# MONTREAL REAL ESTATE APPRISAL USING MACHINE LEARNING

A PREPRINT

**Agata Kopyt**
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Poland

**Zuzanna Kotlińska**
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Poland

**Szymon Matuszewski**
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Poland

**Patryk Słowakiewicz**
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Poland

March 17, 2022

***Keywords*** Machine Learning · Real Estate Apprisal · Montreal · Tree-Based Models

## 1 Data

The data we have analysed originates from https://github.com/npow/centris/blob/master/scraper/data/all$_data.csv$ and concerns properties in Montreal - primarily their prices, as well as other factors (location, year of construction, square footage, among others) that affect them. The data set consists of 9013 rows and 138 columns.

The first thing that strikes the eye is the excessive number of variables. Some of them seem to be redundant for us and will be omitted in further analysis. We also checked the type of our data - each of the variables is numeric, which is certainly conducive to further analysis. In most of the columns we have not encountered NaN values, only ConstructionYear and LivingArea have 1415 and 3930 missing data respectively. It is worth noting that there are a fair number of columns with multiple null values. If they are not categorical, they are the ones that will be omitted from the data analysis as they do not contribute much information. Such variables are NbEquipements, NbAnimals, NbCultures, Lile-Dorval, FE and FER.

We looked at what the distributions of the continuous variables contained in our dataset look like. Various distributions appear, including Gamma and exponential. Some of the variables may need to be transposed. We also looked at which columns are highly correlated. We obtained a very high correlation coefficient (about 0.986) for the variables Bach and University, as well as for Population and TotalFam (about 0.973) or totalHouse and totaldwelling (approximately 1). Having analysed each of the strong correlations, we can immediately conclude that some of the variables (e.g. totaldwelling) should be discarded, as they do not add any additional information. It will certainly be more interesting to find out with which variables the property price is most correlated, as it is the main object of our analysis. It turns out, which is not surprising, that the real estate price is most influenced by the square footage. Interestingly, the next significant variables are the number of bathrooms, type of development, whether the property was bought for rent, number of rooms and bedrooms. However, these variables are not so strongly correlated with the price (the coefficient is around 0.24-0.30). We have made some charts showing the influence of some factors on the property price. Given that the square footage has the strongest influence on the price, we checked how the correlation between them is presented. Here the conclusions are obvious - the bigger the area, the more expensive the property. We can also see that there are not many very expensive properties on the market, and the majority of them are cheaper and in the middle price range. We also compared the price with the average income of the buyer. Interestingly, buyers with the highest incomes do not tend to purchase the most expensive properties - quite the opposite. It turns out that the average income of buyers of expensive properties is not as high as one might expect.