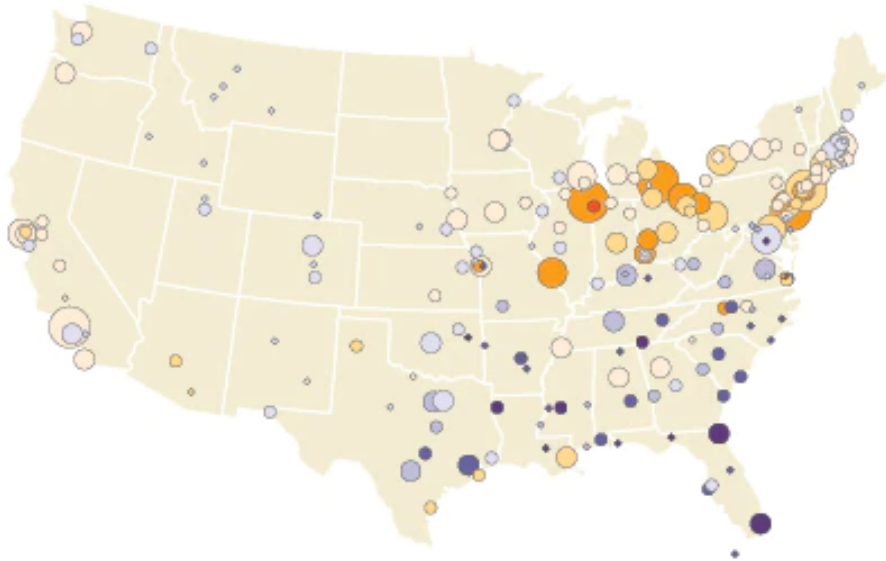
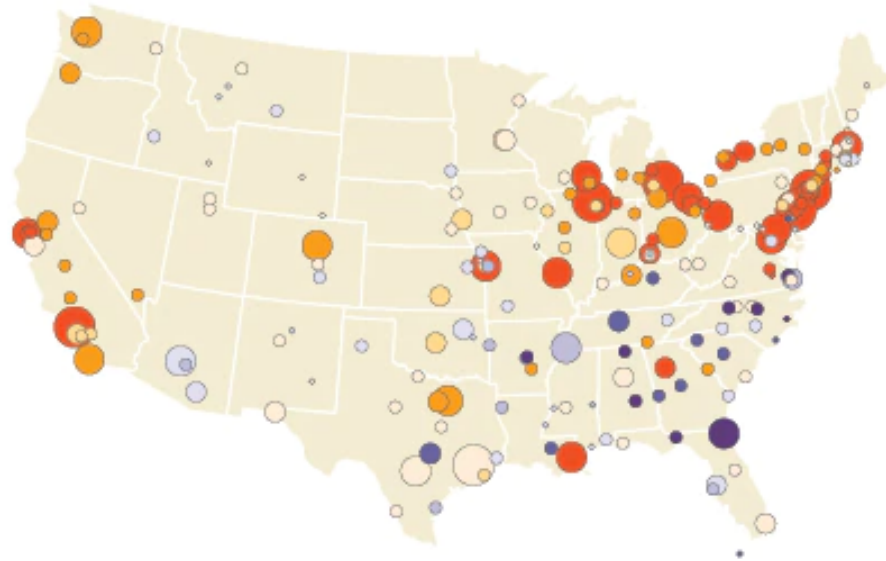
Analiza danych dotyczących migracji na terenie USA

