Module 1 Final Project – Predicting King County House Sales

BEVERLY DELAROSA

SELF – PACED (TUESDAY, MAY 7 @ 4PM PST) | | | BRANDON LEWIS

What we’re dealing with

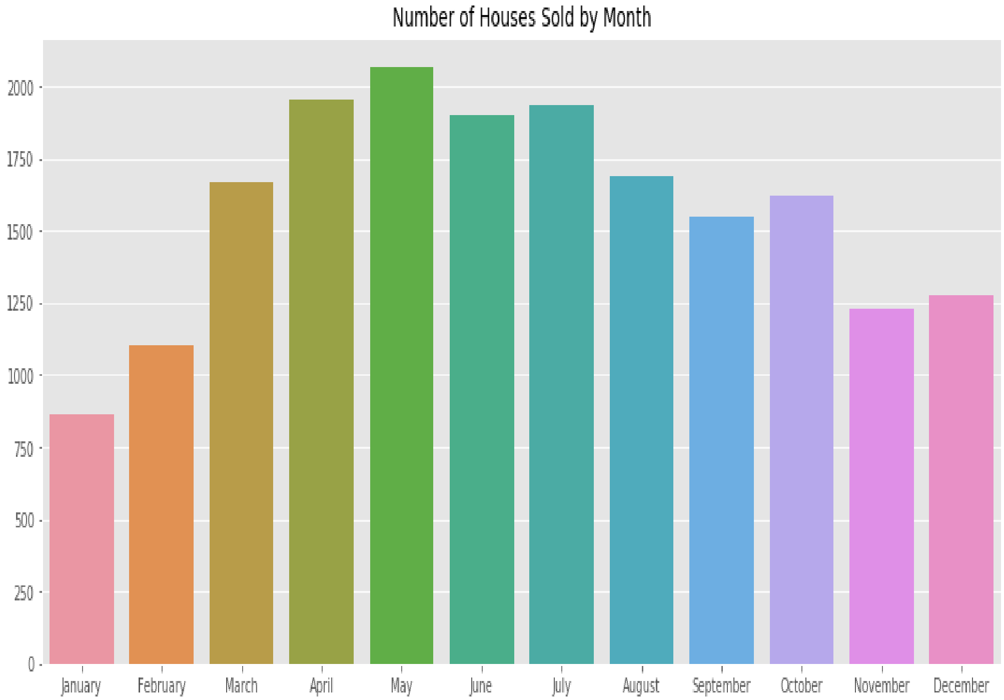
- Objective: Predict the sale price of houses as accurately as possible.

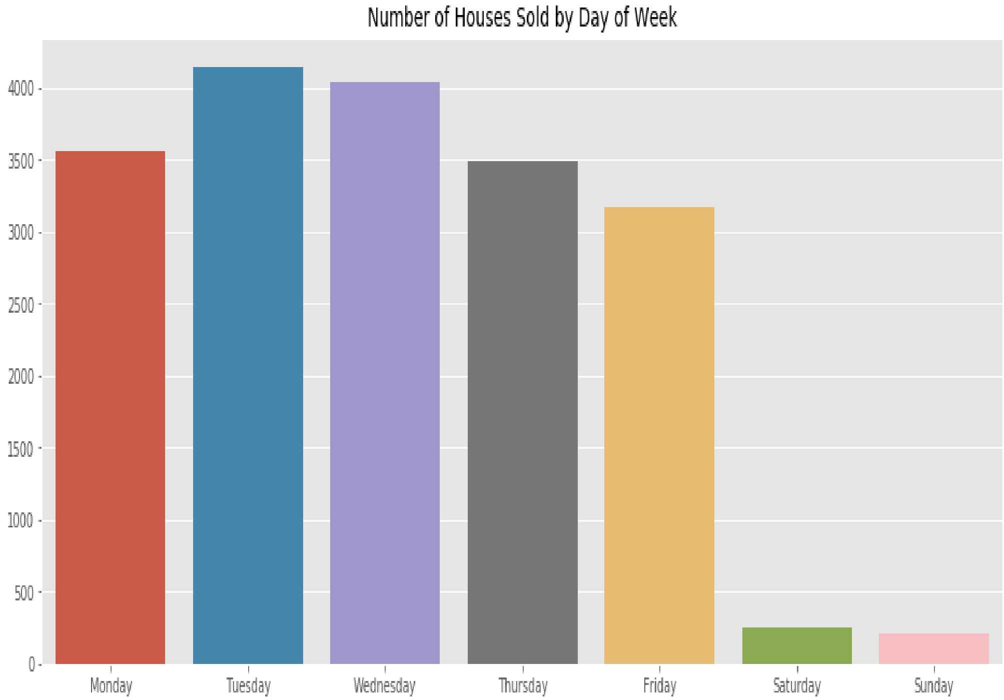
- Our original, raw data: King County House Sales.
  - 20 features that we can use as possible predictors for our target (Price).
    - Some features were the wrong type of data (Date, Square Footage of Basement), or redundant (Square Footage of the Living Space, and “Neighborhoods”).
  - Data for 21,597 houses sold.
    - Some houses that were sold were incomplete (Waterfront, View, Year Renovated), or had drastic outliers (Bathrooms, Bedrooms, Square Footage of the Living Space and Whole Lot, and Price).

