Real Estate Taxation

tax evaluation through Machine Learning

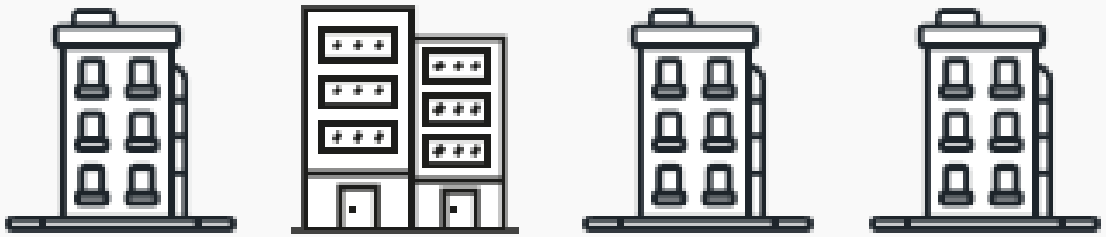
Content

California's Real Estate Tax Scheme

A two sided problem

The property tax value is based on the property value

For Sale



$245,000 **$255,000** **$245,000** **~~$245,000~~**

$285,000

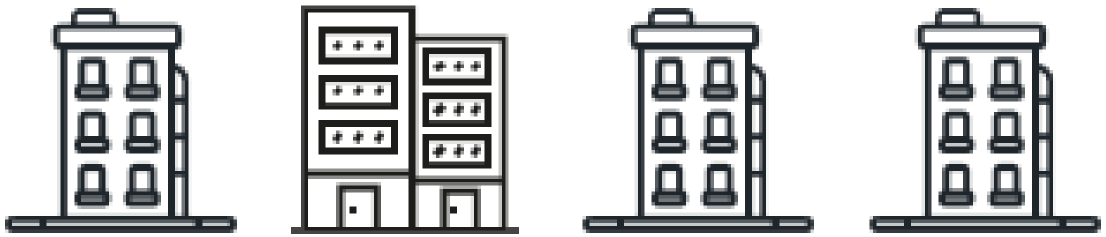
- Local real estate property is **assessed at acquisition** value and adjusted upwards each year by 2%.
- Most recent purchase price is the new tax value
- Creates unbalanced and **unfair tax weight** for new property owners as they have to pay a higher effective tax rate.

California's Real Estate Tax Scheme

A two sided problem

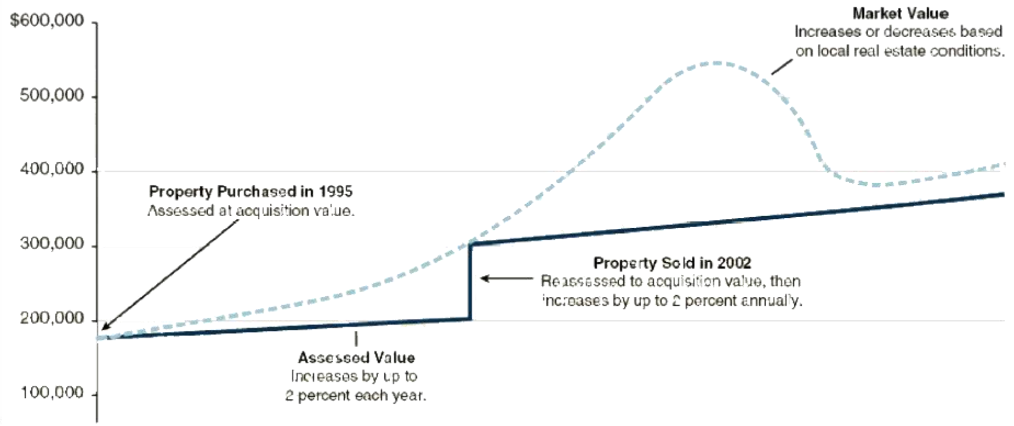
The property tax value is based on the property value

For Sale



$245,000 | $255,000 | $245,000 | ~~$245,000~~ $285,000

The property tax value is not reassessed every year



- Local real estate property is **assessed at acquisition** value and adjusted upwards each year by 2%.
- Most recent purchase price is the new tax value
- Creates unbalanced and **unfair tax weight** for new property owners as they have to pay a higher effective tax rate.

- **Loss** in Tax revenue due to increase in market value over time
- **Heavy effective** tax load during decline in market value

Question

"Can machine learning algorithms help in the prediction of real estate tax value and thereby reduce the unbalance in tax load and governmental loss of tax revenue?"

