Predicting the Housing Market based on Demographic Drifts

EC503: Learning from Data

Ruben F. Carbajal + Axel S. Toro Vega

Background

- Across the United States, gentrification has emerged as a significant topic of discussion, particularly concerning its effects on various communities

- Media predominantly focuses on how specific demographic groups, particularly low-income and minority populations, are disproportionately impacted

Houston, you have a *Problem*

What are we doing?

- *Exploring Relationships*: Analyzing the correlation between growth rates in the Housing and Renting Markets and demographic shifts in various counties across the United States

- *Predictive Modeling*: Utilizing binary classification and regression models to forecast growth rates in the Housing Market based on demographic changes in respective counties

Raw

Raw Data

U.S. Census

- Demographics Data by Zip Code:
  - 398748 data points
  - 358 Features
    - 2 Grouping Variables & 356 Numerical

GEO_ID	NAME	DP05_0001E	DP05_0001M	DP05_0002E	DP05_0002M	DP05_0003E	DP05_0003M
Geography	Geographic Area Name	Estimate!!SEX AND AGE!!Total population	Margin of Error!!SEX AND AGE!!Total population	Estimate!!SEX AND AGE!!Total population!!Male	Margin of Error!!SEX AND AGE!!Total population!!Male	Estimate!!SEX AND AGE!!Total population!!Female	Margin of Error!!SEX AND AGE!!Total population!!Female
860Z200US00601	ZCTA5 00601	16834	506	8337	227	8497	318
860Z200US00602	ZCTA5 00602	37642	205	18405	84	19237	163
860Z200US00603	ZCTA5 00603	49075	963	23813	585	25262	561
860Z200US00606	ZCTA5 00606	5590	264	2723	216	2867	135
860Z200US00610	ZCTA5 00610	25542	344	12317	225	13225	169
860Z200US00611	ZCTA5 00611	1315	382	667	225	648	204
860Z200US00612	ZCTA5 00612	63312	1805	29745	943	33567	1053
860Z200US00616	ZCTA5 00616	9625	1319	4515	661	5110	803
860Z200US00617	ZCTA5 00617	22573	241	10709	144	11864	124
860Z200US00622	ZCTA5 00622	7577	979	3334	470	4243	619
860Z200US00623	ZCTA5 00623	39406	979	18869	470	20537	619
860Z200US00624	ZCTA5 00624	21648	516	10612	323	11036	248
860Z200US00627	ZCTA5 00627	32733	*****	15525	*****	17208	*****

Zillow

- Housing Data by Zip Code:
  - 26351 data points
  - 300 Features
    - 9 Grouping Variables & 291 Numerical

RegionID	SizeRank	RegionNar	RegionTyp	StateNam	State	City	Metro	CountyNa	1/31/2000	2/29/2000	3/31/2000
91982	1	77494	zip	TX	TX	Katy	Houston-T	Fort Bend	211762.0785	211945.8689	212436.9787
61148	2	8701	zip	NJ	NJ	Lakewood	New York-	Ocean Cou	136347.9094	136910.6557	137291.012

- Renting Data by Zip Code:
  - 6994 data points
  - 120 Features
    - 9 Grouping Variables & 111 Numerical

RegionID	SizeRank	RegionNar	RegionTyp	StateNam	State	City	Metro	CountyNa	1/31/2015	2/28/2015	3/31/2015
91982	1	77494	zip	TX	TX	Katy	Houston-T	Fort Bend	1485.699498	1491.481639	1498.76234
61148	2	8701	zip	NJ	NJ	Lakewood	New York-	Ocean County			

trust the *PreProcess*

Preprocessing Data: U.S. Census & Zillow

U.S. Census

- Get Demographics Percentages by Year
- Clean missing data
- Generate Demographics Drifts
  - Compare with previous year
- Keep 2 Grouping Variables
  - Year and Zip Code

GEOGRAPHY	YEAR	WHITE	BLACK_OR_AFRICAN_AMERICAN	AMERICAN_INDIAN_AND_ALASKA_NATIVE	ASIAN	NATIVE_HAWAIIAN_AND_OTHER_PACIFIC_ISLANDER	HISPANIC_OR_LATINO
601	2012	0.757575758	-28.57142857	0	0	0	0.100908174
601	2013	1.611170784	-20	0	0	0	0.302419355
601	2014	1.691331924	25	0	0	0	0.301507538
601	2015	-1.975051975	-40	0	0	0	0
601	2016	-13.99787911	0	0	0	0	0
601	2017	-4.069050555	16.66666667	0	0	0	-0.200400802
601	2018	-2.956298201	14.28571429	0	0	0	0.100401606
601	2019	-2.38410596	37.5	100	0	0	-0.100300903
601	2020	-1.085481682	36.36363636	-50	0	0	0
601	2021	15.9122085	-6.666666667	0	0	0	-0.100401606
601	2022	-0.355029586	50	0	0	0	-0.100502513
602	2012	-12.80558789	0	0	100	0	0.64171123
602	2013	-15.08678238	-18.60465116	0	0	0	-0.531349628
602	2014	-8.018867925	-22.85714286	0	0	0	-0.427350427
602	2015	-1.709401709	-37.03703704	-100	0	0	0.429184549
602	2016	0.52173913	41.17647059	0	0	0	-0.106837607
602	2017	15.74394464	16.66666667	0	-100	0	0.213903743
602	2018	18.68460389	0	0	0	0	-0.213447172
602	2019	7.304785894	-14.28571429	0	0	0	-1.069518717
602	2020	-10.21126761	12.5	200	0	0	0.216216216
602	2021	-19.86928105	-33.33333333	-66.66666667	0	0	1.510248112
602	2022	-19.90212072	-16.66666667	0	0	0	0.531349628
603	2012	-5.269058296	3.125	0	66.6667	0	-1.139896373
603	2013	-8.284023669	0	0	60	0	0.524109015

