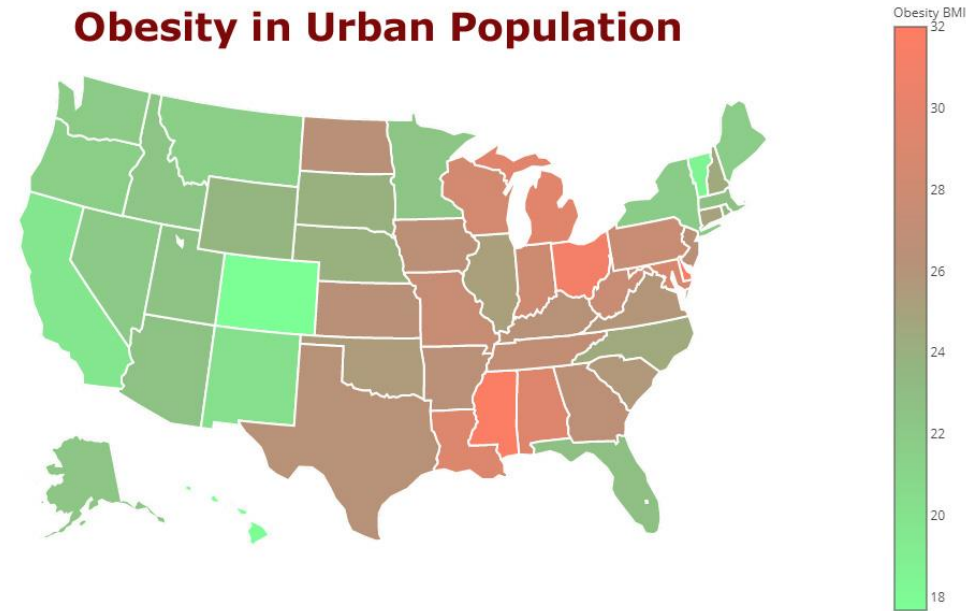


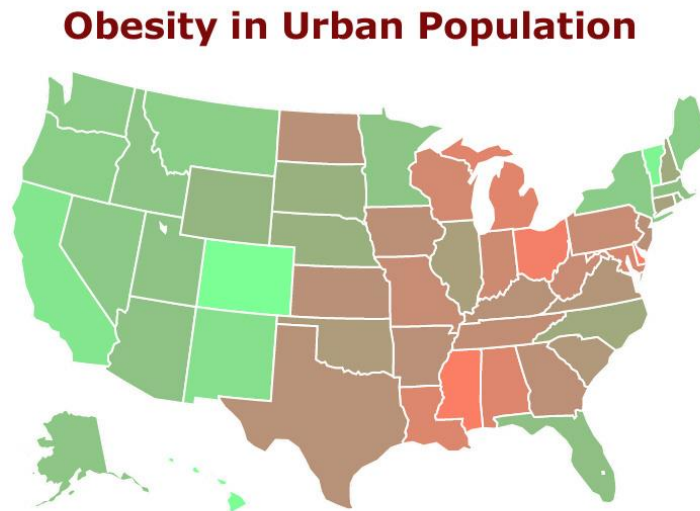
Correlation Study and Linear Classifier Performance Analysis for Obesity from CDC and Census Data Merge for Urban Population



Soumyendu Sarkar

States with Lowest and Highest Obesity in Cities

Rank	State	State	BMI
1	CO	Colorado	17.7
2	HI	Hawaii	17.7
3	VT	Vermont	18.5
4	CA	California	19.6
5	NM	New Mexico	20.2
6	MT	Montana	21.6
7	OR	Oregon	21.6
8	ME	Maine	21.8
9	NY	New York	21.8
10	WA	Washington	21.9
11	NV	Nevada	22.0
12	AK	Alaska	22.3
13	MN	Minnesota	22.3
14	ID	Idaho	22.3
15	UT	Utah	22.6
16	AZ	Arizona	22.7
17	FL	Florida	22.8
18	MA	Massachusetts	22.8
19	RI	Rhode Island	23.5
20	WY	Wyoming	23.7
21	NE	Nebraska	23.9
22	SD	South Dakota	24.0
23	NH	New Hampshire	24.5
24	NC	North Carolina	24.5
25	CT	Connecticut	25.0



Rank	State	State	BMI
1	DE	Delaware	32.0
2	MS	Mississippi	31.7
3	OH	Ohio	31.1
4	MI	Michigan	29.6
5	AL	Alabama	29.3
6	LA	Louisiana	29.2
7	MD	Maryland	28.8
8	WI	Wisconsin	28.4
9	IN	Indiana	27.9
10	PA	Pennsylvania	27.5
11	WV	West Virginia	27.5
12	MO	Missouri	27.4
13	TN	Tennessee	27.0
14	GA	Georgia	26.6
15	IA	Iowa	26.5
16	NJ	New Jersey	26.4
17	AR	Arkansas	26.3
18	ND	North Dakota	26.3
19	KS	Kansas	26.3
20	TX	Texas	26.2
21	KY	Kentucky	26.2
22	VA	Virginia	25.9
23	SC	South Carolina	25.7
24	OK	Oklahoma	25.5
25	IL	Illinois	25.2

Healthy Cities with Least Obesity

Rank	City	State	BMI	Rank	City	State	BMI	Rank	City	State	BMI	Rank	City	State	BMI
1	Milpitas	CA	12.2	21	Lake Forest	CA	15.6	41	Thousand Oaks	CA	16.9	61	Folsom	CA	17.9
2	Fremont	CA	12.9	22	Santa Fe	NM	15.6	42	Carlsbad	CA	17	62	Plymouth	MN	17.9
3	Irvine	CA	13.4	23	Huntington Beach	CA	15.8	43	Orange	CA	17	63	Glendale	CA	18
4	Union City	CA	13.6	24	Loveland	CO	15.8	44	Simi Valley	CA	17	64	San Diego	CA	18
5	Daly City	CA	13.9	25	Arvada	CO	15.9	45	Sparks	NV	17.3	65	Santa Clarita	CA	18
6	Westminster	CA	13.9	26	Buena Park	CA	15.9	46	West Covina	CA	17.3	66	Arlington Heights	IL	18.1
7	Alhambra	CA	14.2	27	Denver	CO	15.9	47	Boca Raton	FL	17.4	67	Newton	MA	18.2
8	Boulder	CO	14.3	28	Mountain View	CA	16	48	San Mateo	CA	17.4	68	Roseville	CA	18.2
9	Longmont	CO	14.3	29	Bellevue	WA	16.3	49	Schaumburg	IL	17.4	69	Santa Rosa	CA	18.2
10	Sunnyvale	CA	14.5	30	Centennial	CO	16.3	50	Lakewood	CA	17.5	70	Cary	NC	18.3
11	Torrance	CA	14.5	31	Fullerton	CA	16.3	51	Burbank	CA	17.7	71	Rancho Cucamonga	CA	18.3
12	Pleasanton	CA	14.8	32	Tustin	CA	16.3	52	Costa Mesa	CA	17.7	72	Berkeley	CA	18.4
13	Garden Grove	CA	15	33	San Jose	CA	16.4	53	Honolulu	HI	17.7	73	Hayward	CA	18.4
14	Chino Hills	CA	15.1	34	Fort Collins	CO	16.5	54	Sugar Land	TX	17.7	74	Burlington	VT	18.5
15	Santa Clara	CA	15.1	35	Redondo Beach	CA	16.5	55	Westminster	CO	17.7	75	Chula Vista	CA	18.6
16	Mission Viejo	CA	15.5	36	Johns Creek	GA	16.6	56	Quincy	MA	17.8	76	Clovis	CA	18.6
17	Newport Beach	CA	15.5	37	Lakewood	CO	16.6	57	Reno	NV	17.8	77	San Marcos	CA	18.6
18	San Francisco	CA	15.5	38	San Leandro	CA	16.7	58	Santa Monica	CA	17.8	78	Colorado Springs	CO	18.7
19	Alameda	CA	15.6	39	Livermore	CA	16.9	59	Anaheim	CA	17.9	79	Redlands	CA	18.7
20	Lake Forest	CA	15.6	40	San Ramon	CA	16.9	60	Elk Grove	CA	17.9	80	San Buenaventura	CA	18.7

