Edmonton Property Assessment and Crime Analysis

Edmonton property assessments are calculated on a set of metrics resulting in *municipal government* and *provincial education* property taxes. Do residential property owners in Edmonton have a transparent third party assessment model for their annual property tax notice? Do these property owners have the tools to assess all available data when assessing their property assessment? Do these property owners even trust the process that produces their property assessment?

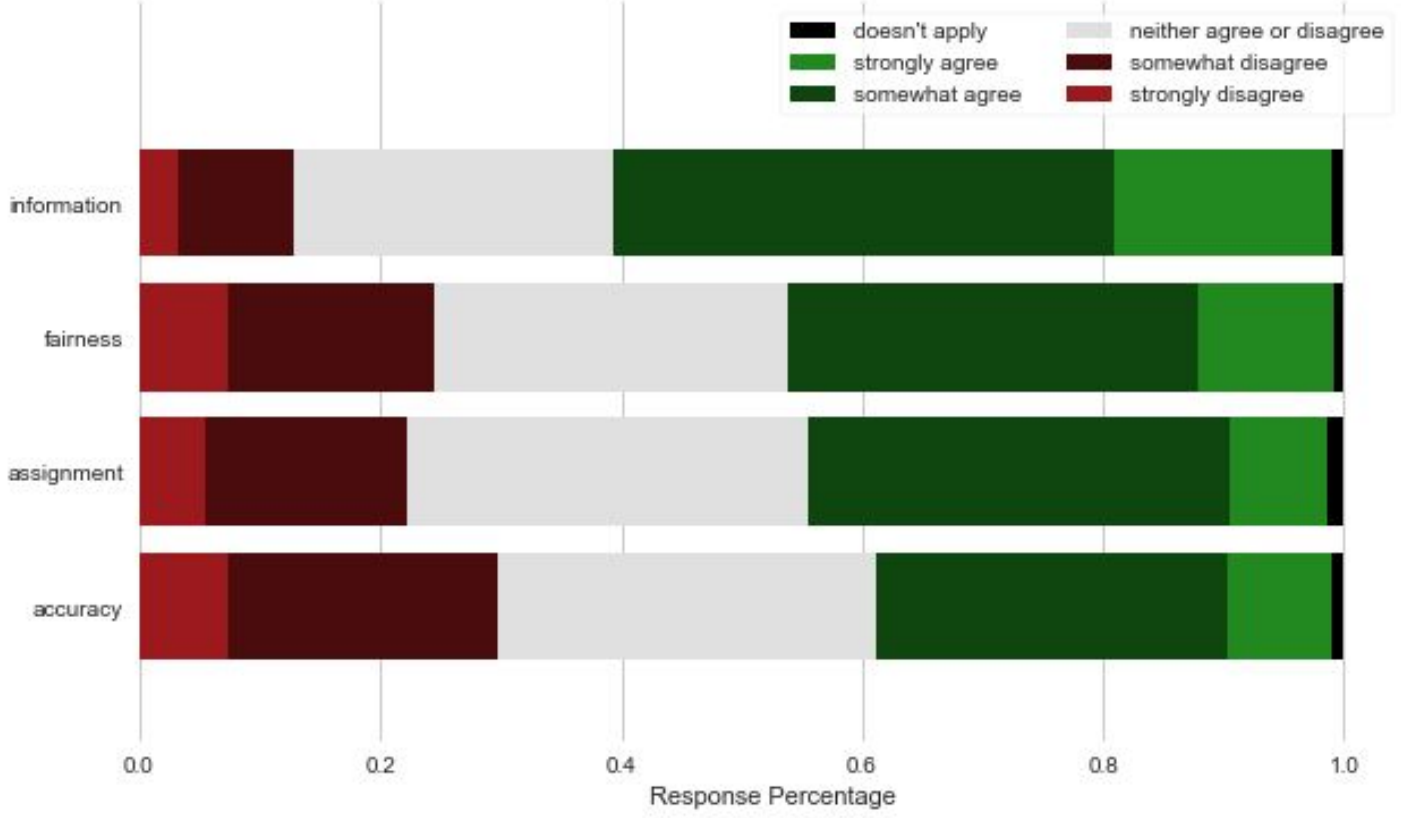
Edmonton residential property owners are the last of three levels of property assessment review. Namely, the first level is the City's internal checks and balances, then the Alberta government's annual assessment audit process and finally, **individual property owners' review of their notice** ("Purpose by Legislation" edmonton.ca, 2018-07-26).