Methodology

How to answer such a research question

Cleaning, Impuation and Generalization

Dateset from Kaggle competition contains a large number of missing values

Exploratory Descriptive Statistics

Interesting patterns in the value for the structure and land taxes

Hypotheses formulation

Formulating hypotheses about relations in order to reduce variables

Hypotheses Testing

Testing relationships in order to predict the structure and land tax value

Data Cleaning Process

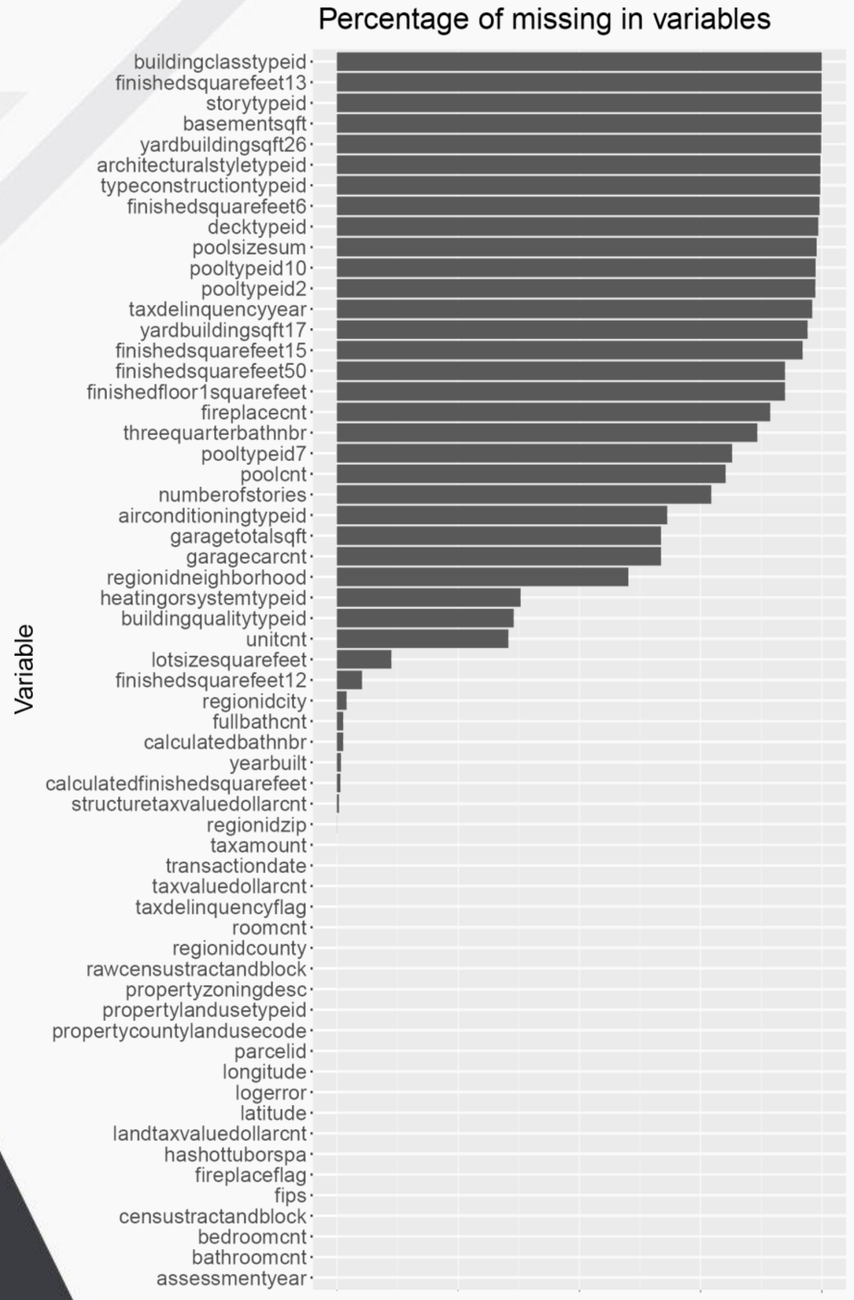
From Missing to Zero

- Mostly missing values: 41.3%

  90275 observations in the dataset

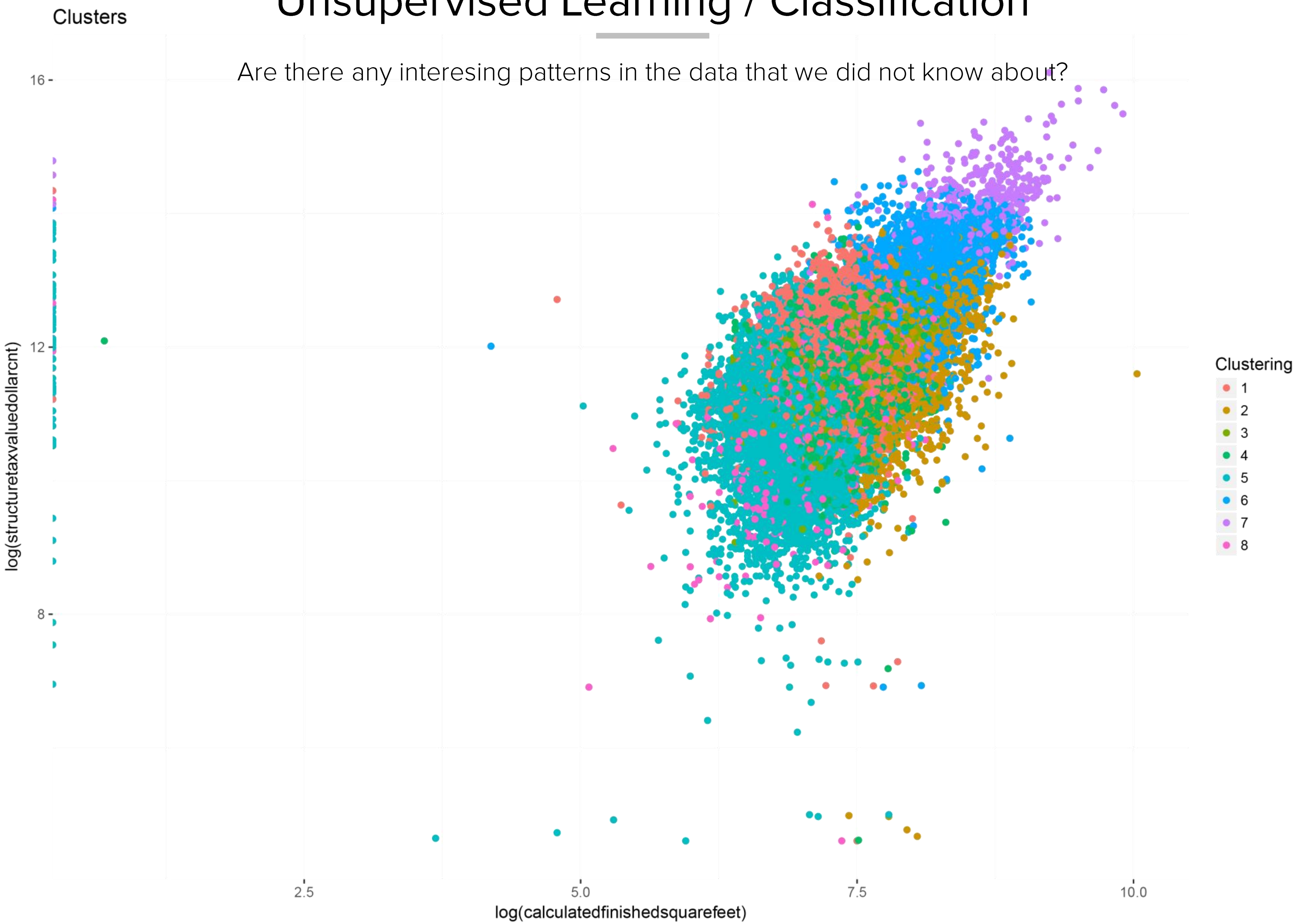
- Special Imputation Rules:
  - Year built = mean(year built)
  - Taxamount = 2% of totaltaxvalue
  - Structural house tax = Total tax – Land tax
  - Land tax = Total tax – Structural house tax

- Other missing values set to zero
- Blank spaces are not detected by is.na()



Unsupervised Learning / Classification

Are there any interesing patterns in the data that we did not know about?



Generalizable or Transferable

Will the conclusions be valid for the county of Los Angelas?

- *A representative sample is a sample which, for a specified set of variables, resembles the population … (Kruskal and Mosteller, 1979)*
- Teddlie and Yu (2007) substantiate this statement as they mention that such as sample (*purposive sample)* may seeks the form of generalizability and consequently could be considered to have the characteristics of transferability.

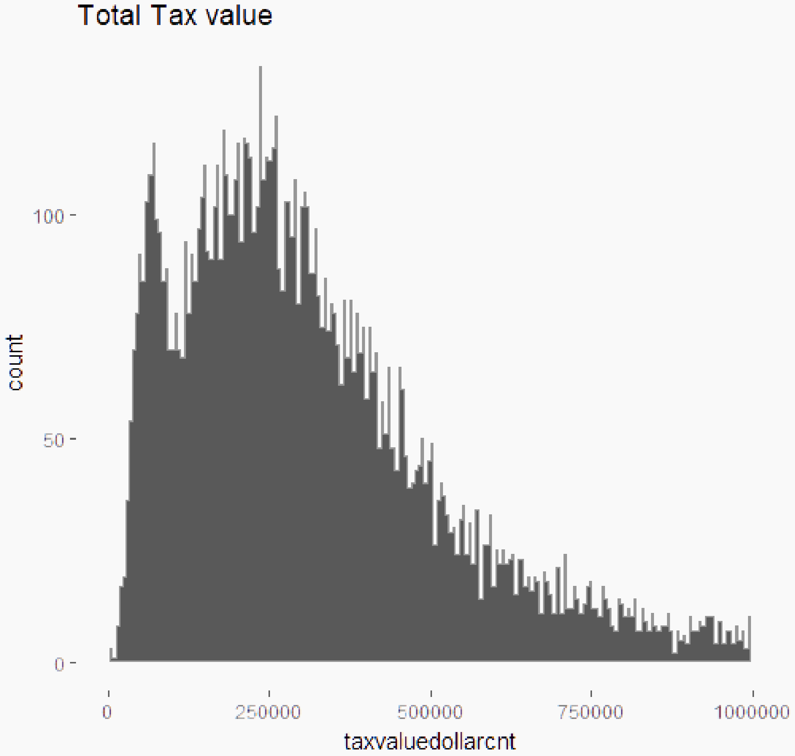
- *Is the data generalizable?* **Yes**
  - *Chi-Square p-value on Construction year:* 0.2313
  - *Chi-Square p-value on House Value / Tax Value:* 0.2289
- *Is the data randomly sampled?* **No**

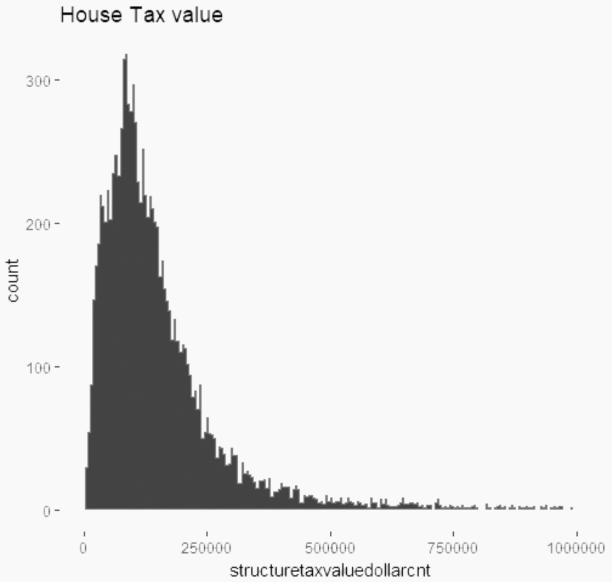
→ ***Transferability***

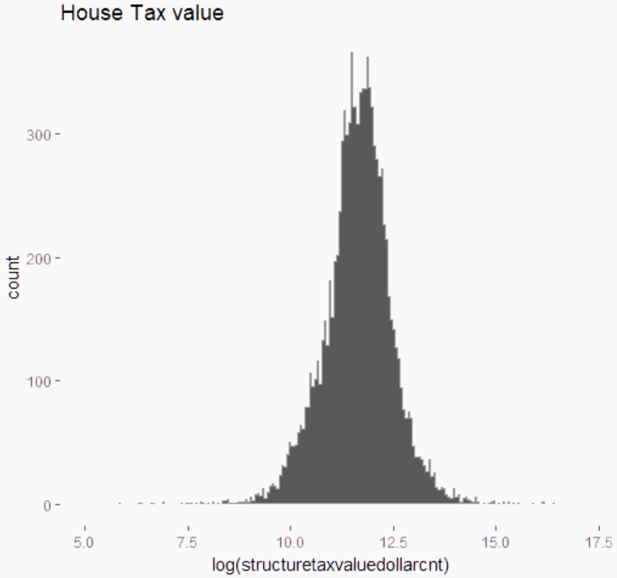
- Kruskal and Mosteller (1979) (taken from pp. 31-32 of Stephan, Frederick F., and McCarthy, Philip J., Sampling Opinions: An Analysis of Survey Procedure, New York: Wiley, 1958)
- Teddlie, C., & Yu, F. (2007). Mixed Methods Sampling: A typology with examples. Journal of Mixed methods research, 77 - 100.

