**##Final Project Name: Data Analysis of Vancouver Land Property**

**Instructor: Dr. Maher Elshakankiri**

Group Name: Group 10

Group Member: Xiaojing Liu, ShiFan Sun, Ruochen Zhao

**##Table of Content**

Project Description

Guideline to run the Program

Program Requirement

Part 1: Descriptive Analysis

Part 2: Diagnostic Analysis

Part 3: Predictive Analysis

**## Project Description**

This project is a Python program that reads the Vancouver Property Assessment dataset from a file into a Google Colab document and performs comprehensive data analysis, including descriptive analysis, diagnostic analysis and predictive analysis.

**## Guideline to run the program**

Prepare the dataset

FilePath for Descriptive Analysis: /content/drive/My Drive/INF1340/Sources/property.csv

FilePath for Diagnostic Analysis:

'/drive/MyDrive/inf1340/property.csv'

FilePath for Predictive Analysis: /content/drive/My Drive/INF1340/Sources/property.csv

Install the necessary libraries, including pandas, numpy, matplotlib.pyplot, Seaborn, sklearn.model\_selection, statsmodels.api, sklearn.linear\_model, sklearn.metrics, sklearn.naive\_bayes

Set the working directory and adjust the working directory path in the first cell

Run each cell sequentially to execute the analysis steps by clicking the top right button ‘run’

Review the output, including numbers and diagrams

**## Program Requirement**

Users can access the colabortory or install Python.3x on the computer.

Google Colab file path:

<https://colab.research.google.com/drive/1jAjpSQBJZUMJl4xRrNApcAR3bzD4sOUh?usp=sharing>

**Part 1: Descriptive Analysis**

**1.1 Descriptive Analysis Explanation**

The descriptive analysis includes the analysis of both continuous variables and categorical variables. The first part analyses two key continuous variables of Current Land Value and Tax Levy. The analysis includes the measure of tendency and variability. The second part is the Frequency analysis of the two key categorical variables, legal type and zone classification. The third part examines the tax rate by legal type and the current land value growth rate by zone classification. These descriptive analyses help understand the features of the key variables through the measure of data and frequency analysis. The descriptive analysis also helps to identify the patterns and relationships between variables. This analysis will enable the audience to understand what happened in the past for the property values and provide the direction for the following diagnostic and predictive analyses.

***Analysis Breakdown***

1) Continuous variables: Current Land Value and Tax Levy

· Measures of Tendency:

Calculate measures such as mean, median, and mode to understand central tendencies.

· Measures of Variability:

Computation of measures like standard deviation and interquartile range to assess data spread.

· Visualization:

Creation of histograms and box plots illustrating the distribution and variability of these variables.

2) Frequency Analysis of Categorical Variables

· Frequency Distribution:

Examination of the occurrence and distribution of different legal types and zone classifications.

· Visualization:

Creation of Pie chart illustrating the distribution legal type and Zone classification variables.

3) Tax Rate by Legal Type and Land Value Growth Rate by Zone Classification

· Tax Rate by Legal Type:

Investigating the variation in tax rates among different legal types.

Assists in understanding why specific legal types might have higher tax rates.

· Current Land Value Growth Rate by Zone Classification:

Examining the growth rate of land values based on zone classifications.

Helps identify which zone classifications experience higher land value growth rates.

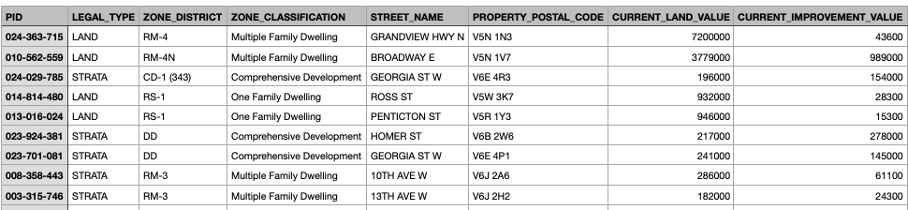
· Visualization:

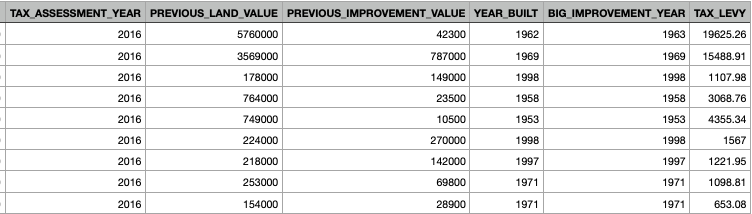
Creation of bar chart to compare average tax rate among different legal types and current value growth rate among different zone classification.

**1.2 Sample Input and Sample Output**

Sample input: property.csv

```





```

Sample output

```

Current Land Value Mean: 1185929.90

Current Land Value Median: 742000.00

Current Land Value Mode: 1020000.00

```

```

**Part 2: Diagnostic Analysis**

**2.1 Diagnostic Analysis Explanation**

This analysis methodically examines property value trends by categorizing data based on legal types, outlier occurrences, regional variations, and property types. It starts with a comparative analysis by legal type, using line charts, bar graphs, and histograms to identify and visualize trends and data distributions across different legal categories. The analysis then shifts focus to outliers, employing box plots to detect and scrutinize extreme values in land and improvement costs, aiming to discern whether these anomalies stem from data inaccuracies or unique property characteristics. A regional comparison follows, leveraging location data such as ZIP codes or districts to compare property value trends across different areas, thereby uncovering potential regional disparities in property value changes. Finally, the analysis considers various property types, including residential, commercial, and industrial, to understand their respective appreciation or depreciation trends over time. This comprehensive approach offers a multi-faceted view of the factors influencing property values, facilitating a deeper understanding of market dynamics and trends.

**Analysis Breakdown**

1. Comparative Analysis by Legal Type:

* Objective: To investigate whether there are distinct trends among various legal types of properties.
* Methodology: The data is categorized by legal type, followed by a comparative analysis of trends using line charts or bar graphs.
* Visualization Approach: Histograms are created for each variable to illustrate the distribution of data. This aids in visualizing the frequency of data points within specified value ranges, thereby facilitating an understanding of trends across different legal types.

1.1. Two-Sample T-Test Analysis on Legal Types:

* Objective: The objective is to analyze the differences in 'Current Land Value' between properties classified as 'Strata' and those classified as 'Land'. This analysis aims to understand how the legal type of a property influences its land value, providing insights into the valuation disparities within the real estate market.
* Analytical Approach: The approach involves segregating properties into two groups based on their legal classification—'Strata' and 'Land'. A two-sample t-test will be conducted to compare the means of 'Current Land Value' for these groups. This statistical test will determine if there is a significant difference in the average land values between 'Strata' and 'Land' properties. The outcome of this analysis is crucial for understanding how legal type classification impacts property valuation, guiding stakeholders in making informed decisions in the real estate sector.

