Behavioural Risk Factors in Mental and Physical Health: High Impact Predictors and High Risk Groups for Negative Outcomes in the U.S.

Elizabeth Cyr

Selected Topic: Behavioural and Environmental predictors of physical and mental health

Environmental/Demographic factors:

- Geographical location (e.g., US state, urban vs rural location)
- Socio-economic status.
- Education
- Race
- Age

Behavioural factors:

- Positive health behaviours (those associated with positive health outcomes)
  - Frequency, duration or type of physical activity
  - Consumption of fruits or vegetables
- Negative health behaviours (those associated with poor health):
  - Smoking
  - Drinking alcohol.

Selected Topic: Behavioural and Environmental predictors of physical and mental health

Measures of health:

- Subjective measures
  - How a respondent rates their general health on a Likert scale
  - Reports of bad mental or physical health days per month
- Objective measures
  - The existence or severity of illnesses (e.g., heart disease, asthma).

Primary Objective:

Determine which predictors have the greatest impact on health outcomes in the United States

Why I Selected This Topic

Mental and physical health outcomes are currently dire in North America.

- 42.4% obesity prevalence in the US as of 2018 (cdc, 2021)
- The Major Depressive Disorder affects > 6.7% of U.S. adults in a given year (ADAA, 2021)
  - Leading cause of disability in the U.S. for ages 15 to 44.3 (ADAA, 2021).
- Approximately 39.5% of men and women in the U.S. will be diagnosed with cancer at some point during their lifetimes (based on 2015–2017 data; National Cancer Inst., 2020)

The current health landscape is incredibly complex. There is an abundance of information on health online with advice for optimizing health, losing weight, or curing diseases with varying degrees of reliability. It's challenging to discern how to best take care of ourselves with so many novel factors at play.

The more clarity we can provide on improving outcomes, the better we can empower the public to take positive steps for their quality of life.

Data Source: BRFSS

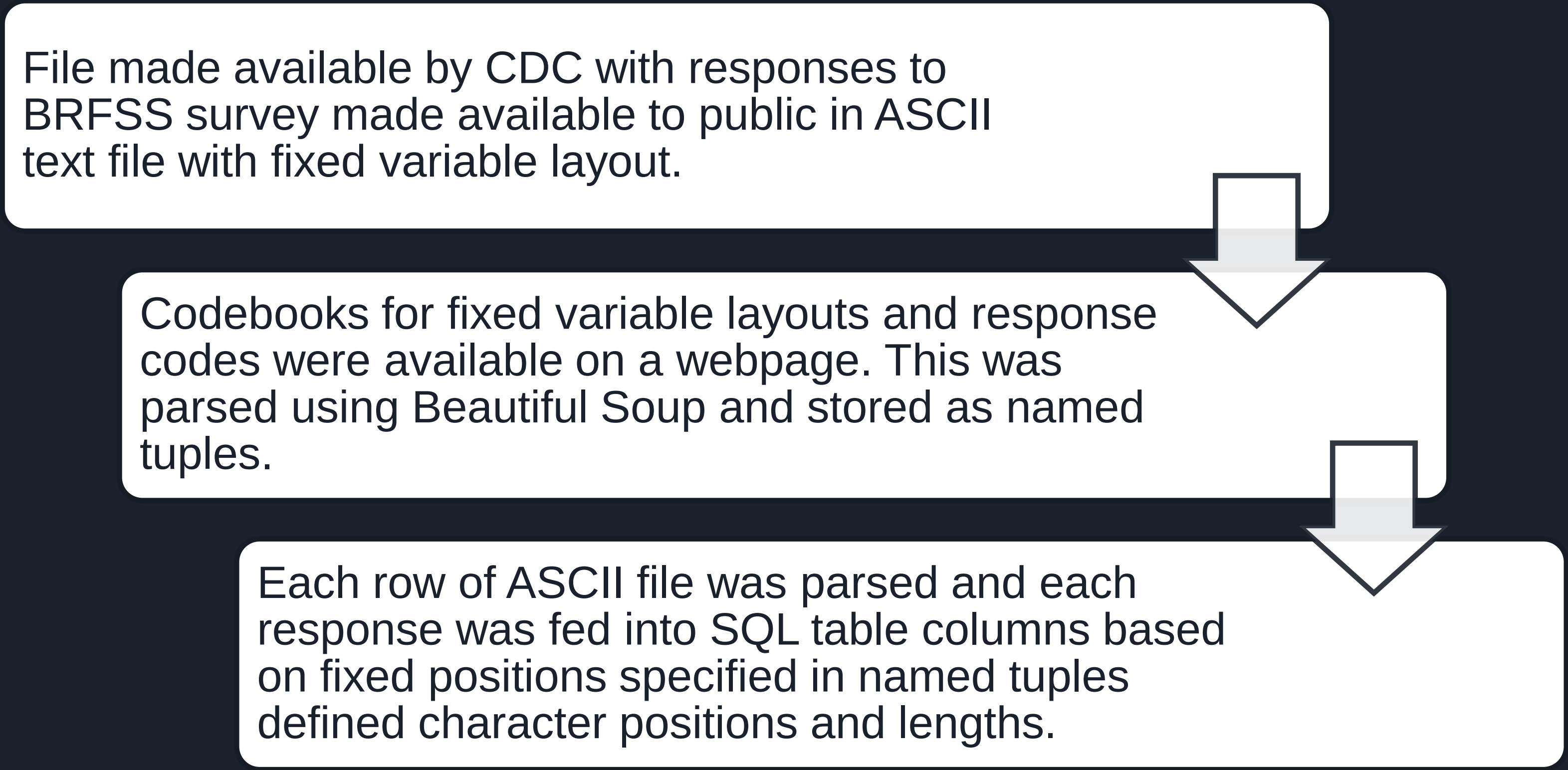
Behavioural Risk Factor Surveillance System (BRFSS):

- A system of ongoing health-related telephone surveys designed to collect data on health-related risk behaviors and chronic health conditions
- A collaborative project between all of the states in the United States (US) and participating US territories and the Centers for Disease Control and Prevention (CDC).
- Analysis performed on data from the 2019 Annual Survey:
    - Data collected from 49 states, the District of Columbia, Guam, and Puerto Rico
        - Exclusion: New Jersey (unable to collect enough BRFSS data in 2019 to meet the minimum requirements for inclusion)
    - Data collected from 418,268 noninstitutionalized adults (18 years or older) U.S. citizens

Primary Research Questions

1. What does mental and physical health look like in the United States today?
2. What impact do environmental/demographic factors have on subjective health and negative health outcomes?
3. What impact do health behaviours have on subjective health and negative health outcomes?
4. If someone only has the energy or ability to make one change in their life to improve their health, what would be the best thing for them to focus on?

