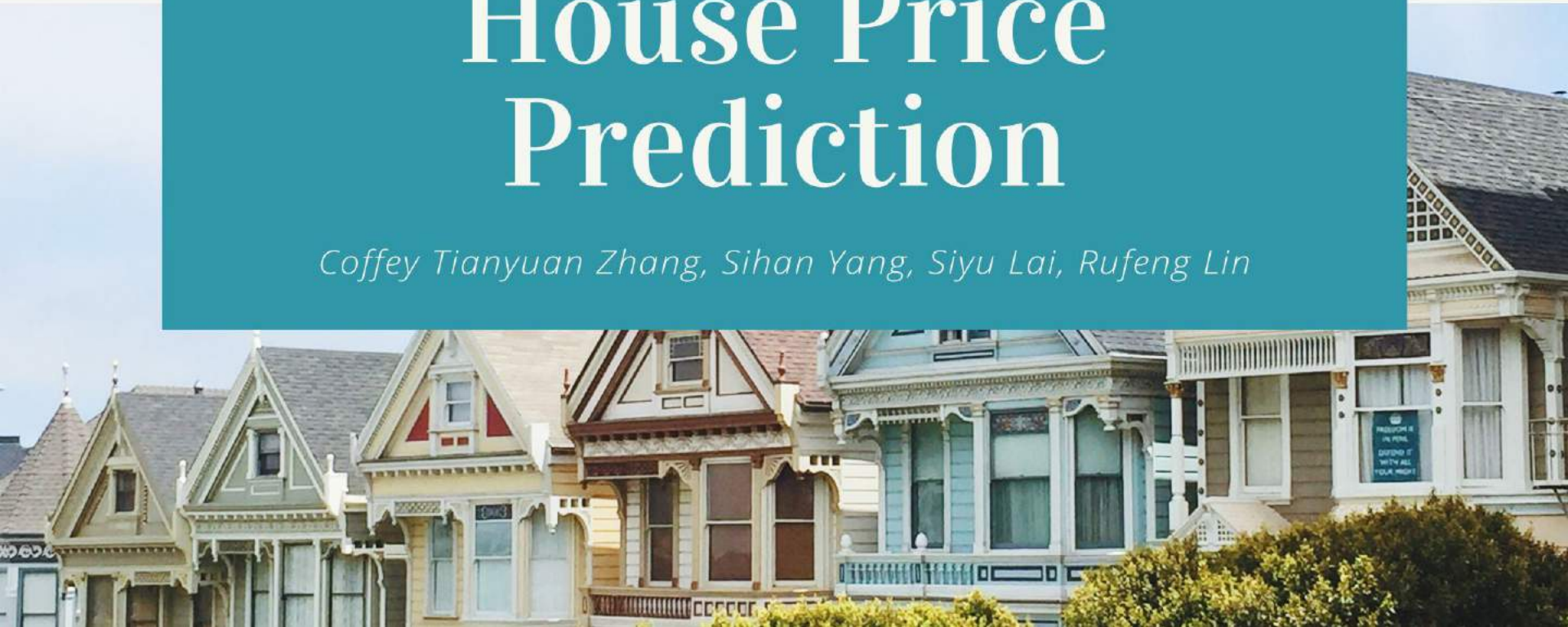
House Price Prediction

Coffey Tianyuan Zhang, Sihan Yang, Siyu Lai, Rufeng Lin



Objectives

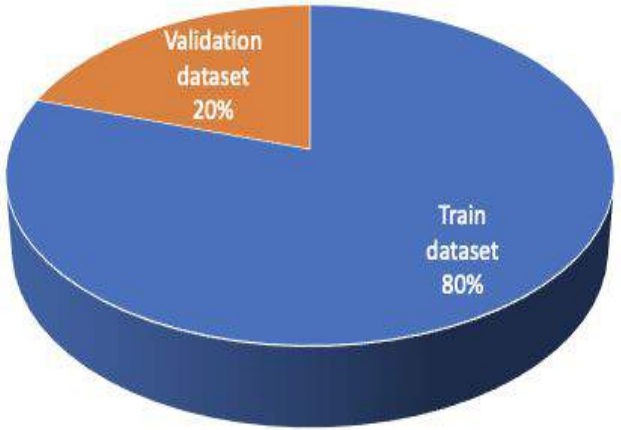
PURPOSE OF PROJECT

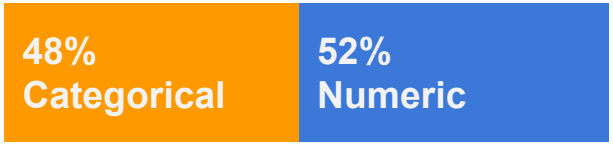
- Use necessary tools and course frameworks to analyze dataset of real estate in Ames, Iowa
- Provide first-time home owners basic knowledge about the domestic housing market
- Build a decent regression model with a low MSE
- Accurately predict the closing price using our model so that a potential buyer knows the true value of a property

Dataset & Analytical Approach

Data Preparation

- **Train.csv** data into:
  - 80% Train dataset
  - 20% Validation dataset
- **Test.csv data**
- Dimensions: 1460 * 81
- Type: Numeric & Categorical
- Source: Kaggle competition



