Housing market in the Laurentides region

Table of content

1. Introduction

Context:

We are in January 2021 and the Covid-19 pandemic his driving the housing market in the Montreal region to new highs. My partner and I have the desire to get "on the train" while we can and before the prices are to steep. This study is meant for us to understand the market and help us make a good decision regarding the house we will buy. It also lays the foundations for further possible business developments.

The study will be biased towards **our dream house characteristics**:

- Minimum **three bedrooms**
- Minimum **two bathroooms**
- A **decently sized house - between 1200 and 1500 sq.ft**
- A **decently sized land - between 10000 and 20000 sq.ft** away from road noise
- A **garage** so I can fullfill my handyman side

Business questions this study intends to answer:

- Study the Laurentides region market characteristics: i.e. statistical analysis of current listings
- Be able to provide a market price for a potential new listing using its characteristics
- Be able to tell if any of the current listings are considered under-priced, fairly-priced or over-priced

Data characteristics:

123 house listings were manually (because a lot of variables are subjective and needed analysis) gathered between the last week of December 2020 and the first week of January 2021. They represent all the listings for houses with **prices between 200000 and 500000$** and located within the **following municipalities**: Piedmont, Sainte-Adèle, Saint-Sauveur, Sainte-Agathe-des-Monts, Val-David, Saint-Adolphe-d'Howard, Morin-Heights and Val-Morin.

2. Data Preparation

2.1 Importation of the libraries to be used during this study

	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	Renovations	W
0	2	1	oui	non	non	non	non	1999	215000	Sainte-Agathe-des-Monts	non	
1	3	3	oui	non	oui	non	non	2019	320000	Sainte-Adèle	oui	

2.2 Data manipulations prior to analysis

	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	Renovations	W
0	2	1	1	0	0	0	0	1999	215000	Sainte-Agathe-des-Monts	0.0	
1	3	3	1	0	1	0	0	2019	320000	Sainte-Adèle	1.0	

2.3 Variables Description

Quantitave data:

- **Bedrooms:** The number of bedrooms in the house
- **Bathrooms:** The number of bathrooms in the house
- **Price:** The house price
- **Land:** The land dimensions in sq.ft
- **Rooms:** The number of rooms in the house
- **Living_area:** The house living_area dimensions in sq.ft
- **Prix_per_house_size:** The house price in terms of dollars / sq.ft
- **Price_per_land_size:** The land price in terms of dollars / sq.ft

Categorical data:

- **Village:** Is the house located in the hearth of the village or not?
- **Road:** Is the house located close to a busy road or not?
- **Water-access:** Is the house located near a river or a lake?
- **Garage:** Does the house have a garage or not?
- **Year:** When was the house built?
- **Municipality:** In which municipality is the house located?
- **Renovations:** In the house renovated?
- **Warranty:** Is the house covered under a warranty?

3. Statistical evaluation of current listings (January 2021)

3.1 Statistical distribution of quantitative variables

```
It is interesting to see that the average current house on sale in the Laurentides has:
3.0 bedrooms
2.0 bathrooms
10.0 rooms
a living area of 1705.0 square feet
a land of 30190.0 square feet
a goes for an average price of 354974.0

All of this pretty much in line with our expectations described in the intro! yeah!
```

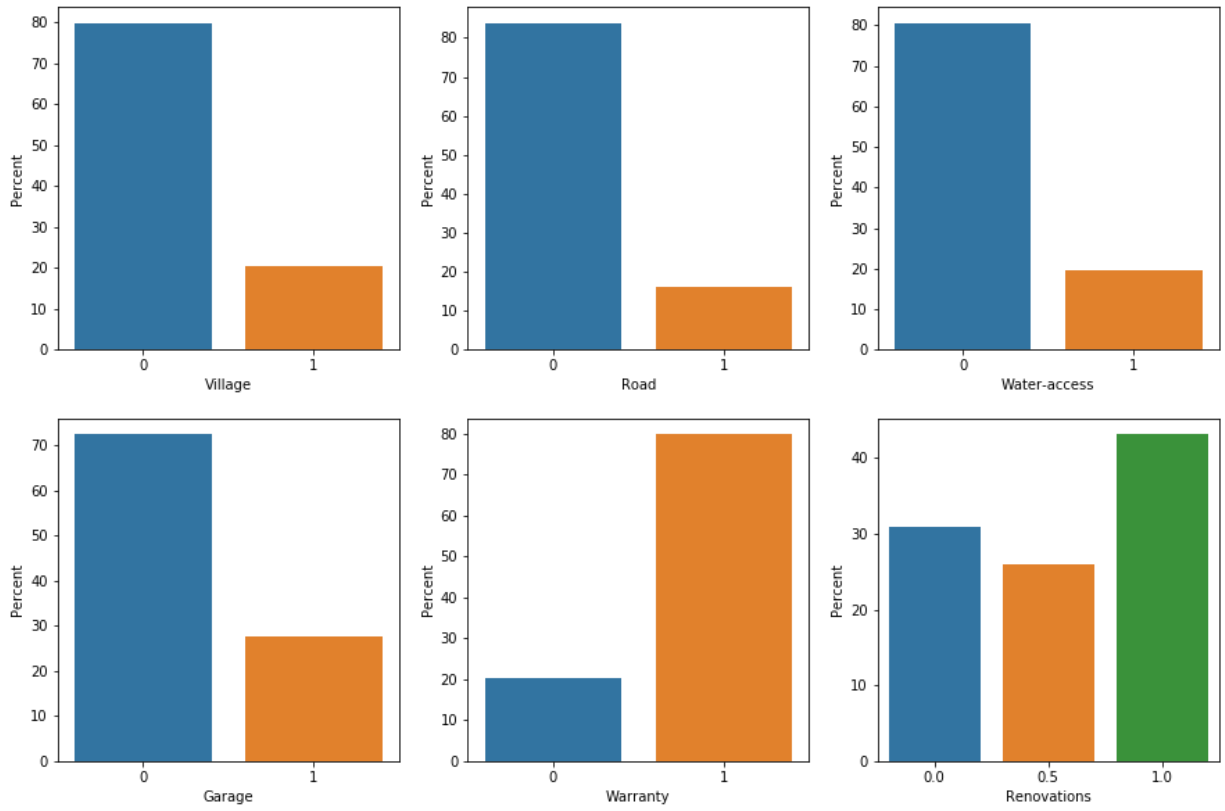
3.2 Univariate Analysis

In this section we will study the distribution of categorical and quantitative variables:

