QST BA 222 Final Group Project Team 10

House Price Analysis

Insights and Predictions

Presented By: Quique De La Cruz, Tina Hoang, Marshall Naegelin, Ania Shaheed, Julius Yang

Agenda

1. Introduction

2. Data Overview

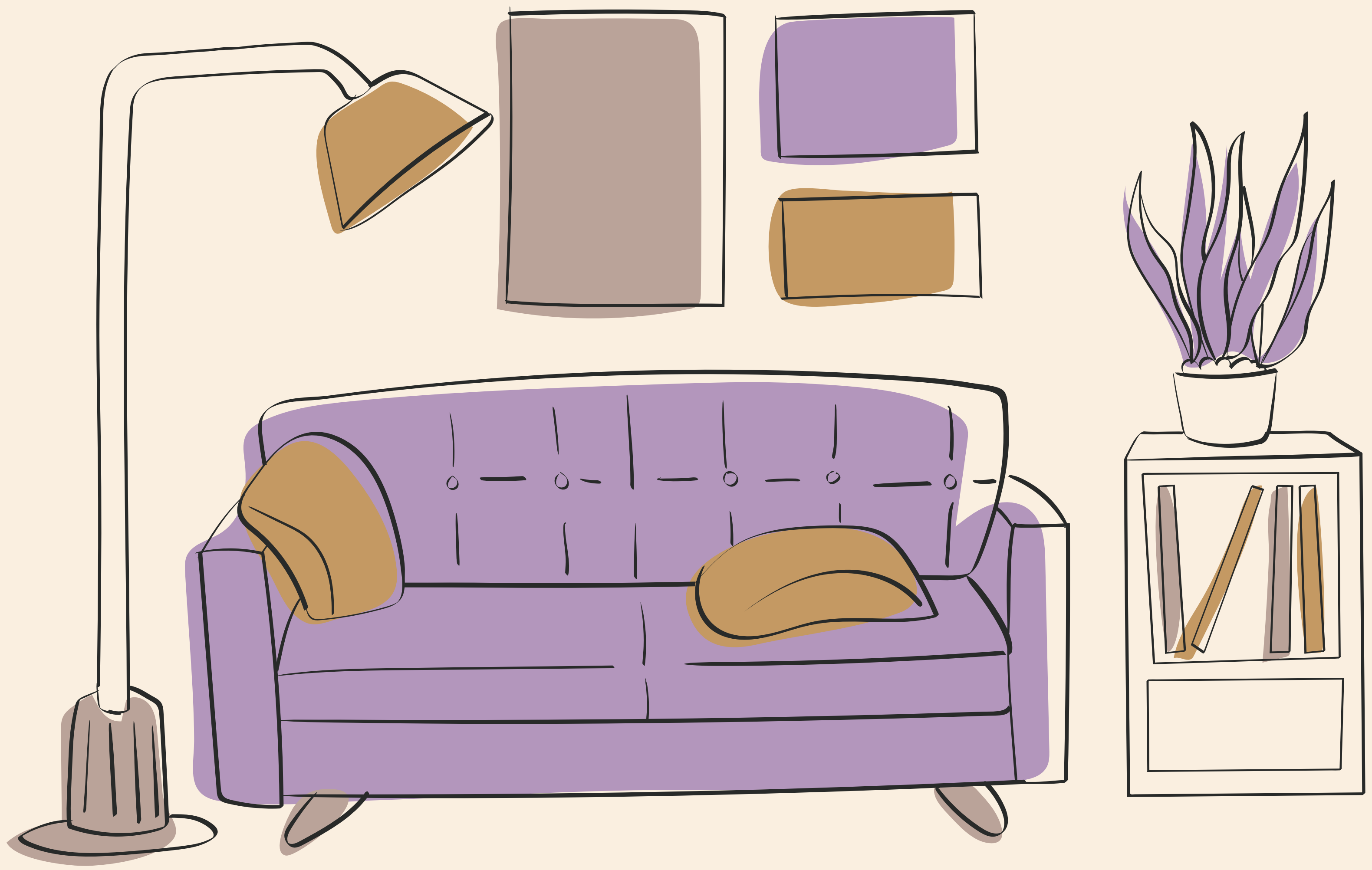
4. Causal Analysis

5. Data Challenges

6. Predictive Model

7. Results & Conclusion

Introduction



