Regression-omics

Project 2

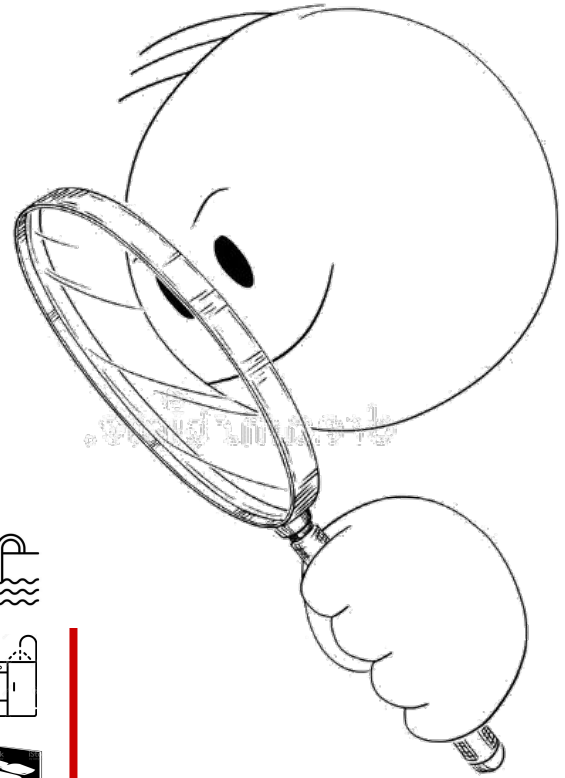
Have you ever looked at a nice house and said,

“I wonder how much that’s worth?”

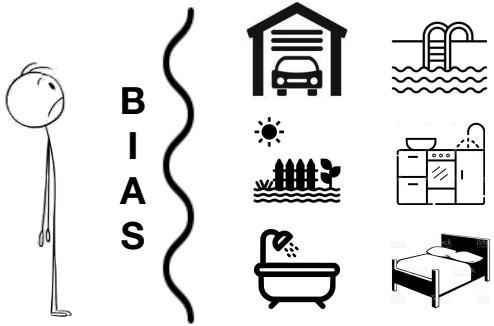
Perhaps you’ve **fretted about the value of your own property**, or got into a **heated argument with a relative/potential buyer** on the perceived price of your property.

There are are **many features** that determine how much a home can fetch!

An ***Outside View**** involves ignoring these details and using an estimate based on a class of roughly similar previous cases.



Your **Inside View** involves making predictions based on your understanding of the details of the process - not that it's wrong, it's just that it's often biased!



Excerpt from Daniel Kahneman - Thinking Fast and Slow: The outside view asks if there are similar situations that can provide a statistical basis for making a decision. Rather than seeing a problem as unique, the outside view wants to know if others have faced comparable problems and, if so, what happened

Problem Statement

There are are many variables that determine how much a home can fetch.

Using the Ames (IA) dataset (train, test), we want to find out which variables matter for home sale prices and produce accurate sale price predictions.

This model will help provide the *Outside View**, helping to reduce information asymmetry between potential home-buyers, home-sellers and real estate agents.

Overview

- Defining our problem statement (Zhewei)
- Overview of our datasets / Exploratory data analysis (EDA) (Wee Hong)
- Cleaning our data (Darren)
- Processing our data (DJ)
- Model Selection (Darien)
- Conclusion (Darien)

Overview of our dataset

Dataset Snapshot

- 80 features + 1 target variable (‘sale-price’)
- 42 qualitative columns vs. 38 quantitative columns (incl. ‘id’, ‘pid’)

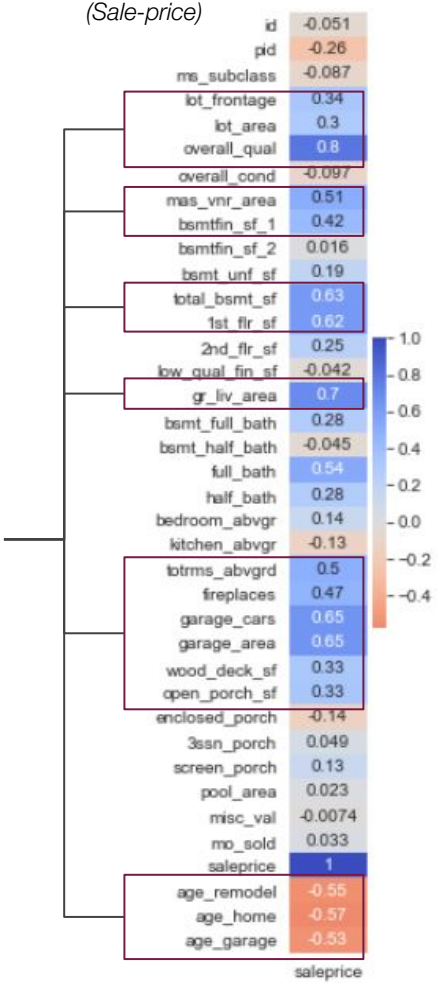
Brief Process Taken

- Data Cleaning and Discovery
- Descriptive Statistics to ensure we have good dataset to develop a linear model
- Features Engineering and Selection (illustrated in diagram on the right)
- Model Development

Selected features

saleprice	1.000000
overall_qual	0.800207
gr_liv_area	0.697038
garage_area	0.650270
garage_cars	0.648220
total_bsmt_sf	0.628925
1st_flr_sf	0.618486
full_bath	0.537969
mas_vnr_area	0.512230
totrms_abvgrd	0.504014
fireplaces	0.471093
bsmtfin_sf_1	0.423519
lot_frontage	0.341842
open_porch_sf	0.333476
wood_deck_sf	0.326490
lot_area	0.296566
age_garage	-0.533962
age_remodel	-0.551716
age_home	-0.571881

Correlation matrix (Sale-price)



