Ames Housing - Sale price prediction

By Zackarias Chia, Farah Liyana, Seth Ang

Problem statement

Most banks' businesses involve real estate acquisitions and mortgages.

As being part of bank's real estate risk assessment team, we often look into property sales pricing to evaluate potential and risks of properties. The ability to predict property's sale price allows us to provide a better analysis and evaluation to risks managers and management.

We are tasked to create a regression model that provide the most accurate prediction on price of a property at sale. The model will be built using Ames Housing Dataset. Models will be fine-tuned through analysis of features utilised, type of modelling and parameters, and will be evaluated through array of scoring such as RMSE, R^2 before a final model is selected.

Methodology

EDA & Data Cleaning

1. Determine missing values and identify
2. Understand categorical values
3. Identify outliers
4. Multicollinearity
5. Log Transformation

Data Visualisation

1. Use of scatter plot for numerical data
2. Use of violin plots/bar plots for categorical data

Pre-Processing

1. Features/Output Split
2. Train/Test Split
3. Standard Scalar
4. Hyper-parameter tuning

Model

1. Ridge
2. Linear Regression
3. Lasso
4. Elastic Net

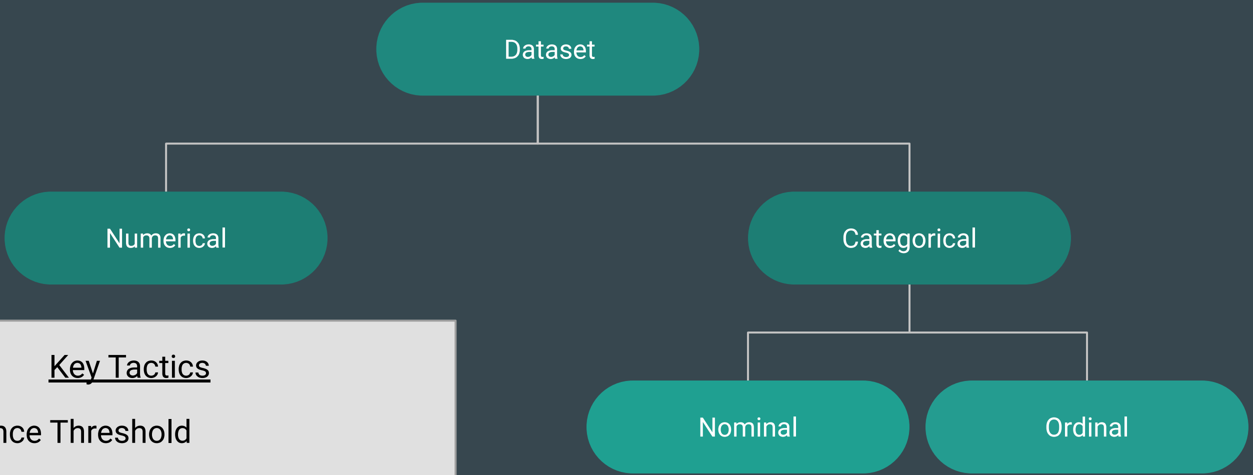
Business Recommendation

1. Model Decision
2. Highest accuracy (R2)
3. Lowest RMSE

Data cleaning

Type	Method
Null values	● Categorical: cross reference with data dictionary, impute with missing rating or mode if not available ● Numerical: impute with Median/Mean
Outliers	Drop obvious outliers
Features	Combine/drop similar features that provide similar data
Collinear features	Drop when identified via python function / algorithms

Exploratory Data Analysis (EDA)