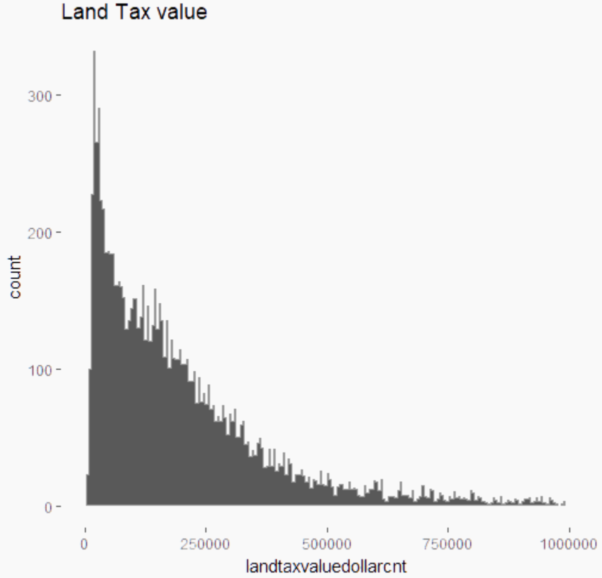
Descriptive Statistics

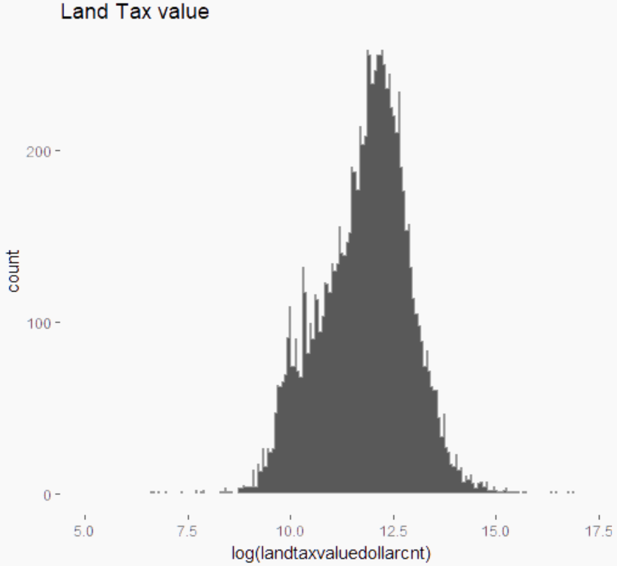
Are there any interesing patterns in the independent variable?