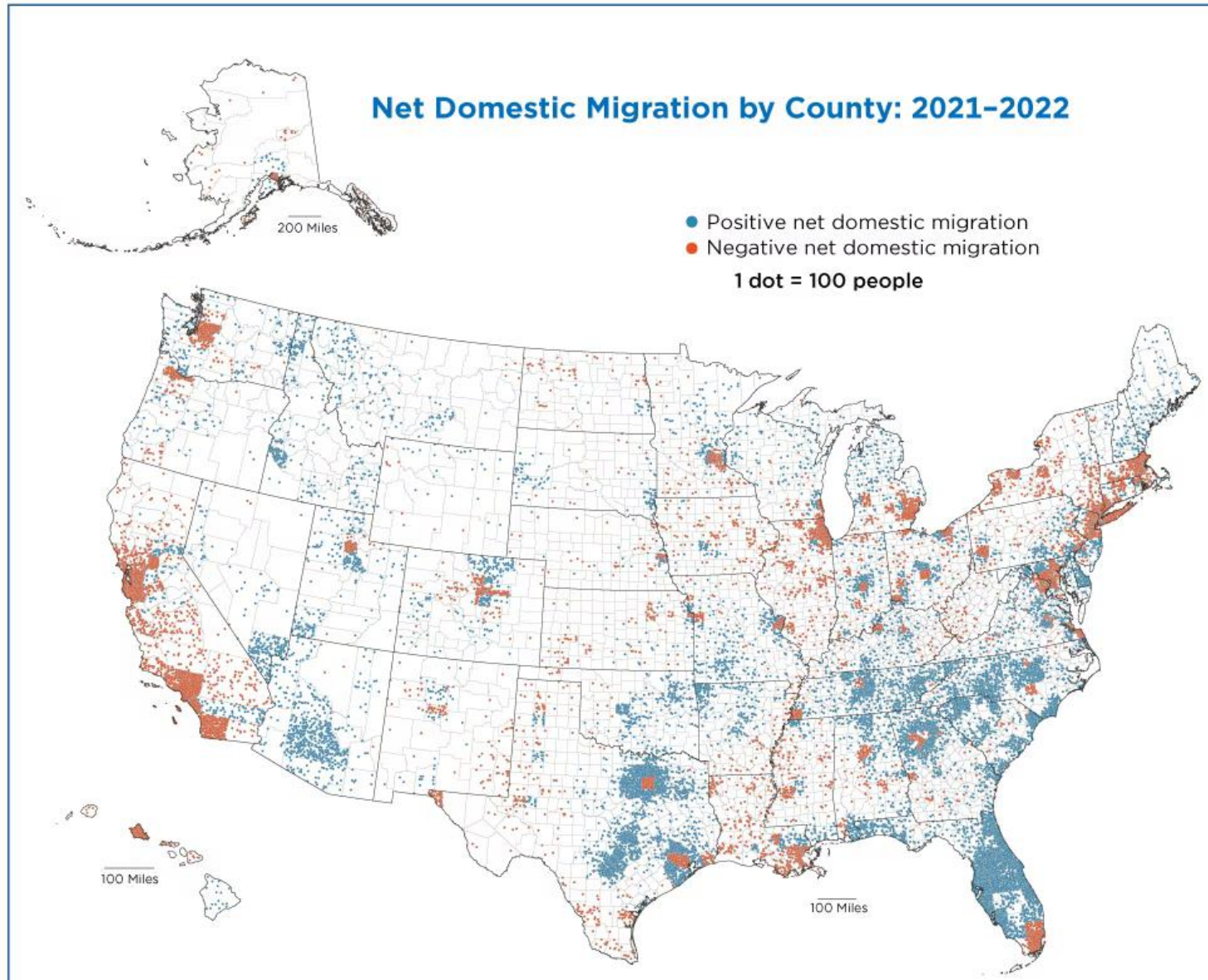
The First Great Migration: 1910-1940



The Second Great Migration: 1940-1970

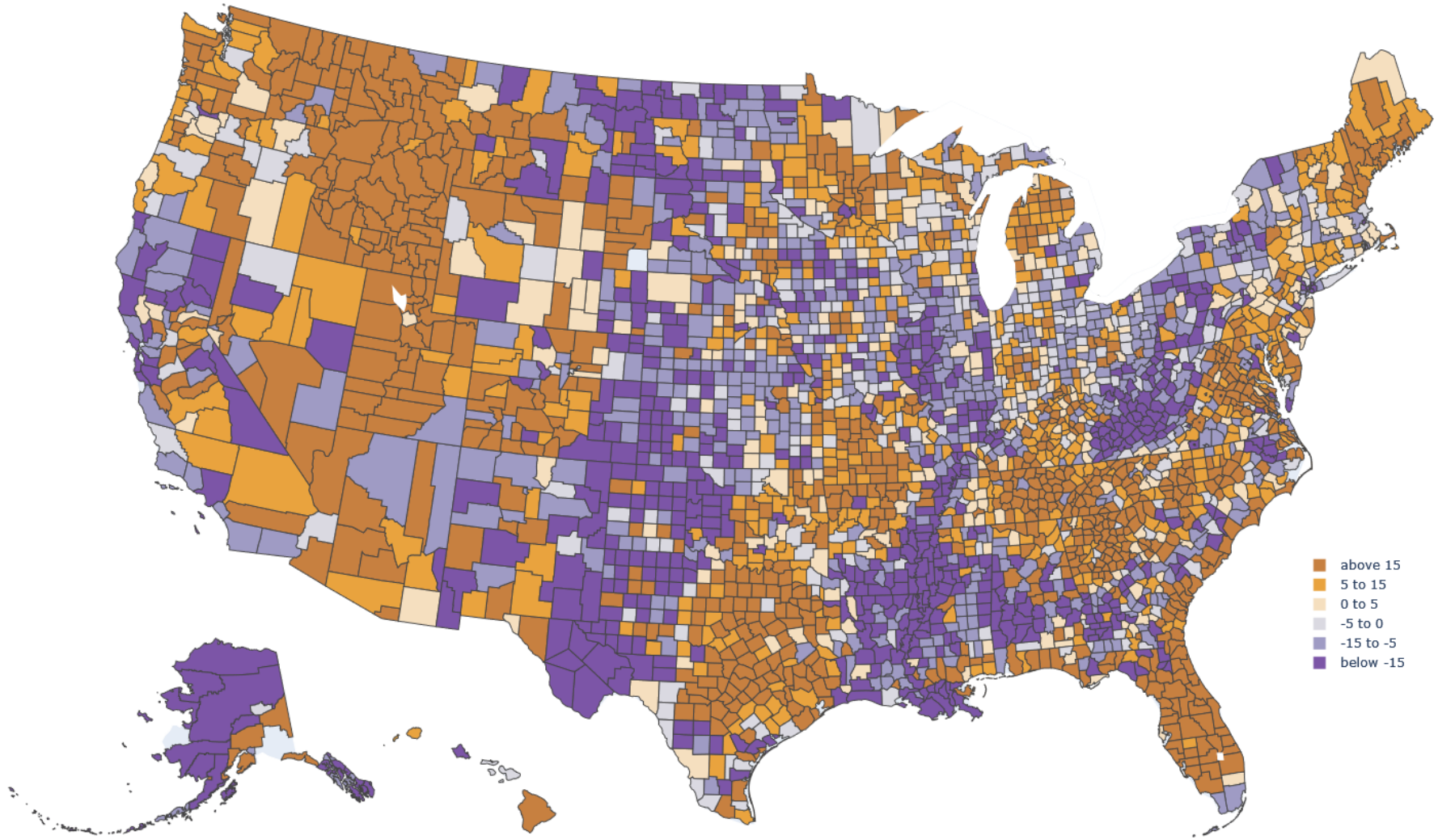


The change in share of Blacks in cities is based on the percentage point difference in the percent of population that was Black in the later time period compared to the earlier. For example, 18.3 percent of the population in Gary, IN was Black in 1940 but was just 2.3 in 1910, which