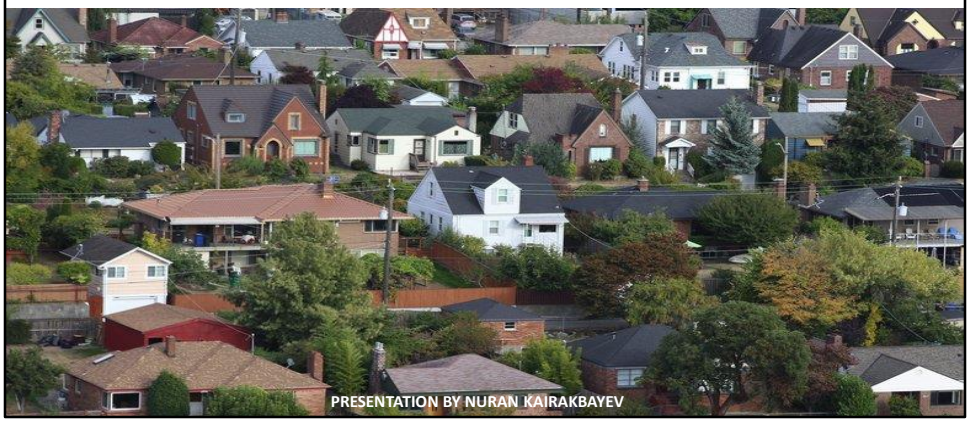
Module 1 Project

Flatiron School. Data Science program

# King County House Sale

Research on price prediction.

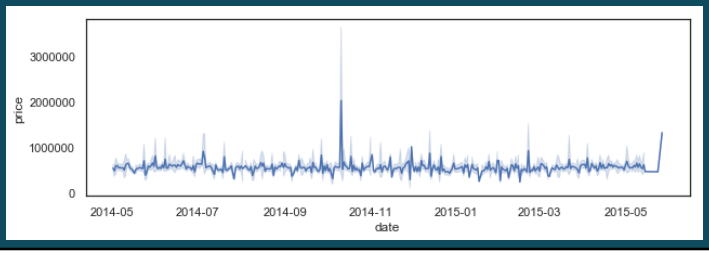



Hi! Let me present to your attention a brief overview presentation prepared as part of a statistical study of home pricing mechanisms.

## PROBLEM STATEMENT

- HIGH RANGE in HOUSES PRICES
- UNCLEAR PRICING MECHANISM
- EXTERNAL FACTORS
- NO CLEAR BUSINESS STRATEGY

Houses' prices volatility illustration (2014-2015)



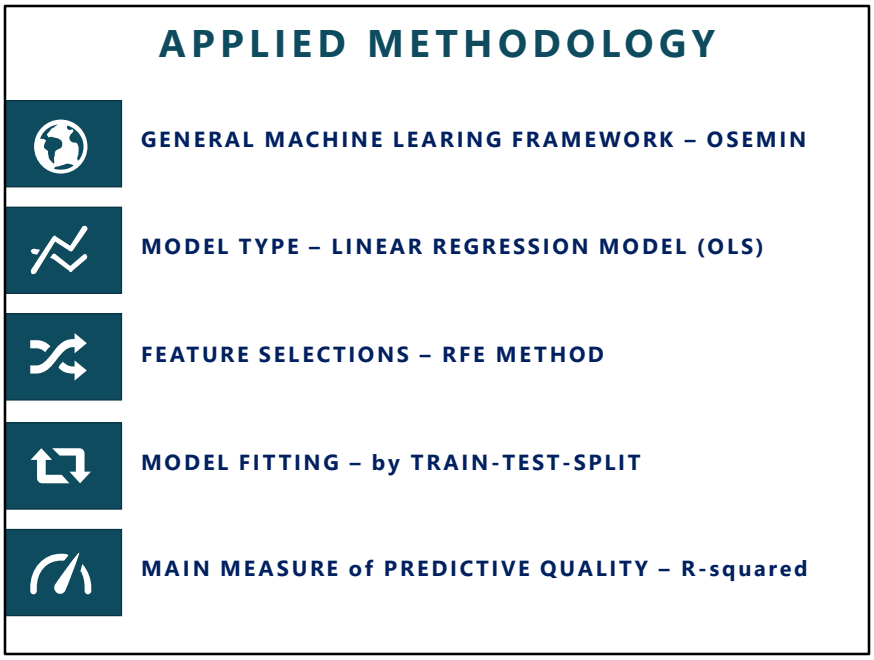

Property prices in King County have a high range, which creates uncertainty when dealing with specific transactions. Understanding fair market price for specific transactions is the key to success the profitability of investment KPI. Given the changing environment is the task of the house trader – how to assess the market value of the home, what factors need to be considered, how not to go at risk and buy the house lower than the real cost. What new houses can have good investment prospects? To answer such questions we need to have special tool.

