Exploratory Data Analysis on House Prices in King County

...

A new **Business Model** based on
Identifying Underestimated Houses
and
Predicting realistic Market Prices

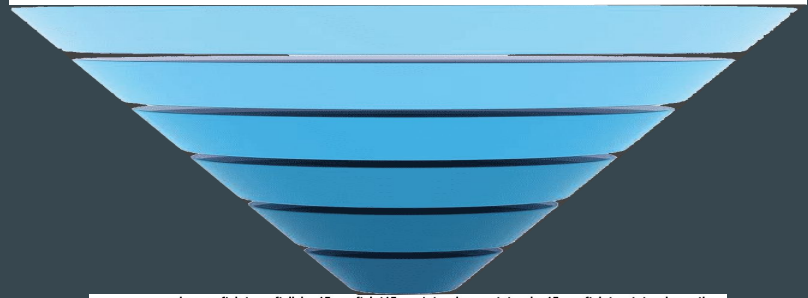
OLS Regression Model

- based on the "**King County House Price Dataset**"
- contains **21 different variables** (e.g. price, bedrooms, yr_built....)
- of **21,597 houses** in King County
- aim to find the **best predictors** for house prices
- starting with **a model with all 21 variables** and **fitting the model by narrowing down** to the least possible number of variables **(Top-Down-Approach)**

→ **6 price predictors**

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	...
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	...	7	1180	0.0	1955	
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	...	7	2170	400.0	1951	
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	...	6	770	0.0	1933	
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	...	7	1050	910.0	1965	
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	...	8	1680	0.0	1987	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
21592	263000018	5/21/2014	360000.0	3	2.50	1530	1131	3.0	0.0	0.0	...	8	1530	0.0	2009	
21593	6600060120	2/23/2015	400000.0	4	2.50	2310	5813	2.0	0.0	0.0	...	8	2310	0.0	2014	
21594	1523300141	6/23/2014	402101.0	2	0.75	1020	1350	2.0	0.0	0.0	...	7	1020	0.0	2009	
21595	291310100	1/16/2015	400000.0	3	2.50	1600	2388	2.0	NaN	0.0	...	8	1600	0.0	2004	
21596	1523300157	10/15/2014	325000.0	2	0.75	1020	1076	2.0	0.0	0.0	...	7	1020	0.0	2008	