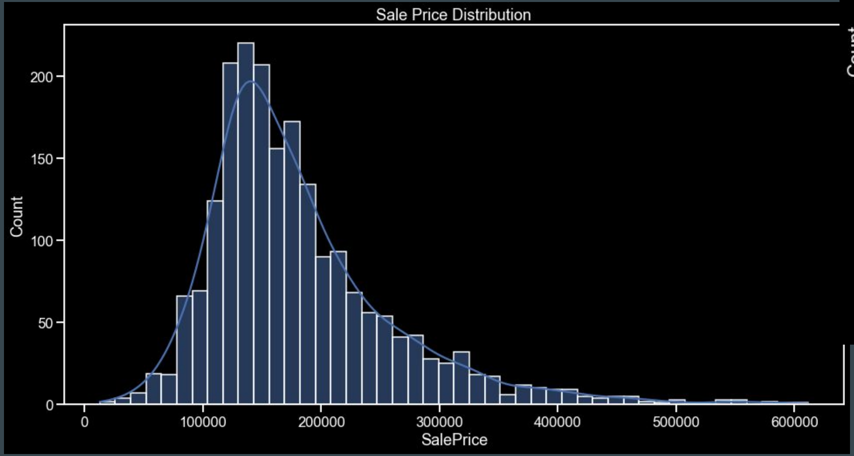


Key Tactics

- Variance Threshold
- Multicollinearity
- Null Hypothesis (p-values)/ Backward Regression
- Pearson Correlation