represented a 16.0 percentage-point change in the share of Blacks in the city. It was the largest change in share during the First Great Migration. By the end of the Second Great Migration, Newark, NJ had realized the largest increase in Black population share, with the Black proportion of the city rising from 10.6 in 1940 to 54.2 in 1970.

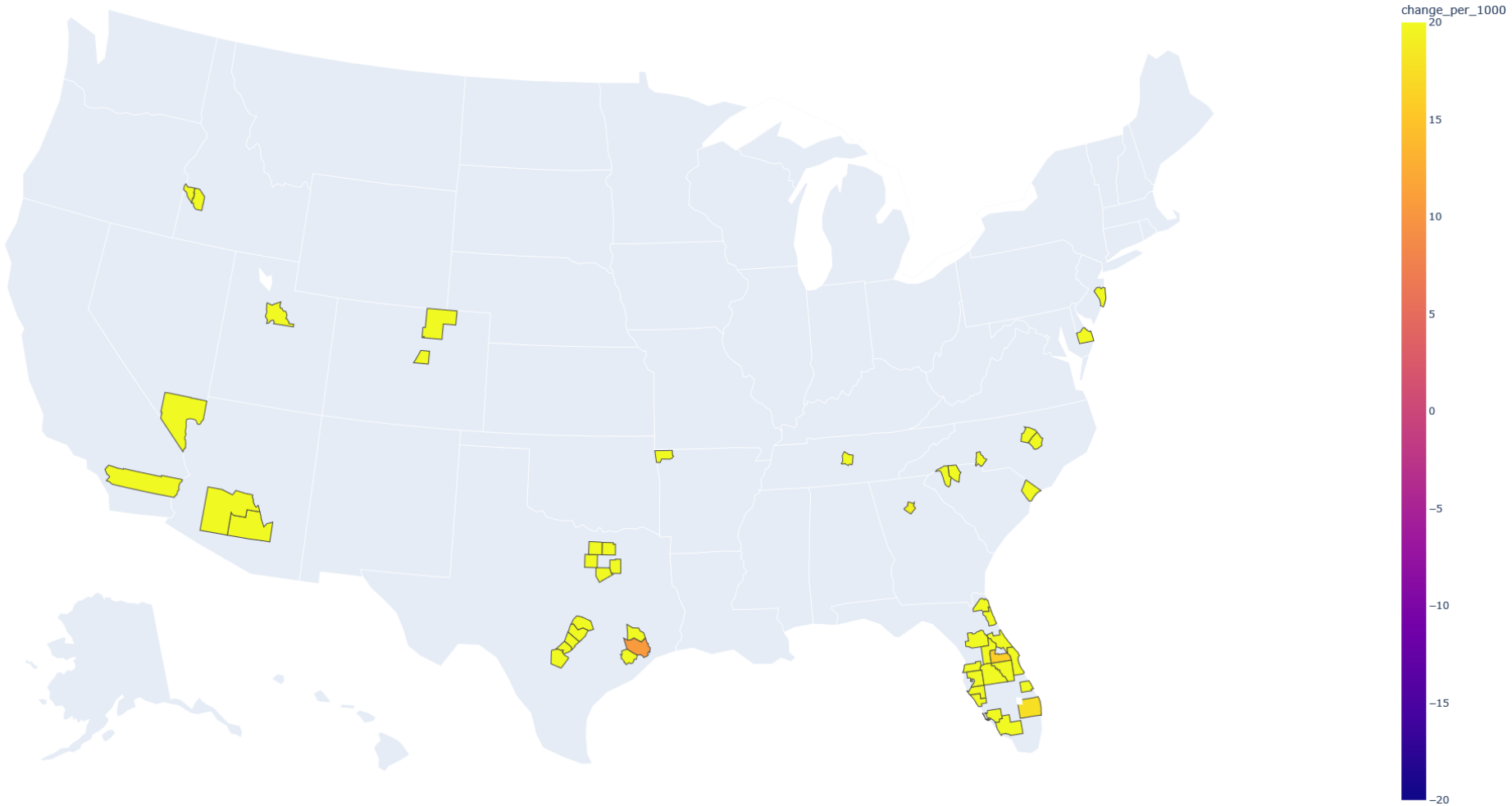


Source: U.S. Census Bureau, Vintage 2022 Population Estimates.

