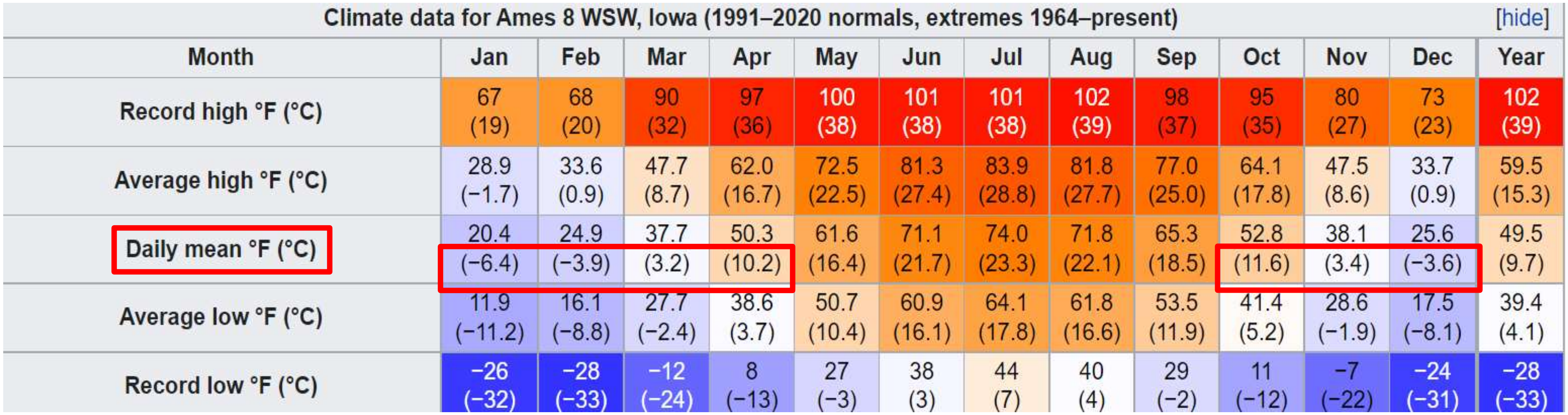
Price Prediction for Ames Iowa Housing Kaggle Dataset

CHANTANEE MANONOM

Background

- Population of **66,000** in 2020 (around half are university students)
- Low temperature climate area



Climate data for Ames 8 WSW, Iowa (1991–2020 normals, extremes 1964–present) [hide]

