IBM Unsupervised Machine Learning

Course Final Project:
Major Projects Clusters by Region

By Javier A. Jaime-Serrano

June 15, 2021

Abstract

For this Project we used the Major Project dataset from the province of Alberta in Canada [1], it contains more than 700 currently active projects on the province, this Data set if filtered for Projects valued at $5 million or greater.

This Dataset contain a lot of valuable information on the Major Projects.
The Estimated Cost, Sector & Type and the Location data will be used.

We will like to find the features that can predict the cost of the project for a given project type in an industry sector and a geographical region.

We will be using to unsupervised machine learning algorithms to cluster the location data into regions that will be used to estimate cost of the projects.

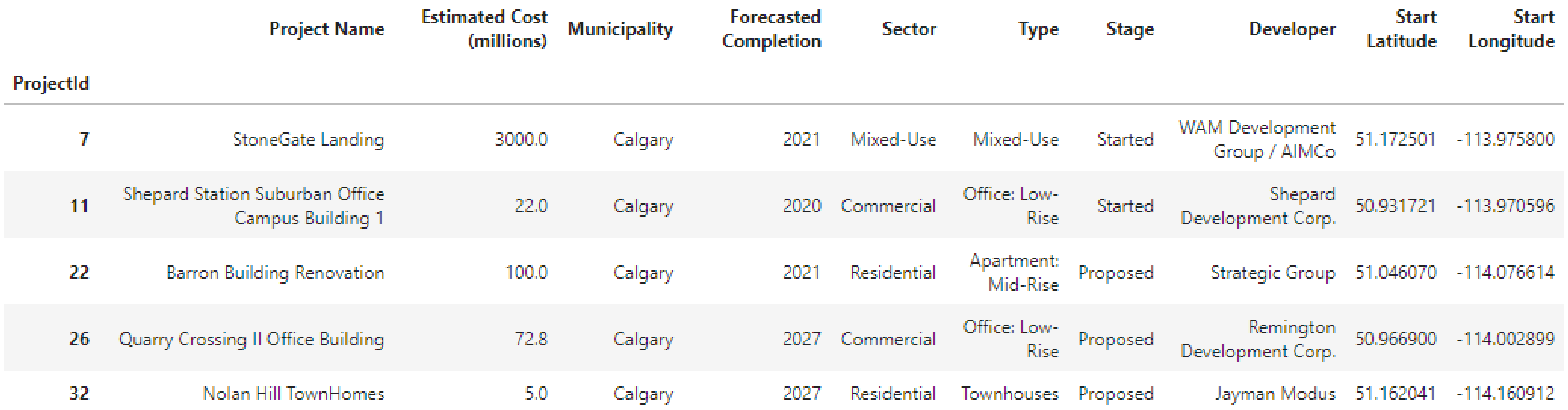
Data Cleaning

In order to prepare and clean the dataset:

- We drop the projects where there is no estimated cost.
- Made some assumptions about schedule completion and status.
- Drop not required columns and renamed the remaining columns.
- Made corrections on project types and sectors.
- A problem encountered was how to extract the location coordinates (Longitude & Latitude) from a GeoJASON column. The problem was solved with Python code that loop over all rows and extract the start (first) locations by a type condition.

Data Cleaning

Figure 1. Cleaned Dataset



	Project Name	Estimated Cost (millions)	Municipality	Forecasted Completion	Sector	Type	Stage	Developer	Start Latitude	Start Longitude
ProjectId										
7	StoneGate Landing	3000.0	Calgary	2021	Mixed-Use	Mixed-Use	Started	WAM Development Group / AIMCo	51.172501	-113.975800
11	Shepard Station Suburban Office Campus Building 1	22.0	Calgary	2020	Commercial	Office: Low-Rise	Started	Shepard Development Corp.	50.931721	-113.970596
22	Barron Building Renovation	100.0	Calgary	2021	Residential	Apartment: Mid-Rise	Proposed	Strategic Group	51.046070	-114.076614
26	Quarry Crossing II Office Building	72.8	Calgary	2027	Commercial	Office: Low-Rise	Proposed	Remington Development Corp.	50.966900	-114.002899
32	Nolan Hill TownHomes	5.0	Calgary	2027	Residential	Townhouses	Proposed	Jayman Modus	51.162041	-114.160912

