1, Business Question:

How do we improve the accuracy of Zestimate?

2, A brief overview of dataset and key processing steps I took:

Data overview:

- properties_2016 data(only has home-features info) with (2985217rows, 58cols).
- train_2016_v2 data(only has transaction info) with (90275rows, 3cols). Key processing steps:

- Merged two datasets.
- Dropped na and eliminated missing values by filtering the home feature related

  columns with values>0 only(i.e. number of bedroom).

- Transformed 'transactiondate' to 'transaction_year_month' by:

  X['transaction_ym'] = 100 * X['transactiondate'].dt.year +
  X['transactiondate'].dt.month -(i.e. from 2016-01-27 to 201601, easier to work with.)

• Performed feature engineering and decided on the key variables for the later calculation.

-(i.e. 'bathroomcnt', 'bedroomcnt', 'taxamount', 'yearbuilt', 'calculatedfinishedsquarefeet', 'transaction_yearmonth_i', 'log_error')

• Split the data into train and test.

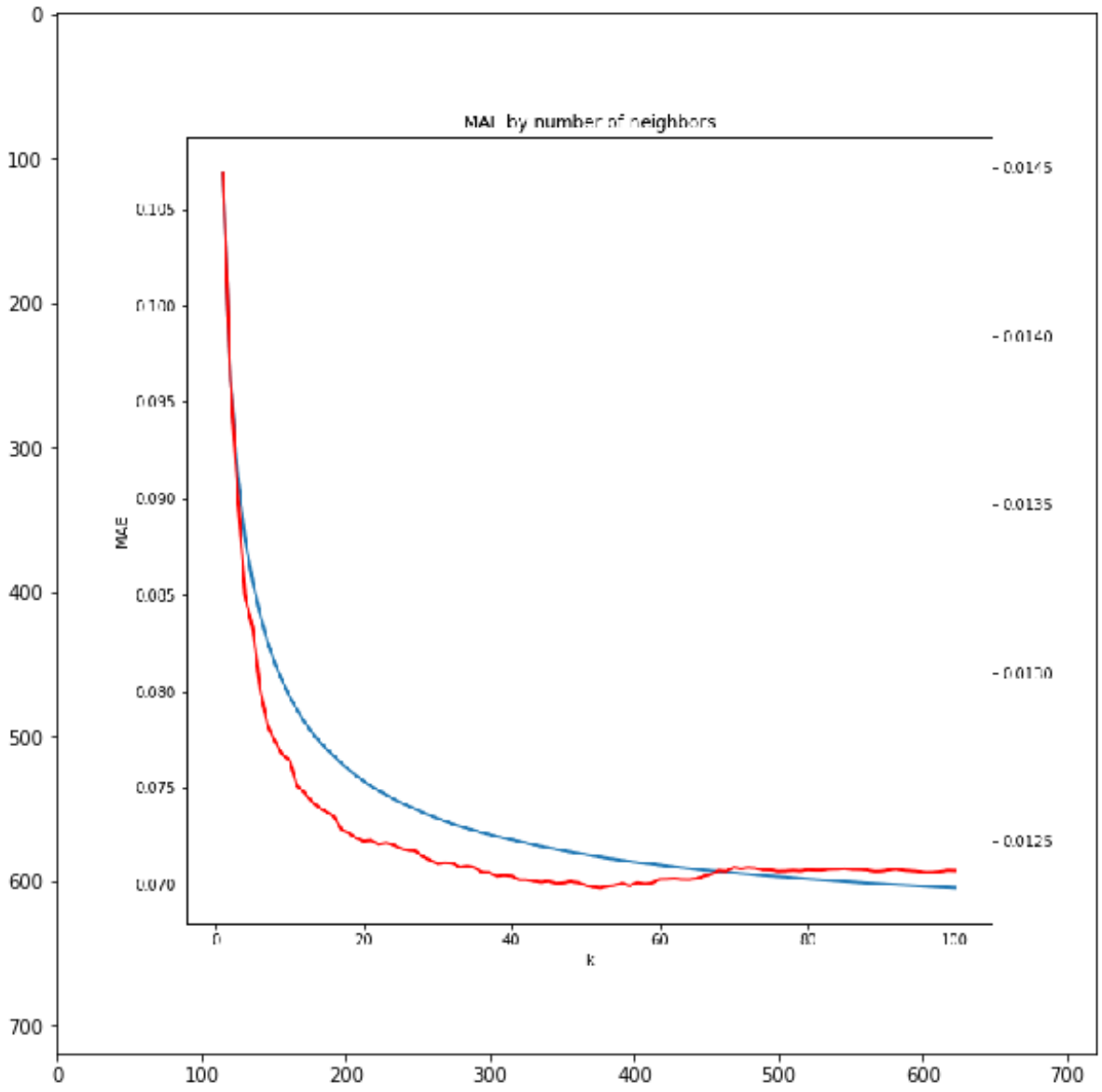
3, Overview of modelling approach implemented:

Chose K-Nearest-Neighborhood as the final model.

Reason: Tobler's first law of Geography, "everything is related to everything else, but near things are more related than distant things."

Next: Calculated cross_val_score on number of neighbors from 1 to 100, finally decided number of k to be 50 (i.e. neighbors).

Fit the model.



[footnote:the lower the score the better the performance, since our scoring='neg_mean_absolute_error'.]

4, Key findings and recommendations:

- Larger houses are easier to predict than the smaller ones.
- Built year does not have too much impact on the log_error prediction.

• The tax amount is the most important variable.

5, How would business use the model and model output to answer the business question:

Now we are able to predict log_error correctly. How is Zillow benefiting from this?

These residuals are the unexplained parts of the current model used by Zestimate. Predicting the unexplained part correctly will help them just add the new predictions on top of the existing ones, resulting in prediction improvement.

Furthermore, since June 2018, Zillow started buying homes with an average price of 250,000 dollars and resell within 3 months. Let's say if we can help them improve Zestimate's accuracy by 1%, Zillow can either save up to 2500 dollars on one property or just change their decision and stop wasting extra money.

6, **Potential next steps or further research topics**:
Study the log_error outliers, since these are the most interesting datapoints. More importantly, this is where the Zillow algorithm fails. If we can predict these failures/log_error, it would improve the Zestimate significantly.