2. Analysis of Outliers:

* Objective: To scrutinize extreme values in land and improvement values.
* Methodology: Box plots are employed to identify outliers, which are then examined in detail to ascertain their origins.
* Analytical Focus: The analysis seeks to determine whether outliers are attributable to data entry errors, exceptional property characteristics, or other underlying factors.

3. Regional Comparison:

* Objective: To compare property value trends across different regions, leveraging location data such as ZIP codes, districts, or neighborhoods.
* Methodology: The methodology involves grouping the data based on a location identifier and then aggregating it to ensure that all entries within the same location are collated. Subsequently, the mean property values for each location group are calculated.
* Analytical Outcome: This comparison may reveal regional disparities in property value changes, offering insights into spatial variations in property market dynamics.

4. Property Type Analysis:

* Objective: To analyze trends for each property type, including residential, commercial, and industrial, and to understand how these property types have appreciated or depreciated over time.
* Analytical Approach: The analysis involves categorizing properties based on their type and examining the trends in their values over time. This assessment aims to uncover variations in market behavior and valuation across different types of properties.

4.1. Zone Classification Impact Analysis:

* Objective: To investigate the impact of zone classifications, such as residential, commercial, and industrial, on mean current land values. The goal is to understand how different zoning categories influence property valuation, and whether certain types of zones are associated with higher or lower land values.
* Analytical Approach: The analysis entails categorizing properties based on their zone classification and then examining the mean current land values for each category. An ANOVA test is utilized to determine whether there are statistically significant differences in land values across these zoning categories. This approach aims to highlight variations in property valuation that are attributable to zoning, offering insights into how different types of areas (residential, commercial, industrial, etc.) are valued in the real estate market. The outcome of this analysis is crucial for understanding zoning's role in shaping property values and market dynamics.

**2.2 Sample Input and Sample Output**

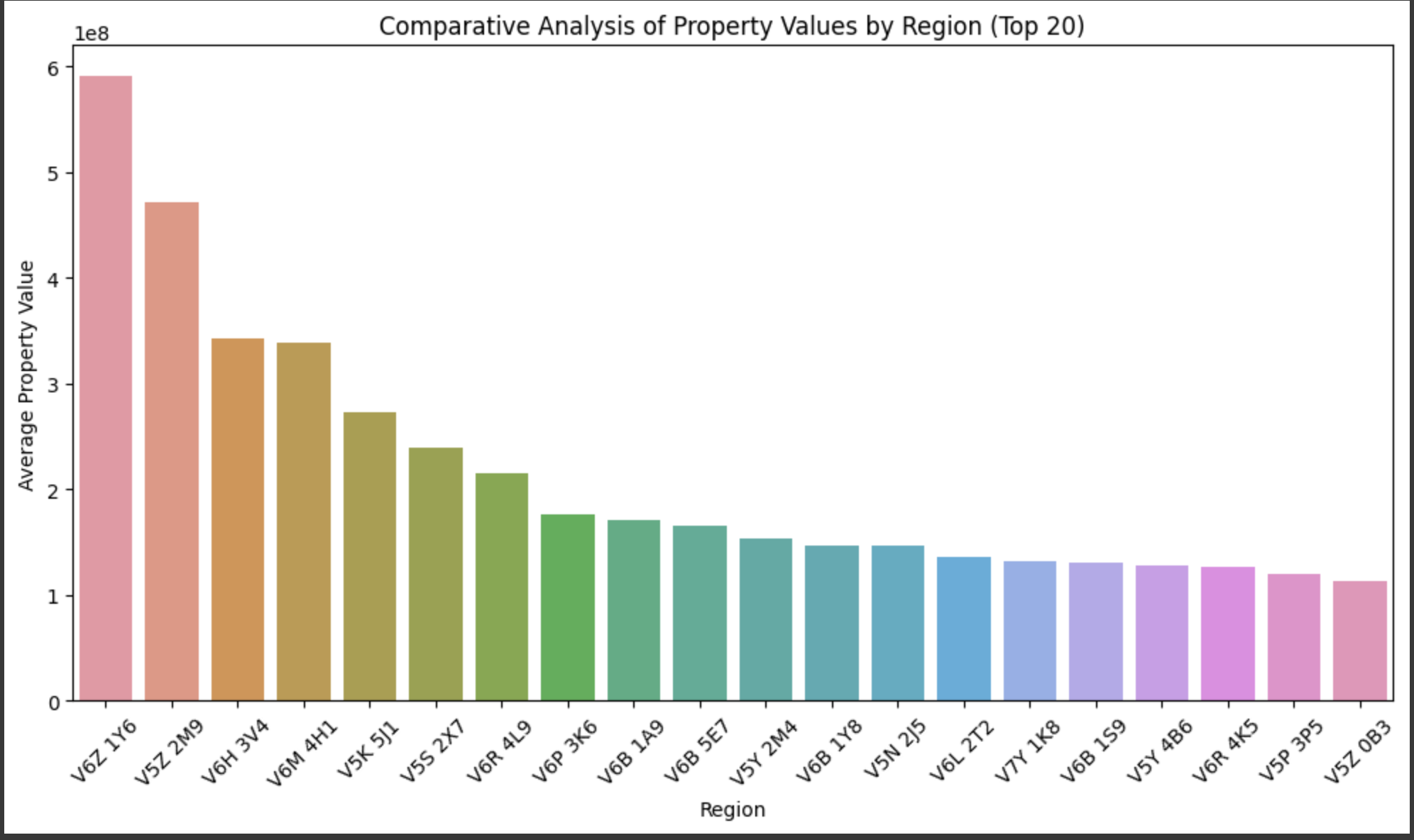
Sample input: property.csv

Sample output:

```

TtestResult(statistic=-92.33089216178084, pvalue=0.0, df=87800.12580237564)

```



**Part 3: Predictive Analysis**

**3.1 Predictive Analysis Explanation**

Predictive Analytics: In this part three methods (linear regression, logistic regression and Naive Bayes Classifier) will be performed to analyse and predict the housing price as well as the legal type for properties. The details are as follows

* The first method is to use multiple linear regression to predict the properties’ current market value. In this task we examined the relationship between three independent variables (current land value, previous land value and previous improvement value) and the dependent variable (current improvement value), hoping to see a clear linear link. However, the previous improvement value is the only variable who has a coefficient significantly different from zero. So it is used as the independent variable to form a linear relationship, and then datasets are splitted into train sets for prediction and test sets to check the model’s accuracy.
* The second method is to use logistic regression to predict housing price fluctuation. In this task we use the first three postal codes (binary), housing ages, tax levy, previous land value and current land value to predict the change of improvement price. We start by calculating the difference in current improvement price and current land value, and we assign 1 for a price increase and 0 for price decrease. Then we also transfer the first 3 postal codes into binary variables for simplification, and a logistic regression is performed and predicted to see if properties with different patterns have an increase or decrease in improvement value.
* The third method is to use a supervised machine learning method, as we already know the labels (its legal type) for each observation, the Naive Bayes Classifier to model and predict properties’ legal type (STRATA or LAND) and test for its accuracy. After loading the data and performing data cleaning, we split the zone district and zone letter column and perform the Naive Bayes Classifier to make a prediction.

Then all the analysis results (including data visualization) will be included in the output file, and these results will also be displayed on the screen.

**Analysis Breakdown**

1) Using linear regression to predict properties’ current market value

* Executing the 1st chunk can load the input data file into the Google Colab; it also include data cleaning process to select the properties with year built range from 2006 to 2014, drop the tax assessment year column as it is not related to our analysis, transfer year related column into integers and create a new column called ‘YEAR\_BUILT’ to calculate the ages of properties.
* Executing the 2nd chunk will show the data visualization of each numerical features’ relationship with current improvement value, from the graph it seems that only previous improvement value has a clear linear relationship with the current improvement value, but we need further evidence to confirm that.
* Executing the 3rd and 4th chunk will perform an initial OLS regression and the 5th chunk shows a summary of the test. We can see from the test that the coefficient for previous improvement value is significantly different from zero, which supports our guess.