- **Categorical variables** are qualitative and can either be ordinal (rank) or nominal (no rank)
- **Quantitative variables** can either be continuous (price) or discrete (number or rooms)

3.2.1 Categorical Data

```
let's look at the distribution (in %) of the main categorical variables across all listi
ngs
0 means 'no', 1 means 'yes'
```



3.2.1.1 Discussion on categorical data

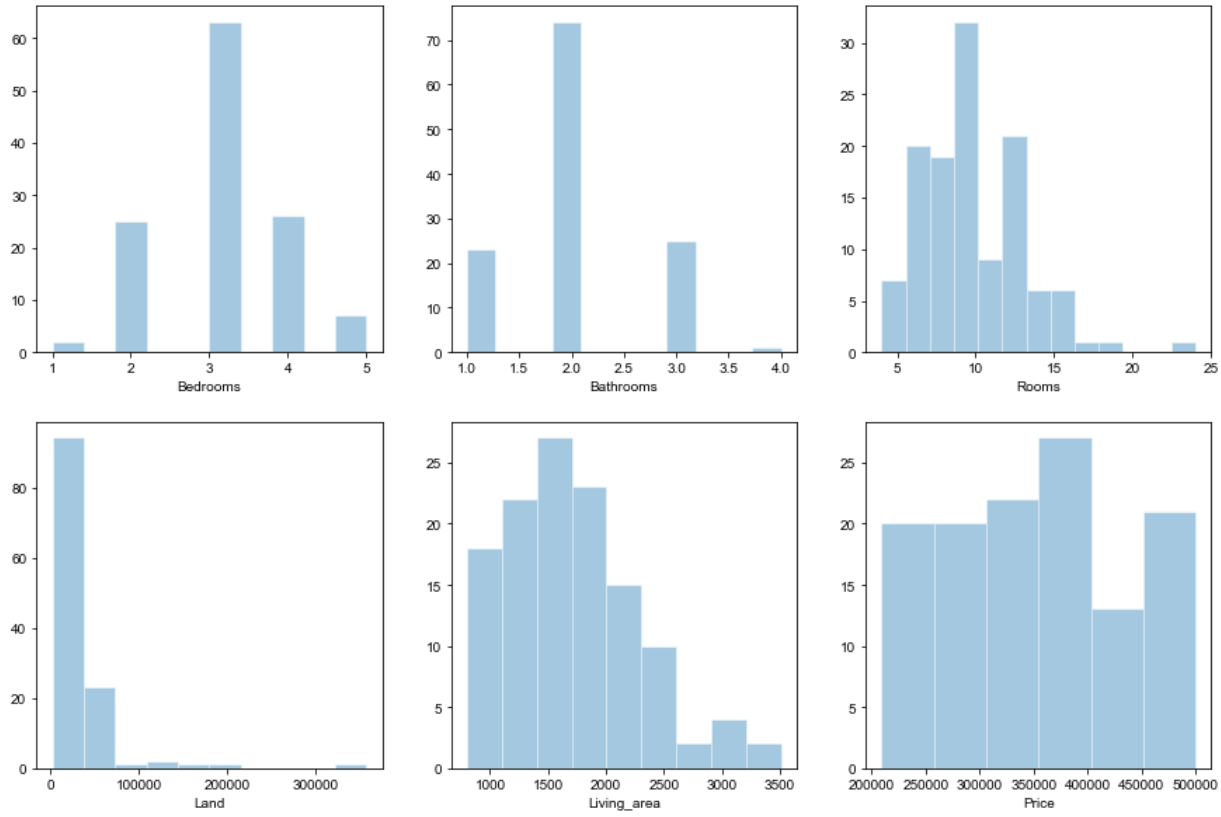
As it can be seen above, the vast majority of house currently on sale are:

- not located in the hearth of a village (more or less 80%)
- not close to a major road (more or less 80%)
- do not have access to a river or a lake (more or less 80%)
- do not have a garage (more or less 70%)
- come with a warranty (more or less 80%)
- partly or fully renovated (more or less 70%)

At the light of this, it is understood that finding a house with a garage and with water access might be more difficult and this is to be kept in mind.

3.2.2 Quantitive Data

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f0396bf508>
```



3.2.2.1 Discussion on quantitative data

As it can be seen above and as explained previously in the report, the vast majority of house currently on sale have three bedrooms, two bathrooms and ten rooms with a living area around 1500 sq.ft.

We can now assert that those variables obey to the normal distribution with some being skewed to the left or to the right.

It is also interesting to see that the price distribution seem to follow something like an uniform distribution with small peeks between 300000 and 400000$.

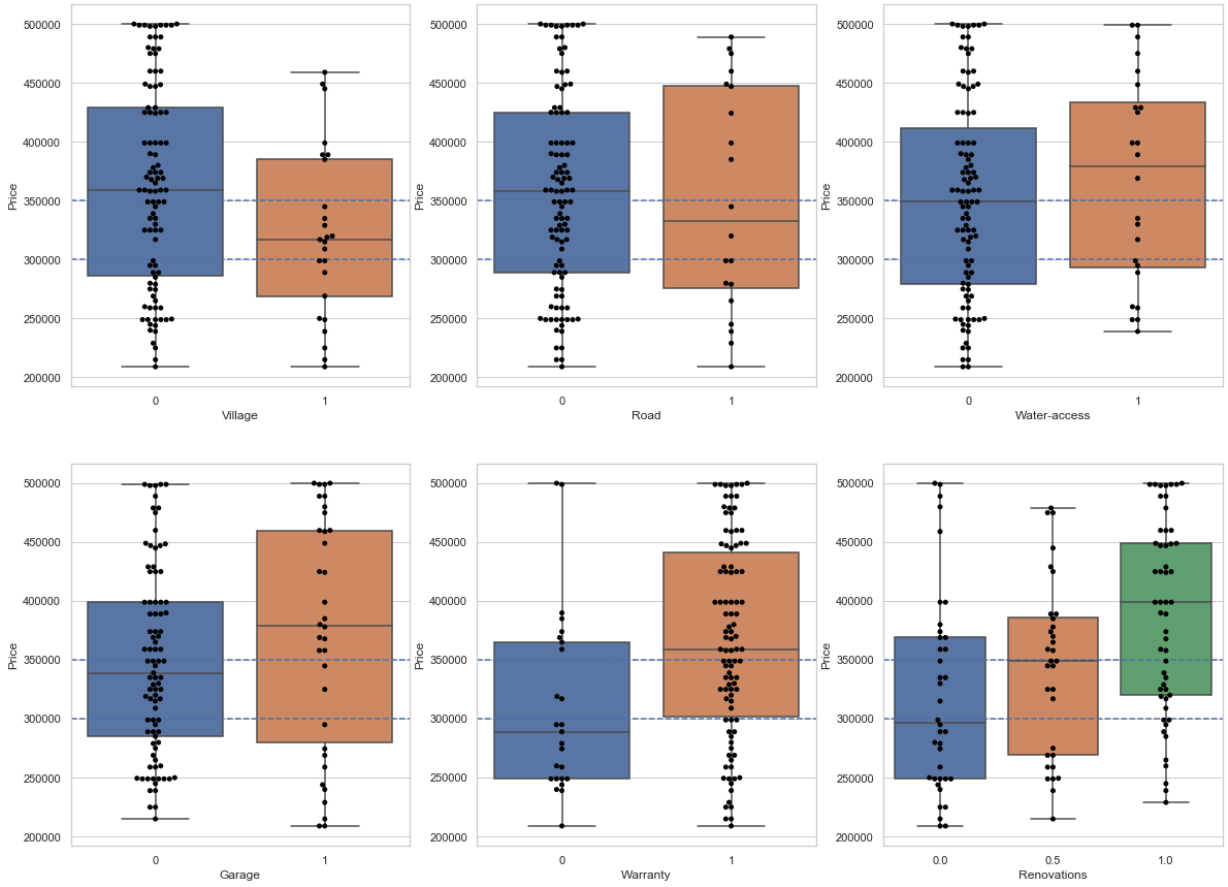
3.3 Multivariable Analysis

In this section we will study the distribution of categorical and quantitative variables against other variables to see if we can establish which ones have influence on others.

We are obviously mostly interested by a given variable influence on price.

3.3.1 Categorical variables

