# Real Estate Sales Pricing Prediction

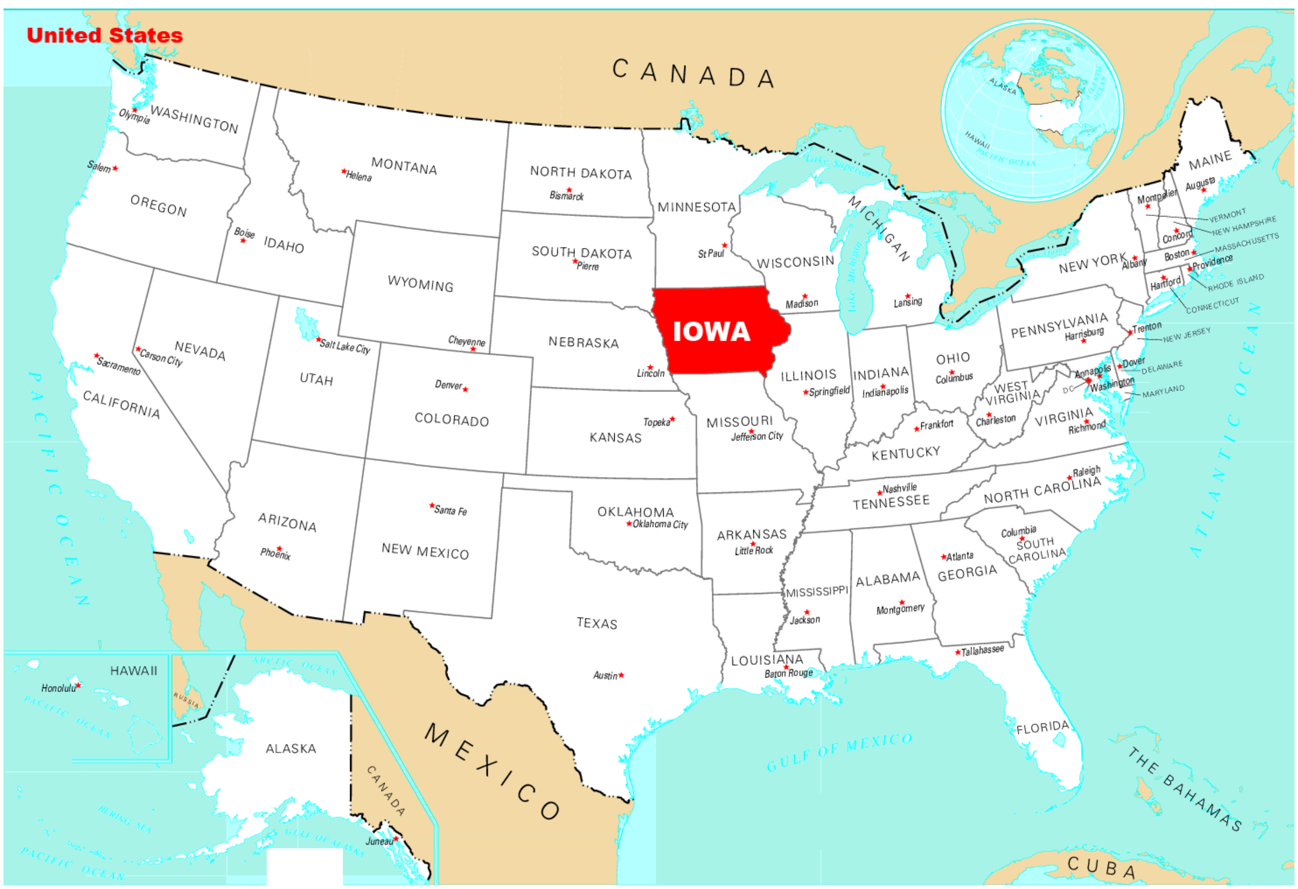
## Litter Box


- **Ruizhe (Jack) Dong**
- **Xu Liu**
- **Esme Luo**


# Are size, neiborghood the only key drivers of real estate sale prices? What's Missing?

**Using the Ames Housing dataset which is an expanded and modernized version of the often cited Boston Housing dataset, we are tempted to address the question mentioned above.**

