Multi-Linear Regression

Adding More Features

Polynomial Regression

m=1

m=2

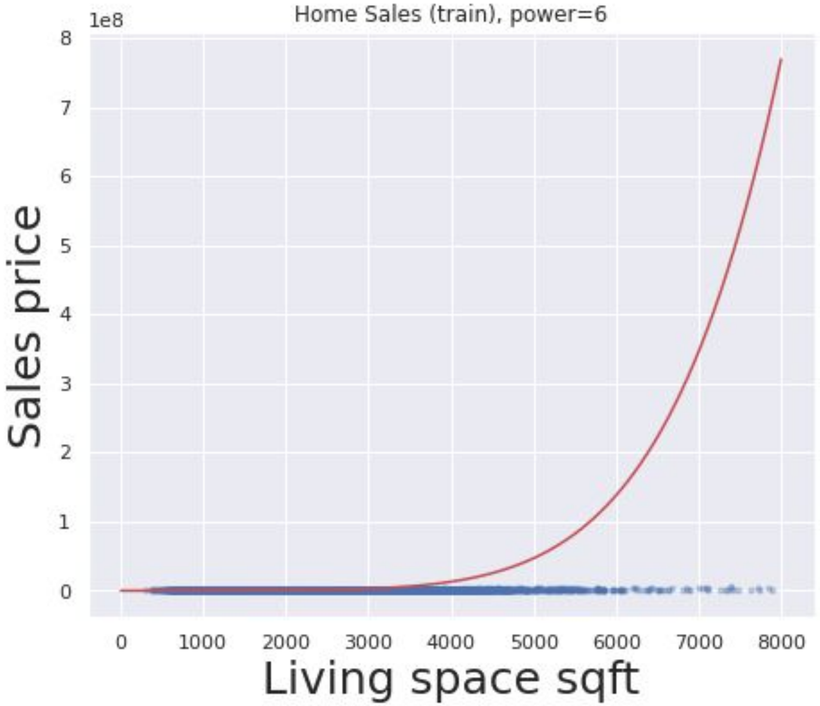
m=3

Polynomial Regression