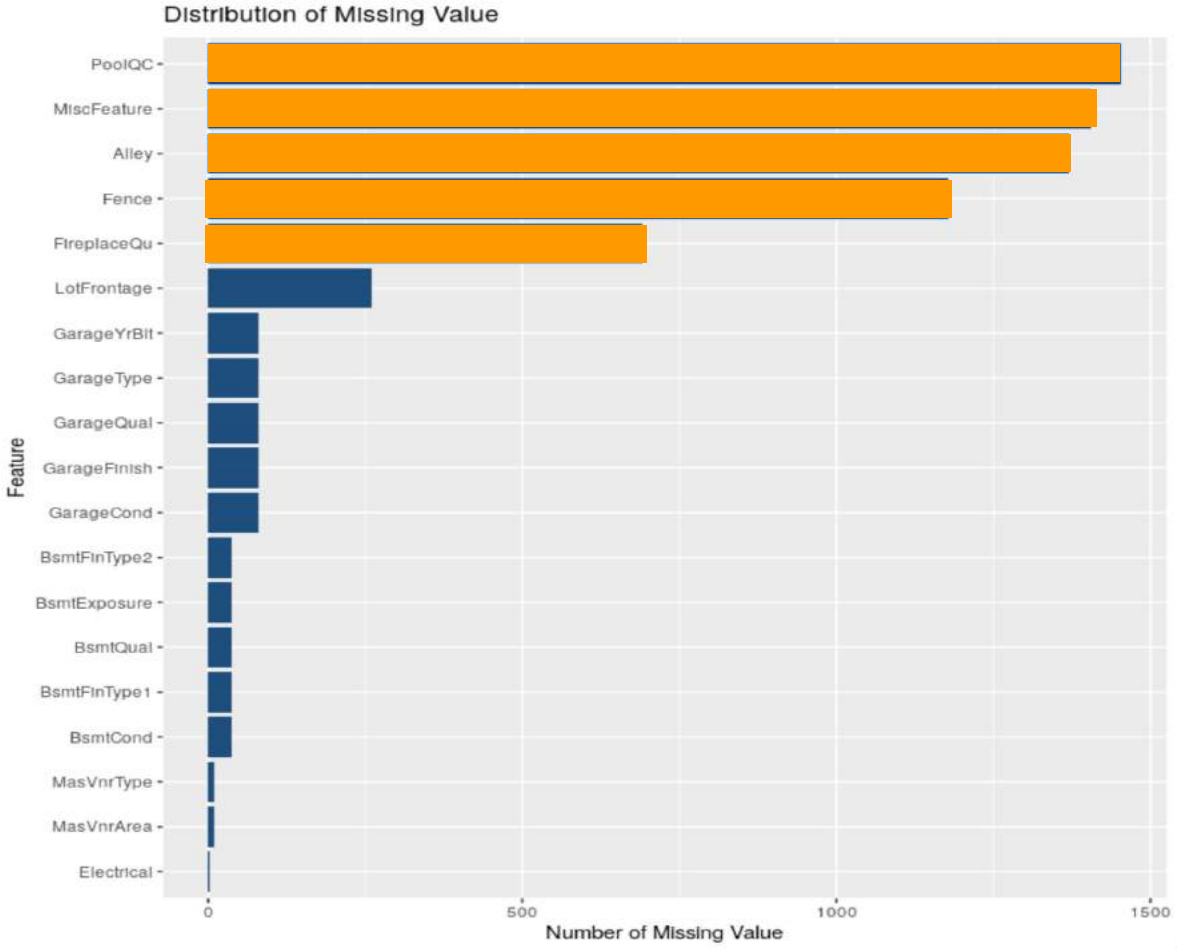


Data Processing

Checked
missing values

Data Processing

(a) Check Missing Values



Data Processing

Fill missing values

Checked missing values

Data Processing

(b) Fill the missing values

1) PoolQC
2) Miscfeature
3) Alley
4) Fence
5) FireplaceQu

Data Processing

Create dummies for categorical features

Fill missing values

Checked missing values

Data Processing

(c) Dummy variables

All factors were transformed to dummy variables:

- MSZoning, Street, BldgType, HouseStyle,Neighborhood,
- Heating, HeatingQC, CentralAir, Electrical, Alley, PavedDrive, PoolQC, Fence, MiscFeature, Utilities, LotConfig, LandSlope, FireplaceQu
- Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond

Data Processing

Data Processing

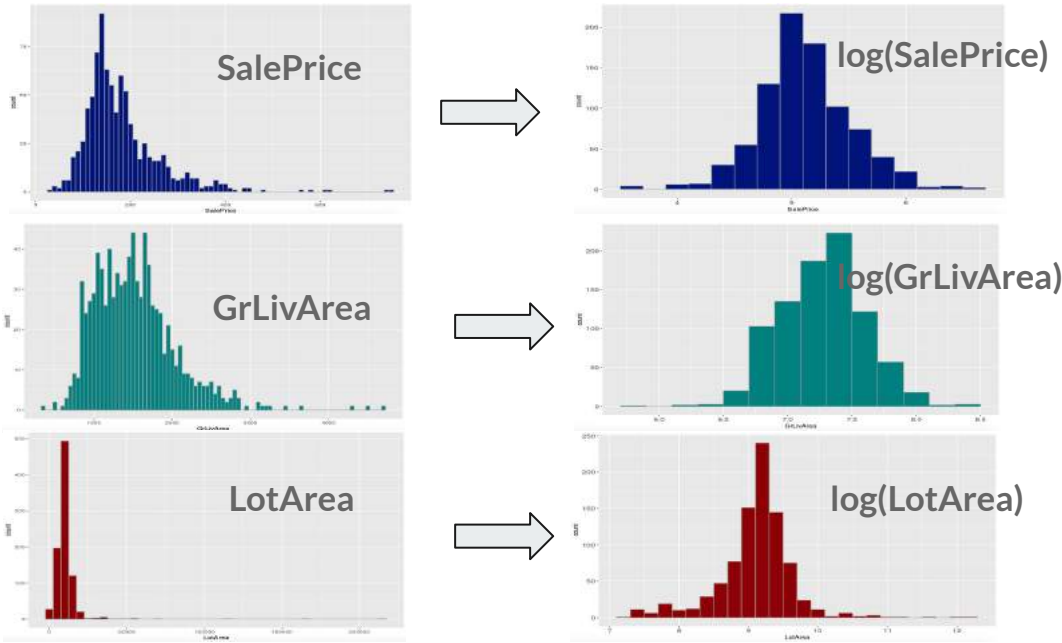
(d) Log Transformation

Skewness threshold (0.75) to maintain the lowest MSE, if the skewness > threshold, use log transformation.

