Analysing and predicting housing prices

Presentation of the EDA-project

cgn-data-21-1

Christian Steck

Contents

1. What are we looking for?
2. What‘s the deal with the data?
3. Insights
    - What can we change?
    - When do we change it?
    - Location matters!
4. (Counterfactual) prediction
5. What should we do next?

What are we looking for?

- Exploratory Data Analysis of the real estate market
- Build a model to predict the price of a house

Leading questions:

- What does the data tell us about housing prices?
- How can we profit from the data?
- Is it in the interest of our stakeholder, a Texas construction company that wants to branch into the real estate market?

What’s the deal with the data?

- Dataset on houses in Kings County, TX
- n = 21,597 houses
- Target variable for analysis: price of the house in USD
- Features:
    - Size of the house (square footage, # of bedrooms, # of bathrooms, …)
    - Condition of the house (year built, year renovated, grade of housing unit, …)
    - Location of the house (size of neighbours houses, coordinates, zip code, …)

What’s the deal with the data?

Data cleaning process:

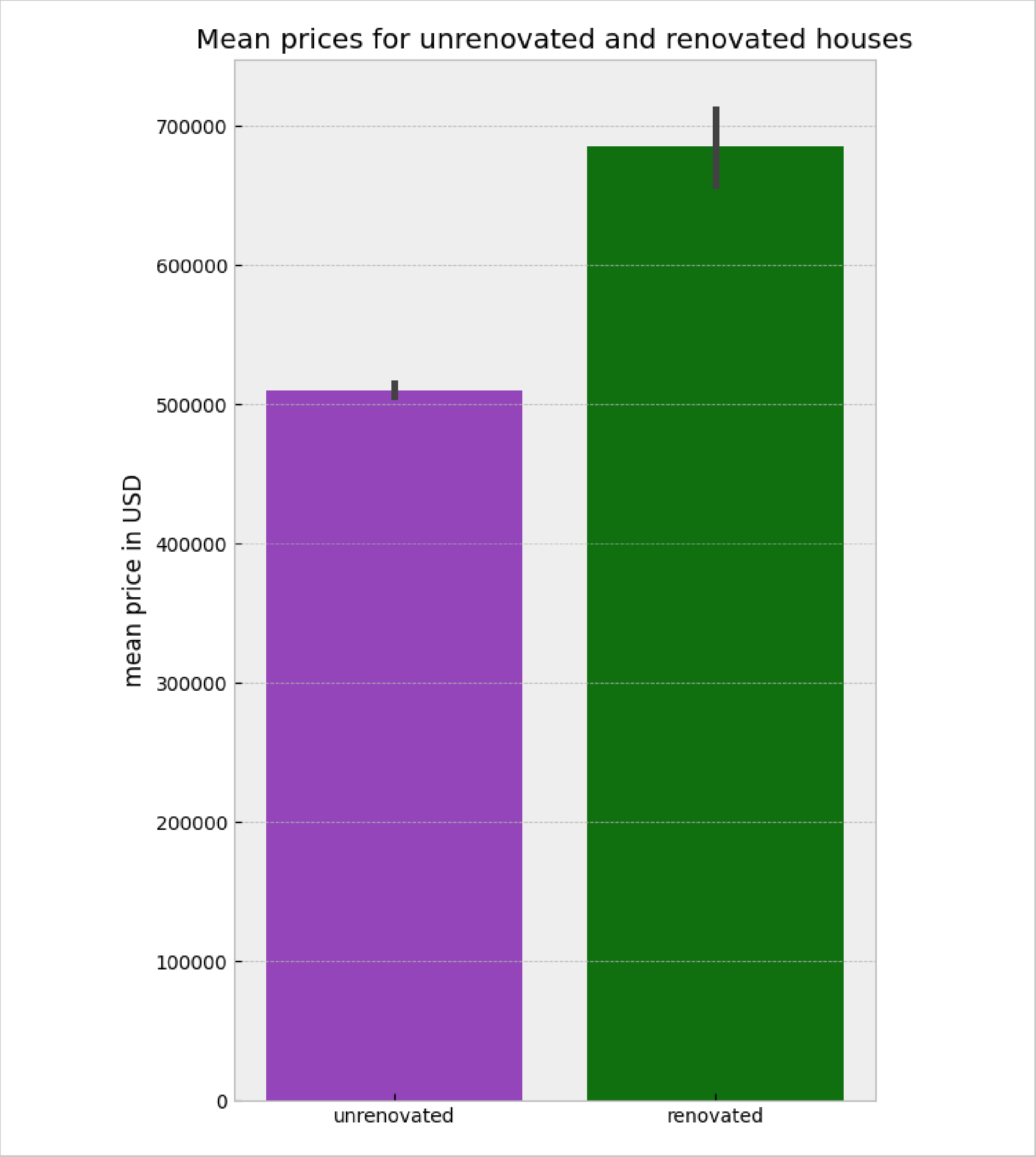
- Filtering the data for really large and expensive houses (> $2M, >10 bedrooms, > 6000 sqft)
- Dropping missing data
- Creating new features (age, renovation status)
- Only include houses that are built before 1985

Insights – What can we change?

