Predicting Electricity Usage

Linear Regression & Web Scraping

Jason Dunleavy, Data Scientist


Introduction
Predictors
Background and Purpose

Introduction

Case Study
- Louisiana

Use Cases
- Infrastructure planning
- Corporate budgeting
- Rate cases
- Electric market operations

Time Periods
- Hourly
- Daily
- Monthly*
- Seasonal
- Annual

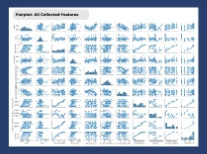
*Monthly data used for this project

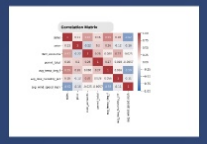

Predictors
Weather
• Temperature
• Humidity
• Wind speed
Population
• State population growth
• Number of customer accounts
Economic
• Real GDP
• Personal income
• Consumer spending
• New building permits
• Price of electricity


Data Gathering
Feature Engineering
Methodologies
Exploring Feature Relationships
LASSO Regression

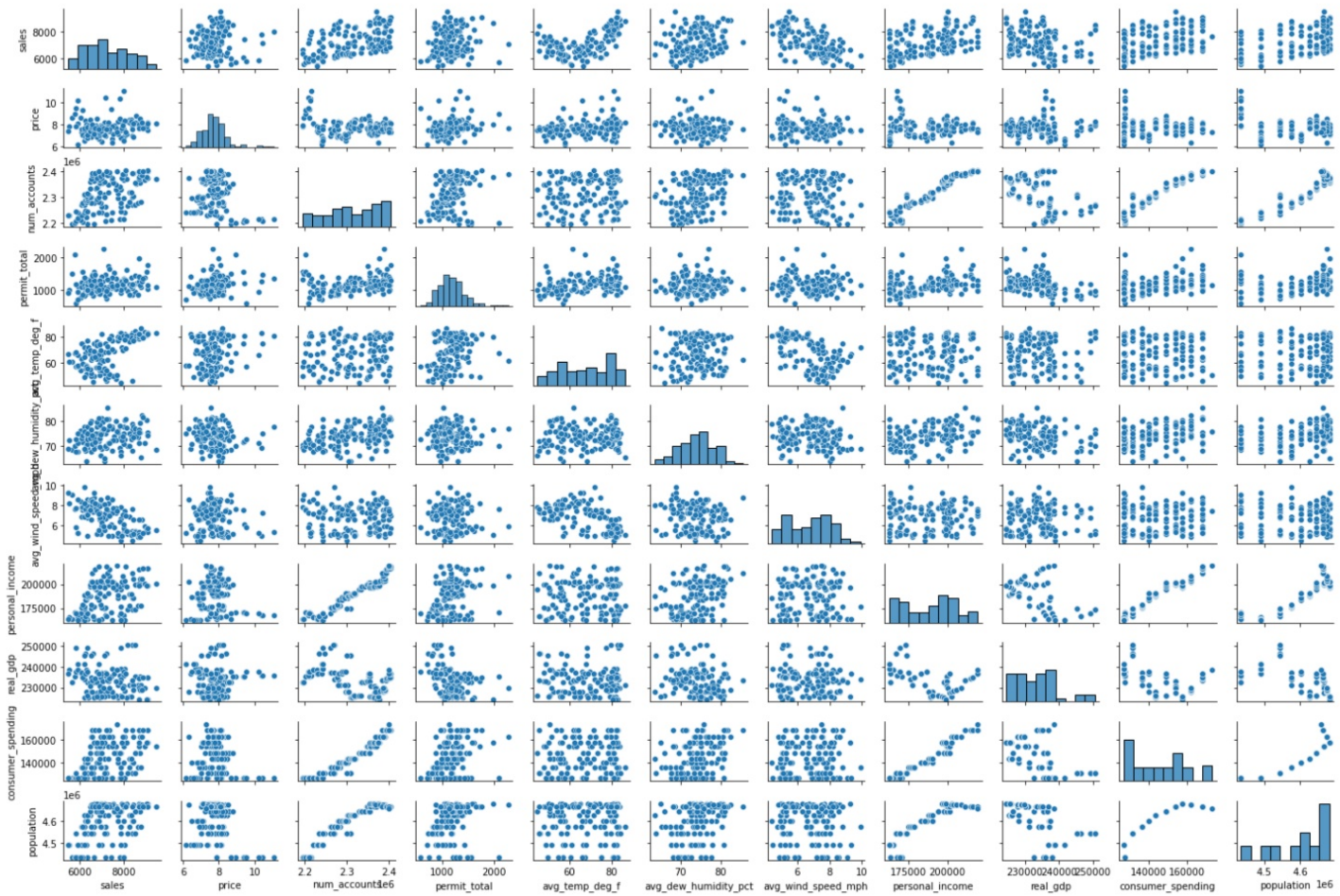

Data Gathering
Beautiful Soup
United States Census Bureau
Se
WEATHER UNDERGROUND
*see appendix for all data sources