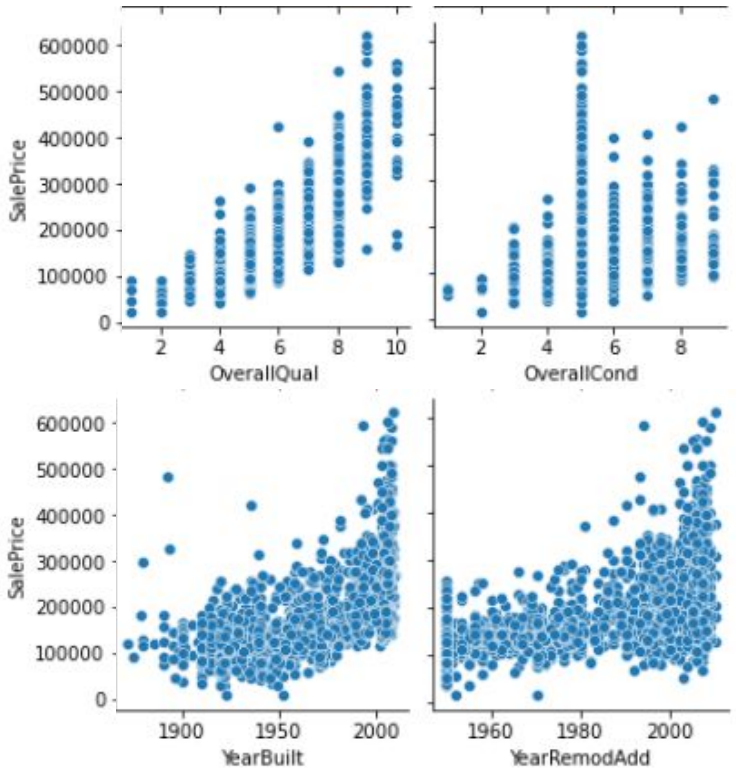
EDA Overview - Missing Value Summary

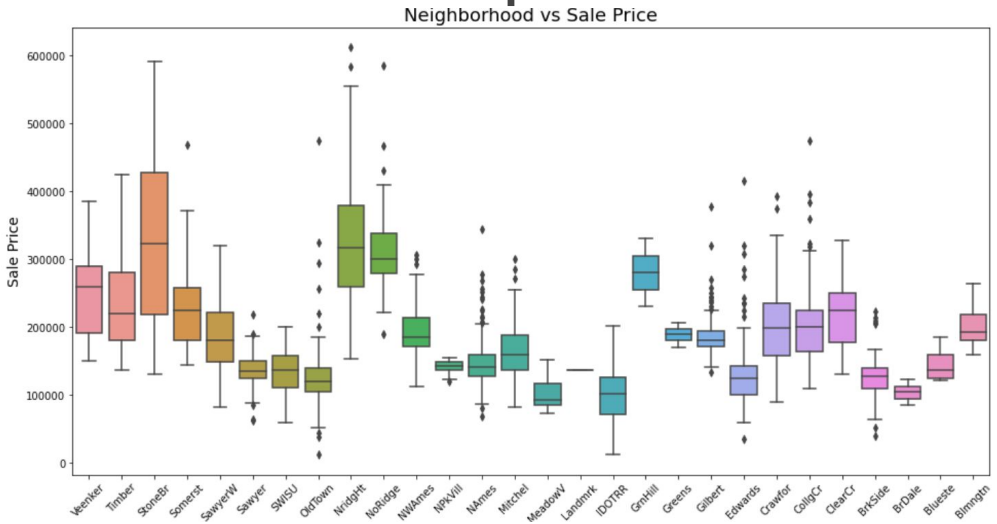
	col	num_nulls	%nulls
73	PoolQC	2042	99.561190
75	MiscFeature	1986	96.830814
7	Alley	1911	93.174061
74	Fence	1651	80.497318
58	FireplaceQu	1000	48.756704
4	LotFrontage	330	16.089712
60	GarageYrBlt	114	5.558264
65	GarageCond	114	5.558264
64	GarageQual	114	5.558264
61	GarageFinish	114	5.558264
59	GarageType	113	5.509508
33	BsmtExposure	58	2.827889
36	BsmtFinType2	56	2.730375
32	BsmtCond	55	2.681619
31	BsmtQual	55	2.681619
34	BsmtFinType1	55	2.681619
27	MasVnrArea	22	1.072647
26	MasVnrType	22	1.072647
49	BsmtHalfBath	2	0.097513
48	BsmtFullBath	2	0.097513
39	TotalBsmtSF	1	0.048757
38	BsmtUnfSF	1	0.048757
37	BsmtFinSF2	1	0.048757
62	GarageCars	1	0.048757
63	GarageArea	1	0.048757
35	BsmtFinSF1	1	0.048757

- Some columns has significant percentage of missing value like PoolQC and others

- Further investigation was done and found that they aren't exactly missing value but accurate representation of certain categorical feature not being available

EDA Overview - Bivariate Relationship

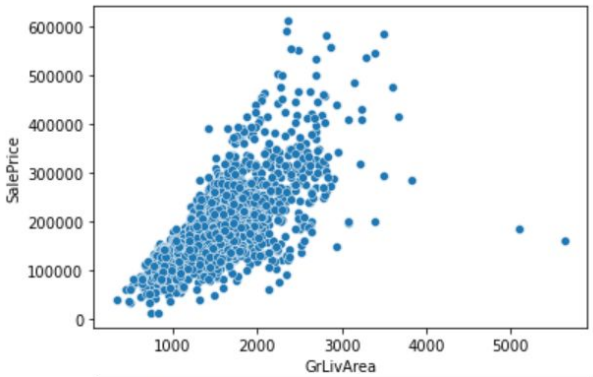


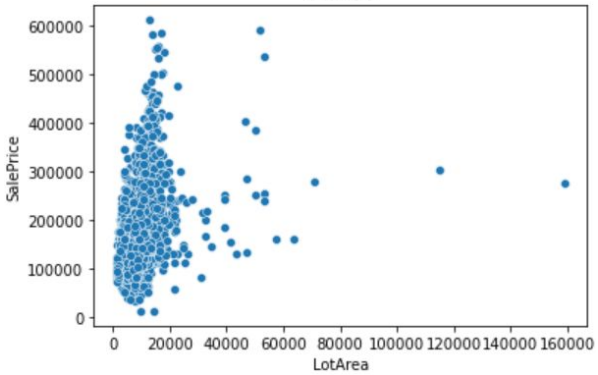


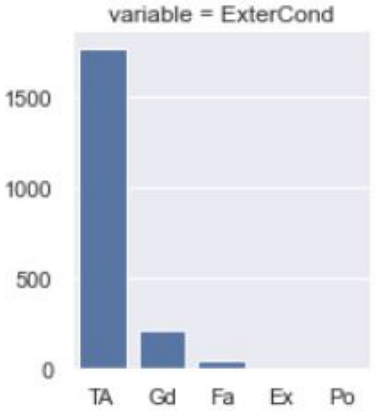
- Doing bivariate analysis, we were able to identify relationship between some columns against sale price

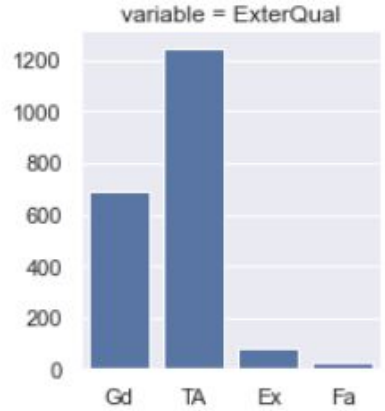
- We were able to identify interesting relationship, not only for continuous variable but also for categorical one like above

EDA Overview - Outlier & Non-Numerical Data









- We also observed outlier in some features which must be resolved to avoid problem during model training

- We have quite a number of categorical column that needs to be converted to numerical since model can only understand numbers

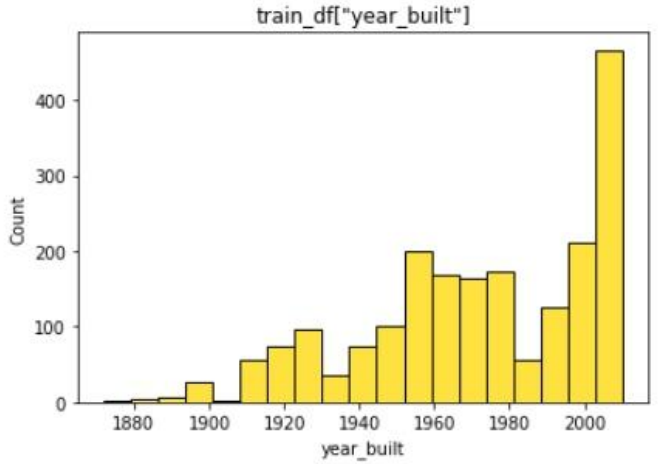
Cleaning up datasets

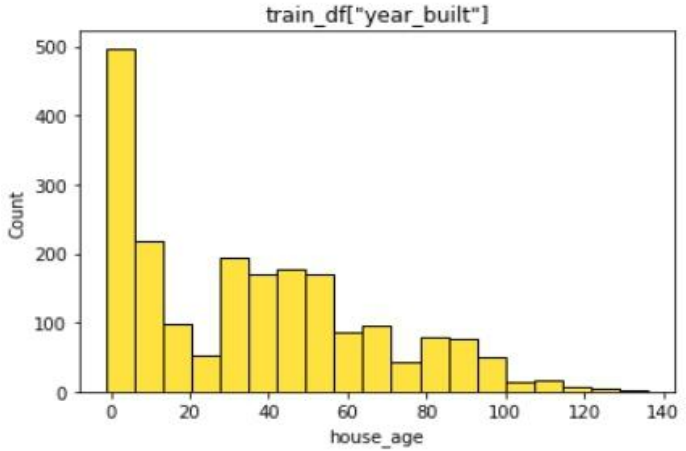
Convert features for both train and test dataset

