King County House Prices

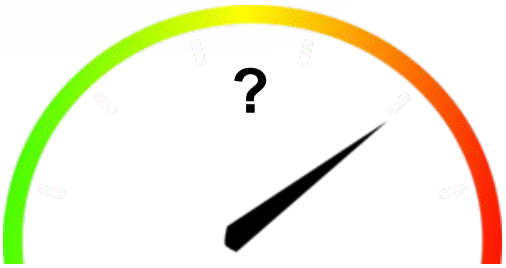
- Recommendations for Home Sellers & Buyers -

by Karima Chakroun
11/04/19

- welcome to my presentation about King County House Prices - recommendations for home sellers & buyers

- image source: https://storage.googleapis.com/kaggle-datasets-images/128/270/d149695d1f9a97ec54cf673be6430ad7/dataset-card.jpg

Business problem



$ 78 000 **$ 7 700 000**

renovated grade waterfront
bathrooms floors basement
living area bedrooms location

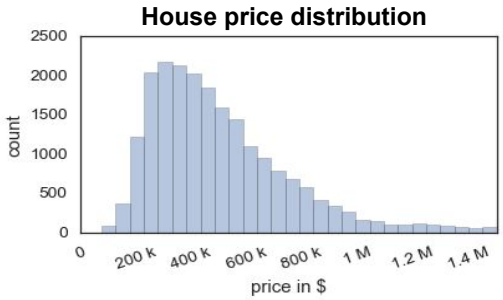
- houses come in different flavors, leading also to an extreme range in selling price
- e.g. house sale prices in King County ranged from $ 70 000 up to >$ 7 million (in the years 2014/2015)
- home sellers/buyers face the question: Where is my house located in this extreme price range?
- there are several features that determine the sales price of a house and can be used for price prediction

- **Business problem**: How can house features be used to predict selling price?
- **Business goal**:
  - explore the impact of different features on the house price
  - come up with some useful recommendations for home sellers/buyers in King County

- image sources:
  - https://www.flickr.com/photos/dyamasaki/463569490,
  - https://rdcnewscdn.realtor.com/wp-content/uploads/2018/02/KY-home-02-628x354.jpg

Methodology

- Dataset:
  - >20 000 houses in King County
  - 20 features to predict price
- Exploratory data analysis
- Linear Regression (simple, multiple)

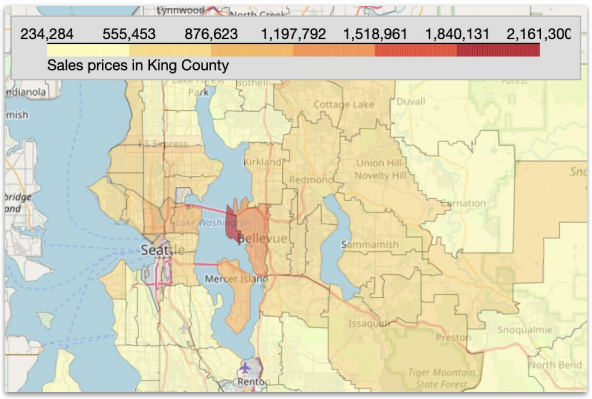


- dataset used for this project:
  - included historical sales prices of >20 000 houses in King County, sold between 2014 and 2015
  - included 20 features to predict price
  - most houses were located in the price range between $ 100 000 and $ 1 million, but some of them also ranged up to several million $

- project methodology:
  - included exploratory data analysis and data visualization
  - included both simple and multiple linear regression, i.e. models that linearly predict price using one or several features as predictors

- map source: https://de.wikipedia.org/wiki/King_County_(Washington)

Location (zipcode)



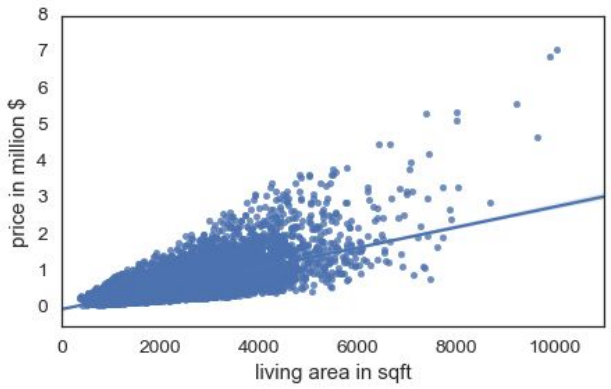
- explains **>40%** of price
- up to 10-fold price difference!