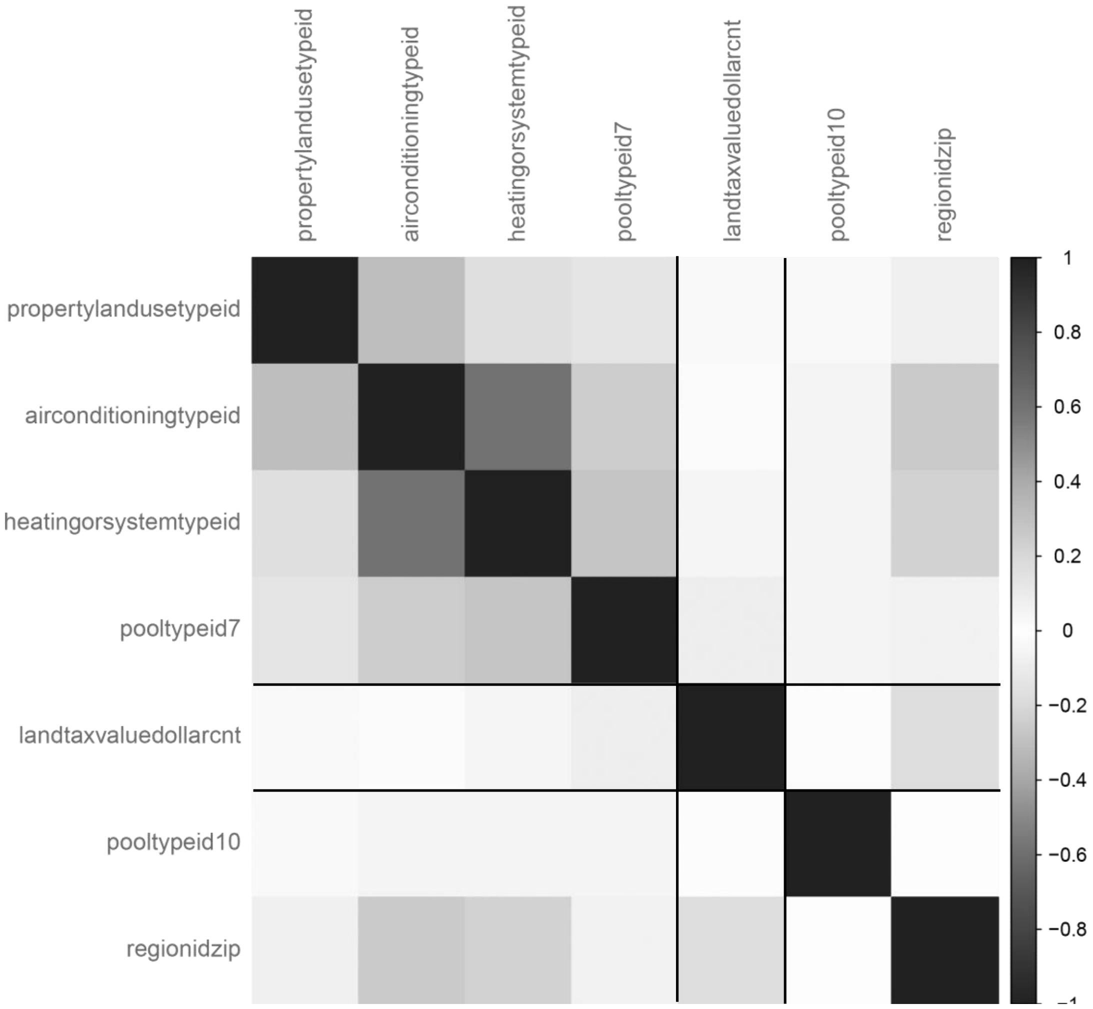


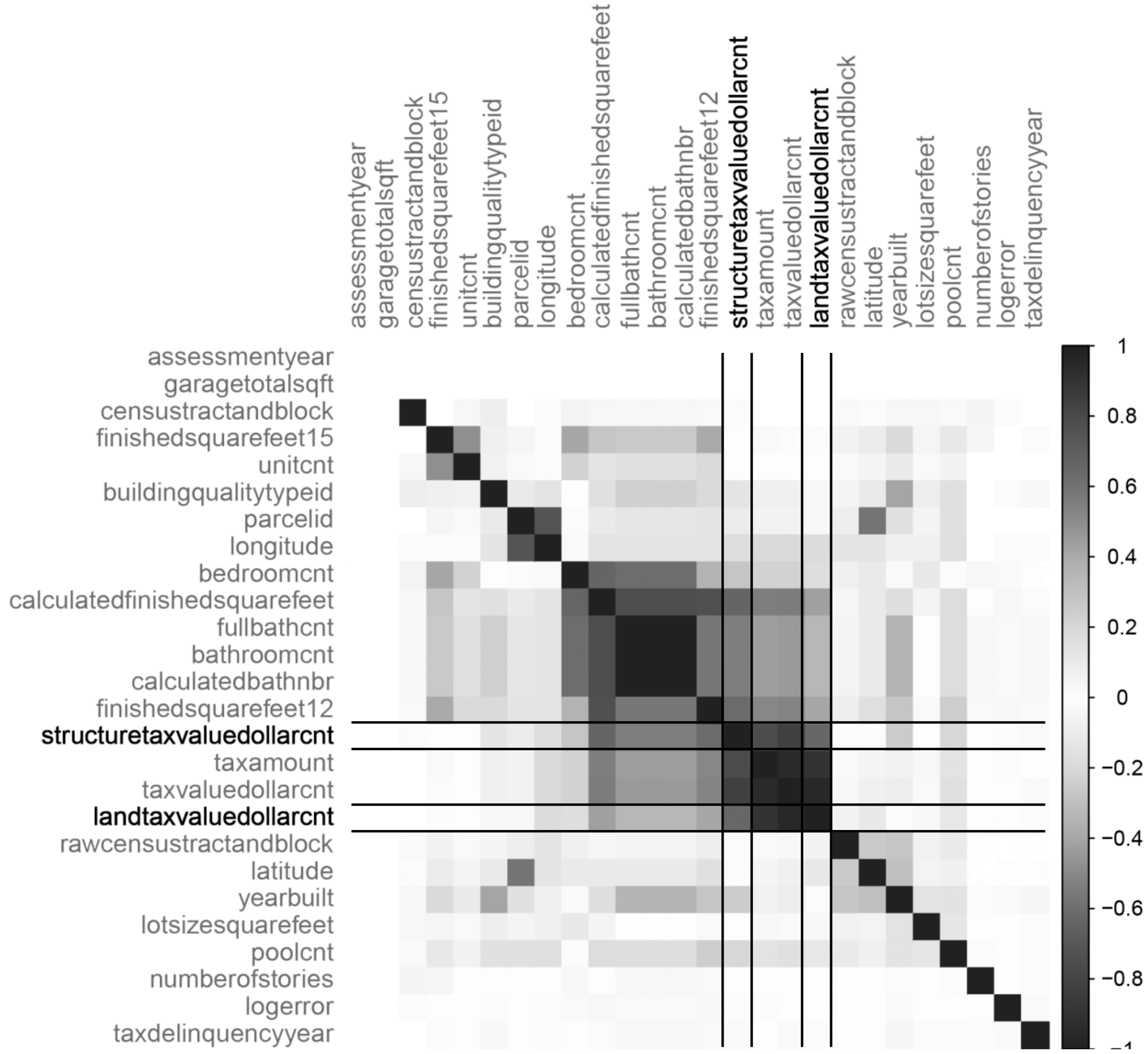




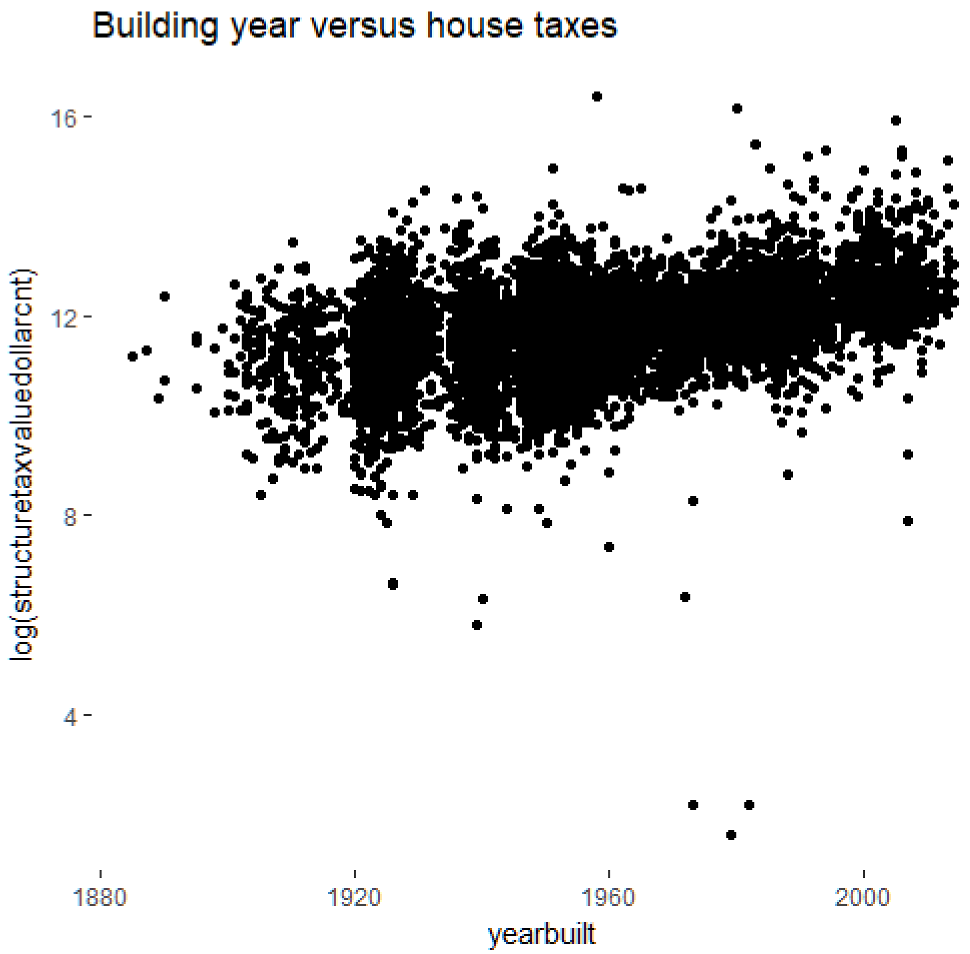
Descriptive Statistics

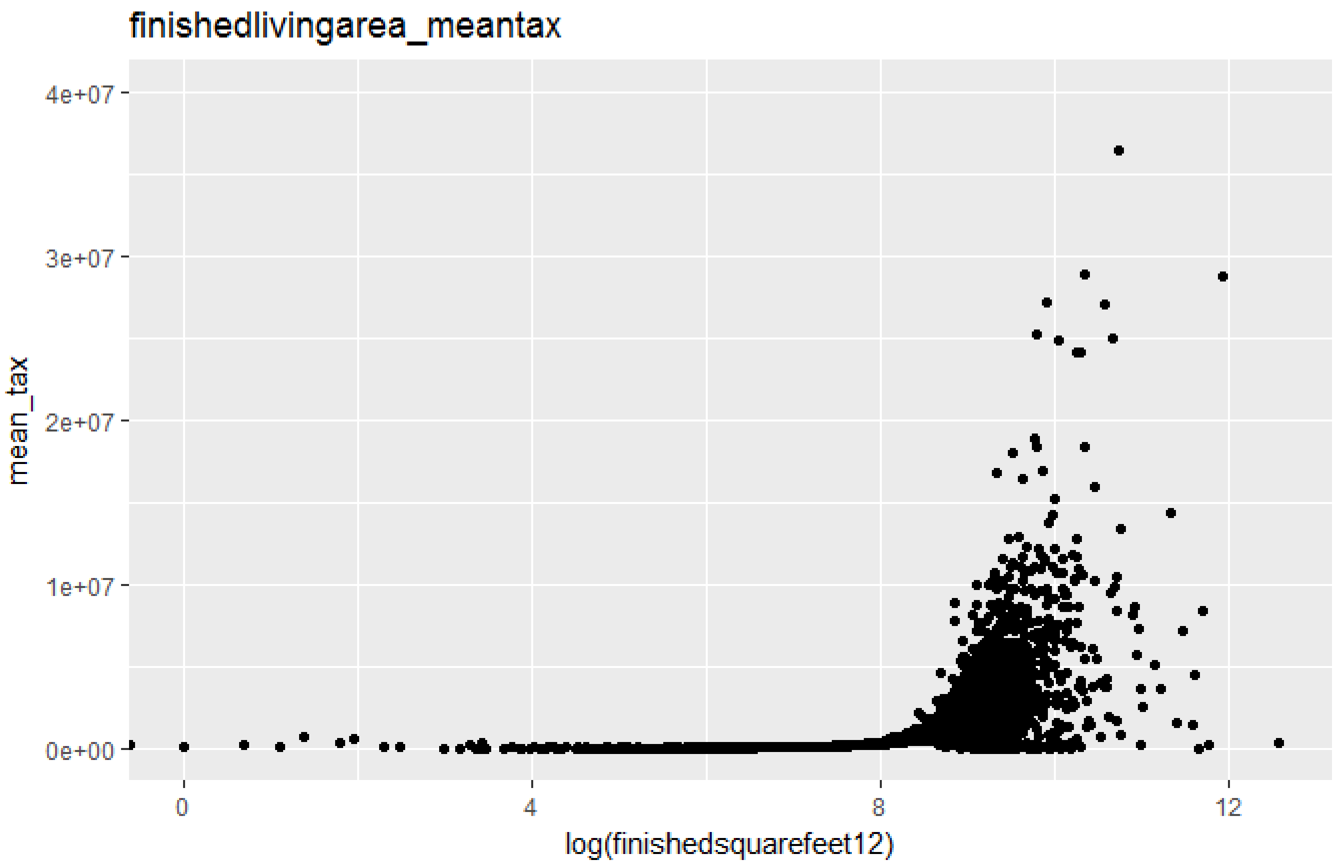
Are there any bivariate relations with the dependent variable?



