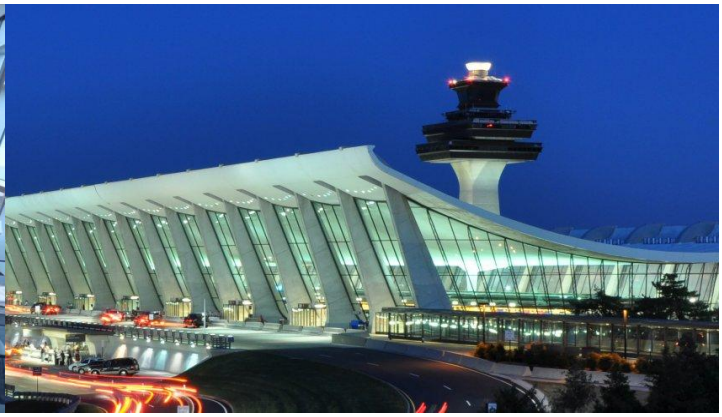
Data Science Project Report:

Determining the best large US airport to open a Fashion Retail Outlet

Allen Wu

July 2019

Introduction

The United States is the most visited country in the world, attracting international residents with economic opportunities, tourist destinations, and cultural experiences. Central to the majority of international travel are more than 5000 public use and 14000 private use airports which help move over 849 million annual passengers.

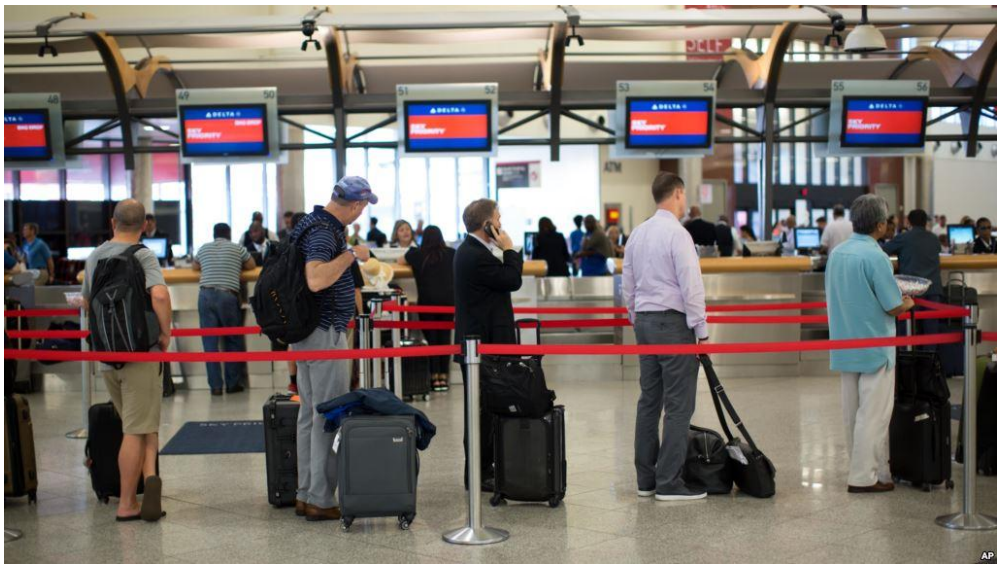


Figure 1: Inside Atlanta Hartsfield-Jackson International Airport, the largest airport by passenger traffic in the United States

As a result, large airports across the United States are often hubs for consumer spending, especially in the fashion retail sector. With so many customers and the potential for huge annual sales, airports are an enticing place for fashion retail brands both large and small to set up shop and make a profit. However, the important question is which large airport would be the ideal location to open a new business? Throughout the report, this question will be investigated using a variety of data sources and tools.

The target audience of this report includes:

1. Owners of small fashion retail franchises seeking to serve a large number of domestic and international customers through opening an airport location
2. Existing brand-name franchises and chains which are seeking to expand and diversify operations by opening new locations within airports
3. Investors seeking to determine which airports may present valuable opportunities for retail in future years

Business Understanding

US international airports offer vendors heavy foot traffic with millions of yearly customers. If a good location is selected, a US airport fashion retail outlet can be incredibly successful and profitable.

