STAT 450 Project: Real Estate

Yuting Wen, 44625168

3/2/2020

Summary

Our client is a real estate data analyst in Metro Vancouver who analyzes the market trend to make a market prediction for his company and future client. He is interested in a good prediction model that can accurately predict future mill rates, given the past data like the past mill rates and the past assessment values, in Metro Vancouver. Data cleaning, exploratory data analysis, and linear models with different kinds of regulazations were used to analyze the relationship between mill rate and other factors and predict future mill rates. Data cleaning is performed to aggregate our data into summary statistics and also reduce the number of data entries. Exploratory analysis has shown that there is a strong correlation between mill rate and municipal budget and strong correlation between mill rate and assessment in some municipalities. The results of exploratory analysis have also shown that there is no clear difference among full linear Regression model with all features (Year, TaxClassCode, Municipalities, assessTotal_pct, tax (budget), landTotal, improvementTotal, numProperties) and reduced linear regression models (Ridge, Lasso and Elastic net).

Introduction

Our goal is to accurately predict the mill rate in 2020 based on past data from 2016 to 2019, for each of the following 21 sub-regions in the Greater Vancouver area:

- Burnaby
- Coquitlam
- Delta
- Langley City,
- Langley Township
- Maple Ridge
- Maple Ridge Rural
- North Vancouver City
- North Vancouver Dist
- Pitt Meadows
- Port Coquitlam
- Port Moody
- Richmond
- Surrey
- Vancouver
- White Rock
- West Vancouver
- Bowen Island
- Anmore
- Belcarra
- Lions Bay

Correlations between mill rates, assessment values, municipal budgets, land assessment values, improvement assessment values, number of properties, tax class codes, municipalities, and year are performed to see the significance of these factors in our predictive model.

A full linear model with all features, a few reduced linear models with regularzations like the LASSO, Ridge Regression are used to train our predictive model. Once we have finished fitting those models, we will use the mean squared training error to evaluate the performance of our model.

Data Description

To address these problems, we used past assessment data, which is provided by our client. We obtained information about commercial and residential properties in all cities in Metro Vancouver from 2016 to 2019. There are approximately 2.4 million data entries across these 4 years. We selected variables that we believe were relevant to the property tax prediction. These selected variables are:

- Mill Rates (2016-2019)
- Assessment Total (2016-2019)
- Land Assessment Total (2016-2019)
- Improvement Assessment Total (2016-2019)
- Number of Properties (2016-2019)
- Tax Class Code (01, 05, 06)
- Area Code (21 Municipalities)
- Year (2016-2019)

We also collected external data from the Government of British Columbia which could be associated with the tax rate:

- Municipal Budget of Cities in Metro Vancouver (Taxes Imposed & Collected, Schedule 706)

The listing data was cleaned to reduce its dimension. We decided to calculate the sum of total assessment value, total land assessment value, total improvement assessment value, budget, and number of properties for each combination of municipality and their corresponding tax class. The mill rate is unique for each combination of municipality and tax class, we have merged this information into our data for each combination of municipality and tax class.

Methods

Exploratory Analysis

We did some exploratory analysis to visualize the main characteristics of our dataset and found relationships between mill rate and other factors, before we made any prediction on the future mill rate of Metro Vancouver's real estate market.

We assume that the government aims to match its budget and its income by adjusting the mill rates, so we have plotted a graph of tax income for three different tax classes (01, 05, and 06) vs. budget through time to see the trend of tax income in different municipalities to make sure our assumption about the data is correct. Vancouver, Richmond, and Burnaby were chosen to illustrate such relationship as they are the most representing city in metro vancouver.

We also made scatterplot of mill rate vs. assessment, mil rate vs. municipal budget, mill rate vs. landTotal (total land assessment value), mill rate vs. improvementTotal (total improvement assessment value), and mill rate vs. numOfProperties (number of properties) to see the correlation between each of the two factors. ANOVA analysis was also performed to see the correlations between mill rate vs. tax class, mill rate vs. year and mill rate vs. municipality.

These exploratory analysis and plots helped us to find trends in data within each of its own categories.

Measure of goodness of fit and prediction power

In this study, we first built various models then evaluated their performances by goodness of fit and prediction power, that is, how well the model explains the data and how well it can predict future values.

Goodness of fit is defined as the extent to which the sample data are consistent with the model, which examines how well the model explains the data. R squared and adjusted R squared are the most well known measures of goodness of fit and higher R squared and adjusted R squares indicate better goodness of fit.

Prediction power measures how well models can predict the future values. In this study, we use mean squared prediction error to compare the prediction performance across all the models. To do that, we will divide the data set into training data, used as building models, and test data, used as computing mean squared prediction error.

Ordinary Linear Model

The explanatory variables we are interested in are Year, TaxClassCode, Municipalities, assessTotal, tax (budget), landTotal, improvementTotal, numOfProperties. We first explored the full linear model using all the available features; then the linear models with regularzations were performed since they potentially have better prediction power.

There are potential outliers in the data set, it might be helpful to refit the model without the outliers to check if the fit of the model is improved. Futher prediction models will be built here.

Reduced Ordinary Linear Model (Variable Selection)

A list of significant variables are included in the result of the full model. In the next step, we will refit the model with only the significant variables. Goodness of fit will decline with fewer features, however, the performance of prediction power might be improved.

Ridge, Lasso and Elastic Net

Except for ordinary linear regression, we are also interested in the performance of Ridge, Lasso, and Elastic Net models as they are also in the linear regression family, but more advanced than ordinary linear regression. In these three models, weights are assigned to features. Ridge regression and Lasso take penalty in sum of absolute values and sum of squared absolute values of weights respectively. Elastic net is a combination of Ridge and Lasso.

We build the three models using all features and compare their goodness of fit, and then will examine their prediction power on testing data.

