IBM Exploratory Data Analysis for Machine Learning

Course Project

By Javier A. Jaime-Serrano

June 4th, 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.

From this dataset, we will like to find the features that can predict the cost of the project for a given project type in a sector.

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

ProjectId	Project Name	Estimated Cost (millions)	Municipality	Forecasted Completion	Sector	Type	Stage	Developer	Start Latitude	Start Longitude
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

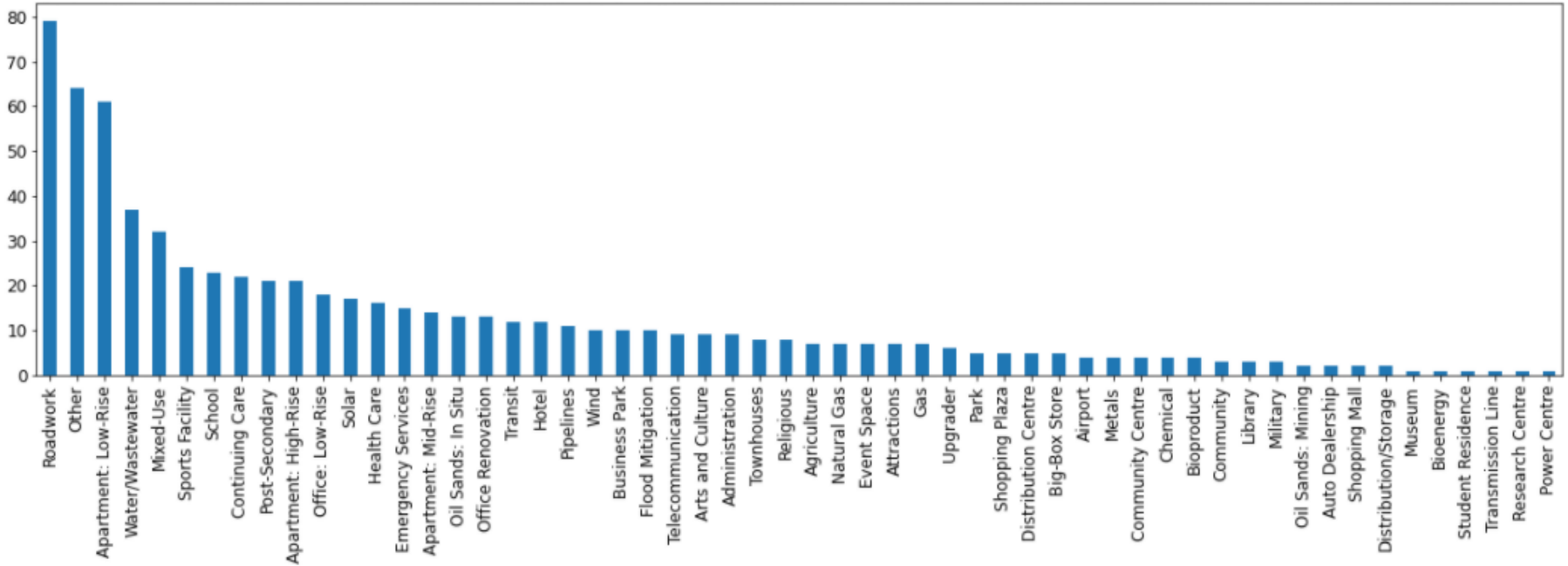
We explored the data, first with descriptive statistics and bar charts (see Figure 2).

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).

