Analysis for Best Cities to Live in California

Sudip , Tamasree & Hima

About the Project.

California is widely regarded as the coolest state in the United States. The tech capital of the world, Silicon Valley, and there's just about everything. It is one of the few places in the world where you can go from skiing to surfing, to sand-boarding in the desert – all on the same day.

It's a little wonder everyone wants to move here, and in actual fact, this happens to be the most populous state – and state with the largest economy in the U.S

Motivation

The New York Times

The Market Tectonics of California Real Estate

Some residents have left the state, but many others have just moved out of big cities in search of more space and lower prices, creating hot spots in the suburbs and the once-sleepy exurbs.

Facebook Twitter Email Like this story? Leave a tip!

Fast-Rising Housing Prices Changing South County Real Estate Market

Thursday, April 29, 2021

By **Alexandra Rangel**

Moving to California? The 10 Best Places to Live in The Golden State

Moving, Real Estate | 03/16/2018 | Marian White

america's best, best places to live, california, city moving, interstate move

Share this:

Facebook LinkedIn Twitter More

Need for analysis..

1. Understanding the Barriers of Affordable Housing

How Housing Prices related to other factors?

What effects the housing value much?

Understanding the housing price correlation between different terms.

Need for analysis..

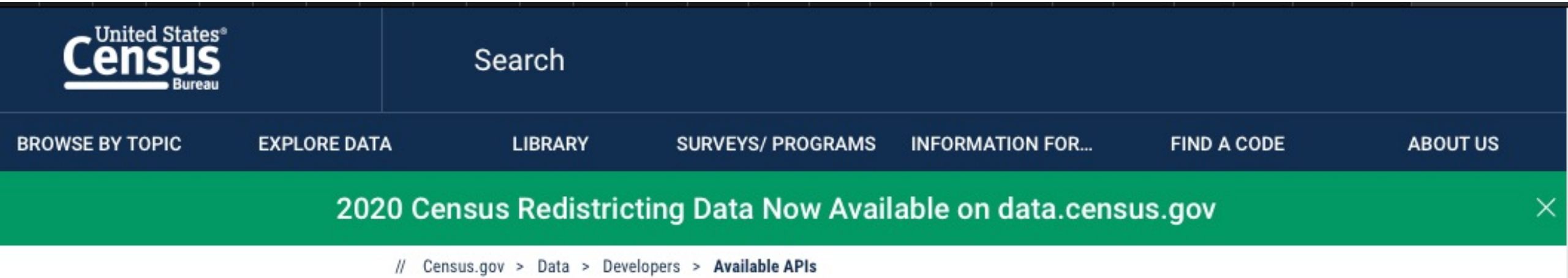
2. Most Affordable 5 Cities of California

We chose to analyze most common factors:

- **House Value**
- **Income**
- **Cost/Rent**
- **Poverty**
- **Public Transport**
- **Employment/unemployment rate**

Approach

Collecting the Data



Downloaded 5 years of US Census data for every zip code

Rows: 33000
Columns: 28

```
#shape of the dataframe
usa_2014.shape
```

```
(33120, 30)
```

Data Cleaning in Python

using pandas

Code — Python 3 (ipykernel)

[34…
```
#columns of 2017  dataframe
usa_2014.columns
```

[34…
```
Index(['Zipcode', 'Population', 'Median Age', 'Household Income',
       'Per Capita Income', 'Poverty Rate', 'Unemployment Rate', 'House Value',
       'House Construction Year', 'Monthly Owner Cost', 'Monthly Rent',
       'Public Transport Rate', 'Personal Transport Rate',
       'Commute Time Public', 'Commute Time Car', 'High School Rate',
       'College Rate', 'Uneducated Rate', 'English Language Rate',
       'Spanish Language Rate', 'White Population Rate',
       'Black Population Rate', 'Hispanic Population Rate',
       'Asian Population Rate', 'City', 'County', 'Lat', 'Lng',
       'Housing_units', 'State'],
      dtype='object')
```

[35…
```
#shape of the dataframe
usa_2014.shape
```

[35…
```
(33120, 30)
```

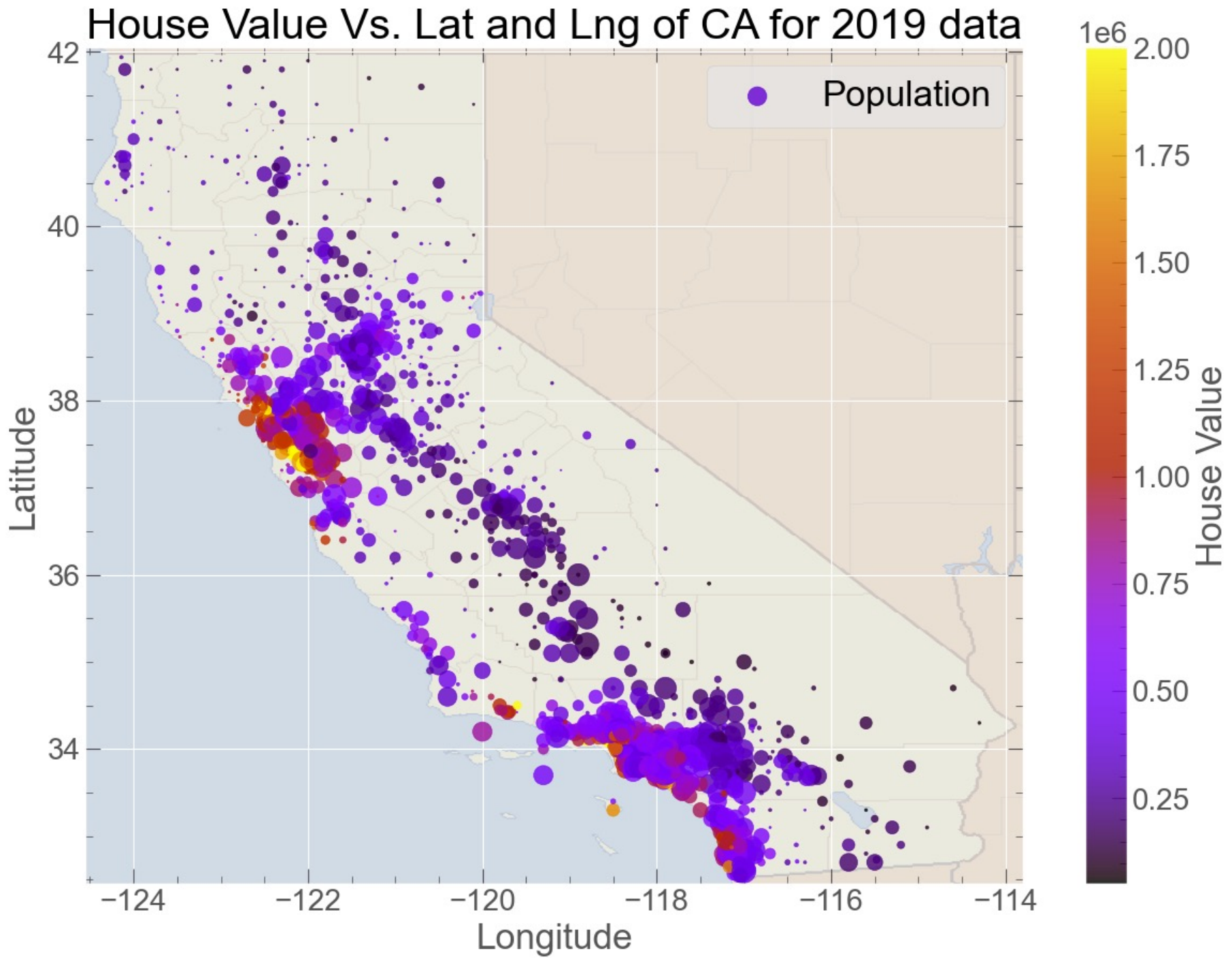
nshot 6…
```
#removing the rows with house value less than 1
usa_2014=(usa_2014[(usa_2014['House Value']>0)&
                   (usa_2014['Household Income']>0)&
                   (usa_2014['Monthly Owner Cost']>0)&
                   (usa_2014['Monthly Rent']>0)])
```

[37…
```
#looking for null value
usa_2014.isna().sum()
```

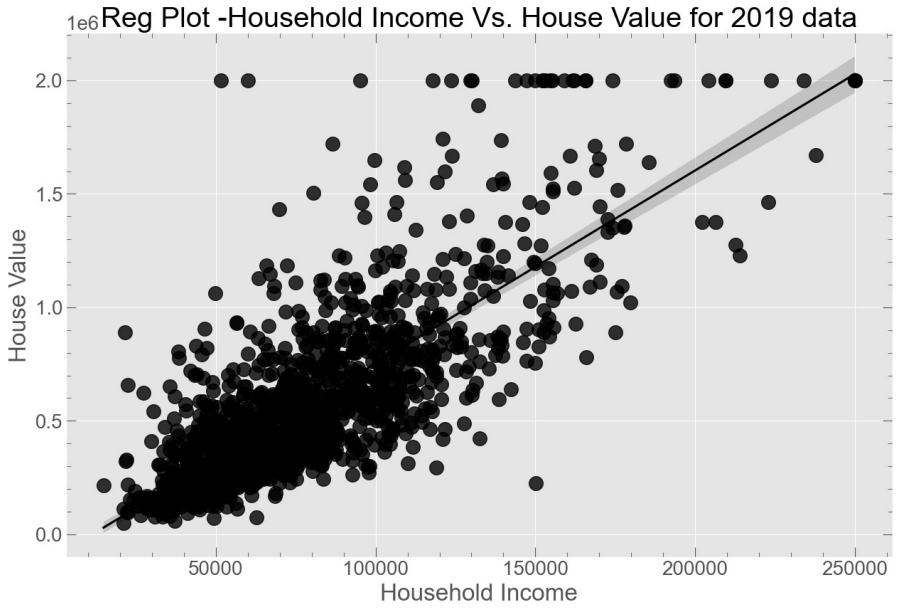
Data Analysis of CA State for 2019

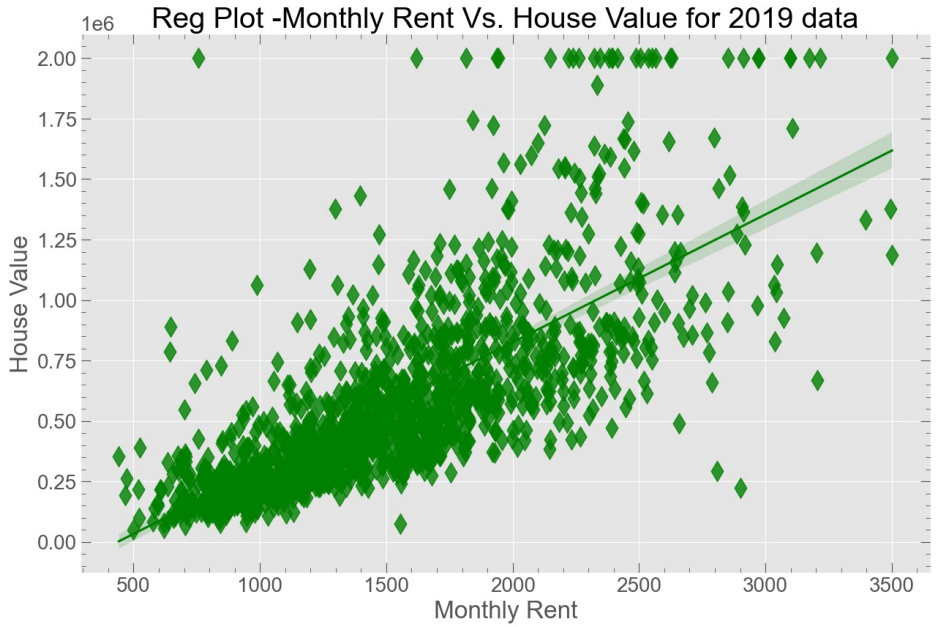
1. **Economy**
   'Household Income', 'Poverty Rate', 'House Value', 'Monthly Owner Cost', 'Monthly Rent'
2. **Education**
   'College Rate', 'Uneducated Rate', 'High School Rate', 'Housing units', 'Median Age'
3. **Transportation**
   'Public Transport Rate', 'Personal Transport Rate', 'Commute Time Car'
4. **Race**
   'Asian Population Rate', 'White Population Rate', 'Hispanic Population Rate', 'Black Population Rate'

