You have been tasked with undertaking a multi-part analysis of homes in Detroit, Michigan. You are provided with a database to facilitate this analysis. This database was constructed from the Detroit Open Data portal and numerous FOIA requests. More information is included in the database section below. Note that the database must be downloaded.

PART 1

R Markdown Requirements

Please include `code_folding: hide` as a yaml option and `knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)` in your setup chunk so that your code can be seen in your knitted output but is initially hidden.

dbplyr

Starter code, replace with path.

```
con <- DBI::dbConnect(RSQLite::SQLite(), "PATH")

# sales tbl

dplyr::tbl(con, 'sales')
```

```
# convert to tibble
#dplyr::tbl(con, 'sales') %>%
dplyr::collect()

# sql query

dplyr::tbl(con, 'sales') %>%
count(year(sale_date))

#dplyr::tbl(con, 'sales') %>%
count(year(sale_date)) %>% show_query()
```

Database

I have provided data via a sqlite database. It can be found on OneDrive.

Five tables are provided:

assessments

Built from numerous FOIA requests, this table includes information on assessments for residential properties from 2011 to 2021. 2022 tentative assessments are also included.

blight

See: https://data.detroitmi.gov/datasets/blight-violations/explore

parcels

See: https://data.detroitmi.gov/datasets/parcels-2/explore

parcels_historic

Parcel data from 2009

sales

See: https://data.detroitmi.gov/datasets/property-sales-1/explore

foreclosures

See: https://portal.datadrivendetroit.org/datasets/detroit-tax-foreclosures-2002-2019/about

Assignment

Submit the rmd and knitted output

- Section A: Conduct an exploratory data analysis of homes in Detroit. Offer an overview of relevant trends

in the data and data quality issues. Contextualize your analysis with key literature on properties in Detroit.

- Section B: Use `cmfproperty` to conduct a sales ratio study across the relevant time period. Note that `cmfproperty` is designed to produce Rmarkdown reports but use thedocumentation and insert relevant graphs/figures into your report. Look to make this reproducible since you'll need these methods to analyze your assessment model later on. Detroit has many sales which are not arm's length (sold at fair market value) so some sales should be excluded, but which ones?
- Section C: Explore trends and relationships with property sales using simple regressions
- Section D: Explore trends and relationships with foreclosures using simple regressions

PART 2

Objective: Now that you have a decent understanding of the landscape in Detroit, create a new file (part_2.Rmd) which builds upon part_1 in the html report Rmarkdown style.

Submission: submit to Blackboard both your code and the knitted Rmarkdown output.

Part A

Create an ‘introduction’ to your report. Generally, only include stylized output (do not use base R print). This could mean using stargazer to show regressions, DT::datatable to show data.frames, and adding titles/labels to plots. Your introduction should include:

1. Brief background (2-3 sentences) on issues in the Detroit assessment space
2. 3 to 4 graphs with descriptive captions which include information on sale price, assessment accuracy, foreclosures, and outliers. Generally focus on single family homes and arm’s length transactions. While it is notable that so many properties are sold for small amounts, we typically only want to look at properties which are class 401, taxable (e.g. assessed over 2000 or so), and sell above $4,000.

Part B

We have two separate (but very related) problems we want to model. First, we want to find a way to identify if a home is likely to be overassessed in a given year. We will analyze homes and assessments from 2016. We will use tidymodels to create a workflow.

1. Create your workflow
2. Add to your workflow a classification model

3. Add to your workflow a recipe of preprocessing steps. Use 2016 sales and assessments with the parcels property characteristics (note that we only know if a home was overassessed if it sold). Create a classification metric of overassessment based on properties which sold and use this as your **dependent** variable. Explain how you decided to construct this metric and how many classes it has.
4. Create testing/training data and evaluate your model using the classification metrics from tables 8.3 and 8.4 from the textbook and the classification probability metric ROC curves.

Part C

Second, building off of the workflow from part B. Create a second model to create your own 2019 assessments. (Note that I am choosing this year to avoid impacts from the pandemic and data quality issues. You may, if you'd like, create 2022 assessments. Limited sales data is released here.)

1. Create your workflow
2. Add to your workflow a model
3. Add to your workflow a recipe of preprocessing steps. Use sales and assessments from before 2019 with the parcels property characteristics.
4. Create testing/training data and evaluate your model using numeric metrics RMSE and MAPE.

Grading Overview

For each assignment, you will be graded on substantial completion of the assignment (demonstrated by an attempt of all parts). When submitting parts 2, 3, and 4, you will be additionally graded on your incorporation of feedback, new concepts from the course, or the correction of any flagged issues.

The assignment will culminate in a final submission of code/report and presentation. Code will be graded based on reproducibility, conceptual understanding, and accuracy. The report will be an Rmarkdown file which knits together graphs, tables, and ethical frameworks. It should be concise (include only relevant information from Parts 1-4). This report will be used to give a five minute presentation to the class on your model and ethical/technical issues with Detroit property assessment.