Project Overview

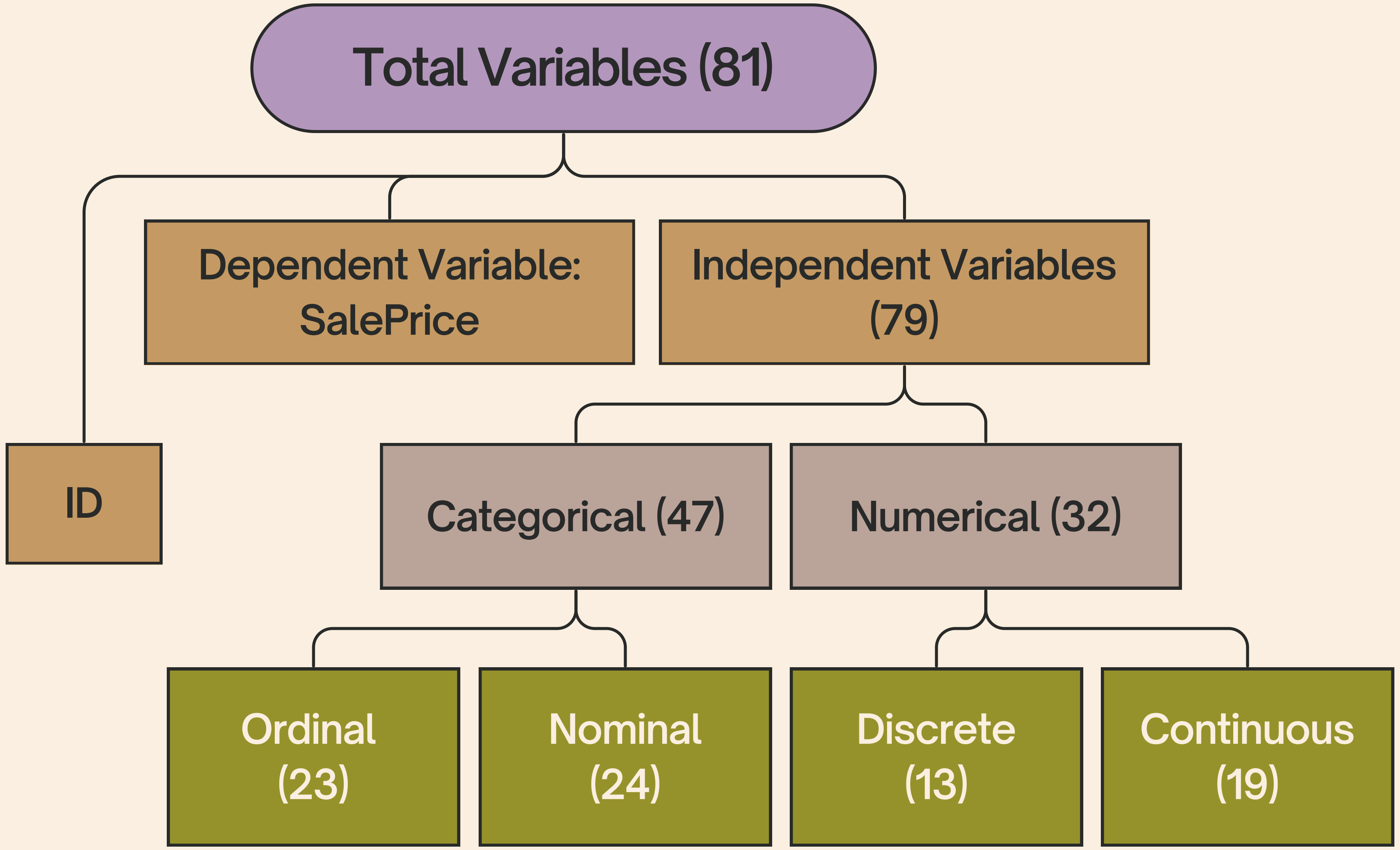
Purpose: build a predictive model for home sale prices

Dataset: 1,460 homes sold in Ames, Iowa (2006–2010)

Key Goals

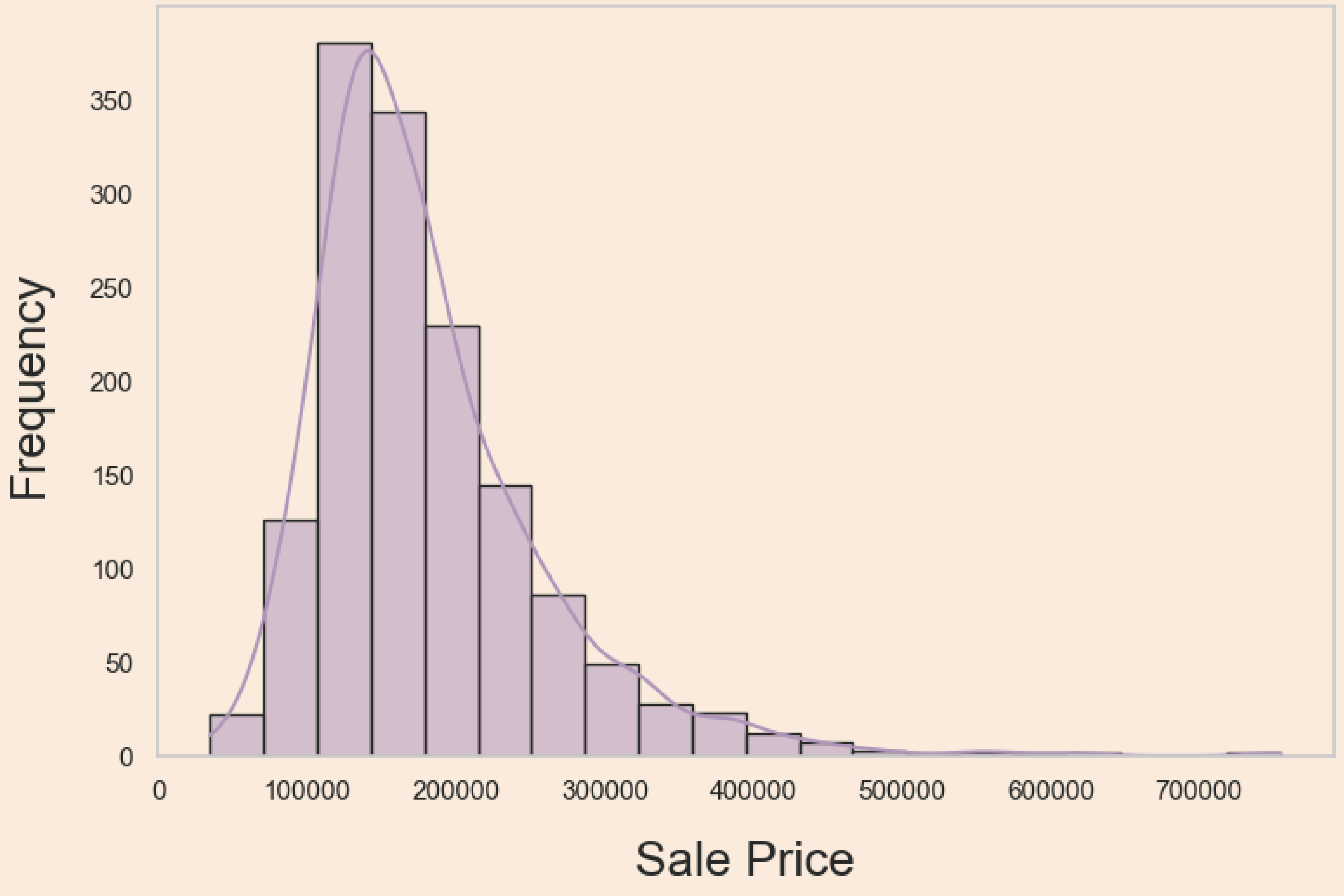
- Analyze data structure
- Pre-process dataset
- Explore causal effect
- Select important predictors
- Train and evaluate models

