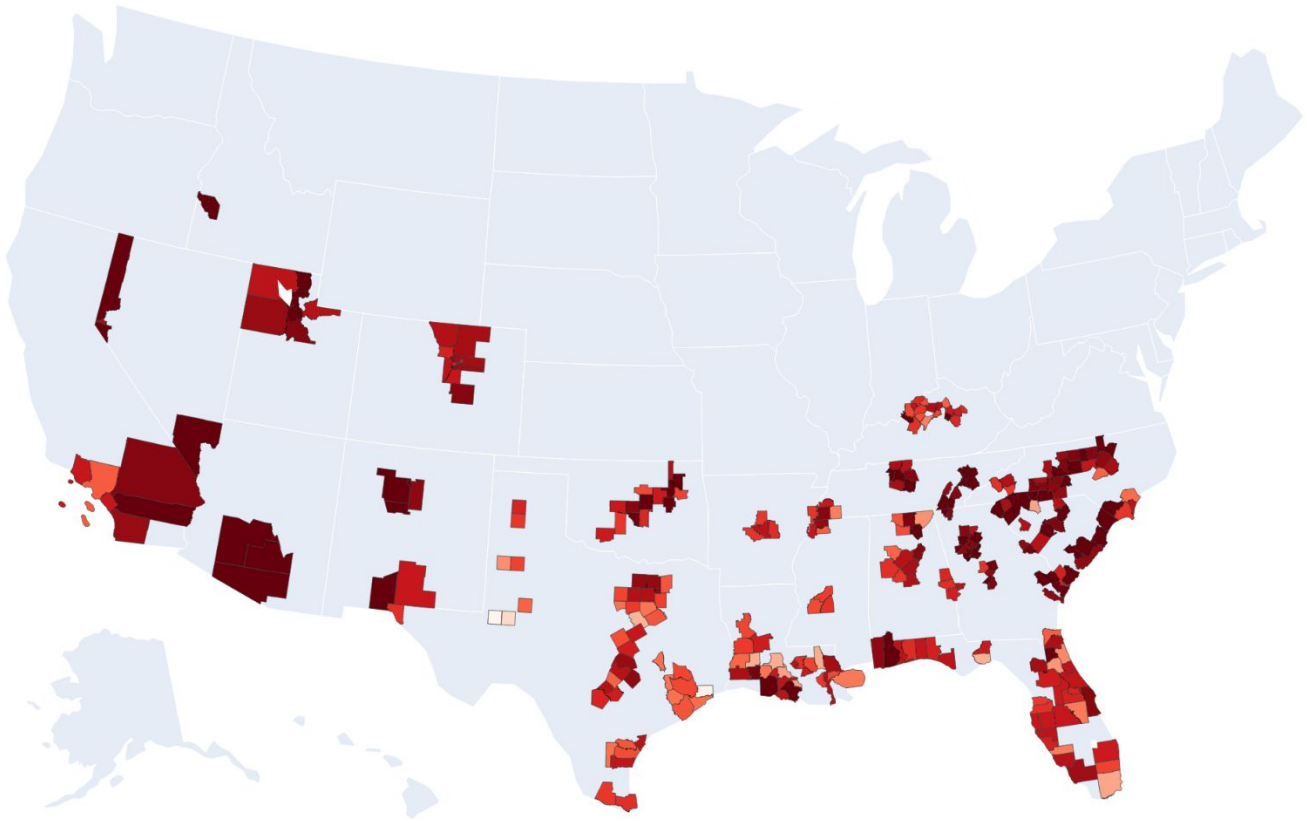
USC Marshall
School of Business

Predicting Real Estate Hot Spots (DSO 597)

Nattawut Kananusorn, Yi-Ching Lin, Chinka Huang, Siheng Rong, Thomas Nguyen, and Keegan O'Neill



USC Marshall School of Business

Introduction

Goal: Identify the 10 areas (US Census Tract) in the country with the highest expected rent growth in the next three years (2025).

Methodology: Use property level rent data in conjunction with additionally gathered data to train models that predicts the change in rent for the entire country from 2022 to 2025.

Machine Learning Models Used:

- **Traditional ML Models:** Linear Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and Long Short Term Memory.
- **TIme/Series Specific Models:** Pooled OLS, Fixed Effects and Random Effects.

USC Marshall School of Business

Overview of Methodology

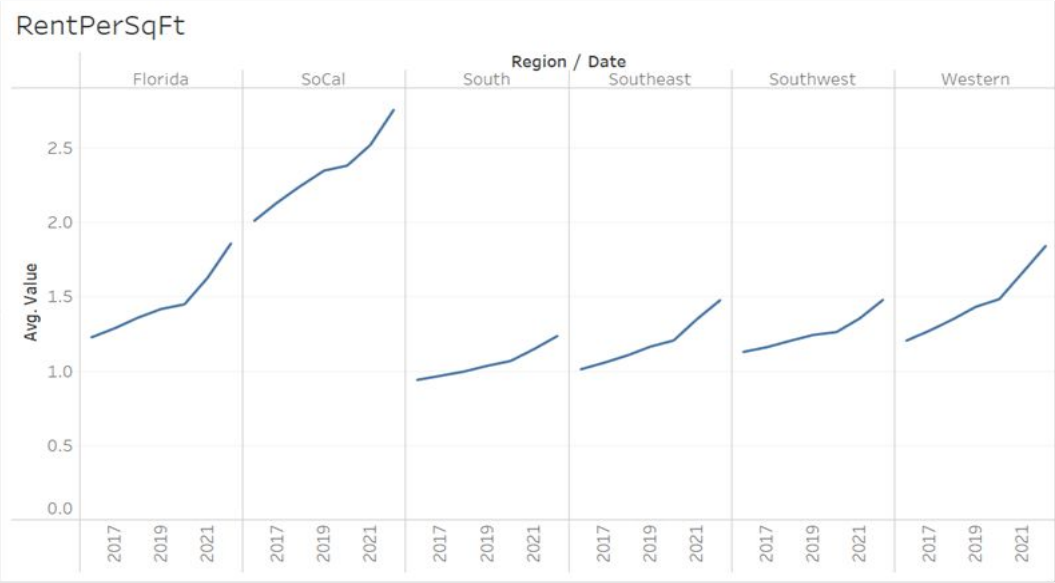
USC Marshall School of Business

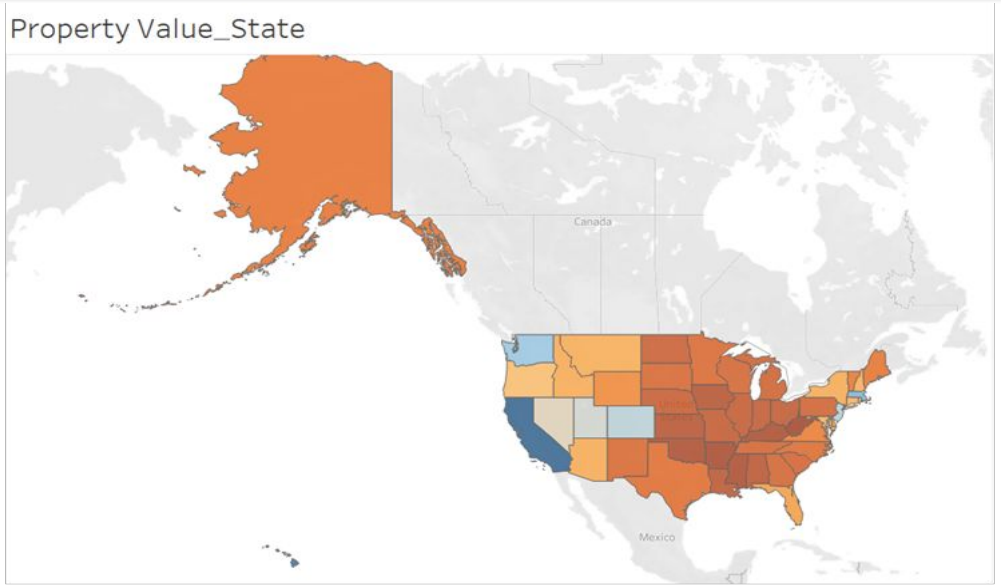
Factors Affecting Rent Growth

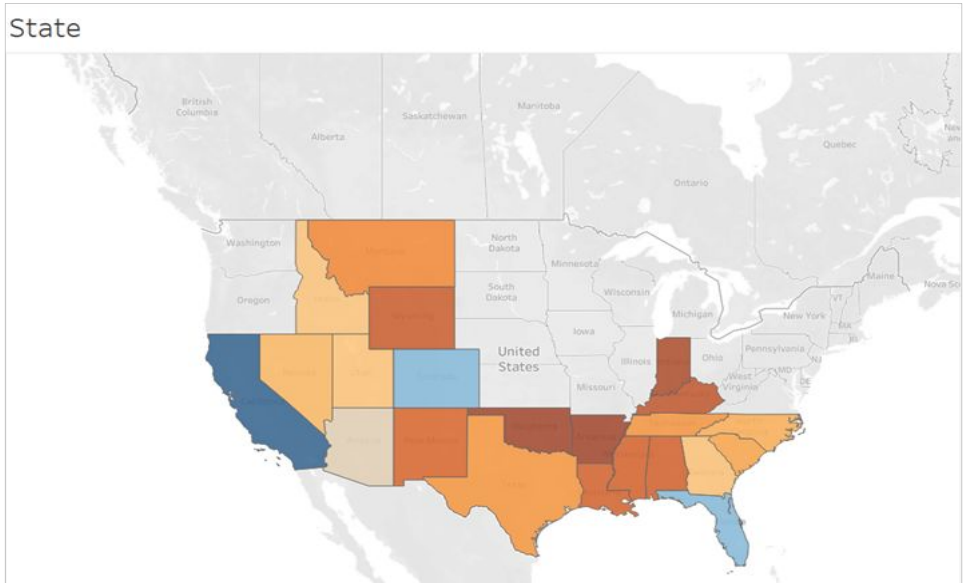
- **Housing data**
  - Occupancy Rate
  - Housing Supply
  - Property Values
- **Economics data**
  - Tax Rates (Sales, Income, Corporate, Combined)
  - Debt
  - National Indicators (Bond, SP 500, Inflation Rate)
- **Location data**
  - Starbucks, Banks, Public School, Private Schools, Hospitals (Normalized by Location Density)
  - Zip Code Area
- **Demographic data**
  - Population (gender, age, race)
  - Median Household Income
  - Employment Rate (16+ employed population / 16+ labor force)
- **Crime data**
  - Crime Rate

