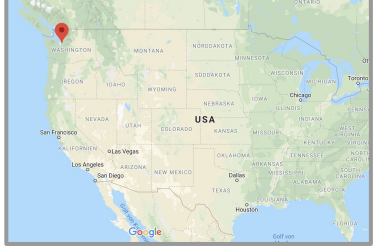
King County House Prices

Is the Region important for Modelling?

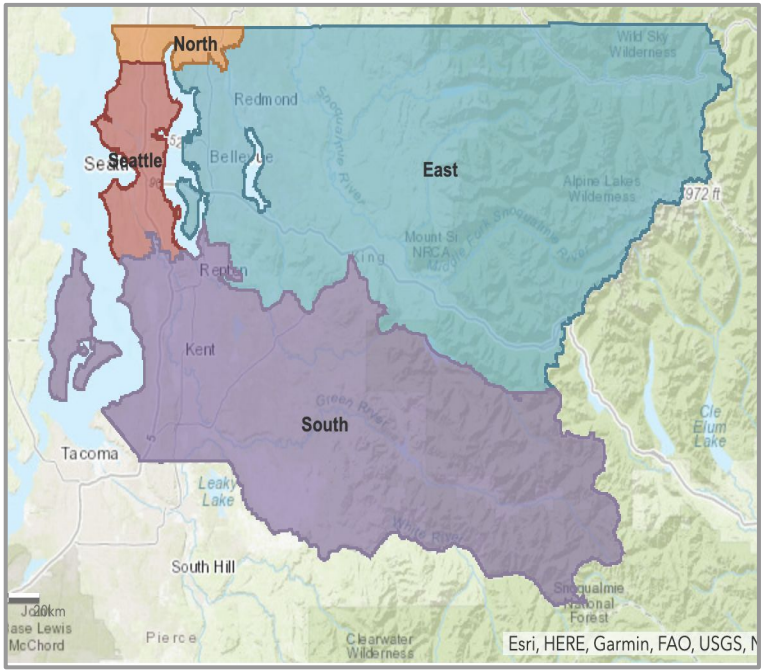
Goal: Predict House Prices!

- Data: prices of 21523 houses
- The houses are described by different characteristics (21 in total), e.g.
  - Number of bedrooms/bathrooms/floors
  - Interior living space/lot space of the house
  - Living area in the basement and in higher floors
  - The year they were built/renovated
  - A general grade (based on King County grading system)
  - Location!

Location and Regions



- North-West of USA
- State: Washington
- ~2Mil. residents, ~6,000km² area
- Regions with different characteristics
  - Seattle: city,~700,000 residents
  - East: mountains
- Divide in 4 regions



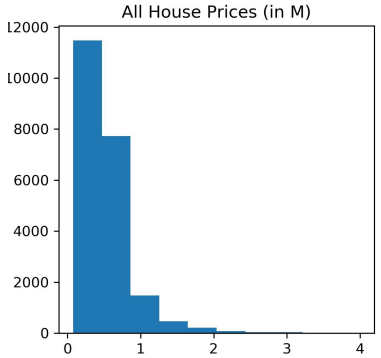
➔ Do we need a specific prediction model for each region?

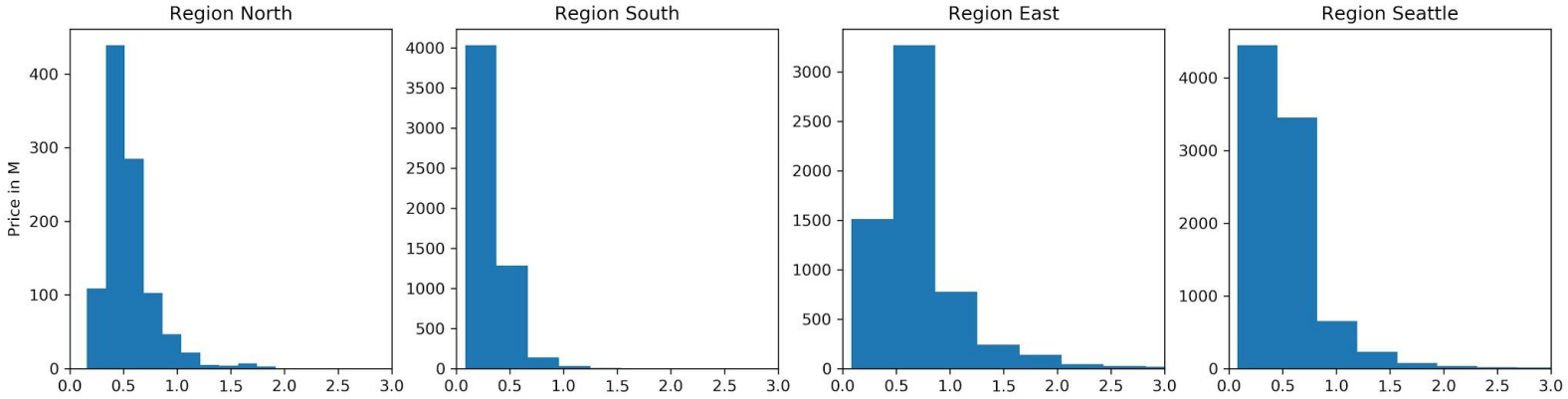
Source: https://www.kingcounty.gov/depts/health/data/community-health-indicators/definitions.aspx

Overview Prices (in Mil.$)