* Executing the 6th and 7th chunk create a simple linear regression model with previous improvement value as independent variable, and then we train the model with the splitted training data set.
* Executing the 8th, 9th and 10th chunk show the coefficient and intercept of the linear regression model, and then we test the model with the splitted testing dataset to see its accuracy, the plots for expected value (from test dataset) and predicted value using log transformation (to see the results more clearly) show that most of the dots fit a clear linear relationship, while some of the outliers still exist.
* Finally, executing the 11th chunk shows the R2 score of the model, the value 0.71 indicates that approximately 71% of the variability in the dependent variable is explained by the independent variables, which means the model captures a substantial portion of the variation in the target variable.

2) Using logistic regression to predict housing price fluctuation

* Like the first chunk in the above task, executing the 1st chunk will perform a simple data cleaning.
* Executing the 2nd chunk adds a new column and transfers it into binary variables to represent the change for improvement value, where 0 stands for a decrease in improvement value and 1 stands for an increase in improvement value.
* Executing the 3rd and 4th chunk use the split() method to separate postal codes and zone districts into 4 new columns, and then we transfer the objects from the ‘First\_Three\_Code’ column into binary variables in the 5th chunk.
* Executing the 6th, 7th, 8th and 9th chunk performs a logistic regression and it gives summary information in the 10th chunk, where we saw that all the X variables are significant. Then the 10th chunk splits X and y variables into train and test dataset, and from the 11th to 16th chunk the intercept and coefficients are given, together with the probit and histogram shows the predicted probabilities for class 1.
* Executing the 17th to 19 th chunk examines the accuracy for the model, we first set the cutoff as 0.5, so any probability above 0.5 will be considered as 1. Then the confusion matrix was performed and the defined function shows that the accuracy is 81.3%.

3) Using Naive Bayes classifier to predict properties’ legal type based on Zone district, Zone classification and Postal code

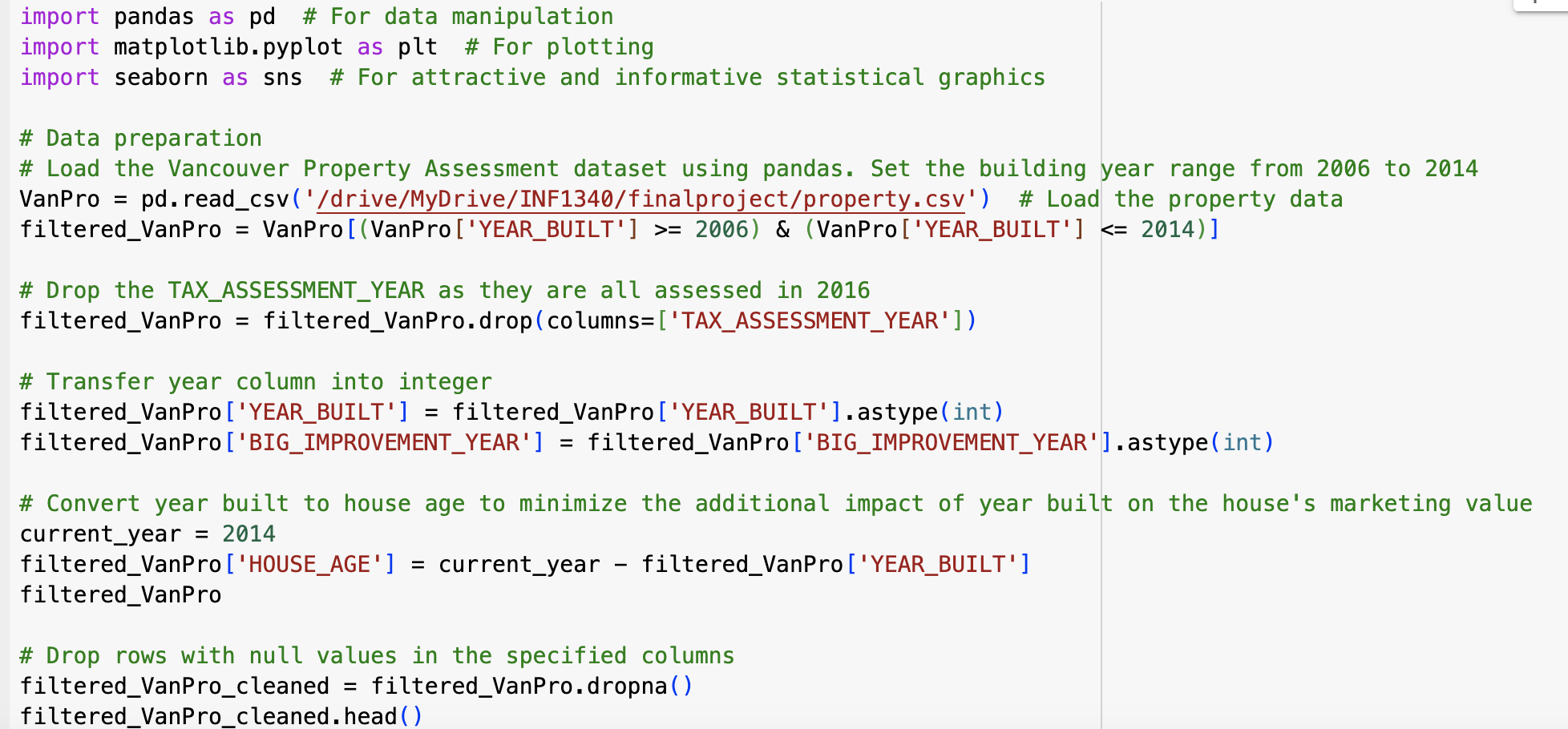
* Like the first task, executing the 1st chunk helps clean the raw data file, and executing the 2nd and 3rd chunk split the postal code and zone district column.
* The 4th chunk explains housing types and creates a new dataframe to only contain the features we want to study for the type of properties.
* Executing the 5th chunk converts categorical variables into dummies and the 6th and 7th chunk initialize the multinomial Naive Bayes classifier (for discrete data) and make predictions on the testing data.
* Executing the 8th and 9th chunk gives the accuracy of this model, where it shows an accuracy of 98% with some minor mistakes.

**3.2 Sample Input and Sample Output**

**## Example from Task 1 chunk 1:**

Sample input: property.csv