Year_built -> house_age

year_remod/add -> reno_age

Garage_yr_blt -> garage_age



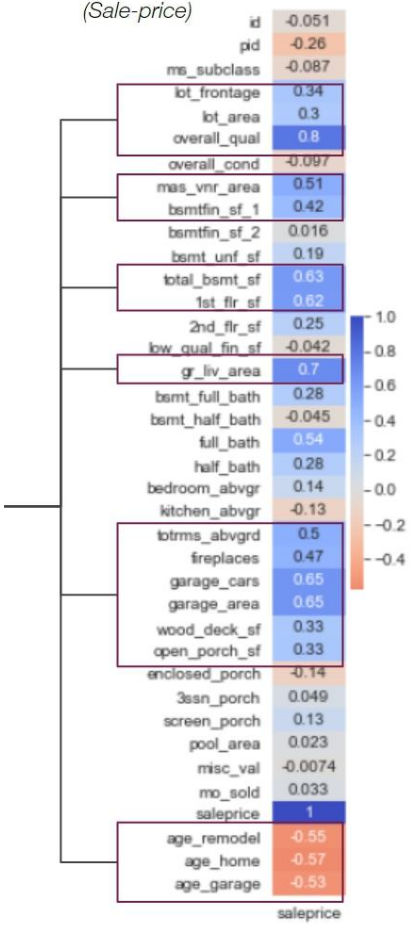


Drop Poor Correlation

Selected features

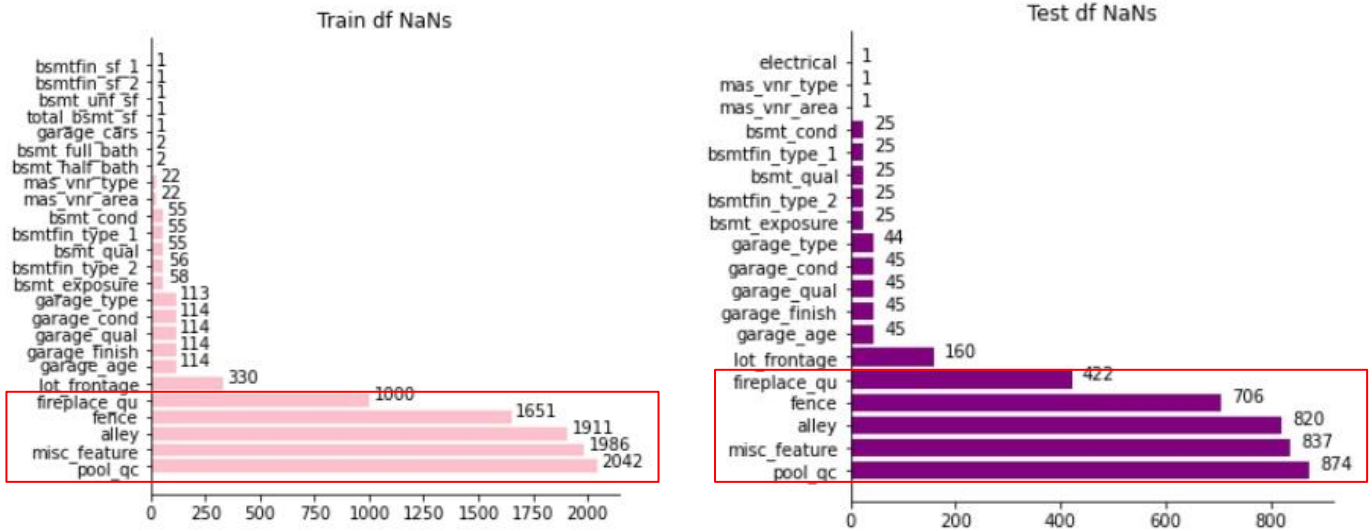
saleprice	1.000000
overall_qual	0.800207
gr_liv_area	0.697038
garage_area	0.650270
garage_cars	0.648220
total_bsmt_sf	0.628925
1st_flr_sf	0.618486
full_bath	0.537969
mas_vnr_area	0.512230
totrms_abvgrd	0.504014
fireplaces	0.471093
bsmtfin_sf_1	0.423519
lot_frontage	0.341842
open_porch_sf	0.333476
wood_deck_sf	0.326490
lot_area	0.296566
age_garage	-0.533962
age_remodel	-0.551716
age_home	-0.571881

Correlation matrix (Sale-price)



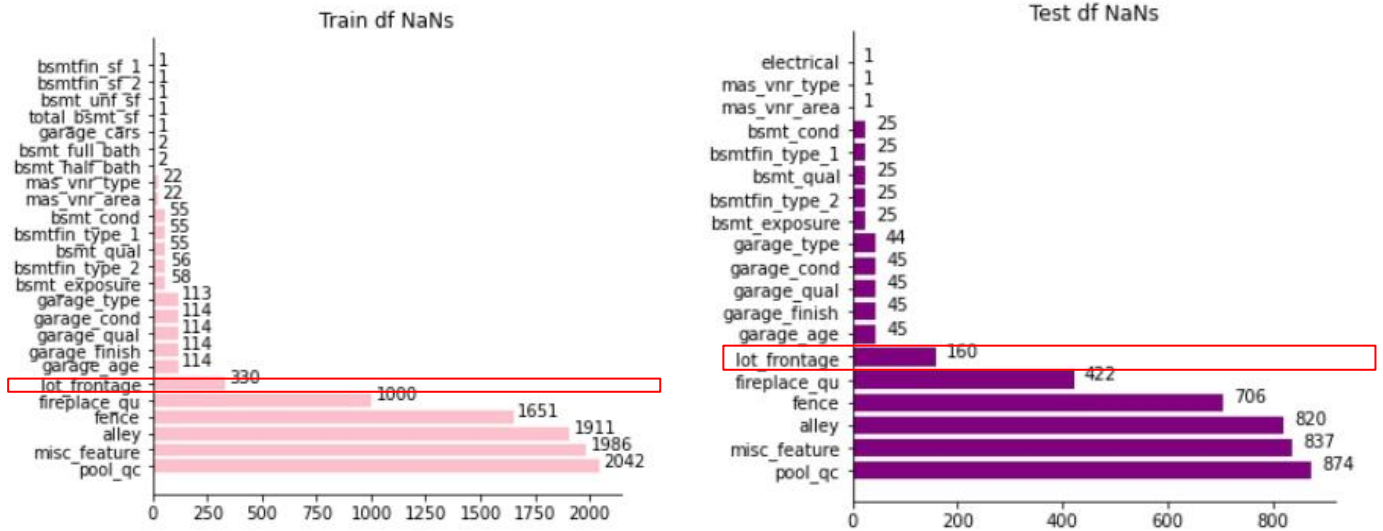
NaNs

Drop categorical features with large percentage of NaNs.



