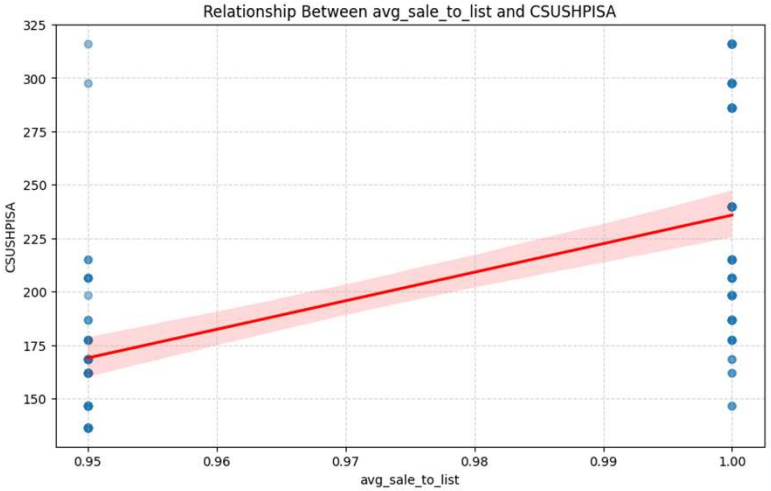
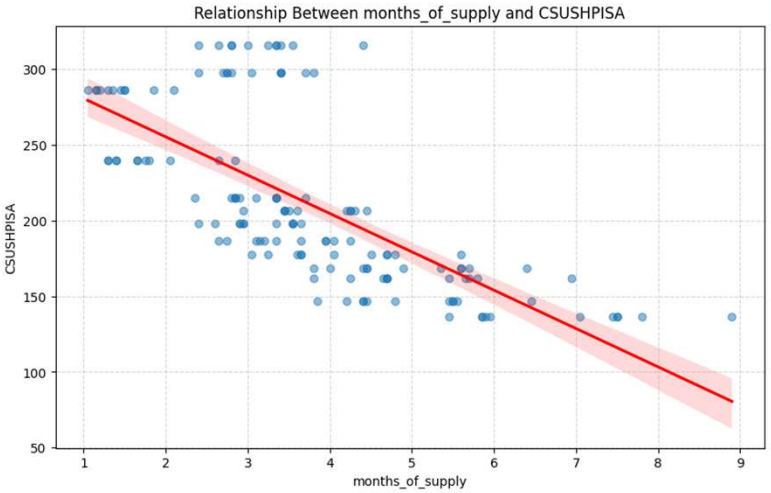
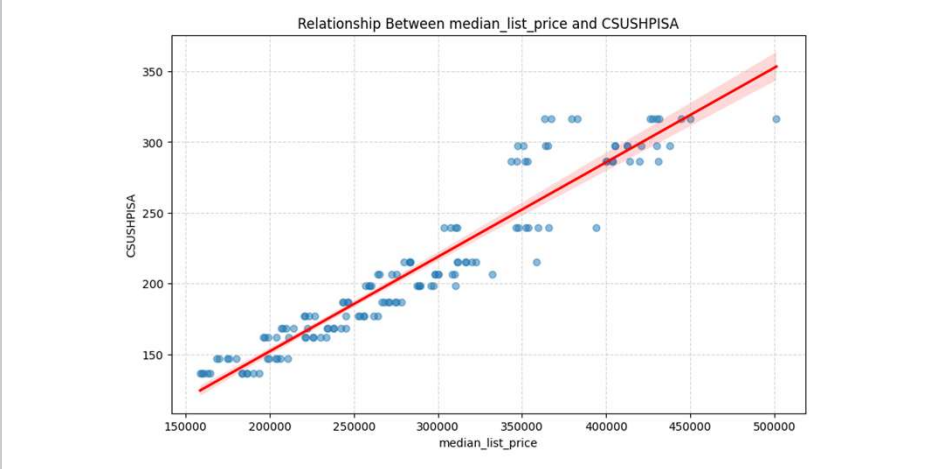

