angeles, California'.



Based on some trial and error and looking at the source code, I automated the process of collecting the 10 or less gas prices on a location's webpage (sometimes '---' is shown instead of a decimal value) and use its average as an estimate of the gas price at the destination. Some draw backs of my automated process include not checking whether the gas price is really from the city -- I am taking faith that after searching for a specific city, the gas prices are representative of that city (i.e. the cities in those blocks are consistent with the city I searched for). Also, even if the city is correct, the prices probably describe an area whose center is at a different latitude and longitude than the airport and it is unclear where is that point of reference.

I use the 'yesterday_max_temperature' and 'yesterday_min_temperature' found in the source code of a Weather Underground webpage describing a location.

I assume this is the max temperature and min temperature for the day before. I am not confident about the accuracy of this.

I use Foursquare to calculate the venue type frequency around a destination (almost a repeat of the analysis in the ungraded lab example).

Methodology

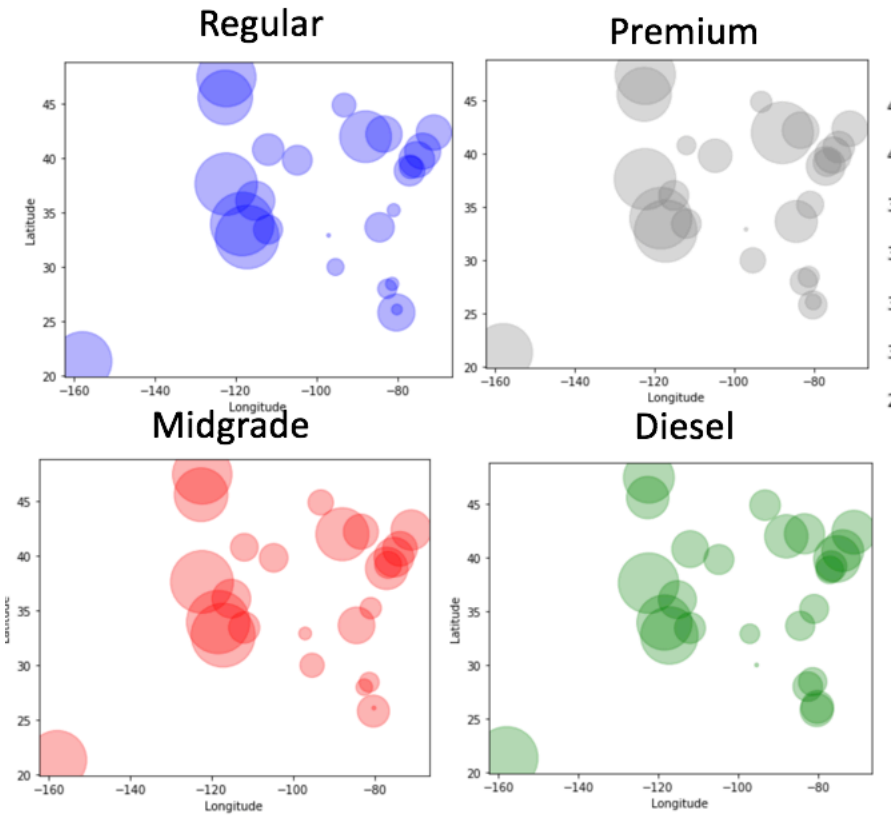
I created boxplots and scatter plots for both the temperature columns and the gas price columns to get an idea of how those quantities vary across the nation.

I used two clustering methods (DBSCAN and KMeans) so I could compare the labels derived by each of them. I clustered the destinations 6 different ways - 3 using DBSCAN and 3 using KMeans. Furthermore, I clustered either by (regular/midgrade/premium/diesel) gas price estimates, yesterday max/min temperature estimates, or venue category information.

I used folium maps to visualize the clusters.

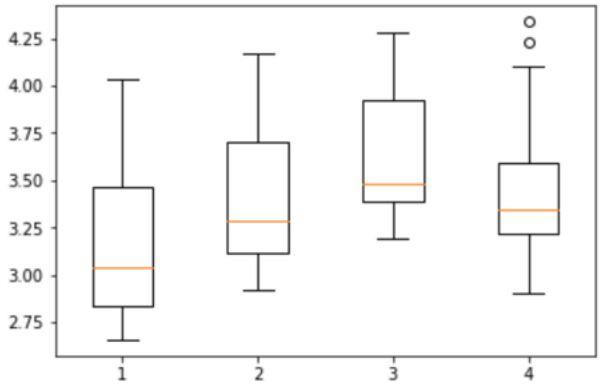
Results

The following plots illustrate the relationship between geographical coordinates and estimated gas price.



On a superficial level, it seems like the West tends to have higher gas prices judging by the size of the circles in comparison to elsewhere.

