Are we what we tweet?

Final Capstone

Nadine Ruecker

April 15th 2019

Obesity

Body-mass-index >30kg/m^2

Overweight: BMI 25-30

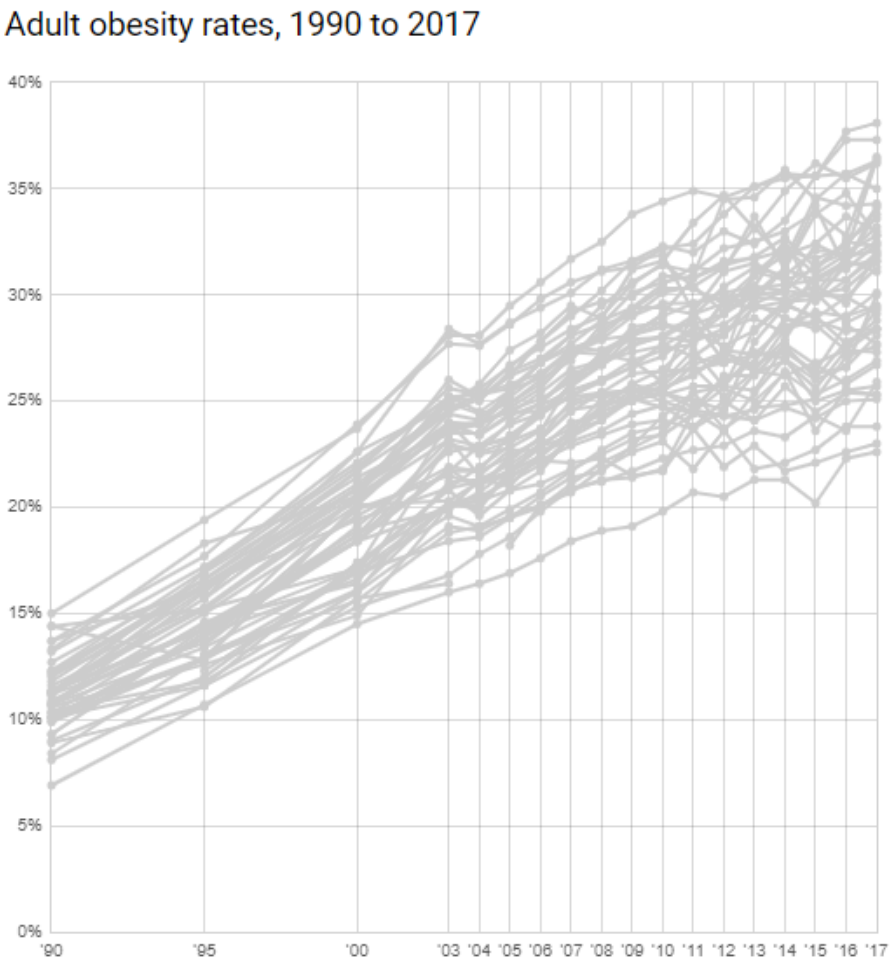
Medical condition, associated with:

- cardiovascular disease
- type-2-diabetes
- some forms of cancer
- osteoarthritis
- mental illness

Huge burden for the medical system

and the economy.

Reasons: Genetics, behavior and society



- https://www.stateofobesity.org/adult-obesity/

Hypothesis

Can twitter data predict obesity rates and associated health indicators?

Can NLP determine a healthy from an unhealthy tweet?

Data – Search terms

	healthy	unhealthy
Restaurants 15 vs. 19	Panerabread,jasonsdeli,aubonpain,Noodles and Company, Chipotle, AtlantaBread, EinsteinBros, LePainQuotidien, Justsalad,Mymarthas,krunch,chopt, sweetgreen, cava,olivegarden	kfc,tacobell,burgerking,cinnabon,chickfila, PandaExpress,dunkindonuts,pizzahut,Waffle House, cinnabon, AutieAnnes, cheesecake,arbys,wendys,FiveGuys,Shake shack,WhiteCastle,DairyQueen,quiznos
Activities 44 vs. 38	run,running,ran,walk,walking,walked,hike, hiking,hiked,surf,surfing,yoga,exercise, climb,climbed,soccer,tennis,volleyball,baseball,softball,swim,swimming,swam,dance,ballet,mountainbike,marathon,triathlon,boxing,kickboxing,gymnastics,ski,skiing,snowboarding,snowboard,kanu,kayak,row,rowing,sail,sailing,sailed, body building, spinning, cardio	couch,sofa,nap,sleep,TV,watch,watching,watched,HBO,Netflix,binge watch,binge watched,HULU,Amazon Video,season,primevideo,television,slinc,CBS,philo,fuboTV,direct TV,Youtube TV,Youtube,playstation, xbox,wii,ESPN,Showtime,ABC,Starz,Fox,binge,Pluto TV, lazy,cozy,blanket,pillow
Food 66 vs. 28	banana,blackberries,blueberries,cherry,coconut,cranberry,date,fig,goji, grape,grapefruit,kiwi,lemon,lime,lyche,mango,melon,watermelon,nectarine, orange,papaya,passionfruit,peach,pear,plum,pineapple,pomegranate,raspberry, star fruit,strawberry,cantaloupe,artichoke,asparagus,beans,legumes,broccoli, brussels sprouts,cabbage,cauliflower,celery,endives,fennel,kale,spinach,lettuce,salad,mushrooms,okra,garlic,chives,beetroot,beets,ginger,radish,squash,tomato	lemonade,coke,soda,sprite,pepsi,pizza,fries,burger,cheeseburger,cheese,cream,sauce,cupcake,cake,cookie,donut,chips,syrup,candy,fudge,pie,pudding,brownie, all you can eat,frozen yoghurt,chicken nuggets,waffle,pancake

For each query collect a sample of 2000 tweets.

Data – Scrapping twitter


```
[{
  "created_at": "Thu Jun 22 21:00:00 +0000 2017",
  "id": 877994604561387500,
  "id_str": "877994604561387520",
  "text": "Creating a Grocery List Manager Using Angular, Part 1: Add &amp;amp; Display Items https://t.co/xFox78juL1 #Angular",
  "truncated": false,
  "entities": {
    "hashtags": [{
      "text": "Angular",
      "indices": [103, 111]
    }],
    "symbols": [],
    "user_mentions": [],
    "urls": [{
      "url": "https://t.co/xFox78juL1",
      "expanded_url": "http://buff.ly/2sr60pf",
      "display_url": "buff.ly/2sr60pf",
      "indices": [79, 102]
    }]
  },
  "source": "<a href=\"http://bufferapp.com\" rel=\"nofollow\">Buffer</a>",
  "user": {
    "id": 772682964,
    "id_str": "772682964",
    "name": "SitePoint JavaScript",
    "screen_name": "SitePointJS",
    "location": "Melbourne, Australia",
    "description": "Keep up with JavaScript tutorials, tips, tricks and articles at SitePoint.",
    "url": "http://t.co/cCH13gqeUK",
    "entities": {
```

Data – Scrapping twitter - Constraints

	Standard API
Rate	2000 / h
Li	
Tw	
U	
Historic tweets	no
Spe	
Cos	
Main Problem	No historic tweets

