DSI Project 2

Linear Regression on Ames Housing Dataset to Predict Sale Price

Krisgun & Scent

Team Member

Krisgun Chirasanta (Kris)

Senior Data Scientist

Kanitin Sukdit (Scent)

Junior Data Scientist

Agenda

Ideal Workflow of Data Science Project

1

Project Overview

Brief understanding of Ames IOWA and the dataset

2

Data Cleaning

Fill missing values

3

Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

4

Model Iteration

Iterate combination of cleaned features vs model benchmark

5

Model Production

Select model for production

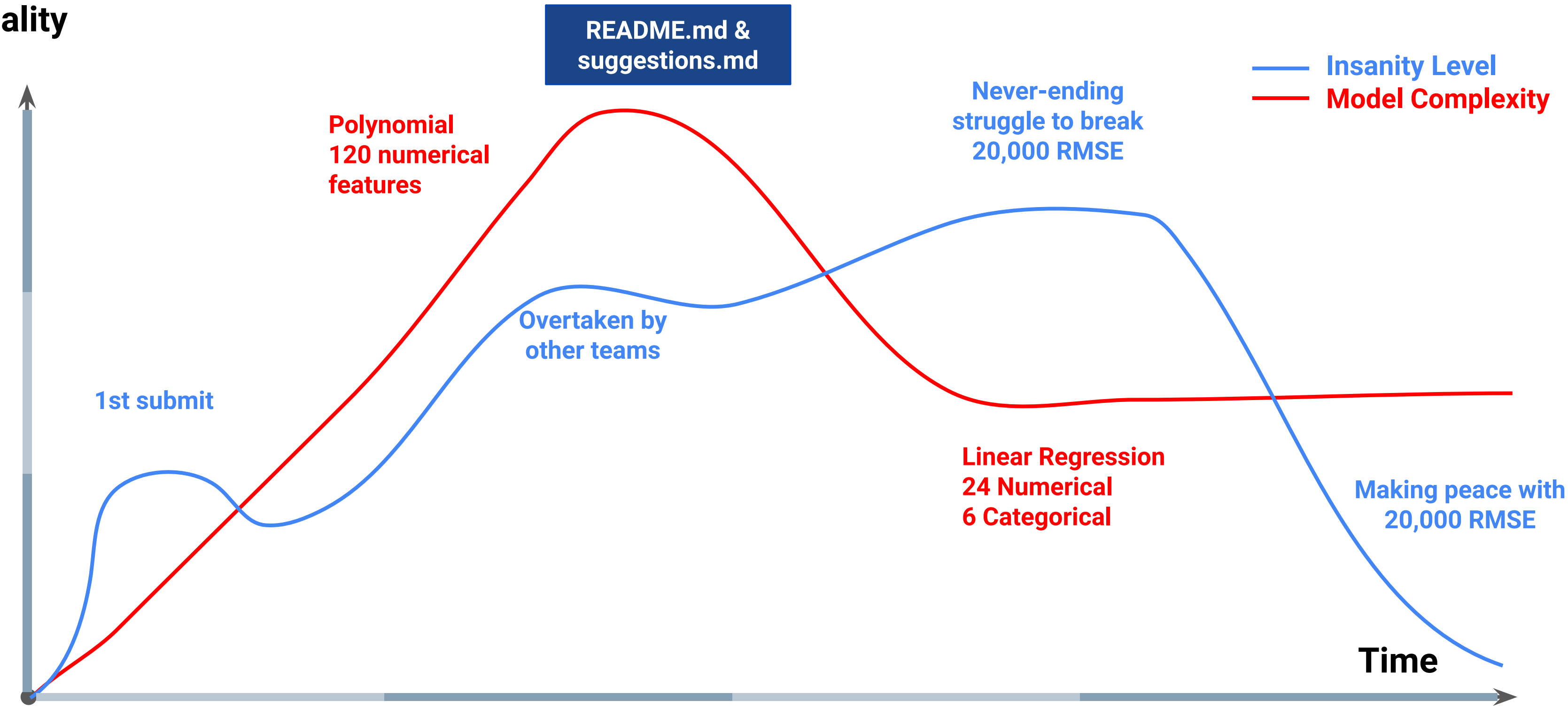
6

Further Investigation

Areas to explore in the future

Emotional Rollercoaster of Kaggle Competition

In reality



README.md & suggestions.md

1st submit

Polynomial
120 numerical
features

Overtaken by
other teams

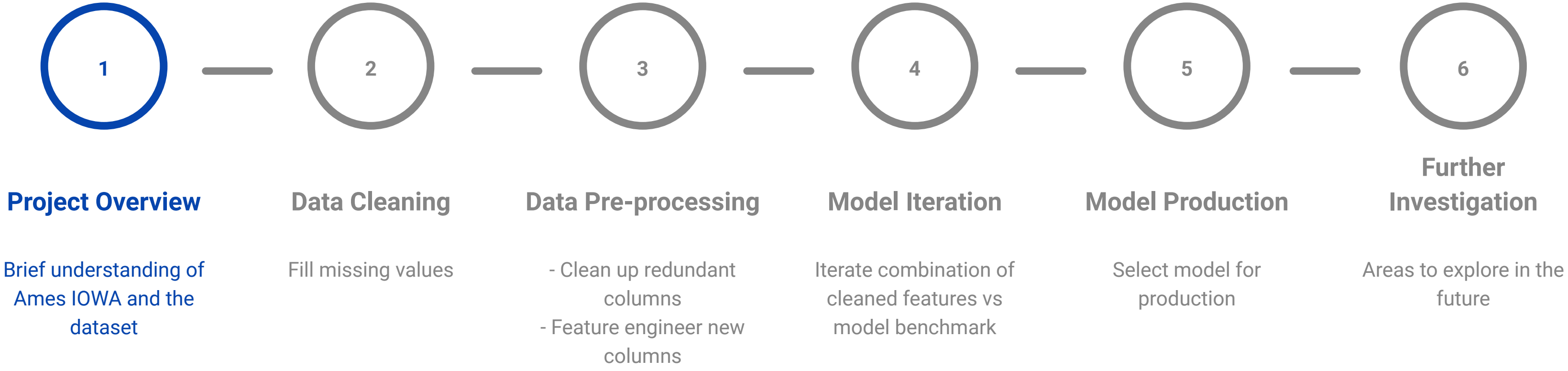
Never-ending
struggle to break
20,000 RMSE