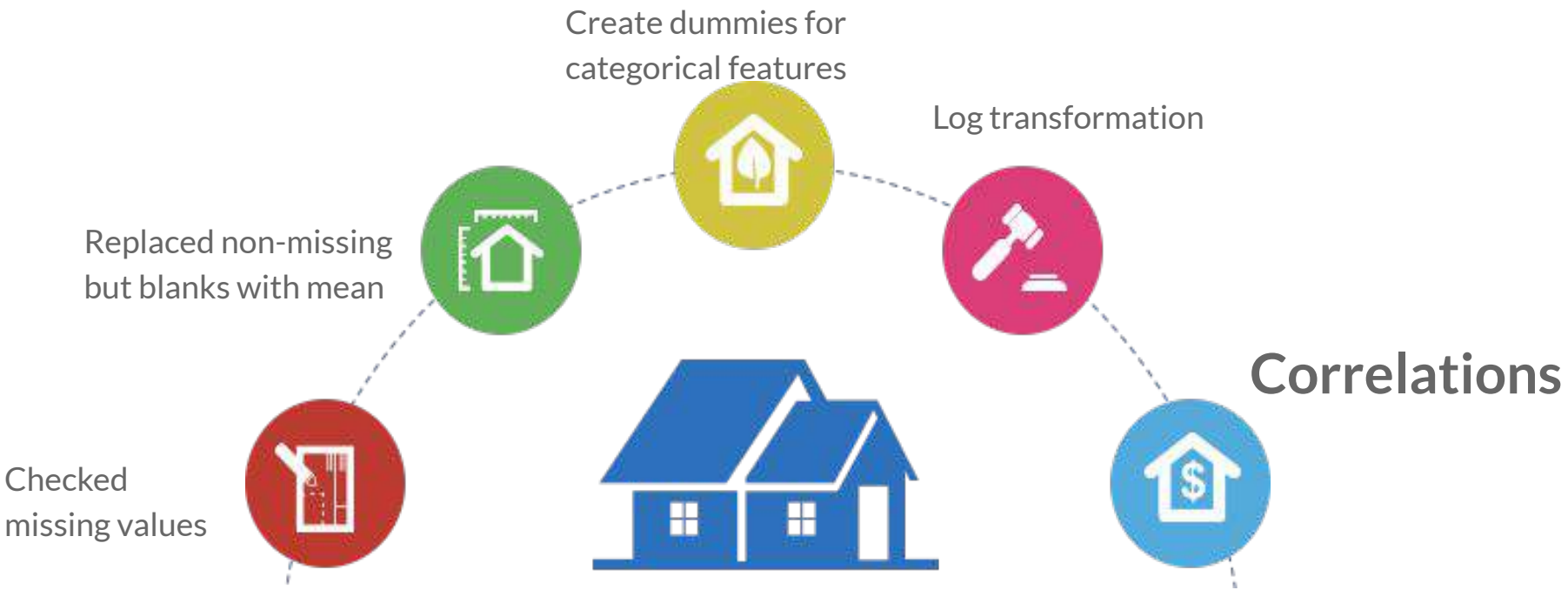
log(SalePrice)

log(GrLivArea)

log(LotArea)