Data Exploration

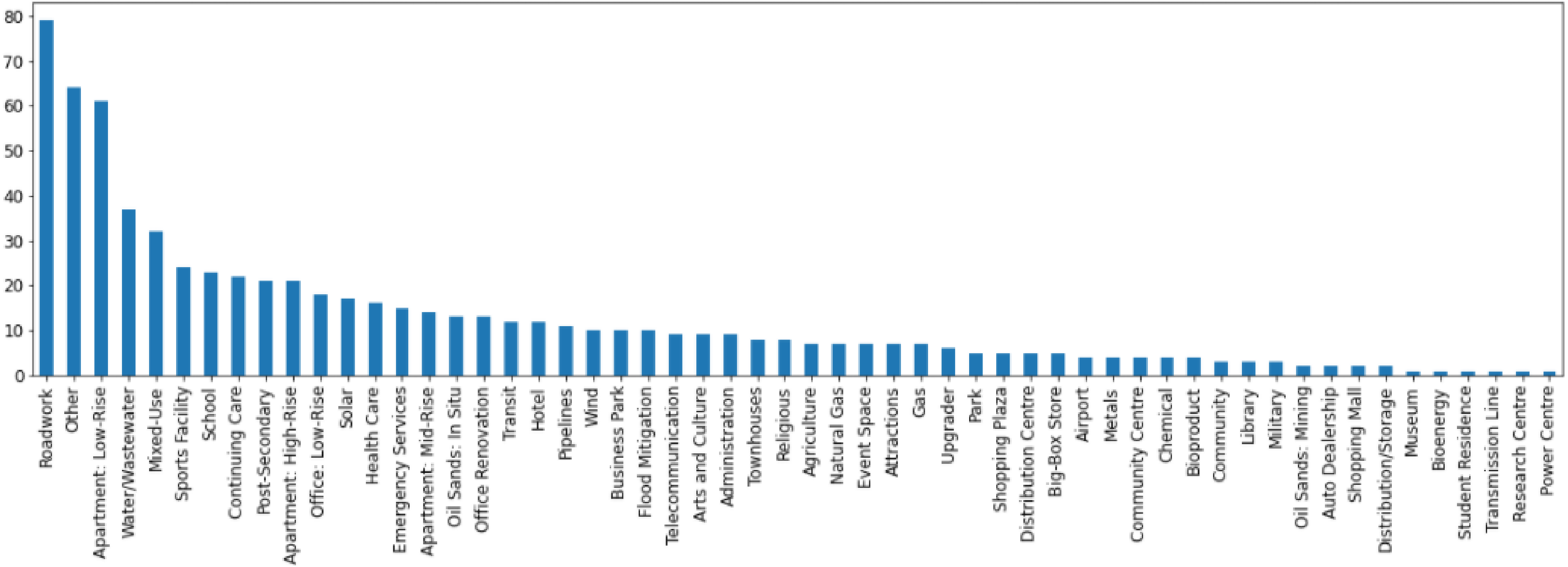
We explored the data, first with descriptive statistics and bar charts (see Figure 2).

Second, we used box plots for the cost estimate ranges by type (see Figure 3).

Third, we used folium library to create a map using latitude and longitude values (See Figure 4).

