Location recommendations for a poor Data Scientist

How to successfully relocate to King County, WA

Table of Content

Business Case

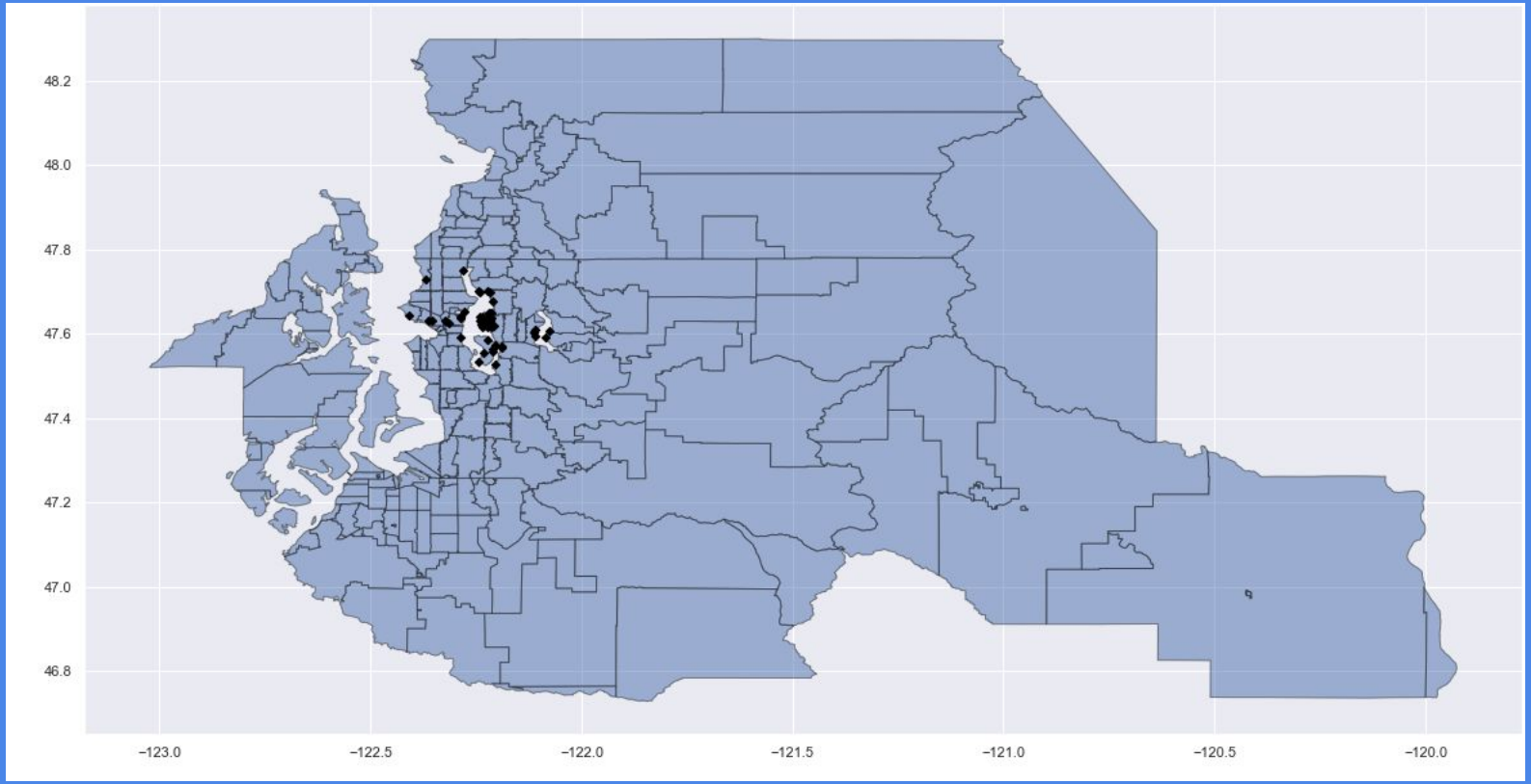
- my friend is an extraordinary good, but poor Freelance Data Scientist - lets call him Bendix ;)
- looking for a partner to found a Data Science Company in King County, WA.
- wants to work from home
- wants to help rich people to accomplish private goals with data
- he wants to test me (his potential partner), if I can come up with recommendations

