Department of City and Regional Planning
University of Pennsylvania

MUSA 508, Fall 2021
Profs. Michael Fichman & Matt Harris

Mid-term project:
Hedonic home price prediction

(Predictions due 10/20 by 5PM, Markdown due prior to class 10/22)

Zillow has realized that its housing market predictions are not as accurate as they could be because they do not factor in enough local intelligence. As such they have asked you and your partner (as well as several other teams) to build a better predictive model of home prices for Boulder, CO.

To do this, you will gather as much data as you can from websites like Boulder's Open Data portal and Tigris. Use Chapters 3 & 4 of *Public Policy Analytics* to inform your analysis. **You must stick to OLS regression.** There are many more powerful predictive algorithms out there – but for this exercise these models are strictly off limits.

You must work in **teams of 2** (signup here). This is a competition. The team with the lowest average error will win. The first deliverable for your assignment is a brief **R Markdown** (driven by dataviz) that explains to me, the non-technical manager, how you undertook your analysis. I am a visual learner – so make it pretty and make sure your narrative is clear and concise. This is what your project grade will be based on. Please submit your reports **digitally** as html documents to a folder that will be linked on Piazza. Please use code folding and make sure there is limited additional output from code chunks besides the intended visualizations.

The second deliverable is around predictions. The primary *training* dataset is **'studentData.geojson`** which you use to train your model. You will notice a field called `toPredict`. Where that field is set equal to '1', I have set SalePrice equal '0'. This is the "challenge set" - the sales you are predicting for. You should remove these from your training set **but you need to have create features for these sales so you can predict**. You will receive instructions on how to prepare your predictions. Once you do, you will submit those predictions and we will learn as a group who the winners are.

The winning team will be the one that is able to find the best predictive 'features' or variables and pour enough predictive power into the model to predict well without overfitting to the training data. Remember, the key is to have a model that 'generalizes' to both new data and to different neighborhood contexts.

Any dataset that might include sales information or be derived from sales information **cannot be used**. One example is assessment data used for taxes. That data is generated from a predictive model – so predicting on predictions is cheating.

About the writeup - Written in R Markdown

You should assume your audience is a manager not a data scientist. Try to break down the technical details to a more general audience. The report will have the following deliverables:

Introduction: What is the purpose of this project? Why should we care about it? What makes this a difficult exercise? What is your overall modeling strategy? Briefly summarize your results.

Data:

-Briefly describe your methods for gathering the data.

-Present a table of summary statistics with variable descriptions**.** Sort these variables by their category (internal characteristics, amenities/public services or spatial structure). Check out the `stargazer` package for this.

-Present a correlation matrix

-Present 4 home price correlation scatterplots that you think are of interest. I'm going to look for **interesting open data** that you've integrated with the home sale observations.

-Develop 1 map of your dependent variable (sale price)

-Develop 3 maps of 3 of your most interesting independent variables.

-Include any other maps/graphs/charts you think might be of interest.

Methods:

-Briefly describe your method (remember who your audience is).

Results: Briefly interpret each in the context of the Zillow use case

-Split the 'toPredict' == 0 into a separate training and test set using a 75/25 split.

-Provide a **polished table** of your (training set) lm summary results (coefficients, R2 etc).

-Provide a **polished table** of mean absolute error and MAPE for a single **test set.** Check out the "kable" function for markdown to create nice tables.

-Provide the results of your cross-validation tests. This includes mean and standard deviation MAE. Do **100** folds and plot your cross-validation MAE as a histogram. Is your model generalizable to new data?

-Plot predicted prices as a function of observed prices

-Provide a map of your residuals for your **test set**. **Include a Moran's I test and a plot** of the spatial lag in errors.

-Provide a map of your predicted values for where 'toPredict' is **both** 0 and 1.

-Using the **test set** predictions, provide a map of mean absolute percentage error (MAPE) by neighborhood.

-Provide a scatterplot plot of MAPE by neighborhood as a function of mean price by neighborhood.

-Using tidycensus, split your city into two groups (perhaps by race or income) and test your model's generalizability. Is your model generalizable?

Discussion: Is this an effective model? What were some of the more interesting variables? How much of the variation in prices could you predict? Describe the more important features? Describe the error in your predictions? According to your maps, could you account the spatial variation in prices? Where did the model predict particularly well? Poorly? Why do you think this might be?

Conclusion: Would you recommend your model to Zillow? Why or why not? How might you improve this model?