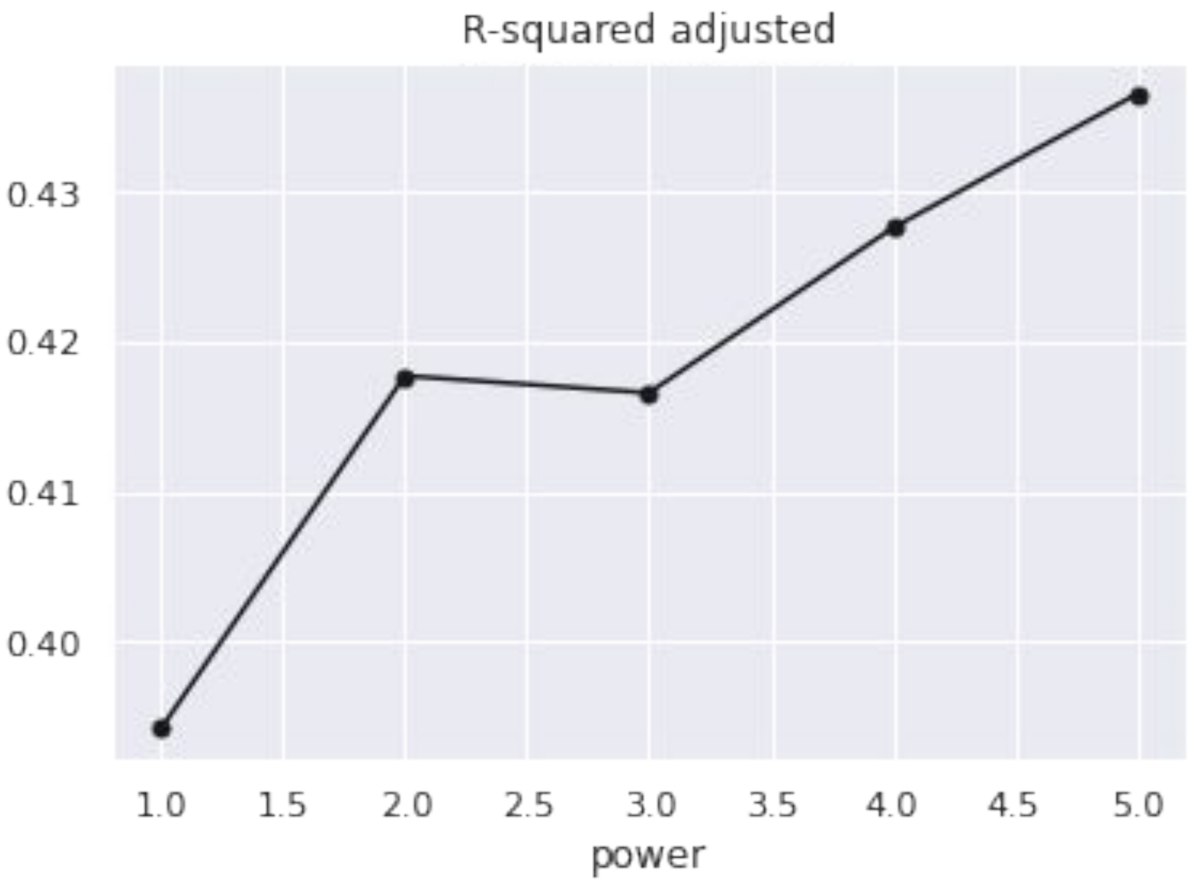
m=4

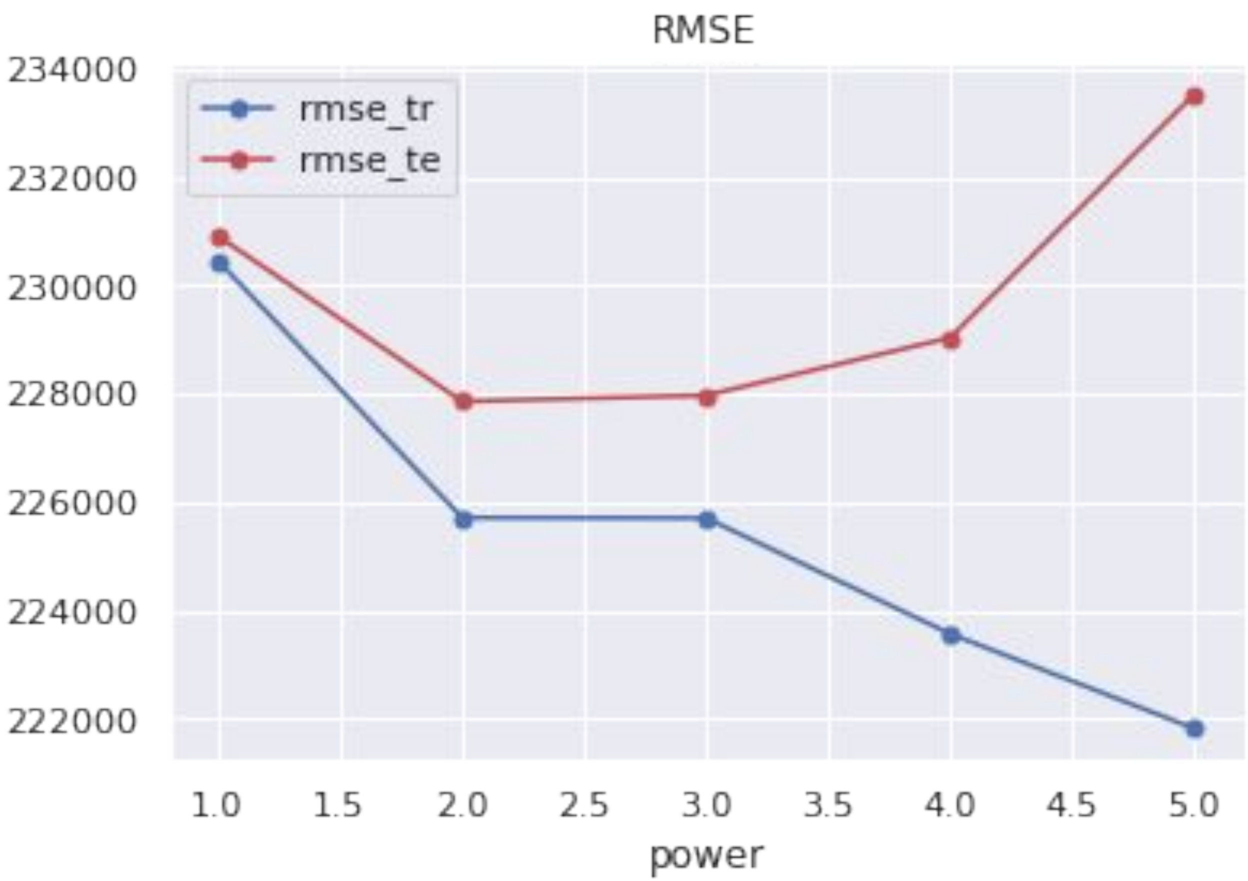
m=5

m=6

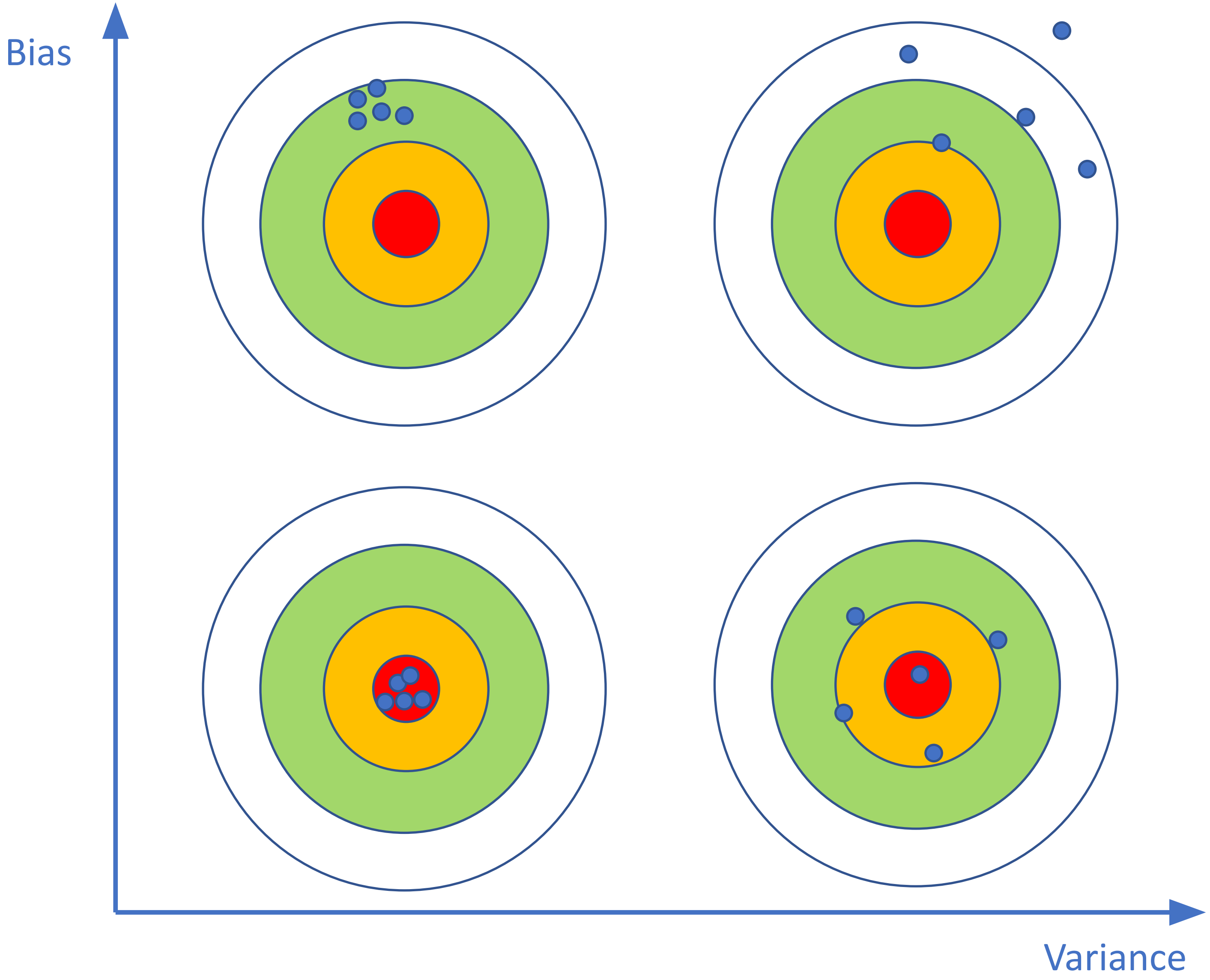


Where to stop?