Data Overview



- Total Variables (81)
  - ID
  - Dependent Variable: SalePrice
  - Independent Variables (79)
    - Categorical (47)
      - Ordinal (23)
      - Nominal (24)
    - Numerical (32)
      - Discrete (13)
      - Continuous (19)

Fig. 1 - Histogram of Sale Price



Exploratory Data Analysis

Correlation to SalesPrice

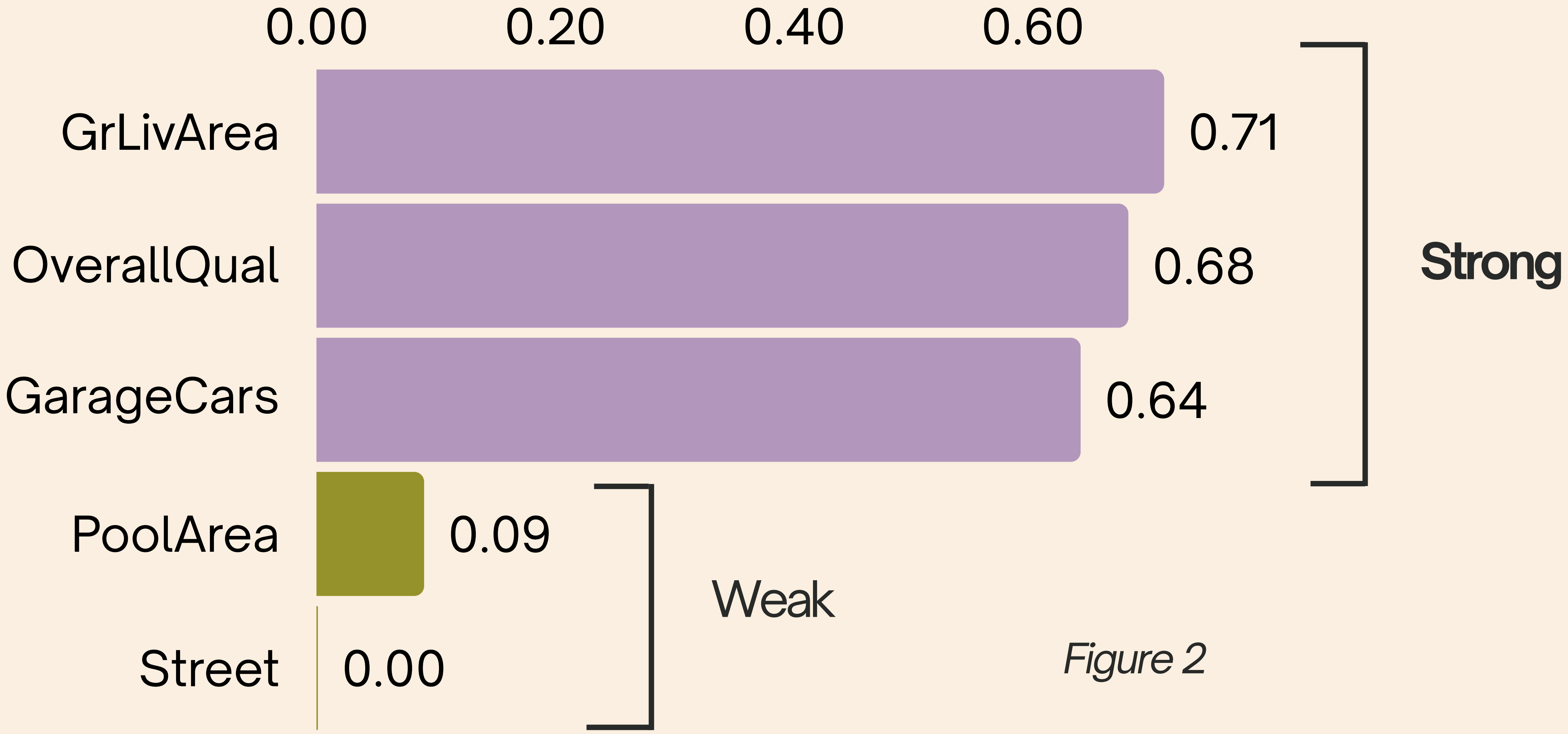


Figure 2

Figure 3. GrLivArea vs SalePrice

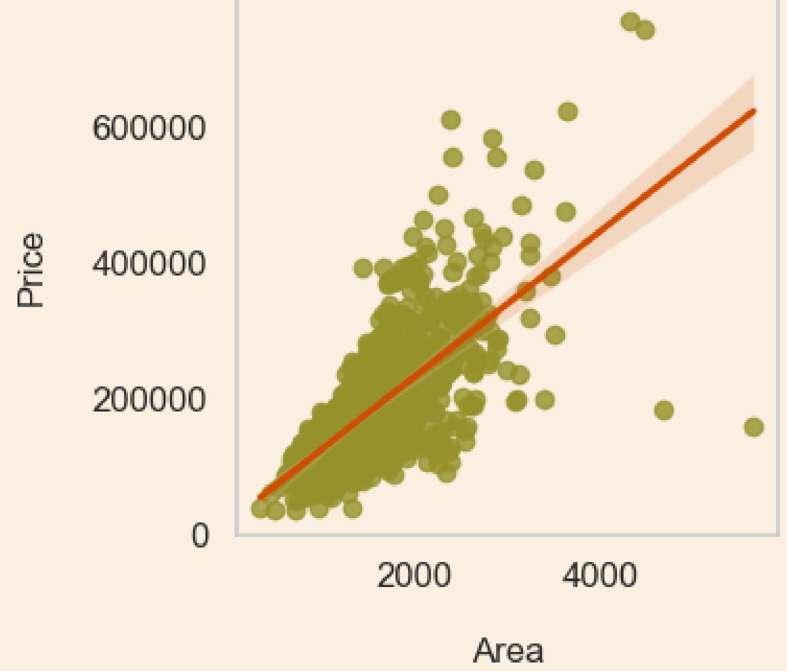
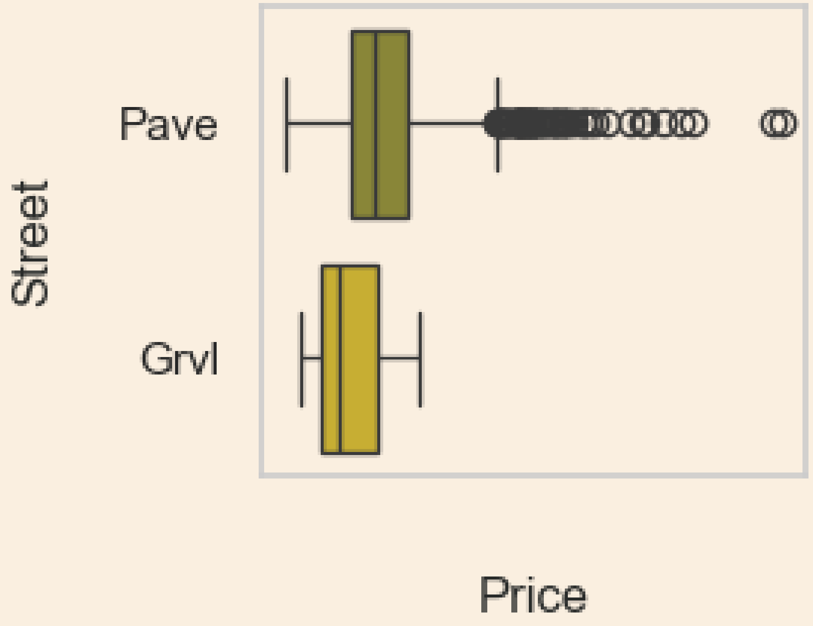


