Comparing statistic learning methods with spatial analysis methods:
In the case of crimes in NYC

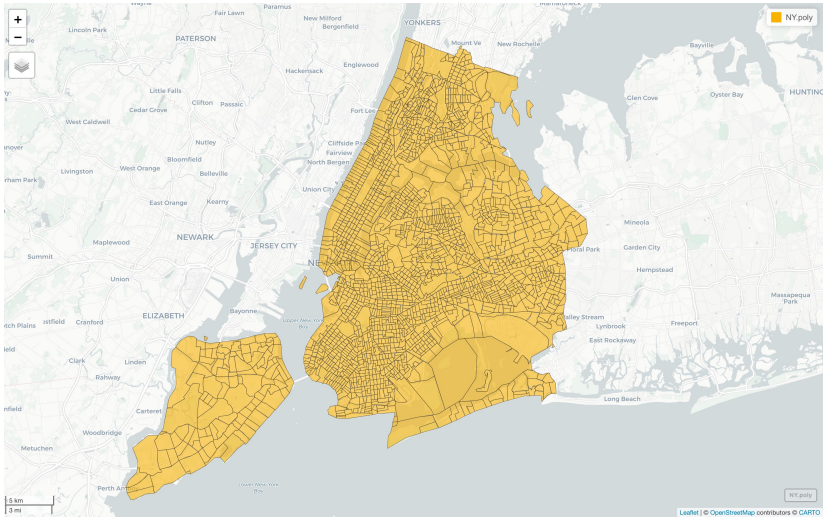
Yun Yueh (Sunny)
Department of Geography
MS Geographic information Science

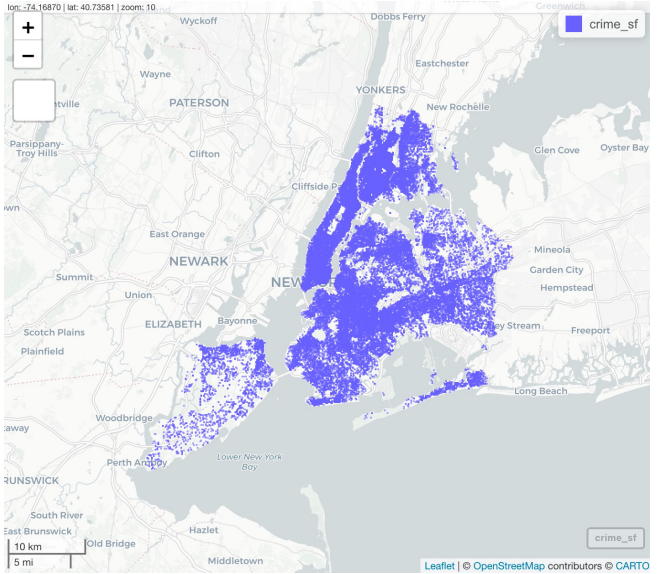
Crimes In NYC

Study area: New York County, Kings County, Queens County, Bronx County Richmond County

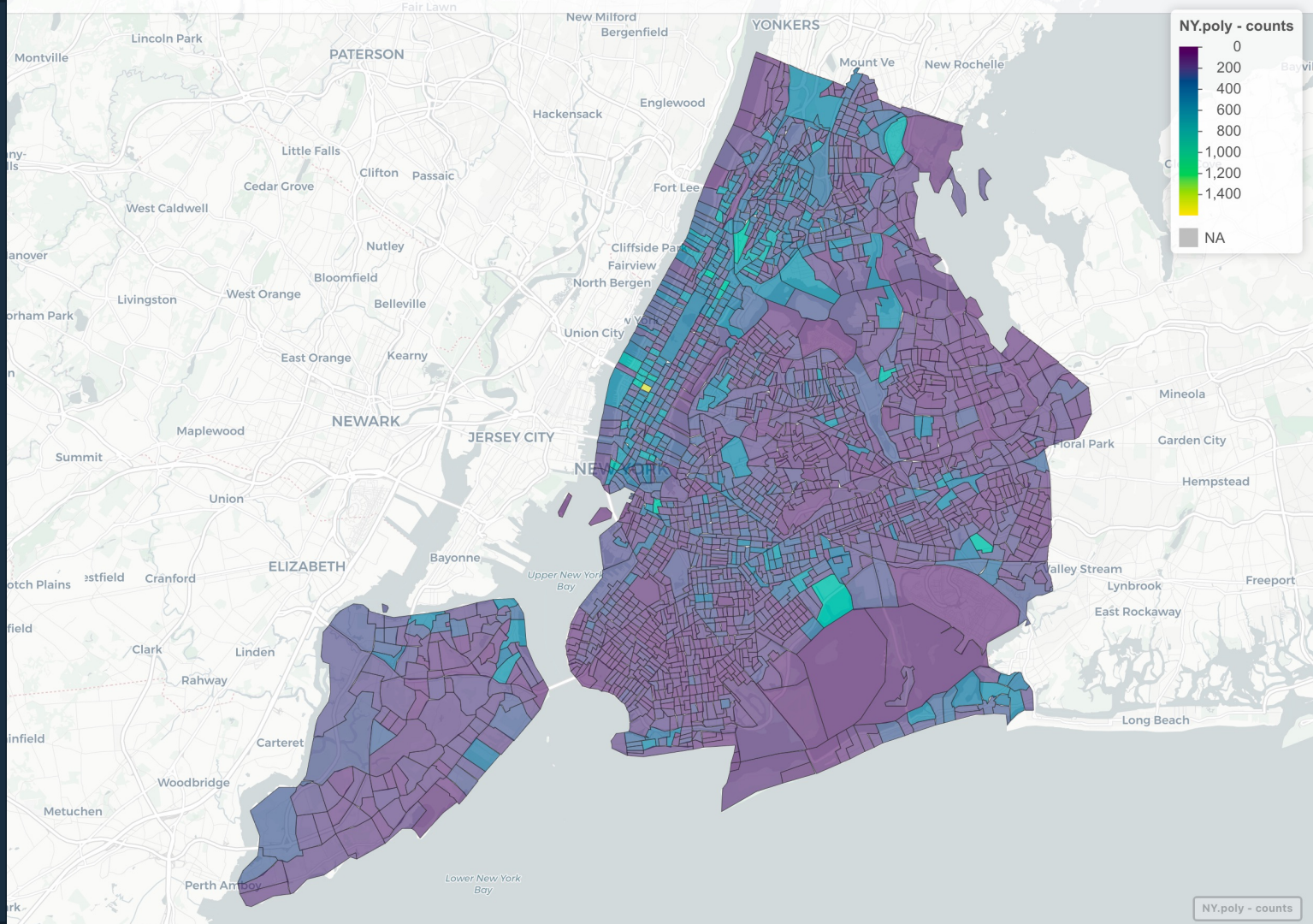
Geographic unit: census tract (2167)

Crime data: Point data of crime in 2020-2021





Crimes In NYC



Crimes
In
NYC
Spatial
analysis

Local Analysis of G-statistic (G*)

- Identifying spatial concentration with low and high value
- Include neighborhood effect and the distance between each other

$$G_i(d) = \frac{\sum_j W_{ij}(d)x_j}{\sum_j x_j}; j \neq i$$

W_{ij}: neighborhood definition
d: distance between target region and neighbor

- Standardize Gi* values> alpha=0.05, indicates cluster regions

Crimes
In
NYC
Spatial
analysis

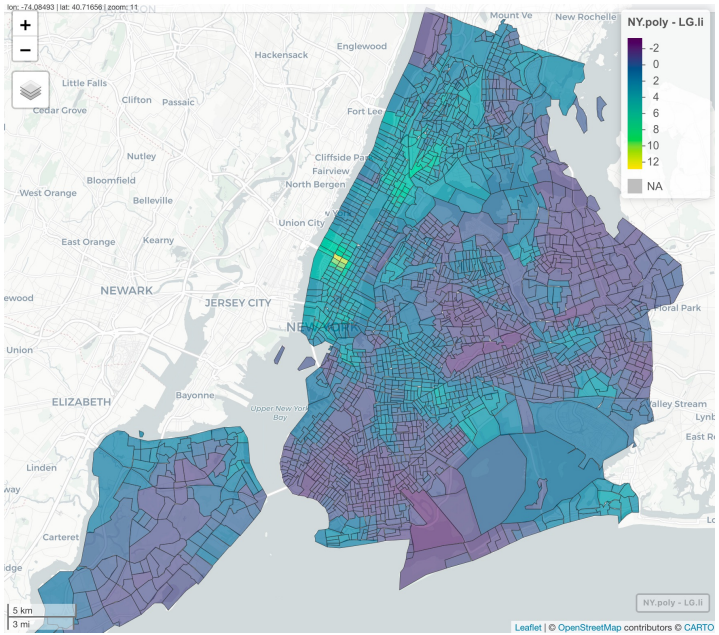
Local Analysis of G-statistic (G*)