These property owners should be confident that the city is using all possible data points and analytics to ensure that property values are being assessed with reliable metrics.

The following investigation will start by looking at the public's view on property assessments and a model will identify key housing and neighborhood factors that are considered for property assessments.

Statistical Analysis - Community Response Survey 2018

An in depth analysis of the 2018 community survey has found some compelling insights. When focusing on questions related to the community's view on property assessments, I found that nearly 50% of property owners may feel their property assessments are inaccurate. This is made even more interesting by the significantly high percentage of property owners who feel that the information provided to them by the City is sufficient.



```
def confidence(a):
    '''Takes survey question responses—pandas Series—
    and provides a 99% confidence interval.
    '''

    count = a.value_counts(sort=False)
    agree = np.sum(count[['5 Strongly agree', '4 Somewhat agree']])
    disagree = np.sum(count[['1 Strongly disagree', '2 Somewhat disagree']])
    sample_size = agree + disagree

    x = (1 * disagree + 0 * agree) / sample_size
    var = (disagree * (1 - x) ** 2 + agree * (1 - x) ** 2) / (sample_size - 1)
    dev = np.sqrt(var)
    sigma = dev / np.sqrt(sample_size)

    return print('99% chance that a random sample mean for Edmonton property owners \nthat disagree is
between:
                 \n{} and {}'.format(x - 2.58 * sigma, x + 2.58 * sigma))
confidence(accurate)
99% chance that a random sample mean for Edmonton property owners that disagree is between:
0.399 and 0.480
```

A 99% confidence interval finds that between 39.9 and 48.0% of property owners likely disagree with the accuracy of their property assessment.

Survey Analysis Notebook

Data Wrangling

The following datasets were pulled from Edmonton's Open Data Catalogue:

- Edmonton Property Assessments 2018
- Edmonton Property Assessments 2012-17
- Edmonton Property Information 2018
- Population by Citizenship Census Data 2016
- Edmonton Police Services Neighbourhood Criminal Incidents Report 2018
- Property Assessment Customer Service Survey 2018

The cleaning process started with analyzing the columns for relevancy in the investigation. Both property assessment datasets were filtered for missing values and only one of nearly 400,000 records was missing significant information. This property was removed from the dataset because it was labeled an outlier.

The next step was to begin merging the datasets. The 2018 property assessments and the 2018 property information datasets were merged on the City's unique account identifier. This newly merged dataset was analyzed more closely and properties that are not residential were removed. Filtering for residential properties was not as simple as filtering on the class variable. Some significantly low assessments are attributed to walkways, parks, and small community building which I did not want to include. Also, zoning was found to be a secondary variable for filtering out residential property records. Only zones with an "R" are zoned for residential use. Furthermore, the remaining outliers to residential property assessments were filtered by setting a lower bound of $20,000 and an upper bound of $10,000,000 Canadian Dollars on the property assessment value. Finally, records without street data were removed.

A very similar technique was applied to the property assessment dataset from 2012-2017 and this data was used to calculate median assessment value each year for every neighbourhood.

The criminal incidents report was merged with the population dataset to engineer a crime per capita variable. A significant amount of pivoting and reshaping was undertook to align this crime/population data with the assessment/information dataset. This crime/population dataset was then merged to the assessment/information dataset and prepared for machine learning through dummy variable encoding.