Cities with Highest Obesity

Rank	City	State	BMI	Rank	City	State	BMI	Rank	City	State	BMI	Rank	City	State	BMI
1	Dayton	OH	38.8	21	Pharr	TX	30.9	41	Hammond	IN	29.5	61	McAllen	TX	28.6
2	Gary	IN	38.2	22	Mobile	AL	30.8	42	Kenner	LA	29.5	62	Winston-Salem	NC	28.6
3	Detroit	MI	36.4	23	Southfield	MI	30.8	43	Newark	NJ	29.5	63	Columbia	SC	28.5
4	Youngstown	OH	35.1	24	Hampton	VA	30.7	44	Buffalo	NY	29.4	64	Kansas City	MO	28.5
5	Macon	GA	34.7	25	Trenton	NJ	30.7	45	Racine	WI	29.4	65	Longview	TX	28.5
6	Kalamazoo	MI	34.5	26	Canton	OH	30.6	46	Lansing	MI	29.3	66	Brockton	MA	28.4
7	Flint	MI	34.1	27	Hartford	CT	30.4	47	Memphis	TN	29.3	67	Chattanooga	TN	28.4
8	Cleveland	OH	33.8	28	St. Louis	MO	30.4	48	Muncie	IN	29.3	68	Mission	TX	28.4
9	Jackson	MS	33.3	29	Baton Rouge	LA	30.3	49	Edinburg	TX	29.2	69	Fort Smith	AR	28.3
10	Toledo	OH	33.2	30	Kansas City	KS	30.3	50	Savannah	GA	29.1	70	Indianapolis	IN	28.3
11	Camden	NJ	33.1	31	Laredo	TX	30.3	51	Springfield	MO	29.1	71	Kenosha	WI	28.3
12	Reading	PA	32.6	32	Beaumont	TX	30.2	52	Paterson	NJ	29	72	Lawton	OK	28.3
13	Albany	GA	32.4	33	Montgomery	AL	30.2	53	Waco	TX	29	73	New Orleans	LA	28.3
14	Wilmington	DE	32	34	Rockford	IL	30.1	54	Allentown	PA	28.8	74	Topeka	KS	28.3
15	Tuscaloosa	AL	31.8	35	South Bend	IN	30.1	55	Augusta	GA	28.8	75	Decatur	IL	28.2
16	Birmingham	AL	31.7	36	Akron	OH	30	56	Baltimore	MD	28.8	76	El Paso	TX	28.2
17	Milwaukee	WI	31.6	37	Cincinnati	OH	30	57	Columbus	OH	28.8	77	North Charleston	SC	28.2
18	Brownsville	TX	31.5	38	Shreveport	LA	30	58	New Haven	CT	28.8	78	Sioux City	IA	28.2
19	Rochester	NY	31.3	39	Corpus Christi	TX	29.9	59	Erie	PA	28.7	79	Tyler	TX	28.2
20	Portsmouth	VA	31.1	40	Lake Charles	LA	29.7	60	Grand Rapids	MI	28.7	80	Fayetteville	NC	28.1

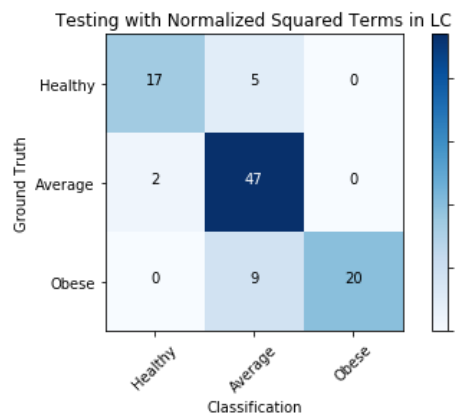
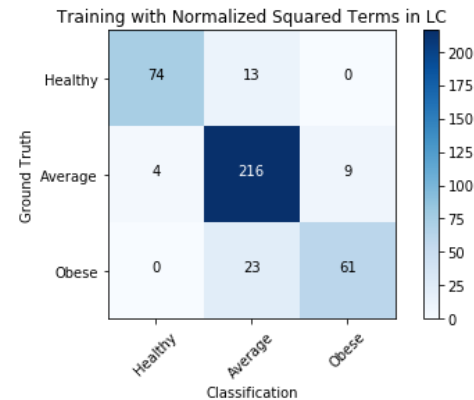
Linear Classifier for Obesity Classification

Objective

- Use Linear Classification and Principal Component Analysis to implement a Machine Learning Model to Classify Obesity from CDC data for 500 cities and Census Tract & County data
- Research on Variation of Classification Performance with variation of feature set
- Research on Classification Performance with additional non-linear terms in Linear Classification

Methods

- Merge Census Tract County data with CDC 500 cities data with population weights to augment feature set
- Correlation Analysis of Features from Census data with Obesity and include Top 30 Features for classification
- Obtain 400 Training & 100 Test Feature vectors with shuffling from 500 cities database to remove bias
- Implement Linear Classifier with Squared and Cubed Non-Linear terms to enhanced accuracy and use data normalization to optimize performance
- Implement Principal Component Analysis for Obesity Classification



BMI Class Definition:

Healthy : BMI < 20
 Average : 20 >= BMI < 27
 Obese : BMI >= 27

Linear Classifier with Normalized SQUARED Terms with 30 features

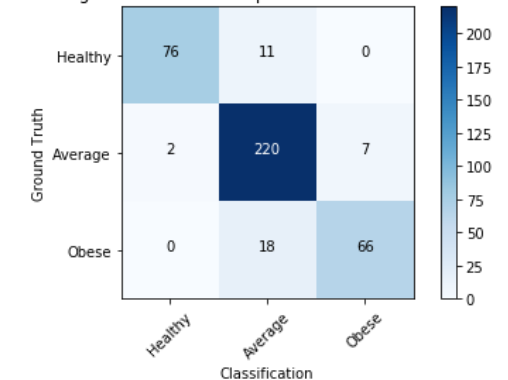
	Training	Testing
Accuracy	0.88	0.84
Highest PPV	0.95	1.00
Lowest PPV	0.86	0.77
PPV for Healthy Classification	0.95	0.89
PPV for Average Classification	0.86	0.77
PPV for Obese Classification	0.87	1.00

Over-Fitting ?

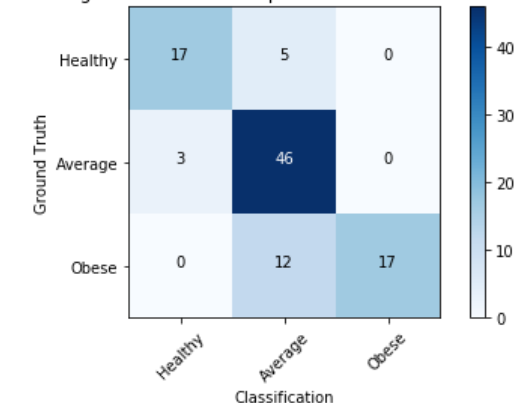
Linear Classifier with Normalized CUBED and Squared Terms with 30 features

	Training	Testing
Accuracy	0.91	0.80
Highest PPV	1.00	1.00
Lowest PPV	0.88	0.73
PPV for Healthy Classification	0.97	0.85
PPV for Average Classification	0.88	0.73
PPV for Obese Classification	0.90	1.00

