The Battle of the Neighborhoods

Using location data to predict house sale prices

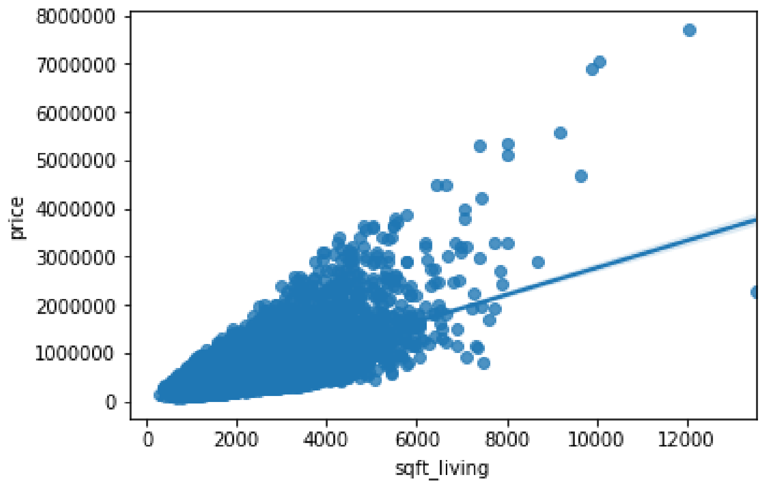
Making better house sale price predictions using location data

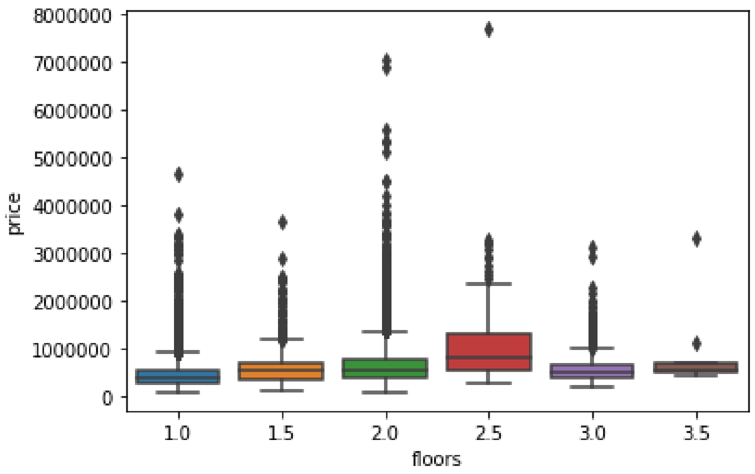
- An accurate prediction on the house price is important to a lot of different stakeholders as
  - prospective homeowners and real estate agencies
  - developers and investors
  - mortgage lenders, insurers etc.
- Existing models usually contain a lot of information about the house but none about the neighborhood
- Including location data in the prediction model could:
  - Enhance the prediction
  - Reduce the number of models needed for different areas

Data

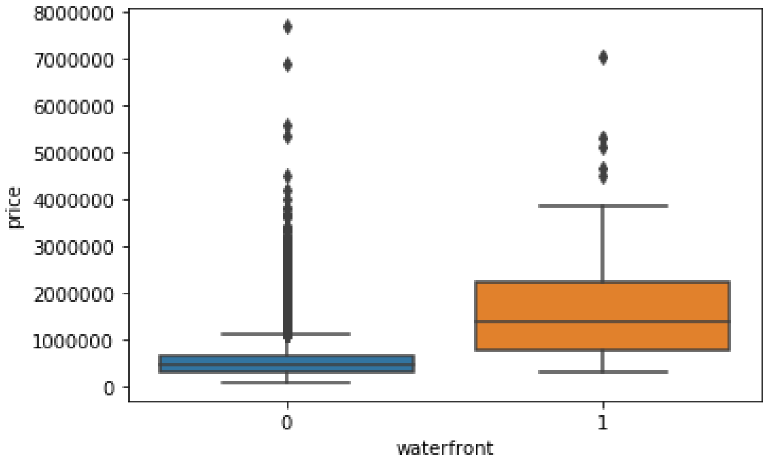
- House sales data from King County from 2014-2015 from Kaggle including:
  - house sale prices
  - information on the sold objects as square footage, number om floors, number of bathrooms etc.
  - 21 613 houses
- Foursquare location data using Foursquare API
  - Collected information on venues within 500 meters for each house
  - Summarized the data in to 203 variables indicating total number of venues from each cathegory in the neighborhood
  - PCA was used to reduce these 203 highly correlated variables in to 4 new uncorrelated variables that account för xx% of the variance in the dataset.