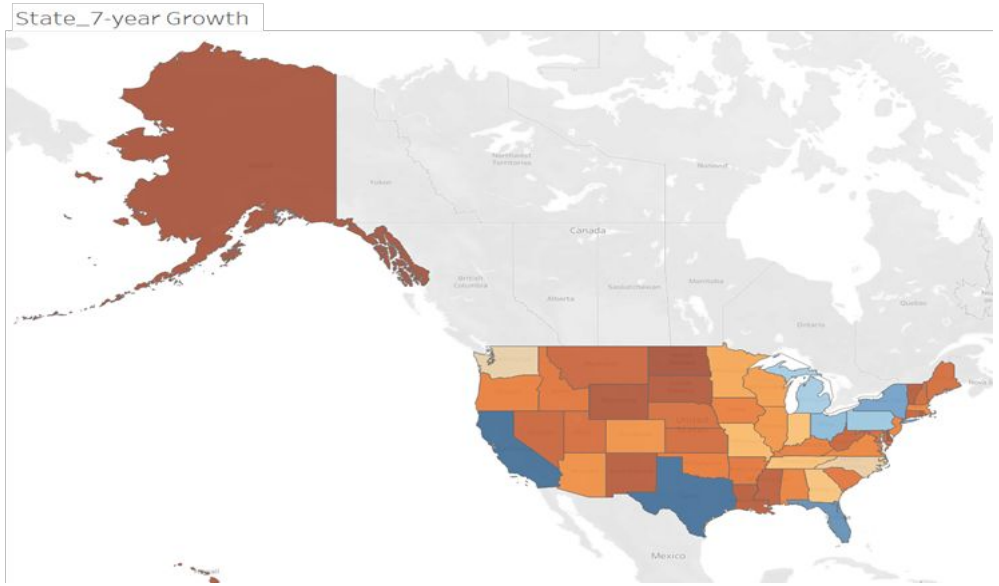
USC Marshall School of Business

EDA: Rent and Property Value



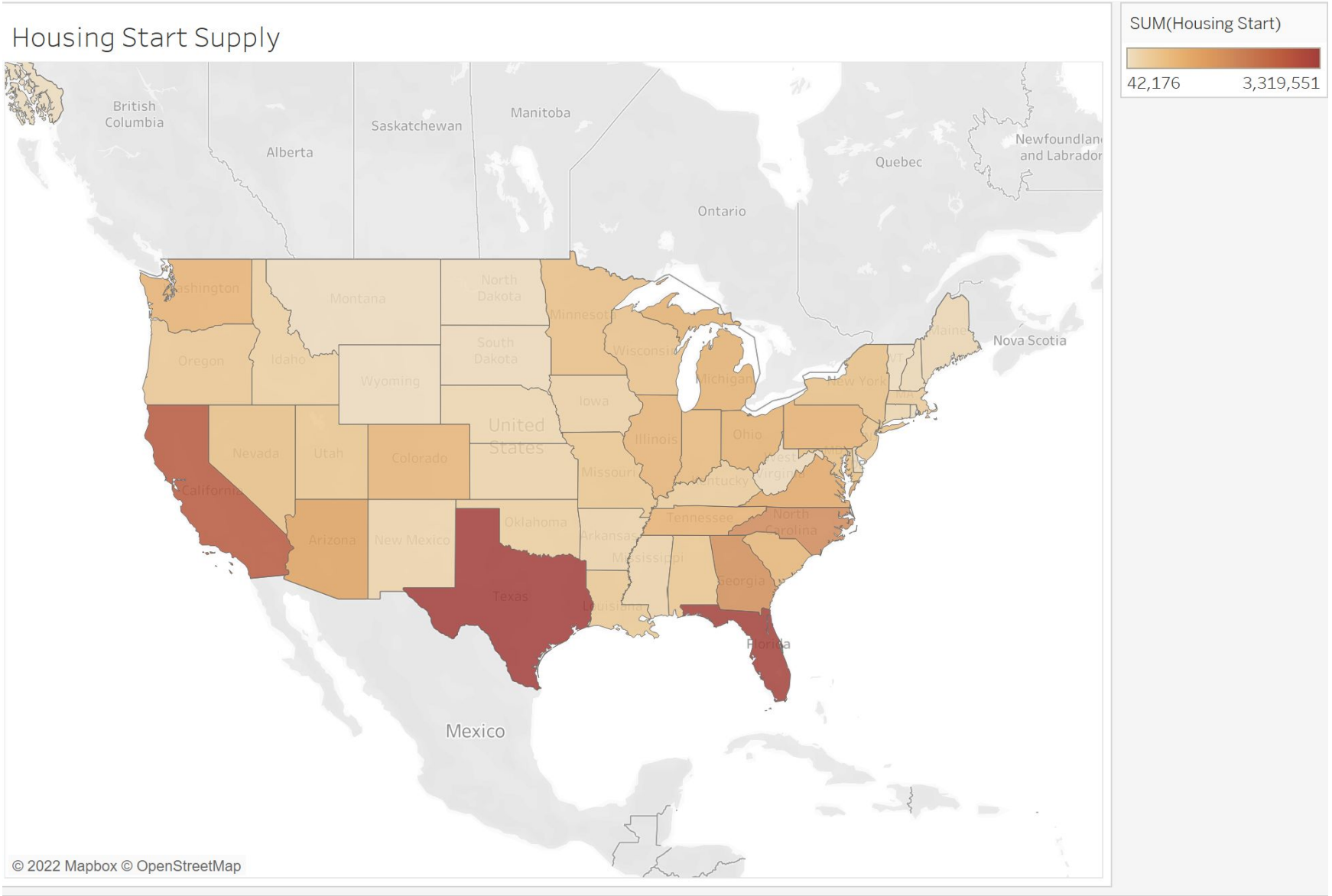






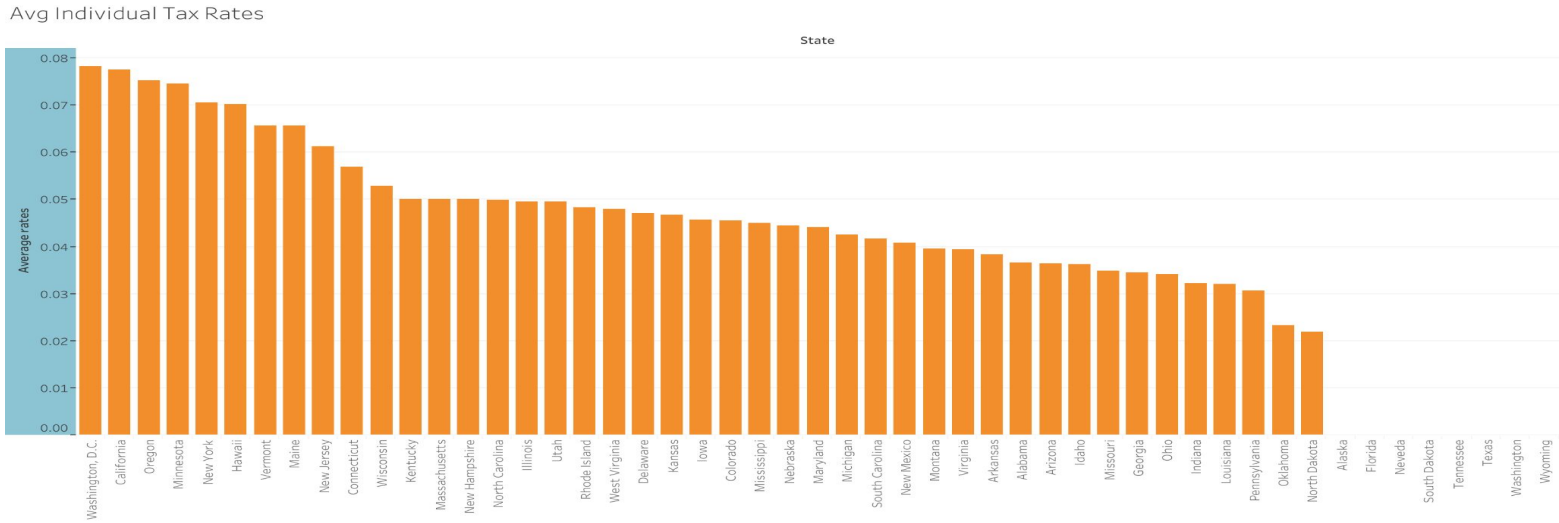
USC Marshall School of Business

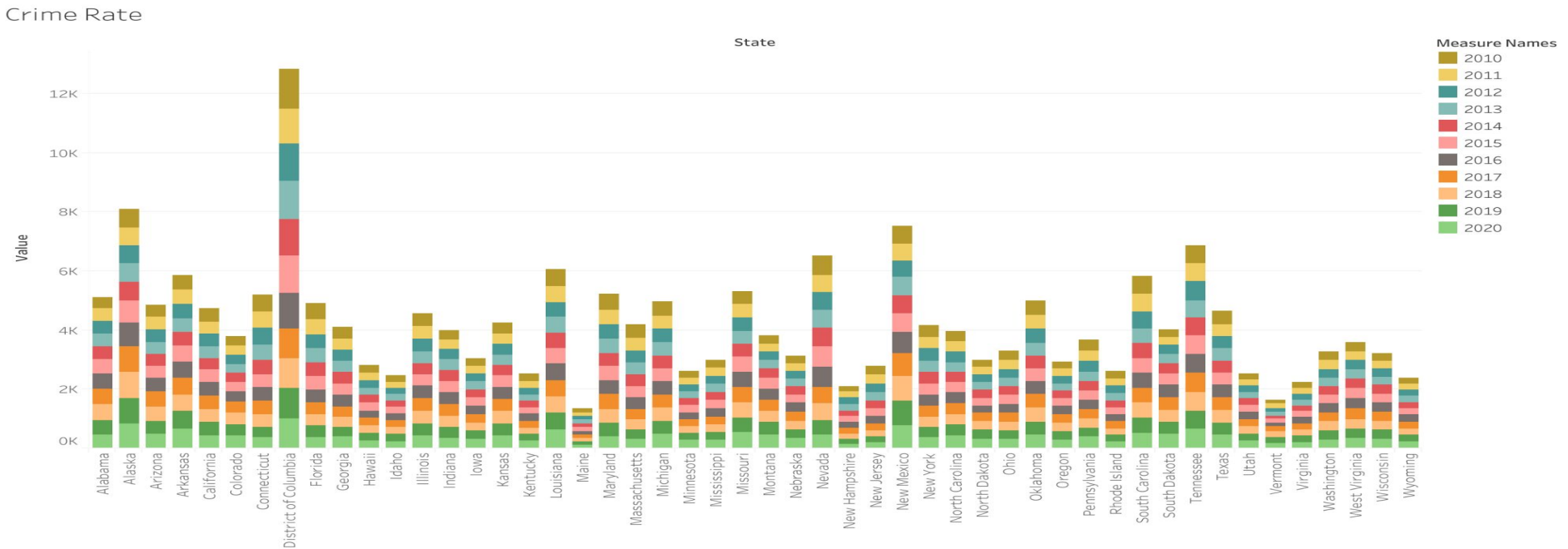
EDA: Housing Supply



USC Marshall School of Business

EDA: Economics Factors

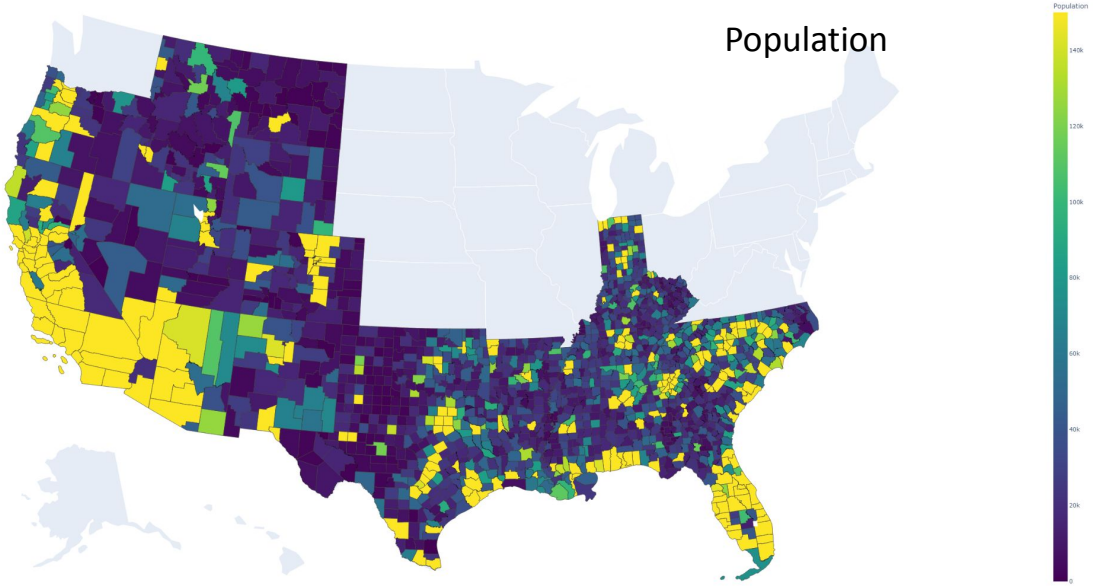




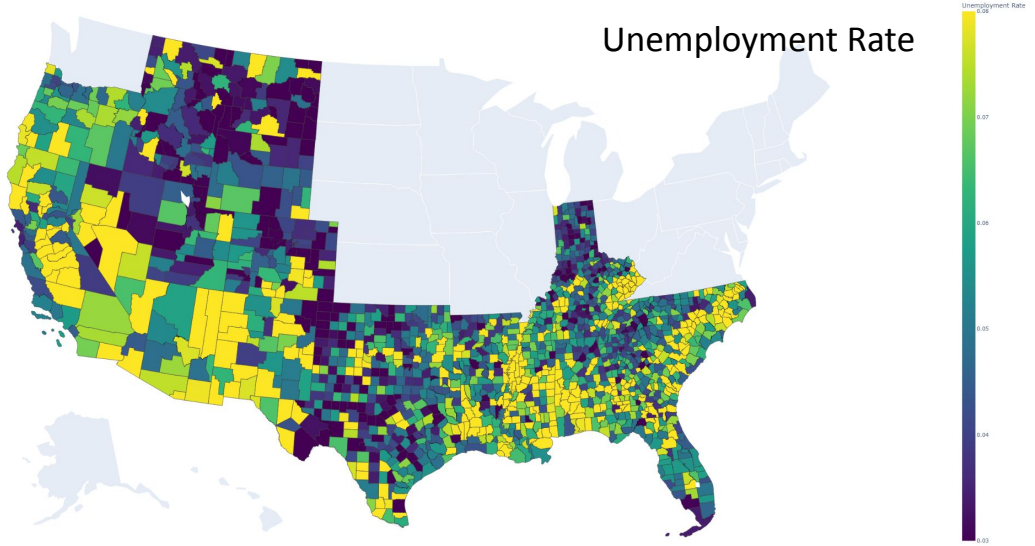
USC Marshall School of Business

EDA: Demographic Factors

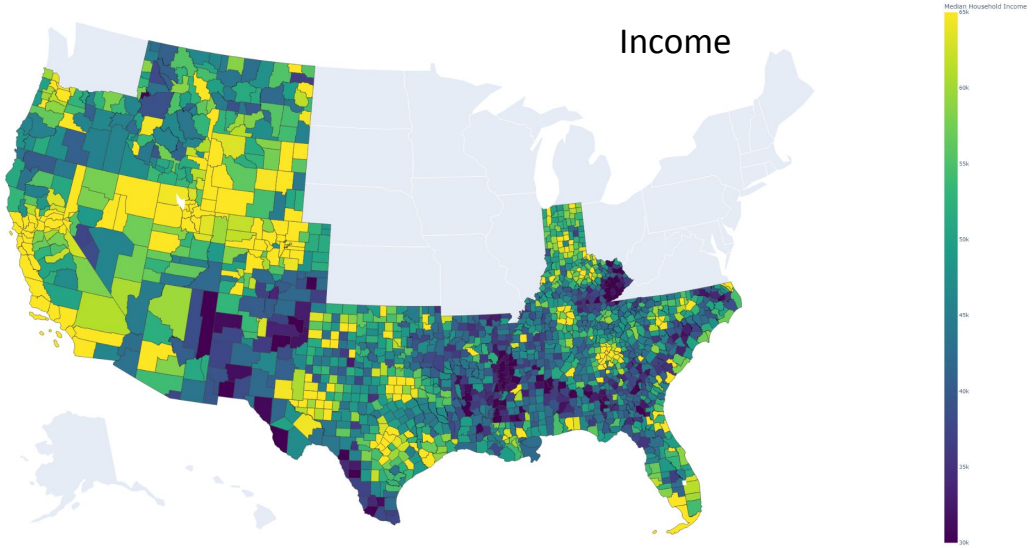
Population



Unemployment Rate



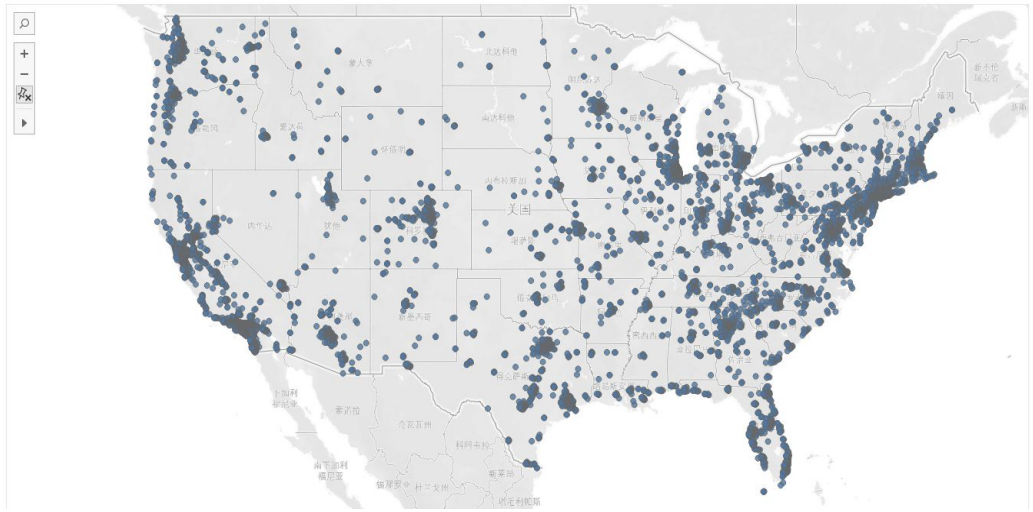
Income



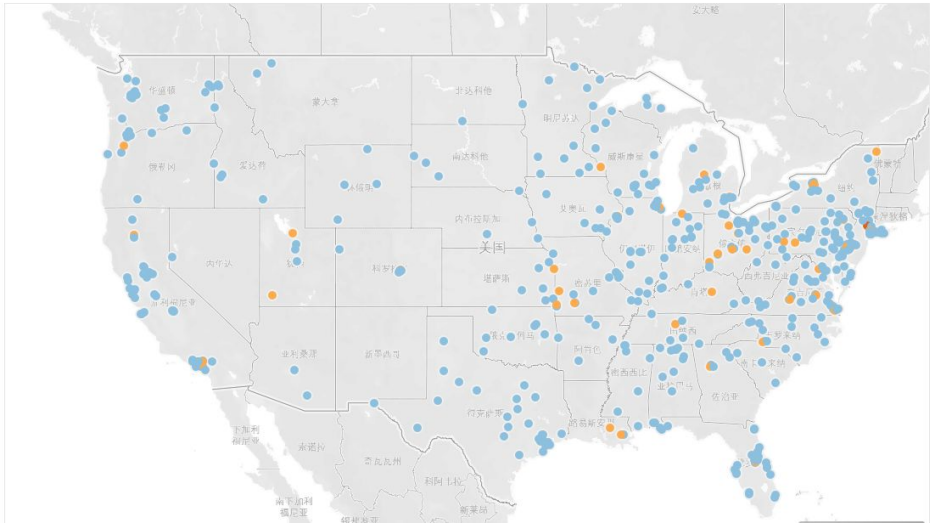
USC Marshall School of Business

EDA: Location Factors

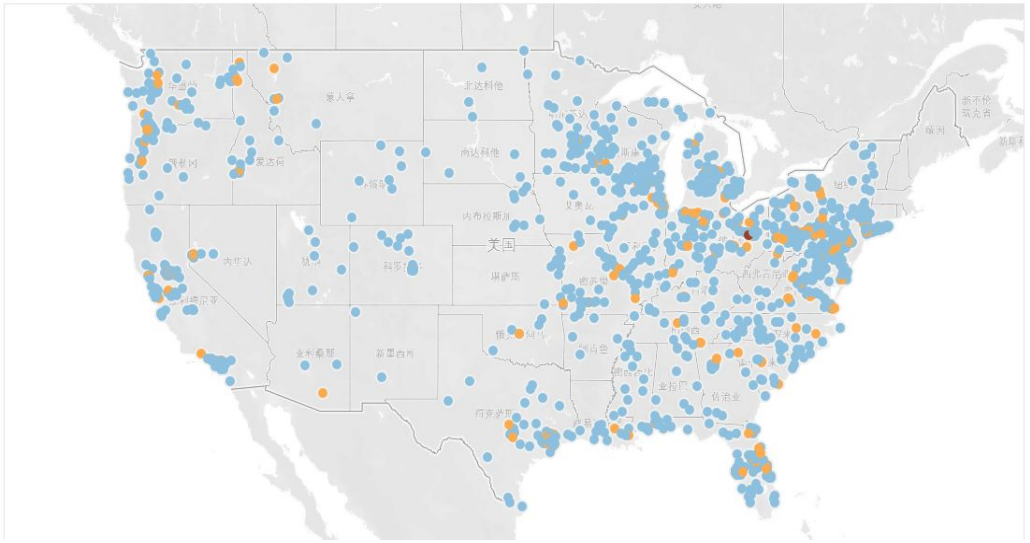
starbucks



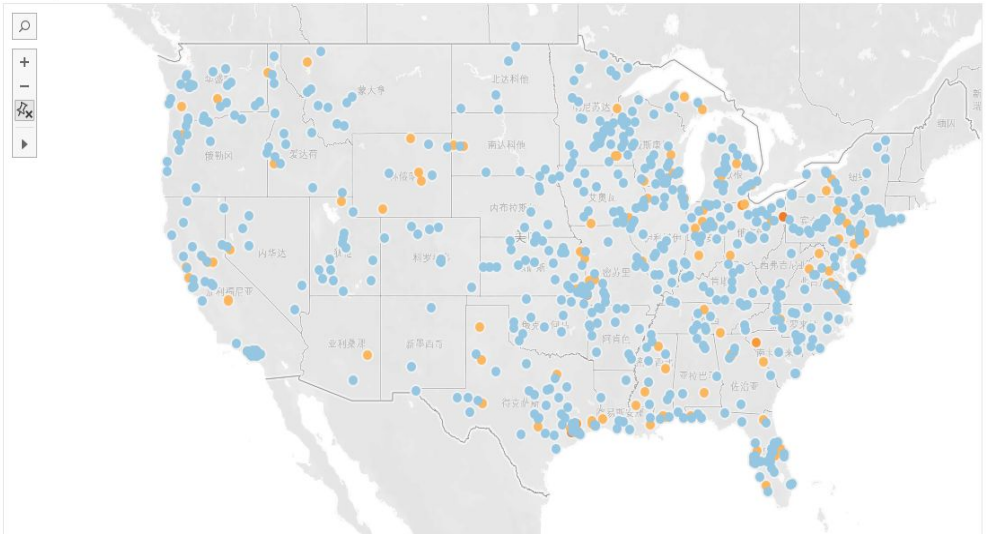
college



private school



hospital



- Starbucks stores
- Public schools
- Private schools
- Colleges
- Hospitals
- Banks
- Grocery visits
- Manufacturing facilities

USC Marshall School of Business

Data Cleaning

- Map Property Lat/Lon to Census tracts
- Add Third Party Data
  - Location data, Economics data, etc.
- Group Dataset by Year and Tract
- **Feature Enhancement**
  - Ex: Employment Rate, Male to Female Ratio
- **Missing Values**
  - Drop data
    - Missing time series rent data (15% of tracts)
  - Linear Interpolate