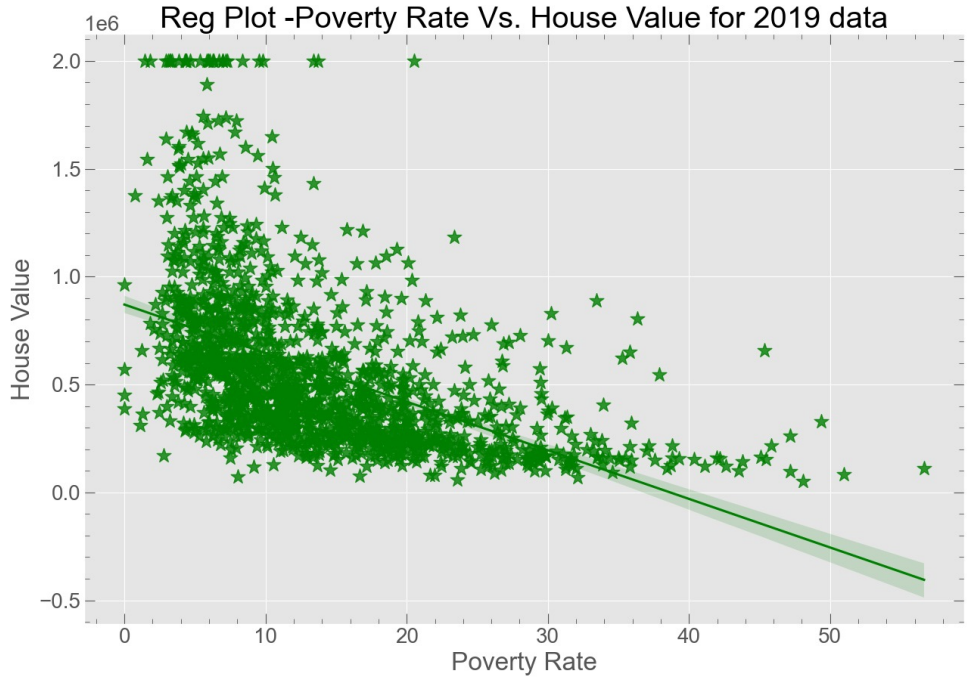
House Price

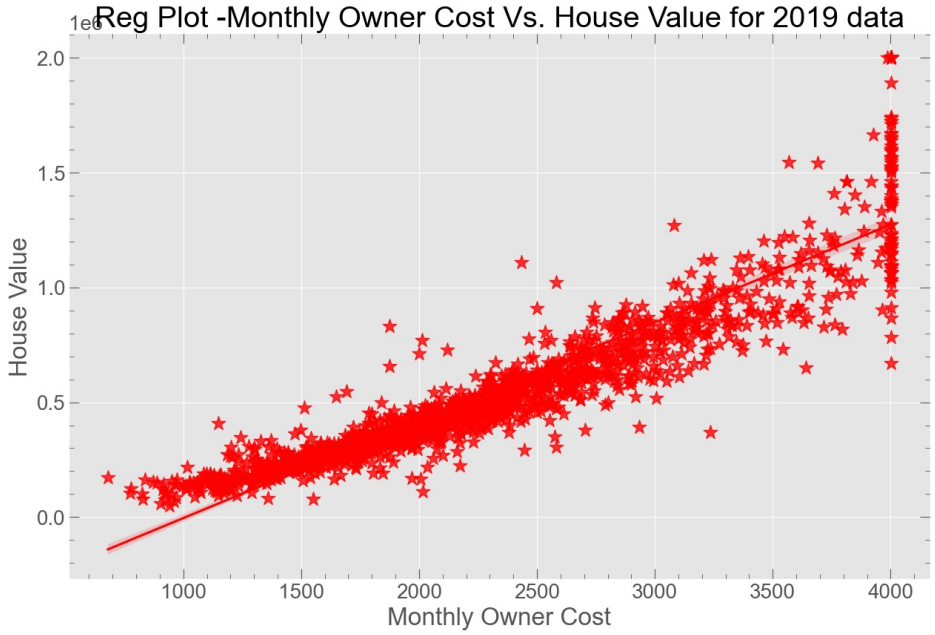

House Value Vs. Lat and Lng of CA for 2019 data
Population
Latitude
Longitude
42
40
38
36
34
−124
−122
−120
−118
−116
−114
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
House Value

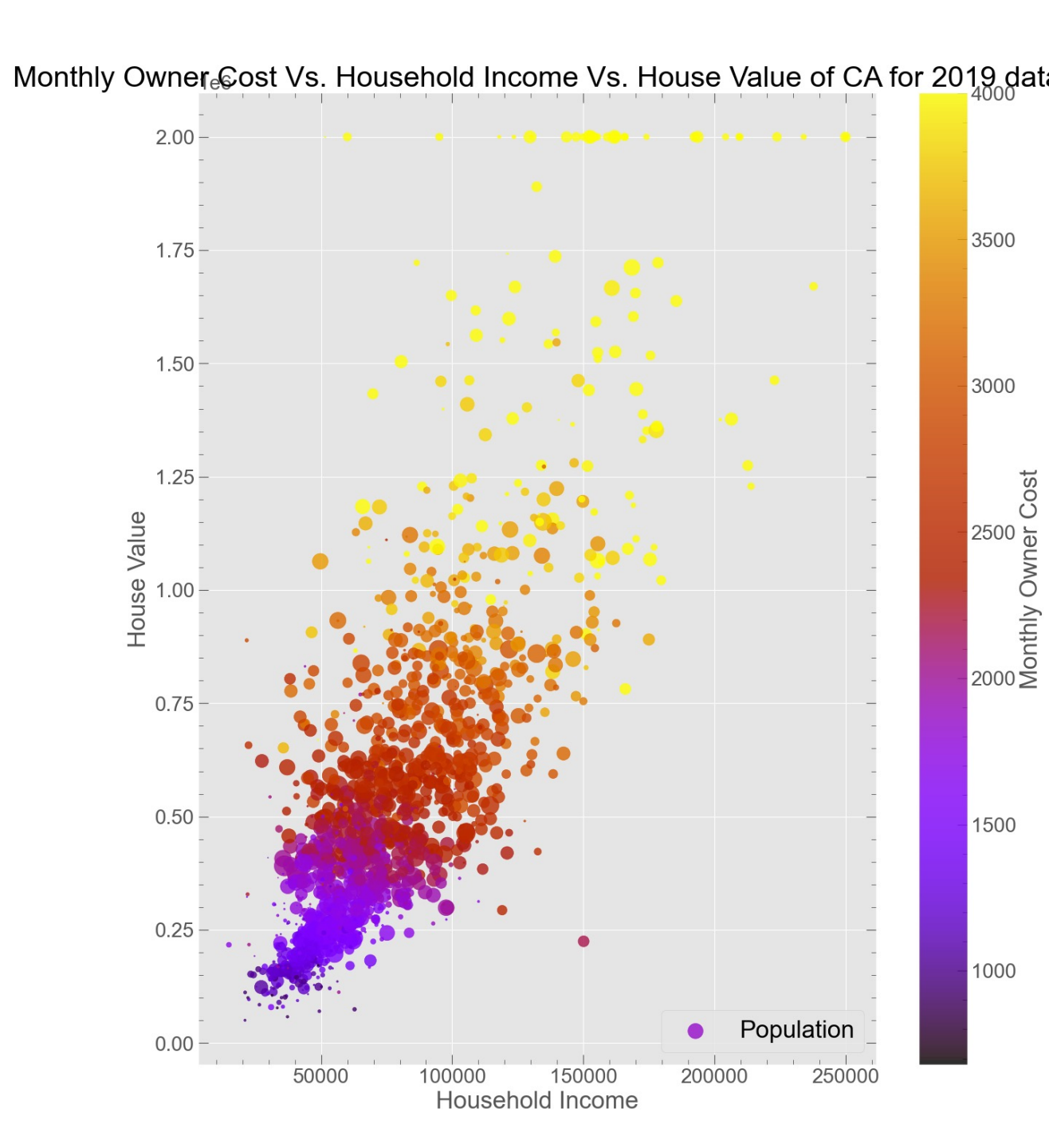
1. Economy

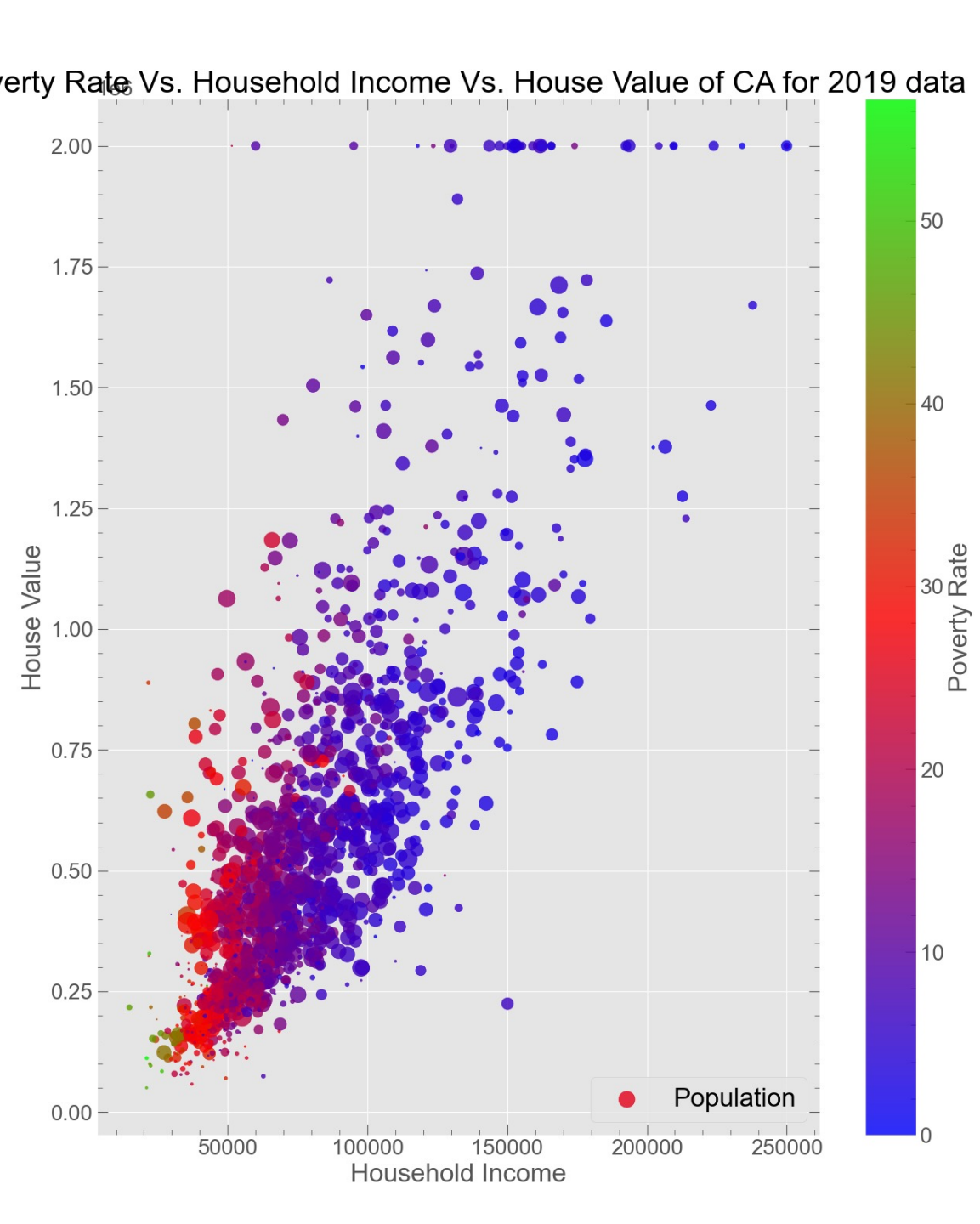

Reg Plot -Household Income Vs. House Value for 2019 data
1e6
House Value
Household Income
0.0
0.5
1.0
1.5
2.0
50000
100000
150000
200000
250000


