Final Project Submission

December 17, 2021

MGMT 590: Computing for Analytics

Jackson Bronkema, Kai-Duan Chang

Dr. Vardi

Brief Description of Our Dataset

The dataset we chose was from ArcGIS Hub, and it is a dataset that includes comprehensive information about every city in the United States that has a population greater than 7,900 people. This includes 3,886 cities in total. The dataset also includes a total of 51 different variables with specific information about each of these cities. Some of these include longitude, latitude, name of the city, state, population, population numbers by race, number of males and females, the number of people in the city across different age groups (e.g., ages 5-9, ages 35-44), median ages of people in those cities, number of households, number of families, average family size, and other characteristics as well. We did not use all of these variables in our analysis, but we used many of them in order to gain different insights about the data. We have attached the dataset if you want to see all the variables that were included. We have also given a link to the website, which goes into more depth about each of the variables in that data set.

We thought we could be most effective in this project and in gaining insights from our dataset if we used the data we had and created a business scenario. A brief description of our business scenario is below.

Business Scenario:

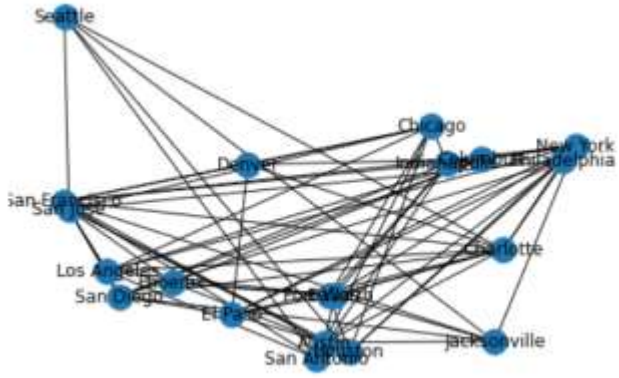
Our business scenario is that we are an airlines company that offers flights to different cities in the United States. We are based in New York City, and we currently offer flights from the 20 largest cities in the United States by population. Along with providing flights, our company also builds their own planes as well. Our goal is to use the techniques we learned in class to gain business insights about two different aspects of the business. First, our company is considering expanding and building a new plane production facility in one of the 20 cities where we offer flights. We will use information from our dataset, to decide which cities are the best candidates to build our factory in. We have used Linear Programming, K-means, and SVM as tools to help us identify which cities would be the best candidates for building our new factory. Second, we have used our dataset to track different costs and variables associated with travel. For example, in the graph algorithm, we are calculating the total distance traveled and fuel costs for each city that we fly out of in each week.

Insights from our Data

We used 5 different techniques that we learned in class in order to obtain the many insights that we were able to draw from our dataset. We did one graph algorithm, we did two different linear programming techniques (both were binary programs), and we did k-means clustering and support vector machines (SVM) from the machine learning portion of the class. Within each of these separate areas, we were able to get multiple insights. Our insights are in each section below.

Graph Algorithm

Our first program creates a graph where the nodes represent each of the cities that our company flies to/from and the edges are the flights that our company services each week. As you can see, we fly out of 20 different cities. There is a 40% chance that we fly from each of these cities to another city (to simulate different flight patterns each week). The graph is undirected, because each time we fly to one city, we will fly back to city that we flew out of. Also, we have used the latitude and longitude of each city to plot them in the place that they would be relative to other cities in the US.



totalDistance() uses BFS on the above graph (or any graph) in order to calculate and return 3 different business insights.

1. All flights that are scheduled out of each city and in which city each flight will be landing. Ultimately, these are the neighbors on the graph.
2. The total distance traveled out of each city (in miles), the expected fuel cost, and the number of planes needed/used in each city. In our research we found that jet fuel costs roughly $1.70 per gallon and that Boeing 737 get roughly 65 mpg. These numbers have been factored into our calculations.
3. The total distance traveled in flights across all cities for the entire week and the total cost in fuel for that week. Our source for how to calculate distances using latitude and longitude is included in our works cites page.

As we run BFS, our source node will always be New York City, since it is the first city in our dataset. We have designed it this way because New York City is the headquarters of our company.