Training with Normalized Squared and Cubed Terms in LC



Testing with Normalized Squared and Cubed Terms in LC



CDC & Census Data Correlation Analysis with Obesity

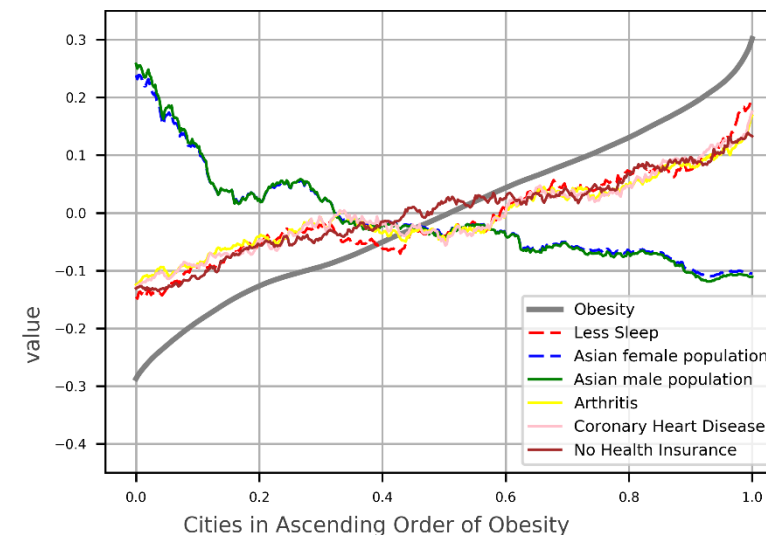
Very Highly Correlated



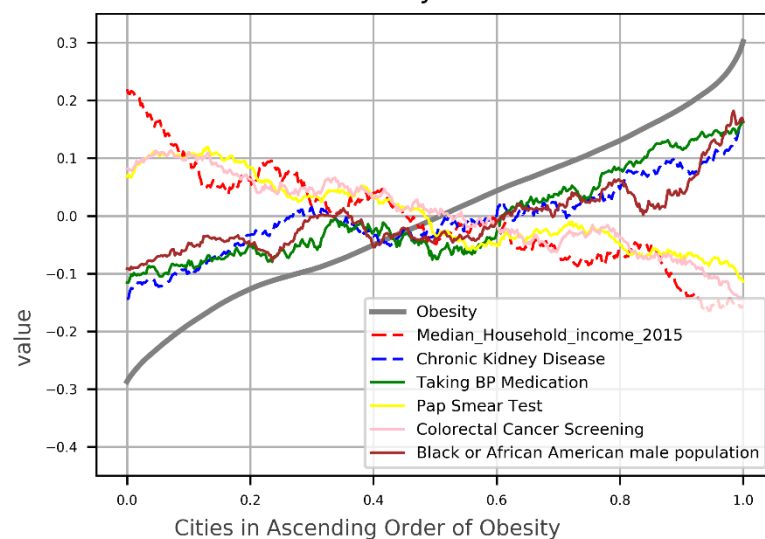
Highly Correlated



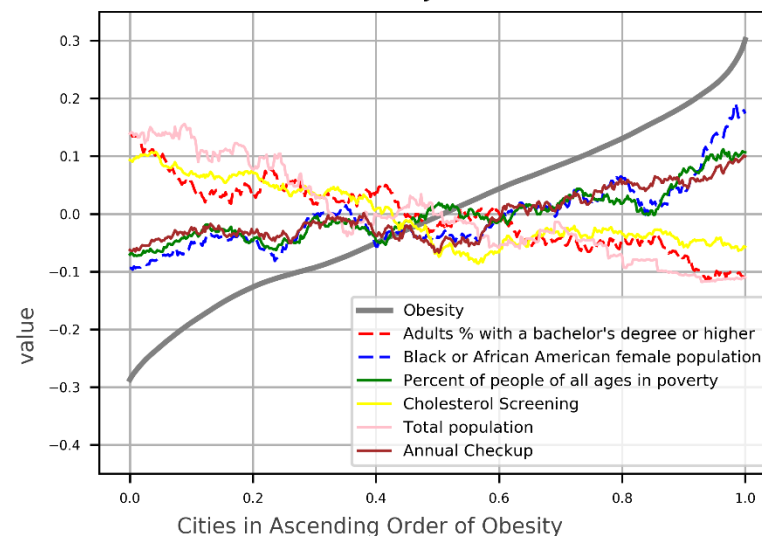
Well Correlated



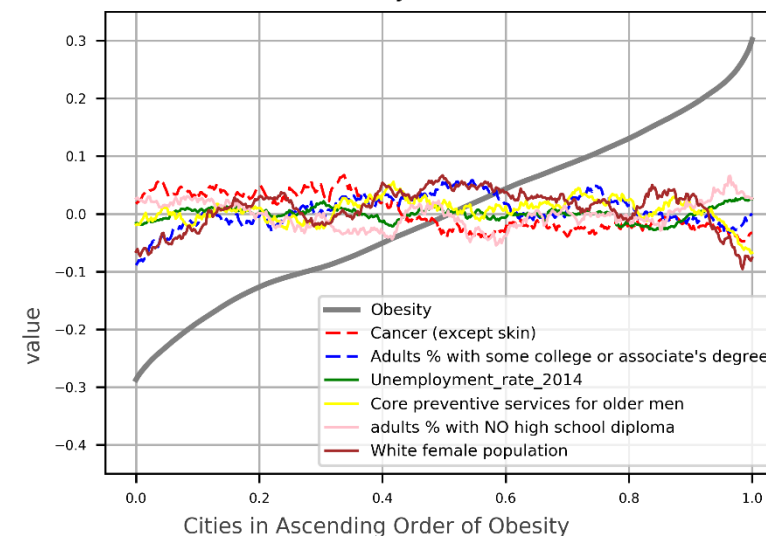
Moderately Correlated



Moderately Correlated



Poorly Correlated



Feature Correlation with Obesity

Rank	FEATURE	CORRELATION	Rank	FEATURE	CORRELATION
1	Teeth Loss	0.8136687437263442	25	Adults % bachelor's degree or higher	-0.37409617528765426
2	Smoking	0.7782953004281681	26	African American female population	0.37191965521328857
3	Low Physical Activity	0.7088155129076699	27	Percent of people in poverty	0.3375526128971053
4	Stroke	0.6884054217581411	28	Cholesterol Screening	-0.3372678541648043
5	COPD	0.6520786256266082	29	Total population	-0.3245786518713684
6	Dental Visit	-0.651864343360142	30	Annual Checkup	0.29761058387099665
7	High Blood Pressure	0.6514067957254205	31	Mammography	-0.2854508003727619
8	Asthma	0.6476837471816237	32	% of People in poverty	-0.2701119883267754
9	Diabetes	0.6450348353789633	33	High Cholesterol	0.25805175528402285
10	Poor Mental Health	0.6341907564608653	34	Pacific Islander Female population	-0.25486328530053304
11	Poor Physical Health	0.6029283356020563	35	Pacific Islander Male population	-0.2475011132522548
12	Adults % with high school diploma	0.5760699676500974	36	Binge Drinking	-0.2354599926965364
13	Less Sleep	0.5680532130394792	37	Total male population	-0.22279697497623033
14	Asian female population	-0.5633977183347055	38	Total female population	0.2227969749762296
15	Asian male population	-0.5517973396661352	39	Preventive services for older women	-0.217631636493823
16	Arthritis	0.5202128697892304	40	American Indian and Alaskan Male population	-0.1733907182844397
17	Coronary Heart Disease	0.5150142669268579	41	American Indian and Alaskan Female population	-0.16113420140074838
18	No Health Insurance	0.5044810771017083	42	Cancer (except skin)	-0.1259704222265354
19	Median_Household_income_2015	-0.49924542549151063	43	Adults % with college or associate's degree	0.11066137961902359
20	Chronic Kidney Disease	0.49136753647522907	44	Unemployment_rate_2014	0.0722820170201452
21	Taking BP Medication	0.4835522676520599	45	Preventive services for older men	-0.03667305959706792
22	Pap Smear Test	-0.40802112000662577	46	adults % with NO high school diploma	0.0029059468315383538
23	Colorectal Cancer Screening	-0.40387566366546335	47	White female population	0.002221613640036274
24	African American male population	0.3782030479176095	48	White male population	-0.0015587877326375108




Linear Classifier (Training)

The number of features were varied from 2 to 30 for the Classifier
Both Square and Cube Non-Linear functions were added in addition to feature terms

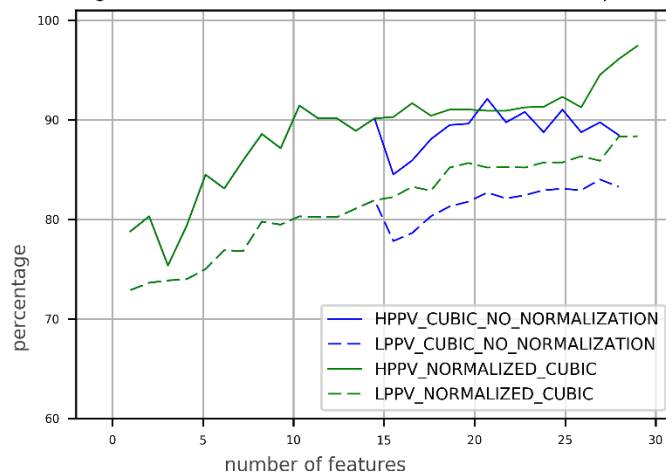
How **TRAINING Accuracy** and **PPV** varies with ?

