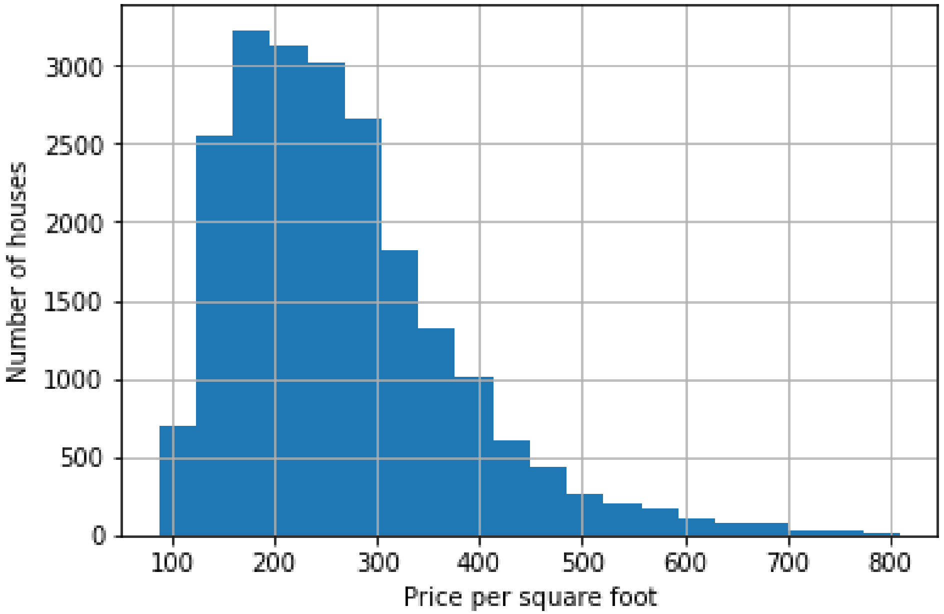
Applied Data Science Capstone

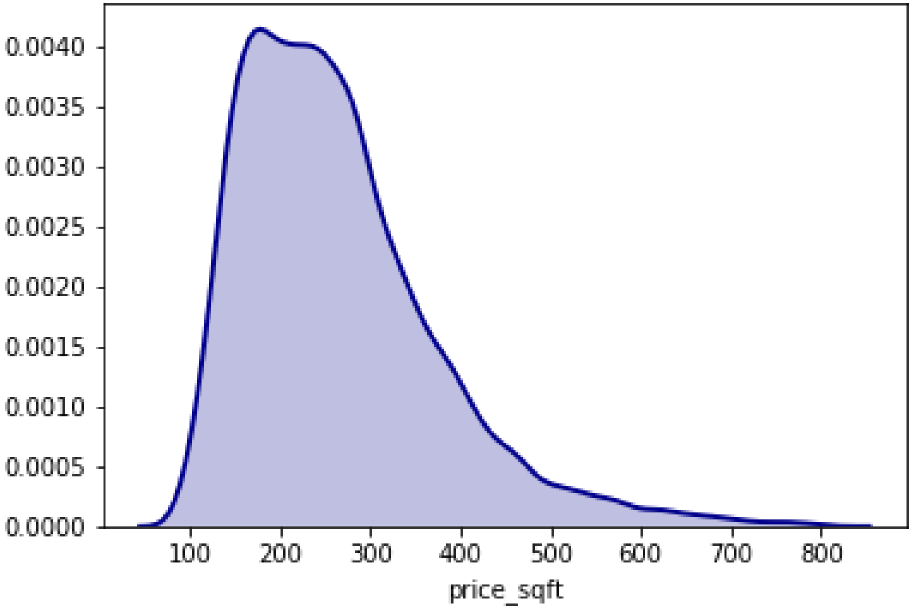
Final Assignment

General information

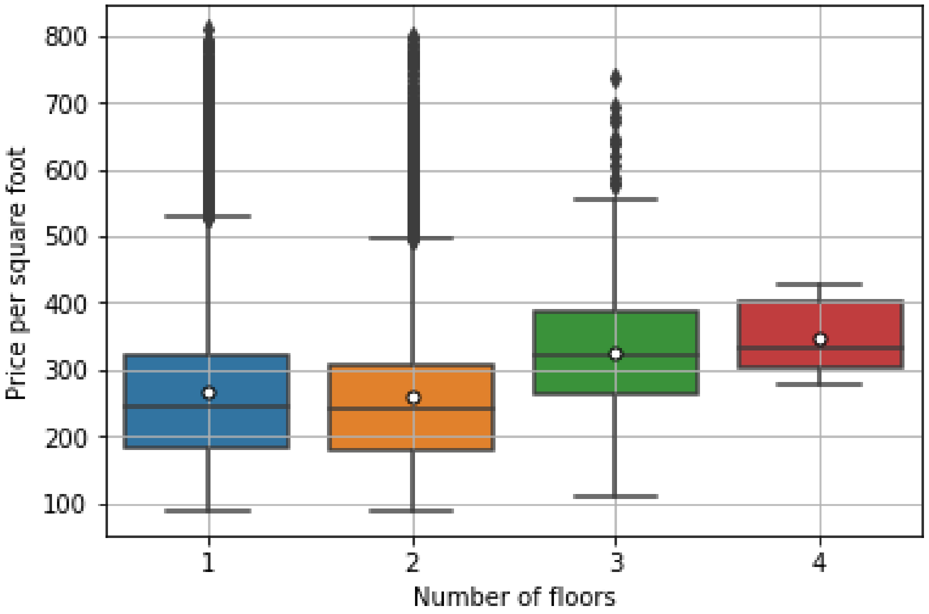
- The goal of this project is to build a predictive model for house prices with using both, standard variables as well as additional location-related features
- Kaggle repository as a source of data