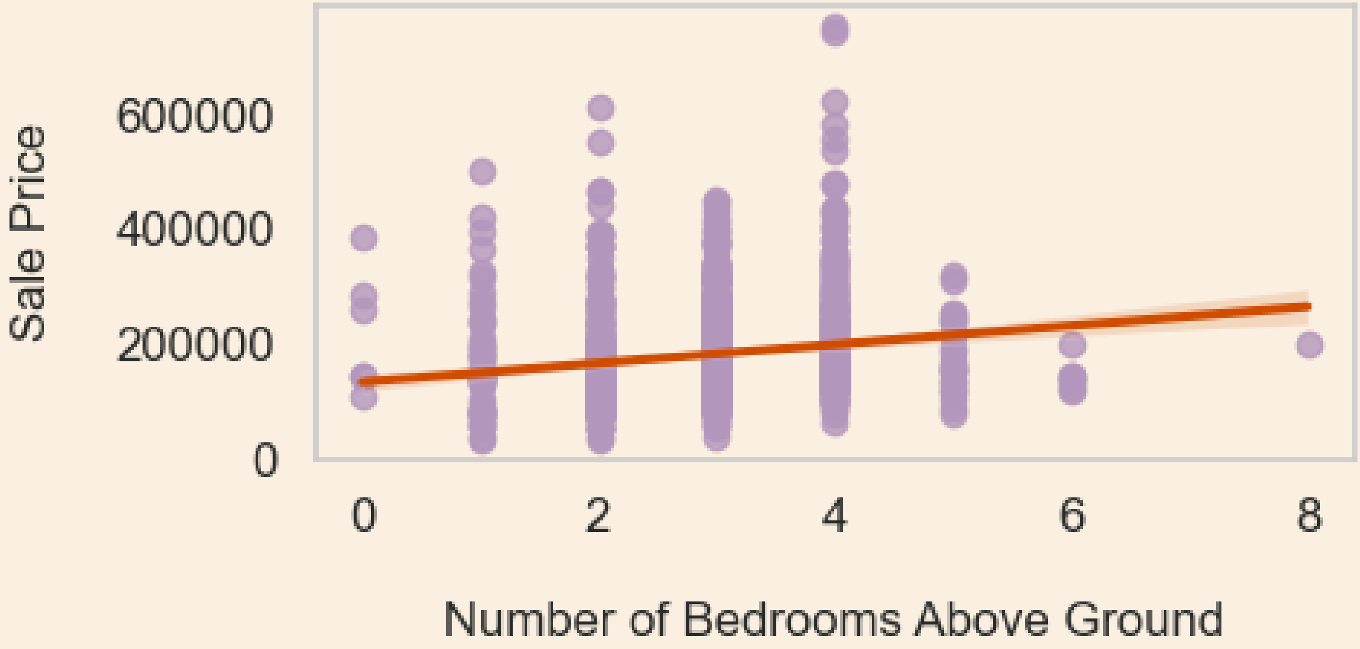
Figure 4. Street vs SalePrice



Causal Analysis

What is the estimated *causal effect* of each additional **bedroom** on **price**?

Figure 5. BedroomAbvGr vs SalePrice



Univariate Regression

Dependent Variable: Sale Price

Independent Variable: Number of Bedrooms

Result: β = $16,381.02

Limitations

- Adding a bedroom increases Sales Price by $16,381.02 on average? Incorrect
- Cannot infer causality due to **omitted variable bias**
- Estimation is biased

Causal Analysis

Confounding Variables

- Univariate regressions with both price and bedrooms
- Statistically significant predictors (p < 0.05) for both