## BUSINESS VALUE

**Effective predictor of house price**

- **key factors for prices**
- **profitable bargain ranges**
- **effective investments on high-valued houses**
- **clear strategy for business sustainable growth**

Predictive tool for broker and investment departments of the Company will provides an abilities for understanding pricing mechanism and key factors. With some certainty level it will be possible to evaluate expected market prices for targets. As a result, Company can be effective in auctions, bargain processes in deals, and moreover, can reveal target groups for high-value potentials. Knowing it will gives opportunities to make good operational and investment strategies for the Company and increase its market capitalization.



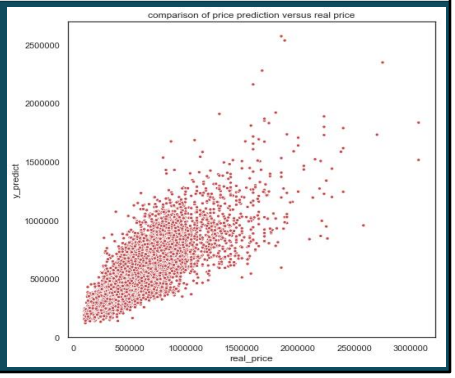

Building an effective predictive model of prices for houses based on the use of a database of prices and possible factors influencing prices, as well as machine learning methods. Used process - OSEMIN is a widely used framework for such problems. As a specification model is selected linear regression calculated by the OLS method. Given the uncertainty factors applied to the recursive selection of factors. The calculation of the final coefficients produced using the separation of samples into learning and test subsets. Key quality parameter prediction –R-squared.

## FINDINGS (1/3)

- **MAIN KEY FACTORS for HOUSE PRICES**
- **PREDICTION QUALITY - 70%-75%**

R-squared

**70%-75%**




Based on currently available data, the model defined set of predictors for house prices and get 70-75% of prediction quality (R-squared). As you can see in right plot, here comparison between real prices and predicted prices, which works well for most of examples excluding some single outliers.

## FINDINGS (2/3)

**TOP 3 POSITIVE FACTORS:**

- **Living space area of a house**
- **Geographical Latitude**
- **View on Waterfront**

If we look to the scaled factors and rank their coefficients, we can reveal most important factors which have positive influence on prices. They are: living space (which is logically expected), geographical latitude and view of waterfront.

## FINDINGS (3/3)

**TOP 3 NEGATIVE FACTORS:**

- **Number of bathroom (<0,5 per bedroom)**
- **Low grade (< 7)**
- **Lot square area**

The top 3 negative factors on prices from the model are: low number of bathrooms (less than 0,5 per bedroom), low grade in King County grade system (less than 7), and lot square area.

## BUSINESS RECOMMENDATIONS

**TARGETS to BUY / INVEST (with discount) and SELL (with premiums):**

- **houses located closer to WEST COAST and with Waterfront view**

**TARGETS to SELL (with discount) and avoid INVESTMENTS:**

- **houses with low grade, with low number of bathroom and with high lot square**

Depending on what the Company planning to do with specific house: sell or buy, different strategies could be recommended. But general recommendations are: to concentrate investments on high-valued houses, which mainly located closer to WEST COAST and with view on Waterfront. Houses, which are graded lower than 7 according to King County grading system, with small numbers of bathroom and with high square of lot – are not highly demanded and should be sold as assets with low liquidity.

## FUTURE WORK

- **Updated data**
- **Geographical clustering**
- **Outliers and anomalies**
- **Alternative sets of binned predictors**

All specific outcomes are the result of input data and applied methodologies. Important point here is than any ML-model should learn after new real data added to the dataset in order to be actual on any point of time. So, maintaining and updating the model is crucial. And there are some areas which could give improvement to current version of the model. First, clustering location of houses could provide clearer picture of what specific locations are good for prices. Second, working deeper with outliers is another way to improve R-squared. And finally, there are a lot of combinations of binned and regrouped factors, which need to be tested on the model.