Exploring Feature Relationships

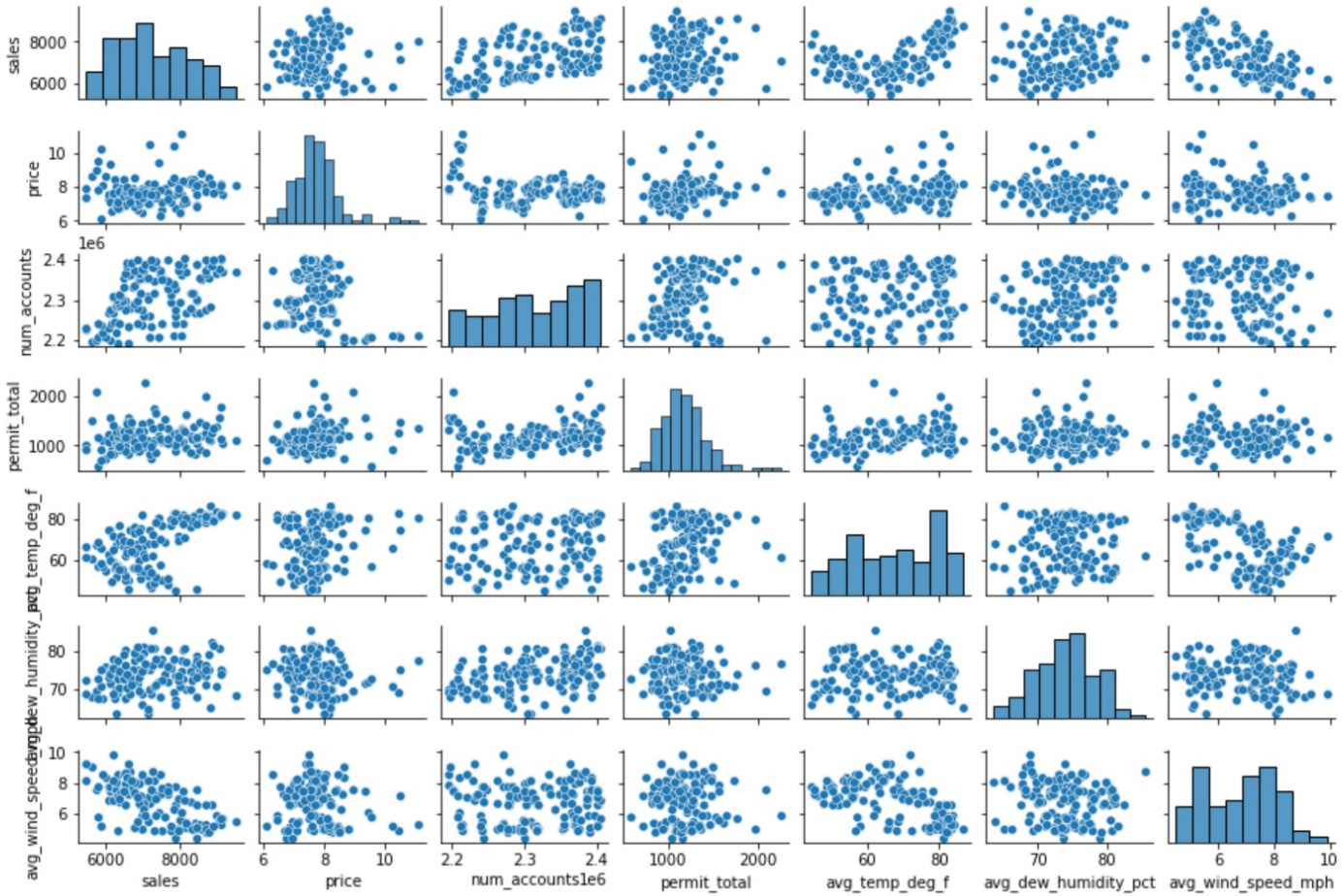




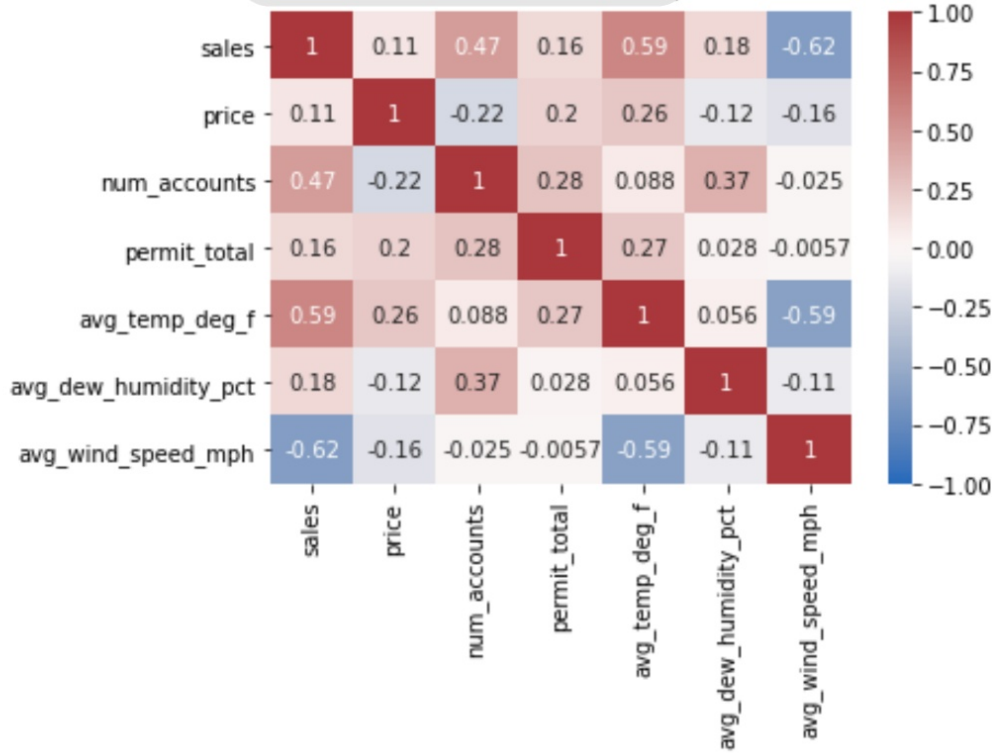
Pairplot: All Collected Features



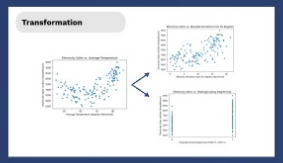
Pairplot: Avoiding Multicollinearity

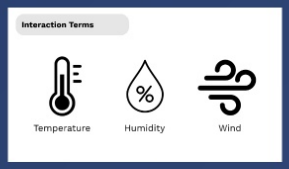


Correlation Matrix

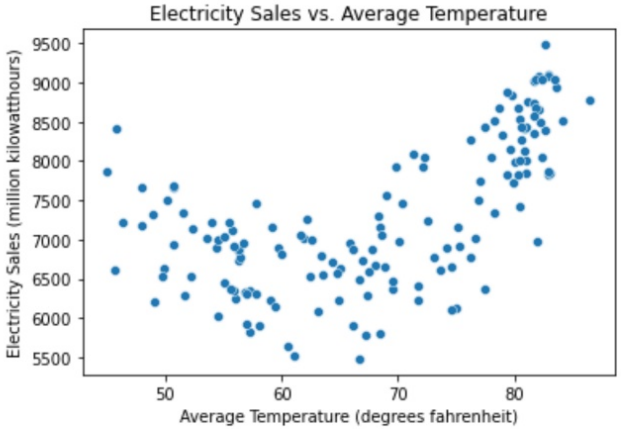


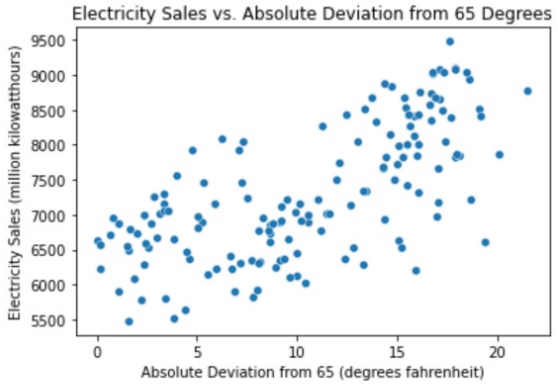
Feature Engineering