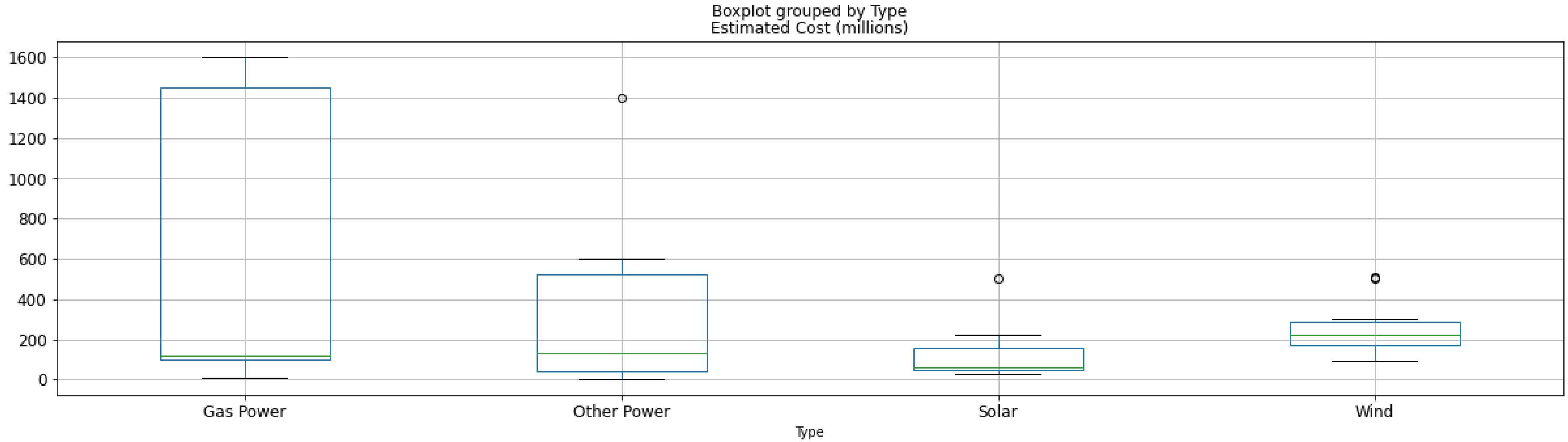
Data Exploration

Figure 2. Project types bar chart



Data Exploration

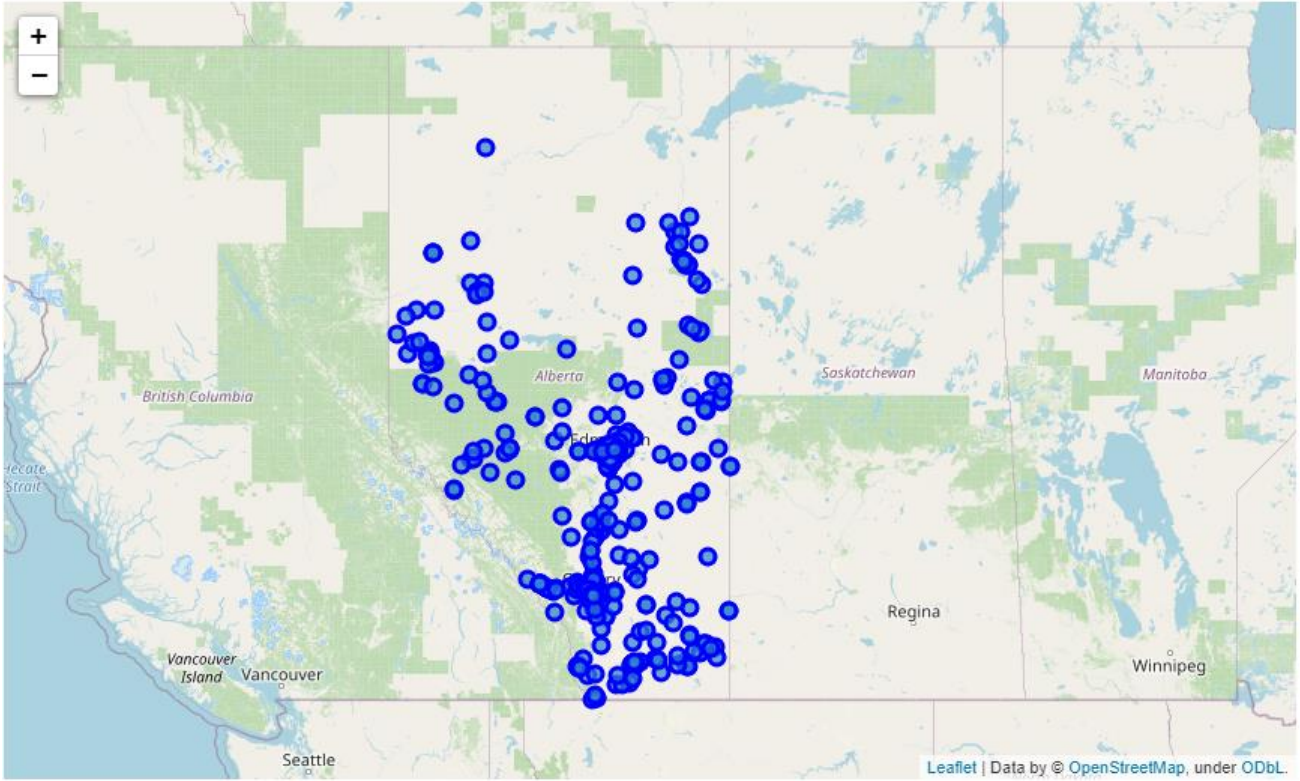
Figure 3. Box Plot Estimated Cost for Power Sector



Data Exploration

Figure 4.

