Econ 184b, Fall 2021, Final Project

Note on submission format:

- You are expected to turn in the problems in a **report format**: formulate succinct and coherent arguments and use econometric outputs as your evidence. Cite any reference if necessary. Just showing programming outputs will receive zero credit.
- You need to submit (1) the compiled pdf/html file; (2) the RMarkdown (.rmd) file; (3) any additional data you used on LATTE. You will get a bonus equal to 5% of your performance if your RMarkdown file is completely reproducible with minimal altercation (install packages, load data, etc.). That is, our team (or anyone else) should be able to recompile your Rmarkdown file and reach *the same* result.

Note on collaboration:

- You are allowed to work in a group of no more than two people. If you choose to work individually, you will receive a 10% bonus added to your final grade. Your grade can exceed 100% if you choose to work individually. State your team member clearly on the first page of your report.
- You are expected to complete this project individually/between your team members. Any help from others (excluding permitted individuals, details below) are not allowed. Any dishonest behaviors will result in an automatic zero on the final project, in addition to being referred to the Department of Student Rights and Community Standards.
- You are allowed to direct questions towards myself, the TAs, and/or other resources available at the institutional level, e.g., a research libriarian. Please be noted although our team is happy to offer high-level guidance, **we are under no obligation to help you debug your code and/or check your answers.**

The Question

You are provided with a random sample of actual housing market transactions in the state of New York. The dataset contains almost everything that is reported to tax authorities: buyer and seller names, sales price, year, structural information, etc.

The dataset is available at: https://github.com/f1kidd/Econ184/raw/main/Final%20Project/housing_transaction.csv. Legal notice: this is a sample from a proprietary data set. Any unauthorized use other than for instructional purposes is strictly prohibited. [1]

Choose one of the two following options:

Option (1) You are tasked by a leading real estate company to predict housing prices, mimicking the effort of Zestimate.[2] Use the provided data, construct a model that predicts historical housing market prices as close as possible. **You will be evaluate based on the out-of-sample performance of your model.** That is, your model will be used to predict a new set of housing transaction data.

Creativity in the modeling approach/choice of predictors will be rewarded.

Option (2) Use the provided data, answer an economic question of your choice. One prominent framework is the hedonic regression framework, which quantifies the value of public goods using observed housing market

[1] Although Brandeis does have restricted access to the Zillow transaction data, this sample is NOT associated with any Zillow product/data set.

[2] Also see https://www.bloomberg.com/news/articles/2021-11-08/zillow-z-home-flipping-experiment-doomed-by-tech-algorithms for a recent cautionary tale from Zillow.

price differences (Bishop et al. 2020 for a guideline). Some of the examples using this framework include quantifying school district quality (Black 1999), flood risk (Bakkensen and Barrage 2021), or hazardous waste sites(Greenstone and Gallagher 2008).

You could also embark on other journeys that are directly related to the housing market, for example race and discrimination in the housing market (Myers 2004) or the origin and extent of housing market speculation (Burnside et al. 2016).

You will be evaluated by (1) **the novelty and creativity of the research question** (estimating the value of an extra bedroom won't do the job); and (2) **the rigor of the modeling approach** (if you anticipate a problem with omitted variable bias, using different estimation strategies, e.g. panel model/instrumental variable/regression discontinuity, will come in handy).

General tips In order to make a successful prediction/inference, you will probably need to find additional information to augment the data set provided. For example, neighborhood characteristics, school districts, or crime rates will capitalize into the housing price, but are not currently included in the data set provided by the tax authority. A good place to start is the US census. A complete dataset of the 2018 census at the block group level can be downloaded here. A sample code on extracting information from this dataset is available here. The dataset can be merged into the housing transaction dataset using the census block group number (*census_bg*).

A successful report will need to present:

1. A brief description of the context
2. Choices you've made during the modeling process
3. Your results, illustrated by words and any supporting graphs when necessary
4. Attach a) the complete set of codes; b) any additional data you used, excluding the census dataset.