Zillow

- Housing Data & Renting Data:
  - Calculate the annual growth rate
    - Finding annual average
    - Get the percentage growth between years
  - Clean missing data
  - Strip down to 3 Grouping Variables
    - Year, Zip Code, and County

Zipcode	CountyName	Year	GrowthRate
1001	Hampden County	2012	-2.451772316
1001	Hampden County	2013	0.973817444
1001	Hampden County	2014	0.441570136
1001	Hampden County	2015	1.026203594
1001	Hampden County	2016	4.639393165
1001	Hampden County	2017	3.905005302
1001	Hampden County	2018	3.729916617
1001	Hampden County	2019	3.549387713
1001	Hampden County	2020	5.584402025
1001	Hampden County	2021	12.27828987
1001	Hampden County	2022	10.40270516

Data on *Data*

Combined Dataset

U.S. Census + Zillow

- Merged datasets using matching Year and Zip Code
- 1 Grouping Variable, County; removed Year and Zip Code post-merge
- 7 Numerical features
- Housing: 250,127 points
- Renting: 13,407 points

CountyName	GrowthRate	WHITE	BLACK_OR_AFRICAN_AMERICAN	AMERICAN_INDIAN_AND_ALASKA_NATIVE	ASIAN	NATIVE_HAWAIIAN_AND_OTHER_PACIFIC_ISLANDER	HISPANIC_OR_LATINO
Abbeville County	7.461725809	-2.503681885	2.950819672	0	-50	0	66.66666667
Abbeville County	8.630875159	-1.208459215	5.732484076	0	0	0	-40
Abbeville County	4.710787865	-0.458715596	2.108433735	-100	-100	0	-33.33333333
Abbeville County	14.16990524	0.153609831	0	0	0	0	300
Abbeville County	12.18864395	0.306748466	0.589970501	0	0	0	-37.5
Abbeville County	0.963353585	-2.293577982	1.759530792	0	0	0	100
Abbeville County	6.681921278	5.007824726	-8.933717579	0	0	0	-40
Abbeville County	4.286182504	-4.470938897	9.17721519	0	0	0	50
Abbeville County	7.293745603	2.496099844	-6.666666667	800	0	0	-22.22222222

Binarized Dataset*

Classes!

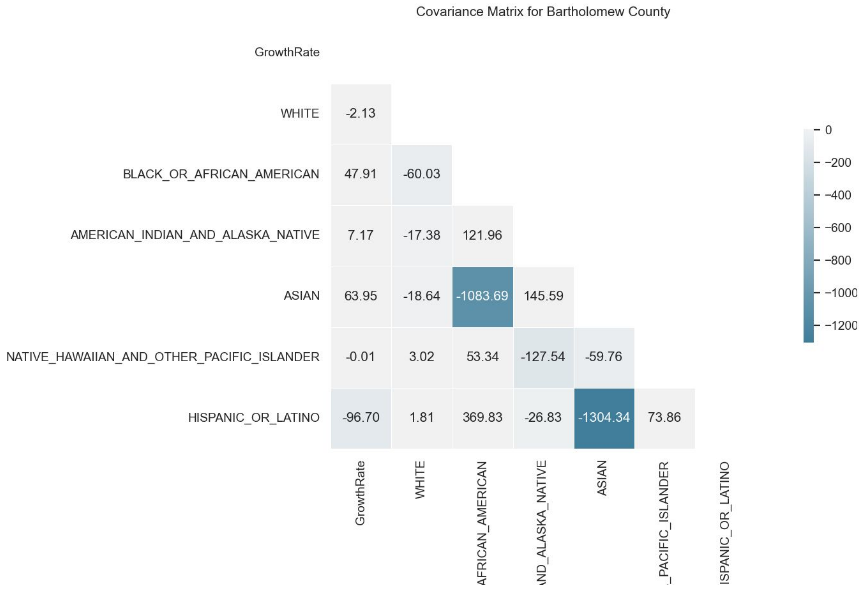
- Binned all features and labels
  - -1 = drift less than -1%
  - 0 = drift between -1% and 1%
  - +1 = drift greater than 1%
- Allows for Classification Problem!

