Peppy Poké Pipper Pals

COMP 333 - Github

FINAL PROJECT

1.

Topic & Defined ML Tasks

Topic

Goal: Housing Price Predictor

- Predict house prices in Montreal:
  - Based on Bedrooms, location, bath, sqft, etc...
- Apply the following tools and disciplines for Data Analytics:
  - Data Collection
  - Data Integration
  - Data Cleaning
  - Data Transformation
  - Data Visualization
  - Machine learning outcomes and Modeling

Machine Learning

- Create machine learning model to predict new unsold house prices based on sold properties.
- Datasets to evaluate:
  - Model without data cleaning (DC) and transformation (DT).
  - Model with applied Data Analytic techniques.
- Performance measures
  - RMSE, MAE, R^2

2.

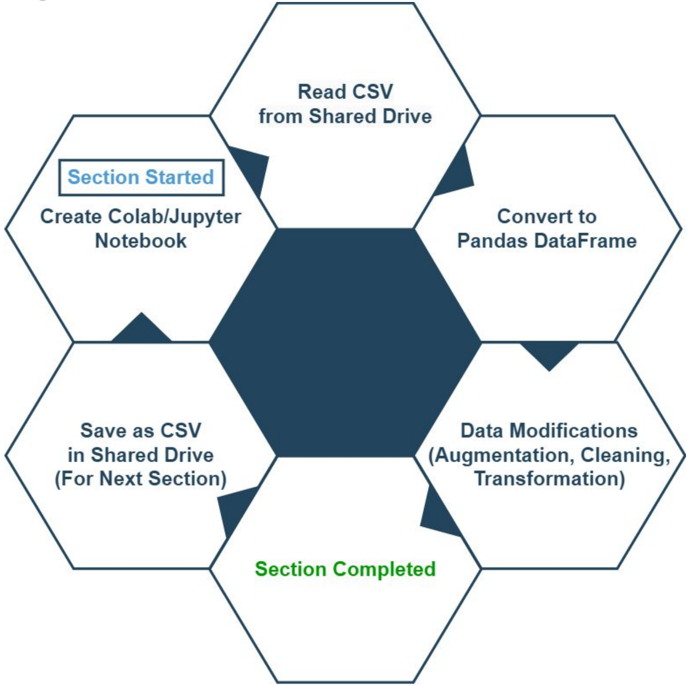
Data Collection

Datasets

1. Residential property prices from duProprio scraped in March 2022 from:
   - www.kaggle.com/montreal-property-price/
     Contains price, region, address, bedrooms, bathrooms, living area, lot dimension
2. Canada Revenue Agency - income tax filed per postal code region from 2021 Tax year:
   - www.canada.ca/individual-tax-statistics-fsa/
     FSA (forward sortation Area), total tax reports filed, total income, net income, taxable income
3. Statistics Canada (from 2021 census) population count and number of private dwellings per postal code region:
   - www12.statcan.gc.ca/2021/population/
     Postal code, population, total private dwellings

Storage System

- Datasets are stored as Excel and csv files on Google Drive between data cleaning and merging phases.
- Workflow:



3.

Data Integration

Schema Integration

Add income, population count and number of private dwellings as features to the property dataset.

Property dataset
Street address and borough without postal code

Income Data & Census Data
First three characters of the Postal Code
Forward Sortation Area (FSA)

region	address
Mercier / Hochelaga / Maisonneuve	5185-5187 RUE DESMARTEAU
Villeray / St-Michel / Parc-Extension	417-8635 RUE LAJEUNESSE
Villeray / St-Michel / Parc-Extension	8517-A-8515-8517 avenue de l'esplanade

FSA	Total Income
G0A	3.444018e+09
G0C	1.618508e+09
G0E	1.489380e+08

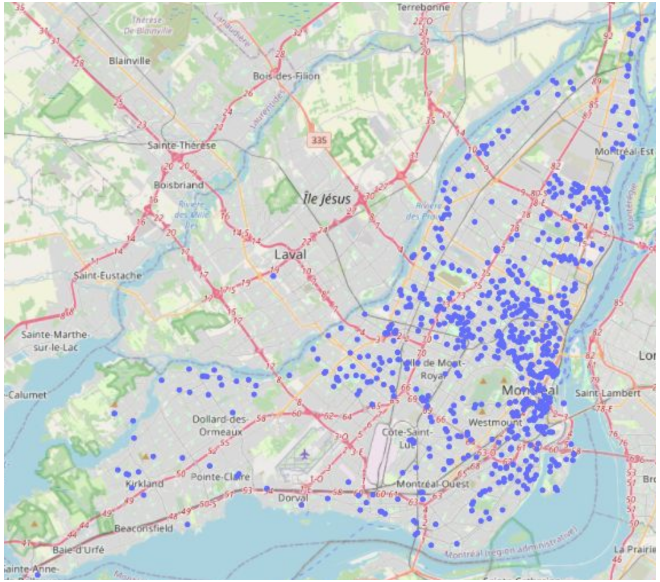
Geographic code	Population	Total pvt dwlgs
H0M	1202	763
H1A	32516	14287
H1B	20160	9400

Data Mapping and ER

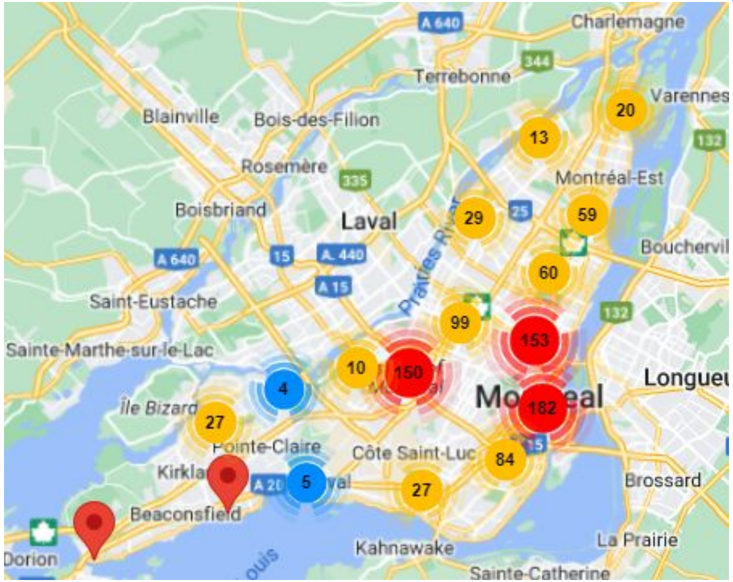
Idea: Regional datasets are based on FSA.
Goal: Get first three characters of postal code for housings.

