REPORT

TASK: Find publicly available data for key factors that influence US home prices nationally. Then, build a data science model that explains how these factors impacted home prices over the last 20 years. Use the S&P Case-Schiller Home Price Index as a proxy for home prices: [fred.stlouisfed.org/series/CSUSHPISA]

DATA: The date range of all datasets was taken from 01-07-2003 to 01-07-2023 (20 years). To keep the data consistent, datasets focusing on single family units was considered.

Based on the literature review done, the following data was collected:

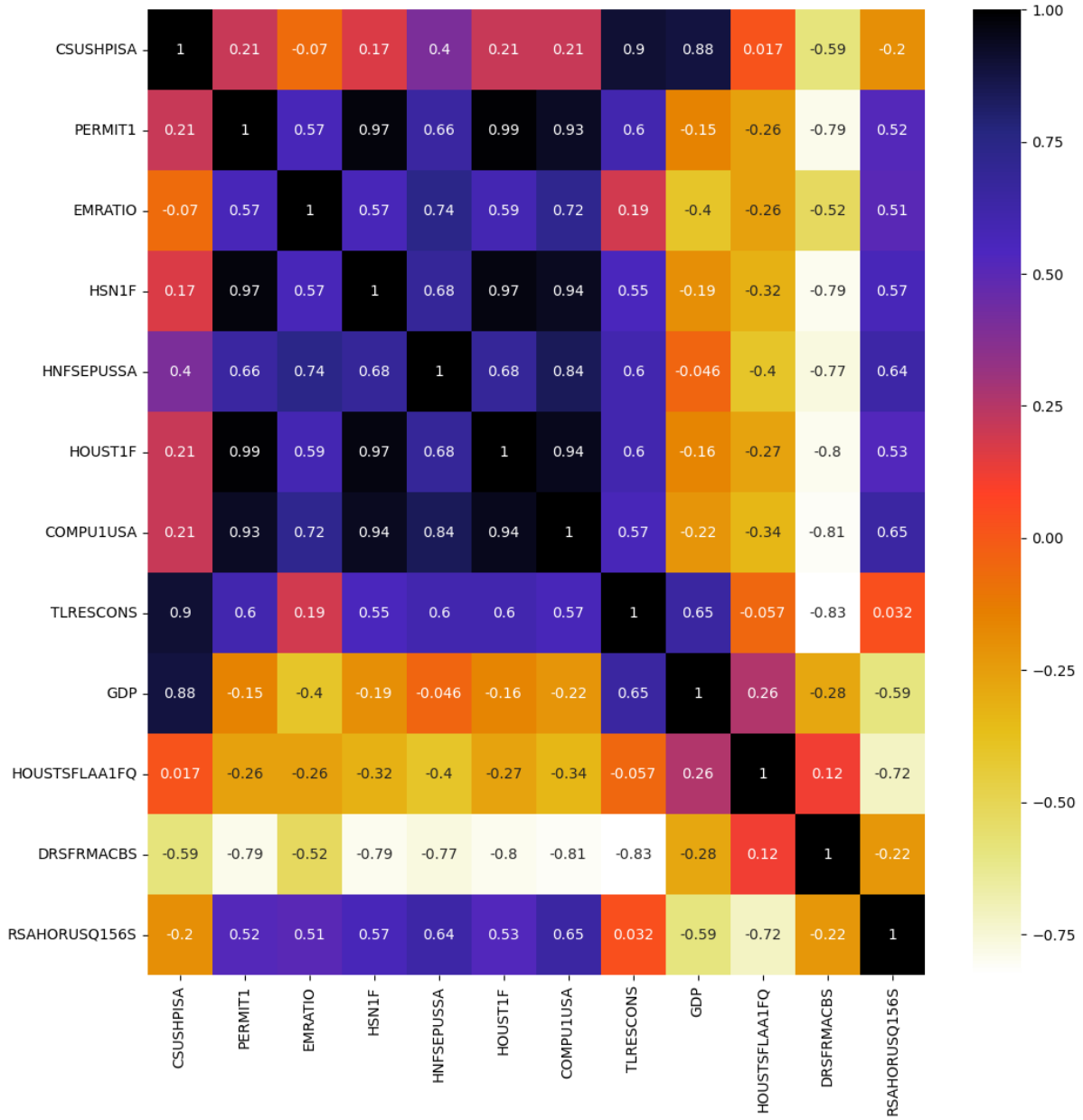
KEY	DESCRIPTION	TYPE
CSUSHPISA	S&P/Case-Shiller U.S. National Home Price Index	Monthly
GDP	Gross Domestic Product	Quarterly
HOUSTSFLAA1FQ	New Privately Owned Housing Starts in the United States, Average Square Feet of Floor Area for One-Family Units	Quarterly
PERMIT1	New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Single-Family Units	Monthly
EMRATIO	Employment-Population Ratio	Monthly
HSN1F	New One Family Houses Sold: United States	Monthly
DRSFRMACBS	Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks	Quarterly
RSAHORUSQ156S	Homeownership Rate in the United States	Quarterly
HNFSEPUSSA	New One Family Homes for Sale in the United States	Monthly
HOUST1F	New Privately-Owned Housing Units Started: Single-Family Units	Monthly
COMPU1USA	New Privately-Owned Housing Units Completed: Single-Family Units	Monthly
TLRESCONS	Total Construction Spending: Residential in the United States	Monthly