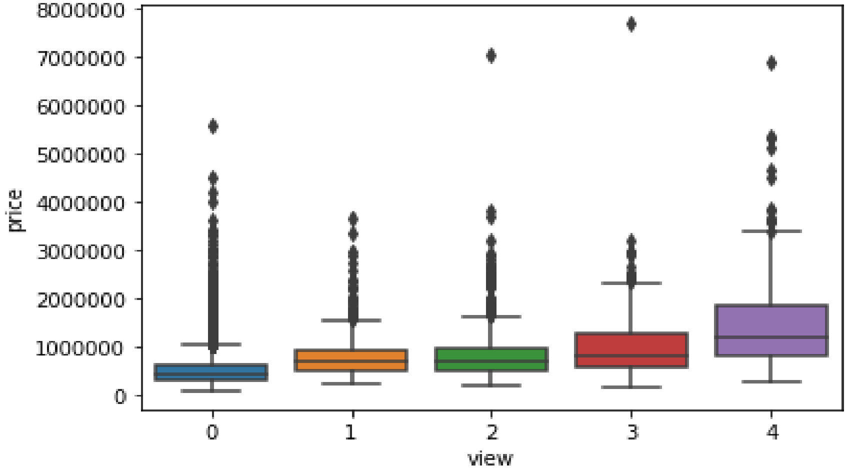
Square foot living area and number of floors affect the house sale price



