**IBM Data Science Professional Certificate**

Applied Data Science Capstone

**Battle of the Neighborhoods, (Week 2) Final Project Full Report**

By Steven Harrison

April 22, 2020

**Introduction**

My project is developing a **data science application (using location data)** where the data would be used to support the business plan and decision to **relocate and establish a new Computer Service and Repair business in the northern part of Colorado Springs, Colorado**. The data would consist of a proposed market area (location), business competition from other computer service businesses (proposed market), hotels located in target market to support our ‘mobile computer service’ niche, and finally income data for our neighborhoods in order forecast a sustainable revenue stream and develop prices based on various competitor services.

The data from the four sources would give a **bird’s eye view of our target market** giving the entrepreneur the decision-making information necessary to reduce the business and financial risk of starting a new business in a new state, while ensuring and increasing the chances of building, maintaining, and growing a successful business in Colorado Springs, CO. People interested in the data output and analysis would be the entrepreneur, their family, investors, creditors, banks, and insurance companies who will provide the financial capital and share the business risk.

**Data Sources**

I used two sources of online information to ascertain and query the data I needed to build my data science application. The application’s data retrieval was not built into the application. It was first queried directly through the browser where the online data service provided the JSON location data file then imported into the application. The application uses Python to upload, parse, and clean the raw JSON data files isolating the location data we need for our mapping and analysis.

These are the two sources of online data used in my application:

1. **FourSquare Places API** – used to query the business competition data and hotel ‘niche’ data for mobile computer services. I used zip code data to isolate the target market where I wish to locate my business and deliver goods and services. The FourSquare data needed to be cleaned in two areas. The first was latitude and longitude format (consistency in lengths), and semantic data values where the data not have the correct meaning in the dataset. An example is Hotels, where various Pet Hotels, Cat Hotels came up in the result set.
2. **IncomeByZipCode.com** – used to query the median income data for our five target markets in North Colorado Springs. The Site is an Online Data Science application business that provides income data for all US Zip Codes and provides map data services providing shape files for each zip code. The Zip Codes, Neighborhoods, Latitude, Longitude, and Median Income data were all uploaded into our Jupyter Lab (notebook) via a cleaned Excel CSV file and imported into our application for rendering.

**Data Science Methodology**

I was thorough in writing the Week 1 **business understanding** part of the project. It was from that CRISP-DM format that I was able to isolate the locations I was interested and identify the specific data I needed to analyze my business markets and competition. Since the data sets I needed were very specific, I could use Zip Codes to find and query the location data I needed for my application and analysis.

1. **Analytic Approach** – I am using the **Data Science visualization tools** to identify the market, business competition, hotel location for my mobile computer service niche, and income data to ensure a healthy economy where I want to locate my new business. Since the market location and data are very specific, I **did not** need to utilize any of the supervised or unsupervised machine learning models. Using the folium visuals, I can see from a bird’s eye view what my markets look like and their characteristics. I can see where my competition is located. For example, I could see there is **much opportunity in the north part of Colorado Springs** for a new computer repair and service business. Most competition is located to the downtown south, business/airport area, and the university location of Colorado Springs.
2. **Data Collection, Preparation, and Modeling** – I used the FourSquare Places API directly thru the web browser. Since what I was looking for was specific, the dataset and JSON file returned a small finite set of location data. I imported my JSON files into my labs.cognitiveclass.ai Jupyter Labs IDE and used Python the import, clean, and format the data into a Pandas Data Frame for map rendering and visualization. I wrote Input Output Machines (Python Functions) to process the incoming FourSquare JSON files. I had to manually clean the data and fix the inconsistent Latitude and Longitude location format (length). I also had to remove semantic data that was returned in the result set.

I modeled my data using specific location and Zip Code information. The Zip Codes were also mapped to their mean latitude and longitude in Colorado Springs. I mapped each data set individually, then combined my data sets into one single map. I also mapped Income data based on queried data from IncomeByZipCode.com. This income data was also based on zip code/latitude longitude information. Again, I had to manually clean the latitude and longitude formats returned in the result set.

I color-coded my folium map data objects to identify my market, business competition, hotels, and income data. The color visuals were used to analyze my markets from a bird’s eye.. all in one view.

**Result Section**

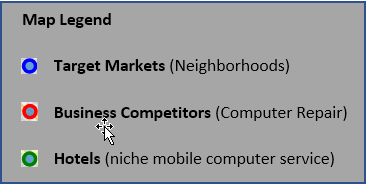
After looking at the rendered data and folium maps, **it is a good decision to relocate and start a new computer service and repair business in the north part of Colorado Springs**. Most of the business competition is located in the downtown/university area and the airport business market area. Most residents live in the north part of Colorado Springs **and there is much opportunity** in the households and Hotels along the I-25 interstate and the North Academy business area. Also, the resident’s median income is at a level that provides reasonable pricing of services (a stable economy/income stream) and located in an area that’s readily accessible (without traffic) for mobile service delivery.

**Discussion**

The FourSquare location data service is not perfect. For instance, if you query Hotel using the API client\_id and client\_secret, you get listings that have ‘hotel’ in the name, not ‘hotel’ as identified as metadata or characteristic of the **hotel location data record**. For instance, when I query Hotel, it gives me example **The Raddison Hotel** at Broadmoor as ‘Hotel’ is part of the name. The result set does not include Quality Inn or Comfort Inn as the term ‘Hotel’ is not part of the name. Also, the FourSquare location data returns data that is not part of the ‘logical meaning’ of what you are looking for. For example, in the result set, you get location data on ‘Pet Hotel’ and ‘Cat Hotel’ and Hostels.

I also learned that there is a variety of methods and API’s available that provide access to large data sets at almost no cost. An example is the United States Census API, where they provide a downloadable Python API dependency to access and query US Census data. There are also websites that are actually data science applications selling formatted data as a business. An example is the IncomeByZipCode site.

I tried to create **Legends** for my visual analysis (maps) in my application. I used the following method but had trouble rendering the legend on my folium maps…



1. **Legend.png**

Uploaded to

Jupyter labs

I used the following code to render my .png graphic. I used the **print(folium.\_\_version\_\_)** to ensure version 0.5.

**from** folium.plugins **import** FloatImage

image\_file = 'legend.PNG' # my png file above

**FloatImage**(image\_file, bottom=5, left=5).add\_to(Colorado\_Springs\_Map6)

Folium did not render the legend on my map. It just showed ‘image\_float’… After hours of troubleshoot, I just moved on. However, **I understand a legend is necessary for a good data analysis and understanding of the map data.**

**Conclusion**

I learned a ton of new tools and methods to find and present data and use that data to make decisions. I imagine this is just the tip of the iceberg (so to speak) of data science and there is much more to learn and practice. Yes.. PRACTICE PRACTICE and more PRACTICE. I would like to write a data science application that is based on epidemiology and how the COVID-19 virus moves thru our part of Arizona.. the Navajo Nation in north eastern Arizona. You get many data science presentations on the number of infections and recoveries, but you can’t see how it travels into and thru our section of the country.

I would also like to revisit the machine learning parts of data science. What was presented was quite overwhelming for a first-time exposure, so we have go back, study, and keep practicing. I look forward to more software engineering training and **COURSERA and IBM did a great job** of designing the course material, how it’s presented and taught, and the hands-on programming of the projects. It was also great practice using the Python language.

I recently completed the Duke University Data Science Math Skills. I did not know that a math algebra function in Data Science is referred to an Input Output Machine. How cool!!!

Again, thank you COURSERA and IBM and the students for all your hard work making the class available and a success to those who want to work hard and learn new complex things. And yes.. We need a vacation.. lol…