Multivariate Regression

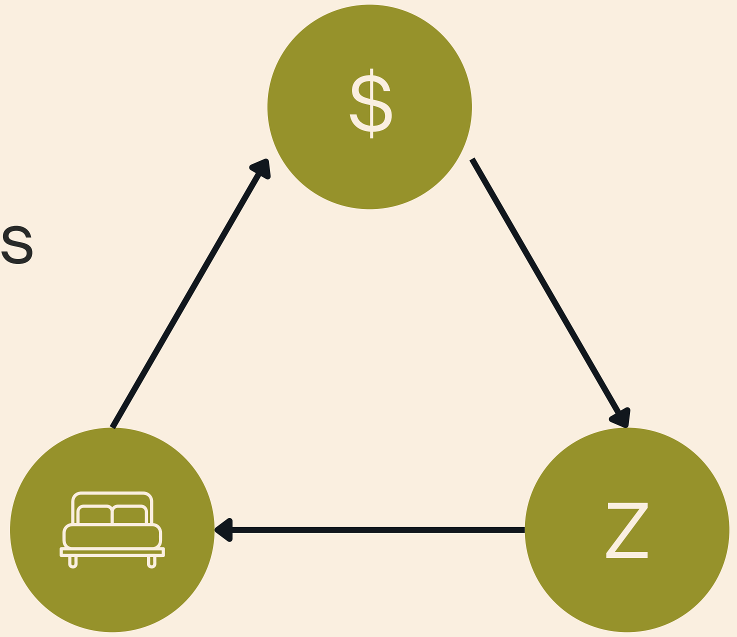
Dependent Variable: Sale Price

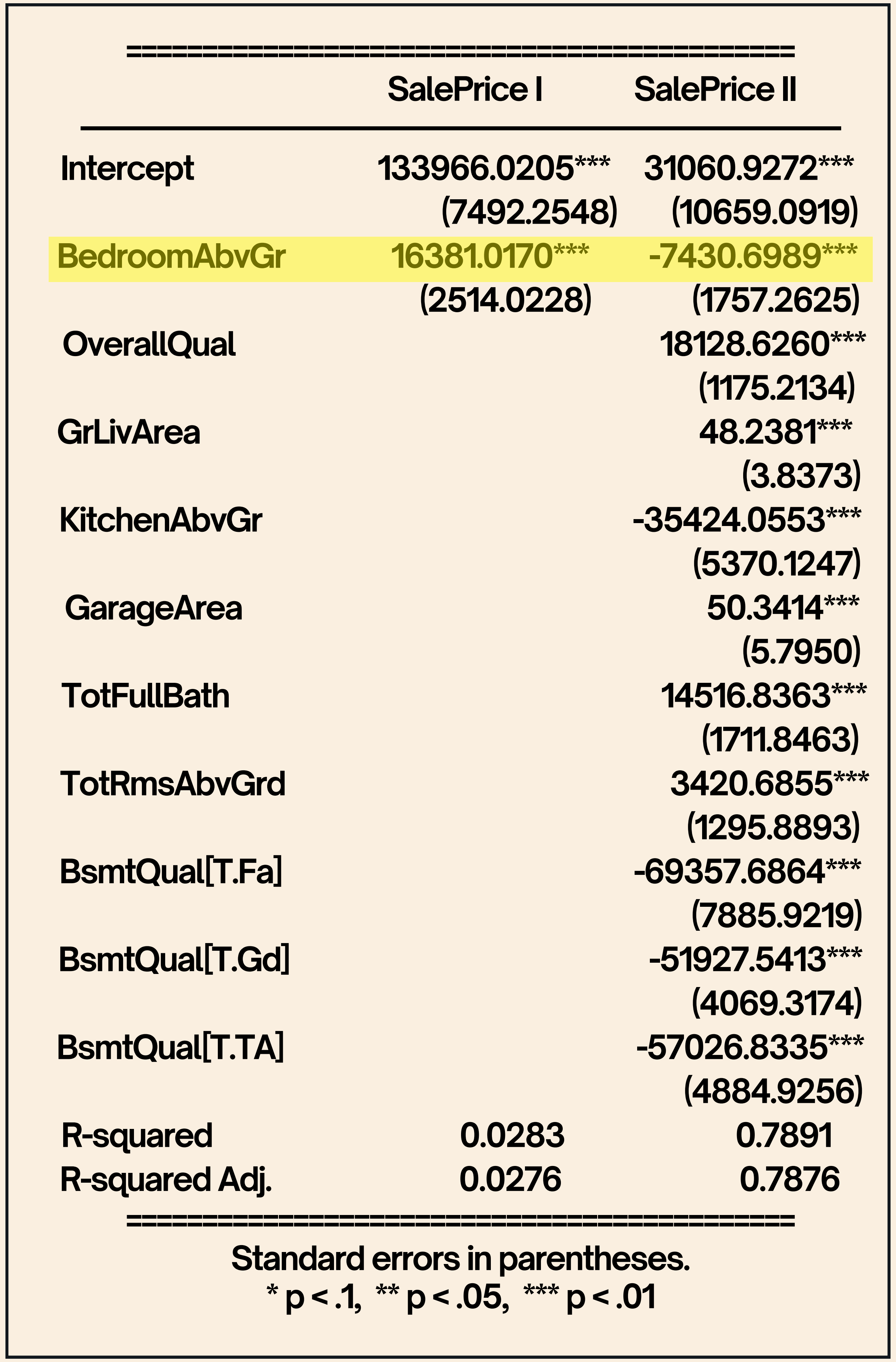
Independent Variables: Bedrooms + Controls

Result: β = -$7,430.70

Takeaways

- Less biased but still not definitive
- Potential for unmeasured confounding variables





	SalePrice I	SalePrice II
Intercept	133966.0205***	31060.9272***
	(7492.2548)	(10659.0919)
BedroomAbvGr	16381.0170***	-7430.6989***
	(2514.0228)	(1757.2625)
OverallQual		18128.6260***
		(1175.2134)
GrLivArea		48.2381***
		(3.8373)
KitchenAbvGr		-35424.0553***
		(5370.1247)
GarageArea		50.3414***
		(5.7950)
TotFullBath		14516.8363***
		(1711.8463)
TotRmsAbvGrd		3420.6855***
		(1295.8893)
BsmtQual[T.Fa]		-69357.6864***
		(7885.9219)
BsmtQual[T.Gd]		-51927.5413***
		(4069.3174)
BsmtQual[T.TA]		-57026.8335***
		(4884.9256)
R-squared	0.0283	0.7891
R-squared Adj.	0.0276	0.7876

