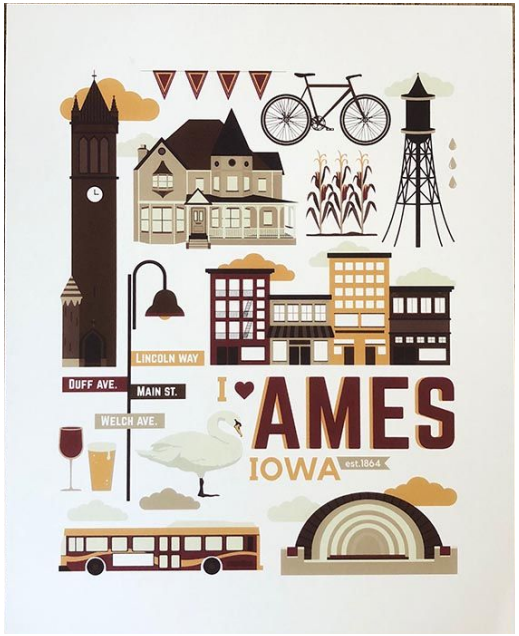
Ames Housing:

Is it Worth it?

Presented by:
- Johnny Tseng
- Ronnette Chan
- June Teo
- Eric Lim &
- Shu Kai Goh

Content



1. Statement
2. Background
3. Data Cleaning
4. Exploratory Data Analysis
5. Feature Engineering
6. Modelling & Interpretations
7. Summary & Conclusion
8. Q&A

Background

- Who are we?
- New to Ames, Iowa?
  - Population
  - Attraction
  - Fun Facts

Statement

- What to look out for
- Price point
- Features in data

Data Cleaning

Standardise names

- Uppercase → lowercase
- Spaces → underscores
- E.g. ‘Garage Area’ → ‘garage_area

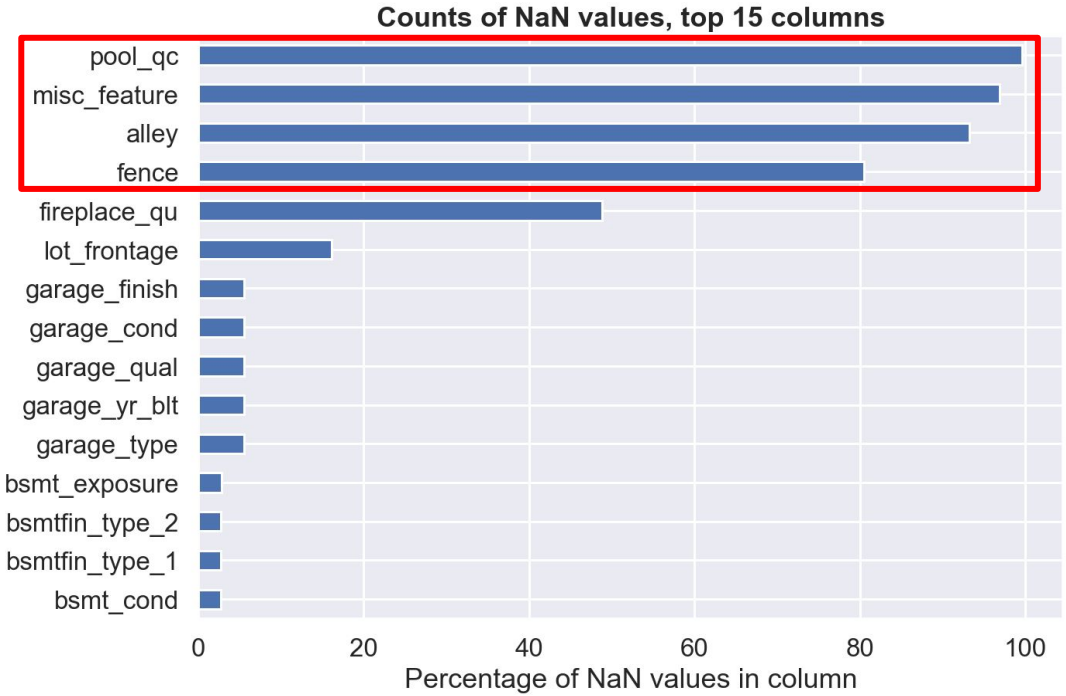
Unrealistic data removed

- Garage year built in 2207
- House was built after it was sold
- House was remodelled after it was sold

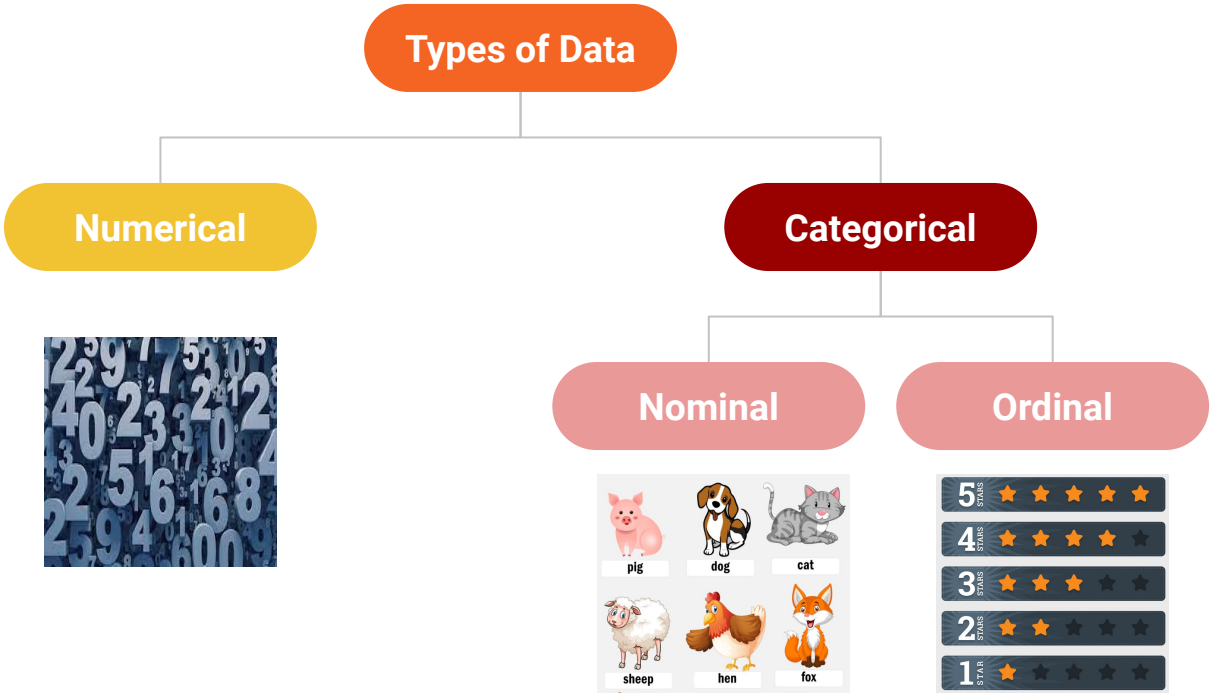
Data Cleaning

Removed features with a very **high proportion** of **missing** values

- Alley (**93%**)
- Pool QC (**99%**)
- Fence (**80%**)
- Misc Feature (**96%**)



Data Cleaning - Types of Data



Data Cleaning

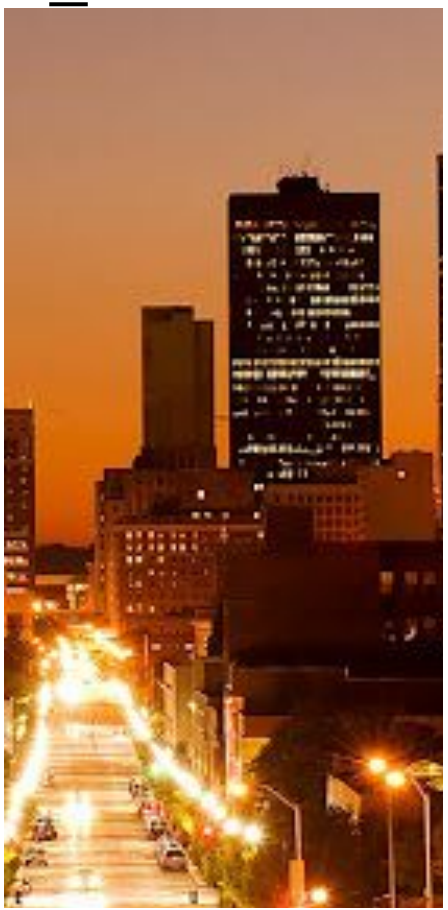
- **Numerical** Features with Null/Missing Values
    - Lot Frontage
        - Filled with the median lot frontage value taken from each lot configuration category
    - Masonry Veneer Area
        - Replaced with 0, the most common value
    - Garage-related Features
        - Replaced with 0 as the missing values in the column indicates ‘no basement’

Data Cleaning

- **Categorical** Features with Null/Missing Values
  - Fireplace Quality Feature
    - Filled with ‘No Fireplace’ as ‘NA’ indicates that the house has no fireplace
  - Basement-related Features
    - Filled with ‘No Basement’ as ‘NA’ indicates that the house has no basement
  - Garage-related Features
    - Filled with ‘No Garage’ as ‘NA’ indicates that the house has no garage

Data Cleaning

- Changing ordered categorical (**ordinal**) data to numerical data
  - Converted to a corresponding number that reflects its order
  - Example:
    - ‘Excellent’ → 5
    - ‘Good’ → 4
    - ‘Average’ → 3
    - ‘Fair’ → 2
    - ‘Poor’ → 1



Exploratory Data Analysis (EDA)

North Ames is the **most popular** out of 28 neighbourhoods

Houses in **Stone Brook** generally fetch a **higher price**.

