Department of City and Regional Planning — Fall 2019
University of Pennsylvania — Dr. Ken Steif

MID-TERM PROJECT:
HEDONIC HOME PRICE PREDICTION

(Due October 17 at 5pm for the competition & October 20th for the written submission)

Zillow has realized that its housing market predictions aren't 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 San Francisco.

To do this, you will gather as much data as you can from websites like San Francisco's Open Data portal. The primary dataset is **midterm_data_sf_studentVersion.** You should use two Boston chapters, and the coffee shop markdown to guide your model. **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 with someone in your lab section** (signup here). This is a competition. The team with the lowest average error will win $150 each and the runner up team will win $100 each. We will give $100 each to the team that wins 'Best in Show', by having a written submission that best communicates their algorithm in both narrative and dataviz form.

The first deliverable for your assignment is a report that explains to me, the non-technical manager, how you undertook your analysis. This should be written in **R Markdown.** 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 in this folder.

Winning the cash is all about predictive accuracy. There is a field called 'holdOut which denotes observations for sale prices that have been changed to '0'. These are the sales you are predicting for. You should obviously remove these from your training set **but you need to have features associated with these prices 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.

Remember: The winning team will be the one that is able to do two things. First, to find the best predictive 'features' or variables. Second, pour enough predictive power into the model to predict well without overfitting to the data I've withheld. You want to create a model that is 'generalizable' to the many different places in and around SF.

Any dataset that might include sales information or be derived from sales information **cannot be used**. One example is assessment data such as this one. If you have a question about a dataset then ask, or risk being disqualified.

About the writeup

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).

-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:

-Provide a polished table of your in-sample (training set) model results.

-Provide a **polished table** of R^2, mean absolute error and MAPE for the **test set.** Check out the "kable" function for markdown to create tables.

-Provide the results of your cross-validation tests on the training set. This includes mean and standard deviation MAE. Do **100** folds and plot your cross-validation MAE as a histogram. Is your model overfit?

-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.**

-Provide a map of your predicted values for the **entire** dataset

-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 in to 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 predictor variables? 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?

GOOD LUCK!