NaNs

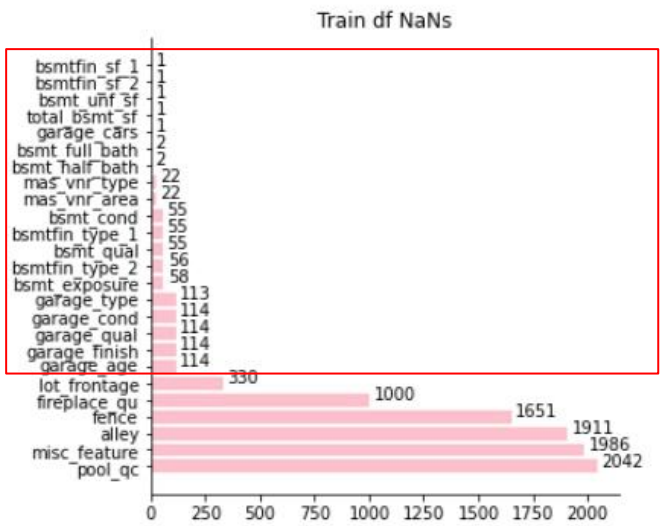
Impute Lot Frontage NaNs with mean lot frontage

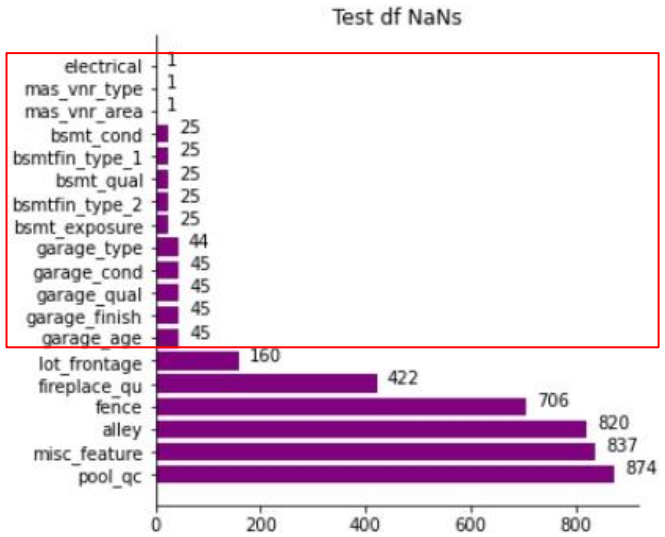


NaNs

Data description states that NaNs represent the non-existence of a feature.

Replace quantitative NaN with 0, qualitative with None.

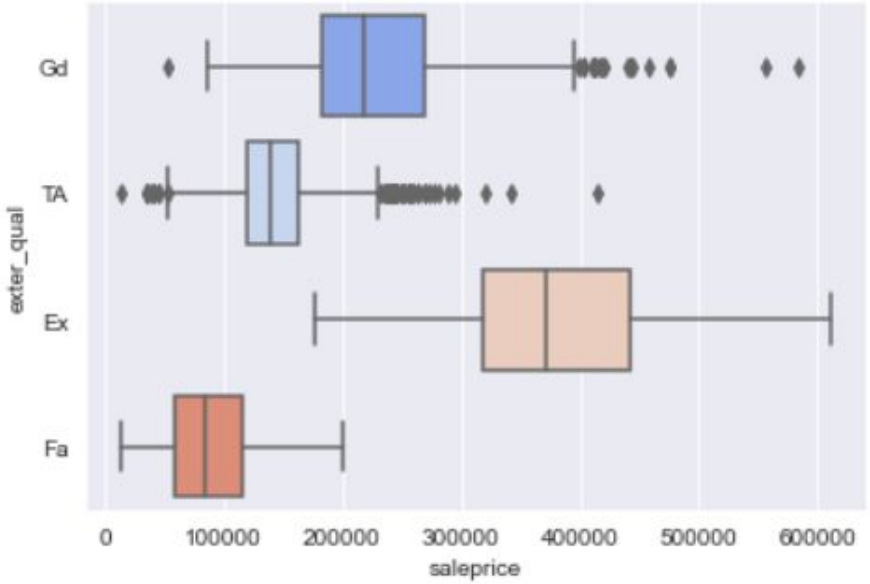




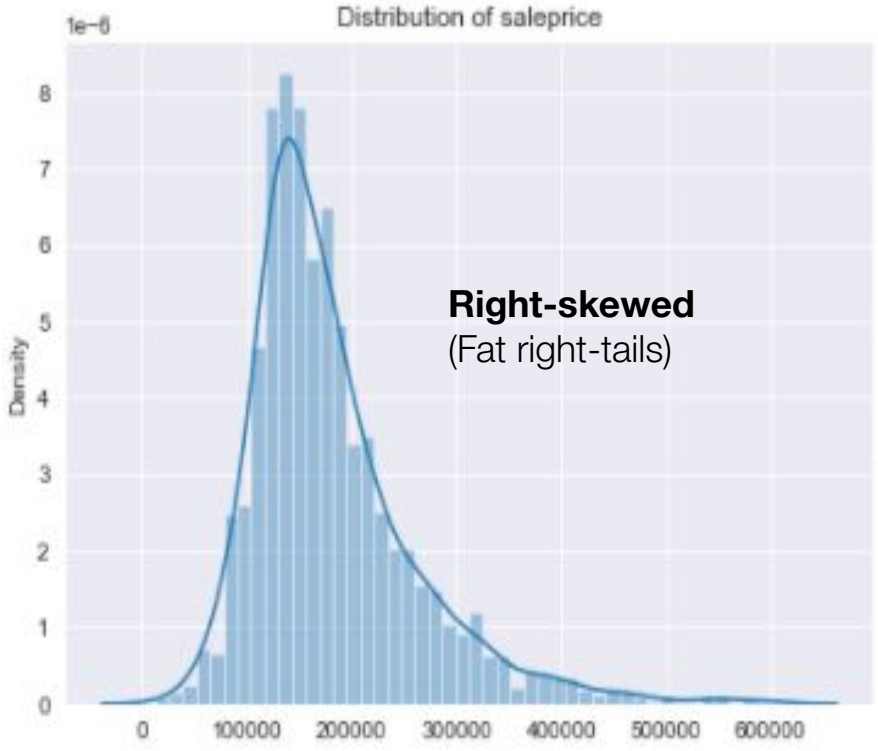
Do we account for ordinal variables?

We have found that converting ordinal features to a numerical scale (0-5) does not significantly improve the model (in terms of R-squared or RMSE).

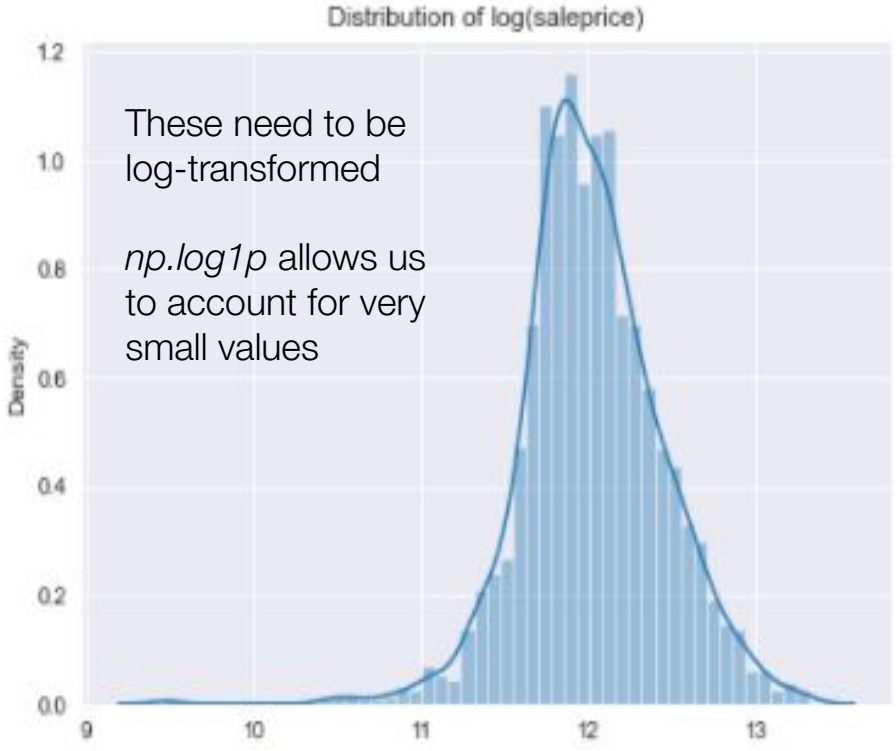
We posit that ordinal features here… do not assume a linear relationship with sale-price. One-hot encoding them may also a be suitable (perhaps more efficient) approach