Reg Plot -Monthly Rent Vs. House Value for 2019 data
1e6
House Value
Monthly Rent
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
500
1000
1500
2000
2500
3000
3500

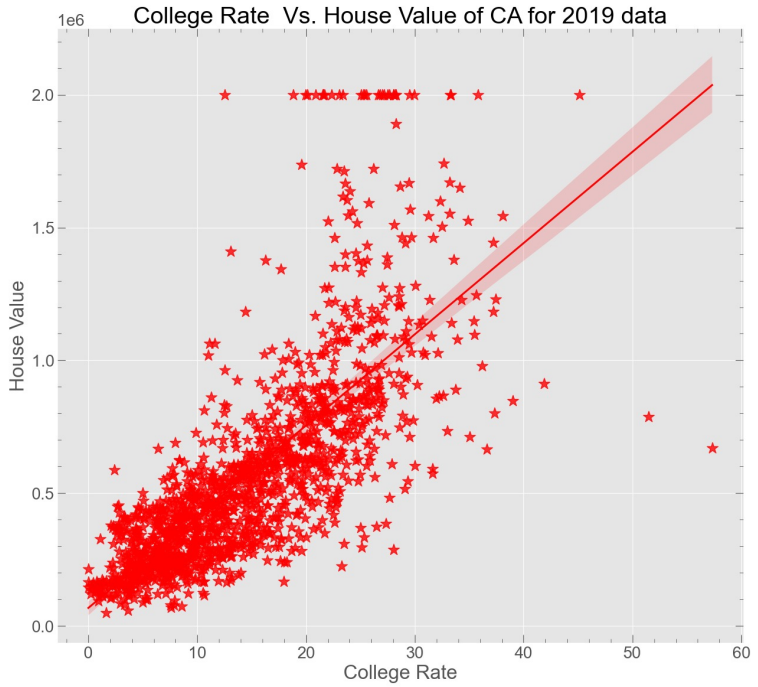

Reg Plot -Poverty Rate Vs. House Value for 2019 data
1e6
House Value
Poverty Rate
−0.5
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50

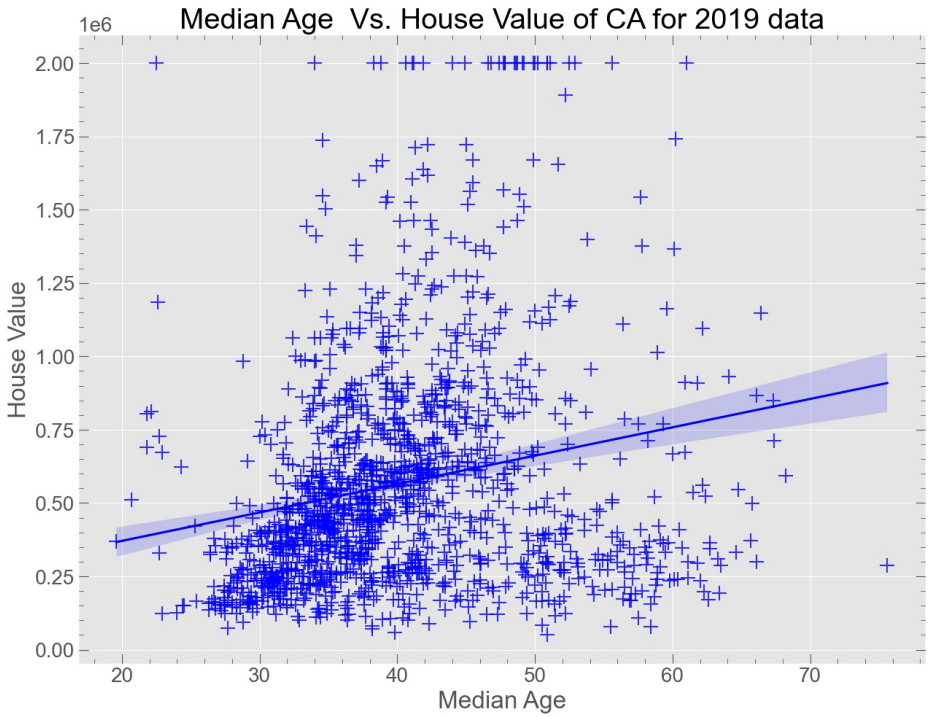

Reg Plot -Monthly Owner Cost Vs. House Value for 2019 data
1e6
House Value
Monthly Owner Cost
0.0
0.5
1.0
1.5
2.0
1000
1500
2000
2500
3000
3500
4000

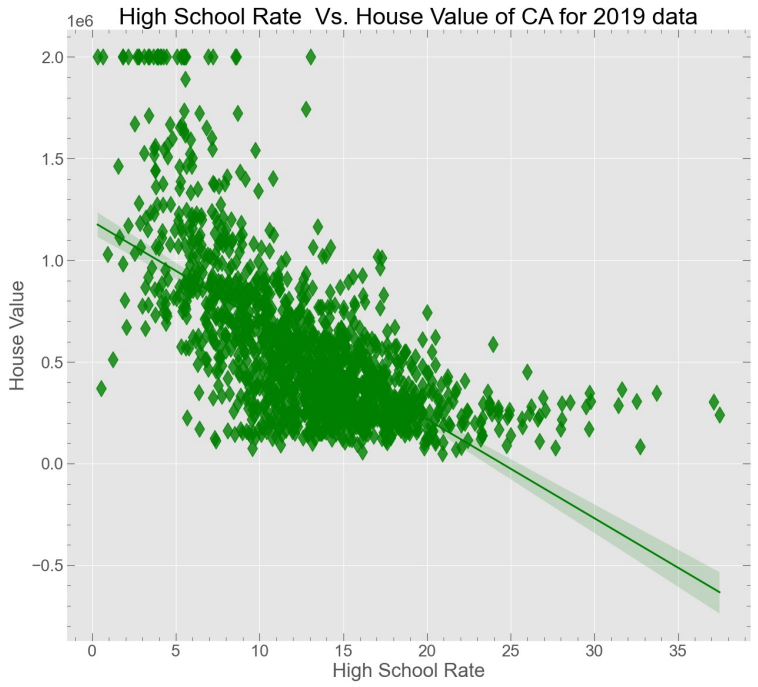

Monthly Owner Cost Vs. Household Income Vs. House Value of CA for 2019 data
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
House Value
50000
100000
150000
200000
250000
Household Income
Population
4000
3500
3000
2500
2000
1500
1000
Monthly Owner Cost

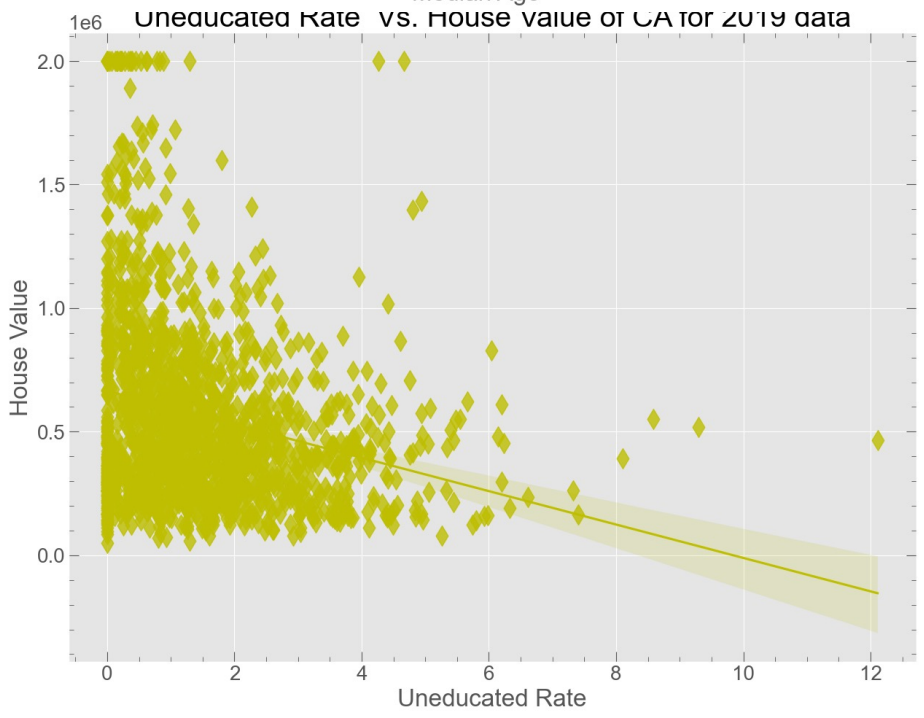

Poverty Rate Vs. Household Income Vs. House Value of CA for 2019 data
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
House Value
50000
100000
150000
200000
250000
Household Income
Population
50
40
30
20
10
0
Poverty Rate