77% of sales are classified as **low density residential** areas

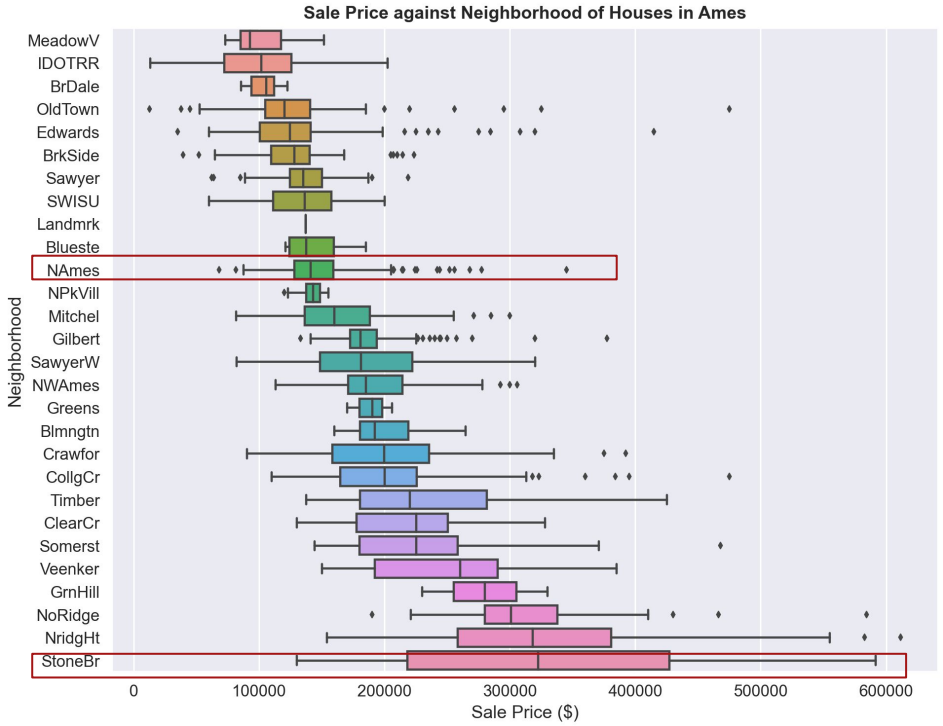
Low density residential areas generally fetch a **higher price**

Based on sample size of 2044.

Exploratory Data Analysis (EDA)

Clear differences in **median sale price**

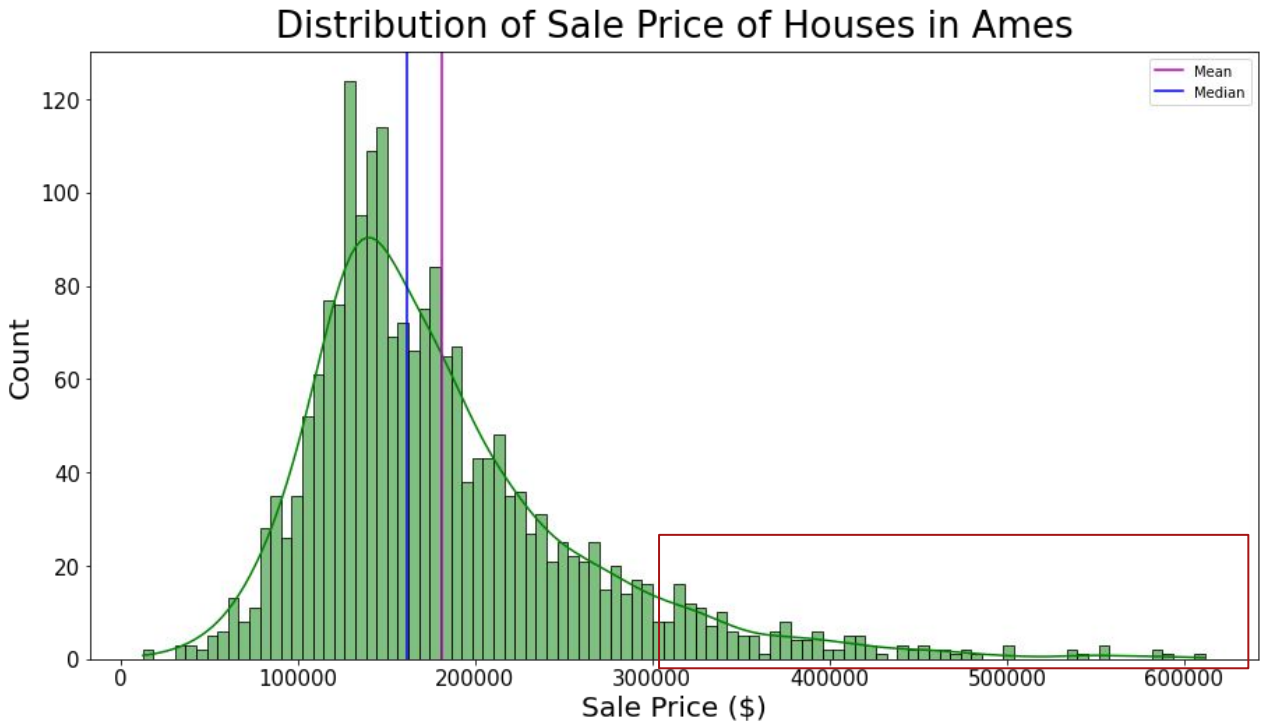
Some overlap in distribution of sale price



Exploratory Data Analysis (EDA)

