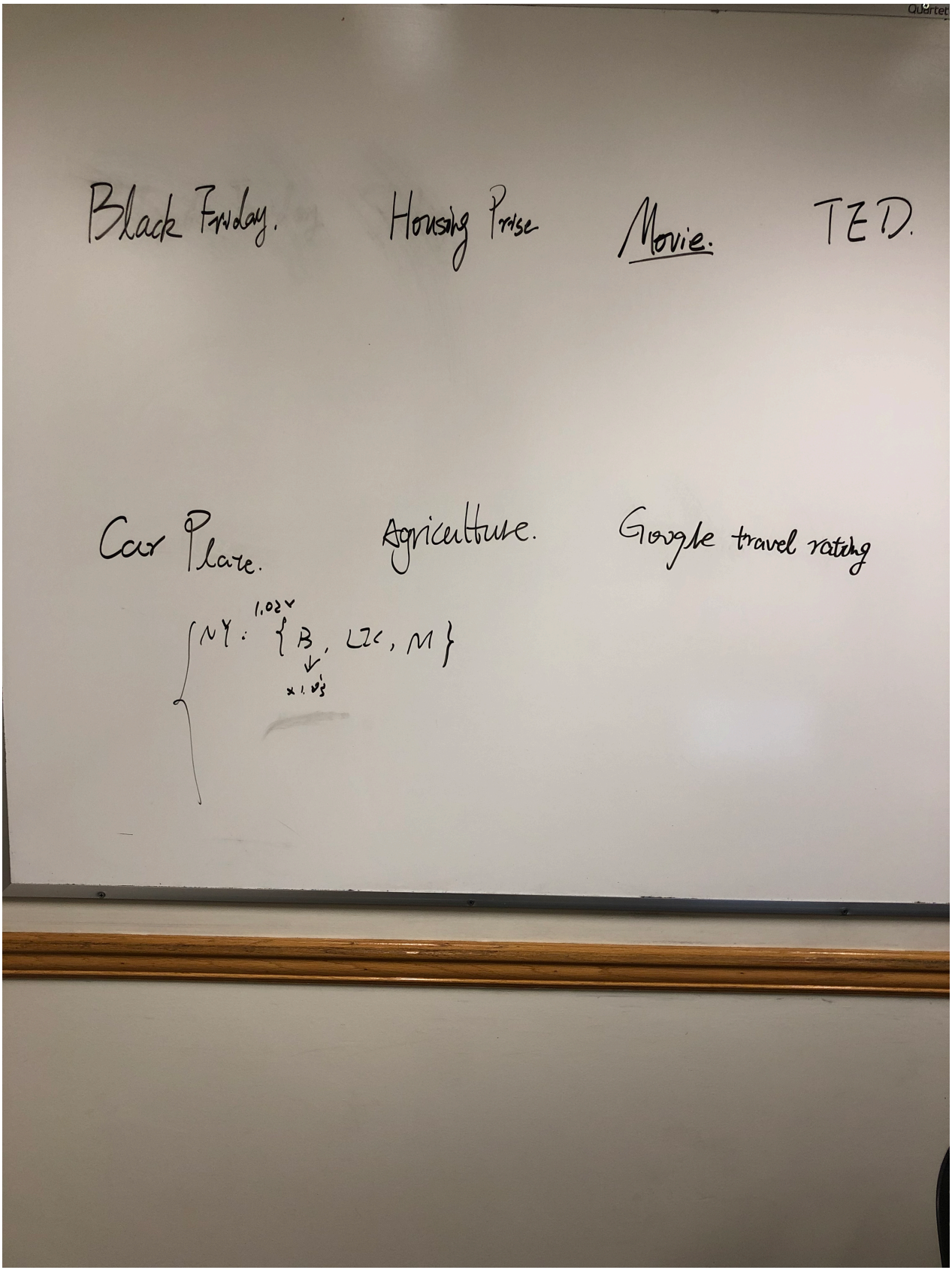
