Exploratory Data Analysis - Project

King County House Sales

Timorsha Rafiq-Dost

June 1st 2021

Framework

- **Stakeholder:** Timothy Stevens (Seller)
- **Agenda:** owns expensive houses, needs to get rid, best timing within year, open for renovation when profits rise
- **Hypothesis:**

  1) It is wiser to renovate in months with least sunshine duration & offer real estates in months with highest sunshine duration

  2) It is better to sell during school holidays

  3) The renovation of houses with grade <= 8 & condition <= 2 is more profitable than grade > 8 & condition > 2

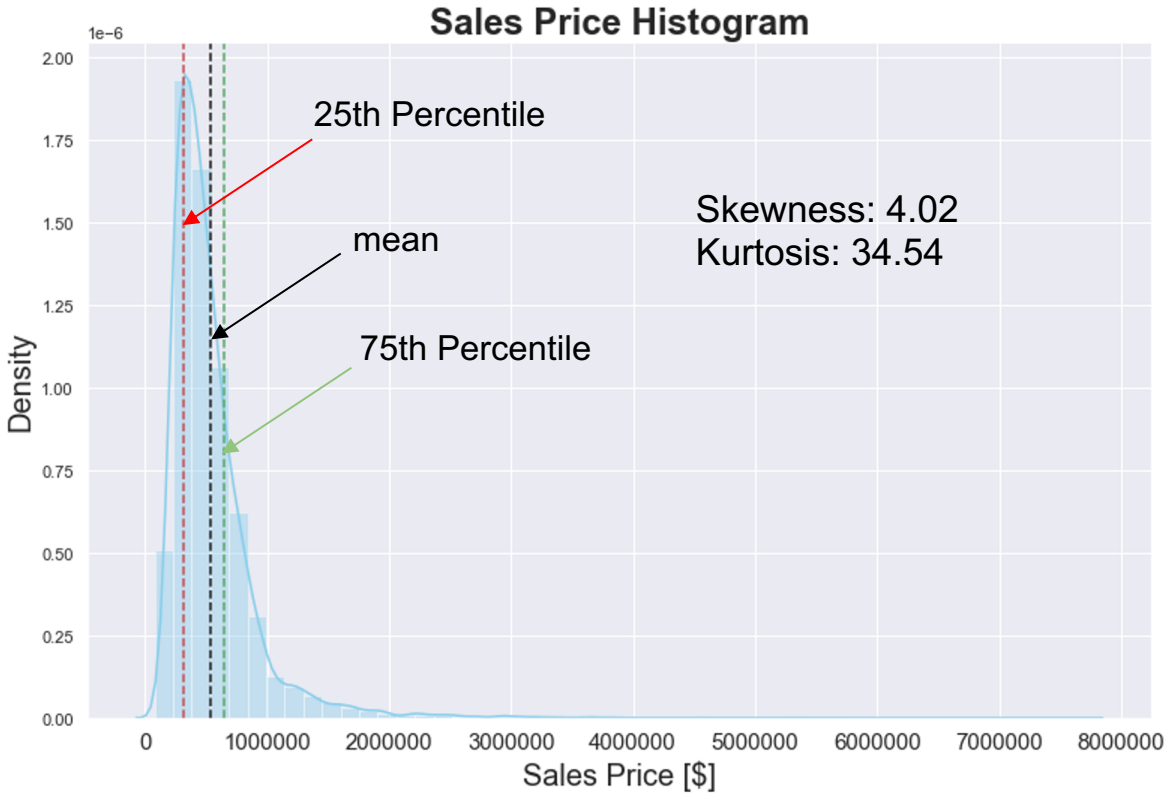
Dataset

dependent variable

	date	price	condition	grade	sqft_living	sqft_above	yr_renovated	bathrooms	bedrooms	floors	sqft_lot	sqft_basement
0	10/13/2014	221900.0	3	7	1180	1180	0	1.00	3	1.0	5650	0.0
1	12/9/2014	538000.0	3	7	2570	2170	1991	2.25	3	2.0	7242	400.0
2	2/25/2015	180000.0	3	6	770	770	0	1.00	2	1.0	10000	0.0
3	12/9/2014	604000.0	5	7	1960	1050	0	3.00	4	1.0	5000	910.0
4	2/18/2015	510000.0	3	8	1680	1680	0	2.00	3	1.0	8080	0.0