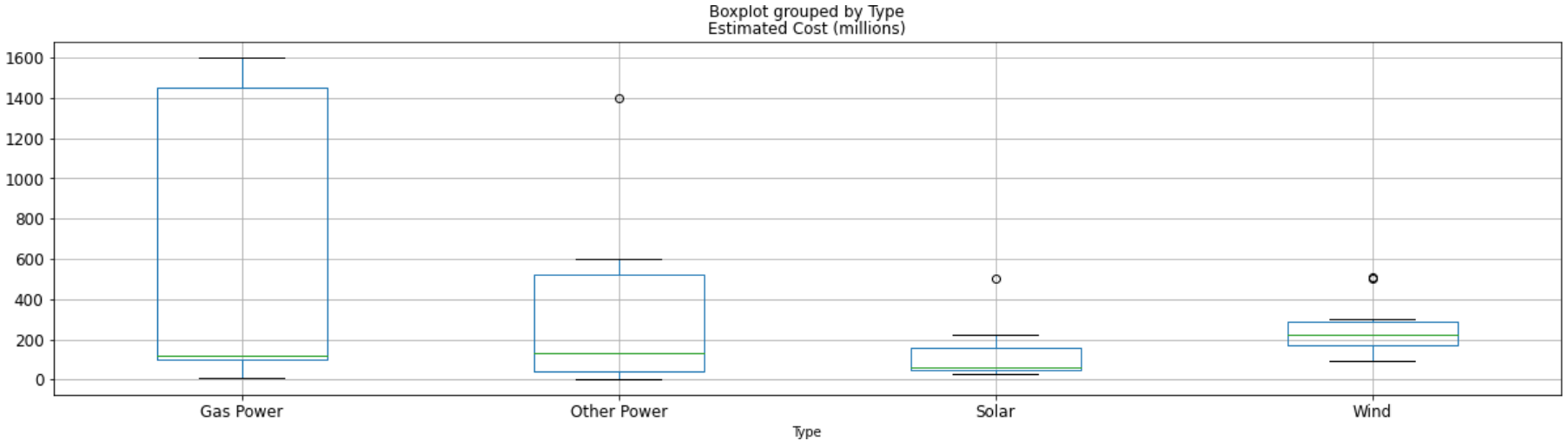
Data Exploration

Figure 2. Project types bar chart



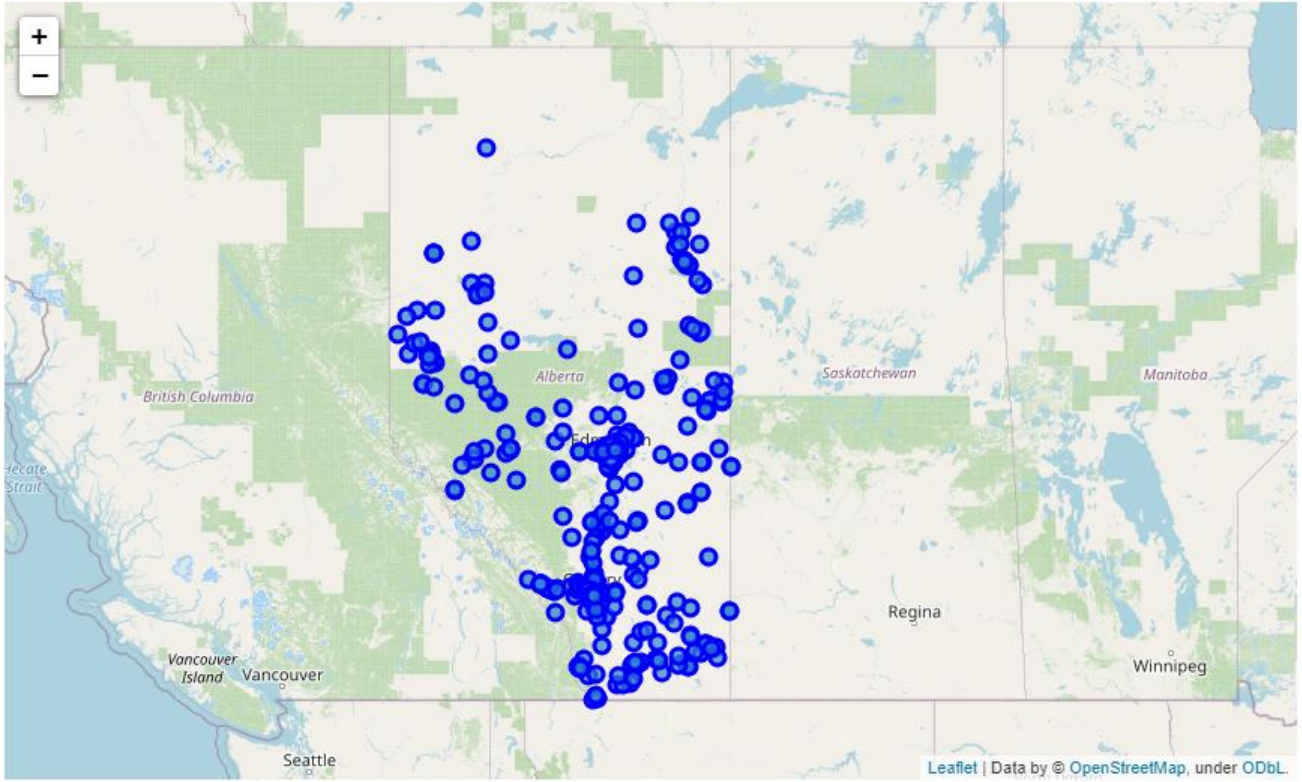
Data Exploration

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 6).