Exploratory Visualizations

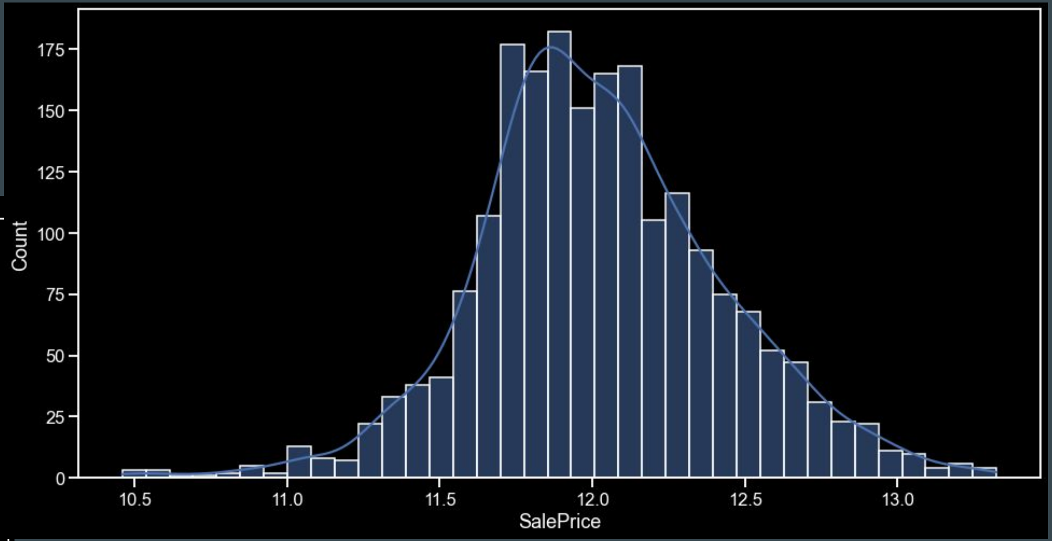
Sale Price Distribution



Before

- Not normally distributed based on the graph