- Number of Features
- Non-Linear Functions

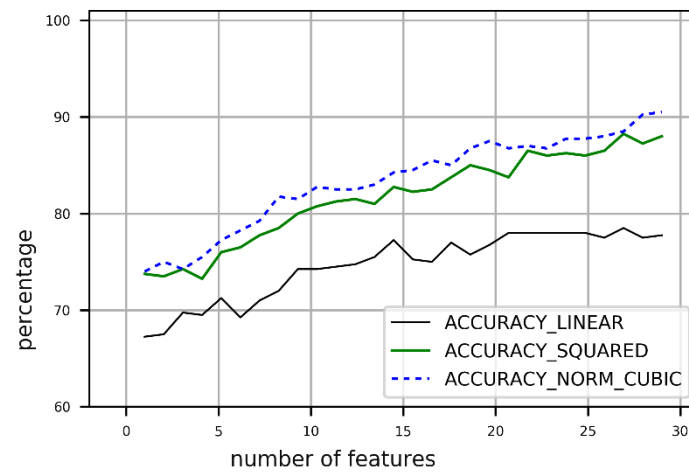
Observations

- Accuracy and PPV goes  up with inclusions of more **FEATURES**
- Accuracy and PPV goes  up with inclusions of **non-linear SQUARED** terms of Feature values
- Accuracy and PPV goes  up FURTHER with inclusions of **non-linear CUBED** and SQUARED terms of Feature values
- Normalization of features improves performance when non-linear functions specially Cubes and also Squares are included in Linear Classifier

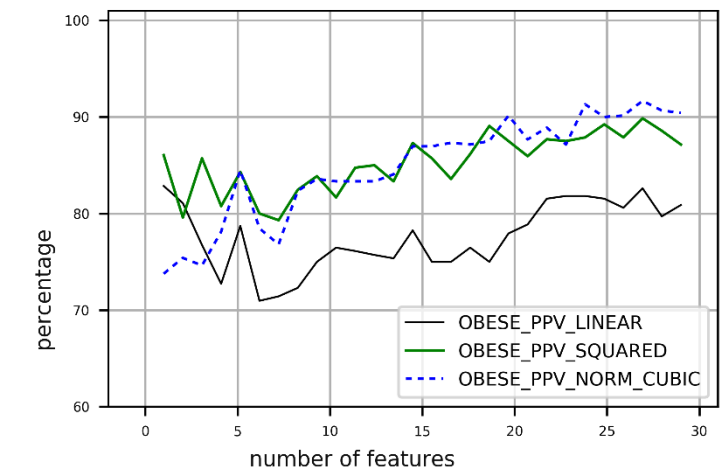
Training PPV: Normalized vs Non-Norm Cubic & Sq Data in LC



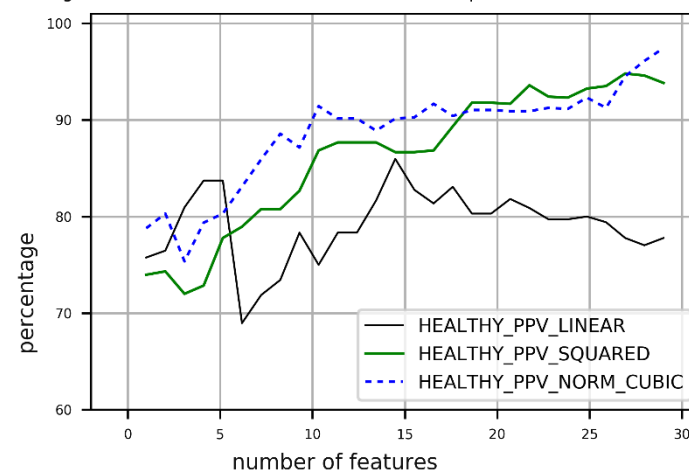
Training ACCURACY: Linear VS Squared VS Normalized Cubic Terms



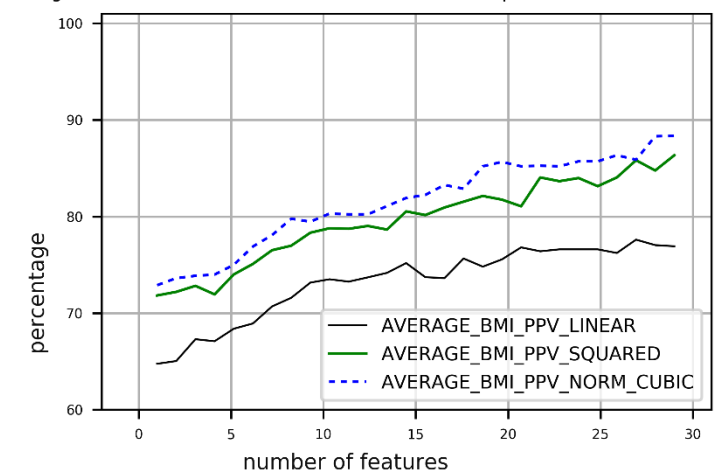
Training: OBESE Classification PPV: Linear VS Squared VS Normalized Cubic Terms



Training: HEALTHY Classification PPV: Linear VS Squared VS Normalized Cubic Terms



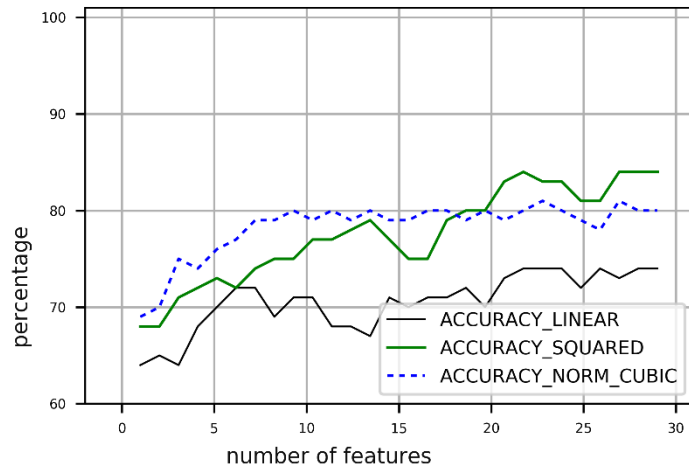
Training: AVERAGE BMI Classification PPV: Linear VS Squared VS Normalized Cubic Terms



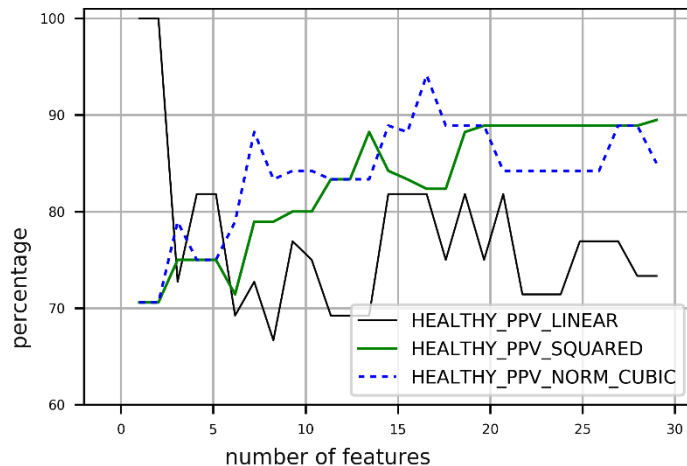
Linear Classifier (Testing)