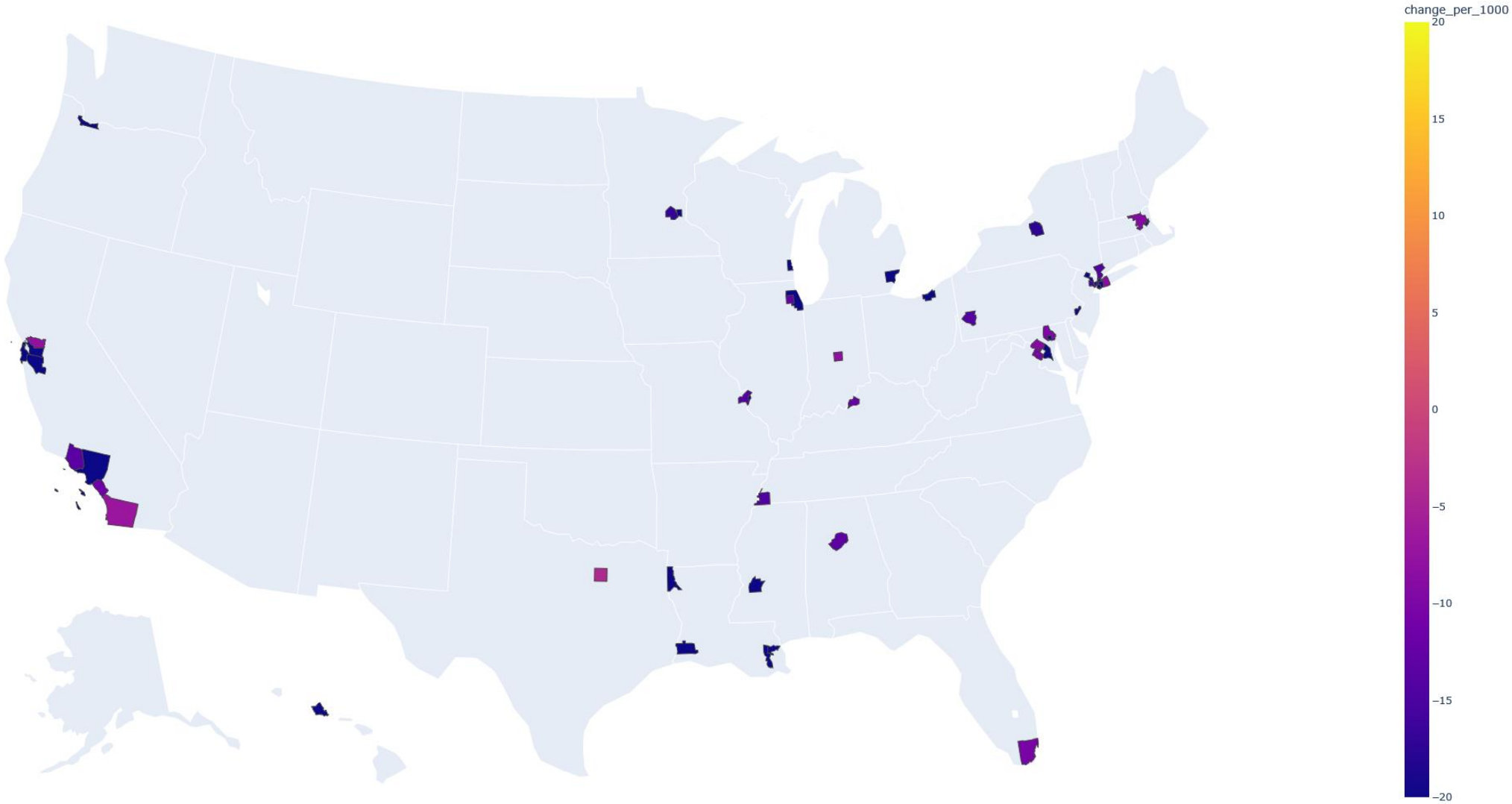
Net domestic migration per 1000 residents by county



50 hrabstw z największym przypływem ludności



50 hrabstw z największą emigracją



Dane

- CENSUS
- Bezrobocie
- PKB i zarobki
- House price index
- Dane dotyczące zdrowia
- Edukacja

Badania korelacji

..\research\SWEETVIZ_REPORT.html

Badania wpływu zmiennych

Sklearn feature selection

RandomForrestRegressor

```python
from sklearn.ensemble import RandomForestRegressor
import numpy as np

model = RandomForestRegressor(random_state=0)
model.fit(X, y)

importance = model.feature_importances_
indices = np.argsort(importance)[::-1]

print("Selected features:")
for f in range(10):
    print(importance[indices[f]],end='\t')
    print(X.columns[indices[f]])
```

```
Selected features:
0.15933233328624144     High school diploma only, 1990
0.05164274791925221     2022 HPI Change
0.05086790671881488     Less than a high school diploma, 1980
0.047528534264601935    Driving alone to work raw value 2016
0.046495907313077435    High school diploma only, 1980
0.03361601998534051     Unemployed_2021
0.028327164057220452    Less than a high school diploma, 1990
0.026309504296987148    High school diploma only, 2017-21
0.019983789638549847    Some college or associate's degree, 2017-21
0.018819059203776176    Some college (1-3 years), 1970
```

RFE - recursive feature elimination

```python
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression


lr = LinearRegression()
rfe = RFE(lr, n_features_to_select=10)

rfe.fit(X, y)

print(f"Optimal number of features: {rfe.n_features_}")

selected_features = X.columns[rfe.support_]

print("Selected features:")
for feature in selected_features:
    print(feature)
```

```
Optimal number of features: 10
Selected features:
Dentists raw value 2016
Other primary care providers raw value 2017
Primary care physicians raw value 2018
Dentists raw value 2018
Mental health providers raw value 2018
Dentists raw value 2019
Mental health providers raw value 2019
Other primary care providers raw value 2019
Primary care physicians raw value 2020
Other primary care providers raw value 2020
```

Lasso - linear model trained with L1 prior as regularizer (do tego wrócimy jeszcze)

```python
from sklearn.linear_model import Lasso
from sklearn.feature_selection import SelectFromModel

lasso = Lasso(max_iter=15000)
lasso.fit(X, y)

sfm = SelectFromModel(lasso, threshold=0.1)
sfm.fit(X, y)

selected_feat= X.columns[(sfm.get_support())]
print("Selected features:")
print(selected_feat)
```

```
Selected features:
Index(['Percent of adults with less than a high school diploma, 1970',
       'Percent of adults with a high school diploma only, 1980',
       'Percent of adults with a bachelor's degree or higher, 2008-12',
       'Percent of adults with a bachelor's degree or higher, 2017-21',
       '2001 HPI Change', '2002 HPI Change', '2004 HPI Change',
       '2005 HPI Change', '2006 HPI Change', 'Unemployment_rate_2013'],
      dtype='object')
```

SelectKBest - removes all but the highest scoring features

```python
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.preprocessing import StandardScaler


scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

selector = SelectKBest(score_func=f_regression, k=10)
selector.fit(X_scaled, y)

selected_features = X.columns[selector.get_support()]
print("Selected features:")
print(selected_features)
```

```
Selected features:
Index(['Percent of adults with a bachelor's degree or higher, 2000',
       'Percent of adults with a bachelor's degree or higher, 2008-12',
       'Percent of adults with a bachelor's degree or higher, 2017-21',
       'Median household income raw value 2016',
       'Children in poverty raw value 2016', 'Some college raw value 2016',
       'Premature death raw value 2016',
       'Premature age-adjusted mortality raw value 2016',
       'Unemployment_rate_2001', 'Median_Household_Income_2020'],
      dtype='object')
```