200 queries * 4 years * 2000 tweets

⇒ 800.000 tweets

But only 500 tweets per request

200 queries * 4 year * 2000 tweets = 1.6 M tweets

1.600.000/2000= 800h to look up user data

=> 33 days

Data – Locate Tweet to US State

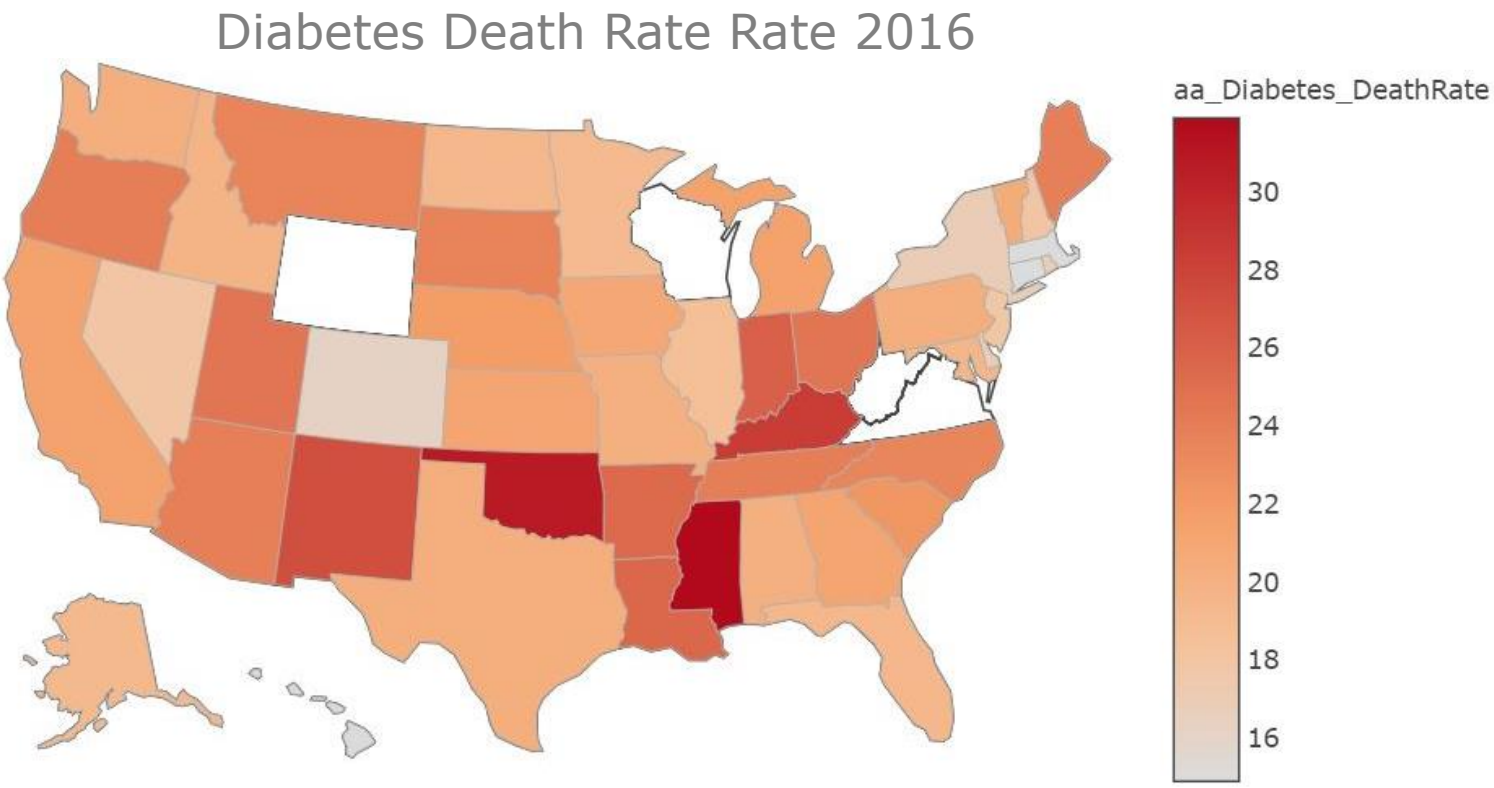
	Coordinat	Date	Location		Place	Query	ScreenNa	Text	TweetId	UserId	Cat
0		Wed Mar				fruit	Sofie_Lov	ðŸ˜ðŸ˜Introducing Soulfood	1.11E+18	4.49E+08	healthy
1		Wed Mar				fruit	marvinbry	If a fruit has to have seeds, what	1.11E+18	3.66E+08	healthy
2		Wed Mar				fruit	intemittn	colleges (Yale, etc.) because	1.11E+18	3.3E+08	healthy
3		Wed Mar				fruit	fruit_brea	@expogamerdev 0_0	1.11E+18	1.07E+18	healthy
4		Wed Mar	New Orleans			fruit	_cbiscuit	@cousinwayne Club in the FQ ca	1.11E+18	27380994	healthy
5		Wed Mar	West Java			fruit	dophamin	RT @techinsider: Here's how to r	1.11E+18	1.07E+18	healthy
6		Wed Mar	Cynthiana, IN			fruit	waynenal	Today's "Abide In Christ" by Andı	1.11E+18	18128585	healthy
7		Wed Mar	Yorkshire UK			fruit	LizWalker	@saranewman321 I also have a s	1.11E+18	4.37E+08	healthy
8		Wed Mar	Republik Federasi Indonesia			fruit	Herrreza	RT @techinsider: Here's how to r	1.11E+18	25979312	healthy
9		Wed Mar	Sna Francisco			fruit	IDFRQK	RT @jiggyJurmy: Petition to chan	1.11E+18	4.21E+09	healthy
10		Wed Mar	Thataway			fruit	RhodyRep	@IanDon @StopandShop Bags of	1.11E+18	9.41E+17	healthy
11		Wed Mar	Dublin City, Ireland			fruit	Strictlysug	@suggbuswell_ @Joe_Sugg @db	1.11E+18	9.92E+17	healthy
12		Wed Mar	805			fruit	Eazy22	Got some fruit to eat on my lunch	1.11E+18	38101845	healthy
13		Wed Mar				fruit	MinionCal	Is it bad @preston_scherr that I'r	1.11E+18	1.11E+18	healthy
14		Wed Mar	camp half blood			fruit	chxrasriel	RT @hattiesoykan: oliver:elio: i c	1.11E+18	6.06E+08	healthy
15		Wed Mar	Worldwide			fruit	TopTwts	Here's how to remove pesky pes	1.11E+18	4.89E+08	healthy
16		Wed Mar	STL			fruit	mirwads	I had to make a personality type	1.11E+18	1.94E+08	healthy

~17% can be located to a US State.

Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

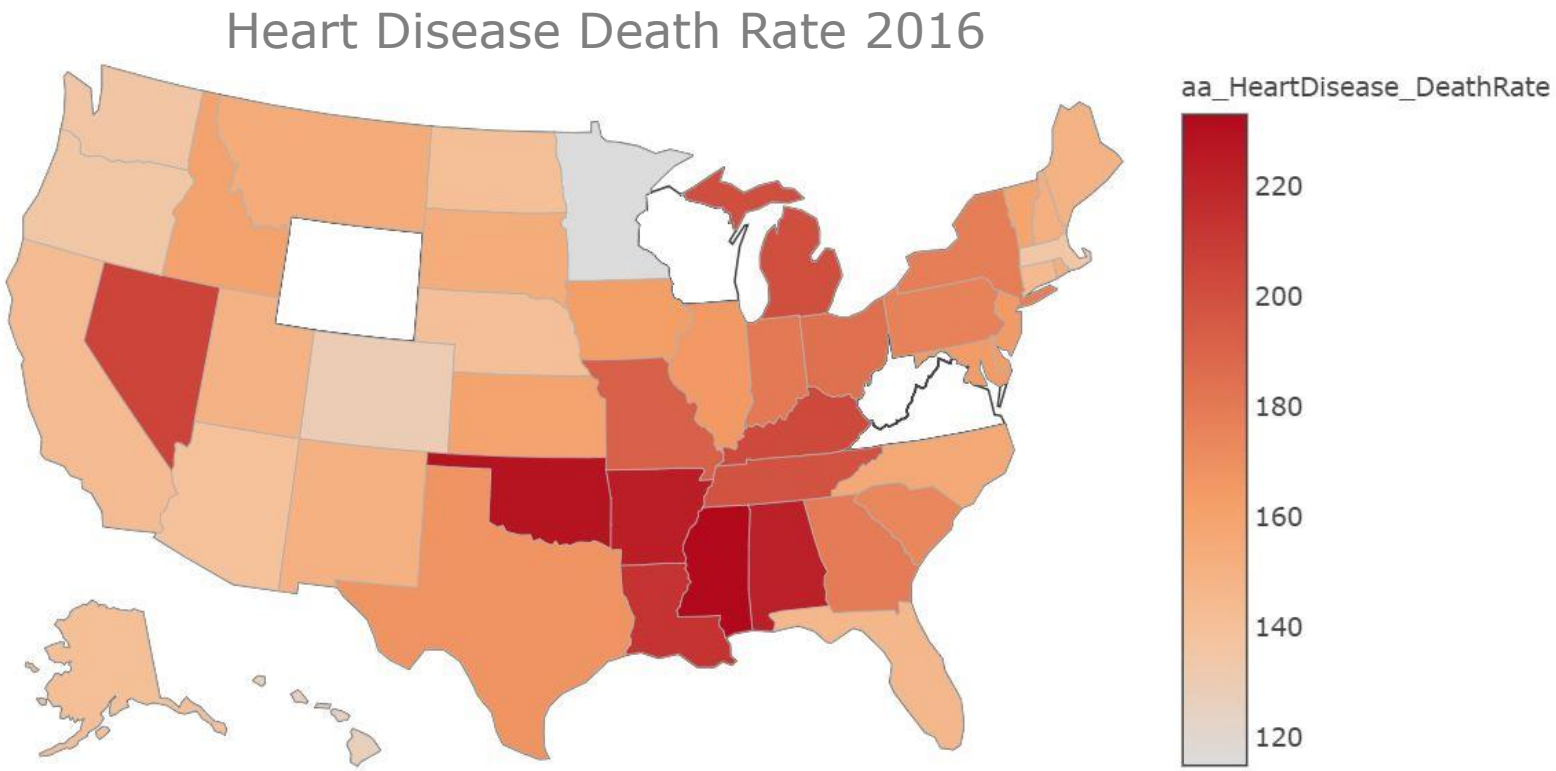
Collected data for several health indicators
-> age_adjusted Diabetes death rate