21597 rows × 21 columns

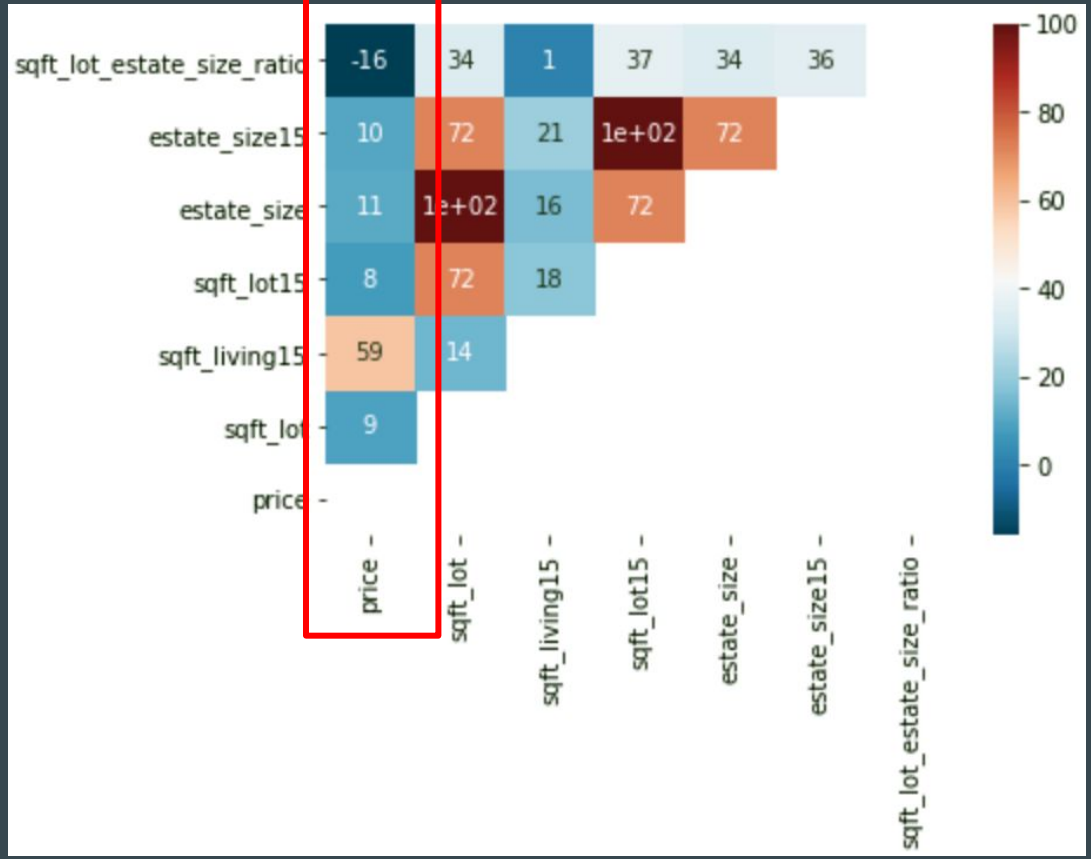


	price	sqft_lot	sqft_living15	sqft_lot15	estate_size	estate_size15	sqft_lot_estate_size_ratio
0	221900.0	5650	1340	5650	6830	6990	0.827233
1	538000.0	7242	1690	7639	9812	9329	0.738076
2	180000.0	10000	2720	8062	10770	10782	0.928505
3	604000.0	5000	1360	5000	6960	6360	0.718391
4	510000.0	8080	1800	7503	9760	9303	0.827869
...	...	...	...	...	...	...	...
21592	360000.0	1131	1530	1509	2661	3039	0.425028
21593	400000.0	5813	1830	7200	8123	9030	0.715622
21594	402101.0	1350	1020	2007	2370	3027	0.569620
21595	400000.0	2388	1410	1287	3988	2697	0.598796
21596	325000.0	1076	1020	1357	2096	2377	0.513359

Six Price Predictors

- **6 final variables** seem to be good predictors for house prices
- **sqft_lot_estate_size_ratio** is best negative correlator
- **sqft_living15** is best positive correlator

- **sqft_lot_estate_size_ratio = sqft_lot / (sqft_lot + sqft_living)**
- **percentage of lot from total estate**

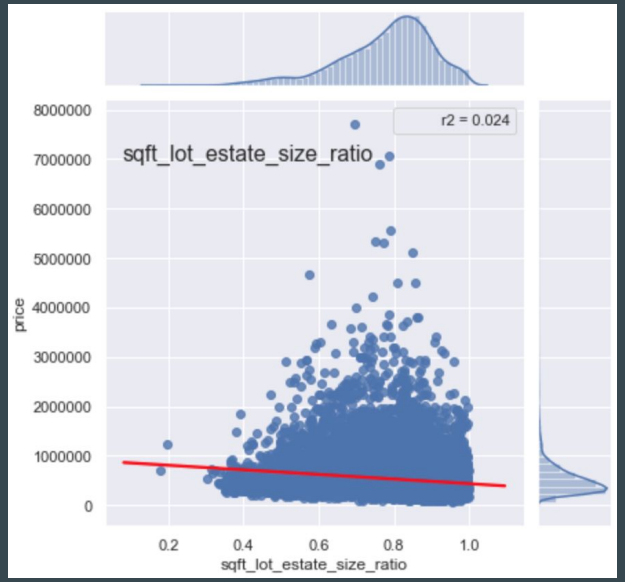


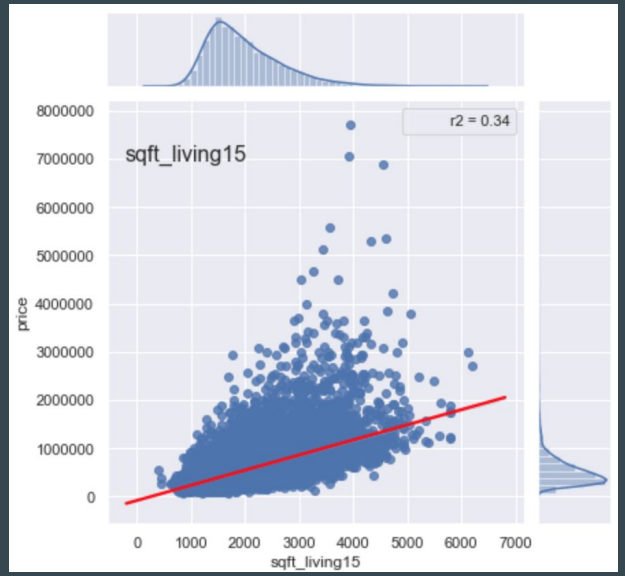
Single Correlation of best Price Predictors