Linear Regression
24 Numerical
6 Categorical

Making peace with
20,000 RMSE

Agenda

Ideal Workflow of Data Science Project



1

Project Overview

Brief understanding of Ames IOWA and the dataset

2

Data Cleaning

Fill missing values

3

Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

4

Model Iteration

Iterate combination of cleaned features vs model benchmark

5

Model Production

Select model for production

6

Further Investigation

Areas to explore in the future

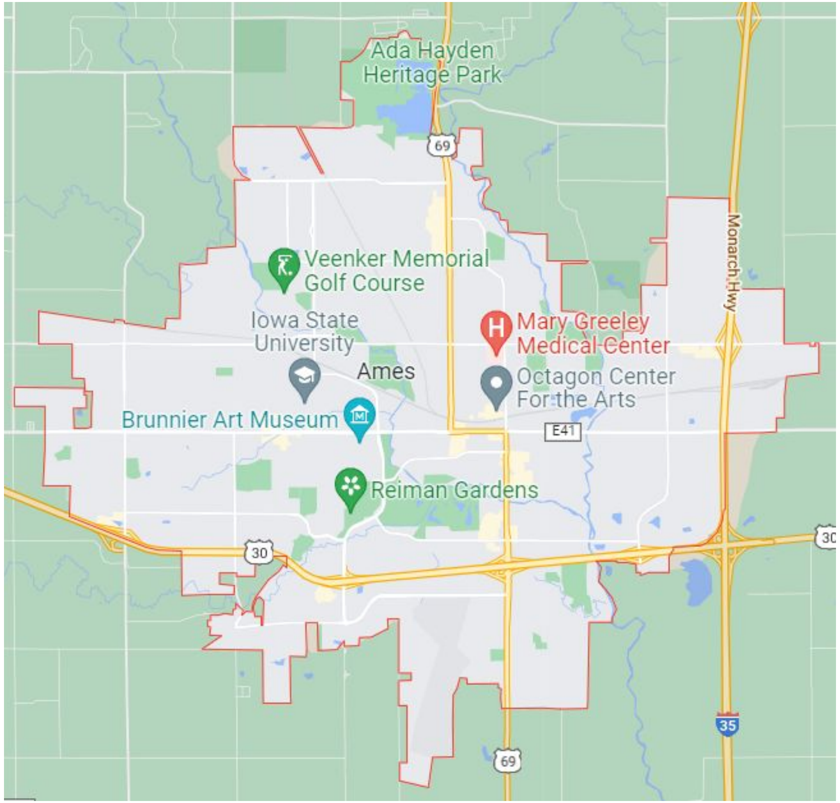
Project Overview - Ames, Iowa

Ames, City

- Country : United States
- State: Iowa
- County : Story
- Area: ~ 143.75 km^2

Housing style

- House (1-3 Floors)
- Townhouse
- Condo



Source : Ames, Iowa - Wikipedia

Project Overview - Dataset

Datasets

Shape : 2197 rows, 82 columns

Numerical : 39

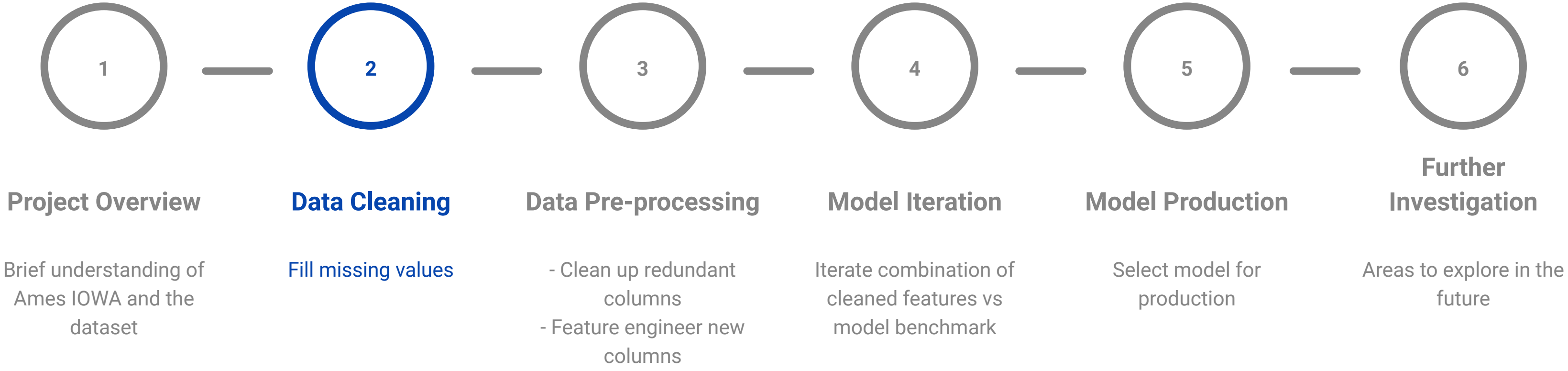
Categorical : 43

Column Groups

- Lot
- Quality
- Masonry
- Garage
- Basement
- Square feet
- Year
- Bathroom
- Rooms
- Porch
- Fireplace
- Wood Deck

Agenda

Ideal Workflow of Data Science Project



1

2

3

4

5

6

Project Overview

Brief understanding of Ames IOWA and the dataset

Data Cleaning

Fill missing values

Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

Model Iteration

Iterate combination of cleaned features vs model benchmark

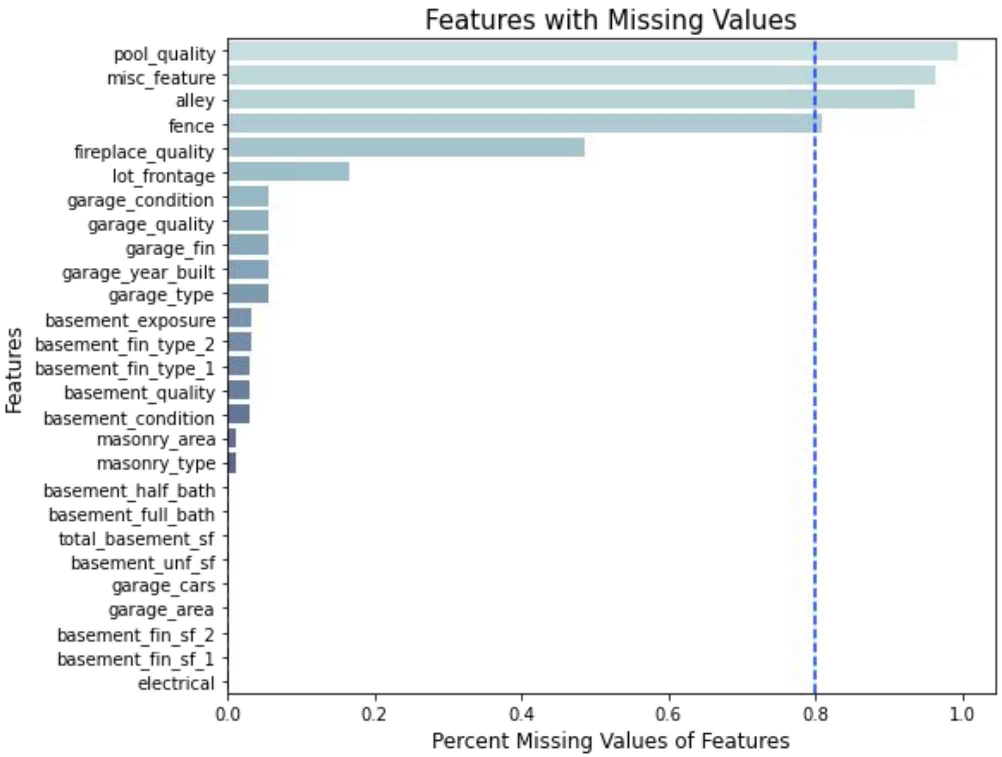
Model Production

Select model for production

Further Investigation

Areas to explore in the future

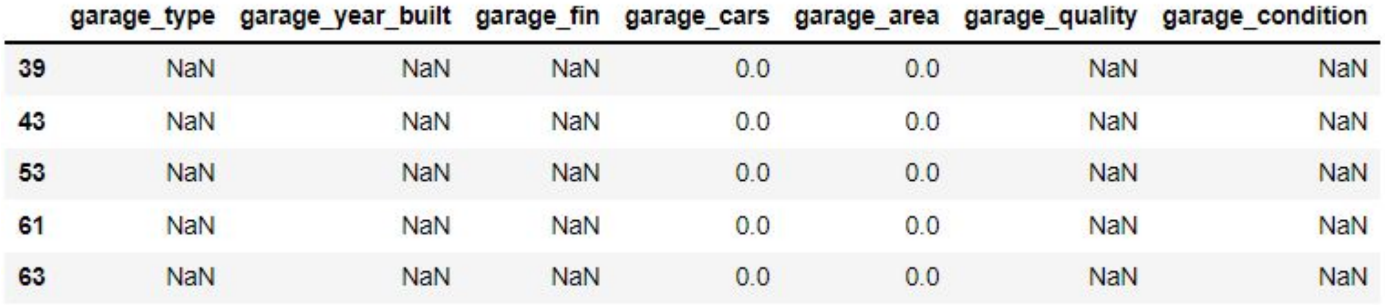
Data Cleaning - Missing Values - “Drop”



- **Drop** (over 80% missing)
  - Pool Quality
  - Miscellaneous Features
  - Alley
  - Fence

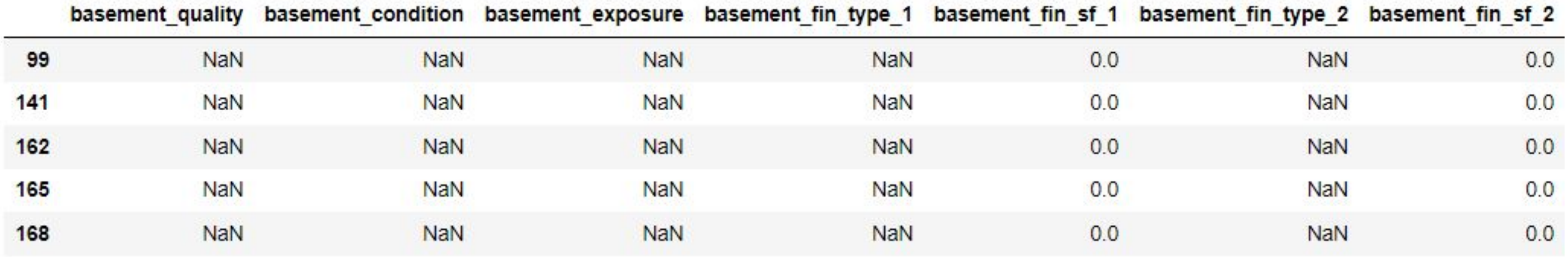
Data Cleaning - Missing Values - “None”

