Predicting Home Values : Using a Linear Regression Modeling Approach

Presenter: Ganesh Bala

AGENDA

OVERVIEW

APPROACH

DATA COLLECTION

MODEL ASSUMPTIONS

MODELING

VALIDATIONS

LEARNINGS & NEXT ACTIONS

Q&A

NEIGHBORHOODS

1. **Potomac**, Maryland
(Median Zestimate® is $915,724)

2. **Rockville**, Maryland
(Median Zestimate® is $576,118)

3. **North Potomac**, Maryland
(Median Zestimate® is 564,402)

4. **Gaithersburg**, Maryland
(Median Zestimate® is $ 555,016)

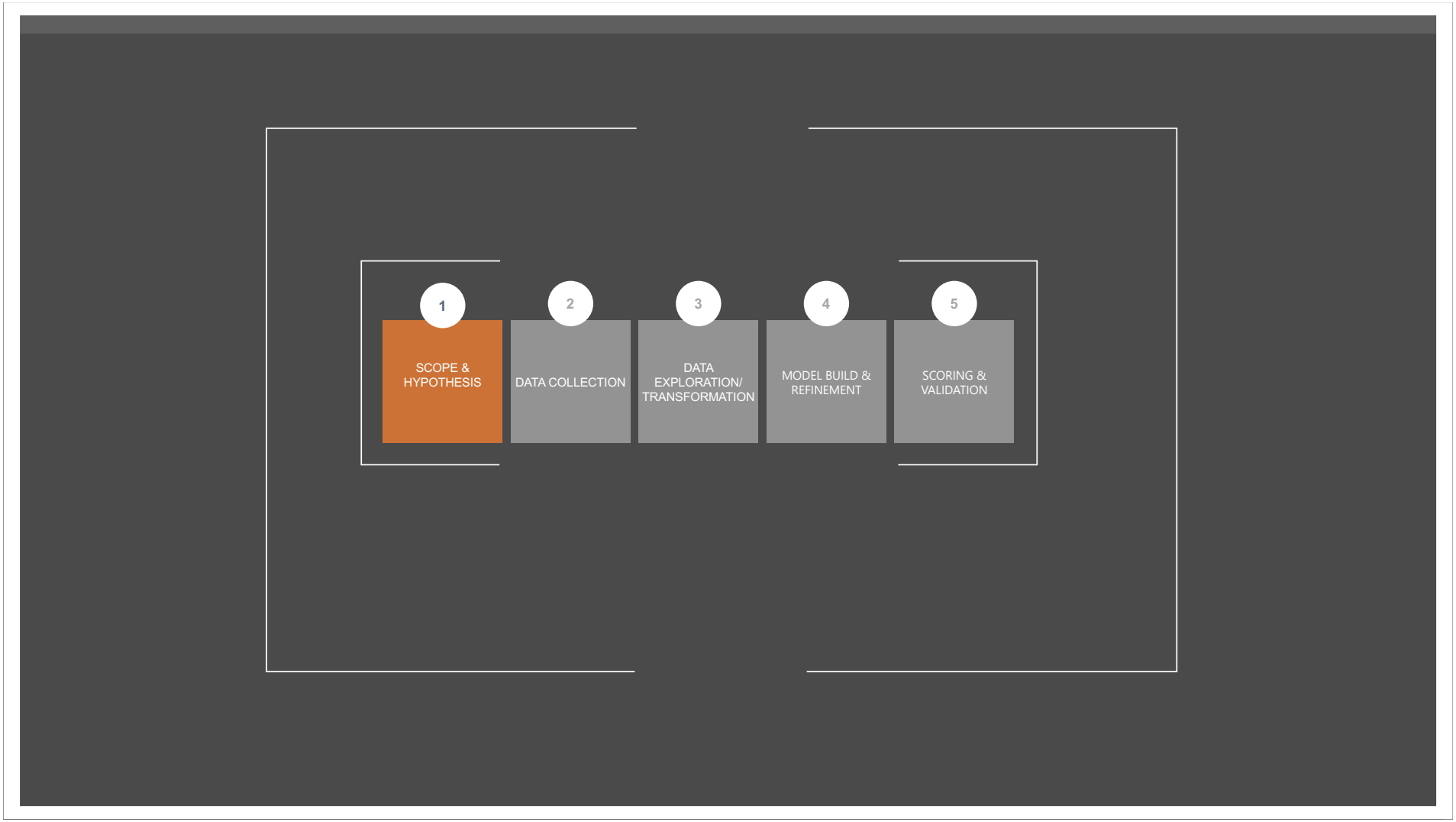
HIGH AFFLUENT NEIGHBORHOODS

CLOSE TO D.C. & MAJOR HUBS

HIGHLY RANKED SCHOOL DISTRICT

INDEX LOW IN UNEMPLOYMENT

INDEX LOW IN CRIME


1
SCOPE & HYPOTHESIS
2
DATA COLLECTION
3
DATA EXPLORATION/ TRANSFORMATION
4
MODEL BUILD & REFINEMENT
5
SCORING & VALIDATION