However, opening a new store is no easy feat, and especially not in an airport. Unlike regular fashion retail outlets, those setting up shop in airports must contend with a myriad of other factors to operate. For instance, at many US airports including the busiest: Atlanta's Hartsfield Jackson International Airport, vendors are required to remain open until the last flight leaves. In addition, airports charge on average 12% of yearly gross sales, on top of stores paying already high rental costs.

As a result, all of these factors lead to higher costs of operation. Thus, not carefully selecting a location can have incredibly high stakes for the owner, and it is imperative that a data-driven approach is required to find the best location.

Data Requirements

The project will determine an airport or cluster of airports where it is most ideal to open a new retail outlet based on features such as yearly passenger traffic, rental prices, and existing number of clothing store venues.

First, web data will be of importance. The BeautifulSoup library will be used to scrape Wikipedia in order to obtain the name of each airport, the metropolitan area they belong to, as well as yearly passenger traffic. The table provides information on the 46 busiest US airports by total passenger traffic in 2016, and would be appropriate for use in this analysis. Further filtering can be conducted to only conduct analysis on airports which had a positive change from 2015-2016 in passenger growth, as outlets would seek to open in areas with growing foot traffic.

Rank	Airport name	Location	IATA Code	Traffic		Aircraft	
				Passengers	% chg. 2015/16	Movements	% chg. 2015/16
1	Hartsfield–Jackson Atlanta International Airport	Atlanta, College Park, and Hapeville; Georgia	ATL	104,171,935	▲ 2.6	898,356	▲ 1.8
2	Los Angeles International Airport	Westchester, Los Angeles, California	LAX	80,921,527	▲ 8.0	697,138	▲ 6.3
3	Chicago O'Hare International Airport	Chicago, Illinois	ORD	77,960,588	▲ 1.3	867,635	▼ 0.9
4	Dallas/Fort Worth International Airport	Coppell, Euless, Grapevine, and Irving; Texas	DFW	65,670,697	▲ 0.2	672,748	▼ 1.3
5	John F. Kennedy International Airport	Queens, New York	JFK	59,105,513	▲ 3.9	452,415	▲ 3.0
6	Denver International Airport	Denver, Colorado	DEN	58,266,515	▲ 7.9	565,503	▲ 4.5
7	San Francisco International Airport	San Mateo County, California	SFO	53,099,282	▲ 6.1	450,388	▲ 4.8
8	McCarran International Airport	Paradise, Nevada	LAS	47,496,614	▲ 4.5	541,428	▲ 2.1
9	Seattle–Tacoma International Airport	SeaTac, Washington	SEA	45,736,700	▲ 8.0	412,170	▲ 8.1
10	Miami International Airport	Miami-Dade County, Florida	MIA	44,584,603	▲ 0.5	414,234	▲ 0.3
11	Charlotte Douglas International Airport	Charlotte, North Carolina	CLT	44,422,022	▼ 1.0	545,742	▲ 0.3
12	Phoenix Sky Harbor International Airport	Phoenix, Arizona	PHX	43,302,381	▼ 1.6	440,643	▲ 0.1
13	Orlando International Airport	Orlando, Florida	MCO	41,923,399	▲ 8.0	316,981	▲ 2.9
14	George Bush Intercontinental Airport	Houston, Texas	IAH	41,622,594	▼ 3.3	470,780	▼ 6.4
15	Newark Liberty International Airport	Newark and Elizabeth, New Jersey	EWR	40,563,285	▲ 8.2	435,907	▲ 5.3
16	Minneapolis–Saint Paul International Airport	Hennepin County, Minnesota	MSP	37,413,728	▲ 2.3	412,872	▲ 2.0
17	Logan International Airport	Boston and Winthrop, Massachusetts	BOS	36,356,917	▲ 8.5	372,930	▲ 2.5
18	Detroit Metropolitan Airport	Romulus, Michigan	DTW	34,401,254	▲ 2.9	393,427	▲ 3.7
19	Philadelphia International Airport	Philadelphia and Tinicum Township, Pennsylvania	PHL	30,155,090	▼ 4.1	394,022	▼ 4.2
20	LaGuardia Airport	Queens, New York	LGA	29,786,769	▲ 4.7	369,987	▲ 2.7

From here, the airport names can be passed into a Python geocoder API which will determine latitude and longitude figures for individual airports. The coordinates of an airport can be passed into the Foursquare API, which can determine information about existing clothing stores near and within each airport. In addition, BeautifulSoup was used to scrape Wikipedia to determine the median per capita income for metropolitan areas.