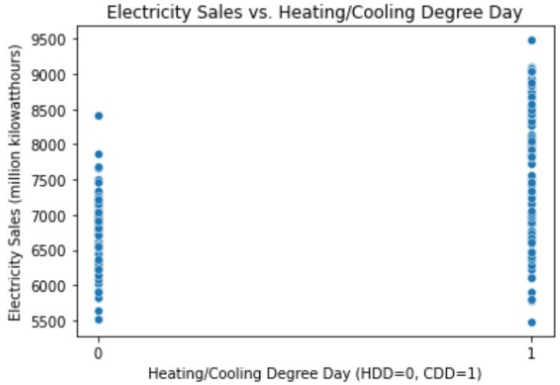




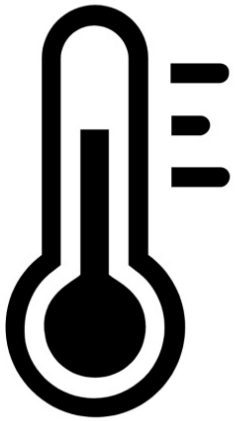
Transformation



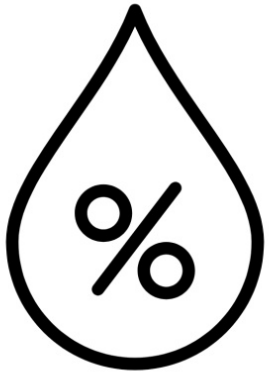




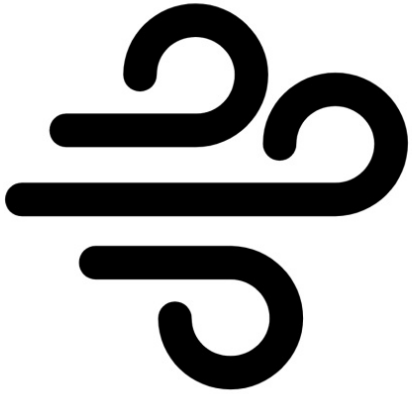
Interaction Terms



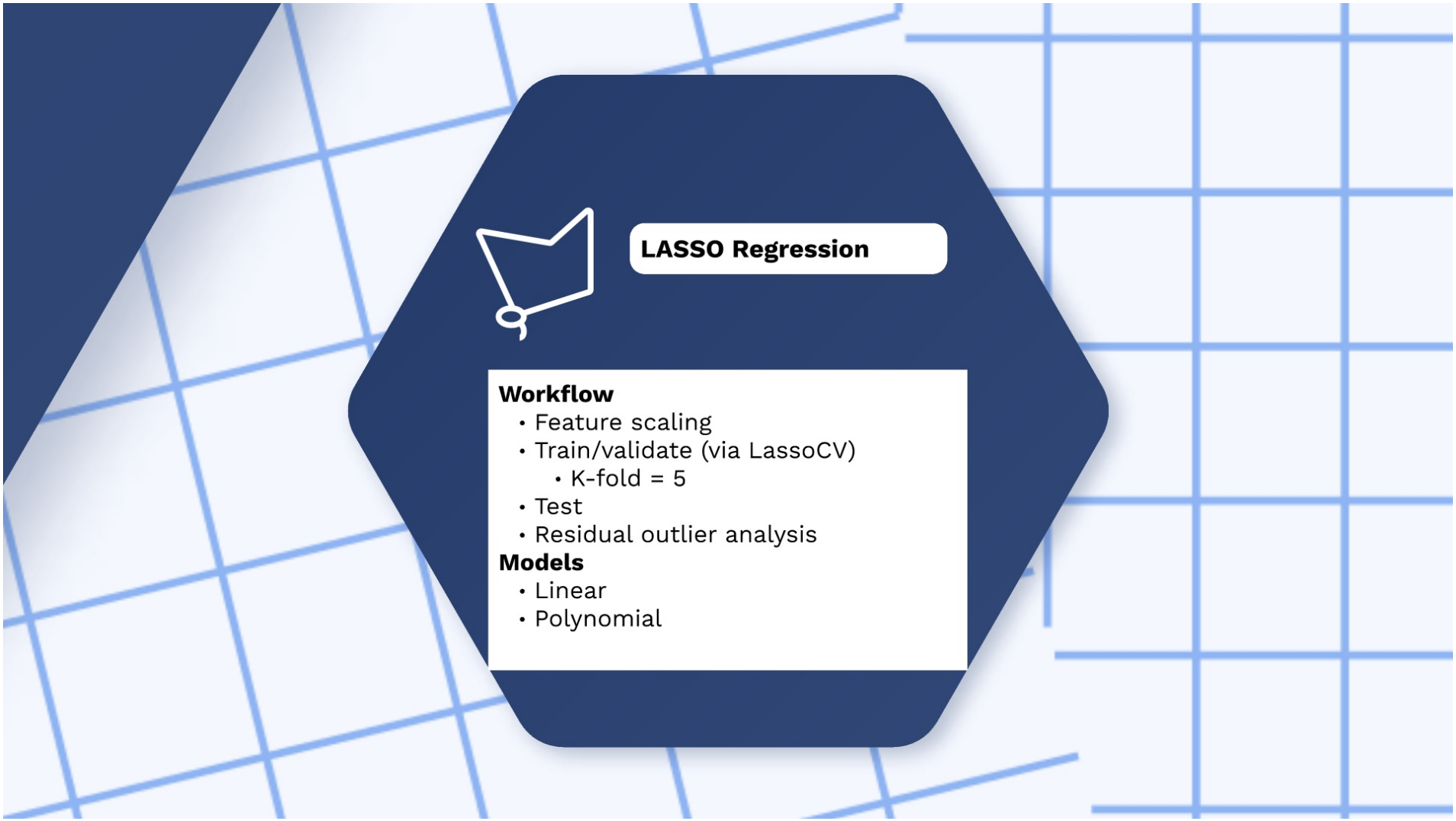
Temperature



Humidity



Wind

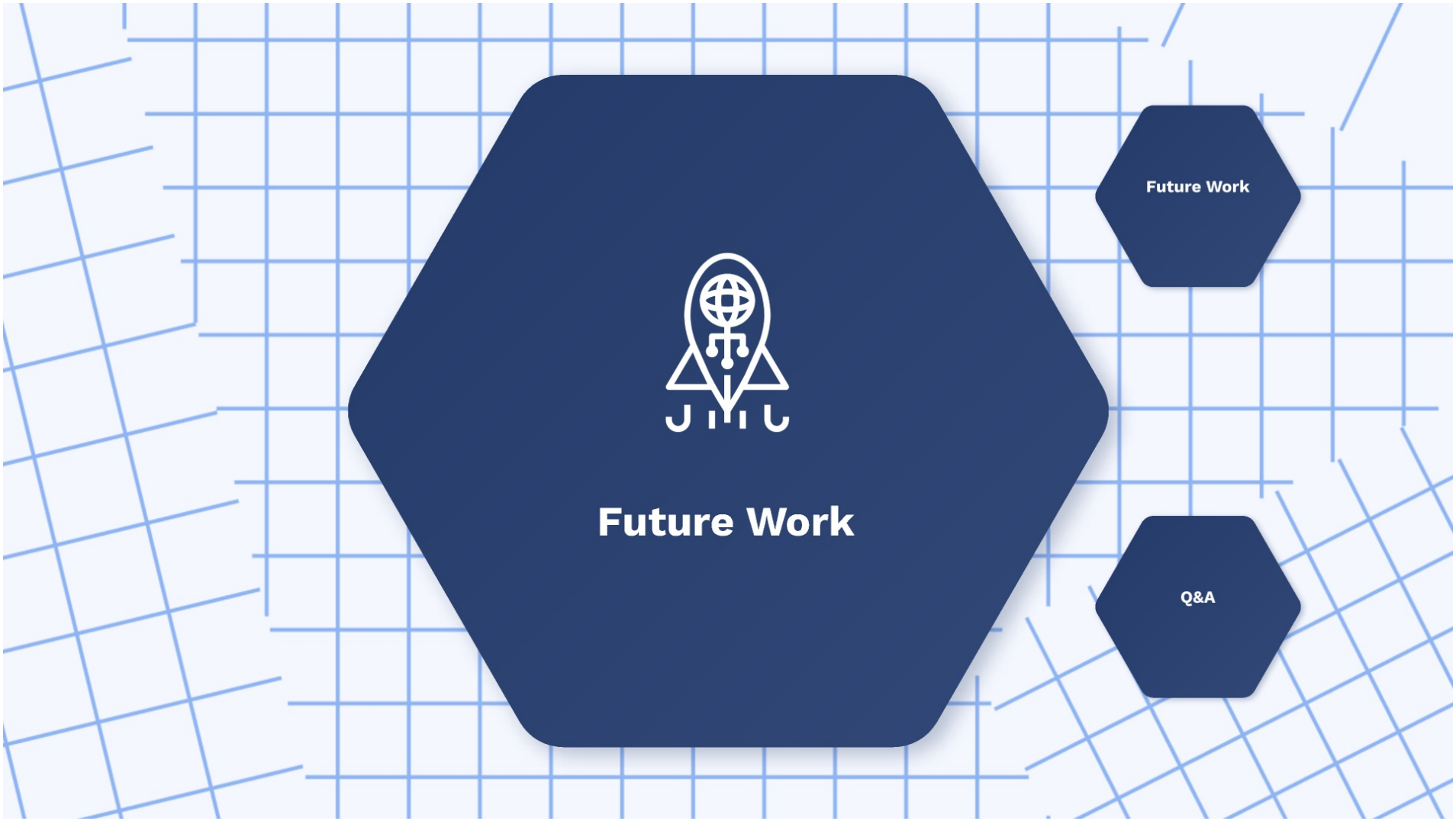

LASSO Regression
Workflow
• Feature scaling
• Train/validate (via LassoCV)
• K-fold = 5
• Test
• Residual outlier analysis
Models
• Linear
• Polynomial


Findings
R-Squared
Coefficients


R-Squared
Training
0.863
Testing
0.856

Coefficients

feature	coefficient
price	70.79031
num_accounts	441.91271
permit_total	-29.52717
avg_dew_humidity_pct	-364.36851
avg_wind_speed_mph	-834.69014
temp_deviation	383.48239
HDD_CDD	51.06585
hum_wind	801.70915
hum_temp	392.57398
wind_temp	-137.31290