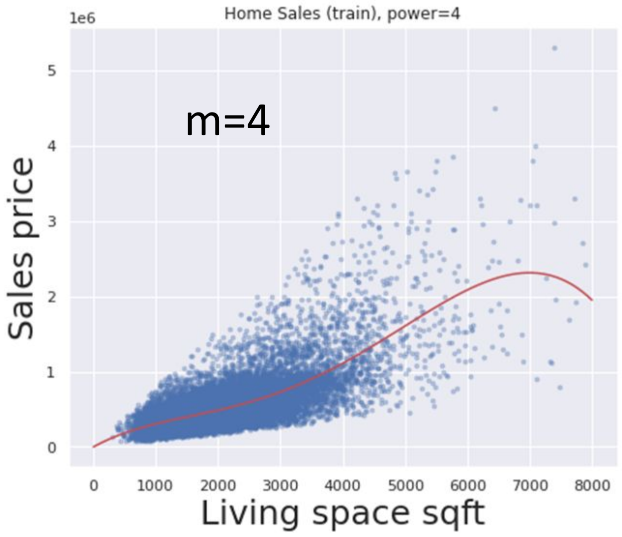


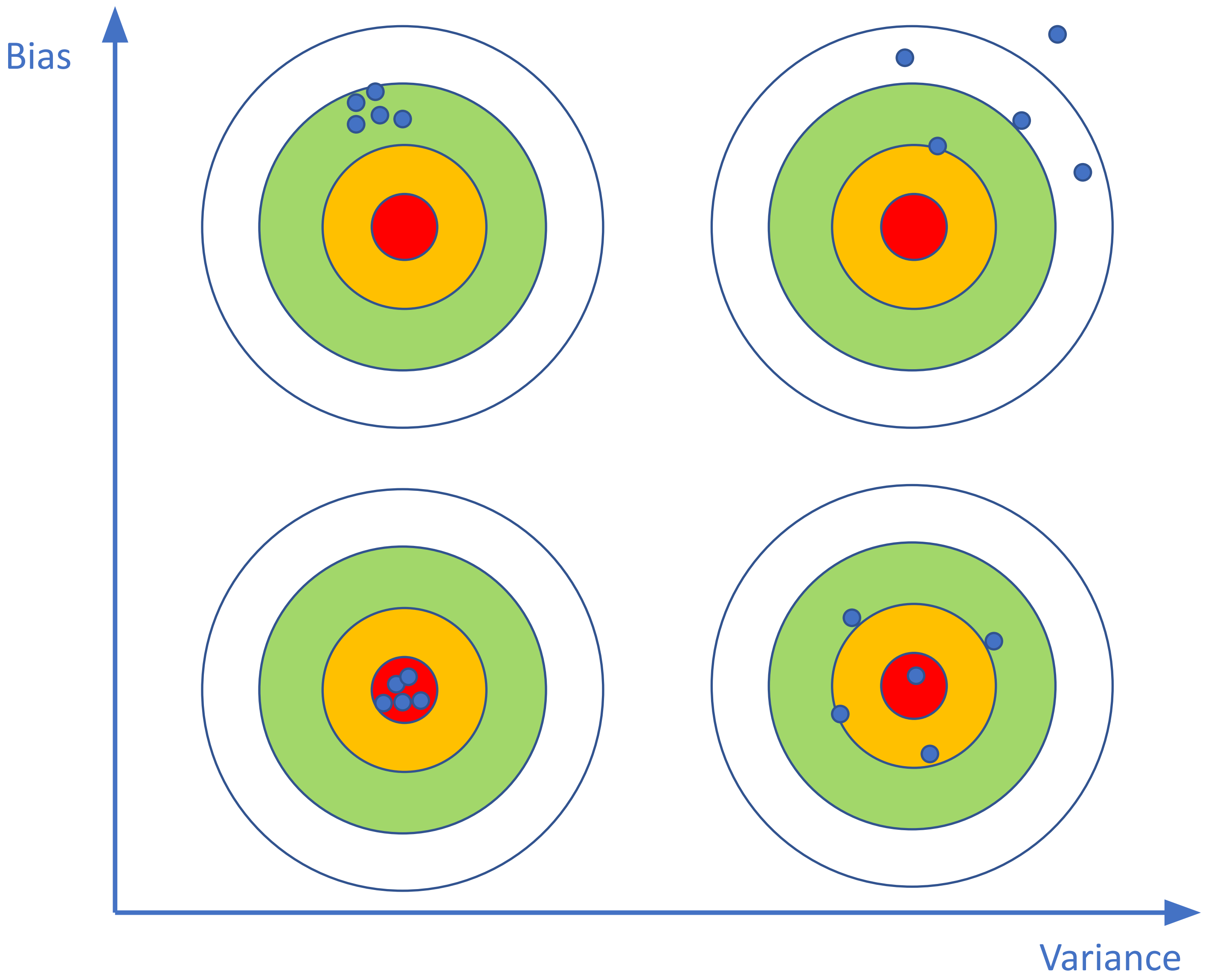


Bias-Variance Trade-off

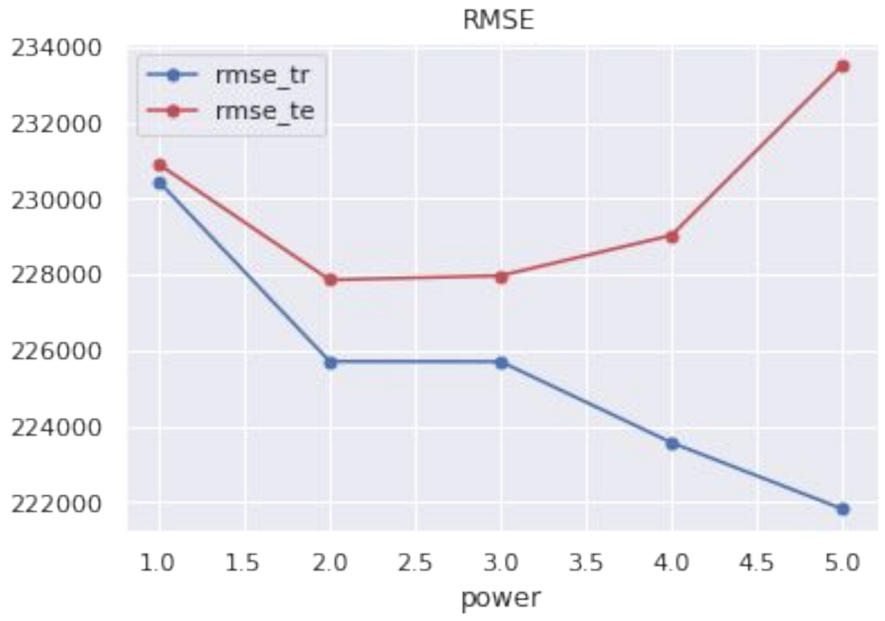


Bias-Variance Trade-off





Bias-Variance Trade-off and Test Error



Multi-Linear Regression part 2

Outline

- Multilinear regression model
- Model coefficients and significance
- How to select features
- Highly correlated features and (multi)collinearity
- Other things to consider when selecting features
- When there are interactions

Multilinear regression model

All predictors(variables) X_1~X_p are linear to Y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_j Average effect of X_j to Y when all other predictors fixed

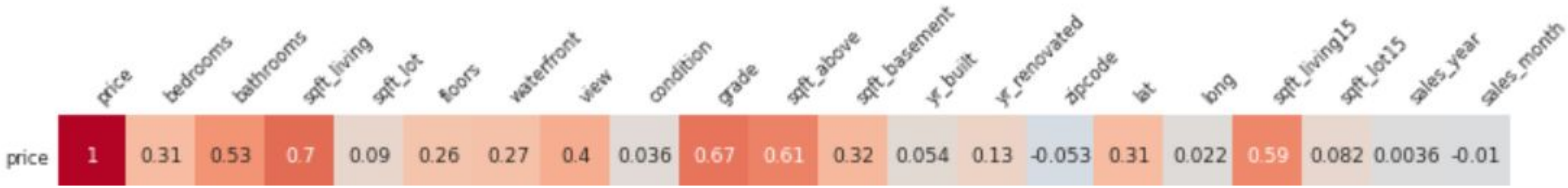
Caution 1: In general, predictors might be correlated

