STAT 450 Project: Real Estate

Peter Han, 14912166

3/3/2020

Summary

Our client is a real estate analyst who analyzes the market trend and make market prediction for his company and future client. Data cleaning, exploratory data analysis were used in this project to analyze the relationship between mill rate and other factors. Data cleaning is performed to aggregate our data into summary statistics and also reduce the number of data entries. Exploratory analysis has shown there is a linear trend in mill rates over the years; it has also shown there is a strong correlation between mill rate and municipal budget and strong correlation between mill rate and assessment in some municipalities.

Introduction

Our client is only interested in accurate predictive models.The goal is to predict the property tax in 2020, for each of the following 21 sub-regions in the Metro 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 mil rate, assessment value (Land and Imporvement) and municipal budget were computed to see the significance of these factors in our predictive model. An ANOVA analysis had also been carried out to see the significance of categorical factor such as Year, Municipality and Tax Class Code. A Simple linear model, a linear model with the interaction term, LASSO, Ridge Regression and Elastic Net were used to train our predictive model. Once we have finished the above tasks, we will use the mean square test error to evaluate the performance of our model.

Data Description

For exploratory data analysis and building predictive model, we used past property assessment data, which is provided by our client. It contains information about some commercial and residential properties in all cities in Metro Vancouver from 2016 to 2019. We selected variables that we believe were relevant to the property tax prediction. These selected variables are:

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

As the client suggested, we have 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)

We decided to calculate the sum of TotalAssessedValue, LandAssessedValue, ImprovementAssessedValue, and Budget for each combination of municipality and the tax class to reduce the dimension of our data. The mill rate is unique for each combination of municipality and tax class, so we have added this information into our data for each municipality and its tax classes.

Methods

Exploratory Analysis

Before we made any prediction on the future mill rate of Metro Vancouver's real estate market, we did some exploratory analysis to explore and visualize the main characteristics of our dataset and found relationships between Mill Rate vs. Assessment, Mill Rate vs. Municipal Budget, Mill Rate vs. Land Total, Mill Rate vs. Imrpovement Total and Mill Rate vs. Number of Properties.

The municipal government wants their tax income to match their annual spending in order to balance their bedget. Given this piece of information, we want to find and model the relationship between municipal tax income and municipal budget. We 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. We chose to illustrate such relationship using scatter plot below for Vancouver, Richmond and Burnaby as they are the most representing city in Metro Vancouver.

Mill rate is mainly affected by assessment and municipal budget, so we have plotted scatter plot of mill rate vs. assessment and mil rate vs. municipal budget to see the correlation between each pair of the two factors. ANOVA analysis is also performed to see the correlation between mill rate vs. tax class, mill rate vs. year and mill rate vs. municipality.

Measure of goodness of fit and prediction power

In this study, we built a few linear models and 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. The definition of goodness of fit and prediction power is given below.

- **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. We use mean squared prediction power to compare the prediction performance across all of our fitted linear predictive models. To do that, we will divide the data set into training data - used to build our models, and testing data - used to evaluate the prediction power of our models.

Linear model

For our linear model, Year, TaxClassCode, Municipalities, Assess Total, Land Assessment Total, Improvement Assessment total, Number of properties and tax (budget) are considered. We first explore the full linear model using all the available features. From our exploratory data analysis, we have found Vancouver, Richmond, Surrey and Burnaby are outliers. Further prediction model will be built based on these findings. We are also intesterd in how different tax class can affect our predictive model, so models based on different tax classes will also be considered in the future. For our reduced model, we build our linear predictor using only the significant variables from the full model (See table 4). The reduced model's performence in terms of goodness of fit will decrease with fewer features, however, the performance of prediction power might be improved.

