BOOMERANG: Greater Vancouver House Price Analysis

Hyelim Moon and Joanne Yoon

***Abstract*—Boomerang is an extensive Greater Vancouver house price analysis with a focus on Vancouver and Surrey. Using big data, we studied historical datasets and current trend to find out what features make house prices fluctuate so much. Analysis results and our tools are reachable via a Google Cloud compute machine. Ask us what you want. We will swift through time to deliver the answers to your fingertips like a boomerang.**

I. Introduction

Boomerang is an all-in-one Greater Vancouver property value assessment program that swifts through past, present and future to deliver the answers to your fingertips like a boomerang. We grouped houses by similar features and contrasted value and its fluctuation with different groups. This will inform users when looking for certain features of a house or contrasting two houses with some differences. Since Surrey and Vancouver has many schools, we compared their houses' relationship with nearby schools and statistically analyzed correlation of these features and property prices. We have also launched a 16.04 Ubuntu instance under the Google cloud platform so that from the web, users can analyze the house market by postal area and by specific house features. In this paper, we are going to explain our motivations, data preparations for the analysis, analysis results, and our final products.

II. Motivation and Background

Should I buy this house? Is it worth selling my current house? Is the new house in a good neighbourhood? As investment decision is a big decision for a family, everyone is interested in these questions. To make a good investment, it is vital to know which features formulate and fluctuate the prices. From another perspective, banks will be interested in this analysis because it helps them to decide whether they should provide mortgage to a customer or not. Furthermore, the government can use this analysis to detect abnormal fluctuations and enforce new regulations. The sellers, buyers, banks, and government are all interested in the climate of the house prices.

Since this question is frequently asked, there are plenty of research on this area. Boomerang applies these theories and displays an all-in-one overview of the area that the user is interested in. We present analysis in a specific house level, general postal area level that is grouped by the first three characters of the postal code, and city level. As a result, we searched through multiple datasets and combined them in order to understand today's market from diverse perspectives. Our final product analyzes the data via machine learning and statistics to answer the true value of a Greater Vancouver house.

III. Data Science Pipeline

The following sections IV to VII dives into how we collected, integrated, and analyzed the property value data in order to create the final Boomerang product.

IV. Data Collection

In order to get real time house listings, we web scraped Real Estate Wire (REW). REW is a real estate marketplace and information hub in BC. We used REW's listings to understand the current market and profitable features of listed houses.

Historical data will illustrate price evolution of different houses, but past data was hard to find because it was confidential and a profitable asset that was not released to the public. Luckily, City of Surrey and Vancouver had open data of the past property values that they used to derive property tax. Since Vancouver and Surrey were the only municipalities that provided historical data, some of our analysis was restricted to Vancouver and Surrey properties.

House prices does not just depend on the location and features. It is quite sensitive to its surroundings. In order to understand the influence that nearby schools has on the house price, we scrapped each Greater Vancouver secondary school's rating from Fraser Institute's website. We initially gathered school data without any assumptions, but were surprised with the resulting insights. The analysis and the finding are discussed later in the paper.

V. Data Cleaning and Integration

Current house listings and historical municipality open data were all broken up into csv files. We first explored each dataset to understand the meaning and value of each column in different datasets. Then we cleaned each dataset to the same format and joined them on their address. By having multiple data sources, we were able to fill in missing data from a source using data from another source. As a result, we had current listing price of a house and their historical prices.

Each house recorded the secondary schools near its catchment. We combined the school rating data with the house data to link residential postal codes with nearby schools and to record a house's distance to a school and the rating of its nearby schools.

VI. Data Analysis

A. Importance of Features

We compared seven features to find the correlation between each feature and price of the property. Seven features include: numbers of bedrooms (beds), numbers of bathrooms (baths), area of the property in square feet (sqft), first three digits of

postal code (areacode), type of the property (housetype), ranking of the best school in catchment (schoolranking), and distance from the school and property (schoolDistance). This importance computation is based on Shapiro-Wilk test, which evaluates normal distribution, and reject the null hypothesis. To visualize statistical significance of the features, we used Yellowbrick library.