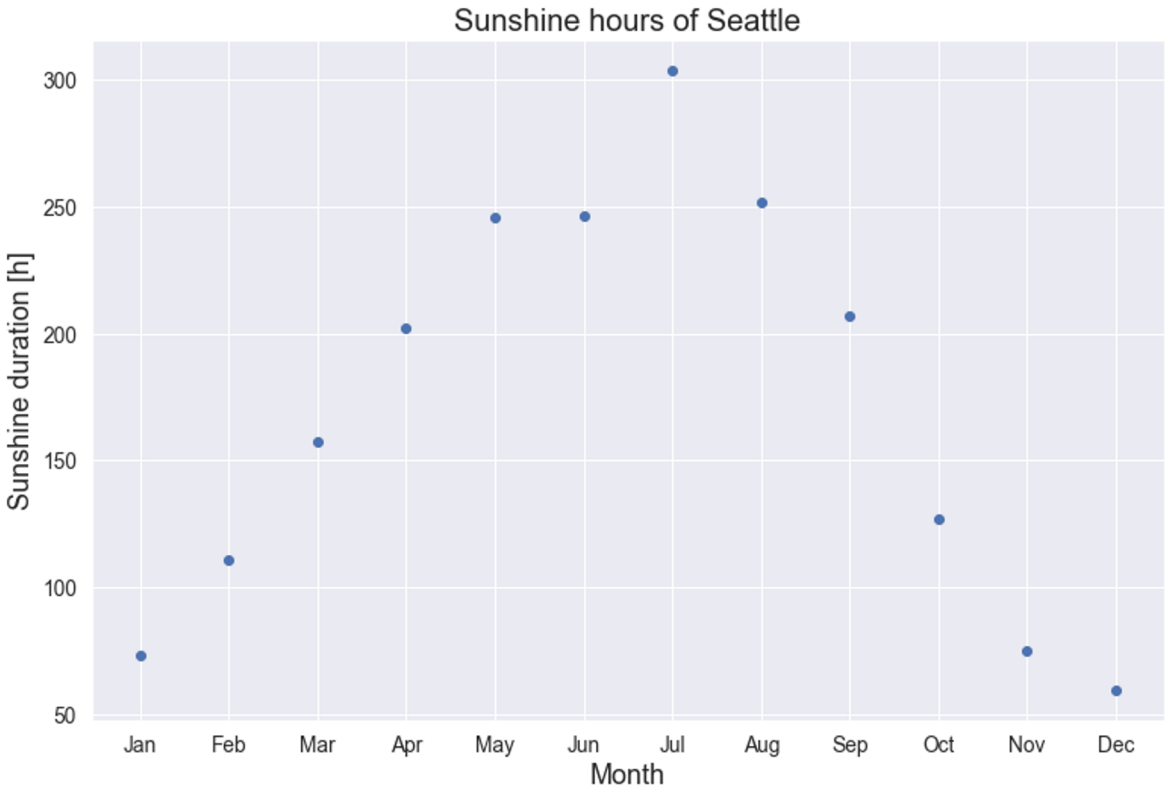
21597 rows x 12 columns

Sales Price Distribution



	price
count	21597.0
mean	540296.6
std	367368.1

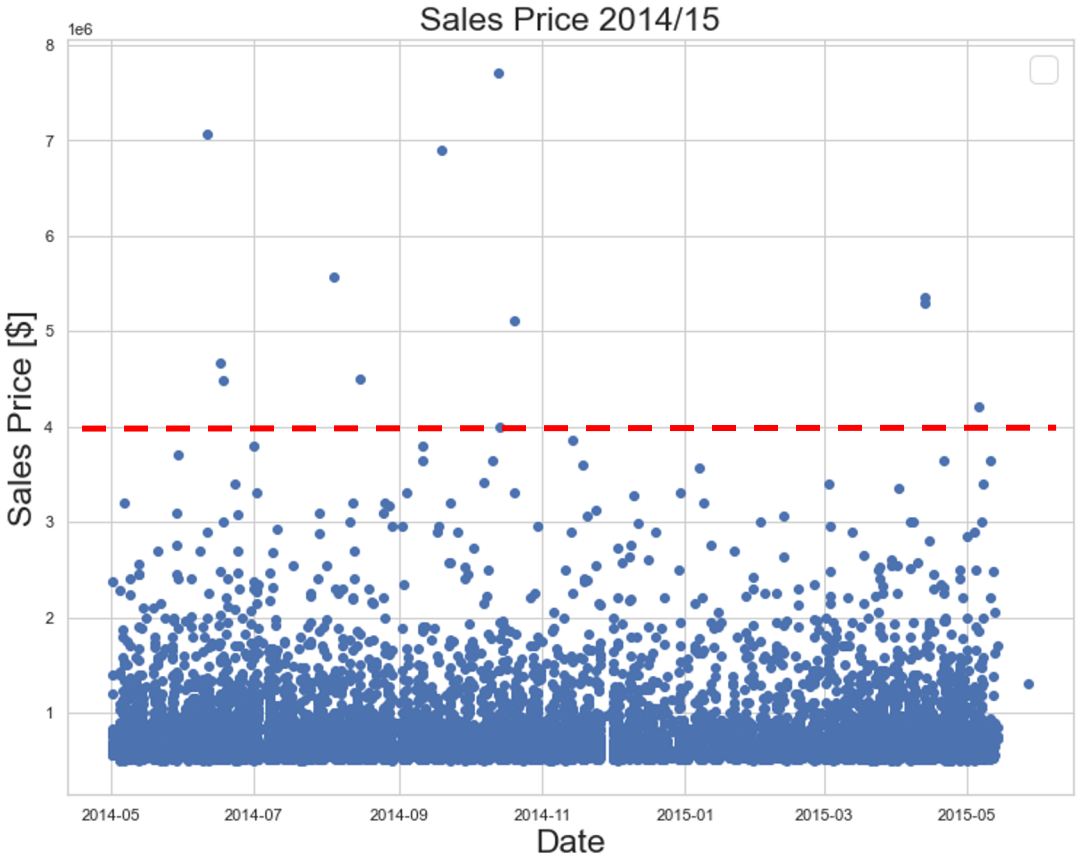
Exploratory data analysis