Ridge, Lasso and Elastic Net

Other than simple linear regression, we are also interested in the performance of Ridge Regression, Lasso, and Elastic Net. They are more advanced than simple linear regression as they reduce the variance of the model. 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 predictive 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

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

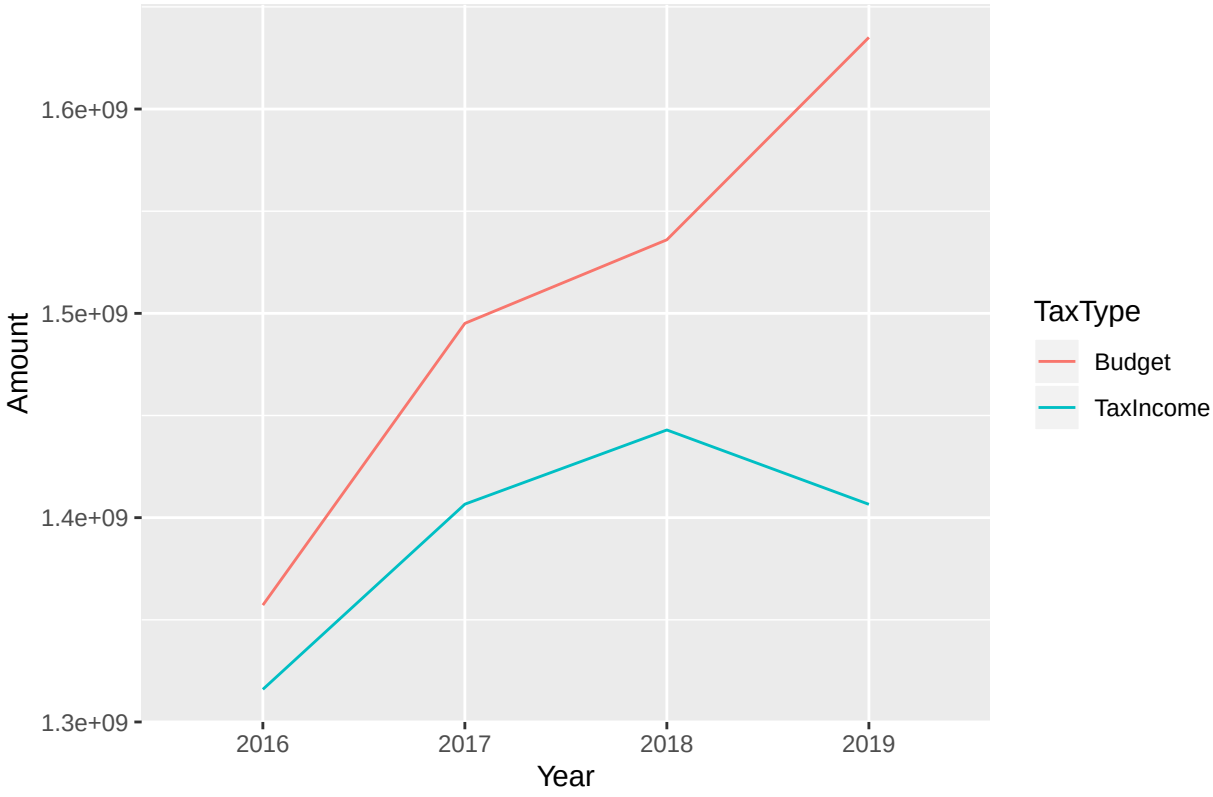


Figure 2: Budget and Tax Income v.s. Year of Richmond

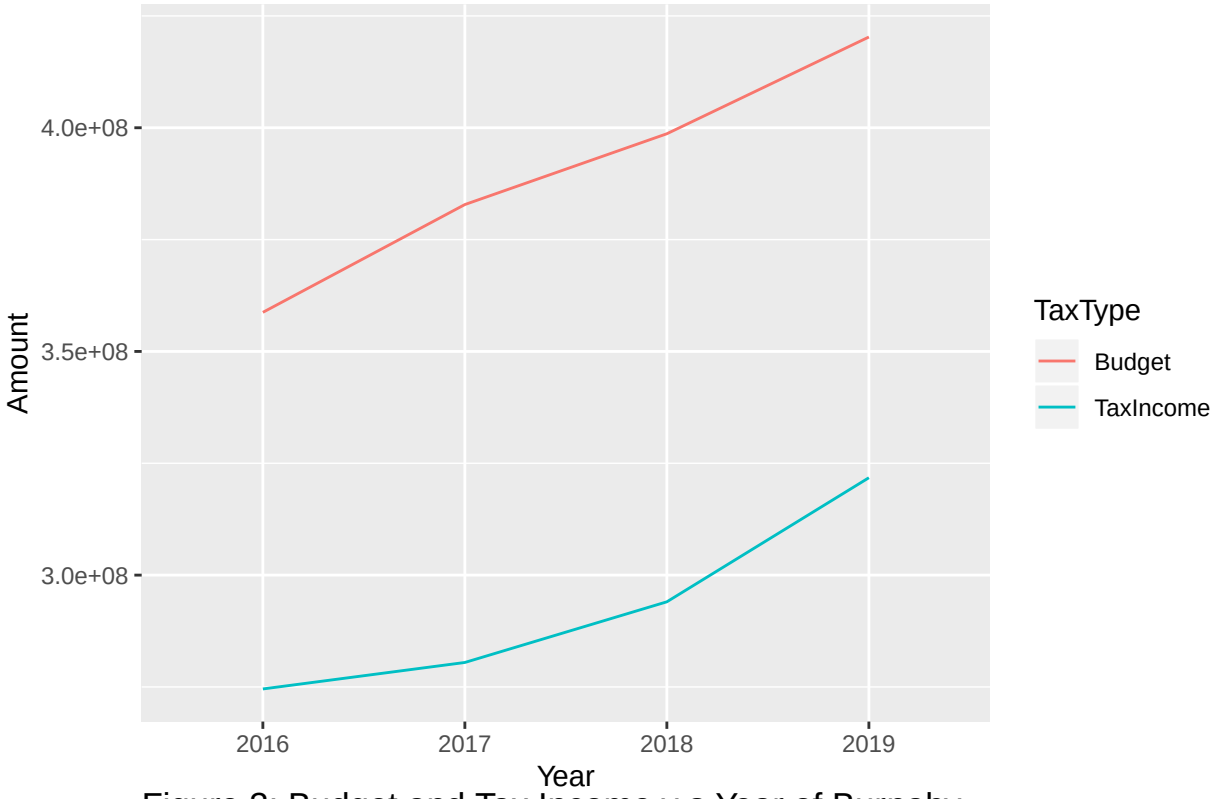
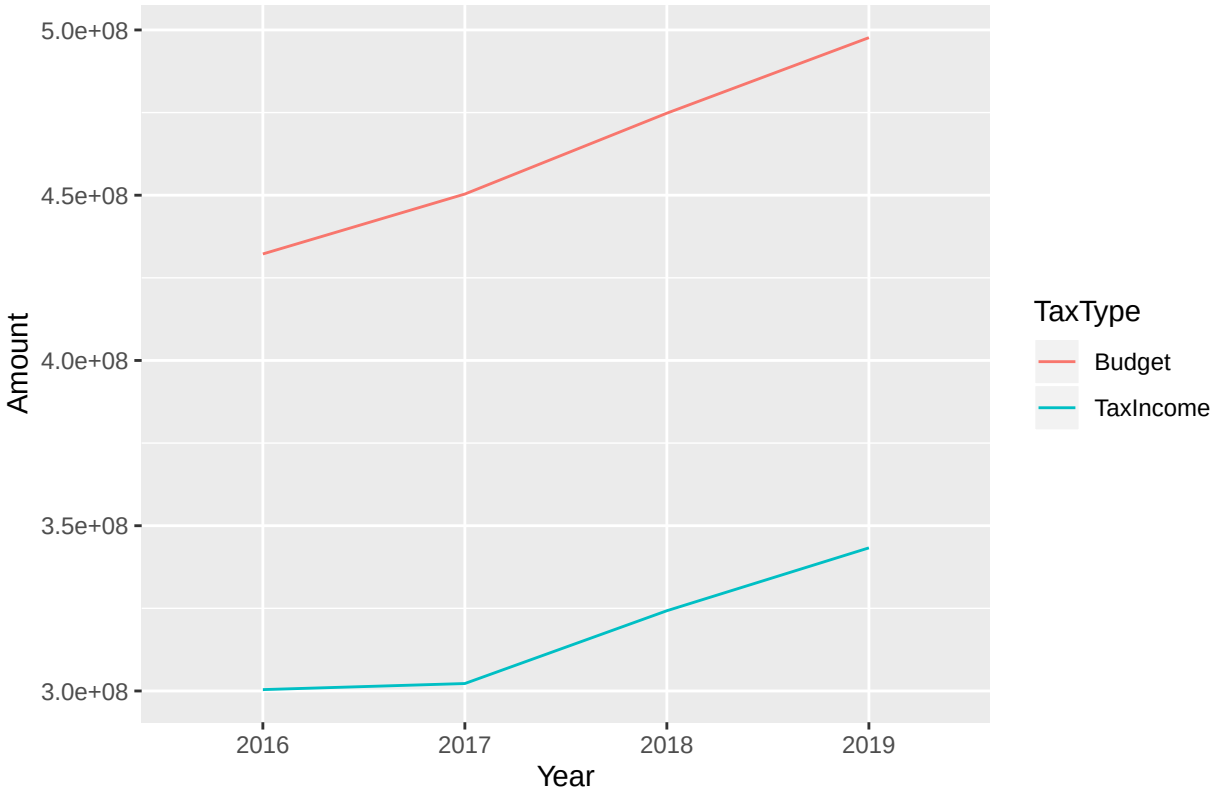