SCOPE & HYPOTHESIS

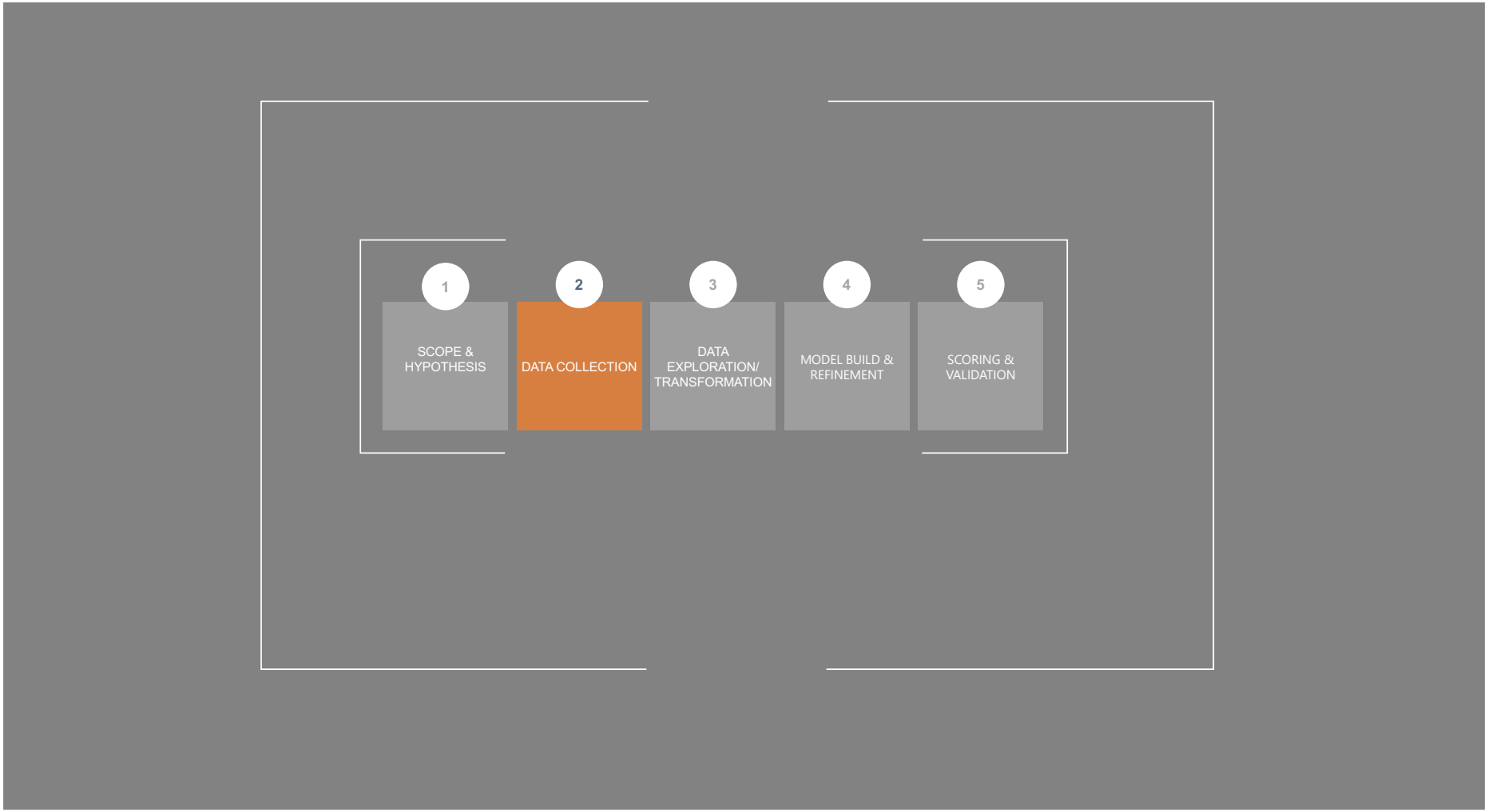
Leverage data on sold homes and zip-level geo coordinate data to predict home value

- Supervised learning – Linear Regression

- Model the linear relationship between between the predictor variables and the target (continuous) variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \varepsilon$$

- Ensure that data satisfies the 4 key assumptions of the Ordinary Least Squares Regression (OLSR) model
  i. Linear functional form
  ii. Residual errors are independent and identically distributed random variables
  iii. Residual errors are normally distributed
  iv. Homoscedasticity – constant variance of residual errors


1
SCOPE & HYPOTHESIS
2
DATA COLLECTION
3
DATA EXPLORATION/ TRANSFORMATION
4
MODEL BUILD & REFINEMENT
5
SCORING & VALIDATION