- mean out of monthly averages from different sources: climate-data.org
  allmetsat.com
  klimatabelle.de
  wetter.de

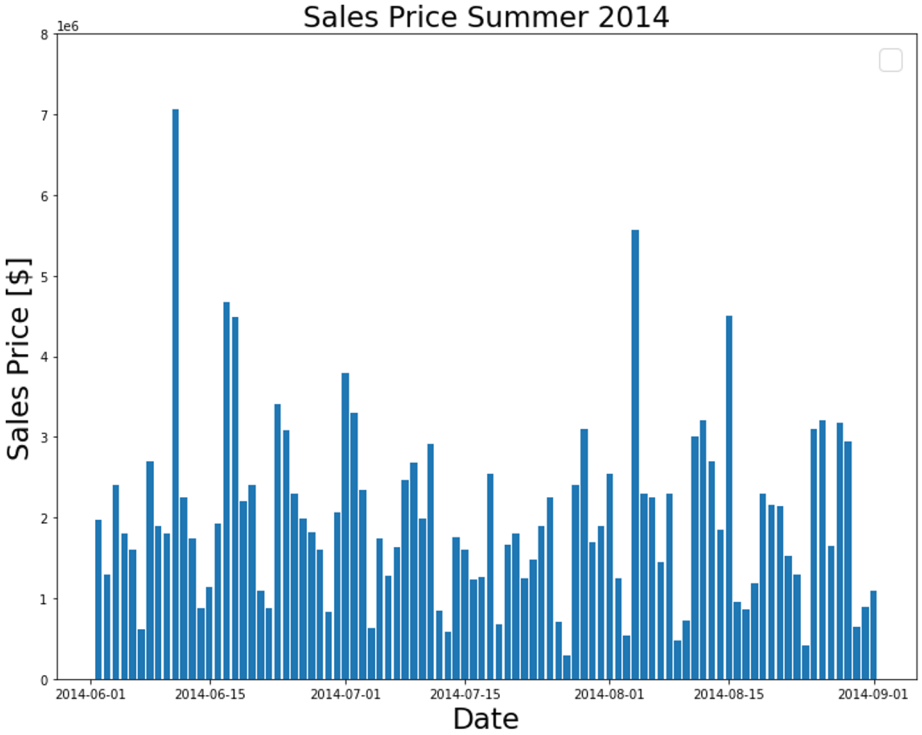
- max: Juli (303.4h)
- min: December (59.7h)