Standard errors in parentheses.
* p<.1, ** p<.05, *** p<.01

Data Challenges

Transformation on Sale Price

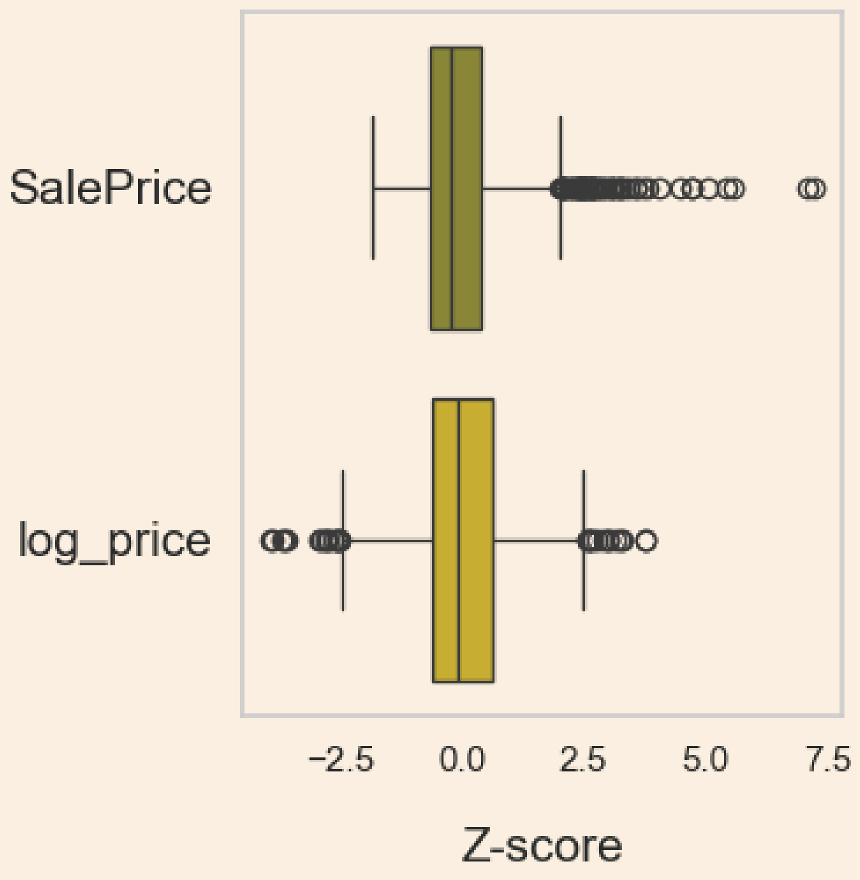


Figure 6

TotalSF Outliers

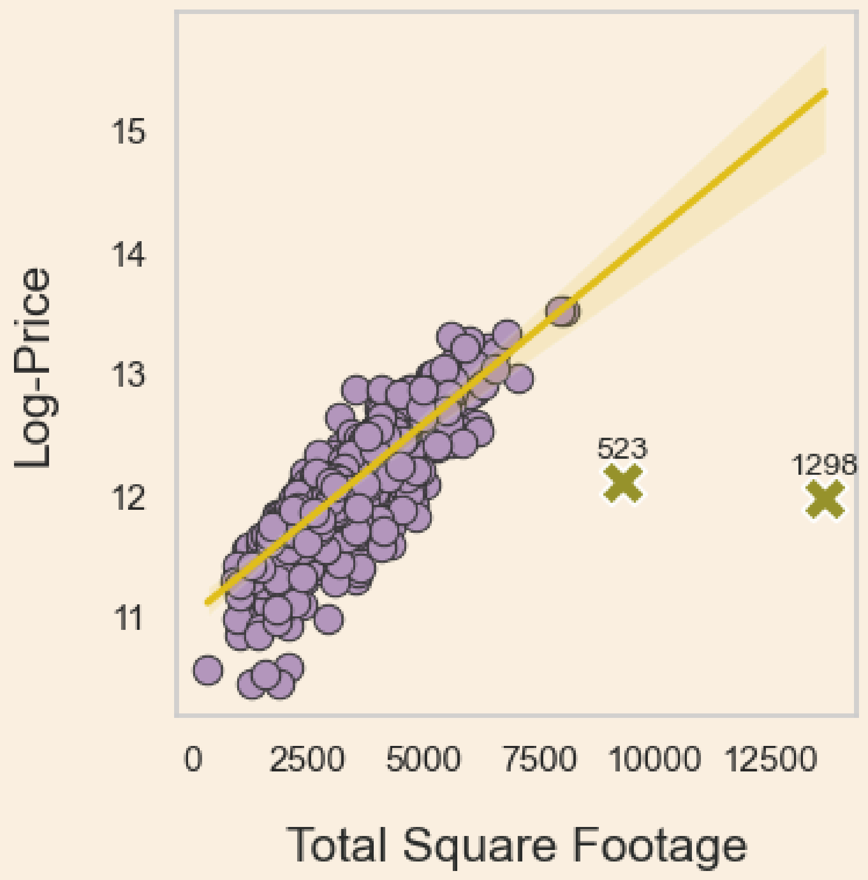
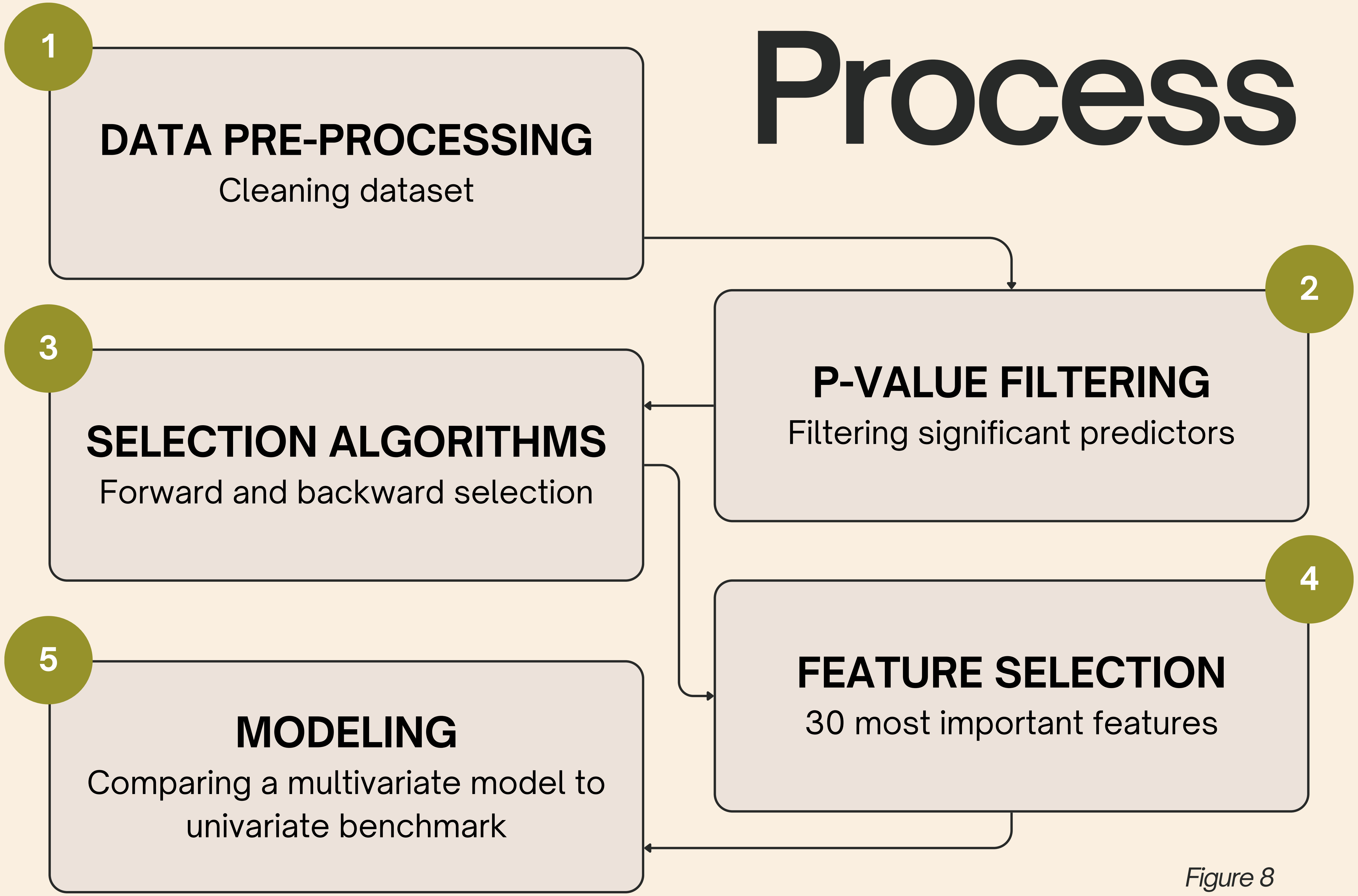