Rank	Metropolitan statistical area	Population	Per capita income
1	Washington-Arlington-Alexandria, D.C-Virginia-Maryland MSA	5,949,178	$47,411
2	San Jose-Santa Clara-Sunnyvale, California MSA	1,918,944	$40,392
3	Seattle-Tacoma-Bellevue, Washington MSA	3,611,644	$39,322
4	San Francisco-Oakland-Hayward, California MSA	4,122,177	$38,355
5	Boston–Worcester–Lawrence, Massachusetts–New Hampshire–Maine–Connecticut CMSA	5,819,100	$37,311
6	Honolulu, Hawaii MSA	921,000	$36,339
7	Minneapolis-St. Paul-Bloomington, Minnesota MSA	3,478,415	$35,388
8	Hartford, Connecticut MSA	1,183,110	$34,310
9	Denver-Aurora-Lakewood, Colorado MSA	2,871,068	$32,399
10	Portland-Vancouver-Hillsboro, Oregon MSA	2,345,318	$31,377
11	Sarasota–Bradenton, Florida MSA	589,959	$30,344
12	Anchorage, Alaska MSA	260,283	$30,129
13	Baltimore-Towson, Maryland MSA	2,700,000	$29,771
14	Atlanta, Georgia MSA	5,544,577	$25,288
15	Madison, Wisconsin MSA	726,526	$25,163

Features which were examined include the ratio of passengers to clothing stores at airports, average income of metropolitan areas, and the types of existing venues within and nearby airports to determine where opportunities may lie to open a new outlet.

Methodology

Calculation of growth rate

A table from Wikipedia was scraped to retrieve data for passenger growth. First, the dataframe obtained from online data was cleaned, and filtered to only include large airports of over 10 million annual passengers to focus the analysis. The data provided passenger numbers for each year from 2012 to 2018. A large number of results from 2018 were missing, thus a 5 year growth rate from 2012-2017 was calculated for each airport as a separate column in the dataframe.

Since an owner would ideally seek to open a store in a location where foot traffic is projected to increase, only airports with a passenger growth rate higher than 10% were included for the analysis. This left out several major airports including Hartsfield-Jackson in Atlanta with a

relatively slow growth rate, and George Bush Intercontinental Airport in Houston which had a decline of passengers from 2012-2017. The resulting dataframe had a total of 21 airports, which were plotted for exploratory purposes.

Using the Geocoder library, latitude and longitude figures were then determined for each airport.

	Name	City	IATA	2012	2017	Latitude	Longitude	Growth
1	Los Angeles International Airport	Los Angeles	LAX	31326268	41232416	33.9422	-118.421	31.62
2	O'Hare International Airport	Chicago	ORD	32171743	38593028	41.978	-87.9093	19.96
3	Dallas/Fort Worth International Airport	Dallas/Fort Worth	DFW	28022877	31861933	32.8965	-97.0465	13.70
4	Denver International Airport	Denver	DEN	25799832	29809091	39.8502	-104.675	15.54
5	John F. Kennedy International Airport	New York	JFK	24520943	29533154	40.6429	-73.7794	20.44

This allowed for airports to be plotted on a map. Within Figure 1, each airport's marker size is indictive of the number of passengers, and the marker color represents the airport's growth rate from 2012-2017