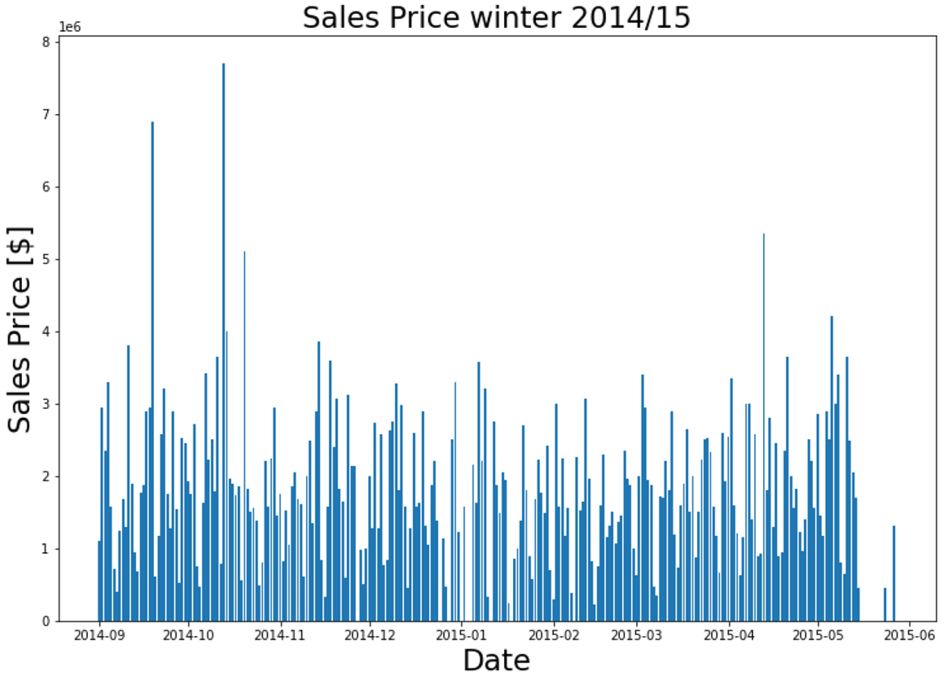
Sales Price over time



- range above 58th percentile (0.5 Mio) considered as “expensive” (9051 observations)
- Sales Price < 4 Mio → no significant tendency sales price/sunshine hours
- Sales price >= 4 Mio (12 observations) → majority sold around summer time/begin of autumn