Recommendation 1: **Choose the right location!**
(e.g. Mercer Island, Bellevue, Lake Washington)

- first feature with major impact on price: location (coded in the dataset by the zipcode)
- zipcode alone already explained >40% of the sales price
- see map:
  - mean price of light yellow areas: ~$ 200 000
  - mean price or dark red area: >$ 2 million
  - → 10-fold price difference explained by location alone

- **Recommendation 1** (to home buyers): Choose the right location for your house!
  - for very expensive objects, choose the areas of Bellevue or Mercer Island

Living area



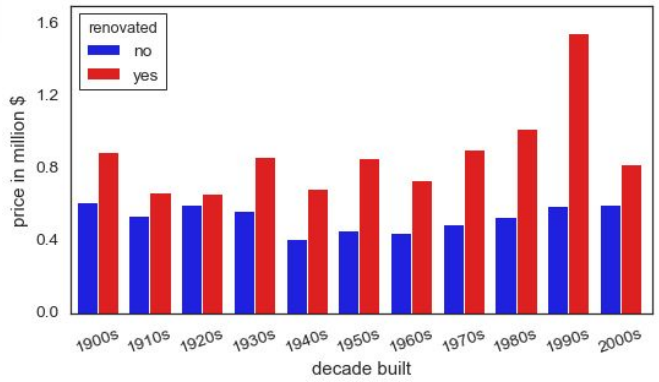
- explains **~50%** of price
- every 10 sqft raise price by **$ 2800**

Recommendation 2: **Increase the interior living area!**
(e.g. by winter garden, additional floors, …)

- second important feature: total interior living area (including both basement and above)
- explained ~50% of price
- for every additional 10 sqft of living space, the mean price increased by ~$ 2 800

- **Recommendation 2** (to increase price): Increase the interior living area!
  - e.g. by building a winter garden, additional floors etc.

Renovation status



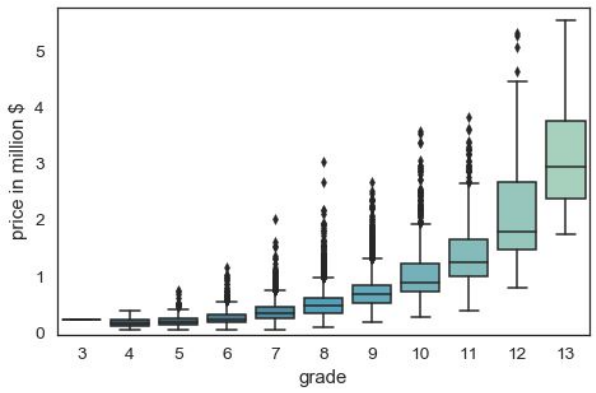
- renovated houses bring **$ 238 000** more
- < 5% of houses renovated

Recommendation 3: **Renovate the house!**
(e.g. electricity, thermal isolation, bathrooms, ...)

- third important feature: renovation status of the house
- renovated houses (red bars) were on average worth 238 000 $ more than non-renovated houses (blue bars)
- renovation raises the price not only for very old houses, but also for houses built only 20-50 years ago
- <5% of houses renovated so far

- **Recommendation 3** (to increase price): Renovate the house!
    - e.g. by modernizing electricity, thermal isolation, bathrooms, ...

Grade



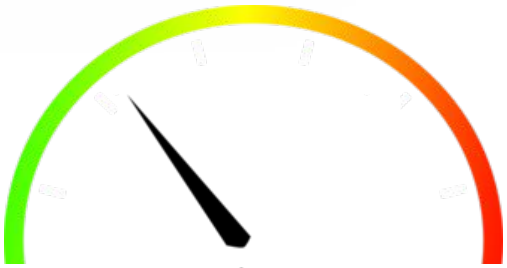
- explains **>50%** of price
- 8 to 9: **+$ 230 000**
- 11 to 12: **+$ 700 000**
- 12 to 13: **+$ 1 500 000**

Recommendation 4: **Improve the house's grade!**
(e.g. add amenities of solid woods, luxurious options, …)

- forth important feature: grade (based on the King County grading system, see appendix)
- grade alone explained >50% of price
- regression modeling showed that the selling price shows not only a linear, but a quadratic increase with grade (see appendix)
- for example:
  - upgrade from 8 to 9: raises price by $ 230 000
  - upgrade from 11 to 12: raises price by $ 700 000
  - upgrade from 12 to 13: raises price by $ 1.5 million

- **Recommendation 4** (to increase price): Try to increase the grade of the house!
  - e.g. an upgrade from 10 to 11 can be achieved by adding amenities of solid woods and luxurious options to the house