- Dataset consists of sale prices of houses located in King County (state of Washington) with characteristic of each property
- Data of a good quality – above 20 000 observations, 21 variables and no missing values, only few suspicious observations
- In addition Forsquare API have been used to get the most common venues of given regions in King County.

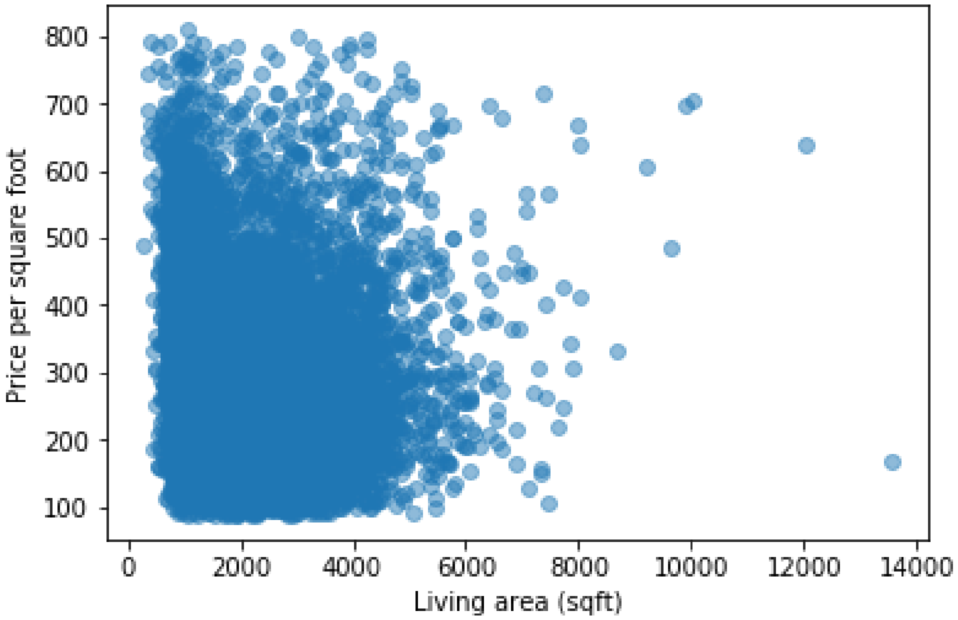
Target variable





Independent variables (e.g.)

