

# King County Prices Prediction

Astrid Archundia  
Marcos Duran  
Hugo Enriquez

# WORKFLOW

- Data extraction
- Data Cleaning
- EDA
- Data pre-processing
- Modeling
- Trying differents models
- Final conclusions

# Data extraction

- Exploring the data
  - .info
  - .shape
  - .describe()
  - .head()
- Understanding the data
- Questioning ourselves



# Data cleaning

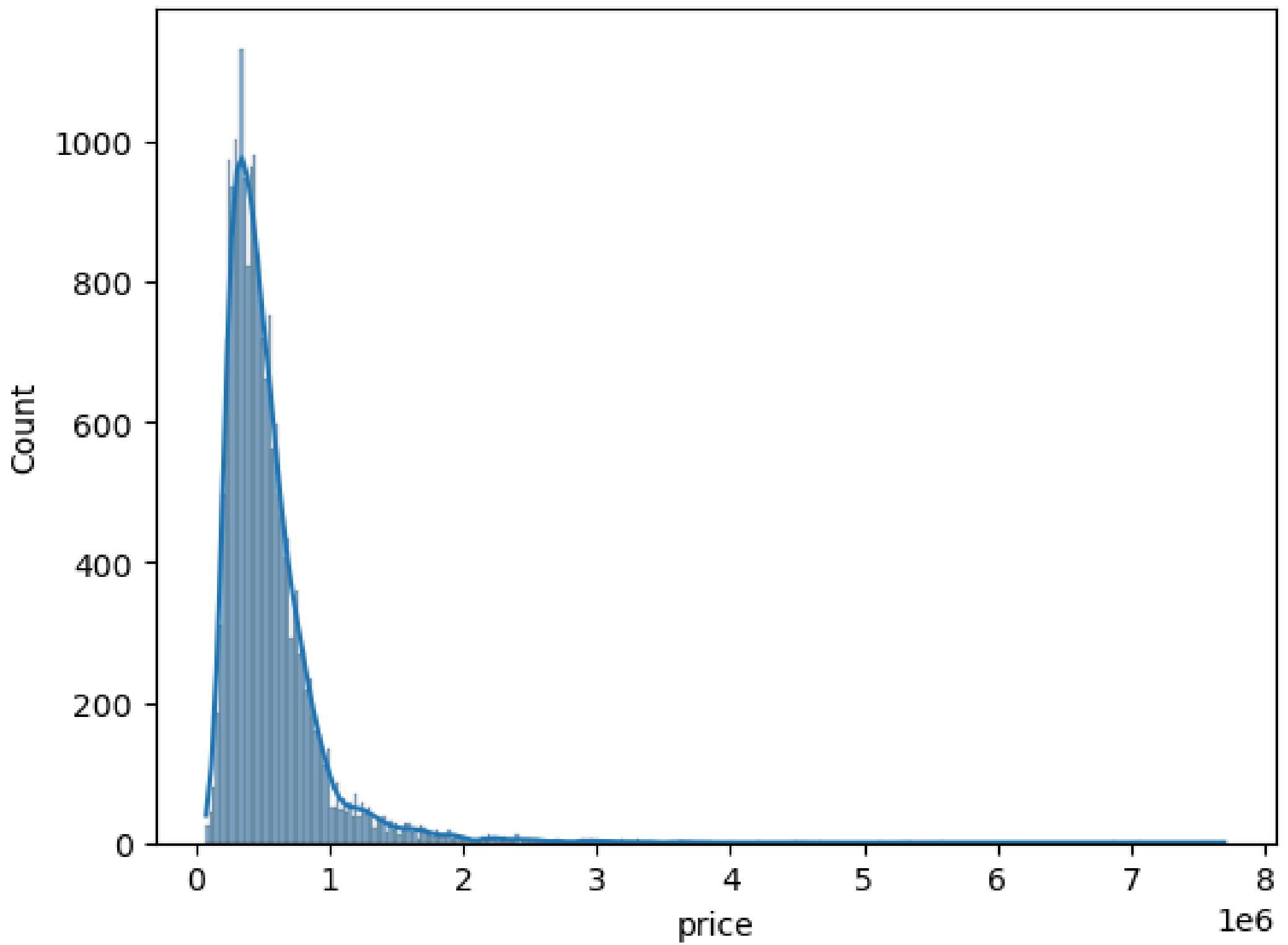
- Dealing with:
  - NaN's
  - Empty Spaces
  - Duplicates
- Changing index. Index → Id's
- Dropping columns:
  - Date
  - Condition