Conclusion



$ 78 000 **$ 7 700 000**

>80%

renovation **grade** waterfront

bathrooms floors basement

living area bedrooms **location**

- **Conclusion**: several features were presented that have a major impact on house selling prices in King County
  - most important ones: location (zipcode), total living area, renovation status, house grade
  - using these and further features in a multiple regression model, these predictors could together explain >80% of sales prices in King County

- **Recommendations** were given on:
  - a) features to consider (see above) when determining an appropriate price for selling or buying a house in King County
  - b) how to significantly boost the selling price of a house

Future work

- Feature engineering:
  - mathematical transformations
  - grouping of categorical features

- Using additional location data:
  - Foursquare, Google maps
  - nearby venues
    (shops, schools, parks, …)

Some prospects for future work include:

- feature engineering, i.e. taking the features we already have to build new features, which might be even better predictors of price
  - e.g. mathematical transformations like taking logarithm or inverse of a feature
  - e.g. grouping of categorical features, like combining zipcodes to larger areas

- using additional location data to build new features:
  - e.g. via Foursquare or Google maps
  - use latitude and longitude to get nearby venues of a house (shops, schools, parks, …)

- image source:
  https://geospatialmedia.s3.amazonaws.com/wp-content/uploads/2018/04/multi-location.jpg

Thank you!

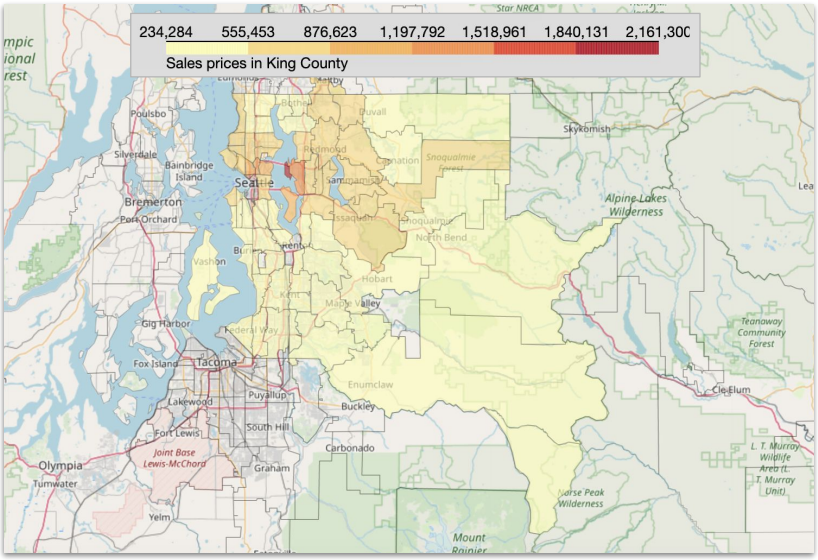
Eli
Dirk, Larissa
Bootcamp team!

- thanks to Eli, Dirk, Larissa and the whole bootcamp team for their support in this project

Appendix

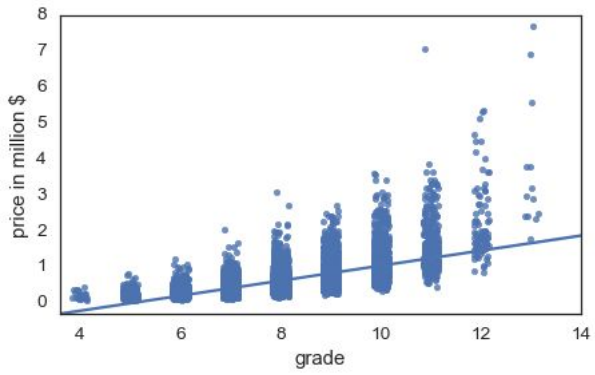
1) Full price map of King County
2) Grade as quadratic predictor
3) King County Grading System
4) Heatmap - Correlation analysis
5) GitHub project repository

Full price map of King County

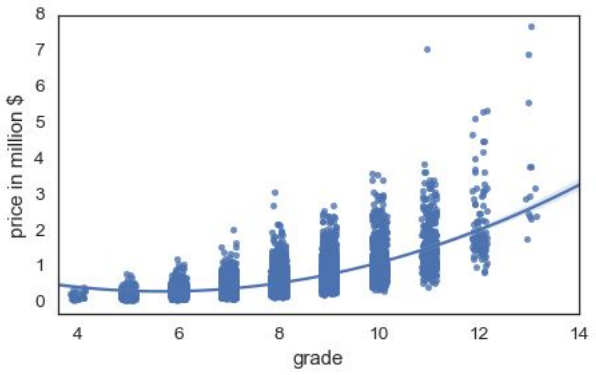


- map shows the mean sales price for each of the zipcodes (included in the dataset)
  - lowest mean price: zipcode 98002 ($ 234 284) - Auburn, Washington
  - highest mean price: zipcode 98039 ($ 2161300) - Medina, Washington