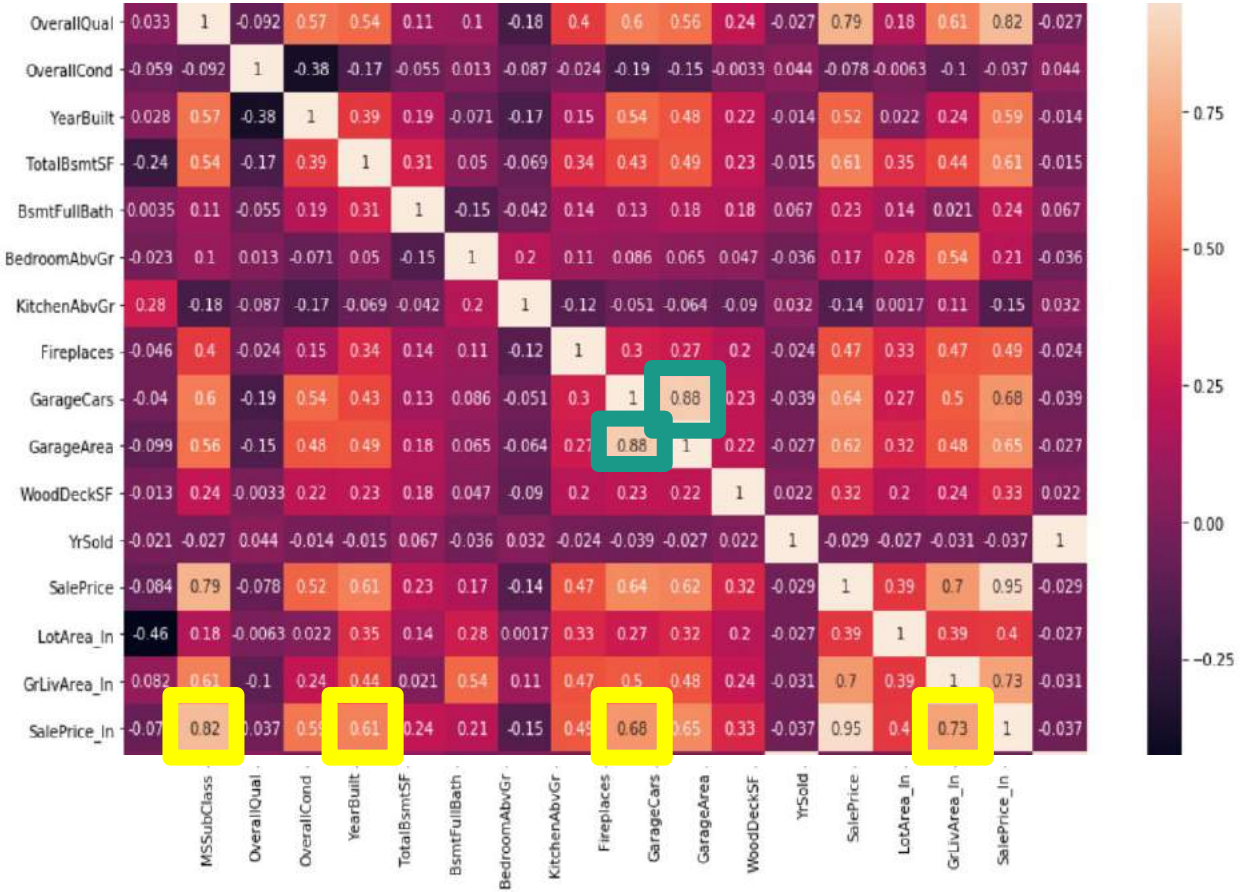
Data Processing



Data Processing

(e) Correlations

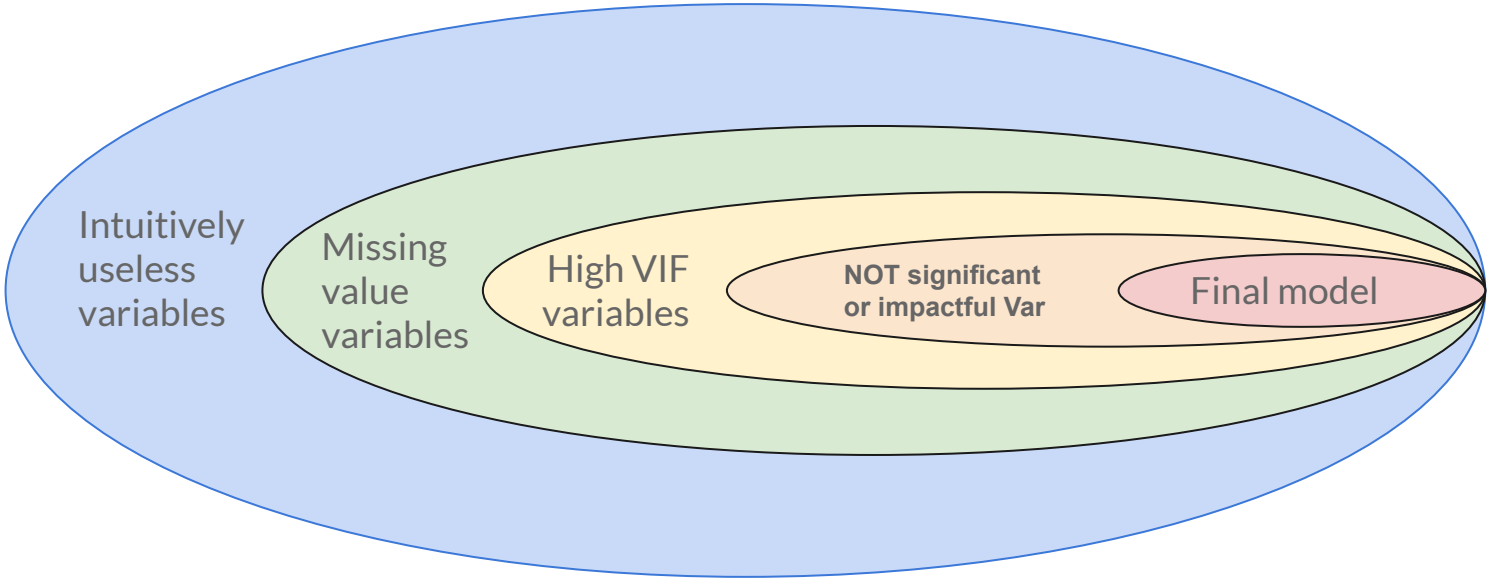
Review to check potential multicollinearity issue



Linear Regression Model - Variable Selection

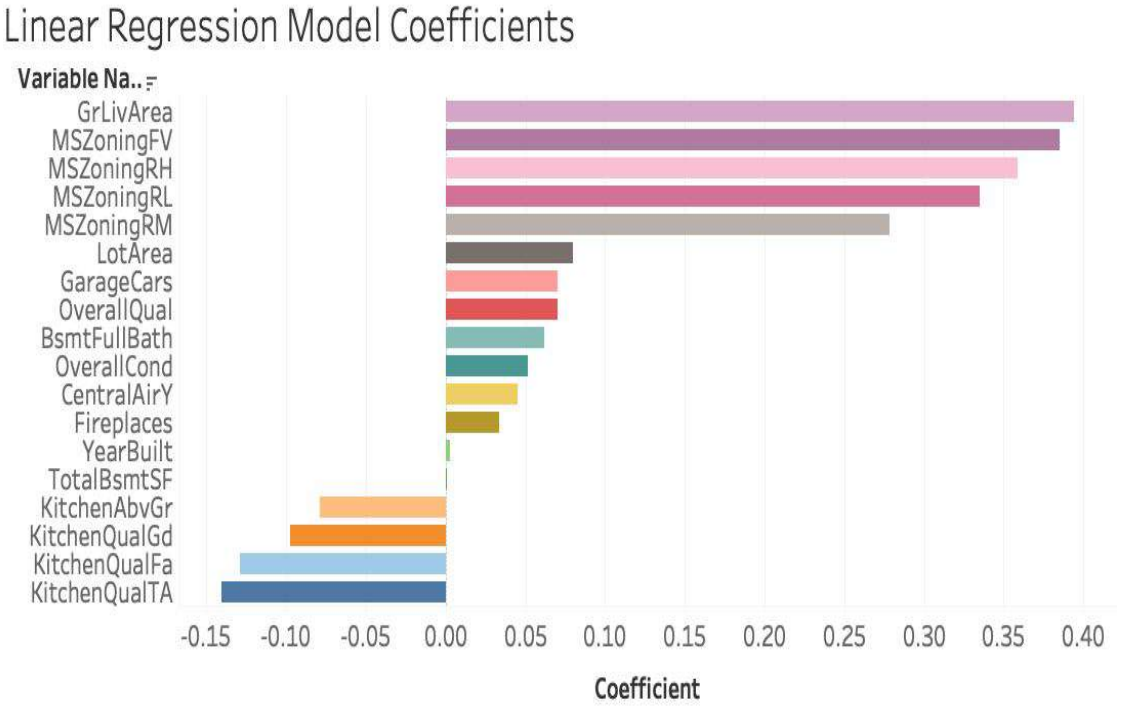
Step 1	Step 2	Step 3	Step 4
Used common sense to select economically significant variables	Used linear regression models to regress price	Decided the final model based on the statistical significance	Created visualization for analytics