Business Case - considerations

- he is poor and wants to put everything back into our business, so no luxury needed for a home
- it is important to be near the clients, if they want a presentation, they want it quick and in person, they dont have much time and they want it in their private homes
- prediction model that helps us to find an affordable house near our clients
- open a box office address for client mail in a good neigbourhood and start a membership in a gym in that area (will force us to drive there and meet potential clients)

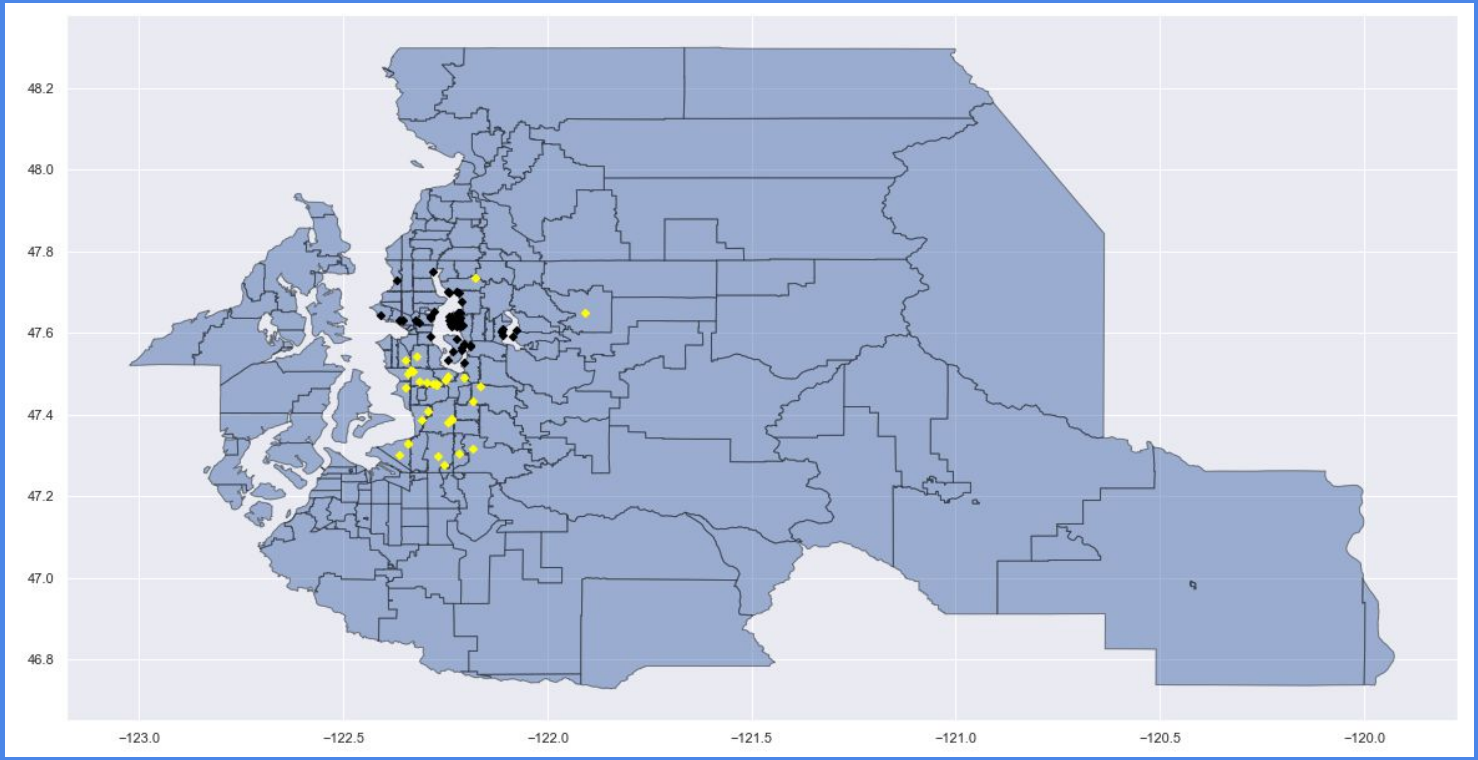
Where do the rich live?