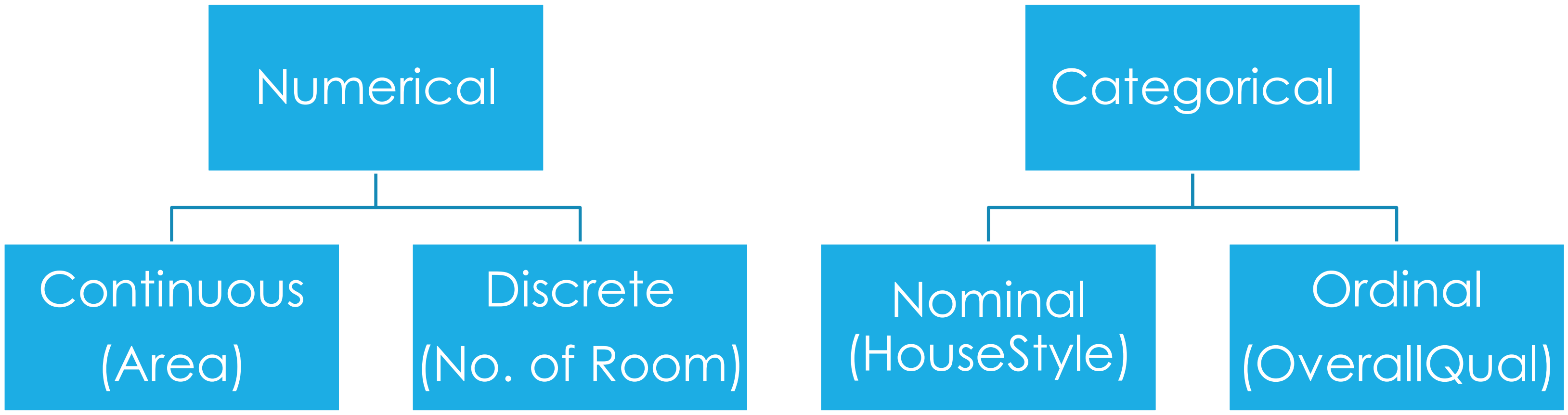
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °F (°C)	67 (19)	68 (20)	90 (32)	97 (36)	100 (38)	101 (38)	101 (38)	102 (39)	98 (37)	95 (35)	80 (27)	73 (23)	102 (39)
Average high °F (°C)	28.9 (−1.7)	33.6 (0.9)	47.7 (8.7)	62.0 (16.7)	72.5 (22.5)	81.3 (27.4)	83.9 (28.8)	81.8 (27.7)	77.0 (25.0)	64.1 (17.8)	47.5 (8.6)	33.7 (0.9)	59.5 (15.3)
Daily mean °F (°C)	20.4 (−6.4)	24.9 (−3.9)	37.7 (3.2)	50.3 (10.2)	61.6 (16.4)	71.1 (21.7)	74.0 (23.3)	71.8 (22.1)	65.3 (18.5)	52.8 (11.6)	38.1 (3.4)	25.6 (−3.6)	49.5 (9.7)
Average low °F (°C)	11.9 (−11.2)	16.1 (−8.8)	27.7 (−2.4)	38.6 (3.7)	50.7 (10.4)	60.9 (16.1)	64.1 (17.8)	61.8 (16.6)	53.5 (11.9)	41.4 (5.2)	28.6 (−1.9)	17.5 (−8.1)	39.4 (4.1)
Record low °F (°C)	−26 (−32)	−28 (−33)	−12 (−24)	8 (−13)	27 (−3)	38 (3)	44 (7)	40 (4)	29 (−2)	11 (−12)	−7 (−22)	−24 (−31)	−28 (−33)

Focus

- Goal: To predict **Sale Price** for each house in the test file based on houses sold during **2006 – 2010** using
  - Model: **Linear Regression/Ridge/Lasso**
  - Performance to be measured by **R^2** & Root Mean Squared Error (**RMSE**)

Data Features

- Data files:
  - train(**80** features)
  - test (**79** features)



Numerical

Continuous (Area)

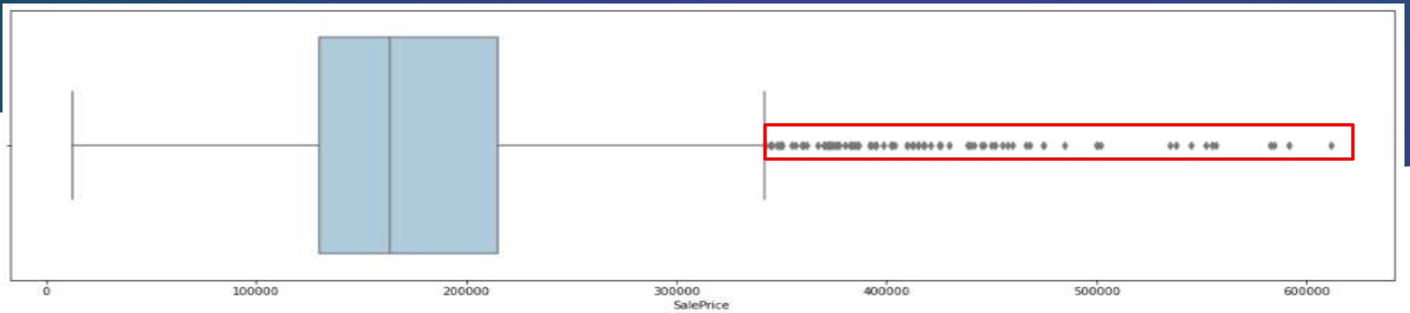
Discrete (No. of Room)