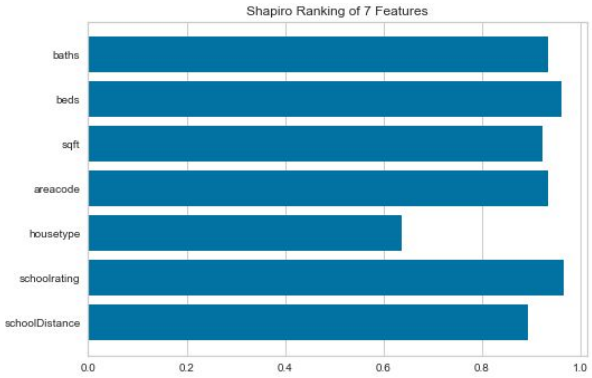

Fig 1. Shapiro Ranking of 7 Features

Figure 1 visualizes importances of each features in determining price of a property. Among the seven features, school ranking is the most important feature that determines the price of a property. Size and location also turned out to highly impact price of a property, as expected. From this, we concluded that we need to perform deeper analysis on the impact that the school rating and distance has on the house prices.

B. Nearby Secondary Schools

A family may be interested in moving into a city but is unsure if investing in an expensive house next to a good ranking school is better than buying a cheaper house next to a lower ranking school. We analyzed the 10-year fluctuation of house prices depending on the rating of schools in its catchment. New, fresh houses are naturally worth more than older houses so the number of new houses being built can be a confounding factor. Thus, we only tracked the fluctuations of houses that existed for the whole ten years and remove the new houses. Our preliminary analysis is in part C and our final results are elaborated later on in the paper.

C. Overall Analysis by Area

We started off by analyzing which area had the most price increase in past the ten years. Figure 2 shows average percentage of price fluctuation by area. Each color bar represents the relative percentage based on the government assessment prices of 1, 2, 3, 5 or 10 years ago. Area code V6N showed significantly large increase rate, as expected because this area is Dunbar Southlands, with good school catchment and large building size (with less condos and more LAND type properties). Other areas listed in top 5 price fluctuations were located in downtown Vancouver(V6E, V6G), East Fairview / South Cambie (V5Z), and North Grandview - Woodlands (V5L).

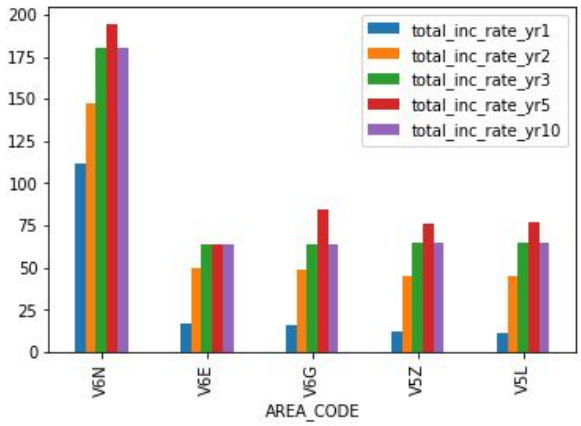

Fig 2. Price Fluctuation Percentage by Area

From the above analysis, we can see that a property near a good school with large building size would have the largest price increase.

VII. Data Product

A. Web Interface

If users are interested in a city but would like to select an area with less price gap between listed and government assessed price, they can use our program. Boomerang has compared assessed value with the listed price and used bootstrap draw out their significance. The users can also see the amount of variance in price in different parts of the same city. We have implemented simple cloud based clusters.

Boomerang has a web interface that is based on a 16.04 Ubuntu instance under the Google cloud platform. It has added firewall rules so that any user can access port 7000 which we configured the web to run on. Files are transferred using buckets. Static resources (images, csv files) are also uploaded on the cloud.

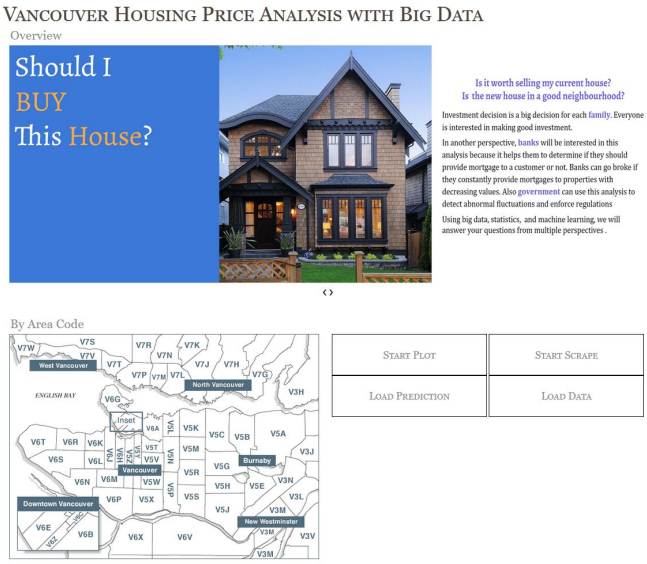

Fig 3. Initial web page

Figure 3 is capture of our main page at http://35.190.157.131:7000/. On the main site, we displayed a slideshow that contains brief introduction and analysis results of our project. Also, users can get an overview of a postal area by clicking on *Load Data* button.

