

# Real Estate Price Prediction

MINH PHUONG

# STEP 1: IDENTIFY PROBLEM

- ▶ Context: A local Vermont/New Hampshire real estate firm is looking into modeling closed prices for houses.
- ▶ Question: CAN YOU PREDICT CLOSED PRICE OF A PROPERTY USING THEIR LISTING FEATURES

# STEP 2: OBTAINING DATA

- ▶ Data was obtained from Kaggle.com
- ▶ This dataset contains features of houses in three towns in Vermont, which make up a sizable chunk of the real estate firm's business.
- ▶ MLS.com is the real estate information platform that is publicly available. Features were exported from an MLS web platform. (Perhaps webscraping)

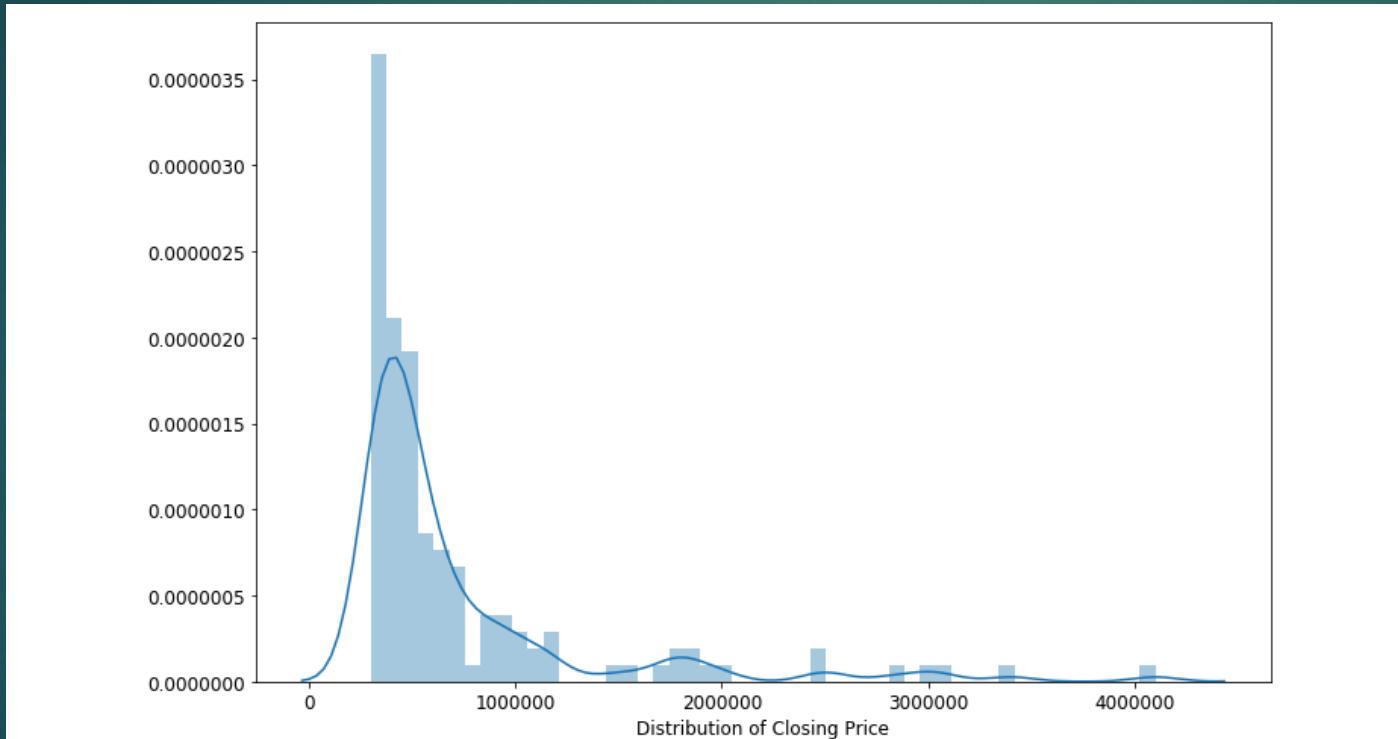
# STEP 3: UNDERSTANDING THE DATA

- ▶ Features in Data Frame:

Features	Type
Id, bedrooms_total, baths_total, acres, sq_ft_tot_fn, tax_gross_amount, assessment_value_town, garage_capacity, year_built, total_stories, water_frontage_length, rooms_total,	Numeric
Address, city ,garage_type , surveyed , seasonal , water_body_type , short_sale , flood_zone , easements , current_use , covenants , common_land_acres, basement_access_type,basement,price_closed	String

# STEP 3: UNDERSTANDING THE DATA

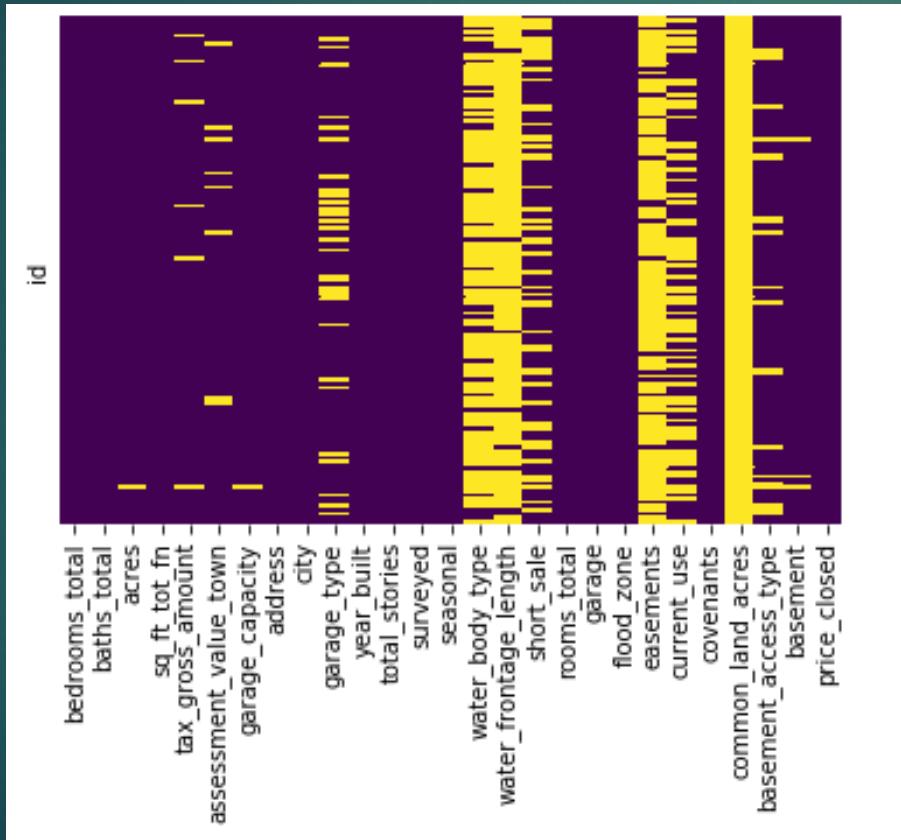
## ► Distribution of Closing Price



Almost normal distribution  
with a long tail

# STEP 3: UNDERSTANDING THE DATA

- ▶ Missing values: Using Seaborn heatmap



Most prominent:

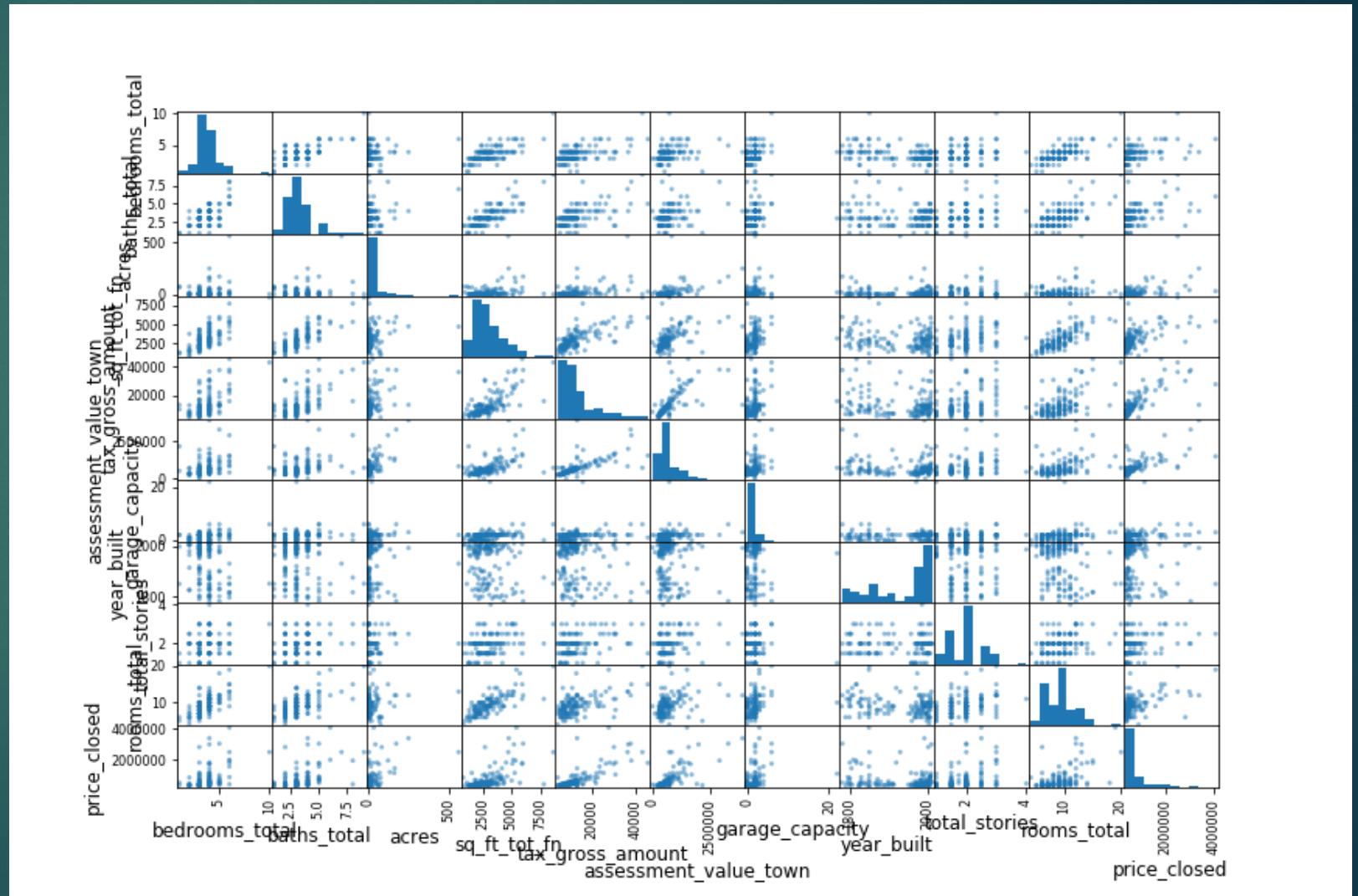
- common\_land\_acres
- water\_frontage\_length
- easements
- water\_body\_type

# STEP 4: PREPARING THE DATA

- ▶ Drop features that are missing from more than 80% of the data set
  - ▶ df =  
df.drop(['common\_land\_acres','water\_frontage\_length','easements','water\_body\_type'],axis=1)
- ▶ Filled the rest with mean (numerical)
  - ▶ for col in  
['acres','garage\_capacity','tax\_gross\_amount','assessment\_value\_town']: df[col] = df[col].fillna(df[col].mean())