1. Retrieve longitude and latitude for each address.
    - Geocoders by Awesome Table.
        - uses Google maps to set coordinates using street address.
2. Get postal code from the coordinates.
    - Nominatim from Geopy.Geocoders
        - Reverse geocoding to get only the postcode using Longitude and Latitude.
3. String operations to split the postal code to get the first three characters.
    - In python notebook.

Distribution of Listings



General View



Aggregated View

4.

Data Cleaning

Conversion of Data Types

1. Numerical data extraction - all entries were strings

bedrooms	bathrooms	living area
Bedrooms\n 3	Bathrooms + Half baths\n 1	Lot dimensions\n2,176 ft²
Bedrooms\n 2	Bathrooms + Half baths\n 1	Living space area (basement exclu.)\n957 ft²
Bedrooms\n 8	Bathrooms + Half baths\n 3	Lot dimensions\n2,850 ft²

	price
82	$545,000\n\n\n\n\n\n\n\n\n ...
192	$399,000\n\n\n\n\n\n\n\n\n ...
280	$500,000\n\n\n\n\n\n\n\n\n ...

2. Some entries were shifted to the left. (In the wrong column)

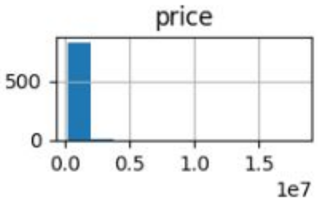
Bedrooms 3	Bathrooms + Half baths 1 +1	Lot dimensions2,259 ft²
Bedrooms 4	Bathrooms + Half baths 2 +1	Living space area (basem
Bathrooms + Half baths	Living space area (basement exclu.)357 f	NaN

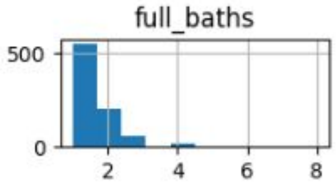
Cleaning of Null Values

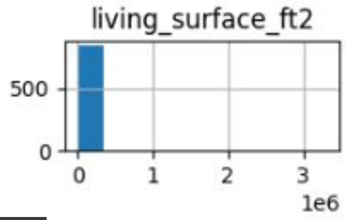
- Removal
    - Entire null rows and entire null columns were removed from all three sets:

	price	region	address	bedrooms	bathrooms	living area	lot_dimension
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	NaN	NaN	NaN	NaN	NaN	NaN	NaN
27	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- Imputation
    - The median is better than the mean as a replacement of null values since the null columns are heavily skewed. (Outliers!)



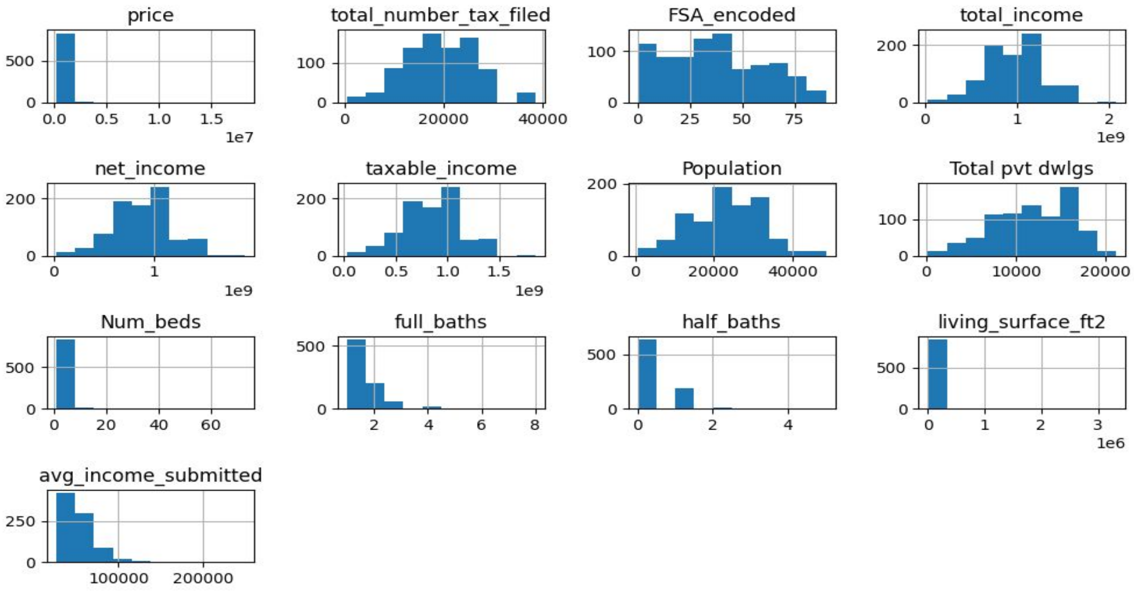




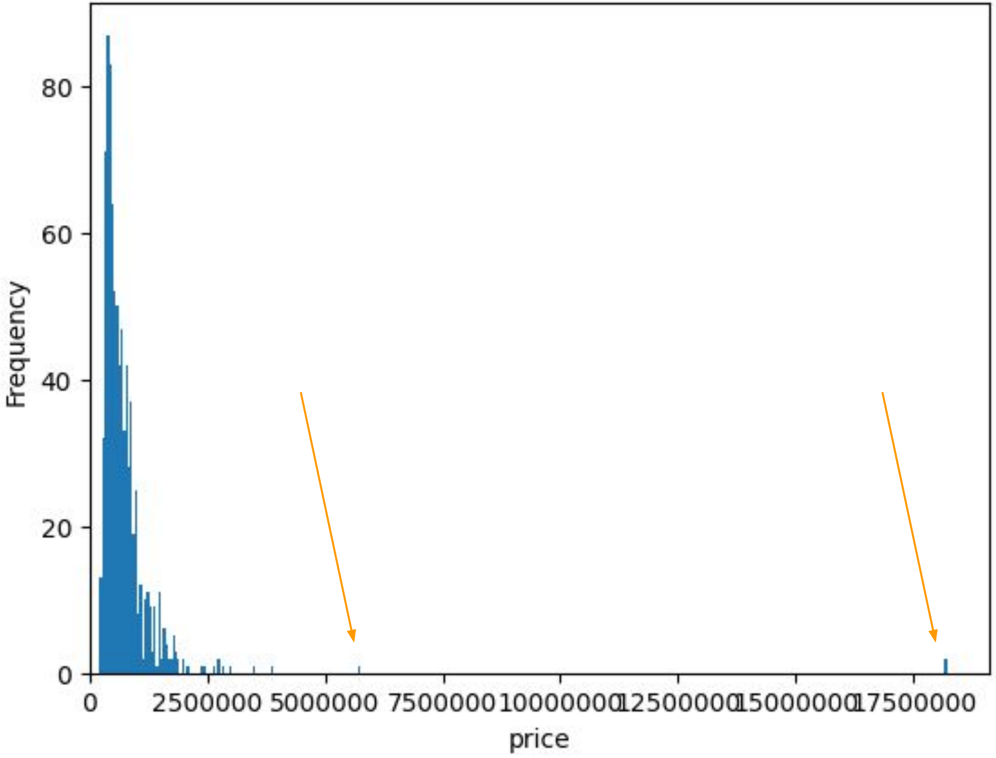

```
Median price: 580000.0
Median full_baths: 1.0
Median living_surface_ft2: 1060.5
```