Adjusting for skewness



Right-skewed
(Fat right-tails)



These need to be log-transformed

np.log1p allows us to account for very small values

Model Selection & Evaluation Metrics

Multiple Linear Regression with Regularization

- 175 Features were included in the model
- Many iterations of feature which may cause:
  - Overfitting
  - Multicollinearity
  - Computationally intensive
- Performed Linear Regression, Ridge and Lasso

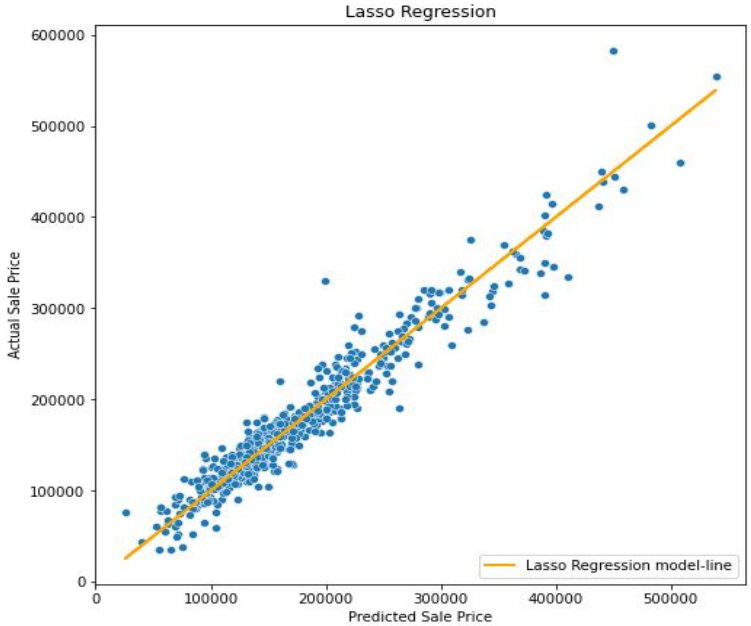
Evaluation Metrics

	Model Type			
Scoring Metric	**Simple Linear**	**LASSO**	**RIDGE**	**Baseline**
R2	**0.93**	**0.94**	**0.94**	**0**
RMSE	**19909.82**	**19574.17**	**19250.99**	**77435.05**

R2: Coefficient of Determination
RMSE: Root Mean Square Error

Lasso regularization

- Lasso can set coefficients to absolute zero compared to Ridge which never set the value of coefficient to zero.
- Helps to reduce features that are not relevant.



Top 5 Features	Positive Coefficient
Gr Liv Area	0.12
Overall Qual	0.09
Functional (Typical)	0.05
BsmtFin SF 1	0.04
MS Zoning (RL)	0.03

Top 5 Features	Negative Coefficient
Remodel Age	-0.04
Exter Cond (Poor)	-0.04
KitchenQual (TA)	-0.05
Home Age	-0.03
Kitchen Qual (Gd)	-0.03

Conclusion:

1. The model will perform best when having to predict houses within the range of $100k to $300k
2. To fetch a better sale price, we will need to maintain/improve features that have high coefficients.

Recommendations:

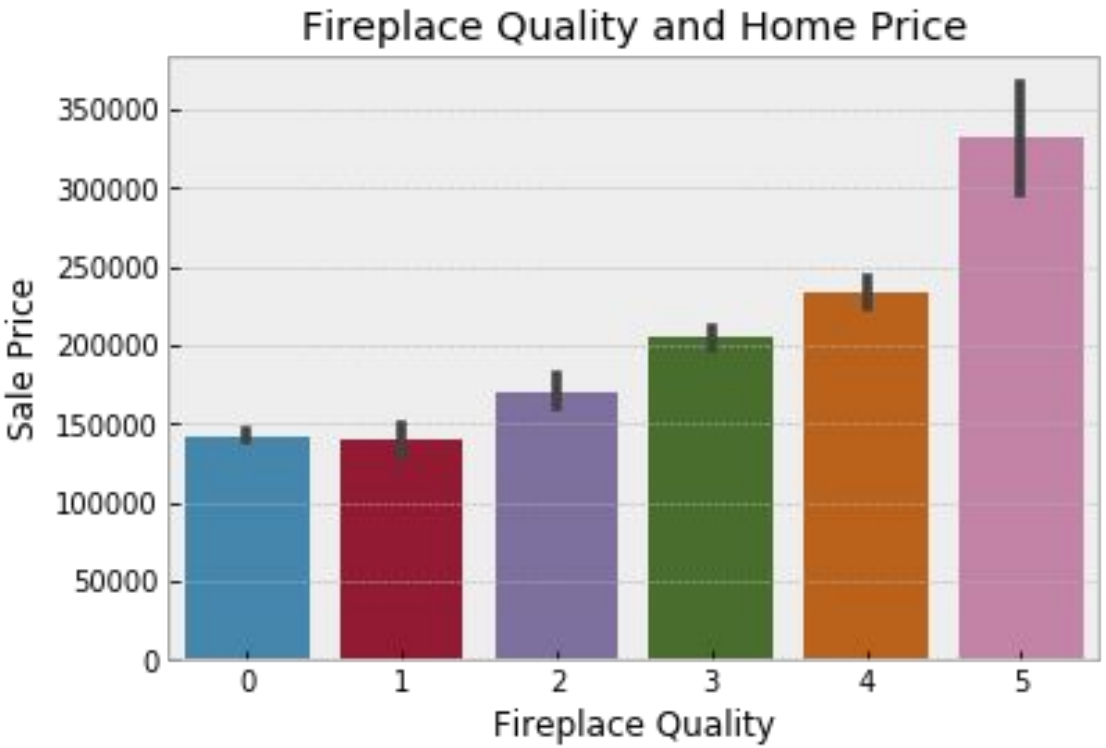
1. Distance to amenities.
2. Crime rate / accident around the neighbourhood.