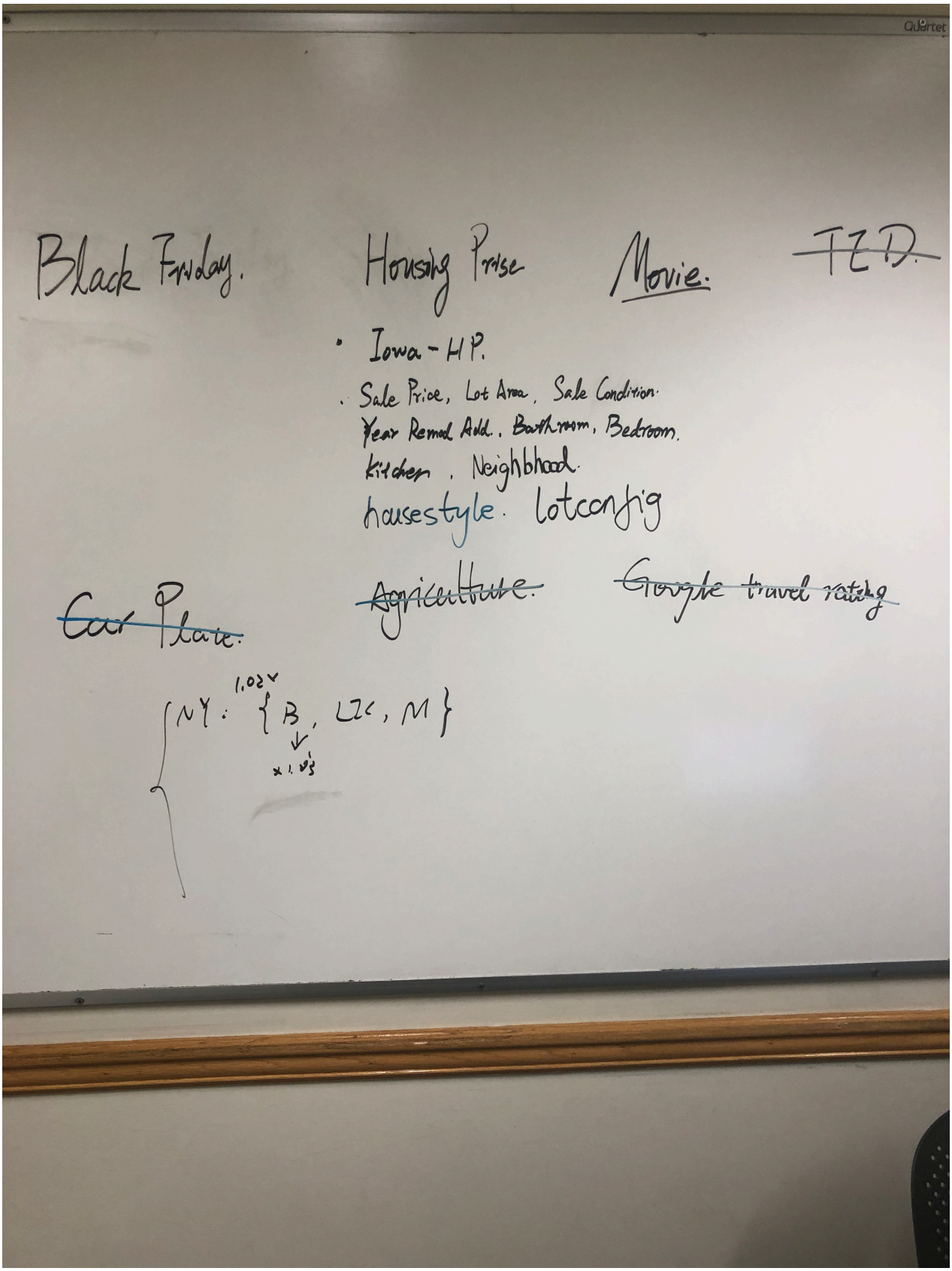
Topic Selection