Future Work
Future Work
Q&A

Future Work

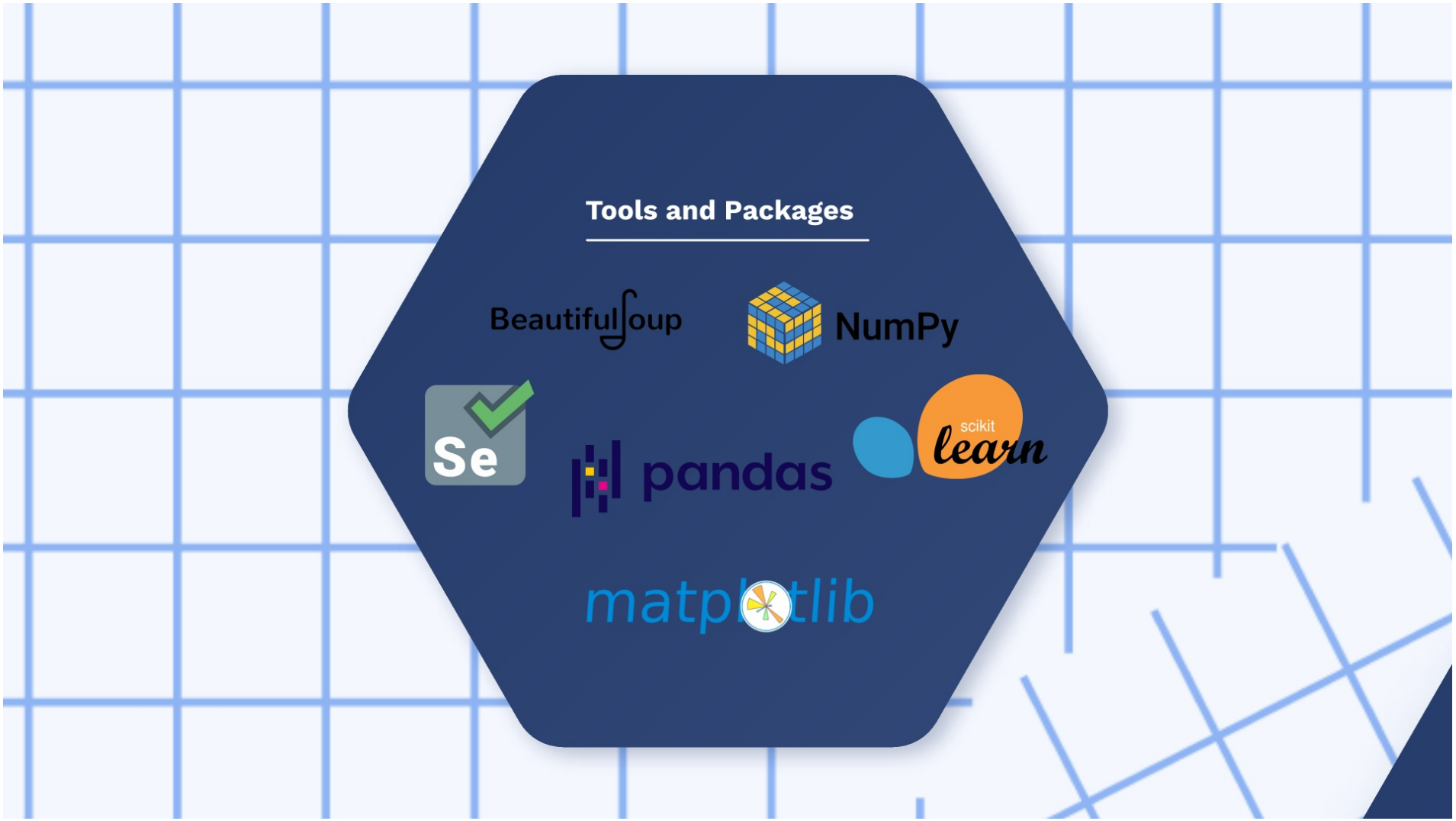
Additional Data

- In-house data
- More discrete time-periods

Other Models

- Artificial Neural Networks (ANN)
- Support-Vector Machines (SVMs)
- Random Forests (RFs)
- Autoregressive Integrated Moving Average (ARIMA)


Tools and Packages
Data Sources
Linear - Selected Features
Appendix
Linear - All Features
R-Squared


Tools and Packages
BeautifulSoup
NumPy
Se
pandas
scikit
learn
matplotlib

Data Sources

U.S. Energy Information Administration
- https://www.eia.gov/

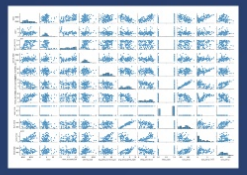
WeatherUnderground
- https://www.wunderground.com/

