**The Project: House Prices Prediction**

**Project purpose**

Our goal is to produce a model which may estimating the price of a given real-estate asset. in order to achieve our goal, we intend to use the following regression models:

(a) Linear Regression

(b) Decision-Tree

(c) KNN

**The Dataset**

(a) Data was taken from Kaggle - <https://www.kaggle.com/harlfoxem/housesalesprediction>

(b) This dataset provides information about real estate assets features associated with the price of sale in King County, USA

**The Team**

**- Anat Shulman (**[**sanatash@gmail.com**](mailto:sanatash@gmail.com)**)**

**- Alon Avramson (**[**alon1122@gmail.com**](mailto:alon1122@gmail.com)**)**

**- Nadav Gordon-Bar (**[**nadavgordonbar@gmail.com**](mailto:nadavgordonbar@gmail.com)**)**

**- Shuky Persky(**[**shukyp@gmail.com**](mailto:shukyp@gmail.com)**)**

## ****Geographic Illustration****

Map

Description automatically generated

Map

Description automatically generated

**Data Set Features**

A screenshot of a computer

Description automatically generated with medium confidence

**Shape:** rows: 21613   
 columns: 21 (Features)

**NaN/NA:** Found to be 0

## Duplication: Found to be 0

## ****Features Explanation****

* ′id′ - ID value attached to the sale event (no specification)
* ′date′ - Unclear whether it is the date of sale or date the asset was registered fro sale (date) (date)
* ′price′ - House sale price (numeric)
* ′bedrooms′ - Number of house bedrooms (numeric)
* ′bathrooms′ - Number of house bathrooms (numeric)
* ′sqft living′ - Size (Ft\*\*2) of house living region (numeric)
* ′sqft lot′ - Size (Ft\*\*2) of house yard area (numeric)
* ′floors′ - Number of house floors (numeric)
* ′waterfront′ - Wether or not the house has a waterfront (categorial/ binary)
* ′view′ - grade of the house sorrounding view (categorial/ordinal)
* ′condition′ - grade the house condition(categorial/ordinal)
* ′grade′ - grade of house general condition (categorial/ordinal)
* ′sqft above′ - Size (Ft\*\*2) of house living area above the ground(numeric)
* ′sqft basement′ - Size (Ft\*\*2) of house living area below the ground(numeric)
* ′yr built′ - The year of the housebuild (numeric)
* ′yr renovated′ - The year the house was renovated in (numeric)
* ′zipcode′ - Zipcode location feature (no specification, can be categorial/ordinal)
* ′lat′ - Geographical latitude location feature (no specification, can be categorial/ordinal)
* ′long′ - Geographical longitude location feature (no specification, can be categorial/ordinal)
* ′sqft living15′ - Size (Ft\*\*2) of house living area of 15 neighboring assets (numeric)
* ′sqft lot15′ - Size (Ft\*\*2) of house yard area of 15 neighboring assets (numeric)

## ****Drop non-contributing Columns/Features****

1 - **id** : non-contributing Feature

2 - **date**: to our understanding a date doesn't affect the the sale price

3 - **SqFt living** overlaps other features which will be used (a) above size (b) basement size

4- **zip code**  
 (a) not sure we have enough coverage of the problem region  
 (b) will be used on further testing after having 1st cycle models

5 - **Yr\_renovation** overlaps condition feature

6 - **lat** and **long**  
 (a) seems like an overlap of zip information.  
 (b) not sure we have enough coverage of the problem region

7 – **VIEW** most entries stamped as Zero, unclear whether anyone borthered filling it

**Further Manipulation:**

1- **Zip Code** will be replaced by the log of the price mean for all assets that reside in a given zip code area

**After reduction : Dataset is based on 14 features**

## ****Transform Prices to log scale****

1 - **price** :

2 - **price-based-zip**:

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

## ****Transform sqft\_xxxx features****

**The following features are transformed as described below**

**1. sqft\_lot   
2. sqft\_above   
3. sqft\_basement   
4. yr\_built   
5. sqft\_living15   
6. sqft\_lot15**

**Transformation**

**(a) Log based**

**(b) Normalization: MaxAbs**

**(c) Normalization: MinMax**

**(d) StandardScaler**

**\* LR was done w\wo ElasticNet (StandardScaler was better)**

## ****Remove Outliers****

**We located outliers by using:  
(1) boxplot - to understand there are actually outliers  
(2) cumsum – to decide where to put the limit which defines the range to removed**

**Example**

Chart, histogram

Description automatically generated Chart

Description automatically generated

**After deciding where the limits are – we set HYPER PARAMS for the cut**

MAX\_BEDROOMS = 10

MAX\_BATHROOMS = 5

MAX\_SQFT\_LOT =  500000

MAX\_SQFT\_ABOVE =  6000

MAX\_SQFT\_BASEMENT =  2500

MAX\_SQFT\_LIVING15 = 5000

MAX\_SQFT\_LOT15 = 300000

* We tried more outlier removal (minimum and none).