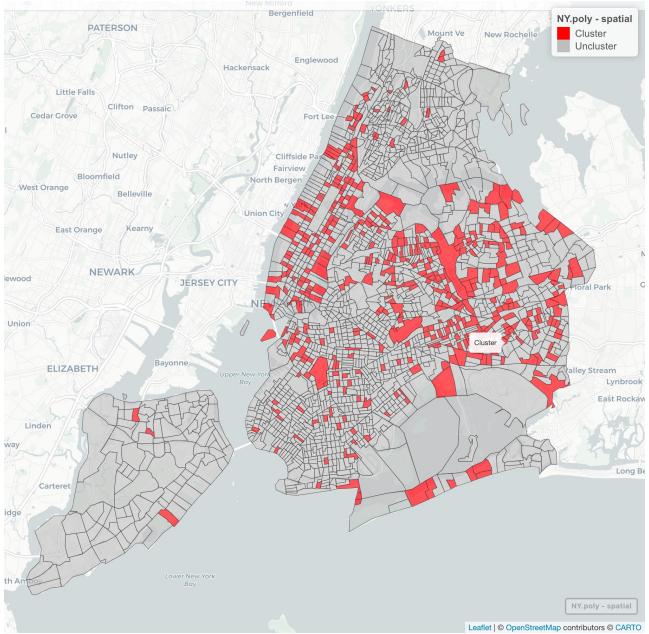
- Standardize Gi* values> alpha=0.05, indicates cluster regions

$$G_i(d) = \frac{\sum_j W_{ij}(d)x_j}{\sum_j x_j}; j \neq i$$

W_{ij}: neighborhood definition

d: distance between target region and neighbor





Crimes In NYC **Spatial analysis**

Local Analysis of G-statistic (G*)

- Continuous Gi* in each polygon

```
> NY.poly$LG.li
  [1]  2.250929457  1.291977368 -0.043152971  0.237658381  1.815188266  3.055983228  2.356253154
  [8]  1.189292075  2.762793467  3.063641719  3.607394611  1.101162256  2.604132233  7.316590661
 [15]  0.560927478  0.872343559  3.155795583  3.451671952  1.393107312 -0.171868271  2.553075624
 [22]  2.286076411  3.011086788  4.770938904  1.253955930 -1.466297622  5.861049822  3.878732983
 [29] -0.776017175  0.533496853  3.008476098  3.936030611  6.258269089  0.296373482  0.186601772
 [36]  2.729036079  2.673058656  0.424015006  6.476776067  1.154124523  2.993744149  1.231803104
 [43]  0.986567000  2.625445436  7.158340070  3.318924767  3.329632576  2.404038496  1.821081045
 [50]  1.824027435  1.555905972  2.651962943  2.364607304  1.834664344  1.701423120  1.772363040
 [57]  3.493271073  1.687780828  4.880077243  0.135746132  7.324776037 -0.653236544  4.222626756
 [64]  4.403877720  1.627754741  2.035707902  4.895686631  2.118877267  1.234749493 -0.424046032
 [71] -0.209027375  0.769219919  0.813415765  2.871182212 10.085976013  1.682323911  3.978471231
 [78]  1.164232902  3.456067709  4.870594394  2.053595516  1.498860043  3.176420311  1.237585179
 [85]  3.301054954  3.155543616  3.461278836  2.127062642 -1.114686922  2.320783194  3.099381346
 [92]  4.857168399  3.199346823  0.348107717  4.473536930  1.867859087 -0.020202733  0.743195863
 [99] -1.077630960  2.705495839  2.622018511 -0.198349556  3.270698717  1.225603776  6.614562736
[106]  1.220017545  2.940709135  3.038113415  1.323141185  4.222626756  0.353649993  1.999234272
```

- Classification of cluster in each polygon

```
> NY.poly$spatial
  [1] "Uncluster" "Uncluster" "Uncluster" "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Cluster"
  [9] "Cluster"   "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Uncluster" "Cluster"
 [17] "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Cluster"   "Uncluster" "Cluster"
 [25] "Uncluster" "Uncluster" "Uncluster" "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"
 [33] "Cluster"   "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Cluster"   "Cluster"
 [41] "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Uncluster" "Cluster"
 [49] "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Uncluster"
 [57] "Cluster"   "Cluster"   "Uncluster" "Uncluster" "Cluster"   "Uncluster" "Uncluster" "Uncluster"
 [65] "Cluster"   "Uncluster" "Uncluster" "Cluster"   "Uncluster" "Uncluster" "Cluster"   "Uncluster"
 [73] "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Uncluster"
 [81] "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Cluster"
 [89] "Uncluster" "Uncluster" "Cluster"   "Cluster"   "Uncluster" "Cluster"   "Uncluster" "Uncluster"
 [97] "Uncluster" "Cluster"   "Uncluster" "Uncluster" "Uncluster" "Uncluster" "Uncluster" "Uncluster"
[105] "Cluster"   "Uncluster" "Cluster"   "Cluster"   "Cluster"   "Cluster"   "Uncluster" "Uncluster"
```

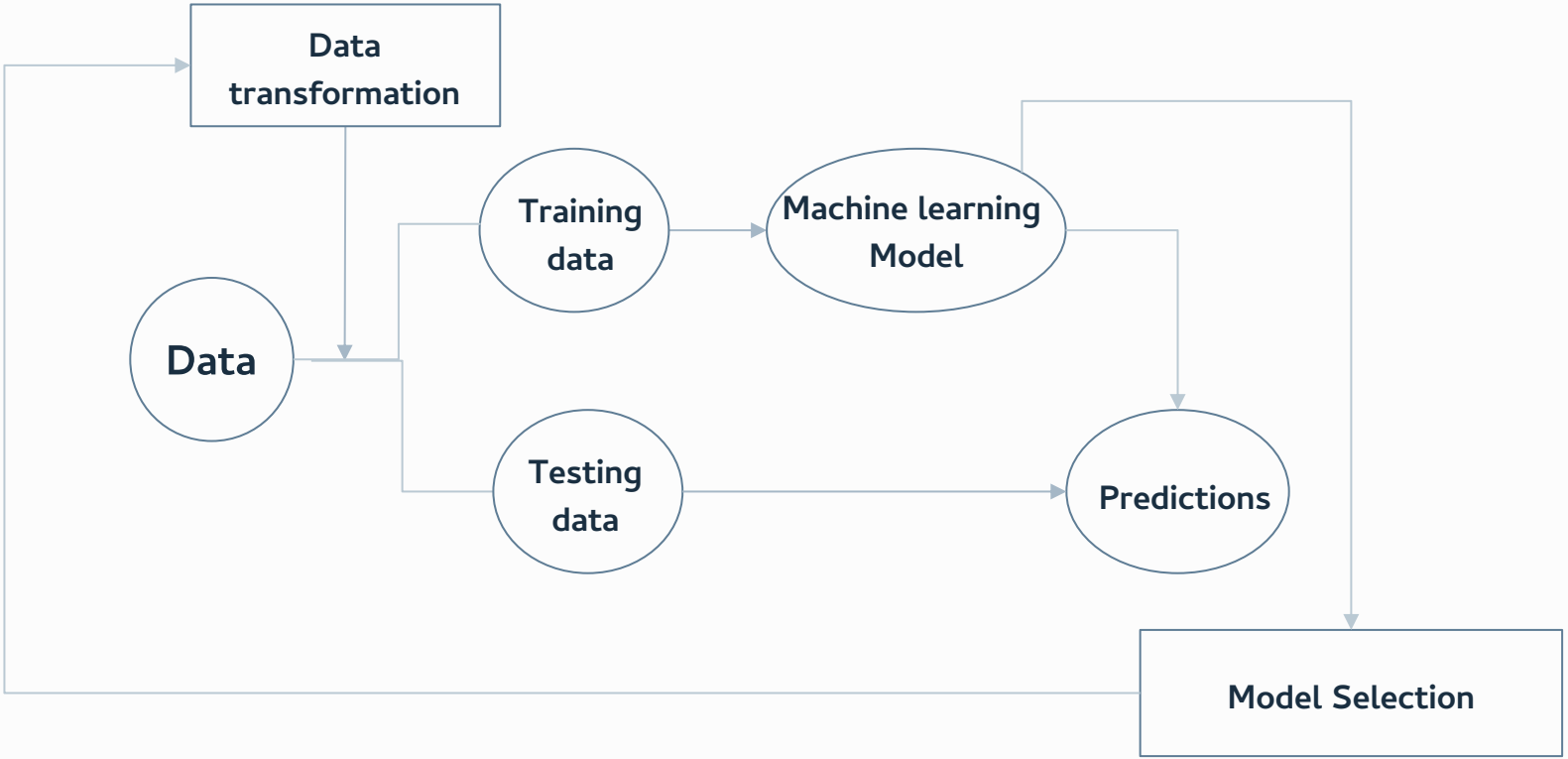
Census Data

Social-economic Variables

- Education level in each age group
- Housing variables
- Gini Index (income and wealth equality)
- Median income
- Age
- Ethnicities proportion
- Unemployed rate