# Data Set and Background - Location of Sales




# Data Set and Background

- **Data Sourced in 2009**

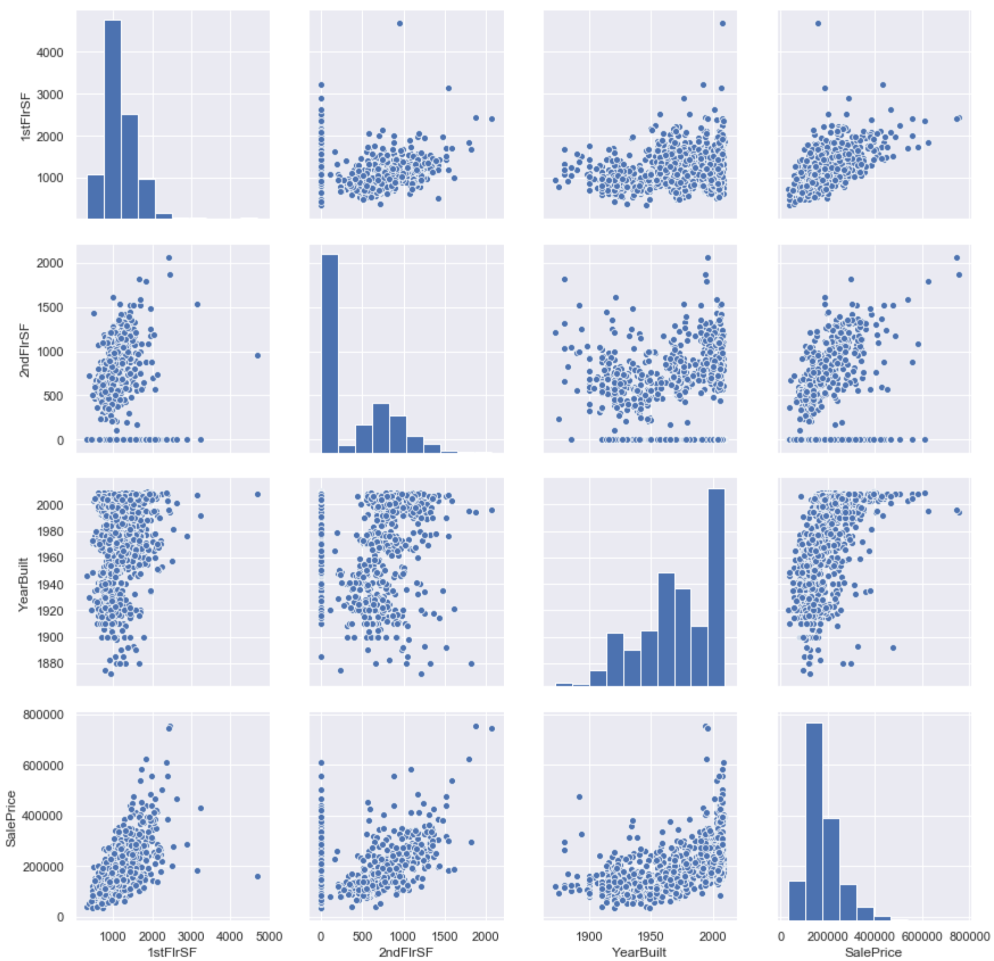
- **Range of Sale Price: $34,900 to $755,000**
- **36 Numerical Features**
- **43 Categorical Features**