The number of features were varied from 2 to 30 for the Classifier
Both Square and Cube Non-Linear functions were added in addition to feature terms

Testing ACCURACY: Linear VS Squared VS Normalized Cubic Terms in Linear Classifier



Testing: HEALTHY Classification PPV: Linear VS Squared VS Normalized Cubic Terms



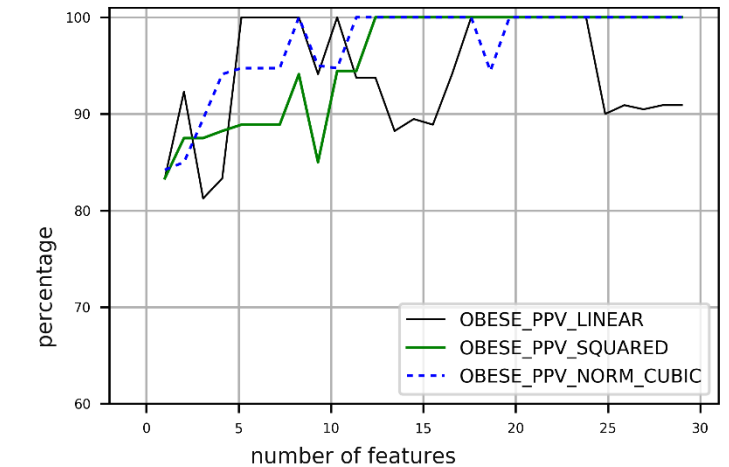
How **TESTING Accuracy** and **PPV** varies with ?

- Number of Features
- Non-Linear Functions

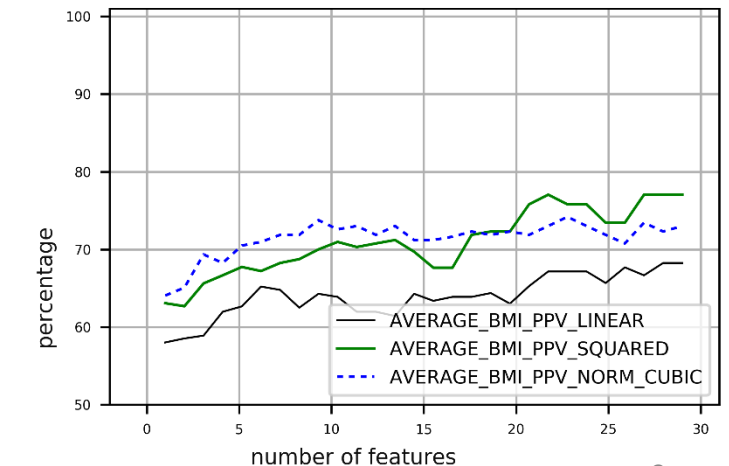
Observations

- Accuracy and PPV goes **up** with inclusions of more **FEATURES**
- Accuracy and PPV goes **up** with inclusions of **non-linear SQUARED** terms of Feature values
- Accuracy and PPV goes **up FURTHER** with inclusions of **non-linear CUBED** and SQUARED terms with lower number of Features, but then goes lower than that of LC with SQUARED terms with inclusion of more features. This is likely due to OVER-FITTING
- Normalization of features improves performance when non-linear functions specially Cubes and also Squares are included in Linear Classifier

Testing: OBESE Classification PPV: Linear VS Squared VS Normalized Cubic Terms



Testing: AVERAGE BMI Classification PPV: Linear VS Squared VS Normalized Cubic Terms

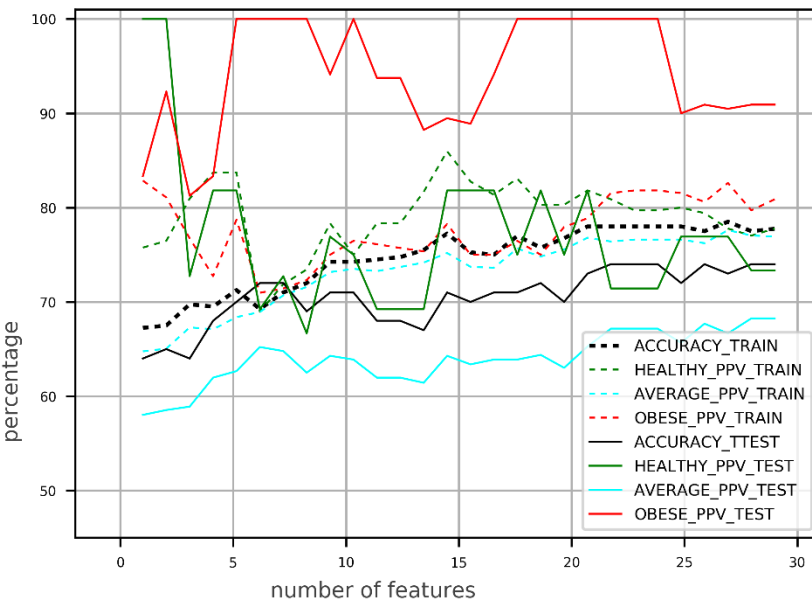


Linear Classifier

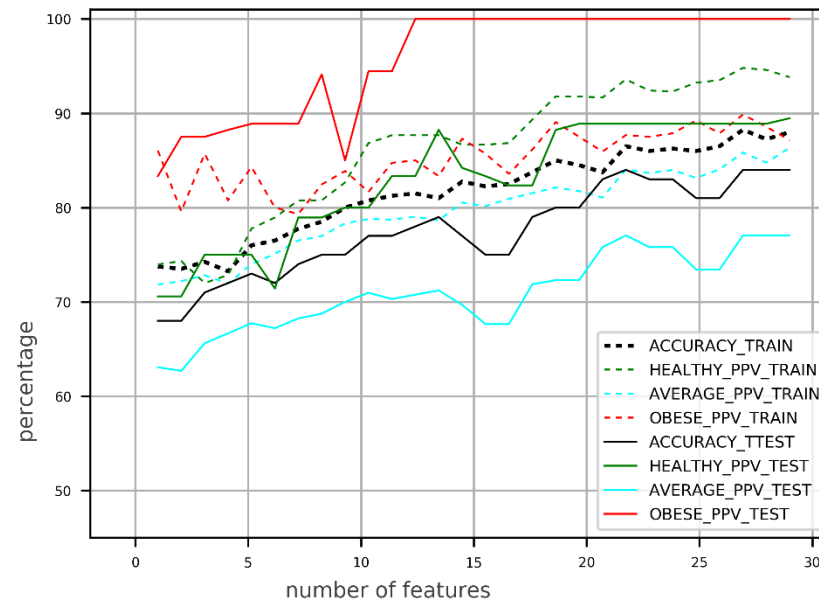
Testing VS Training Classification Performance

- Accuracy of classification of Testing is less than that of Training but mostly within 5% for all three classifiers for the range of number of features
- Obesity classification PPV is more for Testing than that of Training and reaches 100% when more number of features are included specially with non-linear Squared and Cubed terms
- Average BMI classification PPV is less for Testing than that of Training
- Even though classification performance for test data is significantly better with non-linear Squared terms in the Linear Classifier, inclusion of Cubed terms does not significantly help classification with Test data performance