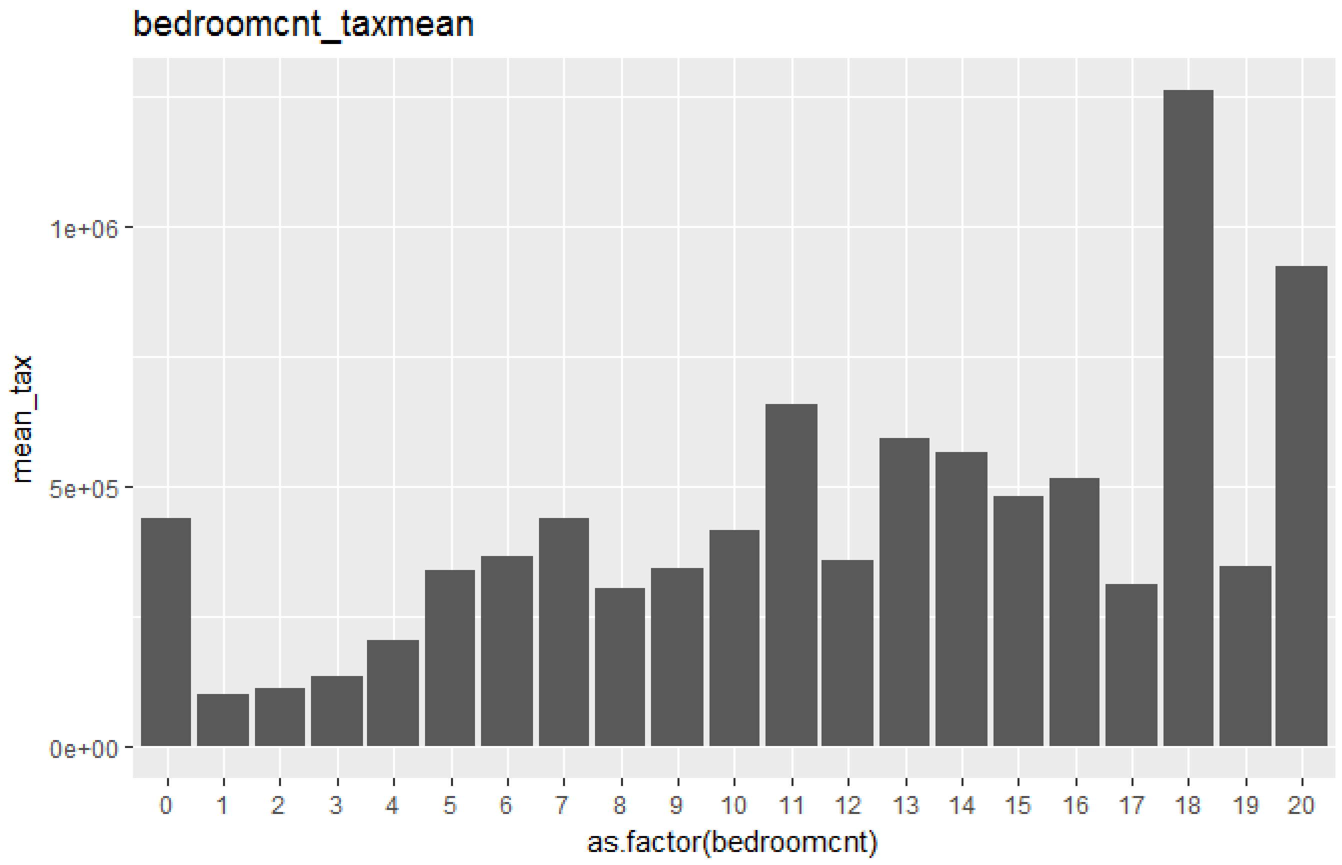


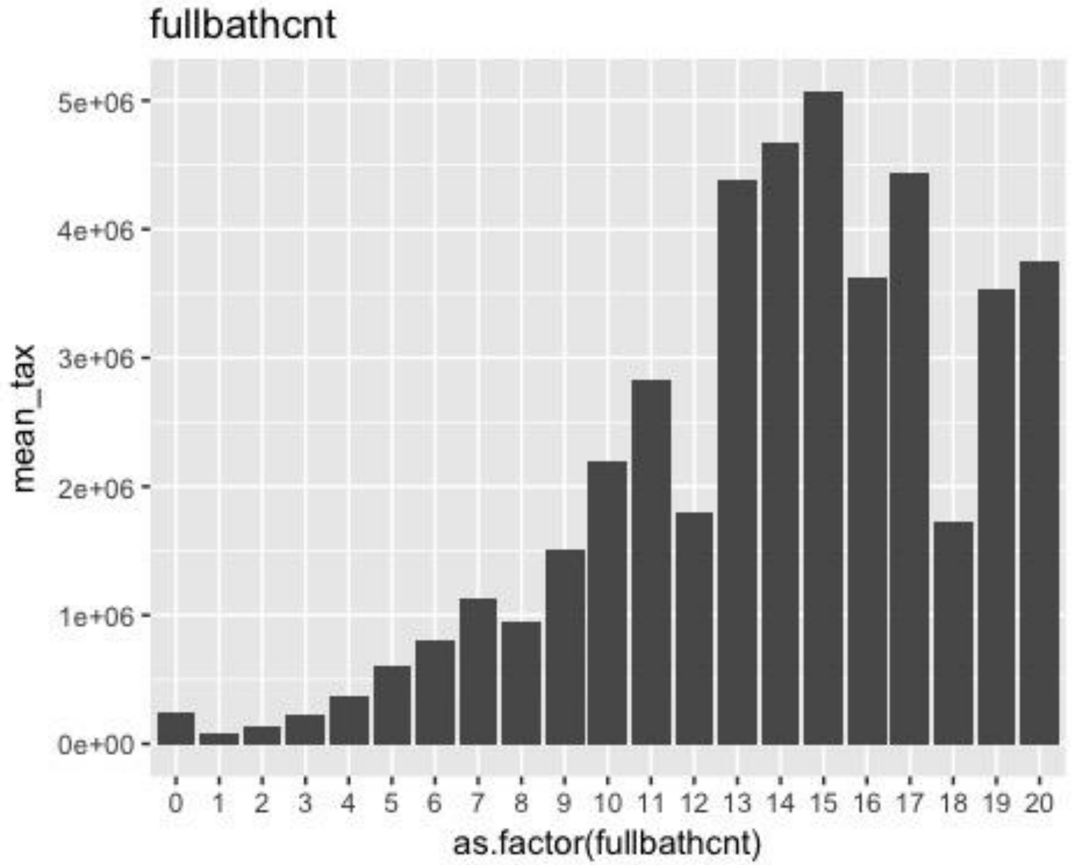
Descriptive Statistics



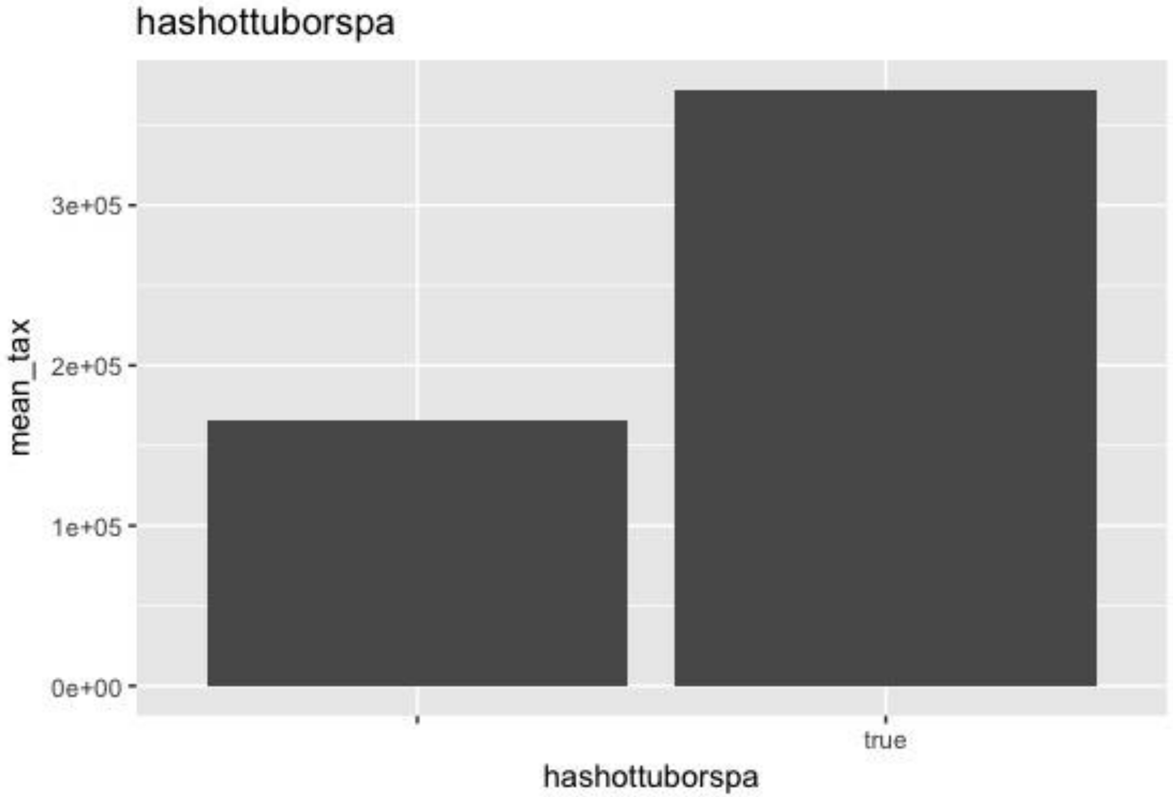


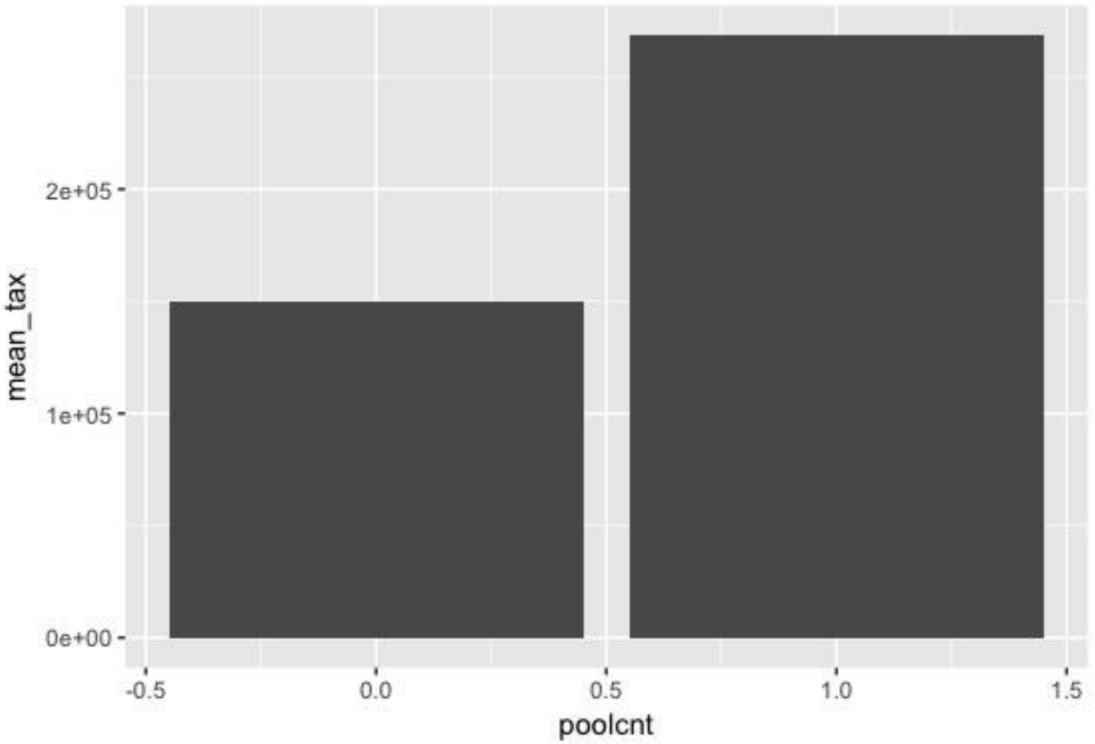
Descriptive Statistics





Descriptive Statistics





Hypotheses Formulation

- In order to provide a clear overview of the relations, a causal relations diagram can be presented. However, a causal relations diagram is a model that expresses more than correlation (Judea, 2000), as correlation does not imply causation.

- Literature cannot be considered a fool proof technique to determine if such a relation actually exists (Berry & Feldman, 1985). In order to proof that a relation exists, a bivariate analysis may be performed.

Judea, P. (2000). *Causality: Models, Reasoning, and Interference*. Cambridge University Press.
Berry, W. D., & Feldman, S. (1985). Multiple regression in practice (No. 50). Sage.

Causel Model

Housing Tax

Airconditioning +		+ Heating system
Number of Bathrooms +		~ Land Type
Number of Bedrooms +		+ Unit count
Building Quality +		+ Construction year
Square Feet +		+ Number of Pools

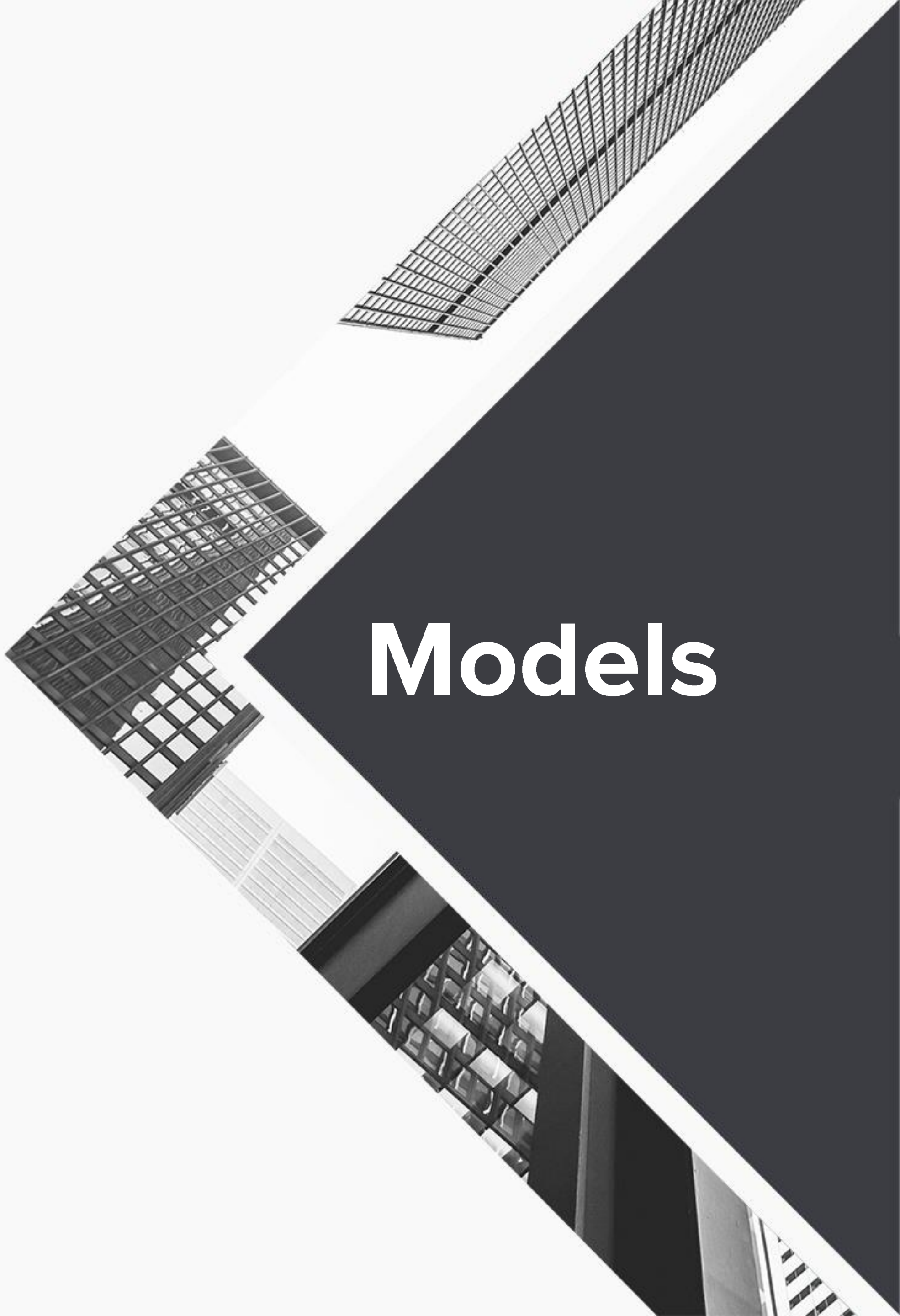
Bivariate Analysis

Hypotheses Testing of expected relations

Variable	Test	Test Statistic	P Value	Cor	Alt Hyp	Con	Use in model
airconditioningtypeid	Welch Two Sample t-test	54.269	p-value < 2.2e-16		Positive	Reject H0	Yes
