**Project name: Property Tax Report in Vancouver**

**Date: 18-Jun-2019**

**Author: Niloofar Zarifi**

This dataset is free and available to everyone from the official BC government website (<https://data.vancouver.ca/datacatalogue/propertyTax.htm>). The data is updated in the normal course of business, however priorities and resources determine how fast a change, in reality, is reflected in the database. Please note that the property tax data for the current year is updated weekly and for this project the data was downloaded on June 8, 2019. The dataset contains over 209000 observations and 28 attributes. Here I briefly explain the attributes used in this project

* Numerical: CURRENT/PREVIOUS\_LAND\_VALUE, CURRENT/PREVIOUS\_IMPROVMENT\_VALUE, NEIGHBOURHOOD\_CODE, YEAR\_BUILT
* Categorical: ZONE\_CATEGORY, PROPERTY\_POSTAL\_CODE

Definition of current, improvement and previous value:

* In tax assessment speak, the 'improvement' is the house itself, as opposed to the lot. So, your 100k house with an improvement percent of 80% means that of the total or current value of 100k, the improvement (house) is worth 80k, with the lot being worth 20k. The previous value is for the previous assessment year.

Definition of the neighbourhood, zone category and property postal codes:

* These attributes are attached to a property address. Note that there are several columns referencing data to the address in the BC dataset.

Cleaning data**:**

* Unfortunately, 99% of the observations have at least one missing data. I decided to clean the data for each individual question.

Statistical analysis:

Number of properties for each year

* The scatter and bar chart plots were shown which count frequency of YEAR\_BUILT. As expected, over a century the number of properties are increasing and also there is no data point between 1800 to 1886.

Land and improvement Value in Current Year

* It shows the frequency of current land value and its improvement value in two plots. I removed those data points that show the value of the property is zero or one dollar. As shown the majority of the data are below 1 million dollars.

Zone Categories division

* The pie chart shows top zone categories with a percentage over the total number of properties. Here I merged four categories to a new observation called ‘Others’ since they only present less than 3% in the dataset.

Rate of growth for land value assessed in 2018 and 2017

* Here I noticed that data has unbelievable outliers. To my point of view, it is possible to lose almost 100% of your property with wildfire within a year but it will be extremely rare condition that one land value raises up to 3000% profit in a year! Therefore I decided to remove extremely positive outliers.

Summary of Land Values over Years

* Here I compared median and mean data of current land value over the years. This plot cannot meet our expectation. Generally speaking, the property value over a century should be increased while the plot shows outliers around 1900. The reason is only one data point for years around 1900 has been stored.

Count neighbourhood code and correlation with land values.

* It simply shows some codes have more values and some codes have more properties.

Land value vs Zone category vs neighbourhood code

* Although data points could easily overlap in this colorful plot, it somehow shows commercial and industrial area is more valuable.

Rate of Growth vs Zone Category

* Here I labelled zone category to smaller groups. For example, three groups of families were merged into one label called “Family”. As shown, commercial and industrial groups have a higher rate of growth comparing to 2017.

Conclusion

* The dataset attributes were mostly nominal rather than numerical.
* To ask a meaningful question I need to understand the attribute in detail
* Proper data cleaning and analyzing the outliers was quite challenging

My codes were stored in two files, cleaning.R and groupby.R