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

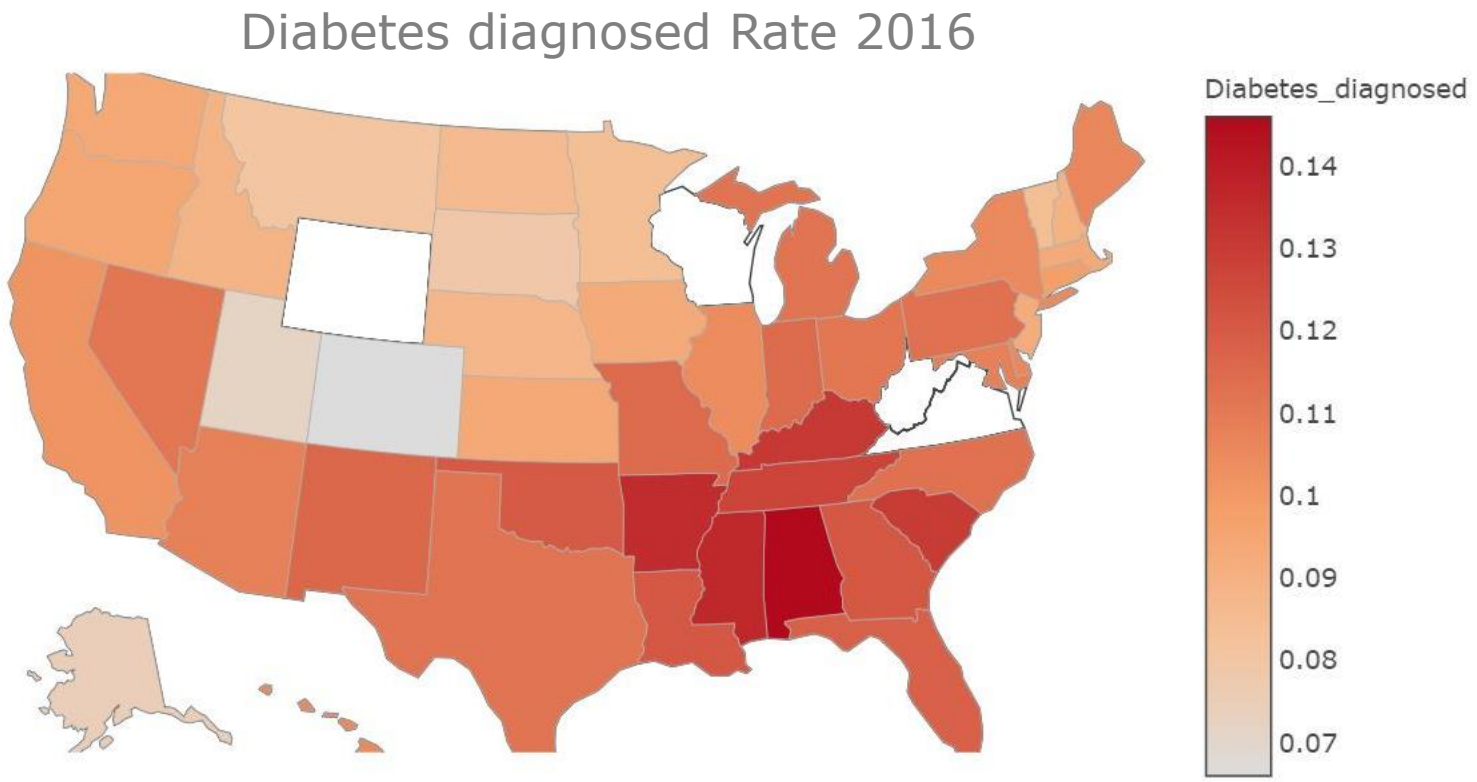
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

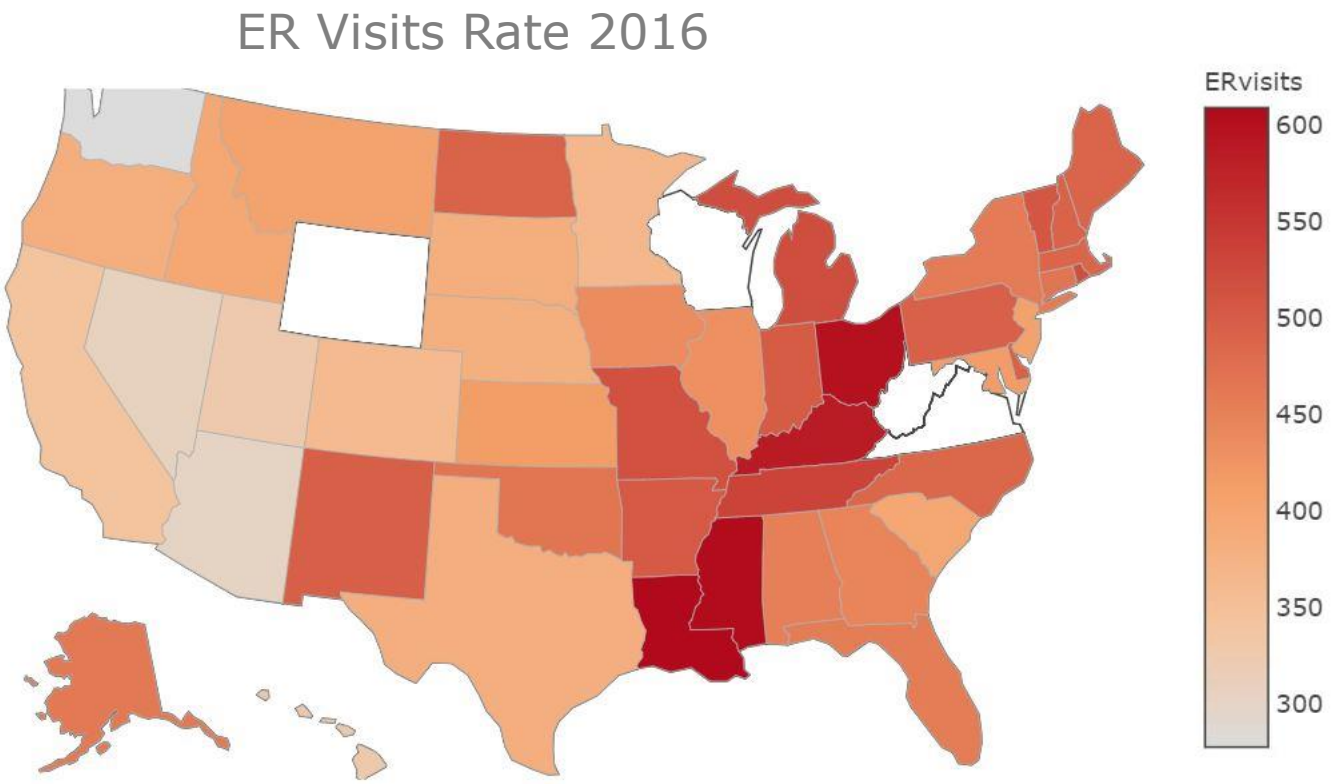
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

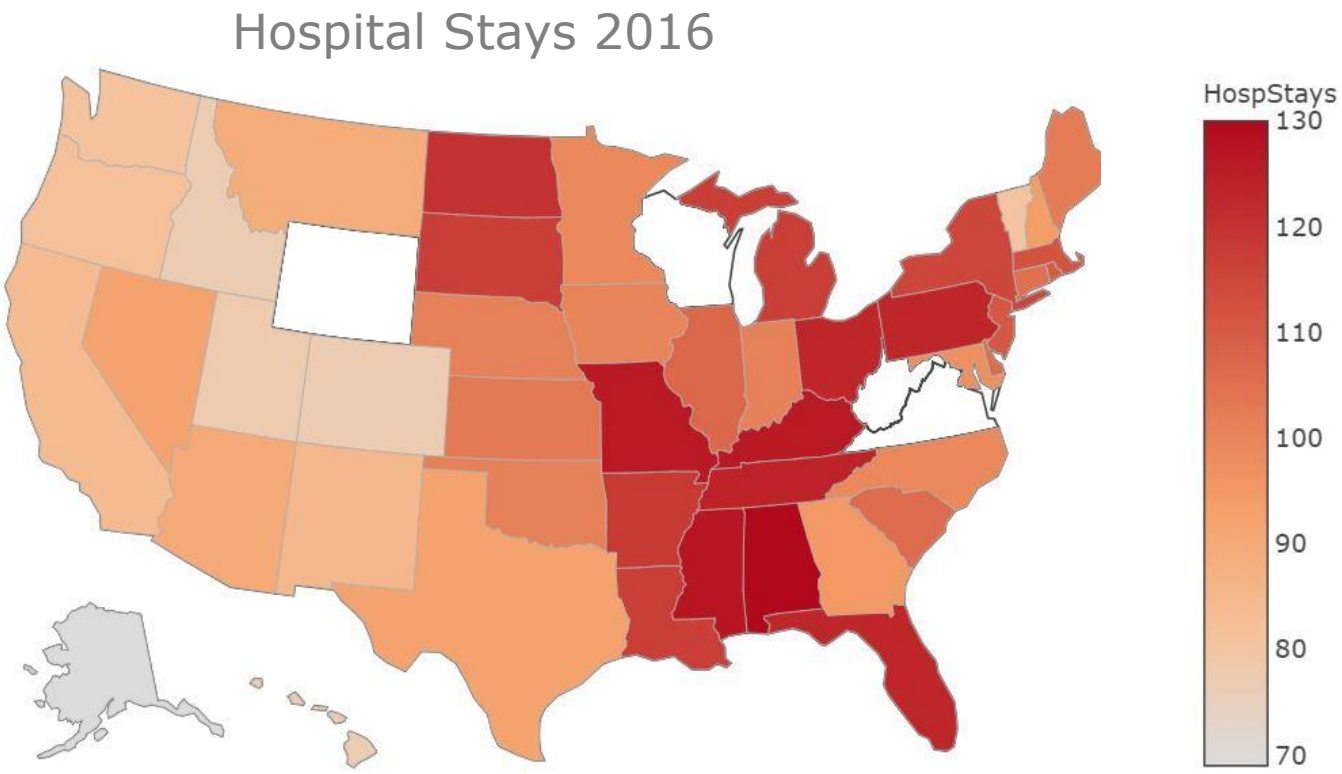
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed
-> ER visits