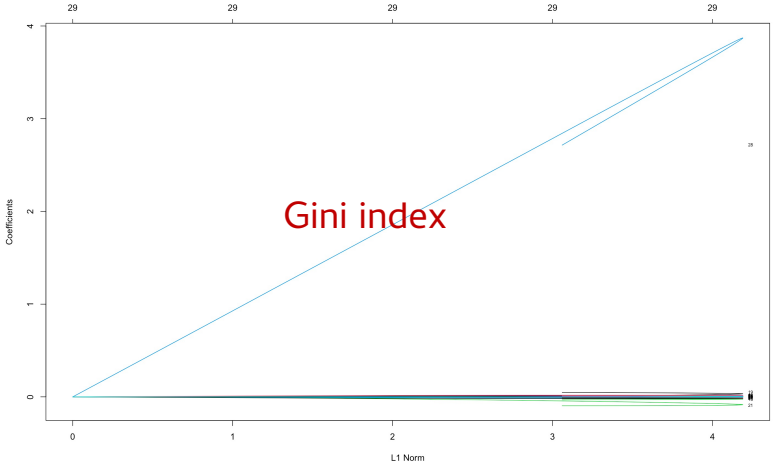
```
> colnames(NY)
 [1] "X..Population.25.Years.and.Over..Less.than.High.School"
 [2] "X..Population.25.Years.and.Over..High.School.Graduate..Includes.Equivalency."
 [3] "X..Population.25.Years.and.Over..Some.College"
 [4] "X..Population.25.Years.and.Over..Bachelor.s.Degree"
 [5] "X..Population.25.Years.and.Over..Master.s.Degree"
 [6] "X..Population.25.Years.and.Over..Professional.School.Degree"
 [7] "X..Population.25.Years.and.Over..Doctorate.Degree"
 [8] "Vacant.Housing.Units."
 [9] "X..Vacant.Housing.Units..for.Rent"
[10] "X..Vacant.Housing.Units..for.Sale.Only"
[11] "X..Vacant.Housing.Units..Other.Vacant"
[12] "Average.Gross.Rent.for.Renter.Occupied.Housing.Units"
[13] "Population.for.Whom.Poverty.Status.Is.Determined"
[14] "Total.Population"
[15] "Median.Age."
[16] "X..Total.Population..Not.Hispanic.or.Latino"
[17] "X..Total.Population..Not.Hispanic.or.Latino..White.Alone"
[18] "X..Total.Population..Not.Hispanic.or.Latino..Black.or.African.American.Alone"
[19] "X..Total.Population..Not.Hispanic.or.Latino..American.Indian.and.Alaska.Native.Alone"
[20] "X..Total.Population..Not.Hispanic.or.Latino..Asian.Alone"
[21] "X..Total.Population..Not.Hispanic.or.Latino..Native.Hawaiian.and.Other.Pacific.Islander.Alone"
[22] "X..Total.Population..Not.Hispanic.or.Latino..Some.Other.Race.Alone"
[23] "X..Total.Population..Not.Hispanic.or.Latino..Two.or.More.Races"
[24] "Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed"
[25] "X..Civilian.Population.in.Labor.Force.16.Years.and.Over..Employed"
[26] "X..Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed"
[27] "Per.Capita.Income..In.2019.Inflation.Adjusted.Dollars."
[28] "Gini.Index"
[29] "LG.li"
[30] "meidan_income"
[31] "GEOID"
[32] "counts"
[33] "LG.li.1"
[34] "spatial"
```

Work Flow


Data transformation
Training data
Machine learning Model
Data
Testing data
Predictions
Model Selection

Linear Model Selection

Lasso



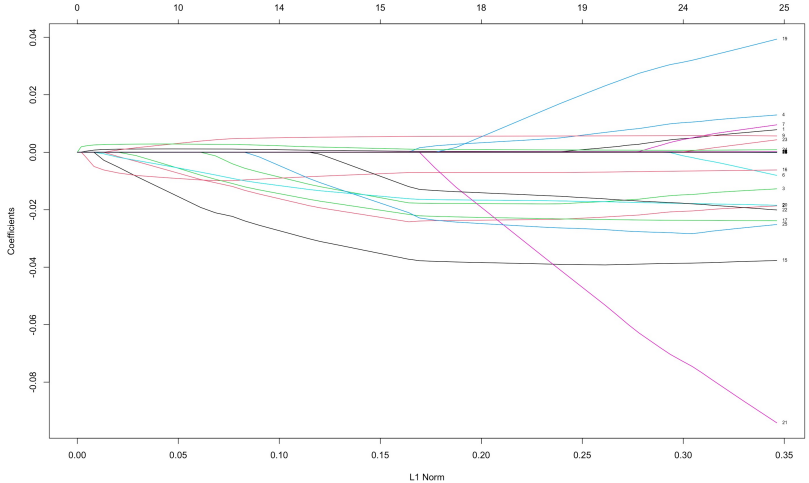
Full dataset:

Test error: 2.109729

Lasso	1 3 4 5 9 10 15 16 17 18 21 23 25 26 27 28 29 30

Without Gini index:

Test error: 2.16249



Linear Model Selection

Subset Selection

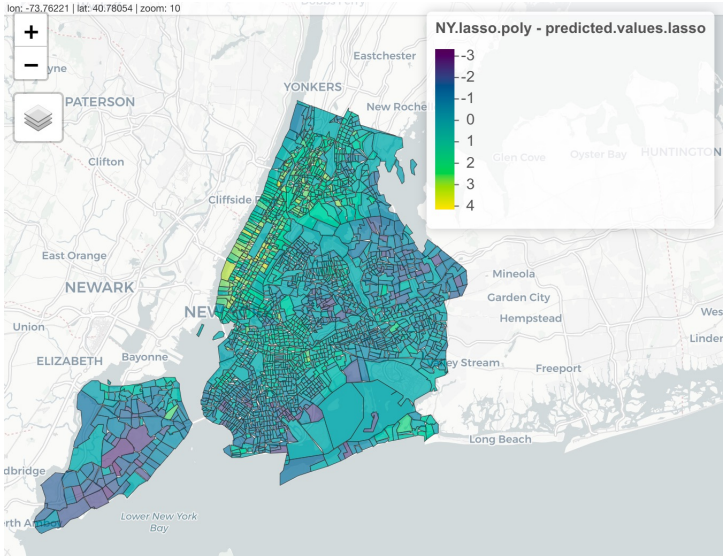
Subset Selections

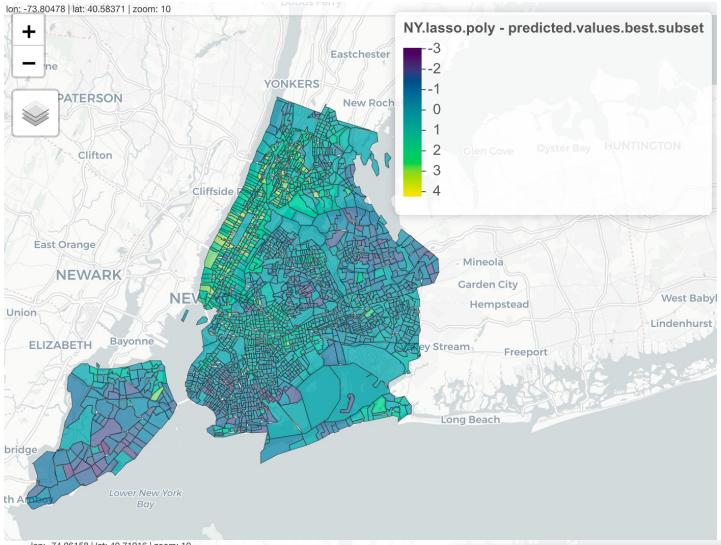
- Examine the 15 best variables in **29** variables
- Best subset, Forward and Sequential have the same results in variable selection
- Backward selection has different result than others

Best Subset	2 3 4 8 9 13 14 15 16 17 18 24 27 28 29
Forward	2 3 4 8 9 13 14 15 16 17 18 24 27 28 29
Backward	2 3 8 10 11 14 15 16 18 20 22 24 27 28 29
Sequential	2 3 4 8 9 13 14 15 16 17 18 24 27 28 29