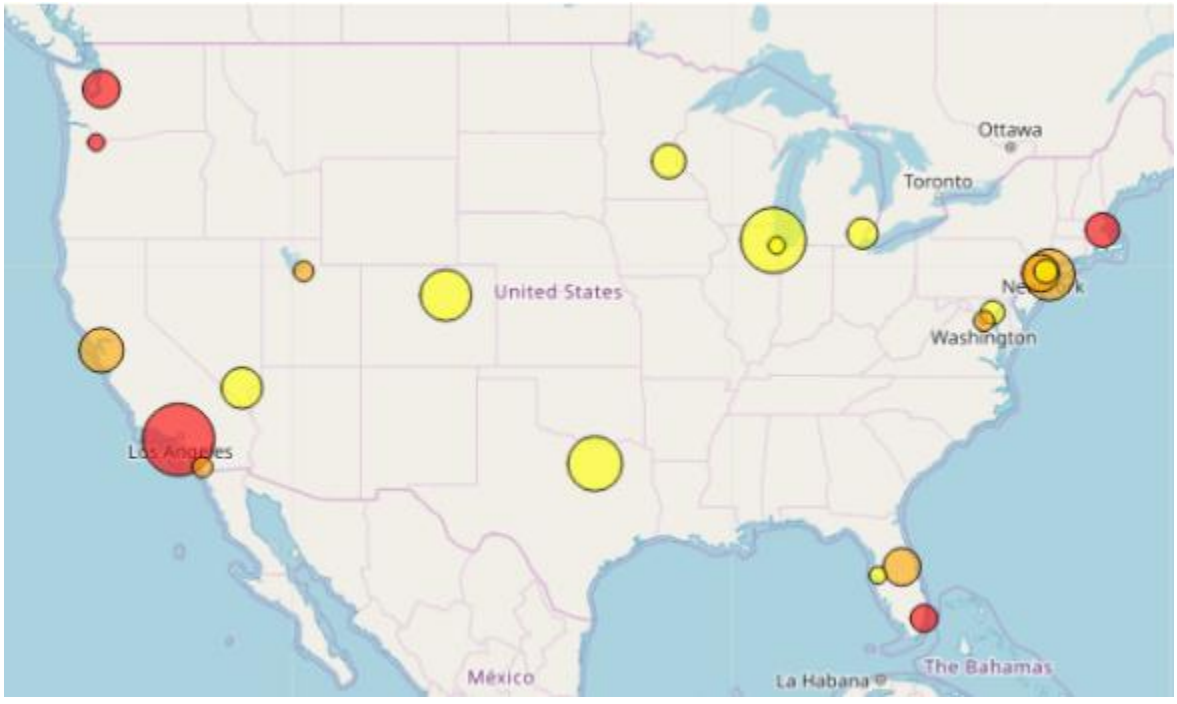


Figure 1: Map of airport locations, passenger volume and growth

Using Foursquare API to search venues for each airport

The Foursquare API's search endpoint was utilized to determine all of the clothing venues within each airport location. A category id relating to clothing stores was passed as a parameter to search only for existing clothing venues. A resulting dataframe was obtained, with each row as a clothing venue at an airport.

Performing an exploratory visualization, it was determined that clothing stores were the most common category, followed by boutiques and accessories stores.

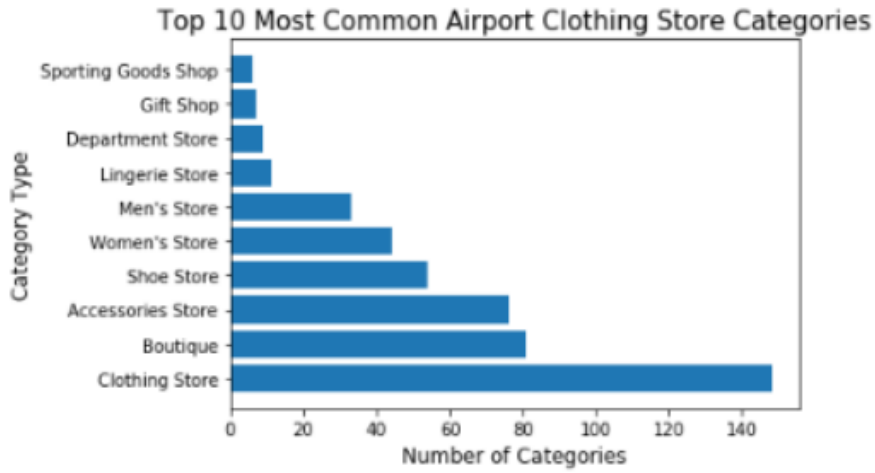


Figure 2: Most common airport clothing store categories

A count of all clothing venues was then obtained for each airport within a separate dataframe.

Airport	Store Count
Tampa International Airport	50
Ronald Reagan Washington National Airport	50
McCarran International Airport	48
San Diego International Airport	42
John F. Kennedy International Airport	35
Dallas/Fort Worth International Airport	31
Los Angeles International Airport	28
San Francisco International Airport	22
LaGuardia Airport	22
Minneapolis–Saint Paul International Airport	19
Orlando International Airport	18
Denver International Airport	16
Midway International Airport	15
O'Hare International Airport	15
Detroit Metropolitan Airport	15
ort Lauderdale–Hollywood International Airport	13
Seattle–Tacoma International Airport	13
Logan International Airport	12
Newark Liberty International Airport	11
Baltimore–Washington International Airport	10
Portland International Airport	9
Salt Lake City International Airport	3

Figure 3: Number of clothing store venues per airport location

Determining median per capita income for metropolitan areas

Wikipedia data regarding the mean individual income for all metropolitan areas was compiled. BeautifulSoup was then utilized to scrape data from a Wikipedia table. The data was obtained from a page depicting data from the 2010 US census for median per capita income by metropolitan area, which was the most recent comprehensive dataset found online. Upon performing a merge with the original dataset, a final comprehensive dataframe was generated with all of the featuresets required.