# THANK YOU

# Q&A

PRESENTATION BY NURAN KAIRAKBAYEV

Thank you for attention. Any questions?

# APPENDICES

PRESENTATION BY NURAN KAIRAKBAYEV

On the next slides you can see specific and detailed information related to MODEL

## A1. DESCRIPTION of INITIAL VARIABLES

### PREDICTION TARGET – "PRICE"

### NON-NUMERICAL VARIABLES

- "ID" – unique identifier for House (but not unique in Data Set);
- "Date" – date of House sold;
- "zipcode" – House's zip adress;

### ORDINAL CATEGORICAL VARIABLES and DUMMIES

- "waterfront" – is a House has view to waterfront (1-yes, 0-no);
- "conditions" – how good is condition (Ordinal variable);
- "grade" – overall grade based on King County grading system;

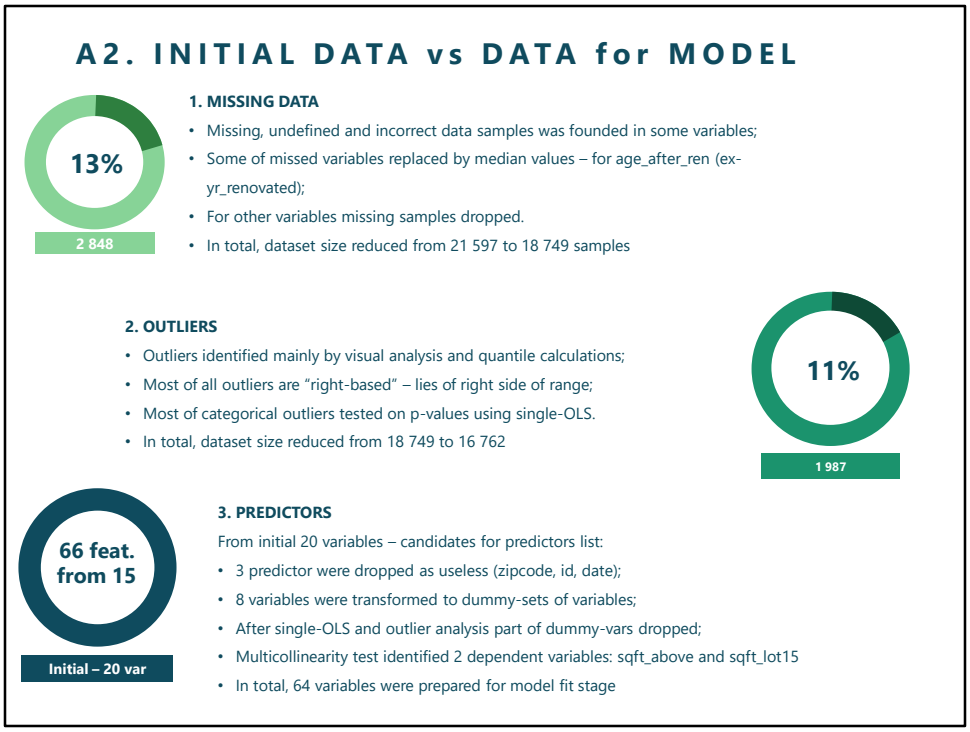
### NUMERICAL VARIABLES WITH DISCRETE VALUES

- "bedroom" – number of bedrooms per house;
- "bathroom" – number of bathroom per bedroom;
- "floors" – number of floors in House;
- "view" – how many times a House has been viewed;