- Garage group



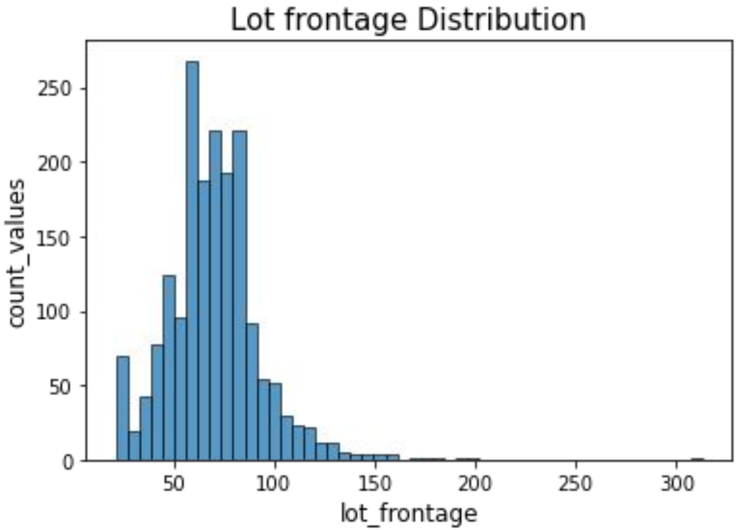
	garage_type	garage_year_built	garage_fin	garage_cars	garage_area	garage_quality	garage_condition
39	NaN	NaN	NaN	0.0	0.0	NaN	NaN
43	NaN	NaN	NaN	0.0	0.0	NaN	NaN
53	NaN	NaN	NaN	0.0	0.0	NaN	NaN
61	NaN	NaN	NaN	0.0	0.0	NaN	NaN
63	NaN	NaN	NaN	0.0	0.0	NaN	NaN

- Basement group

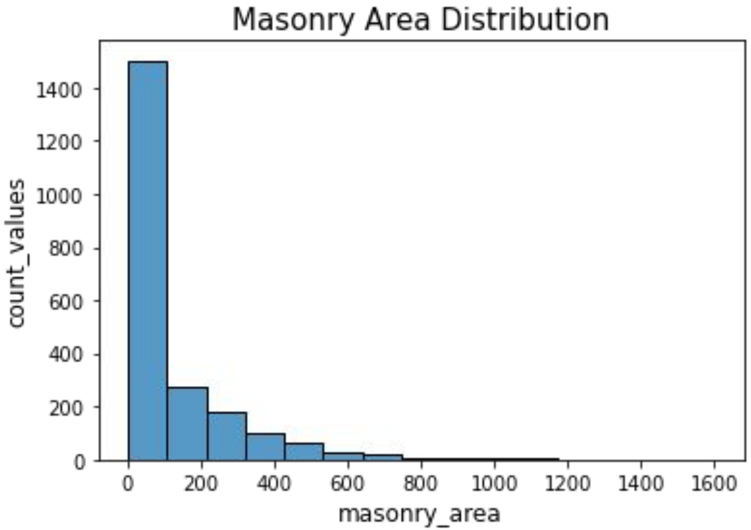


	basement_quality	basement_condition	basement_exposure	basement_fin_type_1	basement_fin_sf_1	basement_fin_type_2	basement_fin_sf_2
99	NaN	NaN	NaN	NaN	0.0	NaN	0.0
141	NaN	NaN	NaN	NaN	0.0	NaN	0.0
162	NaN	NaN	NaN	NaN	0.0	NaN	0.0
165	NaN	NaN	NaN	NaN	0.0	NaN	0.0
168	NaN	NaN	NaN	NaN	0.0	NaN	0.0

Data Cleaning - Missing Values - “Stats/0”



- **Impute**: Mode
- **Reasoning:** Mode of similar groupby property (Lot Area & Lot Shape)