The property assessment survey data was analyzed to generate the visualization and the statistics at the beginning of this report. Responses indicating that property owners had contacted the City regarding their assessments were filtered out. These responses were further filtered into negative, neutral, and positive response bins. Statistical analysis and visualization was performed on the about 400 records with full questionnaire data.

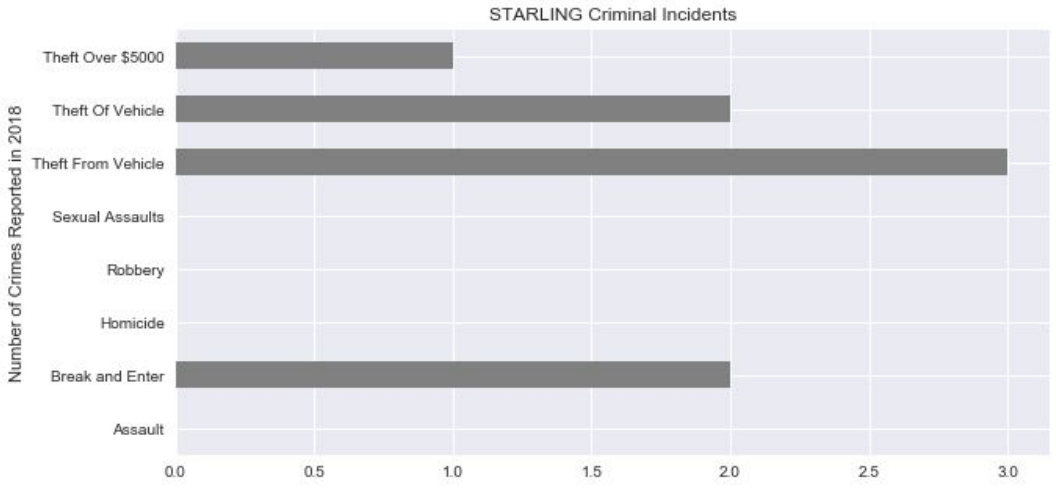
GitHub Data Wrangling Notebook

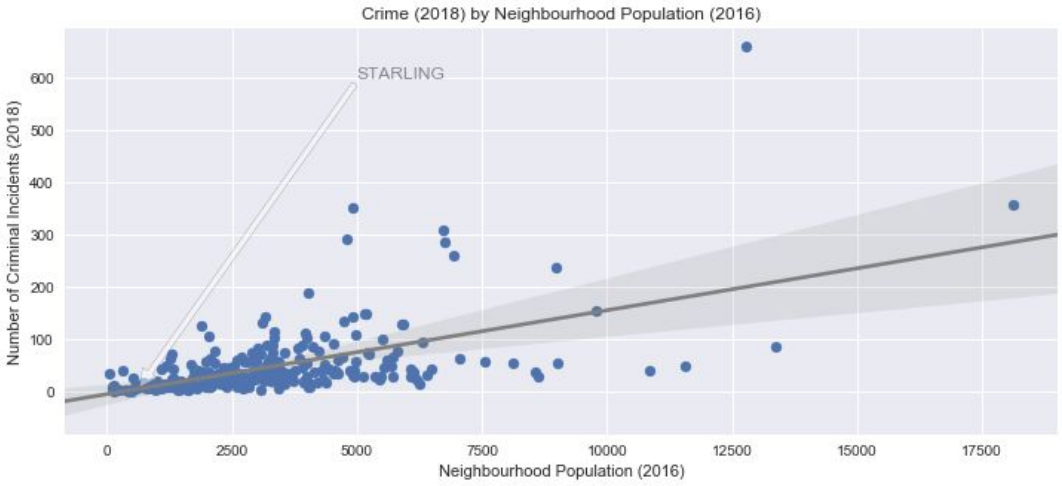
Functional Analytics Tool

These five datasets are cleaned, merged, and/or pivoted to generate the data used in a function my_neighbourhood() that takes one input and has the following description (Python docstring):

```
This function takes a string input for the Edmonton neighbourhood and
returns the criminal incidents, median property value change, and a label
of where the neighbourhood resides in a crime per capita scatterplot.
```