```
<matplotlib.lines.Line2D at 0x1f03896ae08>
```



3.3.1.1 Discussion on categorical variable

These boxplots show the influence of the categorical variables on house price. We can see that a house far from a road, with water access, a garage, a warranty and renovated will command a higher sell price.

Contrary to what we would have expected, a house located in a village sells for less than a house located in the woods. This is probably explained by the lot size.

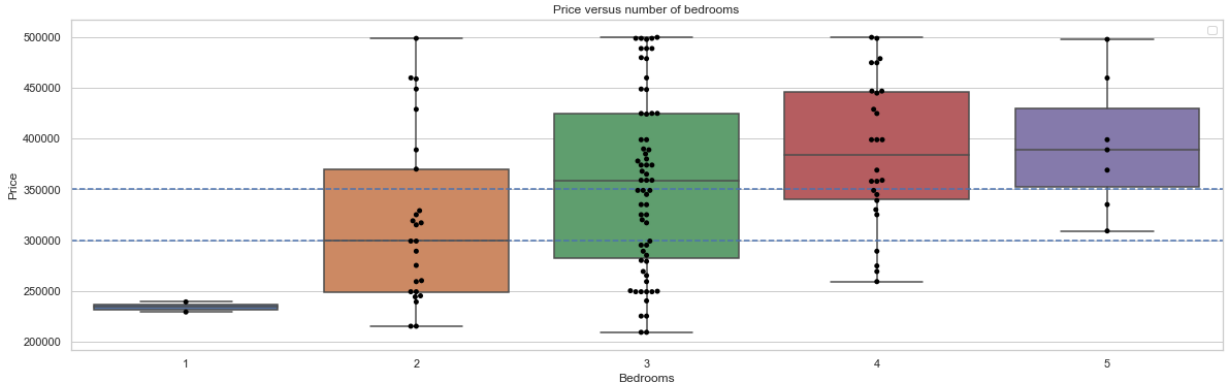
It can be also seen that the effect of a warranty and renovations will command the highest difference in median prices. A house with a warranty will typically sell 50000 dollars higher than one without one.

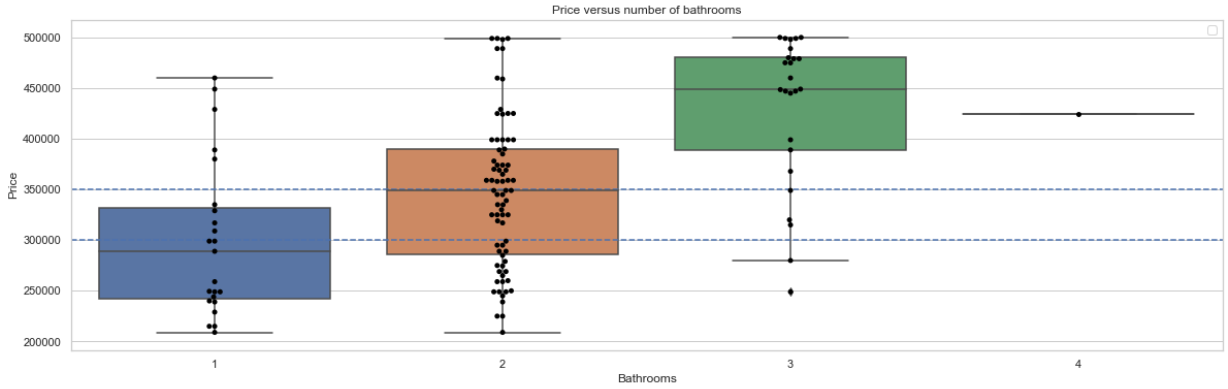
A semi or full renovated houses will for its part sell 50000 and 100000 dollars more than one that is not renovated.

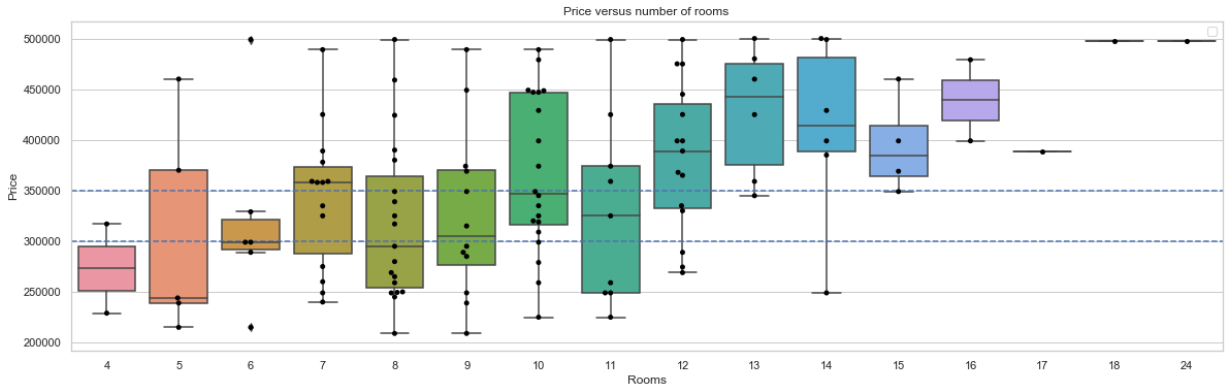
The category will the least impact is the proximity to a major road. A house close to a major road will likely sell a few thousand dollars less than one located further.

3.3.2 Quantitative variables

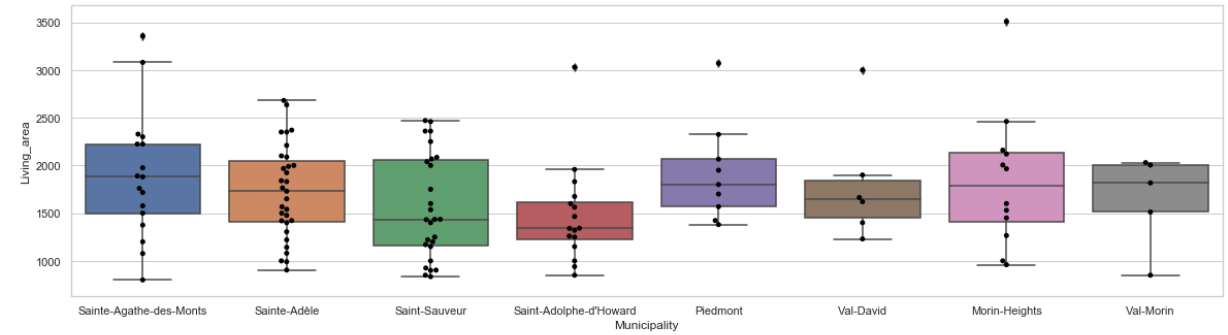
```
No handles with labels found to put in legend.
No handles with labels found to put in legend.
No handles with labels found to put in legend.