GrowthRate	WHITE	BLACK_OR_AFRICAN_AMERICAN	AMERICAN_INDIAN_AND_ALASKA_NATIVE	ASIAN	NATIVE_HAWAIIAN_AND_OTHER_PACIFIC_ISLANDER	HISPANIC_OR_LATINO
-1	-1	1	0	1	0	1
0	-1	-1	0	1	1	1
0	0	-1	0	1	0	1
1	0	1	0	1	0	1
1	0	1	-1	1	-1	1
1	1	-1	0	-1	1	-1
1	0	1	0	0	-1	-1
1	0	1	0	-1	0	-1
1	-1	1	0	1	0	1
1	-1	1	0	1	0	-1
1	0	-1	0	-1	0	1

*** = Based only on housing data**

Covariance Matrices

Using Combined Data



Covariance Matrix Crash Course

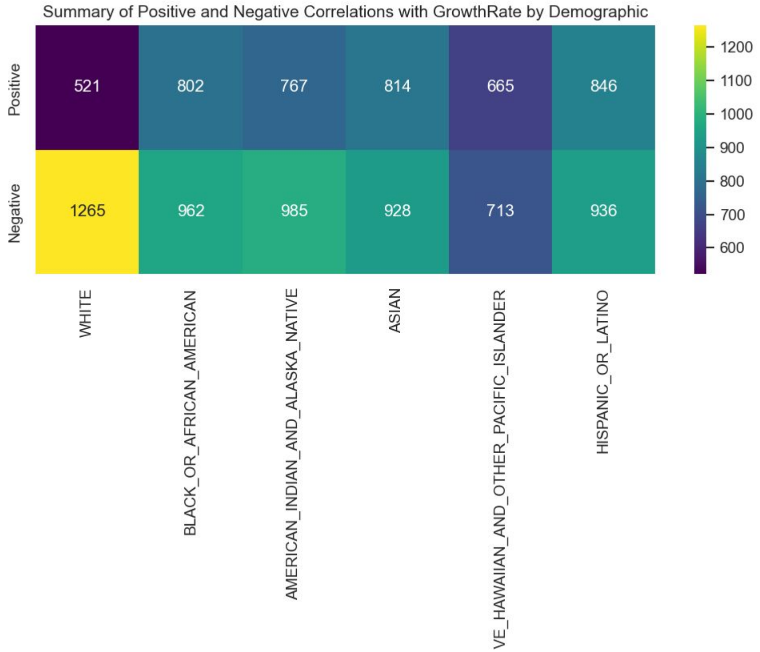
- +++ (Strong Positive Relationship)
- - - - (Strong Negative Relationship)
- 0 (No linear Relationship)

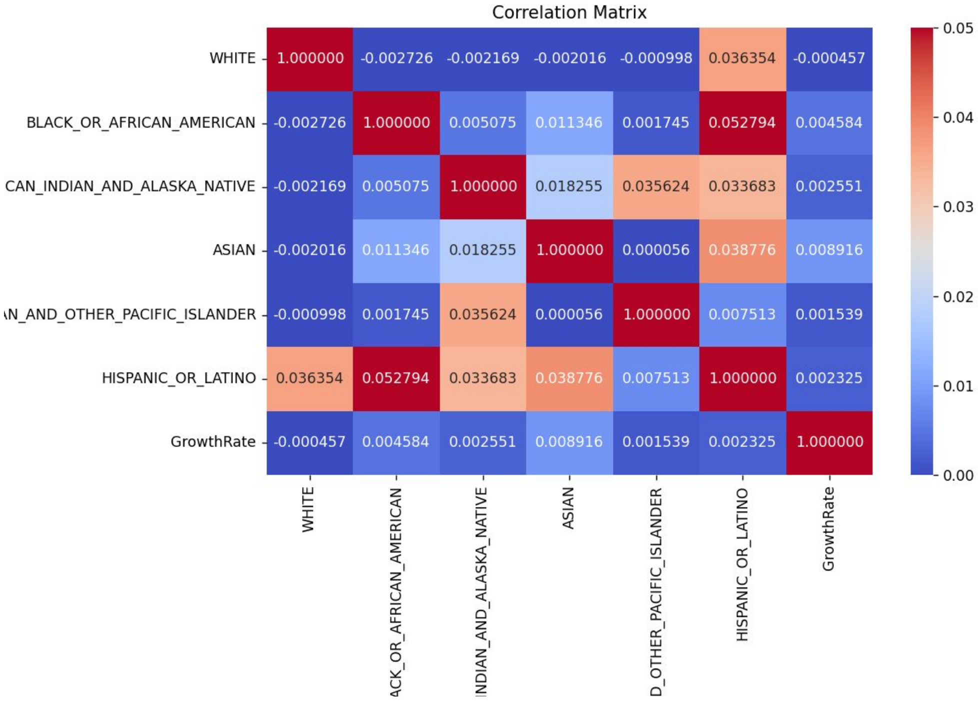
Dashboard

Correlated-*ish*

Using Combined Data

Housing Data:





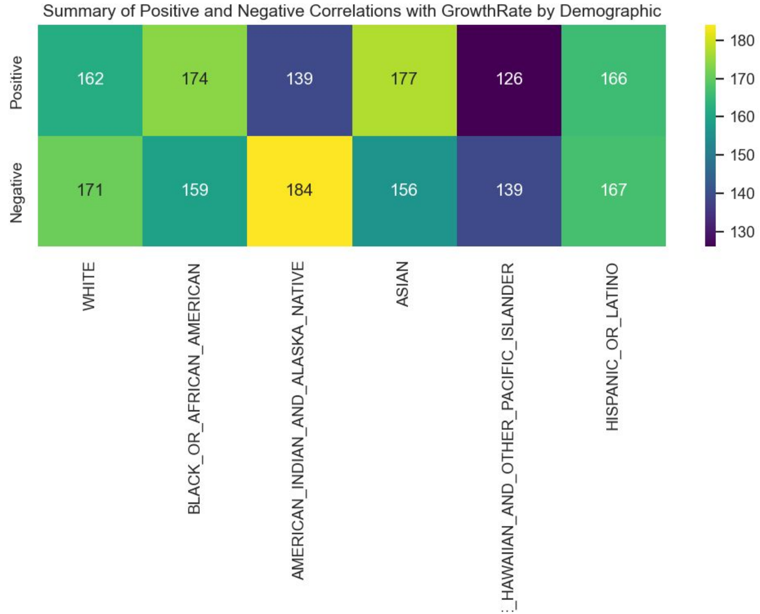
Correlation Matrix Crash Course

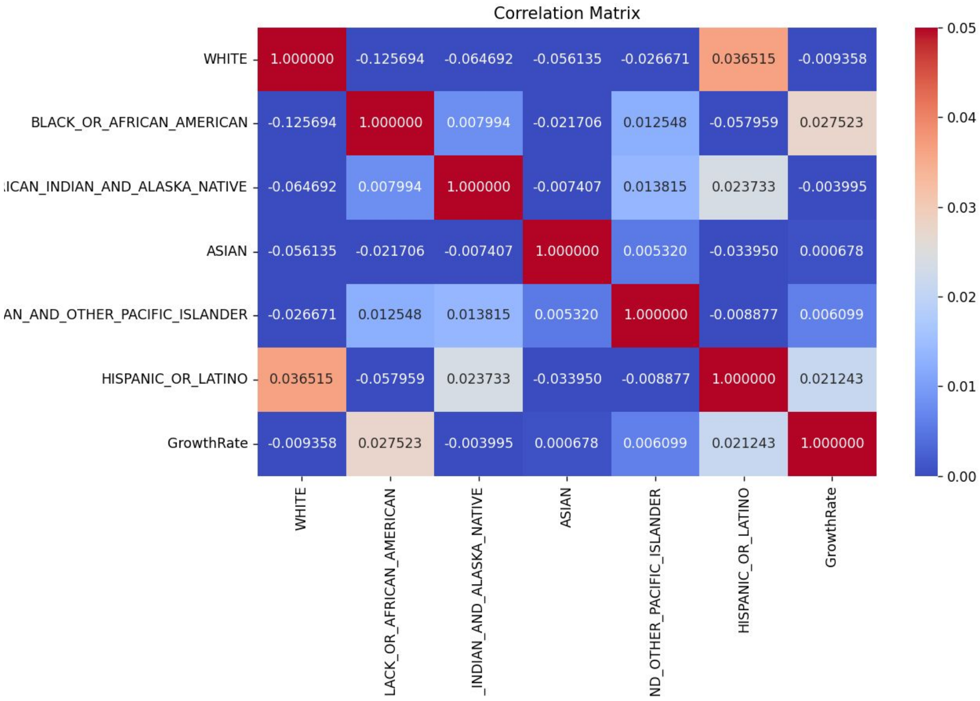
- **+1(Strong Positive Relationship)**
- **-1 (Strong Negative Relationship)**
- **0 (No linear Relationship)**

Correlated-*ish* pt.2

Using Combined Data

Renting Data:





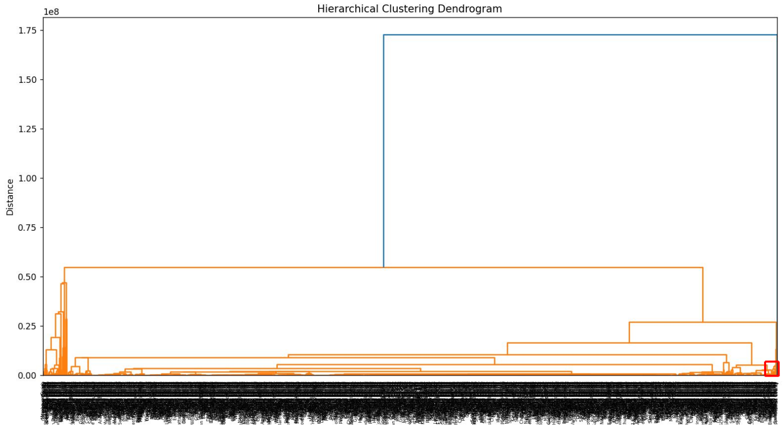
Correlation Matrix Crash Course

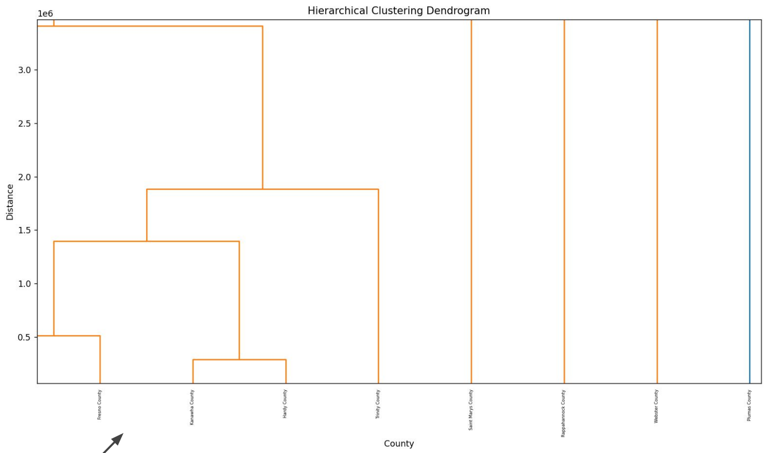
- **+1(Strong Positive Relationship)**
- **-1 (Strong Negative Relationship)**
- **0 (No linear Relationship)**

Dendrograms

Using Combined Data

Housing Data:

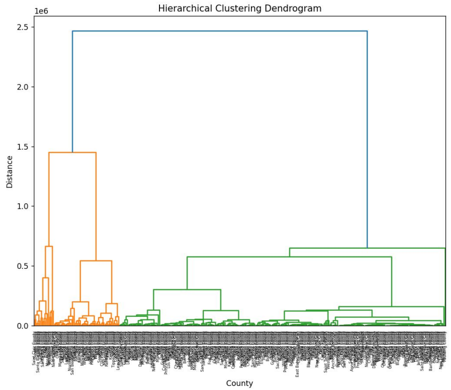




Why is Plumas County so Isolated?

- Extreme Linear Relationships

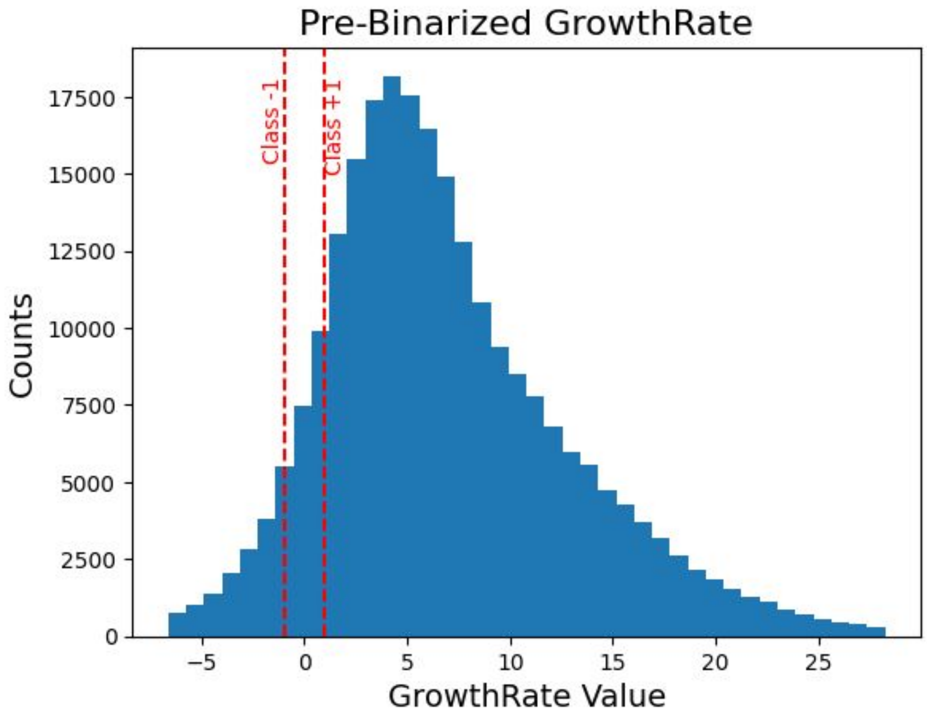
Renting Data:

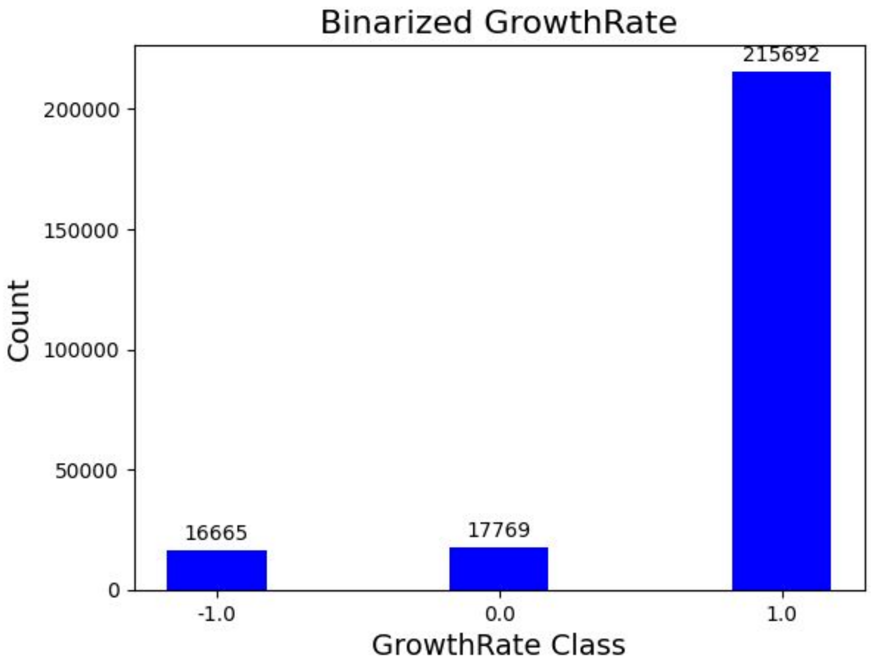


Dendrogram Crash Course

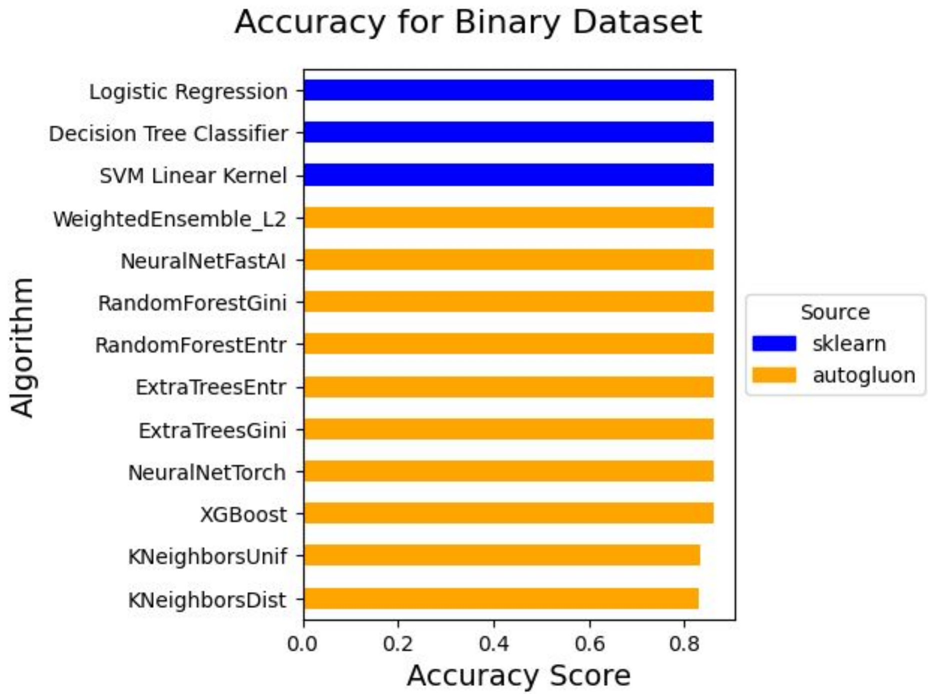
- Shows the clusters produced by hierarchical clustering
- Height of join reflects distance in clusters
- Meant to identify zip codes that are similar