    - Corporate Finance Data (SP500, Corporate Bonds, T Bills), Crime Rate, Demographic Data
  - KNN
    - Employment Rate (2 tracts), Property Value (100 Properties)
  - County Median
    - Population, Income, Employment Data (33%)

USC Marshall School of Business

Final Panel Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	tract	year	id	growth	growth_per_sqft	mean_rent	mean_rent_per	total_units	region	state_code	county	property_sqft	current_rent	state
2	1003010500	2016	1.00301E+13	3.47909791	3.968253968	627.0833333	0.655	68	s	1	3	65200	633	AL
3	1003010500	2017	1.00301E+13	4.598006645	4.071246819	655.9166667	0.681666667	68	s	1	3	65200	633	AL
4	1003010500	2018	1.00301E+13	-5.107356117	-4.88997555	622.4166667	0.648333333	68	s	1	3	65200	633	AL
5	1003010500	2019	1.00301E+13	1.017539162	1.285347044	628.75	0.656666667	68	s	1	3	65200	633	AL
6	1003010500	2020	1.00301E+13	5.990722333	5.837563452	666.4166667	0.695	68	s	1	3	65200	633	AL
7	1003010500	2021	1.00301E+13	-2.450919095	-2.637889688	650.0833333	0.676666667	68	s	1	3	65200	633	AL
8	1003010500	2022	1.00301E+13	2.479169337	2.561576355	666.2	0.694	68	s	1	3	65200	633	AL
9	1003010600	2016	1.00301E+13	2.424568966	2.529761905	633.6666667	0.574166667	56	s	1	3	61800	691	AL
10	1003010600	2017	1.00301E+13	1.407154129	1.45137881	642.5833333	0.5825	56	s	1	3	61800	691	AL
11	1003010600	2018	1.00301E+13	1.997146933	2.00286123	655.4166667	0.594166667	56	s	1	3	61800	691	AL