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

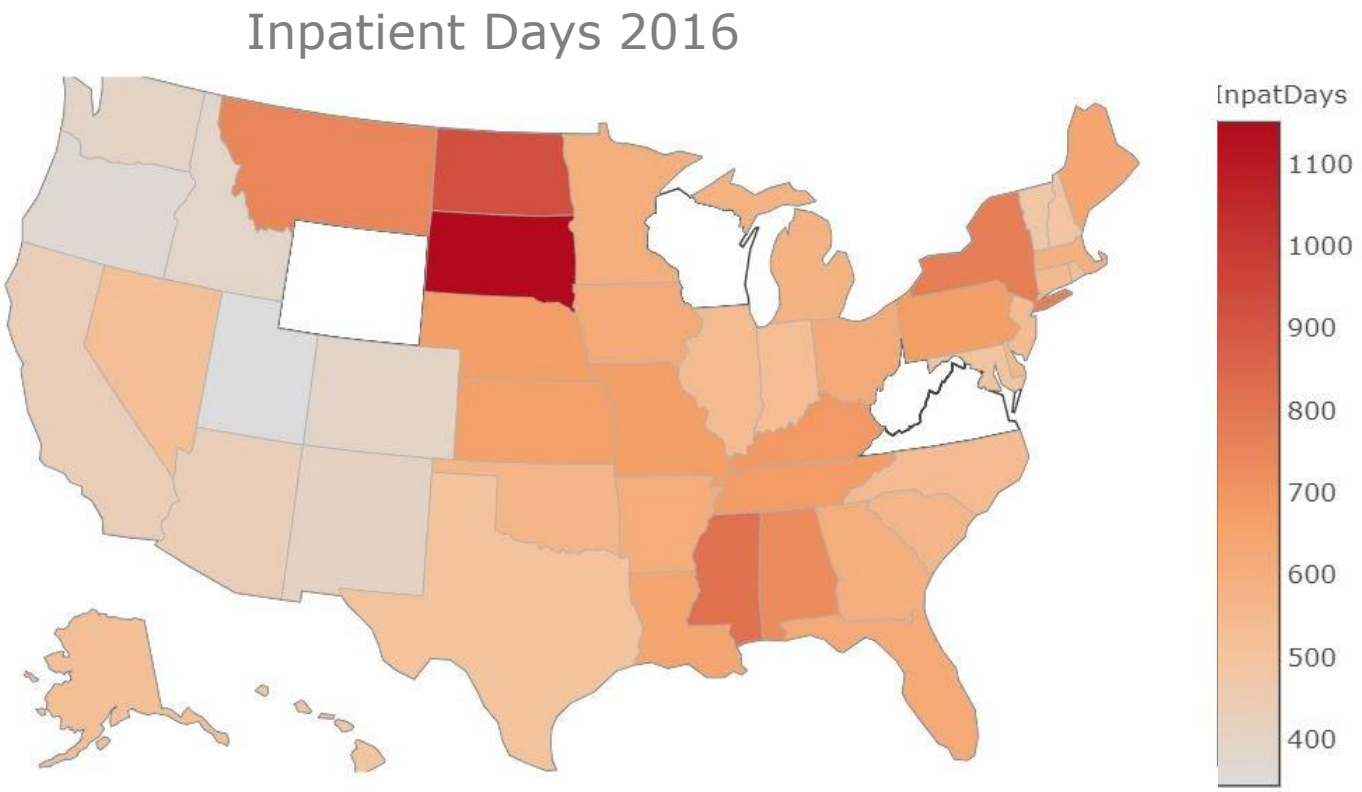
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed
-> ER visits
-> Hospital stays



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

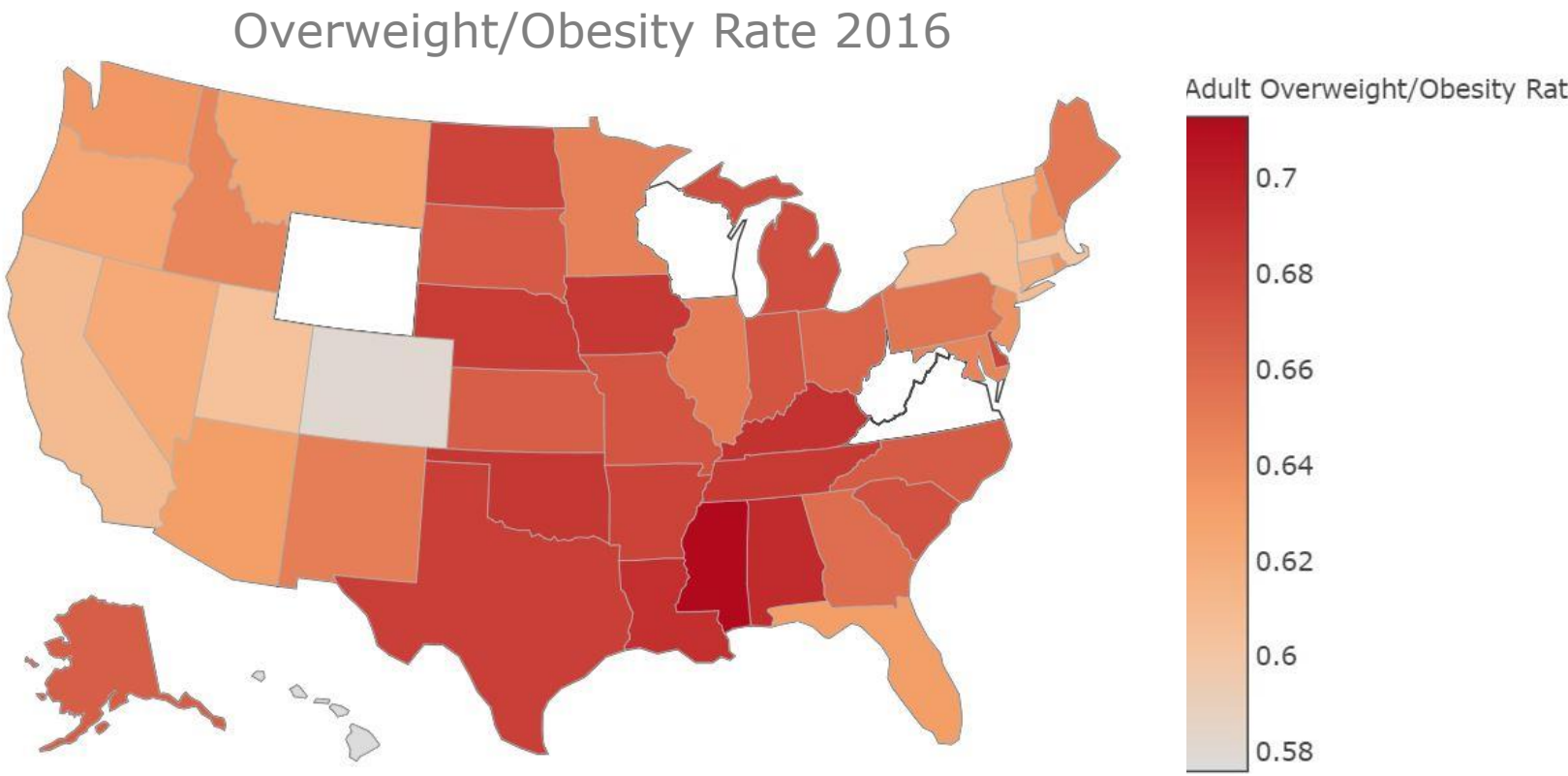
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed
-> ER visits
-> Hospital stays
-> Inpatient days



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

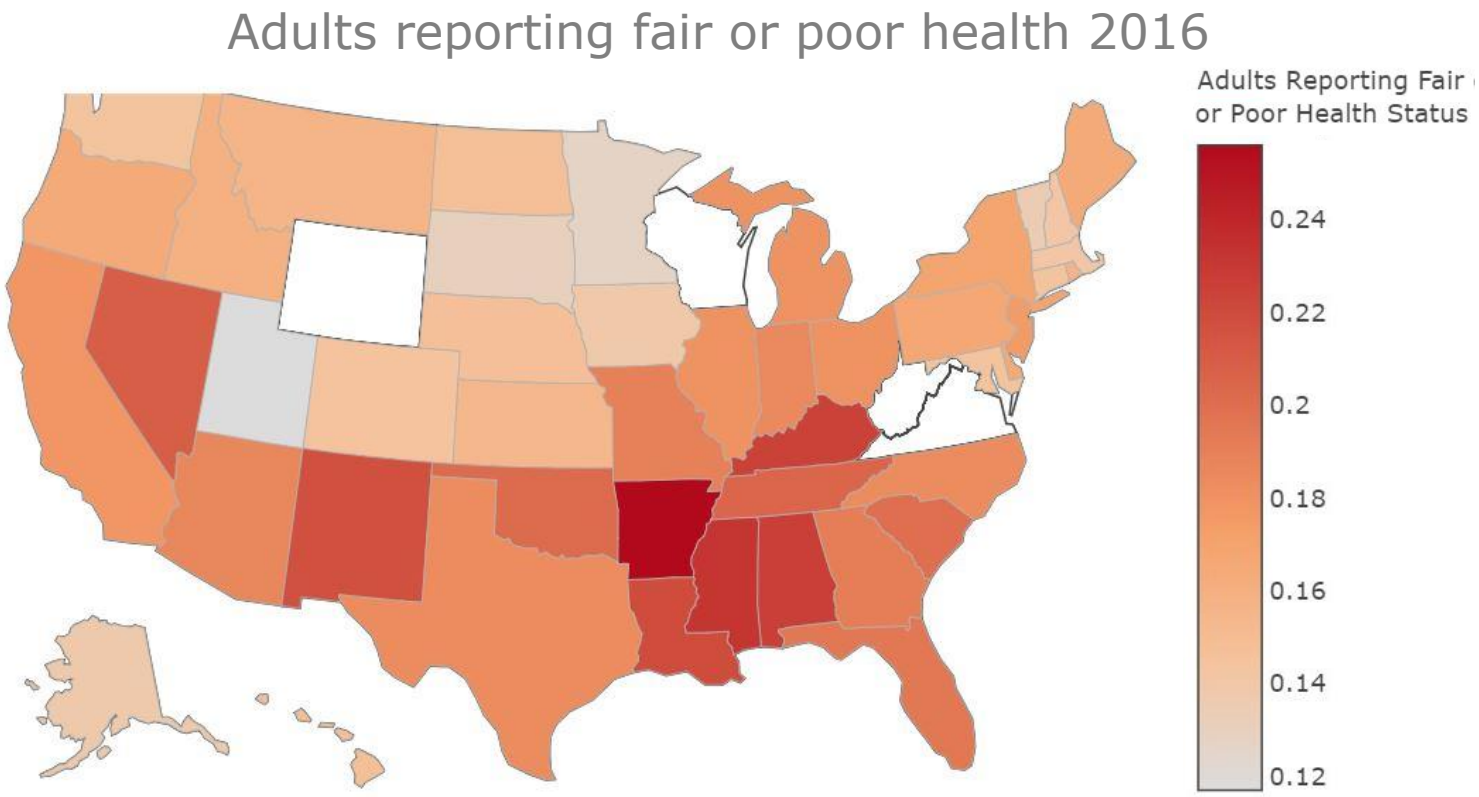
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed
-> ER visits
-> Hospital stays
-> Inpatient days
-> Adult Overweight/Obesity Rate