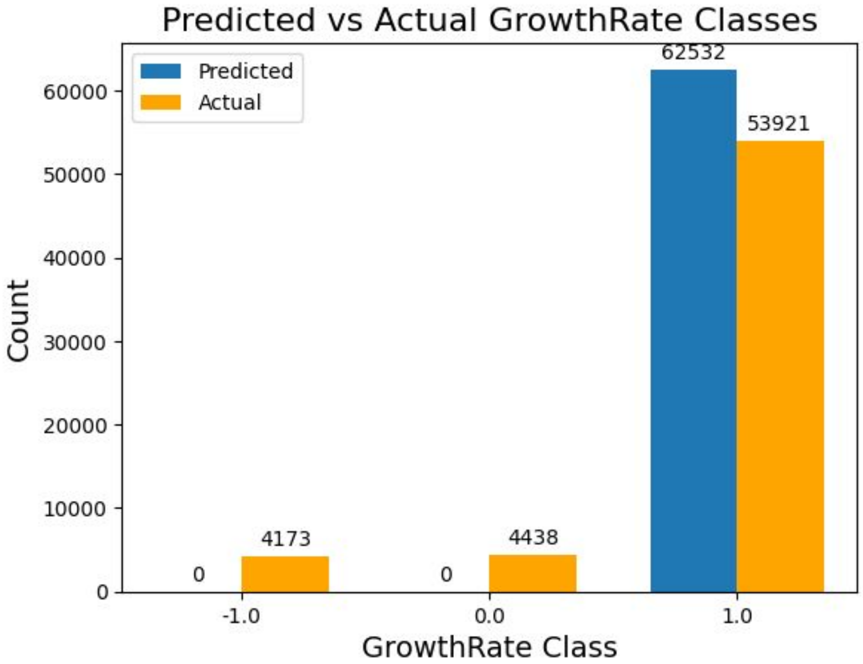
Binarized Dataset





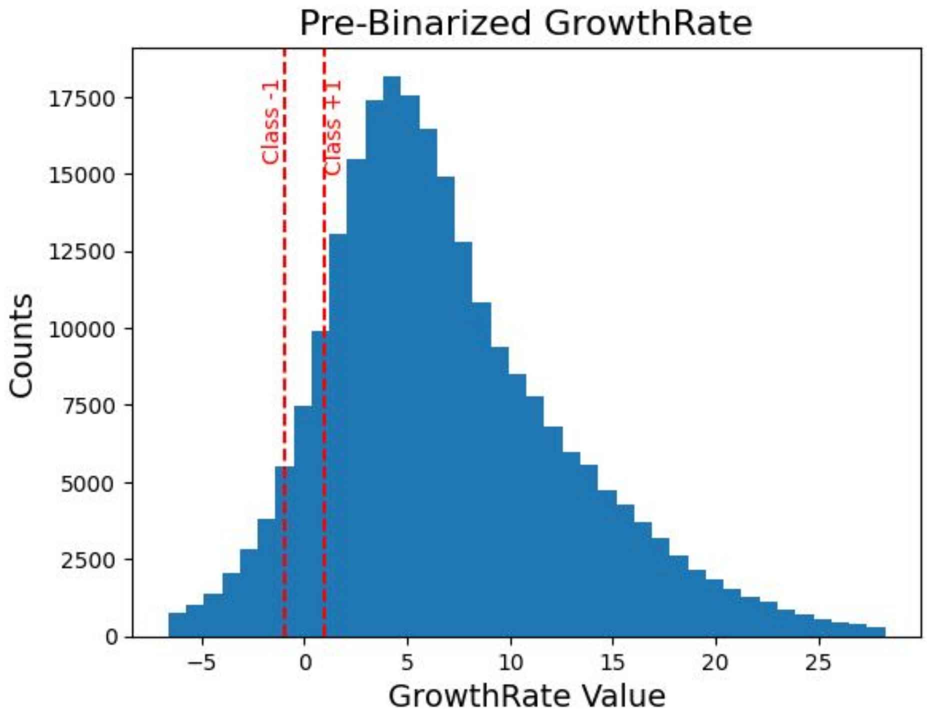
Binary: Failed Successfully!

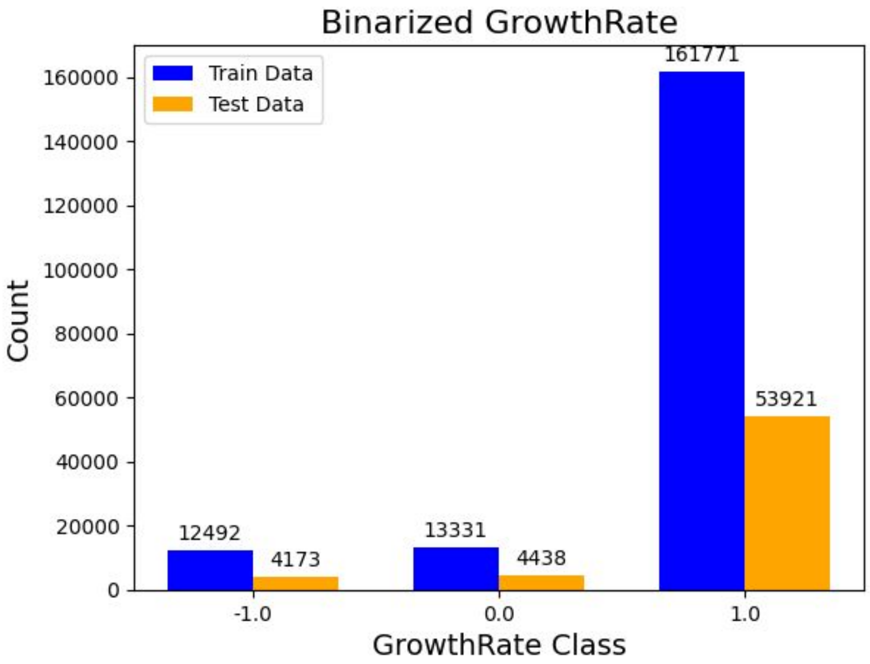




but *why?*

Unbalanced Training Data!