Code Output:

```
There are 9 flights going out of New York this week. The destinations are: ['Los Angeles', 'Houston', 'Dallas', 'Jacksonvill
e', 'Columbus', 'San Francisco', 'Indianapolis', 'Charlotte', 'El Paso'] . A total of  9 planes will be needed.
The total distance of flights traveling out of New York is expected to be: 12185.61 miles


The total distance of flights for this week is: 178339.83 miles. For a total cost of: $ 4664.27 in fuel
```

findDistance() calculates the distance between any two cities in our dataset (out of 3,886 cities) and the fuel cost to fly there.

Code Output:

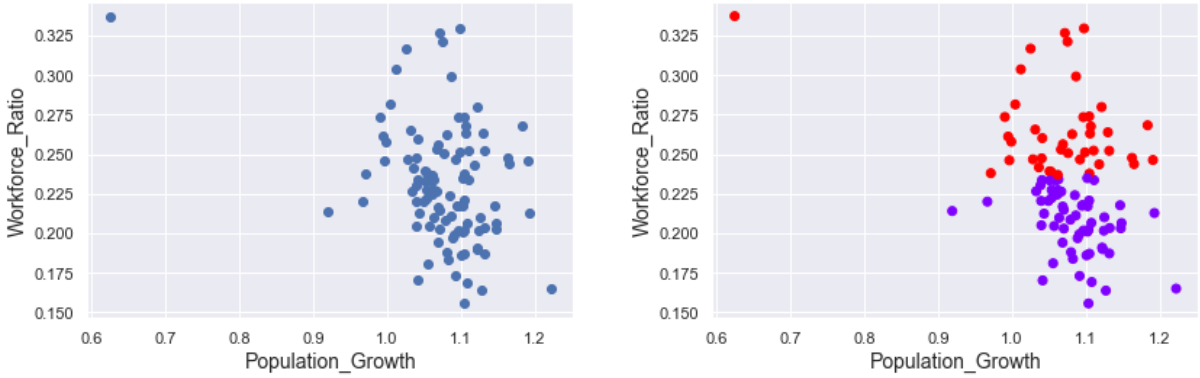
```
The total gas required to fly between Grand Rapids and Salt Lake City is estimated to be: $ 35.38 .
The distance between Grand Rapids and Salt Lake City is 1352.74 miles.
```

The insights from these programs are straightforward. Our company can track fuel costs, total distance traveled, and the number of planes needed in each city for a given week of flights.

K-means Clustering

To find out which city is suitable to build the factory. We have decided to focus on different workforce variables in our dataset because we want to build in a city that will have enough present and future workers who can provide labor in our factory. Thus, we conducted the K-means clustering to see if there have any relationships between workforce ratio and population growth in cities across the US. Here, we defined the workforce ratio as the percentage of people whose ages are between 20 to 34, and we divided it by the total population. Population growth is the total current population divided by the population of that city in 2010. To see the optimal number of clusters to use, we used the Elbow method and decided to use two clusters in our K-means clustering (elbow method included in the .ipynb file).

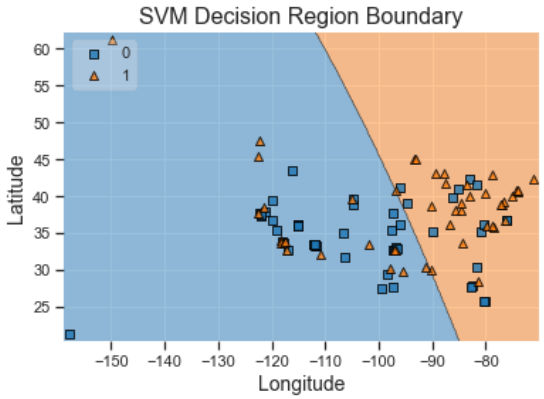
The higher workforce ratio in the top 100 cities in the first cluster includes LA, Chicago, Philadelphia, Dallas, and so on. The second cluster includes the city such as New York, Houston, Phoenix, and San Antonio. An insight we can take away from this is that if a city has a higher population, it does not mean it will have a higher workforce ratio. This result led us to the following model and led us to explore more relationships between workforce ratio and other variables using SVM.