Data – Health data by US State 2013-2016

Source: https://www.kff.org/statedata/

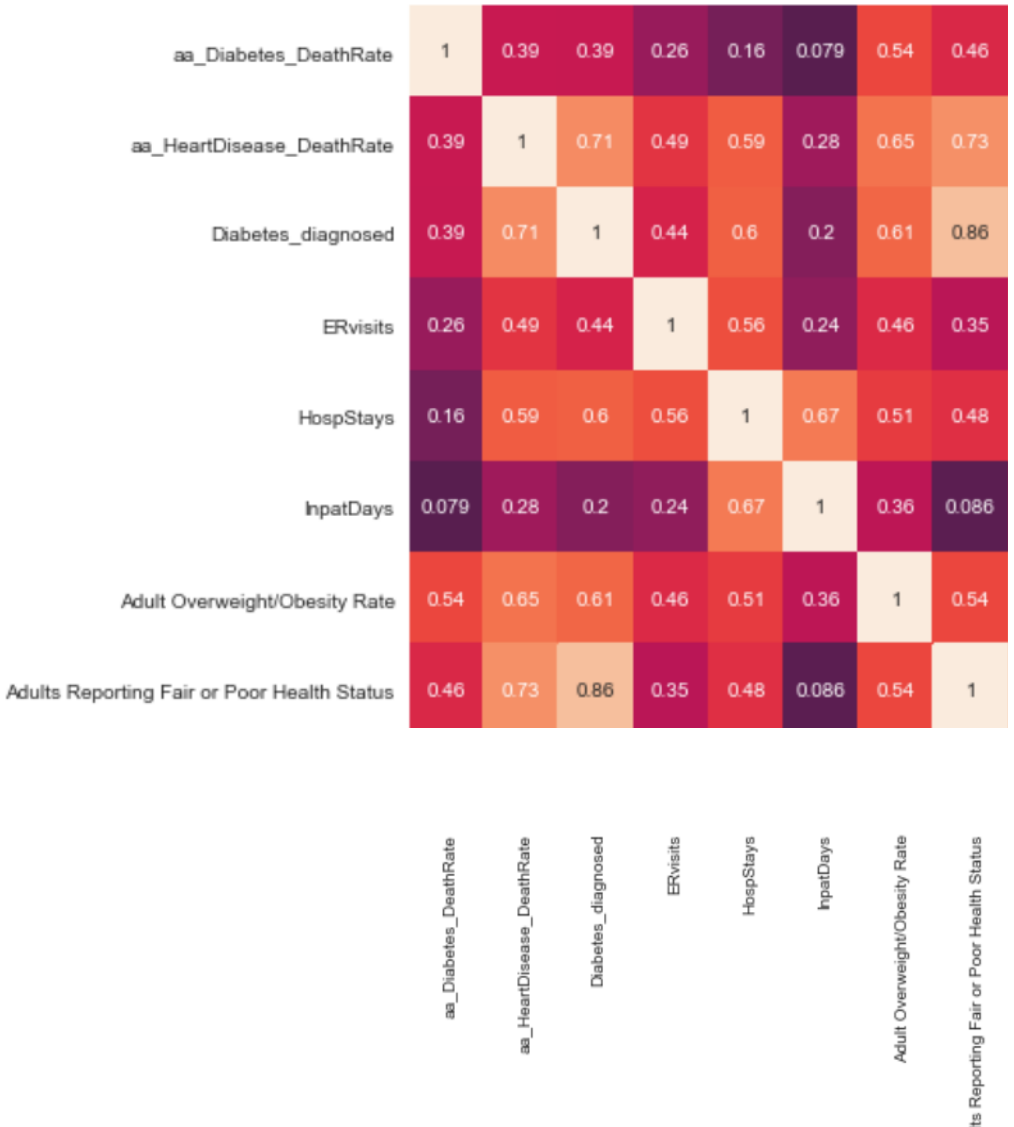
Collected data for several health indicators
-> age_adjusted Diabetes death rate
-> age_adjusted Heart disease death rate
-> Diabetes diagnosed
-> ER visits
-> Hospital stays
-> Inpatient days
-> Adult Overweight/Obesity Rate
-> Adults reporting fair or poor health status



Feature engineering

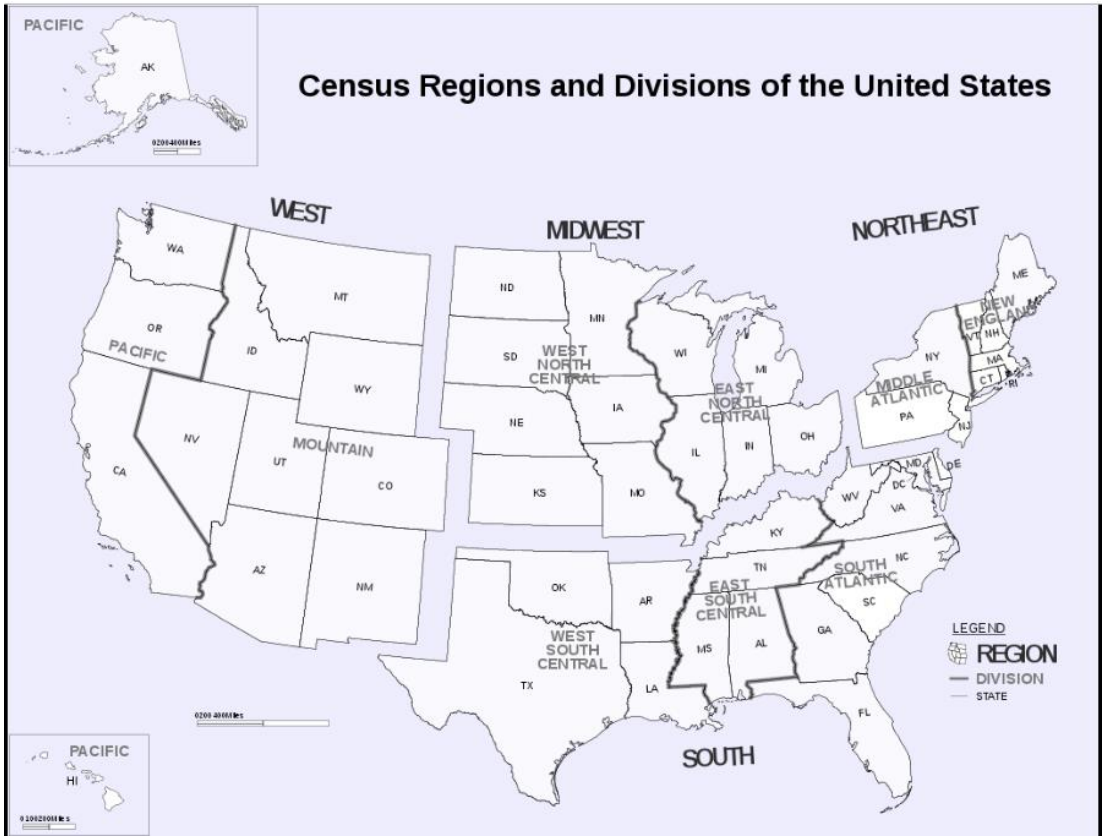
1. Combi_Indi: Add all normalized health indicators into one feature

   The higher Combi_Indi
   the unhealthier.



