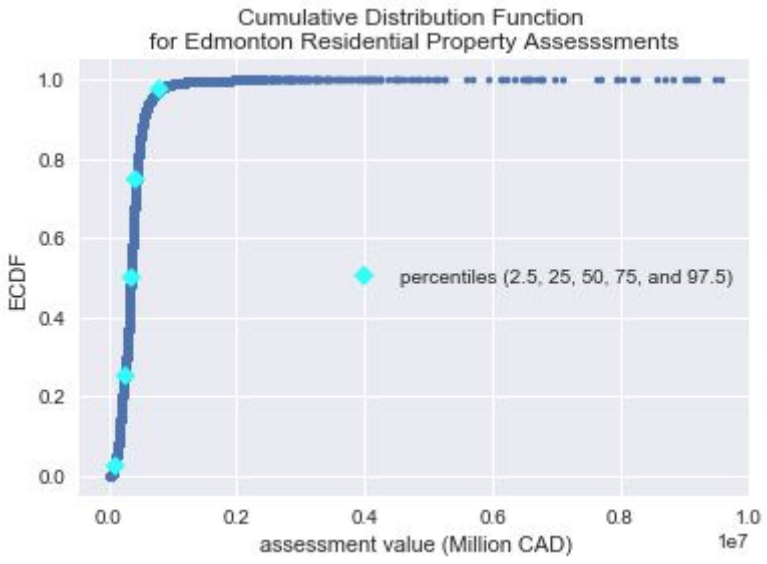
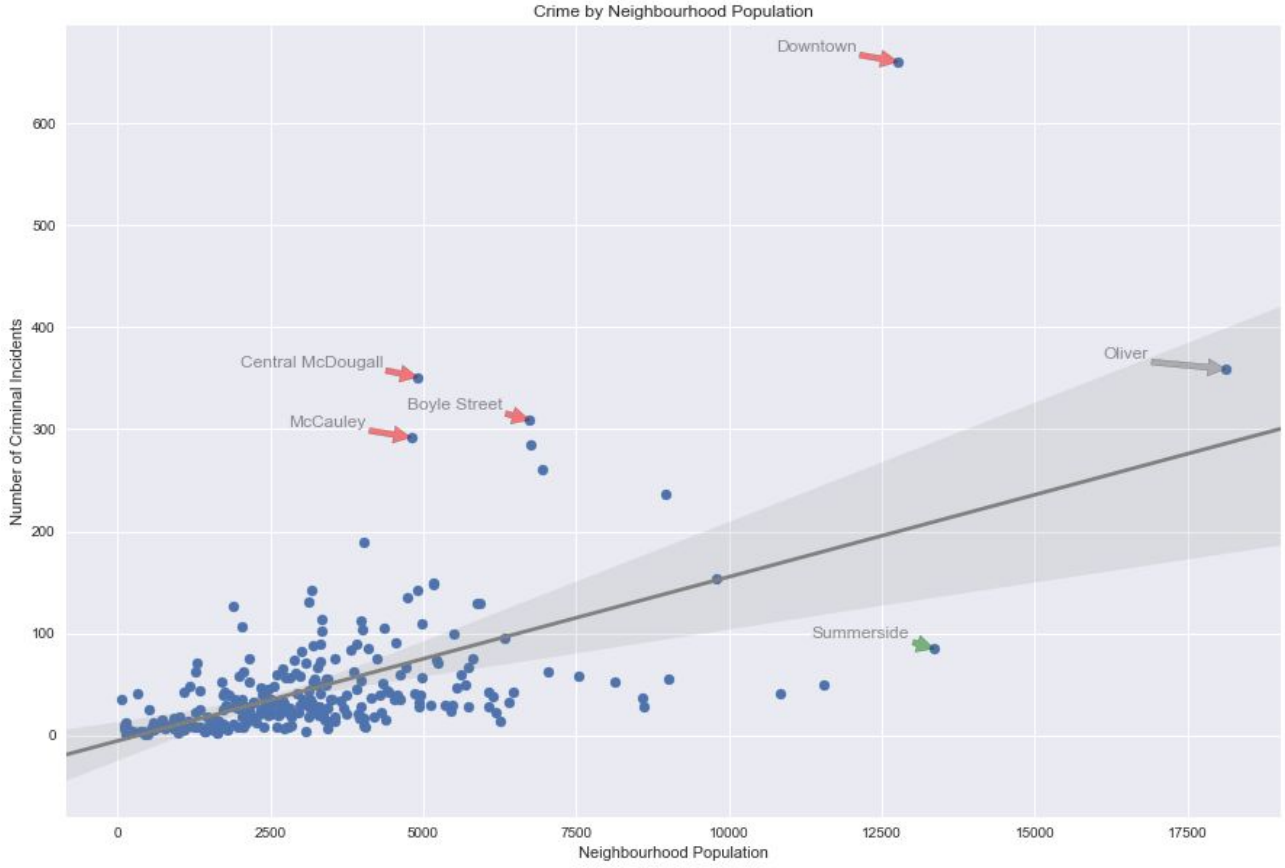
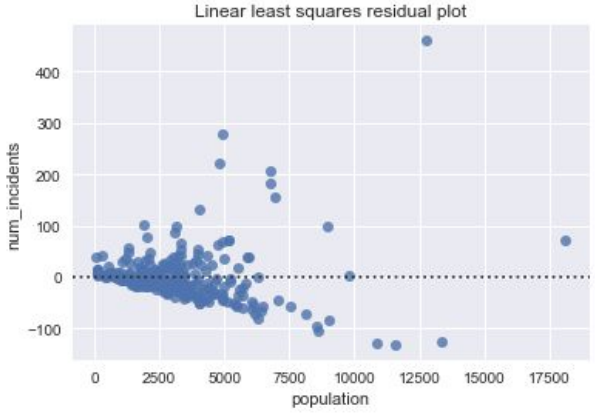
Statistical Analysis - Edmonton Property Assessments and Crime 2018



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: 430575.3893830812
no garage mean assessment: 207761.48337595907
with garage assessment standard deviation: 197672.30372608424
no garage assessment standard deviation: 177406.09280880945
```

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]:

```
736.4367936160216
```

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]:

```
1715.8977291253304
```

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 are 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]:

```
222813.91
```

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: **220913.90 and 224713.91**