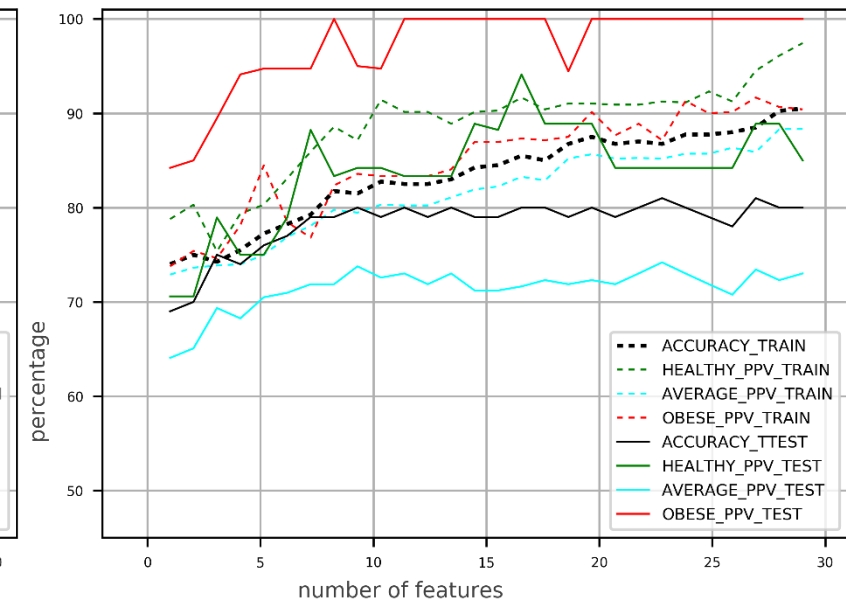
Testing VS Training: Linear Terms in Linear Classifier



Testing VS Training: SQUARE Terms in Linear Classifier

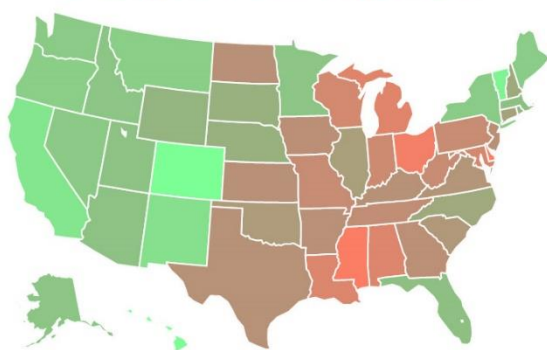


Testing VS Training: NORMALIZED CUBIC Terms in LC



Distribution of Urban Health Factors by States

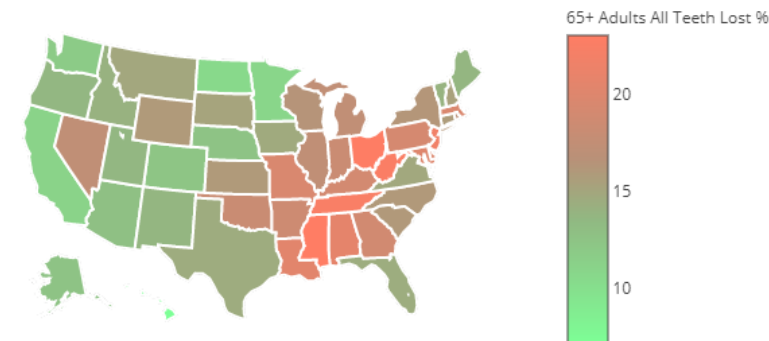
Obesity in Urban Population



Very Highly Correlated

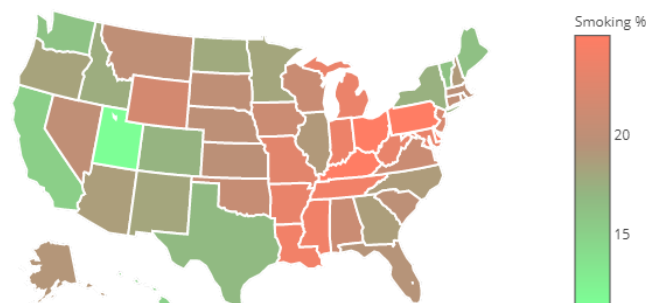


2014 US 65+ Adult Total Teeth Lost % in Urban population by State
(Hover for breakdown)



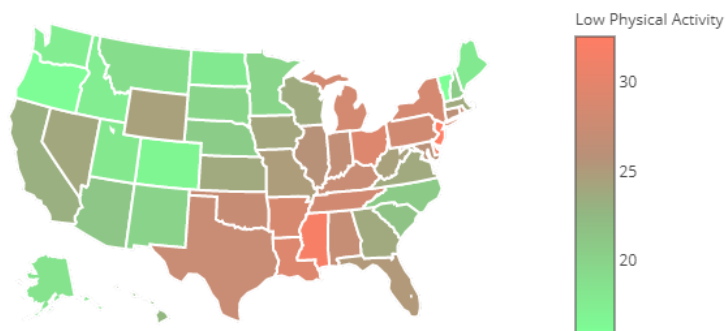
100% Adult Teeth Loss for 65 years old or older

2014 US Smoking in Urban population by State in %
(Hover for breakdown)



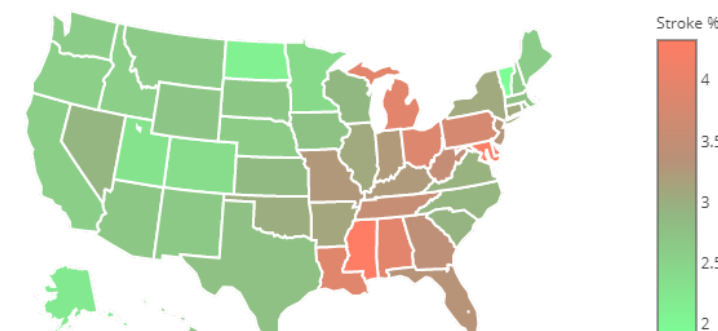
Smoking

2014 US Low Physical Activity in Urban population by State
(Hover for breakdown)



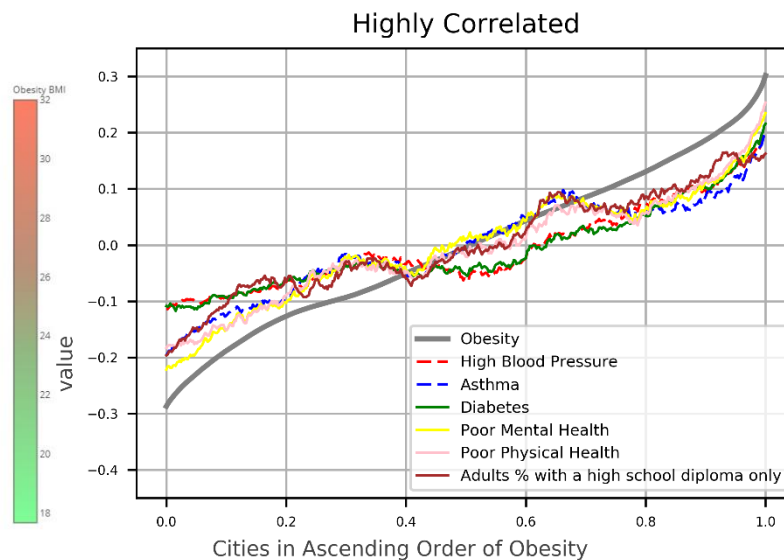
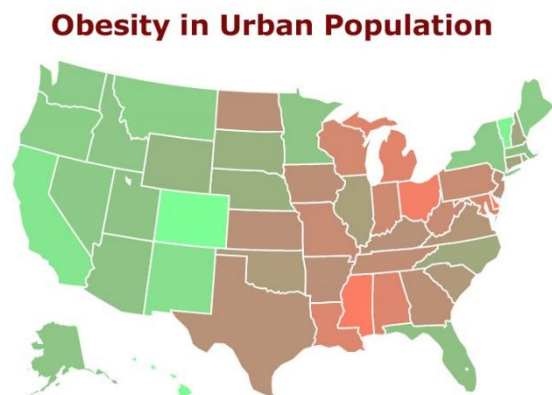
Low Physical Activity