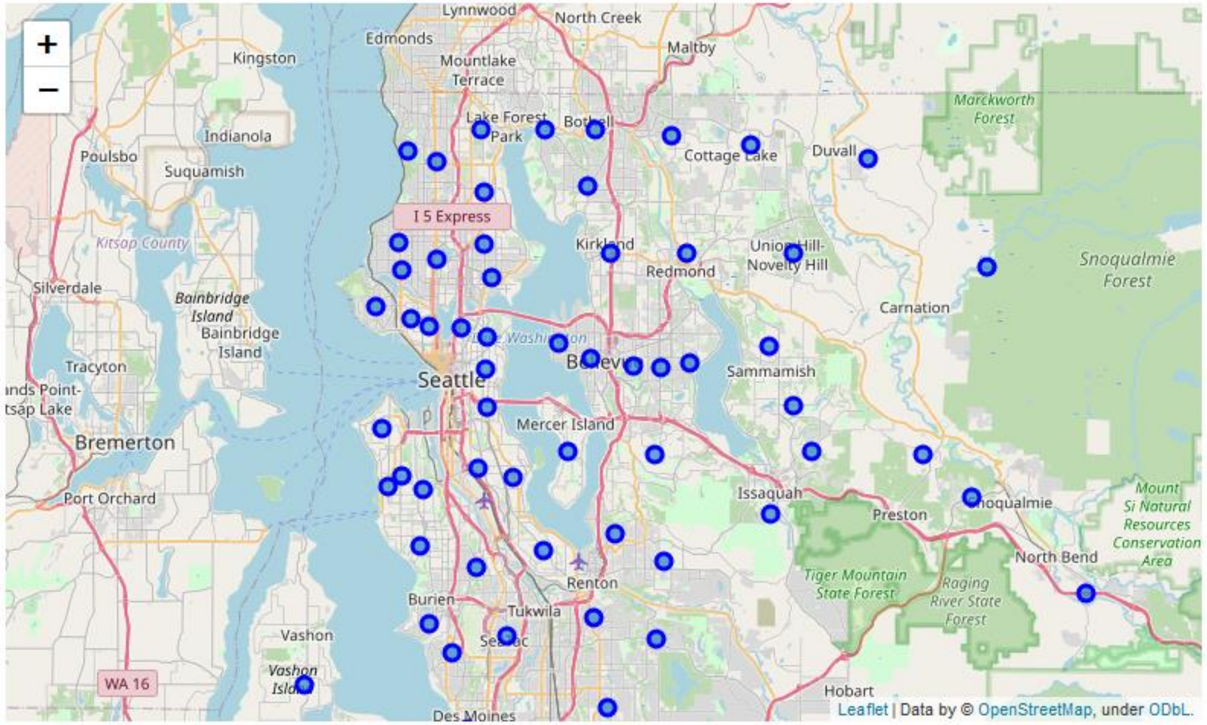


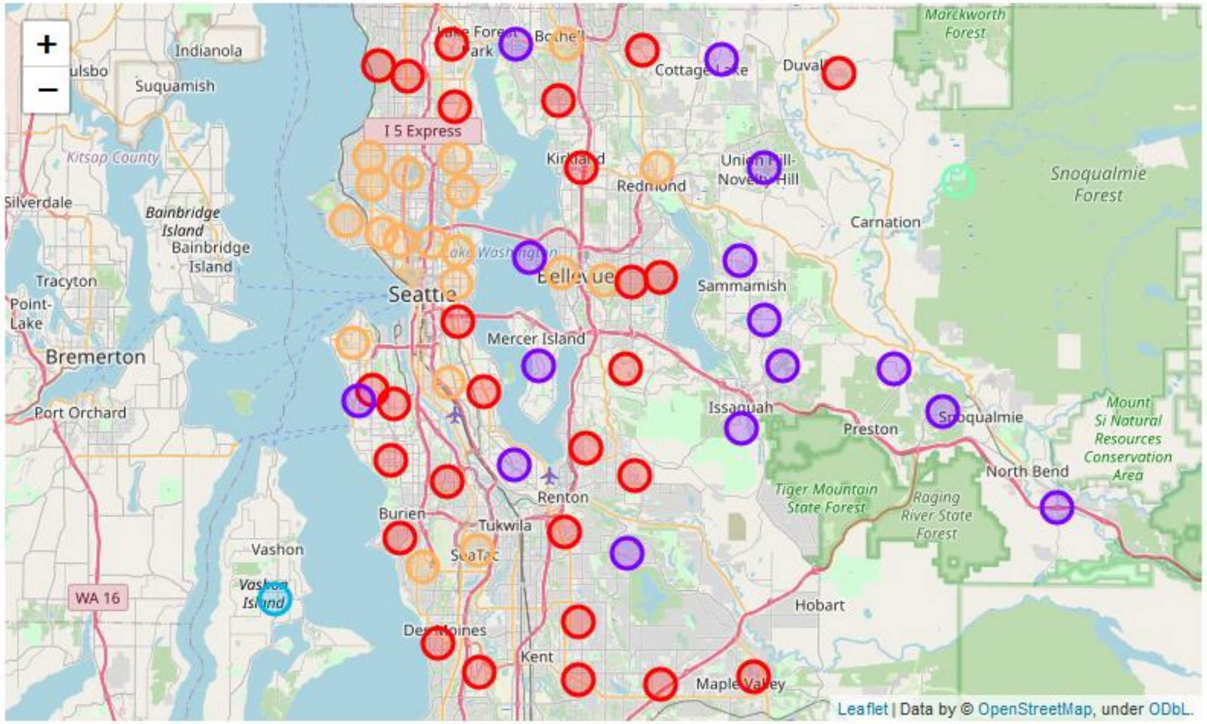


Basic cleaning & feature engineering

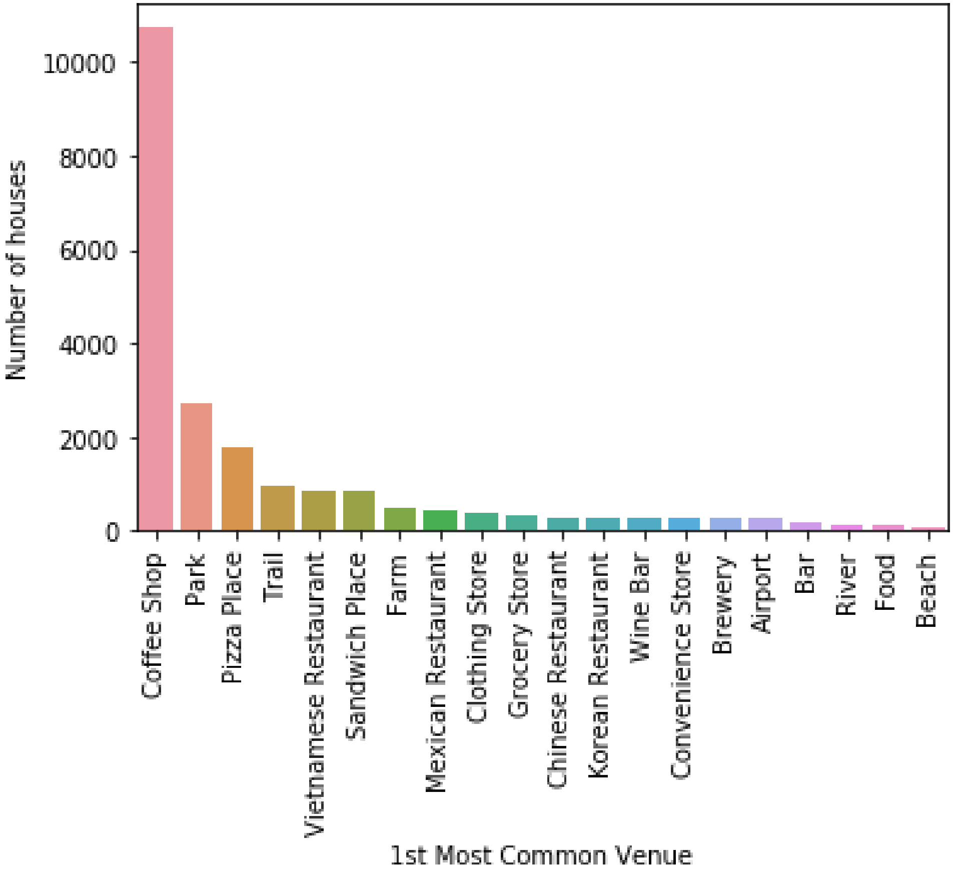
- Removal of few extreme observations – possibly typos (eg. house with 33 bedrooms)
- Treatment of duplicates
- Creation of new features
- And others...