Cleaning of Outliers

- Create histograms to check for outliers.
- Some features are heavily skewed, indicating possible outliers



Property Price **Outliers**



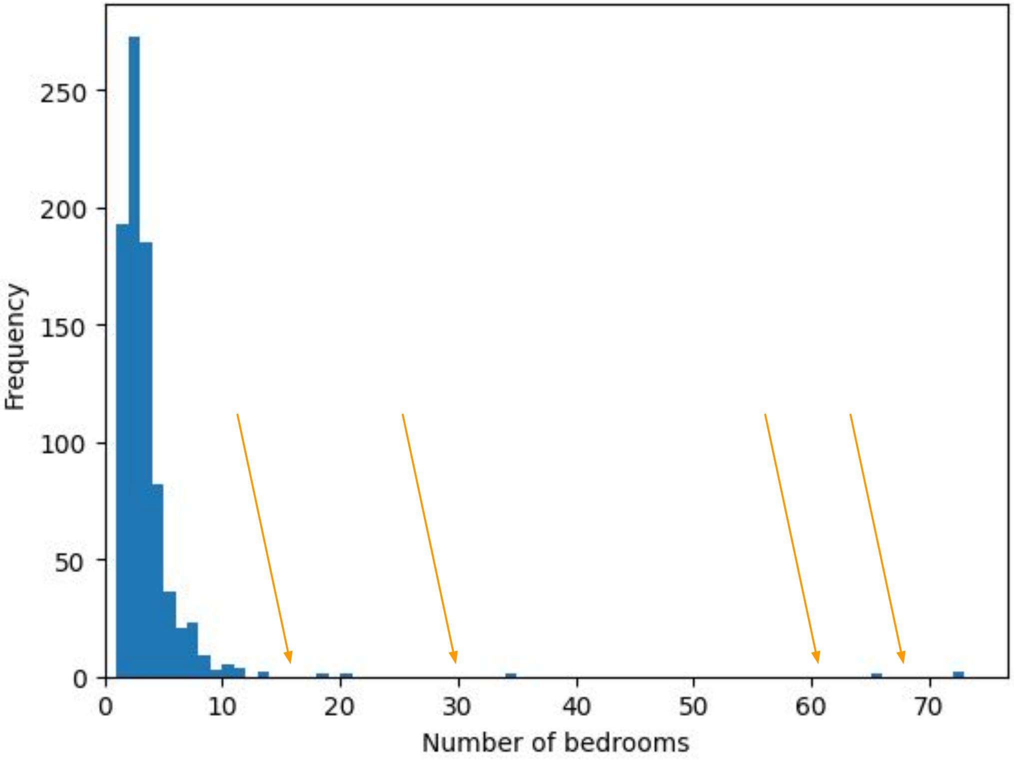
Two outliers!

- Between 5-7.5 million
- Above 17.5 million

Solution:

- Remove any house over $1.5 million and keeping square feet between 300 and 4000.

Number of Bedrooms Outliers

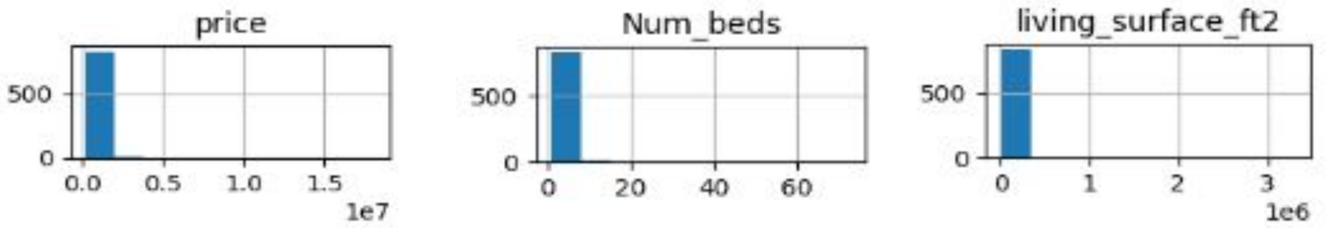


<u>Four</u> **outliers:**

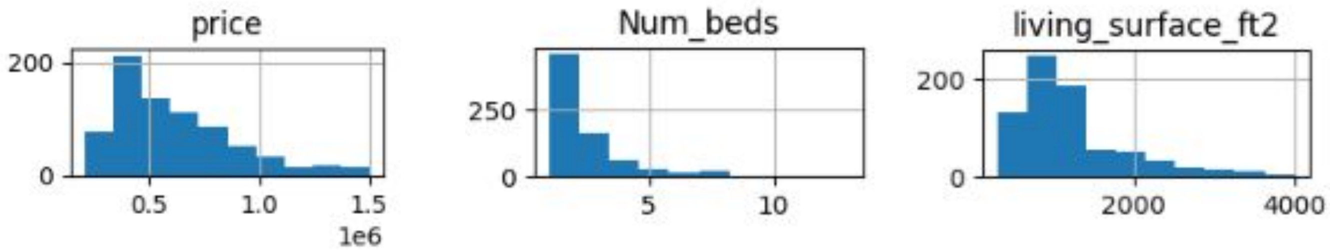
- Remove all houses with over 15 bedrooms
- Absurd properties will highly skew the data

Result of **Outlier Cleaning**

- Before cleaning:



- After cleaning:



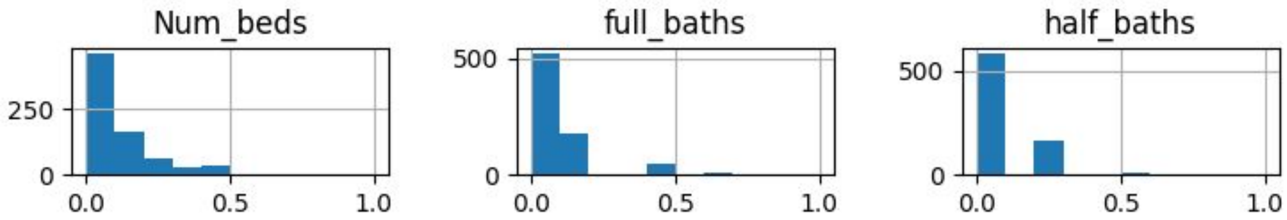
5.