12	1003010600	2019	1.00301E+13	-0.152574698	-0.280504909	654.4166667	0.5925	56	s	1	3	61800	691	AL
13	1003010600	2020	1.00301E+13	2.941551	2.953586498	673.6666667	0.61	56	s	1	3	61800	691	AL
14	1003010600	2021	1.00301E+13	2.090549233	2.322404372	687.75	0.624166667	56	s	1	3	61800	691	AL
15	1003010600	2022	1.00301E+13	7.888040712	7.343124166	742	0.67	56	s	1	3	61800	691	AL
16	1003010704	2016	1.00301E+13	-0.542197077	-0.35318824	703.1666667	0.790833333	180	s	1	3	160288	1417	AL
17	1003010704	2017	1.00301E+13	15.86868926	15.70073762	814.75	0.915	180	s	1	3	160288	1417	AL
18	1003010704	2018	1.00301E+13	5.615219392	5.464480874	860.5	0.965	180	s	1	3	160288	1417	AL
19	1003010704	2019	1.00301E+13	8.15417393	8.376511226	930.6666667	1.045833333	180	s	1	3	160288	1417	AL
20	1003010704	2020	1.00301E+13	0.698424069	0.478087649	937.1666667	1.050833333	180	s	1	3	160288	1417	AL
21	1003010704	2021	1.00301E+13	29.94842611	30.13481364	1217.833333	1.3675	180	s	1	3	160288	1417	AL

Time Series Models: Pooled OLS

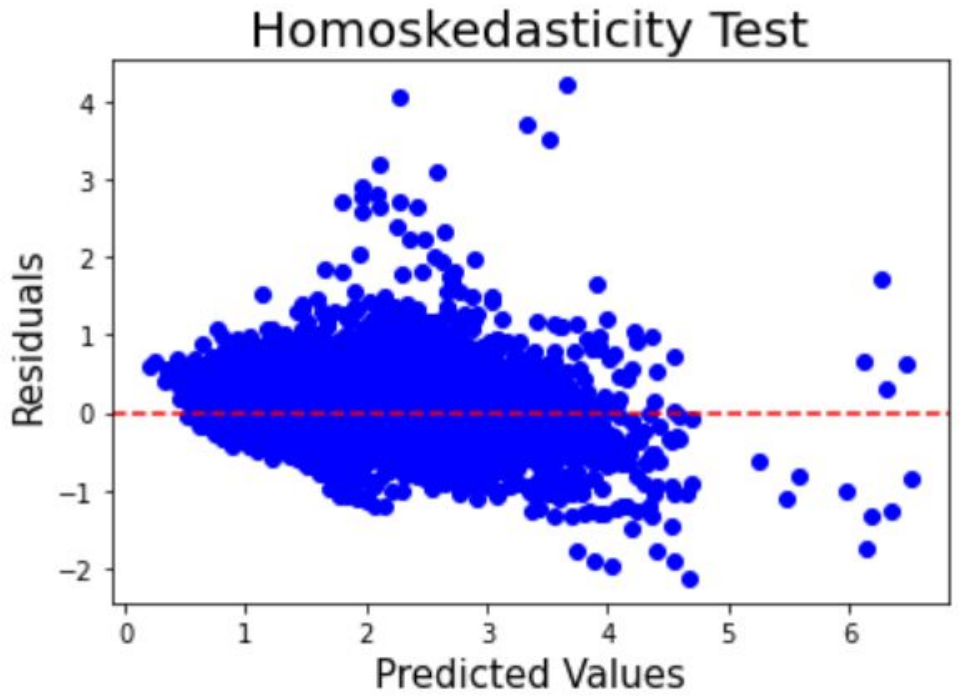
Pooled OLS can be described as simple OLS (Ordinary Least Squared) model that is performed on panel data. It ignores time and individual characteristics and focuses only on dependencies **between** the individuums. It requires that there is no correlation between unobserved, independent variable(s) and the IVs (i.e. exogeneity).