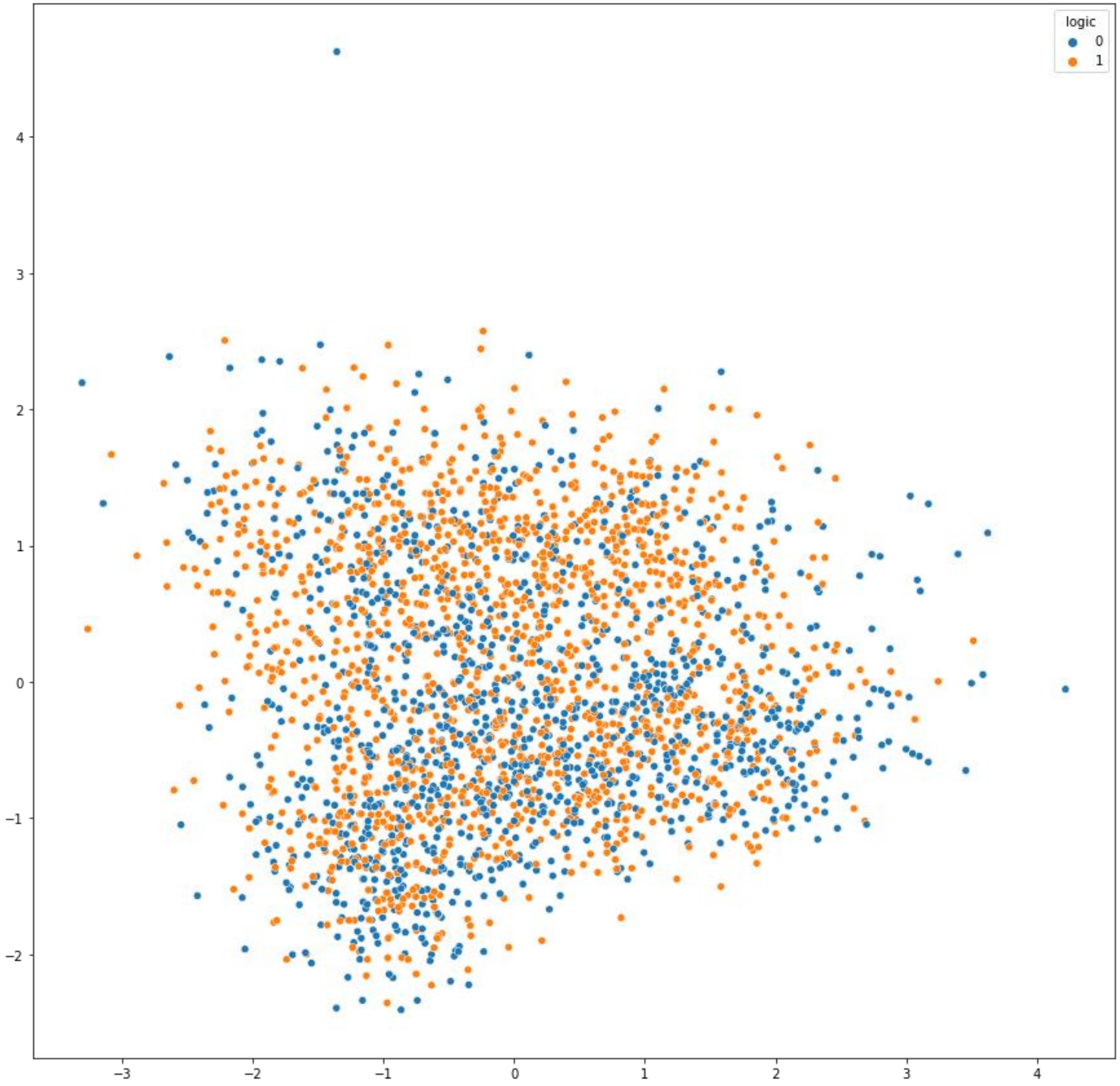
RidgeCV - ridge regression with built-in cross-validation

```python
from sklearn.linear_model import RidgeCV

ridgecv = RidgeCV(alphas=[0.1, 1.0, 10.0], cv=5)
ridgecv.fit(X, y)

print("Selected features:")
counter=0
for coef, feature in sorted(zip(ridgecv.coef_, X.columns), reverse=True):
    if counter<10:
        if coef != 0:
            print("{:.3f}\t{}".format(coef, feature))
            counter=counter+1
```

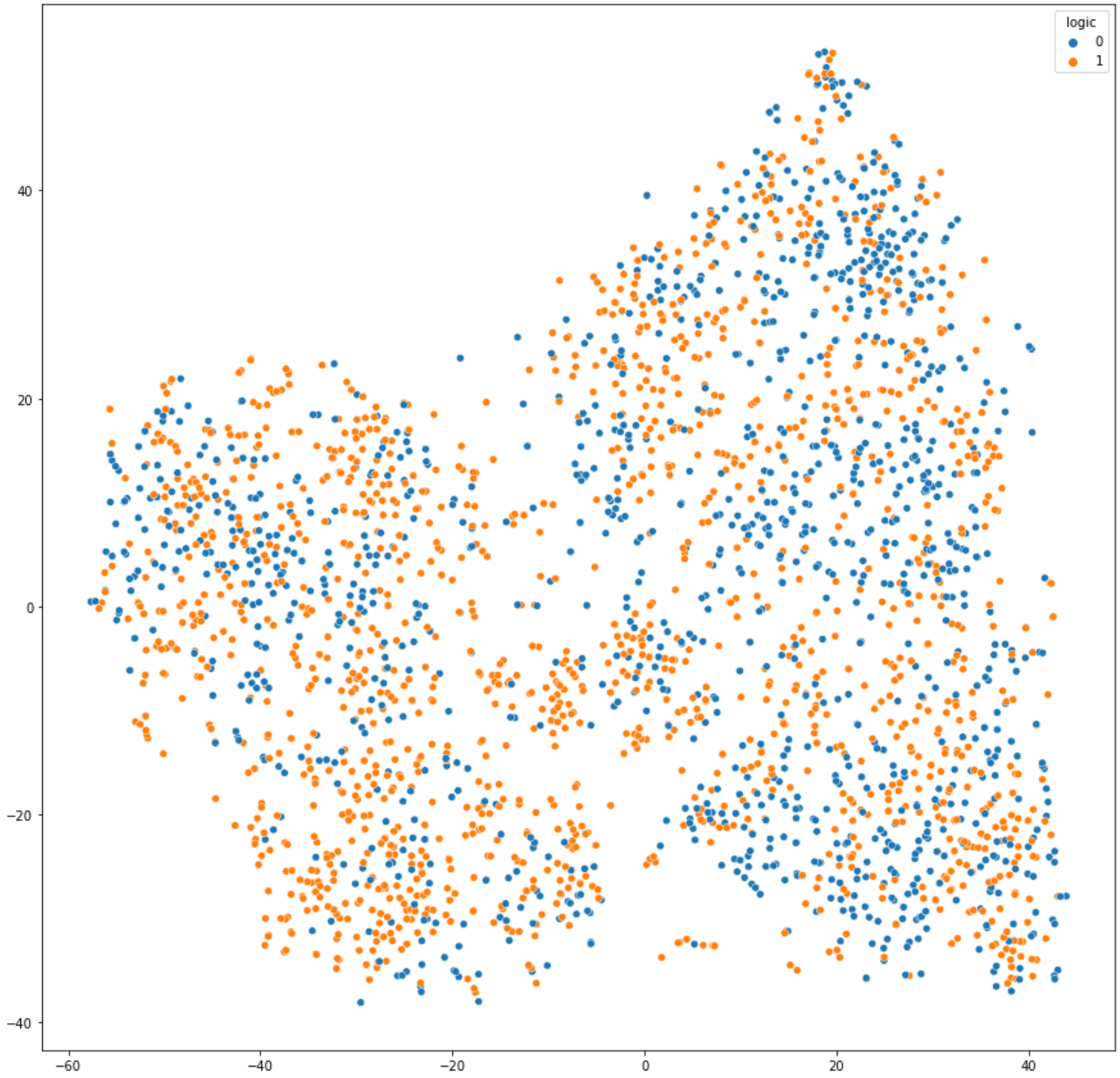
```
Selected features:
670.529 % Females raw value 2016
664.152 % not proficient in English raw value 2016
605.183 Physical inactivity raw value 2019
463.970 Severe housing problems raw value 2017
428.092 Uninsured adults raw value 2017
427.776 Mammography screening raw value 2017
410.106 Long commute - driving alone raw value 2020
393.967 Mammography screening raw value 2020
381.169 Some college raw value 2016
376.457 % 65 and older raw value 2017
```