2. Education

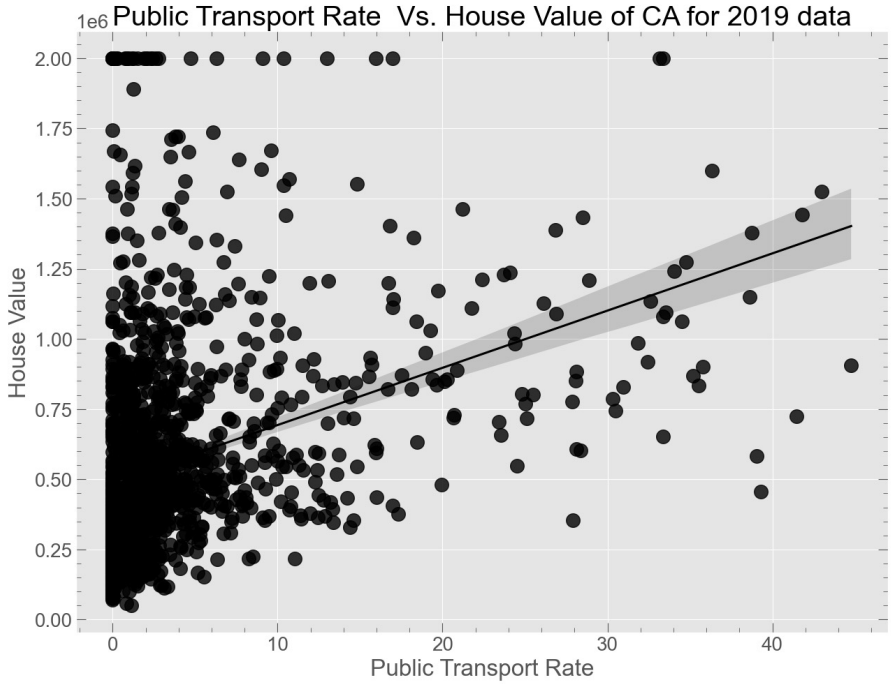

College Rate Vs. House Value of CA for 2019 data
1e6
House Value
College Rate

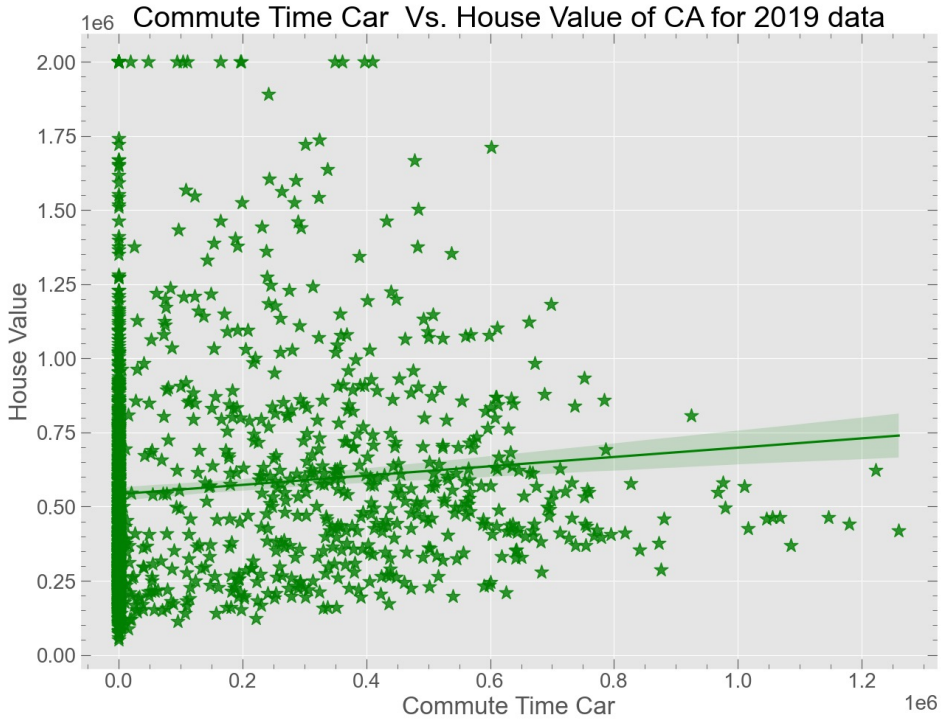

Median Age Vs. House Value of CA for 2019 data
1e6
House Value
Median Age

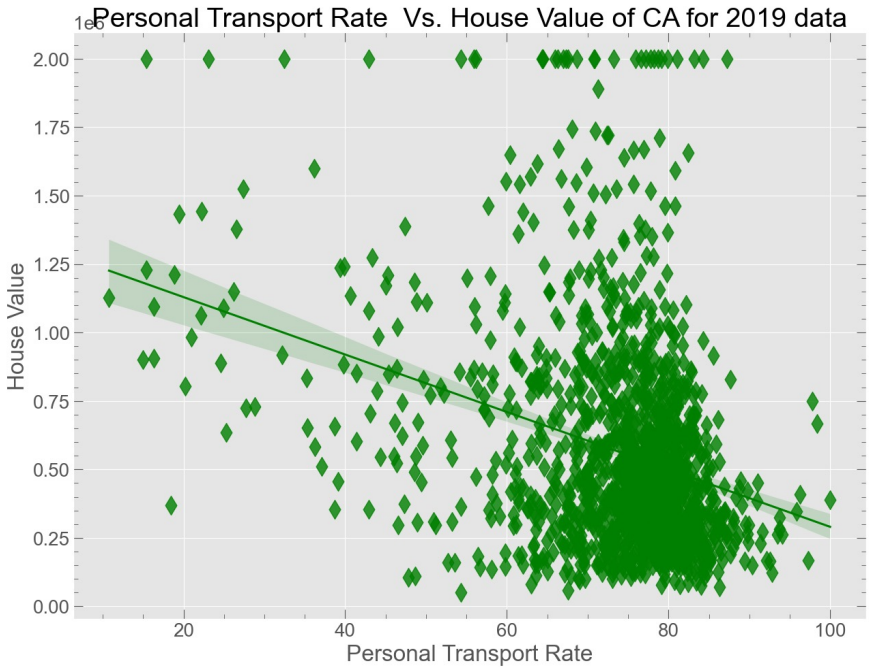

High School Rate Vs. House Value of CA for 2019 data
1e6
House Value
High School Rate


Uneducated Rate Vs. House Value of CA for 2019 data
1e6
House Value
Uneducated Rate