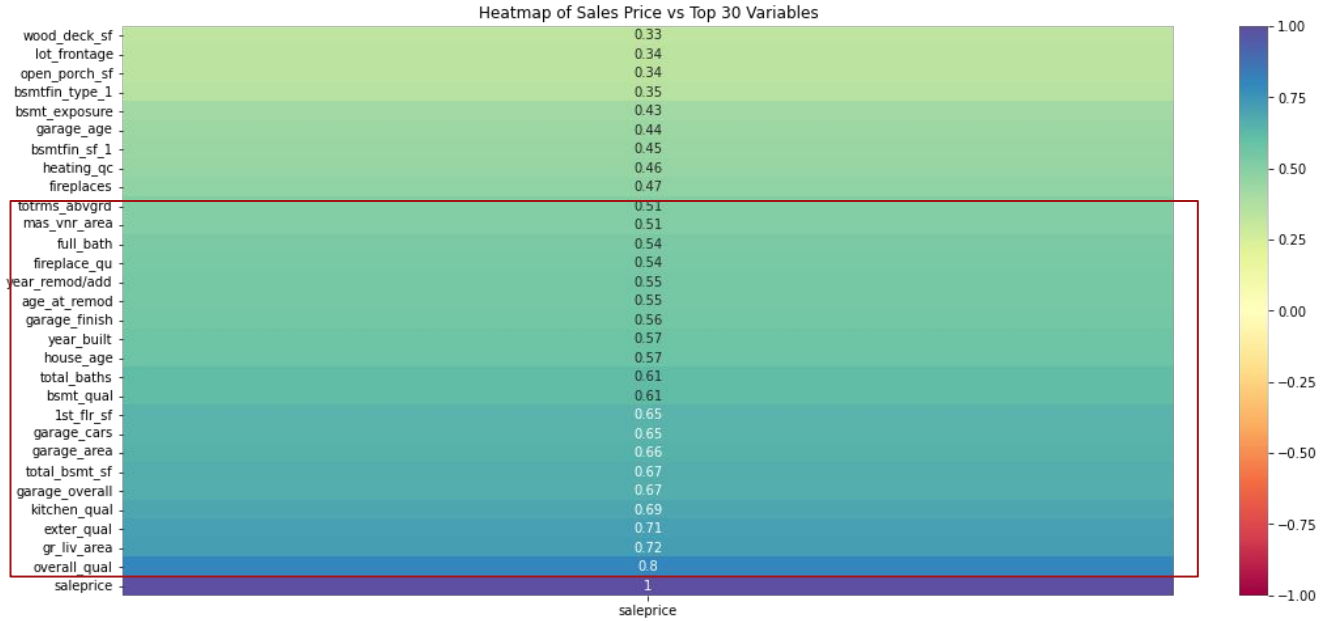
Mean: $181,354

Standard Deviation: $79,331



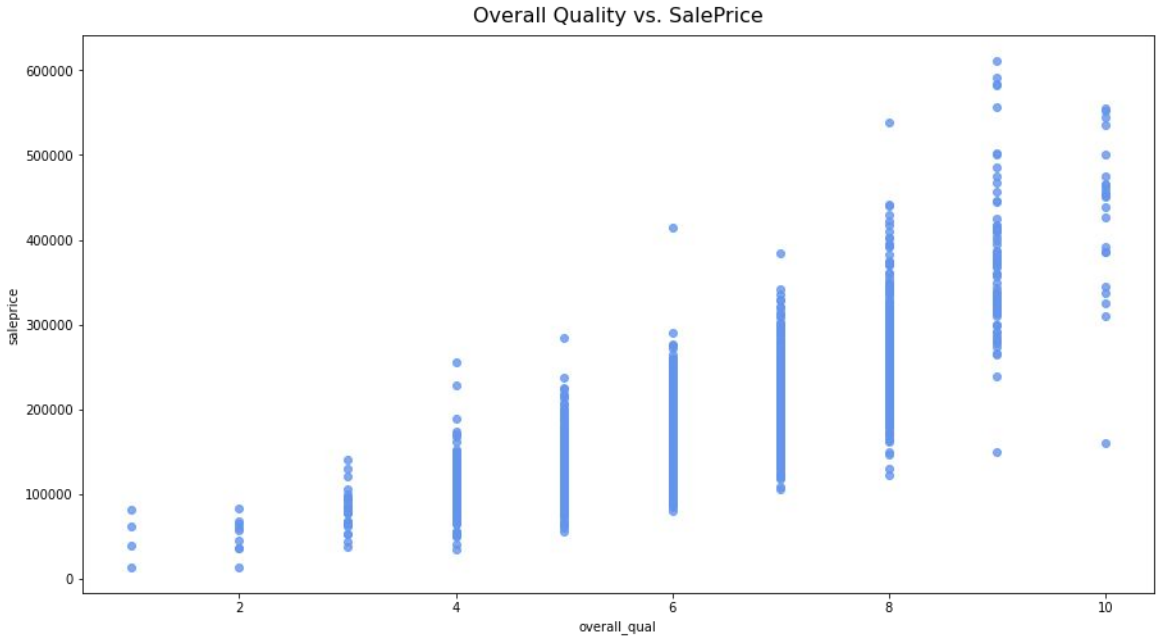
Exploratory Data Analysis (EDA)

Remove columns with correlation of +/- less than **0.45**



Exploratory Data Analysis (EDA)

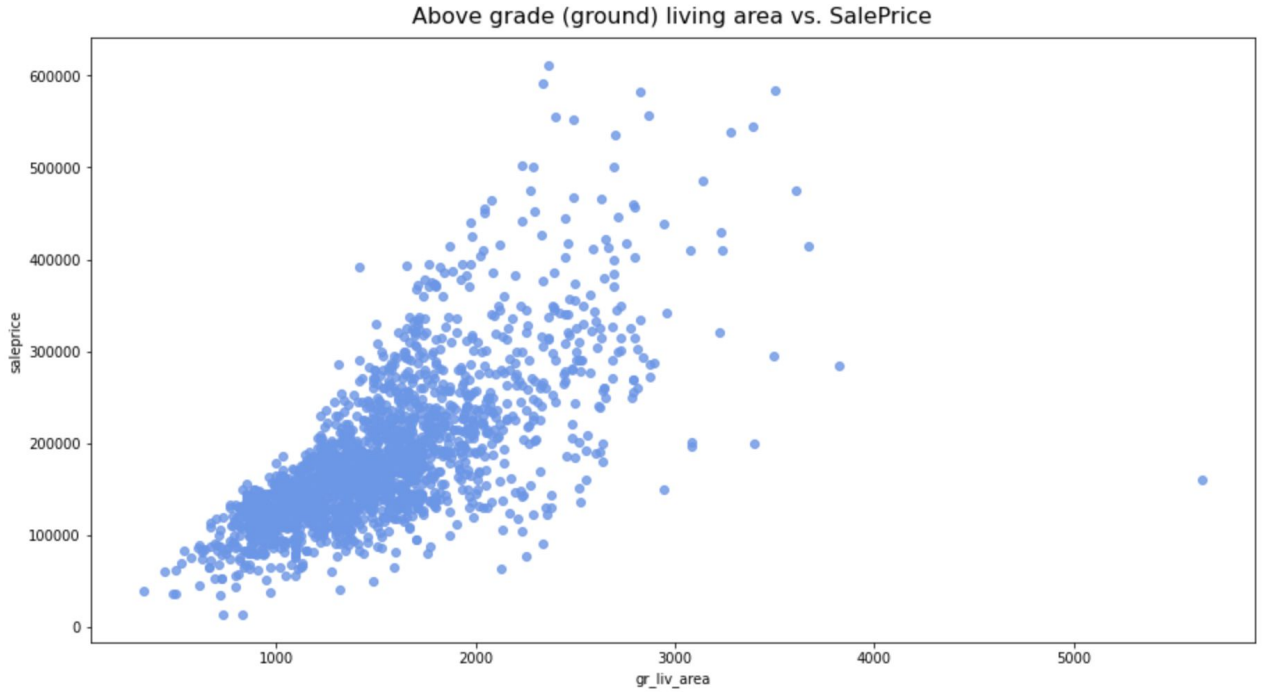
Higher rating
=
higher price



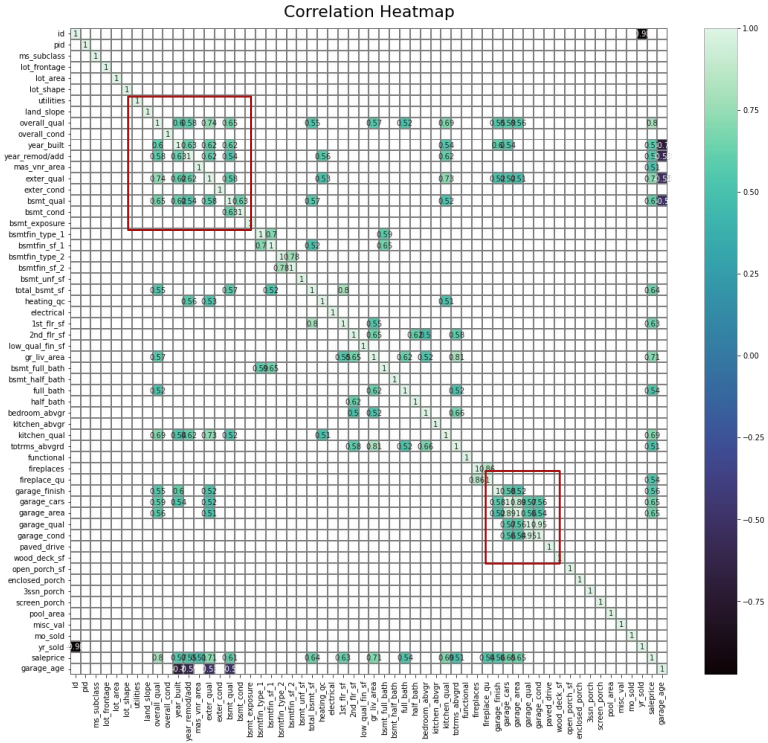
Exploratory Data Analysis (EDA)

Higher area
=
Higher price

Remove Outliers



Exploratory Data Analysis (EDA)



Too many variables, we will **combine a few features** together later, for a manageable analysis

Signs of multicollinearity, which will **affect model effectiveness.**

Adding of columns

Feature Engineering

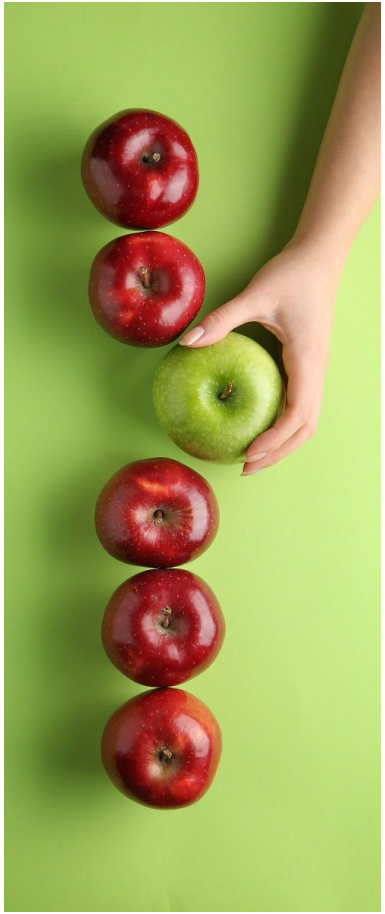
House Age = Year Sold - Year Built

Garage Age = Year Sold - Year Garage was Built

Age at Remodel = Year Sold - Year of Remodelling

Total Baths = (No. of Full Baths
+ No. of Half Baths
+ No. of Basement Full Baths
+ Number of Basement Half Baths)