2014 US Stroke in Urban population by State
(Hover for breakdown)

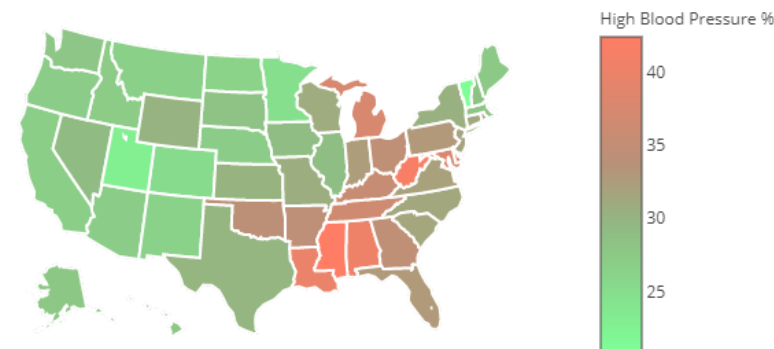


Stroke

Distribution of Urban Health Factors by States

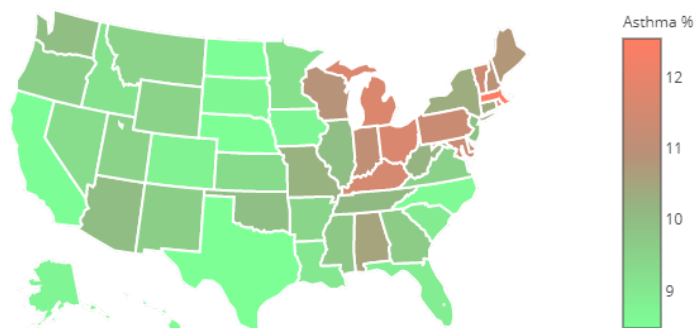


2014 US High Blood Pressure % in Urban population by State
(Hover for breakdown)



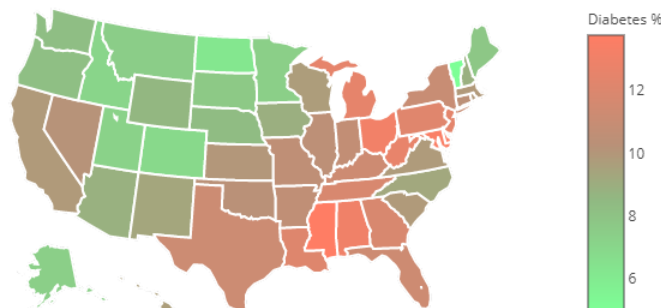
High Blood Pressure

2014 US Asthma in Urban population by State
(Hover for breakdown)



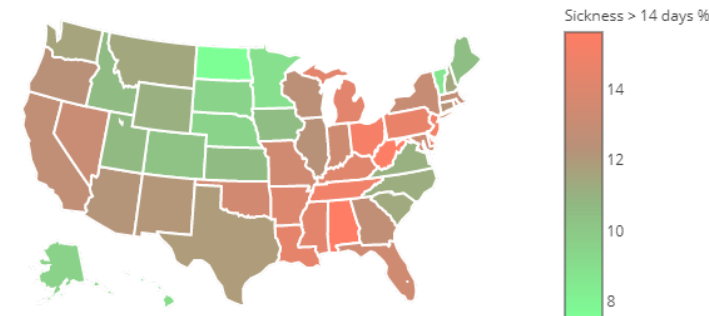
Asthma

2014 US Diabetes in Urban population by State
(Hover for breakdown)



Diabetes

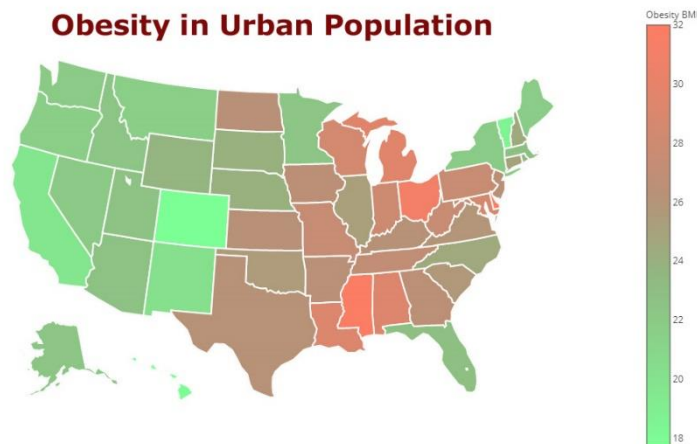
2014 US Poor Physical Health in Urban population by State
(Hover for breakdown)