Home_Price
DATA PREPROCESSING, EDA,
FEATURE ENGINEERING AND
MODEL SELECTION

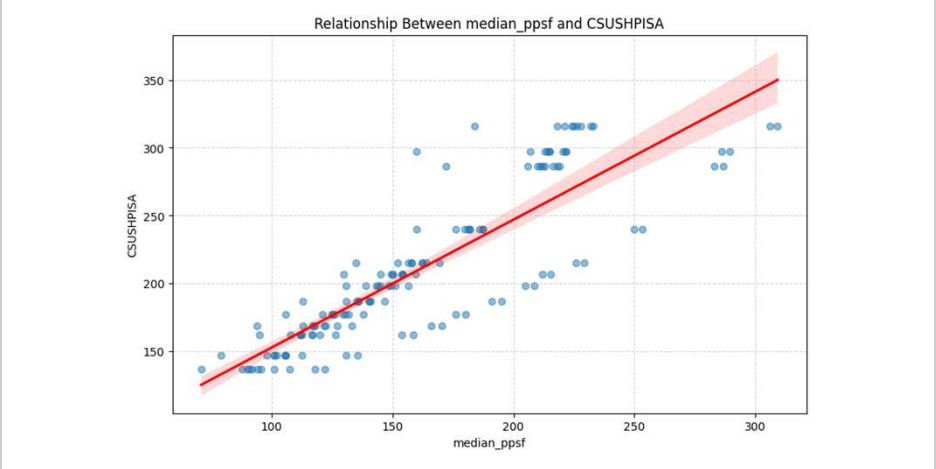


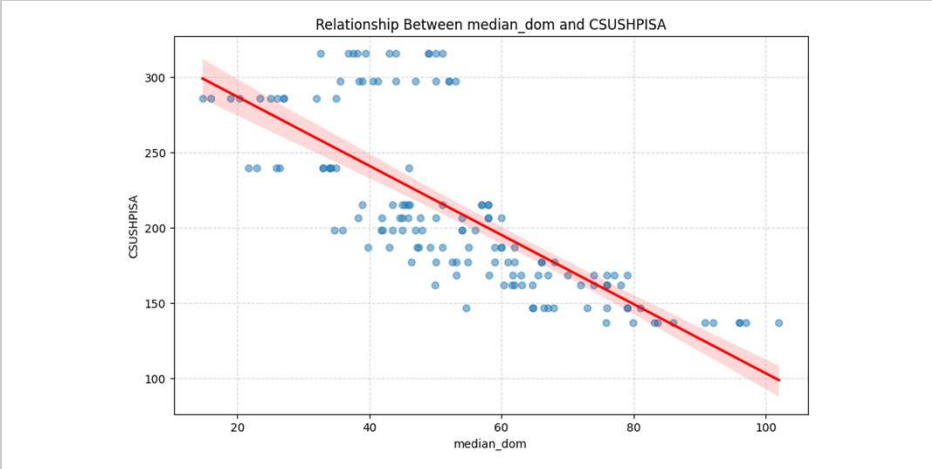


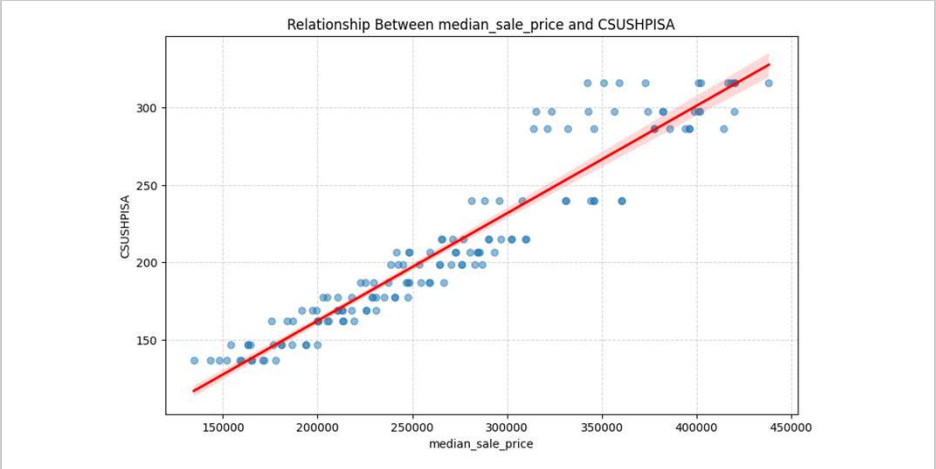
*Explanatory Data Analysis – **EDA** (showcasing the most affecting feature)*

- **Key Insights:**
  - Strong correlations identified between CSUSHPISA and features like median_sale_price, inventory, and homes_sold.
  - Seasonal trends observed in inventory and sales metrics.
  - Outliers detected in median_sale_price and inventory.
- **Visualizations:**
  - Correlation heatmap for feature relationships.
  - Time-series trends for CSUSHPISA and key features.
  - Boxplots for outlier detection.