# EDA

- Checking distribution
- Data
- Target
- Checking outliers
- Correlations

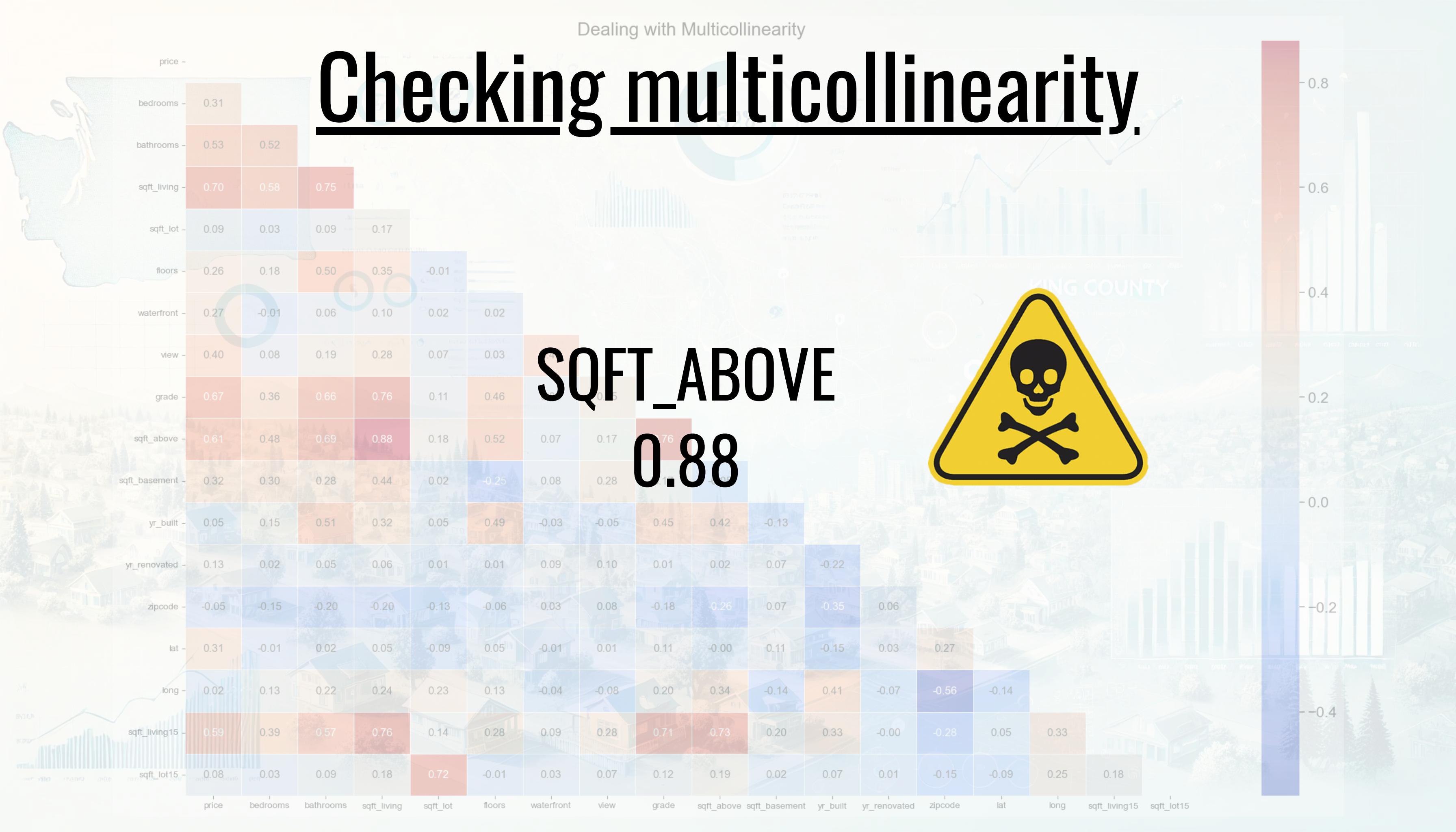


# Target distribution