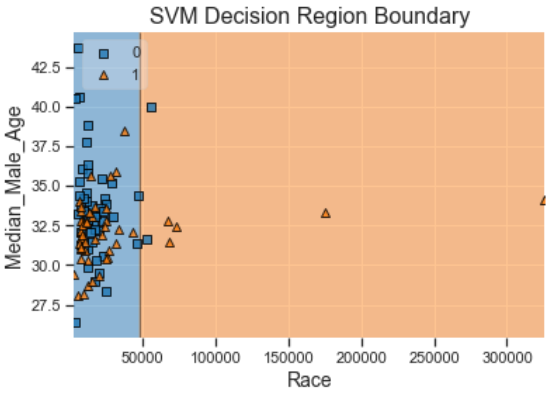


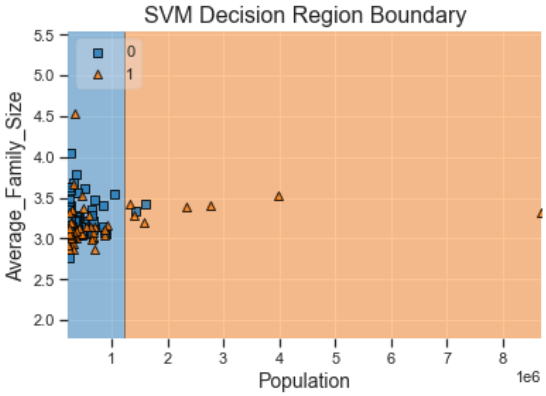
SVM

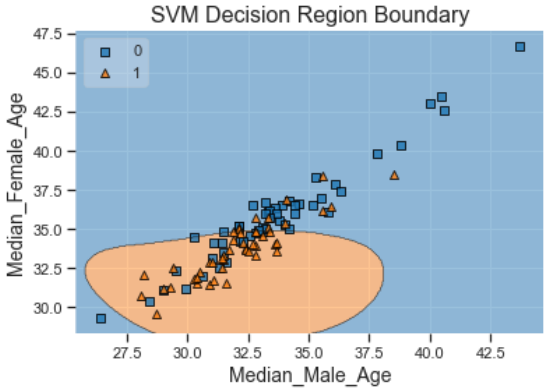
We used SVM to classify and examine possible relationships that existed between the workforce ratio (total number of people in each city age 20-34 / total city population) and some of the key variables in our dataset. In performing this classification, the workforce ratio will be plotted as a triangle (1) if it is greater than average, and as a square (0) if it is below average. Orange regions on the graph represent a workforce ratio greater than average and blue regions represent where the workforce ratio is less than average. We wanted to use SVM to determine which variables had the greatest effect or relationship with the workforce ratio. We also wanted to use this information to give us insights into how to structure the objective function and constraints in the linear programming portion of the project. Having a better understanding of how the key variables in our dataset relate to each other can help us to choose which city is best to build the production factory in as well.

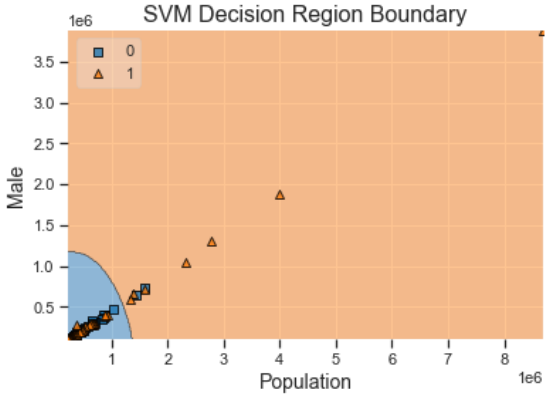
In the first SVM, we used latitude and longitude to see the distribution among the states. The SVM region boundary showed us that eastern states have a higher workforce ratio on average. This means that a larger portion of the population in eastern states is in the age range of 20-34. We also used multi-race and median male age to see if there exists any pattern. We can see that there is a relationship between the number of people who are multi-race living in a city and the workforce ratio. However, we did not consider it as a variable to use in the LP, because we think it could be an outlier, and would not be an effective way to choose a city to build our factory in. The two cities which have the largest number of multi-race people are New York and Los Angeles. Since our company has HQ in New York, we decided to look for other variables. We also looked at average family size and population. It appears that cities with an above average workforce ratio are large cities and those with an average family size around 3.0. These are variables that could help us to choose which city to build in and are ones that we used in the linear programming portion of the project. Next, we compared median male age and median female age. We can see there is a linear relationship because our workforce is based on the age range. Thus, the age range between 20 to 34 would classify as yes, and cities with lower median ages are those that also have a higher workforce ratio. The final relationship we examined was population and total number of males living in that city. The positive linear relationship is intuitive. From this, we can learn that cities with a larger number of males and with a larger population have a higher work force ratio on average. Thus, the total number of males and total population are insightful and useful variables to consider in our linear programs.