- Renovation status as mutable property
- Much higher prices for renovated houses
- For houses older than 30 years:

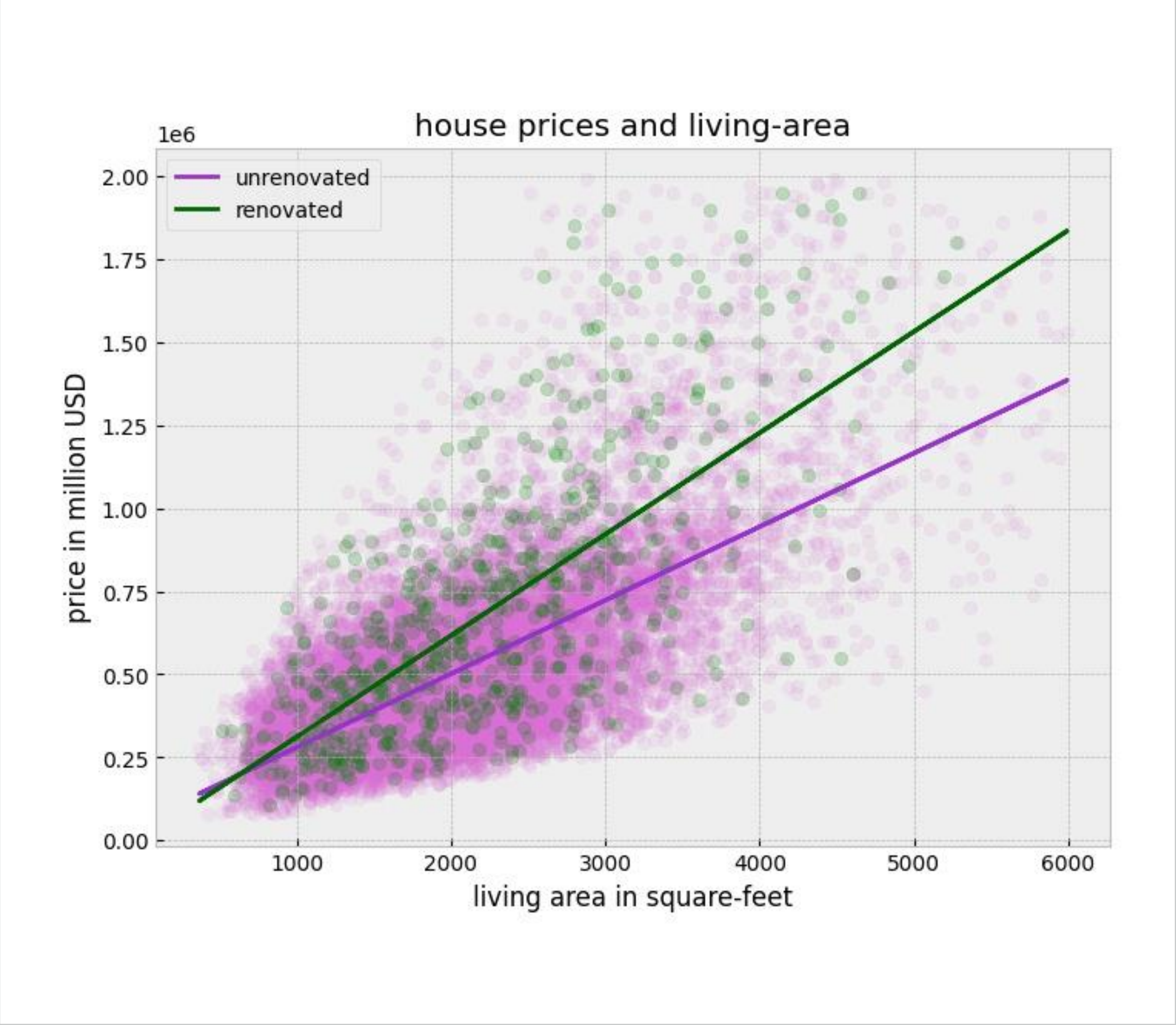
  Δ = $174,558
- Typical range to renovate house $18-75k[1]
- We can profit by renovating houses

[1] Source : https://www.homeadvisor.com/cost/additions-and-remodels/remodel-multiple-rooms/