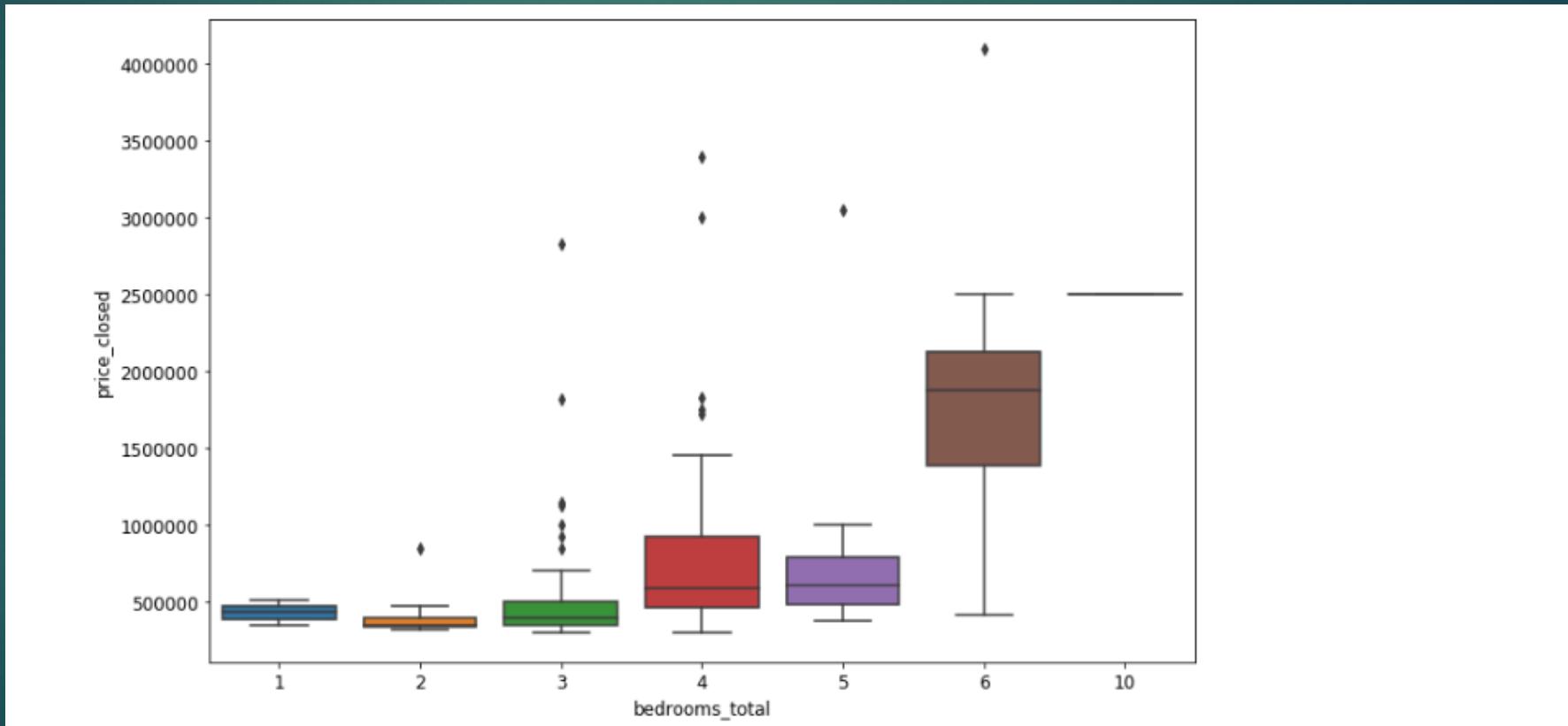
# STEP 5: VISUALIZATION

- ▶ Explore relationships:



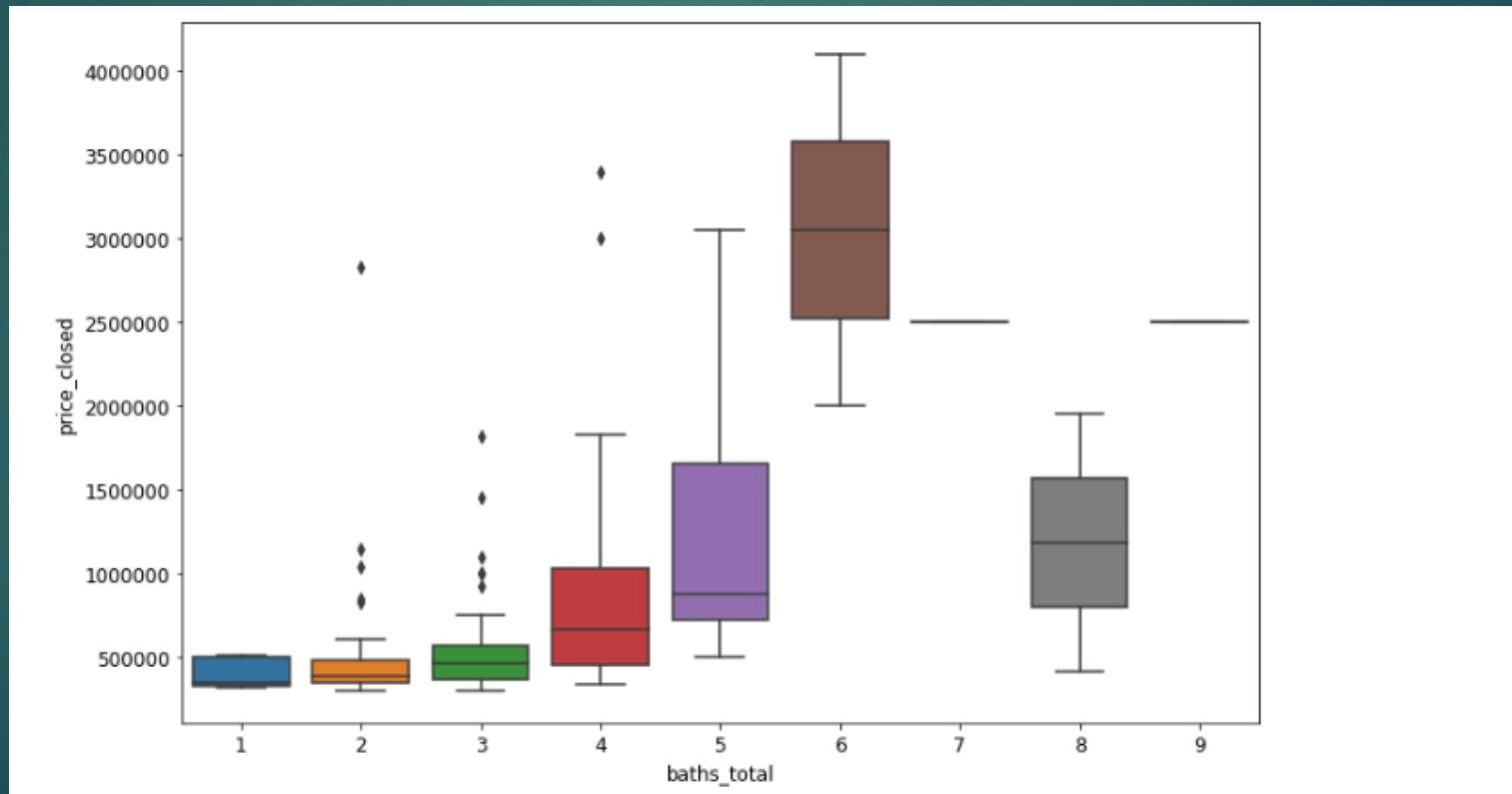
# STEP 5: VISUALIZATION

- ▶ Narrow down to features with regression:



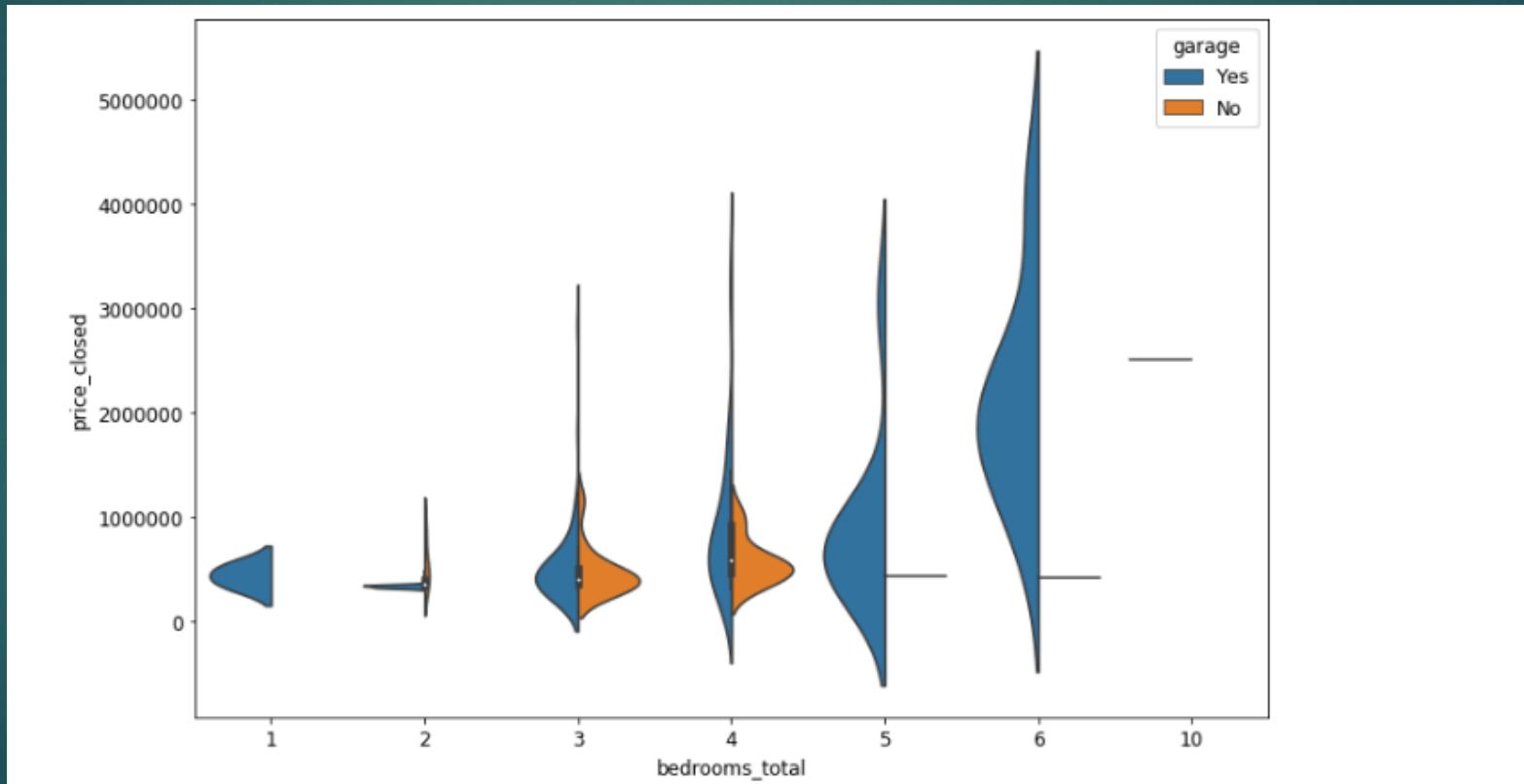
# STEP 5: VISUALIZATION

- ▶ Narrow down to features with regression:



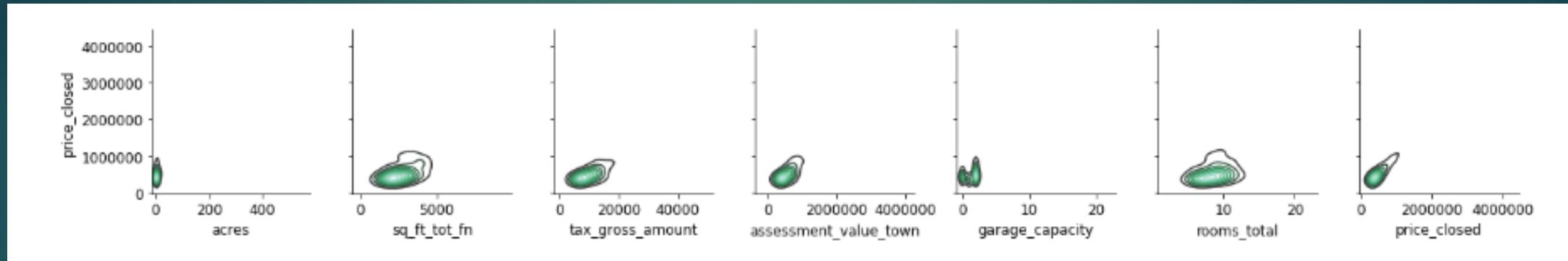
# STEP 5: VISUALIZATION

- ▶ Narrow down to features with regression:



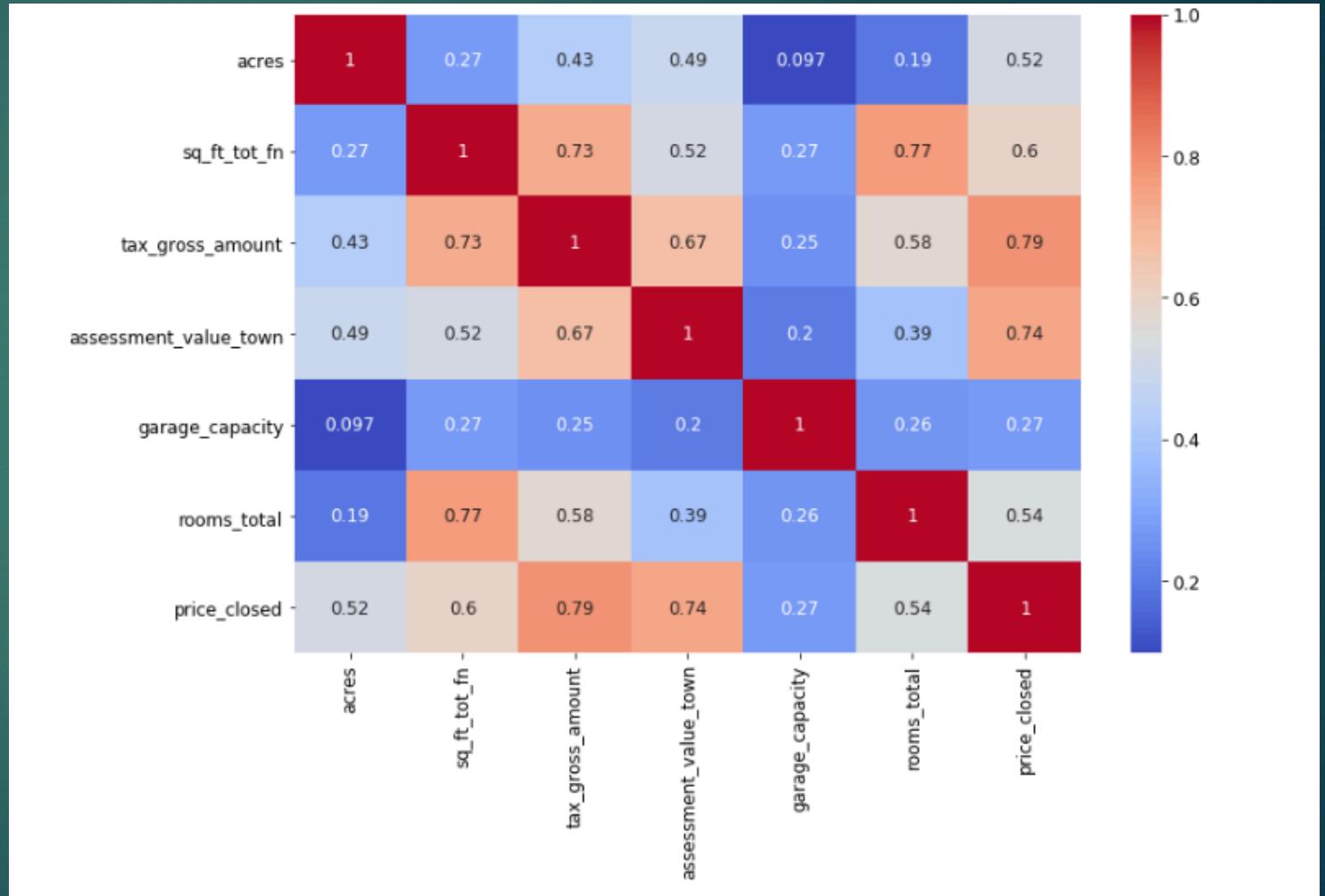
# STEP 5: VISUALIZATION

- ▶ Narrow down to features with no clear regression:



# STEP 5: VISUALIZATION

► Find out correlation:



# Next step:

- ▶ Feature engineering
- ▶ Try creating cluster as a feature
- ▶ Modeling techniques: Regression
- ▶ Experiment with dimensionality reduction and deep learning  
(TensorFlow)