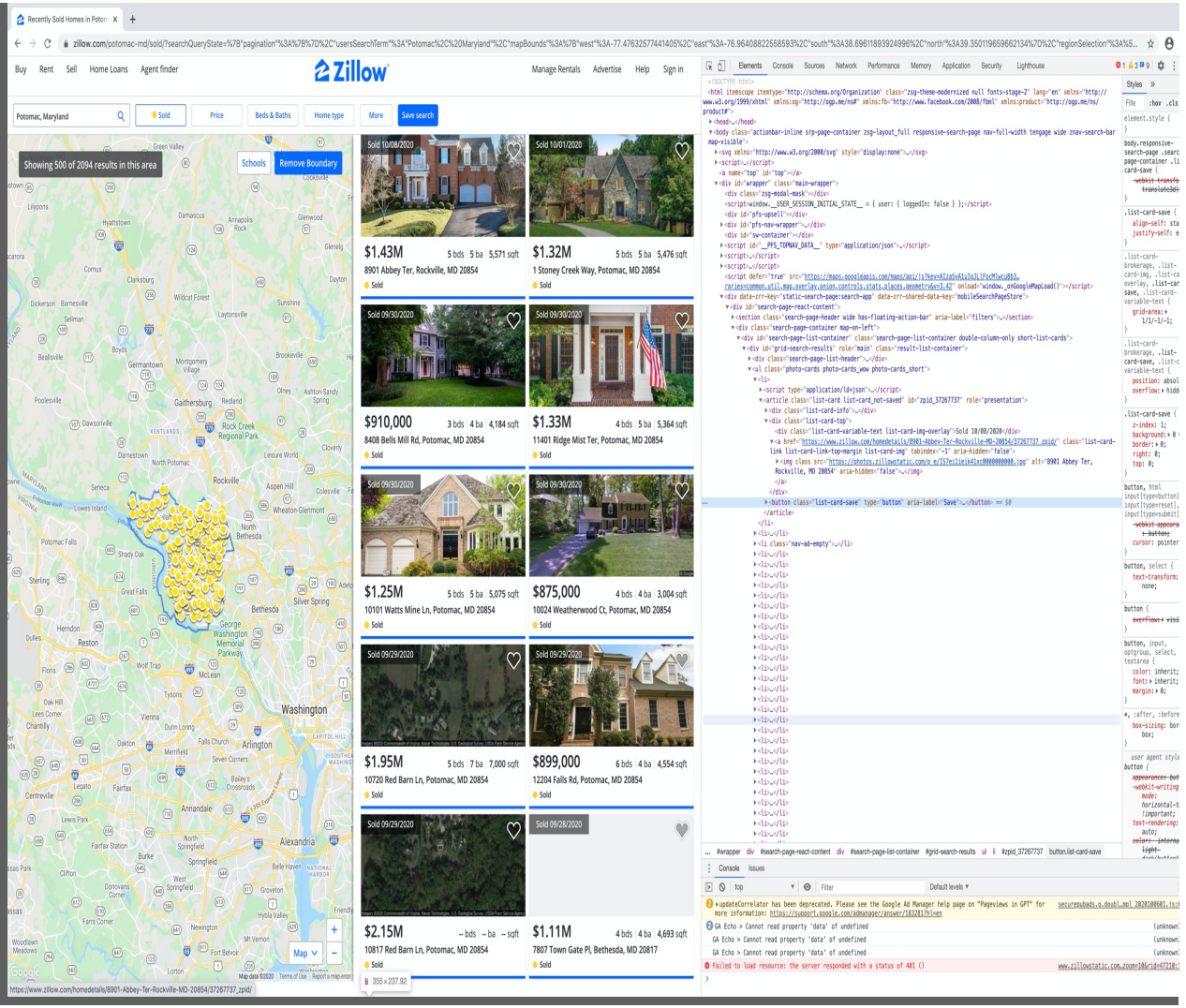

IT’S WAS A JOURNEY..

WEBSCRAPING DATA

Leveraged BeautifulSoup and Selenium to scrape homes sold data off Zillow's website

- **Had to wrestle the complexity (amidst nascent approaches to scraping data!)**

- **On average, ~ 400 records of data on homes sold was extracted for each of the 4 cities of interest**