Linear Regression Model - Variables Filtration



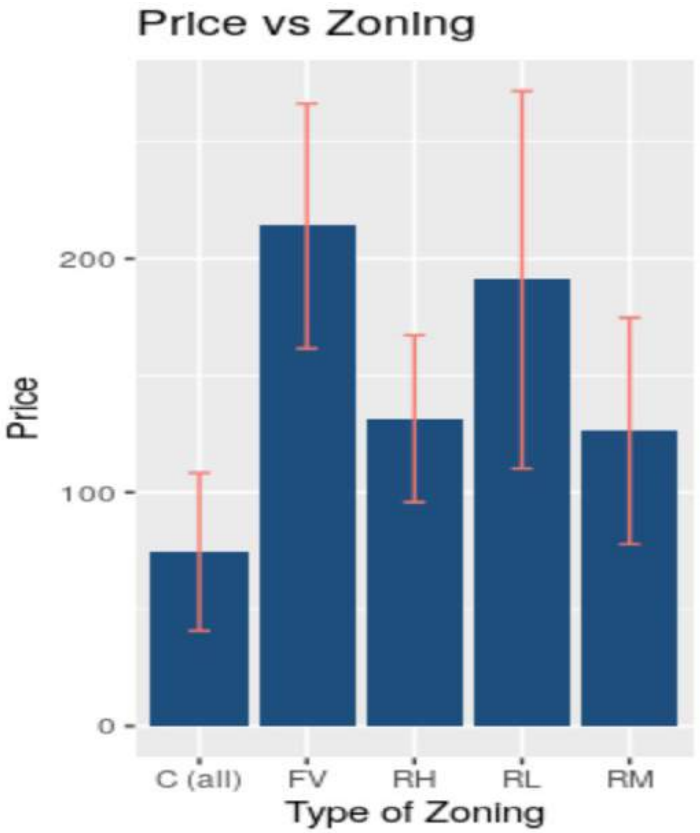
Coefficient Interpretation

- ***Neighborhood*** is essential according to common sense
- BUT ***Neighborhood*** is highly correlated with other predictors
- We choose a similar predictor ***MSZoning*** instead



Coefficient Interpretation

- ***Floating Village (FV)*** has the highest median price
- House price of ***Residential High (RH)*** zone are more stable
- House price of ***Residential Low (RL)*** zone are more expensive

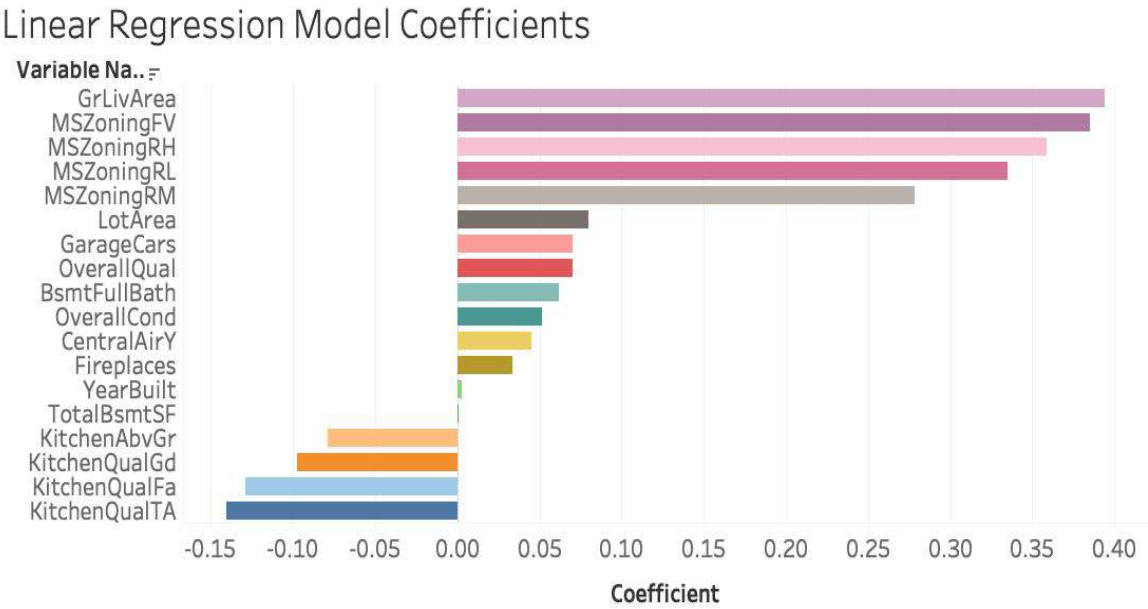


Negative Coefficient

- We chose *Excellent quality* as the base of this dummy variable.

- The number of kitchens show a negative relationship with price.

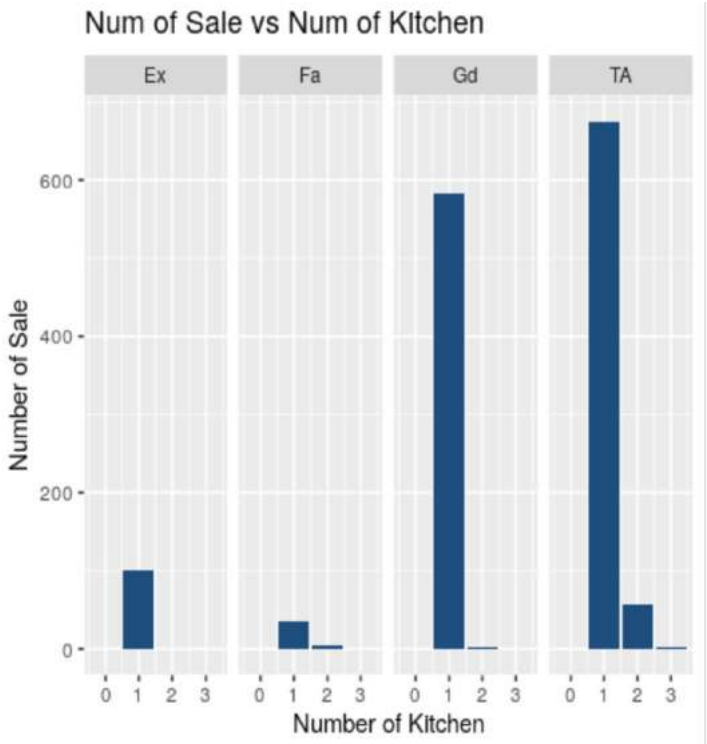
But why?