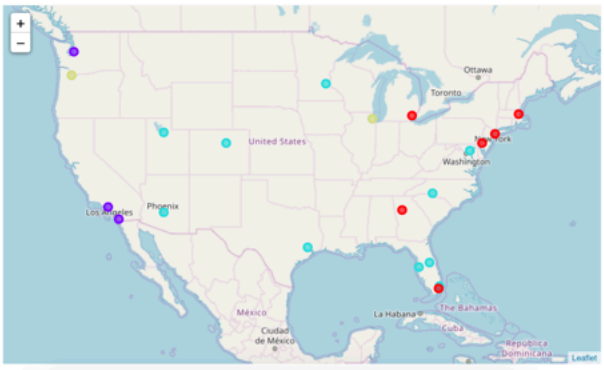
Here are box plots and maps of the clusters using the estimated gas price data.



	Regular Gas Price Estimate [$]	Midgrade Gas Price Estimate [$]
count	26.000000	26.000000
mean	3.169135	3.408512
std	0.455359	0.400248
min	2.655714	2.916667
25%	2.833571	3.115833
50%	3.034786	3.286875
75%	3.468500	3.700562
max	4.032857	4.175000

	Premium Gas Price Estimate [$]	Diesel Gas Price Estimate [$]
count	26.000000	26.000000
mean	3.634353	3.473265
std	0.369644	0.369642
min	3.190000	2.905000
25%	3.390000	3.220714
50%	3.482143	3.347875
75%	3.927500	3.596500
max	4.283333	4.340000

K Means (k = 4)



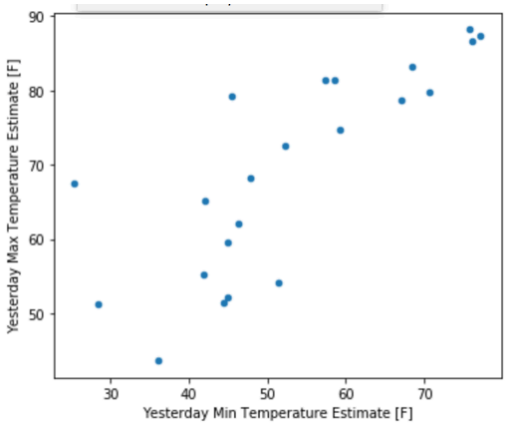
DBSCAN (eps = 0.29, min_samples = 2)

Most of the red dots are outliers
Exceptions: Philadelphia & Boston

Some examples of things that can be pointed out about the box plot are (1) the spread for estimated diesel gas price seems considerably smaller than the other gas types (2) about 11 destinations have an estimated regular gas price less than $3.04.

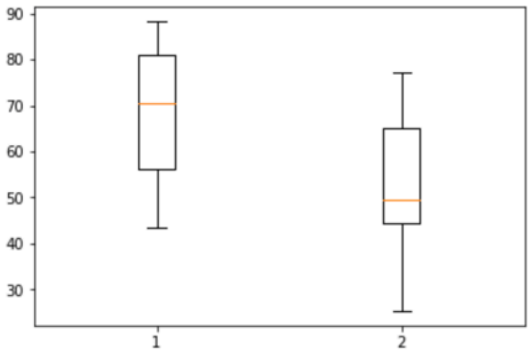
At least three things can be said about the maps. Phoenix, and Denver seem to have similar gas prices. Boston and Philadelphia seem to have similar gas prices. Lastly, Baltimore, Atlanta, Tampa, Minneapolis, and Orlando seem to have similar gas prices.

Here is a scatterplot showing the relationship between estimated minimum and maximum temperatures for yesterday. Each dot represents a destination city.



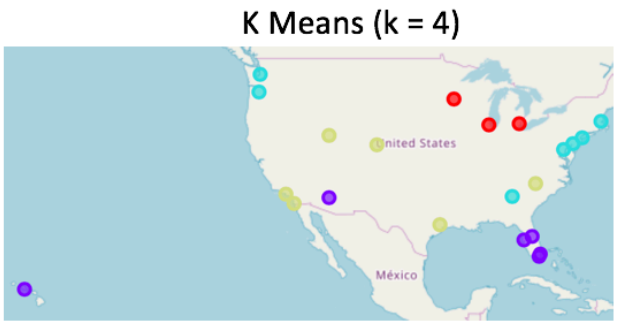
As the minimum temperature increases, the maximum temperature seems to tend to increase.

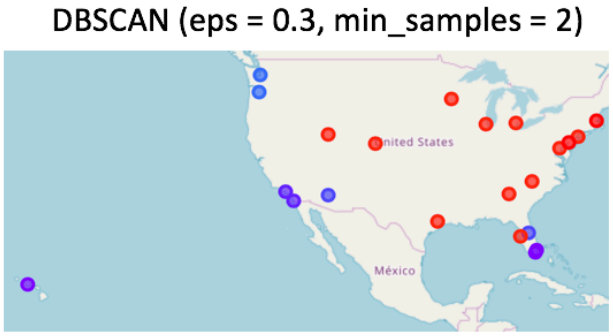
Here are box plots and maps of clusters using the estimated temperature data.