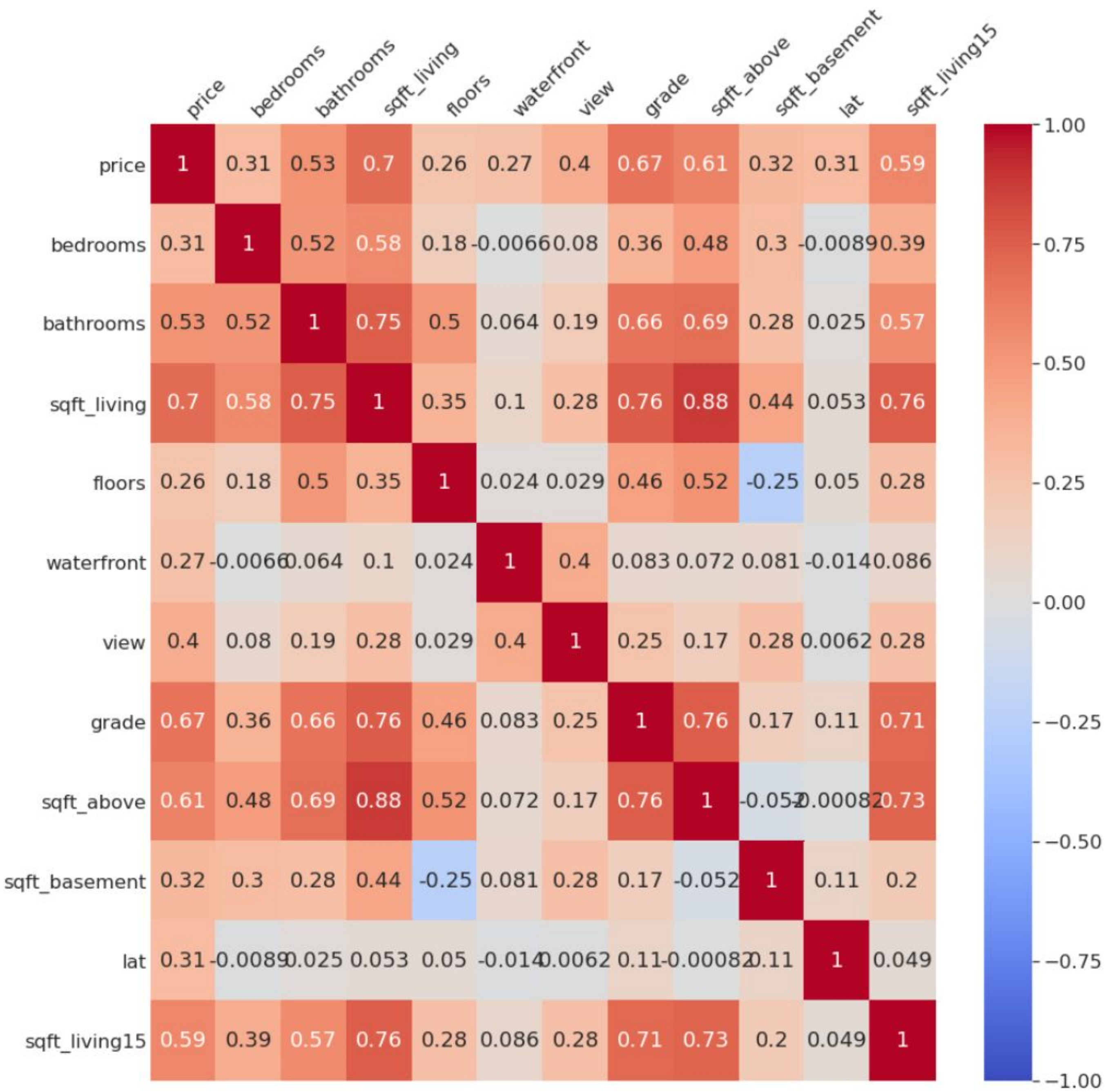
Caution 2: There may be interactions between predictors

Types of variables in a linear regression model

price	650000	1350000	369900	905000	690000
bedrooms	4	3	1	5	3
bathrooms	3	2.5	0.75	3.5	1
sqft_living	2950	2753	760	3100	1090
sqft_lot	5000	65005	10079	10200	4000
floors	2	1	1	1	1.5
waterfront	0	1	1	0	0
view	3	2	4	4	0
condition	3	5	5	3	4
grade	9	9	5	9	7
sqft_above	1980	2165	760	1660	1090
sqft_basement	970	588	0	1440	0
yr_built	1979	1953	1936	1970	1945
yr_renovated	0	0	0	0	0
zipcode	98126	98070	98070	98008	98117
lat	47.5714	47.4041	47.4683	47.6134	47.6846
long	-122.375	-122.451	-122.438	-122.112	-122.386
sqft_living15	2140	2680	1230	2700	1520
sqft_lot15	4000	72513	14267	10455	4000