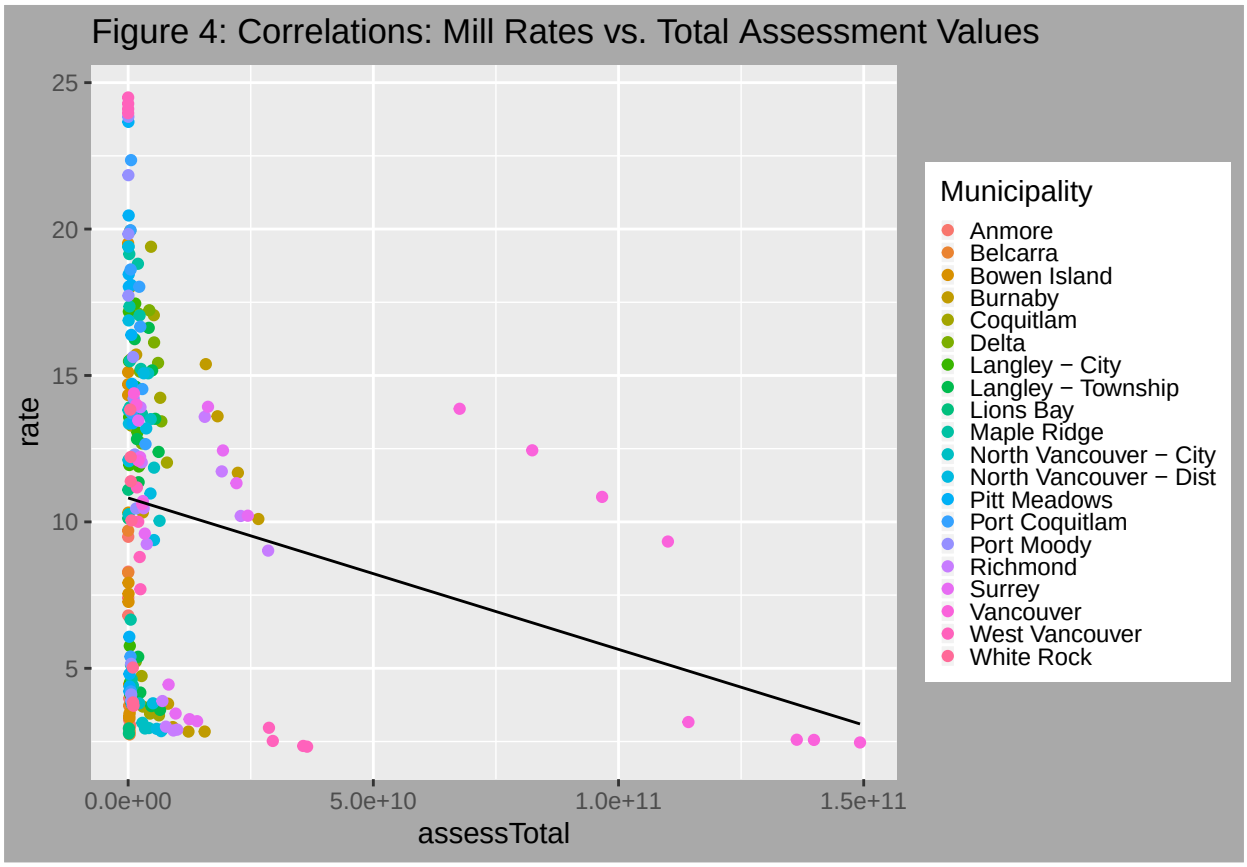
Figure 3: Budget and Tax Income v.s Year of Burnaby



We calculated the total tax income of each municipality by summing the product of assessment total and

mill rate of all tax class. Then we plotted the computed tax income vs municipal budget into line-chart to visualize their relationship. We chose Vancoouver, Richmond and Burnaby as representative because they are the three largest municipalities in Metro Vancouver. The plots showed that the computed tax amount was positively and approximatly linearly related to municipal budget through time.