Neural Network

In the current report, we are mainly focusing on predicting the mill rate using Linear Model. In the future, we will discuss with our supervisor and try to fit a neural network model if possible (R package: neuralnet or keras). Neural Network has the ability to learn and model non-linear and complex relationships. It is flexible and can be used for both classification and prediction. Since we have a large number of data, it makes Neural Network a very good candidate for our prediction model.

Report on the performance of the Neural Network will be included in the next report.

Results: Exploratory Analysis

We assume that the government aims to match its budget and its income by adjusting the mill rates, so we plotted the predicted tax income vs municipal budget through time into line-chart to visualize their relationship, to make sure our assumption is correct.

For each year from 2016 to 2019, we predicted the tax income of three class codes (01, 05, 06) in each municipality by multiplying access total and mill rate. Vancoouver, Richmond, and Burnaby were chosen because they were the three most representing municipalities in Metro Vancouver, and the plots showed that the predicted tax amount was positively related to municipal budget.

Figure 1: Budget and Tax Income v.s. Year of Vancouver

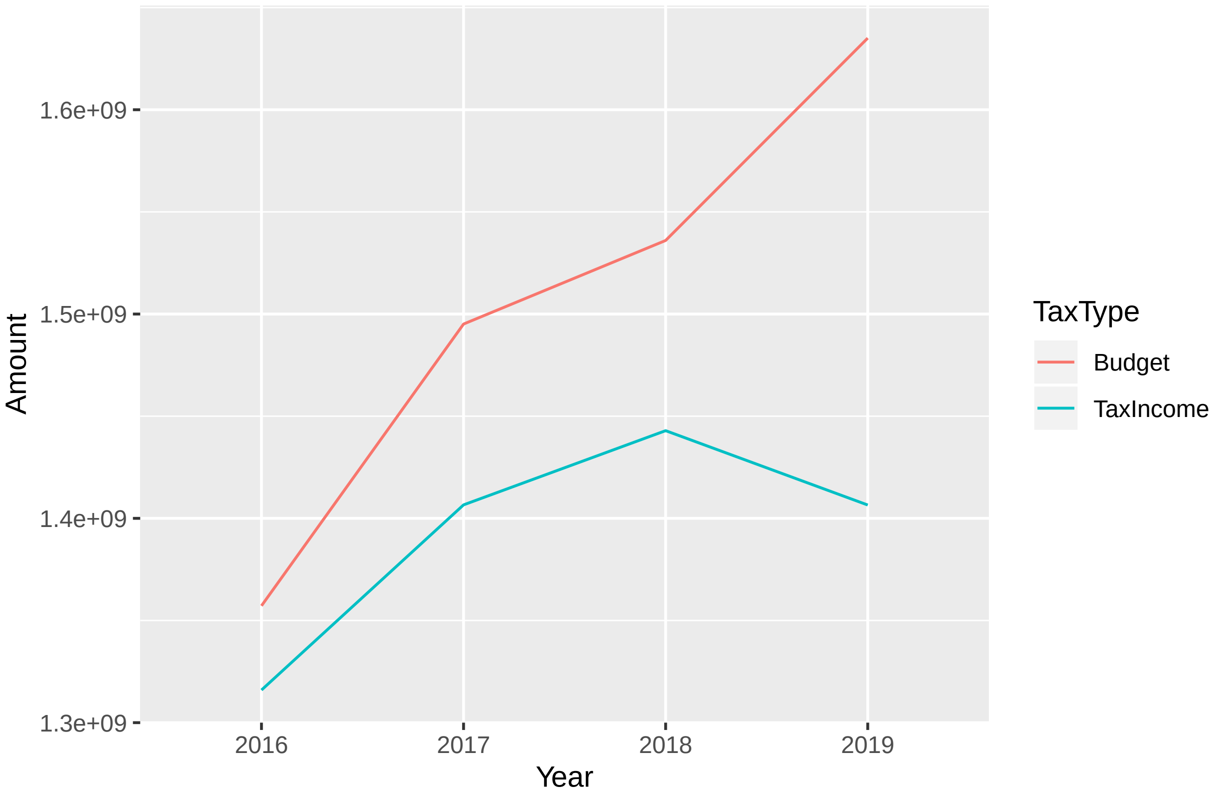


Figure 2: Budget and Tax Income v.s Year of Richmond

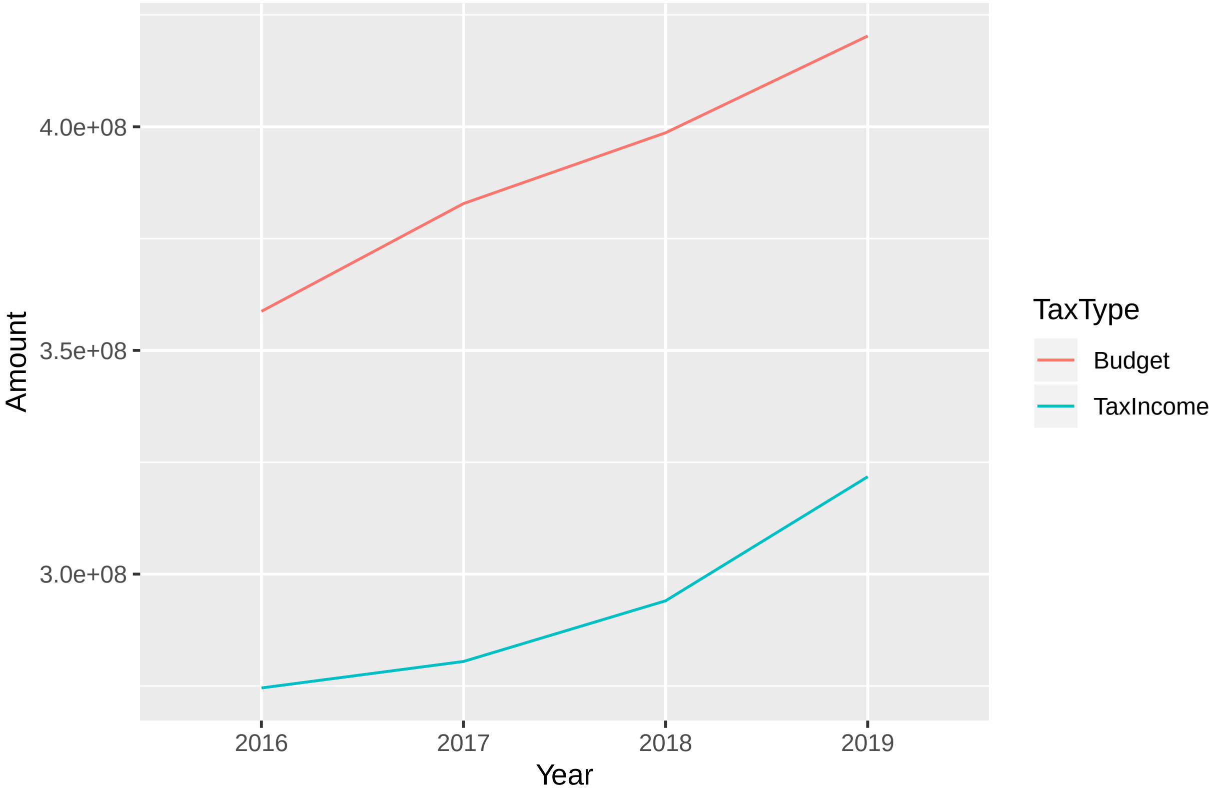
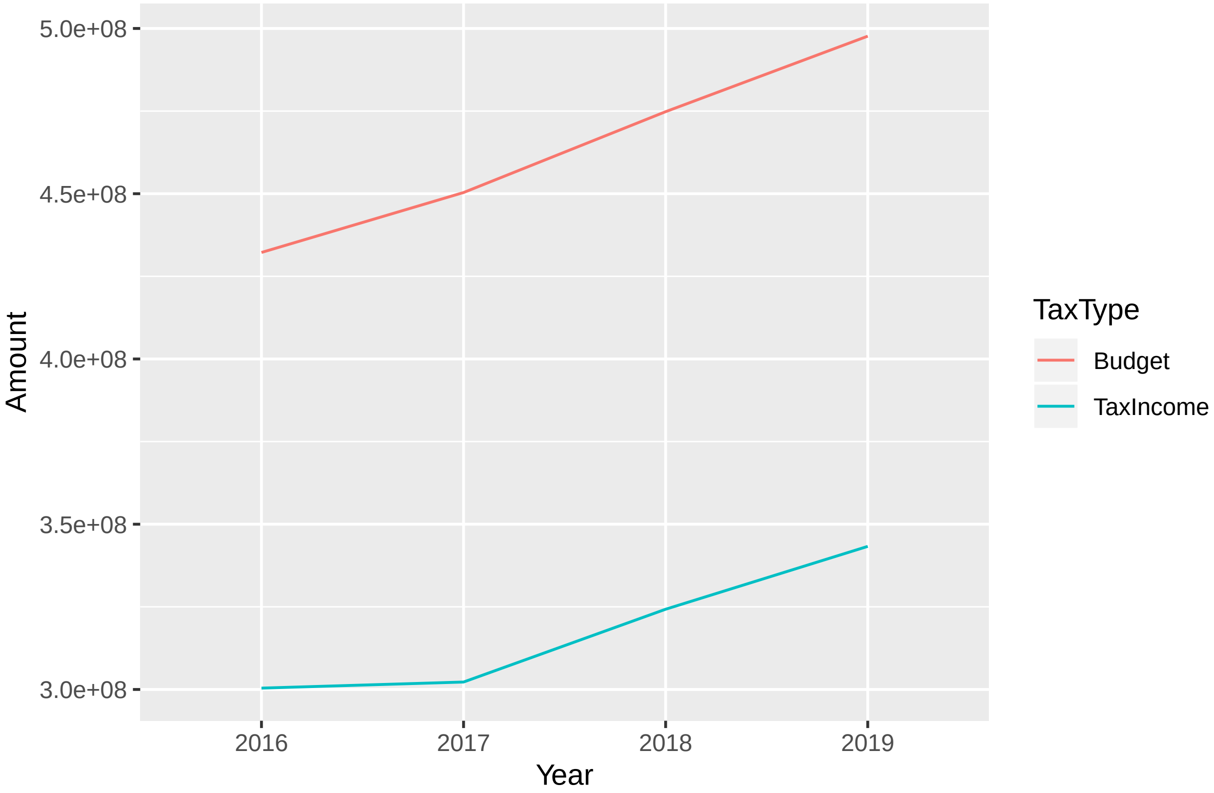
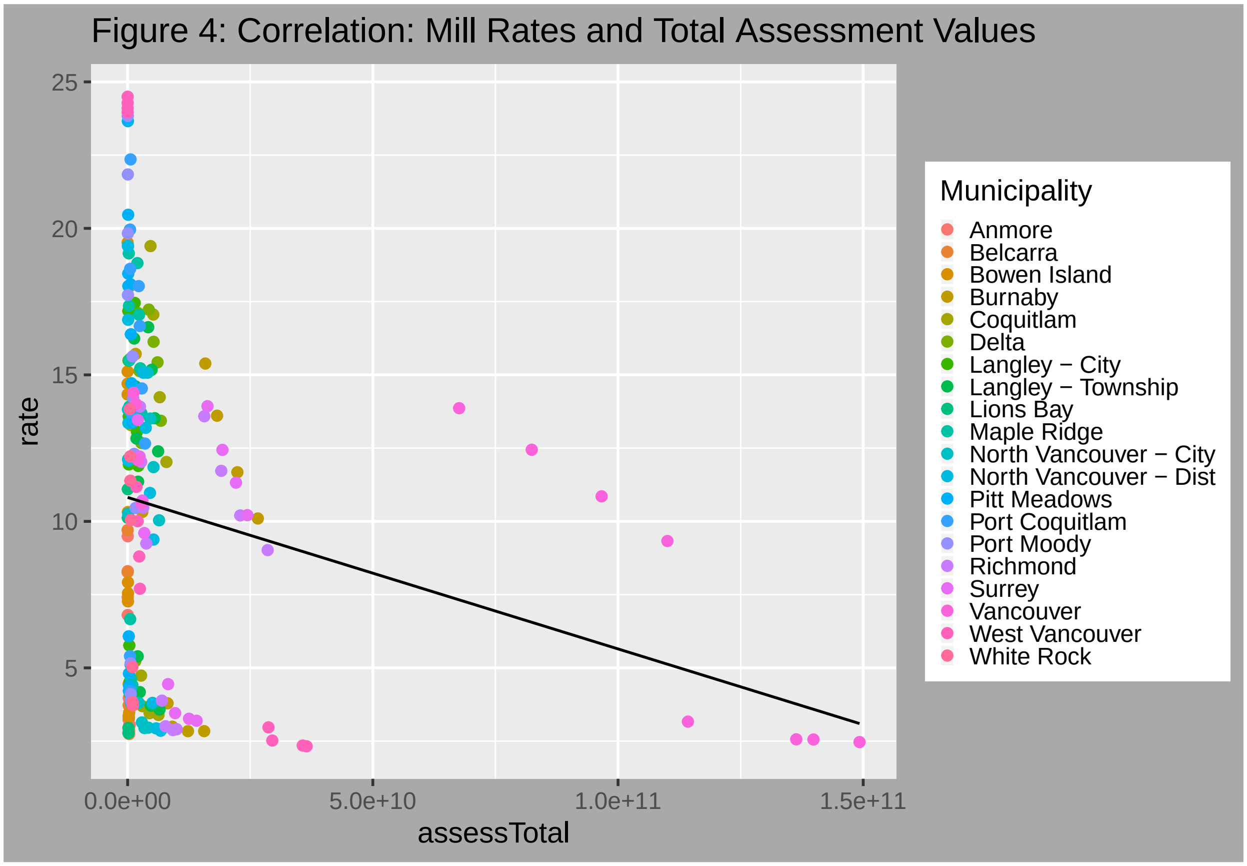