Data Transformation

Data Scaling (MinMaxScaler)

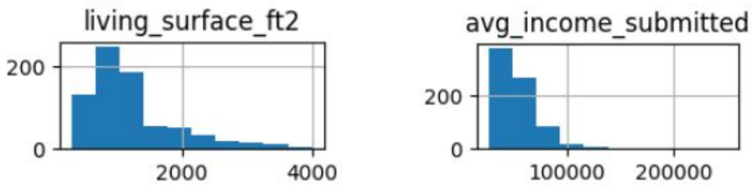
- Had no impact on correlation of features to price.
- Scale features with low numbers between 0 and 1:

- Result: (Only changing range)

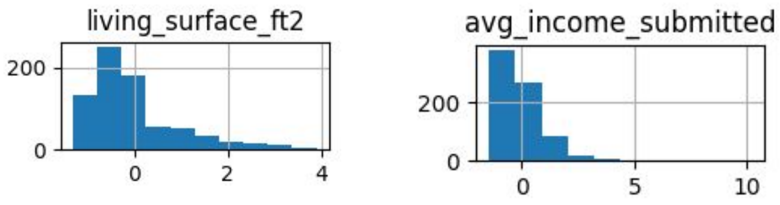


Data Normalization (StandardScaler)

- Had no impact on correlation of features to price.
- Fit high-numbered data into standard distributions:



- Result: (zero mean and standard deviation of 1)



Data Encoding

- Changed FSA (part of Postal Code) from categorical to numerical:

```
#changing FSA from categorical to numerical
encoder = LabelEncoder()
FSA_cat = properties['FSA']
FSA_cat_encoded = encoder.fit_transform(FSA_cat)
```

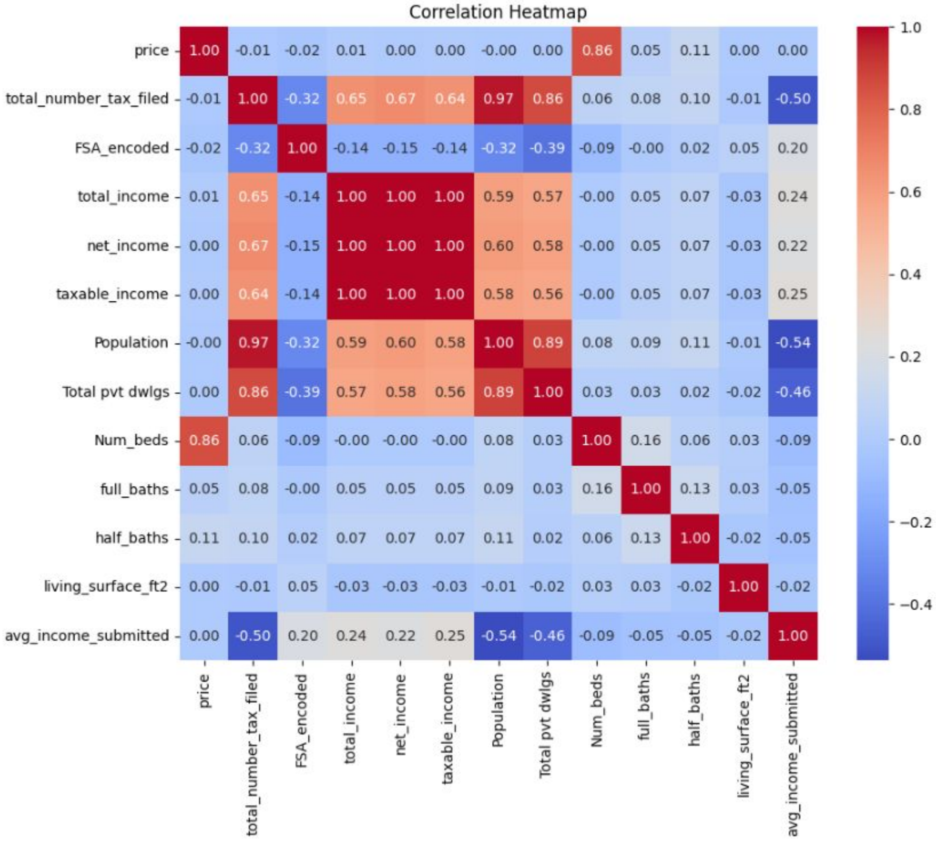
6.

Data Visualization and ETL

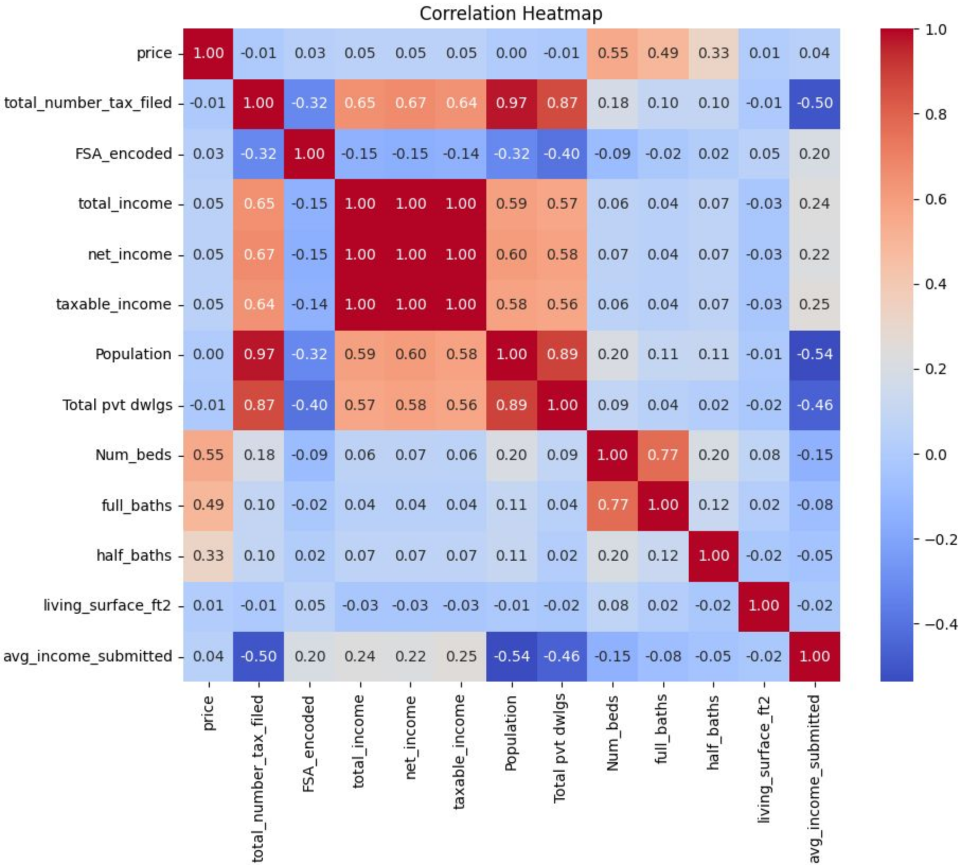
Tools

- Jupyter Notebook
- Google Colab
- Sklearn
- Geopy geocoders
- Pandas
- Draw.io
- Numpy
- Alteryx
- Tableau
- Seaborn
- Plotly