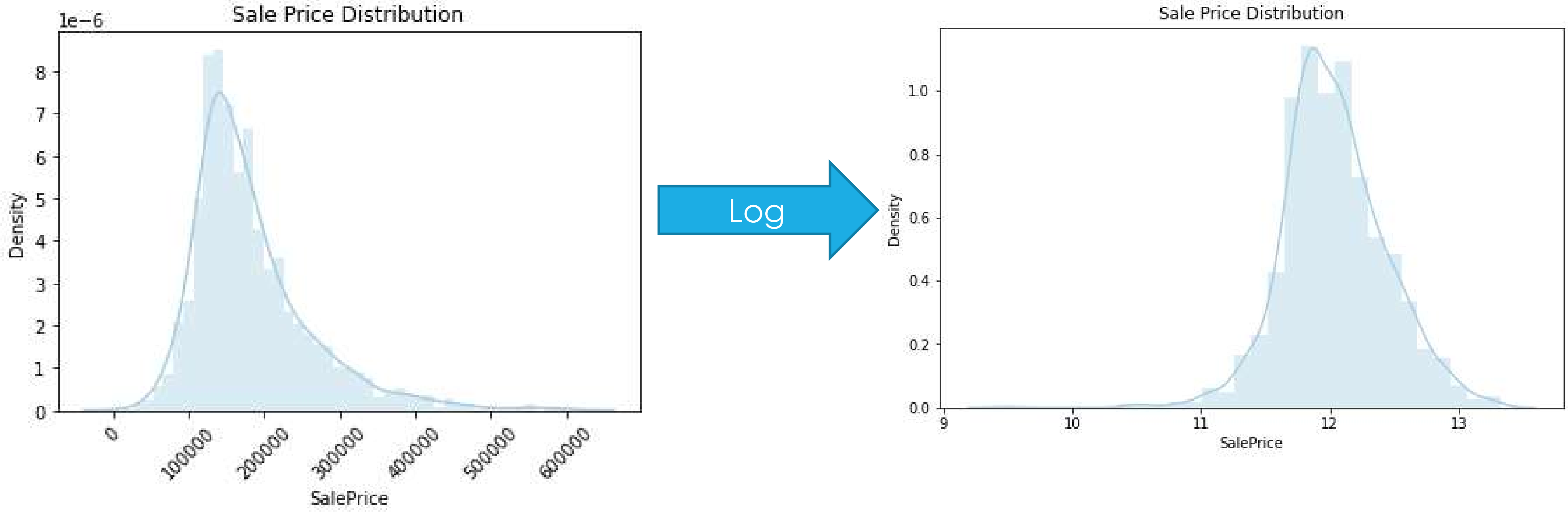
Categorical

Nominal (HouseStyle)