Figure 3: Budget and Tax Income v.s Year of Burnaby



Correlation Analysis

We made 5 scatter plots to visualize the correlation between mill rate and 5 numerical features (assessment total, municipal budget, land assessment total, improvement assessment total, and the number of properties).



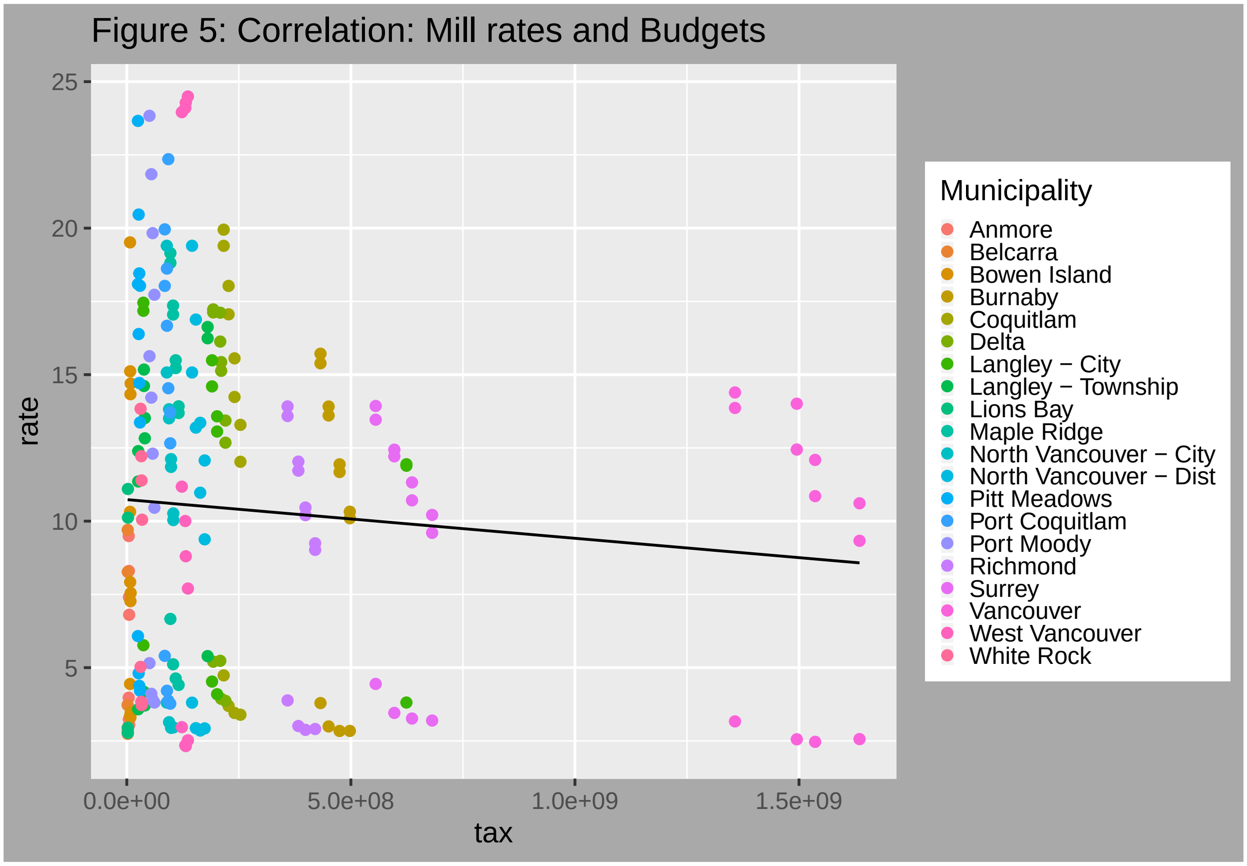


Figure 5: Correlation: Mill rates and Budgets
rate
25
20
15
10
5
0.0e+00
5.0e+08
1.0e+09
1.5e+09
tax
Municipality
Anmore
Belcarra
Bowen Island
Burnaby
Coquitlam
Delta
Langley − City
Langley − Township
Lions Bay
Maple Ridge
North Vancouver − City
North Vancouver − Dist
Pitt Meadows
Port Coquitlam
Port Moody
Richmond
Surrey
Vancouver
West Vancouver
White Rock