Q&A

Appendix

Feature Engineering & Selection

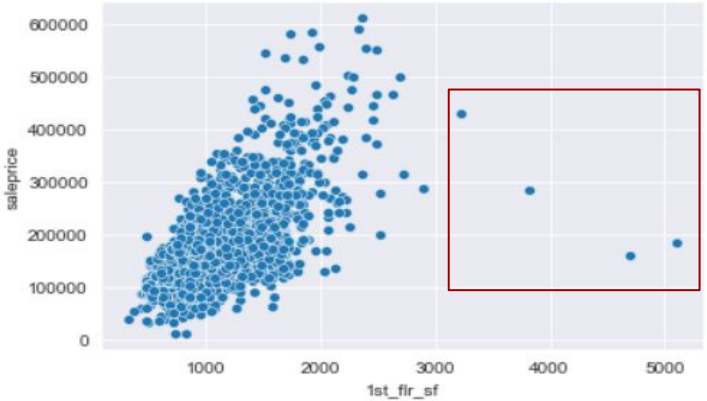
- Logarithmic Distributions
- Categorical Features Handling
- Generation of new features like “House Age”
- One-hot encoding for nominal features

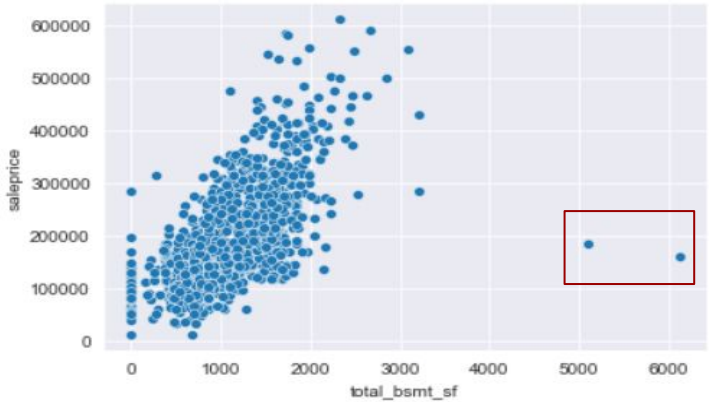


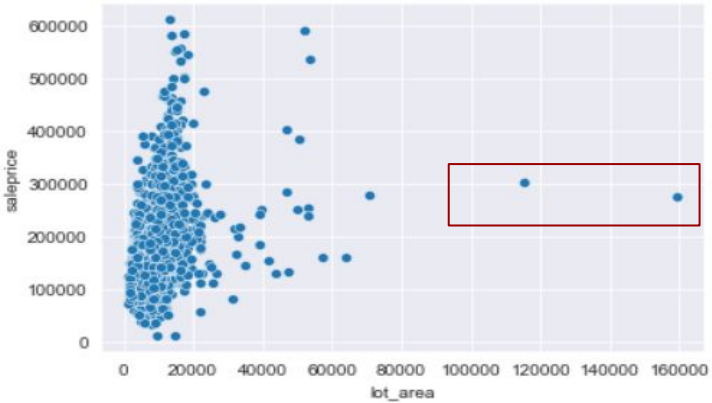
Cleaning our train data; Removing outliers

Outliers will need to be removed in order for our supervised machine learning models to work

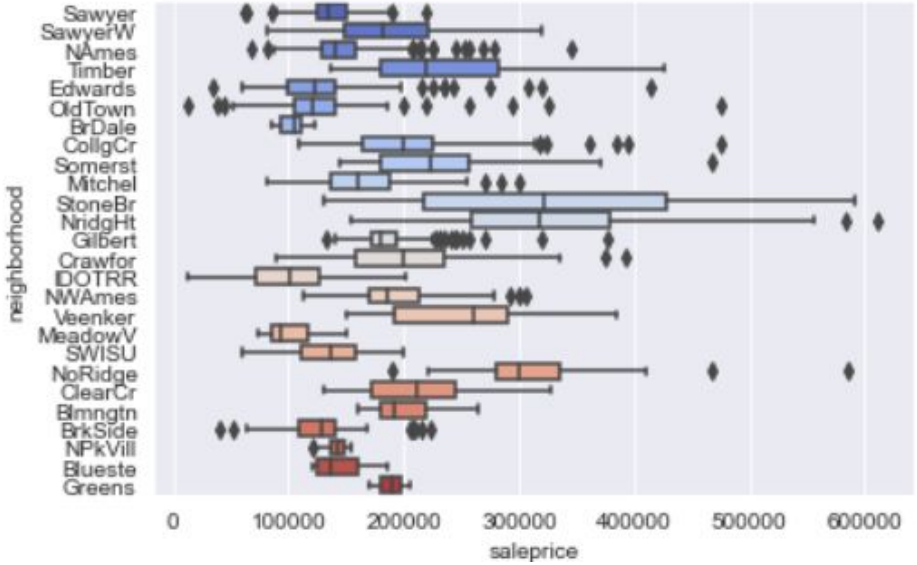
These can be inspected manually *or* filtered via kurtosis/skewness thresholds