At first, we find and come up with some interesting topics, including:

- Black Friday prom
- Housing price prediction
- Movie rating/prediction
- TED topic prediction
- Car license plate recognition
- Agriculture
- Google travel rating



After discussion, we selected the Housing Price Prediction as the final topic.As this is a popular and familar topic for us, and there are abundant resources we can refer to on the web.Then we come up with some interesting features that are related to the housing price.



General Procedure

- Data visualization. Select some features manually from the data and visulize the data to get an intuitive recogtion of the data set.
- Data preprocessing.Do preprocessing like normalization and filling in the blank on the data

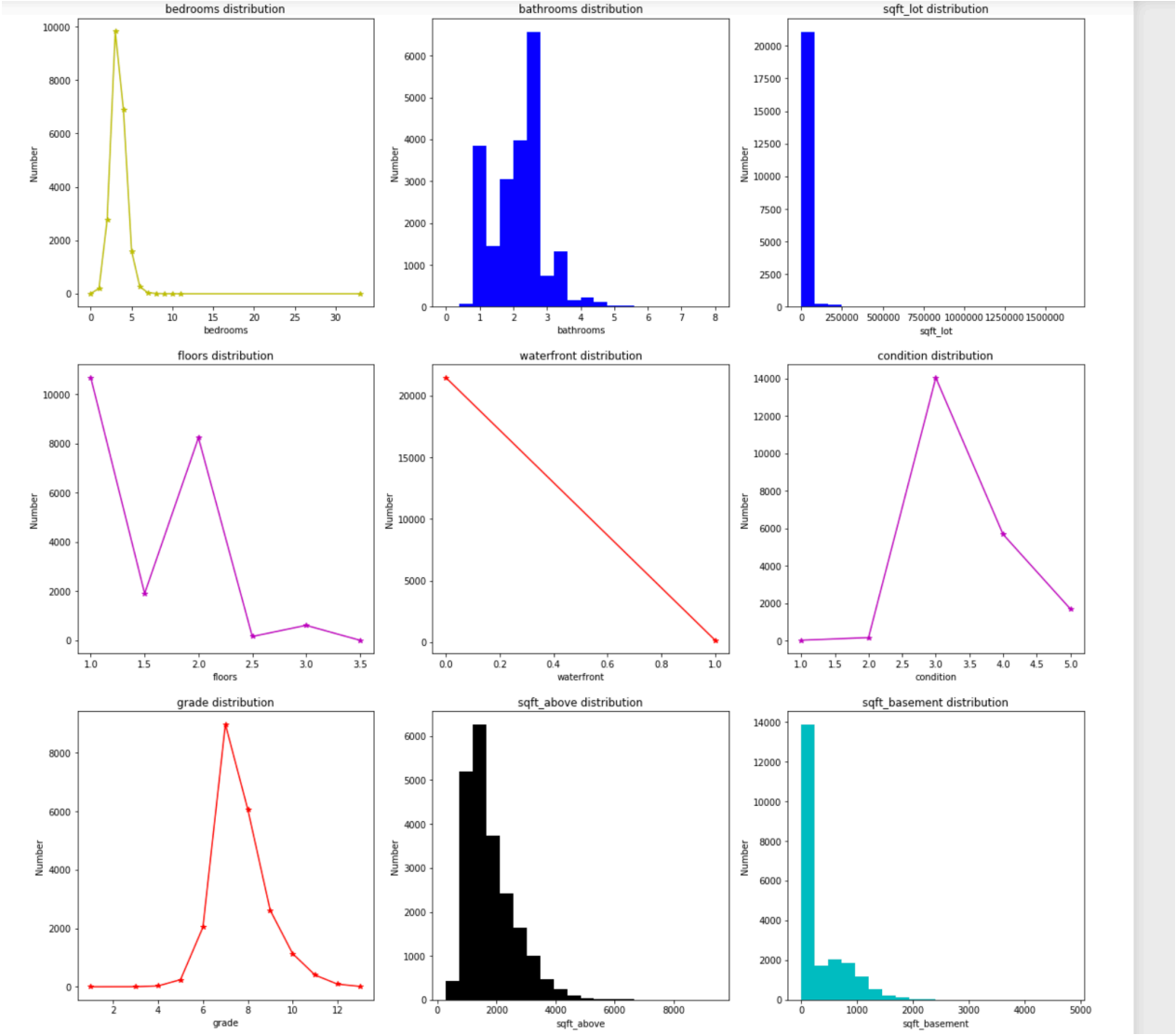
to improve the validation accuracy.
- Model selection.Try to use some models we studied or from external resources to train the data, and use K-fold validation methods to evaulate the error on the test data to select the best model.
- Evaluate the model.Select the best model and retrain the model with some extreme value removed to check whether it imporves the performance.
- Generate the summary.

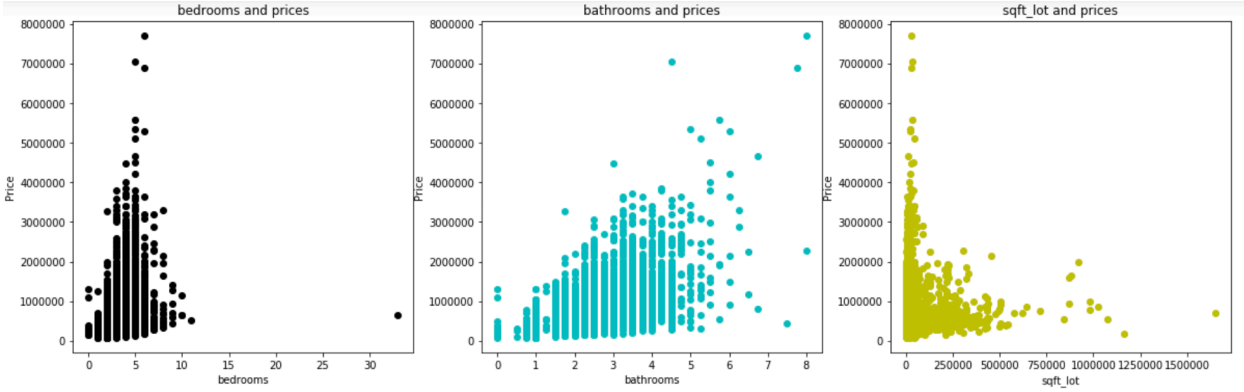
Data Visualization

We selected 9 features that are most relevant to house price from on data set.