Example outputs:

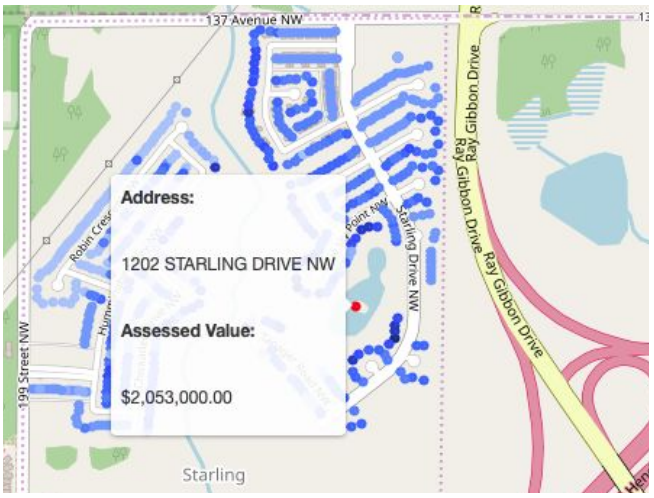




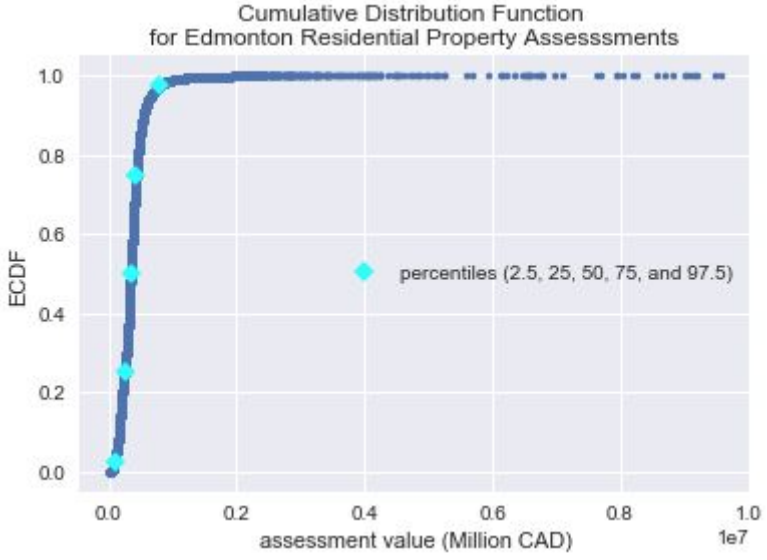
Mapping Analytical Tools

The cleaned dataset is imported into a new notebook and formatted for the folium module. Latitude and Longitude data are isolated for plotting the points on the map (presented by leaflet). Assessment data is functionally formatted into HTML for the hover tooltip information. The final solution presents a *geographical information system* for Edmonton property assessments.

[Mapping Analysis Notebook]

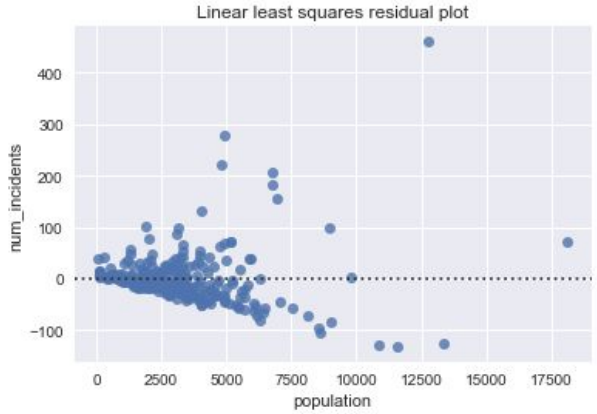


Statistical Analysis - Edmonton Property Assessments and Crime 2018



The first step in exploratory data analysis (EDA) for this dataset was an Empirical Cumulative Distribution Function (ECDF). Through this visualization it is easy to ascertain the true spread and distribution of the data, while avoiding binning bias in a histogram.