DATA COLLECTION

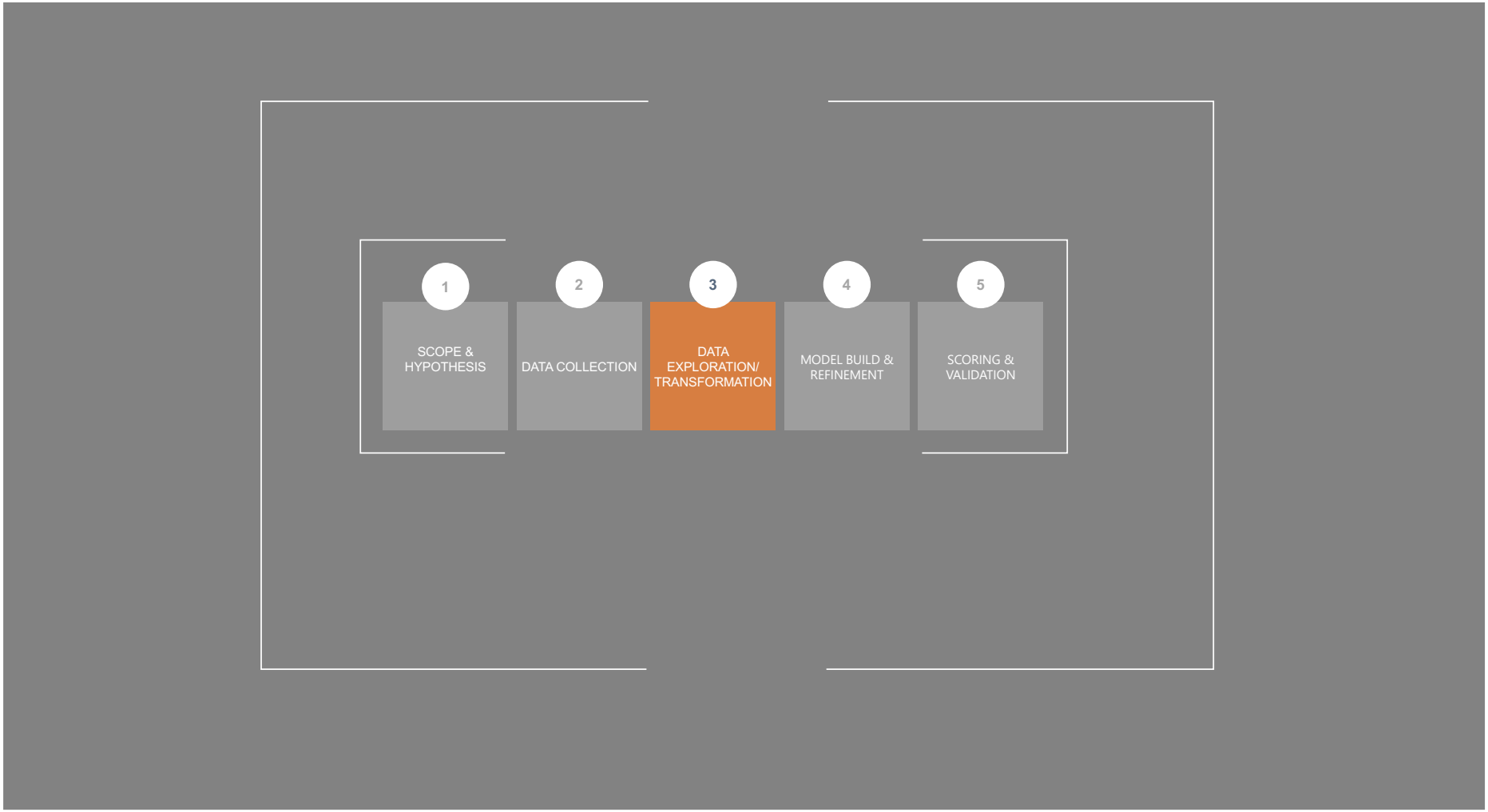
Collected ~1500 rows of data on home sales across 4 cities in the D.C. metro area

- Attributes included for modeling were
    1. Primarily sourced from Zillow.com
    2. Additional publicly available data on geo-coordinates was also augmented

- Besides **Home Sale Price**, key features that were available as model inputs included:

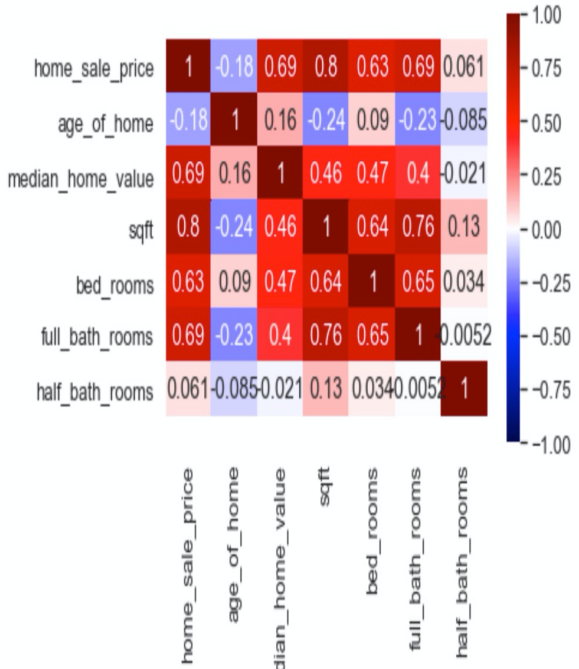
- Square Footage of the house
- Lot Area (in acres)
- Home Type – SFH or Townhouse or Condo
- Median Value of Homes Sold in the neighborhood (by home type)
- Year Built
- New Construction (Y/N)
- Number of bedrooms
- Number of Full bathrooms
- Number of Half bathrooms