	Yesterday Max Temperature Estimate [F]
count	22.000000
mean	69.254545
std	13.810649
min	43.600000
25%	56.275000
50%	70.400000
75%	81.050000
max	88.200000

	Yesterday Min Temperature Estimate [F]
count	22.000000
mean	52.831818
std	14.967374
min	25.400000
25%	44.550000
50%	49.650000
75%	65.150000
max	77.200000



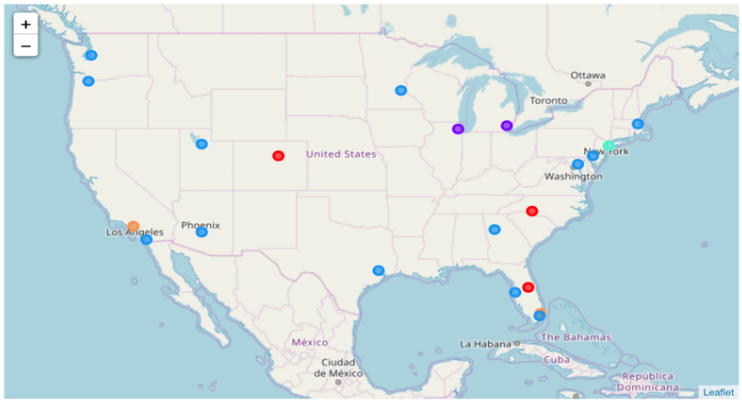


Some examples of things about the box plot that can be pointed out are (1) about 11 destination cities had an estimated minimum temperature of less than 49.65 degrees Fahrenheit yesterday (2) about 11 destination cities had an estimated max temperature of less than 70.4 degrees Fahrenheit yesterday.

At least 4 things can be said about the maps. Seattle and Portland seem to have similar temperature ranges yesterday. Boston and Philadelphia seem to have similar temperature ranges yesterday. San Diego and Los Angeles seem to have similar temperature ranges yesterday. Lastly, Honolulu, Tampa, Miami, and Fort Lauderdale seem to have similar temperatures yesterday.

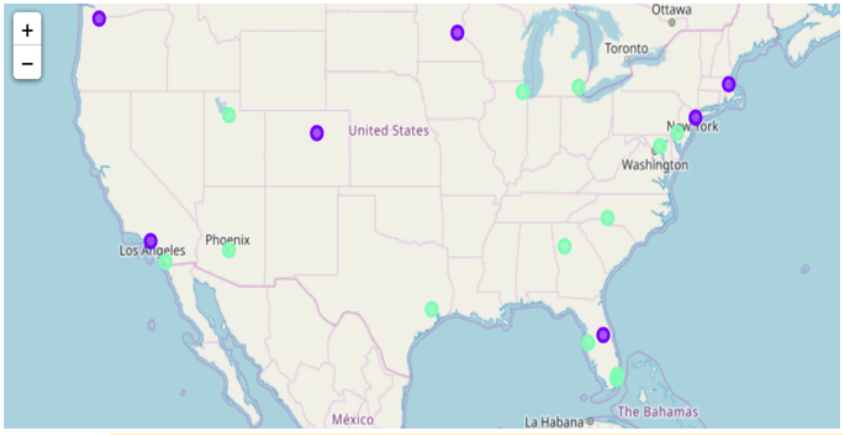
Here are maps of clusters using venue type frequency data.

K Means (k = 6)



Outliers are purple

DBSCAN (eps = 21, min_samples = 4)



The vicinity of each large airport hub from Phoenix, Los Angeles, Salt Lake City, Houston, Atlanta, Tampa, Miami, Baltimore, and Philadelphia seem to have similar venue type frequencies.

Discussion/Recommendations

An example of a recommendation based on the gas prices analysis is to suggest Orlando to someone who wants to visit a place with similar gas prices to Minneapolis (assuming he/she doesn't care about other factors).

An example of a recommendation based on the temperature range analysis is to suggest Honolulu to someone who wants to visit a place with temperature ranges like Miami (assuming he/she doesn't care about other factors).

An example of a recommendation based on the venue type frequency is to suggest Phoenix to someone who wants to visit a place with similar venue type frequency as Miami (assuming he/she doesn't care about other factors).

Conclusion

Clustering was not as effective as I had wished for - for example, DBSCAN sometimes left many outliers. But, I hope this project may motivate further study in classifying destinations for the end goal in improving service to travel company users.

I have ideas for improvement. Maybe, we should into account other variables such as humidity, wind speed, UV, recorded incidents of flu/cold, bus pass costs, traffic, venue frequency of big brand names (e.g. Starbucks), and frequency of slang words in Venue tips. Also, maybe we should make predictions with ML algorithms (e.g. Logistic regression) on historical data (e.g. weather data).