Inspecting features qualitatively



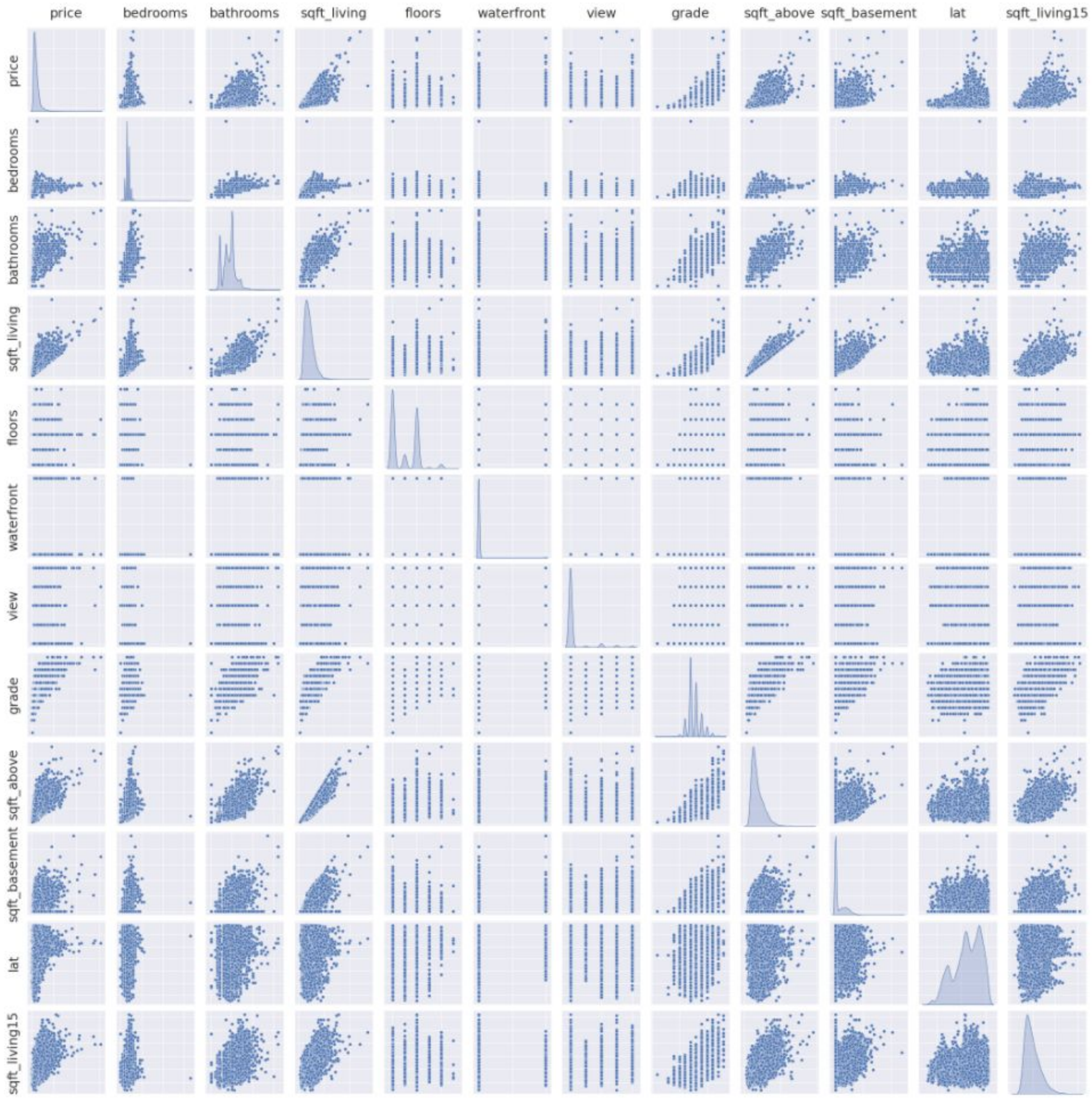


sqft_living	sqft_above	0.876597
sqft_living	grade	0.762704
sqft_living15	sqft_living	0.756420
sqft_above	grade	0.755923
sqft_living	bathrooms	0.754665
sqft_living15	sqft_above	0.731870
sqft_living15	grade	0.713202
sqft_above	bathrooms	0.685342
grade	bathrooms	0.664983
sqft_living	bedrooms	0.576671
sqft_living15	bathrooms	0.568634
sqft_above	floors	0.523885
bedrooms	bathrooms	0.515884
floors	bathrooms	0.500653

```
{'bathrooms',
 'bedrooms',
 'floors',
 'grade',
 'sqft_above',
 'sqft_living',
 'sqft_living15'}
```

	sqft_living	sqft_above	sqft_basement
0	1180	1180	0
1	2570	2170	400
2	770	770	0
3	1960	1050	910
4	1680	1680	0

```python
import seaborn as sns
g = sns.pairplot(df_small,diag_kind='kde')
g.set(xticklabels=[],yticklabels=[])
```



Model fitting

OLS Regression Results (All features)

Dep. Variable:	price	**R-squared:**	0.697
Model:	OLS	**Adj. R-squared:**	0.696
Method:	Least Squares	**F-statistic:**	2204.
Date:	Sun, 21 Mar 2021	**Prob (F-statistic):**	0.00
Time:	14:02:11	**Log-Likelihood:**	-2.3550e+05
No. Observations:	17290	**AIC:**	4.710e+05
Df Residuals:	17271	**BIC:**	4.712e+05
Df Model:	18		
Covariance Type:	nonrobust		

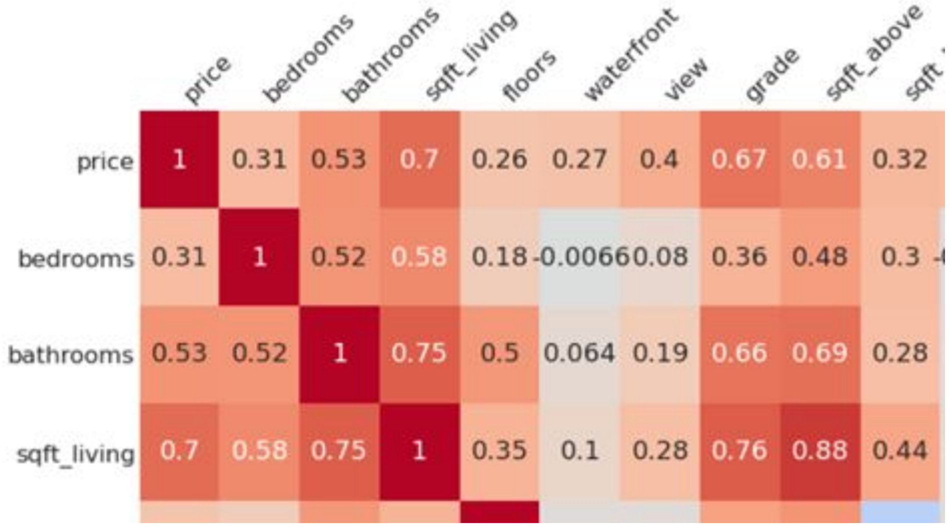
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.104e+08	1.07e+07	-10.355	0.000	-1.31e+08	-8.95e+07
bedrooms	-3.294e+04	2089.640	-15.763	0.000	-3.7e+04	-2.88e+04
bathrooms	4.558e+04	3622.204	12.584	0.000	3.85e+04	5.27e+04
sqft_living	107.5406	2.539	42.350	0.000	102.563	112.518
sqft_lot	0.0819	0.058	1.412	0.158	-0.032	0.196
floors	2084.8394	3959.535	0.527	0.599	-5676.250	9845.929
waterfront	5.687e+05	1.96e+04	29.030	0.000	5.3e+05	6.07e+05
view	5.019e+04	2360.305	21.265	0.000	4.56e+04	5.48e+04
condition	3.019e+04	2589.813	11.657	0.000	2.51e+04	3.53e+04
grade	9.593e+04	2381.529	40.280	0.000	9.13e+04	1.01e+05
sqft_above	69.9565	2.497	28.013	0.000	65.062	74.851
sqft_basement	37.5886	2.950	12.741	0.000	31.806	43.371
yr_built	-2526.7913	79.987	-31.590	0.000	-2683.573	-2370.009
yr_renovated	23.2448	4.079	5.698	0.000	15.249	31.240
lat	5.592e+05	1.16e+04	48.078	0.000	5.36e+05	5.82e+05
long	-1.017e+05	1.33e+04	-7.640	0.000	-1.28e+05	-7.56e+04
sqft_living15	26.9088	3.810	7.062	0.000	19.440	34.378
sqft_lot15	-0.3324	0.082	-4.045	0.000	-0.493	-0.171
sales_year	3.757e+04	5218.374	7.199	0.000	2.73e+04	4.78e+04
sales_month	1425.6858	781.931	1.823	0.068	-106.978	2958.349

(After removing features with p>0.025)

OLS Regression Results

Dep. Variable:	price	R-squared:	0.697
Model:	OLS	Adj. R-squared:	0.696
Method:	Least Squares	F-statistic:	2644.
Date:	Sun, 21 Mar 2021	Prob (F-statistic):	0.00
Time:	14:17:48	Log-Likelihood:	-2.3550e+05
No. Observations:	17290	AIC:	4.710e+05
Df Residuals:	17274	BIC:	4.712e+05
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-9.537e+07	6.79e+06	-14.035	0.000	-1.09e+08	-8.2e+07
bedrooms	-3.309e+04	2087.984	-15.848	0.000	-3.72e+04	-2.9e+04
bathrooms	4.605e+04	3491.887	13.189	0.000	3.92e+04	5.29e+04
sqft_living	107.6540	2.519	42.741	0.000	102.717	112.591
waterfront	5.687e+05	1.96e+04	29.029	0.000	5.3e+05	6.07e+05
view	5.025e+04	2358.449	21.305	0.000	4.56e+04	5.49e+04
condition	2.989e+04	2581.805	11.578	0.000	2.48e+04	3.5e+04
grade	9.606e+04	2372.468	40.491	0.000	9.14e+04	1.01e+05
sqft_above	70.6384	2.319	30.457	0.000	66.092	75.184
sqft_basement	37.0074	2.691	13.752	0.000	31.733	42.282
yr_built	-2525.0393	77.985	-32.379	0.000	-2677.898	-2372.181
yr_renovated	23.1731	4.073	5.689	0.000	15.190	31.157
lat	5.591e+05	1.16e+04	48.406	0.000	5.37e+05	5.82e+05
long	-1.013e+05	1.31e+04	-7.722	0.000	-1.27e+05	-7.56e+04
sqft_living15	26.2850	3.775	6.963	0.000	18.885	33.685
sqft_lot15	-0.2514	0.058	-4.331	0.000	-0.365	-0.138
sales_year	3.012e+04	3250.486	9.265	0.000	2.37e+04	3.65e+04