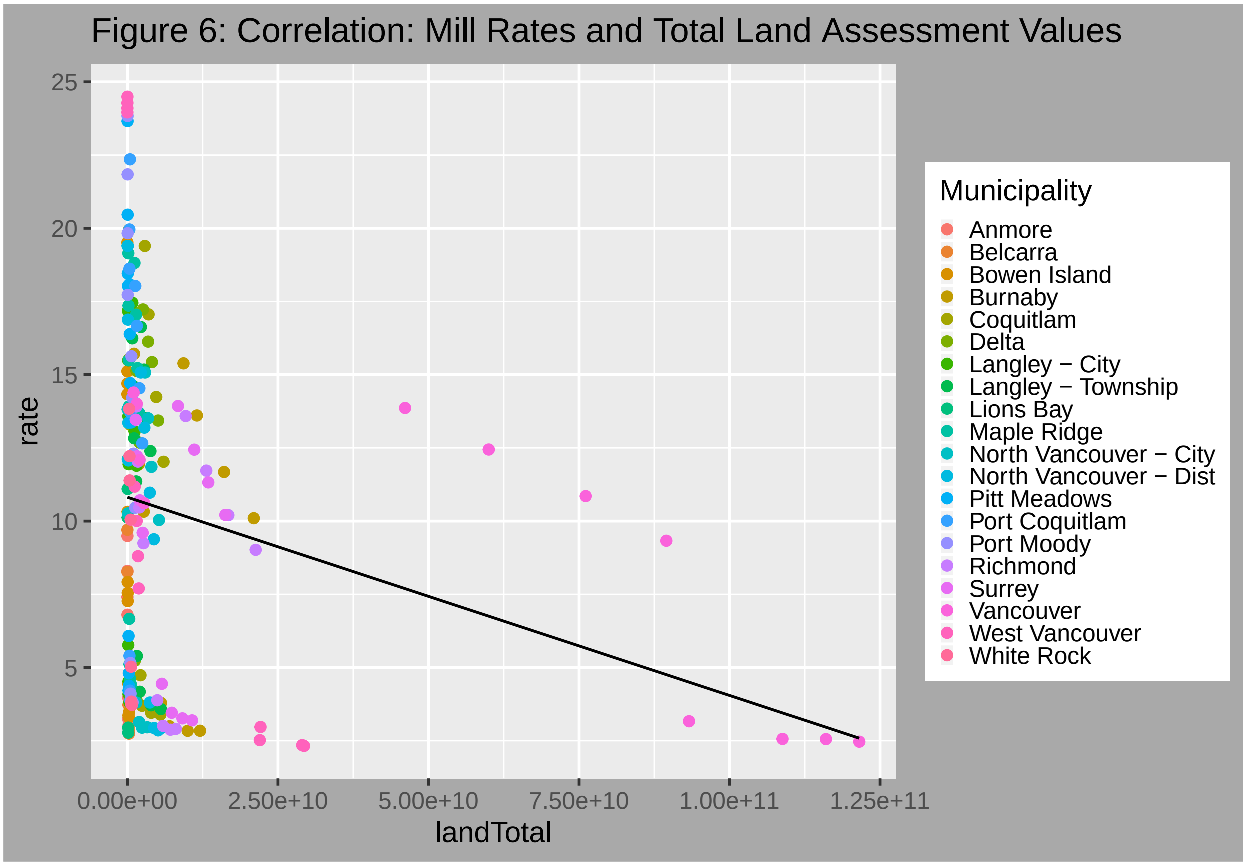


Figure 6: Correlation: Mill Rates and Total Land Assessment Values
rate
25
20
15
10
5
0.00e+00
2.50e+10
5.00e+10
7.50e+10
1.00e+11
1.25e+11
landTotal
Municipality
Anmore
Belcarra
Bowen Island
Burnaby
Coquitlam
Delta
Langley − City
Langley − Township
Lions Bay
Maple Ridge
North Vancouver − City
North Vancouver − Dist
Pitt Meadows
Port Coquitlam
Port Moody
Richmond
Surrey
Vancouver
West Vancouver
White Rock