Projects Map



Key Findings and Insights

After corrections, we ended with 58 unique types in 9 sectors (see figure 5).

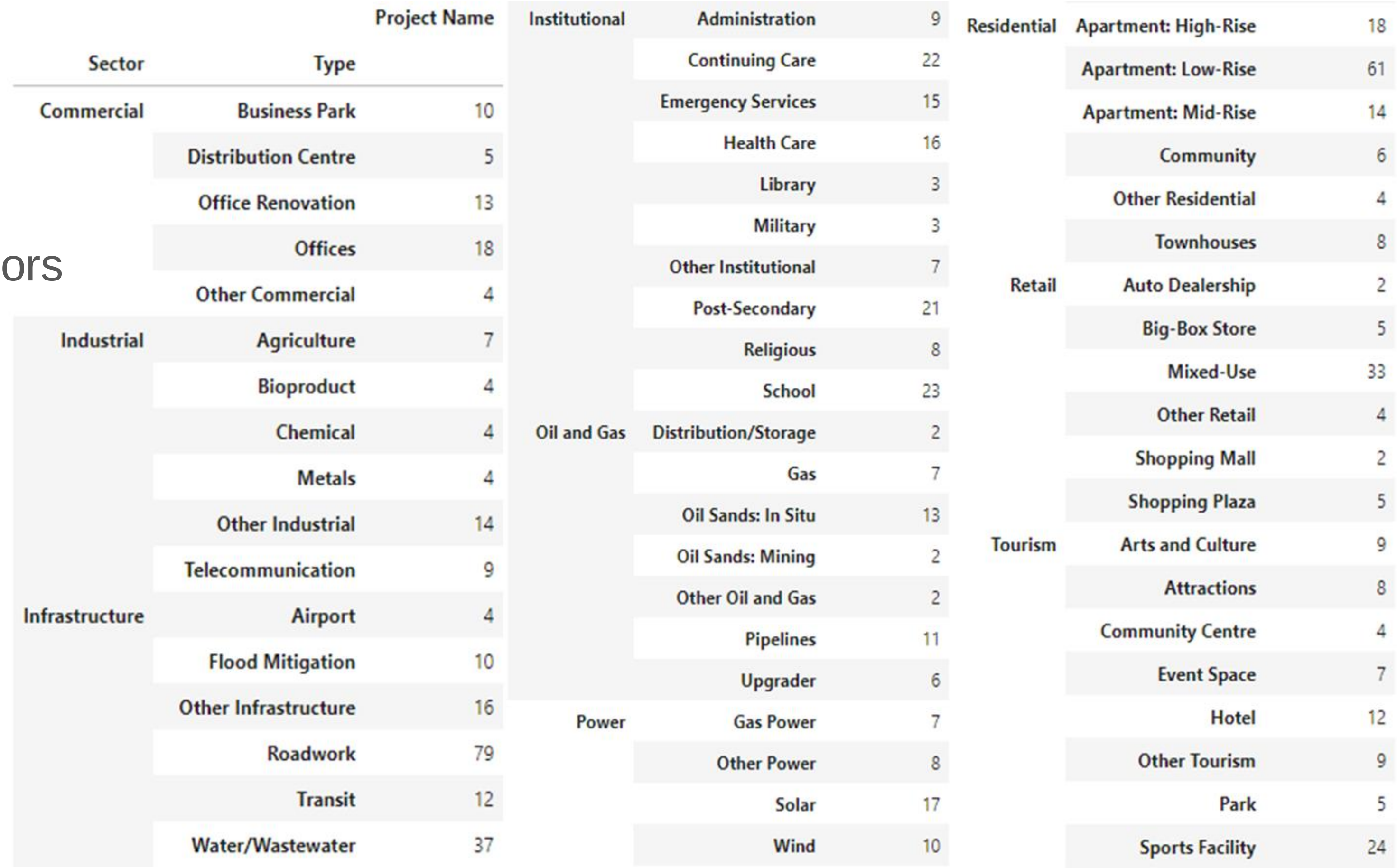
We used KMeans [2] to Cluster the projects by geographical region and type using the location values and one hot encoding of project types (See Figure 6).