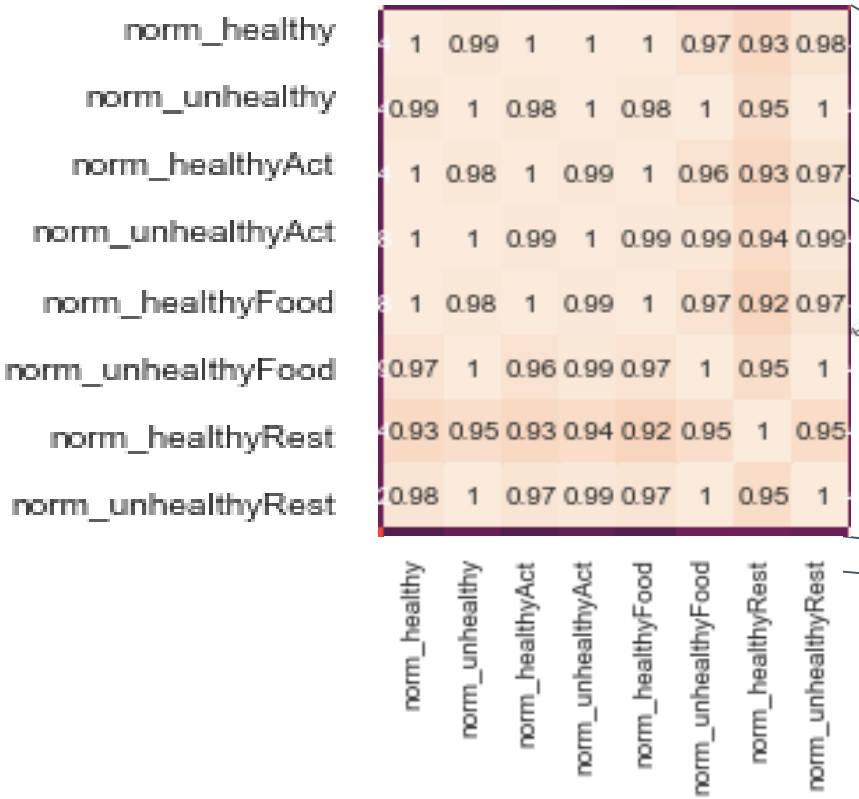
Feature engineering

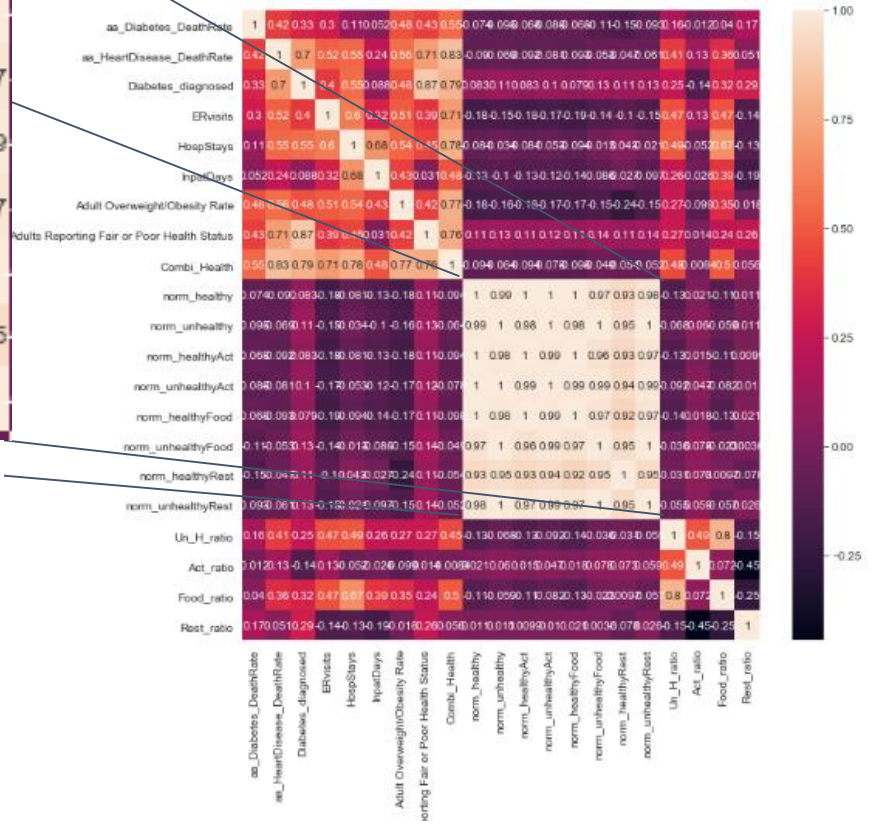
1. Combi_Indi: Add all normalized health indicators into one feature
2. Regions and Divisions: 4 Regions + 9 Divisions
    1. **Northeast:**
        - I. CT, ME, MA, NH, RI, VT
        - II. NJ, NY, PA
    2. **Midwest**
        - III. IL, IN, MI, OH, WI
        - IV. IA, IN, KS, MN, ND, SD, MO, NE
    3. **South**
        - V. DE, FL, GA, MD, NC, SC, WV
        - VI. AL, KY, MS, TN
        - VII. AR,LA,OK,TX
    4. **West**
        - VIII: AZ, CO, ID, MT, NM, NV,UT,WY
        - IX: AK, CA, HI, OR, WA



Feature engineering

1. Combi_Indi: Add all normalized health indicators into one feature
2. Regions and Divisions: 4 Regions + 9 Divisions
3. Drop normalized counts





Feature engineering

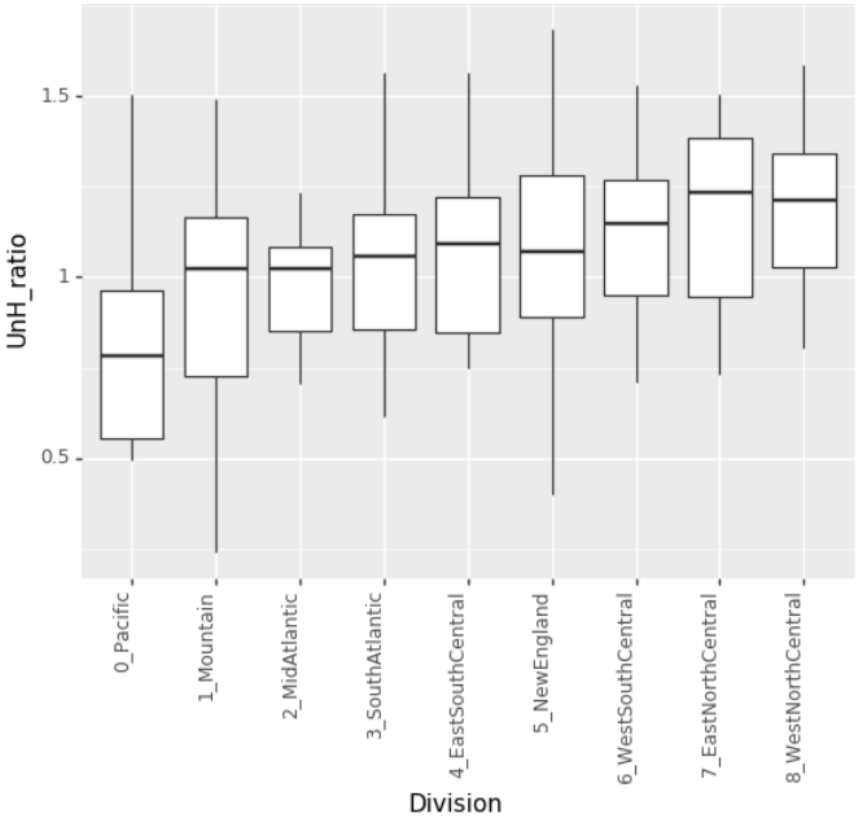
1. Combi_Indi: Add all normalized health indicators into one feature
2. Regions and Divisions: 4 Regions + 9 Divisions
3. Drop normalized counts
4. Keeping only UnH_ratio
   - $Tweet\_ratio = \frac{Sum(unhealthy\ tweets)}{Sum(healthy\ tweets)}$

.