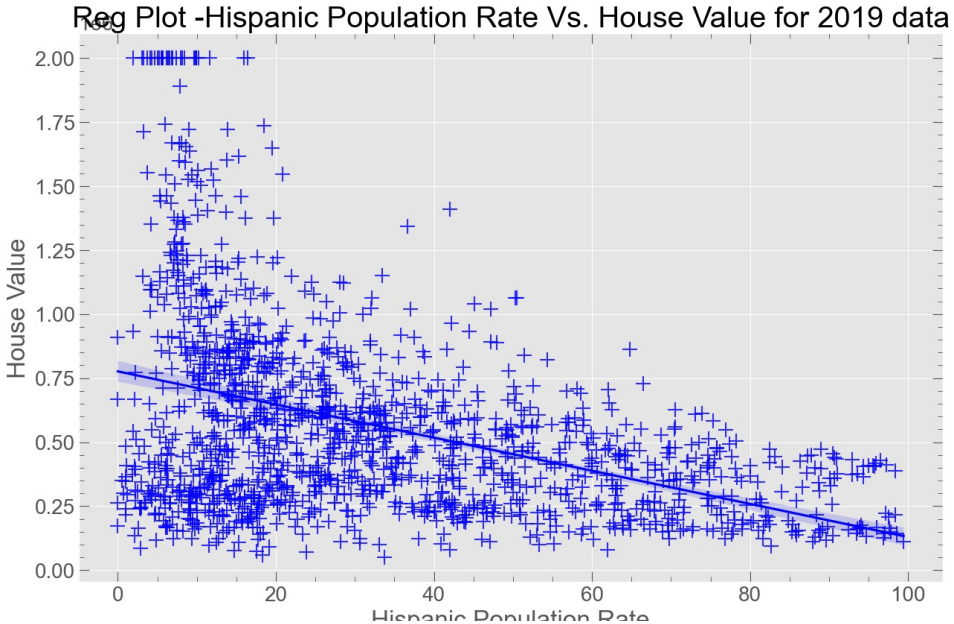
3. Transportation

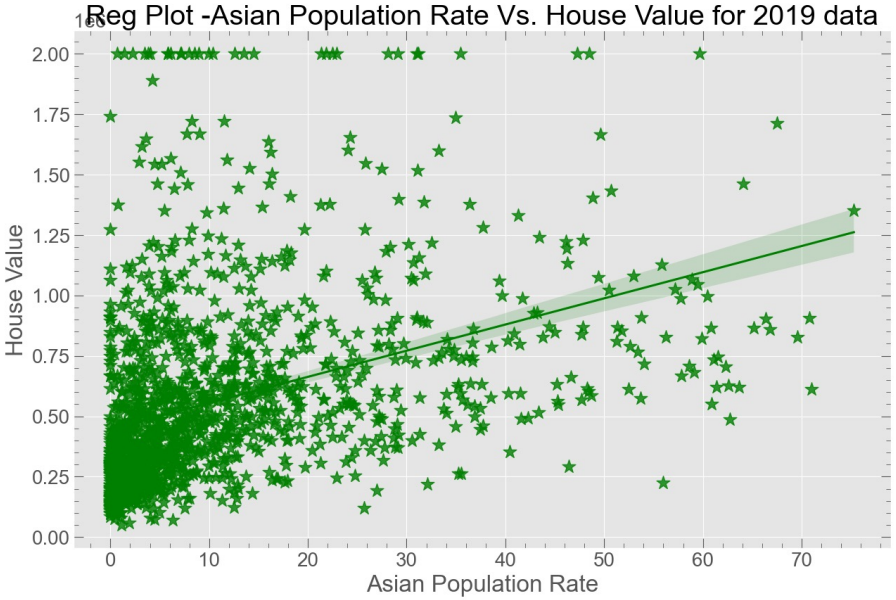


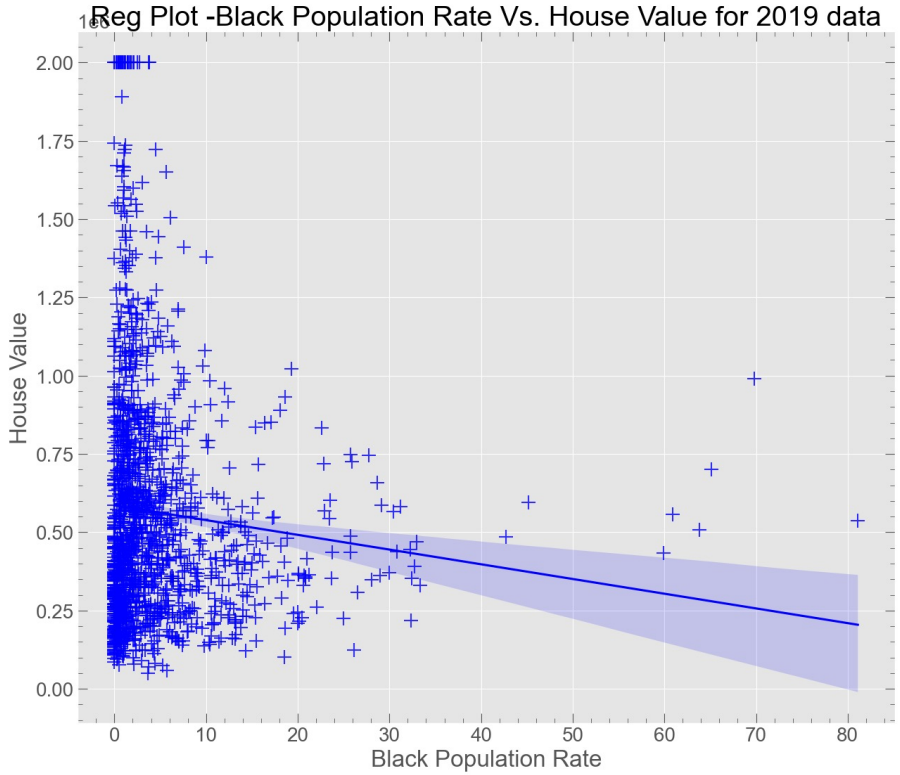


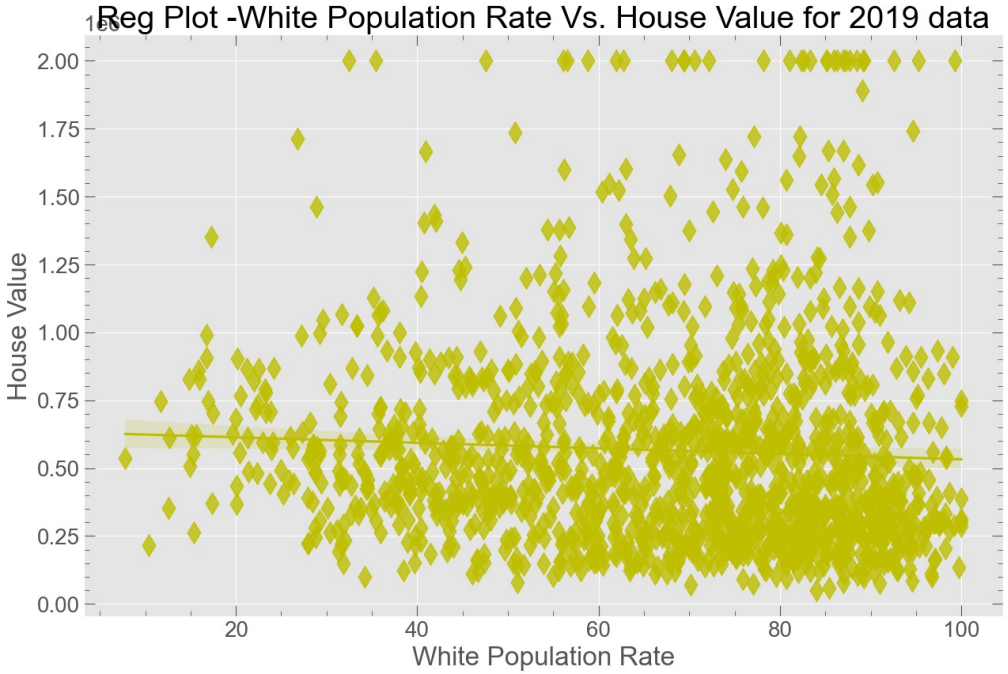


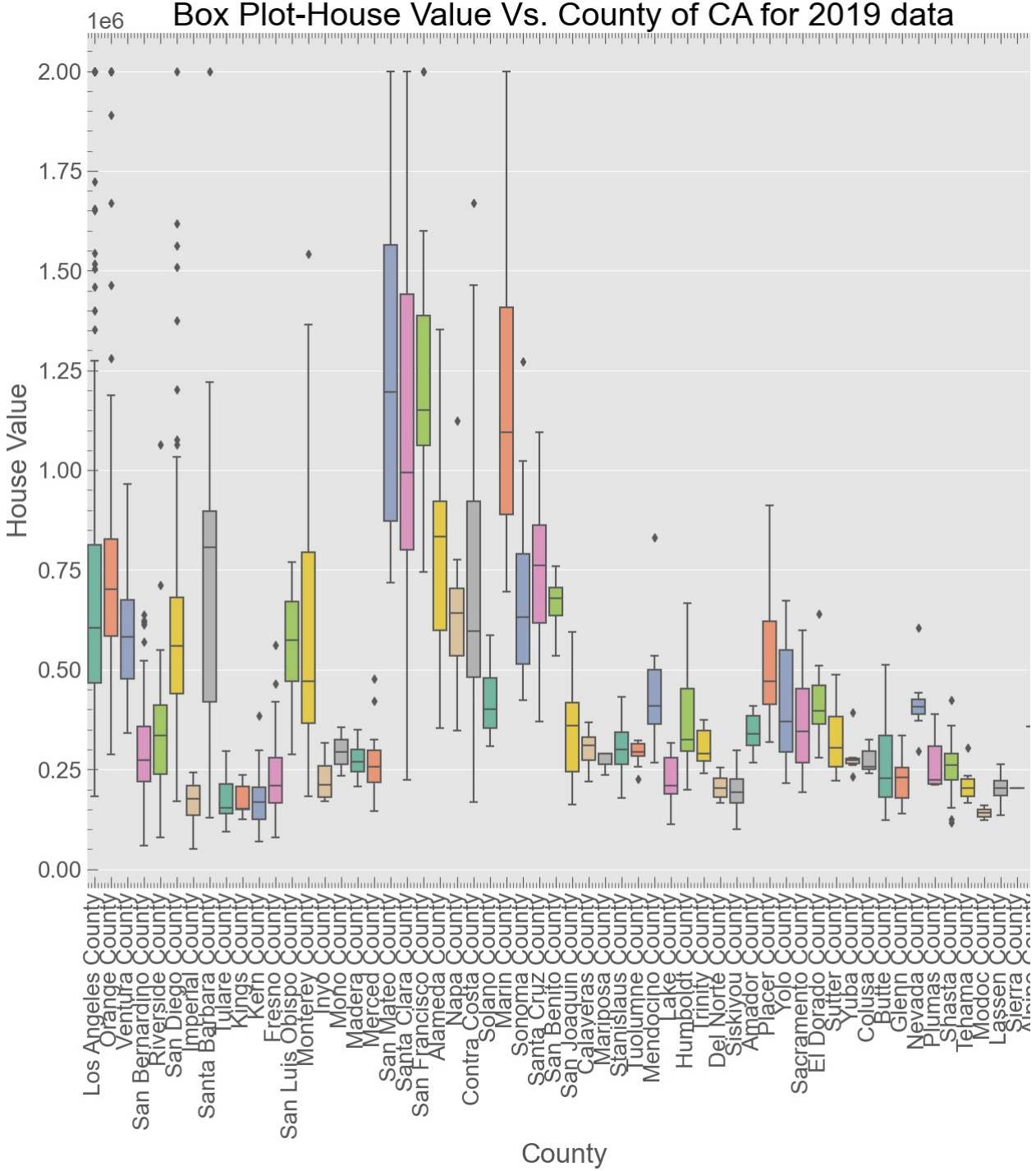
4. Race

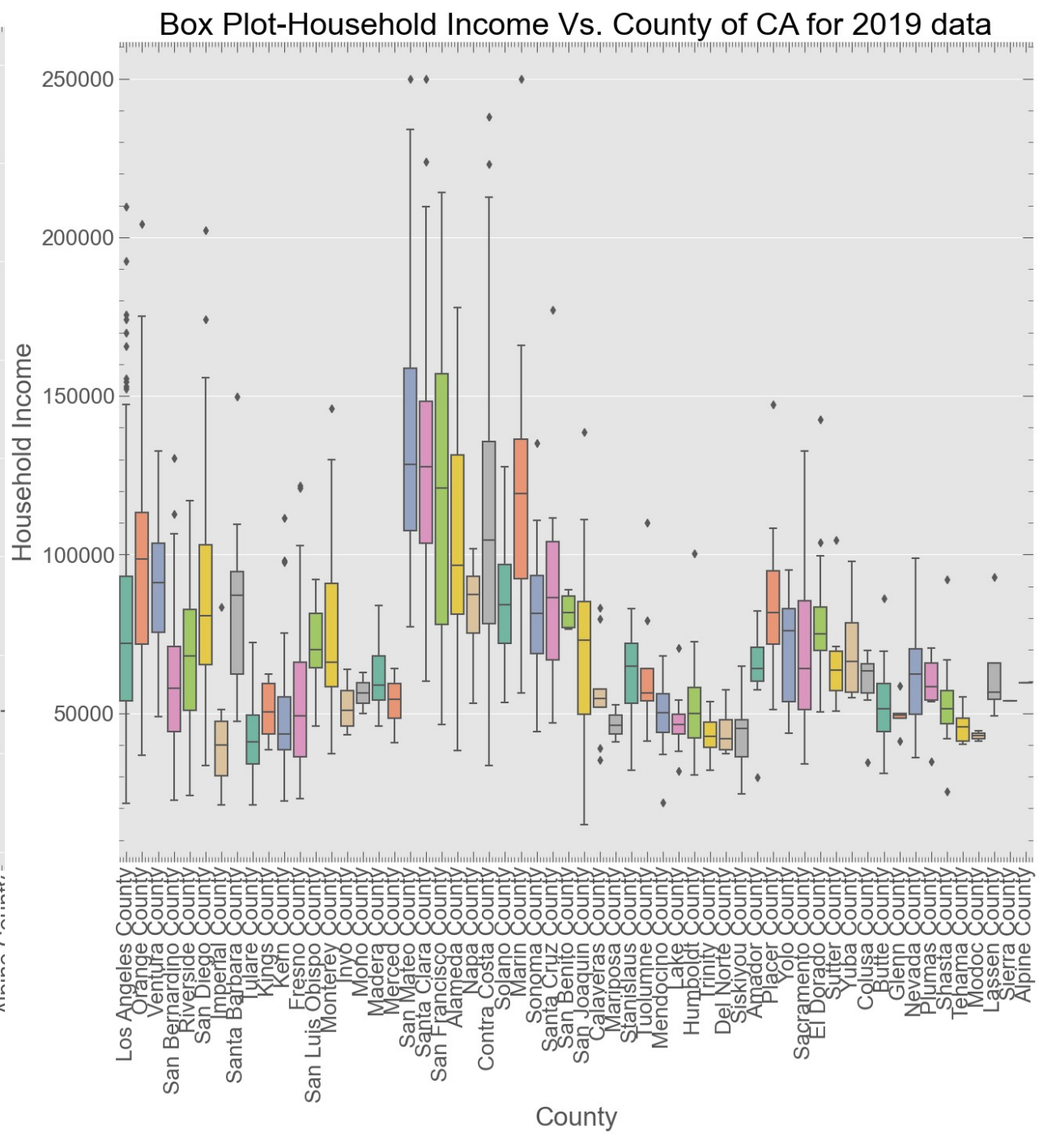

Reg Plot -Hispanic Population Rate Vs. House Value for 2019 data
House Value
Hispanic Population Rate


Reg Plot -Asian Population Rate Vs. House Value for 2019 data
House Value
Asian Population Rate

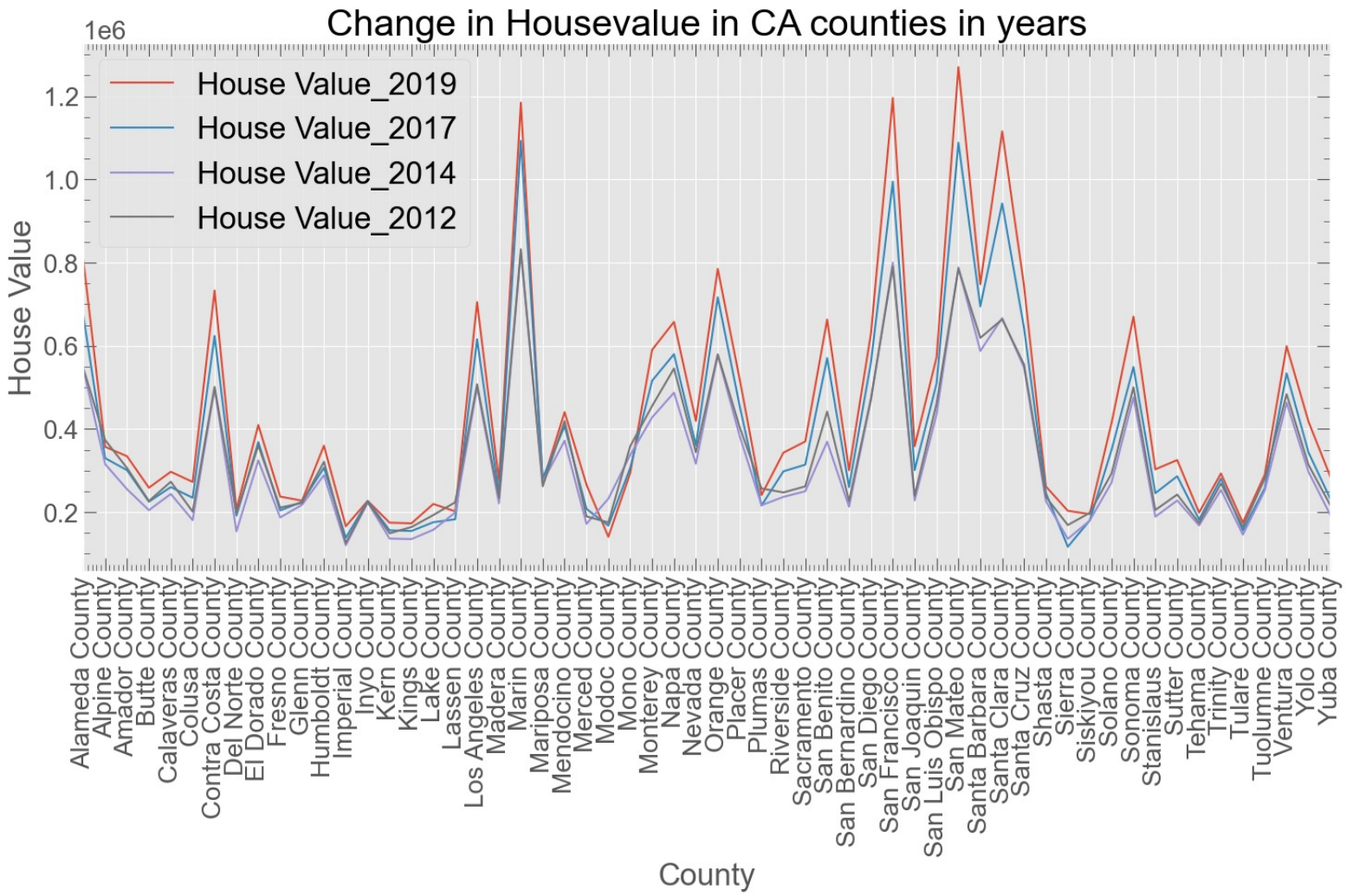

Reg Plot -Black Population Rate Vs. House Value for 2019 data
House Value
Black Population Rate