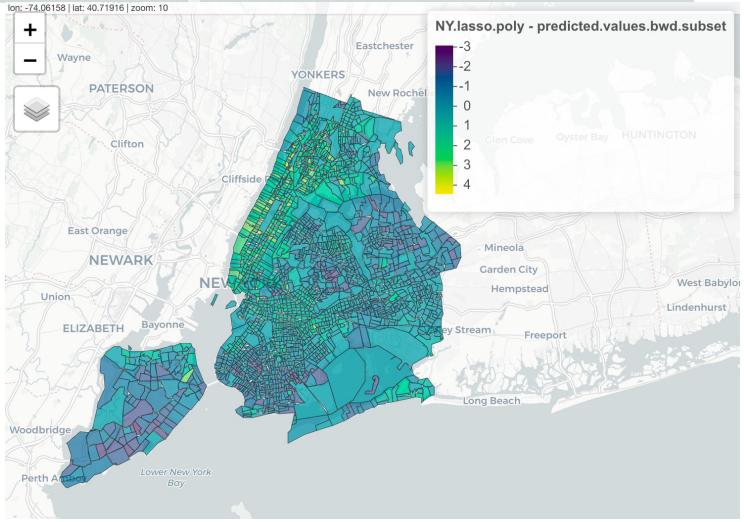
Linear Model Selection

Variable Selection In Lasso



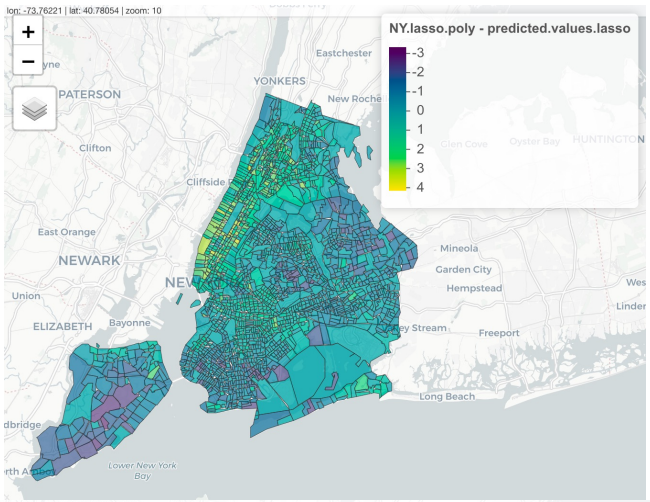


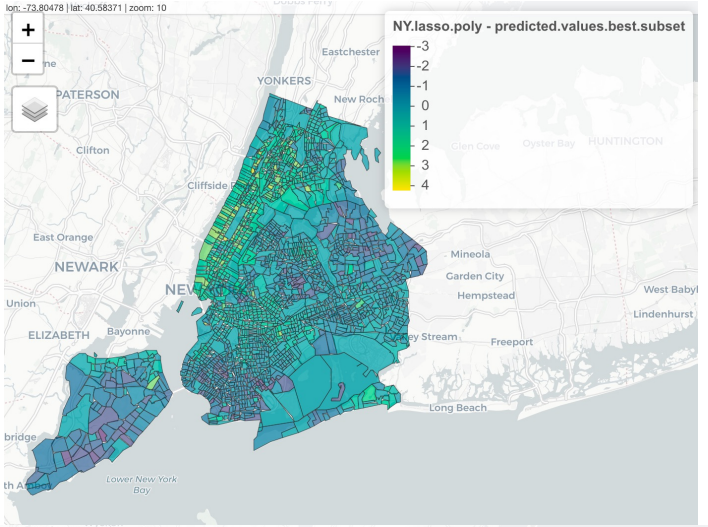
Subset selection	Test Error
Lasso alone	2.109729
Best Subset -in Lasso	2.127144
Backward- In Lasso	3.233424