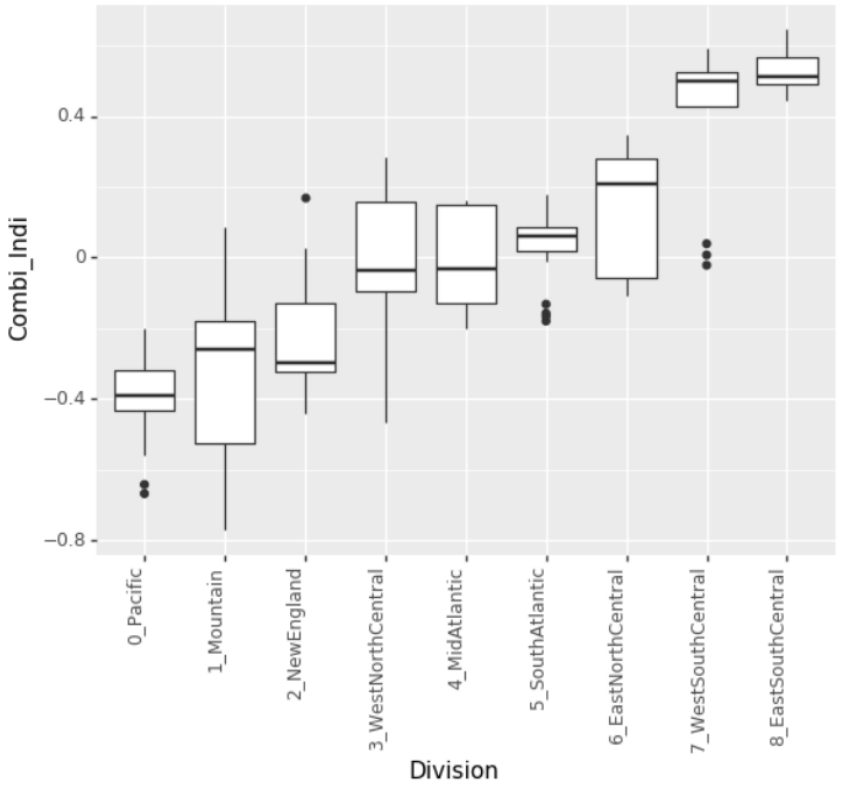
Data for Modeling

Data for Modeling

Tweet count ratio

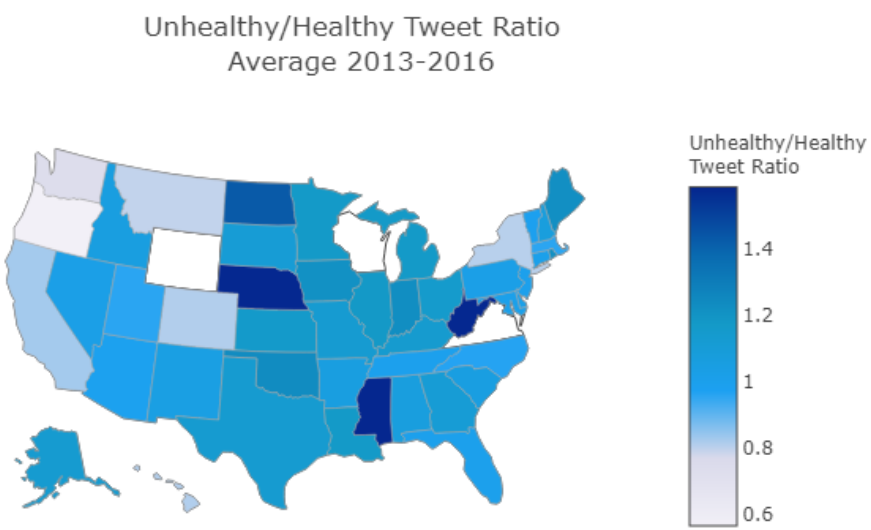


Combi_Indi

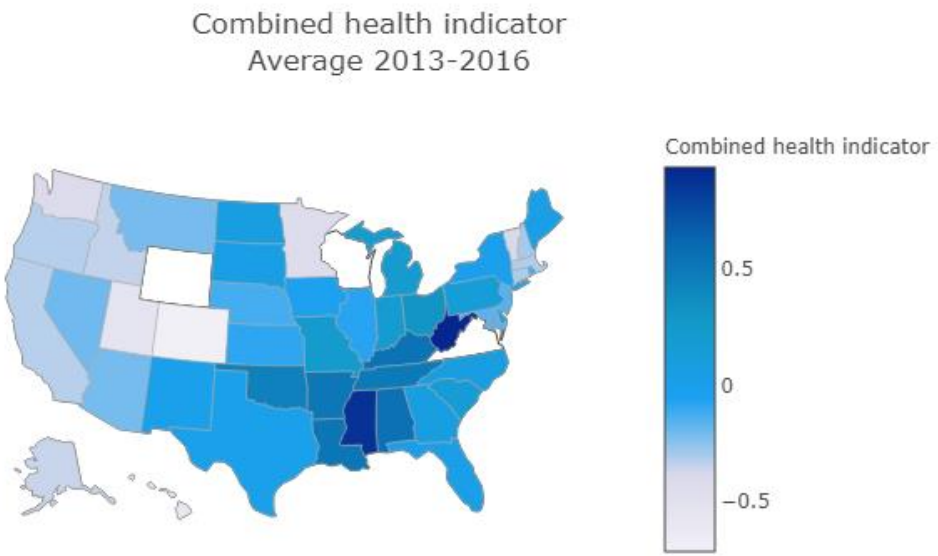


Data for Modeling

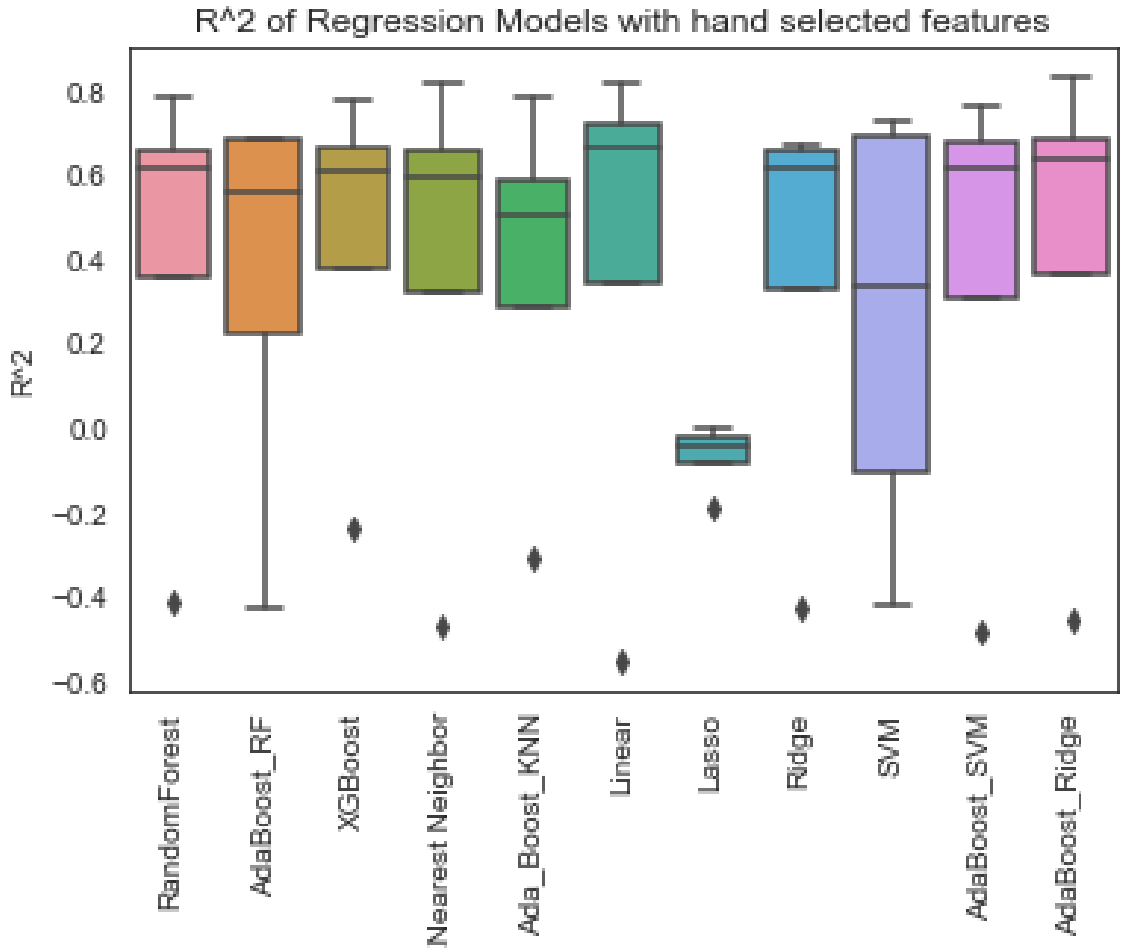
Tweet count ratio



Combi_Indi

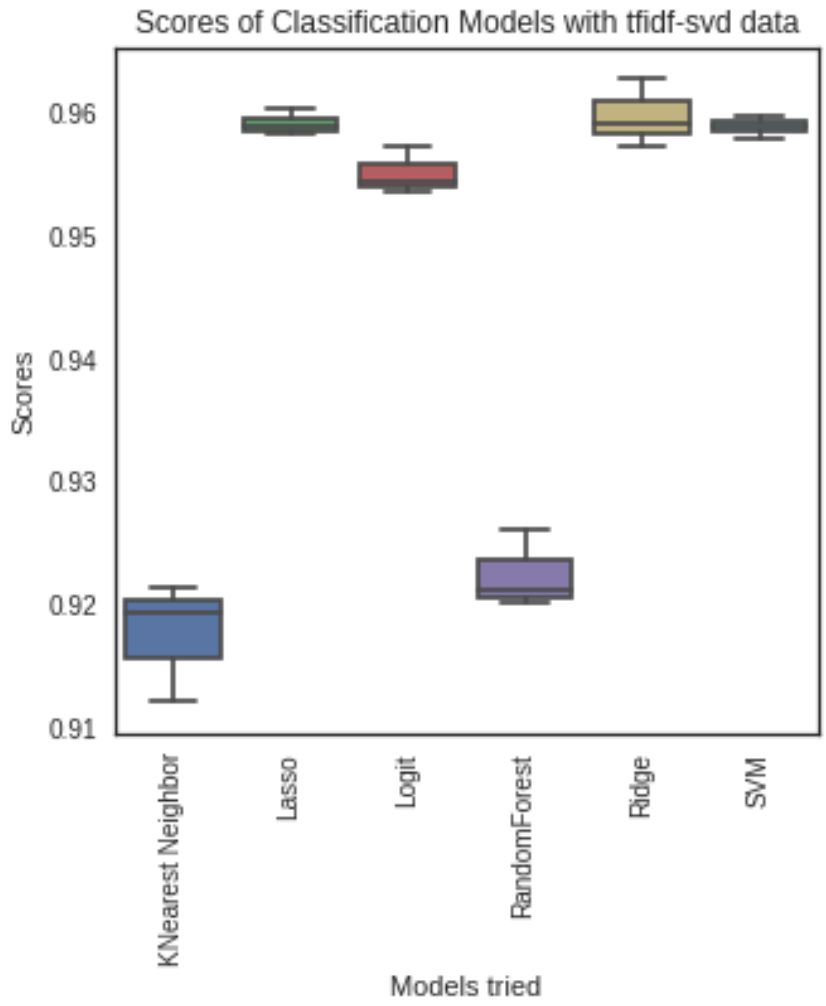


Modeling with hand selected features



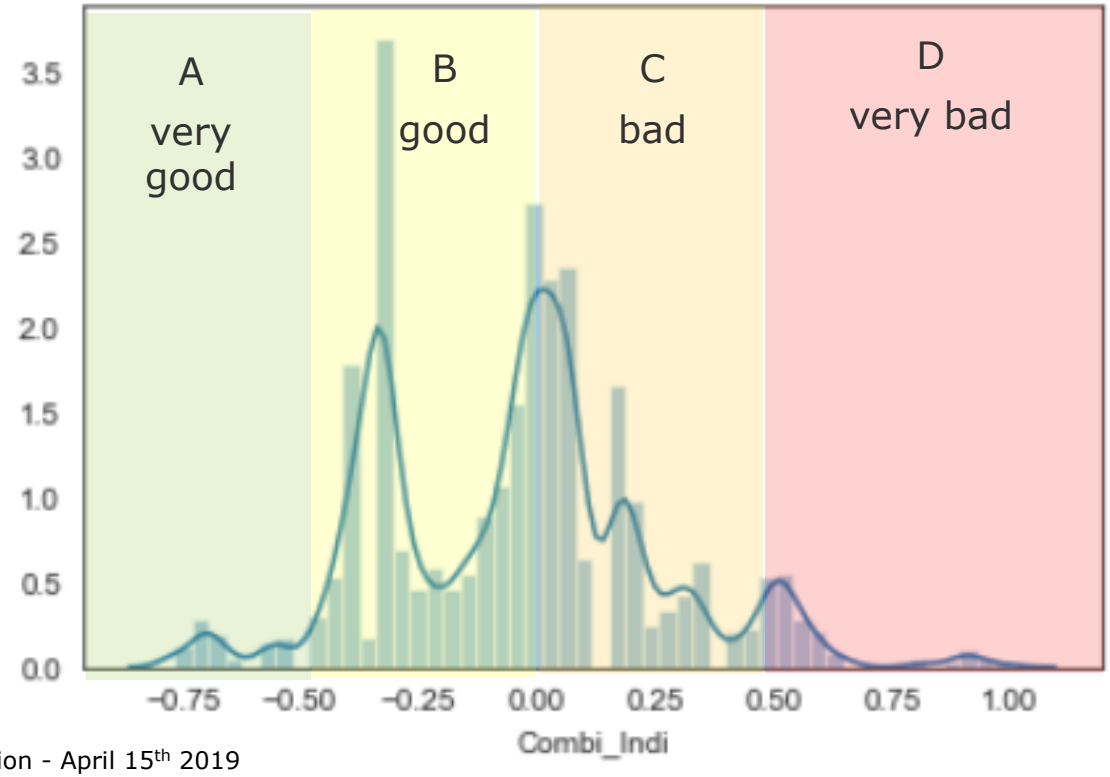
Natural language processing

1. Vectorized tweets into tfidf – matrix
2. Tested models on:
    1. LSA (n=300, ~30% variance)



Classification of Health Cat

1. Vectorized tweets into tfidf – matrix
2. Feature generation:
    1. LSA (n=300, ~30% variance)
    2. Combined LSA with health data and regional data
    3. Categorized Combi_Indi to Health_Cat
    4. Balanced dataset by undersampling



Classification of Health Cat

1. Vectorized tweets into tfidf – matrix
2. Feature generation:
    1. LSA (n=300, ~30% variance)
    2. Combined LSA with health data and regional data
    3. Categorized Combi_Indi to Health_Cat
    4. Balanced dataset by undersampling

Tweet count ratio and regional features per Year- State – Combi_Indi

	UnH_ratio	Combi_Indi	Year	Division_EastNorthCentral	Division_EastSouthCentral	Division_MidAtlantic	Division_Mountain	Division_NewEng
0	0.852940	0.599171	2013	0	1	0	0	
1	1.028597	0.510582	2014	0	1	0	0	
2	0.834628	0.559005	2015	0	1	0	0	
3	1.566837	0.560150	2016	0	1	0	0	
4	0.768955	-0.409322	2013	0	0	0	0	

1

188, 8

LSA data and regional features per tweet

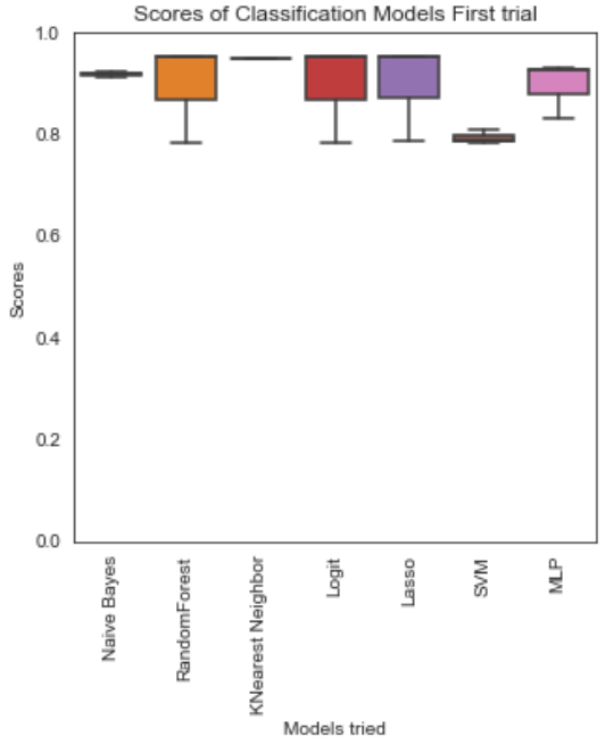
Combi_Indi	lsa_1	lsa_2	lsa_3	lsa_4	lsa_5	lsa_6	lsa_7	lsa_8	lsa_9
0.600182	0.098676	-0.031747	-0.031175	0.002269	-0.040481	-0.018853	-0.036722	0.014207	-0.001484