Correlation Analysis



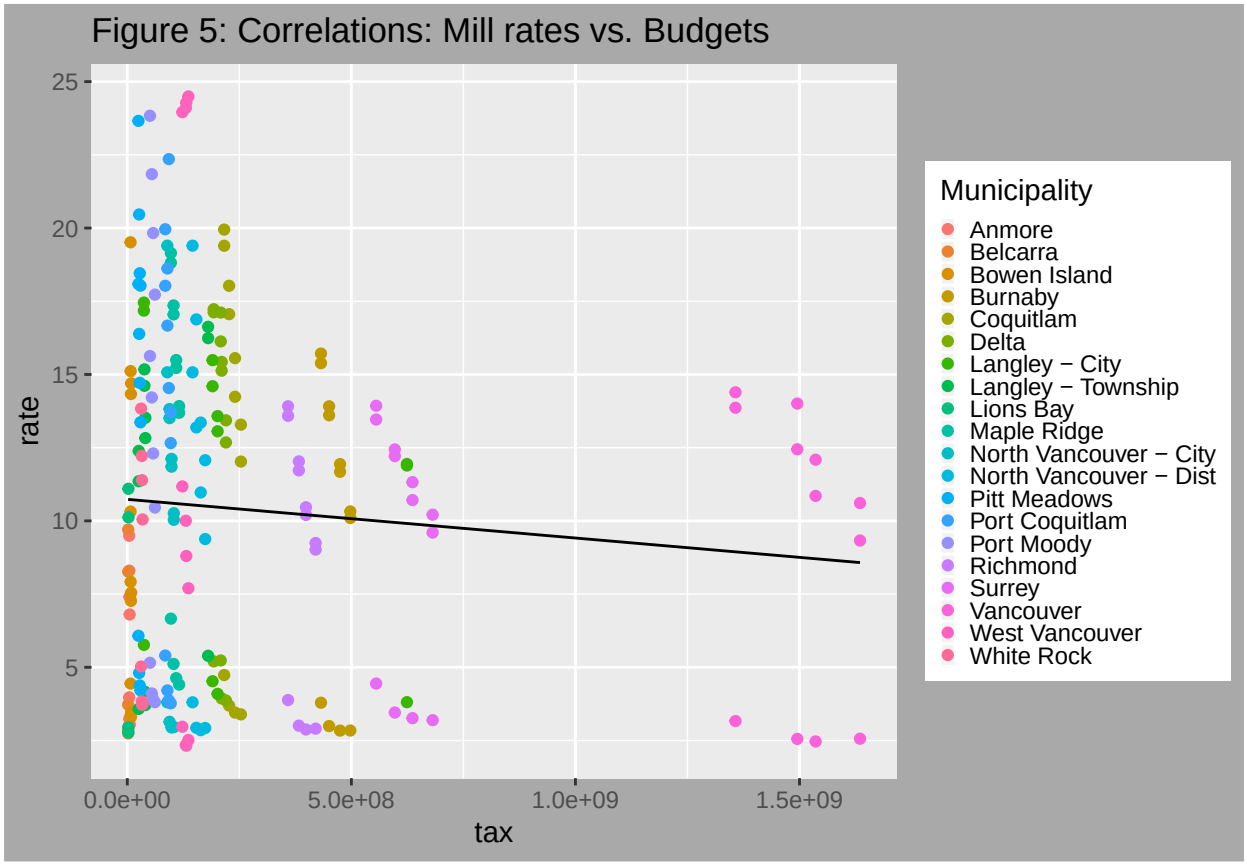


Figure 6: Correlations: Mill Rates vs. Total Assessment Values of Land

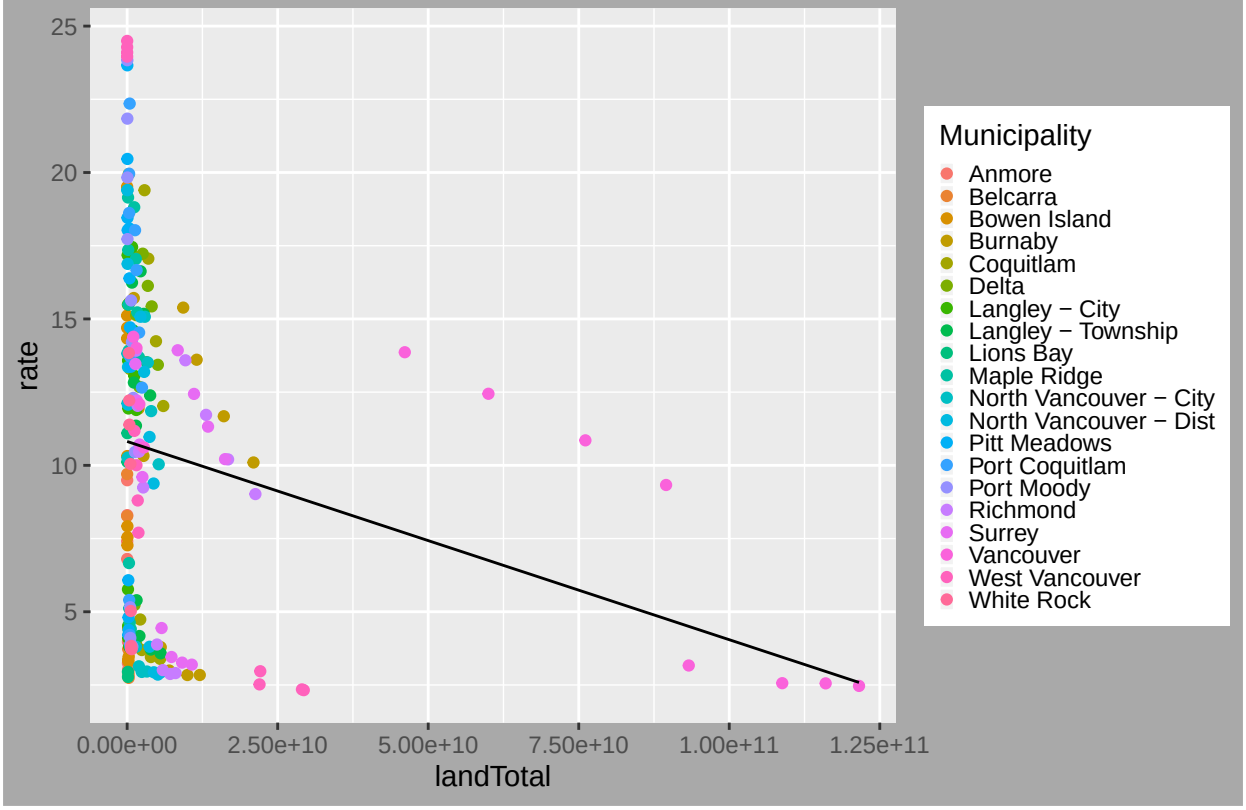