For the feature engineering, we extracted size and capacity data by type from the dropped project details column, adding units and cleaning it manually in excel.

A filter was set to remove extreme values, from a percentage range in relation to average cost, divided by capacity per type of projects, and labeled with conditional formulas in excel.

Only 97 projects were properly classified as “In the Range”, and Only 3 types has more than 10 Projects in this valid range.

Findings

Figure 6. 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

Hypothesis formulation

We want to be able to estimate the cost of similar projects, with a rough order of magnitude (ROOM). In order to test the predictive features with correlation, we need first to pick a sector, for this case we choose the power sector (see figure 7).

Is Size/Capacity in Megawatts is correlated with the Estimated Cost?

State the hypothesis:

- Null Hypotesis (H0): Size/Capacity is not correlated with Estimated Cost
- Alternative Hypotesis (H1): Size/Capacity is correlated with Estimated Cost

Since they are both continuous variables we can use a pearson correlation test and draw a scatter plot (see figure 8).

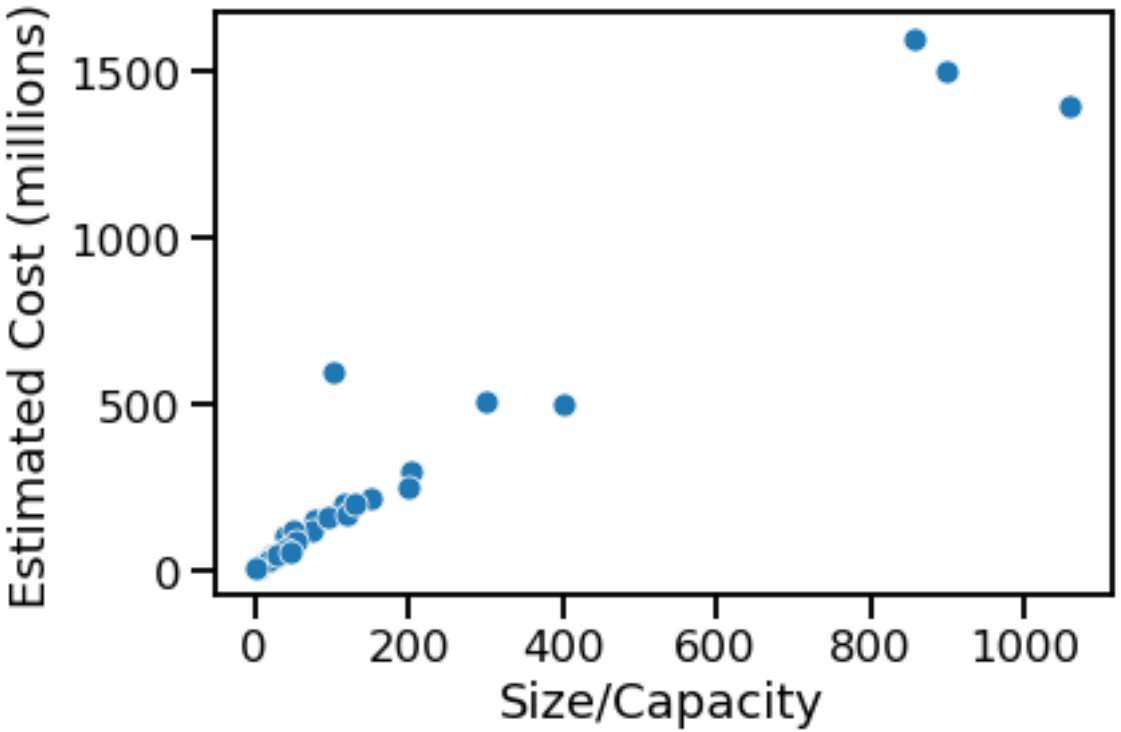
Hypothesis formulation

Figure 7. Power Sector Data

ProjectId	Project Name	Estimated Cost (millions)	Type	Start Latitude	Start Longitude	Size/Capacity	$1000/Capacity
642	Sundance 7 Gas-Fired Power Plant	1600.0	Gas Power	53.547501	-114.445000	856.0	1869.158879
644	Genesee Generating Station Units 4 and 5 Project	1400.0	Gas Power	53.547501	-114.445000	1060.0	1320.754717
649	Peace Butte Wind Farm	200.0	Wind	49.896000	-110.856003	120.0	1666.666667
653	Harvest Operations Gas Fired Power Plant	10.0	Gas Power	52.514198	-111.926003	5.6	1785.714286
2086	Vulcan Solar Project	155.0	Solar	50.093319	-112.848129	77.5	2000.000000

Hypothesis formulation

Figure 8.
Estimated Cost Vs. Size/Capacity Scatter Plot



Hypothesis formulation

A second hypotesis will be: **Does Estimate Cost differ by Project Type?**

An ANOVA test can be used for the Second hypotesis.