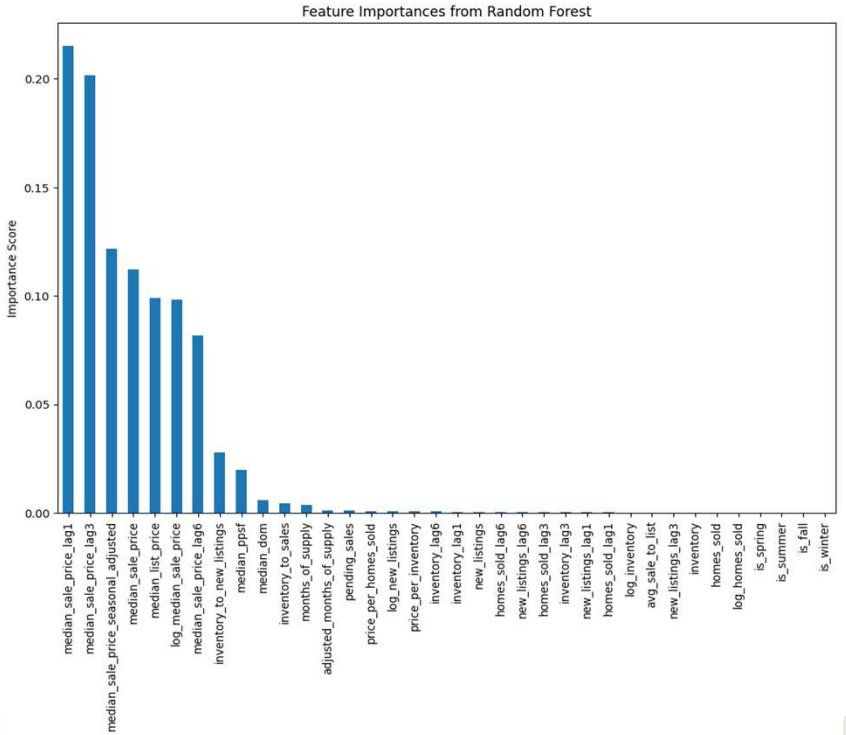

Relationship Between median_list_price and CSUSHPISA
CSUSHPISA
median_list_price
150
200
250
300
350
150000
200000
250000
300000
350000
400000
450000
500000


Relationship Between median_ppsf and CSUSHPISA
CSUSHPISA
median_ppsf
150
200
250
300
350
100
150
200
250
300


Relationship Between median_dom and CSUSHPISA
CSUSHPISA
median_dom
100
150
200
250
300
20
40
60
80
100


Relationship Between median_sale_price and CSUSHPISA
CSUSHPISA
median_sale_price
150
200
250
300
150000
200000
250000
300000
350000
400000
450000

Feature extraction w.r.t CSUSHPISA



Correlation Analysis:

CSUSHPISA	1.000000
year	0.966263
observation_date	0.966262
period_begin	0.966262
period_end	0.966262
median_sale_price_seasonal_adjusted	0.949950
median_sale_price	0.949950
median_sale_price_lag3	0.948853
median_sale_price_lag1	0.947139
median_list_price	0.946645
median_sale_price_lag6	0.943180
log_median_sale_price	0.935417
median_ppsf	0.851070

Feature	VIF
median_sale_price	inf
median_list_price	3.794887e+02
median_ppsf	1.025461e+01
homes_sold	1.341754e+02
pending_sales	3.404031e+02
new_listings	6.711863e+02
inventory	1.357116e+02
months_of_supply	9.530194e+02
median_dom	4.909919e+01
avg_sale_to_list	3.662217e+00
median_sale_price_lag1	1.771170e+01
median_sale_price_lag3	1.772926e+01
median_sale_price_lag6	1.777905e+01
inventory_lag1	3.257678e+01