Exploratory Analysis: From ASCII and HTML to SQL



File made available by CDC with responses to BRFSS survey made available to public in ASCII text file with fixed variable layout.

Codebooks for fixed variable layouts and response codes were available on a webpage. This was parsed using Beautiful Soup and stored as named tuples.

Each row of ASCII file was parsed and each response was fed into SQL table columns based on fixed positions specified in named tuples defined character positions and lengths.

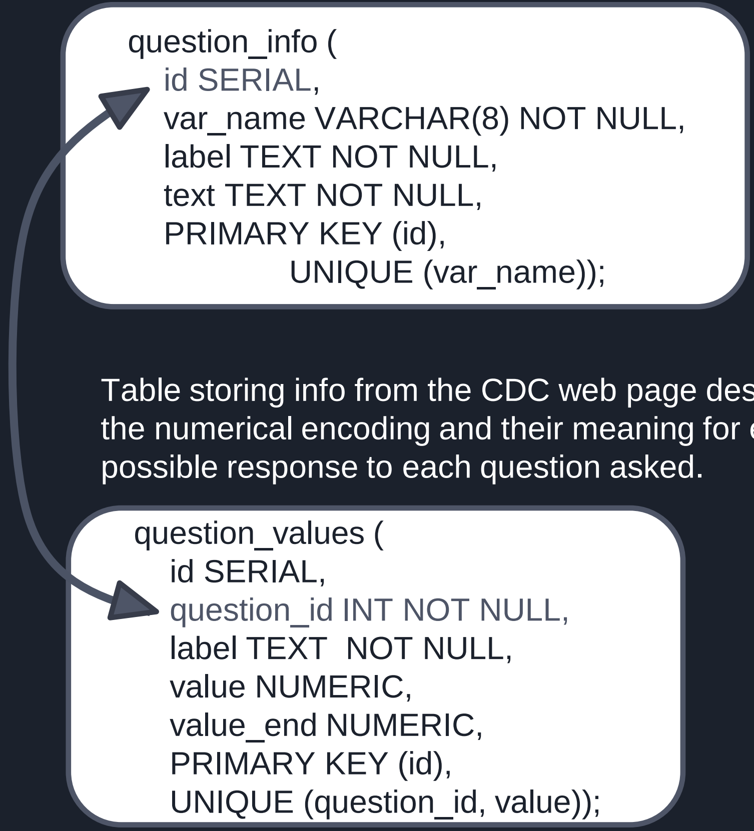
SQL Structure

Table reflecting original ASCII file with row per user and col per question asked.

Columns for responses to each question asked {

```
user_answers (
id SERIAL,
_STATE NUMERIC,
FMONTH NUMERIC,
.
.
.
_VEG23A NUMERIC,
_FRUITE1 NUMERIC,
_VEGETE1 NUMERIC,
_FLSHOT7 NUMERIC,
_PNEUMO3 NUMERIC,
_AIDTST4  NUMERIC,
 PRIMARY KEY (id));
```

Table storing info from the CDC web page describing the SAS variable name, label and wording of each question asked.



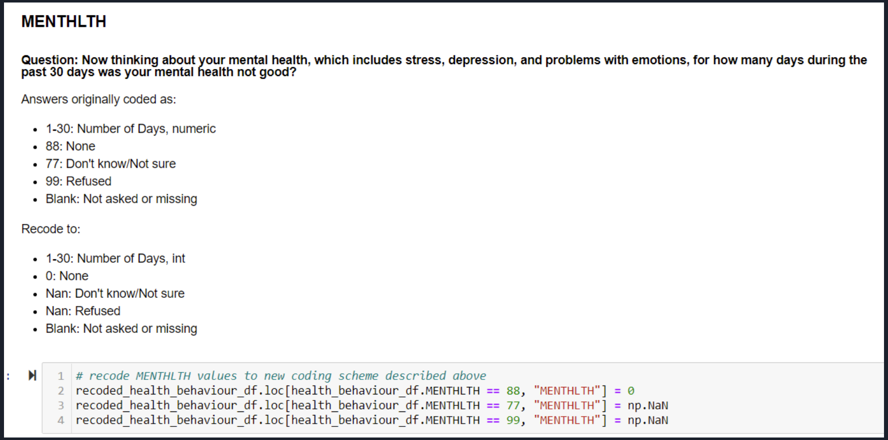
```
question_info (
   id SERIAL,
   var_name VARCHAR(8) NOT NULL,
   label TEXT NOT NULL,
   text TEXT NOT NULL,
   PRIMARY KEY (id),
              UNIQUE (var_name));
```

Table storing info from the CDC web page describing the numerical encoding and their meaning for each possible response to each question asked.

```
question_values (
   id SERIAL,
   question_id INT NOT NULL,
   label TEXT  NOT NULL,
   value NUMERIC,
   value_end NUMERIC,
   PRIMARY KEY (id),
   UNIQUE (question_id, value));
```

Exploratory Analysis: Data Cleaning

In original file, all instances where respondents refused to answer or responded that they did not know the answer were allocated numerical values. These needed to be replaced with a non-numerical value so as to not influence analysis. E.g.:



MENTHLTH

Question: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

Answers originally coded as:

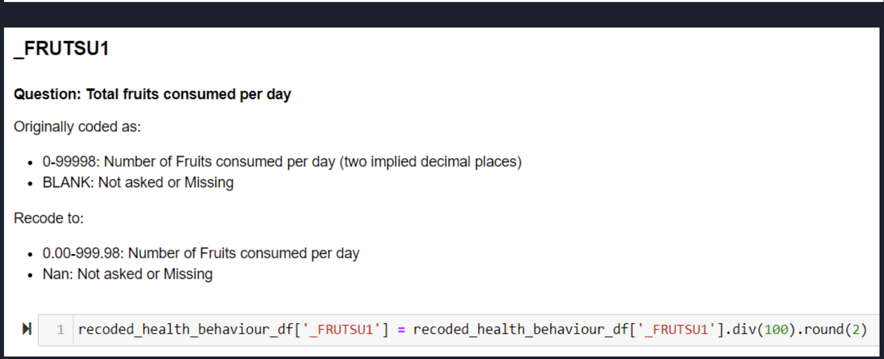
- 1-30: Number of Days, numeric
- 88: None
- 77: Don't know/Not sure
- 99: Refused
- Blank: Not asked or missing