DATA PREPROCESSING

- It was decided to bring all the data into a standard format (monthly). The time variable in some of the datasets were in 'quarterly' format. They were converted into monthly format using the method of interpolation to fill in the missing values.
- These datasets were then all merged and saved into a .csv file (housing_full.csv).
- Standardization was applied on all columns of the dataset, as the range of values taken by different features of the dataset were not consistent.

HEATMAP ANALYSIS

A heatmap was created to find out correlations between all the features of the dataset.



The first row of the heatmap is of great interest to us. CSUSHPISA (S&P/Case-Shiller U.S. National Home Price Index) is out target ‘y’ variable for the regression task. This row shows the correlation of all features in ‘x’ to this target variable. This shows GDP and TLRESCONS has a strong correlation to the CSUSHPISA. DRSFRMACBS has an extremely weak correlation to the target variable.

FEATURE SELECTION

The wrapper method is used to select best features for each model in ‘backward’ direction.

MODEL TRAINING AND SCORING

The dataset was split into training (80%) and testing dataset (20%).

Four models were selected for this project:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net Regression

Selection of the model depends on the R-squared score obtained on the testing dataset.

RESULTS

MODEL	R-SQUARED SCORE	BEST FEATURES
Linear Regression	0.99203	'PERMIT1', 'EMRATIO', 'HOUST1F', 'TLRESCONS', 'GDP', 'RSAHORUSQ156S'
Ridge Regression	0.99065	'PERMIT1', 'HSN1F', 'TLRESCONS', 'GDP', 'DRSFRMACBS', 'RSAHORUSQ156S'
Lasso Regression	-0.07447	'HOUST1F', 'COMPU1USA', 'GDP', 'HOUSTSFLAA1FQ', 'DRSFRMACBS', 'RSAHORUSQ156S'
Elastic Net Regression	0.51527	'COMPU1USA', 'TLRESCONS', 'GDP', 'HOUSTSFLAA1FQ', 'DRSFRMACBS', 'RSAHORUSQ156S'

Thus, Ridge regression was found out to be the best model, having an R-Squared score of 0.99065. The best features for Ridge Regression were found to be 'PERMIT1', 'HSN1F', 'TLRESCONS', 'GDP', 'DRSFRMACBS', 'RSAHORUSQ156S'.

The home prices nationally were explained best by these following features:

- New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Single-Family Units (PERMIT1)
- New One Family Houses Sold: United States (HSN1F)
- Total Construction Spending: Residential in the United States (TLRESCONS)
- Gross Domestic Product (GDP)
- Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks (DRSFRMACBS)
- Homeownership Rate in the United States (RSAHORUSQ156S)

REFERENCES FOR THE DATA

1. S&P/Case-Shiller U.S. National Home Price Index (CSUSHPISA) | FRED | St. Louis Fed (stlouisfed.org)
2. Gross Domestic Product (GDP) | FRED | St. Louis Fed (stlouisfed.org)
3. New Privately Owned Housing Starts in the United States, Average Square Feet of Floor Area for One-Family Units (HOUSTSFLAA1FQ) | FRED | St. Louis Fed (stlouisfed.org)
4. New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Single-Family Units (PERMIT1) | FRED | St. Louis Fed (stlouisfed.org)
5. Employment-Population Ratio (EMRATIO) | FRED | St. Louis Fed (stlouisfed.org)
6. New One Family Houses Sold: United States (HSN1F) | FRED | St. Louis Fed (stlouisfed.org)
7. Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks (DRSFRMACBS) | FRED | St. Louis Fed (stlouisfed.org)
8. Homeownership Rate in the United States (RSAHORUSQ156S) | FRED | St. Louis Fed (stlouisfed.org)
9. New One Family Homes for Sale in the United States (HNFSEPUSSA) | FRED | St. Louis Fed (stlouisfed.org)
10. New Privately-Owned Housing Units Started: Single-Family Units (HOUST1F) | FRED | St. Louis Fed (stlouisfed.org)
11. New Privately-Owned Housing Units Completed: Single-Family Units (COMPU1USA) | FRED | St. Louis Fed (stlouisfed.org)
12. Total Construction Spending: Residential in the United States (TLRESCONS) | FRED | St. Louis Fed (stlouisfed.org)