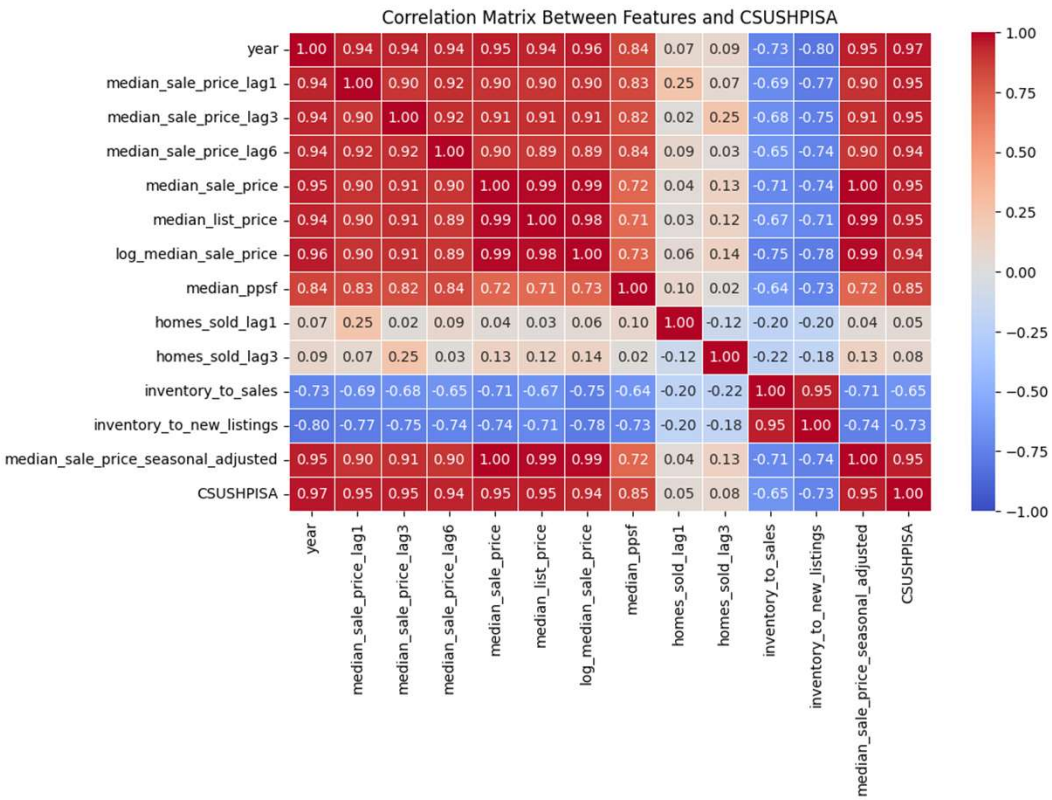
Feature Engineering Steps

- **Lagged Features:**
  - Created 1, 3, and 6-month lags for critical metrics (e.g., median_sale_price, inventory).
- **Seasonality Adjustments:**
  - Added seasonal indicators (e.g., is_spring, is_summer).
  - Seasonally adjusted median_sale_price.
- **Derived Metrics:**
  - Inventory-to-sales ratio, price-per-unit metrics, and adjusted months of supply.
- **Log Transformations:**
  - Applied to skewed features like median_sale_price and inventory.
- **Final Features:**
  - Selected top features using Random Forest feature importance and correlation analysis.

Final dataset after applying eda and feature engineering

year	median_sale_price_lag1	median_sale_price_lag3	median_sale_price_lag6	median_sale_price	median_list_price	log_median_sale_price	median_ppsf	homes_sold_lag1	homes_sold_lag3	inventory_to_sales	inventory_to_new_listings	median_sale_price_seasonal_adjusted	CSUSHPISA
2012	134703.00	170747.80	143553.55	152198.00	160420.00	11.932944	101.00	32672.00	395277.90	7.036442	4.028379	0.564327	136.6
2012	152198.00	165518.40	177702.15	148058.00	164294.00	11.905366	71.00	22090.00	31868.50	8.910852	4.598782	0.548976	136.6
2012	148058.00	134703.00	158876.00	171650.05	186478.55	12.053219	95.70	6865.00	32672.00	5.640796	4.072382	0.636452	136.6
2012	171650.05	152198.00	170747.80	165019.00	190335.00	12.013822	88.00	386652.05	22090.00	7.472938	4.260074	0.611865	136.6
2012	165019.00	148058.00	165518.40	159672.00	183524.00	11.980883	91.00	213083.00	6865.00	7.479624	4.236058	0.592039	136.6

Correlation matrix between the final features for the model and CSUSHPISA



Score comparison for all the Base models

Model	R^2	MAE	RMSE
Linear Regression	0.972525	7.280401e+00	9.396792e+00
Decision Tree	1.000000	9.135549e-15	1.936613e-14
Random Forest	0.999041	1.244732e+00	1.755848e+00
XGBoost	1.000000	4.146031e-04	8.053193e-04

Score on Prediction data after fine tunning the ***Random Forest model***

R^2:	**0.9992023730006738**
MAE:	**1.0771821428567907**
RMSE:	**1.6010706066563858**