From the ECDF we see that >97.5% of the residential property assessments in Edmonton fall below the 1 Million dollar evaluation amount. We can also see that the median assessment for residential properties in Edmonton falls below 500,000 Canadian Dollars. With the true median residential assessment calculated at 346,500 CAD.



The second step in EDA was looking for a correlation between crime in populated neighbourhoods. From an ordinary least squares linear regression we can see that the variance is increasing with population. A residual plot more clearly illustrates the variance as seen here to the left. Moving the model to a second order polynomial did not generate less variance in the residuals plot but only served to move the data point for the Oliver neighbourhood closer to the model.

A plausible hypothesis: Having a garage increases your assessment value with the city. Public data regarding garage is provided for every property by a binary designation (garage or no garage).

In [1]:

```
df_garage = df_clean[df_clean['garage'] == True]
df_no_garage = df_clean[~df_clean['garage'] == True]
print('with garage mean assessment:', np.mean(df_garage.value), '\nno garage mean assessment:',
np.mean(df_no_garage.value))
print('with garage assessment standard deviation:', np.std(df_garage.value), '\nno garage assessment
standard deviation:', np.std(df_no_garage.value))
```

Out[1]:

```
with garage mean assessment: $43,0575.39
no garage mean assessment: $207,761.48
with garage assessment standard deviation: $197,672.30
no garage assessment standard deviation: $177,406.09
```

In [2]:

```
std_mean_diff = np.sqrt(np.std(df_garage.value) ** 2 / len(df_garage.value) +
np.std(df_no_garage.value) ** 2 / len(df_no_garage.value))
std_mean_diff
```

Out[2]:

736.44

Hypothesis Test

H_o: Having a garage does not affect your property assessment value (μ garage - μ no garage = 0)

H_a: Having a garage increases your property assessment value (μ garage - μ no garage > 0) α = 0.01

In [3]:

```
critical_z = 2.33
critical_value = std_mean_diff * critical_z
critical_value
```

Out[3]:

1715.90

If we assume that the null hypothesis is true, there is only a 1% chance that we see a mean difference (between assessments with garages and those without) greater than 1,715.90 CAD. Our observed difference is 222,813.91 CAD and we therefore reject the null hypothesis for the alternative. The mean assessed value of properties with garages is significantly higher than those without garages.

Confidence Interval

In [4]:

```
observed_mean_diff = np.mean(df_garage.value) - np.mean(df_no_garage.value)
observed_mean_diff
```

Out[4]:

$222,813.91

In [5]:

```
conf_int = std_mean_diff * 2.58
```

In [6]:

```
print('We are 99% confident that the true difference of means lies between:', observed_mean_diff -
conf_int, 'and', observed_mean_diff + conf_int)
```

Out[6]:

We are 99% confident that the true difference of means lies between: $**220,913.90 and** $**224,713.91**

Machine Learning

The true power of a predictive model—trained on Edmonton residential property assessment values—would be realized by property owners who are able to access a tool designed to give informative context to their assessment notice. With the availability of additional features (discussed later) a model may be soon available to the public to compare their assessment notice to an output trained on third party data.

Ideally, this solution would allow property owners to input a shortlist of features related to their property —or even look up their address and their neighbours through a web API— and the model would provide them with a benchmark assessment value or a confidence interval of a realistic assessment range. This property owner could then take this contextual knowledge and compare it to the information provided in their assessment notice from the City of Edmonton. This process would empower the public to carry out their right to contest their assessment in an informed and systematic manner.