What we’re working with

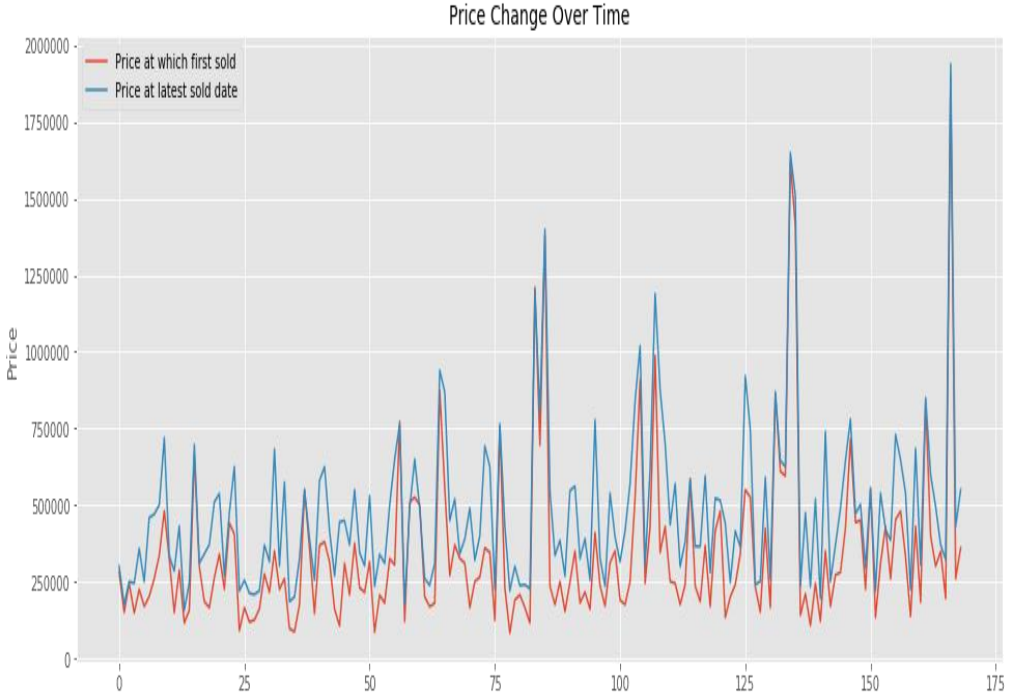
- Our reduced data.
  - 20 → 14 features that we can use as possible predictors for our target (Price).
    - Bedrooms, Bathrooms, Date, Id, Square Footage of Living and Lot, Floors, Waterfront, View, Condition, Grade, Year Built and Renovated, and ZIP Code.
  - Complete data for 21,597→18,843 houses sold.

Results – Does renovating the house or the time of year affect price?





Results – Does renovating the house or the time of year affect price?