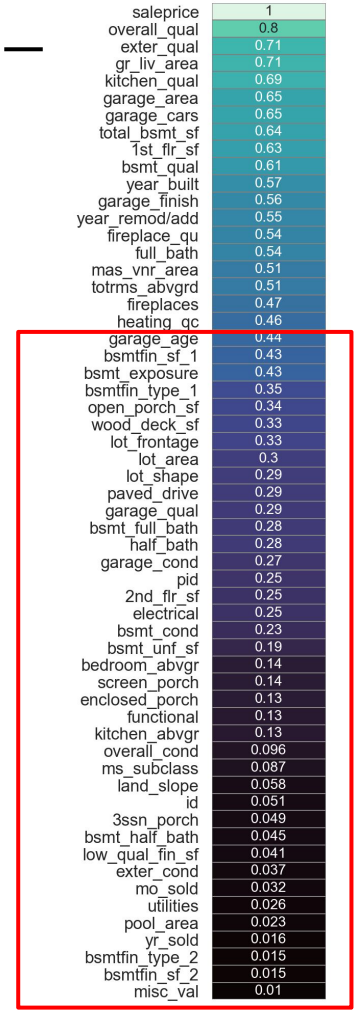
Garage Overall = Garage Quality x Garage Capacity



Extreme outliers

- House with above ground living area of 5642 square feet was removed.
- No other outliers were removed
- Even if we see skewing, removing too many - generalisation ability of model possibly affected
- Only remove very extreme cases

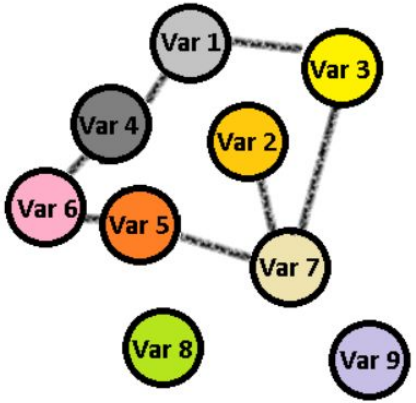
Removal of rows



Low correlation with Sale Price

- Features should have linear relation with target
- Numerical features having absolute correlation values **< 0.45** with sale price - Removed
- Ran default models to determine this threshold
- **39** columns removed, including
  - ID, PID
  - Lot frontage, Garage age
  - Year sold, Month sold
  - Basement-related features
  - Porch-related features.

Removal of columns



Removal of columns

Multicollinearity

- Should remove severely multicollinear features
- Removed one of a pair (lower correlation with target) when absolute correlation value > **0.75**

Year built, Year of remodelling/addition (kept Age features)

Garage car capacity, Garage area (kept Garage overall)

Fireplaces (kept Fireplace quality)

Total rooms above ground (kept Living area above ground)

1st Floor sq. feet (kept Total basement sq. feet)

Excluding categorical features

- 3 categorical columns were removed
- Ran default models to finalise which columns

Exterior covering on house (2nd, if more than one is present)

Proximity to various conditions such as railroads and parks (2nd, if more than one is present

General zoning classification (e.g. Residential, Agriculture)

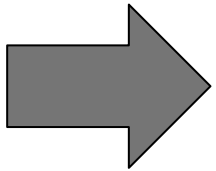
Removal of columns

Transforming columns

Final Preprocessing

- Changed categorical nominal data (text-based data, no particular order) to numerical form
  - ‘Dummy encoding’

central_air
Y
Y
N
Y

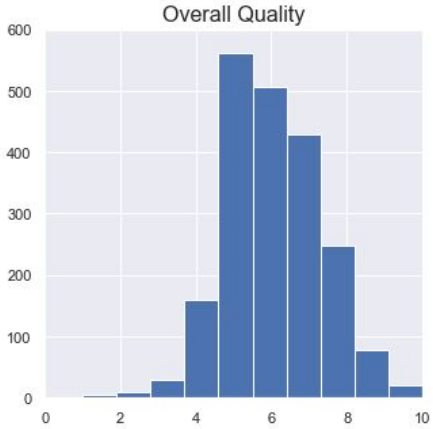


central_air_Y	central_air_N
1	0
1	0
0	1
1	0

Model Preparation

- Dependant variable: **Sale Price**
- Independent Variables: **73 variables** such as house age, overall quality, total baths etc.

- Train-test-split of **70-30**
  - Aligned with industry norms
- **Scaling** so that the model is not impacted because of variables with large magnitude.

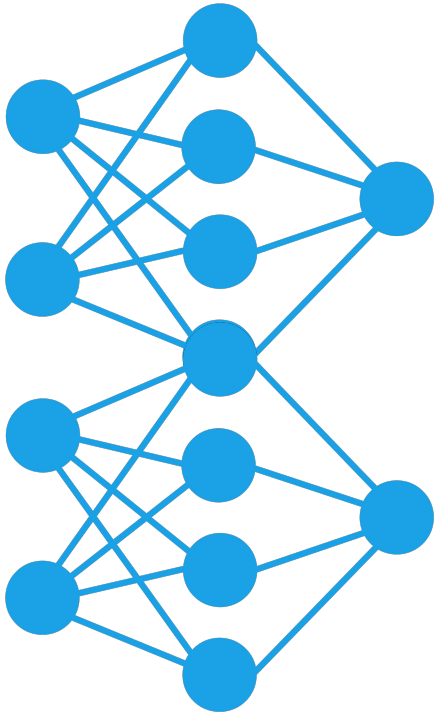


Model Evaluation

	Linear Regression	LassoCV	RidgeCV
Cross Validation	0.8794	0.7794	0.8843
Train R^2	0.9082	0.9068	0.9057
Test R^2	0.8901	0.8962	0.8991
Difference in R^2	0.0181	0.0106	0.0066

- Selected RidgeCV
  - Highest cross validation score
  - Smallest difference R^2 in between train and test set

** Higher the R^2, the better*



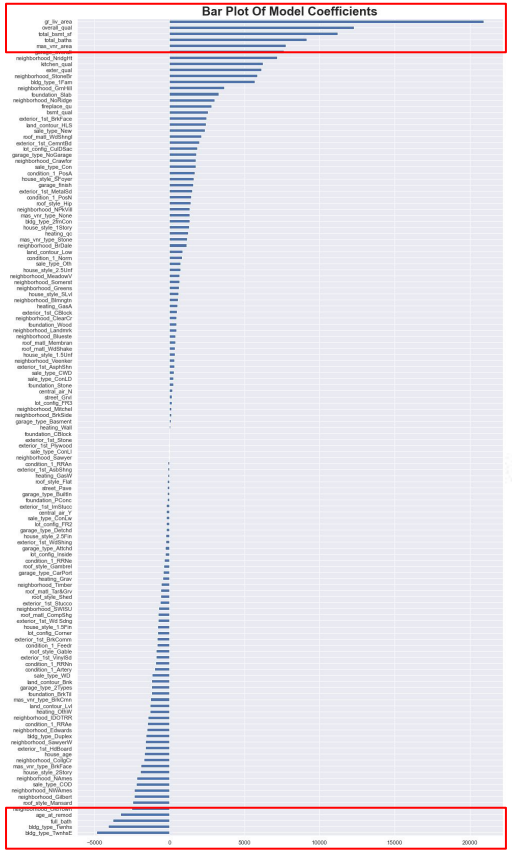
Model Selected: Ridge Model

Ridge RMSE
24198.52

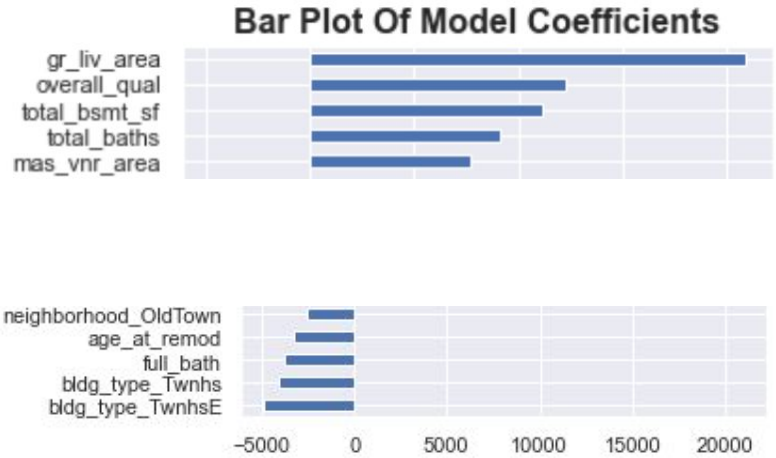
<

Baseline RMSE
80894.05

- Shrinks coefficient estimates towards zero (i.e. eliminating irrelevant variables) to make the model less complex.