- **Impute**: 0
- **Reasoning:** small percentage

Agenda

Ideal Workflow of Data Science Project

1. Project Overview

Brief understanding of Ames IOWA and the dataset

2. Data Cleaning

Fill missing values

3. Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

4. Model Iteration

Iterate combination of cleaned features vs model benchmark

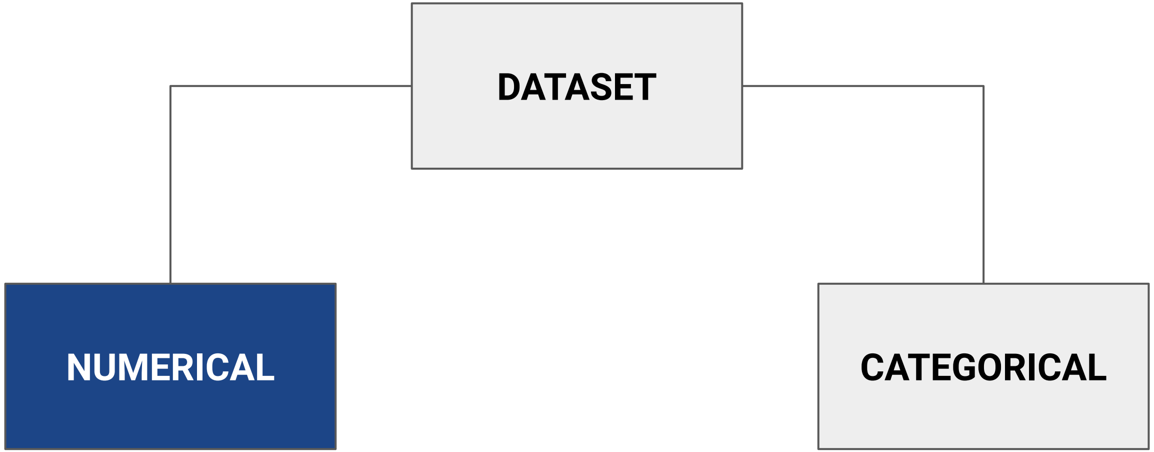
5. Model Production

Select model for production

6. Further Investigation

Areas to explore in the future

Data Cleaning - Pre-processing



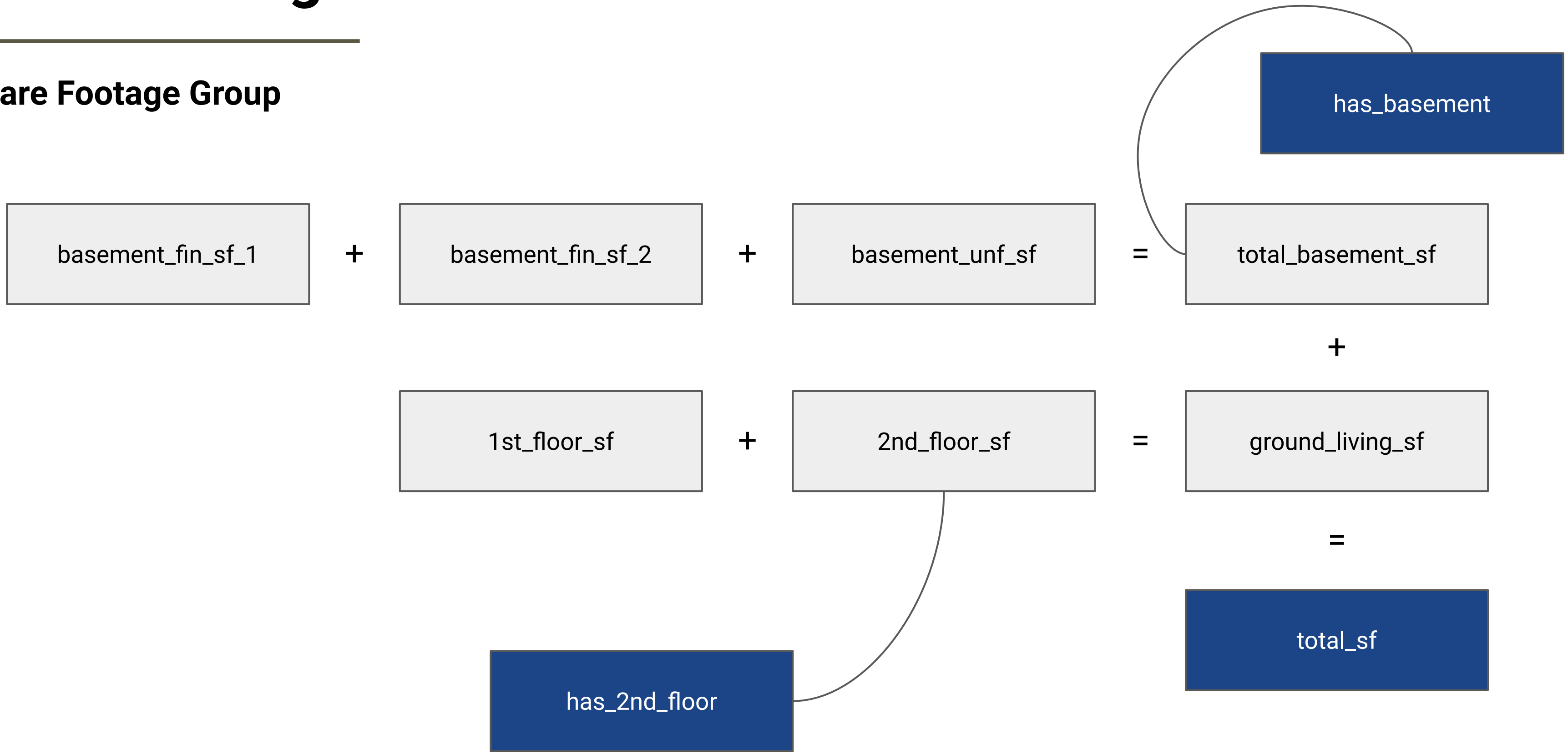
DATASET

NUMERICAL

CATEGORICAL

Data Cleaning - Redundant Columns

Square Footage Group



has_basement

basement_fin_sf_1 + basement_fin_sf_2 + basement_unf_sf = total_basement_sf

+

1st_floor_sf + 2nd_floor_sf = ground_living_sf

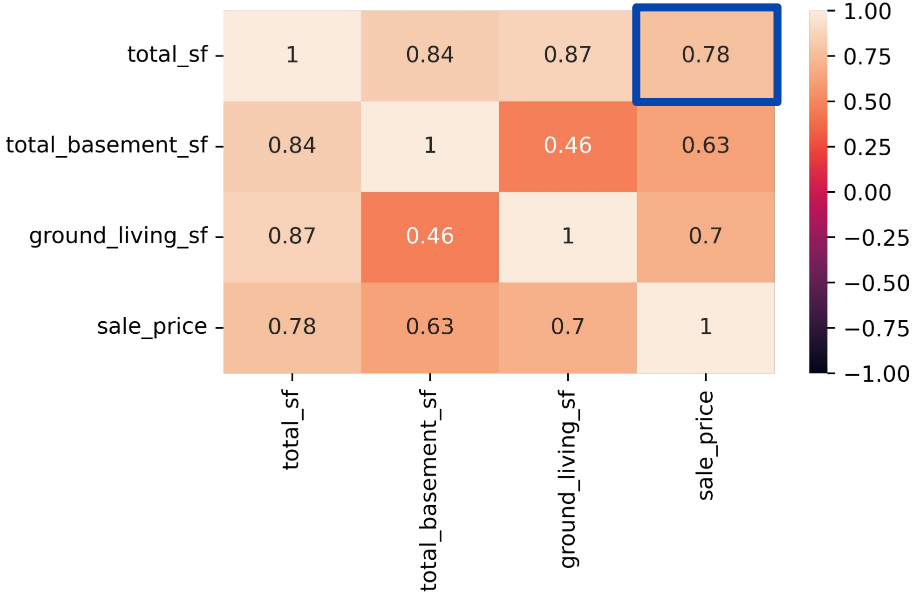
=

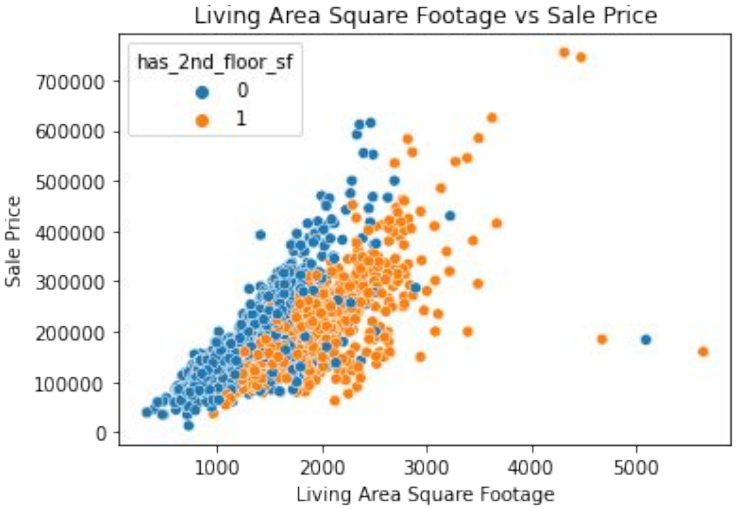
total_sf

has_2nd_floor

Data Cleaning - Redundant Columns

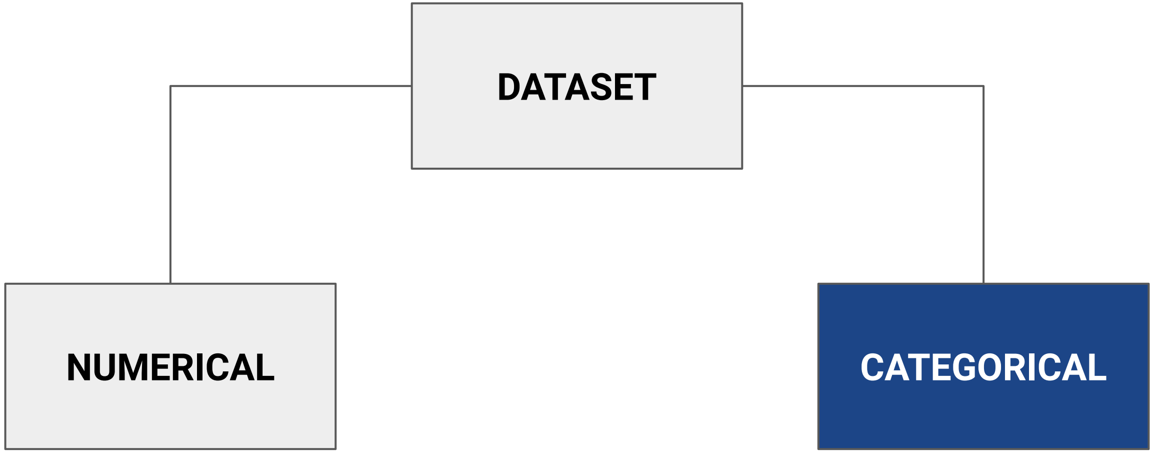
Square Footage Group





- Total square footage is **more correlated to sale price** compared to original two
- Keep redundant column as **boolean feature column**