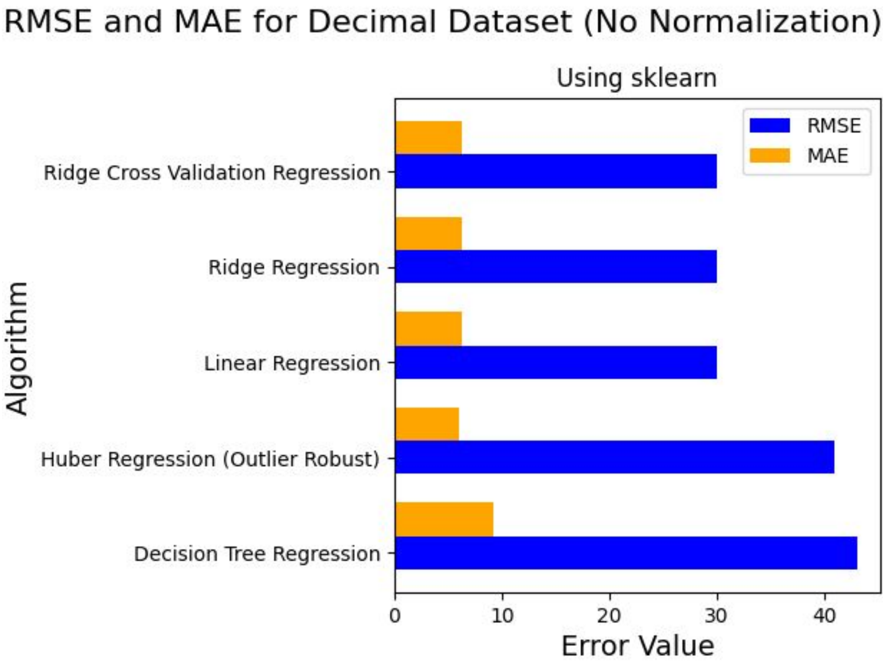




Back to regression: sklearn

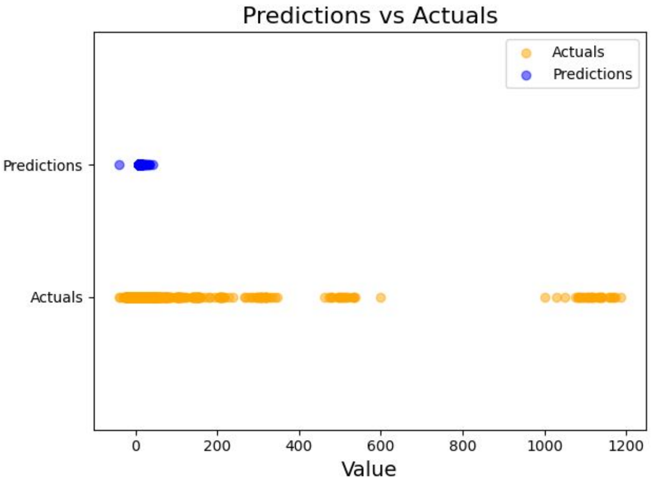
Algorithms

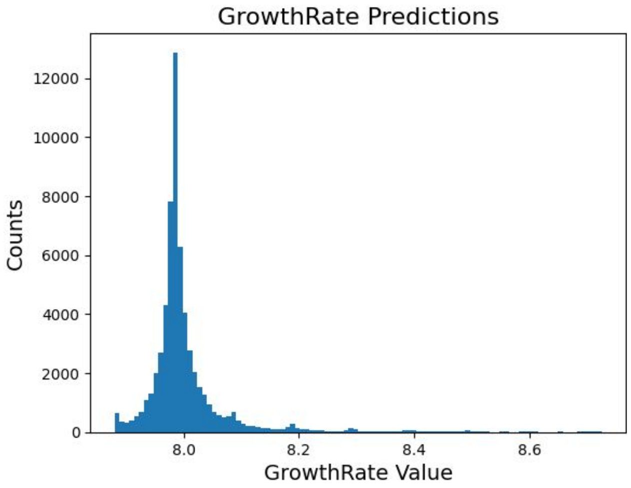
- Linear Regression (LR)
  - Predictive model algorithm
  - Estimates the relationship between one or more independent variables and a dependent variable
- RidgeCV Regression
  - Type of LR with L2 regularization
  - Uses CV to determine the regularization strength automatically to optimize model
- Huber Regression
  - Type of LR with more robust to outliers
  - Combines least squares and least absolute deviations properties in loss function
- Decision Tree Regression
  - Model non-linear data by recursively splitting data

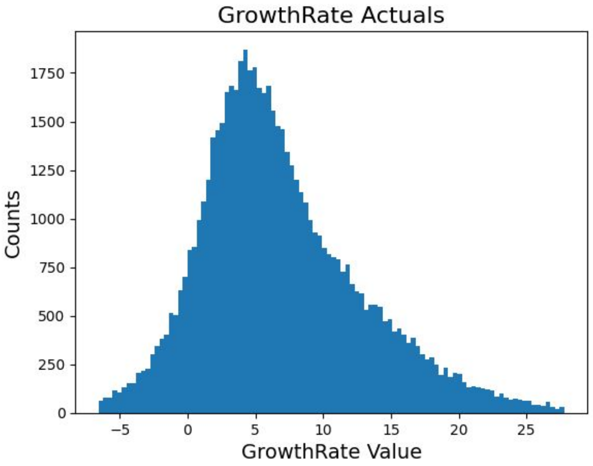


RidgeCV Results

Outliers!







Conclusion

- Our analysis using correlation and covariance matrices revealed minimal relationships between demographic shifts and growth rates across various counties

- The binary classification and regression models, while partially effective, indicated that predicting housing rates based solely on demographic changes might be challenging

Thank You!