<class 'pandas.core.frame.DataFrame'>

Int64Index: **21489 entries**, 0 to 21612

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 price 21489 non-null float64

1 bedrooms 21489 non-null int64

2 bathrooms 21489 non-null float64

3 sqft\_lot 21489 non-null int64

4 floors 21489 non-null float64

5 waterfront 21489 non-null int64

6 condition 21489 non-null int64

7 grade 21489 non-null int64

8 sqft\_above 21489 non-null int64

9 sqft\_basement 21489 non-null int64

10 yr\_built 21489 non-null int64

11 sqft\_living15 21489 non-null int64

12 sqft\_lot15 21489 non-null int64

dtypes: float64(3), int64(10)

**Features: 13 (price-based-zip will be added soon)**

**Rows: 21,489 (initially: 21613**)

**HeatMap**

Chart, treemap chart

Description automatically generated

**Illustration:**

**Price-Based-Zip Map**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**Chart, scatter chart

Description automatically generated

**Log based Transformation**

**Linear Regression**

**RMSE = (0.21) 0.21**

Chart, bar chart

Description automatically generated **Chart, scatter chart

Description automatically generated** Chart, histogram

Description automatically generated

**Decision Tree**

bedrooms : 0.0

bathrooms : 0.0216841724132233

sqft\_lot\_log: 0.007775377676905517

floors : 0.00046794764004093456

waterfront : 0.001088077861612243

condition : 0.0003186684341140099

grade : 0.29410581216148163

sqft\_above\_log: 0.049972845924627314

sqft\_basement\_log: 0.02295474881342498

yr\_built : 0.013225568216013934

sqft\_living15\_log: 0.009510965167501552

sqft\_lot15\_log: 0.00240163117434811

**RMSE = (0.22) 0.24**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Auto Process to find optimal value for Max Leaf Values**

Chart

Description automatically generated

DT\_MAX\_DEPTH = 7

DT\_MIN\_SAMPLES\_LEAF = 7

DT\_MIN\_SAMPLES\_SPLIT=3

DT\_MAX\_LEAF\_NODES = 50 # max

Diagram

Description automatically generated

**K-NN**

**RMSE = (0.26) 0.28**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Auto Process to find optimal value for Max Leaf Values**

Chart, line chart

Description automatically generated

**Scale/MaxAbs Transformation**

**Linear Regression**

**RMSE = (0.21) 0.22**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Decision Tree**

**RMSE = (0.29) 0.30**

Chart

Description automatically generated Chart, histogram

Description automatically generated

**K-NN**

**RMSE = (0.22) 0.28**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Scale/MinMax Transformation**

**Linear Regression**

**RMSE = (0.21) 0.22**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Decision Tree**

**RMSE = (0.29) 0.30**

Chart

Description automatically generated Chart, histogram

Description automatically generated

**K-NN**

**RMSE = (0.22) 0.28**

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Scale/ StandardScaler Transformation**

**Linear Regression**

**RMSE = (0.21) 0.21**

Chart, scatter chart

Description automatically generated

**Decision Tree**

**RMSE = (0.23) 0.24**

Chart, scatter chart

Description automatically generated

**K-NN**

**RMSE = (0.19) 0.20**

**RUN OVER FEATURES:**

############################################################

run for: sqft\_lot15 testing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model: LinearRegression(copy\_X=True, fit\_intercept=True, n\_jobs=None, normalize=False)

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train : 0.2132892787743452

RMSE\_test : 0.21372417865593987

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train\_scaled : 0.21323955359791355

RMSE\_test\_scaled : 0.21363289068243538

############################################################

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model: DecisionTreeRegressor(ccp\_alpha=0.0, criterion='mse', max\_depth=None,

max\_features=None, max\_leaf\_nodes=50,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, presort='deprecated',

random\_state=None, splitter='best')

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train : 0.22638371264276239

RMSE\_test : 0.23848812932434427

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train\_scaled : 0.22638371264276239

RMSE\_test\_scaled : 0.2416426815512984

############################################################

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model: KNeighborsRegressor(algorithm='auto', leaf\_size=30, metric='minkowski',

metric\_params=None, n\_jobs=None, n\_neighbors=18, p=2,

weights='uniform')

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train : 0.32961333577499846

RMSE\_test : 0.35560005855090876

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RMSE\_train\_scaled : 0.1985351714981046

RMSE\_test\_scaled : 0.20997328248571934

############################################################

############################################################