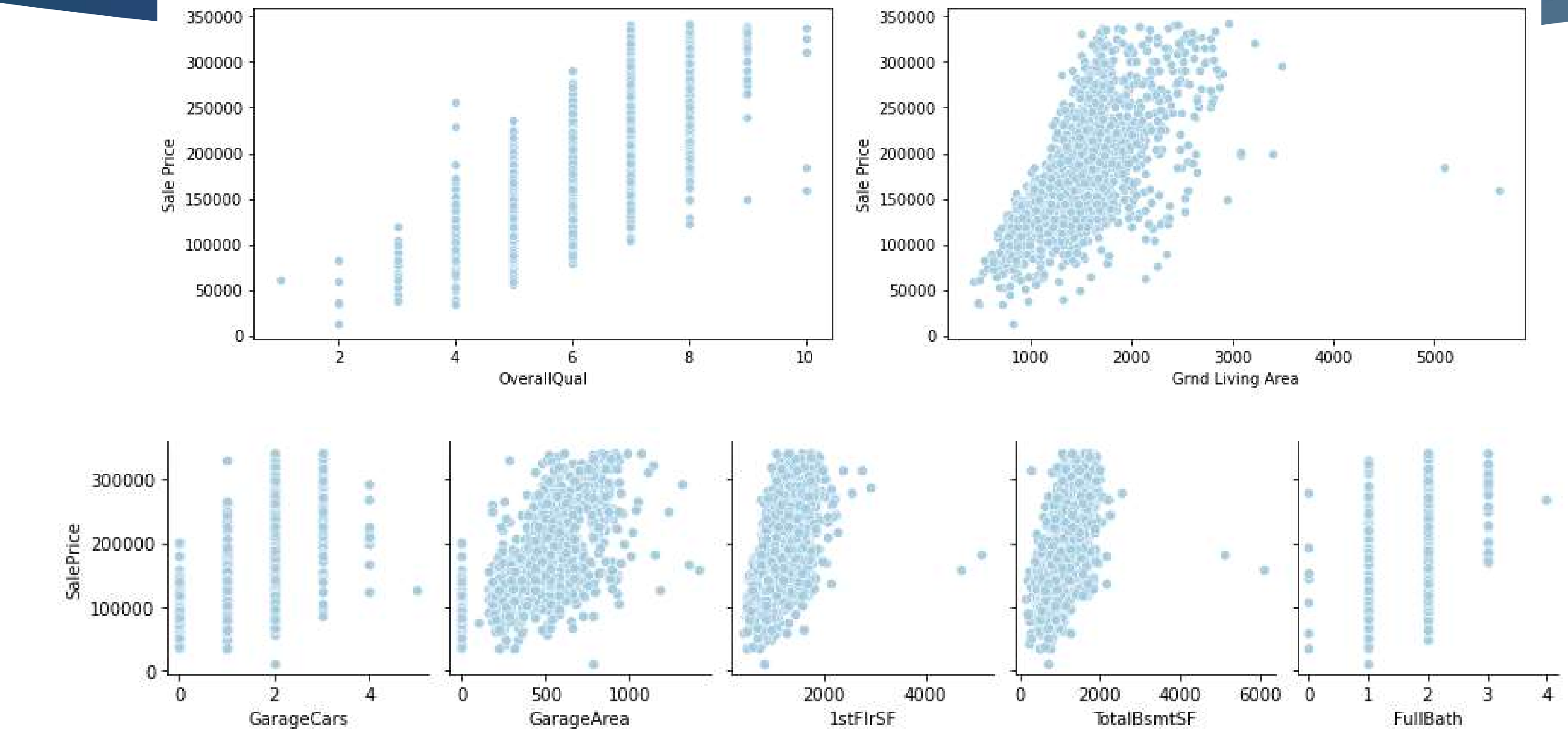
Ordinal (OverallQual)

