Project -2:
Predicting Ames Housing Prices

GENERAL ASSEMBLY

Sileshi Hirpa Dec. 03/21

Content

- Problem statement
- Project overview
- EDA of some features
- Model comparison
- Interpreting a model
- Conclusions/Recommendations

Aim: Identifying features that affect housing market

Problem Statement:

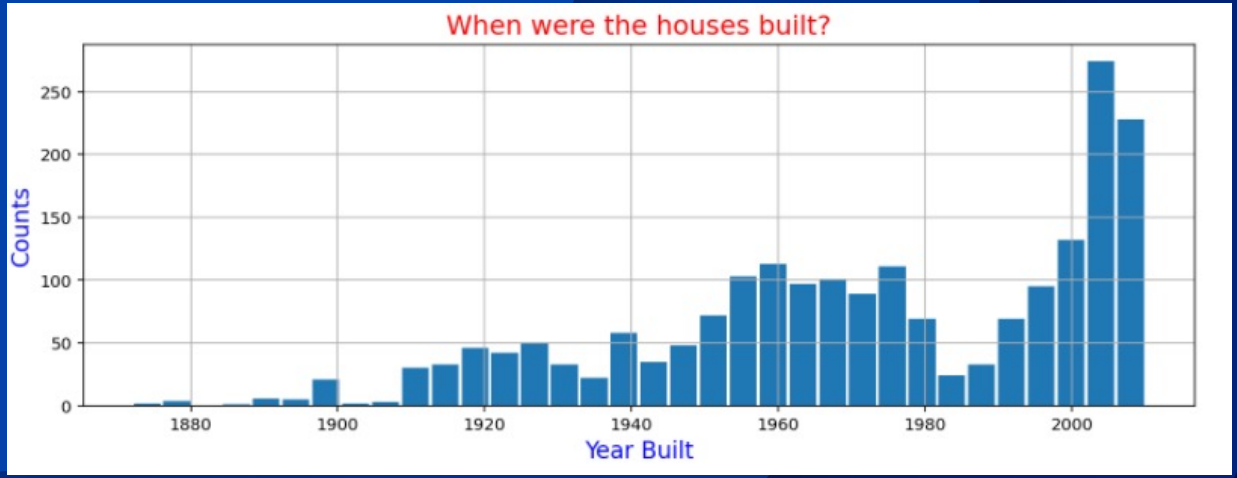
Creating BEST models based on the Ames Housing Dataset which will predict the price of a house at sale in Ames, Iowa.

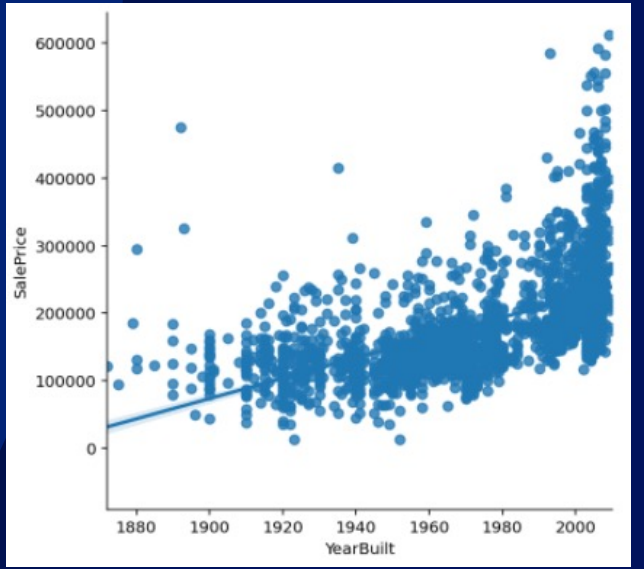
Although individuals who are interested in selling/buying house can benefit from this, the main target audience for my project will be property agents (real estate firms in Ames).

Methodology (Project overview)

Step	Description
1. Obtain Data	Ames Housing Data (2051 homes, each with 81 features)
2. Clean Data	Null Values, Data Types, Values within expected range
3. EDA	Focus on relationship between features and target (Sale Price)
4. Feature Engineering	Informed by EDA
5. Model Preparation	Split the dataset into train & Test where, test size = 20%
6. Modeling	Linear Regression, Ridge Regression, LASSO Regression
Metrics comparison	Mean squared error and the R^2 metrics were used for comparison