Figure 4 shows one of the postal area analysis page from the *Load Data* button. The overview targets all interest groups to give a general idea of each area in Greater Vancouver. Each area is grouped by the first three digits of postal code. By aggregating our data by area, we can display property value fluctuation, different types of properties, and nearby school ratings. The bar graph on the right compares the average property value that is assessed by the government with the listing value. Bootstrap quantifies the uncertainty of property values. The bar chart at the right side of Figure 4 shows that the variance is greater for the listing price than the government assessed value

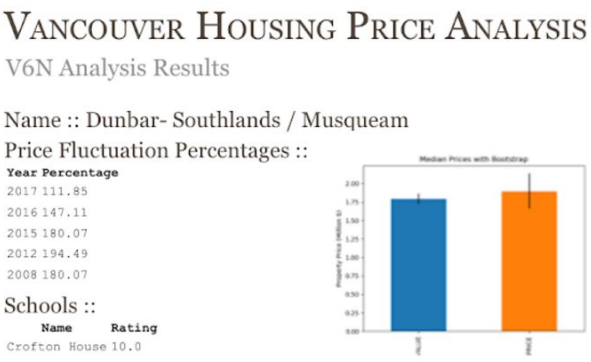

Fig 4. Area overview

If the user is interested in certain attributes of a house such as the house type and location, she can click *Load Prediction*. We used Tensorflow's Deep Neural-Network Regressor (DNNRegressor) to predict the price based on house features. By having deep networks where each layer of nodes trains on a distinct set of features based on the previous layer output, our program can flexibly represent features into a price. On top of that, we integrated 10-Year Property tax report with a realtor site to display price fluctuations over a timeline as shown in figure 6. Please note that historical information was only available for Vancouver and Surrey. On top of predicting the future and past value of a house, Boomerang also crawls a realtor site in real time to show current listings that may interest the users as shown in figure 7. This page provides Boomerang's fundamental goal of swifting from past (figure 6), present (figure 7) and future (figure 5) to deliver analysis back to user's fingertips.

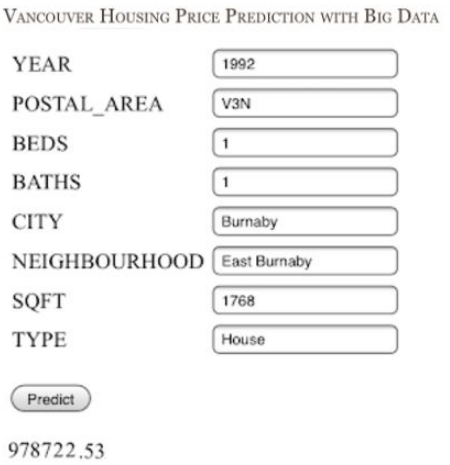

Fig 5. Evolution of property values

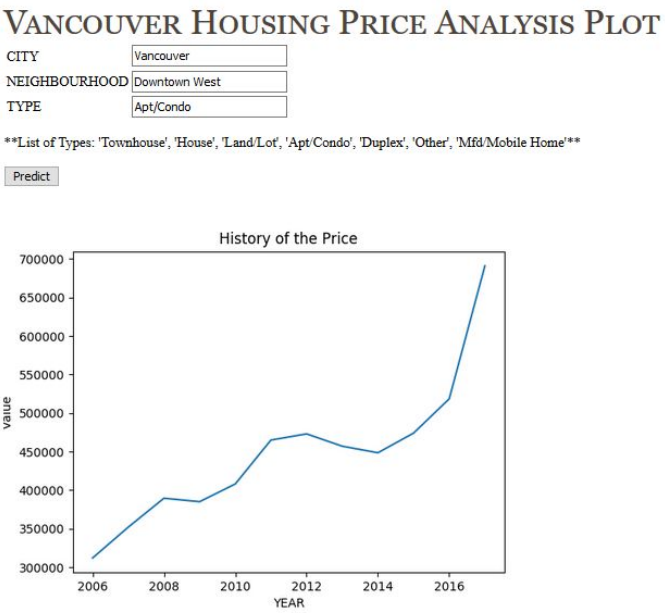

Fig 6. Historical Price Analysis by Neighbourhood and Type of Property

VANCOUVER HOUSING PRICE SCRAPING TOOL

CITY Burnaby
NEIGHBOURHOOD East Burnaby
TYPE House

List of Types: 'Townhouse', 'House', 'Land/Lot', 'Apt/Condo', 'Duplex', 'Other', 'Mfd/Mobile Home'