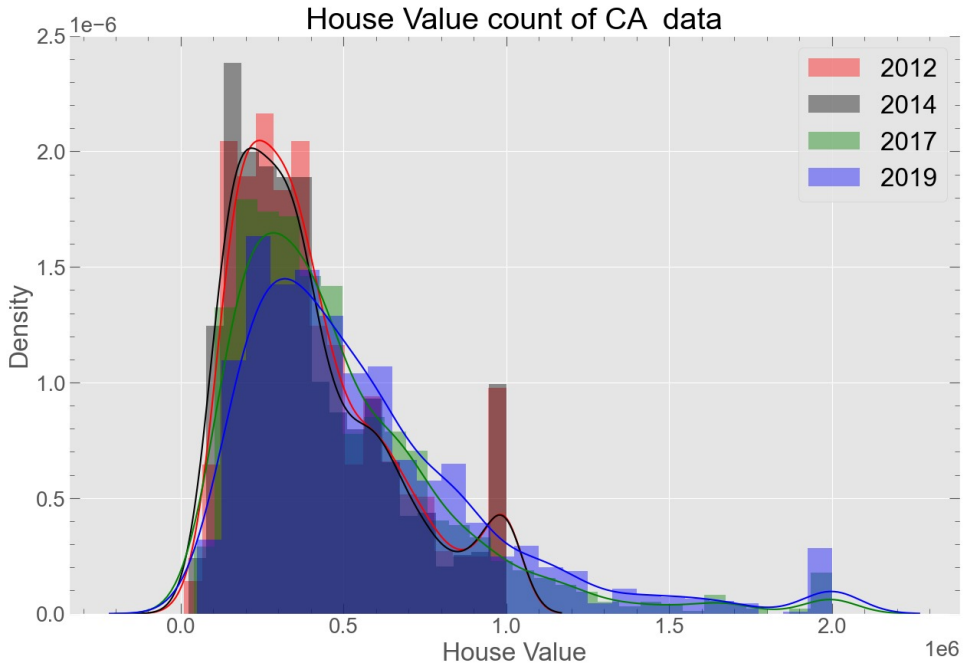

Reg Plot -White Population Rate Vs. House Value for 2019 data
House Value
White Population Rate


Box Plot-House Value Vs. County of CA for 2019 data
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
House Value
Los Angeles County
Orange County
Ventura County
San Bernardino County
Riverside County
San Diego County
Imperial County
Santa Barbara County
Tulare County
Kings County
Kern County
Fresno County
San Luis Obispo County
Monterey County
Inyo County
Mono County
Madera County
Merced County
San Mateo County
Santa Clara County
San Francisco County
Alameda County
Napa County
Contra Costa County
Solano County
Marin County
Sonoma County
Santa Cruz County
San Benito County
San Joaquin County
Calaveras County
Mariposa County
Stanislaus County
Tuolumne County
Mendocino County
Lake County
Humboldt County
Trinity County
Del Norte County
Siskiyou County
Amador County
Placer County
Yolo County
Sacramento County
El Dorado County
Sutter County
Yuba County
Colusa County
Butte County
Glenn County
Nevada County
Plumas County
Shasta County
Tehama County
Modoc County
Lassen County
Sierra County
County

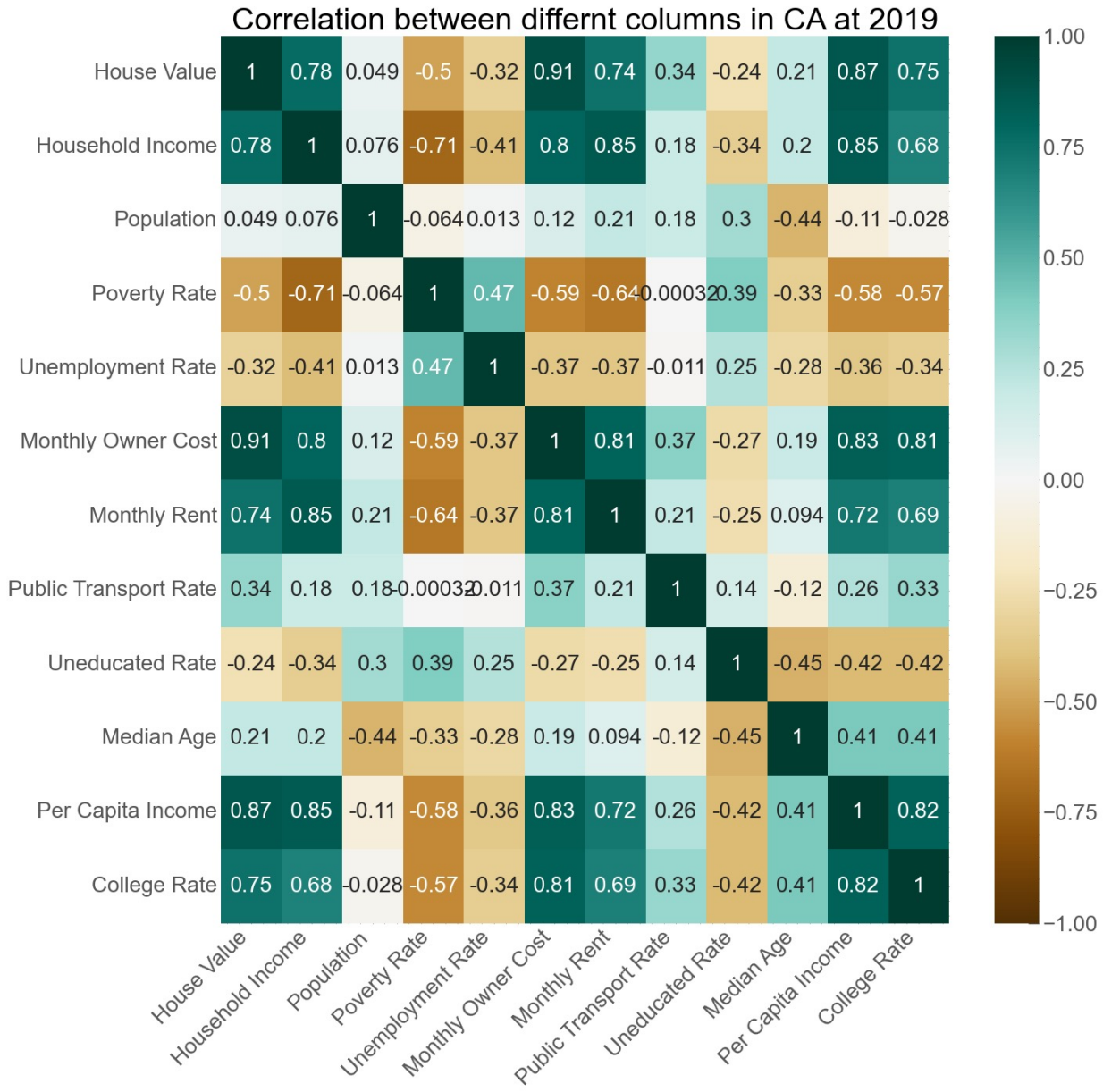

Box Plot-Household Income Vs. County of CA for 2019 data
250000
200000
150000
100000
50000
Household Income
Los Angeles County
Orange County
Ventura County
San Bernardino County
Riverside County
San Diego County
Imperial County
Santa Barbara County
Tulare County
Kings County
Kern County
Fresno County
San Luis Obispo County
Monterey County
Inyo County
Mono County
Madera County
Merced County
San Mateo County
Santa Clara County
San Francisco County
Alameda County
Napa County
Contra Costa County
Solano County
Marin County
Sonoma County
Santa Cruz County
San Benito County
San Joaquin County
Calaveras County
Mariposa County
Stanislaus County
Tuolumne County
Mendocino County
Lake County
Humboldt County
Trinity County
Del Norte County
Siskiyou County
Amador County
Placer County
Yolo County
Sacramento County
El Dorado County
Sutter County
Yuba County
Colusa County
Butte County
Glenn County
Nevada County
Plumas County
Shasta County
Tehama County
Modoc County
Lassen County
Sierra County
Alpine County
County