Recode to:

- 1-30: Number of Days, int
- 0: None
- Nan: Don't know/Not sure
- Nan: Refused
- Blank: Not asked or missing

```
# recode MENTHLTH values to new coding scheme described above
recoded_health_behaviour_df.loc[health_behaviour_df.MENTHLTH == 88, "MENTHLTH"] = 0
recoded_health_behaviour_df.loc[health_behaviour_df.MENTHLTH == 77, "MENTHLTH"] = np.NaN
recoded_health_behaviour_df.loc[health_behaviour_df.MENTHLTH == 99, "MENTHLTH"] = np.NaN
```

Certain numerical factors were coded as integers with "implied decimal places". These were converted to floats and recoded to reflect the implied numbers. E.g.:



_FRUTSU1

Question: Total fruits consumed per day

Originally coded as:

- 0-99998: Number of Fruits consumed per day (two implied decimal places)
- BLANK: Not asked or Missing

Recode to:

- 0.00-999.98: Number of Fruits consumed per day
- Nan: Not asked or Missing

```
recoded_health_behaviour_df['_FRUTSU1'] = recoded_health_behaviour_df['_FRUTSU1'].div(100).round(2)
```

Exploratory Analysis: Tools

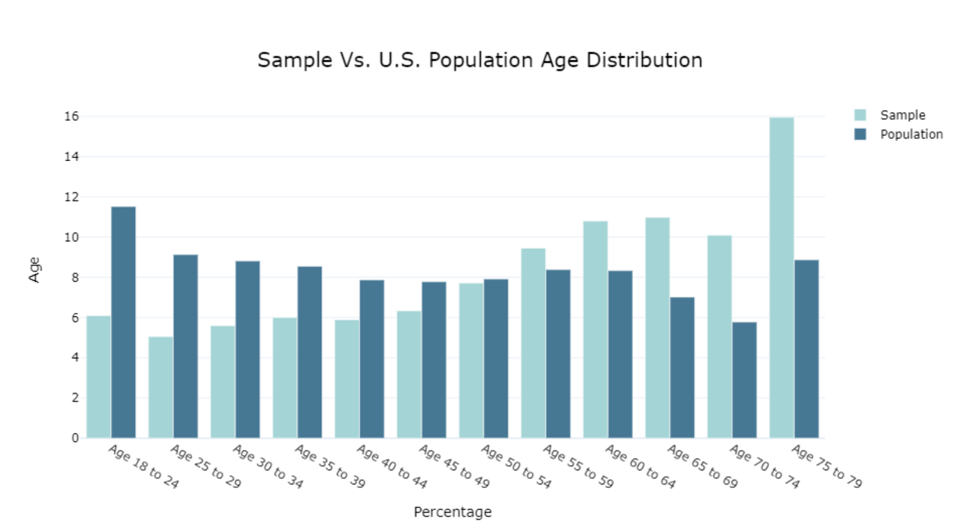
- All data was stored in pandas data frames and analysis was performed in python. Numpy was used to mathematically clean data and calculate aggregates or summary statistics.
- Visualizations were created using either Matplotlib (for quick sample summaries), Plotly graph_objects or plotly express.
- Interactive visualizations were brought online to be shared using plotly chart_studio.

Exploratory Analysis: Sample Breakdown

BRFSS Sample (n = 418,268, M_{age} = 55.38, SD_{age} = 17.62), much older than U.S. Population. Perhaps this reflects greater access to older respondents or an increased likelihood to participate in optional studies with age.

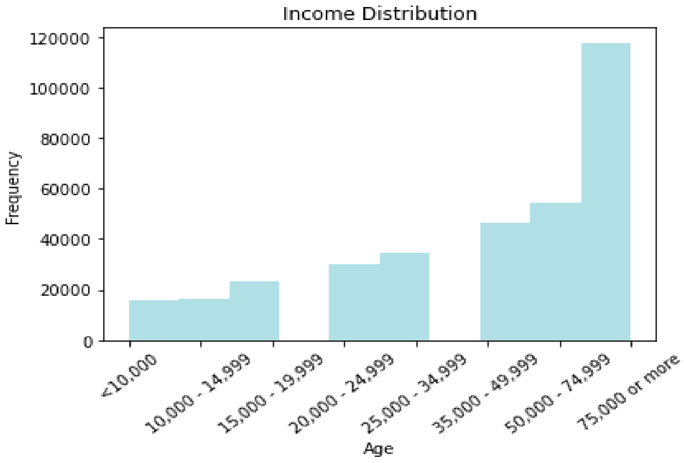
54.6% Female, comparable to U.S. population (50.8%).

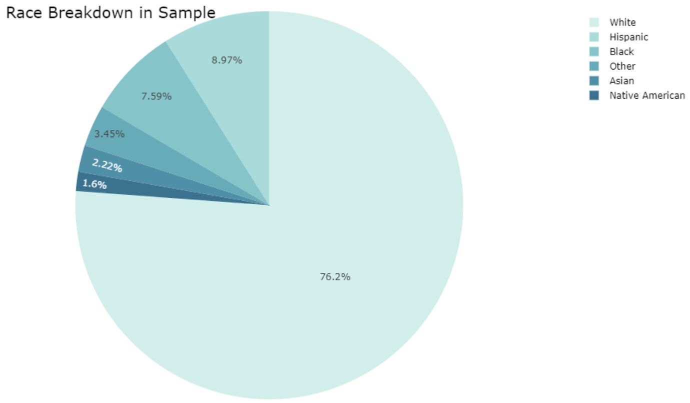
*U.S. population data gathered from [U.S. Census Bureau, 2014-2019]