After

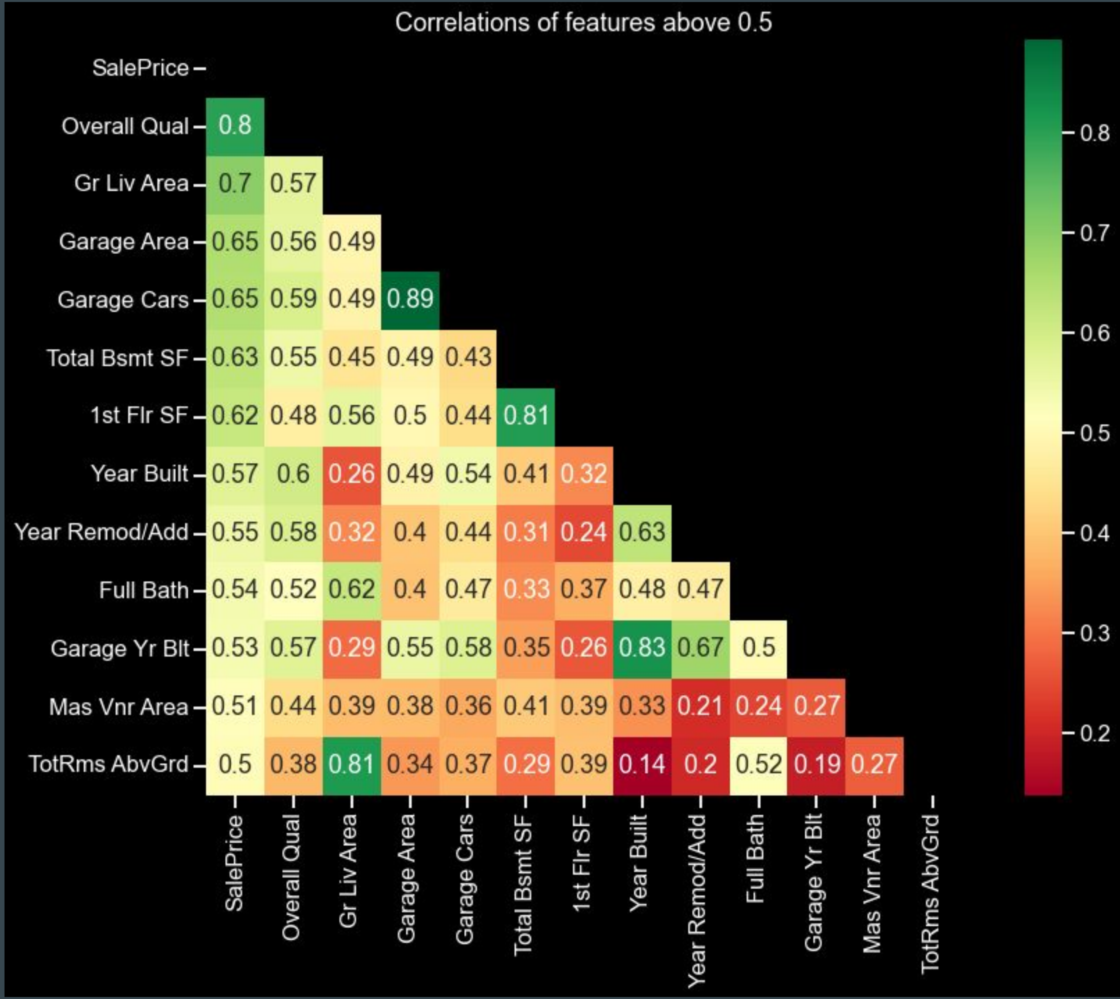


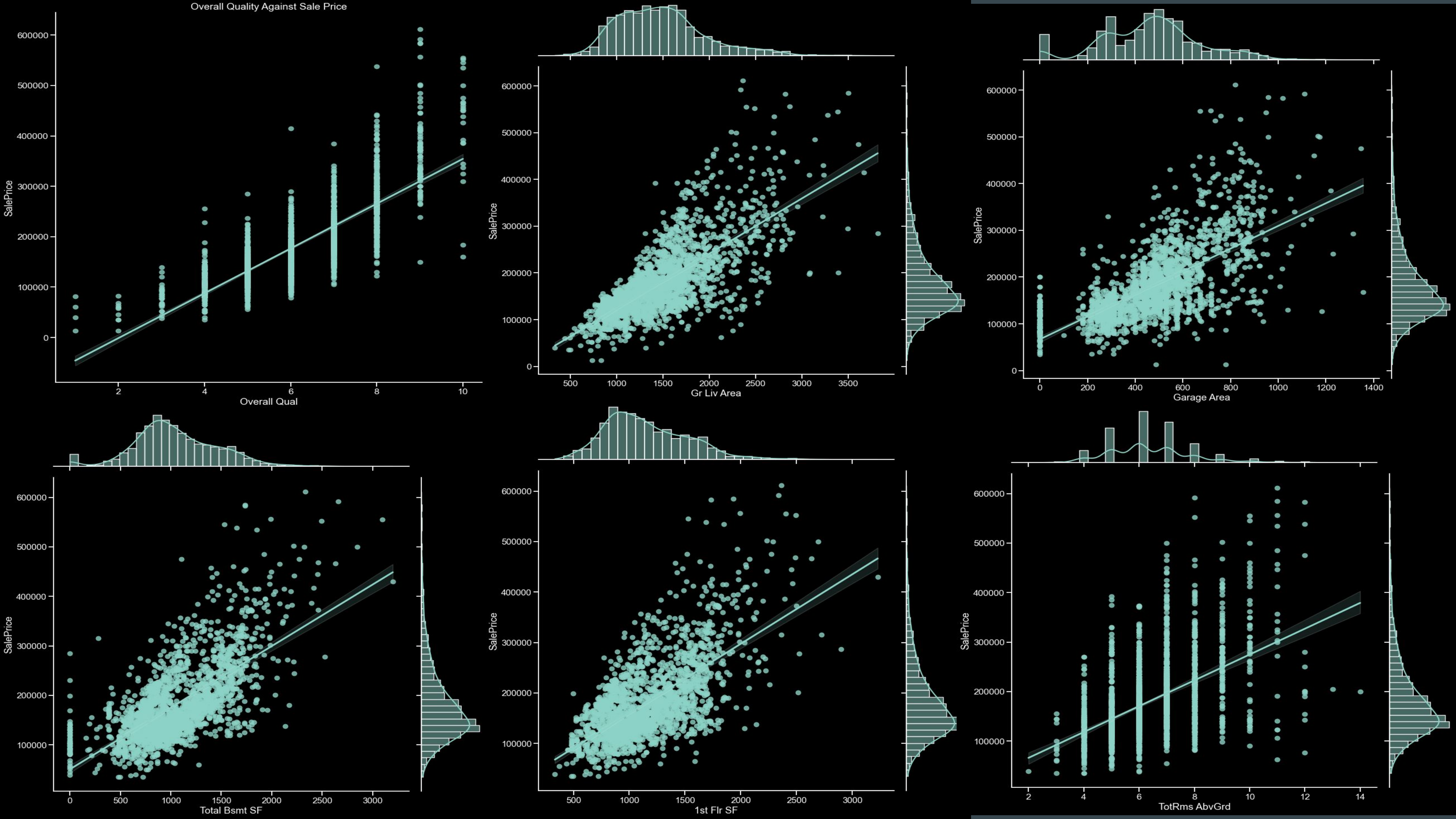
- Will log the value to make the graph more normally distributed