Data Cleaning - Grouping of category



Data Cleaning - Cleaning of category

Quality Group

external_quality

basement_quality

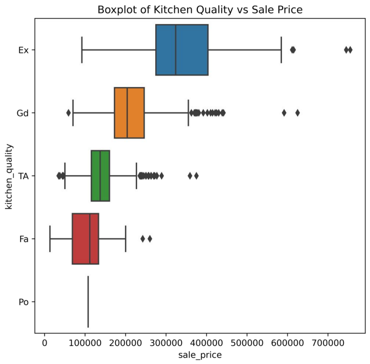
fireplace_quality

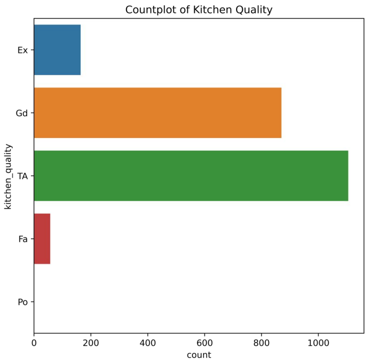
garage_quality

heating_quality

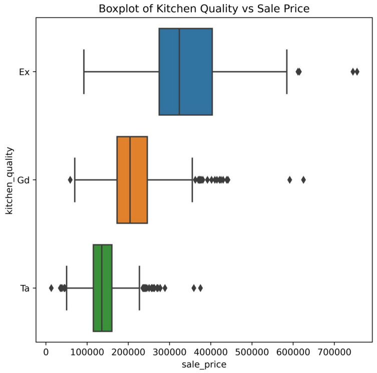
kitchen_quality

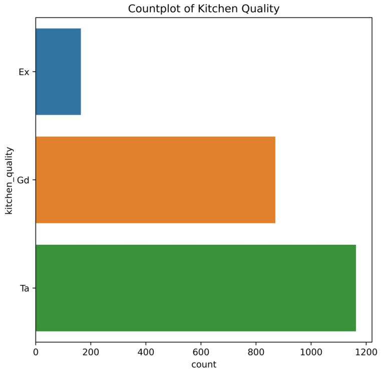
Before



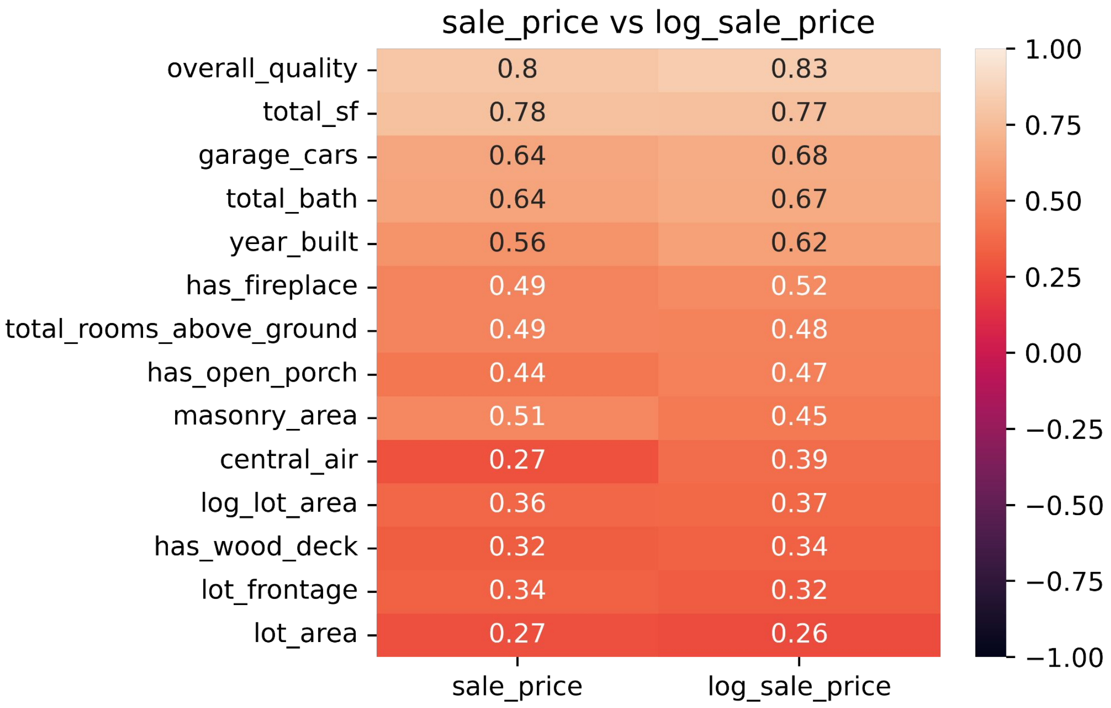


After





Data Cleaning - Log Transform



- Log transforming following features:
    - Target variable y - sale_price

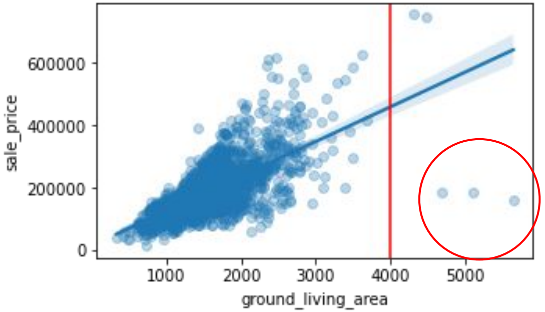
- Features correlates better with log_sale_price

- This will be evidential during modeling

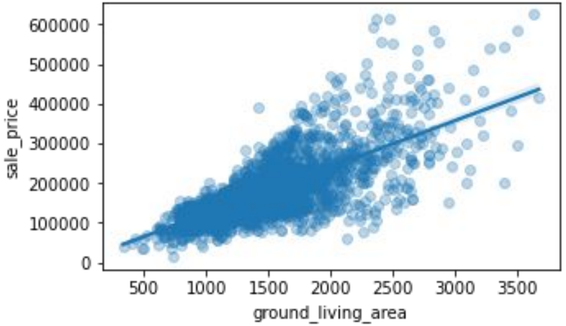
Data Cleaning - Outliers

Before → **After**

Scatterplot for Sale Price and Ground Living Area (corr: 0.6997)

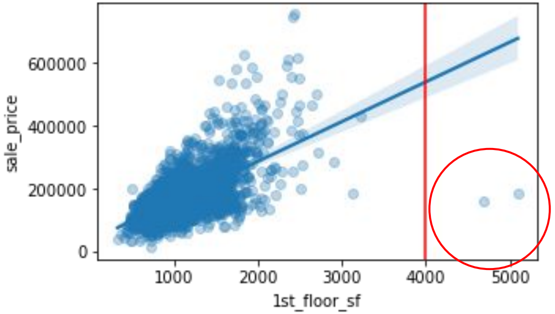


Scatterplot for Sale Price and Ground Living Area after Remove Outliner (corr: 0.7161)

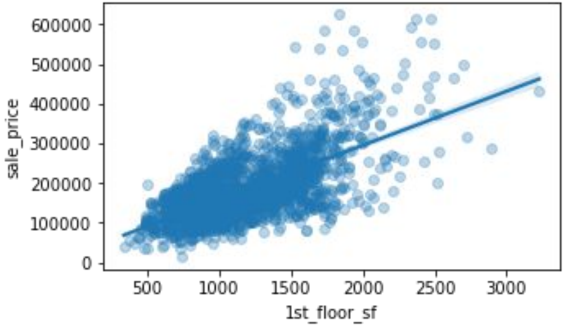


Drop row:
Ground living square footage of >= 4000 sqft. (far from trendline)

Scatterplot for Sale Price and 1st Floor Square Feet (corr: 0.6192)



Scatterplot for Sale Price and 1st Floor Square Feet after Remove Outliner (corr: 0.6471)



Affected:
Features correlates better with sale price

Data Cleaning - Cleaned Dataset

Datasets

Shape : 2092 rows, 60 columns

Numerical : 26

Categorical : 34

Columns Group Up

- **Overall**
- **Total**
- **Lot**
- **Rooms**
- **Date**
- **Other**
- **Feature**

Agenda

Ideal Workflow of Data Science Project

1. **Project Overview**

   Brief understanding of Ames IOWA and the dataset

2. **Data Cleaning**

   Fill missing values

3. **Data Pre-processing**

   - Clean up redundant columns
   - Feature engineer new columns

4. **Model Iteration**

   Iterate combination of cleaned features vs model benchmark

5. **Model Production**

   Select model for production

6. **Further Investigation**

   Areas to explore in the future

Terminology

R-squared (R^2)

- goodness-of-fit measure for linear regression model

  (Higher = Better)

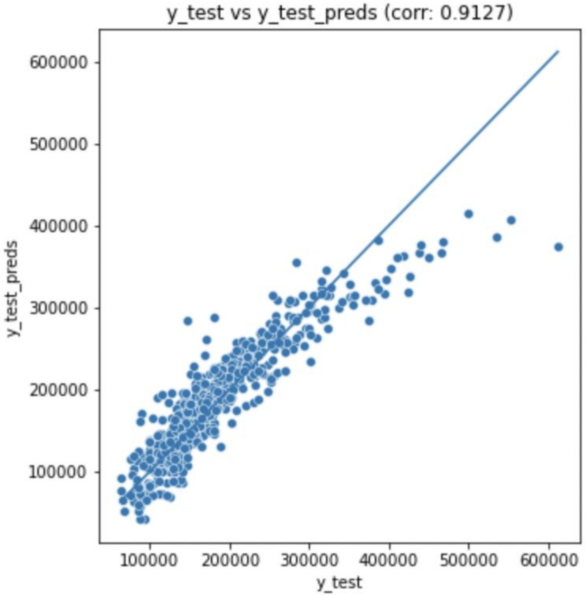
Root Means Squared Error (RMSE)