	Name	City	Growth	Ratio	Income
0	Baltimore–Washington International Airport	Baltimore	18.15	1321418.0	29771
1	Dallas/Fort Worth International Airport	Dallas	13.70	1027804.0	23616
2	Denver International Airport	Denver	15.54	1863068.0	32399
3	Detroit Metropolitan Airport	Detroit	11.06	1155040.0	22319
4	Fort Lauderdale–Hollywood International Airport	Fort Lauderdale	41.89	1247437.0	35828
5	John F. Kennedy International Airport	New York	20.44	843804.0	24581
6	LaGuardia Airport	New York	14.97	669902.0	24581
7	Logan International Airport	Boston	33.94	1595425.0	37311
8	Los Angeles International Airport	Los Angeles	31.62	1472586.0	21170
9	McCarran International Airport	Las Vegas	17.17	486754.0	21210
10	Midway International Airport	Chicago	15.89	727465.0	32258
11	Minneapolis–Saint Paul International Airport	Minneapolis	19.18	1000134.0	35388
12	Newark Liberty International Airport	New York	26.63	1961018.0	24581
13	O'Hare International Airport	Chicago	19.96	2572869.0	32258
14	Orlando International Airport	Orlando	25.68	1198080.0	21232
15	Portland International Airport	Portland	33.54	1059830.0	31377
16	Ronald Reagan Washington National Airport	Washington	26.46	239327.0	47411
17	Salt Lake City International Airport	Salt Lake City	26.29	4032945.0	24277
18	San Diego International Airport	San Diego	27.86	264454.0	22926
19	San Francisco International Airport	San Francisco	29.38	1222728.0	38366

Figure 4: Dataframe including median per capita income of airport's metropolitan area

Analyzing relationships between features

A final dataframe was obtained through merging the three dataframes of growth, venue count and income. Through creating scatter plots, two primary relationships were analyzed.

First, the relationship between income and airport growth rate was determined. The relationship showed a positive but relatively weak correlation. Metropolitan areas with higher incomes tend to have a workforce more concentrated in growth industries such as technology or financial services, and many experience a growing number of visitors as a result.

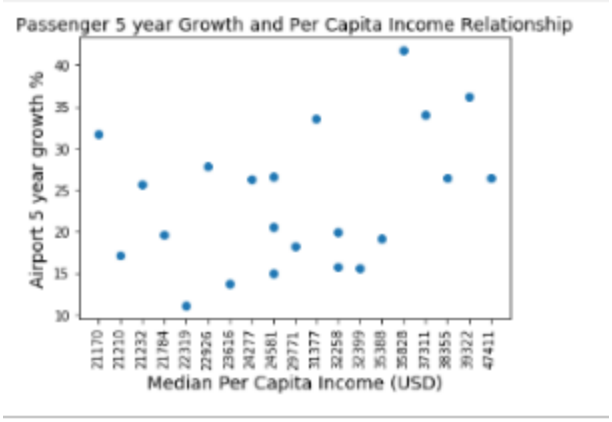


Figure 5: Relationship between per capita income and 5 year airport passenger growth

Moreover, the relationship between income and number of clothing stores was determined. This relationship was again relatively weak.

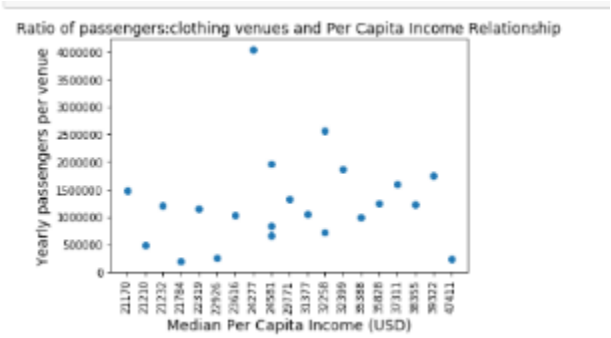


Figure 6: Relationship between per capita income and ratio of passengers per venue.