<matplotlib.lines.Line2D at 0x1f039a4a388>
```

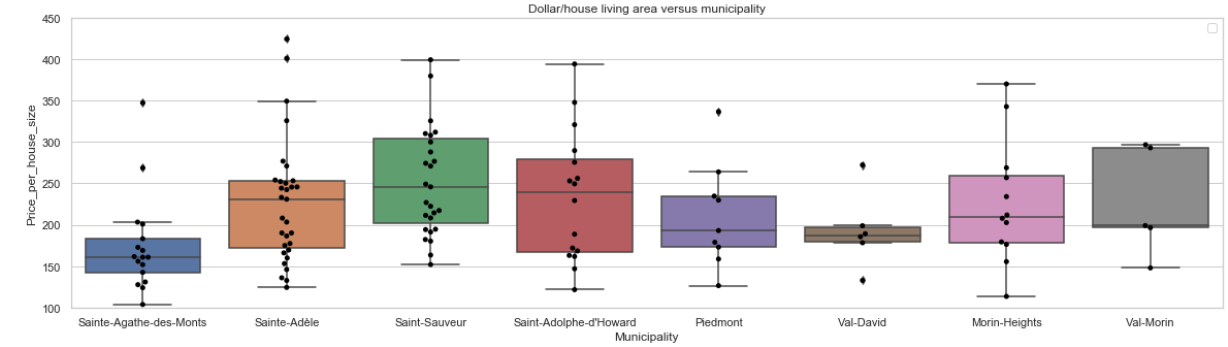


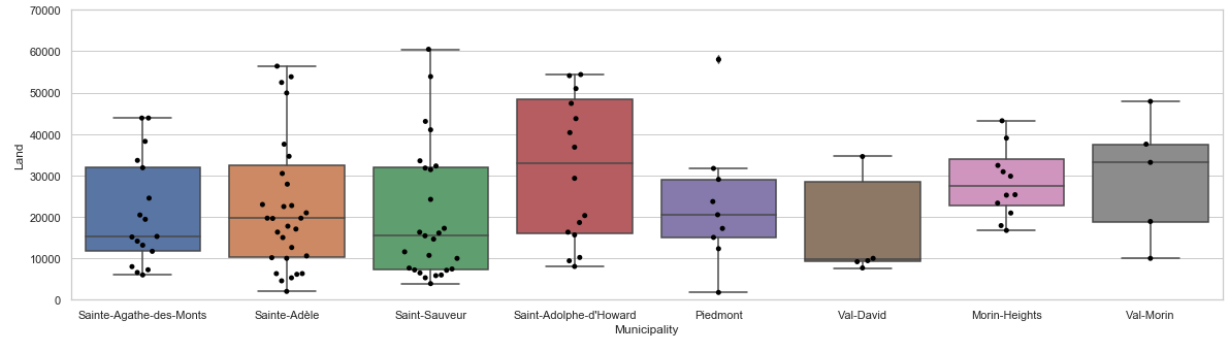


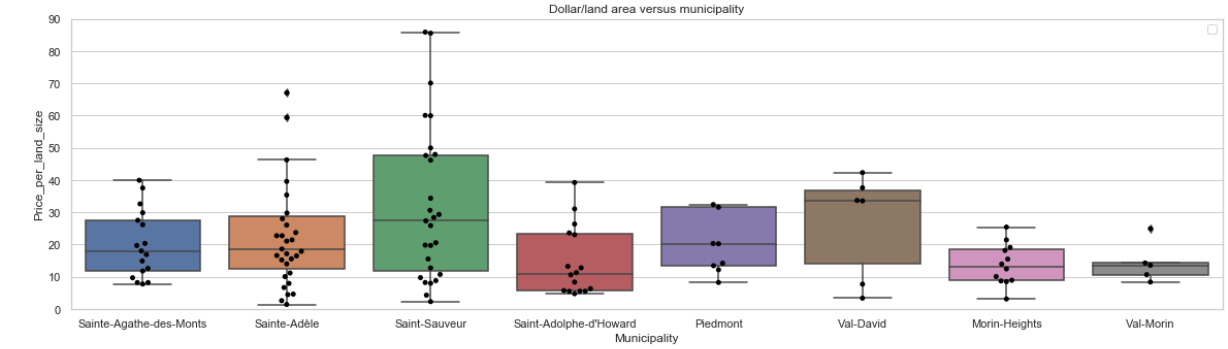


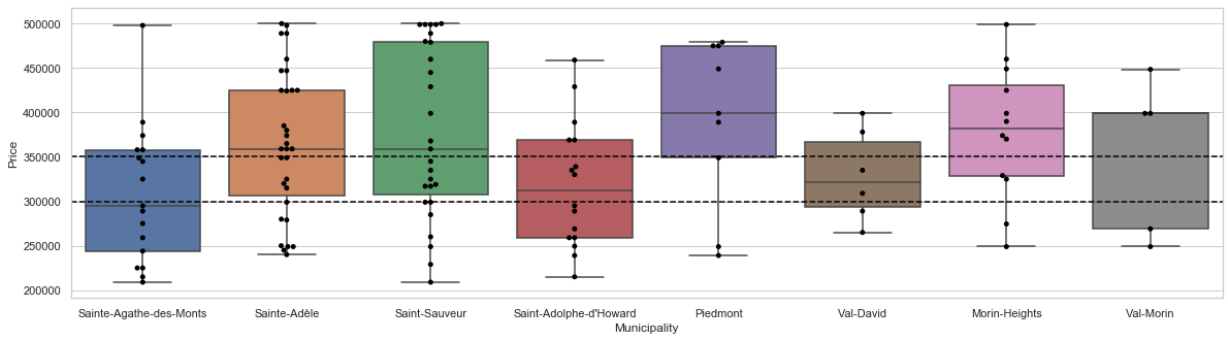
```
No handles with labels found to put in legend.
No handles with labels found to put in legend.