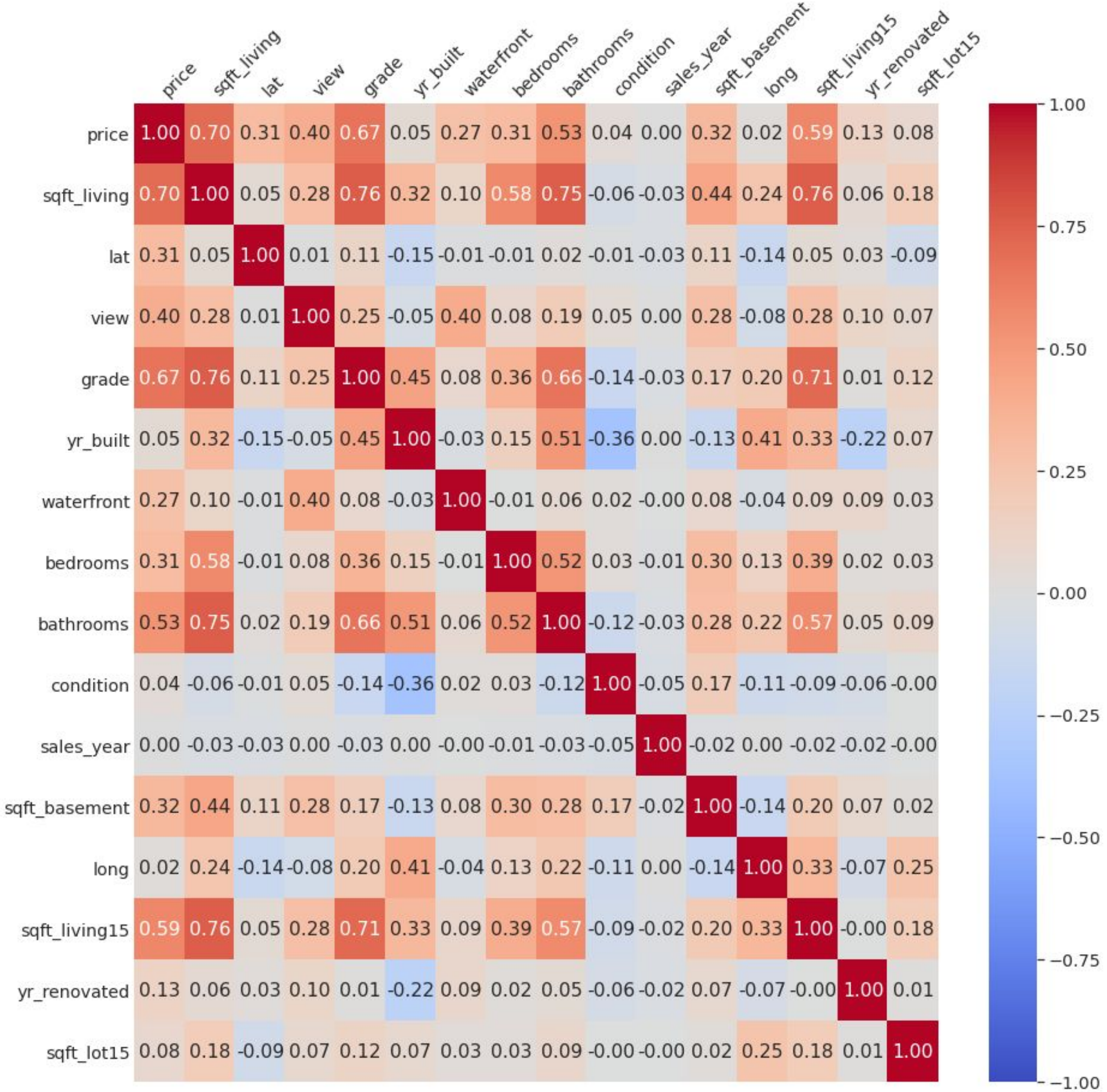
Multi-Linear Regression part 2

Feature selection

Do we include all the features or a subset?

Feature selection

- Forward selection
- Backward selection
- Mixed selection


price
sqft_living
lat
view
grade
yr_built
waterfront
bedrooms
bathrooms
condition
sales_year
sqft_basement
long
sqft_living15
yr_renovated
sqft_lot15
price 1.00 0.70 0.31 0.40 0.67 0.05 0.27 0.31 0.53 0.04 0.00 0.32 0.02 0.59 0.13 0.08
sqft_living 0.70 1.00 0.05 0.28 0.76 0.32 0.10 0.58 0.75 -0.06 -0.03 0.44 0.24 0.76 0.06 0.18
lat 0.31 0.05 1.00 0.01 0.11 -0.15 -0.01 -0.01 0.02 -0.01 -0.03 0.11 -0.14 0.05 0.03 -0.09
view 0.40 0.28 0.01 1.00 0.25 -0.05 0.40 0.08 0.19 0.05 0.00 0.28 -0.08 0.28 0.10 0.07
grade 0.67 0.76 0.11 0.25 1.00 0.45 0.08 0.36 0.66 -0.14 -0.03 0.17 0.20 0.71 0.01 0.12
yr_built 0.05 0.32 -0.15 -0.05 0.45 1.00 -0.03 0.15 0.51 -0.36 0.00 -0.13 0.41 0.33 -0.22 0.07
waterfront 0.27 0.10 -0.01 0.40 0.08 -0.03 1.00 -0.01 0.06 0.02 -0.00 0.08 -0.04 0.09 0.09 0.03
bedrooms 0.31 0.58 -0.01 0.08 0.36 0.15 -0.01 1.00 0.52 0.03 -0.01 0.30 0.13 0.39 0.02 0.03
bathrooms 0.53 0.75 0.02 0.19 0.66 0.51 0.06 0.52 1.00 -0.12 -0.03 0.28 0.22 0.57 0.05 0.09
condition 0.04 -0.06 -0.01 0.05 -0.14 -0.36 0.02 0.03 -0.12 1.00 -0.05 0.17 -0.11 -0.09 -0.06 -0.00
sales_year 0.00 -0.03 -0.03 0.00 -0.03 0.00 -0.00 -0.01 -0.03 -0.05 1.00 -0.02 0.00 -0.02 -0.02 -0.00
sqft_basement 0.32 0.44 0.11 0.28 0.17 -0.13 0.08 0.30 0.28 0.17 -0.02 1.00 -0.14 0.20 0.07 0.02
long 0.02 0.24 -0.14 -0.08 0.20 0.41 -0.04 0.13 0.22 -0.11 0.00 -0.14 1.00 0.33 -0.07 0.25
sqft_living15 0.59 0.76 0.05 0.28 0.71 0.33 0.09 0.39 0.57 -0.09 -0.02 0.20 0.33 1.00 -0.00 0.18
yr_renovated 0.13 0.06 0.03 0.10 0.01 -0.22 0.09 0.02 0.05 -0.06 -0.02 0.07 -0.07 -0.00 1.00 0.01
sqft_lot15 0.08 0.18 -0.09 0.07 0.12 0.07 0.03 0.03 0.09 -0.00 -0.00 0.02 0.25 0.18 0.01 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Correlated features

Caution: In general, predictors might be correlated
Where does correlation come from?