- **negative correlation** of price with **sqft_lot_estate_size_ratio** “confirmed”
- **positive correlation** of price with **sqft_living15** “confirmed”

→ The **bigger** the **percentage of lot**, the **smaller** the **price**

→ The **bigger** the **neigbour houses**, the **higher** the **price.**





Business Idea

- use **sqft_lot_estate_size_ratio** as **identifier for underestimated houses** on the market
- select them with **sqft_living15** as **location independent predictor for realistic house prices**
- **built new houses on the lots**
- **offer them with a realistic price**

→ **buy affordable lots**

→ **build new houses**

→ **increase value on the market**

buy affordable lots

build new houses

increase value on the market

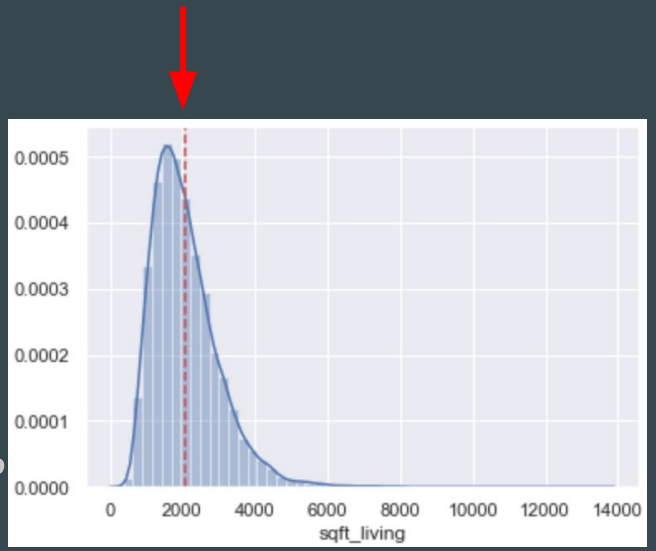
Top List

- mean size of houses (sqft_living) = 2,080 sqft
- focus on houses with lot (sqft_lot) > 2,080 sqft

→ 20,350 houses (~94%)

- focus on sqft_lot_estate_size_ratio > 0.8 to minimize investor risk

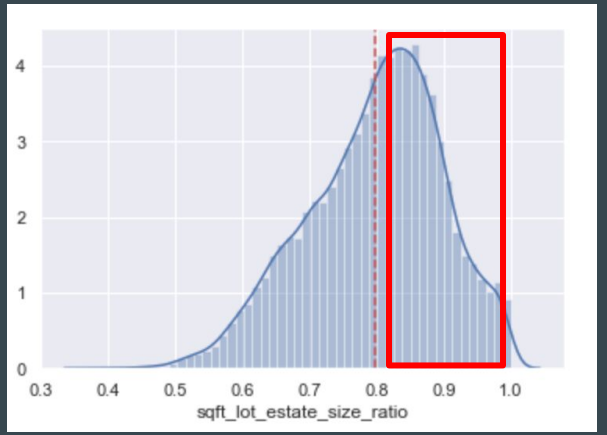
→ 10,350 houses (~48%)



mean house size = 2,080 sqft

= minimum lot size

→ 20,350 houses (~94%)



sqft_lot_estate_size_ratio > 0.8

→ 10,350 houses (~48%)

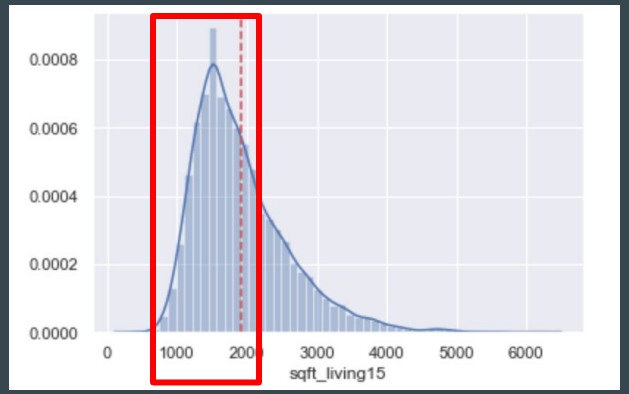
Top List

- **sqft_living15** is a good **predictor** for **realistic house prices**
- look for those in our Top List for **increase in value**
- focus on houses with **sqft_living15 = 1,000 - 2,000 sqft** to **minimize investor risk**