# Challenges

### 1. Too many features, better emphasis on the features that's important to improve accuracy.

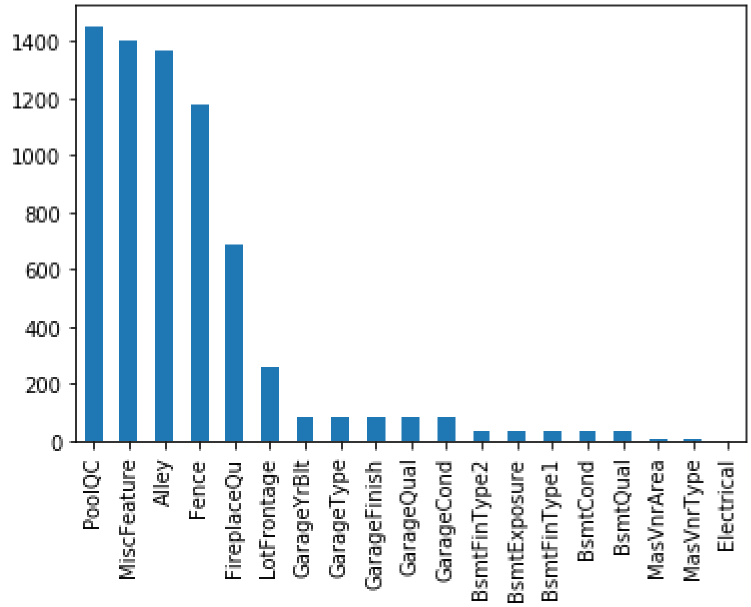
# Challenges

### 2. Multicolinearity:




# Data Processing

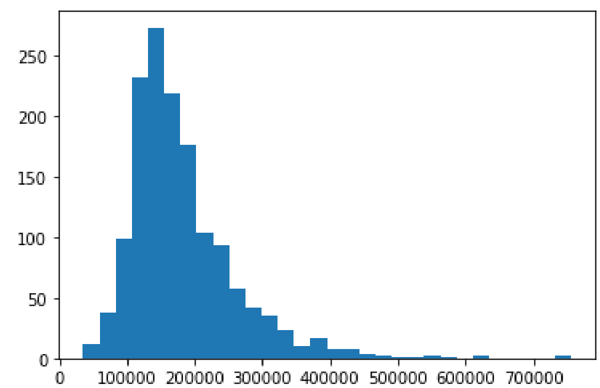
- **Excluding features with more than 10% missing values**



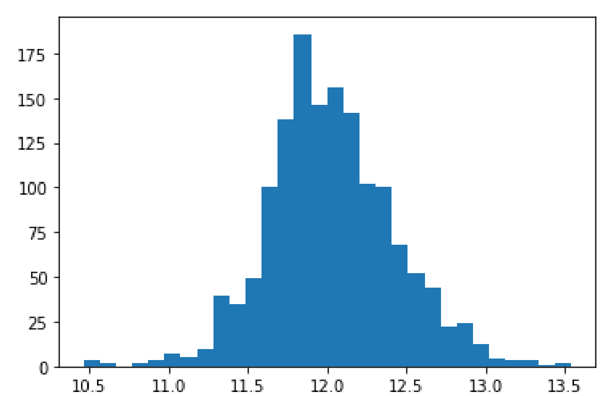

- **Excluding categorical features with too many classes to avoid overfitting**

# Data Processing - Log Transformation of Target

- **Sale Price**




- **log(Sale Price)**




# Pipeline

- **Numerical:** impute with median
- **Categorical:** impute with most common / one hot encoding.

## Feature Selection

- **Naive selection** - selection includes squarefeet, neiborhood and housestyle, year built based on intuition.
- **Full Model** - all features included.
- **Reduced Model** - feature removed based on permutation importance score.