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \quad cov(x_{it}, \alpha_i) = 0, \; for\; t = 1, \dots, T, i = 1, \dots, N$$

USC Marshall School of Business

Time Series Models: Pooled OLS

```
exog = sm.add_constant(dataset[vars])
mod = PooledOLS(dataset.y, exog, check_rank=False)
res = mod.fit(cov_type='clustered', cluster_entity=True)
```




```
from statsmodels.stats.stattools import durbin_watson
pooled_OLS_dataset = pd.concat([dataset[vars], residuals_pooled_OLS], axis=1)
durbin_watson_test_results = durbin_watson(pooled_OLS_dataset['residual'])
print(durbin_watson_test_results)
```

✓ 0.4s

nan

- 3-year period data modeling 2017-2019 for 2020-2022
- Assumptions check (Homoskedasticity & Autocorrelation)
- Solutions: Random-Effects (RE) Model and Fixed-Effects (FE) Model

USC Marshall School of Business

Time Series Models: Fixed Effects Regression

- Panel data method for regression analysis
- All parameters being used are fixed, eliminating any randomness in the data
  - Uses the entity effect to determine model
- 4 Year Panel Period
  - Model trained using 2019-2022 data
- R-square as primary decision factor
- Feature selection was done by deleting the predictor variable with the highest p-value sequentially
- Does not include covariance variables (Drops these variables if present in dataset)

$$y_i = \sum_{j=1}^{J} \alpha_j z_{j[i]} + \beta x_i + \varepsilon_i; \qquad \varepsilon_i \sim N(0, \sigma_y^2).$$

USC Marshall School of Business

Time Series Models: Fixed Effects Regression

PanelOLS Estimation Summary

Dep. Variable:	mean_rent_per	R-squared:	0.6362
Estimator:	PanelOLS	R-squared (Between):	0.7052
No. Observations:	45732	R-squared (Within):	0.6362
Date:	Thu, Jun 30 2022	R-squared (Overall):	0.7015
Time:	15:54:39	Log-likelihood	7.167e+04
Cov. Estimator:	Unadjusted		
		F-statistic:	2119.2
Entities:	11767	P-value	0.0000
Avg Obs:	3.8865	Distribution:	F(28,33937)
Min Obs:	1.0000		
Max Obs:	4.0000	F-statistic (robust):	2119.2
		P-value	0.0000
Time periods:	4	Distribution:	F(28,33937)
Avg Obs:	1.143e+04		
Min Obs:	1.1e+04		
Max Obs:	1.176e+04		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0042	0.0314	0.1325	0.8946	-0.0573	0.0656
total_units	0.0035	0.0001	28.983	0.0000	0.0033	0.0038
property_sqft	-3.566e-06	1.212e-07	-29.423	0.0000	-3.804e-06	-3.329e-06
current_rent	0.0005	1.107e-05	46.318	0.0000	0.0005	0.0005
property_value	1.568e-06	1.8e-08	87.116	0.0000	1.533e-06	1.603e-06
median_income	1.961e-07	6.647e-08	2.9503	0.0032	6.583e-08	3.264e-07
population	4.235e-05	7.621e-06	5.5568	0.0000	2.741e-05	5.729e-05
white_pop	-1.633e-05	2.66e-06	-6.1376	0.0000	-2.154e-05	-1.111e-05
black_pop	-7.713e-06	3.89e-06	-1.9826	0.0474	-1.534e-05	-8.787e-08
asian_pop	-2.023e-05	5.47e-06	-3.6984	0.0002	-3.095e-05	-9.509e-06
pacific_pop	6.475e-05	2.223e-05	2.9132	0.0036	2.118e-05	0.0001
pop_0_19	-4.355e-05	8.339e-06	-5.2224	0.0000	-5.989e-05	-2.72e-05
pop_20_29	-3.693e-05	8.612e-06	-4.2887	0.0000	-5.382e-05	-2.005e-05
pop_30_39	-2.585e-05	9.065e-06	-2.8516	0.0044	-4.362e-05	-8.082e-06
pop_40_49	-5.119e-05	9.326e-06	-5.4887	0.0000	-6.947e-05	-3.291e-05
pop_50_59	-2.015e-05	9.62e-06	-2.0946	0.0362	-3.901e-05	-1.295e-06
pop_60_69	-3.004e-05	1.022e-05	-2.9403	0.0033	-5.006e-05	-1.001e-05
pop_70_over	-6.489e-05	1.073e-05	-6.0488	0.0000	-8.591e-05	-4.386e-05
occupancy_rate	-0.0013	8.934e-05	-14.038	0.0000	-0.0014	-0.0011
house_supply	3.31e-06	1.018e-06	3.2516	0.0011	1.315e-06	5.306e-06
crime_rate	-0.0001	1.663e-05	-6.6664	0.0000	-0.0001	-7.829e-05
ROI_SP500	-0.0800	0.0085	-9.3616	0.0000	-0.0968	-0.0633
ROI_Tbill	1.5735	0.1290	12.200	0.0000	1.3207	1.8263
ROI_Tbond	0.7152	0.0299	23.951	0.0000	0.6567	0.7738
grocery_vists_per	0.1359	0.0360	3.7687	0.0002	0.0652	0.2065
hospitals_per	0.0705	0.0200	3.5263	0.0004	0.0313	0.1097
colleges_per	0.0569	0.0144	3.9412	0.0001	0.0286	0.0852
private_schools_per	-0.0838	0.0315	-2.6576	0.0079	-0.1457	-0.0220
employment_rate	0.1620	0.0194	8.3654	0.0000	0.1241	0.2000

F-test for Poolability: 50.393
P-value: 0.0000
Distribution: F(11766,33937)

Included effects: Entity

Time Series Models: Random Effects Regression

Fixed Effects Model: $$y_i = \sum_{j=1}^{J} \alpha_j z_{j[i]} + \beta x_i + \varepsilon_i; \qquad \varepsilon_i \sim N(0, \sigma_y^2).$$

- Linear Regression of Y on X
- Adds to the specification a series of indicator variables Zj for each unit

Random Effects Model: $y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i; \qquad \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2); \qquad \varepsilon_i \sim N(0, \sigma_y^2).$

- αj are not estimated directly
- Follow a specified probability distribution

USC Marshall School of Business

Time Series Models: Random Effects Regression

- Similar to the Fixed Effect Regression
  - Panel Data methods for regression analysis.
  - Does not use entity effect to determine model

```
exog_vars = X_var_names
exog = sm.add_constant(df[exog_vars])
mod = PanelOLS(df.mean_rent_per, exog, entity_effects = False, check_rank=False, drop_absorbed=True)
RE = mod.fit()
print(RE)
```

- 4 Year Panel Period
  - Model trained from 2019-2022

```
Time periods:                           4   Distribution:
Avg Obs:                        1.143e+04
Min Obs:                          1.1e+04
Max Obs:                        1.176e+04

                                Parameter Estimates
==============================================================
```

- R-square as primary decision factor
- Feature selection was done by deleting the predictor variable with the highest p-value sequentially

USC Marshall School of Business

Time Series Models: Random Effects Regression

$$Y_{2025} = B_1 X_{2022} + B_2 X_{2022} + .. + B_0$$

	tract	year	id	growth	growth_per_sqft	mean_rent	mean_rent_per	total_units	region	state_code	...	county_male_populatio
0	1003010500	2016	10030105002016	3.479098	3.968254	627.083333	0.655000	68.0	s	1	...	2705.50000
1	1003010500	2017	10030105002017	4.598007	4.071247	655.916667	0.681667	68.0	s	1	...	2728.00000
2	1003010500	2018	10030105002018	-5.107356	-4.889976	622.416667	0.648333	68.0	s	1	...	2701.50000
3	1003010500	2019	10030105002019	1.017539	1.285347	628.750000	0.656667	68.0	s	1	...	2991.00000
4	1003010500	2020	10030105002020	5.990722	5.837563	666.416667	0.695000	68.0	s	1	...	2801.00000

year		total_units	property_sqft	current_rent	property_value	income_tax_rate	corporate_tax_rate	state_sales_tax_rate	avg_local_sales_tax_rate	combined_rate
2016	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2017	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2019	3	68.0	65200.0	X:2016 633.0	119546.416667	0.036667	0.065	0.04	0.052386	0.092386
2020	4	68.0	65200.0	X:2017 633.0	122164.416667	0.036667	0.065	0.04	0.052386	0.092386

USC Marshall School of Business

Random Effects Regression Results

RandomEffects Estimation Summary

Dep. Variable:	mean_rent_per	R-squared:	0.6912
Estimator:	RandomEffects	R-squared (Between):	0.7968
No. Observations:	45732	R-squared (Within):	0.6099
Date:	Thu, Jun 30 2022	R-squared (Overall):	0.7937
Time:	15:54:41	Log-likelihood	6.302e+04
Cov. Estimator:	Unadjusted		
		F-statistic:	4263.5
Entities:	11767	P-value	0.0000
Avg Obs:	3.8865	Distribution:	F(24,45707)
Min Obs:	1.0000		
Max Obs:	4.0000	F-statistic (robust):	4170.1
		P-value	0.0000
Time periods:	4	Distribution:	F(24,45707)
Avg Obs:	1.143e+04		
Min Obs:	1.1e+04		
Max Obs:	1.176e+04		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-1.1344	0.0493	-23.014	0.0000	-1.2310	-1.0378
property_sqft	-3.386e-07	1.71e-08	-19.802	0.0000	-3.721e-07	-3.051e-07
current_rent	0.0004	5.238e-06	71.877	0.0000	0.0004	0.0004
property_value	1.034e-06	1.161e-08	89.061	0.0000	1.011e-06	1.057e-06
income_tax_rate	-0.2669	0.1281	-2.0833	0.0372	-0.5180	-0.0158
corporate_tax_rate	0.6044	0.1188	5.0851	0.0000	0.3714	0.8373
state_sales_tax_rate	1.2445	0.3724	3.3417	0.0008	0.5146	1.9744
max_local_sales_tax	0.7456	0.2168	3.4394	0.0006	0.3207	1.1705
avg_household_income	6.867e-07	1.747e-07	3.9303	0.0001	3.443e-07	1.029e-06
population	7.779e-06	3.204e-06	2.4276	0.0152	1.498e-06	1.406e-05
white_pop	-2.614e-05	2.369e-06	-11.034	0.0000	-3.079e-05	-2.15e-05
black_pop	-2.933e-05	3.056e-06	-9.5983	0.0000	-3.532e-05	-2.334e-05
asian_pop	-3.435e-05	4.52e-06	-7.6009	0.0000	-4.321e-05	-2.549e-05
pop_0_19	-1.369e-05	4.498e-06	-3.0434	0.0023	-2.25e-05	-4.873e-06
pop_20_29	3.146e-05	4.339e-06	7.2523	0.0000	2.296e-05	3.997e-05
pop_30_39	3.739e-05	5.221e-06	7.1604	0.0000	2.715e-05	4.762e-05
house_supply	1.059e-05	7.55e-07	14.025	0.0000	9.109e-06	1.207e-05
crime_rate	-7.62e-05	1.562e-05	-4.8771	0.0000	-0.0001	-4.558e-05
Tbond5	0.0001	1.825e-06	59.343	0.0000	0.0001	0.0001
hist_risk	13.408	0.9253	14.491	0.0000	11.594	15.221
SP500	-0.1674	0.0051	-32.833	0.0000	-0.1773	-0.1574
starbucks_vists_per	0.0376	0.0041	9.0829	0.0000	0.0295	0.0457
colleges_per	0.0686	0.0060	11.361	0.0000	0.0568	0.0804
public_schools_per	0.0489	0.0024	20.118	0.0000	0.0441	0.0537
private_schools_per	0.0411	0.0061	6.7698	0.0000	0.0292	0.0530

predict	growth_25
3.166725	8.223470
3.117854	6.651176
3.114385	6.611532
3.226689	6.014542
3.122464	5.990590
3.238094	5.963643
3.527521	5.939385
3.277191	5.733954
3.253315	5.639417
3.306811	5.338936

USC Marshall School of Business

Random Effect Regressions Results (FE Variables)