EDA – Sale Price

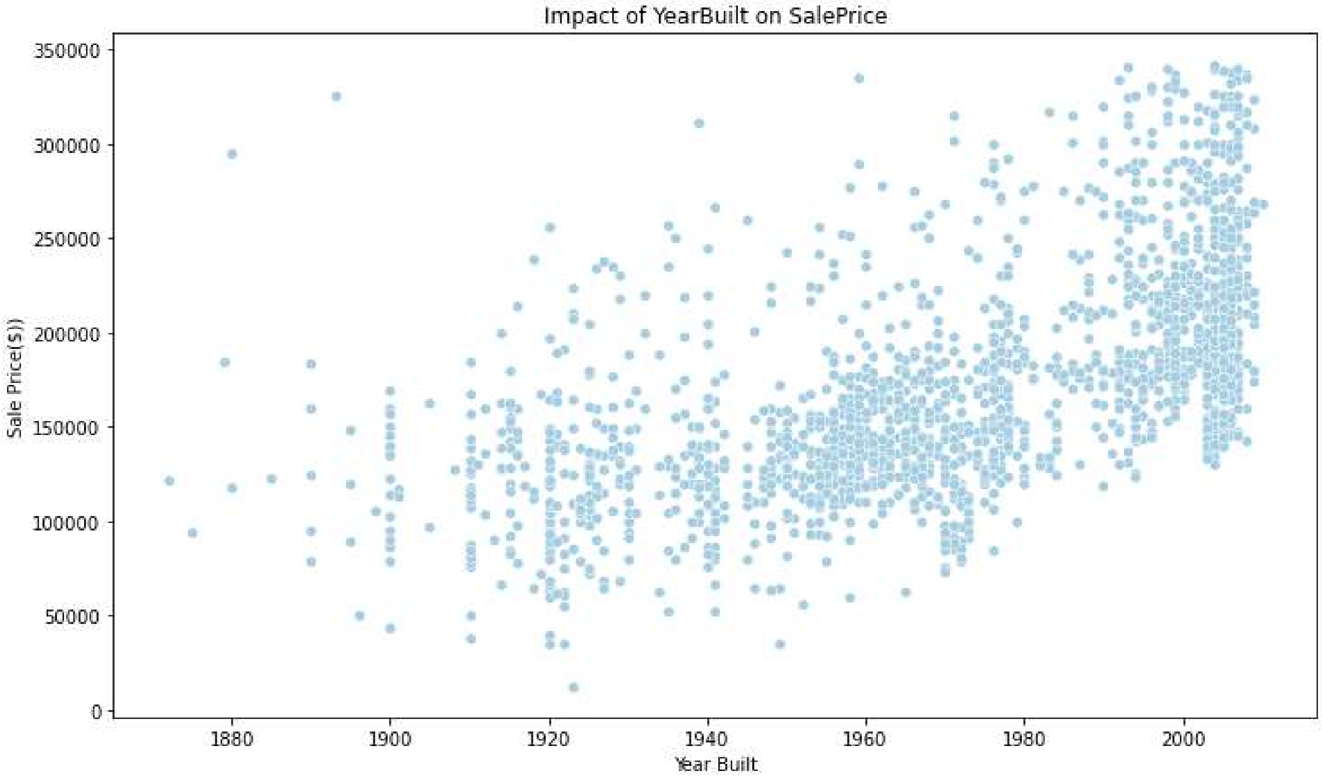




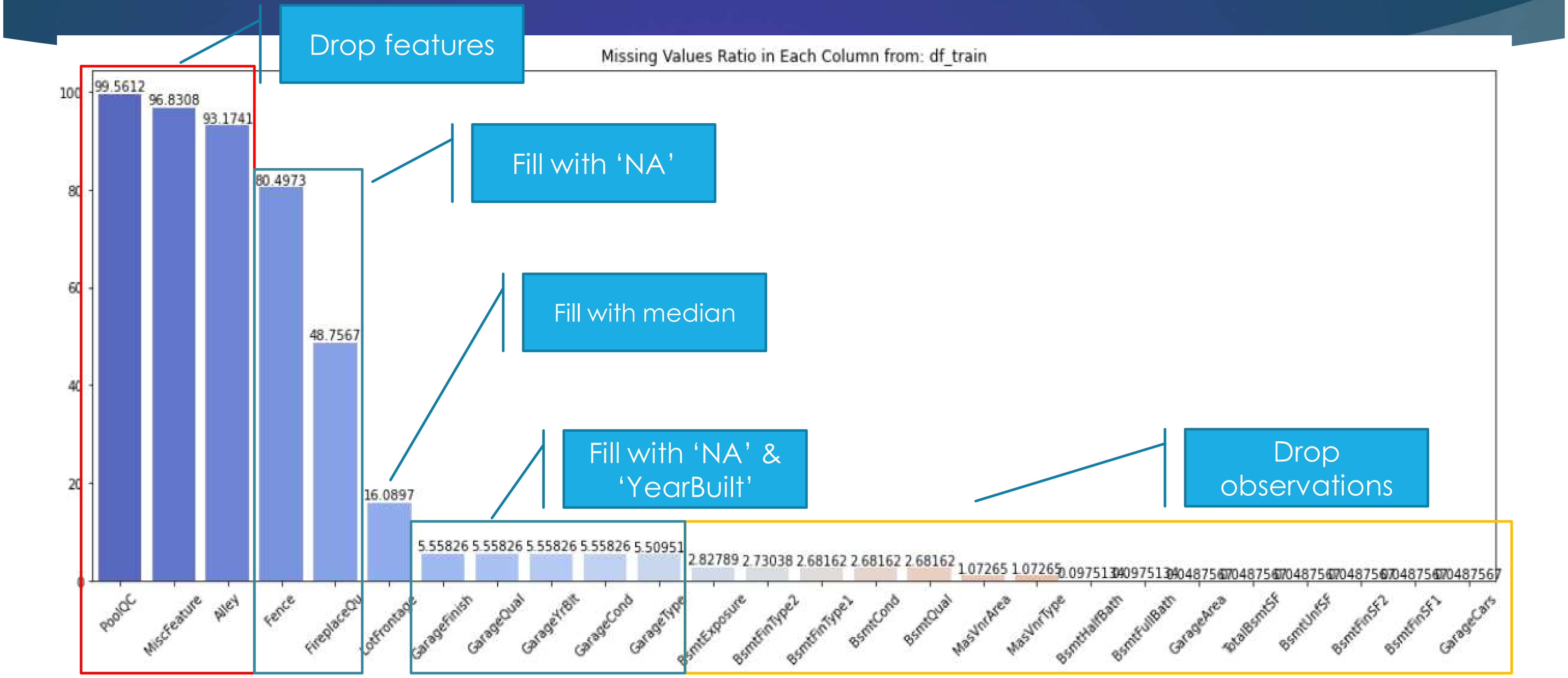
Features with high correlation to Sale Price



EDA – Sale Price & Year Built



Handling missing values in train



Drop features

Fill with ‘NA’

Fill with median

Fill with ‘NA’ & ‘YearBuilt’

Drop observations

Feature Engineering

- Combine train & test DataFrames to
  - Transform ordinal features to numerical scale
  - Create additional features
    - Total Area
    - Property Age
    - Is Remodeled(?)
    - New or Resale
    - Scoring for house features (combination of Quality & Condition)
  - Hot encoding (Dummify nominal features)
- Split DataFrames back to train & test

Feature Selection

- Total features: **253**

- Use forward selection: **105** features
  - Ridge: **104** features
  - Lasso: **19** features