Coefficient Interpretation

The more kitchens in a house, the lower the house price

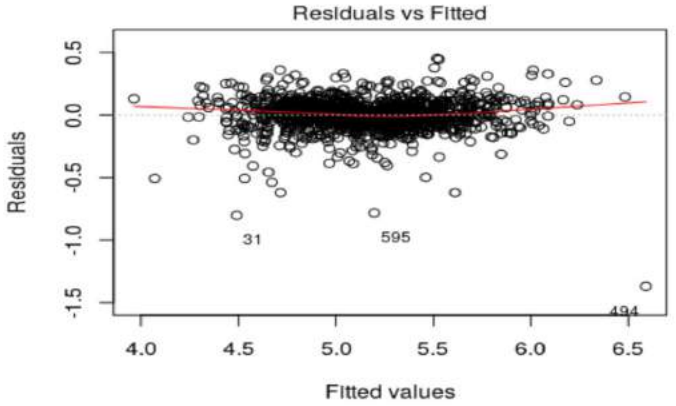
- Additional kitchens are for leasing, but they are not allowed in Iowa
- In rural areas, the restriction is not that strict
- Most houses with multiple kitchens are located in the rural areas
- Houses are cheaper in rural areas of Iowa

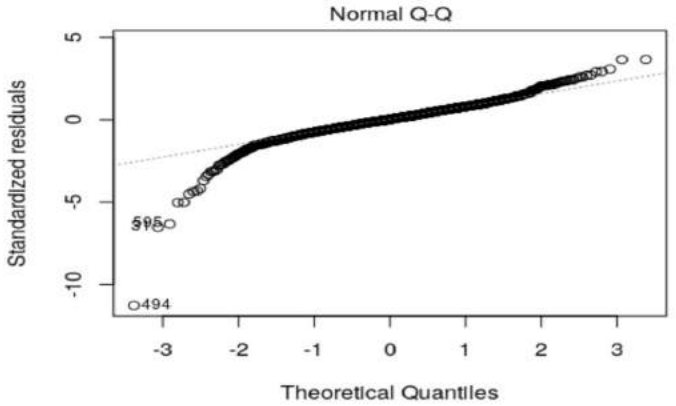


Model Fit

- From the dashboard graphs, the data seems to be driven by outliers, but overall, our model performs well.

- The R square of our final model is 0.9021. When we applied a generalized linear model to our validation data, the Mean Squared Error is only 0.135.





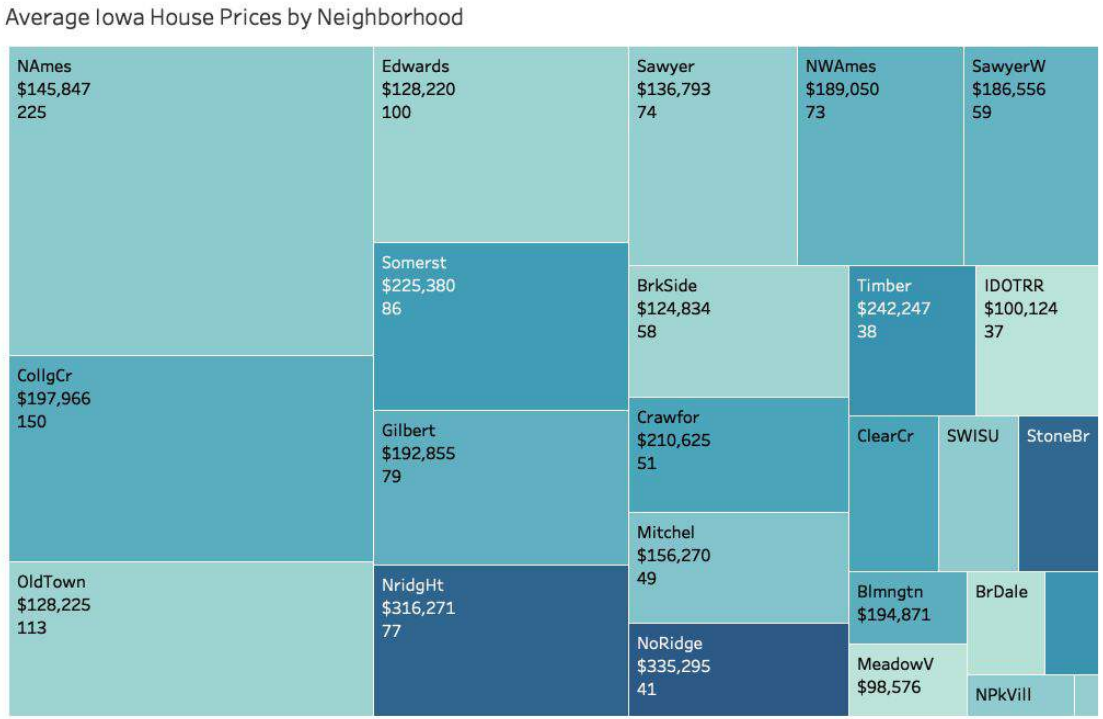
Business Insights

Business Insights #1

Home sales volume varies greatly by neighborhood

Neighborhood vs. Sale Price

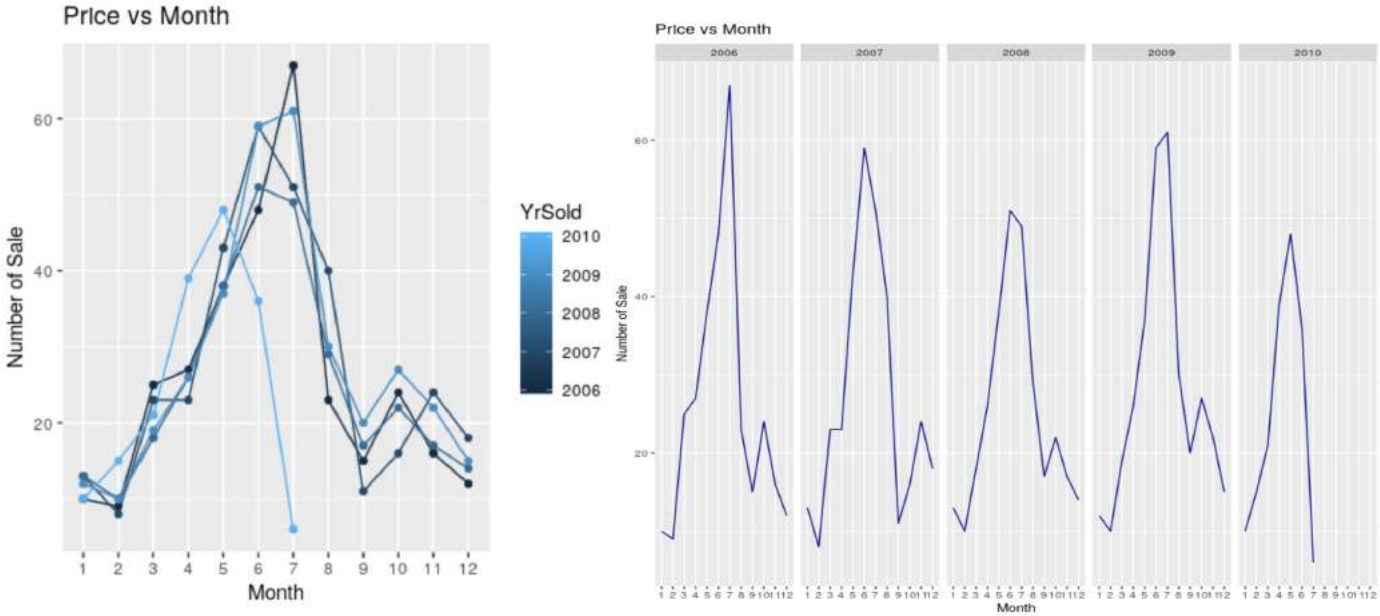
- Neighborhoods with the greatest number of house sales: North Ames, College Creek, Old Town
- The average sales price by neighborhood has range from $98,576 to $335,295