Figure 7

Key Issues	Solution
Missing data - unusually high amount of null values	Replacing with ‘0’ for numerical data, ‘None’ for categorical data
Price distribution - skewed right	Log transformation
Categorical variables - ill-equipped for regression analysis	Recoding data with dummy columns and numerical scales
Outliers - anomalies that are potentially influential	Identifying and removing data points using scatterplots
Redundant variables - several columns pertaining to one metric	Aggregating features

Predictive Model Process



1 **DATA PRE-PROCESSING**
Cleaning dataset

2 **P-VALUE FILTERING**
Filtering significant predictors

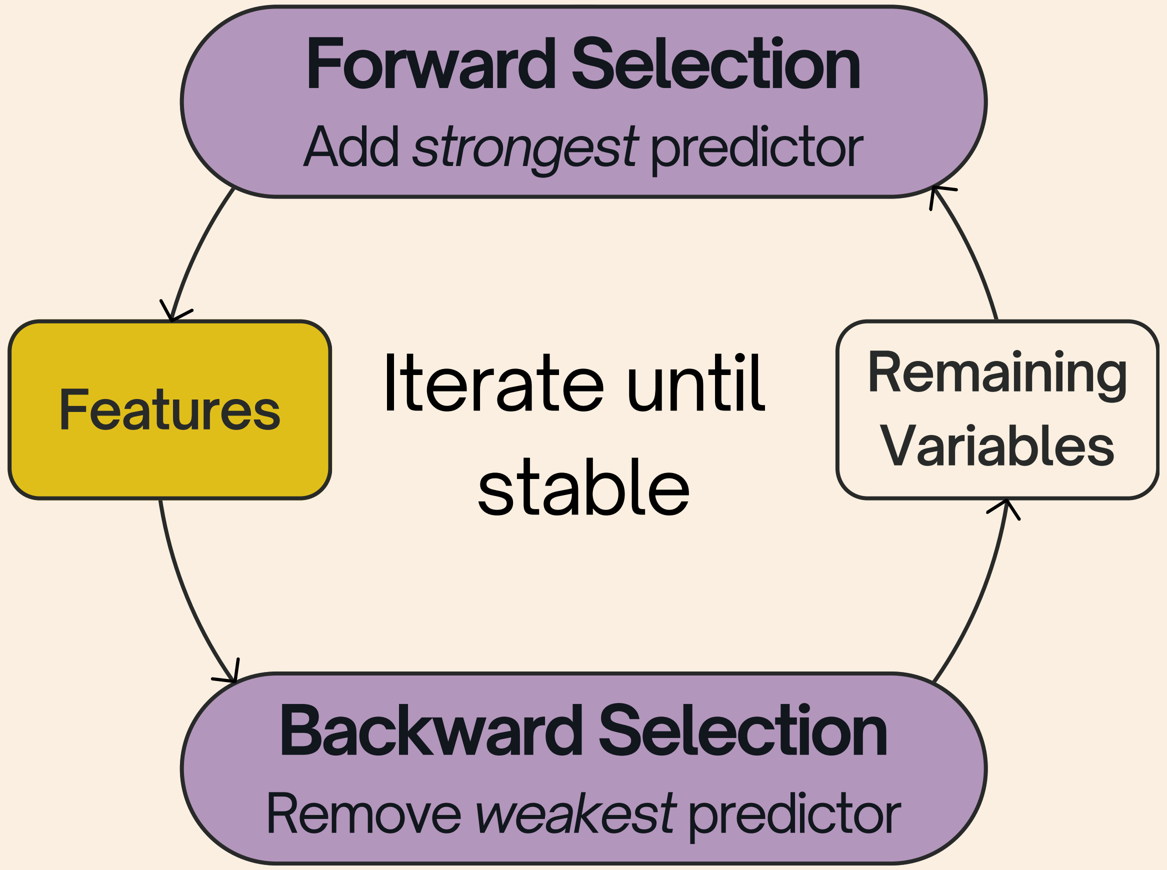
3 **SELECTION ALGORITHMS**
Forward and backward selection