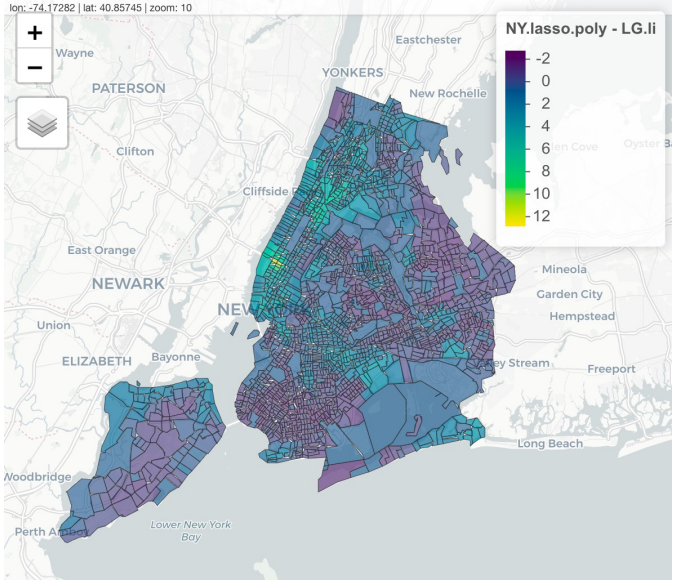


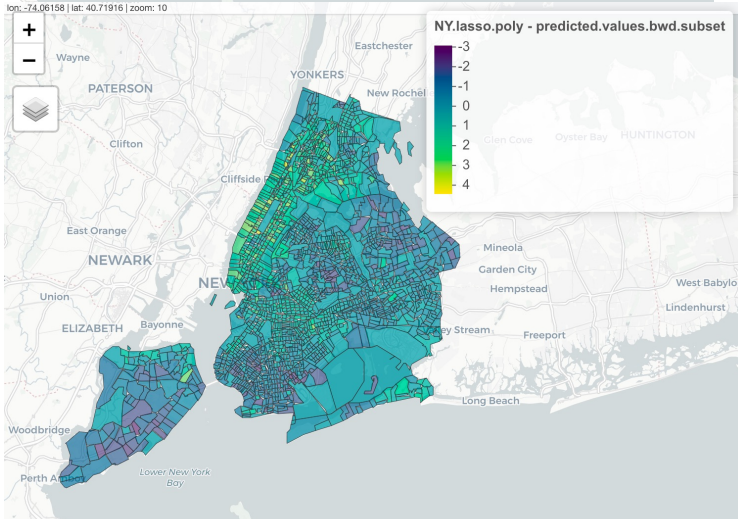
Linear Model Selection

Variable Selection In Lasso









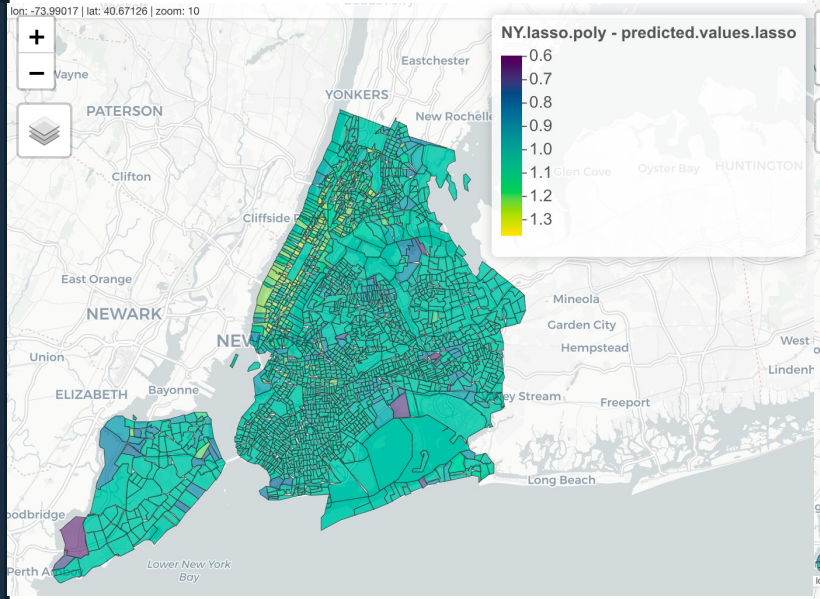
$$new(G_i^*) = G_i^{*1/2}$$

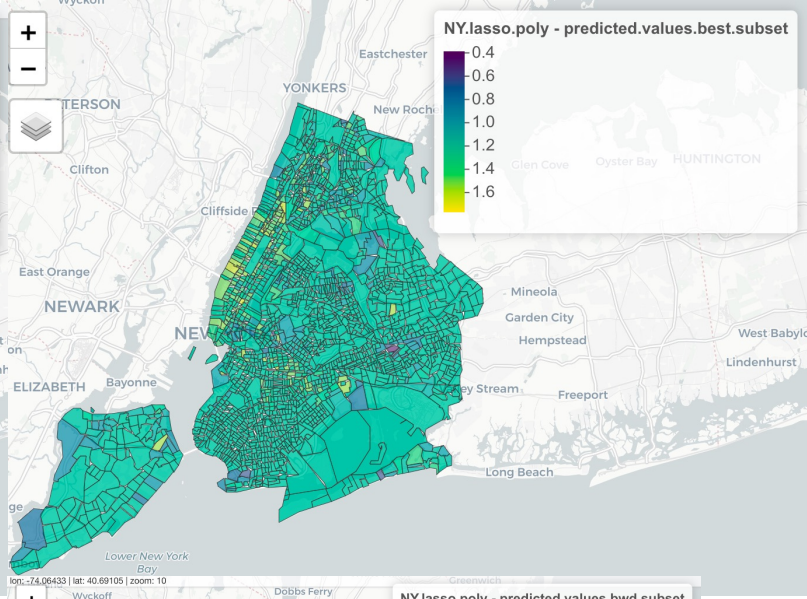
Data Transformation

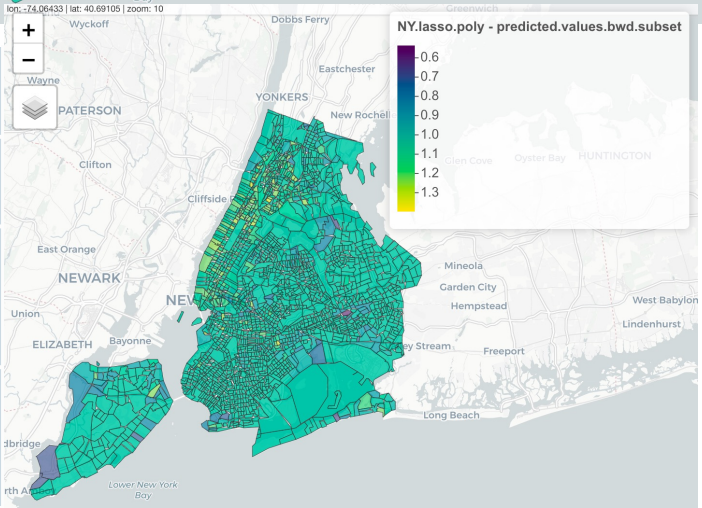
Square-root

Linear Model Selection Variable Transformation

Variable Selection In Lasso



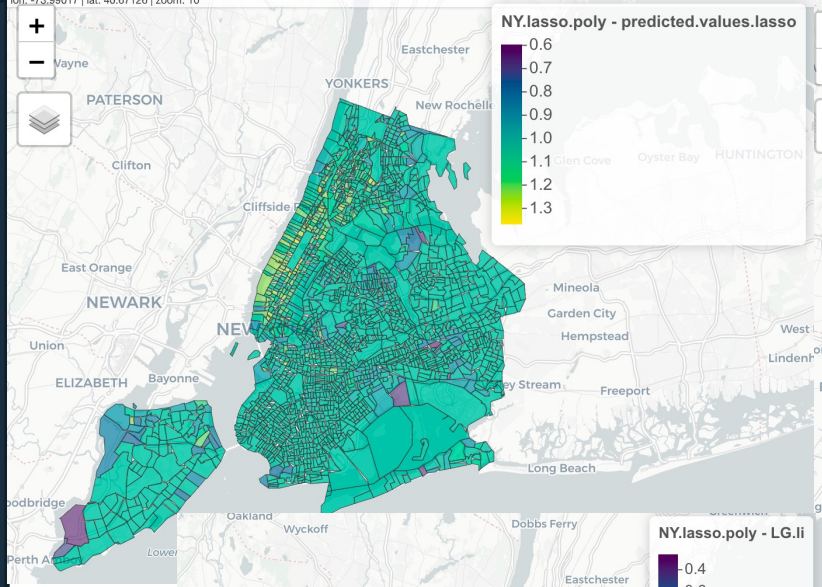


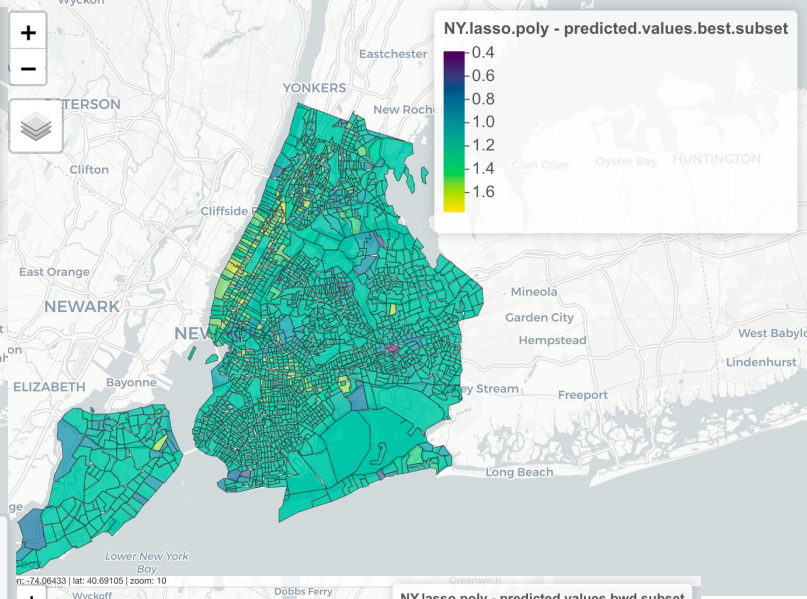


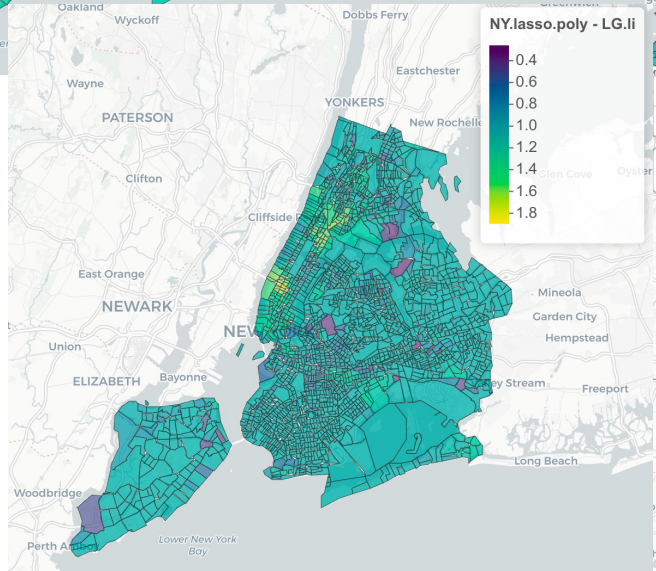
Subset selection	Test Error
Lasso alone	2.109729 → 0.3580408
Best Subset -in Lasso	2.127144 → 0.3719996
Backward- In Lasso	3.233424 → 0.3358176

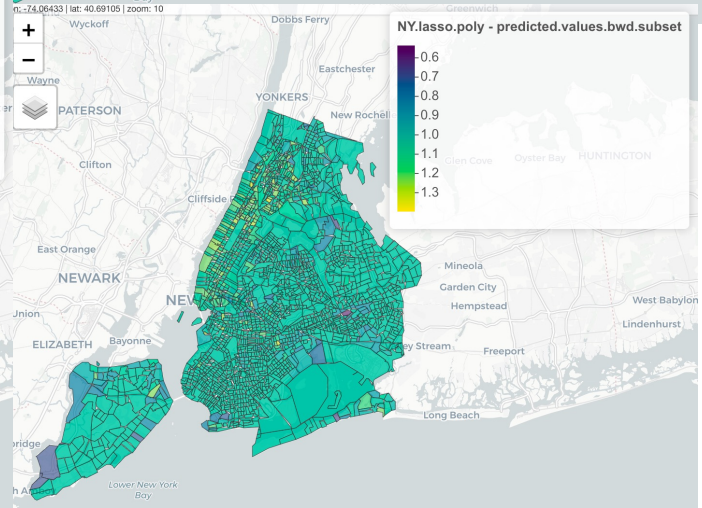
Linear Model Selection Variable Transformation