Sample Breakdown cont’d

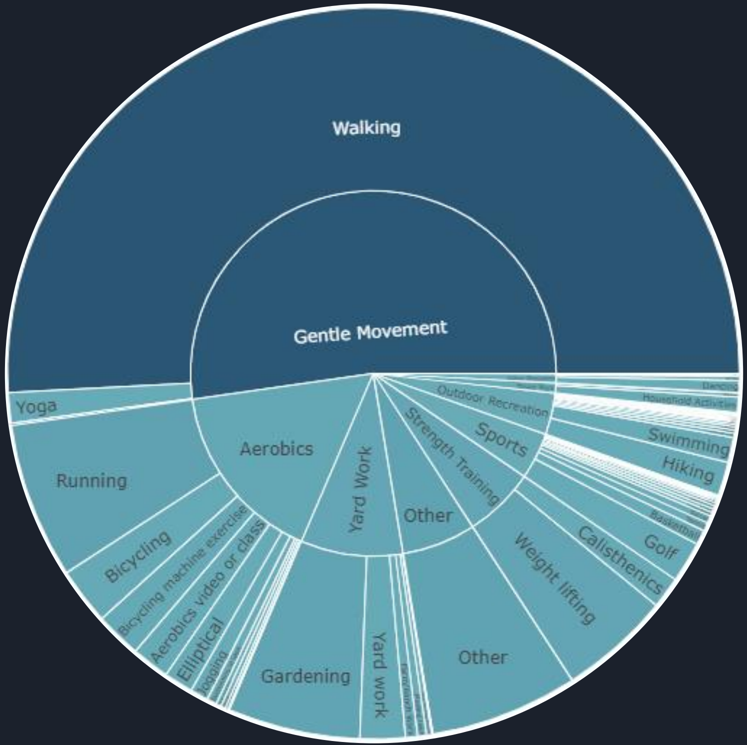
- Income appears in line with U.S. Population (median household income in 2019 = $68,703; U.S. Census Bureau, 2020) but lacking data on income >75k)
- Sample over-represents white portion of U.S. population (60.1%) and under-represents Hispanic (18.5%), black (13.4%) and Asian (5.9%) portions of U.S. population. Native American (American Indian or Alaskan Native) population was accurately reflected at 1.3% (U.S. Census Bureau, 2020).