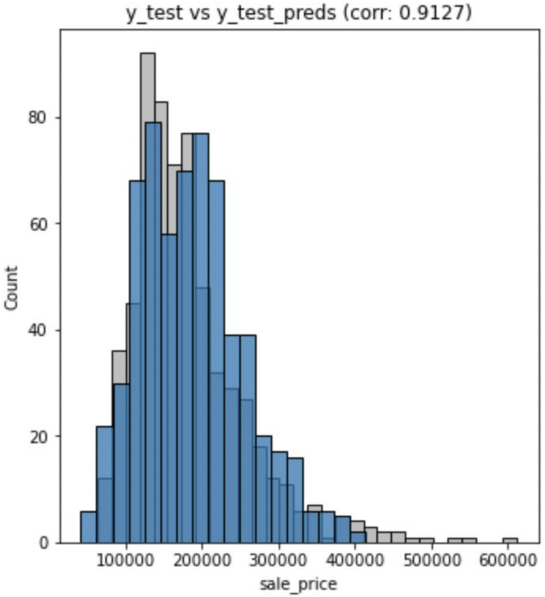
- Standard deviation of the residuals (prediction errors)

  (Lower = Better)

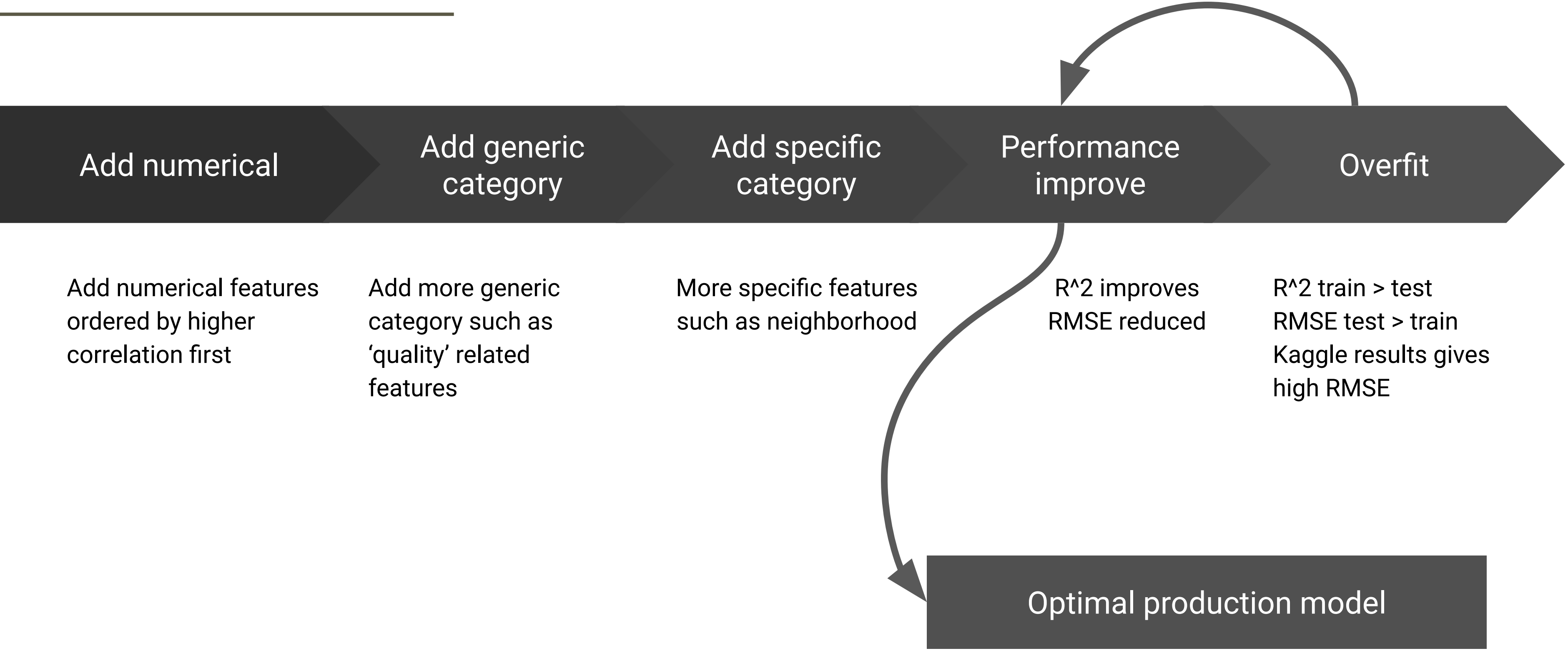
Model 1 (Benchmark) - Top 5 Numerical Features

Model	1
Train R^2	0.83
Test R^2	0.83
Train RMSE	31,982
Test RMSE	31,419





Iteration Approach



Add numerical

Add numerical features ordered by higher correlation first

Add generic category

Add more generic category such as ‘quality’ related features

Add specific category

More specific features such as neighborhood

Performance improve

R^2 improves
RMSE reduced

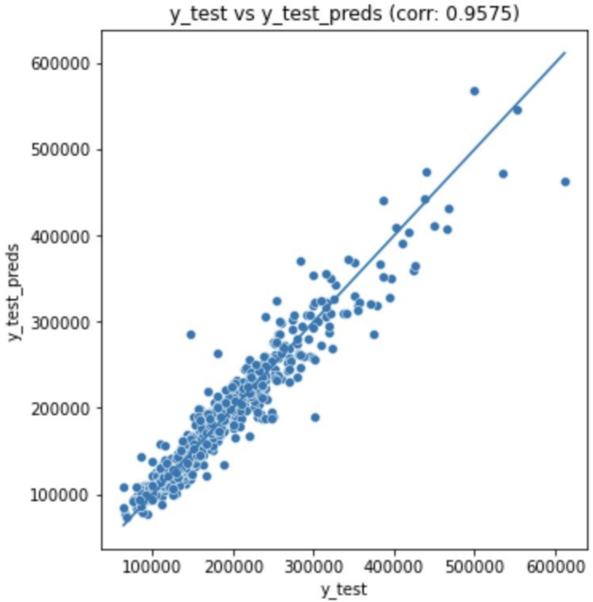
Overfit

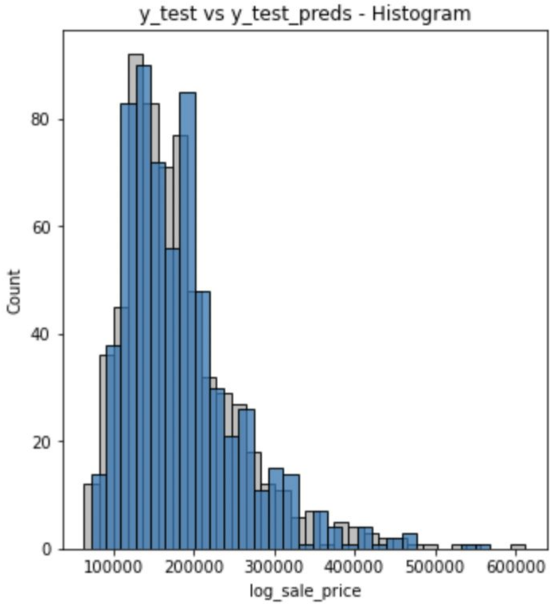
R^2 train > test
RMSE test > train
Kaggle results gives high RMSE

Optimal production model

Model 2 - All 25 Numerical Features

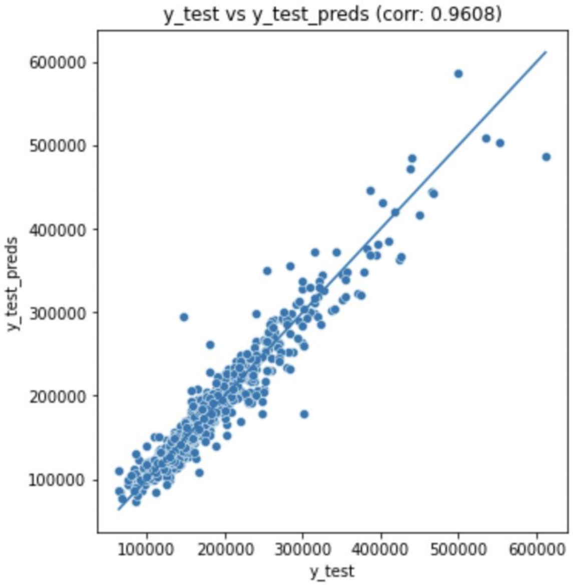
Model	1	2
Train R^2	0.83	0.91
Test R^2	0.83	0.91
Train RMSE	31,982	20,912
Test RMSE	31,419	22,084