house prices >= 3,000,000



How to get near them with little money?

house prices <= 100,000



Observations:

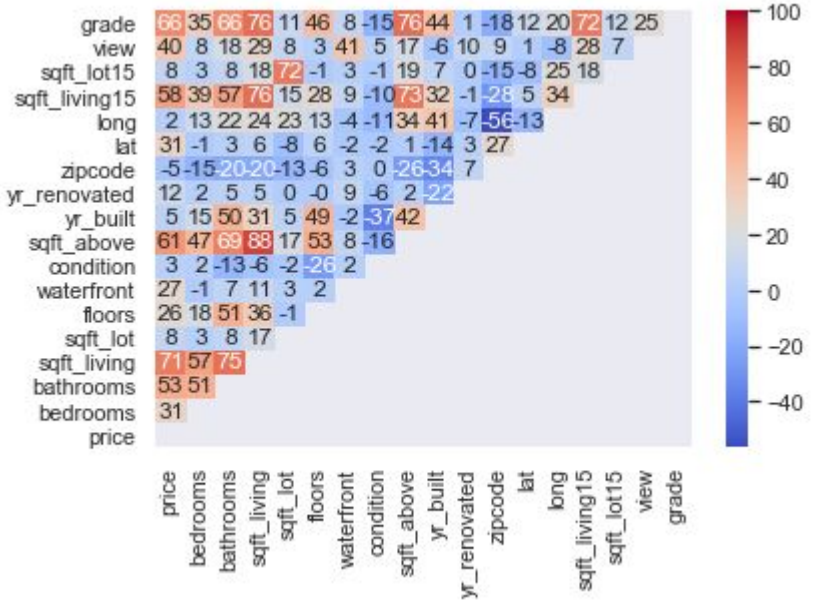
- rich people like to live near the water
- poor people live more in the south
- there are some houses that are cheap and near our potential clients
- why might they be so cheap?
- **expensive houses zipcodes:** 98074, 98034, 98033, 98040, 98004, 98008, 98155, 98144, 98039, 98105, 98109, 98102, 98199, 98112, 98006, 98119, 98075, 98177, 98056
- **cheap houses zipcodes:** 98014, 98001, 98168, 98146, 98198, 98178, 98002, 98166, 98058, 98055, 98023, 98108, 98032, 98034, 98106, 98092

American Red Cross

Mailing Address:
P.O. Box 3097, Seattle, WA 98144

Lets look into the data

- i got rid of all the missing values
- with a heatmap I looked at possible relationships in the data and decided to go with the following features:
  - 'sqft_living'
  - 'grade' (according to the King County grading system)
  - 'waterfront'
  - 'bathrooms'
- p-values of the features showed significance