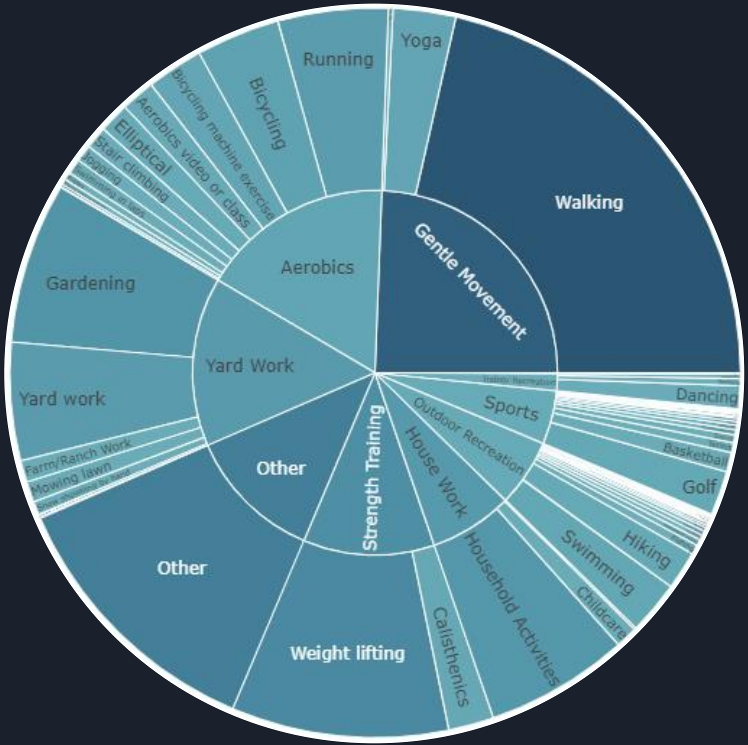


Sources Physical Activity

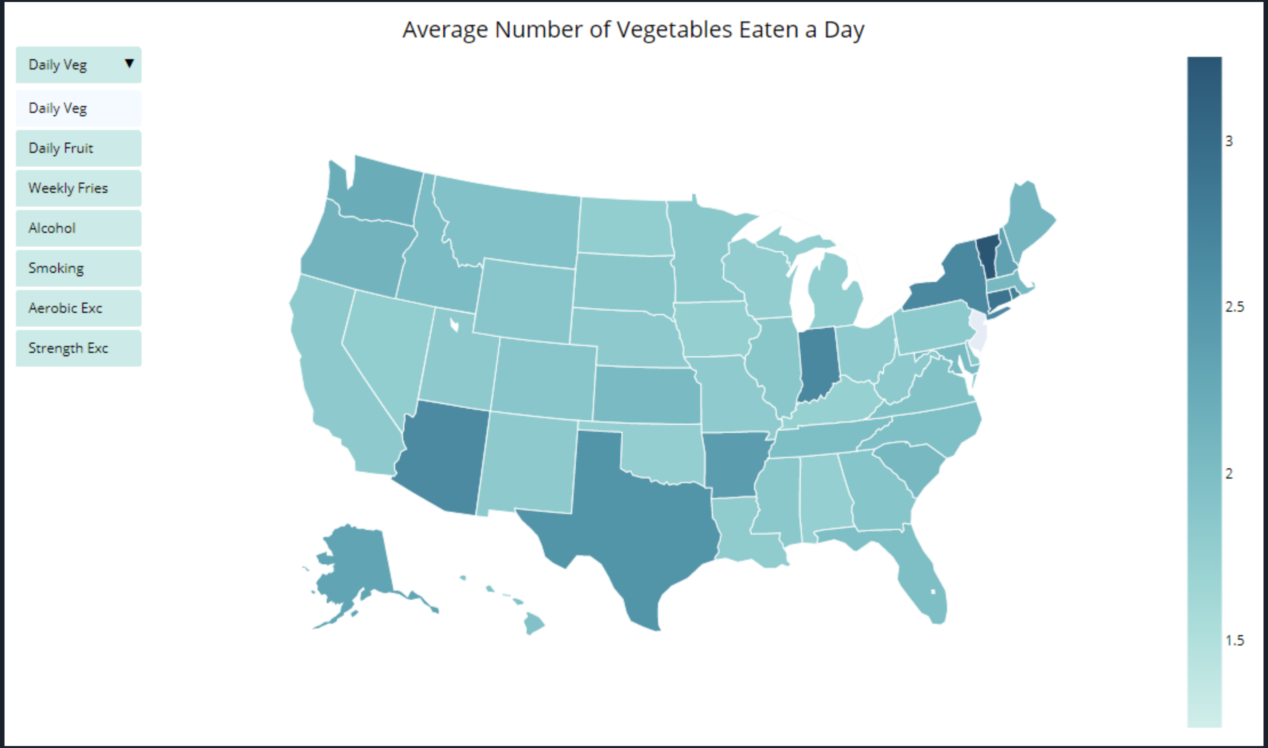
Primary Source



Secondary Source

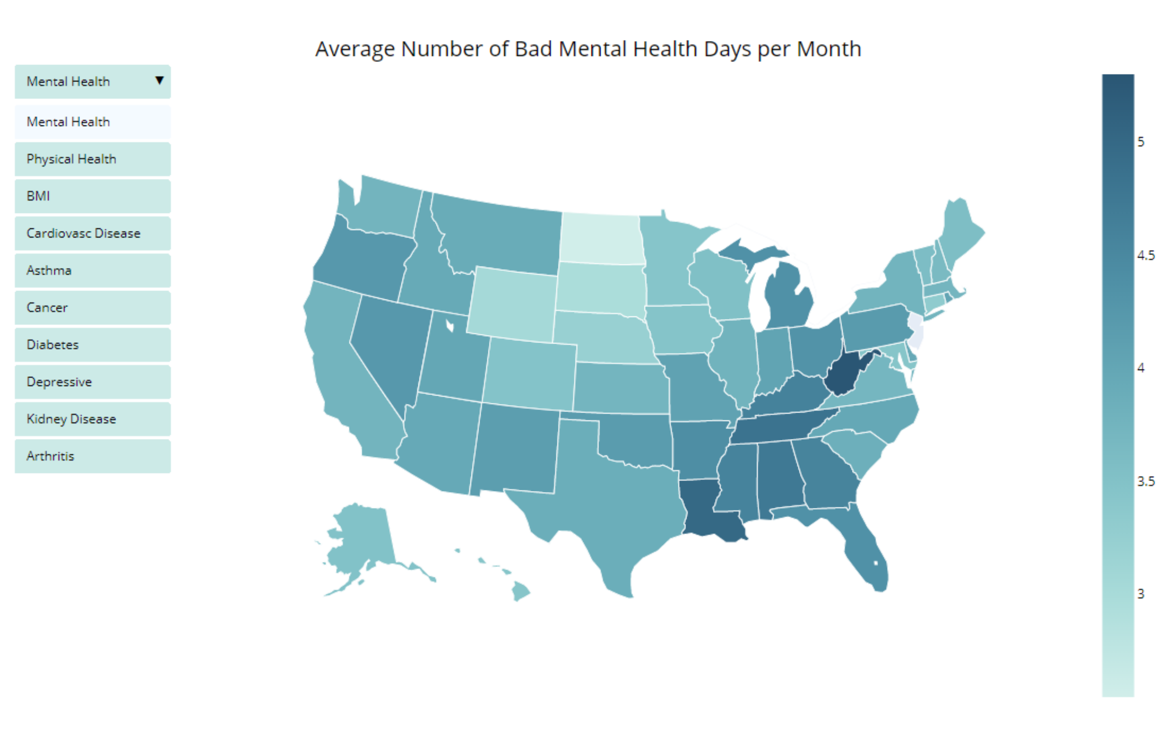


Health Behaviour Across the Country



*Screenshot of interactive map to be made available on github pages dashboard

Health Behaviour Across the Country

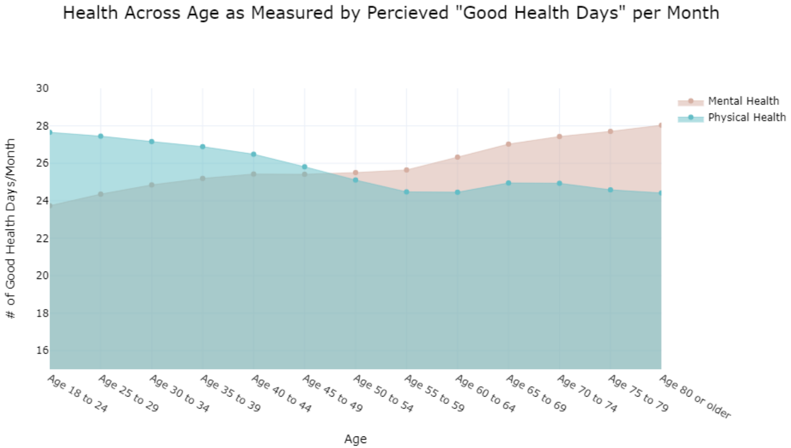


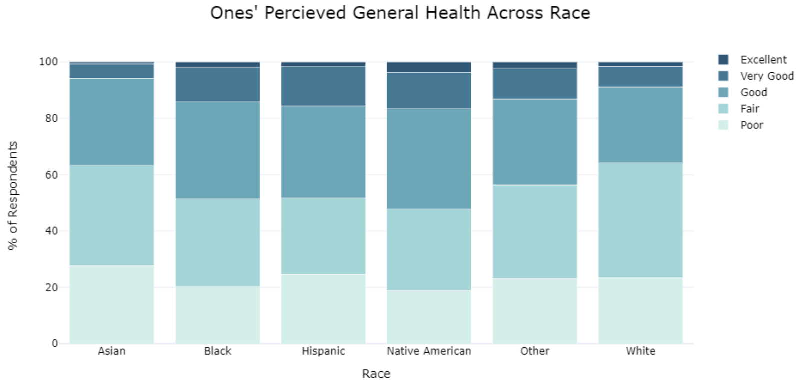
*Screenshot of interactive map to be made available on github pages dashboard

Health Across Age and Race

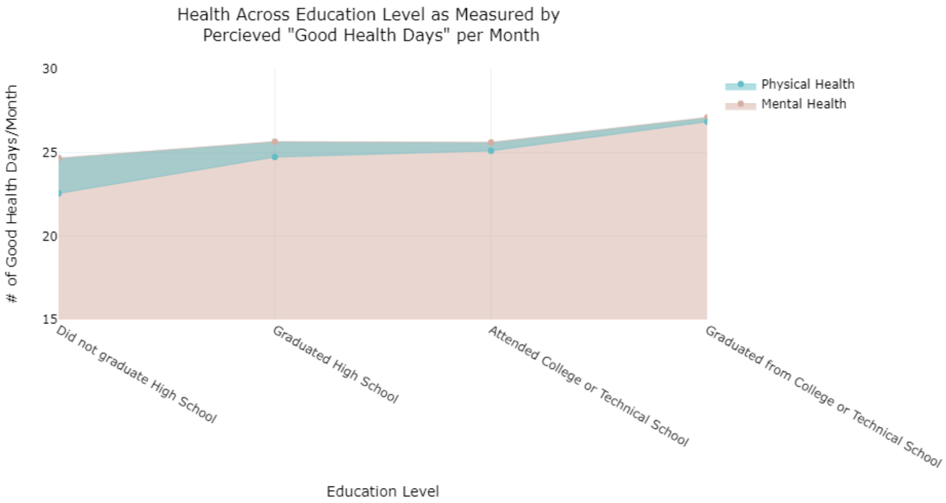
As expected, physical health declines with age. Surprisingly, however, mental health improves with age in a completely inverse relationship to have of physical health and age.