- Redundant Information
- Underlying effect (Confounding/Causality)
- Correlated in nature

Collinearity

- High correlation between features
- Collinearity
- Multicollinearity

Variance Inflation Factor (VIF)

all features

	VIF	feature
0	5.294373e+07	Intercept
1	1.652299e+00	bedrooms
2	3.351125e+00	bathrooms
3	inf	sqft_living
4	2.102643e+00	sqft_lot
5	2.012510e+00	floors
6	1.203920e+00	waterfront
7	1.435544e+00	view
8	1.253893e+00	condition
9	3.418066e+00	grade
10	inf	sqft_above
11	inf	sqft_basement
12	2.430670e+00	yr_built
13	1.151481e+00	yr_renovated
14	1.662368e+00	zipcode
15	1.181326e+00	lat
16	1.825966e+00	long
17	2.980096e+00	sqft_living15
18	2.135827e+00	sqft_lot15
19	2.594041e+00	sales_year
20	2.584148e+00	sales_month

after mixed selection

	VIF	feature
0	2.005295e+07	Intercept
1	5.869886e+00	sqft_living
2	1.110000e+00	lat
3	1.420496e+00	view
4	3.387375e+00	grade
5	2.281479e+00	yr_built
6	1.203120e+00	waterfront
7	1.645621e+00	bedrooms
8	3.124291e+00	bathrooms
9	1.231171e+00	condition
10	1.005421e+00	sales_year
11	1.596443e+00	sqft_basement
12	1.463758e+00	long
13	2.897906e+00	sqft_living15
14	1.147331e+00	yr_renovated
15	1.117361e+00	sqft_lot15

after removing *corr*>0.7

	VIF	feature
0	1.996040e+07	Intercept
1	2.290279e+00	sqft_living
2	1.067585e+00	lat
3	1.364575e+00	view
4	1.696129e+00	yr_built
5	1.200913e+00	waterfront
6	1.558691e+00	bedrooms
7	1.226291e+00	condition
8	1.005199e+00	sales_year
9	1.483550e+00	sqft_basement
10	1.359745e+00	long
11	1.114182e+00	yr_renovated
12	1.113754e+00	sqft_lot15

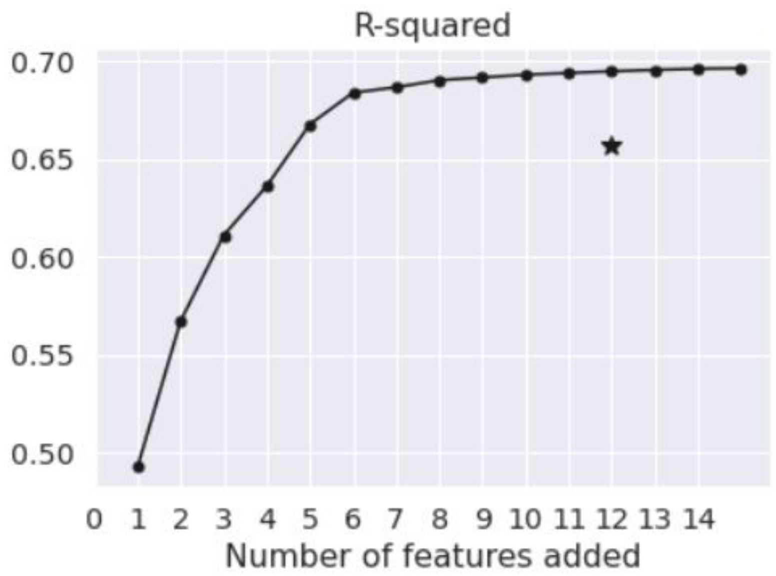
$$\mathrm{VIF}(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}$$

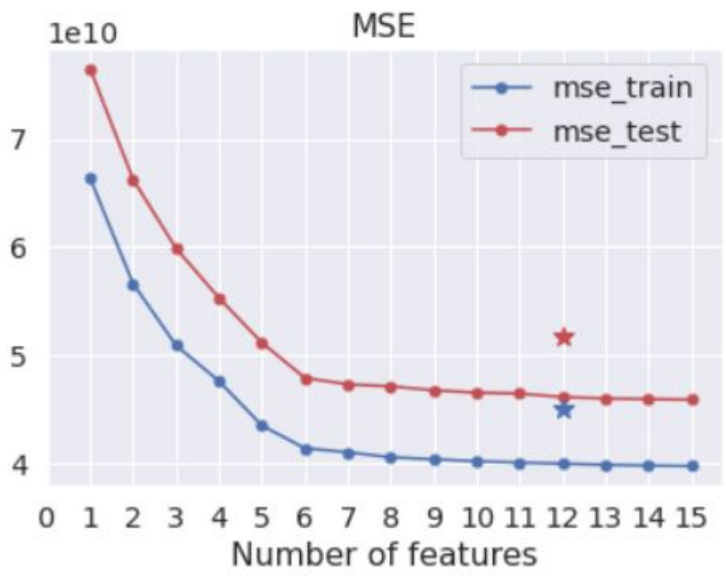
	price	sqft_living	lat	view	grade	yr_built	waterfront	bedrooms	bathrooms	condition	sales_year	sqft_basement	long	sqft_living15	yr_renovated	sqft_lot15
price	1.00	0.70	0.31	0.40	0.67	0.05	0.27	0.31	0.53	0.04	0.00	0.32	0.02	0.59	0.13	0.08
sqft_living	0.70	1.00	0.05	0.28	0.76	0.32	0.10	0.58	0.75	-0.06	-0.03	0.44	0.24	0.76	0.06	0.18

Feature selection considerations

- Model fitness
- Insignificant coefficients
- (Multi)collinearity
- Performance

Feature selection considerations





VIF	feature
144068.664256	Intercept
2.452368	sqft_living
1.075370	lat
1.332444	view
2.831009	grade
1.372389	yr_built
1.193425	waterfront

model	number of features	feature	coef	std err	t	p-value	[0.025, 0.975]	R^2	R^2_{adj}	F
all features	20	sqft_living	107.5406	2.539	42.350	0.000	102.563, 112.518	0.697	0.696	2204
mixed selection	15	sqft_living	178.2924	4.026	44.290	0.000	170.402, 186.18	0.697	0.696	2644
mixed selection and remove high corr	12	sqft_living	313.1627	2.677	116.972	0.000	307.915, 318.410	0.656	0.656	2751
mixed selection (elbow)	6	sqft_living	172.4562	2.657	64.911	0.000	167.249, 177.664	0.684	0.684	6231