Business Insights #2

Home sales volume has obvious seasonality

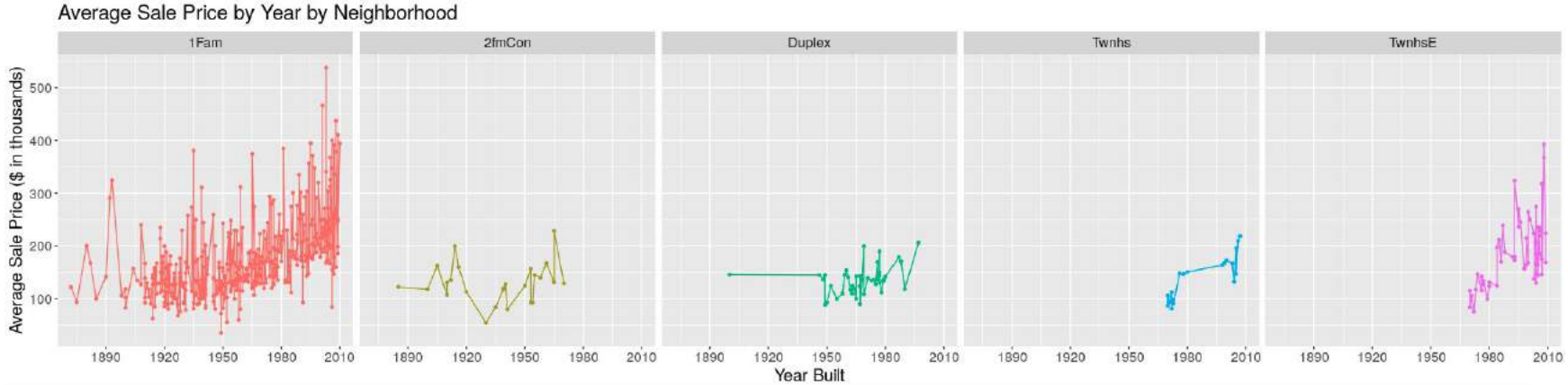
- The volume of house sales peak during the summer
- Why?
  - Easy to remodel
  - Check for house defect, such as leakage



Business Insights #3

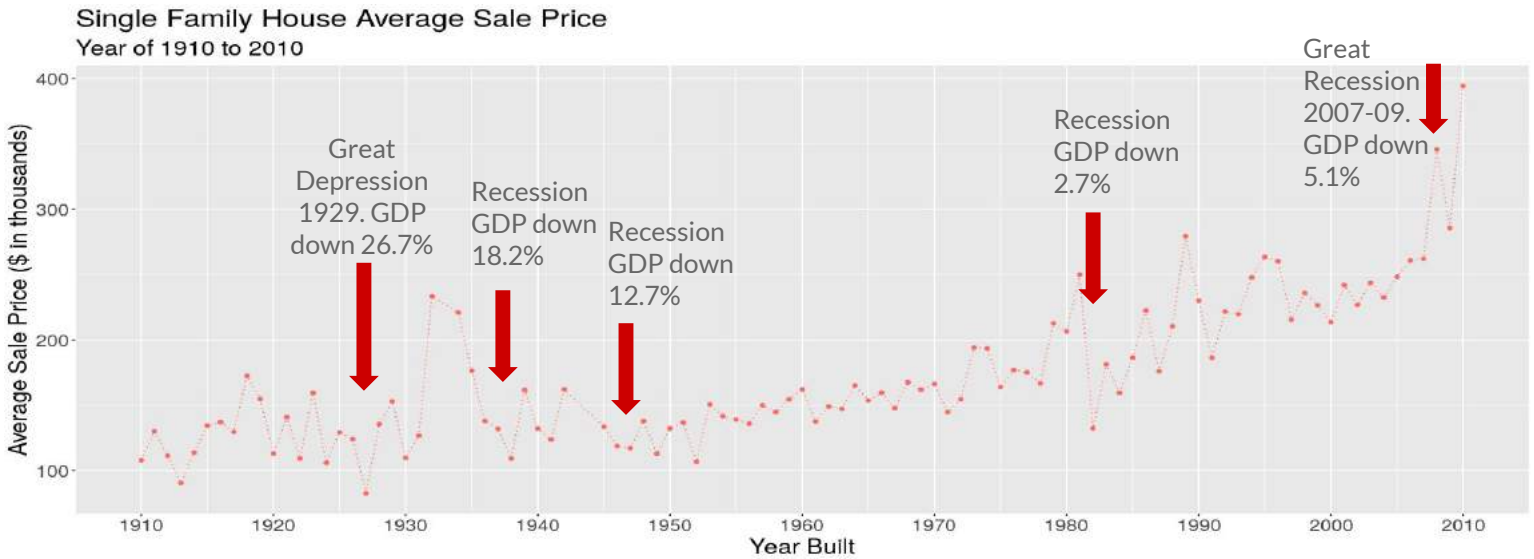
Home price varies significantly by types of buildings