Summer/Winter comparison

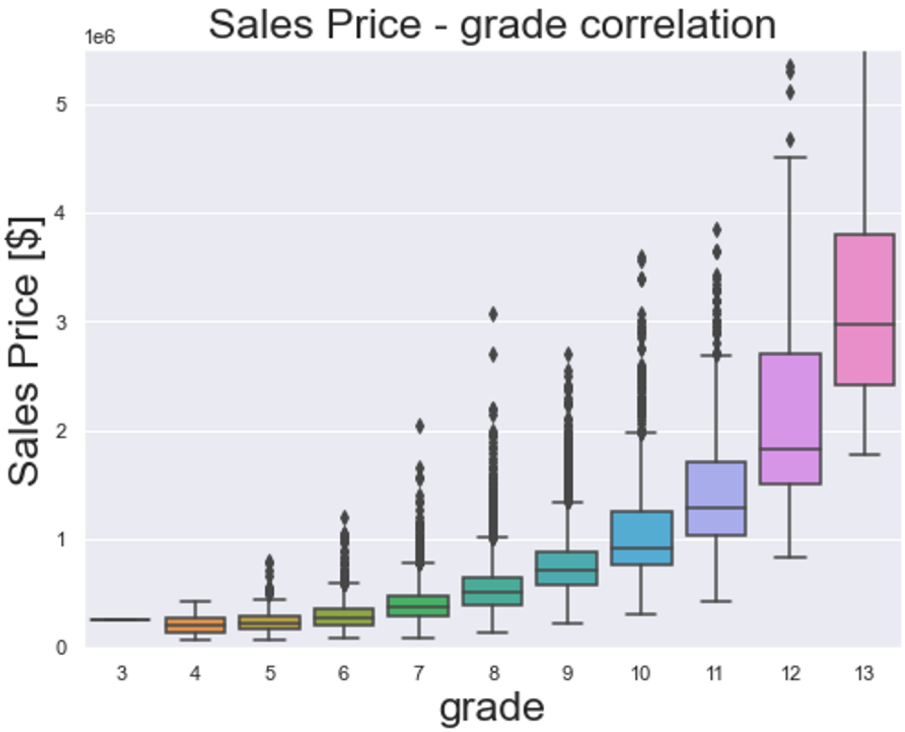


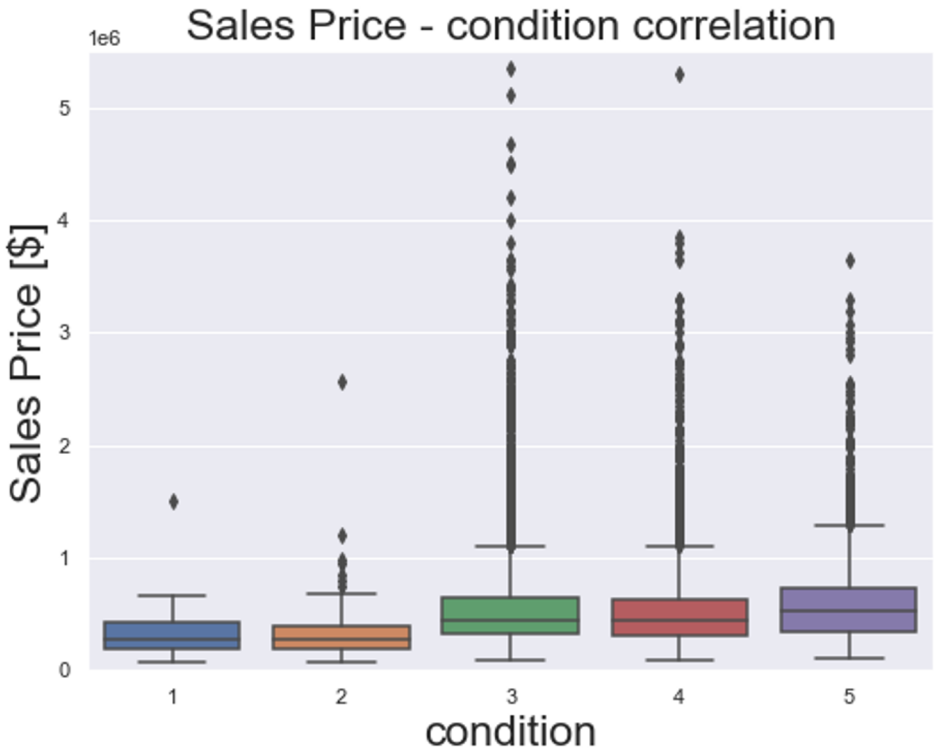


	season	mean	std	50%
0	summer	546520.5	368567.1	455000.0
1	winter	537634.7	366699.9	450000.0