# Checking multicollinearity.

SQFT ABOVE  
0.88



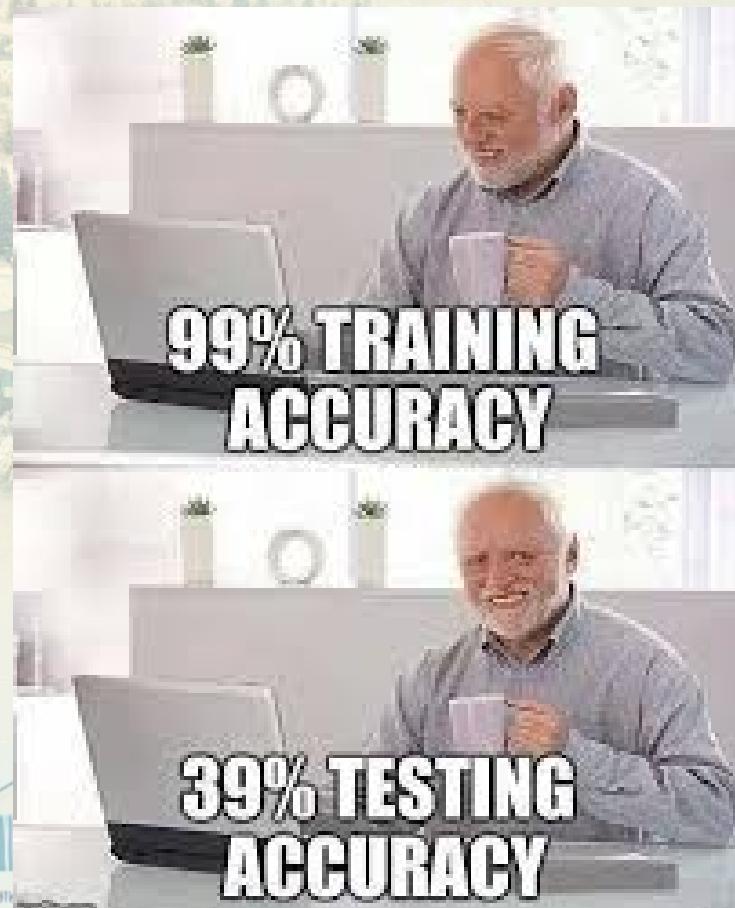
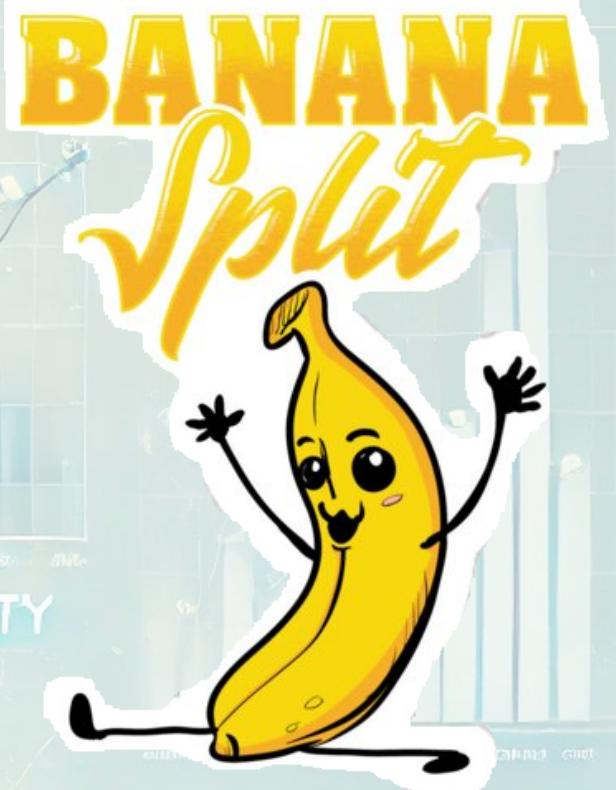
# Data Pre-processing