Start

	ADDRESS	BATHS	BEDS	PRICE	SQFT	TYPE	YEAR
0	8142 W 15th Avenue, Burnaby, BC, V3N 1X5	2.0	3.0	1388000.0	1807.0	House	1954
1	8059 16th Avenue, Burnaby, BC, V3N 1R6	4.0	5.0	1880000.0	2735.0	House	2018
2	7771 19th Avenue, Burnaby, BC, V3N 1E8	4.0	6.0	1850000.0	2956.0	House	1991
3	7767 19th Avenue, Burnaby, BC, V3N 1E8	4.0	6.0	1850000.0	3013.0	House	1991
4	7763 19th Avenue, Burnaby, BC, V3N 1E8	4.0	6.0	1850000.0	3013.0	House	1991
5	7761 19th Avenue, Burnaby, BC, V3N 1E8	5.0	7.0	2035000.0	3353.0	House	1991
6	7469 2nd Street, Burnaby, BC, V3N 3R3	4.0	5.0	1500000.0	2838.0	House	1995
7	8025 16th Avenue, Burnaby, BC, V3N 1R6	5.0	5.0	1848000.0	2716.0	House	2018
8	7744 18th Avenue, Burnaby, BC, V3N 1J2	6.0	6.0	1888000.0	3925.0	House	1998

Fig 7. Real Time Scraping Results

B. Influence of nearby Secondary Schools

The interactive app is great for people who know what they are looking for, but it may not be best for people who just want to get an overall view of different cities, specifically Vancouver and Surrey. Using Vancouver and Surrey open data, we represented the impact of the distance to and the rating of nearby schools.

Vancouver house value were more sensitive to nearby secondary school ranking than that of Surrey. We grouped each city's houses to three sections: houses nearby independent, private schools, houses in the catchment of higher-than-average rating public school, and houses in the catchment of lower-than-average or average rating public schools. Since we compared two cities, we tracked down the median house value of six different house groups in total. We only tracked down houses that existed since 2007 since we wanted the price evolution, not the price change due to new houses being built. The figure 8 shows the property value increase from last year in percentage, and figure 9 displays the property value of that year in million dollars.

Figure 8 shows that houses nearby Vancouver independent schools and higher rating public schools increased by very similar rates. These rates were more aggressively than that of Vancouver houses in the catchment of average and lower than average rating schools. When the market was good in 2012, the values spiked up. When the market dropped at 2013-2014, the values actually dropped negatively. This contrasted with Vancouver houses in the catchment of average and lower than average rating schools where the values did not plunge so negatively when the market dropped. In 2007, the value difference between houses near higher rating and private schools only differed to that near lower rating school by two times, but in 2017, they differed by three times. Figure 9 shows that the value difference went from roughly 0.5 million dollars to 2 million dollars. We conclude that it is worth investing in houses next to higher rating school if the buyer is not going to sell it when the market plummets. It may be the best for the buyer's investment and her child.

On the other hand, Surrey houses did not have significant value increase differences between houses near private, higher rating public, and lower rating public schools. Figure 9 shows that houses in the catchment of higher rating school was consistently more expensive than the rest over the ten years because it may imply that the children in the house will get better education. Houses near private schools came next. Even though residents do not have to be near private schools in order to go there, people still valued living nearby them. Vancouver also had the same reaction to private schools as Surrey. Surrey houses in the catchment of lower rating schools consistently had the lowest value. The value increase also seemed slightly less than the rest as well. In conclusion, if a new buyer is wondering where to buy a house in Surrey, then he should choose what is best for the student because the house investment difference is not extreme as Vancouver.