Health ratings are lower in the Asian and White subgroups. This may reflect differences in health. It may also reflect standards for rating or perceiving one's health.

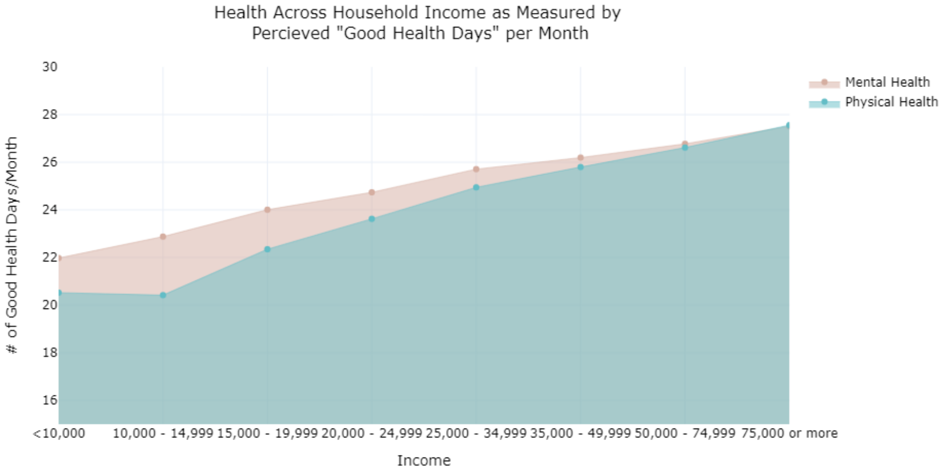




Impact of education and income on reported number of positive mental and physical health days per month



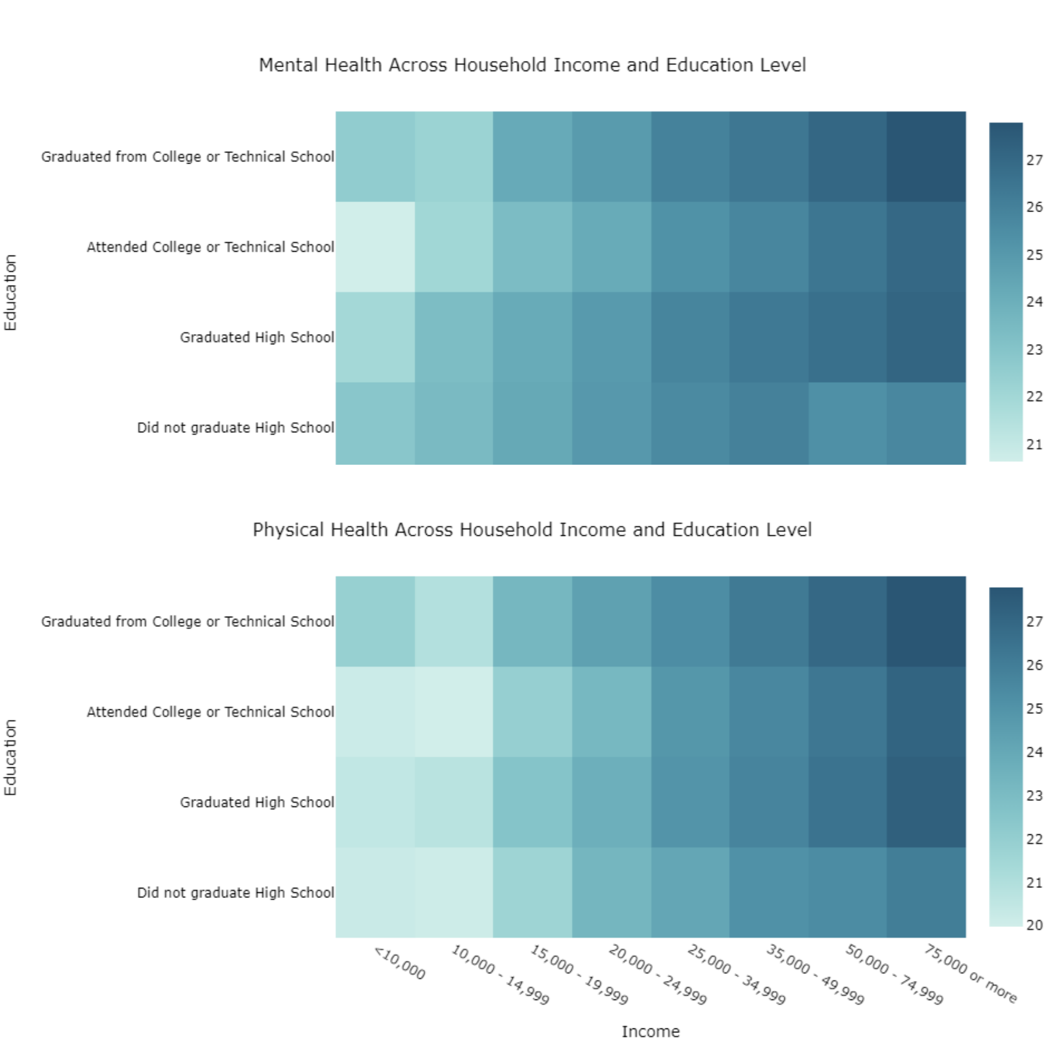
Health Across Education Level as Measured by Percieved "Good Health Days" per Month