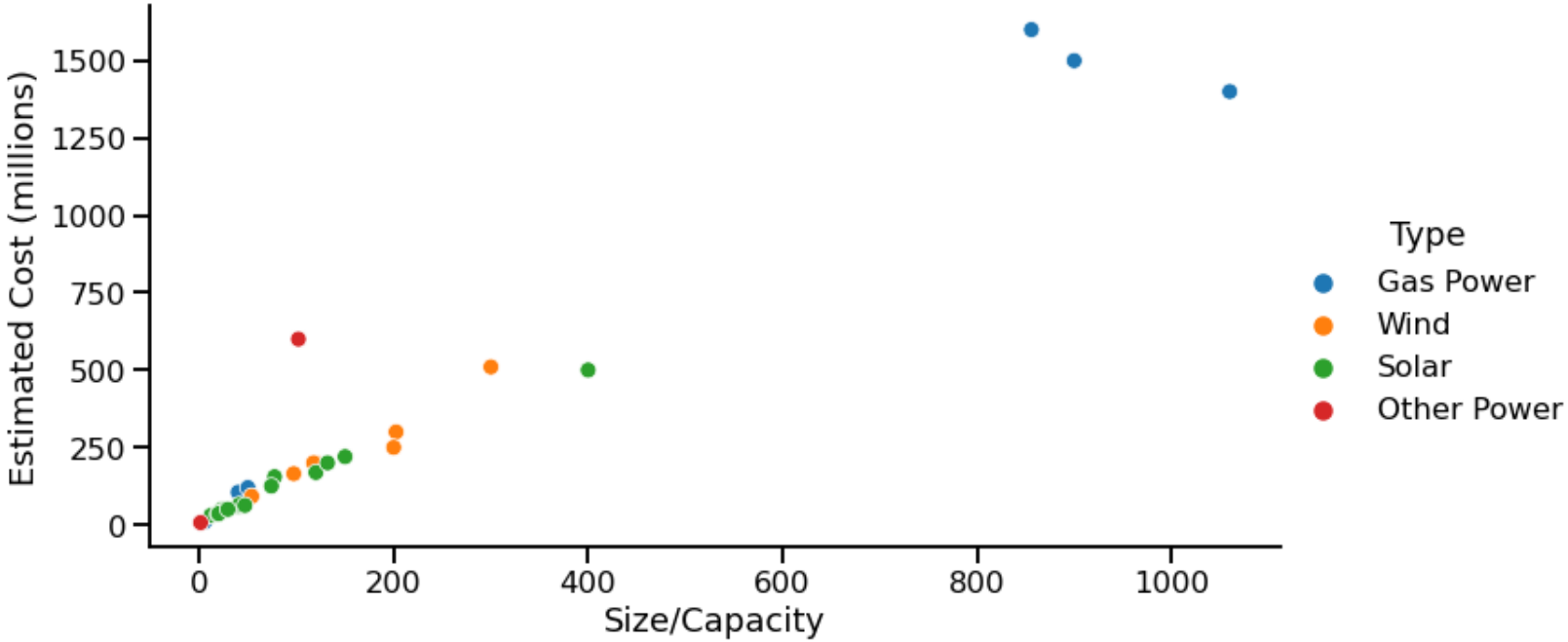
A third hypotesis will be: Is there is an association between Solar and Wind Project Types.

A T-Test can be used for the third hypotesis.

See figure 9 for the correlation by project type.

Hypothesis formulation

Figure 9. Scatter Plot by Project Type



Results

- Null Hypotesis (H0): Size/Capacity is not correlated with Estimated Cost
- Alternative Hypotesis (H1): Size/Capacity is correlated with Estimated Cost

Using the Pearson Correlation from SciPy we obtain:

Correlation Coeficient: `0.967614759429517`8

Two-tailed p-value: `6.958427885850458e-19`

Conclusion: Since the p-value < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between Size/Capacity and the Estimated Cost.

Conclusions

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

After exploring the dataset, we were able to do some feature engineering and filter the data set to extract valid information.

We performed significance testing to prove the correlation between the Capacity in Megawatts in the Power Sector and the Estimated Cost.

Most of the projects has not enough data for the engineered features of size and/or capacity. Further work is required to search for capacity data from other sources (company websites, industry associations, etc.).

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/