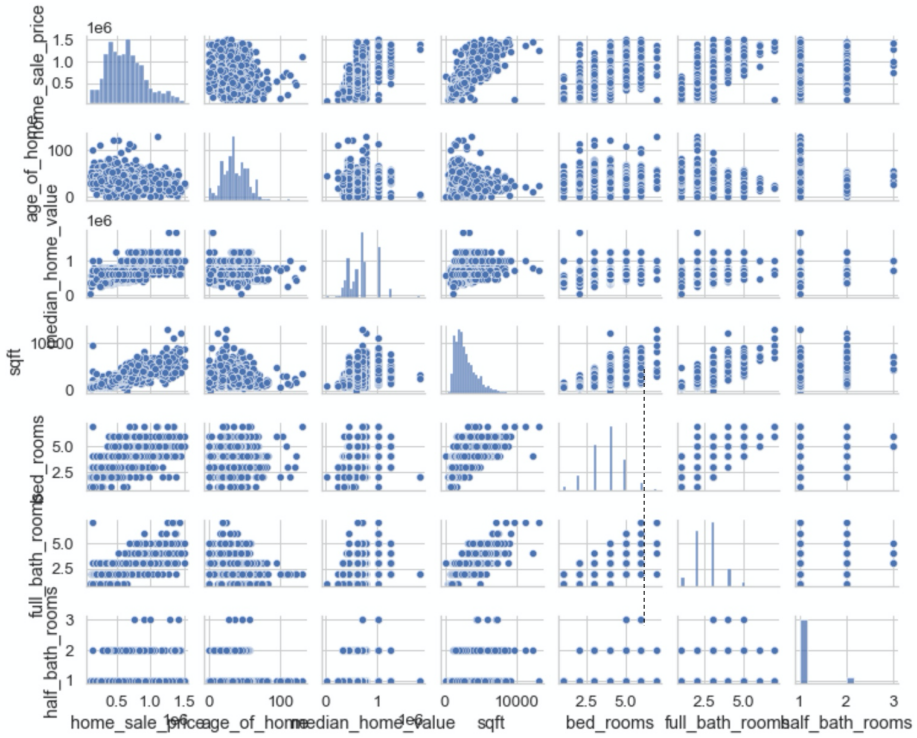
In addition, property related locational attributes like street address, city, state, zip code, latitude and longitude were also gathered.


1
SCOPE & HYPOTHESIS
2
DATA COLLECTION
3
DATA EXPLORATION/ TRANSFORMATION
4
MODEL BUILD & REFINEMENT
5
SCORING & VALIDATION