Population
- https://www.macrotrends.net/states/louisiana/population

Building Permits
- https://www.census.gov/construction/bps/statemonthly.html

U.S. Department of Commerce - Bureau of Economic Analysis
- https://www.bea.gov/

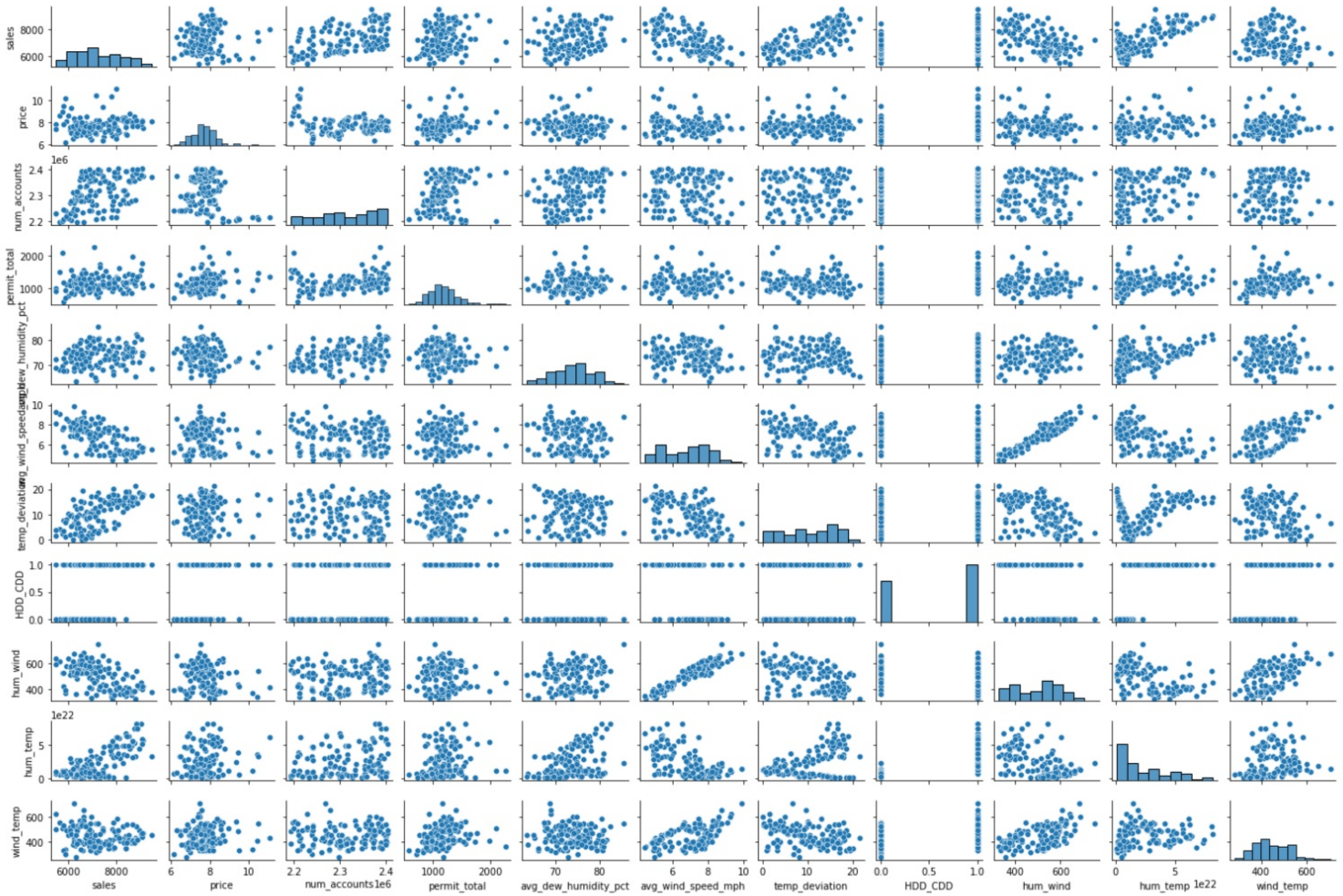
Linear - Selected Features



Additional Information

Mean Absolute Error
- 294.45

Residual Outlier Analysis
- Removed two outliers (residual > 900)
- Resulting r-squared:
  - Test: 0.886
  - Train: 0.859


sales
price
num_accounts
permit_total
avg_dew_humidity_pct
avg_wind_speed_mph
temp_deviation
HDD_CDD
hum_wind
hum_temp
wind_temp
1e6
1e22

