**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 equal to 100 and the furthest to 0.

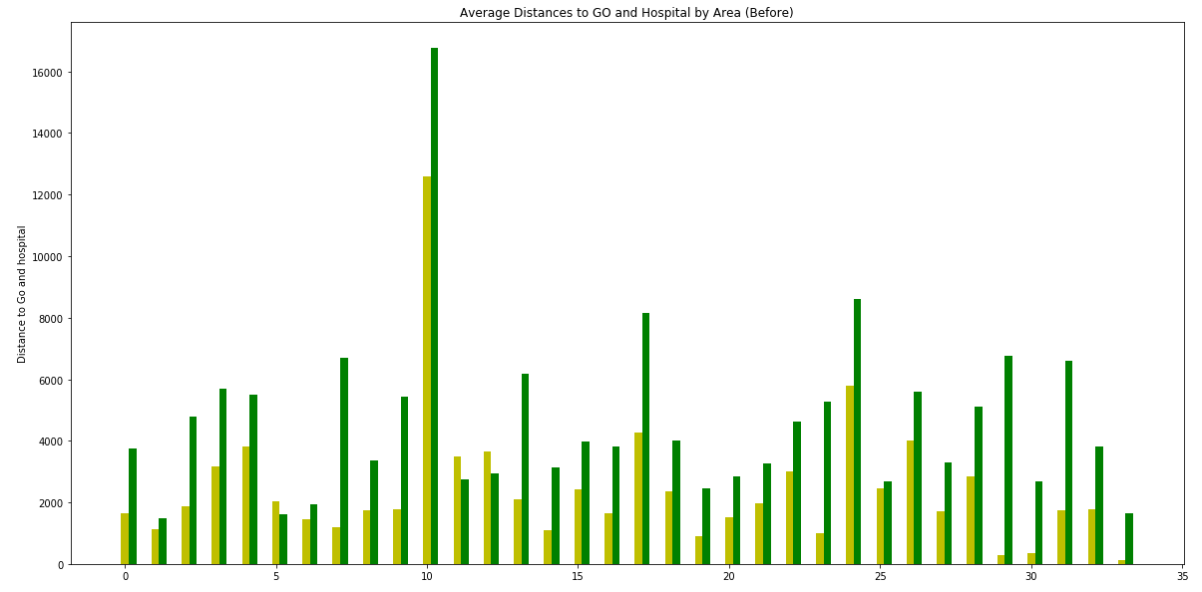
* Rationale: recalling the minimum distances to GO station and hospital can make them comparable to other categorical variables, which range from 0 to 34. This can also avoid having extreme estimated parameters / betas.

**Outliner removal**:

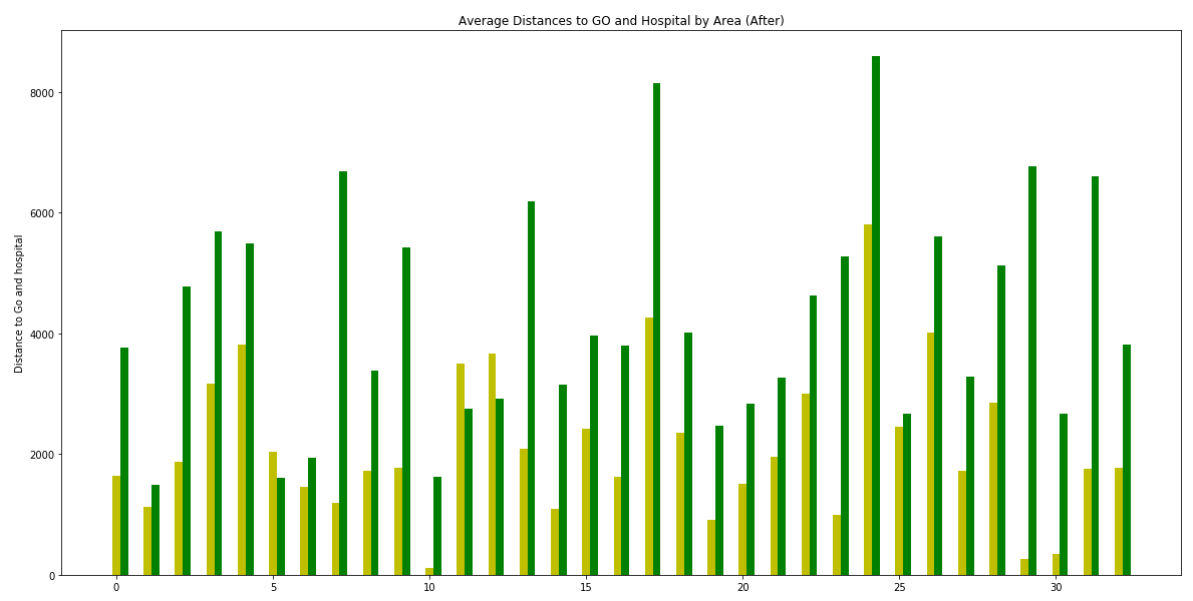
1. Exclude Malton (one of area in Mississauga) from the model

* 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.

Before



After



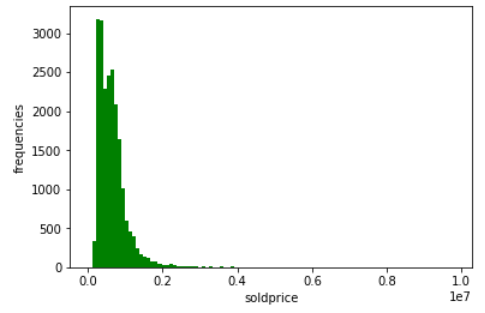
2. Exclude a house sold at zero price

* Rationale: irrelevant data point

3. Exclude houses that were sold at more than three billion.

* Rationale: these houses are outliners as their prices cannot be reasonably projected based on the independent variables. As indicated by the graph below, with these outliners, the entire house prices are highly skewed to the right with a long right tail.

Before



After