→ no difference between seasons

Feature exploration

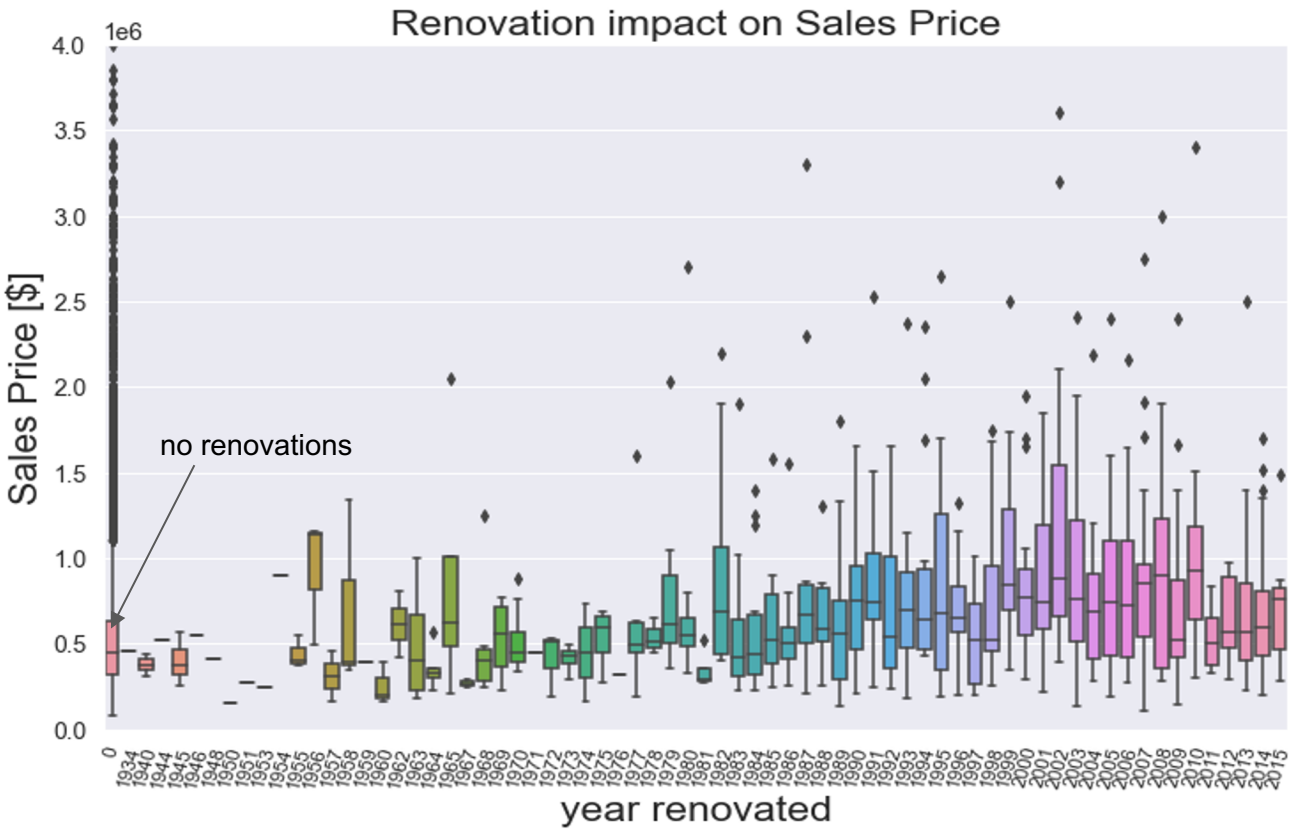




→ if grade goes up, so does the sale price

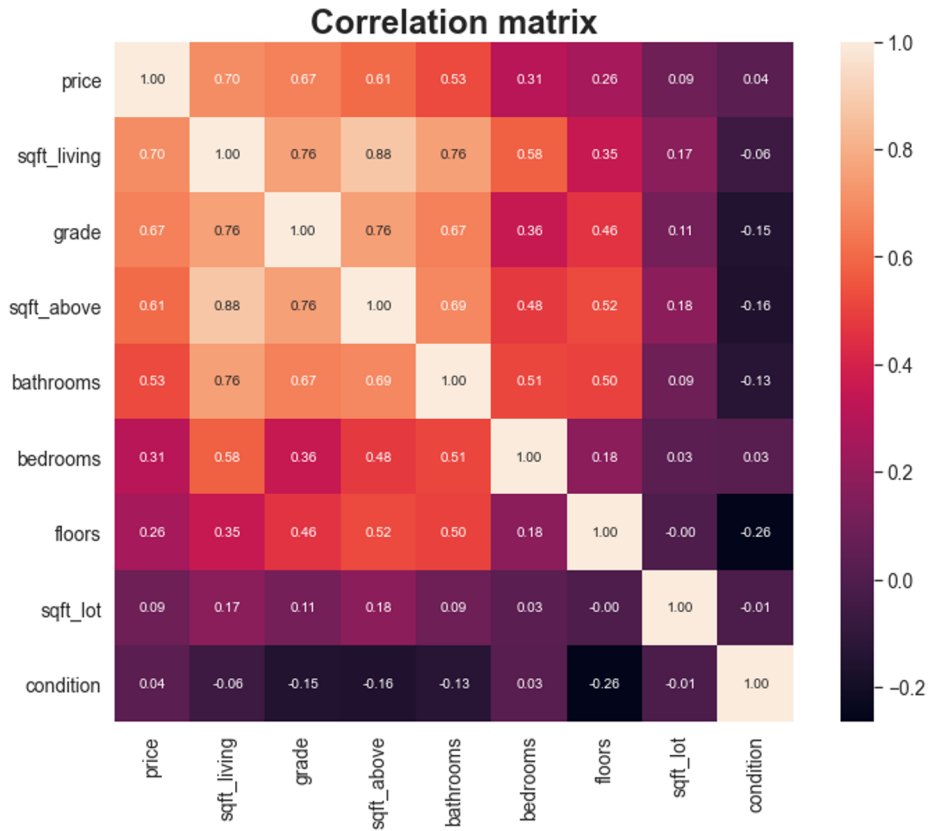
→ no great correlation

Feature exploration



→ moderate rise to higher sales prices when renovating

Feature exploration



- stronger correlation between price &
  - square footage of the home
  - grade
  - square footage of house apart from basement
  - bathrooms

- recommendation:
  - if technically possible, invest in house extension
  - don’t worry about condition/season
  - only invest in renovating bathrooms, not so much bedrooms

Modeling

Fit a multiple linear regression using ordinary least squares (OLS) with 4 independent variables

```
X2 = df_houseSales1[['bedrooms', 'bathrooms', 'sqft_living', 'sqft_above']]
X2 = sm.add_constant(X2)
y2 = df_houseSales1.price
model = sm.OLS(y2, X2).fit()
```

OLS Regression Results

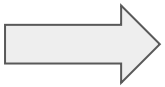
Dep. Variable:	price	**R-squared:**	0.508
Model:	OLS	**Adj. R-squared:**	0.508
Method:	Least Squares	**F-statistic:**	5570.

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.143e+04	6997.070	11.638	0.000	6.77e+04	9.51e+04
bedrooms	-5.978e+04	2353.492	-25.400	0.000	-6.44e+04	-5.52e+04
bathrooms	9606.7903	3534.007	2.718	0.007	2679.875	1.65e+04
sqft_living	331.5011	4.611	71.893	0.000	322.463	340.539
sqft_above	-27.6466	4.429	-6.243	0.000	-36.327	-18.966

Results

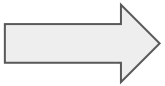
- **Hypothesis:**

1) It is wiser to renovate in months with least sunshine duration & offer real estates in months with highest sunshine duration



- not correct, insignificant trend between sunshine duration & sales price

2) It is better to sell during school holidays



- not correct, sale price independent of season

3) The renovation of houses with grade <= 8 & condition <= 2 is more profitable than grade > 8 & condition > 2

- the higher the grade the higher the sale price
- no significant relation between condition & sale price