Overview about the prices for the entire dataset and for the 4 different regions

Note: Different amount of Observations in each dataset





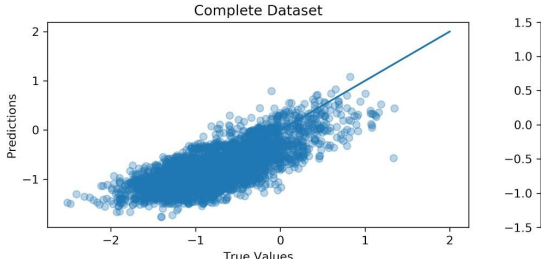
The model

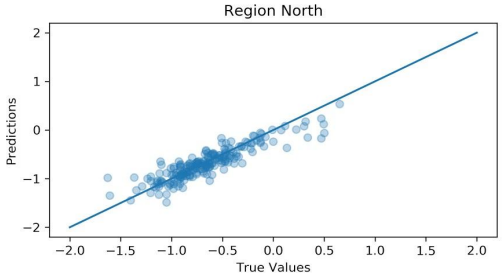
- From the 21 characteristics available in the dataset, chose 2:
  - Interior living space (in squarefeet)
  - Grade (index between 1-13 for building construction and design
    (1: lowest value, 13: highest value)

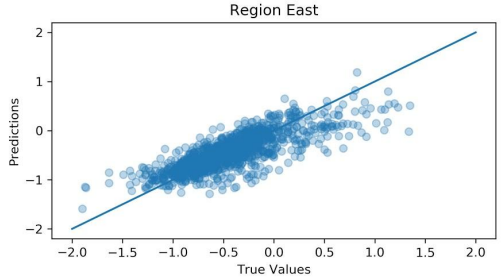
- Predict the price depending on these two characteristics

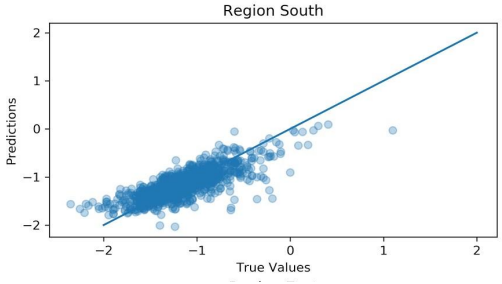
How is the Performance?

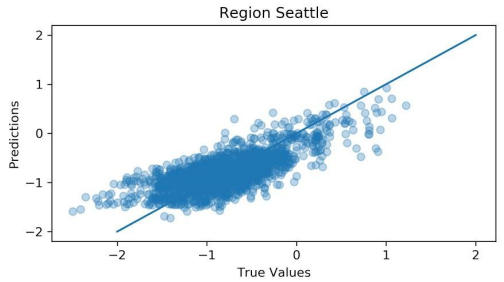
Region	R^2 (adj)	MSE
Complete Dataset	0.54	0.13
North	0.8	0.027
South	0.64	0.052
East	0.64	0.072
Seattle	0.54	0.12











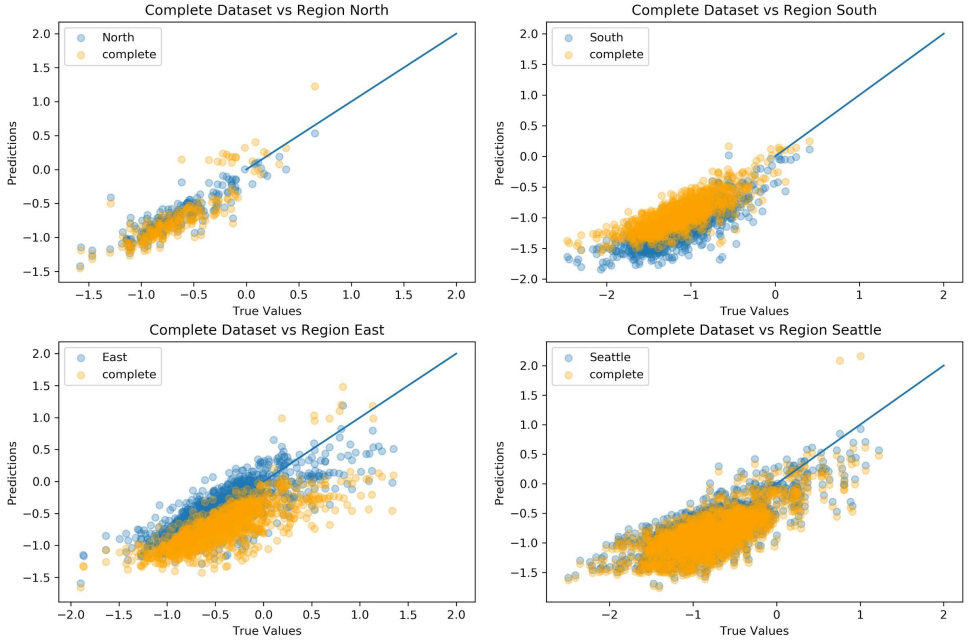
Back to the Beginning ...

Is it necessary to work with different models?

Evaluation of the Models

Predictions using the model for the **entire dataset (orange) vs.** model for **each region (blue)**

- ➔ Different results for region South and East
- ➔ Keep in mind: datasets have different sizes!



Conclusion

There are regions where the prediction is more accurate when adopting the model

Outlook

- Repeat the analysis with datasets of the same sizes
- Improve the model:
    - Consider other characteristics
    - Analyse more complex models
    - Include more characteristics not available from the dataset, which are important for each region (e.g. shopping centers nearby, climate,...)
- Review the regions, are there better criterias for the division
- Analyse the city of Seattle in more detail: Maybe an adaption for different districts is reasonable

Thank you.