<matplotlib.lines.Line2D at 0x1f03bc96088>
```



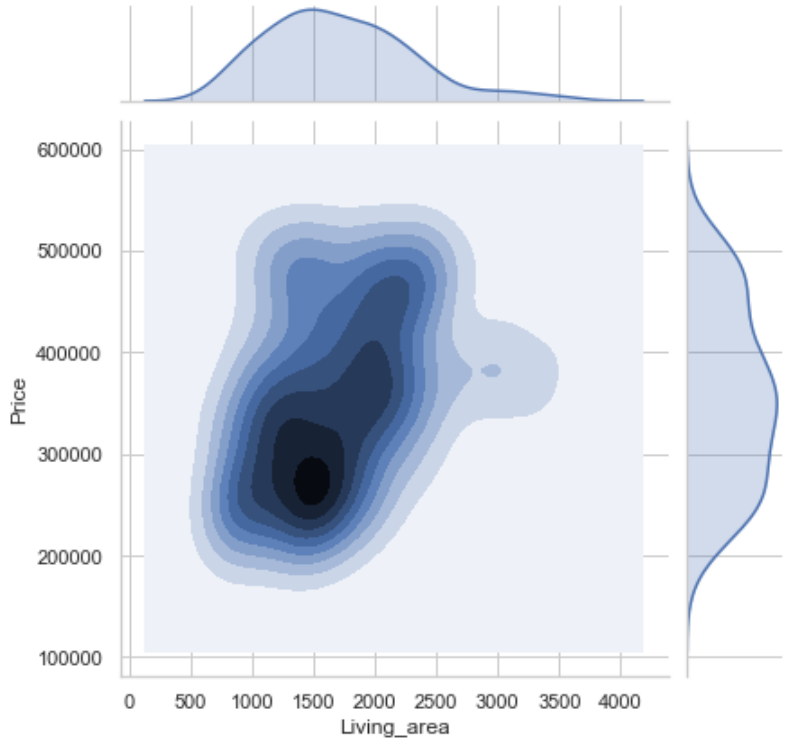




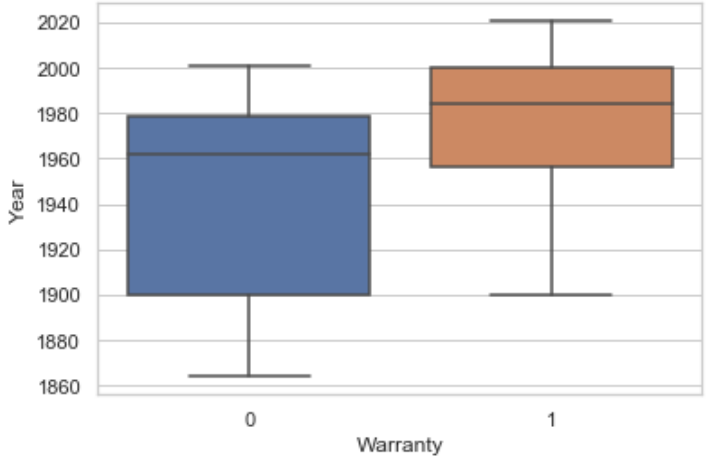




<seaborn.axisgrid.JointGrid at 0x1f03bd325c8>



<matplotlib.axes._subplots.AxesSubplot at 0x1f0389b8348>



3.3.2.1 Discussion on multivariable analysis

The first set of boxplots show the influence of the quantitative variables on house price. As excepted, the more bedrooms / bathrooms and rooms a house has, the highest is price is. The trends seem linear but plateaus in the higher independent variables suggest the trends might more towards logarithmic relationships.

The second set of boxplots represents the distribution of house size (in sq.ft), land size (in sq.ft) dollar/house area, dollar/land area and price versus municipality. The analysis of these graphs show that the house size is typically larger in Sainte-Agathe-des-Monts with a unit cost per area considerably cheaper than in the other towns. The median of the listings in this municipality is the lowest amongst all of them. The median listing in Saint-Adolphe-d'Howard provides a smaller house but a bigger land when compared to the other towns.

It is interesting to note that Val-Morin and Saint-Adolphe-d'Howard are the two towns where the median listing is located between our price range (300000 to 350000 dollars).

Finnaly, we can observe that listings offering a warranty are typically newer constructions. The median listing offering a warranty is built in 1980 compared to 1960 for a listing not offering a warranty.

4. Modeling of a new listing price predicter

Context:

The idea behind this predicter is for us to determine the price of a house based on the attributes we want it to have. We will build a predicter and use the house values below but the user can input the values of its choice:

- Minimum **three bedrooms**
- Minimum **two bathroooms**
- A **decently sized house - between 1200 and 1500 sq.ft**
- A **decently sized land - between 10000 and 20000 sq.ft** away from road noise
- A **garage** so I can fullfill my handyman side

We are not set yet on the municapility we want to live in, the predictor will help find which one would have a house that fits our needs.

Methodology:

Given the previous analysis, we chose the variables that had a clear influence on price to build the predictive model:

- Bedrooms
- Bathrooms
- Rooms
- Living area (sq.ft)
- Land area (sq.ft)
- Garage
- Water access
- Renovations
- Warranty
- Renovations
- Municipality

The model will either be based on a multivariate linear regression, multivariate polynomial regression or a KNN (K-nearest neighbors) algorithm.

Given the fact that we have a lot of variables (11!) and relatively few listings, we will start by finding the random seed that yields the highest "explained variance score" for a multivariate linear regression with a train / split of 0,84 (we want 20 listings to test the algorithm = 123*0,8 = 20).

After that, we will train the three different models described before and see which one yields the best results (i.e. the best variance score). We will pick the best of them to estimate the price of our dream house! :)

4.1 Determination of best random split

we convert the categorical variables to numerical variables: 0s or 1s

	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Renovations	...	Link_https:/
0	2	1	1	0	0	0	0	1999	215000	0.0	...	
1	3	3	1	0	1	0	0	2019	320000	1.0	...	
2	3	2	1	0	0	0	0	1946	334900	1.0	...	
3	2	1	0	0	0	0	1	1900	215000	0.5	...	
4	1	1	0	0	1	0	1	1970	229000	1.0	...	

5 rows × 268 columns

Methodology:

- 1. We loop items 2 to 6:
- 2. we split the data into Train and Test.
- 3. We scale the training set to get better results.
- 4. We fit a multivariate linear regression model
- 5. We estimate the sell price of the training set
- 6. We check the variance
- 7. We select the seed that provides the best variance score")