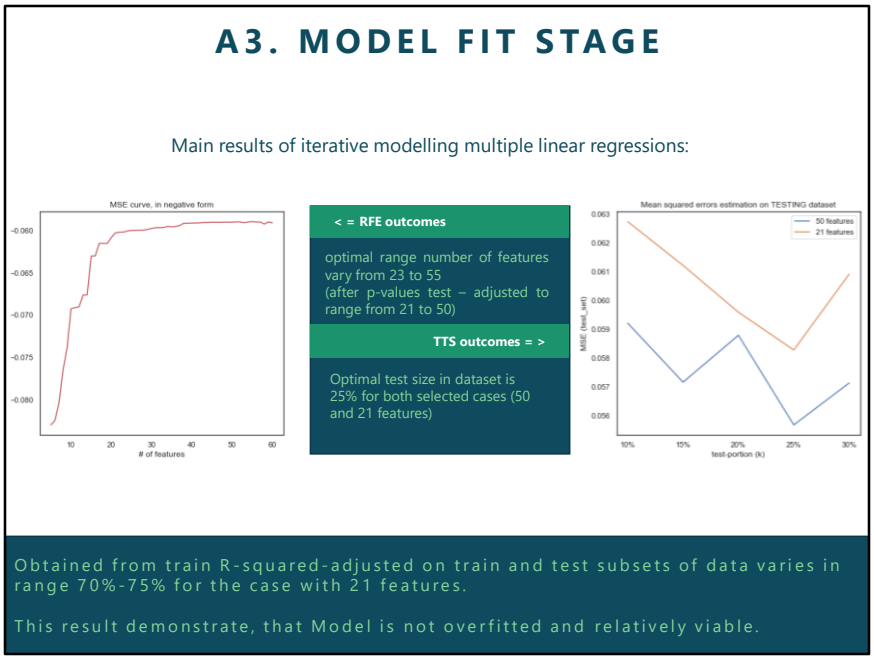
### NUMERICAL VARIABLES WITH CONTINUOUS VALUES

- "sqft_living" – footage of the home;
- "sqft_lot" – footage of the lot;
- "sqft_above" – footage of the home apart from basement;
- "sqft_basement" – footage of the home's basement;
- "sqft_living15" – footage of the nearest 15 homes (average);
- "sqft_lot15" – footage of the nearest 15 lots (average);
- "lat" and "long" – latitude and longitude (location parameters);
- "yr_built" and "yr_renovated" – year, when house was built and renovated

This slide demonstrates initial variables descriptions available in the KS_HOUSE_DATA database. Including the target variable (price), there are 21 variables with different data types with about 21,600 samples of those.




Here, main outcomes of data cleaning and exploring stages are provided.
Of the identified missing data, slightly more than 2,800 samples were removed from the sample (about 13% of the total sample power), about 2,000 observations were removed as outliers (11%). Thus, further modeling was carried out on 16,800 data.
Of the original 20 predictors, 15 were recognized as important for analysis, some of which were converted to dummy-variables. The total number of independent variables increased to 66 and some adjustments within MK-test.




This slide illustrates the main findings of the applied machine learning approaches. So, the graph on the left shows a learning curve that displays the average Mean Squared Error (in negative form) as the number of predictors increases. It can be seen that the form of the graph is logarithmic and becomes quite flat from 23 variables. It gives us optimal range for predictors – from 23 to 55. The graph on the right shows the dependence of this indicator on the volume of the test sample. The optimal value is 25%.

## A4. OUTCOME FORMULA

PREDICTION TARGET:

LN(**PRICE**) =

PREDICTORS (coefficient given as rounded numbers):

= – 49.63 + 2.21 LN(**SQFT_LIVING**) – 0.46 LN(**SQFT_LOT**) + 1.14 LN(**SQFT_LIVING15**) –

– 5.85 **YR_BUILT** + 64.41 **LAT** + 0.52 **WATERFRONT_1.0** – 0.48 **BATHROOM_0.5** +

+ 0.21 **BATHROOM_3.75** + 0.12 **BATHROOM_4.0** + 0.19 **BATHROOM_4.25** +

+ 0.09 **CONDITION_3** + 0.14 **CONDITION_4** + 0.21 **CONDITION_5** – 0.34 **GRADE_4** –

– 0.41 **GRADE_5** – 0.33 **GRADE_6** – 0.18 **GRADE_7** + 0.21 **GRADE_9** + 0.36 **GRADE_10** +

+ 0.48 **GRADE_11** – 0.15 **VIEW_0**

RESULTS INTERPRETATIONS:

- There are 13 predictors with positive influence 8 – with negative effect on price
- We could state, that in general space, latitude, year of building, number of bathroom, condition, grade rank, is house viewed or not and view on waterfront – are the main predictors for prices.
- Negative correlation with SQFT_LOT and YR_BUILD was surprised here, but we need to look deeper for house market specifics for understanding this point.
- In the group of dummy-variables the TOP3 strongest features are: WATERFRONT_1.0, BATHROOM_0.5 and GRADE_11.
- Some subgroups of dummies has clear picture of the point, dividing negative and positive dummies: for example, we can state, that all houses with grade less that 8 traded with discounts, and graded higher than 8 uses some premium on prices.

Here, main regression formula is shown.