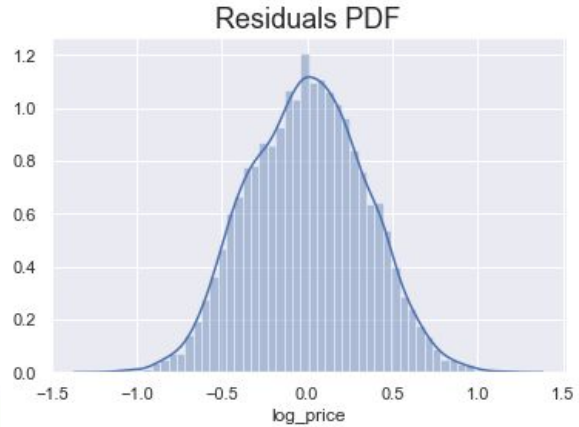


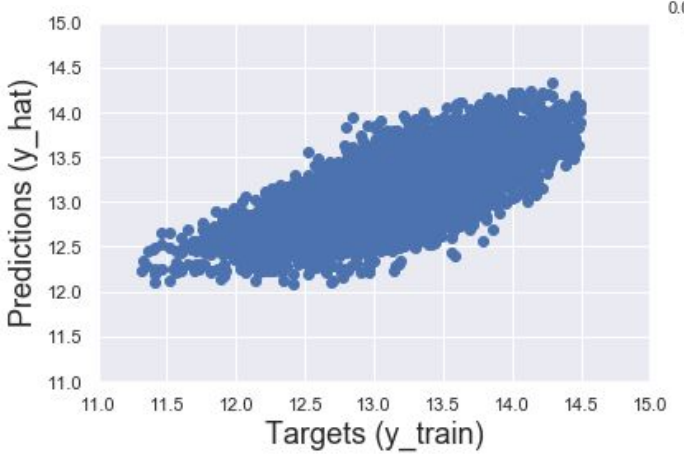
Some words to the resulting linear model

- train test split 80/20
- positive relationship between bathroom/sqft and price

Results

R_quared = 0.49

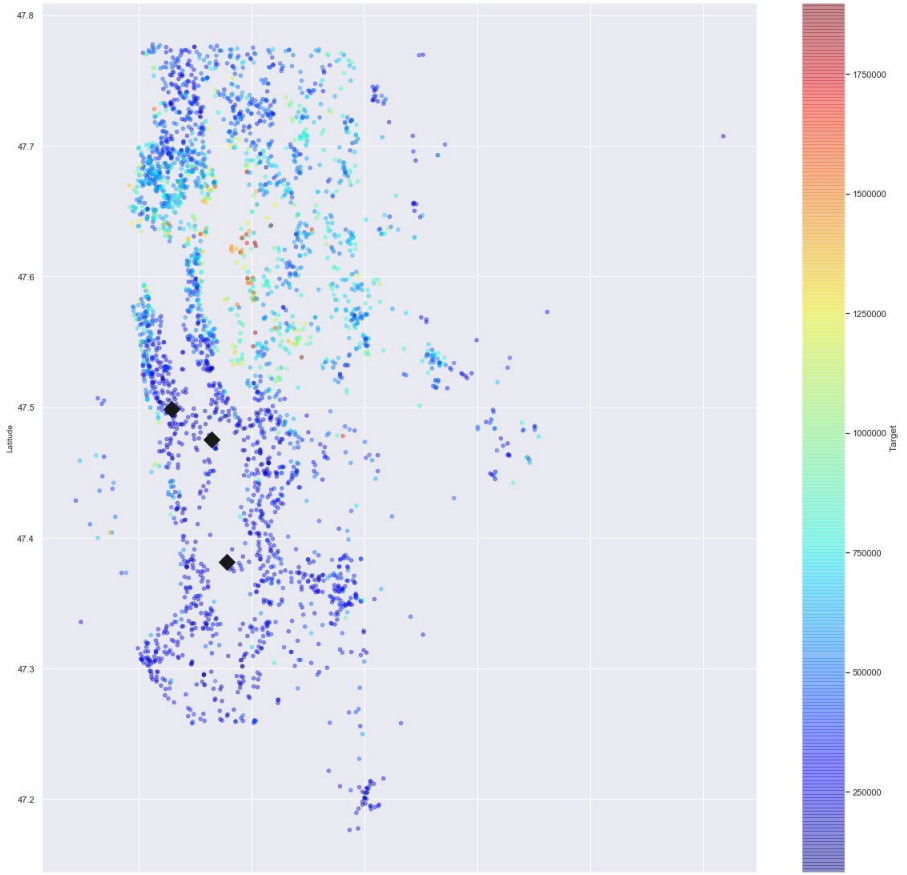




	Prediction	Target	Residual	Diffrence%
count	3043.00	3043.00	3043.00	3043.00
mean	471416.66	492691.96	21275.30	29.06
std	185339.39	259633.26	177447.56	24.08
min	189408.81	82000.00	-536430.64	0.02
25%	351568.28	310000.00	-95563.85	11.58
50%	409036.27	435000.00	-4108.20	23.68
75%	534498.08	611553.00	110733.03	40.06
max	1501853.31	1900000.00	1105135.08	219.50