```



By click the run button, it will first load the data file into the Google Colab, and then filtered\_VanPro = VanPro[(VanPro['YEAR\_BUILT'] >= 2006) & (VanPro['YEAR\_BUILT'] <= 2014)] only keeps rows where column ‘YEAR\_BUILT’ contains value from 2006 to 2014.

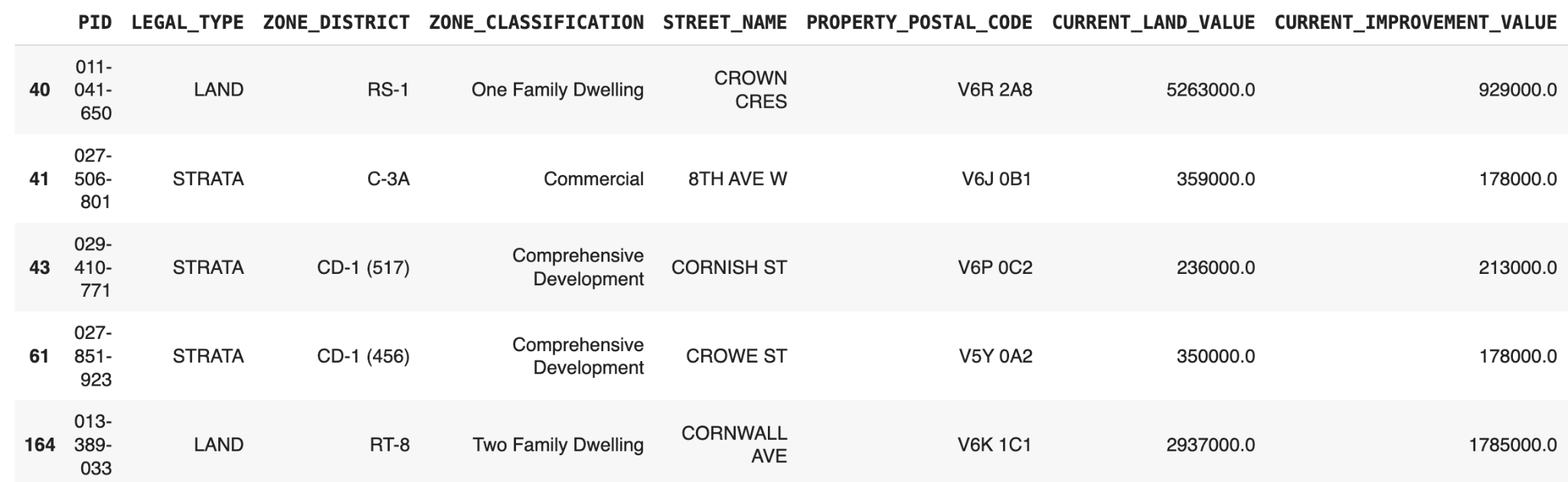
Then the column ‘TAX\_ASSESSMENT\_YEAR’ will be dropped and the year column will be transferred into integer numbers. We also add a ‘HOUSE\_AGE’ column to contain the age of our properties in the data frame.