Figure 7: Correlation: Mill Rates and Total Improvement Assessment Values

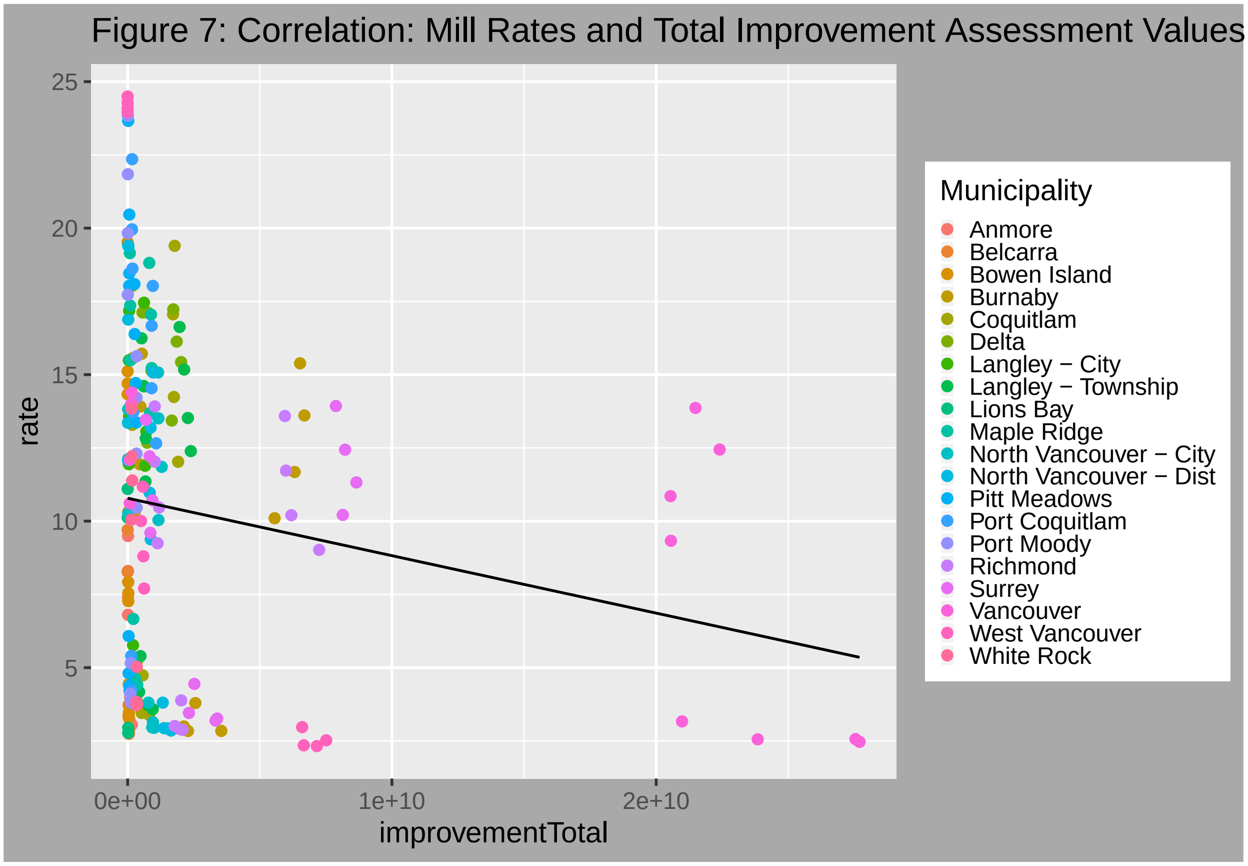
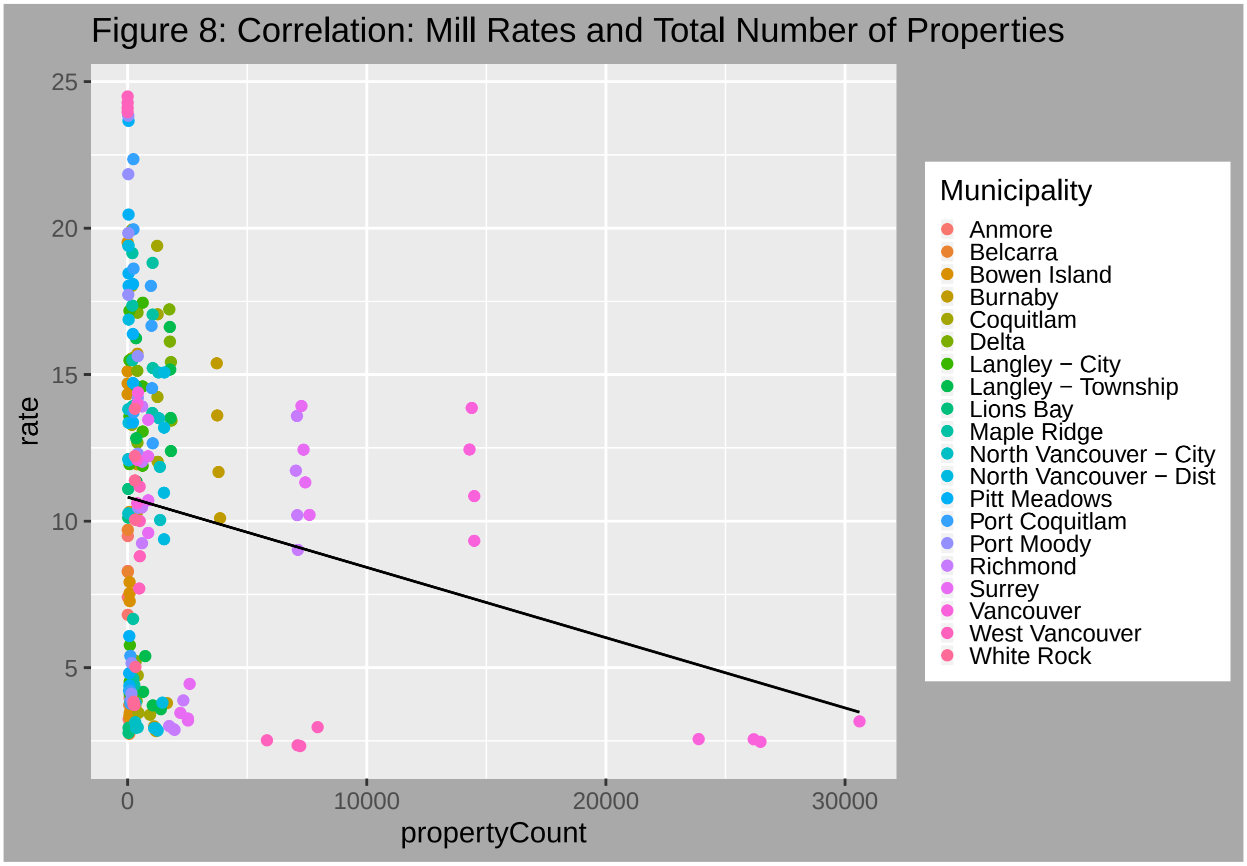


Figure 8: Correlation: Mill Rates and Total Number of Properties



As we can see:

Figure 4, 6, 7, and 8 show that mill rate is negatively correlated with total assessment values, total land assessment values, total improvement assessment values, total number of properties. Budget is not significantly correlated with mill rate as shown in Figure 5.