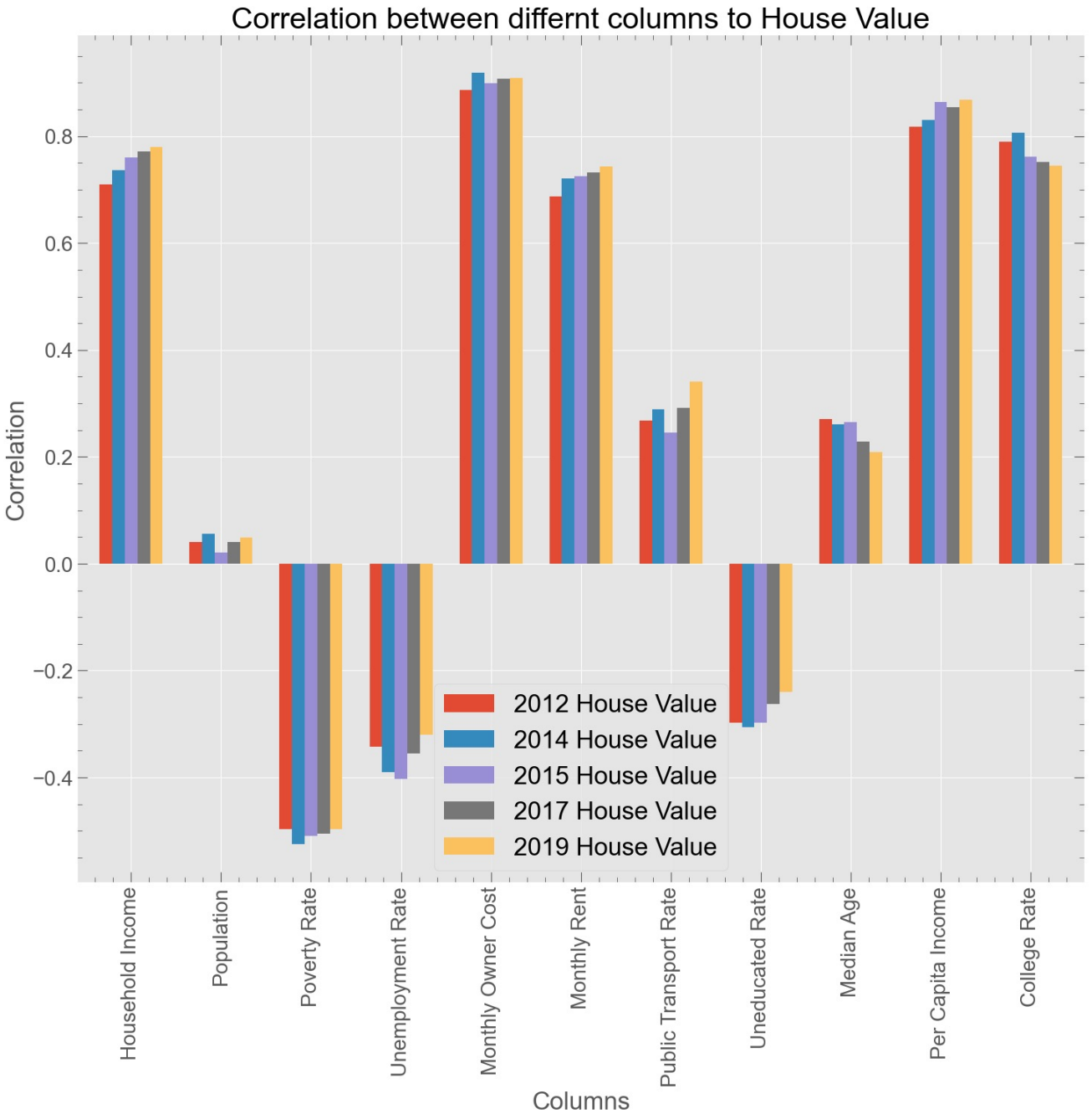
House Value for Different Years





Correlation





Best 5 cities

Columns:

'House Value':
'Household Income':
'Population':
'Poverty Rate':
'Unemployment Rate':
'Monthly Owner Cost':
'Monthly Rent':
'Public Transport Rate':
'Uneducated Rate':
'Per Capita Income':

- We gave each parameters equal weight
- Best city: which satisfy most of this parameters

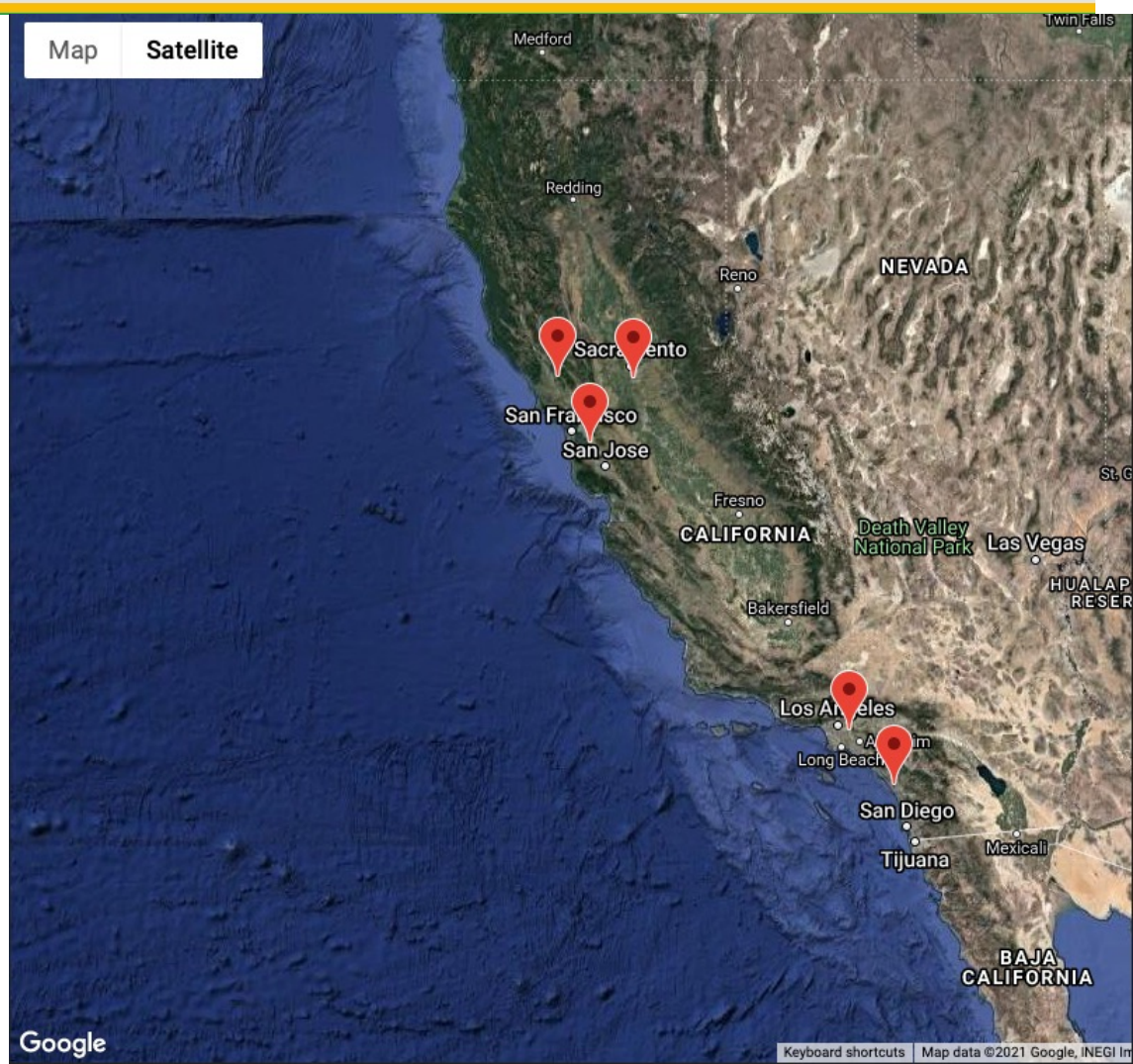
Top 5 Best Cities for Differnt Years

```
#getting best city for each year and appending to cities
cities=[]
all_data=[ca_2012,ca_2014, ca_2015, ca_2017, ca_2019]
for data in all_data:

    #selecting only important columns
    d1=data[['City', 'House Value','Household Income','Population',
       'Poverty Rate', 'Unemployment Rate','Monthly Owner Cost', 'Monthly Rent',
       'Public Transport Rate', 'Uneducated Rate',
        'Median Age','Per Capita Income',
       'College Rate' ]]
    #using groupby method to groupby the data according to city
    d1=d1.groupby('City').agg({'House Value': 'mean',
                               'Household Income': 'mean',
                               'Population': 'sum',
                                'Poverty Rate': 'mean',
                               'Unemployment Rate': 'mean',
                               'Monthly Owner Cost': 'mean',
                               'Monthly Rent': 'mean',
                                'Public Transport Rate': 'mean',
                               'Per Capita Income': 'mean'}).reset_index()
    d2=d1.nlargest(50,"Population")
    #getting nlargest and nsmallest of all columns and selecting their city as a lis
    hv=d2.nsmallest(50, 'House Value')
    hv_c=hv['City'].tolist()
    ue=d2.nsmallest(25, 'Unemployment Rate')
    ue_c=ue['City'].tolist()
    pt=d2.nlargest(25, 'Public Transport Rate')
    pt_c=pt['City'].tolist()
    p=d2.nsmallest(25, 'Poverty Rate')
    p_c=p['City'].tolist()
    mr=d2.nsmallest(25, 'Monthly Rent')
    mr_c=mr['City'].tolist()
    moc=d2.nsmallest(25, 'Monthly Owner Cost')
    moc_c=moc['City'].tolist()
    pci=d2.nlargest(25, 'Per Capita Income')
    pci_c=pci['City'].tolist()
    hi=d2.nlargest(25, 'Household Income')
    hi_c=hi['City'].tolist()

    #Ading largest or smallest city from all columns

    all_city=hv_c + ue_c + pt_c + p_c +mr_c + moc_c +hi_c+pci_c

    #Using counter to count most common
    my_counter=Counter(all_city)
    best_city=my_counter.most_common(5)
    cities.append(best_city)

print(cities)
```

gmaps

Google Maps API

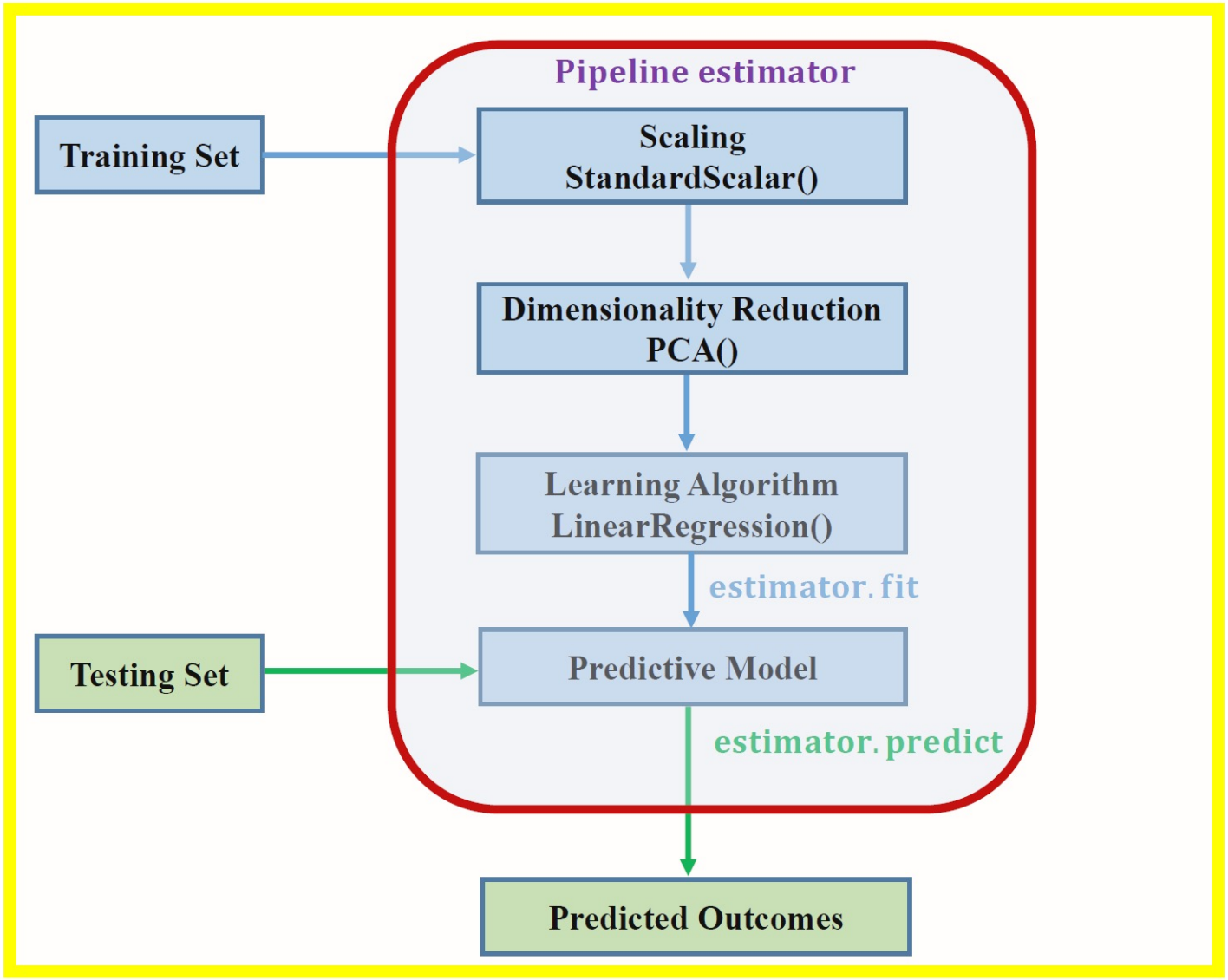
	city_2012	city_2014	city_2015	city_2017	city_2019
0	(Whittier, 9)	(Elk Grove, 8)	(Salinas, 8)	(Roseville, 8)	(Roseville, 8)
1	(Elk Grove, 8)	(Santa Rosa, 8)	(Escondido, 8)	(Salinas, 8)	(Escondido, 8)
2	(Oceanside, 8)	(Hayward, 8)	(Whittier, 8)	(Escondido, 8)	(Salinas, 8)
3	(Hayward, 8)	(Whittier, 8)	(Santa Rosa, 8)	(Whittier, 8)	(Santa Rosa, 8)
4	(Santa Rosa, 8)	(Anaheim, 8)	(Anaheim, 8)	(Santa Rosa, 8)	(Anaheim, 8)