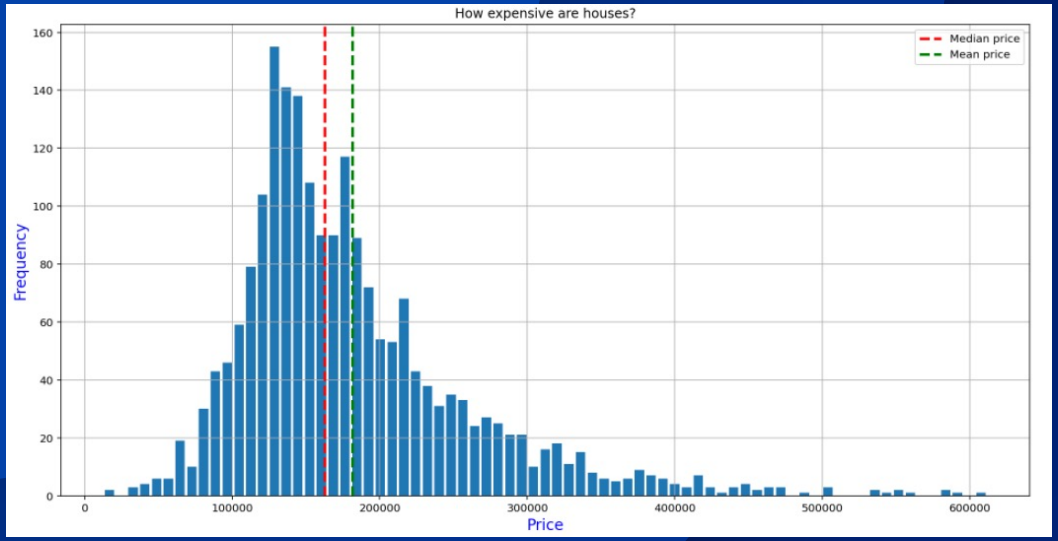
Some of the features:

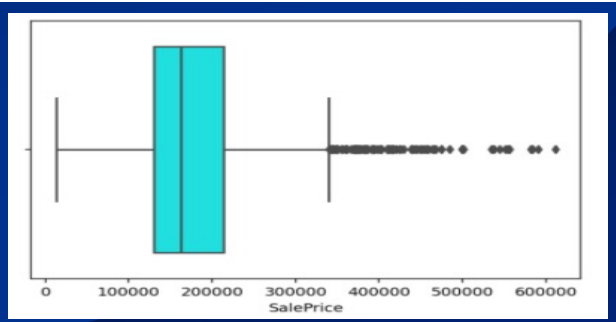




- The oldest house was built in 1872
- The newest was built in 2010

(Cont'd ..Some of the features)






```
train_data['SalePrice'].describe()
```

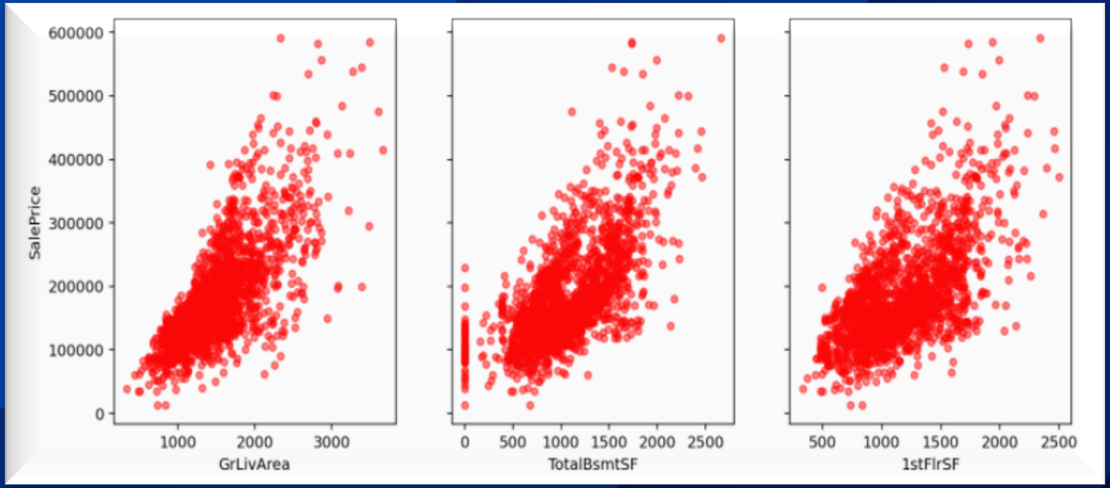
count	2048.000000
mean	181484.252441
std	79248.657891
min	12789.000000
25%	129837.500000
50%	162500.000000
75%	214000.000000
max	611657.000000

Mean price

Q_3 price

- Cheapest house is around $13K
- The most expensive one is around $610K

features EDA (cont'd)