Linear Programming

We have created two different linear programs with the goal of identifying candidate cities to build our new factory. The goal for each of our programs is to choose 4 different cities that we would consider building our production factory in based on certain characteristics of those cities. These are binary programs since each city will be assigned either a 0 (not chosen) or a 1 (chosen).

LP 1

In LP1, we are using different work force characteristics of each city to choose candidate cities that we would want to build our factory in. Our goal is to maximize the total number of people living in each of these 4 cities who are age 20-44, as these are the ages of people most likely to work in our factory. We have also created 4 different constraints to help us choose these cities.

```
#contraints
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'male']) for i in dfsmall.index]) >= 120000
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'mid_age_male']) for i in dfsmall.index]) <= 124
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'age10_14']) for i in dfsmall.index]) >= 200000
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'age15_19']) for i in dfsmall.index]) >= 120000
```

Ultimately, we want the number of males in the 4 cities our LP chooses to average to be greater than 30,000 (120,000/4), the median age of the males in those cities to be less than an average of 31 (124/4), and the number of young people living in those cities to be large as well. We chose these constraints, because we want to build our factory in a city that has a large present work force, for there to be many young males living in that city, and for there to be a large future workforce as well.

Based on our objective function and our constraints, the insight we gained from LP1 is that the four best cities to build a factory in based on work force statistics are Los Angeles, Houston, Austin, and Columbus.

LP 2

In LP2, we are using city population and growth characteristics of each city to choose 4 candidate cities that we would want to build our factory in. Our goal is to maximize the total growth rate (current population / 2010 population) in each of these 4 as we want to build our factory in a city that is growing quickly and can provide enough workers for future years as well. We have also created 4 different constraints to help us choose these cities.

```
#contraints
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'midage']) for i in dfsmall.index]) <= 130
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'avg_fm_size']) for i in dfsmall.index]) >= 12
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'population']) for i in dfsmall.index]) >= 2000000
prob += lpSum([data[dfsmall.loc[i,'city']]*int(dfsmall.loc[i,'age15_19']) for i in dfsmall.index]) >= 120000
```

In using these specific constraints, we want the median age of the people living in the 4 cities our LP chooses to average to be less than 32.5 (130/4), the average family size to be greater than 3.0 (12/4), the average population of each of the 4 cities is greater than 500,000 (2,000,000/4), and where each city has an average number of people of age 15-19 greater than 30,000 people. We chose these specific constraints, because we want to build our factory with a large population, that is growing quickly, we felt that family size, median age, population, the number of young people in those cities were good proxies for population and city growth.

Based on our objective function and constraints, the insight we gained from LP2 is that the four best cities to build our factory in based on city growth statistics are Austin, Fort Worth, Charlotte, and Denver.

Final Decision: Austin was one of the cities chosen in both LPs, we will suggest building that factory in Austin, Texas.

Sources

Dataset: USA Major Cities | USA Major Cities | ArcGIS Hub

Calculations for distance between cities: Power Apps Guide - Formulas - How to calculate the distance between 2 points - Power Apps Guide - Blog

Fuel Economy: How Much Does Jet Fuel Cost? | The Price Of Jet A1 | FlightDeckFriend.com, Fuel Economy In the Sky: Whose Jets Get the Most Mileage? (greencarreports.com)