Correlation Before Cleaning

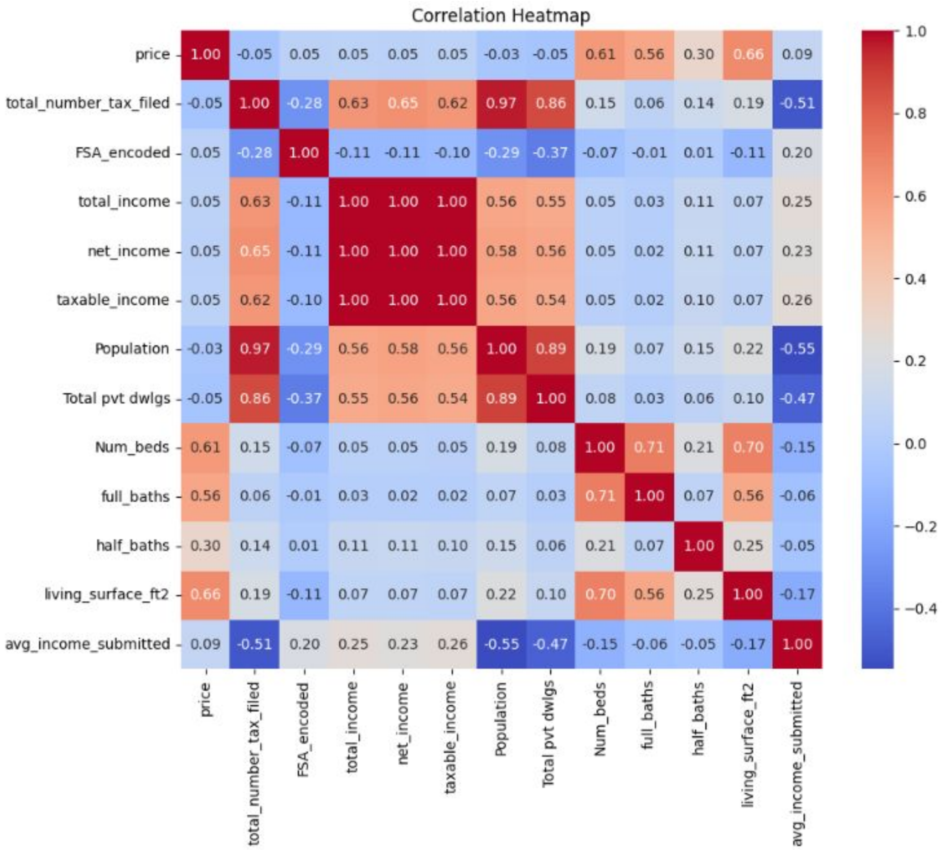


Correlation After Cleaning



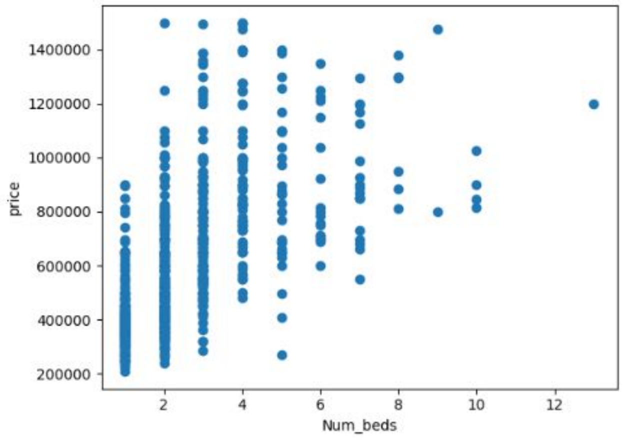
- Number of bedrooms has the strongest correlation with price (0.55)
- Number of bedrooms is a strong predictor

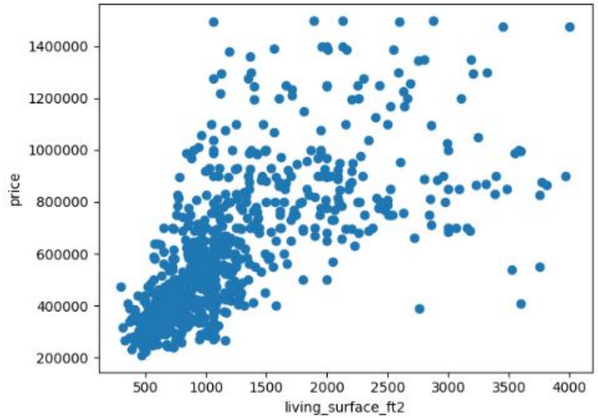
Correlation After More Cleaning



- Living space now has a very strong correlation with price (0.66)