DATA EXPLORATION & TRANSFORMATION

Correlations



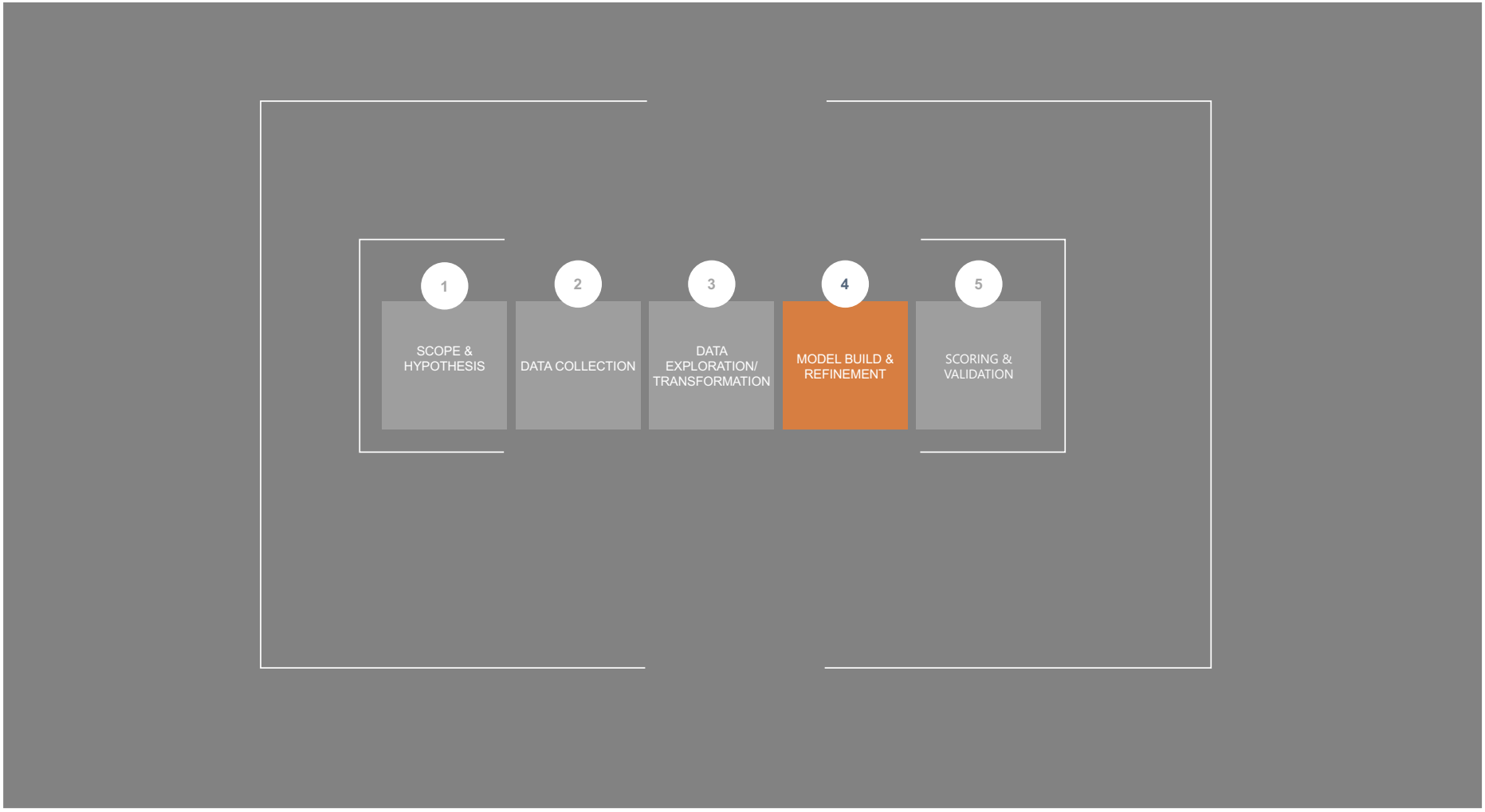


DATA EXPLORATION & TRANSFORMATION

- Removed outliers in the Home Sales Price data
  - some were home sold that were re-listed for Rent (bad picks from scraping)
  - a few others were extreme high values and did not have any 'closest neighbors' for comparison (ex: a home sold for $5.6M)
  - as a result, decided to model based on data where the home value was between $100K-$1.5M

- Further created derived variables by transforming categorical features such as Home Type – each category was its own feature with a binary-response value

- Polynomial transformations were done after establishing the baseline model


1
SCOPE & HYPOTHESIS
2
DATA COLLECTION
3
DATA EXPLORATION/ TRANSFORMATION
4
MODEL BUILD & REFINEMENT
5
SCORING & VALIDATION

MODEL BUILD & REFINEMENT

Baseline Model:

- An initial model was run with all available features on the full sample (prior to test, train and validation splits)
- Opportunities for refinement were identified
- Polynomial transformations were considered as a part of the next iteration
- Also, decided to look closely into potential issues with multicollinearity

Full Sample

OLS Regression Results

Dep. Variable:	home_sale_price	**R-squared:**	0.800
Model:	OLS	**Adj. R-squared:**	0.798
Method:	Least Squares	**F-statistic:**	550.6
Date:	Fri, 09 Oct 2020	**Prob (F-statistic):**	0.00
Time:	07:15:17	**Log-Likelihood:**	-18308.
No. Observations:	1389	**AIC:**	3.664e+04
Df Residuals:	1378	**BIC:**	3.670e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.177e+04	2.09e+04	-0.563	0.574	-5.28e+04	2.93e+04
age_of_home	-2153.5050	244.139	-8.821	0.000	-2632.429	-1674.581
median_home_value	0.5549	0.019	29.401	0.000	0.518	0.592
sqft	80.4087	3.797	21.175	0.000	72.959	87.858
new_const	1.344e+04	2e+04	0.672	0.502	-2.58e+04	5.27e+04
bed_rooms	3.025e+04	5334.080	5.671	0.000	1.98e+04	4.07e+04
full_bath_rooms	2.283e+04	6433.963	3.548	0.000	1.02e+04	3.54e+04
half_bath_rooms	-1.617e+04	1.2e+04	-1.345	0.179	-3.97e+04	7404.426
single_family	-4.358e+04	1.46e+04	-2.995	0.003	-7.21e+04	-1.5e+04
townhouse	1.311e+04	1.36e+04	0.961	0.337	-1.37e+04	3.99e+04
condo	-4.903e+04	1.28e+04	-3.836	0.000	-7.41e+04	-2.4e+04
apartment	6.774e+04	3.96e+04	1.710	0.087	-9956.529	1.45e+05

Omnibus:	272.863	**Durbin-Watson:**	1.674
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5694.456
Skew:	-0.303	**Prob(JB):**	0.00
Kurtosis:	12.901	**Cond. No.**	9.90e+21