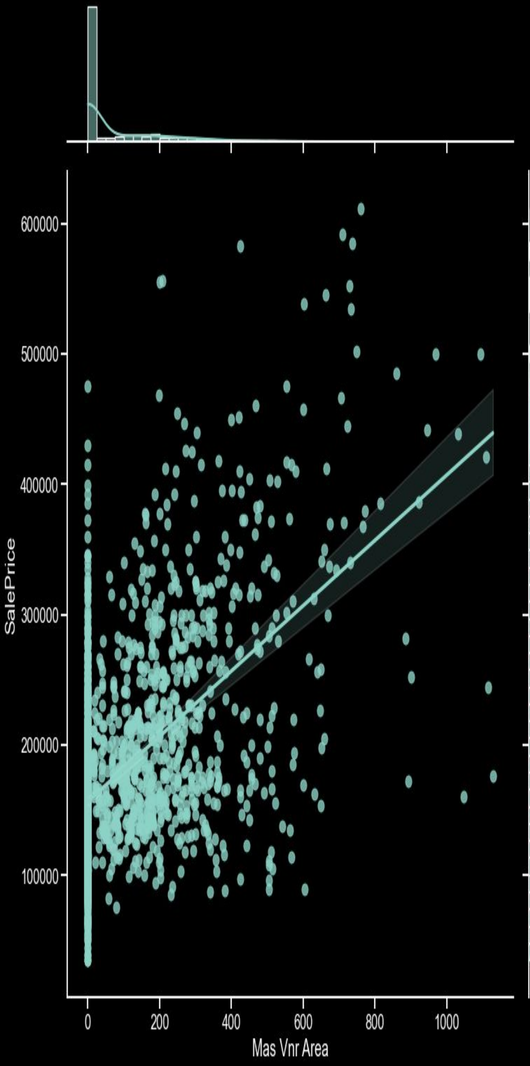
Exploratory Visualizations

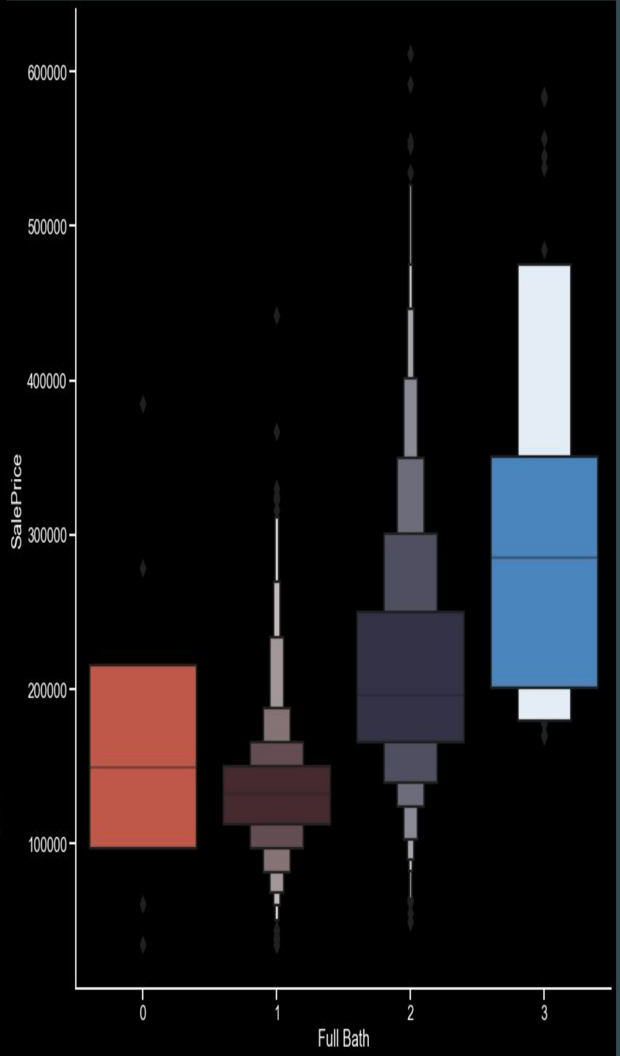
Correlation Features above 0.5

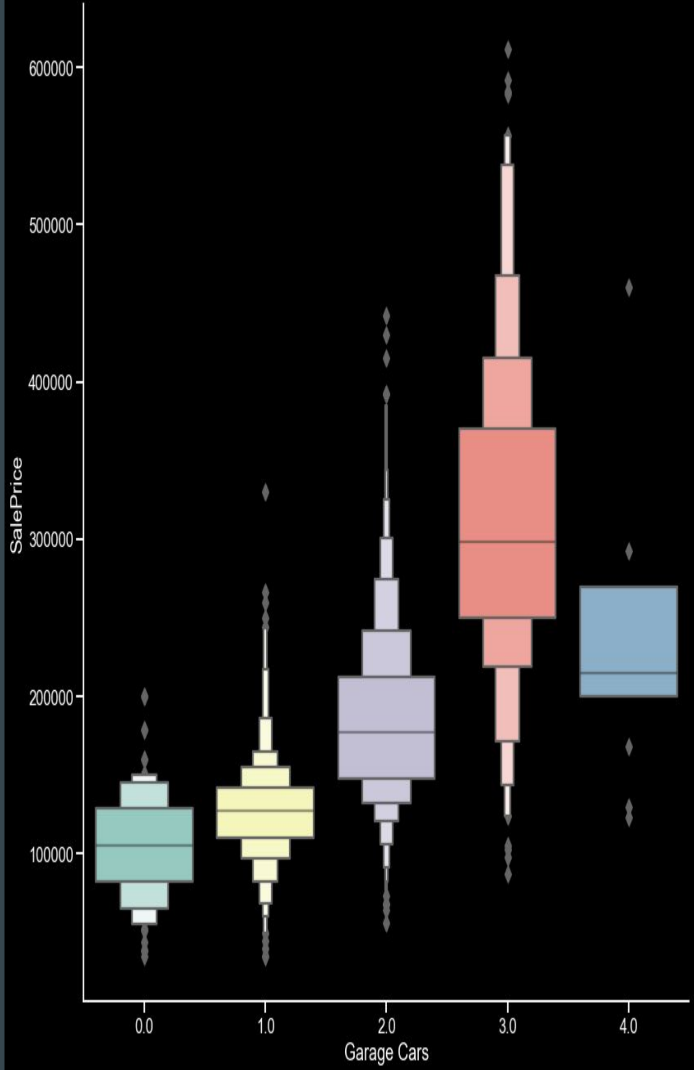
1. Overall Quality
2. Ground Living Area
3. Garage Area
4. Garage Cars
5. Total Basement Square Feet
6. 1st Floor Square Feet
7. Full Bath
8. Masonry Veneer Area
9. Total Rooms Above Ground