Outlier detection:

The plots above also show that Vancouver (the dark pink points) are significantly different from other points and should be regarded as an outlier (futher analysis needed here). Some other municipalities (including Richmond, Surrey and Burnaby) that are also distinct in the scatter plots.

ANOVA Analysis

Table 1: ANOVA analysis of mill rate and three categorical variables

Feature Name	Pr(>F)	significance levels
Tax Class Code	<2e-16	***(very significant)
Year	4.2e-06	**(significant)
Municipality	0.0768	.(not significant)

We used three 3 ANOVA tables to analyze the correlation between mill rate and each of the 3 categorical features (TaxClassCode, Year, Municipalities). The table showed that Tax Class Code and Year have a strong correlation with mill rate while the correlation between Municipality and mill rate is not significant at 5% level.

Results

Variable Selection

Table 2: Significance level of all variables of full linear model

Variable	Pr(>t)	Significance Level
(Intercept)	0.000111	***
factor(Municipality)Belcarra	0.299387	
factor(Municipality)Bowen Island	0.303157	
factor(Municipality)Burnaby	0.604683	
factor(Municipality)Coquitlam	0.001594	**
factor(Municipality)Delta	0.002579	**
factor(Municipality)Langley - City	0.020365	*
factor(Municipality)Langley - Township	0.023030	*
factor(Municipality)Lions Bay	0.091427	.
factor(Municipality)Maple Ridge	4.18e-05	***
factor(Municipality)North Vancouver - City	0.129059	
factor(Municipality)North Vancouver - Dist	0.077124	.
factor(Municipality)Pitt Meadows	2.41e-07	***
factor(Municipality)Port Coquitlam	1.60e-05	***
factor(Municipality)Port Moody	1.16e-05	***
factor(Municipality)Richmond	0.894181	
factor(Municipality)Surrey	0.984723	
factor(Municipality)Vancouver	0.885521	
factor(Municipality)West Vancouver	0.001003	**
factor(Municipality)White Rock	0.034041	*
factor(Year)2017	0.000655	***
factor(Year)2018	1.30e-08	***
factor(Year)2019	4.73e-13	***
factor(TaxClassCode)5	< 2e-16	***
factor(TaxClassCode)6	< 2e-16	***
assessTotal	0.013752	*
landTotal	0.013863	*
improvementTotal	NA	
propertyCount	0.825699	
tax	0.958323	

After fitting the ordinary linear regression model with all features, the summary statistics in Table 2 showed that municipality, year, tax class code, total assessment value, and total land assess are significant, which indicated that total improvement assessment value, number of properties, and budget(tax) might not have strong impact for the linear model. Therefore we built an ordinary linear regression reduced model with only significant variables.

Goodness of Fit

Table 3: Goodness of fit of all models

Model	Mutiple R-Squared	Adjusted R_Squared	MSE
Ordinary Linear Regression full model	0.8874	0.8707	1.9843
Ordinary Linear Regression reduced model	0.8874	0.8721	1.9845
Ridge Regression model	0.8868	NA	1.9896
LASSO Regression model	0.8873	NA	1.9855

Model	Mutiple R-Squared	Adjusted R_Squared	MSE
Elastic Net Regression model	0.8625	NA	2.2366

The table above is the goodness of fit of five models we fitted. All models have relatively high R squared and adjusted R squared values, which indicates that the above models are able to accurately fit the data. Also there is no significant differences among these models in terms of mean squared error, except for Elastic Net model which have higher MSE. We have not used cross validation to reduce model overfitting. Although the results are relatively similar and have high R_squared across all models, we believe using cross validation there will be a clearer difference between the goodness of fit of each model.

Prediction Power

Table 4: Prediction Power of all models

Model	PMSE
Ordinary Linear Regression full model	2.5902
Ordinary Linear Regression reduced model	2.5237
Ridge Regression model	2.5675
LASSO Regression model	2.5280
Elastic Net Regression model	2.5486

The table above is the prediction power of five models we fitted. All models have relatively low error (PMSE), while there is also no clear difference among them. Futher analysis is needed here.

Conclusions

From our exploratory data analysis, We have found that when comparing the correlation between mill rate against features in our models, mill rate is highly correlated with total assessment values, total land assessment values, total improvement assessment values, total number of properties, comparing with budget. Also Vancouver is an outlier in our plot. We would like to further investigate Vancouver as a special case.

We can also see that total assessment value, total land assessment value, tax class code, year and municipality are all significant in our linear regression model. We are also interested in the effect of different tax classes in predicting mill rate. Further analysis and prediction will be performed based on our hypothesis.

The linear regression models were able to accurately predict mill rate with relatively low error, but there is still a very high chance that our model is overfitted. For our next report, we are going to use cross-validation to reduce the effect of overfitting.

References

Links to source of data:

- Schedule 706 (https://www2.gov.bc.ca/gov/content/governments/local-governments/facts-framework/statistics/statistics)

Code repository:

- Data Cleaning (https://github.com/STAT450-550/RealEstate/blob/450/src/Data_Cleaning.Rmd)
- Exploratory Data Analysis and Model Fitting (https://github.com/STAT450-550/RealEstate/blob/450/src/EDA%26mode_fitting.Rmd)

Appendix

Missing Value:

There are 1801 missing values in mill rate(TaxClassTaxRate). We decided to impute these missing values based on client information; all properties in the same region, classcode, and year should have a unique class rate

- For entries with mill rate, we aggregated them into groups by region + classcode + year.
- For entries without mill rate, we found the group that they belong to and assign them mill rate in that group.