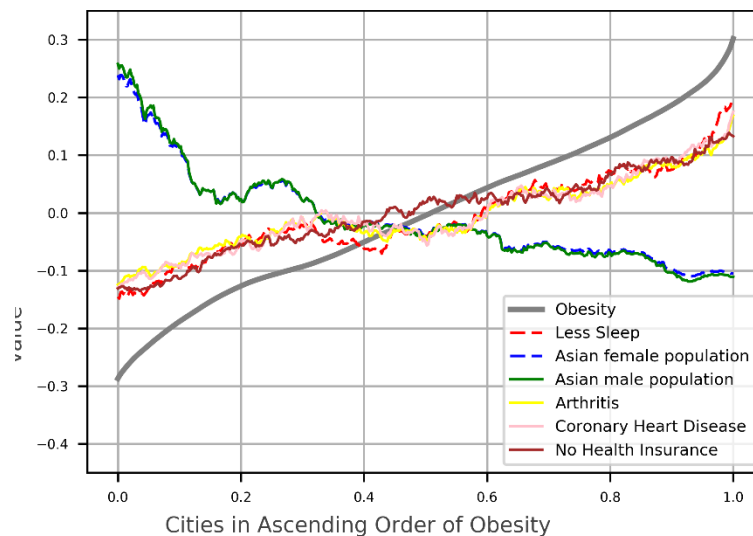
Poor Physical Health

Distribution of Urban Health Factors by States

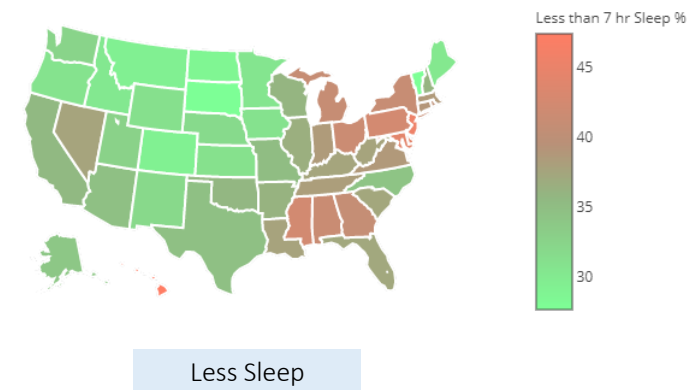
Obesity in Urban Population



Well Correlated

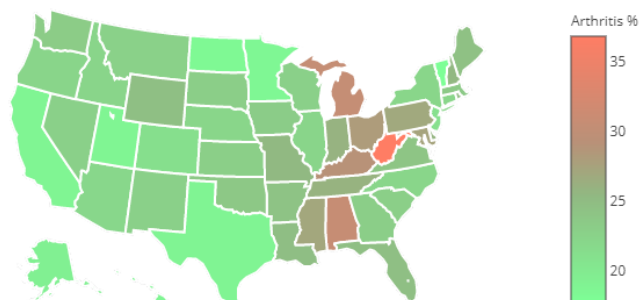


2014 US Less Sleep % in Urban population by State (Hover for breakdown)



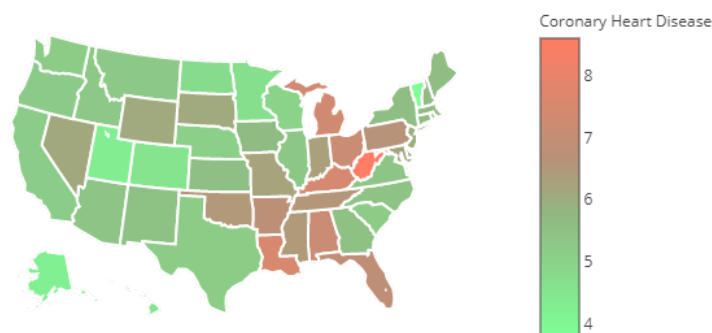
Less Sleep

2014 US Arthritis in Urban population by State (Hover for breakdown)



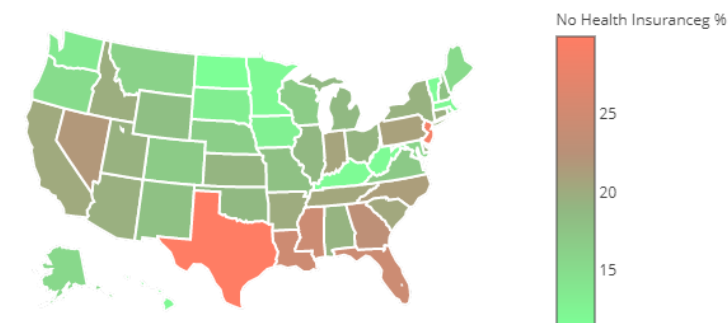
Arthritis

2014 US Coronary Heart Disease in Urban population by State (Hover for breakdown)



Coronary Heart Disease

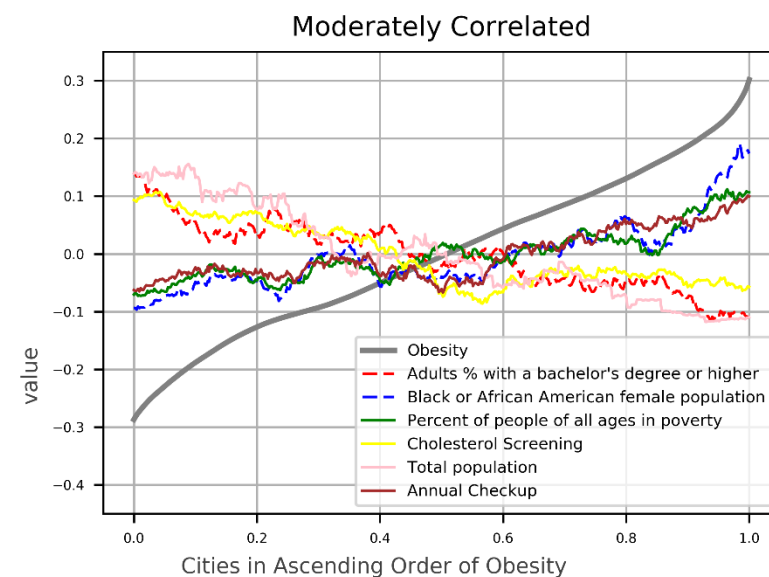
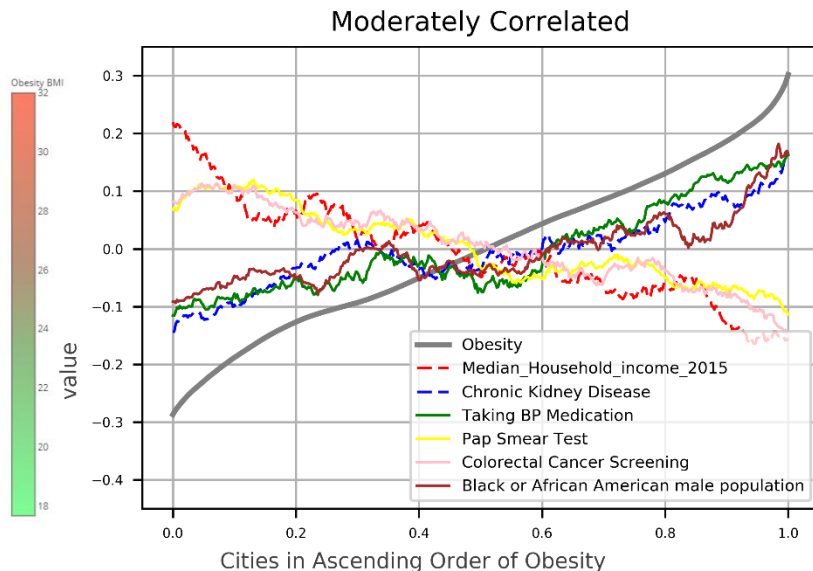
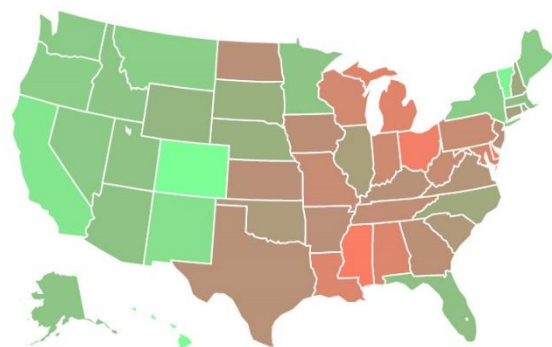
2014 US No Health Insurance in Urban population by State (Hover for breakdown)



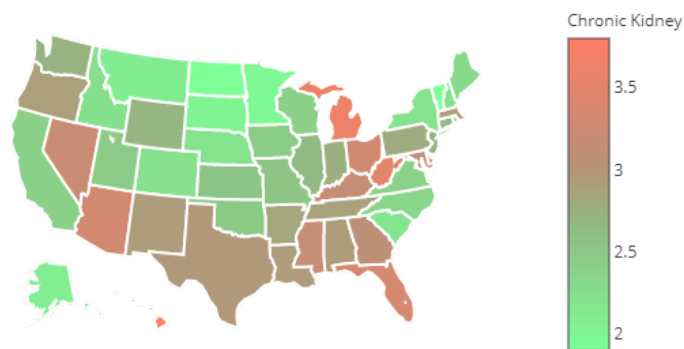
No Health Insurance

Distribution of Urban Health Factors by States

Obesity in Urban Population

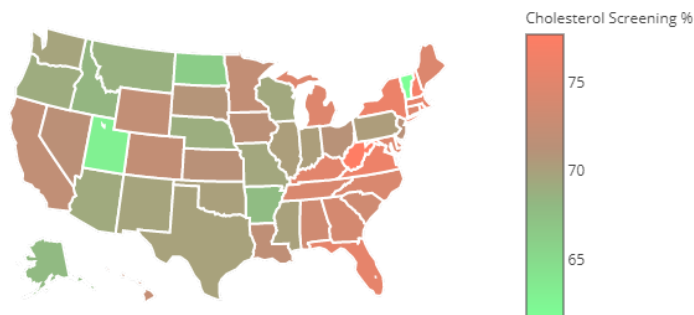


2014 US Chronic Kidney Disease in Urban population by State
(Hover for breakdown)



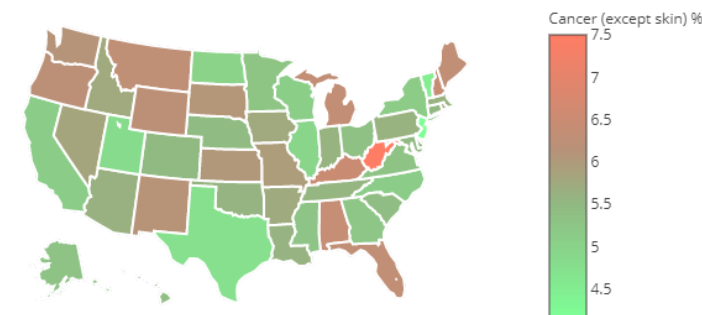
Chronic Kidney Disorder

2014 US Cholesterol Screening in Urban population by State
(Hover for breakdown)



Cholesterol Screening

2014 US Cancer in Urban population by State in %
(Hover for breakdown)



Cancer – Poorly Correlated

References & Acknowledgements

- Overweight & Obesity - <https://www.cdc.gov/obesity/data/adult.html>
- 500 Cities: Local Data for Better Health - <https://www.cdc.gov/500cities/>
- Chronic Disease and Health Promotion Data & Indicators - <https://chronicdata.cdc.gov/health-area/nutrition-physicalactivity-obesity>
- Population Estimates, Unemployment, Education - <https://www.census.gov/data.html>
- **Special thanks to Hakan Egeli** for identifying and collecting some of the CDC and Census Data