# Machine Learning, Spring 2018

**Project Proposal**

Badouh, Asaf
Asaf.Badouh@gmail.com

Manubens, David
DavidManubens89@gmail.com

Rodriguez, Pau
Rodriguez.Pau@gmail.com

April 2018

# 1. House Prices Prediction

Since forever, houses have been the most desirable purchase in humans' lifetime. Prices of houses are determined by many factors since houses differ in many aspects such as location, size, condition etc. In 2016, around 5.45 million of existing houses were sold in the U.S.[1] and there is a steady rise in sales after the "Subprime mortgage crisis" drop in sales in 2008 (Full statistics in Appendix:A). With more 5.45 million transactions per year just in the U.S., without taking into account the new houses transactions, predicting house prices is a hot and interesting topic.

# 2. The Dataset

The "House Sales in King County, USA" dataset is published in Kaggle[2]. It contains 21613 observations with 19 house features and the predication attribute "Price". Aside of "Date" feature, all of the features are numeric. However, there are some features that we might need to pre-process. Latitude and Longitude, for example, are two features that we might consider to transform to one composed feature that will represent the distance from the city center. Full features table in Appendix:B.

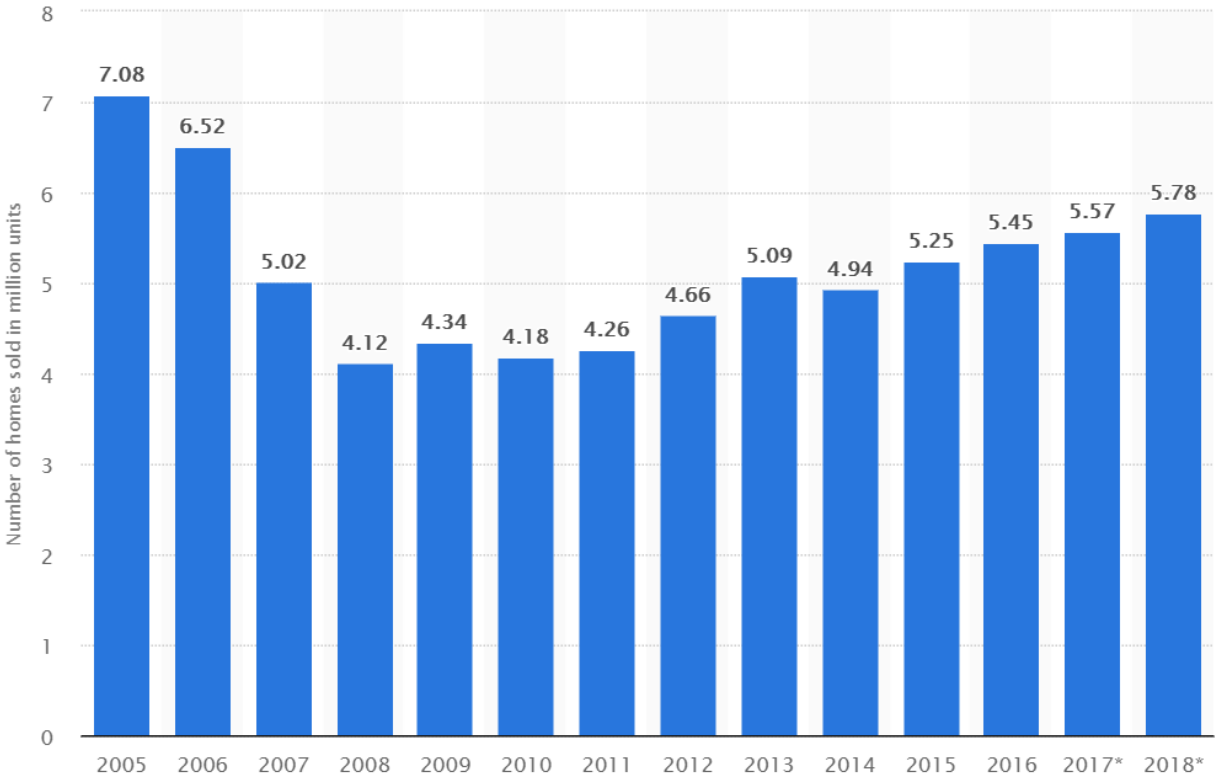
# 3. Goal

The goal of this project is to make the most out of data on house prices to build a model that is able to predict the price of a house based on it's attributes. This includes pre-processing the data, performing feature engineering (feature selection and/or extraction) and training a linear regression model to predict the price.
An important part of our effort will be focused on the decision about what are the most important variables or features that we could extract upon them, as well as what are the best parameter configurations for the model we will train. Different approaches to each phase will be considered and compared appropriately before choosing the final implementation.

# Appendices

## A. House Sales 2005-2018[1]

We can see the raise over the years of house sales, The effect of the "Subprime mortgage crisis" and the recovery over the years. The graph shows the number of existing homes sold in the United States from 2005 to 2016, and a forecast thereof for 2017 and 2018 (in million units).



## B. Features Description Table

| | | |
|---|---|---|
| id | a notation for a house | Numeric |
| date | Date house was sold | String |
| price | Price is prediction target | Numeric |
| bedrooms | Number of Bedrooms/House | Numeric |
| bathrooms | Number of bathrooms/bedrooms | Numeric |
| sqft_living | square footage of the home | Numeric |
| sqft_lot | square footage of the lot | Numeric |
| floors | Total floors (levels) in house | Numeric |
| waterfront | House which has a view to a waterfront | Numeric |
| view | Has been viewed | Numeric |
| condition | How good the condition is ( Overall ) | Numeric |
| grade | overall grade given to the housing unit, based on King County grading system | Numeric |
| sqft_above | square footage of house apart from basement | Numeric |
| sqft_basement | square footage of the basement | Numeric |
| yr_built | Built Year | Numeric |
| yr_renovated | Year when house was renovated | Numeric |
| zipcode | zip | Numeric |
| lat | Latitude coordinate | Numeric |
| long | Longitude coordinate | Numeric |
| sqft_living15 | Living room area in 2015(implies– some renovations) This might or might not have affected the lotsize area | Numeric |
| sqft_lot15 | lotSize area in 2015(implies– some renovations) | Numeric |

Table 1: Dataset's features

## References


[1] The statistic portal. Number of existing homes sold in the united states from 2005 to 2018. `https://www.statista.com/statistics/226144/us-existing-home-sales`, 2018. Accessed: 5th April, 2018.

[2] Kaggle: Your Home for Data Science. House sales in king county, usa. `https://www.kaggle.com/harlfoxem/housesalesprediction`, 2016. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015, Accessed: 4th April, 2018.