bathroomcnt	Pearson's product-moment correlation	179.33	p-value < 2.2e-16	0.5962435	Positive	Reject H0	Yes
bedroomcnt	Pearson's product-moment correlation	80.7	p-value < 2.2e-16	0.3170245	Positive	Reject H0	Yes
buildingqualitytypeid	Pearson's product-moment correlation	-32.534	p-value < 2.2e-16	-0.1335384	Positive	Cannot reject	No
calculatedfinishedsquarefeet	Pearson's product-moment correlation	175.53	p-value < 2.2e-16	0.5880057	Positive	Reject H0	Yes
poolcnt	Welch Two Sample t-test	32.77	p-value < 2.2e-16		Positive	Reject H0	Yes
yearbuilt	Pearson's product-moment correlation	111.12	p-value < 2.2e-16	0.4180533	Positive	Reject H0	Yes
unitcnt	Pearson's product-moment correlation	-0.11402	p-value = 0.9092	-0.0004722	Positive	Cannot reject	No
propertylandusetypeid	Kruskal-Wallis rank sum test	2231.5	p-value < 2.2e-16		Difference	Reject H0	Yes
heatingorsystemtypeid	Kruskal-Wallis rank sum test	11822	p-value < 2.2e-17		Difference	Reject H0	Yes

Models

Predicting the Structure Tax Value



Multiple Linear Regression

Model Parameters:
structuretaxvaluedollarcnt ~ airconditioningtypeid + bathroomcnt + bedroomcnt + calculatedfinishedsquarefeet + heatingorsystemtypeid + poolcnt + yearbuilt + propertylandusetypeid + Clustering

Model Significance:
p-value < 2.2e-16

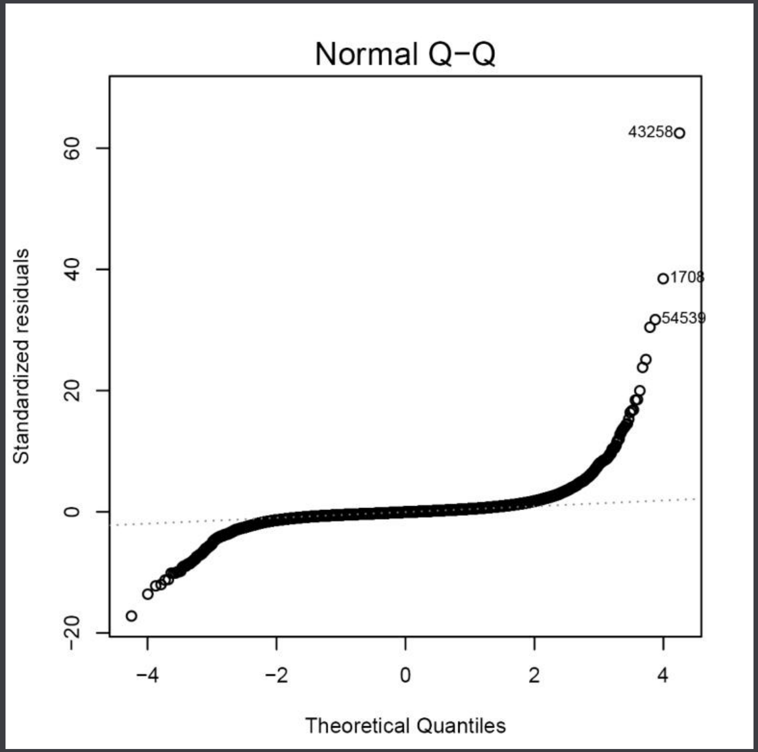
Adj. R^2:
0.664

Higher R2 due to significant variables that will not be significant under White's standard erros. Ergo, overfitting of the data.