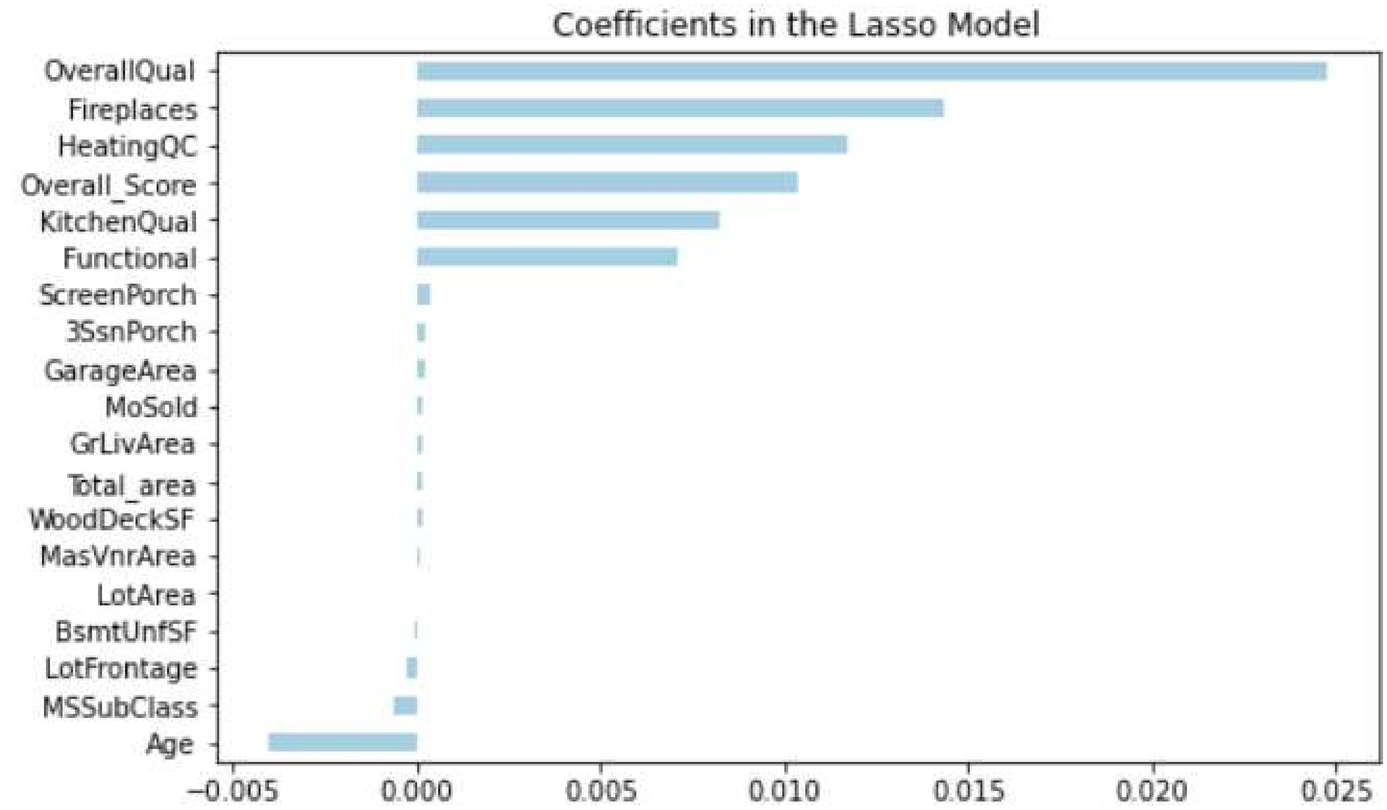
Model Creation & Evaluation

- Split train Data Frame to Train(70%) & Test(30%) set for model training and valuation

Model	Train Score	Test Score	RMSE
Linear Regression	0.94	0.85	0.1614
Ridge (alpha = 1)	0.93	0.87	0.1536
Ridge (alpha = 0.01)	0.94	0.85	0.1611
Lasso (alpha = 1)	0.76	0.73	0.2170
Lasso (alpha = 0.01)	**0.86**	**0.82**	**0.1745**

- **Lasso(alpha = 0.01)** is chosen as the final model for price prediction based on both R^2 and RMSE

Inference


Coefficients in the Lasso Model
OverallQual
Fireplaces
HeatingQC
Overall_Score
KitchenQual
Functional
ScreenPorch
3SsnPorch
GarageArea
MoSold
GrLivArea
Total_area
WoodDeckSF
MasVnrArea
LotArea
BsmtUnfSF
LotFrontage
MSSubClass
Age
−0.005
0.000
0.005
0.010
0.015
0.020
0.025

Inference

- Lasso’s chosen features show overall quality, heating, kitchen quality and functional have positive effects to the price which is sensible due to general cold climate in the area

- House’s age has negative impact to the price

Next Steps/Possible Improvements

- Re-look at Data processing & features engineering to improve performance
- Find out more features that could influence the impact to Sale Price

Thank **you**