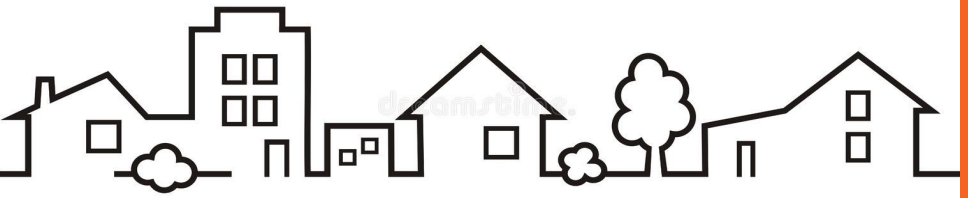
Interpretation



Size and quality of the house are important.

- Above ground living area
- Overall Quality
- Total square feet of basement area
- Total Number of bathrooms
- Masonry veneer area

Summary & Conclusion



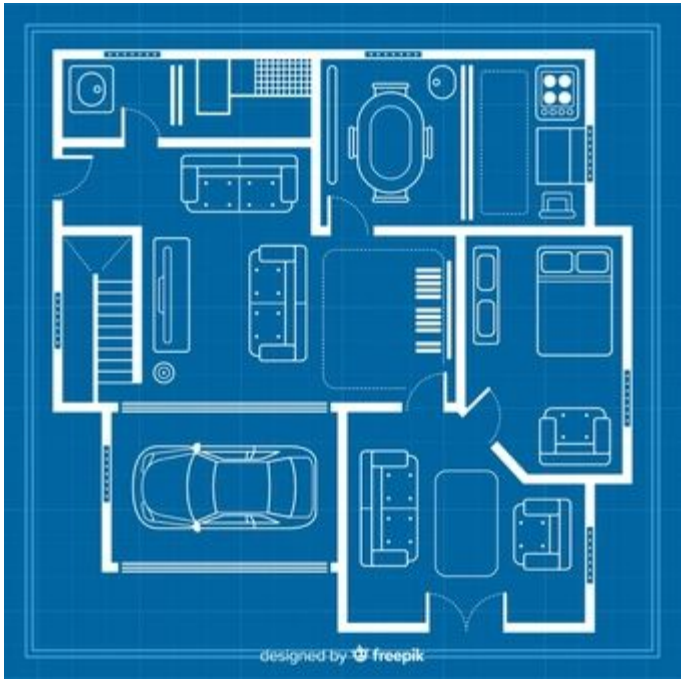
- Based on our findings
- Things to consider
- Further studies
- The Takeaways

Recommendations

- Sellers to pay attention to housing quality and size when setting sale price.
- In addition some of the key factors to note are as follows:
  - **Above ground living area** and **overall quality** influences the sale price more than lot configuration and neighbourhood.
  - The **total area** of basement, and **total number** of bathrooms, etc has a larger effect on the sale price than the condition.
  - Having dwelling types of **Townhouse Inside Unit** and **Townhouse End Unit** will negatively affect on the sale price.

Land Size


Sale Price against Above Ground Living Area of Houses in Ames
Sale Price ($)
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Above ground living area in square feet