- bedrooms
- bathrooms
- sqft_lot
- floors
- waterfront
- condition
- grade
- sqft_above
- sqft_basement

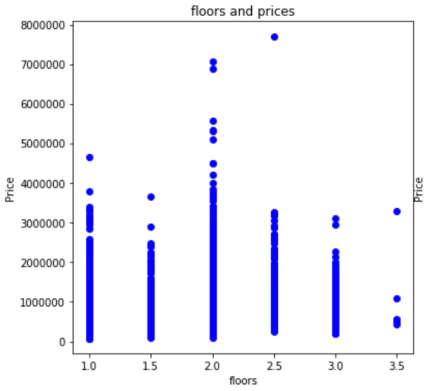
1. Distribution. For every feature, we first compute each distinct value and the number of samples equal to the value in the training data.If there are no more than 20 distinct values, we plot a line graph, otherwise we plot a histogram with 20 blocks.This gives us an intuitive distribution of the feature values.
2. Correlation with the price.For every feature, we plot a scatter graph showing the feature value and the price.Usually all features are in the positive correlation with the price attribute.We can briefly draw a conclusion from the graph which feature has more weight in deciding the price.
3. Correlation of each feature.Calculate and plot the heap map of each pair of attributes.And may exclude the features with the least correlation with the price in future training.

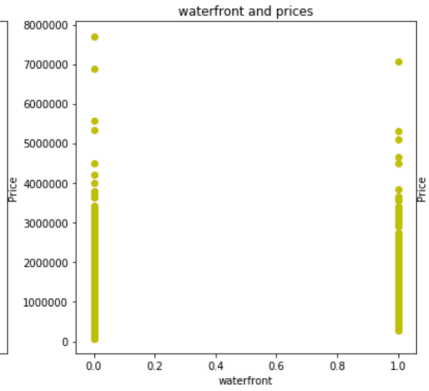

bedrooms distribution
Number
bedrooms
bathrooms distribution
Number
bathrooms
sqft_lot distribution
Number
sqft_lot
floors distribution
Number
floors
waterfront distribution
Number
waterfront
condition distribution
Number
condition
grade distribution
Number
grade
sqft_above distribution
Number
sqft_above
sqft_basement distribution
Number
sqft_basement