X - y split

100% of our data: 21613

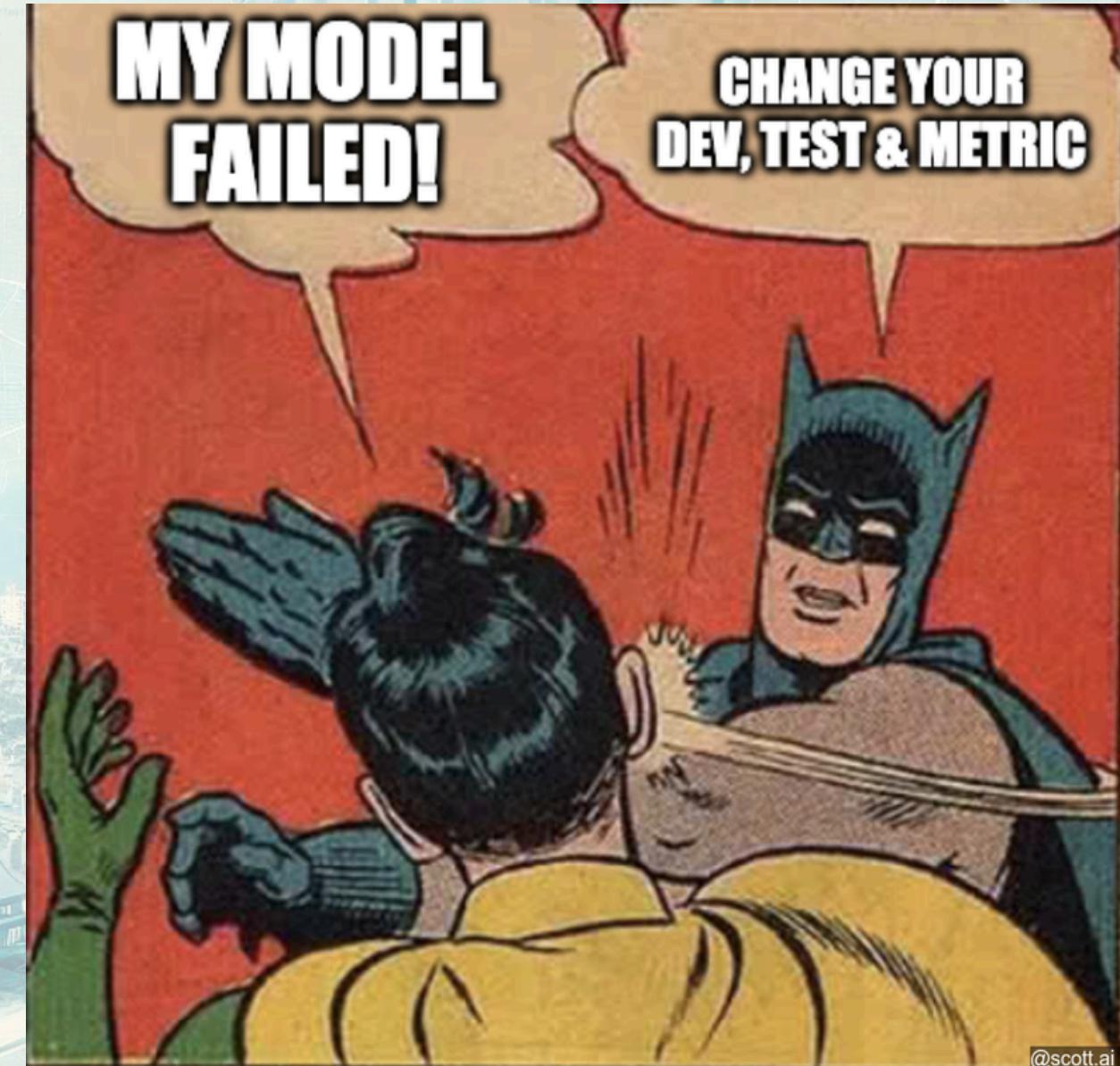
70% for training data: 15129

30% for test data: 6484



# Modeling

- 'Linear Regression'
- 'Ridge',  
'Lasso',
- 'Decision  
Tree'
- 'KNN',
- 'XGBoost'