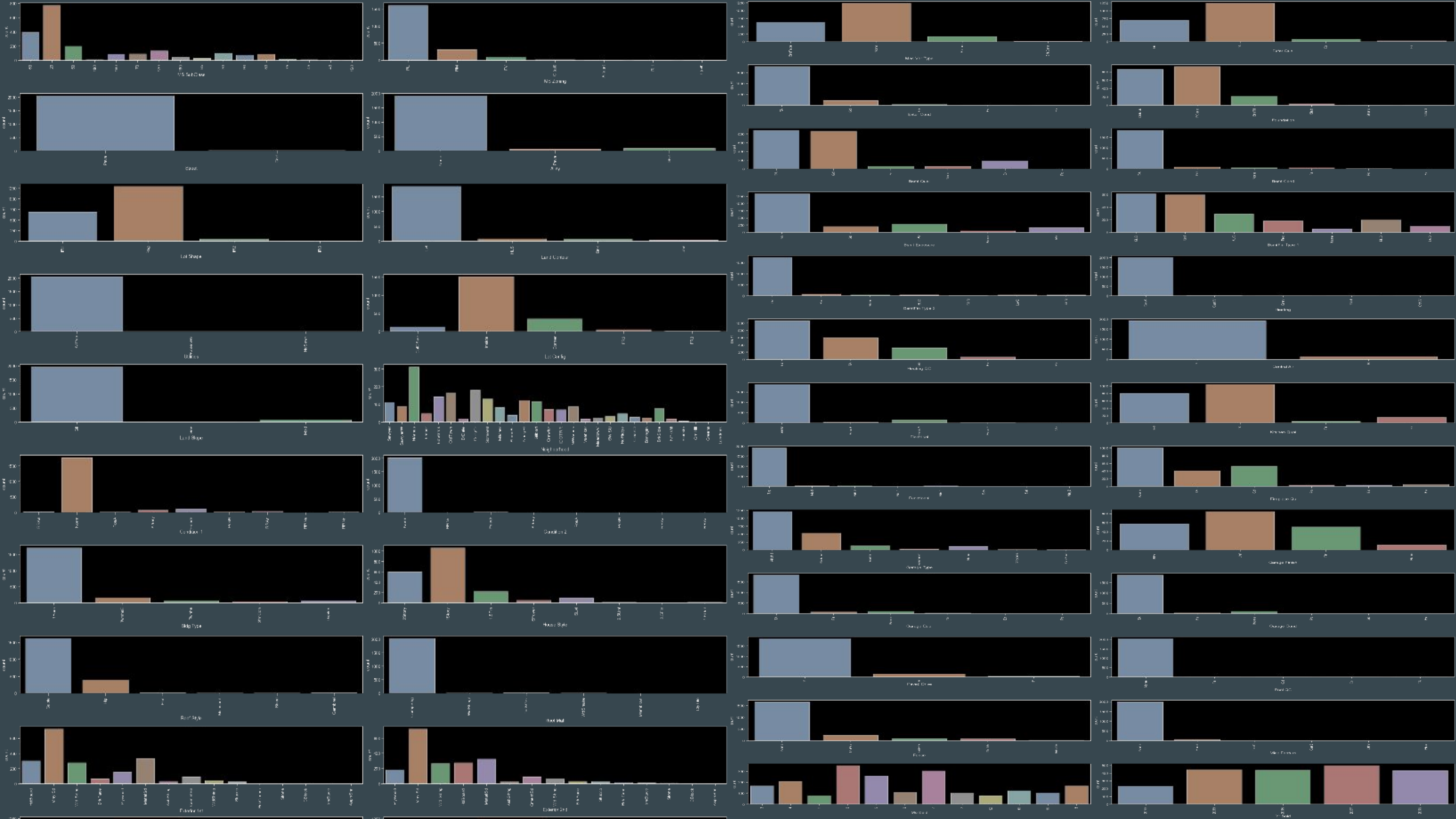



Overall Quality Against Sale Price
SalePrice
Overall Qual
Gr Liv Area
Garage Area
Total Bsmt SF
1st Flr SF
TotRms AbvGrd


SalePrice
Mas Vnr Area
0
200
400
600
800
1000
100000
200000
300000
400000
500000
600000


SalePrice
Full Bath
0
1
2
3
100000
200000
300000
400000
500000
600000


SalePrice
Garage Cars
0.0
1.0
2.0
3.0
4.0
100000
200000
300000
400000
500000
600000



Exploratory Visualizations

Categorical Variables - Drop features that are dominated by one outcome:

1. Street
2. Land Contour
3. Utilities
4. Land Slope
5. Condition 2
6. Roof Material
7. Basement Condition
8. Basement Finish Type 2
9. Heating
10. Central Air Con
11. Electrical