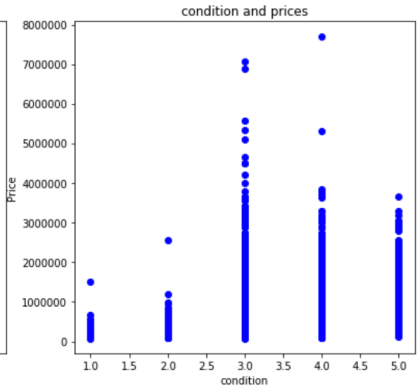

bedrooms and prices
Price
bedrooms

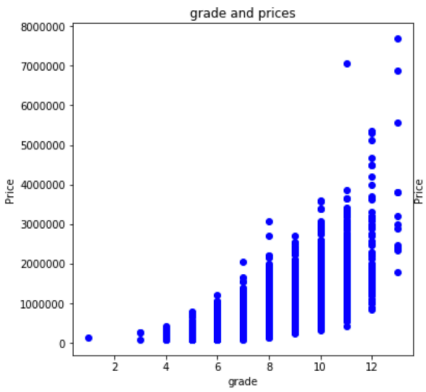
bathrooms and prices
Price
bathrooms

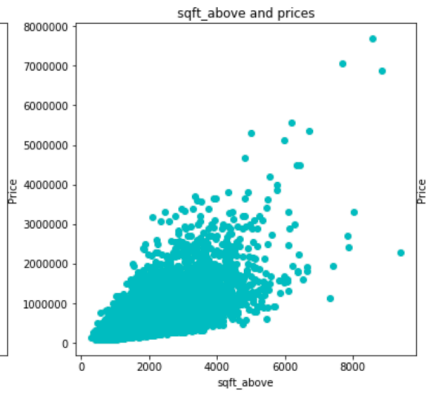
sqft_lot and prices
Price
sqft_lot

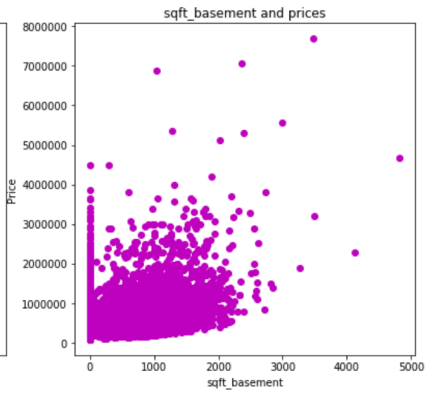

floors and prices
Price
floors

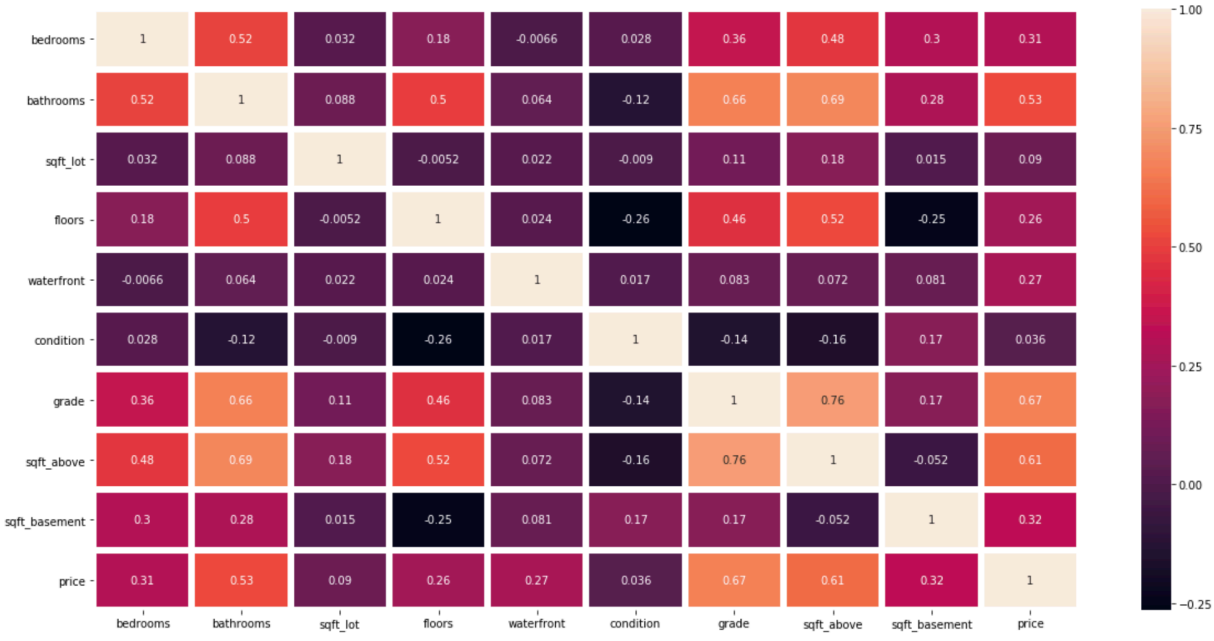

waterfront and prices
Price
waterfront


condition and prices
Price
condition


grade and prices
Price
grade


sqft_above and prices
Price
sqft_above


sqft_basement and prices
Price
sqft_basement


bedrooms 1 0.52 0.032 0.18 -0.0066 0.028 0.36 0.48 0.3 0.31
bathrooms 0.52 1 0.088 0.5 0.064 -0.12 0.66 0.69 0.28 0.53
sqft_lot 0.032 0.088 1 -0.0052 0.022 -0.009 0.11 0.18 0.015 0.09
floors 0.18 0.5 -0.0052 1 0.024 -0.26 0.46 0.52 -0.25 0.26
waterfront -0.0066 0.064 0.022 0.024 1 0.017 0.083 0.072 0.081 0.27
condition 0.028 -0.12 -0.009 -0.26 0.017 1 -0.14 -0.16 0.17 0.036
grade 0.36 0.66 0.11 0.46 0.083 -0.14 1 0.76 0.17 0.67
sqft_above 0.48 0.69 0.18 0.52 0.072 -0.16 0.76 1 -0.052 0.61
sqft_basement 0.3 0.28 0.015 -0.25 0.081 0.17 0.17 -0.052 1 0.32
price 0.31 0.53 0.09 0.26 0.27 0.036 0.67 0.61 0.32 1
bedrooms
bathrooms
sqft_lot
floors
waterfront
condition
grade
sqft_above
sqft_basement
price
1.00
0.75
0.50
0.25
0.00
−0.25