Dekompozycja

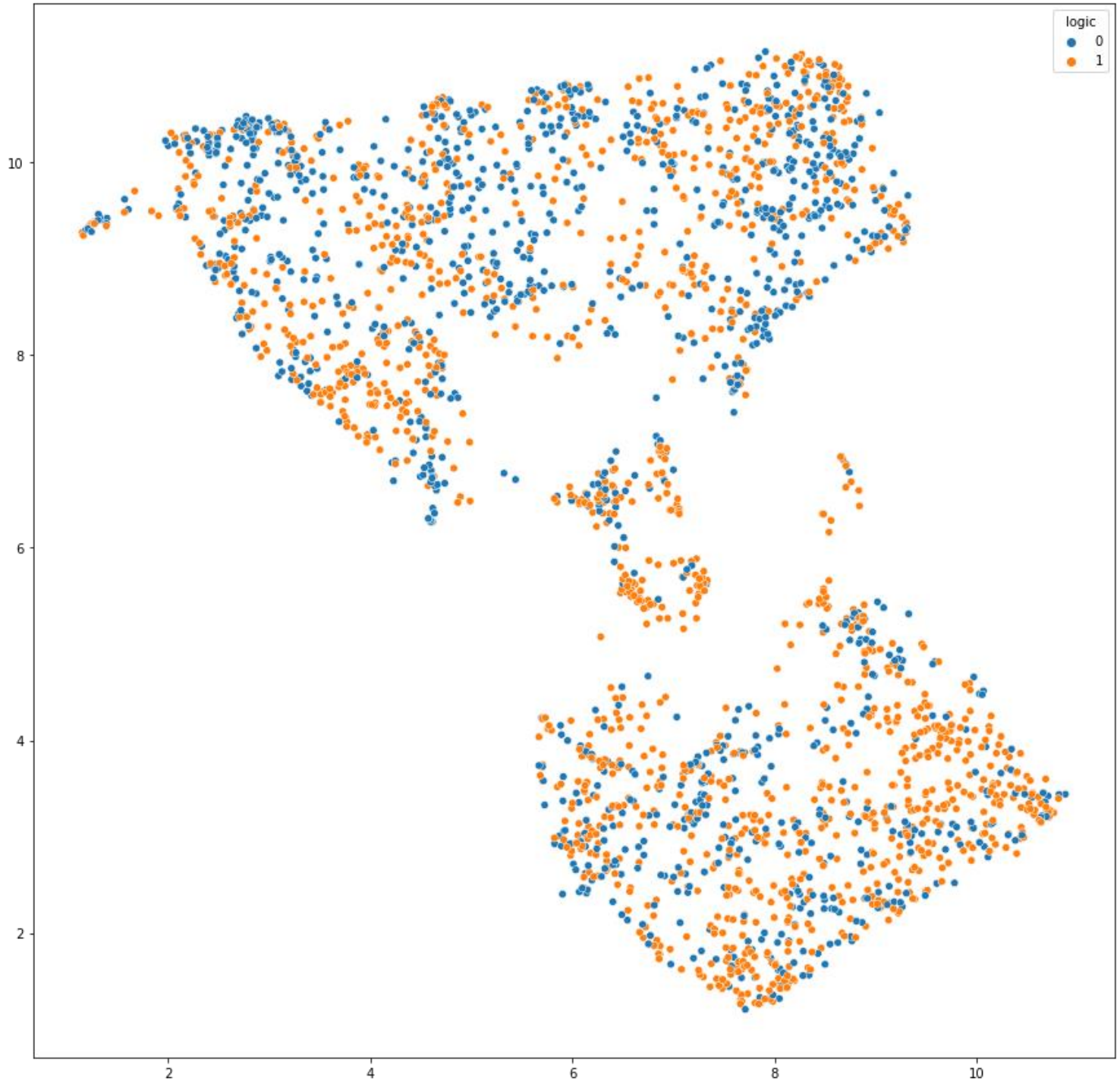
PCA



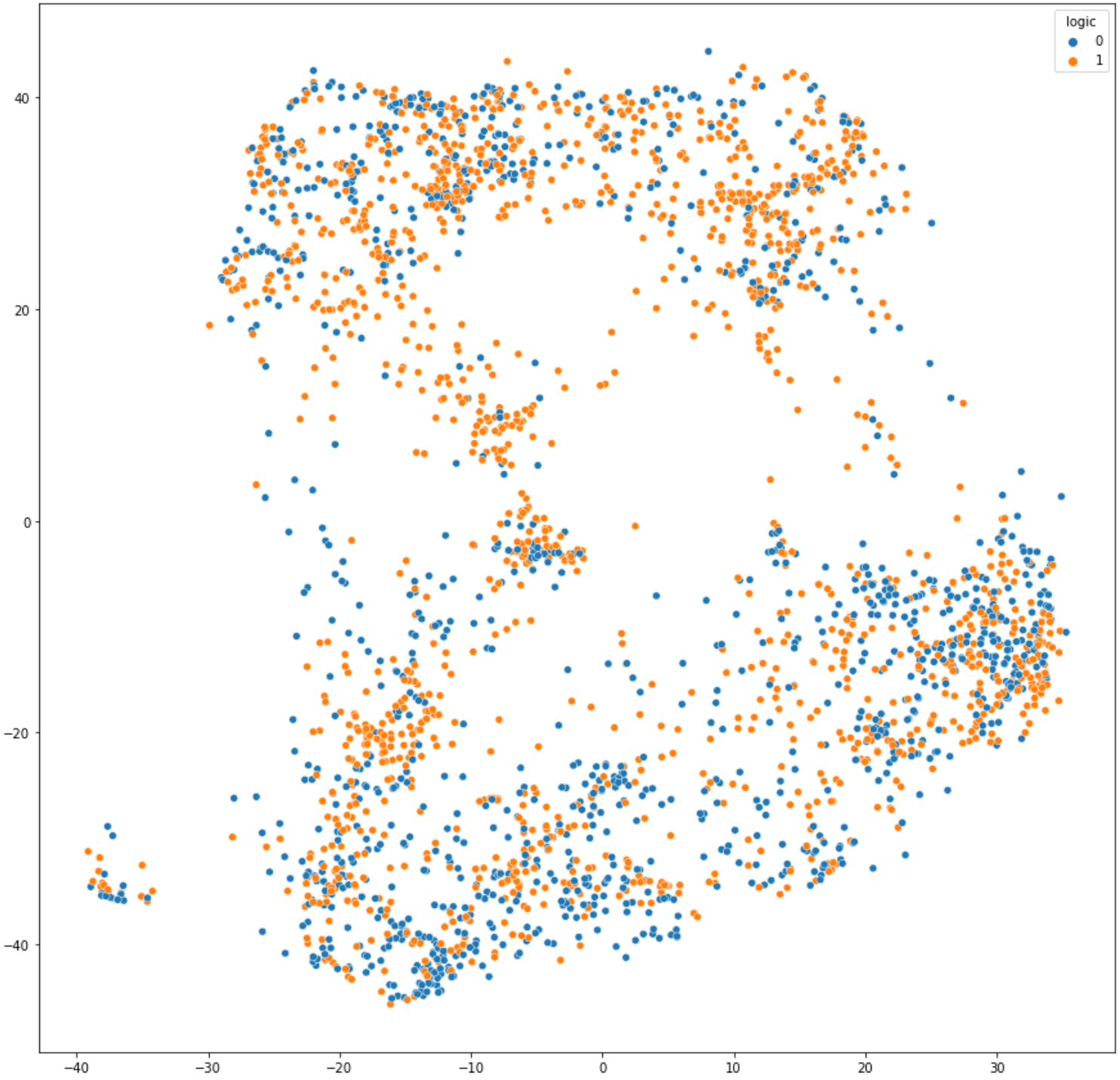
TSNE



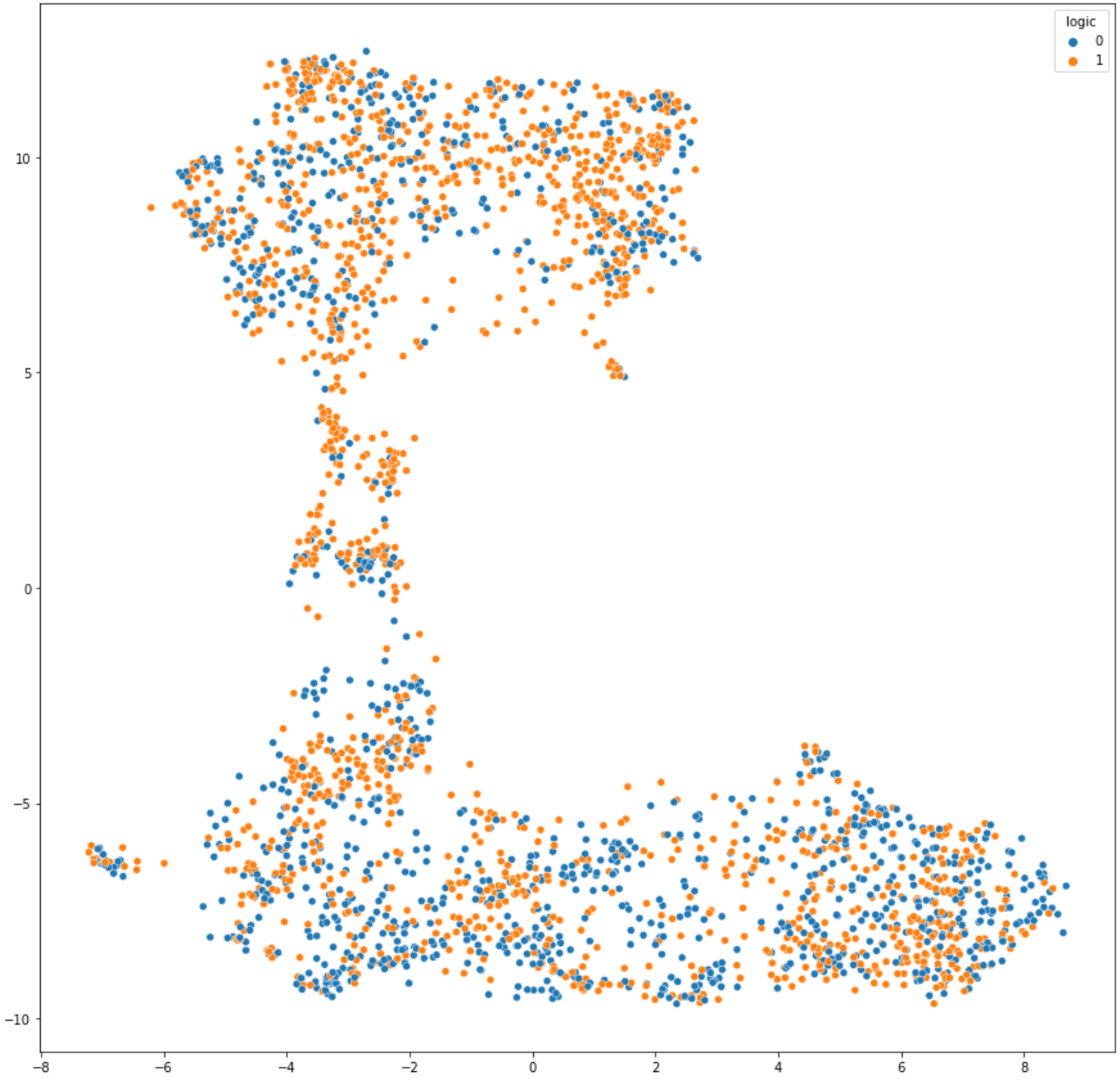
UMAP



TriMAP



PaCMAP



Las drzew decyzyjnych

Trzy podejścia:

- Drzewo decyzyjne na podstawie wszystkich danych
- Drzewo na podstawie wybranego roku
- Drzewo na podstawie „delt” na przestrzeni lat

Wyniki każdego z tych podejść oscylowały w okolicach 75% celności.

Próby regresji na zbiorze dawały mieszane wyniki. Model **RandomForrestRegressor** często dawał bardzo dobre wyniki, jednak w innych próbkach mylił się znacząco.

Metryki:

- R^2 - 0.46
- MAE – 837.11

Badanie ważności cech

Model liniowy
dla poszczególnych lat

2016

```
                                                  Estimate Std. Error t value Pr(>|t|)
X2016.Real.GDP..thousands.of.chained.2012.dollars.  7.748e-01  6.944e-02  11.159  < 2e-16 ***
Median.household.income.raw.value.2016              1.805e-01  3.949e-02   4.571 5.07e-06 ***
X..Hispanic.raw.value.2016                          1.850e-01  3.899e-02   4.744 2.21e-06 ***
Mammography.screening.raw.value.2016                9.389e-02  1.873e-02   5.014 5.69e-07 ***
Diabetes.prevalence.raw.value.2016                 -2.397e-01  5.376e-02  -4.458 8.60e-06 ***
X..65.and.older.raw.value.2016                      1.784e-01  4.485e-02   3.978 7.13e-05 ***
Driving.alone.to.work.raw.value.2016                2.785e-01  2.056e-02  13.543  < 2e-16 ***
Poor.mental.health.days.raw.value.2016             -4.741e-01  1.120e-01  -4.234 2.37e-05 ***
Uninsured.adults.raw.value.2016                     1.474e+00  4.220e-01   3.493 0.000486 ***
Population.raw.value.2016                          -2.583e+01  9.828e-01 -26.279  < 2e-16 ***
Frequent.mental.distress.raw.value.2016             7.110e-01  1.371e-01   5.185 2.32e-07 ***
X..not.proficient.in.English.raw.value.2016        -1.437e-01  3.577e-02  -4.017 6.06e-05 ***
Uninsured.children.raw.value.2016                   4.152e-01  9.225e-02   4.501 7.05e-06 ***
X..Asian.raw.value.2016                            -1.199e-01  2.535e-02  -4.732 2.34e-06 ***
X..Rural.raw.value.2016                            -1.208e-01  2.924e-02  -4.131 3.72e-05 ***
X2016.Personal.income..thousands.of.dollars.       -1.401e+00  1.350e-01 -10.382  < 2e-16 ***
X2016.Population..persons..1.                       2.573e+01  9.998e-01  25.739  < 2e-16 ***
Civilian_labor_force_2016                          -1.802e+01  2.161e+00  -8.340  < 2e-16 ***
Employed_2016                                       1.866e+01  1.994e+00   9.358  < 2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2017

	Estimate	Std. Error	t value	Pr(>|t|)	
Less.than.a.high.school.diploma..2017.21	-1.036e+00	1.094e-01	-9.476	< 2e-16	***
Some.college.or.associate.s.degree..2017.21	2.138e+00	1.756e-01	12.170	< 2e-16	***
Bachelor.s.degree.or.higher..2017.21	-1.411e+00	2.244e-01	-6.287	3.77e-10	***
Percent.of.adults.with.less.than.a.high.school.diploma..2017.21	-8.185e-02	2.119e-02	-3.863	0.000115	***
Percent.of.adults.with.a.high.school.diploma.only..2017.21	-9.312e-02	2.066e-02	-4.507	6.86e-06	***
Percent.of.adults.completing.some.college.or.associate.s.degree..2017.21	-1.416e-01	2.114e-02	-6.700	2.54e-11	***
X2017.Real.GDP..thousands.of.chained.2012.dollars.	8.251e-01	7.666e-02	10.763	< 2e-16	***
X..Native.Hawaiian.Other.Pacific.Islander.raw.value.2017	-7.895e-02	1.896e-02	-4.165	3.21e-05	***
Civilian_labor_force_2017	-2.421e+01	2.808e+00	-8.621	< 2e-16	***
Employed_2017	2.324e+01	2.674e+00	8.693	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2018

	Estimate	Std. Error	t value	Pr(>|t|)	
X2018.Real.GDP..thousands.of.chained.2012.dollars.	8.705e-01	8.198e-02	10.618	< 2e-16	***
X2018.Chain.type.quantity.indexes.for.real.GDP	7.455e-02	1.798e-02	4.147	3.48e-05	***
X..Hispanic.raw.value.2018	2.035e-01	5.105e-02	3.986	6.90e-05	***
Air.pollution...particulate.matter.raw.value.2018	-1.117e-01	2.675e-02	-4.176	3.07e-05	***
Poor.mental.health.days.raw.value.2018	-3.745e-01	1.119e-01	-3.347	0.000828	***
X2018.Personal.income..thousands.of.dollars.	-1.778e+00	1.471e-01	-12.085	< 2e-16	***
X2018.Population..persons..1.	1.837e+00	3.590e-01	5.118	3.31e-07	***
Civilian_labor_force_2018	-4.835e+01	2.951e+00	-16.384	< 2e-16	***
Employed_2018	4.728e+01	2.788e+00	16.958	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2019

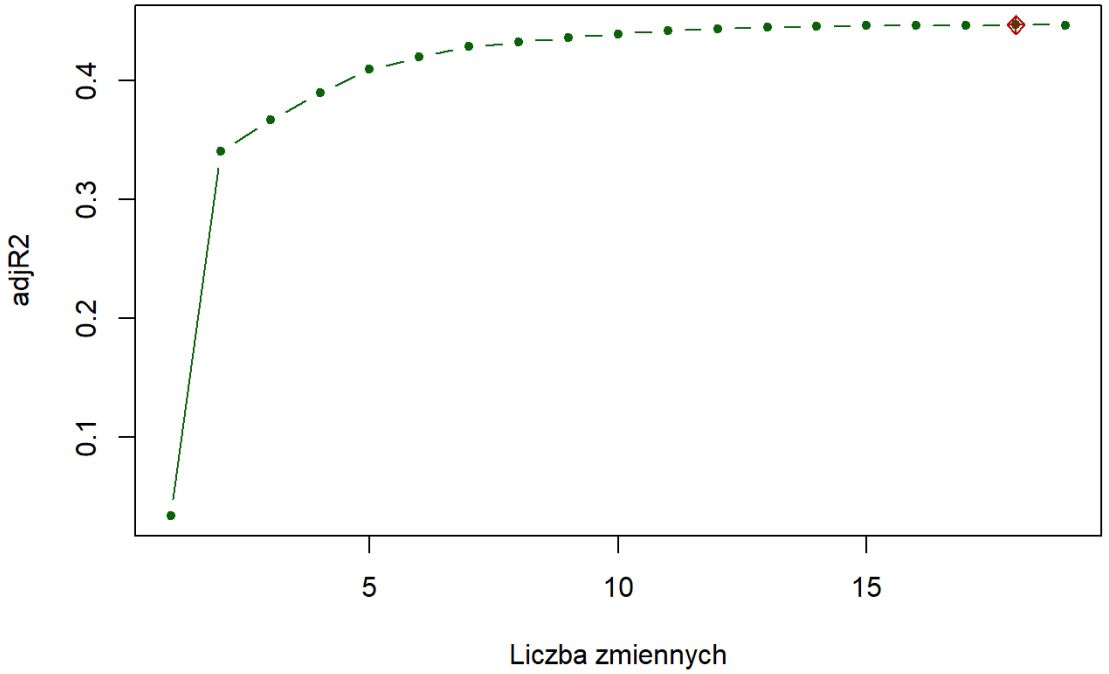
	Estimate	Std. Error	t value	Pr(>|t|)	
X2019.Real.GDP..thousands.of.chained.2012.dollars.	7.258e-01	7.497e-02	9.682	< 2e-16	***
X2019.Chain.type.quantity.indexes.for.real.GDP	7.599e-02	1.787e-02	4.252	2.19e-05	***
X..Hispanic.raw.value.2019	2.466e-01	5.157e-02	4.782	1.83e-06	***
Air.pollution...particulate.matter.raw.value.2019	-8.951e-02	2.528e-02	-3.541	0.000406	***
Drinking.water.violations.raw.value.2019	-6.933e-02	1.826e-02	-3.797	0.000150	***
X2019.Personal.income..thousands.of.dollars.	-1.716e+00	1.420e-01	-12.083	< 2e-16	***
X2019.Population..persons..1.	1.667e+00	3.290e-01	5.067	4.31e-07	***
Civilian_labor_force_2019	-4.966e+01	2.718e+00	-18.268	< 2e-16	***
Employed_2019	4.885e+01	2.594e+00	18.829	< 2e-16	***

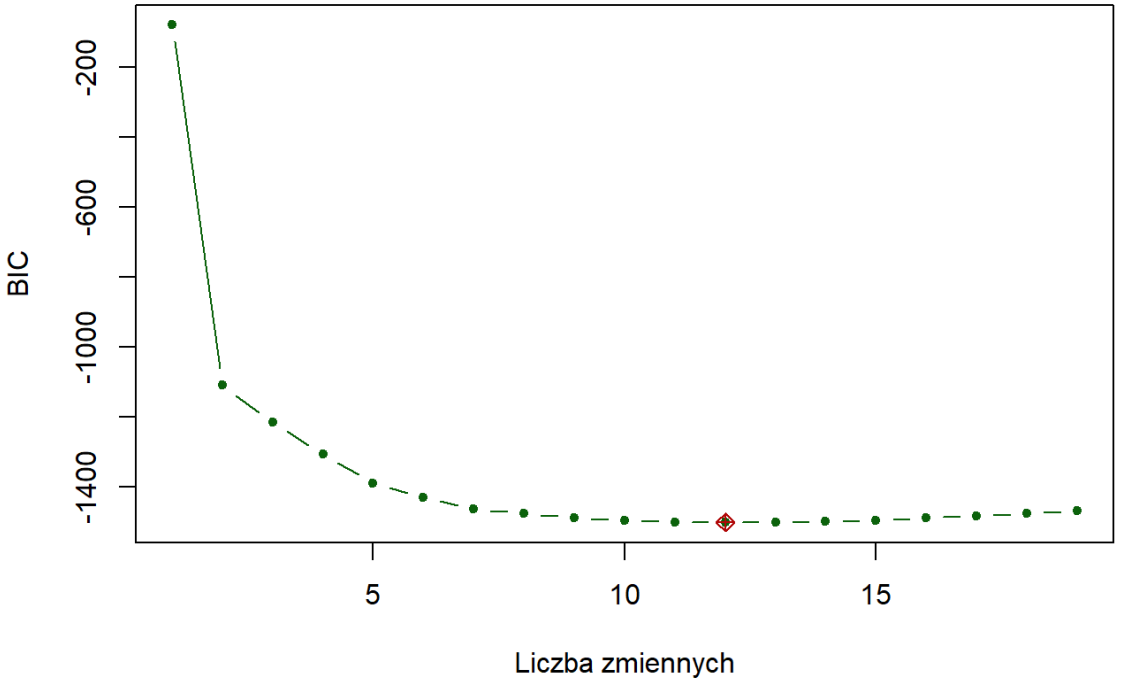
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2020

```
                                                  Estimate Std. Error t value Pr(>|t|)
X2020.Real.GDP..thousands.of.chained.2012.dollars.  7.470e-01  6.105e-02  12.236  < 2e-16 ***
X2020.Chain.type.quantity.indexes.for.real.GDP      5.937e-02  1.519e-02   3.908 9.52e-05 ***
Air.pollution...particulate.matter.raw.value.2020  -8.697e-02  2.122e-02  -4.098 4.29e-05 ***
X2020.Personal.income..thousands.of.dollars.       -1.782e+00  1.192e-01 -14.949  < 2e-16 ***
X2020.Population..persons..1.                       1.899e+00  2.753e-01   6.897 6.62e-12 ***
Civilian_labor_force_2020                          -2.198e+01  7.520e-01 -29.229  < 2e-16 ***
Employed_2020                                       2.098e+01  6.096e-01  34.412  < 2e-16 ***
Unemployment_rate_2020                              9.052e-02  1.839e-02   4.922 9.07e-07 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