designed by freepik

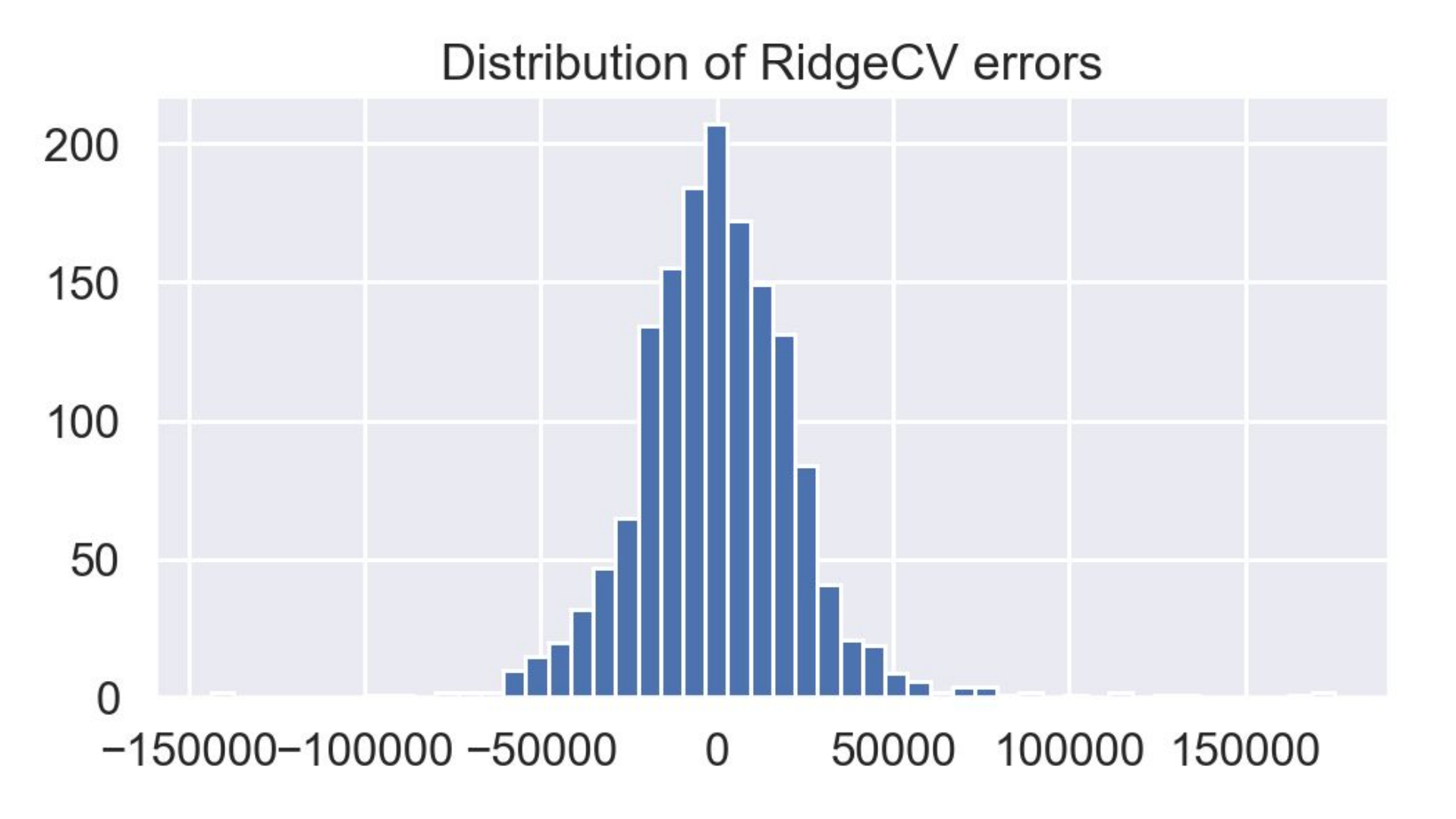
Quality & Quantity

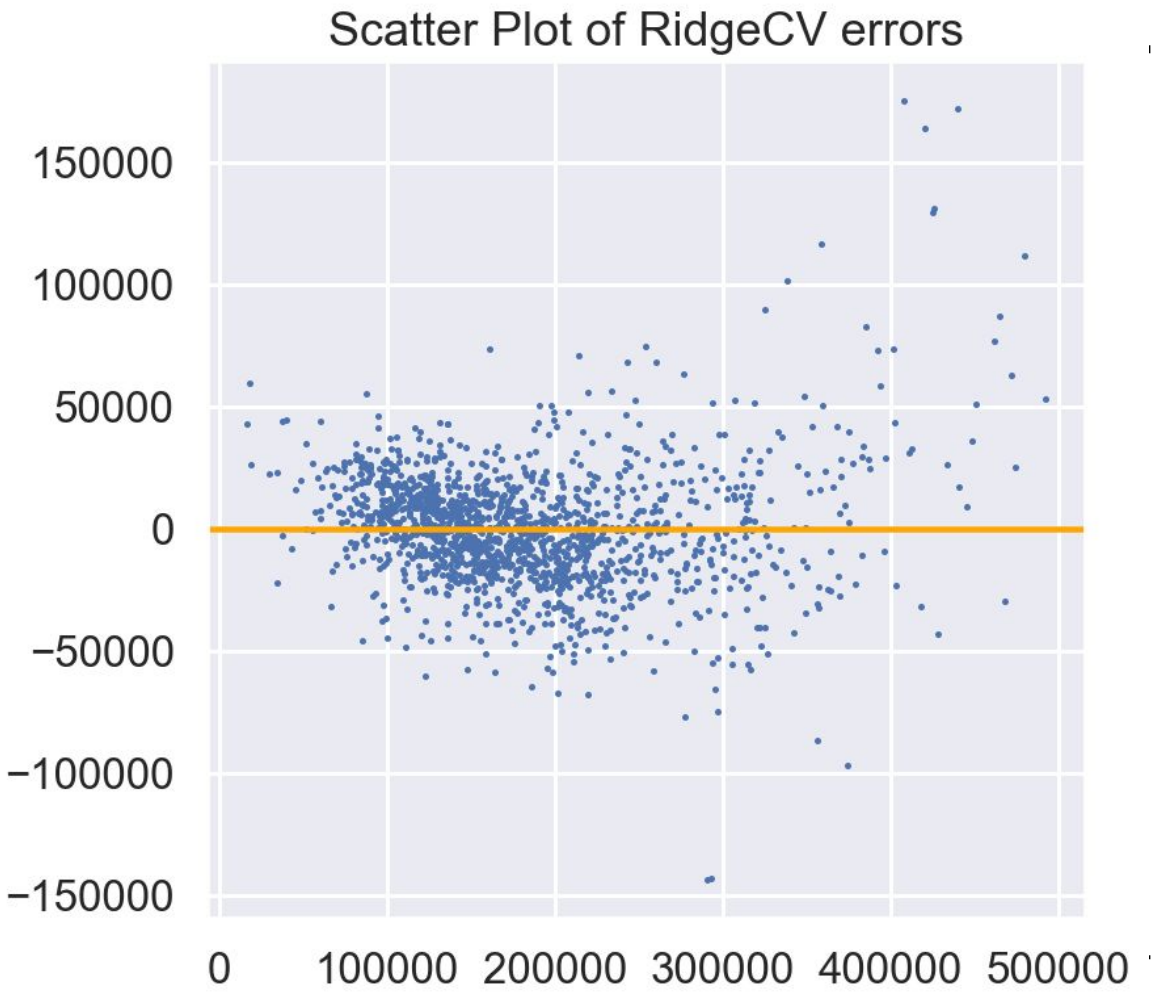
Amenities

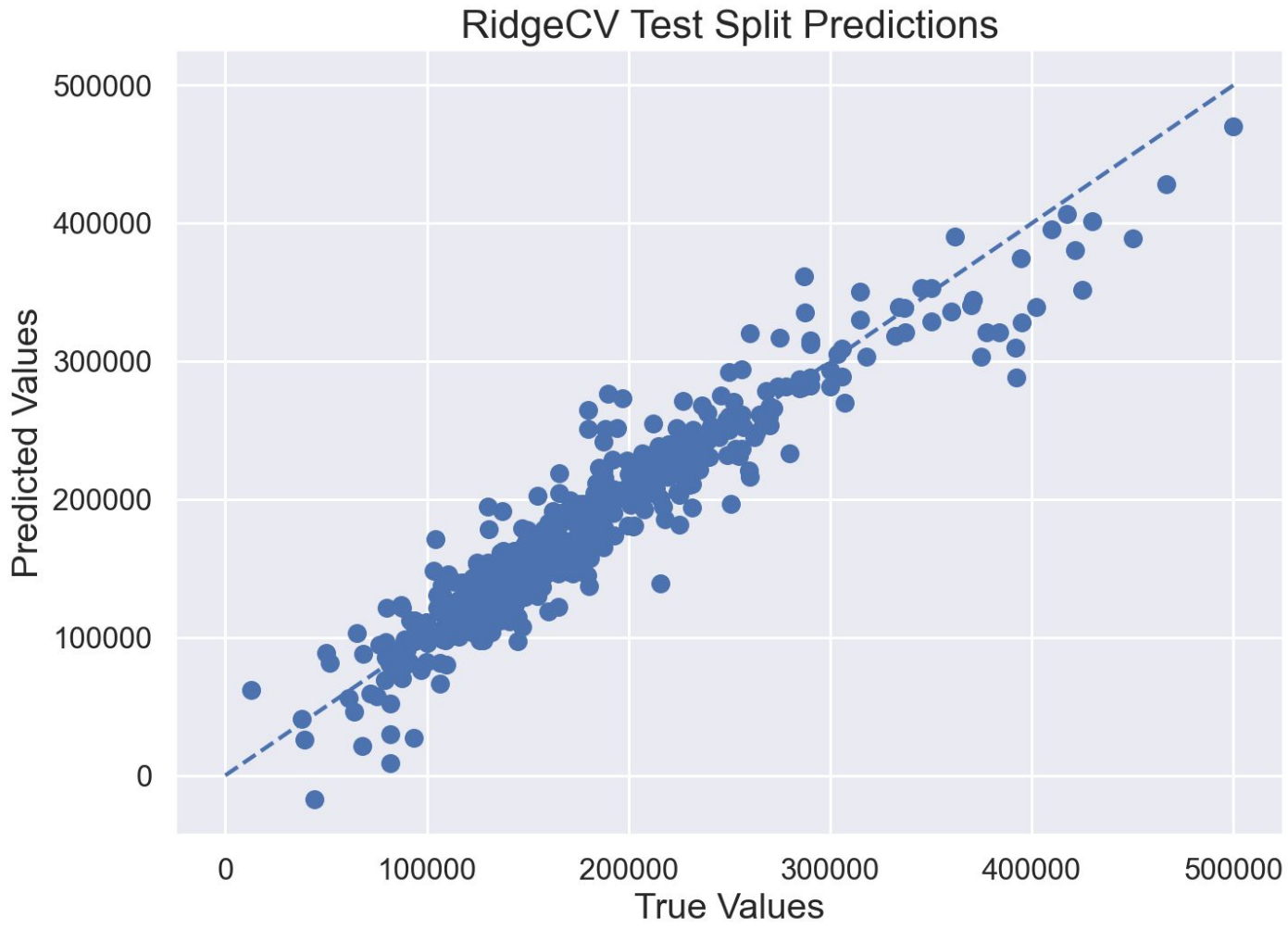
Thank You

Questions?

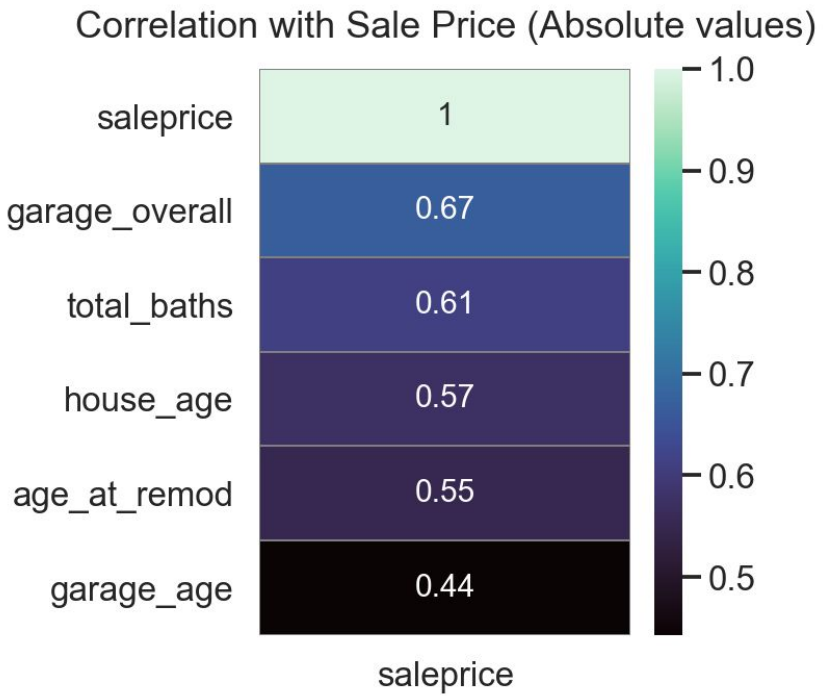
Appendix


Distribution of RidgeCV errors
200
150
100
50
0
−150000
−100000
−50000
0
50000
100000
150000


Scatter Plot of RidgeCV errors
150000
100000
50000
0
−50000
−100000
−150000
0
100000
200000
300000
400000
500000


RidgeCV Test Split Predictions
Predicted Values
True Values
500000
400000
300000
200000
100000
0
0
100000
200000
300000
400000
500000

Correlation of Engineered Columns with Sale Price



Multicollinearity before Dummy Encoding

	feature_1	feature_2	1_2_abs_corr	1_y_corr	2_y_corr
0	house_age	year_built	0.999054	0.572028	0.571976
1	age_at_remod	year_remod/add	0.998053	0.551938	0.550534
2	garage_overall	garage_cars	0.986520	0.667549	0.648383
3	garage_area	garage_cars	0.896963	0.655061	0.648383
4	garage_overall	garage_area	0.896036	0.667549	0.655061
5	fireplace_qu	fireplaces	0.860692	0.539093	0.473624
6	totrms_abvgrd	gr_liv_area	0.812626	0.509762	0.719694
7	1st_flr_sf	total_bsmt_sf	0.792071	0.648273	0.666429