PanelOLS Estimation Summary

Dep. Variable:	mean_rent_per	R-squared:	0.8422
Estimator:	PanelOLS	R-squared (Between):	0.8481
No. Observations:	45732	R-squared (Within):	0.5876
Date:	Thu, Jun 30 2022	R-squared (Overall):	0.8422
Time:	15:54:40	Log-likelihood	5458.1
Cov. Estimator:	Unadjusted		
		F-statistic:	4781.8
Entities:	11767	P-value	0.0000
Avg Obs:	3.8865	Distribution:	F(51,45680)
Min Obs:	1.0000		
Max Obs:	4.0000	F-statistic (robust):	4598.9
		P-value	0.0000
Time periods:	4	Distribution:	F(51,45680)
Avg Obs:	1.143e+04		
Min Obs:	1.1e+04		
Max Obs:	1.176e+04		

Parameter Estimates

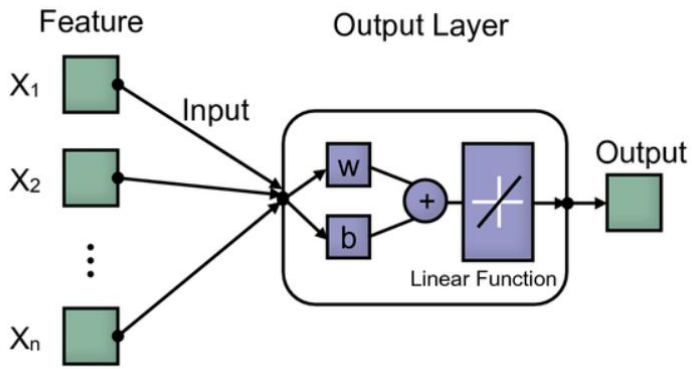
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	7.549e-09					
total_units	0.0033	3.053e-05	108.19	0.0000	0.0032	0.0034
property_sqft	-3.81e-06	3.275e-08	-116.36	0.0000	-3.875e-06	-3.746e-06
current_rent	0.0005	3.096e-06	168.36	0.0000	0.0005	0.0005
property_value	6.105e-07	7.168e-09	85.166	0.0000	5.964e-07	6.245e-07
income_tax_rate	0.8873	0.0705	12.590	0.0000	0.7492	1.0254
corporate_tax_rate	1.1779	0.0616	19.136	0.0000	1.0573	1.2986
state_sales_tax_rate	-1.0142	1.908e+05	-5.317e-06	1.0000	-3.739e+05	3.739e+05
avg_local_sales_tax_rate	1.9438	1.907e+05	1.019e-05	1.0000	-3.738e+05	3.739e+05
combined_rate	0.9296	1.908e+05	4.873e-06	1.0000	-3.739e+05	3.739e+05
max_local_sales_tax	-1.4690	0.1950	-7.5321	0.0000	-1.8512	-1.0867
median_income	-2.607e-07	6.643e-08	-3.9247	0.0001	-3.909e-07	-1.305e-07
population	2.945e-05	7.997e-06	3.6829	0.0002	1.378e-05	4.512e-05
white_pop	-2.586e-05	2.537e-06	-10.190	0.0000	-3.083e-05	-2.088e-05
black_pop	-4.597e-05	2.698e-06	-17.037	0.0000	-5.125e-05	-4.068e-05
asian_pop	-3.218e-05	3.878e-06	-8.2965	0.0000	-3.978e-05	-2.458e-05
pacific_pop	-9.492e-05	3.688e-05	-2.5738	0.0101	-0.0002	-2.264e-05
pop_0_19	-7.898e-05	8.978e-06	-8.7972	0.0000	-9.658e-05	-6.139e-05
pop_20_29	3.27e-05	9.16e-06	3.5699	0.0004	1.475e-05	5.065e-05
pop_30_39	7.632e-05	1.015e-05	7.5231	0.0000	5.644e-05	9.621e-05
pop_40_49	3.331e-05	1.152e-05	2.8919	0.0038	1.074e-05	5.589e-05
pop_50_59	6.645e-06	1.221e-05	0.5440	0.5864	-1.73e-05	3.058e-05
pop_60_69	-1.415e-05	1.313e-05	-1.0774	0.2813	-3.988e-05	1.159e-05
pop_70_over	-6.858e-05	1.037e-05	-6.6100	0.0000	-8.891e-05	-4.824e-05
occupancy_rate	-0.0024	0.0002	-13.457	0.0000	-0.0028	-0.0021
house_supply	1.62e-05	5.47e-07	29.618	0.0000	1.513e-05	1.727e-05
crime_rate	-4.584e-05	1.955e-05	-2.3448	0.0190	-8.416e-05	-7.523e-06
ROI_SP500	-1.621e-07					
ROI_Tbill	7.742e-09					
ROI_Tbond	-1.718e-08	2.945e+05	-5.834e-14	1.0000	-5.772e+05	5.772e+05
corp_bond	-8.639e-08	1.26e+05	-6.855e-13	1.0000	-2.47e+05	2.47e+05
real_estate	-9.241e-09	6.977e+04	-1.324e-13	1.0000	-1.367e+05	1.367e+05
sp500	1.13e-06	0.0434	2.604e-05	1.0000	-0.0850	0.0850
Tbill	4.321e-05	10.727	4.028e-06	1.0000	-21.024	21.024
Tbond5	7.086e-05	4.3967	1.612e-05	1.0000	-8.6175	8.6177
baa_corp_bond	-2.448e-05	0.0927	-0.0003	0.9998	-0.1817	0.1817
real_estate2	0.0002					
stocks_bills	-1.698e-07	6.985e+04	-2.431e-12	1.0000	-1.369e+05	1.369e+05
stocks_bonds	-1.449e-07	4.84e+04	-2.994e-12	1.0000	-9.486e+04	9.486e+04
stocks_corp_bonds	-7.563e-08	1.517e+05	-4.986e-13	1.0000	-2.973e+05	2.973e+05
hist_risk	-1.83e-10	2.136e+05	-8.57e-16	1.0000	-4.186e+05	4.186e+05
inflation	-1.18e-09	6.754e+05	-1.746e-15	1.0000	-1.324e+06	1.324e+06
SP500	-1.576e-07	6.843e+04	-2.304e-12	1.0000	-1.341e+05	1.341e+05
Tbill_3month	8.721e-09	1.315e+06	6.633e-15	1.0000	-2.577e+06	2.577e+06
Tbond_10year	-1.568e-08	3.753e+05	-4.177e-14	1.0000	-7.356e+05	7.356e+05
baa_corp_bonds	-8.359e-08	9.931e+04	-8.417e-13	1.0000	-1.946e+05	1.946e+05
real_estate3	-7.888e-09	6.036e+05	-1.307e-14	1.0000	-1.183e+06	1.183e+06
grocery_vists_per	0.0089	0.0044	2.0059	0.0449	0.0002	0.0175
hospitals_per	0.0603	0.0043	14.146	0.0000	0.0519	0.0686
colleges_per	0.1090	0.0029	38.065	0.0000	0.1034	0.1147
private_schools_per	0.0905	0.0029	31.016	0.0000	0.0848	0.0962
employment_rate	-0.0736	0.0293	-2.5125	0.0120	-0.1311	-0.0162

predict	growth_25
1.521830	1.454564
1.650736	1.311415
0.709246	1.065765
1.146456	0.985205
1.582235	0.906307
0.918871	0.807616
1.625039	0.793971
1.483943	0.786090
0.927042	0.777077
0.816095	0.774119