4 **FEATURE SELECTION**
30 most important features

5 **MODELING**
Comparing a multivariate model to univariate benchmark

Figure 8

BI-DIRECTIONAL ELIMINATION



Forward Selection
Add *strongest* predictor

Features

Iterate until stable

Remaining Variables

Backward Selection
Remove *weakest* predictor

Figure 9

Criteria to finalize features

- ☑ Commonality across methods
- ☑ Non-redundancy
- ☑ Data structure relevance

Results

PREDICTIVE MODEL

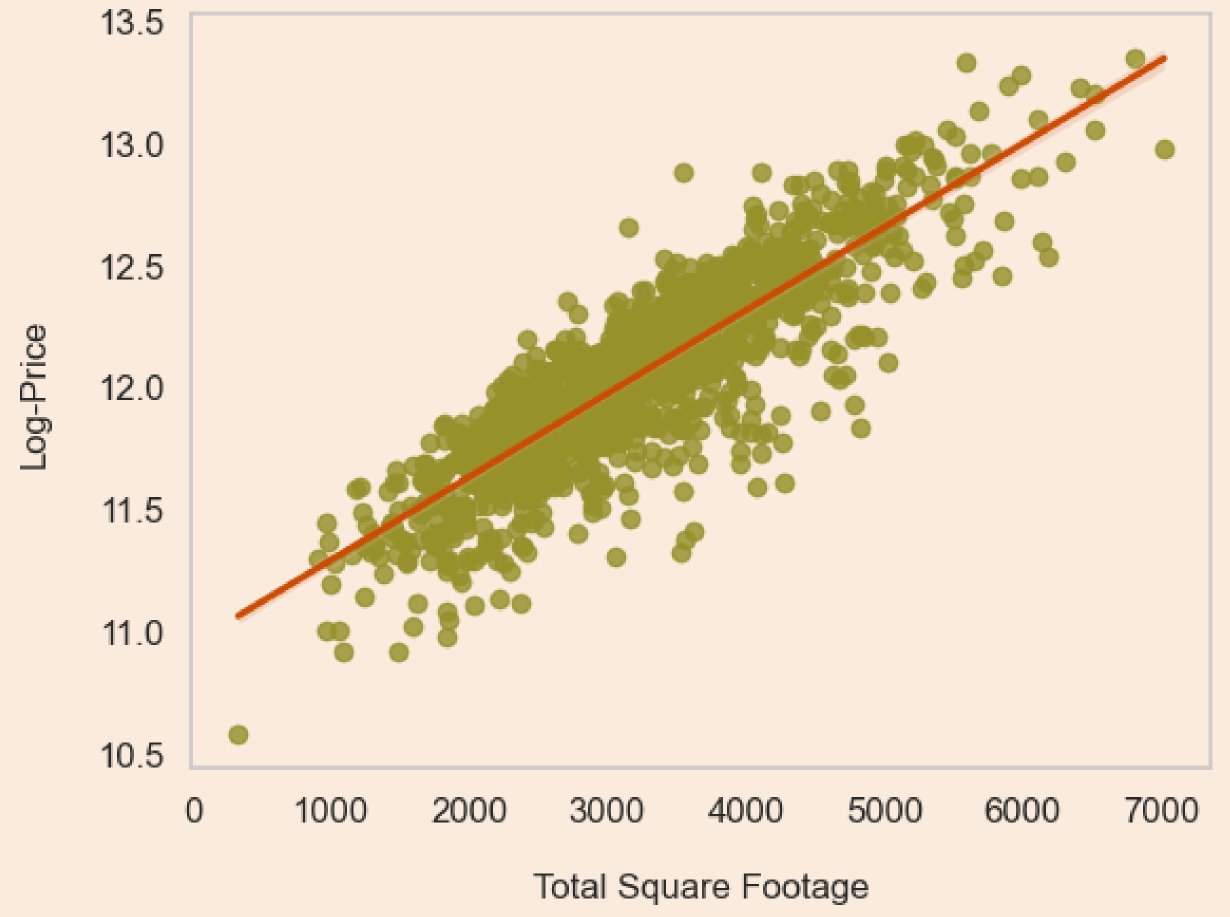
BENCHMARK

30 Predictors / 79 Possible Features

Continuous - area, year
Discrete - bathrooms, fireplaces
Ordinal - quality and condition related variables
Nominal - structural attributes, zoning, land configurations

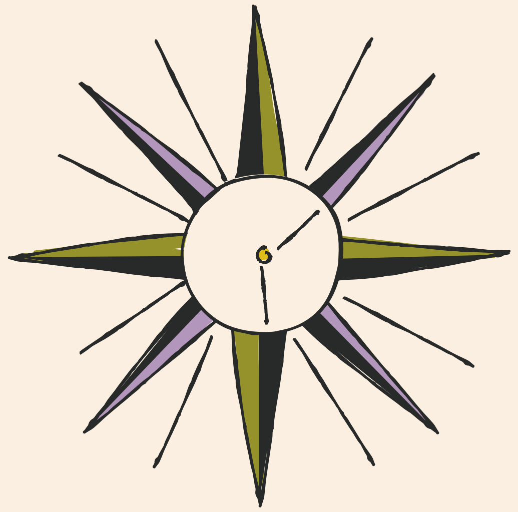
R^2 Adj.
0.93 > 0.75

Figure 10 **TotalSF vs Log Price**

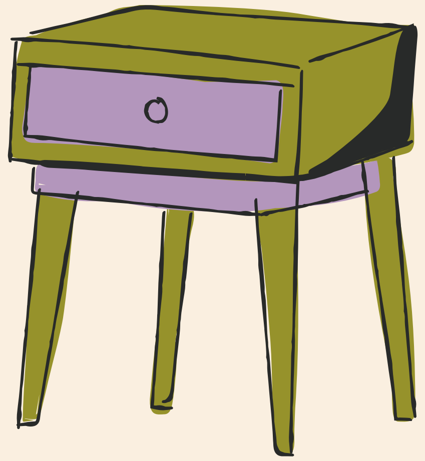


Our Takeaways

Understanding Causal Relationships



Data Challenges and Effective Solutions



Importance of Feature Selection

THANK YOU

For Your Attention

Any Questions?