MODEL BUILD & REFINEMENT

Baseline Model – after finalizing features

- The baseline model was further iterated upon to include polynomial terms
- None of the squared or interaction terms seemed to help with significantly improving the model
- The best-fit model was the baseline model pruned for features that were highly collinear or added very little weight to the model

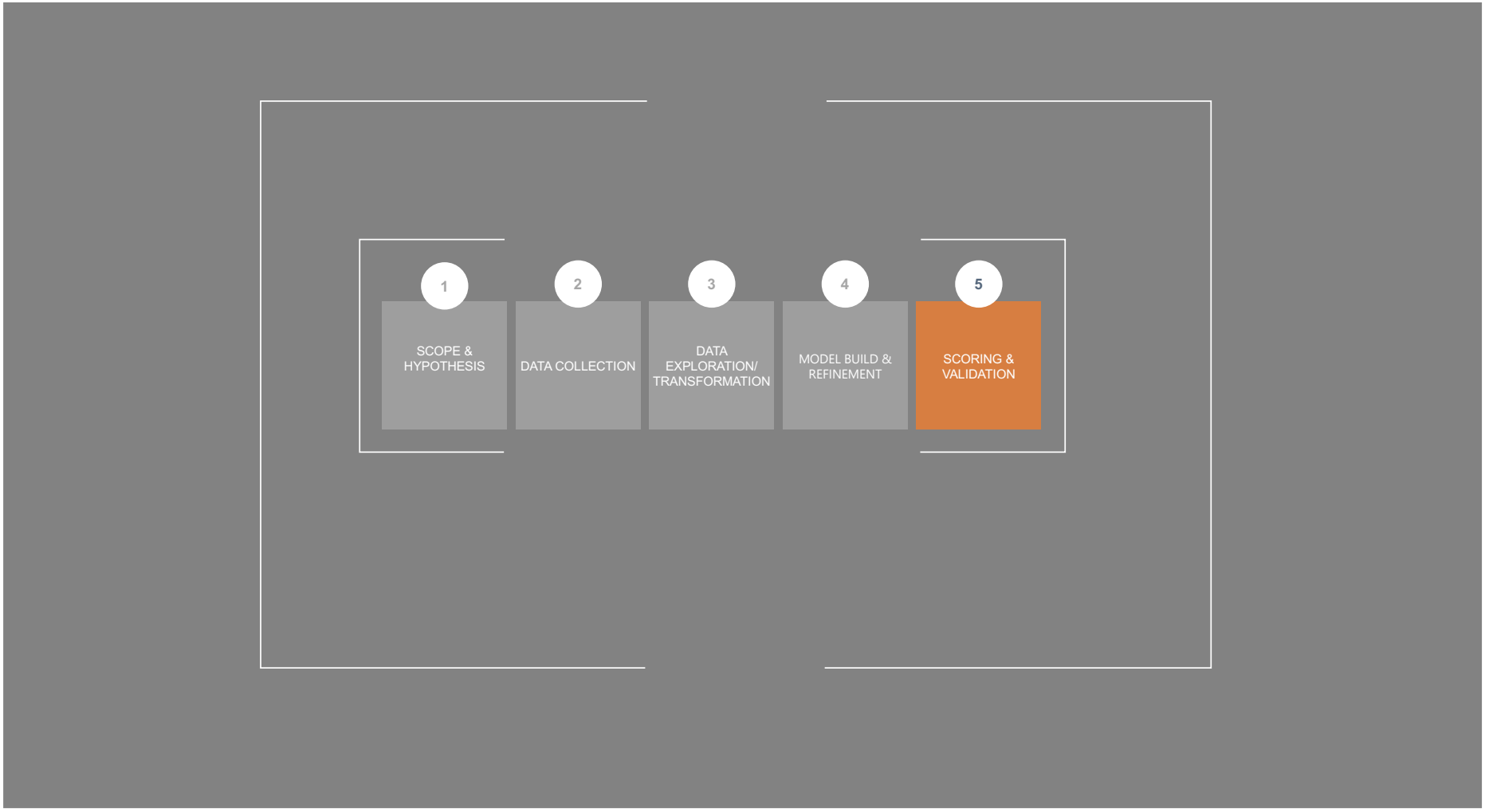
Training Data

OLS Regression Results

Dep. Variable:	home_sale_price	**R-squared:**	0.798
Model:	OLS	**Adj. R-squared:**	0.797
Method:	Least Squares	**F-statistic:**	653.7
Date:	Fri, 09 Oct 2020	**Prob (F-statistic):**	2.23e-284
Time:	06:58:06	**Log-Likelihood:**	-11017.
No. Observations:	833	**AIC:**	2.205e+04
Df Residuals:	827	**BIC:**	2.207e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.799e+04	2.16e+04	-1.756	0.079	-8.05e+04	4475.289
age_of_home	-2527.9101	325.614	-7.764	0.000	-3167.037	-1888.783
median_home_value	0.5487	0.024	22.998	0.000	0.502	0.596
sqft	80.1213	4.938	16.224	0.000	70.428	89.815
bed_rooms	2.757e+04	6363.215	4.333	0.000	1.51e+04	4.01e+04
full_bath_rooms	2.224e+04	8697.534	2.557	0.011	5163.855	3.93e+04