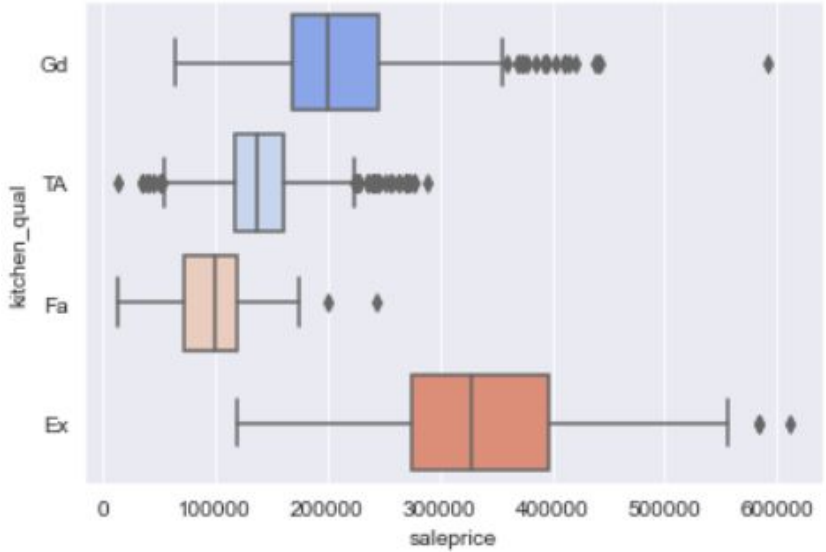




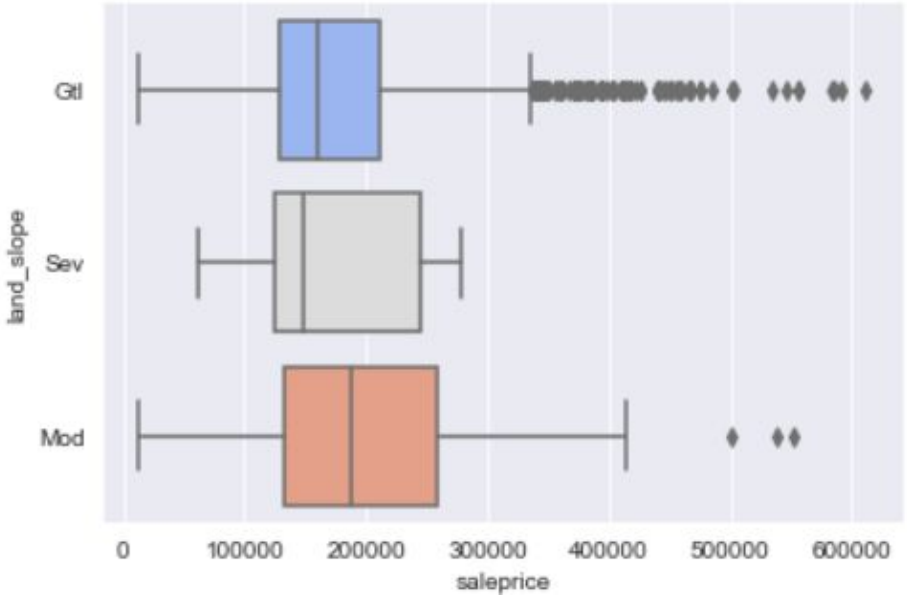


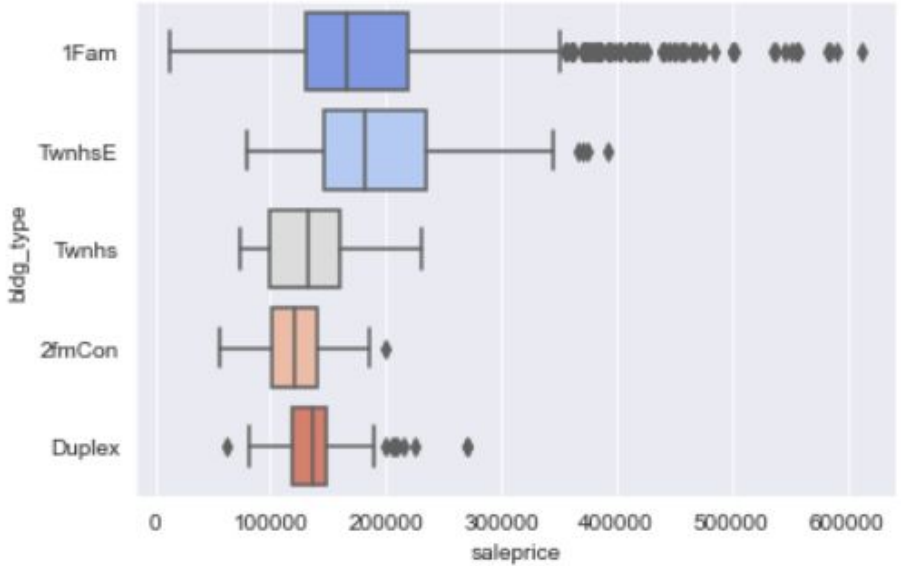
Considerable variance among some categories





And… not so much variance in some





Log-transformations and StandardScaler()

Standardization does not change the skew of the distribution. What it does is to transform the values so that the overall distribution has μ=0 and σ2=1. The shape of the actual distribution remains unchanged.

Many elements used in the objective function of a learning algorithm assume that all features are centered around zero and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

When a feature does not follow a linear distribution, it would be unwise to use the mean and the standard deviation to scale it. **Log-transformation changes the skew of the distribution, and is useful when you deal with right-skewed distributions (fat right tails).**