Variable Selection In Lasso



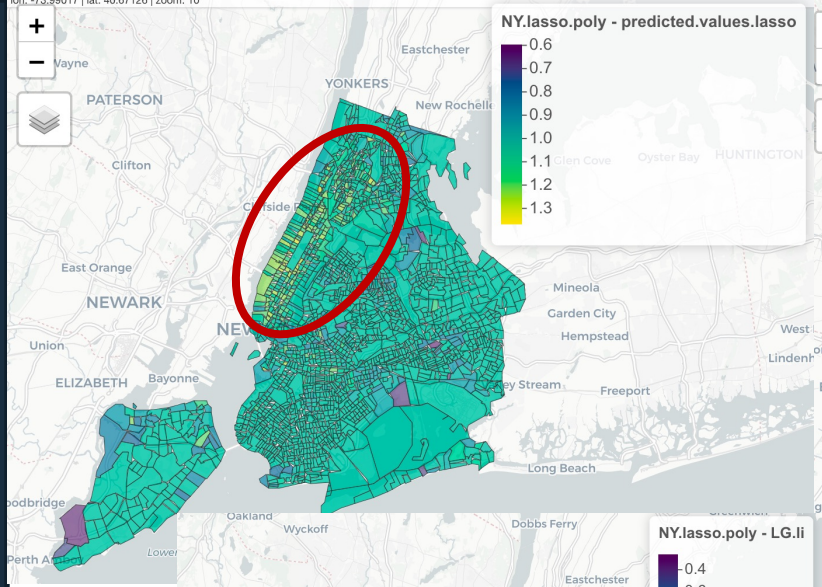


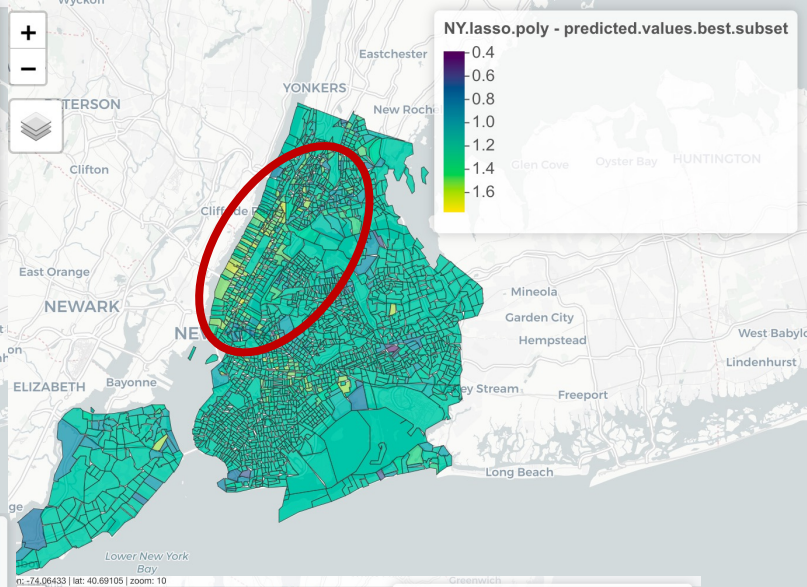


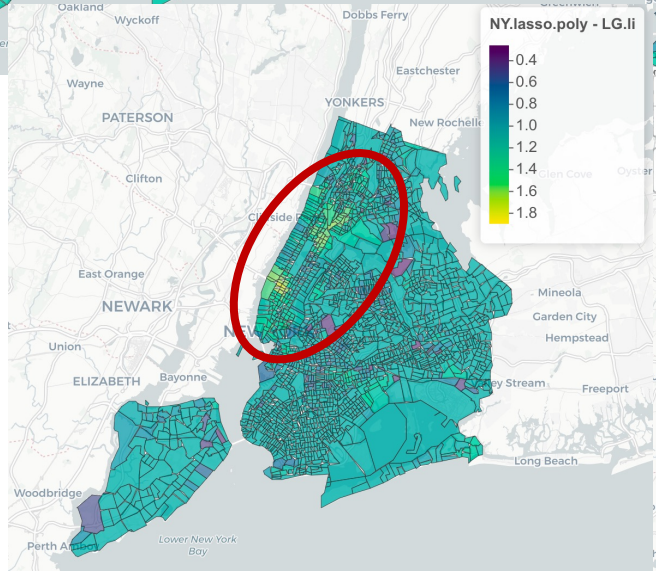


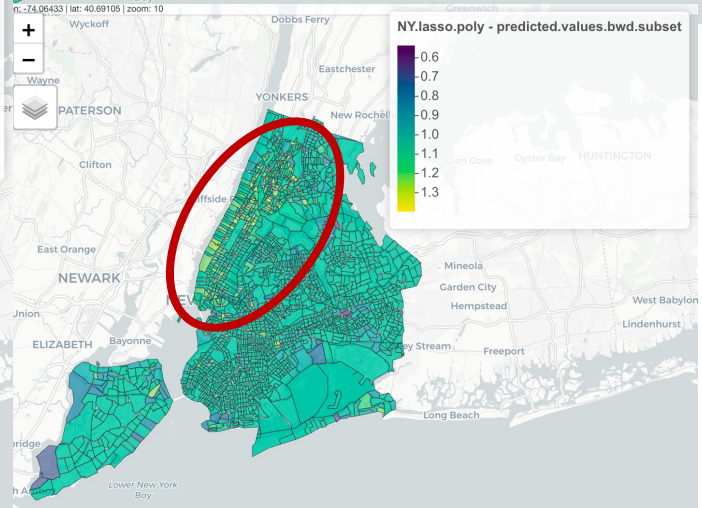
Linear Model Selection Variable Transformation

Variable Selection In Lasso









Variable Transformation

Cross Validation

Model	Test
Ridge	0.3935499
Ridge.cv	0.361645
Lasso	0.3702096
Lasso CV	0.3580408

Linear Model Selection

Lasso Cross Validation

(Intercept)	1.226341e+00	1.031536e+00
X..Population.25.Years.and.Over..Less.than.High.School	.	7.213353e-04
X..Population.25.Years.and.Over..High.School.Graduate..Includes.Equivalency.	-1.028029e-02	-1.004801e-02
X..Population.25.Years.and.Over..Some.College	.	.
X..Population.25.Years.and.Over..Bachelor.s.Degree	.	.
X..Population.25.Years.and.Over..Master.s.Degree	2.964294e-03	5.115259e-03
X..Population.25.Years.and.Over..Professional.School.Degree	.	2.172477e-05
X..Population.25.Years.and.Over..Doctorate.Degree	.	.
Vacant.Housing.Units.	5.276361e-05	4.791561e-05
X..Vacant.Housing.Units..for.Rent	2.360694e-03	2.380265e-03
X..Vacant.Housing.Units..for.Sale.Only	-1.539468e-03	-1.526193e-03
X..Vacant.Housing.Units..Other.Vacant	.	.
Average.Gross.Rent.for.Renter.Occupied.Housing.Units	.	-1.794749e-05
Population.for.Whom.Poverty.Status.Is.Determined	.	.
Total.Population	4.914073e-06	2.446886e-06
Median.Age.	-6.157015e-03	-7.639197e-03
X..Total.Population..Not.Hispanic.or.Latino	.	.
X..Total.Population..Not.Hispanic.or.Latino..White.Alone	-4.339125e-03	-1.001028e-03
X..Total.Population..Not.Hispanic.or.Latino..Black.or.African.American.Alone	.	.
X..Total.Population..Not.Hispanic.or.Latino..American.Indian.and.Alaska.Native.Alone	7.333904e-02	-7.604842e-02
X..Total.Population..Not.Hispanic.or.Latino..Asian.Alone	.	1.121376e-05
X..Total.Population..Not.Hispanic.or.Latino..Native.Hawaiian.and.Other.Pacific.Islander.Alone	.	.
X..Total.Population..Not.Hispanic.or.Latino..Some.Other.Race.Alone	-2.206688e-02	-2.307147e-02
X..Total.Population..Not.Hispanic.or.Latino..Two.or.More.Races	1.859080e-03	4.595172e-03
Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed	2.969046e-04	3.850153e-04
X..Civilian.Population.in.Labor.Force.16.Years.and.Over..Employed	9.877019e-04	4.455666e-03
X..Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed	-2.644463e-16	-1.561947e-15
Per.Capita.Income..In.2019.Inflation.Adjusted.Dollars.	5.337049e-06	6.715336e-06
Gini.Index	6.704151e-01	5.095724e-01
meidan_income	-3.941705e-06	-4.864655e-06