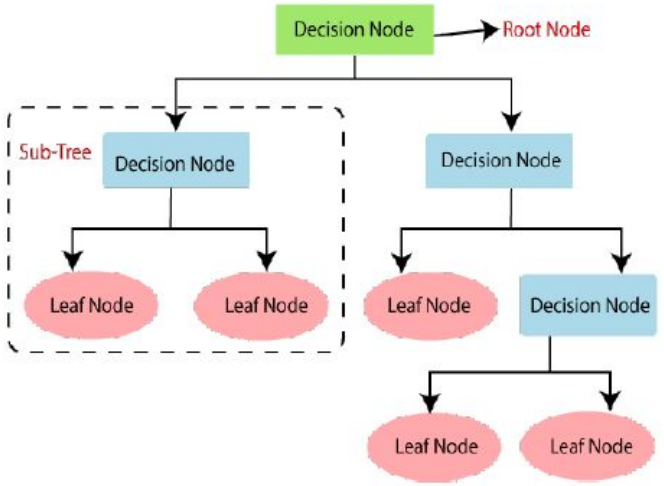
USC Marshall School of Business

Traditional Machine Learning (ML) Models

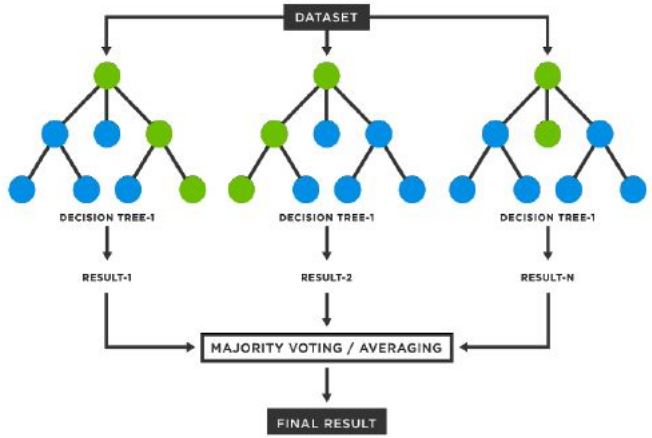
Linear Regression



Decision Tree



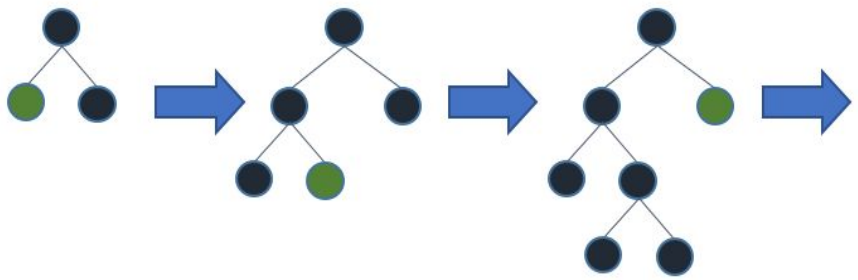
Random Forest



USC Marshall School of Business

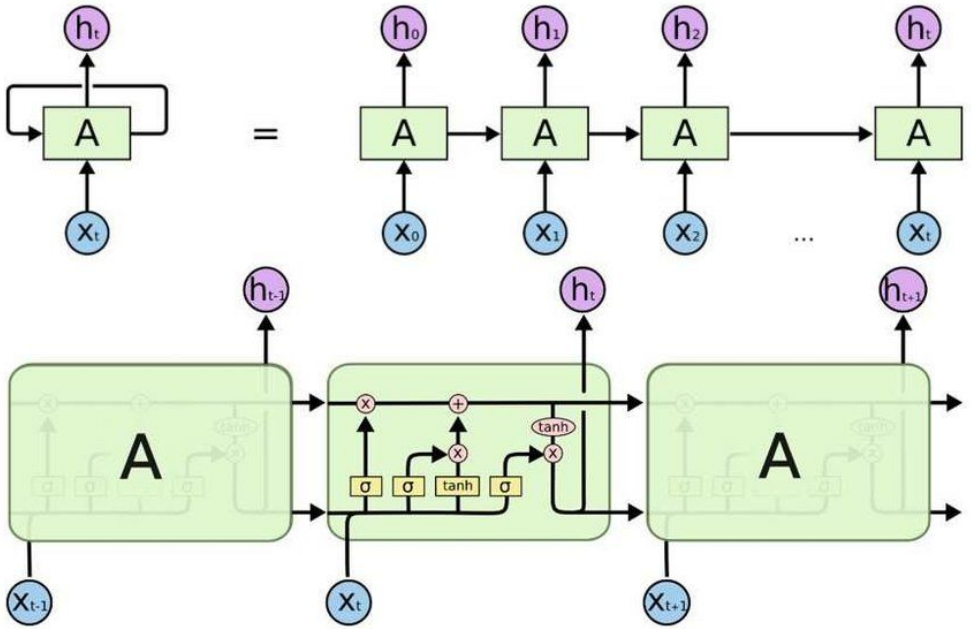
Traditional ML Models

LightGBM



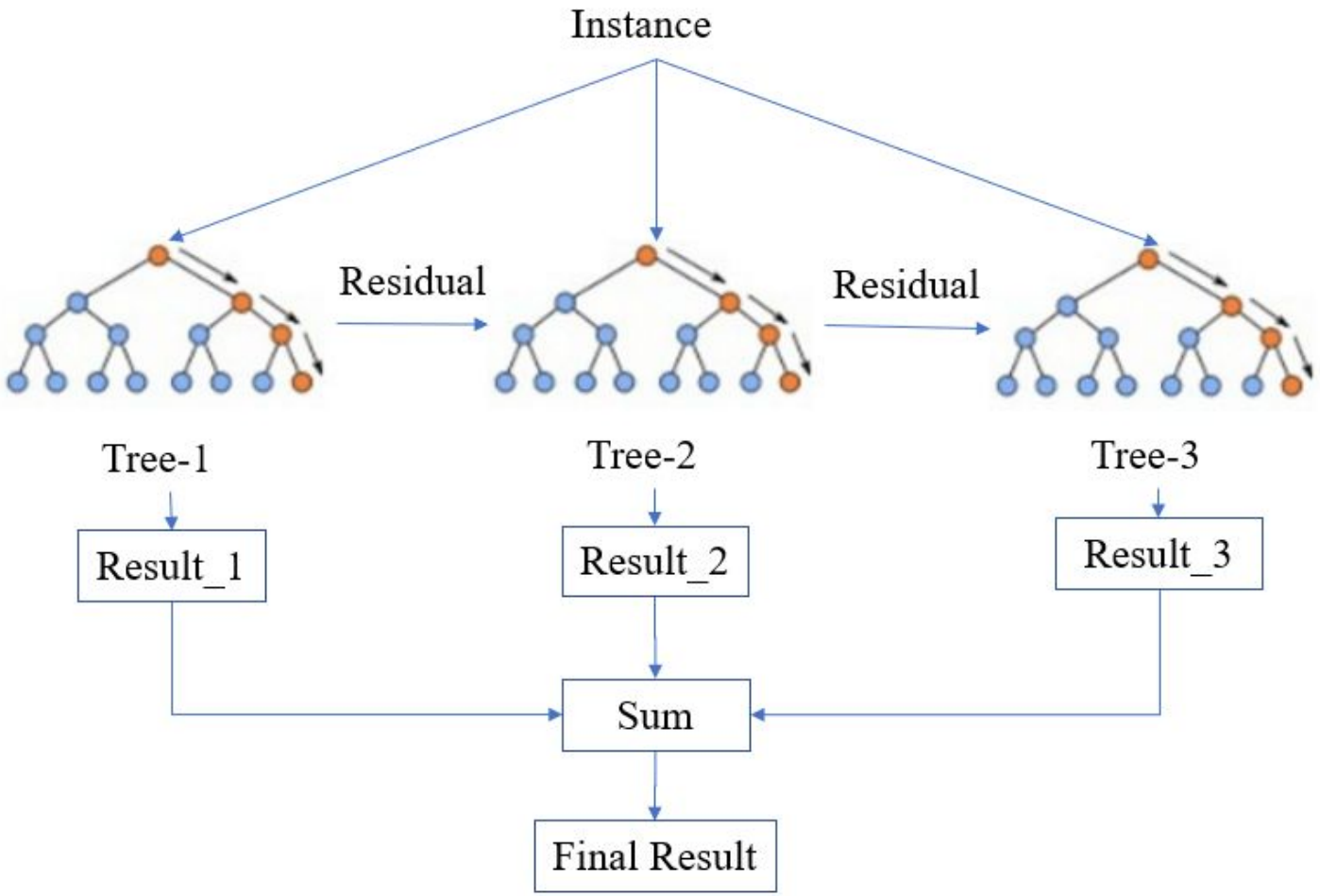
Leaf-wise tree growth

Long short-term memory



USC Marshall School of Business

Traditional ML Model: XGBoost

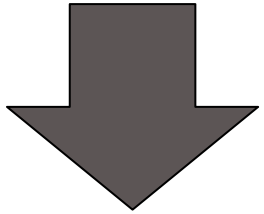


- Decision tree based ensemble model
- Designed for speed and performance
- Suit with small-to-medium structured/tabular data
- Easy for overfitting but overcome by tuning properly
- Training time is pretty high

USC Marshall School of Business

Traditional ML Models: Steps

Xt-5	Xt-4	Xt-3	X	Y
2017	2018	2019	Cross-sectional	2022 mean rent per sqft



Xt-5	Xt-4	Xt-3	X
2020	2021	2022	Cross-sectional

→ Trained Model →

Y
2025 mean rent per sqft

- Y: Mean Rent Per SQFT in year t
- X: Cross- sectional data ; Time-series data in year t-3, t-4, t-5
- Applied the 10-fold cross validation to find the best hyperparameters
- R-squared as criteria

USC Marshall School of Business

Traditional ML Models: Steps

Linear Regression Models

Region	R-Squared Value
All Regions	.952
Florida	.923
SouthEast	.926
South	.931
SouthWest	.920
Western	.867
Southern California	.877

USC Marshall School of Business

Traditional ML Models: Results

Model	Parameters				Avg R squared
Linear Regression					0.9517
Decision Tree	max_depth	min_leaf	min_split	splitter	0.9291
	10	10	50	best	
Random Forest	max_depth	min_leaf	min_split	n_estimators	0.9466
	10	5	10	100	
XGBoost	learning_rate	max_depth	n_estimators		0.9644
	0.1	4	800		
LightGBM	learning_rate	max_depth	n_estimators		0.9618
	0.1	4	500		
RNN - LSTM	epochs	batch_size	verbose		0.8640
	30	50	1		

USC Marshall School of Business

Best Model Results and Evaluation

Sort by Fixed Effect

	tract	FE_rent_growth_2025	XG_rent_growth_2025	LGB_rent_growth_2025
0	04013217001	1.684731	0.161208	0.171603
1	06059062645	1.525346	0.136040	0.141538
2	06059062658	1.473987	0.120287	0.123380
3	06037264000	1.325111	0.066382	0.054277
4	48201333204	1.297981	0.083040	0.071199
5	18043070801	1.276547	0.561004	0.605929
6	06059063010	1.217841	0.142234	0.144315
7	06059062649	1.181896	0.100328	0.108008
8	48201421402	1.153177	0.050812	0.068860
9	22051020517	1.034263	0.095851	0.103653

Sort by XGBoost

	tract	FE_rent_growth_2025	XG_rent_growth_2025	LGB_rent_growth_2025
0	18043070801	1.276547	0.561004	0.605929
1	06037800205	0.586632	0.466287	0.452792
2	13021011100	0.369543	0.435106	0.423584
3	40131050105	0.097562	0.390880	0.380768
4	37119005200	-0.255410	0.368388	0.349797
5	22113950600	-0.153783	0.352597	0.374065
6	40109102500	-0.134369	0.344746	0.340528
7	12081002025	0.192506	0.342930	0.330297
8	13135050747	0.019494	0.342184	0.329722
9	40109101300	0.009879	0.339300	0.354523

Sort by LightGBM

	tract	FE_rent_growth_2025	XG_rent_growth_2025	LGB_rent_growth_2025
0	18043070801	1.276547	0.561004	0.605929
1	06037800205	0.586632	0.466287	0.452792
2	13021011100	0.369543	0.435106	0.423584
3	40131050105	0.097562	0.390880	0.380768
4	04013050620	0.364423	0.326314	0.379444
5	22113950600	-0.153783	0.352597	0.374065
6	12091020800	-0.213282	0.328451	0.367889
7	40109101300	0.009879	0.339300	0.354523
8	37119005200	-0.255410	0.368388	0.349797
9	40109102500	-0.134369	0.344746	0.340528