We choose to minimize the inertia without splitting the major cities of the province, after a few runs we found that a k = 10 clusters have a minimum inertia of 866.

We also used DBSCAN to Cluster the project by region using the location coordinates only, but we were not able to separate the major cities (See Figure 7).

Findings

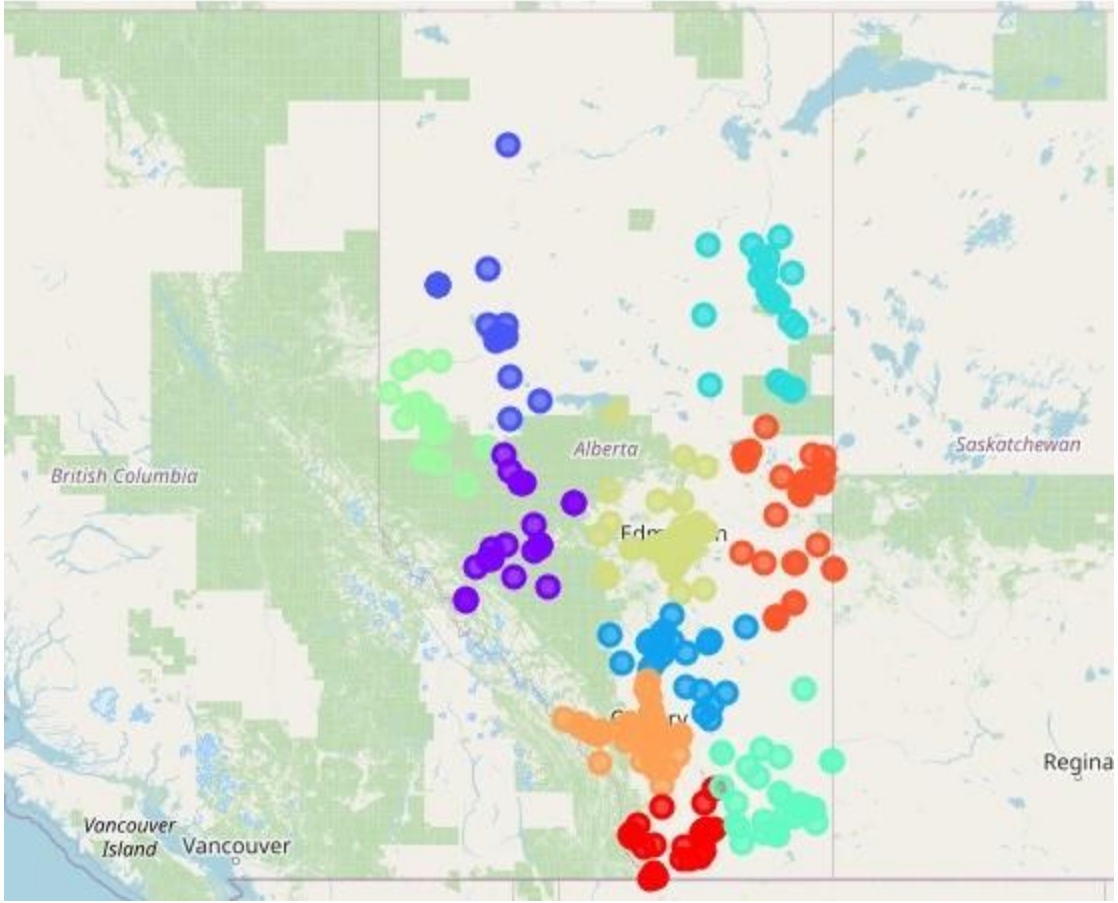
Figure 5. Unique Sectors and Types



Sector	Type	Project Name
Commercial	Business Park	10
	Distribution Centre	5
	Office Renovation	13
	Offices	18
	Other Commercial	4
Industrial	Agriculture	7
	Bioproduct	4
	Chemical	4
	Metals	4
	Other Industrial	14
	Telecommunication	9
Infrastructure	Airport	4
	Flood Mitigation	10
	Other Infrastructure	16
	Roadwork	79
	Transit	12
	Water/Wastewater	37
Institutional	Administration	9
	Continuing Care	22
	Emergency Services	15
	Health Care	16
	Library	3
	Military	3
	Other Institutional	7
	Post-Secondary	21
	Religious	8
	School	23
Oil and Gas	Distribution/Storage	2
	Gas	7
	Oil Sands: In Situ	13
	Oil Sands: Mining	2
	Other Oil and Gas	2
	Pipelines	11
	Upgrader	6
Power	Gas Power	7
	Other Power	8
	Solar	17
	Wind	10
Residential	Apartment: High-Rise	18
	Apartment: Low-Rise	61
	Apartment: Mid-Rise	14
	Community	6
	Other Residential	4
	Townhouses	8
Retail	Auto Dealership	2
	Big-Box Store	5
	Mixed-Use	33
	Other Retail	4
	Shopping Mall	2
	Shopping Plaza	5
Tourism	Arts and Culture	9
	Attractions	8
	Community Centre	4
	Event Space	7
	Hotel	12
	Other Tourism	9
	Park	5
	Sports Facility	24