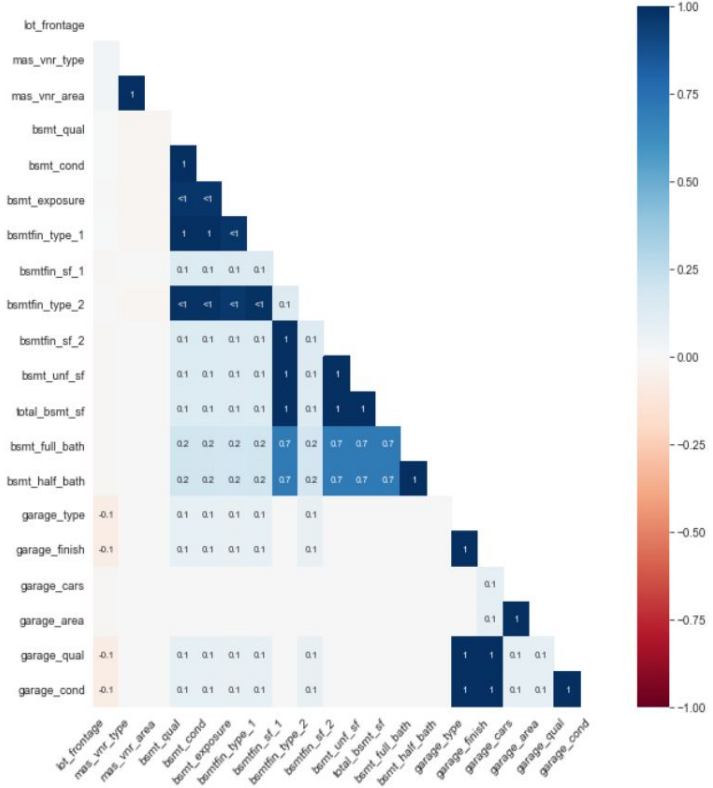
links

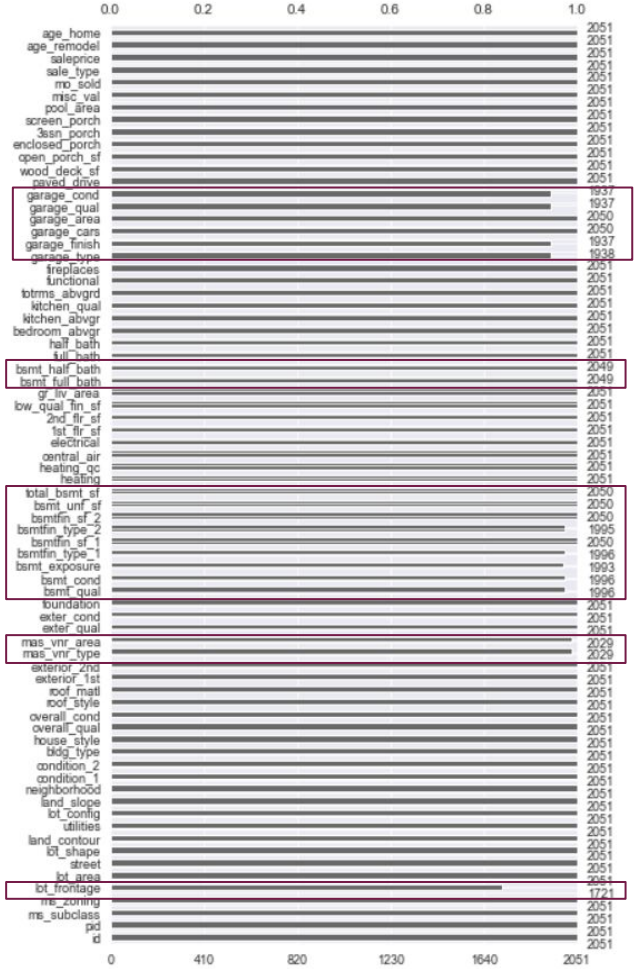
Lot frontage definition

https://www.gimme-shelter.com/frontage-50043/

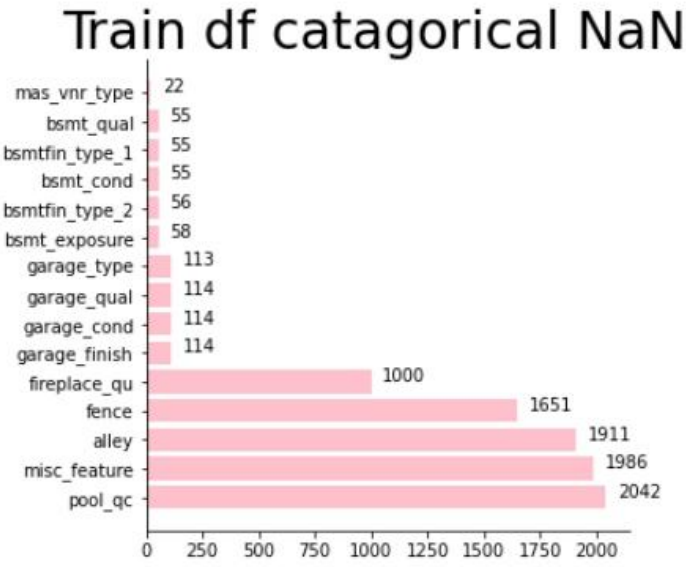
Handling missing values

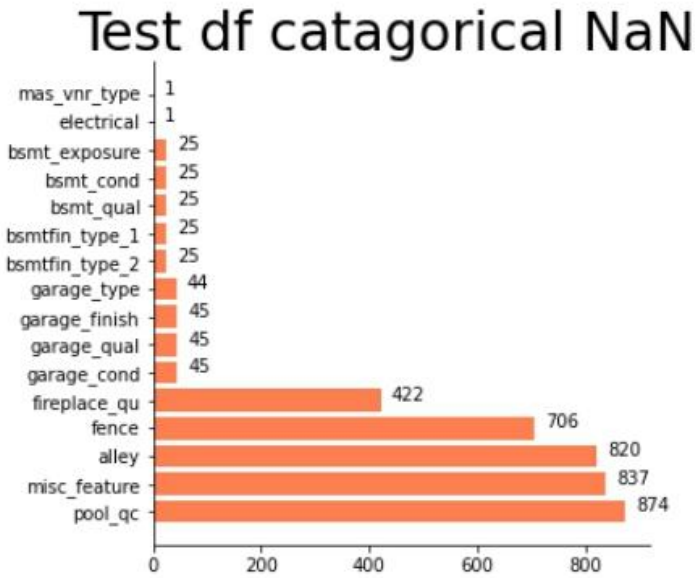
- Several quantitative and qualitative features are missing
- Some of them do not appear to be missing completely at random



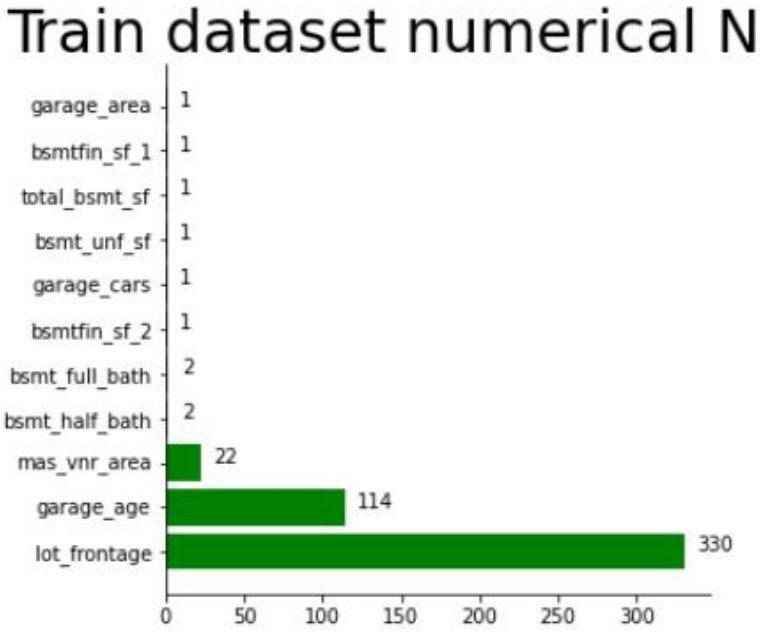


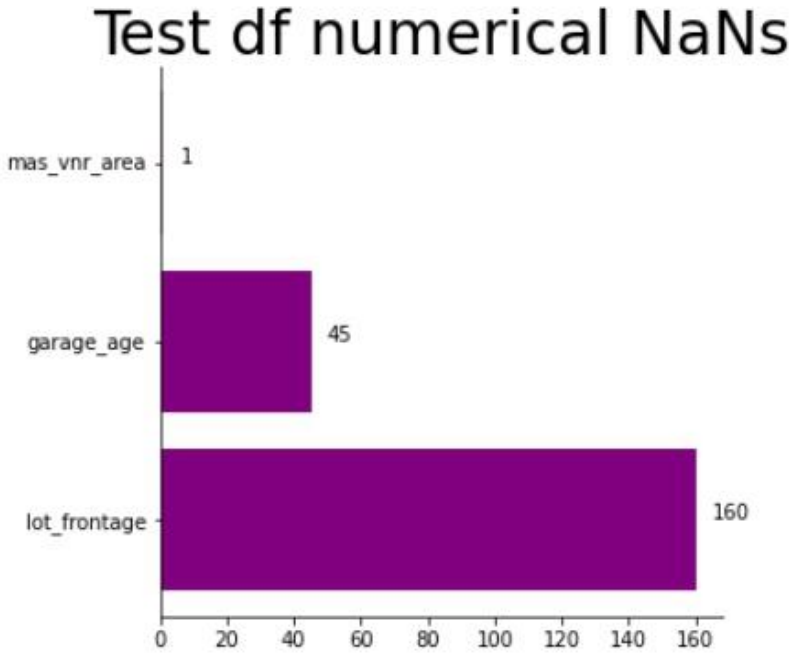
Qualitative Variables





Quantitative Variables





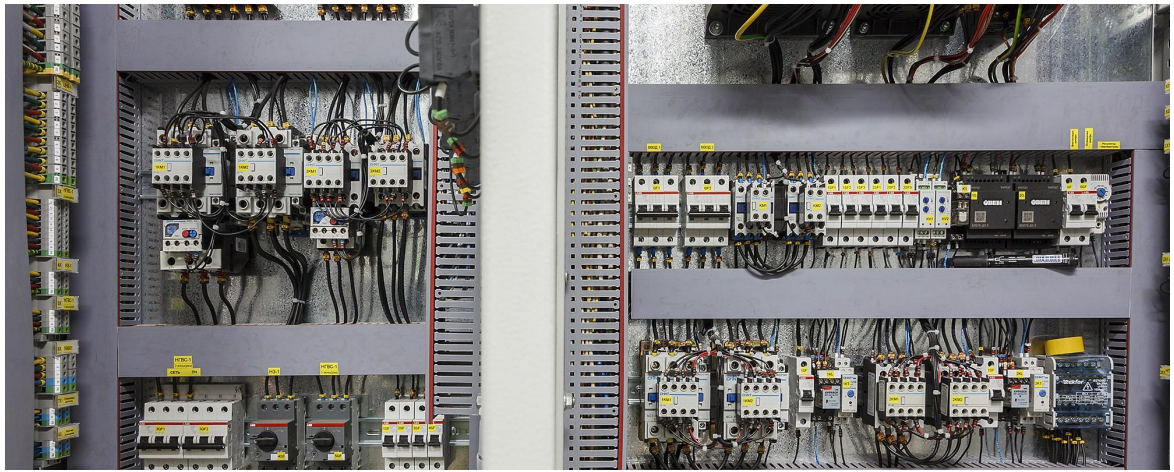
Garage Age in Train and Test dataset

	garage_type	garage_finish	garage_cars	garage_area	garage_qual	garage_cond	garage_age
28	NaN	NaN	0.0	0.0	NaN	NaN	NaN
53	NaN	NaN	0.0	0.0	NaN	NaN	NaN
65	NaN	NaN	0.0	0.0	NaN	NaN	NaN
79	NaN	NaN	0.0	0.0	NaN	NaN	NaN
101	NaN	NaN	0.0	0.0	NaN	NaN	NaN
...	...	...	...	...	...	...	...
1991	NaN	NaN	0.0	0.0	NaN	NaN	NaN
2010	NaN	NaN	0.0	0.0	NaN	NaN	NaN
2027	NaN	NaN	0.0	0.0	NaN	NaN	NaN
2039	NaN	NaN	0.0	0.0	NaN	NaN	NaN
2042	NaN	NaN	0.0	0.0	NaN	NaN	NaN

Electrical in Test Dataset

All modern houses have electricity. It does not make sense that a house does not have any electrical.

Only 1 NaN for electrical, replacing NaN with the mode will not affect the model much.



Replace NaN simply with the mode ‘Sbrkr’

Residuals Plot