@scott.ai

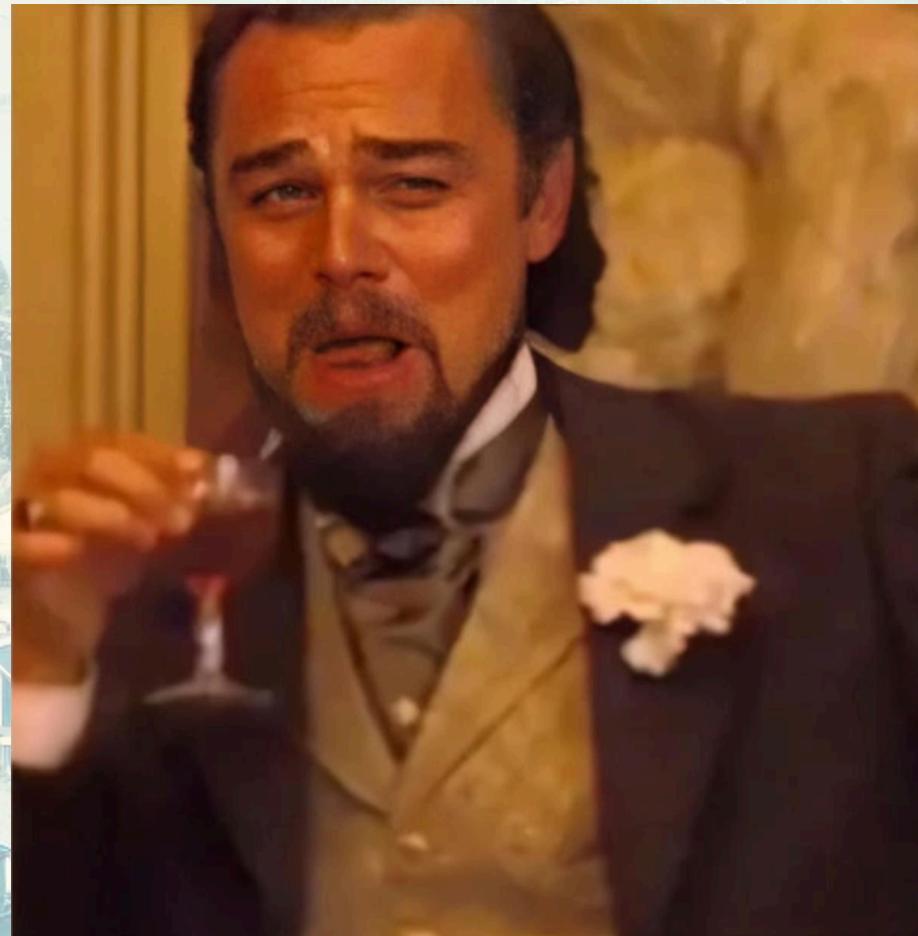
# Results

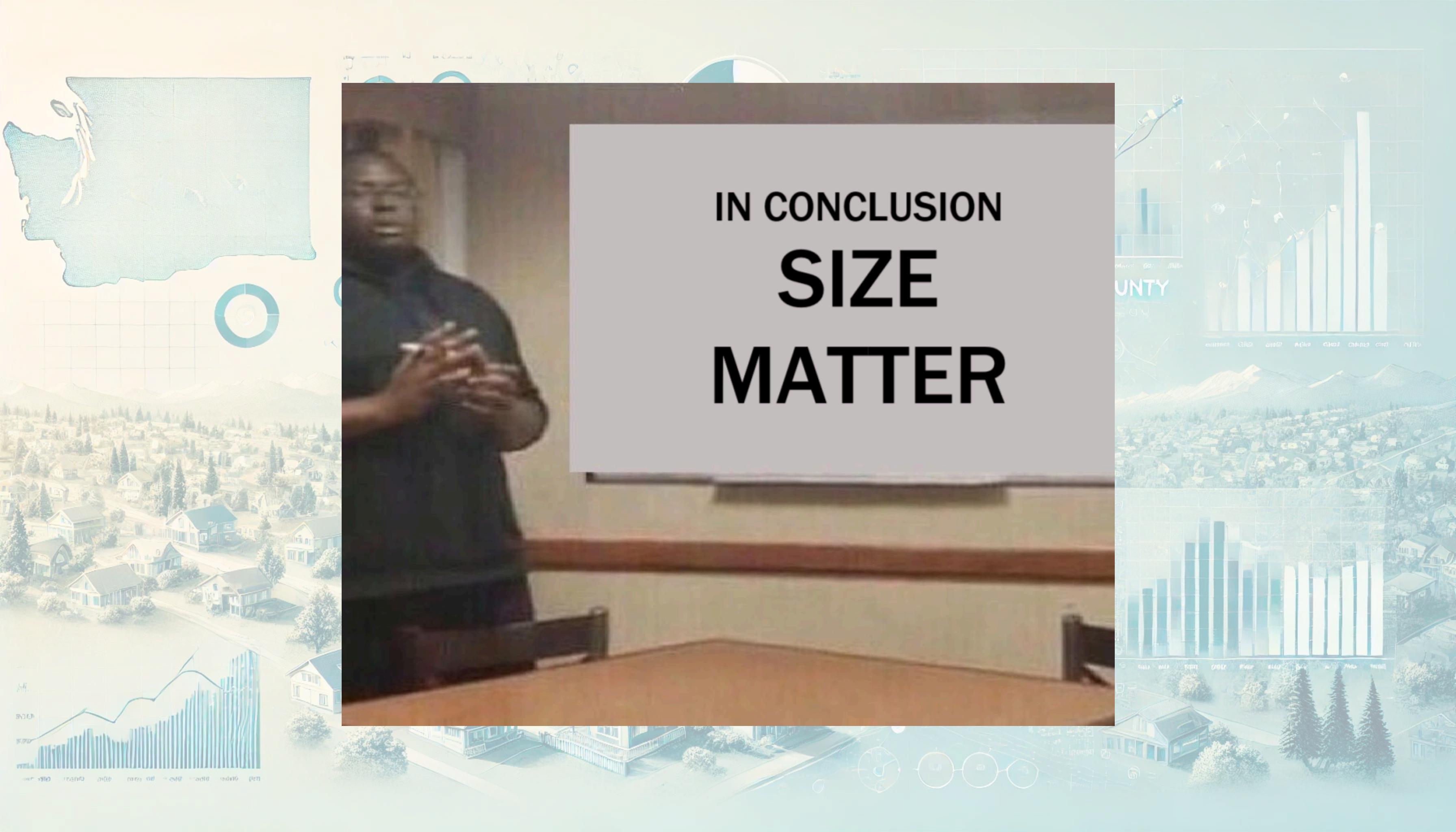
Model	R <sup>2</sup>
Linear Regression	0.6970
Ridge	0.6969
Lasso	0.6970
Decision Tree	0.7557
KNN	0.4932
XGBoost	0.7672

Model	R <sup>2</sup>
MinMaxScaler	0.8633
StandardScaler	0.8633
Log_Transform	0.8300

Model	R <sup>2</sup>
Linear_Regression_Above	0.6970
Linear_Regression_650	0.6034

# More square feet, more zeros to meet !





The background of the collage features a landscape with mountains and houses, overlaid with various data visualizations including a map of Washington state, a pie chart, a line graph, bar charts, and network diagrams.

**IN CONCLUSION  
SIZE  
MATTER**