Data Modeling

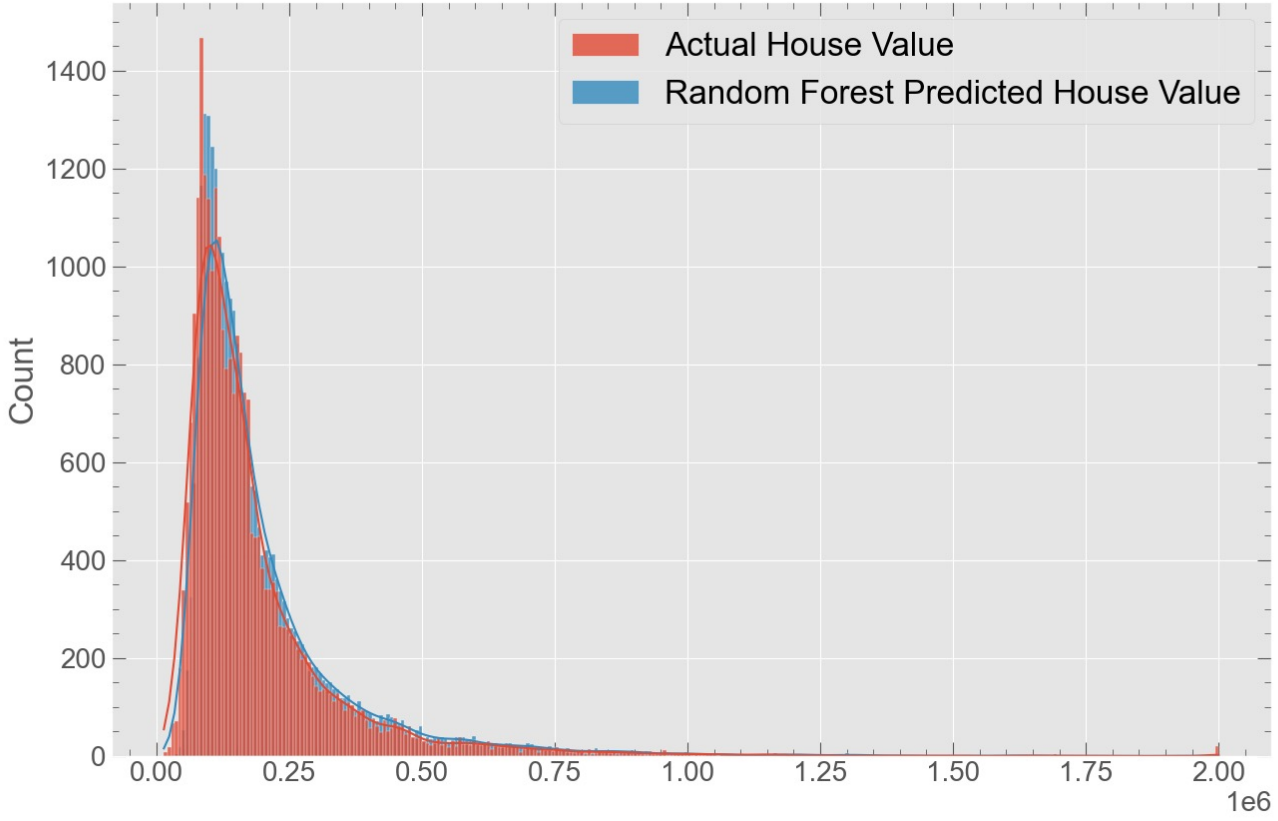
Models

- ❖ Linear Regression
- ❖ Lasso Regression
- ❖ Ridge Regression
- ❖ Support Vector Machine
- ❖ Decision Tree Regressor
- ❖ Random Forest Regressor

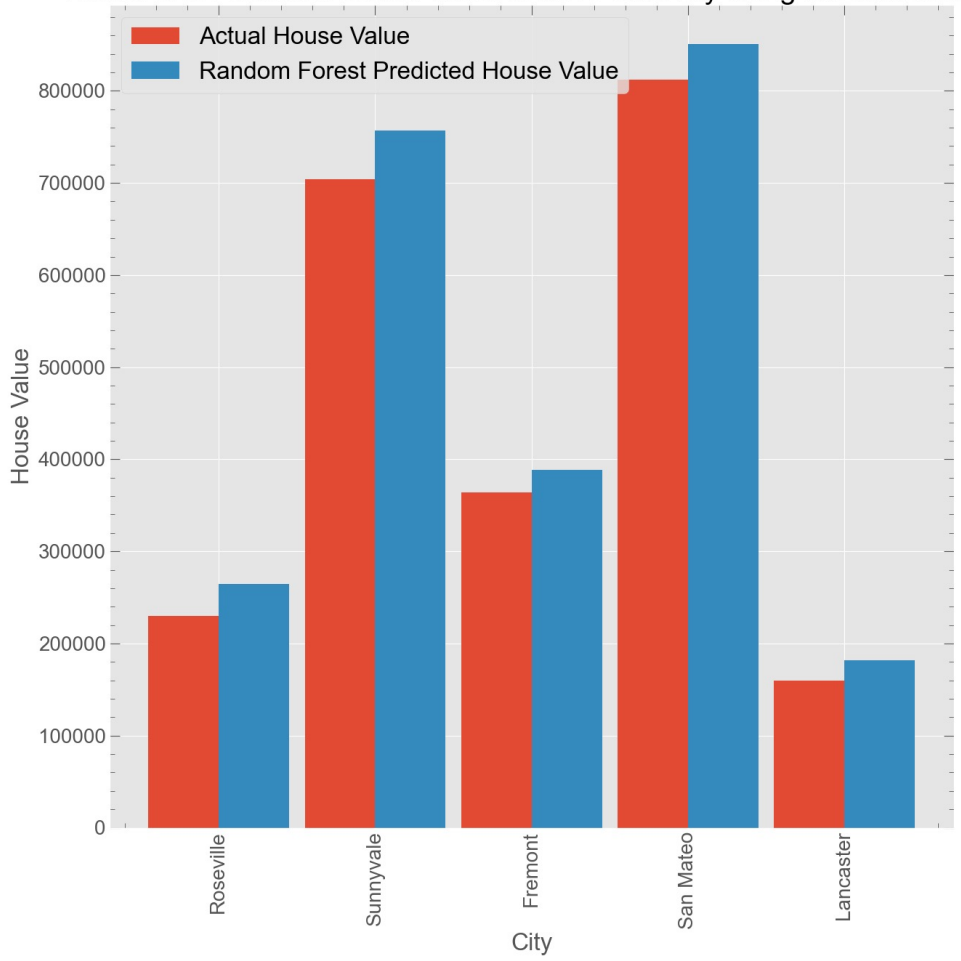


Evaluation Metrices

	Linear Reg	Lasso Reg	Ridge Reg	SVM Reg	Decision Tree	Random Forest
R2	0.834404	0.834404	0.834405	-1.041136	0.943511	0.989842
Mean Absolute Error	45391.511278	45384.616668	45378.015108	190494.756124	34290.569577	26995.017287
Root Mean Squared Error	77458.945832	77459.150249	77457.446438	271194.169585	64724.792754	51740.346617



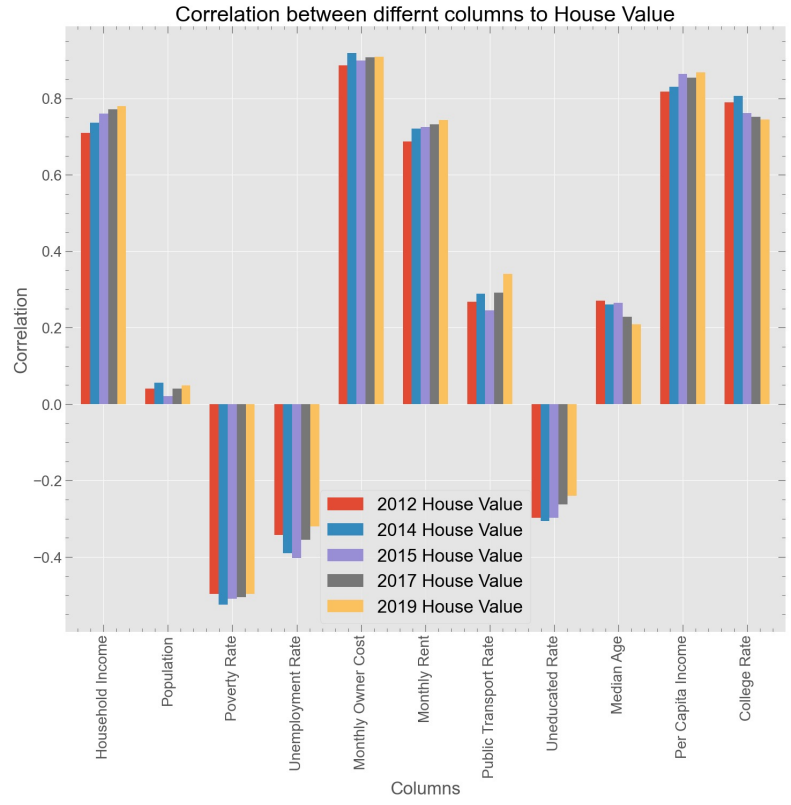
Actual and Predicted house value of 2017 best city using our best model



Our model successfully predict the housing value of unseen data from 2017 Census with an prediction error of around $27,000

Conclusion

- Used Census API to download the 5 years of US data
  - 33000 rows
  - 30 columns
- Looked the relationship of house price to other factors
- Found some interesting trend of house value on different years
- Analyzed the data and recommended the best city
- What parameters other cities need to improve to be the best city
- Build a model to predict the house value
- Our best model predict the house value perfectly with an error of around $30,000.



Important scores of each columns

```
feature_importances=rf_b.feature_importances_
col=X.columns
sorted(zip(feature_importances, col), reverse=True)
```

```
[(0.574906843853325, 'Monthly Owner Cost'),
 (0.18693950369523973, 'Per Capita Income'),
 (0.07941670762889026, 'Lng'),
 (0.05467911949320379, 'Household Income'),
 (0.023908771370979562, 'College Rate'),
 (0.018488581682145788, 'Lat'),
 (0.015871550850818122, 'Personal Transport Rate'),
 (0.014429561450837883, 'High School Rate'),
 (0.010878545524422072, 'Public Transport Rate'),
 (0.010332031244502536, 'Median Age'),
 (0.01014878320563534, 'Population')]
```


THANK YOU