Multicollinearity after Dummy Encoding

	feature_1	feature_2	1_2_abs_corr	1_y_corr	2_y_corr
0	central_air_Y	central_air_N	1.000000	0.277284	0.277284
1	street_Pave	street_Grvl	1.000000	0.069821	0.069821
2	roof_style_Hip	roof_style_Gable	0.948206	0.267016	0.251054
3	mas_vnr_type_None	mas_vnr_type_BrkFace	0.824365	0.411834	0.257498
4	roof_matl_Tar&Grv	roof_style_Flat	0.786279	0.007137	0.010337
5	heating_GasW	heating_GasA	0.775994	0.037714	0.093938
6	foundation_PConc	foundation_CBlock	0.773273	0.529180	0.355644
7	roof_matl_Tar&Grv	roof_matl_CompShg	0.772685	0.007137	0.071869

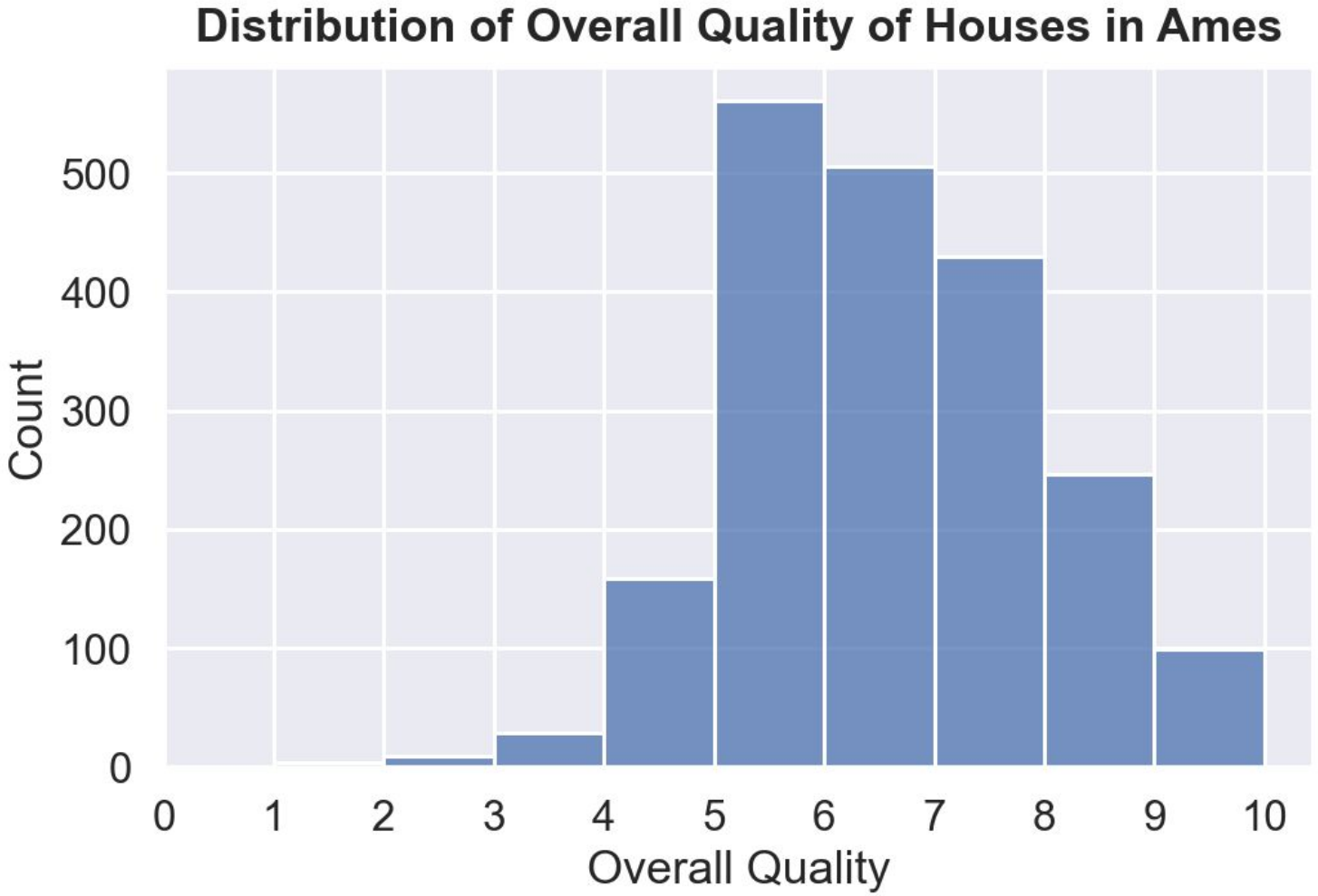
badeloftusa.com/buying-guides/bathrooms/