Pre-processing

One-hot encode categorical variables

Log transformation variables > 0.5 skew

Train/test split

Standard Scale date

Drop non-statistical significant

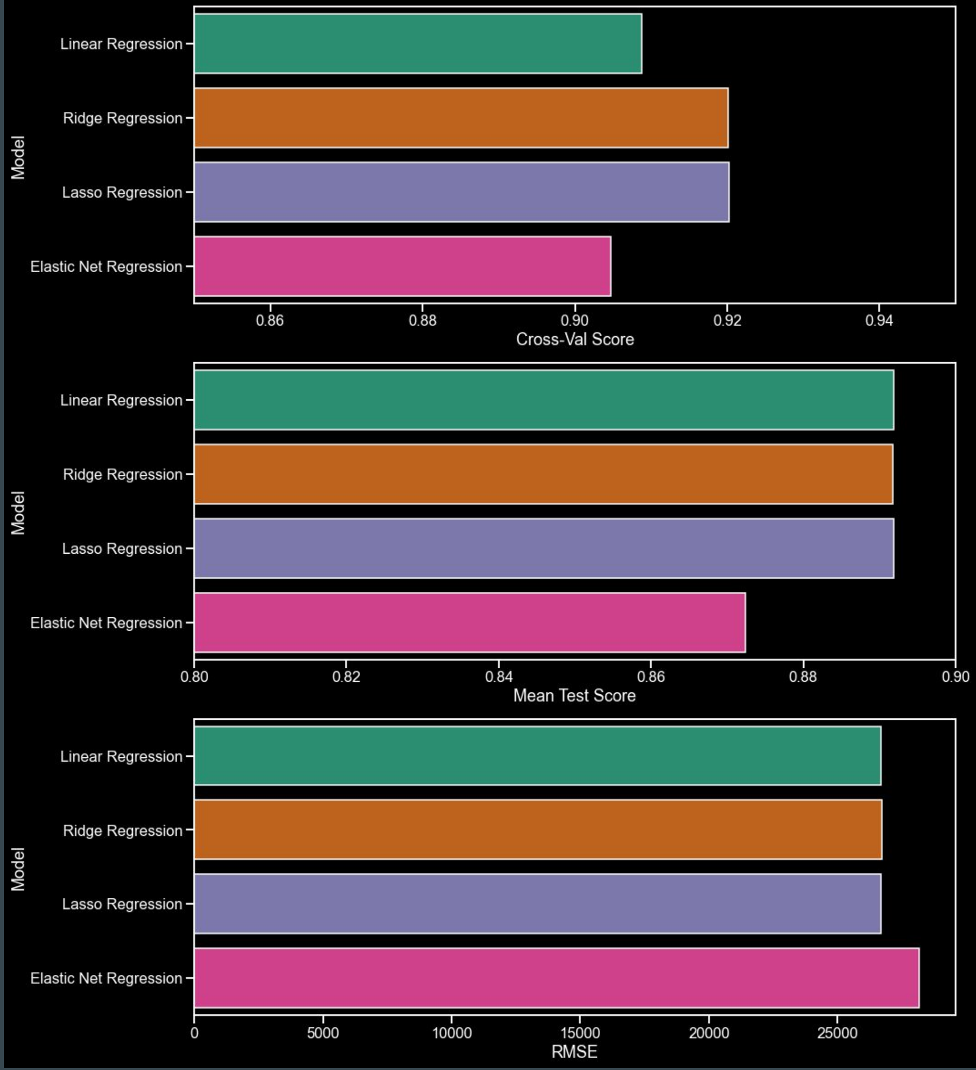
Check multicollinear data

Cook Distance to check for outliers

Review and revisit

Model Results

Model	Cross-Val Score	Mean Test Score	RMSE
Lasso Regression	0.920700	0.897334	26001.582
Ridge Regression	0.920675	0.897268	26027.659
Linear Regression	0.910649	0.897340	26000.857
Elastic Net Regression	0.906654	0.874658	27889.666



Top 5 Coefficients:

Top 5 Positive Features:

1. Ground Living Area
2. Total Basement Square Feet
3. Overall Quality
4. Lot Area
5. Kitchen Quality Excellent

Top 5 Negative Features:

1. NAmes Neighborhood
2. OldTown Neighborhood
3. Edwards Neighborhood
4. CollgCr Neighborhood
5. Gilbert Neighborhood

Feature	Coefficient
Top Positive Features	
Gr Liv Area	18625
Total Bsmt SF	11317
Overall Qual	9759
Lot Area	9512
Overall Condition	7173
Top Negative Features	
Neighborhood_Gilbert	-12526
Neighborhood_CollgCr	-12549
Neighborhood_Edwards	-13384
Neighborhood_OldTown	-14394
Neighborhood_NAmes	-18210

Model recommendation

Objective: Best prediction of property sale price

Criteria: Highest accuracy (R^2 score) with the lowest RMSE

Model chosen: Lasso regression with standard scaling (Z-score)

Shortlisting based on above criteria ensures chosen model provide best predict property value, allowing bank give a safe valuation for closest mortgage loan or real estate acquisition consideration.

Business recommendation

Contributing features:

- Gr Liv Area - Above grade (ground) living area square feet

For every unit increase in living area square feet, sales price is predicted to increase by ~$18,000

- Overall Quality

For every unit increase in Overall Quality, sale price is predicted to increase ~$11,500

- Neighborhood

Properties in certain neighborhoods have positive or negative impact on sale price.

E.g. properties in Northridge Heights increase sales price, while properties in Edwards Neighborhood lowers

Limitations

- Limited to features and scaling used in model such as neighbourhood, type of housing
- Model is feasible on assumptions that future environment remain constant/same, i.e. national real estate regulations, economic status at time of sales.
- Sales price can change drastically should there be changes in macroeconomy.
- Accuracy is limited to type of modelling and parameters used.

  There may be better machine learning models or hypertuning methods available currently and in future that can provide better prediction.