Lasso alone	0.3702096
Lasso CV	0.3580408

Lasso

Using min lambda from CV

Lasso. CV

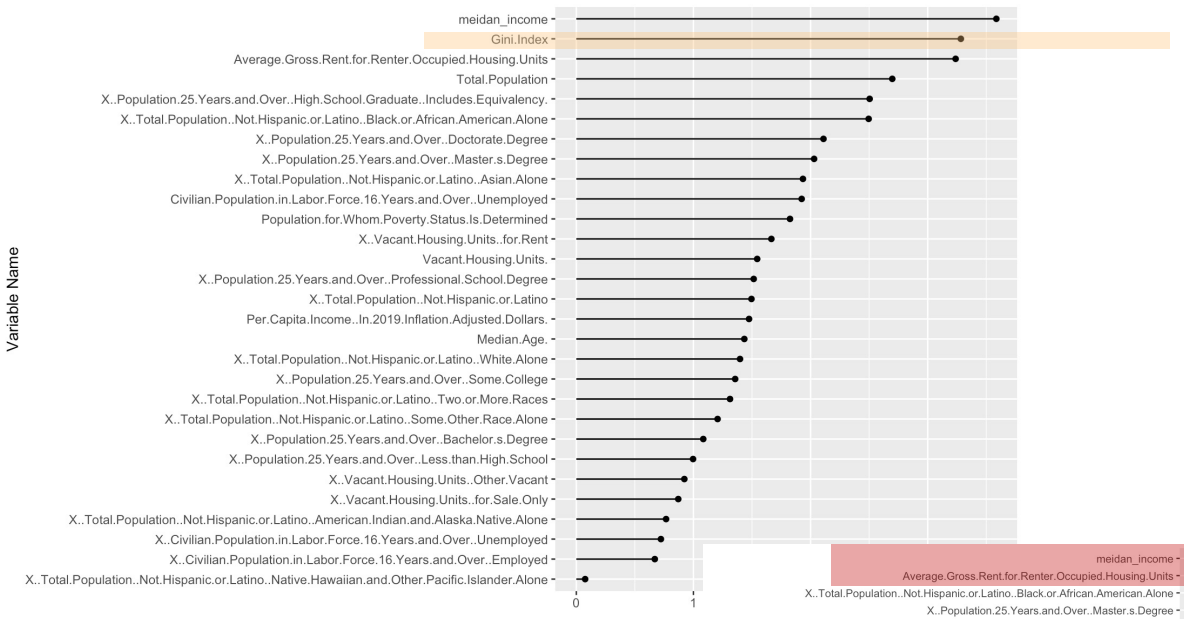
Variable Transformation

Random Forest

Test error: 0.06552897

Variable Transformation

Bagging

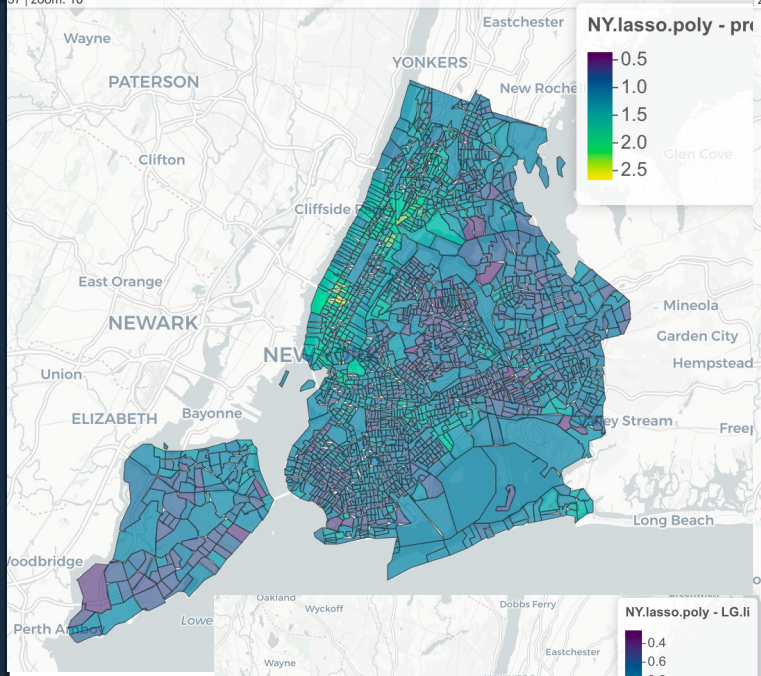


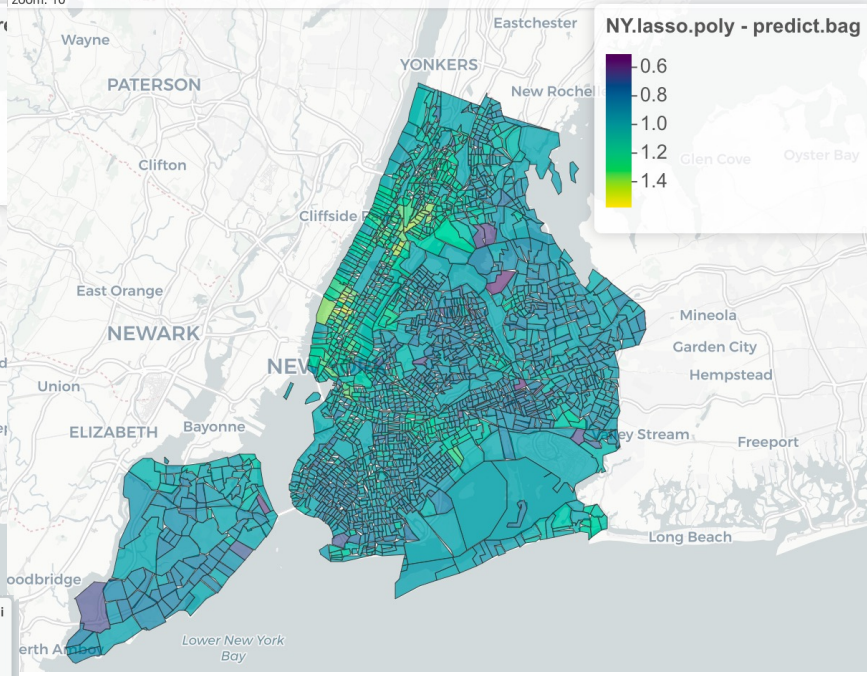
Test error: 0.06566407

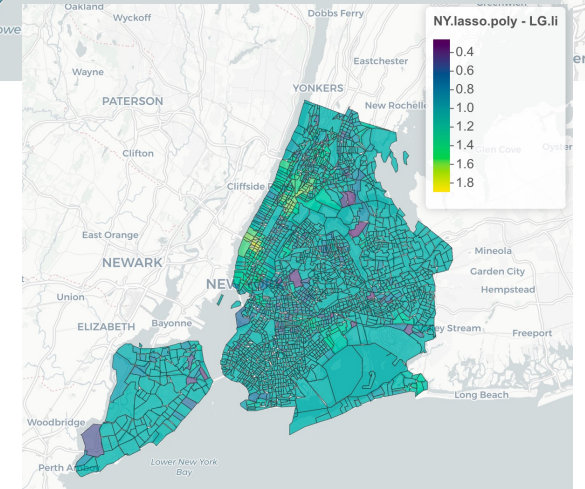
OOB: 0.05754728

Variable Transformation

Random Forest & Bagging



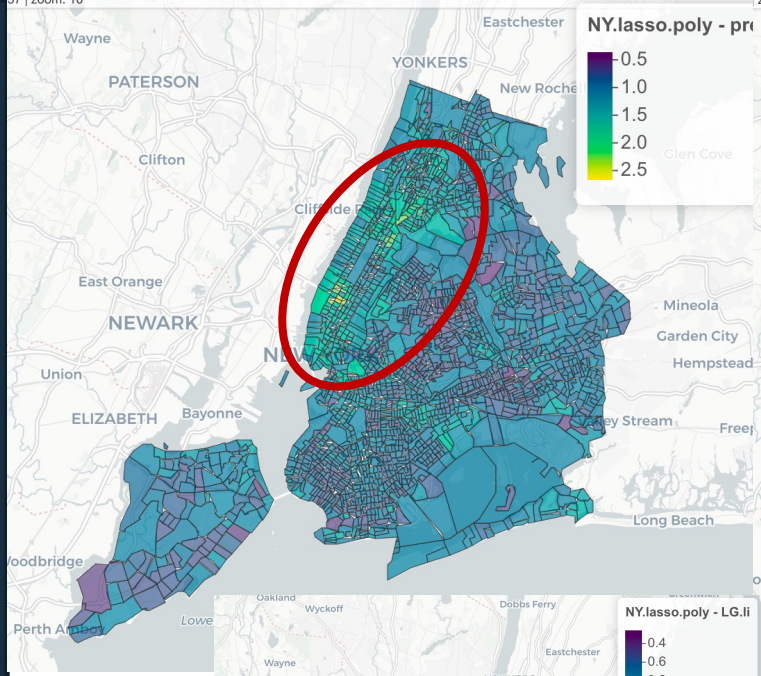


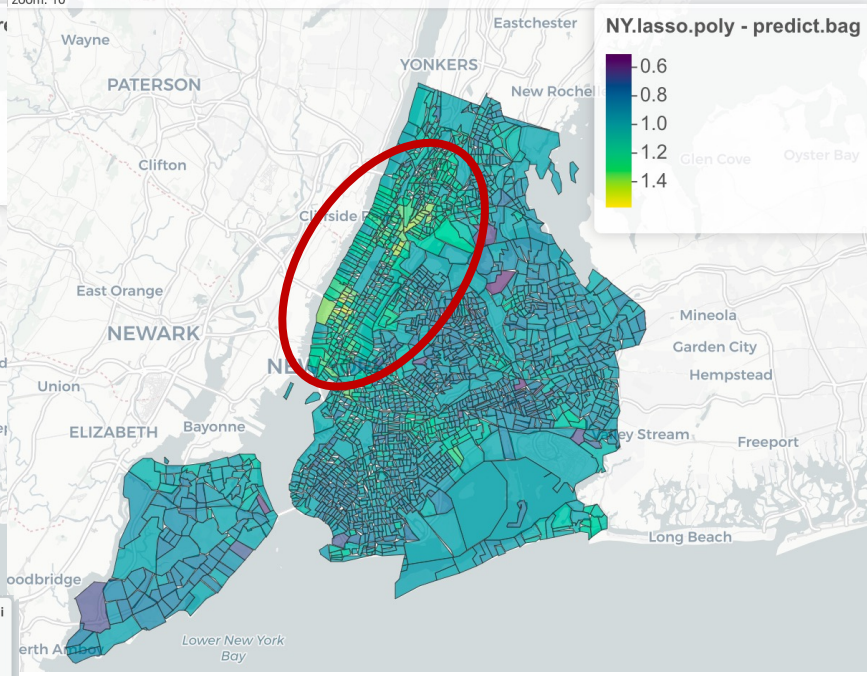


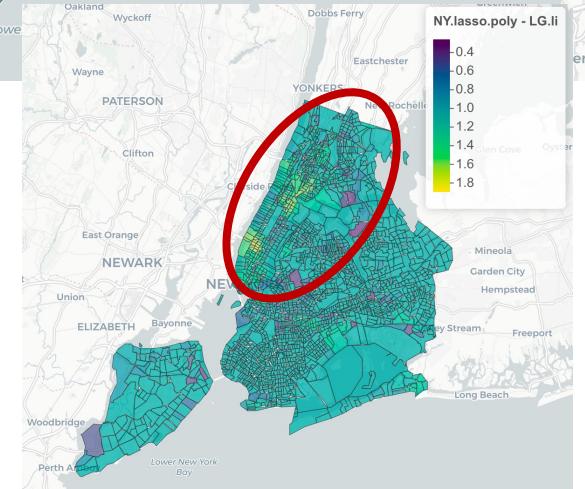
RF	0.06552897
Bagging	0.06566407

Variable Transformation

Random Forest & Bagging



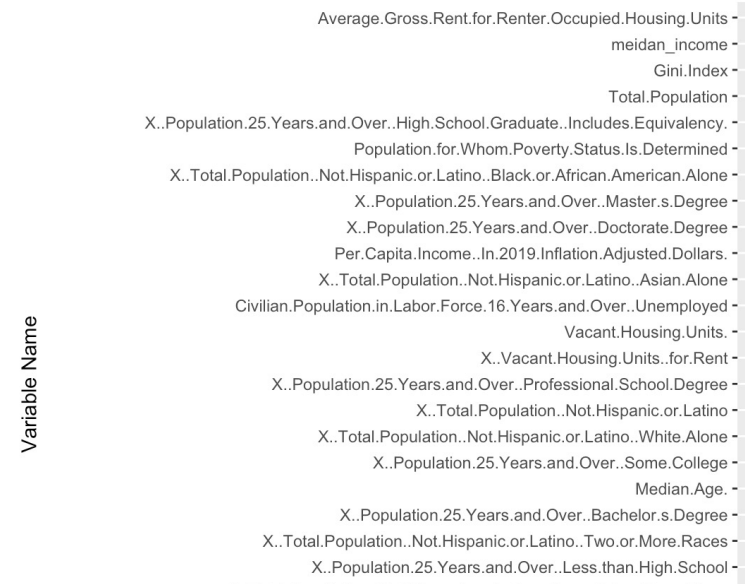




RF	0.06552897
Bagging	0.06566407

Variable Transformation

Random Forest & Backward Selection



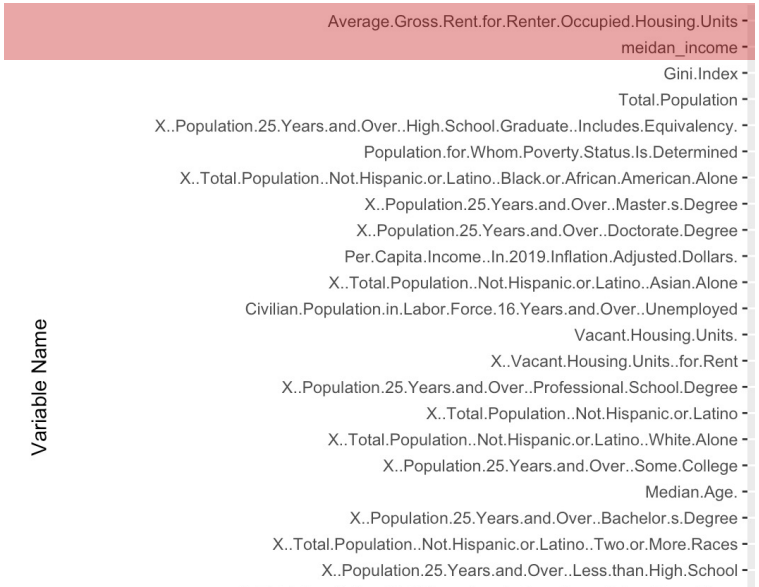
Random Forest

```
[1] "X..Population.25.Years.and.Over..High.School.Graduate..Includes.Equivalency."
[2] "X..Population.25.Years.and.Over..Some.College"
[3] "X..Vacant.Housing.Units..for.Rent"
[4] "X..Vacant.Housing.Units..for.Sale.Only"
[5] "X..Vacant.Housing.Units..Other.Vacant"
[6] "Total.Population"
[7] "Median.Age."
[8] "X..Total.Population..Not.Hispanic.or.Latino"
[9] "X..Total.Population..Not.Hispanic.or.Latino..White.Alone"
[10] "X..Total.Population..Not.Hispanic.or.Latino..American.Indian.and.Alaska.Native.Alone"
[11] "X..Total.Population..Not.Hispanic.or.Latino..Some.Other.Race.Alone"