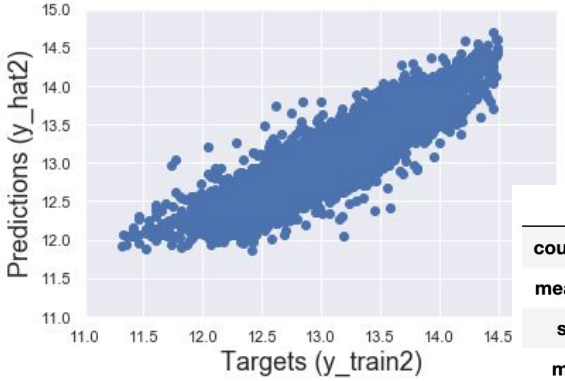
How can I improve my model?

- houses where our price model totally failed (3 worst predictions in map)
- overestimates the price (potential bargains in these areas?)
- there must be points that we have missed:
  - Age of the House
  - Area where the House is located

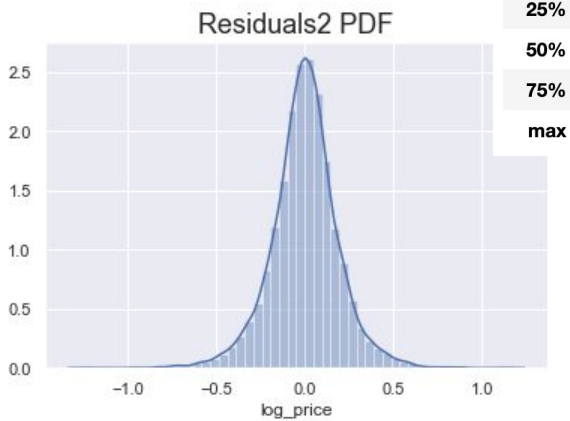


New model with results!

- included 2 new features called:
  - yr_built_renovated (is the year built or the year of renovation if available)
  - zipcode
- p-values of the features showed significance
- R_quared = 0.84

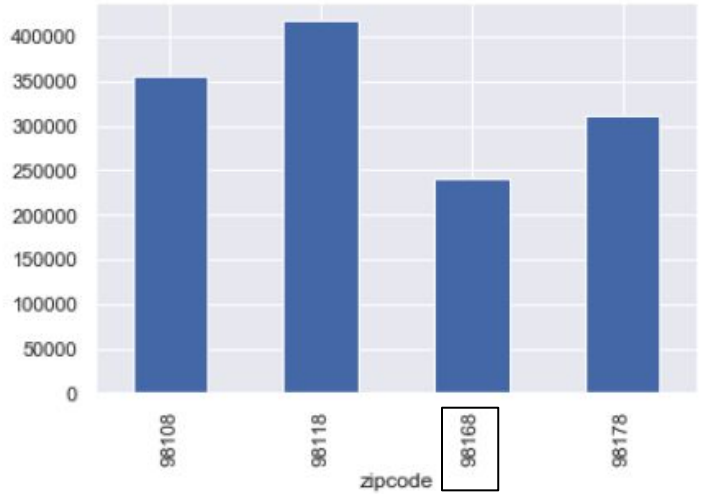


	Prediction	Target	Residual	Diffrence%
count	3042.00	3042.00	3042.00	3042.00
mean	486102.92	492767.80	6664.88	14.65
std	240772.85	259642.23	100141.31	15.26
min	146886.57	82000.00	-429768.82	0.00
25%	314873.16	310000.00	-43833.71	4.84
50%	433943.95	435000.00	1320.27	10.46
75%	584710.52	611726.50	46557.19	19.43
max	2027144.23	1900000.00	639376.08	182.24



Recommendations:

- more in the south
- just one bathroom
- not too young
- small
- definetly not by the water
- choose the right zipcode



Example for the predictive power of the model:

1.5 bathrooms / year 1963 / 1120 sqft / no waterfront / zipcode 98031

model: 228,127 $ / real price: 239,950 $.

Future Work:

- polynomial regression
- more features (condition, etc.)
- better data cleaning
- detailed analysis of all zipcodes
- better code efficiency
- improvement of map visualization

Thank you!