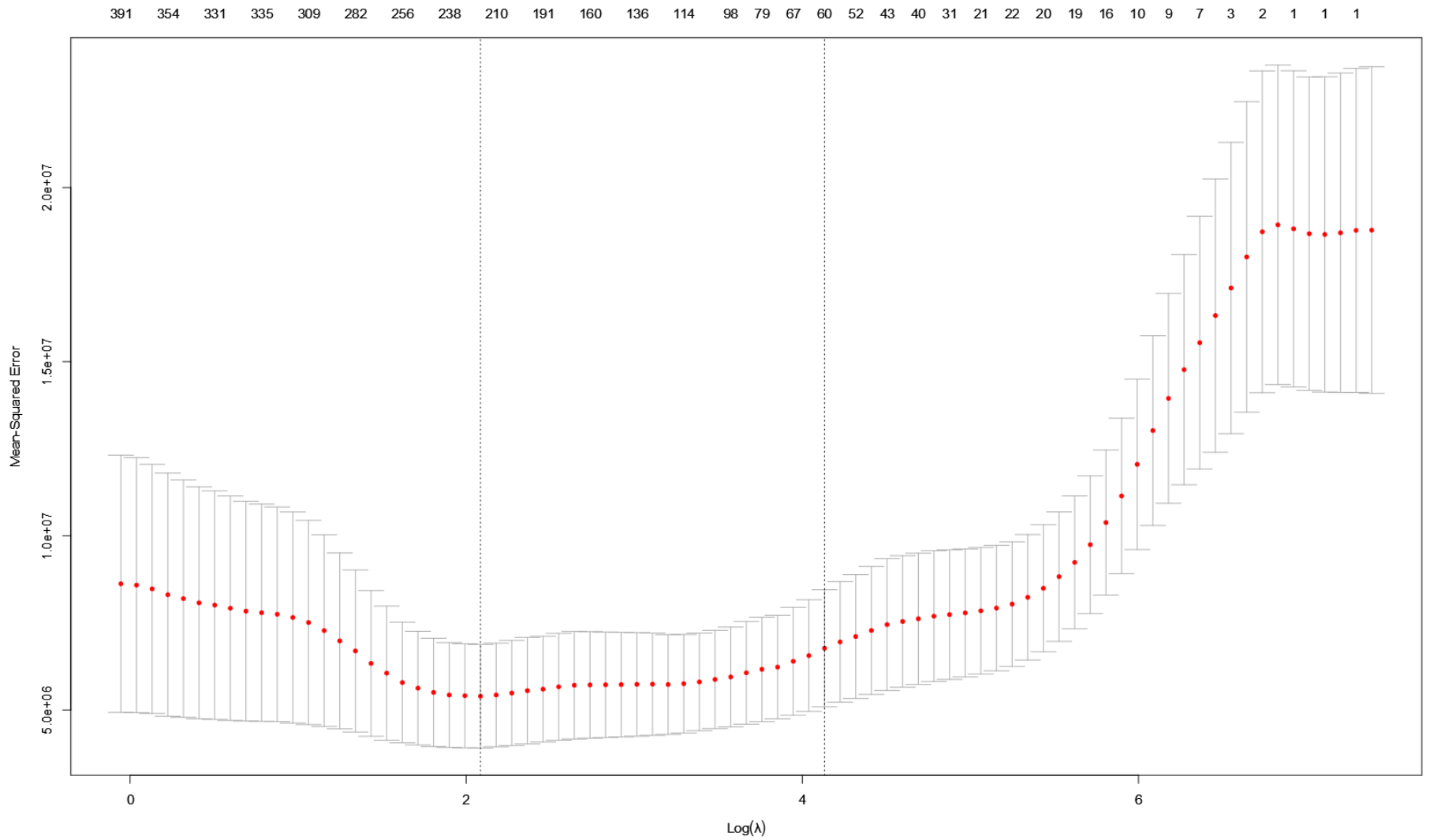
Modele liniowe z krokową selekcją cech





- X2020.Real.GDP..thousands.of.chained.2012.dollars
- Preventable.hospital.stays.raw.value
- Hispanic.raw.value
- Long.commute…driving.alone.raw.value
- Air.pollution…particulate.matter.raw.value
- Driving.alone.to.work.raw.value
- Uninsured.raw.value (top wybór, wszystkie gwiazdki)
- Rural.raw.value
- Adult.obesity.raw.value
- Personal.income..thousands.of.dollars
- Population.
- Civilian_labor_force_2020
- Employed_2020
- Unemployment_rate_2020
- Med_HH_Income_Percent_of_State_Total_2020
- Drinking.water.violations.raw.value

Regresja Lasso



Driving.alone.to.work.raw.value.2016	3310.269
X..Females.raw.value.2017	3247.603
Adult.smoking.raw.value.2016	3116.371
Low.birthweight.raw.value.2017	2658.853
Insufficient.sleep.raw.value.2020	2154.412
Uninsured.children.raw.value.2016	2015.501
Food.insecurity.raw.value.2017	1883.588
X..Females.raw.value.2016	1838.178
X..Hispanic.raw.value.2020	1400.511
Physical.inactivity.raw.value.2019	1330.217
Excessive.drinking.raw.value.2018	1201.816
Mammography.screening.raw.value.2016	1178.545
Food.insecurity.raw.value.2019	1051.122
Severe.housing.problems.raw.value.2017	963.2346
Insufficient.sleep.raw.value.2016	950.3841
Insufficient.sleep.raw.value.2019	940.1676
Frequent.physical.distress.raw.value.2018	891.4655
Frequent.physical.distress.raw.value.2019	702.7024
Mammography.screening.raw.value.2017	680.97
Food.insecurity.raw.value.2016	624.7364
County.Ranked..Yes.1.No.0..2018	623.0218
X..below.18.years.of.age.raw.value.2016	615.6824
Mammography.screening.raw.value.2020	604.277
Uninsured.children.raw.value.2017	475.7803
Alcohol.impaired.driving.deaths.raw.value.2017	425.8148
Some.college.raw.value.2016	317.6833

Badania odchyleń dla „najlepszych” i „najgorszych” hrabstw

Największe emigracje

2020 Real GDP (thousands of chained 2012 dollars)	7.6424828527
2020 Personal income (thousands of dollars)	7.4366807076
Unemployed_2020	7.2677043858
2020 Population (persons) 1/	6.6413192911
2020 Population (persons) 1/.1	6.6413192911
Civilian_labor_force_2020	6.6117609586
Employed_2020	6.5141340291
Median_Household_Income_2020	1.9148449633
Unemployment_rate_2020	1.8459795649
% Asian raw value 2020	1.5440346007
Med_HH_Income_Percent_of_State_Total_2020	1.0102107314
Severe housing problems raw value 2020	0.810119597
Median household income raw value 2020	0.7787129713
Drinking water violations raw value 2020	0.7613720368
% not proficient in English raw value 2020	0.7056226622
2020 Chain-type quantity indexes for real GDP	0.6559157041
2020 Chain-type quantity indexes for real GDP .1	0.6559157041
Food environment index raw value 2020	0.5749230324
% Hispanic raw value 2020	0.5596783341
Excessive drinking raw value 2020	0.5476467673

Uninsured raw value 2020	-0.8548655522
Uninsured adults raw value 2020	-0.8194170493
Uninsured children raw value 2020	-0.7650381077
Injury deaths raw value 2020	-0.7555425645
% Rural raw value 2020	-0.7245067555
2020 HPI Change	-0.6941789197
Adult smoking raw value 2020	-0.6153505951
Premature death raw value 2020	-0.5382424294
Low birthweight raw value 2020	-0.5262072954
Food insecurity raw value 2020	-0.5240794607
Teen births raw value 2020	-0.5126418457
Premature age-adjusted mortality raw value 2020	-0.4879288302
Driving alone to work raw value 2020	-0.4453434244
Physical inactivity raw value 2020	-0.4274551518
Frequent physical distress raw value 2020	-0.3817235023
Limited access to healthy foods raw value 2020	-0.369388556
Children in poverty raw value 2020	-0.3591172056
Poor physical health days raw value 2020	-0.3485026067

Największe imigracje

2020 Population (persons) 1/	4.2527233614
2020 Population (persons) 1/.1	4.2527233614
Employed_2020	4.2402358084
Civilian_labor_force_2020	4.1821351394
Unemployed_2020	3.5393220137
2020 Personal income (thousands of dollars)	3.4807559196
2020 Real GDP (thousands of chained 2012 dollars)	2.9169561052
% Hispanic raw value 2020	1.8929390655
Uninsured raw value 2020	1.262289885
Uninsured adults raw value 2020	1.2575735817
Med_HH_Income_Percent_of_State_Total_2020	1.2295875121
Median_Household_Income_2020	1.1837113663
Uninsured children raw value 2020	1.1277196096
2020 Chain-type quantity indexes for real GDP	1.0379114846
2020 Chain-type quantity indexes for real GDP .1	1.0379114846
% not proficient in English raw value 2020	0.9333777519
Unemployment_rate_2020	0.6355418016

Mammography screening raw value 2020	-0.7463832526
% Females raw value 2020	-0.6441203911
% Rural raw value 2020	-0.5699002465
Air pollution - particulate matter raw value 2020	-0.5265986075
Adult obesity raw value 2020	-0.5002087521
Other primary care providers raw value 2020	-0.4772143895
Social associations raw value 2020	-0.4632659606
Adult smoking raw value 2020	-0.4454431036
Some college raw value 2020	-0.4087108331
Mental health providers raw value 2020	-0.3755705565
Diabetes prevalence raw value 2020	-0.3646963212
% 65 and older raw value 2020	-0.3107064783
Physical inactivity raw value 2020	-0.276425076
Dentists raw value 2020	-0.2485090836
Poor mental health days raw value 2020	-0.2315036663
Premature death raw value 2020	-0.2035777614
Frequent mental distress raw value 2020	-0.1986087024
Food insecurity raw value 2020	-0.1950033153

Wnioski końcowe

Co udało się zrobić w projekcie?