→ **6, 775 houses (~31%)**

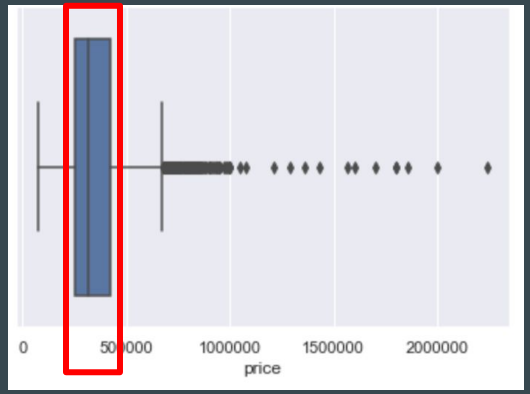
- focus on **normal price range of 249,900 - 421,250 USD (interquartil range)** to **minimize investor risk**

→ **3,390 houses (~15%)**



houses with sqft_living15 of 1,000 - 2,000 sqft

→ 6, 775 houses (~31%)



houses with price range of 249,900 - 421,250 USD

→ 3,390 houses (~15%)

Top List

- lot (sqft_lot) > 2,080 sqft
- sqft_lot_estate_size_ratio > 0.8
- sqft_living15 = 1,000 - 2,000 sqft
- price range of
  249,900 - 421,250 USD

→ minimal investor risk

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	grade	...	waterfront_wonan	view_wonan	yr_renovated_
7	2008000270	1/15/2015	291850.0	3	1.50	1060	9711	1.0	3	7	...	0.0	0.0	
12	114101516	5/28/2014	310000.0	3	1.00	1430	19901	1.5	4	7	...	0.0	0.0	
13	6054650070	10/7/2014	400000.0	3	1.75	1370	9680	1.0	4	7	...	0.0	0.0	
16	1875500060	7/31/2014	395000.0	3	2.00	1890	14040	2.0	3	7	...	0.0	0.0	
23	8091400200	5/16/2014	252700.0	2	1.50	1070	9643	1.0	3	7	...	2.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20833	7137800310	2/25/2015	329950.0	4	2.50	2300	9690	2.0	3	8	...	0.0	0.0	
21027	9276200220	7/17/2014	375000.0	1	1.00	720	3166	1.0	3	6	...	0.0	0.0	
21063	3449000010	3/12/2015	294570.0	3	1.00	1140	8400	1.0	4	7	...	0.0	0.0	
21327	2924079034	9/25/2014	332220.0	3	1.50	2580	47480	1.0	3	7	...	0.0	0.0	
21370	774101755	4/17/2015	320000.0	3	1.75	1790	66250	1.5	3	7	...	0.0	0.0	

3390 rows × 28 columns

Top List with 6, 775 houses (~31%) for investment

Example

- **house ID = 1222069089** from **Top List**
- **price 375,000 USD**
- size of **lot = 533,610 sqft**
- **sqft_lot_estate_size_ratio = 0.998503** which hints for an **underestimate price** on the market
- we build a **new additional house on the lot**
- we use s**qft_living15 =1790** to predict **realistic house price**

→ **prediction of the realistic house price is 478,603 USD** (+103,603 USD in comparison)

	17562		
id	1222069089	long	-121.986
date	9/4/2014	sqft_living15	1790
price	375000	sqft_lot15	216057
bedrooms	1	date_encoded	366
bathrooms	1	waterfront_wonan	2
sqft_living	800	view_wonan	0
sqft_lot	533610	yr_renovated_wonan	0
floors	1.5	sqft_basement_float_wonan	0
condition	5	estate_size	534410
grade	5	estate_size15	217847
sqft_above	800	Bath_Bed_Ratio	1
yr_built	1950	last_construction	1950
zipcode	98038	sqft_living_floors_ratio	533.333
lat	47.4134	sqft_lot_estate_size_ratio	0.998503

Future Work

Business Model

- apply **Business Model** to other regions than King County

→ **Data Mining and Implementing in Predictive Modeling**

Regression Model

- improve Regression Model

→ **non-linear Approaches for Regression Modeling**

Validation

- validate Regression Model with **Test Data Set**

→ **use Train-Test-Split Approach**

Thank you for your Attention!

Feel free to ask Questions

Appendix

https://github.com/Patrick-Neubert/Neuer_Fisch/blob/master/EDA_on_House_Prices_in_King_County_final.ipynb