```
we see that seed # 2090 yields an explained score of 0.8375967184326683
```

We then split our dataset (train & test) using the seed above 84% of all listing being in the train dataset

4.2 Determination of which model provides the best variance

4.2.1 Multivariate Linear Regression

We fit our multivariate linear regression model and find the corresponding variance score

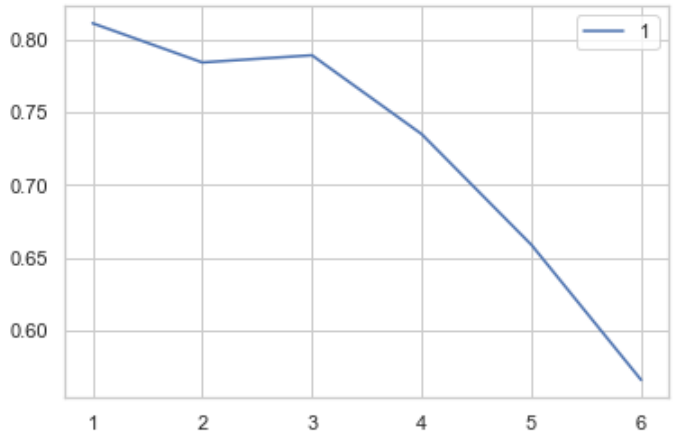
```
Residual sum of squares: 1130236921.32
Variance score: 0.84
```

4.2.2 Multivariate Polynomial Regression

Using the seed found above, we test multivariate polynomial order from 1 to 7 to find which one yield the best variance score.

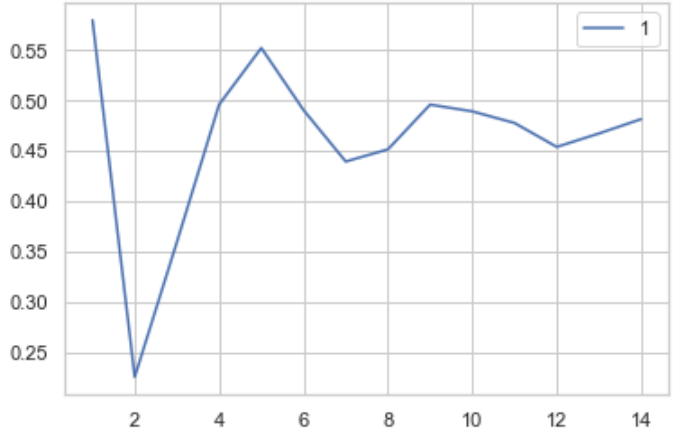
```
Residual sum of squares: 1316166025.30
Variance score: 0.81
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f03bf1e948>
```



4.2.3 K-Nearest Neighbors

```
Residual sum of squares: 2923277142.86
Variance score: 0.58
```

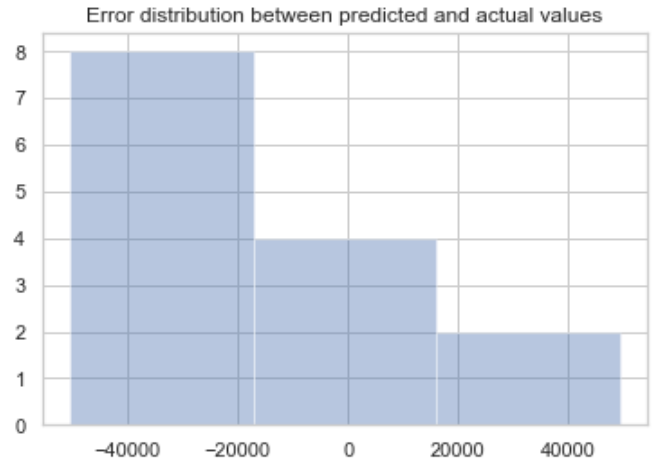


4.2.4 Discussion

We can see that the multivariate linear regression model yielded better variance results than the other two model (linear: 84, polynomial order 1: 0.81 and knn: 0.58). Obviously, the seed was selected to optimize the variance of the linear model but the fact that most inter-variable relationships explored above seemed linear also helps to explain the results.

Looking at the chart and table below, we see that the error distribution between predicted and actual values is skewed to the left with min. and max. differences being around 50000 dollars. The standard error being around 30000 dollar satisfies us.