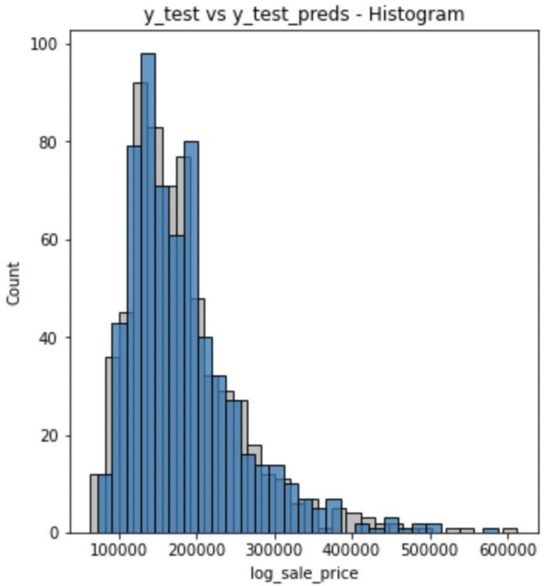




All 25 Numerical Features + 15 Categories

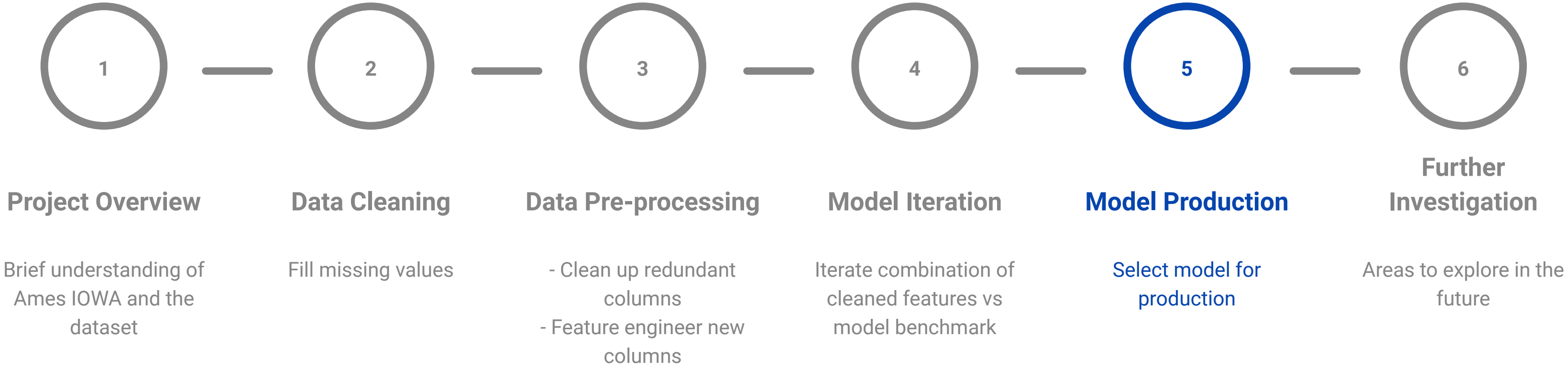
Model	1	2	3
Train R^2	0.83	0.91	0.93
Test R^2	0.83	0.91	0.91
Train RMSE	31,982	20,912	18,690
Test RMSE	31,419	22,084	21,251


y_test vs y_test_preds (corr: 0.9608)


y_test vs y_test_preds - Histogram

Agenda

Ideal Workflow of Data Science Project



1

Project Overview

Brief understanding of Ames IOWA and the dataset

2

Data Cleaning

Fill missing values

3

Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

4

Model Iteration

Iterate combination of cleaned features vs model benchmark

5

Model Production

Select model for production

6

Further Investigation

Areas to explore in the future

Production - All 25 Numerical + 6 Quality Categories

Numerical Features

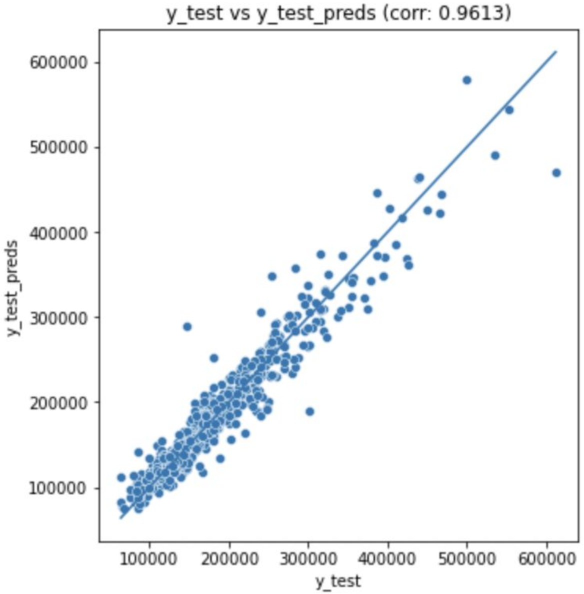
- **Overall Group**
  - Overall Quality
  - Overall Condition
- **Total Group**
  - Total Square Footage
  - Total Rooms Above Ground
  - Total Bath
- **Lot Group**
  - Lot Frontage
  - Lot Area (Natural Log)
  - Lot Slope (bool)
  - Lot Contour (bool)
  - Lot Shape (bool)
- **Rooms**
  - Bedroom
  - Kitchen
- **Date Group**
  - Year Built
  - Months Sold
  - Year Sold
- **Other**
  - Garage Cars
  - Masonry Area
  - Street (bool)
  - Central Air (bool)
  - Functional (bool)
- **Feature Group**
  - Fireplace (bool)
  - Open Porch (bool)
  - Wood Deck (bool)
  - Basement (bool)
  - 2nd Floor (bool)

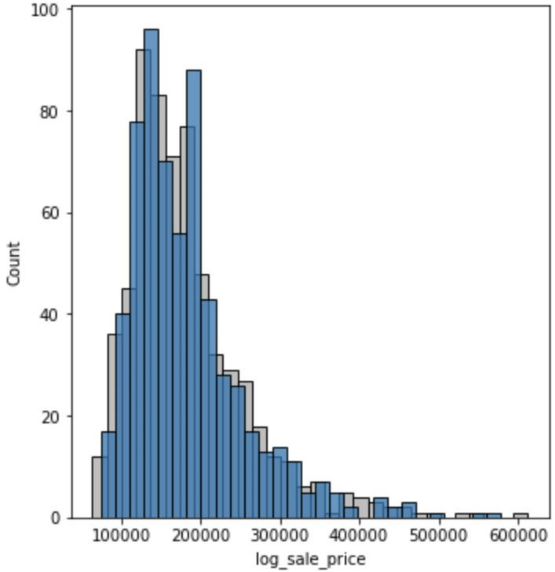
Categorical Features

- **Quality Group**
  - External Quality
  - Basement Quality
  - Heating Quality
  - Kitchen Quality
  - Fireplace Quality
  - Garage Quality

Production - All 25 Numerical + 6 Quality Categories

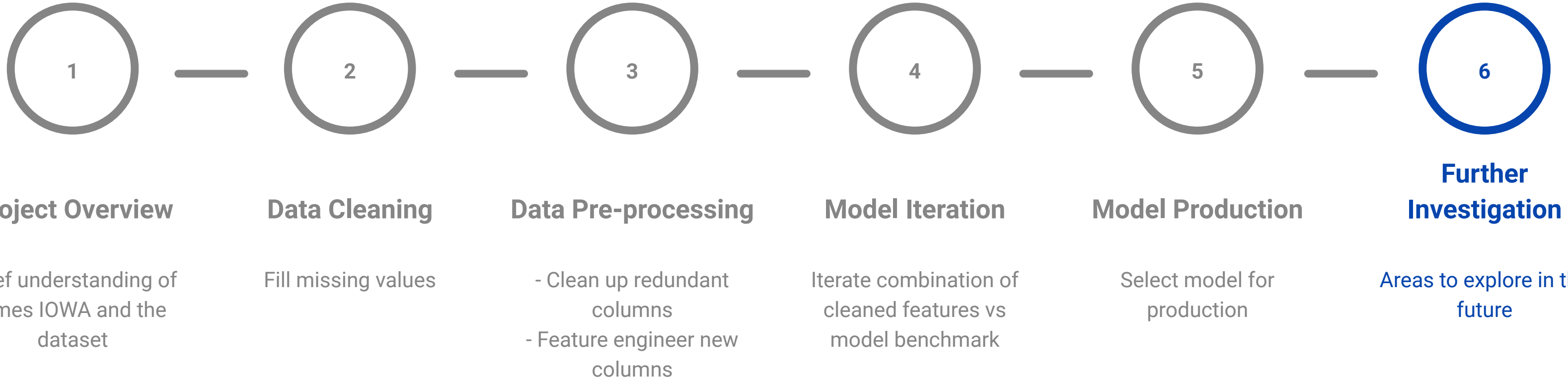
Model	1	2	3	Final
Train R^2	0.83	0.91	0.93	0.92
Test R^2	0.83	0.91	0.91	0.92
Train RMSE	31,982	20,912	18,690	19,704
Test RMSE	31,419	22,084	21,251	21,084