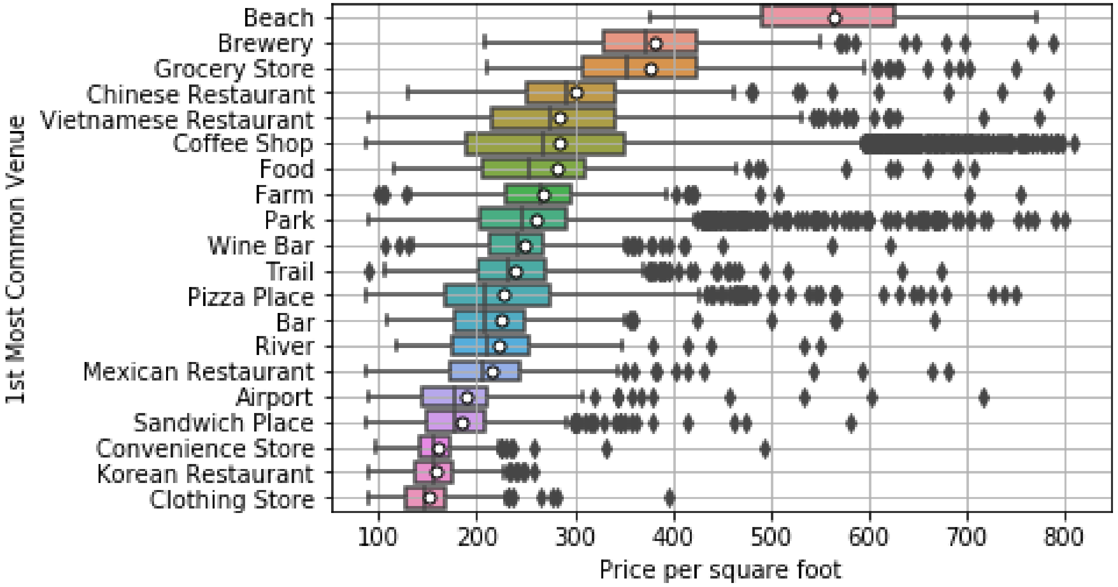
Regions segmentation and clustering





New features based on Foursquare location data





Preparation for model estimation

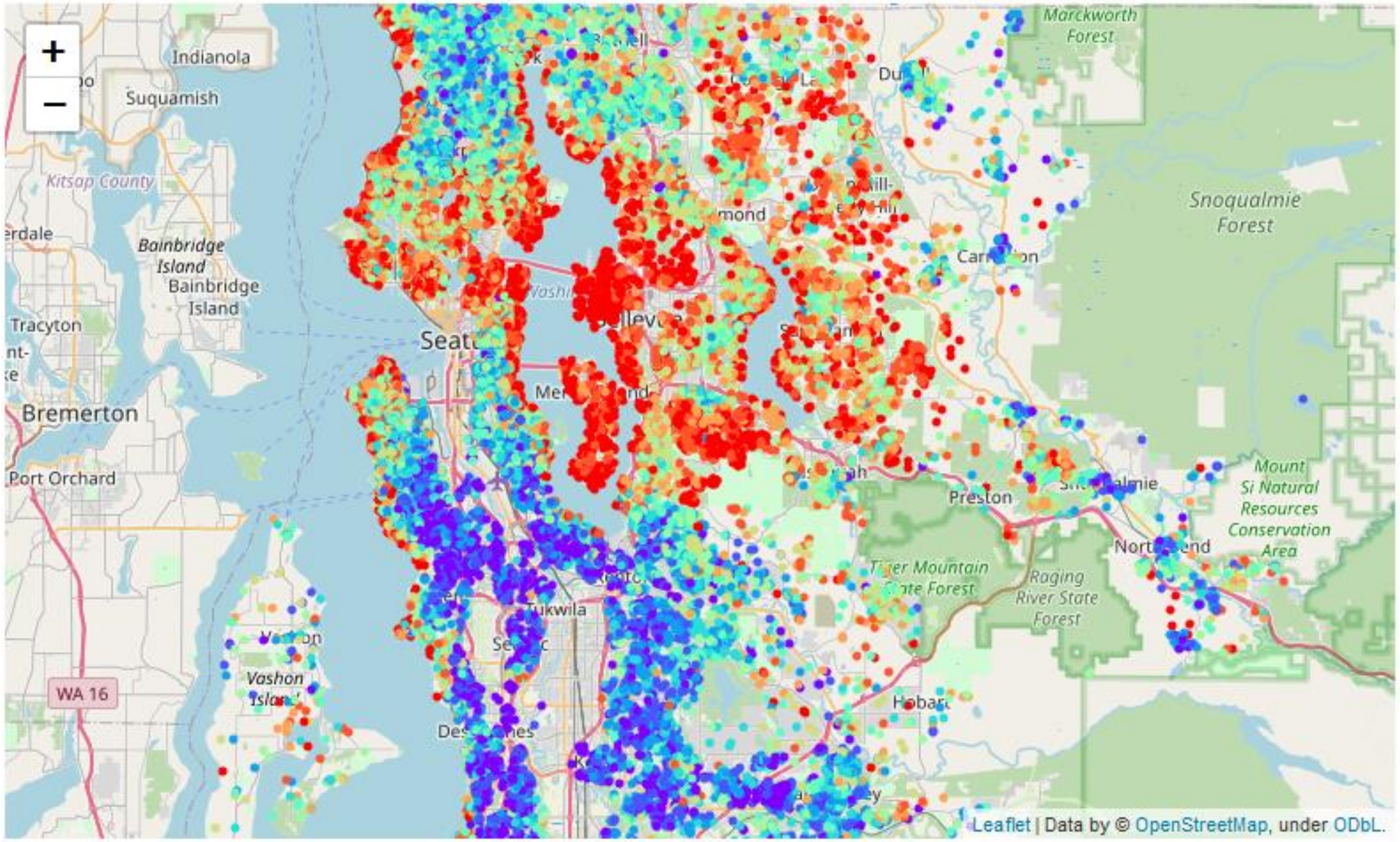
1. Creation of function that calculates Mean Absolute Percentage Error (MAPE), which will be used for model evaluation.
2. Transformation of categorical variables into dummies for linear regression.
3. Preparation of numeric/categorical variables for catboost algorithm.
4. Data split into train / test sample (80% / 20%)

Model comparison

Model type	MAPE (train sample)	MAPE (test sample)
Linear regression	15.43%	15.67%
Catboost	9.74%	11.55%

Visualization of house prices

(red color – high price; blue color – low price)



Conclusions

- Catboost model provides significantly better results than Multiple Linear Regression in case of these data.
- One of the new variables created based on Foursquare data (Cluster Labels) is among 5 most imporant features in our model (based on feature importance). Hence we can consider that Foursquare data gave some value added to our prediction.

Thank you