```
Text(0.5, 1.0, 'Error distribution between predicted and actual values')
```



	0
count	14.000000
mean	-17007.138026
std	30094.618079
min	-50465.545521
25%	-44714.374659
50%	-21527.427042
75%	3523.207092
max	49184.673884

4.3 Determination of our dream house price

Let's now find the price of our dream house in the different towns

For the estimate, we will take the following characteristics for our dream home:

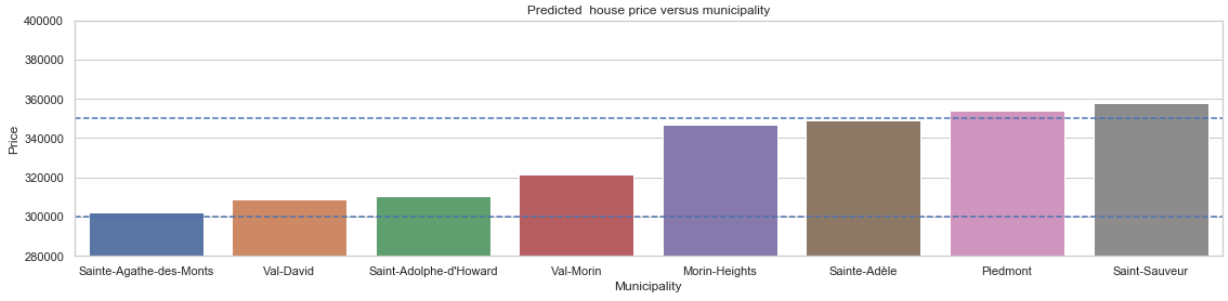
- Three bedrooms
- Two bathroooms
- Nine rooms
- A house of 1300 sq.ft
- A land of 15000 sw.ft away from road noise
- No water access
- Partly renovated
- With a warranty
- A garage

```
user prompts here the values of his choice
How many bedrooms?: 3
How many bathrooms?: 2
How many rooms?: 9
House area in sq.ft?: 1300
Land area in sq.ft?: 15000
Water access? yes:1, no:0: 0
Renovated? yes:1, semi:0.5, no:00.5
Warranty? yes:1, no:0: 1
Garage? yes:1, no:0: 1

We prepare the matrix to feed the predictive model
```

	Bedrooms	Bathrooms	Living_area	Garage	Water-access	Renovations	Warranty	Rooms	Municipality	Land
5	3	2	1300	1	0	0.5	1	9	Sainte-Agathe-des-Monts	15000
6	3	2	1300	1	0	0.5	1	9	Val-David	15000
2	3	2	1300	1	0	0.5	1	9	Saint-Adolphe-d'Howard	15000
7	3	2	1300	1	0	0.5	1	9	Val-Morin	15000
0	3	2	1300	1	0	0.5	1	9	Morin-Heights	15000
4	3	2	1300	1	0	0.5	1	9	Sainte-Adèle	15000
1	3	2	1300	1	0	0.5	1	9	Piedmont	15000
3	3	2	1300	1	0	0.5	1	9	Saint-Sauveur	15000

We scale the matrix and feed the predictive model with it. We then plot the prices by municipality

```
(280000, 400000)
```



4.4 Discussion on predicted house prices

We see from the previous barplot graph that the house we want has more chance to be in our price range in the municipalities of Sainte-Agathe-des-Monts, Val-David, Saint-Adolphe-d'Howard and Val-Morin.

Obviously this represents just an estimate of a sell price, we have to remember that the standard deviation of the difference between the predicted values and the actual values is around 30000 dollars. This is to say that we could be lucky and find a house with the characteristics we want in Morin-Heights with a price within the standard error and still respect our price range / capabilities.

This brings the next section where will determine which listings are under-priced and righlty-priced.

5. Price evaluation of current listing

As described in the previous section, we now will evaluate each listing on the market and determine if they are properly priced.

This will enable us to see which listings are considered good opportunities and which ones are considered rightly priced.

- An opportunity will be a case where the estimated price will be at least 10000 dollars LESS than the actual sell price
- A righlty priced listing will be a case where the estimated price will be plus or minus 10000 than the actual sell price

To keep things simple, we will search for listings using only three parameters: three bedrooms, two bathrooms, a warranty and a sell price above 275000 dollars and below 360000 dollars.

Under-priced listings

	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	Renovations
2	3	2	oui	non	non	non	non	1946	334900	Saint-Sauveur	oui
6	3	2	non	non	non	non	non	1973	285000	Saint-Sauveur	oui
41	3	2	non	non	non	non	non	1994	349000	Piedmont	oui
47	3	2	non	non	non	non	non	1988	325000	Sainte-Adèle	semi
73	3	2	non	non	non	non	non	2007	325000	Sainte-Agathe-des-Monts	oui
80	3	2	non	non	non	non	non	1900	359000	Saint-Sauveur	non

Rightly-priced listings

	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	Renovations
13	3	2	non	non	non	non	non	1973	345000	Saint-Sauveur	semi
24	3	2	non	non	non	non	non	1975	359000	Sainte-Adèle	oui

Let's obtain the latitudes and longitudes for each of team

```
C:\Users\simon\anaconda3\lib\site-packages\pandas\core\indexing.py:844: SettingWithCopyW
arning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  self.obj[key] = _infer_fill_value(value)
C:\Users\simon\anaconda3\lib\site-packages\pandas\core\indexing.py:965: SettingWithCopyW
arning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  self.obj[item] = s

581, Rue du Plateau, Sainte-Agathe-des-Monts, Quebec, Canada HTTPSConnectionPool(host='n
ominatim.openstreetmap.org', port=443): Max retries exceeded with url: /search?q=581%2C+
Rue+du+Plateau%2C+Sainte-Agathe-des-Monts%2C+Quebec%2C+Canada&format=json&limit=1 (Cause
d by ReadTimeoutError("HTTPSConnectionPool(host='nominatim.openstreetmap.org', port=44
3): Read timed out. (read timeout=1)"))