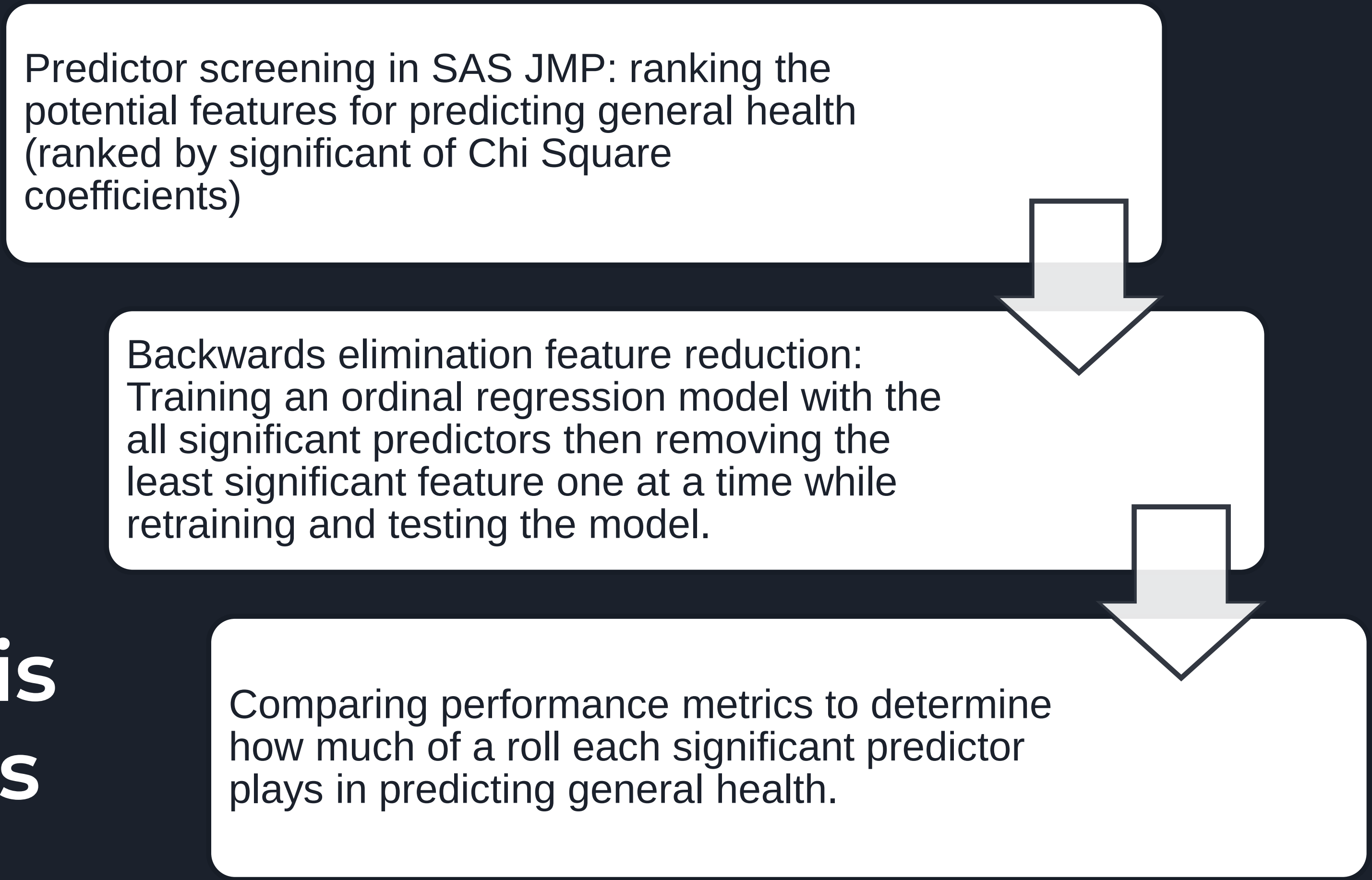
Health Across Household Income as Measured by Percieved "Good Health Days" per Month

Income X Education Interaction

The positive relationships between both income and education with reported health appear to compound with income having a greater effect. This leads to a sharp increase in both positive mental and physical health days for those with greater financial and educational resources. A notable deviation to these trends is a decreased health for those who have attended some post-secondary school. This may reflect the health strain and lack of resources available to students currently enrolled in college or technical school.



Analysis Process



Predictor screening in SAS JMP: ranking the potential features for predicting general health (ranked by significant of Chi Square coefficients)

Backwards elimination feature reduction: Training an ordinal regression model with the all significant predictors then removing the least significant feature one at a time while retraining and testing the model.

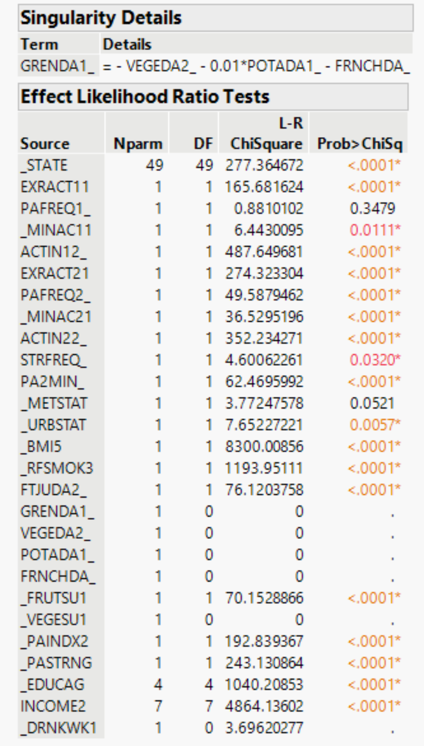
Comparing performance metrics to determine how much of a roll each significant predictor plays in predicting general health.

Predictor Screening

Collinearity between nutrition features rendered all but one useless.

Counterintuitively, primary exercise choice and frequency were weakly linked to general health if at all. Secondary exercise info was more promising.

Income and BMI were, by far and large, the least independent features from the target and expected to have the greatest impact on model performance.