When there are interactions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Summary

NCI-ALMANAC data

NIH NATIONAL CANCER INSTITUTE

Home Spaces New Account

NCI DTP Data

Search Space

- AIDS Antiviral Screen Data
- Chemical Data
- Compound Sets
- In Vivo Antitumor Assays
- Molecular Target Data
- NCI-60 Growth Inhibition Data
- **NCI-ALMANAC**
- Yeast Anticancer Drug Screen

Dashboard / DTP NCI Bulk Data for Download

NCI-ALMANAC

Created by Unknown User (zaharevd), last modified on Dec 22, 2017

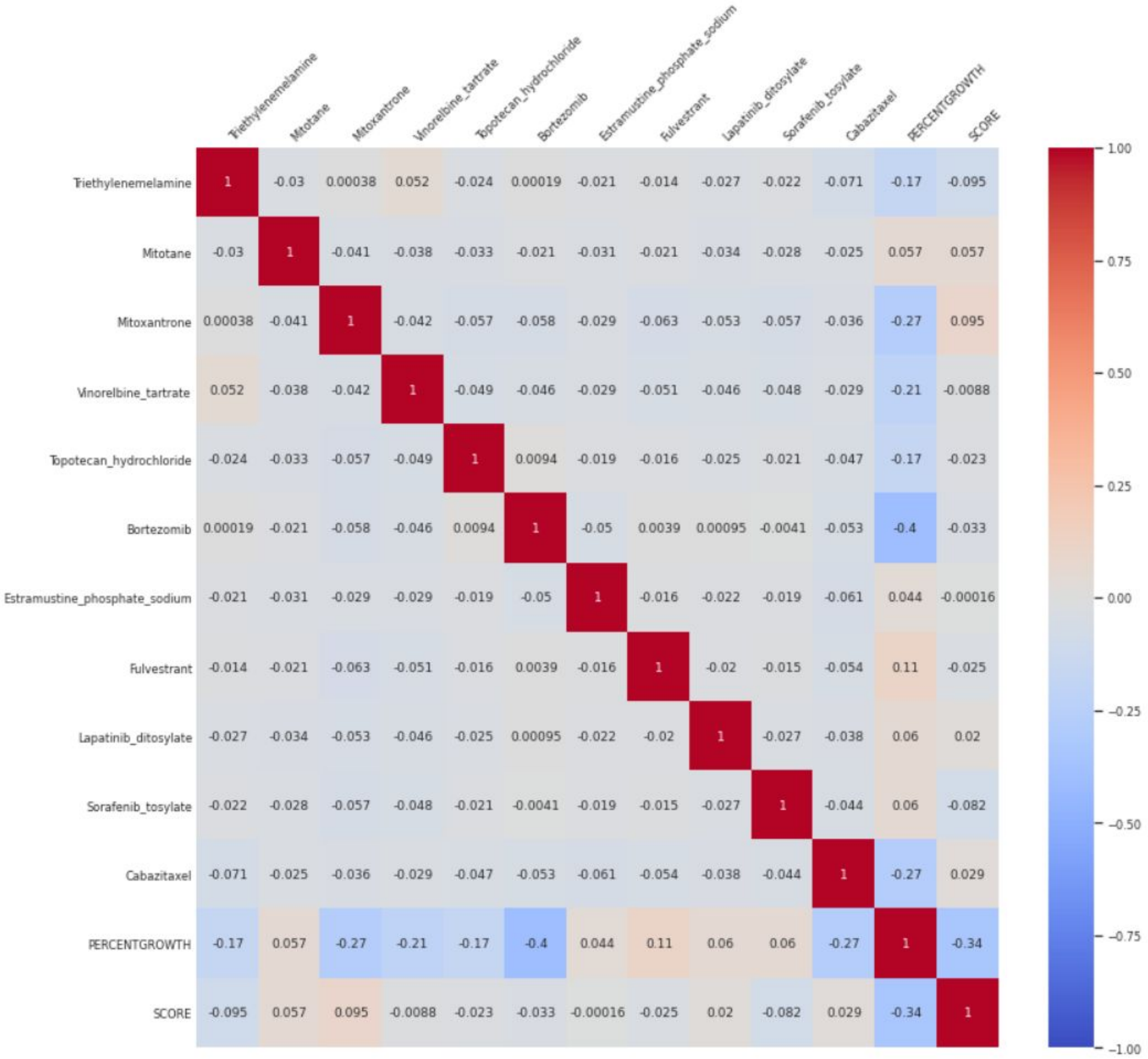
Compound Data

2D structures - ComboCompoundSet.sdf

Chemical Names - Single Name file, file with all available names

Growth Inhibition Data

ComboDrugGrowth_Nov2017.zip


Triethylenemelamine
Mitotane
Mitoxantrone
Vinorelbine_tartrate
Topotecan_hydrochloride
Bortezomib
Estramustine_phosphate_sodium
Fulvestrant
Lapatinib_ditosylate
Sorafenib_tosylate
Cabazitaxel
PERCENTGROWTH
SCORE
Triethylenemelamine 1 -0.03 0.00038 0.052 -0.024 0.00019 -0.021 -0.014 -0.027 -0.022 -0.071 -0.17 -0.095
Mitotane -0.03 1 -0.041 -0.038 -0.033 -0.021 -0.031 -0.021 -0.034 -0.028 -0.025 0.057 0.057
Mitoxantrone 0.00038 -0.041 1 -0.042 -0.057 -0.058 -0.029 -0.063 -0.053 -0.057 -0.036 -0.27 0.095
Vinorelbine_tartrate 0.052 -0.038 -0.042 1 -0.049 -0.046 -0.029 -0.051 -0.046 -0.048 -0.029 -0.21 -0.0088
Topotecan_hydrochloride -0.024 -0.033 -0.057 -0.049 1 0.0094 -0.019 -0.016 -0.025 -0.021 -0.047 -0.17 -0.023
Bortezomib 0.00019 -0.021 -0.058 -0.046 0.0094 1 -0.05 0.0039 0.00095 -0.0041 -0.053 -0.4 -0.033
Estramustine_phosphate_sodium -0.021 -0.031 -0.029 -0.029 -0.019 -0.05 1 -0.016 -0.022 -0.019 -0.061 0.044 -0.00016
Fulvestrant -0.014 -0.021 -0.063 -0.051 -0.016 0.0039 -0.016 1 -0.02 -0.015 -0.054 0.11 -0.025
Lapatinib_ditosylate -0.027 -0.034 -0.053 -0.046 -0.025 0.00095 -0.022 -0.02 1 -0.027 -0.038 0.06 0.02
Sorafenib_tosylate -0.022 -0.028 -0.057 -0.048 -0.021 -0.0041 -0.019 -0.015 -0.027 1 -0.044 0.06 -0.082
Cabazitaxel -0.071 -0.025 -0.036 -0.029 -0.047 -0.053 -0.061 -0.054 -0.038 -0.044 1 -0.27 0.029
PERCENTGROWTH -0.17 0.057 -0.27 -0.21 -0.17 -0.4 0.044 0.11 0.06 0.06 -0.27 1 -0.34
SCORE -0.095 0.057 0.095 -0.0088 -0.023 -0.033 -0.00016 -0.025 0.02 -0.082 0.029 -0.34 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00