C:\Users\simon\anaconda3\lib\site-packages\ipykernel_launcher.py:13: SettingWithCopyWarn
ing:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  del sys.path[0]
```

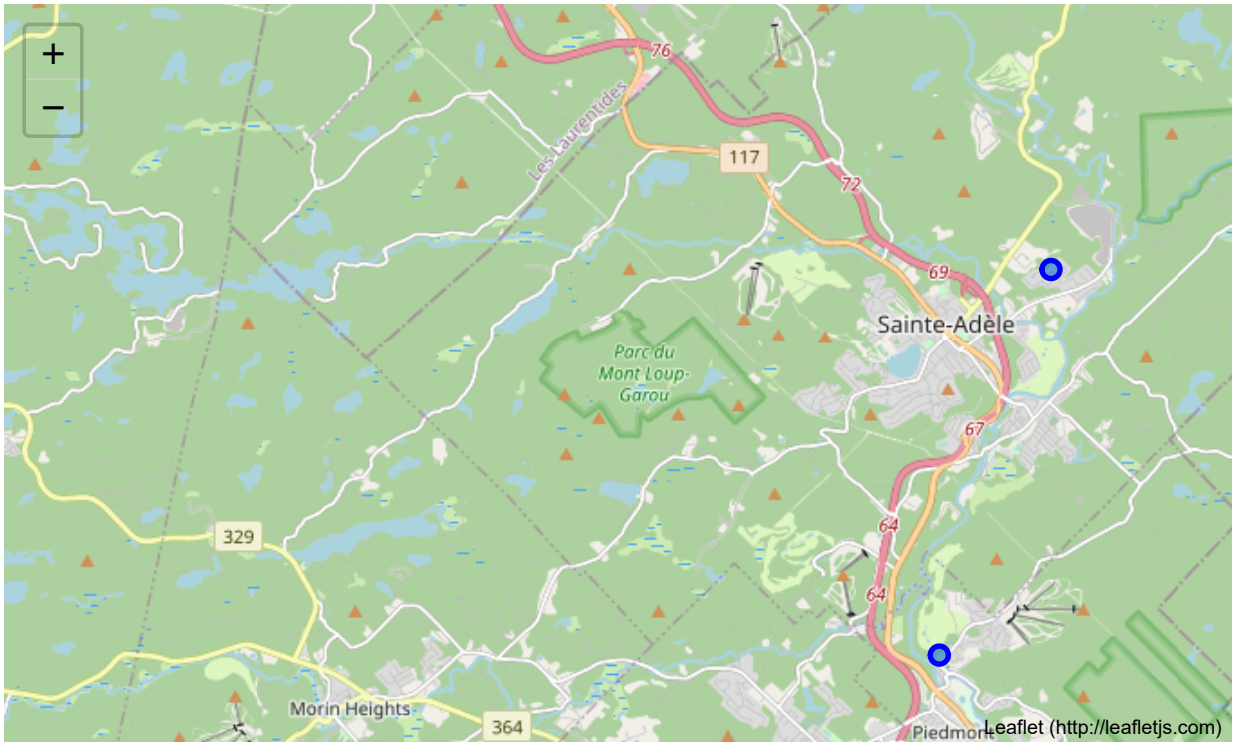
	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	...	Warranty
2	3	2	oui	non	non	non	non	1946	334900	Saint-Sauveur	...	oui
6	3	2	non	non	non	non	non	1973	285000	Saint-Sauveur	...	oui
41	3	2	non	non	non	non	non	1994	349000	Piedmont	...	oui
47	3	2	non	non	non	non	non	1988	325000	Sainte-Adèle	...	oui
80	3	2	non	non	non	non	non	1900	359000	Saint-Sauveur	...	oui

5 rows × 21 columns

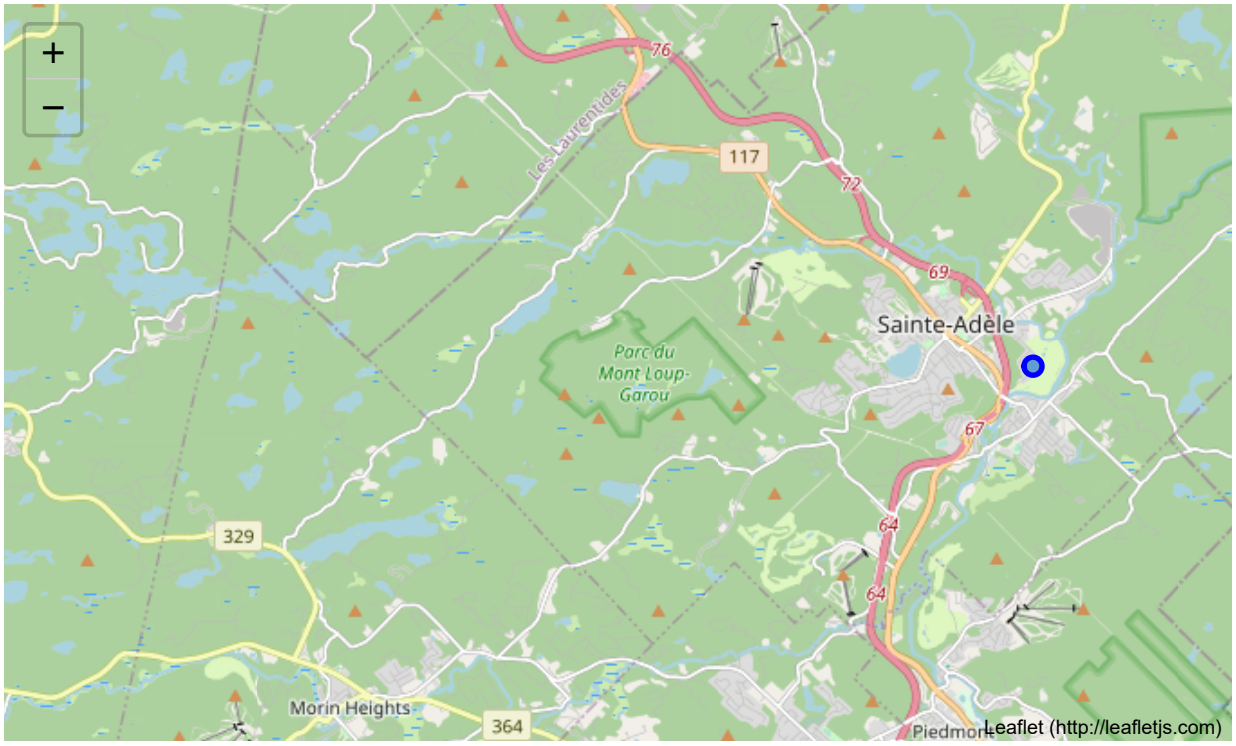
	Bedrooms	Bathrooms	Village	Proxi-ski	Road	Water-access	Garage	Year	Price	Municipality	...	Warranty
13	3	2	non	non	non	non	non	1973	345000	Saint-Sauveur	...	oui
24	3	2	non	non	non	non	non	1975	359000	Sainte-Adèle	...	oui

2 rows × 21 columns

Map of good opportunities



Map of rightly-priced houses



The End!