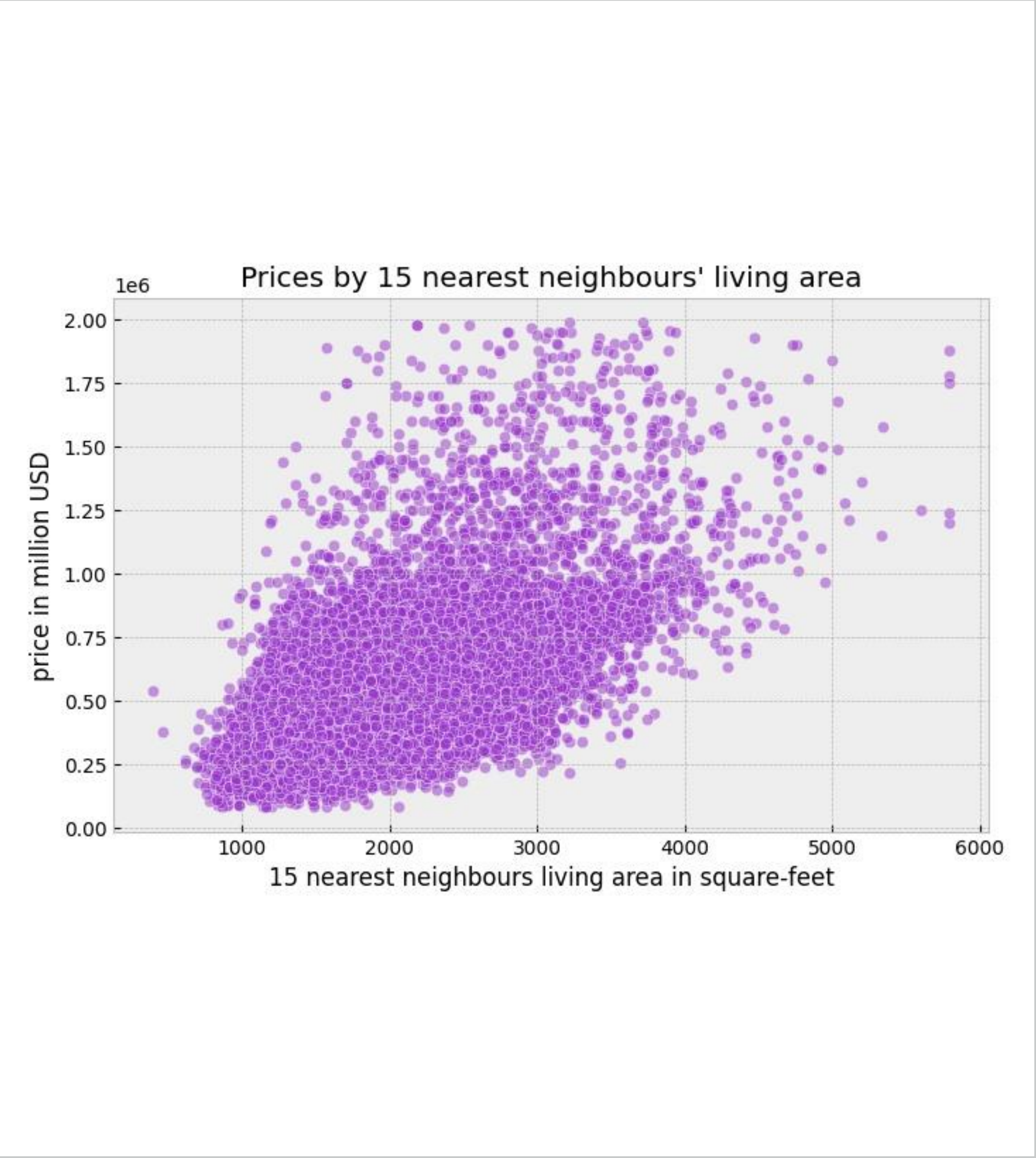
Insights - When do we change it?

- Price margins vary over features
- Tells us how profitable a house is
- Features that interact with renovation status:
    - Living are
    - # bathrooms
    - age
    - grade



Insights – Location matters!

- Prices vary by location
- Neighbours’ living area of highly correlates with price
- Zip code as a proxy for neighbourhood characteristics
    - ➢ Excellent predictor (28% RMSE-reduction)



Prediction

Prediction model based on EDA insights:

- Target variable: *price*
- Numerical variables: *bathrooms, bedrooms, sqft_living, sqft_living15, age, grade*
- Categorial variables: *floors, zipcode, ren*
- Interaction terms: *bathrooms*ren, sqft_living*ren, sqft_living15*ren, age*ren, grade*ren*
- Prediction metrics: $\mathrm{RMSE}_{\mathrm{all}} = 124527.94$
$\mathrm{RMSE}_{\mathrm{unren}} = 123137.91$
$\mathrm{RMSE}_{\mathrm{ren}} = 170586.08$

Counterfactual Prediction

- With our model we can predict what unrenovated and renovated are worth
- We can predict the potential profit of a house renovation if we change the value of renovation from 0 to 1
- Prototype that …

  … takes a dataset of unrenovated houses and a manually set minimum profit

  … returns dataset of profitable houses as csv

  … prints the proportion of profitable houses

What can we do next?

Our RMSE is large

➔ Solution: Further feature engineering

The RMSE for renovated houses is much larger

➔ Solution: Mine more data on renovated houses

Zipcodes aren’t meaningful and prone to overfitting

➔ Solution: Get data behind zipcodes (avg. income, infrastructure, etc.) from municipal statistics

The cost of renovations isn’t a fixed value

➔ (Longterm-)Solution: Mine data for a model predicting renovation costs