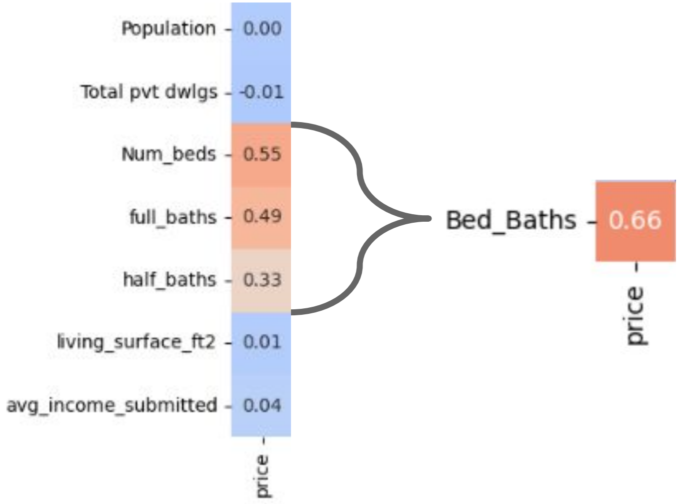
Scatter plot: price vs features



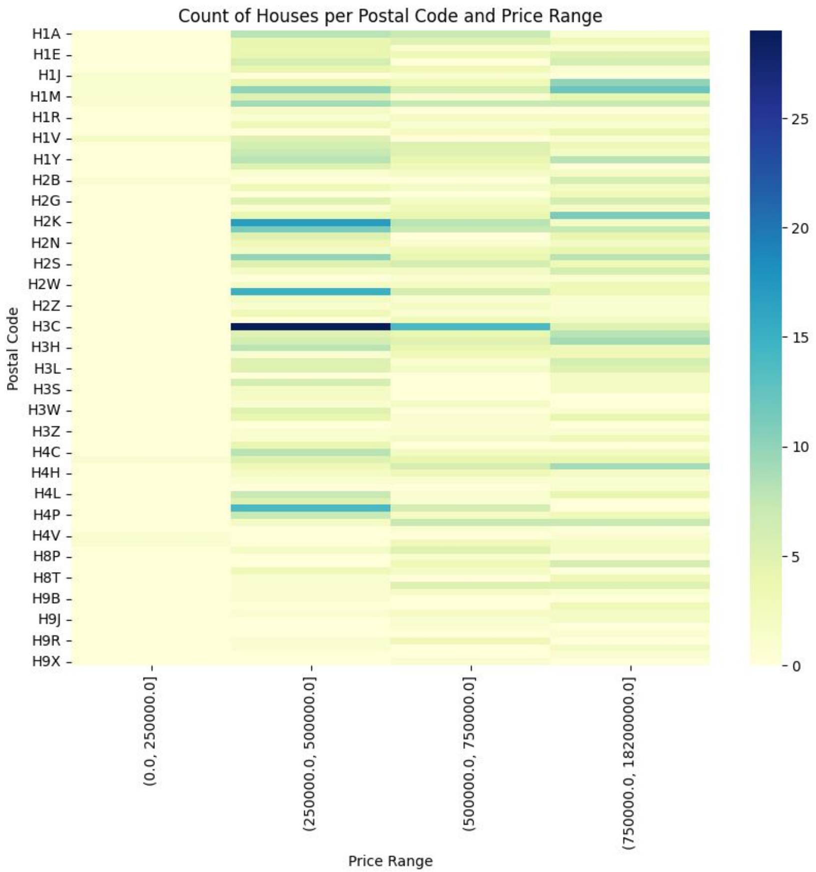


ETL to Find Features

- Create Heatmaps to isolate strongest correlations
- Create new feature to obtain higher correlation:



Heatmap: price vs postal code



Price range and postal code

- It can be seen that Anjou (H1M) has the most expensive houses

Feature Engineering

- Create new feature bed_baths

```
properties_noOutliers['Bed_Baths'] =
        properties_noOutliers['Num_beds'] +
        properties_noOutliers['full_baths'] +
        properties_noOutliers['half_baths']
```

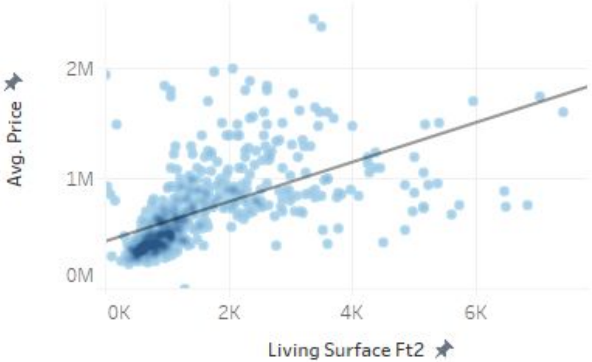
- Recreate heatmap
  - Bed_baths is now feature with higher correlation to price (0.66)
  - 1.1 greater than number of beds

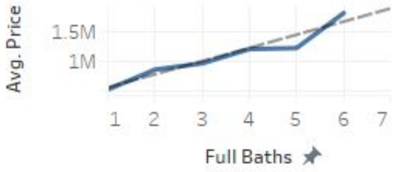
Dashboards

1. Initial impressions on the properties dataset at a glance
    a. Detailed view of selected charts

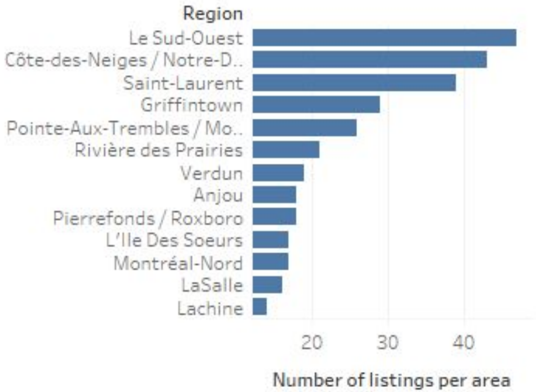
2. General view of demographics based on regions and FSA

1. Initial impressions on the properties dataset at a glance





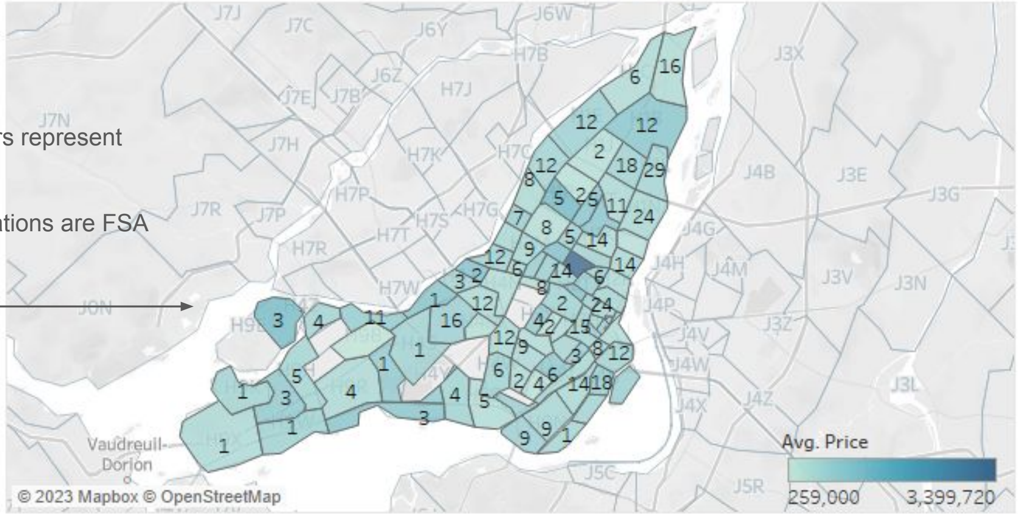
Regions with most listings



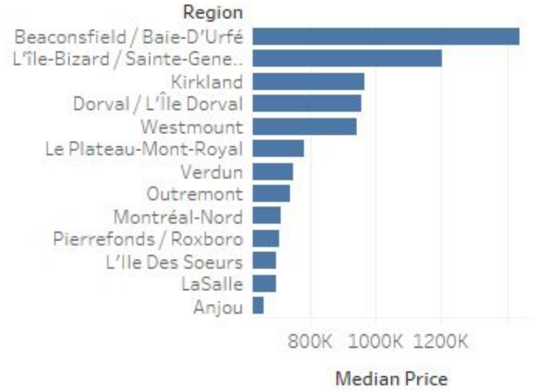
Properties for sale at a glance with average price per region