USC Marshall School of Business

Final Model Selection and Analysis

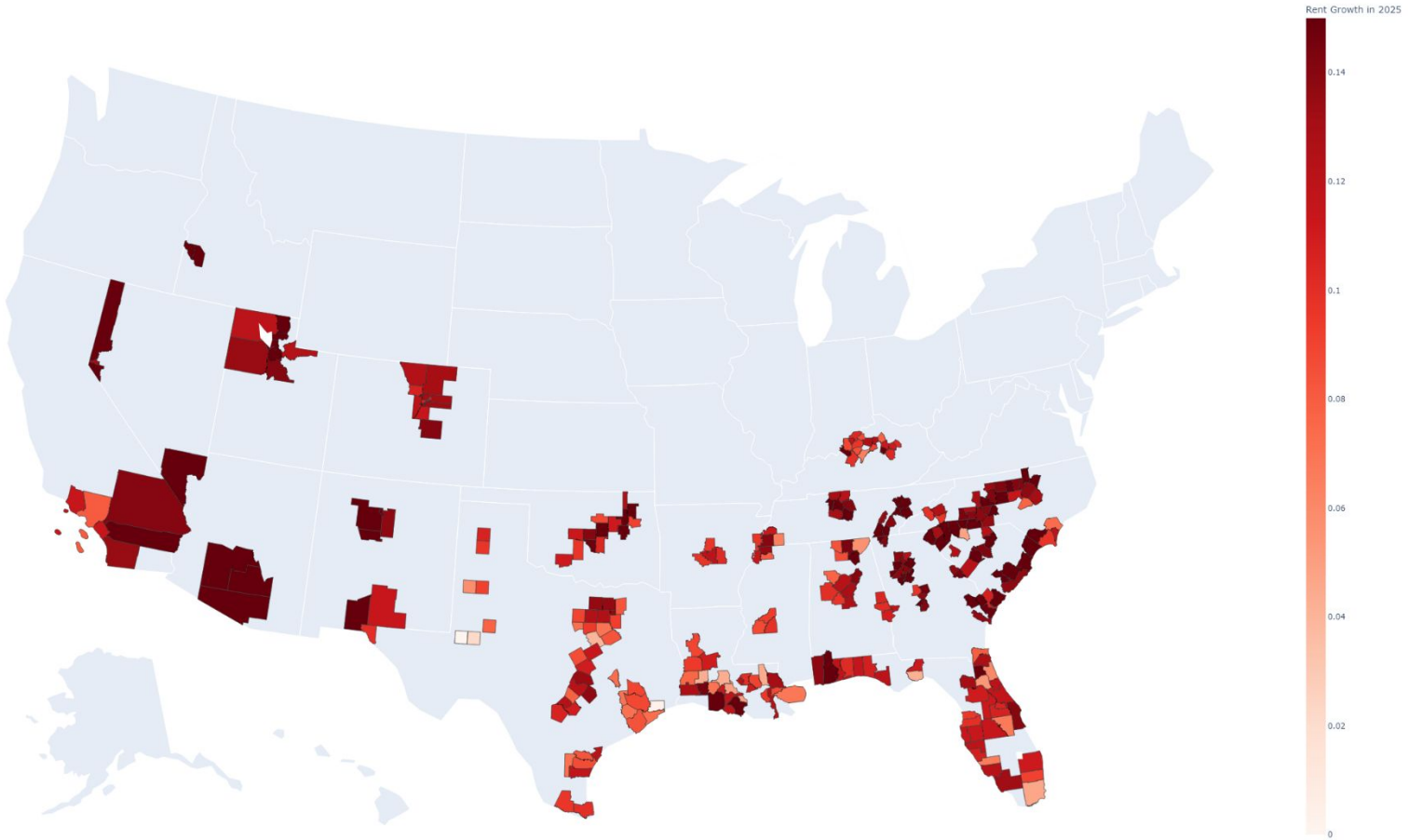
- Final model: **XGBoost**
- R-Squared Value: **0.9644**
- Expected 3 Year ROI* ($10m): **$9.17 Million Profit**
- Causes of Time/Series Model Limitations:
  - Short Time Period?
  - Linearity?

* Assuming 4% cap rate, 50% expense rate, and $1M invested in each of the top 10 tracts.

USC Marshall School of Business

Overview of Results

1. Tract: 18043070801 Expected Growth: 0.561
2. Tract: 06037800205 Expected Growth: 0.466
3. Tract: 13021011100 Expected Growth: 0.435
4. Tract: 40131050105 Expected Growth: 0.391
5. Tract: 37119005200 Expected Growth: 0.368
6. Tract: 22113950600 Expected Growth: 0.353
7. Tract: 40109102500 Expected Growth: 0.345
8. Tract: 12081002025 Expected Growth: 0.343
9. Tract: 13135050747 Expected Growth: 0.342
10. Tract: 40109101300 Expected Growth: 0.339



USC Marshall School of Business

Qualitative Significance of Tracts

1. Tract: 18043070801
   Location: (Floyd County, Indiana)
   Expected Growth: 0.561
   Summary: Hispanic population increased, access to many public schools, near Indiana University Southeast and museums

2. Tract: 22113950600
   Location: (Vermilion County, Louisiana)
   Expected Growth: 0.353
   Summary: Total housing units increased by 10%, 9.5% population growth, hospitals are closely accessible

3. Tract: 13021011100
   Location: (Bibb County, Georgia)
   Expected Growth: 0.435
   Summary: Near attractions, access to many publics schools, university, banks, restaurants

4. Tract: 13135050747
   Location: (Gwinnett County, Georgia)
   Expected Growth: 0.342
   Summary: Total housing units increased by 31.8%, population growth 55.5%, near attractions, access to many schools and hospitals

5. Tract: 37119005200
   Location: (Mecklenburg County, North Carolina)
   Expected Growth: 0.368
   Summary: white population increased by 90%, Hospitals, clinics or other healthcare facilities are closely accessible, neighborhood walkability is good

USC Marshall School of Business

Qualitative Significance of Tracts

6. Tract: 06037800205
   Location: (Los Angeles County, California)
   Expected Growth: 0.466
   Summary: Malibu area, close to city, high income people, almost 10% WFH and 40% migration from last year

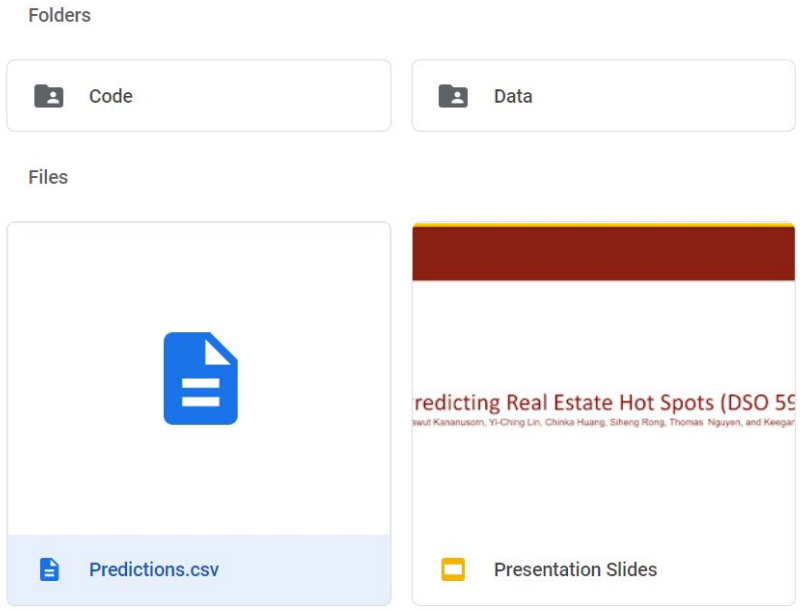
7. Tract: 40131050105
   Location: (Rogers County, Oklahoma)
   Expected Growth: 0.391
   Summary: In small city near Rogers State University, and many hospitals,

8. Tract: 40109102500
   Location: (Oklahoma County, Oklahoma)
   Expected Growth: 0.345
   Summary: High population growth in Oklahoma (~17%), new jobs, low unemployment rate(~1.6%), ~30% movie-in last year

9. Tract: 40109101300
   Location: (Oklahoma County, Oklahoma)
   Expected Growth: 0.339
   Summary: High population growth in Oklahoma (~17%), new jobs, low unemployment rate(~1.6%)

10. Tract: 12081002025
    Location: (Manatee County, Florida)
    Expected Growth: 0.343
    Summary: In the golf area, near many attractions, recently government investment in infrastructure, education, communities

Prediction and Code Submission



Prediction CSV

	Tract	XG_rent_growth_2025	ROI
2	04013815401	0.568875749	1.068875749
3	18043070801	0.561004425	1.061004425
4	06037800205	0.466286662	0.966286662
5	13021011100	0.435106393	0.935106393
6	40131050105	0.39088031	0.89088031
7	37119005200	0.368387849	0.868387849
8	22113950600	0.352597301	0.852597301
9	40109102500	0.344745964	0.844745964
10	12081002025	0.342929524	0.842929524
11	13135050747	0.342184381	0.842184381
12	40109101300	0.339300308	0.839300308
13	1097002502	0.330973507	0.830973507

https://drive.google.com/drive/folders/1Fj6biog-kUq5GLSRYcDUvklzCTc-RlKj?usp=sharing

USC Marshall School of Business

Data Sources

Data	Factors	Sources
Economic Data	- Tax Rates (Sales, Income, Corporate, Combined) - Debt - Bond - National Indicators (SP 500)	- Individual Tax Rate: State Individual Income Tax Rates - Corporate Tax Rate: State Corporate Income Tax Rates - Sales Tax Rate: State Sales Tax Rate - Debt: Debt in America - US Bond:US Bond and Returns
Location Data	- Starbucks, Banks, Public School, Private Schools, Hospitals (Normalized by Location Density) - Zip code area	- Starbucks dataset: Starbucks Locations Worldwide | Kaggle - Public schools: Public Schools | Public Schools | HIFLD Open Data (arcgis.com) - Private schools: Private Schools | Private Schools | HIFLD Open Data (arcgis.com) - Colleges: Colleges and Universities | Colleges and Universities | HIFLD Open Data (arcgis.com) - Banks:FDIC Insured Banks | FDIC Insured Banks | HIFLD Open Data (arcgis.com) - Hospitals: Hospitals | Hospitals | HIFLD Open Data (arcgis.com)
Demographic Data	- Population (gender, age, race) - Median Household Income - Employment Rate	- ACS - TIGER
Housing Data	- Occupancy Rate - Housing Supply - Property Values	- Yardi Property - FRED New Private Housing Units Authorized by Building Permits: 1-Unit Structures - Zilliow Rent Index https://www.zillow.com/research/data/
Crime Data	- Crime Rate	Crime Data Explorer

Research Sources

Research for modeling:

- Filling missing values
  - https://www.machinelearningplus.com/time-series/time-series-analysis-python/
- Fixed Effects, Random Effects, Pooled OLS
  - https://www.ucl.ac.uk/~uctqiax/PUBLG100/2016/week8/seminar8.html
  - https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8
- LSTM
  - https://github.com/melanieshi0120/COVID-19_global_time_series_panel_data/blob/master/Time_series_panel_data_model_part_confiremd_cases.ipynb
  - https://melaniesoek0120.medium.com/covid-19-global-data-time-series-prediction-with-lstm-recurrent-neural-networks-f7825c4a1f6f
- Feature Selection
  - https://towardsdatascience.com/application-of-feature-selection-techniques-in-a-regression-problem-4278e2efd503

Questions?