Additional Information

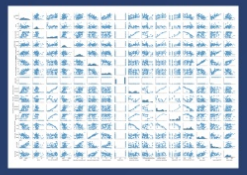
Mean Absolute Error

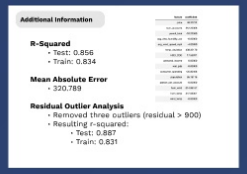
- 294.45

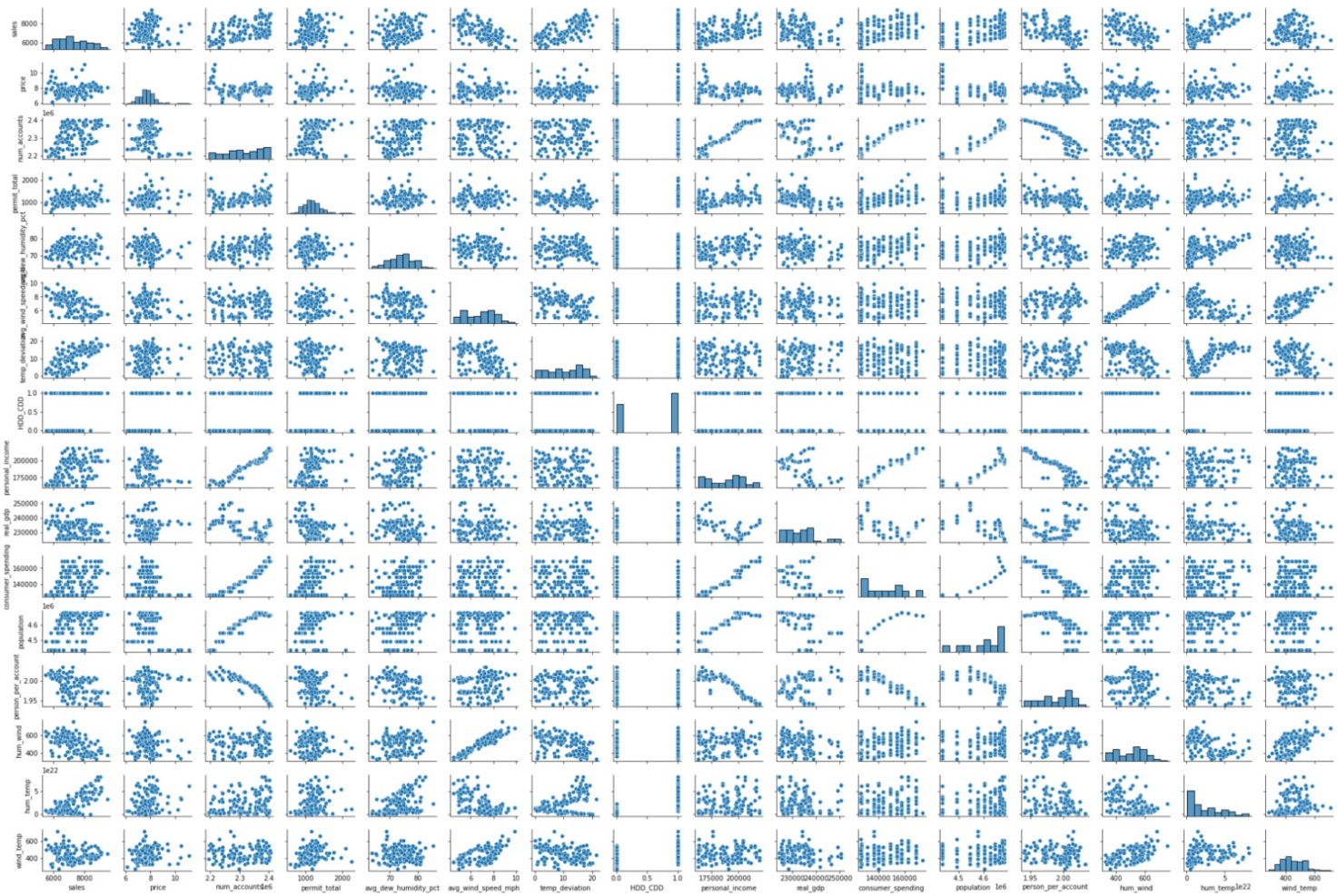
Residual Outlier Analysis

- Removed two outliers (residual > 900)
- Resulting r-squared:
    - Test: 0.886
    - Train: 0.859

Linear - All Features






sales
price
num_accounts
permit_total
avg_dew_humidity_pct
avg_wind_speed_mph
temp_deviation
HDD_CDD
personal_income
real_gdp
consumer_spending
population
person_per_account
hum_wind
hum_temp
wind_temp