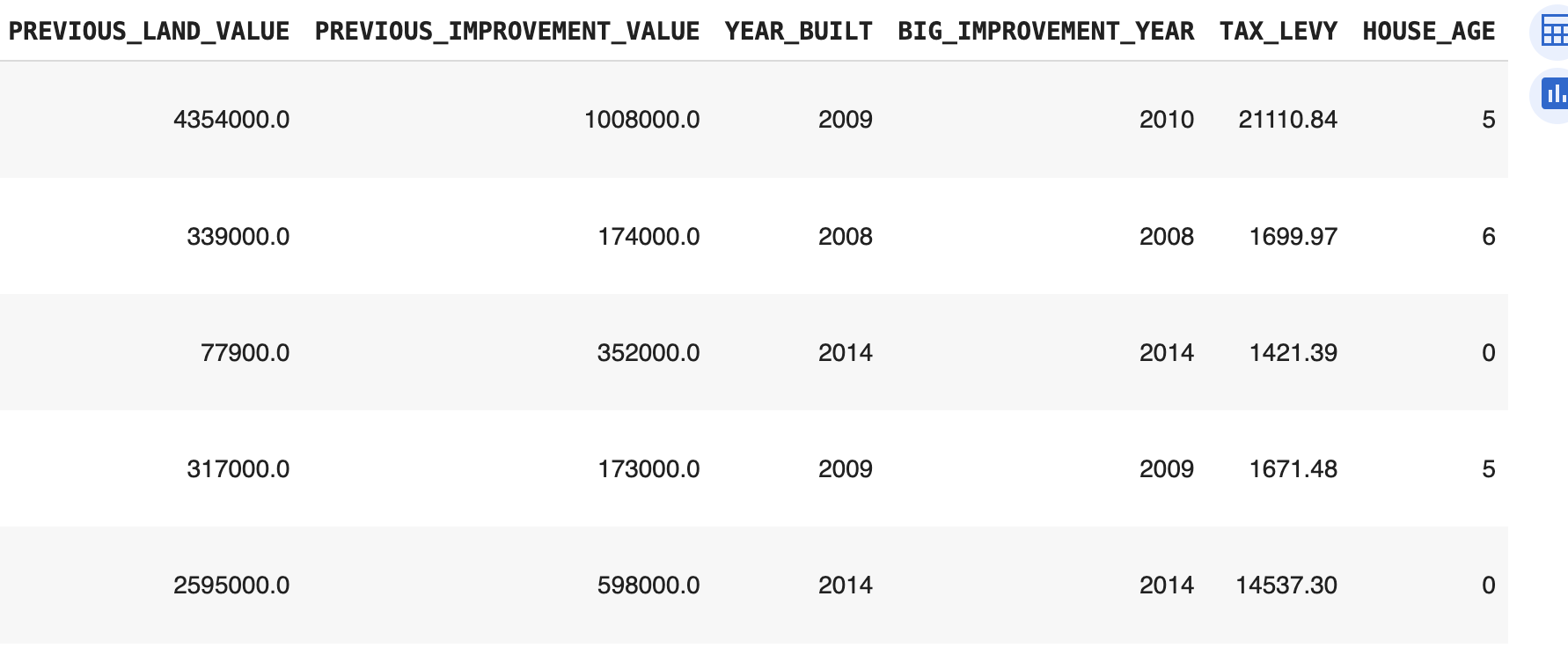
Finally we cleaned the data file by dropping NaN values and the first few rows were displayed on the screen.

```

Sample output

```





```

```