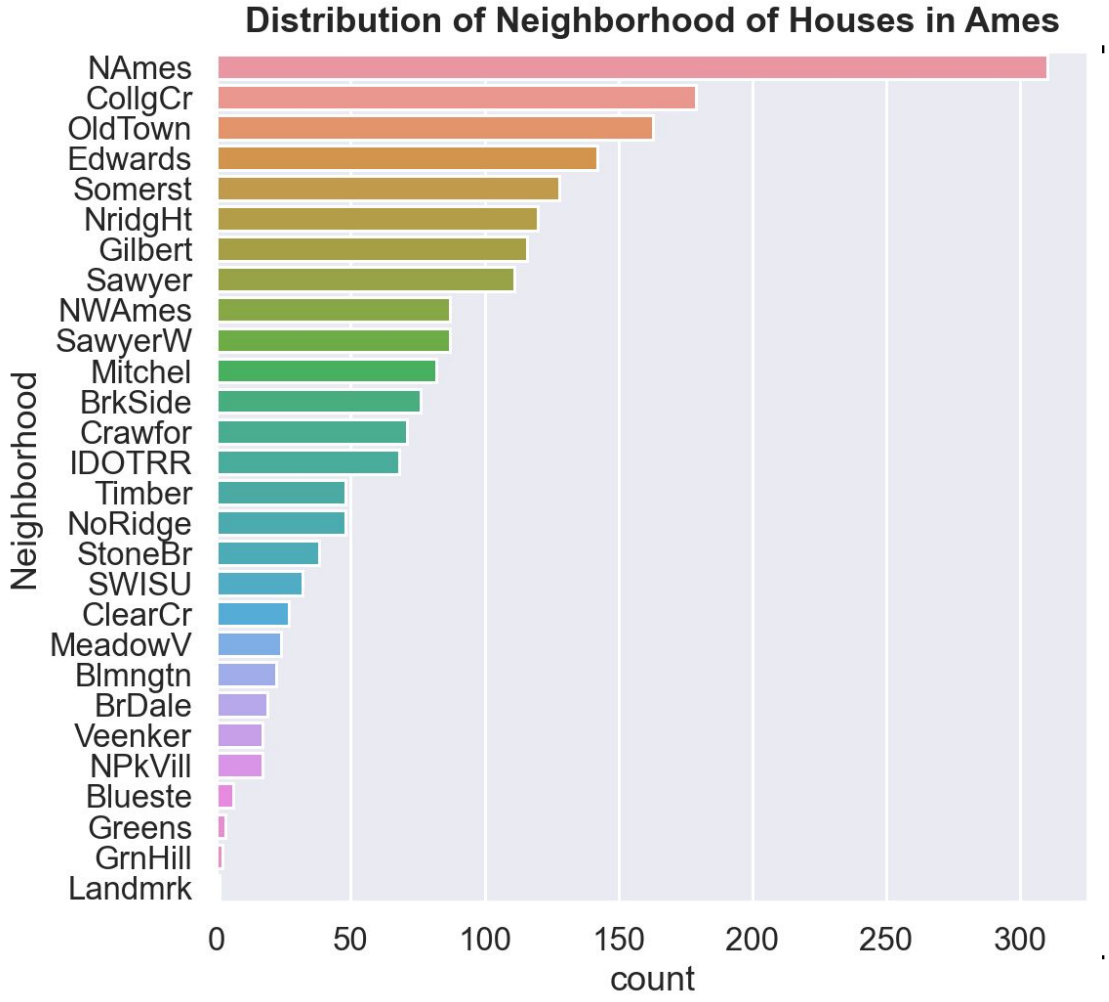
Free Material Sample

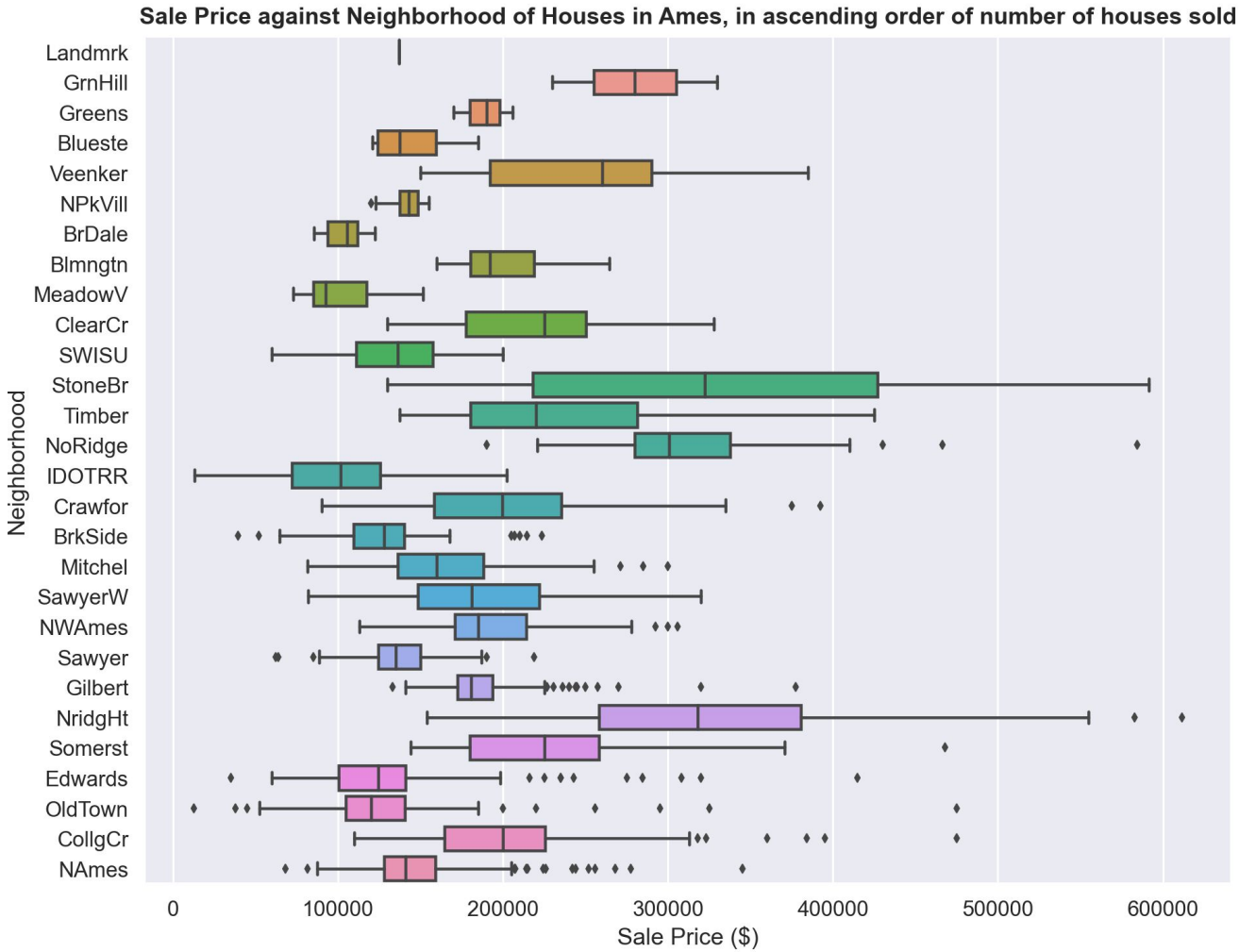
What is a full bathroom?

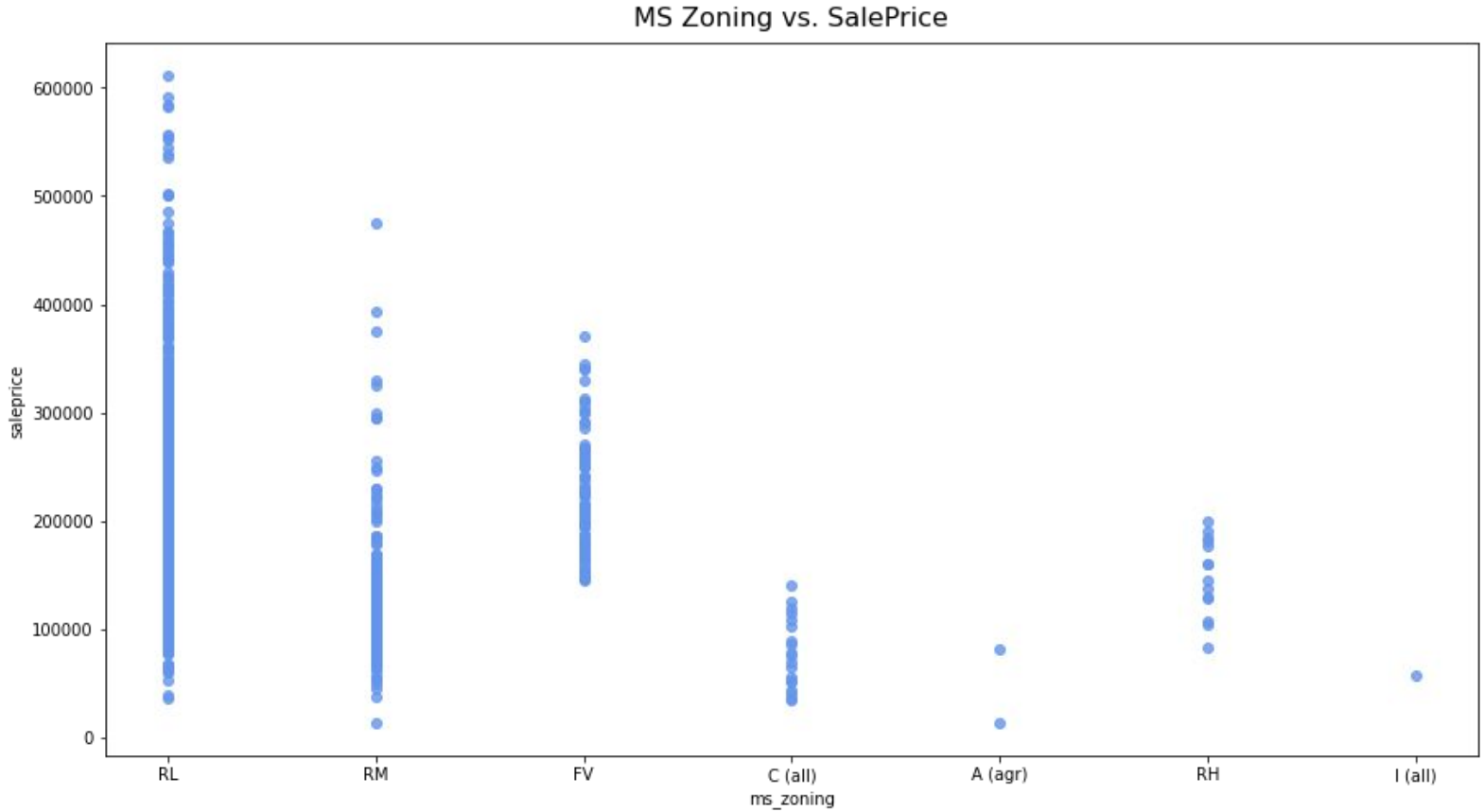
A full bathroom is a bathroom that contains exactly 1 sink, 1 bathtub, 1 shower and 1 toilet. Of course you can have more than just 1 of each of these items in your bathroom, but for it to be called a full bath, it must contain at least 1 of each of these items. And because of the space required, a full bath is rare in most homes, with usually just one in most homes, most likely in the master bedroom of the home.

Townhouse Units


Distribution of Overall Quality of Houses in Ames
Count
0
100
200
300
400
500
0
1
2
3
4
5
6
7
8
9
10
Overall Quality


Distribution of Neighborhood of Houses in Ames
Neighborhood
NAmes
CollgCr
OldTown
Edwards
Somerst
NridgHt
Gilbert
Sawyer
NWAmes
SawyerW
Mitchel
BrkSide
Crawfor
IDOTRR
Timber
NoRidge
StoneBr
SWISU
ClearCr
MeadowV
Blmngtn
BrDale
Veenker
NPkVill
Blueste
Greens
GrnHill
Landmrk
0
50
100
150
200
250
300
count


Sale Price against Neighborhood of Houses in Ames, in ascending order of number of houses sold
Neighborhood
Landmrk
GrnHill
Greens
Blueste
Veenker
NPkVill
BrDale
Blmngtn
MeadowV
ClearCr
SWISU
StoneBr
Timber
NoRidge
IDOTRR
Crawfor
BrkSide
Mitchel
SawyerW
NWAmes
Sawyer
Gilbert
NridgHt
Somerst
Edwards
OldTown
CollgCr
NAmes
0
100000
200000
300000
400000
500000
600000
Sale Price ($)


MS Zoning vs. SalePrice
saleprice
600000
500000
400000
300000
200000
100000
0
RL
RM
FV
C (all)
A (agr)
RH
I (all)
ms_zoning