## Modelling Approach

- **Linear Regression (Lasso)**
- **Random Forest Regressor**
- **Graient Boosting Regressor**

## Model Evaluation Preliminary Results (Naive Selection):

```
Naive Feature Selection Training:
```

| | MAE_train | MSE_train | R^2_train | MAPE_train | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.08 | 0.03 | 0.83 | 0.937456 | 0.162844 |
| Lasso | 0.10 | 0.04 | 0.72 | 1.181609 | 0.210453 |
| GradientBoostingRegressor | 0.08 | 0.02 | 0.88 | 0.847233 | 0.137015 |

```
Naive Feature Selection Testing:
```

| | MAE | MSE | R^2 | MAPE | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.09 | 0.04 | 0.79 | 1.084460 | 0.189584 |
| Lasso | 0.10 | 0.04 | 0.77 | 1.157128 | 0.197627 |
| GradientBoostingRegressor | 0.09 | 0.03 | 0.82 | 1.042674 | 0.176126 |

---

## Model Evaluation Preliminary Results (Full Model):

```
Full Model Training:
```

| | MAE_train | MSE_train | R^2_train | MAPE_train | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.05 | 0.01 | 0.91 | 0.653651 | 0.119048 |
| Lasso | 0.07 | 0.03 | 0.80 | 0.953432 | 0.179829 |
| GradientBoostingRegressor | 0.05 | 0.01 | 0.96 | 0.496677 | 0.080476 |

```
Full Model Testing:
```

| | MAE | MSE | R^2 | MAPE | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.07 | 0.02 | 0.86 | 0.869544 | 0.155604 |
| Lasso | 0.08 | 0.03 | 0.81 | 1.001921 | 0.179815 |
| GradientBoostingRegressor | 0.06 | 0.02 | 0.89 | 0.758071 | 0.139942 |

# Permutation Importance

- **Most Significant Features**

| Feature | Importance |
|---|---|
| Overall Quality | 0.5543 |
| Area above Basement | 0.1611 |
| Basement Area | 0.0399 |
| Basement Finished Area | 0.0303 |
| Size of Garage | 0.0209 |
| First Floor Area | 0.0192 |
| Second Floor Area | 0.0166 |
| Number of Rooms | 0.0154 |
| Lot Area | 0.0120 |
| Year Built | 0.0111 |

### Permutation Importance

- **Least Significant**

| Feature | Importance |
|---|---|
| Kitchen Area | 0.000325 |
| Porch Area | 0.000307 |
| Swimming Pool Area | 0.000279 |
| Low Quality Finished Area | 0.000081 |
| Miscellaneous Feature Value | 0.000078 |

### Model Evaluation Preliminary Results (Reduced Model):

```
Reduced Model Training:
```

| | MAE_train | MSE_train | R^2_train | MAPE_train | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.05 | 0.01 | 0.91 | 0.653651 | 0.119048 |
| Lasso | 0.07 | 0.03 | 0.80 | 0.953432 | 0.179829 |
| GradientBoostingRegressor | 0.05 | 0.01 | 0.96 | 0.496677 | 0.080476 |

```
Reduced Model Testing:
```

| | MAE | MSE | R^2 | MAPE | rmsle |
|---|---|---|---|---|---|
| RandomForestRegressor | 0.07 | 0.02 | 0.86 | 0.869544 | 0.155604 |
| Lasso | 0.08 | 0.03 | 0.81 | 1.001921 | 0.179815 |
| GradientBoostingRegressor | 0.06 | 0.02 | 0.89 | 0.758071 | 0.139942 |

# Limitations

**1. Not suited for broad applications, data is collected in one specific state.**

**2. Macros such as inflation, impact on the larger economic scale is difficult to guage and be taken into account in the model.**

# Future Work

**1. Hyperparameter fine-tuning**

**2. Further investigation in feature importance of categorical variable (Label Encoding, One hot encoding, Drop subset)**

**3. Dimension reduction and feature elimination for precision**

# THANK YOU!!


In [ ]:

```

```

## EDA

Graphs to be included placeholders (challenges to address in challenges)

- 

## Modelling Approach

1. Linear Regression Lasso
2. Random Forest Regressor
3. Graient Boosting Regressor

## Data Cleaning and Pipeline

1. Numerical Data
2. Categorical Data

## Load Data