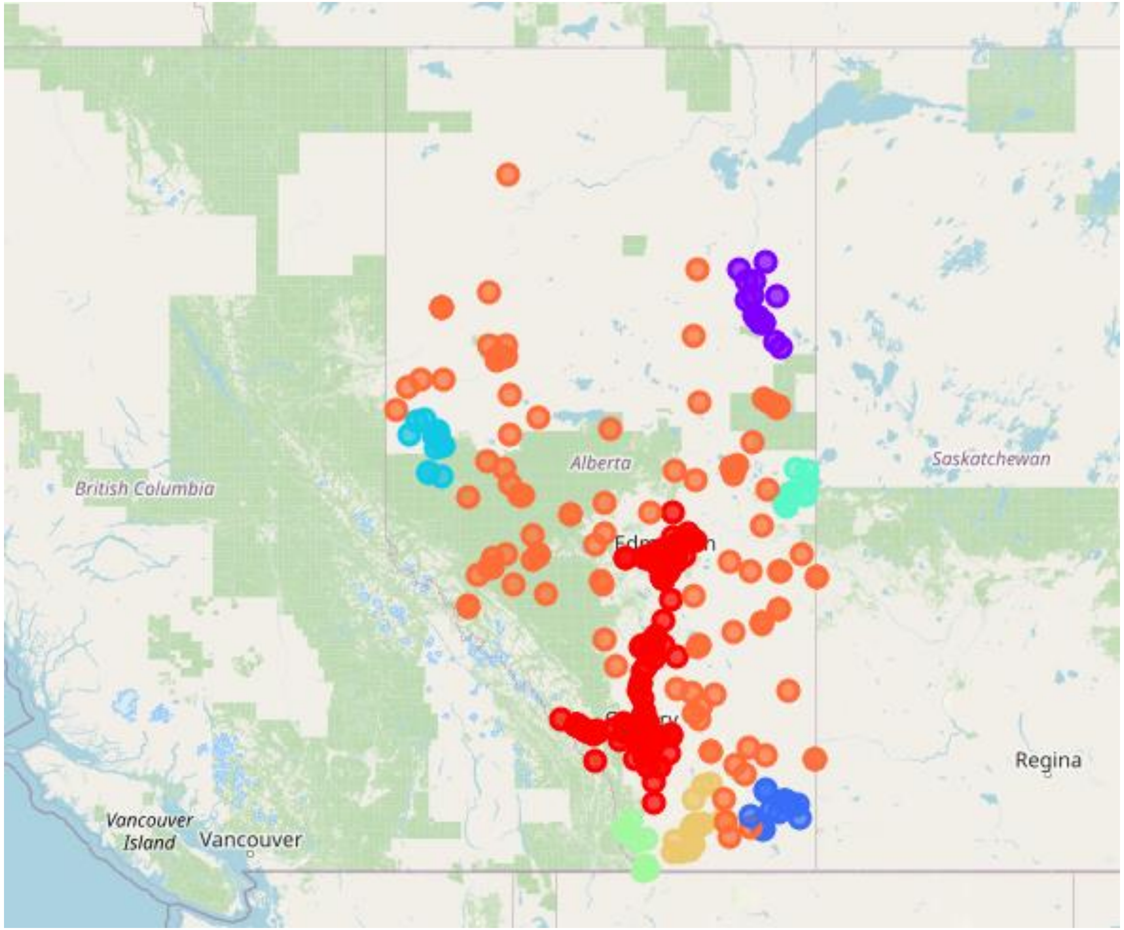
Findings

Figure 6. Projects Map KMeans Clusters by Region with Project type



Findings

Figure 7. Projects Map DBSCAN Clusters by Region



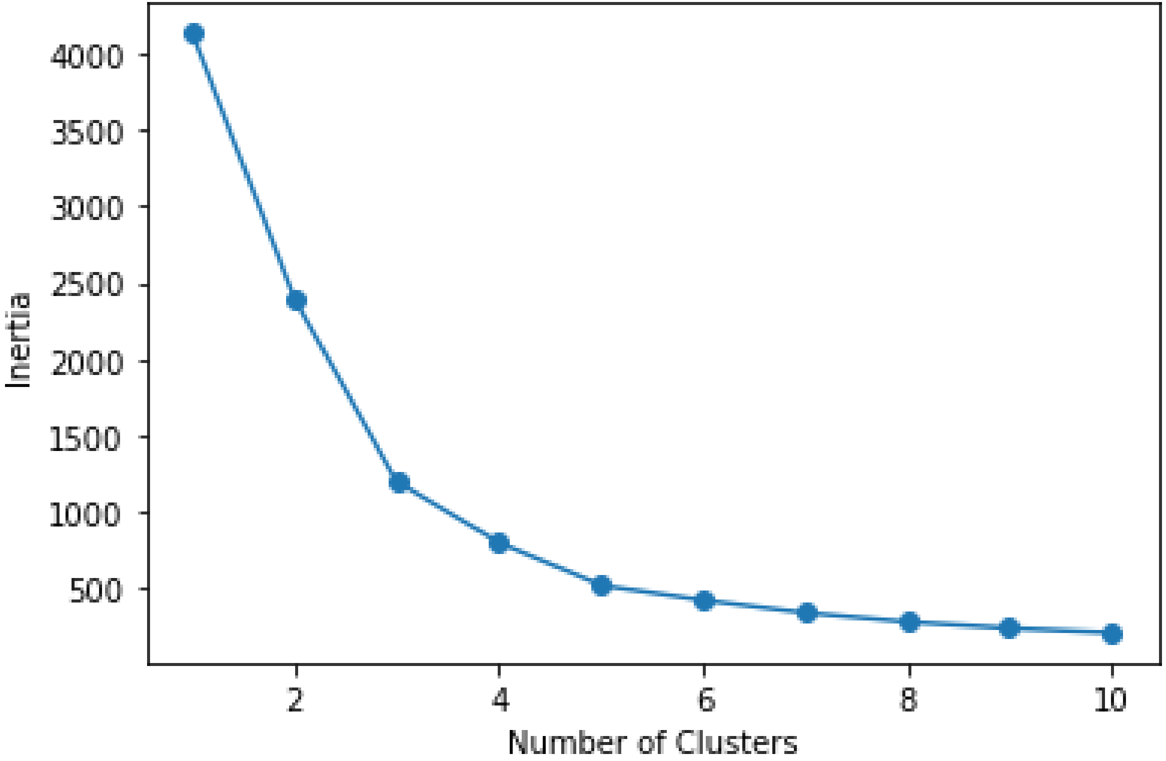
Results

We used KMeans with only the location coordinates and loop through the number of cluster and found the minimum inertia of 211 with k = 10 clusters, also keeping all the major cities in one cluster each (See Figure 7).

The resulting clusters are identical at the ones found using the one hot encoding of the project type (See Figure 6), so there was no need to use the project type.

Results

Figure 8. Inertia Vs. Number of Clusters



Conclusions

Using the Major Project dataset from the province of Alberta, after preparation and cleaning, we were able to extract valuable but limited information.

We found that KMeans was better able to split the Projects location coordinates in regions keeping the major cities in separate clusters, with the minimum inertia.

The use of DBSCAN created only one cluster for the main industrial corridor, and did not comply with the requirement of separating major cities.

This project was done with only one Province dataset, and other Canadian Provinces has datasets available for Future work in similar formats.

References

[1] Alberta Major Projects: https://majorprojects.alberta.ca/

[2] Scikit-learn library: https://scikit-learn.org/stable/

[3] Jupiter Notebook: https://github.com/javier-jaime/IBM-Machine-Learning-Capstone/