Additional Information

feature	coefficient
price	48.37212
num_accounts	212.22306
permit_total	-10.22580
avg_dew_humidity_pct	-0.00000
avg_wind_speed_mph	-4.02690
temp_deviation	408.35170
HDD_CDD	17.59011
personal_income	0.00000
real_gdp	-0.00000
consumer_spending	125.80466
population	85.18115
person_per_account	-0.00000
hum_wind	-214.99167
hum_temp	247.09307
wind_temp	-0.00000

R-Squared

- Test: 0.856
- Train: 0.834

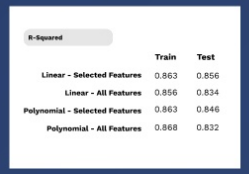
Mean Absolute Error

- 320.789

Residual Outlier Analysis

- Removed three outliers (residual > 900)
- Resulting r-squared:
  - Test: 0.887
  - Train: 0.831

R-Squared

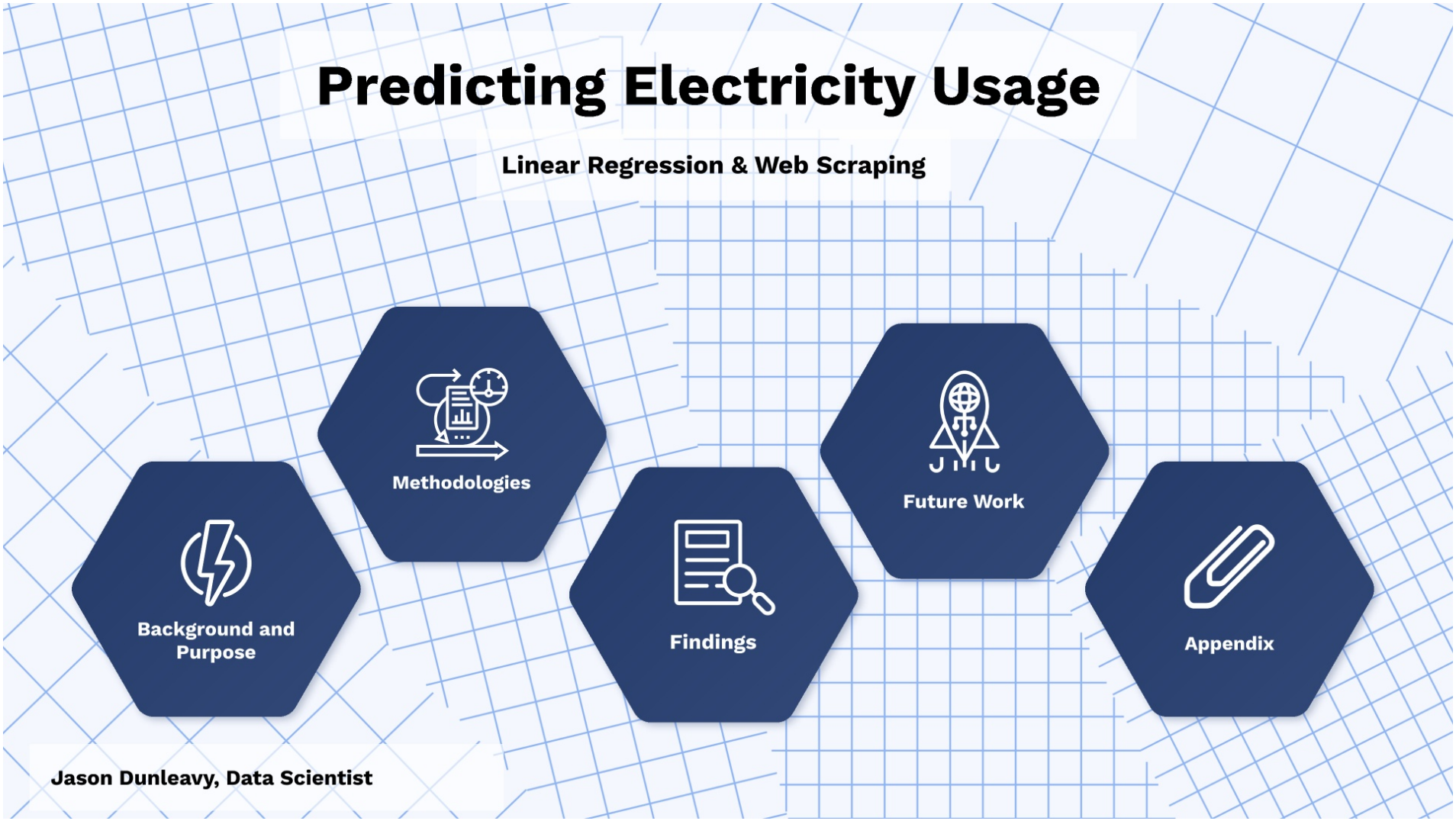


R-Squared

	Train	Test
Linear - Selected Features	0.863	0.856
Linear - All Features	0.856	0.834
Polynomial - Selected Features	0.863	0.846
Polynomial - All Features	0.868	0.832

R-Squared

	Train	Test
Linear - Selected Features	0.863	0.856
Linear - All Features	0.856	0.834
Polynomial - Selected Features	0.863	0.846
Polynomial - All Features	0.868	0.832


Predicting Electricity Usage
Linear Regression & Web Scraping
Background and Purpose
Methodologies
Findings
Future Work
Appendix
Jason Dunleavy, Data Scientist