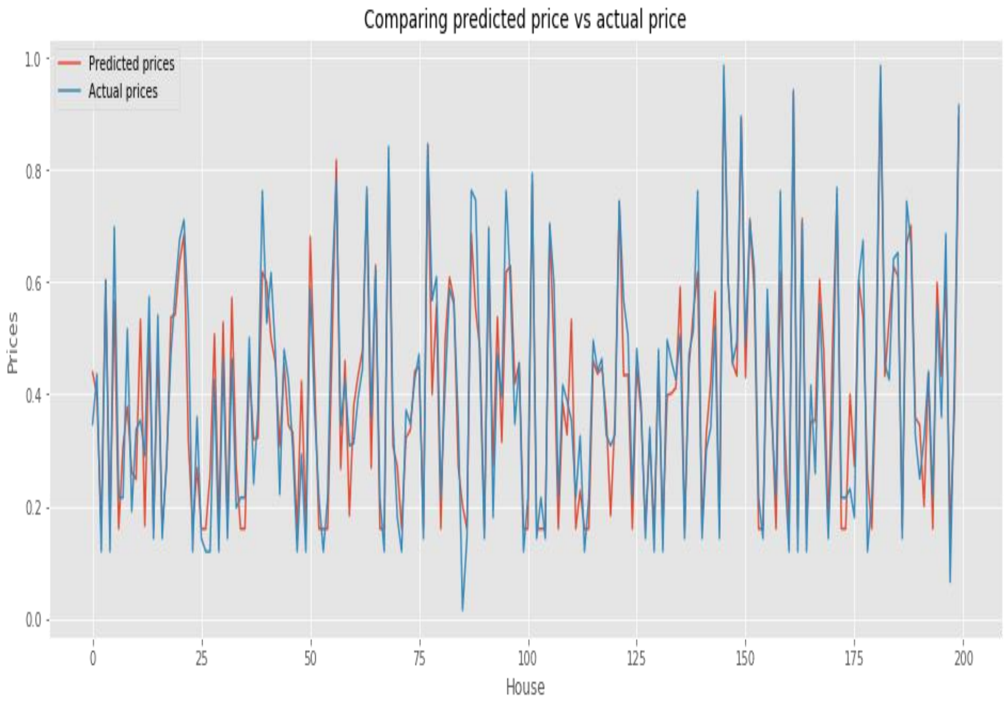


- Number of houses sold once: 18504.
  Number of houses sold twice: 168.
  Number of houses sold thrice: 1.
- Average price percent change between the first and latest date sold: 29 %.
- Average living square footage percent change between the first and latest date sold: 0 %.
  Average lot square footage percent change between the first and latest date sold: 0 %.

Predicting our model

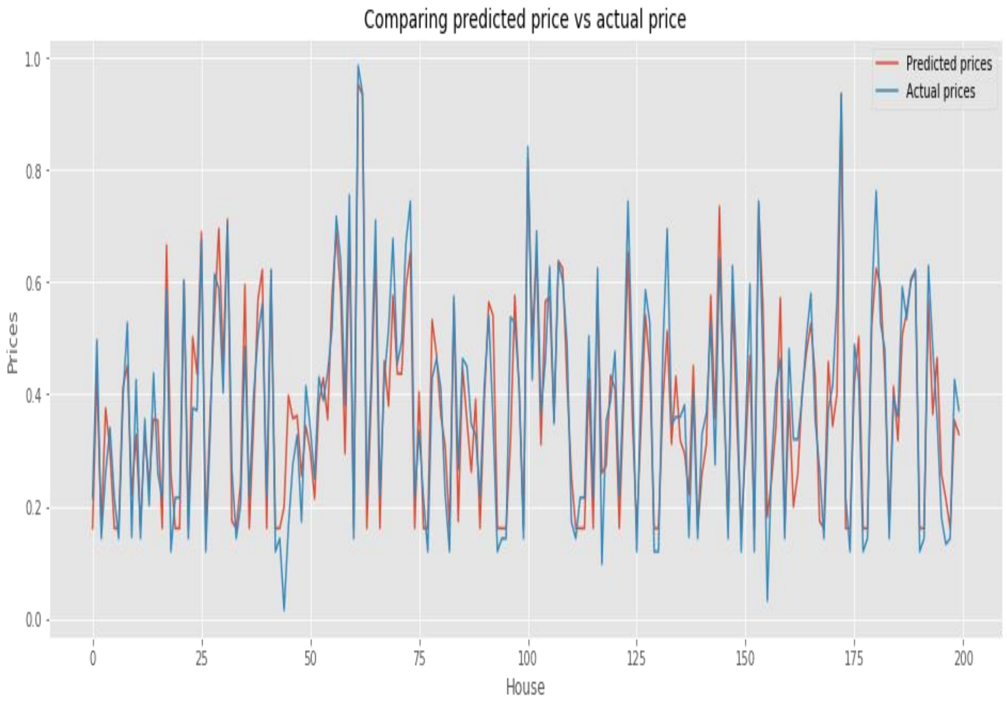
- Standardizing our data.
  - Normalizing, scaling, transforming our continuous variables (Price, Square Footage of Living and Lot).
  - Creating dummy variables for our discrete variables (Bedrooms, Bathrooms, Floors, Waterfront, View, Condition, Grade, Year Built, and ZIP Code)
- Doing a full multiple linear regression on our expanded data.
  - 249,585 rows X 220 columns

Run 1 Results – Is our predictive model a good fit to estimate a house price in King County?



- (Adjusted) R^2: 87.6%
- Accuracy: -5.347115145732844e+19% not within appropriate [0, 1 range]
- Our OLS Summary report showed that we have a large conditional number, meaning we may have multicollinearity among our features.
- Feature Selection (α = .005)

Run 2 Results – Is our predictive model a good fit to estimate a house price in King County?



- (Adjusted) R^2: 87.2%
- Accuracy: 68.4%
  (with > 99% confidence)
- The most influencing features affecting the price are location (ZIP code) and square footage of the house.