Single-family Detached (1Fam)	Two-family Conversion	Duplex (Duplx)	Townhouse Inside Unit (Twnhs)	Townhouse End Unit (TwnhsE)
				

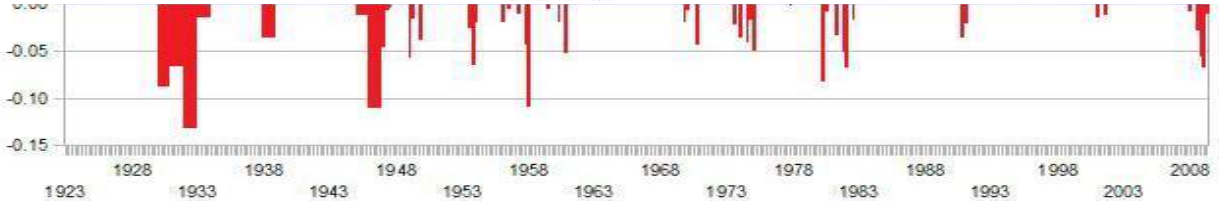


Business Insights #4

Single-family house prices are highly correlated with economy



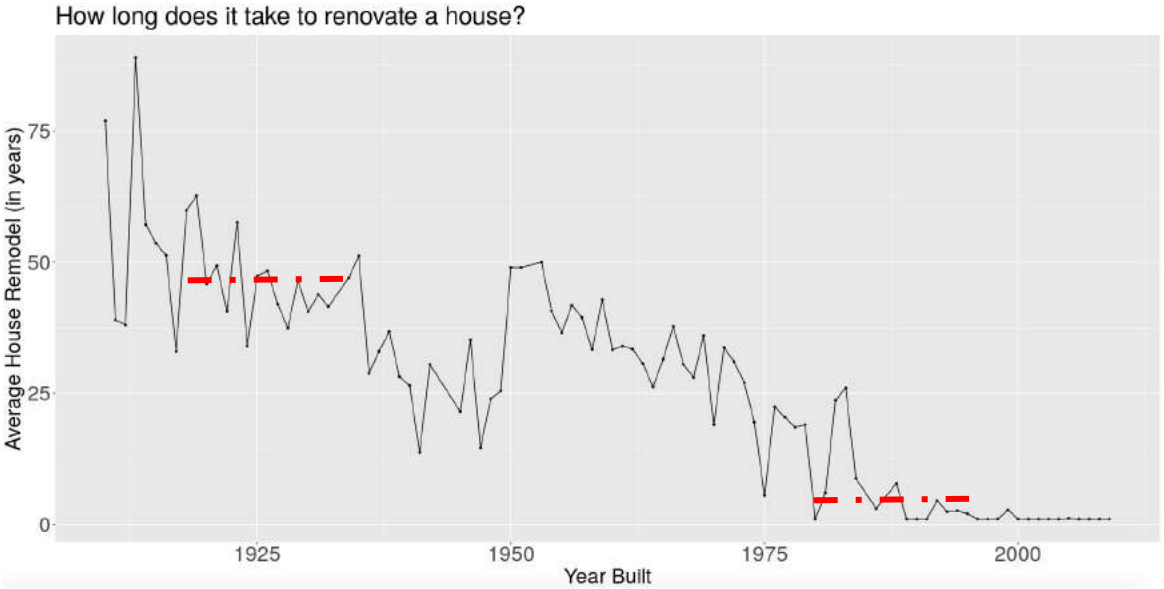
Historical Recessions Magnitude



Business Insights #5

Average house renovation occurs sooner over time

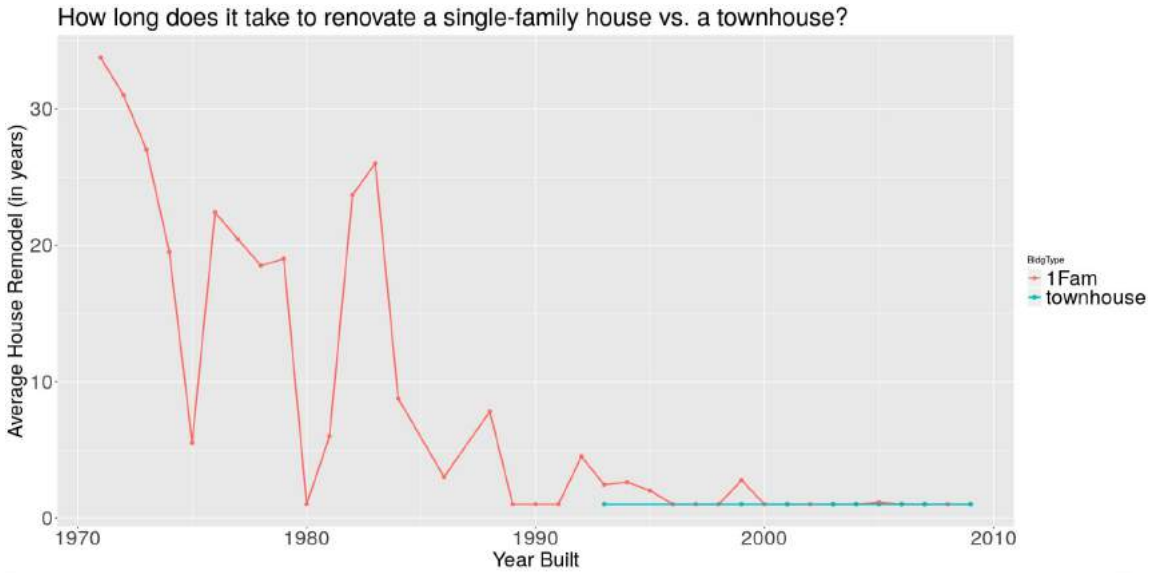
- **Why?**
  - Old days, quality was better so properties preserved longer
  - Nowadays, reselling happens more often so renovate to increase value



Business Insights #5 (cond't)

Townhouses owners do not renovate as single-family home owners

- **Why?**
  - More townhouses than single-family houses are for rentals, so owners tend not to renovate
  - Single-family house are normally purchased by families, with children, they renovate to make houses better



Thanks for listening!

GROUP3: COFFEY, RUFENG,SIYU,SIHAN