[12] "Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed"
[13] "Per.Capita.Income..In.2019.Inflation.Adjusted.Dollars."
[14] "Gini.Index"
[15] "LG.li"
```

Backward Subset selection

Variable Transformation

Random Forest & Backward Selection



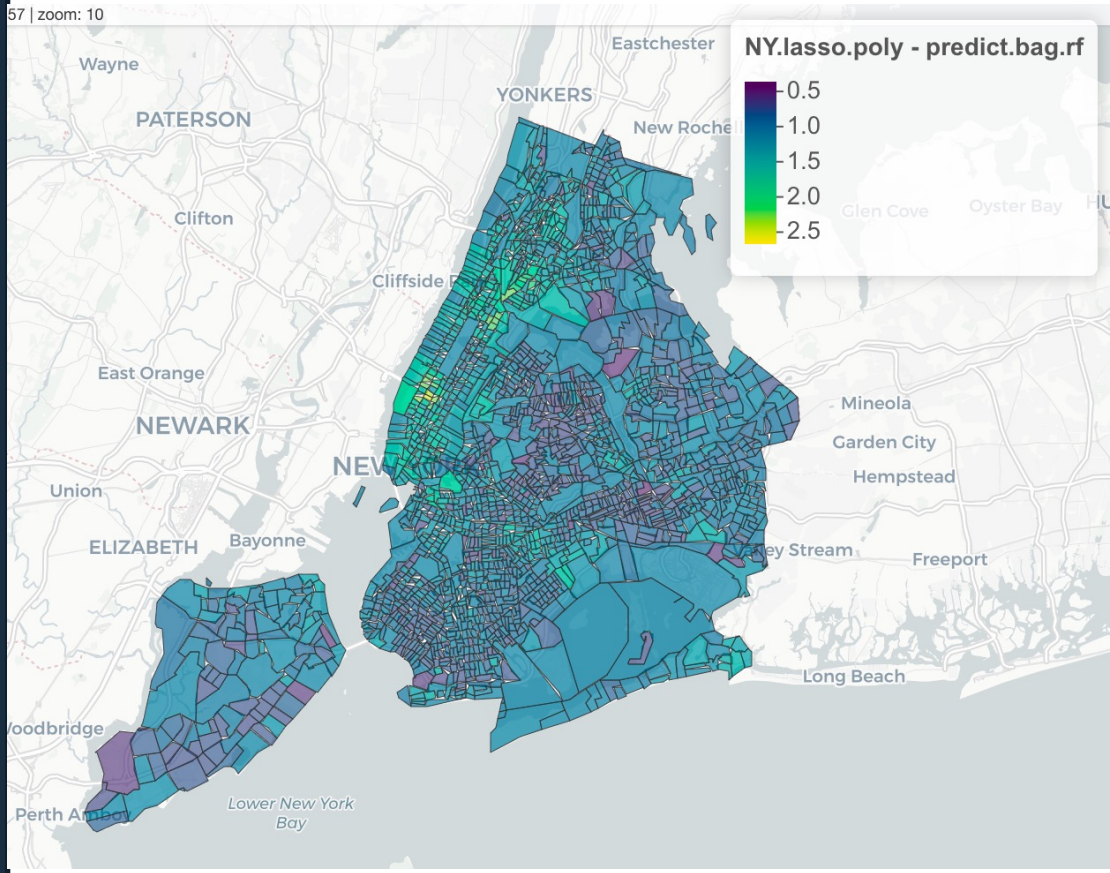
Random Forest

```
[1] "X..Population.25.Years.and.Over..High.School.Graduate..Includes.Equivalency."
[2] "X..Population.25.Years.and.Over..Some.College"
[3] "X..Vacant.Housing.Units..for.Rent"
[4] "X..Vacant.Housing.Units..for.Sale.Only"
[5] "X..Vacant.Housing.Units..Other.Vacant"
[6] "Total.Population"
[7] "Median.Age."
[8] "X..Total.Population..Not.Hispanic.or.Latino"
[9] "X..Total.Population..Not.Hispanic.or.Latino..White.Alone"
[10] "X..Total.Population..Not.Hispanic.or.Latino..American.Indian.and.Alaska.Native.Alone"
[11] "X..Total.Population..Not.Hispanic.or.Latino..Some.Other.Race.Alone"
[12] "Civilian.Population.in.Labor.Force.16.Years.and.Over..Unemployed"
[13] "Per.Capita.Income..In.2019.Inflation.Adjusted.Dollars."
[14] "Gini.Index"
[15] "LG.li"
```

Backward Subset selection

Summary

- Data transformation is the key point in this project
- Using Backward selection helps improve the variable selection
- Random forest and Bagging have the best performance among all models
- Gini_index is the most important variable
- Median income and average gross rent might the key variable



Thank you!