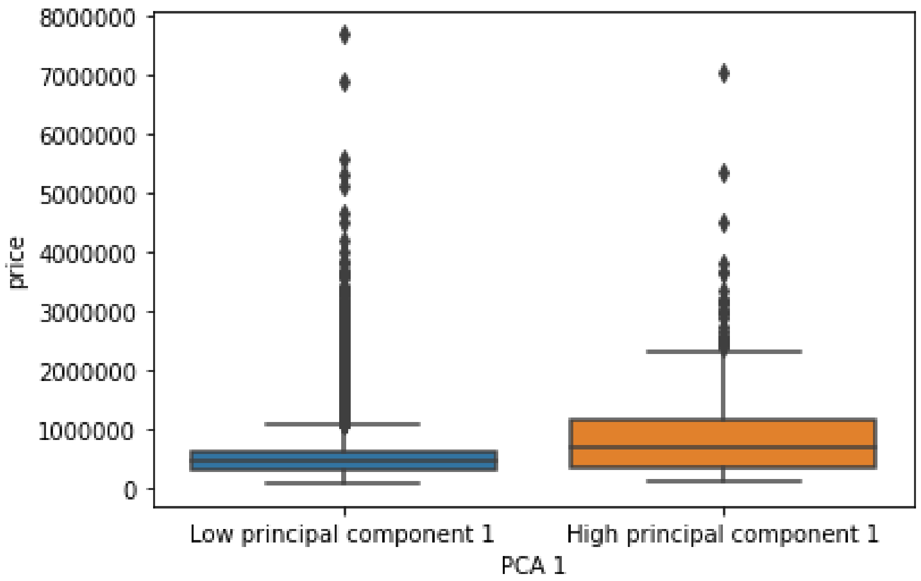


Having a waterfront or a good view increases the house sale price



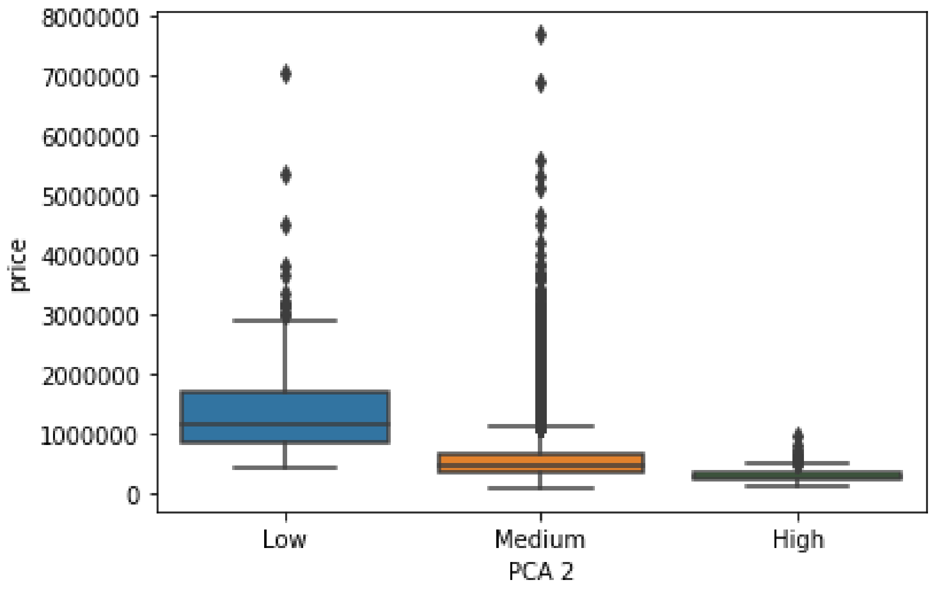


Areas with a wide range of services available have higher house sales prices



- Houses in an area with lots of restaurants, shops, services and entertainments generally have higher house sale prices

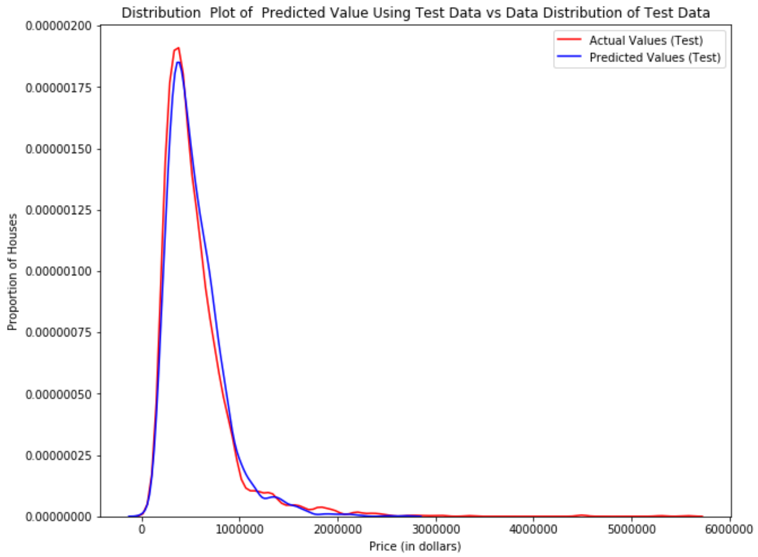
Houses in an area with cheap restaurants and discount stores have lower house sale prices



- Houses in an area with cheap restaurants and burger joints with discount stores and no mall or nice restaurants have lower house sale prices.
- And the ones in an opposite area have much higher house sale prices then other houses

Results

Model	R-square
Model without Principal Components	0.56
Model with Principal Components	0.61



- By adding location data (PC 1 and PC2) to the model I achieve a 0.05 improvement of the R-square value
- The distribution of predicted house prices follows the distribution of actual house prices quite well, indicating a good model fit.
- Houses with a house sale price over 4 million dollars are however underestimated by the model

Conclusions and further directions

- Location data can be used along with historical house sales data to predict house sale prices.
- Further studies are needed
  - Room fore improvement of the model
  - Can the results be generalized?
  - Could adding ratings of venues give us more information?