Numbers represent listings

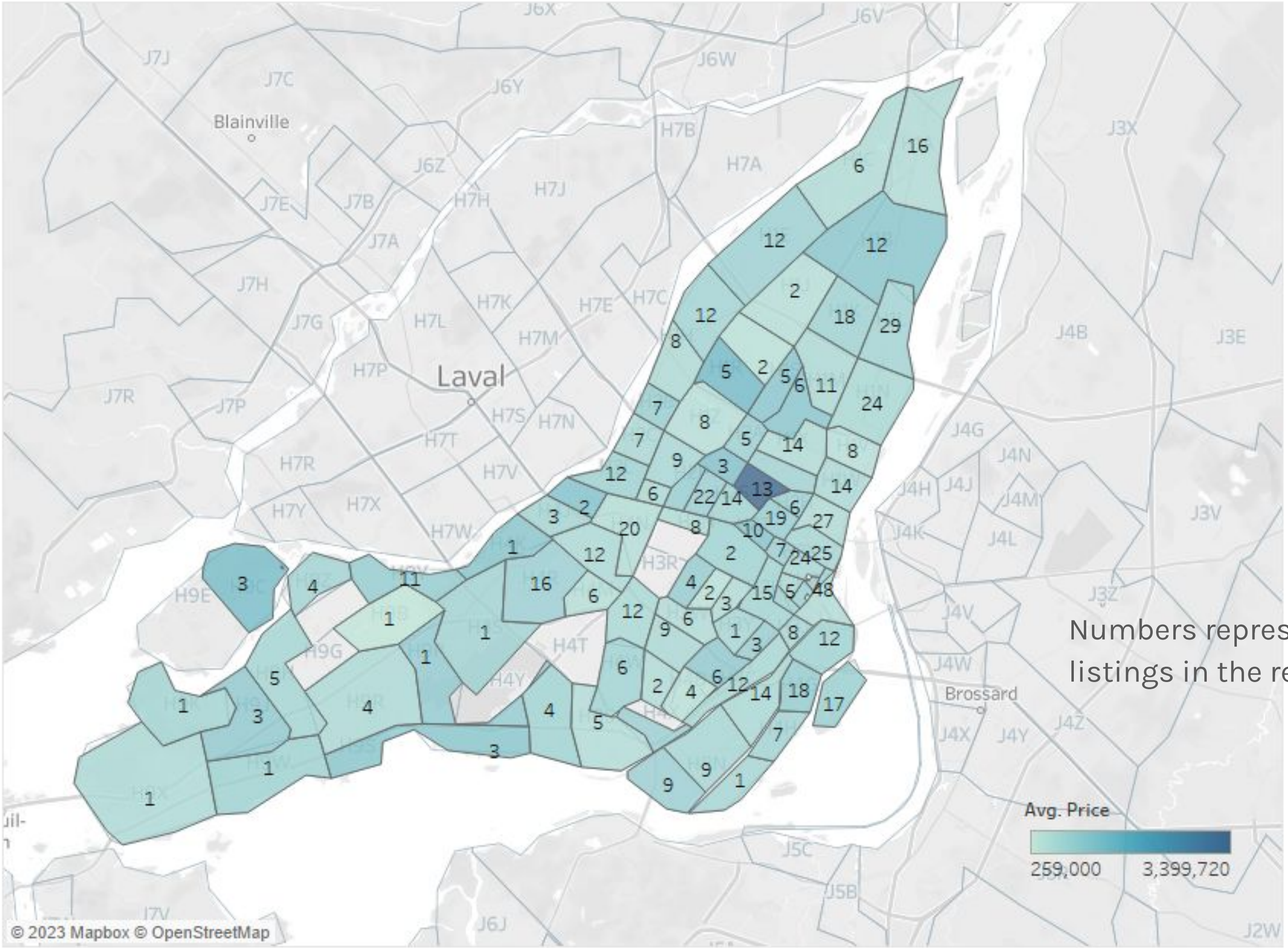
Delimitations are FSA



Most expensive regions



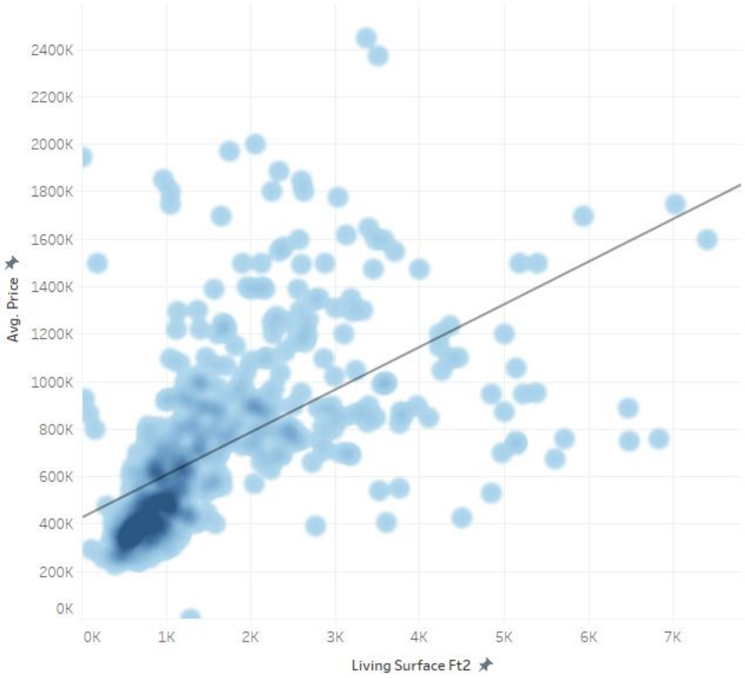
A more granulated distribution with average price per FSA