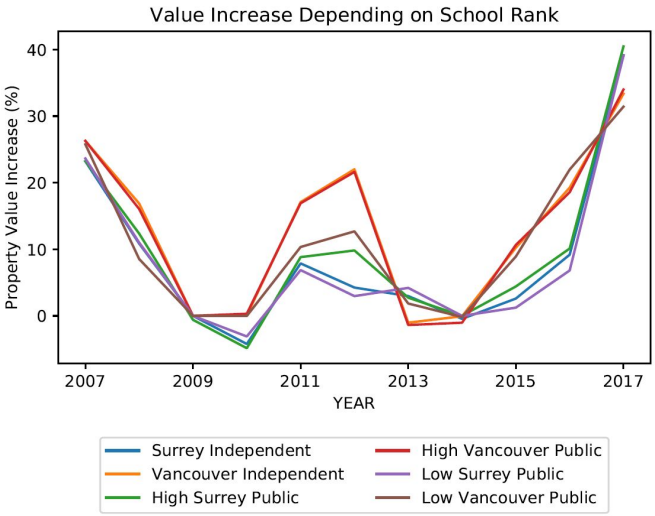

Fig 8. Rate of property value depending on nearby school ranking

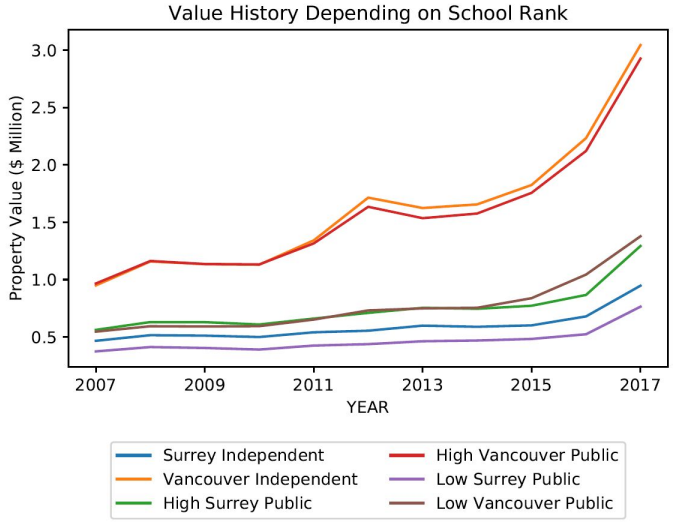

Fig 9. Property value depending on nearby school ranking

Using P-value testing, we found out that when a property is under the catchment of a school with a higher than average rating, the price difference between its assessed value and

listing value is significantly lower than that near a lower rating school. This is true for Vancouver and Surrey houses. Banks will be interested in this finding because even though the government assessed value is the more stable description of the house, clients asking for mortgage would have to borrow money to pay the listing price. With our finding, a bank can have an overall perception of the relationship between house prices and nearby school ratings and react appropriately when giving mortgages.

The correlation between distance to a school and house price varied significantly across Vancouver and Surrey. If an area has a negative correlation that means that as the distance to a school decreased, the price of the house increased. In other words, closer the house is to a school, the more expensive it is. Some areas had high positive correlation, which may mean that the neighbors do not value the school and would rather stay distant from the school. The government can look at the map in figure 10 and pinpoint which schools do not have a good impression to the neighbors and find a way to reverse that. If a house buyer has an area in Surrey in mind, he can look at this map easily visualize how the house value will change if he sacrifices some distance between the school and the house. A house seller can look at this map to understand how competitive he may be compared to his neighbors.

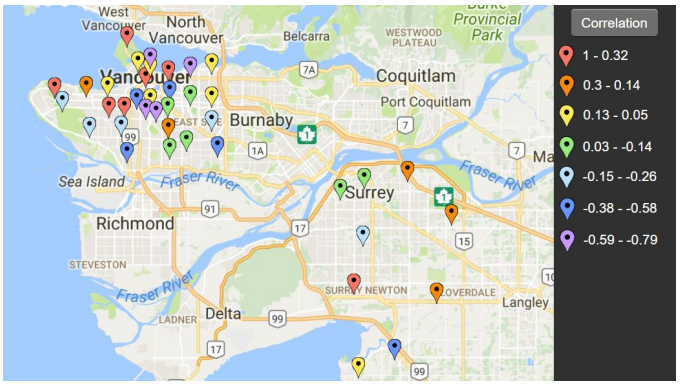

Fig 10. Correlation between distance to public school and house price

VIII. Lessons Learnt

Data is money and power. A dataset alone is poor especially in a volatile market. We had to search through many data sources, explore their features, and integrate them to answer vague questions.

Google Cloud compute machine is not only powerful but also very usable. It provides easy access to virtual machines via custom console user interface. Configurations are easily understandable and modifiable with descriptive official guides on the web. We could finish web deployment in a reasonable amount of time even though it was our first time using Google Cloud services.

IX. Summary

By web scraping, we have collected realistic, up-to-data data. By referring to municipal open data, we obtained historical house values. We then merged data from multiple sources, and used machine learning, statistics, and analytics skills to assess the value of each house and area. We displayed our findings on the web using Google Cloud Services. It includes a prediction tool to estimate a property's future price.