Grade as quadratic predictor



Linear relationship: explains 44.6% of price



Quadratic relationship: explains 51.1% of price

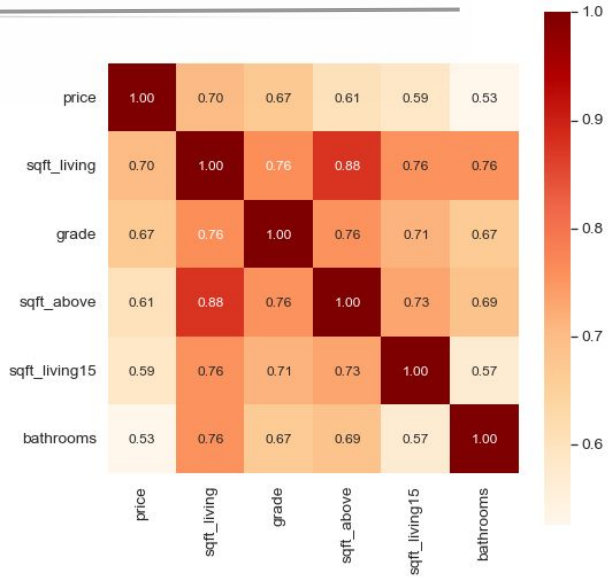
- linear regression modeling shows that the relationship between grade and price is quadratic rather than linear
  - model including only linear relationship between grade and price: explains 44.6% of price
  - model including quadratic relationship between grade and price: explains 51.1% of price

King County Grading System

- **Grade 1-3:** Falls short of minimum building standards. Normally cabin or inferior structure.
- **Grade 4:** Generally older low quality construction. Does not meet code.
- **Grade 5:** Lower construction costs and workmanship. Small, simple design.
- **Grade 6:** Lowest grade currently meeting building codes. Low quality materials, simple designs.
- **Grade 7:** Average grade of construction and design. Commonly seen in plats and older subdivisions.
- **Grade 8:** Just above average in construction and design. Usually better materials in both the exterior and interior finishes.
- **Grade 9:** Better architectural design, with extra exterior and interior design and quality.
- **Grade 10:** Homes of this quality generally have high quality features. Finish work is better, and more design quality is seen in the floor plans and larger square footage.
- **Grade 11**: Custom design and higher quality finish work, with added amenities of solid woods, bathroom fixtures & more luxurious options.
- **Grade 12:** Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.
- **Grade 13:** Generally custom designed and built. Approaching the Mansion level. Large amount of highest quality cabinet work, wood trim and marble; large entries.

Link:
https://www.kingcounty.gov/depts/assessor/~/media/depts/Assessor/documents/AreaReports/2018/Residential/015.ashx

Heatmap - Correlation analysis



- heatmap showing the Pearson correlation coefficients between sales price and the five numerical (non-categorical) features that correlate highest with price
    - ‘grade’: overall grade given to the housing unit, based on King County grading system
    - ‘sqft_living’: square footage of total interior housing living space (both basement and above)
    - ‘sqft_above’: square footage of interior housing living space above ground
    - ‘sqft_living15’: square footage of interior housing living space for the nearest 15 neighbors
    - ‘bathrooms’: number of bathrooms

GitHub project repository

This repository contains a data science project based on the King County House Sales dataset. The dataset can be found in the file "kc_house_data.csv" in this repository, and a description of the corresponding column names can be found in the “column_names.ipynb” file.

The Data Science Life Cycle goals for this project include: Data Cleaning, Data Exploration, Data Visualization and Predictive Modeling (by linear regression).

All parts of the data analysis are documented in the notebook “House_Prices_Project_Karima.ipynb” and include the following steps:

1) Load Packages and Dataset
2) Data Cleaning:
    a) Data types (deals with conversion of data types)
    b) Missing data (handles missing values)
3) Data Exploration & Visualization
    a) Overview of all features (using scatter_matrix plots)
    b) Exploring features one by one (histograms, pie/bar charts, scatter/regression plots)
    c) Correlation analysis (heatmaps)
    d) Folium map (mean price per zipcode)
4) Multiple Regression Model
    a) Model fitting and summary statistics
    b) Multicollinearity (checking variance inflation factors)
5) Customize Plots for Business Presentation

- readme of the GitHub project repository
- link to the GitHub project repository: https://github.com/KarimaCha/House_Prices_Project