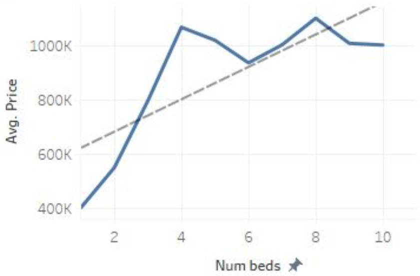
Numbers represent the total listings in the region

Features that seem to be correlated with the price

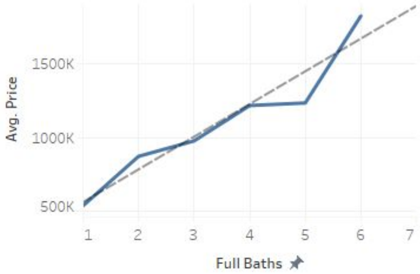
Average price vs living surface



Number of bedrooms vs Average price

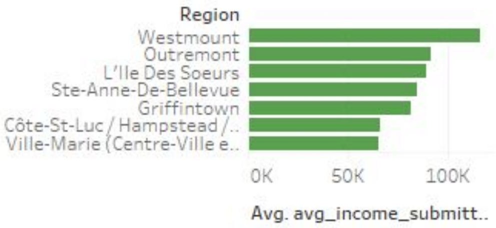


Number of bathrooms vs Average Prices

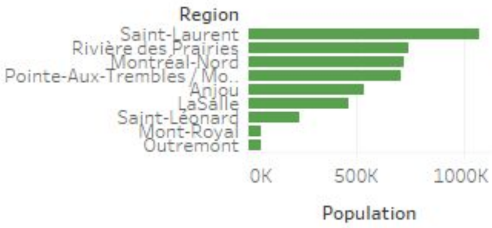


2. General view of demographics based on regions and FSA

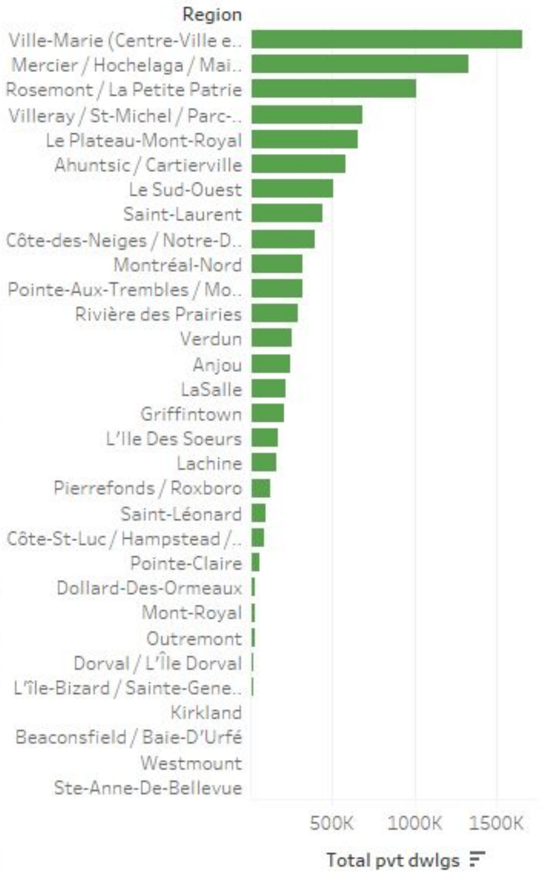
Regions of highest income



Top most populated regions



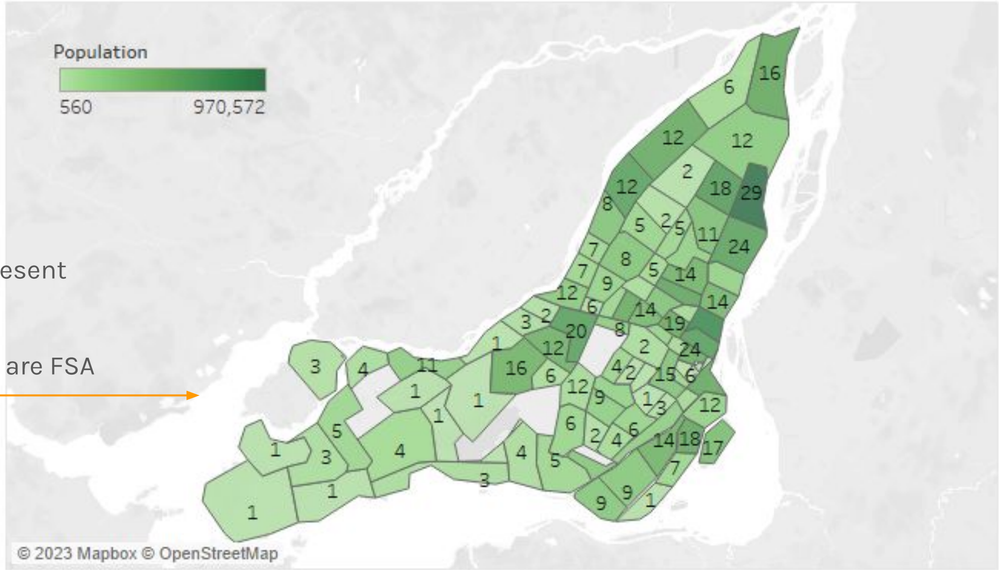
Average number of households



Number of listings and population count per FSA

Numbers represent listings

Delimitations are FSA



7.

Machine Learning Outcomes

Performance Metrics

- Models
  - Random Forest and Linear Regression
  - Linear Regression performed better than Random Forest Model
- Metrics
  - RMSE (Root Mean Square Error)
    - Aim for lowest value possible
  - MAE (Mean Absolute Error)
    - Aim for lowest value possible
  - R^2 (Correlation Coefficient)
    - Aim for closest value to 1

Model Performance (Linear R.)

Scores **Before Cleaning:**

- Accuracy score: 35.01
- RMSE: 391571
- MAE: 216488
- R2: 0.35

Scores **After Cleaning:**

- Accuracy score: 57.28
- RMSE: 178311
- MAE: 137133
- R2: 0.57

Machine Learning Conclusions

Inaccuracies could be due to:

- Dataset with lots of inconsistent values after outlier detection.
    - Lots of housings with really large surfaces selling for average prices. (3324082sqft worth $365’000 in Cote-des-Neiges. That’s 57 football fields!)
    - Lots of housings with really small surfaces selling for really high prices. (A really small one bedroom: 563sqft for $1’800’000)
- We have lots of info about the properties’ regions, but we lack descriptive data about the contents of the properties themselves. (No distinction between apartment, condo, house, mansion, etc.)