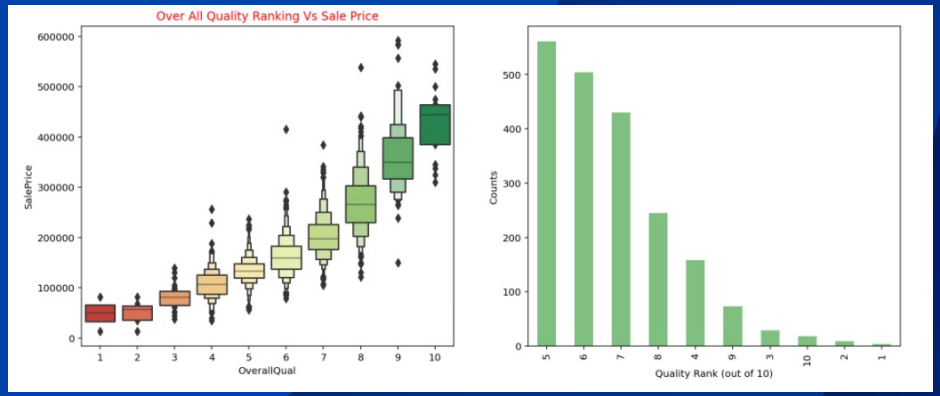


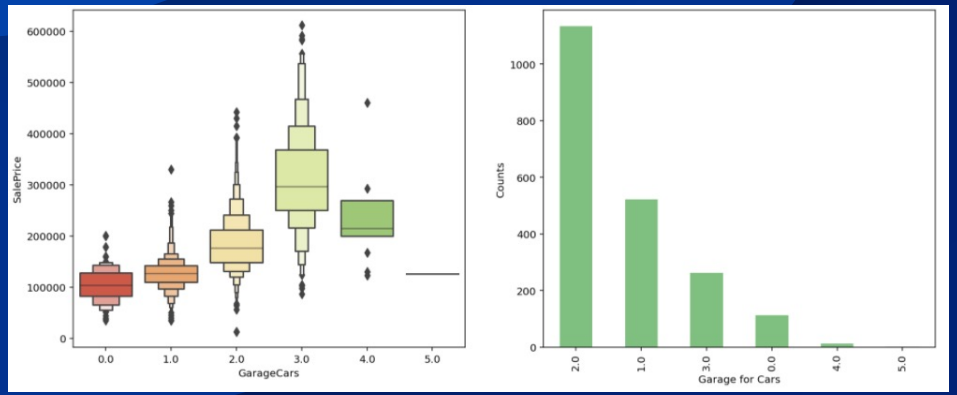
Above grade (ground) living area square feet

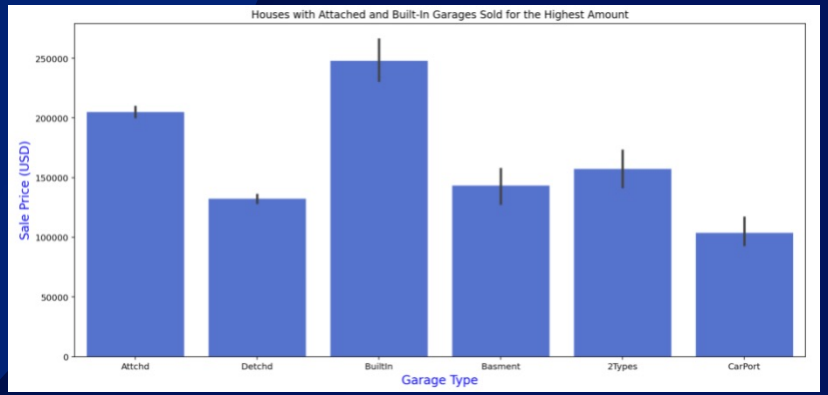
Total square feet of basement area

First Floor square feet

features EDA (cont'd)







Built-in car garages have more price

Model Evaluation

Model Name	Training Score	Testing Score
Linear Regression	94.6%	91.62%
Ridge	91.73%	90.05%
LASSO	88.21%	85.77%

- 91.62% of the variability in sale price can be explained by the house features in our model

The top(best) 5 coefficients(factors) of the Linear regression model in descending order.

	features	coefficients
139	RoofMatl_WdShngl	71240.891003
203	GarageType_None	50179.054036
207	MiscFeature_Othr	48339.020695
90	Condition2_PosA	42571.059112
179	SaleType_Oth	26683.848571

	Feature	Coefficient
113	Neighborhood_dummies_Edwards	-35308.300156
141	Exterior1st_BrkComm	-38622.116798
171	Heating_OthW	-38854.642635
91	Condition2_PosN	-41502.157870
136	RoofStyle_Shed	-44589.583101

The worst 5 coefficients(factors) of the Linear regression model in descending order.

