**Business problem:**

1. To justify a sold price with given features (number of bed room, free hold or not, distance to GO station etc.)
2. To predict a fair and real price of a house with given features

**Time window:** July 1st 2015 to June 30th 2017

**Dependent variable:** real sold price (adjusted by CPI)

**Independent variables**:

**Feature engineering / data transformation:**

1. transform categorical data into numerical data (house type, house fee, and house area)

* Rationale: categorical data needs to be quantified to be eligible as model inputs.

2. adjust sold prices by CPI index

* Rationale: under our cross-sectional regression model, the independent variables are time invariant (i.e. the distance to GO station and the number of bedrooms do not change with the passage of time). However, sold prices could change due to inflation. Adjusting the sold prices by CPI index ensures that the prices are measured by the same dollar value across the entire period from July 1st 2015 to June 30th 2017.

3. incorporate real estate selling seasonality based on sold dates

* Rationale: seasonality is an observable pattern in real estate market where its business tends to prosper in summer but cool down in winter.

4. model real prices for selective house types (town house, condo, detached, and semi-detached)

* Rationale: the raw file contains other property types out of the scope of our designed model including vacant land, mobile trailer etc.

5. scale the minimum distance to GO station and hospital to the range of 0 to 100 with the shortest distance to GO station and hospital close to 0 and the furthest close to 100.

* Rationale: recalling the minimum distances to GO station and hospital can make them comparable to other categorical variables, which range from 0 to 34. This can also avoid having extreme estimated parameters / betas.

6. transform the real sold price using log function

* Rationale: the histogram of real sold prices indicates a pattern of log normal distribution. Using log function to transform the real sold prices can make them behave as if a normal distribution. In addition, log transformation ensures that the predicted real sold price is always positive, for example, exp(X) > 0.

**Outliner removal**:

1. Exclude Malton (one of area in Mississauga) from the model

* Rationale: the distance between a nearest hospital/GO station and a house in Malton is two times longer than in other areas in Mississauga. This is indicated by the below graph on group 10.

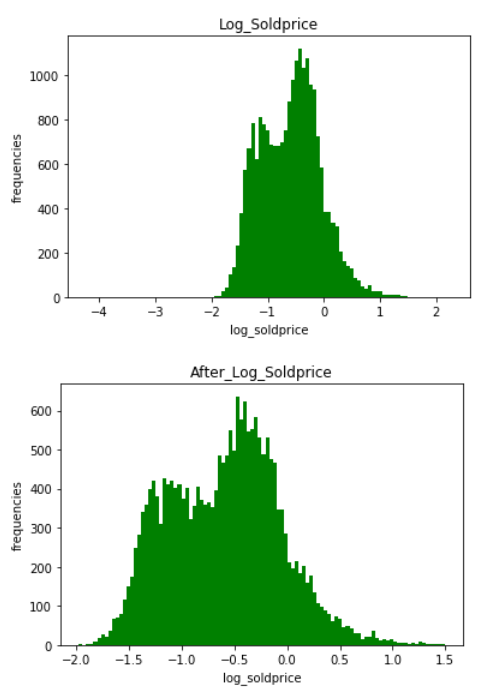


2. Exclude a house sold at zero price

* Rationale: irrelevant data point

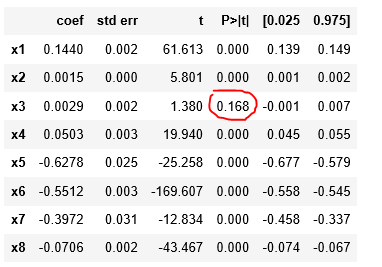
3. Exclude the outliners of log real sold prices.

* Rationale: these outliners cannot be reasonably projected based on the independent variables.



4. Exclude seasonality as an independent variable

* Rationale: the P value for seasonality is high, indicating that it is not affecting the dependent variable statistically and significantly.



**Statistics summary**:

1. High R-squared (89.1%): this indicates that 89.1% changes of dependent variables (log real sold price) can be explained by the independent variables (minimum distance to GO/Hospital, number of bedroom etc.).

2. High F statistic: this indicates that all the coefficients are jointly significant in explaining the dependent variable (in other word, it is unlikely all the coefficients are zero).

3. High individual T statistics: this indicates that all the selective coefficients are individually significant in explaining the dependent variable.

4. Autocorrelation: not a concern for a cross sectional regression

5. The error skewness (0.611): this indicates that the error terms are lightly skewed.

6. The error Kurtosis (4.7): this indicates that the error term distribution has a fatter tail.