0.600182	0.028051	0.016936	-0.016182	-0.001850	-0.012468	[illegible]	-0.008152	-0.008289	0.009551
0.600182	0.210179	0.027338	-0.029103	-0.112215	0.020549	0.009320	-0.042097	-0.084879	0.021942
0.600182	0.153372	-0.090189	-0.018333	0.028724	-0.065652	-0.089376	-0.115554	0.022423	-0.015860
0.600182	0.091618	0.003121	-0.033694	-0.047823	-0.022902	-0.010381	-0.002685	-0.053919	-0.010912

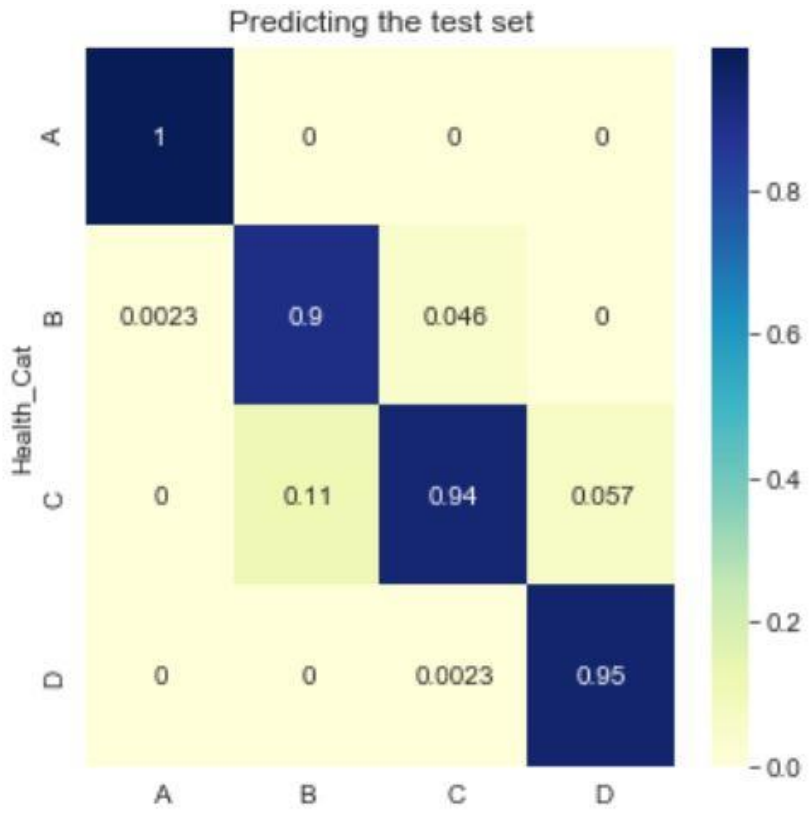
~63

11820, 363

Classification of Health Cat

1. Vectorized tweets into tfidf – matrix
2. Feature generation:
    1. LSA (n=300, ~30% variance)
    2. Combined LSA with health data and regional data
    3. Categorized Combi_Indi to Health_Cat
    4. Balanced dataset by undersampling





Discussion and Outlook

Can twitter data predict obesity rates and associated health indicators? -yes

Can NLP determine a healthy from an unhealthy tweet? - yes

Discussion:

Region data carried a lot predictive power.

Biggest problem- Size of the dataset.

Better use tfidf-svd data set with categorical variable.

Discussion and Outlook

Can twitter data predict obesity rates and associated health indicators? -yes

Can NLP determine a healthy from an unhealthy tweet? - yes

Discussion:

Region data carried a lot predictive power.

Biggest problem: Size of the dataset

Better use tfidf-svd data set with categorical variable.

Future:

Refine the model to city level

Collect ten years of maximum amount of data: since twitter went online

Use data for time series prediction

Thank You!

Tweepy: http://docs.tweepy.org/en/v3.5.0/

Twint: https://github.com/twintproject/twint

Code:
https://github.com/NaRuecker/Final-Capstone/blob/master/Final%20Capstone%20Regression%204Years.ipynb