Omnibus:	242.973	**Durbin-Watson:**	1.969
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5248.330
Skew:	-0.776	**Prob(JB):**	0.00
Kurtosis:	15.198	**Cond. No.**	3.50e+06

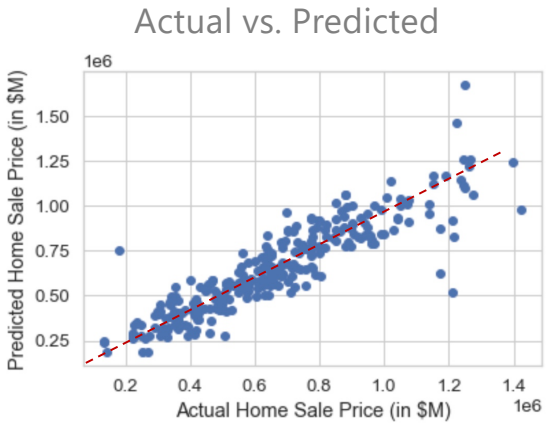

1
SCOPE & HYPOTHESIS
2
DATA COLLECTION
3
DATA EXPLORATION/ TRANSFORMATION
4
MODEL BUILD & REFINEMENT
5
SCORING & VALIDATION

MODEL BUILD & REFINEMENT

Model Cross-Fold Validation & Comparisons to a Hold-out*

Metric	Value
MAE	83604
MSE	14666568925
RMSE	121106

Test Data

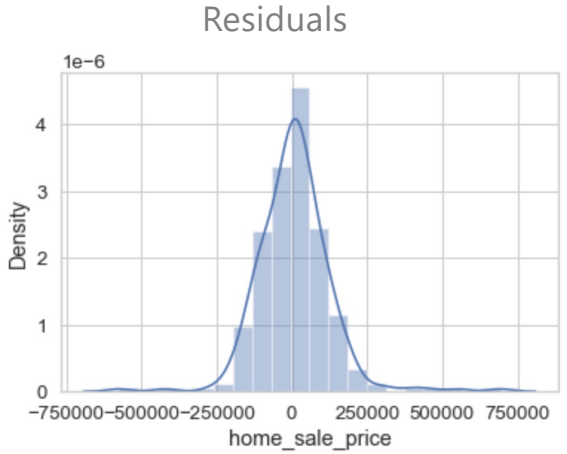


A 5-fold approach to cross-validation – very useful due to relatively low sample size of analysis data

Results from 5-Fold CV Tests

Results show the relative accuracy of the training model when compared to the cross-validation results using a 5-fold method.

Metric	Mean	Range
Test r2 :	0.79	(0.6832, 0.8468)
Train/Test r2 Ratio :	1.018	(0.9257, 1.205)



* ***Hold-out Sample*** *: A 20% random sample set aside prior to start of model training; only intended as a final validation step*

MODEL BUILD & REFINEMENT

Model Cross-Fold Validation (CV) & Comparisons to a Hold-out

Interpreting the coefficients:

•Holding all other features fixed, a 1 unit increase in **Age of Home** is associated with a **decrease of about $2,728** in home sale price.

•Holding all other features fixed, a 1 unit increase in **Median Home Value** is associated with an **increase of $0.57** in home sale price.

•Holding all other features fixed, a 1 unit increase in **Avg. Square Footage of the Home** is associated with an **increase of $77.1** in home sale price.

•Holding all other features fixed, a 1 unit increase in **Total Bedrooms** is associated with an **increase of $28,495** in home sale price.

•Holding all other features fixed, a 1 unit increase in **Total Full Baths** is associated with an **increase of $15,673** in home sale price.

Validation Data

Best Model Features

	Feature	Coefficient SIgn	Coefficient Value
4	Age of Home (in years)	-	2726.76
2	Median Home Value (by Zip Code & Home Type)	+	0.57
1	Home Square Footage	+	77.10
2	Total Bedrooms	+	28495.11
3	Total Full Baths	+	15673.35

* ***Hold-out Sample*** *: A 20% random sample set aside prior to start of model training; only intended as a final validation step*

LEARNINGS & NEXT ACTIONS

- Sample size to be increased so that we can have better confidence from the K-Fold CV approaches
  - There is a 'treasure trove' of data waiting to be scraped on Zillow.com – need smarter and efficient ways to gain competency with BeautifulSoup, Selenium and other tools

- Opportunity exists to improve the model with data overlays from other public available data sources (via APIs)

- Over-fitting is not a significant issue yet due to smaller feature-set. Regularization will become a necessity as the feature-set increases

QUESTIONS?

Thank You!