Interpreting the model

RoofMatl_WdShngl: 71240.891003; where RoofMatl_WdShngl is the Roof material(=wood shingles)

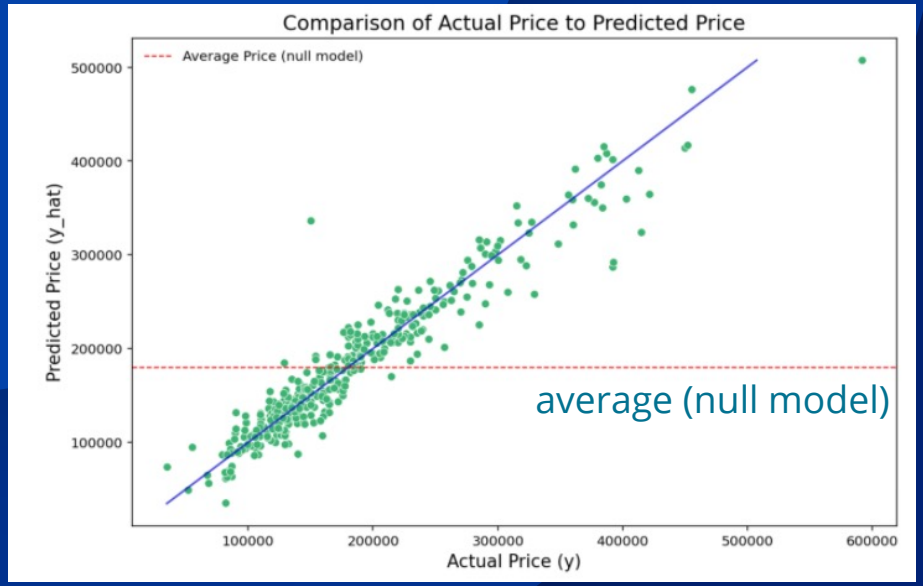
$$\hat{y} = \beta_0 + \beta_1 * X + \varepsilon$$

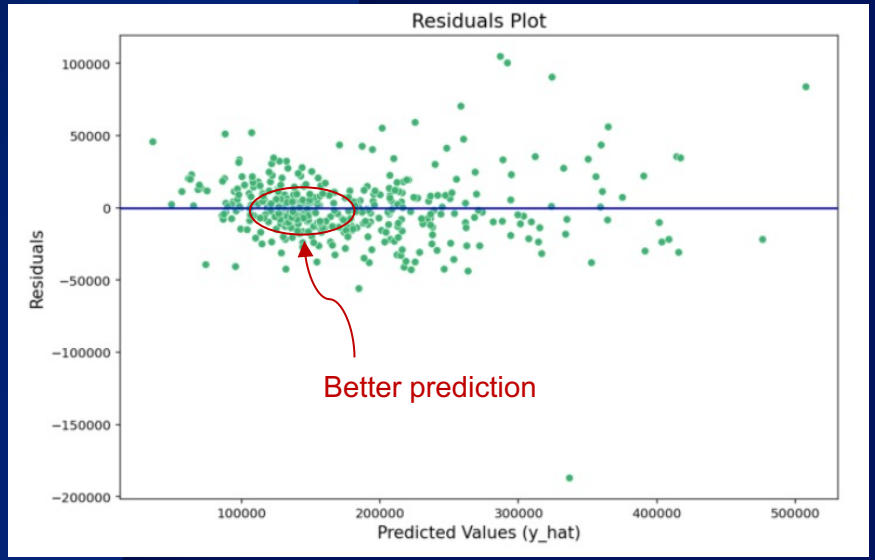
$$\widehat{SalePrice} = lr * intercept\_ + \beta_1 * [house\ features]$$

$$= -5282825.849045787 + 71{,}240.89 * [\text{RoofMatl\_WdShngl}]$$

- Holding all else constant, for every 1 unit ($100\ sqft$) increase in RoofMatl_WdShngl, the price of the house increases approximately by $71,240.

Train dataset Vs residue(difference from actual value) evaluation





This model seems to be good at predicting the price of properties within the range: $\$130K \leq \hat{y} \leq \$214K$.

- *The model's poor performance for* $\hat{y} \leq \$130K$ *and* $\hat{y} \geq \$214K$.
- *Residuals do not show equality of variances, so we need more training data for the extrems*

Conclusion and Recommendations

- Each feature can affect sale price of a house in Ames, IA.
- One can also identify which features exert the most influence on our predictions, and we even have the relative magnitude and direction of their influence, as given by the associated coefficients.
- Linear regression model performed better among the three models.
- Real estate firms or individuals who want good returns from their home investment can focus on improving worst house features identified in this project

- While this model generalizes well to the city of Ames, it's probably not generalizable to other cities, since each city/state differ greatly in terms of external factors like geographical features, seasonal weather or the economic climate of that particular city.
- Another point to keep in mind that **this model doesn't consider the inflation of housing prices**.
- Our model would need significant retraining to predict the current house prices in Ames.
- More feature engineering and deploying more robust predictive models improves predictions of the dataset.

Thank You