Here is some exceptions found:

Some groups' mill rate is not unique:

- In Delta, properties in different neighbourhoods have slightly different mill rates. Since the variance is not significant, we take the mean as the overall mill rate in groups.
- In Vancouver, 2019, Class 01, one property's mill rate is different from others. It is regarded as an outlier.
- In Burnaby, 2019, Class 06, six properties' mill rate are different from others. They are regarded as outliers.
- In Langley, 2019, Class 06, mill rate is different between assessment type. After talking with the client, the mill rate for assessment type "land" is regarded as the overall mill rate in that group.
- In some groups, all entries' mill rates are missing. Entries in these groups are removed. Here is the list of the groups:

Table 4: Groups that miss mill rate

Year	Region	Class	Number of Properties
2016	Belcarra	06	9
2016	Lions Bay	01	40
2016	Lions Bay	06	25
2016	Maple Ridge Rural	05	36
2017	Belcarra	06	9
2017	Lions Bay	01	39
2017	Lions Bay	06	24
2017	Maple Ridge Rural	05	36
2018	Maple Ridge Rural	05	36
2019	Maple Ridge Rural	05	38

Model Fitting

Ordinary Linear Regression Full Model

```
linear_full<-lm(rate~factor(Municipality)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+improv
#summary(linear_full)

library(broom)
linear_full_fit<-augment(linear_full)
mse_full <- sqrt(sum((linear_full_fit$.resid)^2)/nrow(assessment_aggregate))
```

Ordinary Linear Regression Reduced Model

```r
reduced<-lm(rate~factor(Year)+factor(TaxClassCode)+factor(Municipality)+assessTotal+landTotal, data=asse
#summary(reduced)

reduced_fit<-augment(reduced)
mse_reduced <- sqrt(sum((reduced_fit$.resid)^2)/nrow(assessment_aggregate))
```

Ridge

```r
library(glmnet)
library(dummies)
dummy_year<-dummy(assessment_aggregate$Year)
dummy_municipal<-dummy(assessment_aggregate$Municipality)
dummy_taxclass<-dummy(assessment_aggregate$TaxClassCode)
# build x matrix
x<-cbind(dummy_municipal,dummy_year,dummy_taxclass,assessment_aggregate$assessTotal,assessment_aggregate

y<-assessment_aggregate$rate
lambdas <- 10^seq(2, -3, by = -.1)

lambdas <- 10^seq(2, -3, by = -.1)
ridge_reg = glmnet(x, y, nlambda = 25, alpha = 0, family = 'gaussian', lambda = lambdas)
set.seed(450)
cv_ridge <- cv.glmnet(x, y, alpha = 0, lambda = lambdas, nfolds=10)
optimal_lambda <- cv_ridge$lambda.min

predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)

# Compute R^2 from true and predicted values
eval_results <- function(true, predicted) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  MSPE = sqrt(SSE/nrow(predicted))
# Model performance metrics
data.frame(
  MSPE = MSPE,
  Rsquare = R_square
)

}

predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)
ridge_mse <- eval_results(y, predictions_train)
```

LASSO

```r
# Setting alpha = 1 implements lasso regression
set.seed(450)
lasso_reg <- cv.glmnet(x, y, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 10)

# Best
lambda_best <- lasso_reg$lambda.min;#lambda_best

lasso_model <- glmnet(x, y, alpha = 1, lambda = lambda_best, standardize = TRUE)
```

```
predictions_train <- predict(lasso_model, s = lambda_best, newx = x)
lasso_mse <- eval_results(y, predictions_train)
```

Elastic Net

```
library(caret)
#tibble::as_tibble(assessment_aggregate)
cv_10 = trainControl(method = "cv", number = 10)
elastic_net = train(
  rate~factor(Municipality)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+improvementTotal+pre
  method = "glmnet",
  trControl = cv_10
)
```

Prediction Power

```
set.seed(450)
train_ind<-sample(218,218-50)
train<-assessment_aggregate[train_ind,]
test<-assessment_aggregate[-train_ind,]

# Full linear model
newx<-test[,-c(8,9)]
y<-test[,c(8)]
linear_1<-lm(rate~factor(Municipality)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+improveme
resid<-predict(linear_1,newdata = newx) - y
full_mspe <- sqrt(sum(resid^2)/nrow(test))

# Reduced model
linear_2<-lm(rate~factor(Year)+factor(TaxClassCode)+factor(Municipality)+assessTotal+landTotal,data=tra
resid<-predict(linear_2,newdata = newx) - y
reduced_mspe <- sqrt(sum(resid^2)/nrow(test))

# Lasso
# create the whole matrix
y<-as.matrix(assessment_aggregate$rate)
#dim(x) # 165  29
#dim(y)
# creat x_train matrix and y_train
x_train<-x[train_ind,]
y_train<-y[train_ind,]
# create x_test matrix
x_test<-x[-train_ind,]
y_test<-y[-train_ind,]

# Setting alpha = 1 implements lasso regression
set.seed(450)
lasso_reg <- cv.glmnet(x_train, y_train, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 10)

# Best
lambda_best <- lasso_reg$lambda.min;#lambda_best

lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_best, standardize = TRUE)
```

```r
predictions_test <- predict(lasso_model, s = lambda_best, newx = x_test)
lasso_mspe <- eval_results(y_test, predictions_test)

# Ridge
ridge_reg = glmnet(x_train, y_train, nlambda = 25, alpha = 0, family = 'gaussian', lambda = lambdas)
set.seed(450)
cv_ridge <- cv.glmnet(x_train, y_train, alpha = 0, lambda = lambdas, nfolds=10)
optimal_lambda <- cv_ridge$lambda.min
#optimal_lambda
predictions_test <- predict(ridge_reg, s = optimal_lambda, newx = x_test)
ridge_mspe <- eval_results(y_test, predictions_test)

# Elastic Net
#tibble::as_tibble(assessment_aggregate[train_ind,])
cv_10 = trainControl(method = "cv", number = 10)
elastic_net = train(
 rate~factor(Municipality)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+improvementTotal+prop
 data = assessment_aggregate[train_ind,],
  method = "glmnet",
  trControl = cv_10
)

elastic_reg = glmnet(x_train, y_train, nlambda = 25, alpha = 1, family = 'gaussian', lambda =  0.0654920
predictions_test <- predict(elastic_reg, newx = x_test)
elastic_net_mspe <- eval_results(y_test, predictions_test)
```