Overall, a higher number of clothing stores per passenger was generally positively correlated with both growth rate and income. The clustering algorithm would need to detect anomalies and isolate clusters where there were a relatively low number of clothing stores per passenger while still having relatively high passenger growth and passenger income.

Predictive Modelling

To prepare the data for clustering, the dataset values for growth, venue count and income were normalized using the z-score method, with the mean of each column set to 0.

	Name	City	Growth	Ratio	Income
0	Baltimore–Washington International Airport	Baltimore	-0.698716	0.063611	0.069118
1	Dallas/Fort Worth International Airport	Dallas	-1.257255	-0.287276	-0.781181
2	Denver International Airport	Denver	-1.026308	0.710916	0.432170
3	Detroit Metropolitan Airport	Detroit	-1.588613	-0.135221	-0.960358
4	Fort Lauderdale–Hollywood International Airport	Fort Lauderdale	2.255893	-0.024801	0.905878
5	John F. Kennedy International Airport	New York	-0.411288	-0.507167	-0.647868
6	LaGuardia Airport	New York	-1.097852	-0.714990	-0.647868
7	Logan International Airport	Boston	1.283156	0.391066	1.110751
8	Los Angeles International Airport	Los Angeles	0.991963	0.244266	-1.119090
9	McCarran International Airport	Las Vegas	-0.821720	-0.933864	-1.113564
10	Midway International Airport	Chicago	-1.007481	-0.646199	0.412691
11	Minneapolis–Saint Paul International Airport	Minneapolis	-0.569436	-0.320343	0.845093

Figure 8: Final dataframe with normalized features

Using Scikit Learn's K Means clustering algorithm, the sum of squared distance corresponding to the error was calcualated for each k from 1-11, and plotted on an elbow graph.

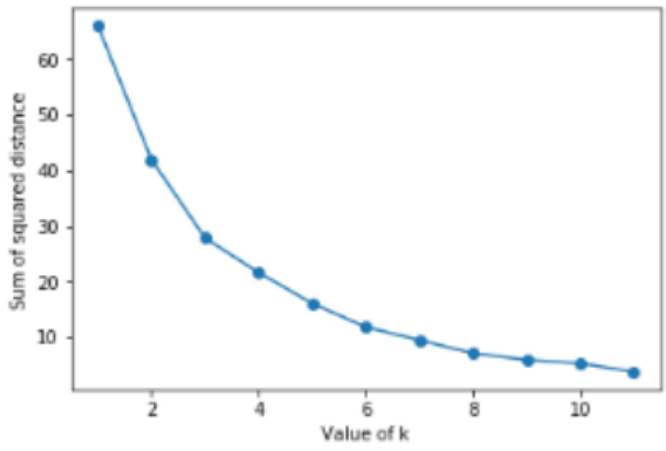


Figure 9: Elbow graph to determine best k value

When plotting the elbow curve for the clustering model, the best value of k was determined to be k=5. After this value, the curve begins to plateau.

Using the value obtained for k, the model has grouped airports into 5 distinct clusters. The resulting clusters were displayed in a map format to determine if clusters also happened to be influenced by geography. As evidence in figure 4, this hypothesis was determined to be unlikely.