The following attempt looks at the currently available features and their ability to predict the assessment value of a new property. It should be noted that the earlier survey analysis finds that most property owners find the process inaccurate, but not necessarily the variables used to inform the assessments.

Technical Considerations

Currently the estimator which performs best is the ExtraTreesRegressor from Scikit-Learn. The optimized model operates with the following parameters:

```
(bootstrap=True, criterion='mse', n_estimators=121, max_depth=None, max_features='sqrt',
max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0, n_jobs=1, oob_score=False, random_state=0,
verbose=False, warm_start=False)
```

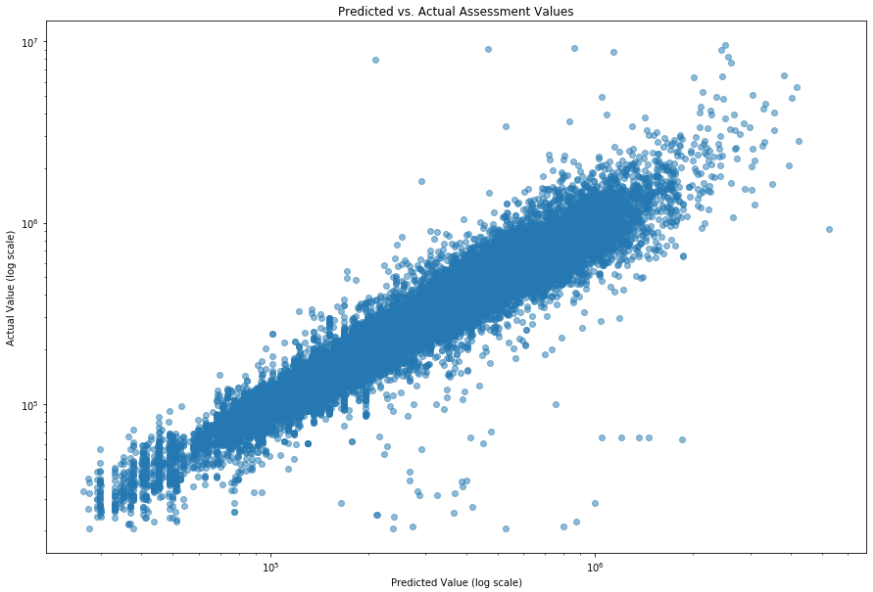
An imputer is used to replace all missing values with the columns mean value, and a scaler is used to standardize the mean and variance of the features.

In order to properly engineer the features and make them available to a training model. Dummy variables were generated for the neighbourhood, garage, and zoning features. This results in 300+ features for training but drastically improved the scoring when using the model to predict on the testing set. (The values for these tests should be included for support of the findings).

The ExtraTrees estimator outputs a mean squared error of 10,044,258,610.27 and therefore a mean error of $100,221.05 which represents a significantly high variance in the ability of the model to predict assessment values. These may not be the strongest features for informing individual property assessments

When similar techniques are applied to the Boston Housing Data—provided by Scikit-Learn—we find that that the model performs significantly better with only 1 feature. Namely, the number of rooms in the dwelling. Unfortunately, Edmonton's Real Estate market does not open their data to the public and therefore we do not have the same feature selection that is available in the Boston, Massachusetts area. With some Canadian cities beginning to soften their policies surrounding Real Estate data, the future of a more precise model for Edmonton may be close at hand.

Visualizations



Analysis

The visualizations show that the ExtraTreesRegressor estimator is a more precise algorithm. The main sources of variance are properties that have actual assessments values below $100,000 or above $1,000,000. This is probably due to extra features such as amenities (swimming pools, terraces, extended season sunrooms) and building permits (basement developments, decks, excessive landscaping) in regards to the actual assessments above $1,000,000. These features are not provided in the public dataset provided by the City of Edmonton.

With a few more targeted features (number of rooms, number of building permits) the model can potentially meet the needs of this project.

Machine Learning Notebook

Presentation