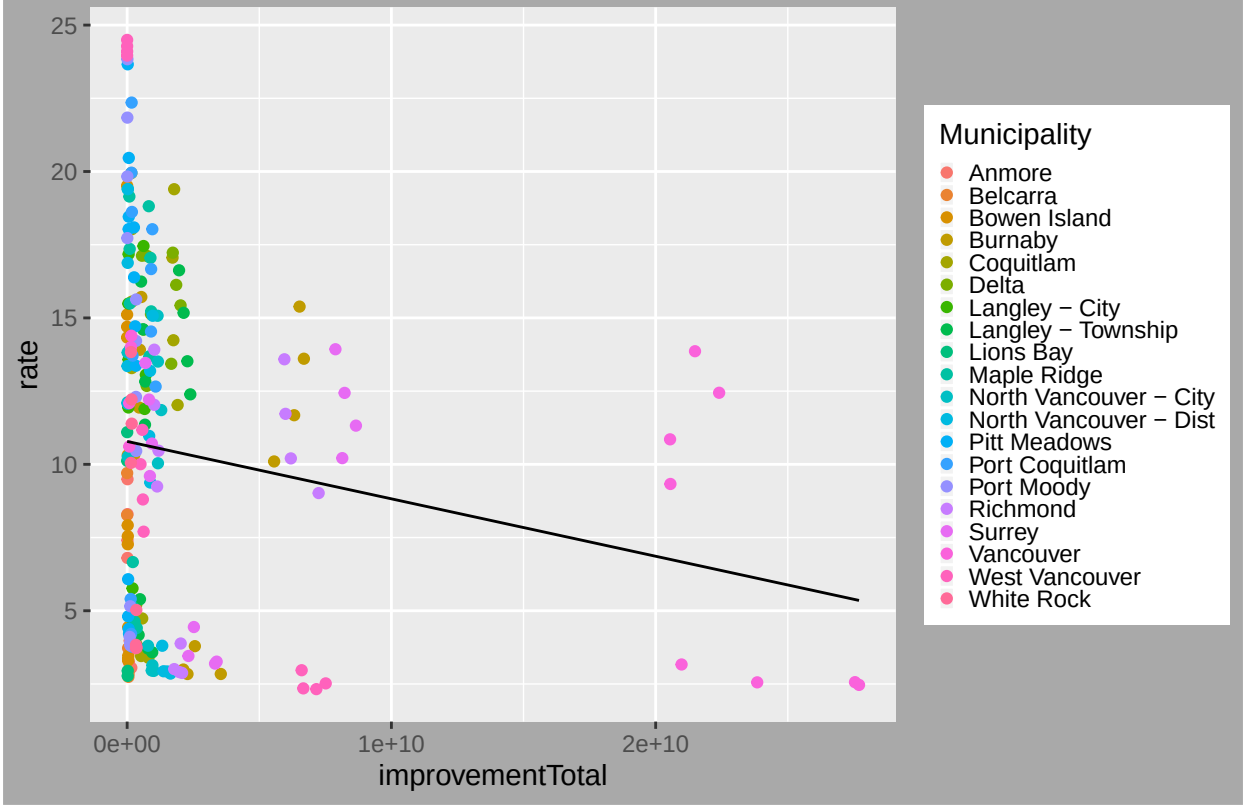
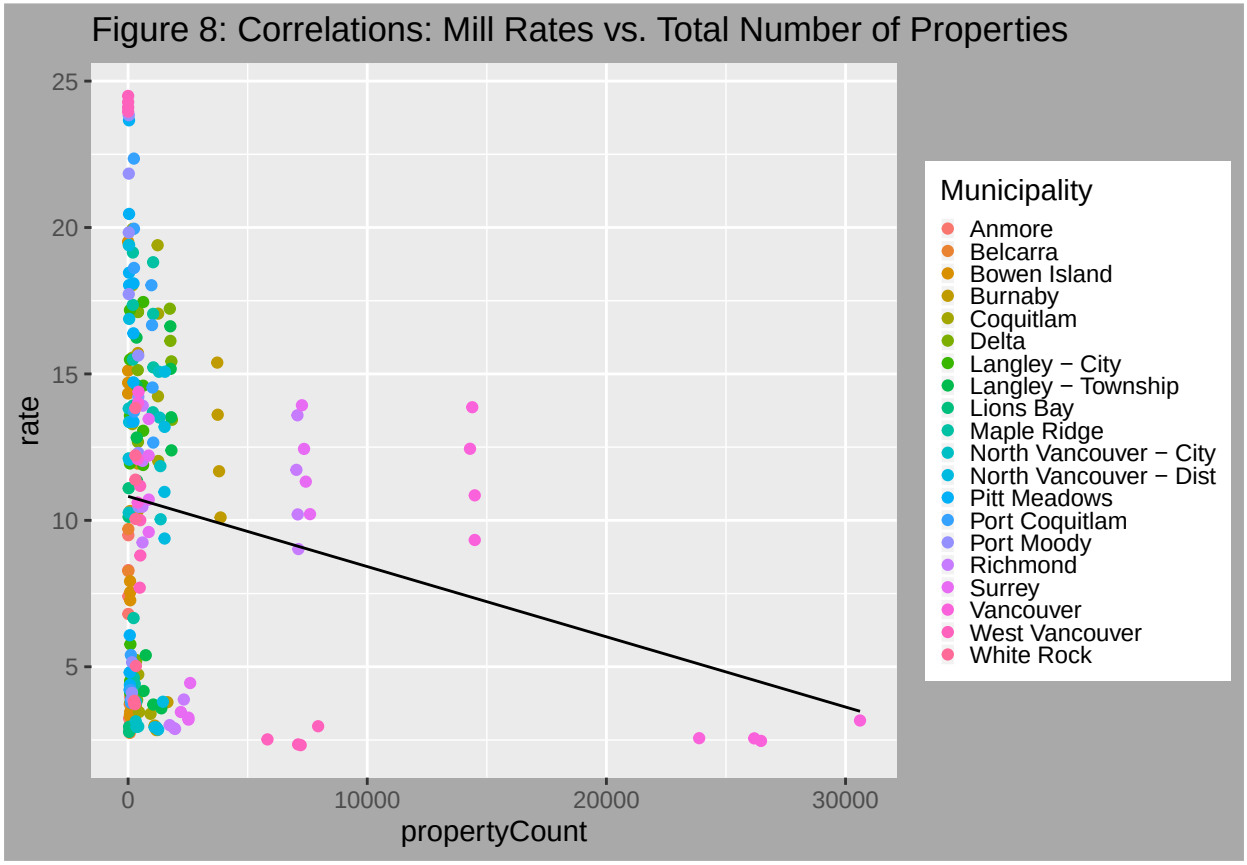