VIF:
all below 2.5

Breusch-Godfrey test for Serial Correlation:
p-value = 0.5906

Studentized Breusch-Pagan test:
p-value < 2.2e-16



Box Cox transformation of the dependent variable

MLR Box Cox Transformation

Lambda:
0.2626

Model Parameters:
structuretaxvaluedollarcnt ~ airconditioningtypeid + bathroomcnt + bedroomcnt + calculatedfinishedsquarefeet + heatingorsystemtypeid + poolcnt + yearbuilt + propertylandusetypeid + Clustering

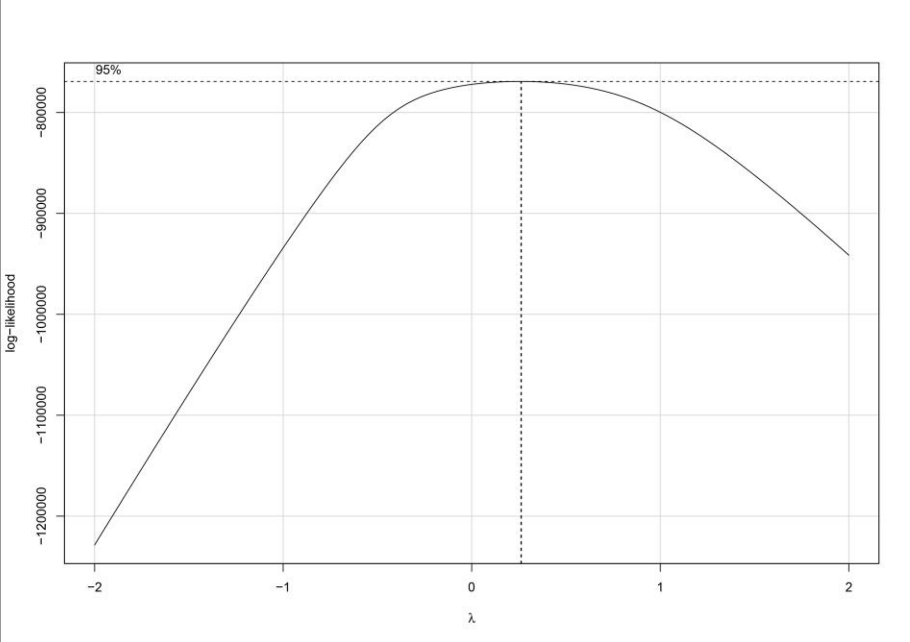
Model Significance:
p-value < 2.2e-16

Adj. R^2:
0.6005

VIF:
all below 2.5

Breusch-Godfrey test for Serial Correlation:
p-value = 0.7893

Studentized Breusch-Pagan test:
p-value < 2.2e-16



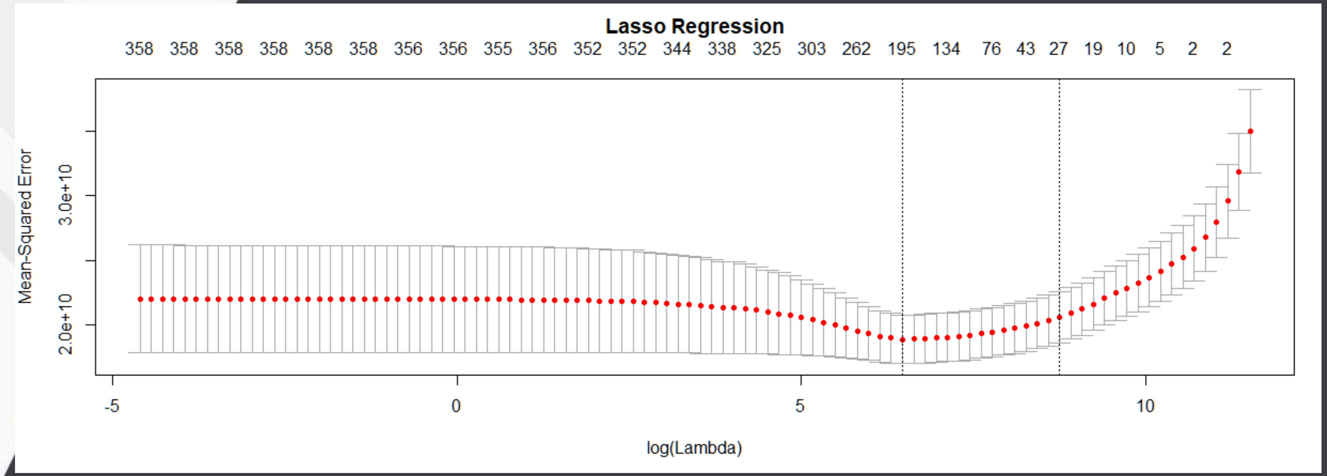
Lasso Regression

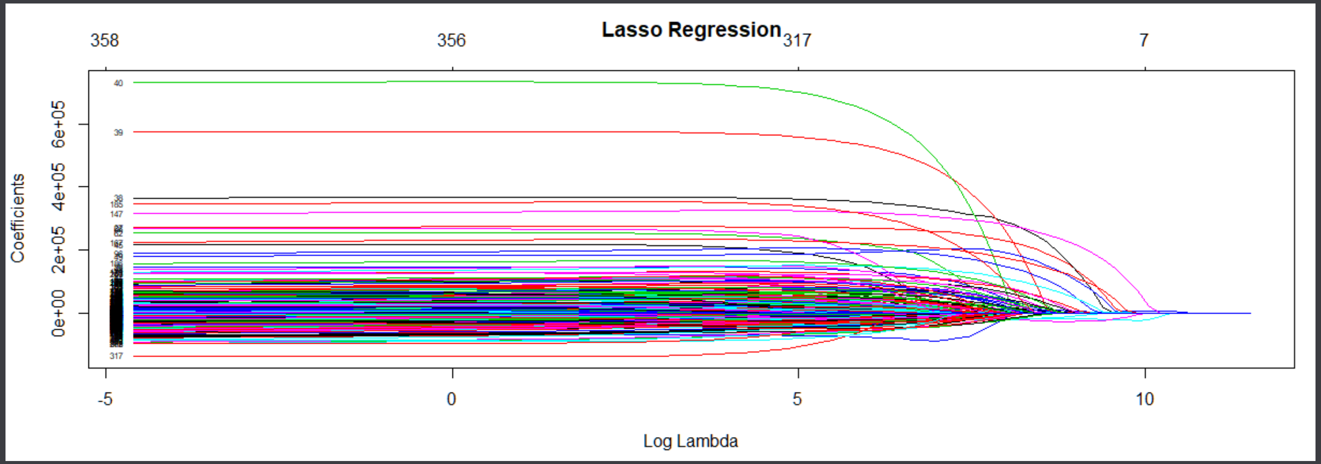
Lambda:
126.1857

Final Model Parameters:
structuretaxvaluedollarcnt ~ calculatedfinishedsquarefeet + latitude + longitude, Regionidzip + clustering

Model Significance:
p-value < 2.2e-16

Adj. R^2:
0.681





Random Forest

Model Parameters:

Structuretaxvaluedollarcnt ~ airconditioningtypeid + bathroomcnt + bedroomcnt + buildingclasstypeid + buildingqualitytypeid + calculatedbathnbr + calculatedfinishedsquarefeet + finishedsquarefeet12 + finishedsquarefeet15 + fullbathcnt + garagetotalsqft + heatingorsystemtypeid + latitude + longitude + lotsizesquarefee + poolcnt + pooltypeid10 + pooltypeid7 + propertylandusetypeid + rawcensustractandblock + yearbuilt + numberofstories + structuretaxvaluedollarcnt + assessmentyear + Clustering + population_level

Mean of squared residuals:

28,364,242,949

R^2:

71.66%

Var explained:

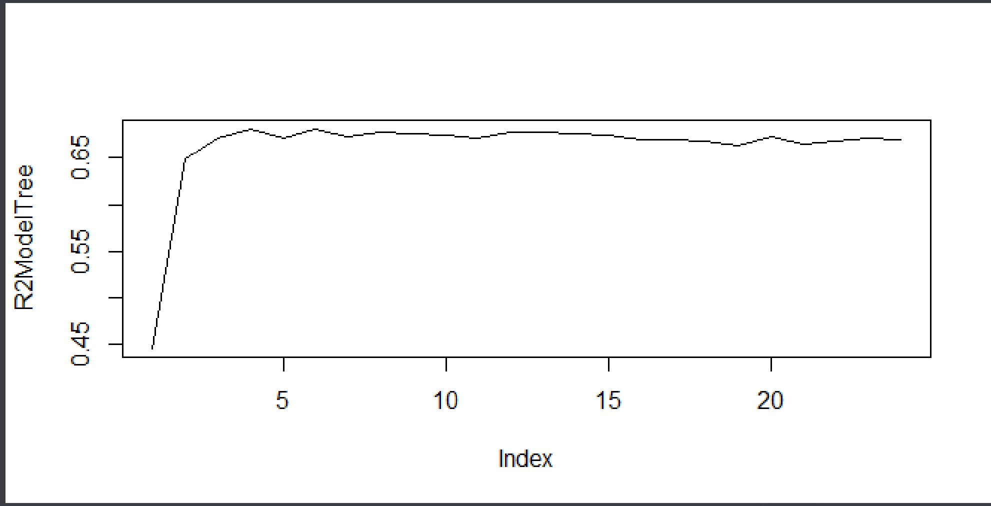
70.76 %

No. of variables tried at each split:

4

TOP 3 Importances:

Calculatedfinishedsquarefeet
Finishedsquarefeet12
Clustering



Boosting

Number of trees:
5000

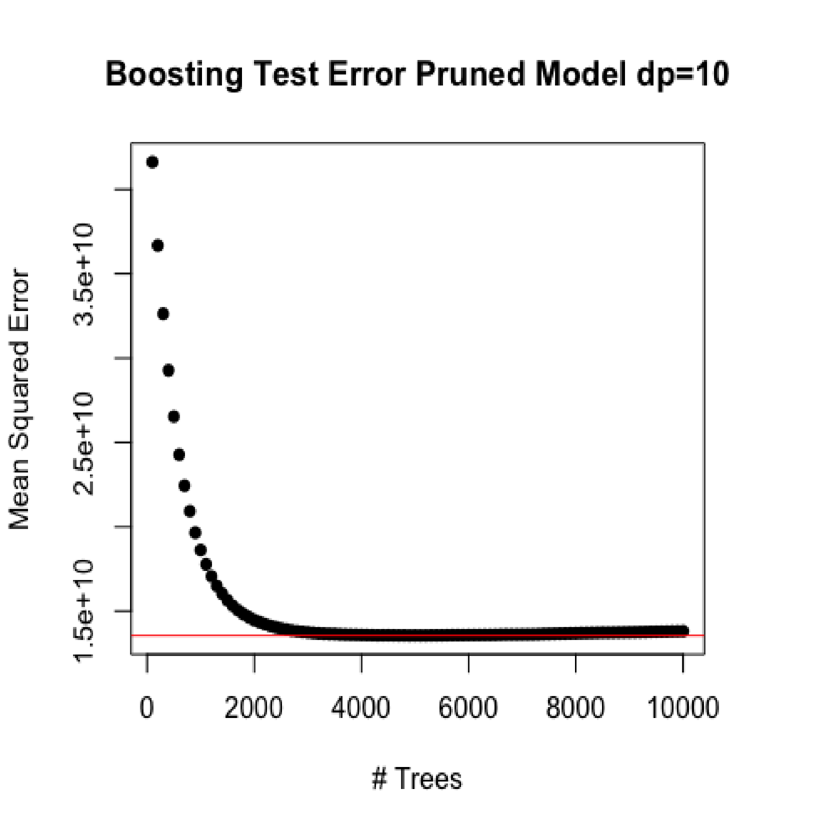
Interaction.depth:
4

Shrinkage:
0.001

Model Parameters:
structuretaxvaluedollarcnt ~ + finishedsquarefeet12 + calculatedfinishedsquarefeet + buildingqualitytypeid + yearbuilt + bathroomcnt + lotsizesquarefeet + bedroomcnt + propertylandusetypeid + poolcnt + airconditioningtypeid + pooltypeid7 + taxdelinquencyyear + pooltypeid10 + hashottuborspa + heatingorsystemtypeid + finishedsquarefeet15 + unitcnt + Clustering

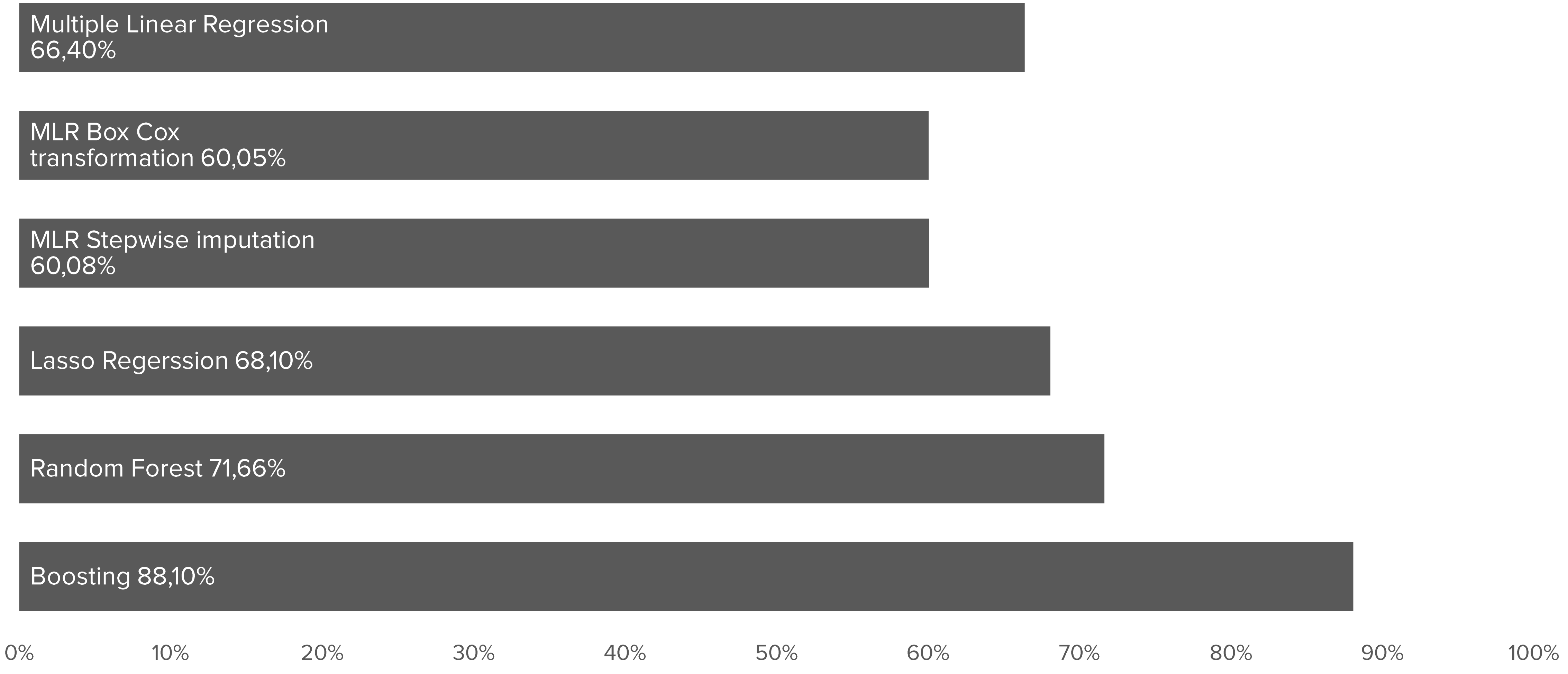
Train R^2:
0.881

Test R^2:
0.378



Model Comparison

Trying to capture most of the variance



Conclusions

What did we learn?

- The Kaggle dataset, which contains a high number of missing values, holds insufficient information to accurately predict the real estate tax value.

- More complex modelling techniques provide higher levels of R2; however also use variables that causally do not have any meaning, which indicates overfitting, or they perform bad in prediction out of sample.

- Overall, complex modelling techniques outperform multiple linear regression only with a few percent, making multiple linear regression techniques preferable due to their dependency on causal relations and interpretability.

- Tax regulations make predicting the tax value / house value complex as two identical properties of equal value can have a great amount of variation in their assessed value, even if they are next to each other. Consequently, with this dataset, it is impossible to capture 100% of the variance, even when models are overfit.

- Predicting the real estate tax value is possible; however contingent on a complete dataset that contains the current market value of the house.

Real Estate Taxation

tax evaluation through Machine Learning

Thank you!