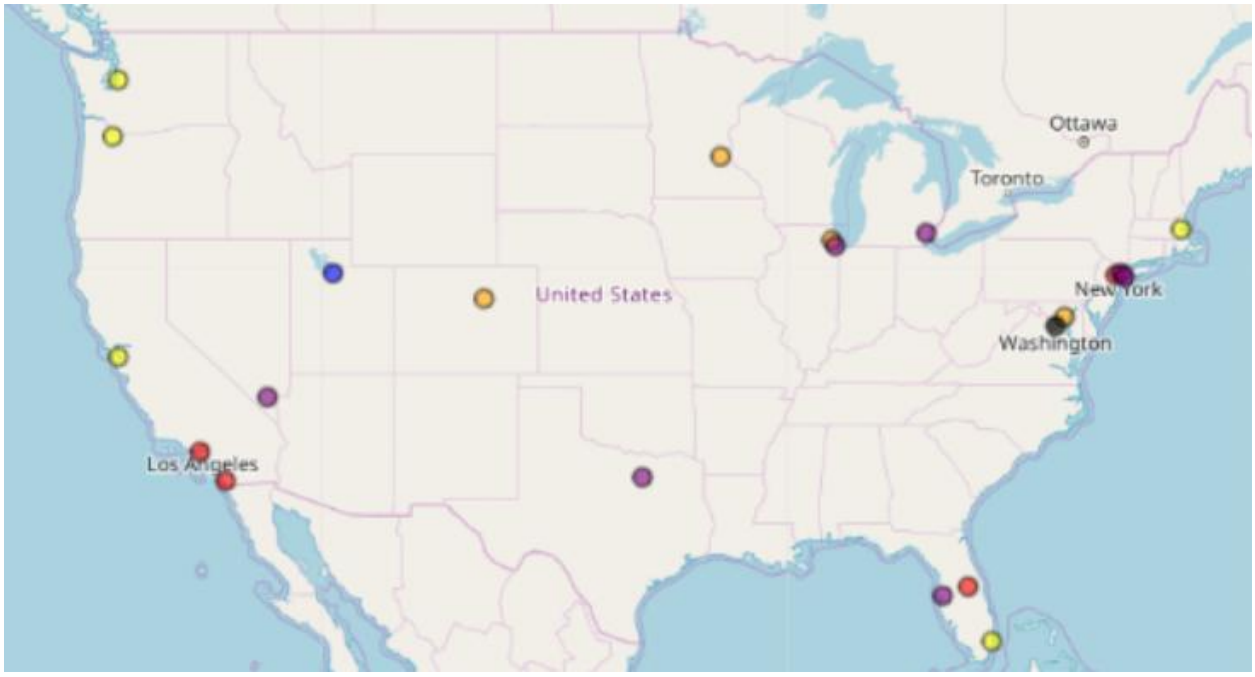


Figure 10: Map of clusters for airports

To determine the cluster containing the best suited airports, the mean values of growth, venue count and income were obtained for each cluster. From there, the cluster which had the greatest sum of mean values for all three features was determined to be the most ideal. This happened to be cluster 4, containing only Salt Lake City International Airport which is significantly under served by clothing stores according to its large passenger volume. While income is slightly lower than mean income of all other cities, it has above an average passenger growth rate and certainly presents a prime investment opportunity.

Also worthwhile of mention are airports in cluster 2, which were Seattle-Tacoma International Airport, Logan International Airport, Portland International Airport, Fort-Lauderdale-Hollywood

International Airport, and San Francisco International Airport. Although many had an average ratio of passengers to clothing stores, they all had very high passenger growth along with high median incomes.

Results

As shown by the data:

- The most common category of clothing stores in airports were not further classified, followed by boutiques, accessories stores and shoe stores. These can all be categorized into higher-end goods and as such, these stores would target higher income passengers.
- There was a positive relationship between income and passenger growth for airports, in addition to a similar relationship observed between income and number of clothing stores
- The best airport to open a new store was identified as Salt Lake City International Airport. Also noteworthy are airports in cluster 2: Seattle-Tacoma International Airport, Logan International Airport, Portland International Airport, Fort-Lauderdale-Hollywood International Airport, and San Francisco International Airport.

Discussion

The analysis conducted utilized a variety of data sources and analysis techniques to reach a final conclusion. The results generated appear to be relatively accurate based on qualitative perception. Salt Lake City is a metropolitan area which has undergone significant change over the past years, and are generally thought of to be well poised economically.

With that being said, the analysis conducted can certainly be improved and the model does have drawbacks. In this case, only three factors are considered in the clustering of airports, whereas there are a multitude more in reality. A more accurate model may incorporate many more feature sets to obtain more refined cluster groups. However, the model and analysis provides a good starting point to base decision making.

Conclusion

In conclusion, for this project I analyzed where the most optimal large airports in America to open a new clothing boutique venue based on factors of income, passenger growth, and ratio of passengers to venues. I utilized k-means classification to determine which group of airports have the highest metrics of all three categories and would be best served as a prime location. This report can be useful for those seeking to take a data-driven approach to determining the best location rather than simply relying on intuition. While the accuracy of this prediction can only be truly tested with time and careful analysis of future store openings at airports, it provides a data-driven approach for clothing store owners to utilize.

References

https://en.wikipedia.org/wiki/List_of_United_States_metropolitan_areas_by_per_capita_income

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