Agenda

Ideal Workflow of Data Science Project



1

Project Overview

Brief understanding of Ames IOWA and the dataset

2

Data Cleaning

Fill missing values

3

Data Pre-processing

- Clean up redundant columns
- Feature engineer new columns

4

Model Iteration

Iterate combination of cleaned features vs model benchmark

5

Model Production

Select model for production

6

Further Investigation

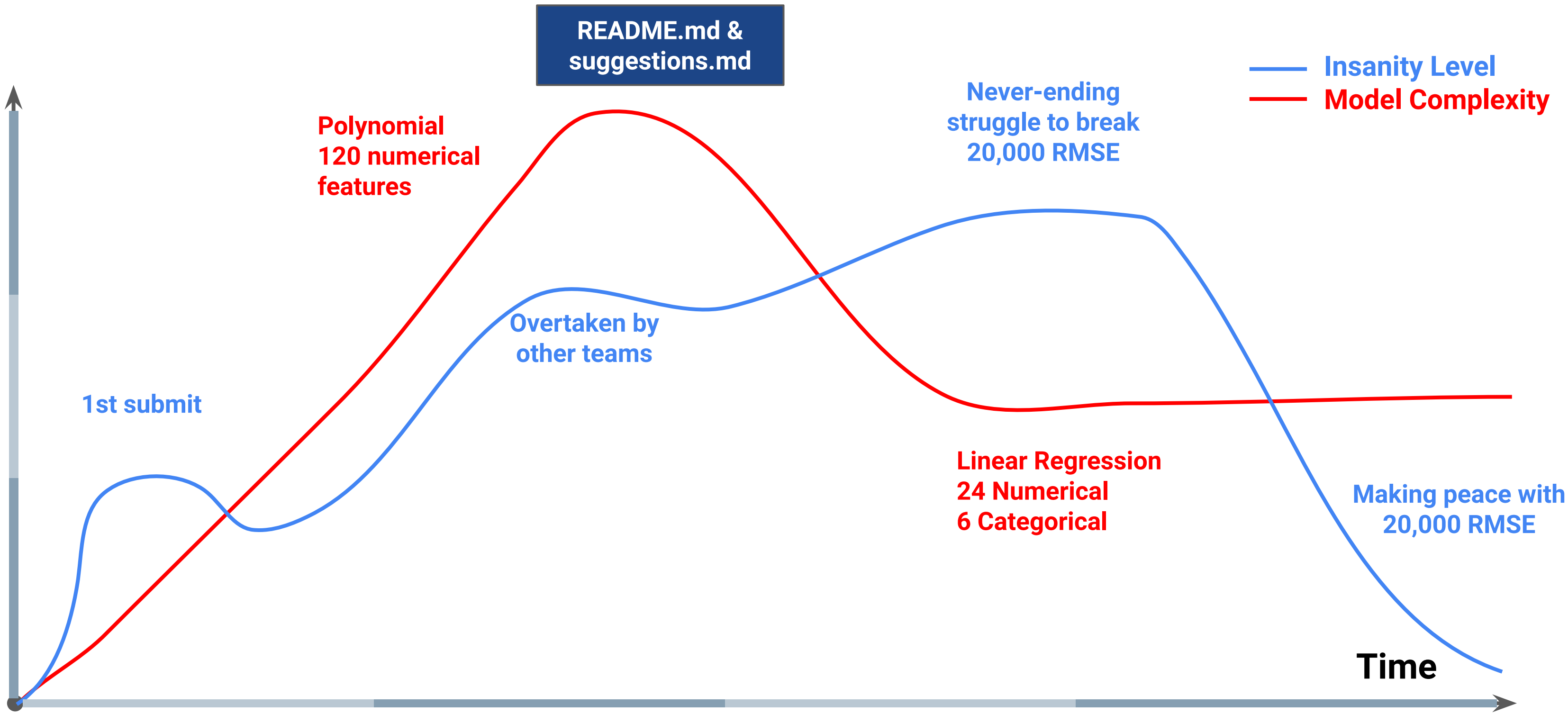
Areas to explore in the future

Further Investigation

- Categorise neighbourhood column to high, medium, and low sale price for use as a categorical feature in our model

- Using Cook’s Distance to identify multivariate outlier in order to optimize our model’s performance (library: yellowbrick)

Emotional Rollercoaster of Kaggle Competition



Key Takeaways

1. Kaggle Competition will drive you crazy

2. README.md should be renamed to ‘README_or_else_you’ll_regret_it.MD’

3. Simplicity is key

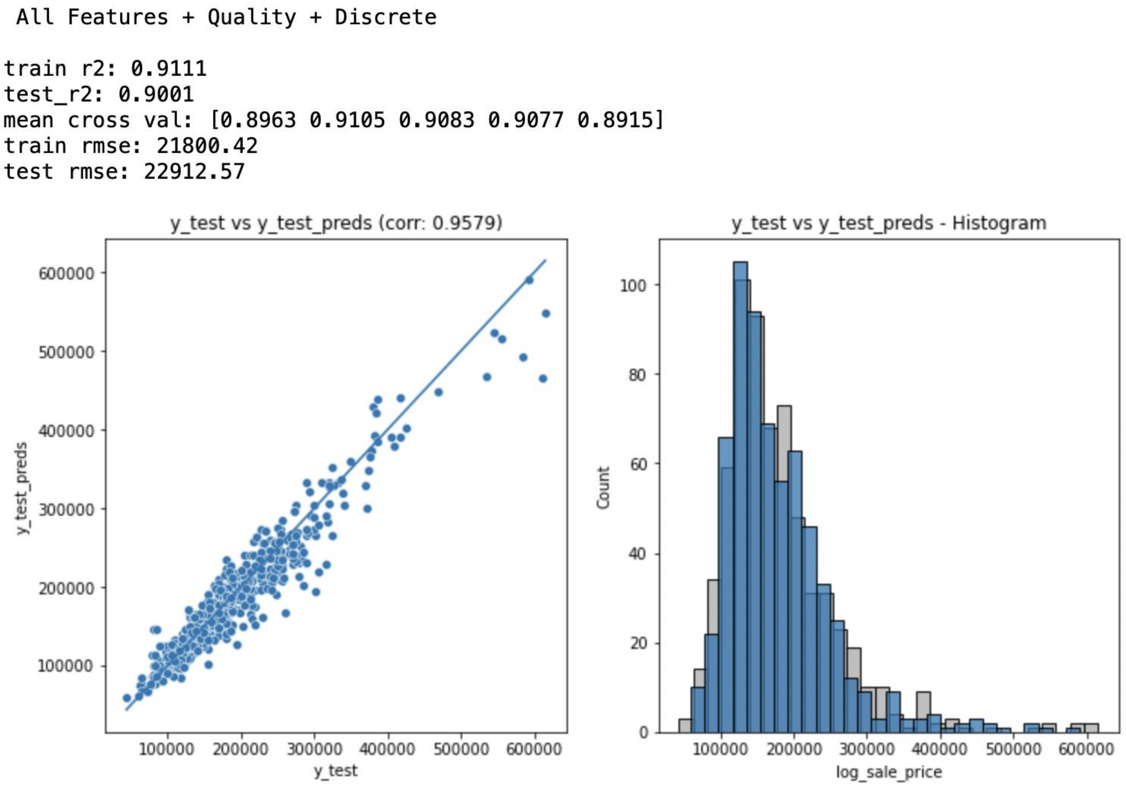
THANK YOU FOR LISTENING

BACKUP

Appendix 1 : Combined Discrete Columns > Category

- Columns converted
  - Total Baths
  - Total Rooms Above Ground
  - Overall Quality

- Results is not as good as leaving it as a discrete value
- Possibly because we’ve combined them into a highly price-correlated column



```
 All Features + Quality + Discrete

train r2: 0.9111
test_r2: 0.9001
mean cross val: [0.8963 0.9105 0.9083 0.9077 0.8915]
train rmse: 21800.42
test rmse: 22912.57
```