Figure 7: Correlations: Mill Rates vs. Total Assessment Values of Improveme





The above scatter plots showed the correlation between each pair of mill rate vs. assessment total, municipal budget, land assessment total, improvement assessment total, and the number of properties. From all the scatter plots above we have noticed that Vancouver (the dark pink points) is significantly different from other points and should be treated as an outlier/special case. Richmond, Surrey and Burnaby are also distinct in the scatter plots and should be treated as outliers. We believe these municiplaties also have impacts on the correlation between mill rate and its features. Further analysis and prediction model will be used to explore and explain such phenomenon.

ANOVA Analysis

Table 1: Result from ANOVA Analysis

Feature Name	Pr(>F)
TaxClassCode	<2e-16
Year	4.2e-06
Municipality	0.0768

We used three 3 ANOVA analysis to analyze the correlation between mill rate and each of the 3 categorical features - TaxClassCode, Year, Municipalities. TaxClassCode and Year showed a strong correlation with mill rate at 5% significant lete, yet the correlation between Municipality and mill rate is not significant at 5% level.

Linear Model

Table 2: Result from Model Used

Model	Mutiple R-Squared	Adjusted R_Squared	MSE	PMSE
OLR full	0.8874	0.8707	1.9843	2.5902
OLR reduced	0.8874	0.8721	1.9845	2.5237
Ridge	0.8868	NA	1.9896	2.5675
LASSO	0.8873	NA	1.9855	2.5280
Elastic Net	0.8625	NA	2.2366	2.5486

Above is the goodness of fit of five models. Except for the Elastic Net model which has higher MSE and PMSE, there are no significant differences between these models. 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 after using cross-validation there will be a clearer difference between the goodness of fit of each model.

For a dtailed look of how the models were fitted and evaluated, please refer to the appendix.

Conclusion

From our exploratory data analysis, we can see that assessment total, land total, tax code, year and municipalities are all significant in our model. The client suggested that we can transform numerical factors such as assessment total into percentage change. This data transformation does not perform as well as we expected, which only yields an R^2 of around 0.2 across all of our fitted models. We believe the method that the client has suggested might leave out some important information about the housing market in each municipality. We have also found that when comparing the correlation between mill rate against features in our models, Vancouver is an outlier. We would like to further investigate Vancouver as a special case. 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.

All fitted linear model were able to make accurate prediction based on means squared 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

Code of Linear Regressoin Model

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

```r
library(broom)

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

```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_aggregat

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


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
#optimal_lambda
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)
```

```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)
```

```r
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)

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+pro
 data = assessment_aggregate[train_ind,],
  method = "glmnet",
  trControl = cv_10
)

elastic_reg = glmnet(x_train, y_train, nlambda = 25, alpha = 1, family = 'gaussian', lambda =  0.065492
predictions_test <- predict(elastic_reg, newx = x_test)
elastic_net_mspe <- eval_results(y_test, predictions_test)
```

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 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 to 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 3: Data with missing 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

Here is the summary statistics from the full linear regression model

Table 4: Summary statistics of the full linear regression 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	