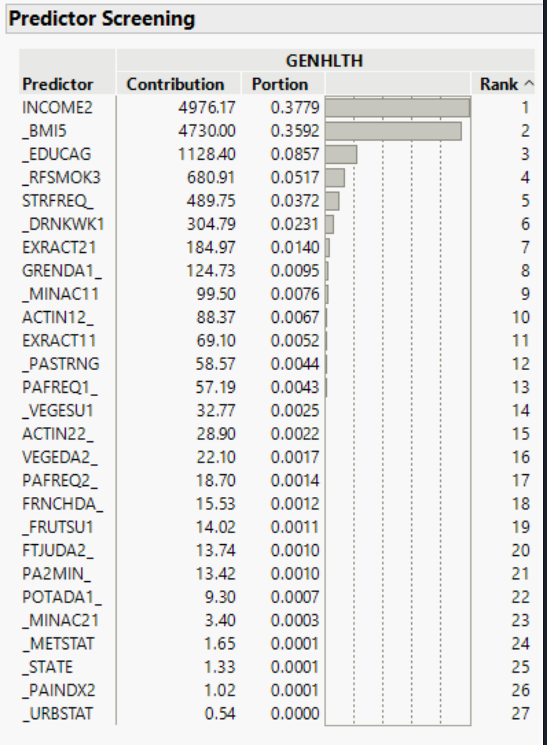


Singularity Details

Term	Details
GRENDA1_	= - VEGEDA2_ - 0.01*POTADA1_ - FRNCHDA_

Effect Likelihood Ratio Tests

			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
_STATE	49	49	277.364672	<.0001*
EXRACT11	1	1	165.681624	<.0001*
PAFREQ1_	1	1	0.8810102	0.3479
_MINAC11	1	1	6.4430095	0.0111*
ACTIN12_	1	1	487.649681	<.0001*
EXRACT21	1	1	274.323304	<.0001*
PAFREQ2_	1	1	49.5879462	<.0001*
_MINAC21	1	1	36.5295196	<.0001*
ACTIN22_	1	1	352.234271	<.0001*
STRFREQ_	1	1	4.60062261	0.0320*
PA2MIN_	1	1	62.4695992	<.0001*
_METSTAT	1	1	3.77247578	0.0521
_URBSTAT	1	1	7.65227221	0.0057*
_BMI5	1	1	8300.00856	<.0001*
_RFSMOK3	1	1	1193.95111	<.0001*
FTJUDA2_	1	1	76.1203758	<.0001*
GRENDA1_	1	0	0	.
VEGEDA2_	1	0	0	.
POTADA1_	1	0	0	.
FRNCHDA_	1	0	0	.
_FRUTSU1	1	1	70.1528866	<.0001*
_VEGESU1	1	0	0	.
_PAINDX2	1	1	192.839367	<.0001*
_PASTRNG	1	1	243.130864	<.0001*
_EDUCAG	4	4	1040.20853	<.0001*
INCOME2	7	7	4864.13602	<.0001*
_DRNKWK1	1	0	3.69620277	.



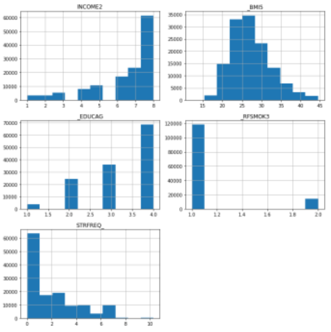
Predictor Screening

	GENHLTH		
Predictor	Contribution	Portion	Rank ^
INCOME2	4976.17	0.3779	1
_BMI5	4730.00	0.3592	2
_EDUCAG	1128.40	0.0857	3
_RFSMOK3	680.91	0.0517	4
STRFREQ_	489.75	0.0372	5
_DRNKWK1	304.79	0.0231	6
EXRACT21	184.97	0.0140	7
GRENDA1_	124.73	0.0095	8
_MINAC11	99.50	0.0076	9
ACTIN12_	88.37	0.0067	10
EXRACT11	69.10	0.0052	11
_PASTRNG	58.57	0.0044	12
PAFREQ1_	57.19	0.0043	13
_VEGESU1	32.77	0.0025	14
ACTIN22_	28.90	0.0022	15
VEGEDA2_	22.10	0.0017	16
PAFREQ2_	18.70	0.0014	17
FRNCHDA_	15.53	0.0012	18
_FRUTSU1	14.02	0.0011	19
FTJUDA2_	13.74	0.0010	20
PA2MIN_	13.42	0.0010	21
POTADA1_	9.30	0.0007	22
_MINAC21	3.40	0.0003	23
_METSTAT	1.65	0.0001	24
_STATE	1.33	0.0001	25
_PAINDX2	1.02	0.0001	26
_URBSTAT	0.54	0.0000	27

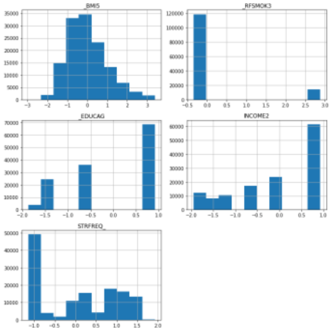
Ordinal Regression: Data Preprocessing

- Features were visualized to determine their distribution. For this that appeared more gaussian, scikit-learns StandardScaler() was used and for those which appeared far from normal had scikit-learns PowerScaler() applied to fit the data closer towards a normal distribution post-analysis.
- The target was fit into a range of -1 to 1
- The data was then split into training and testing sets with a test size of 33%

Pre-Processing



Post-Processing

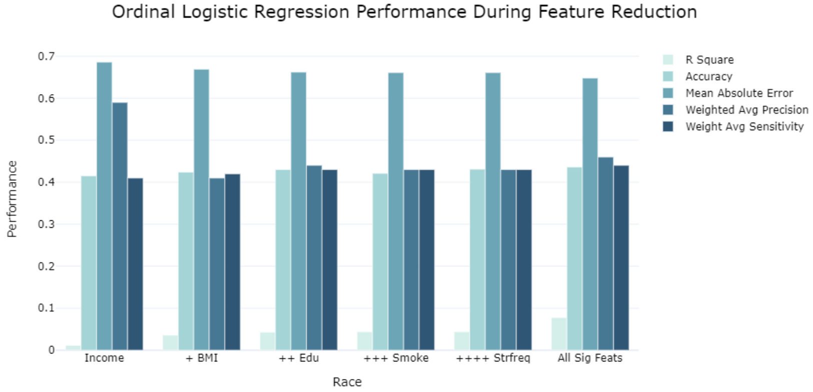


Ordinal Regression: Fitting the model

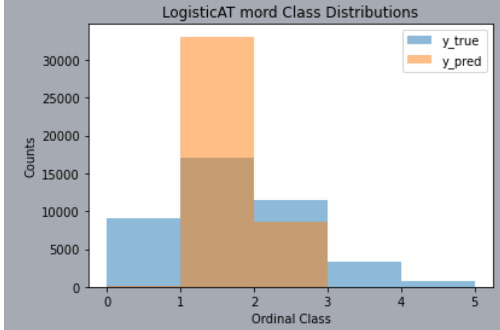
- Target: General health rated on a 5-point Likert scale (ordinal)
- Model: Ordinal Logistic Regression
  - Sci-kit learn does not have a module for performing ordinal logistic regression
  - mord is a python package that implements some ordinal regression methods following the scikit-learn API.
    - mord.LogisticIT (immediate-threshold regression) and mord.LogisticAT (all-threshold) were tried and found to have similar results. LogisticIT was chosen so penalty would relate to distance of y-predictions to actual y values.
  - Scikit learns metric functions were used to produce confusion matrices, classification reports and values for accuracy and mean absolute error so performance could be compared between models.

Ordinal Regression: Results

The addition of features resulted in a small increase of prediction accuracy and reduction of MAE. Precision was highest when only income was used however this is only because no false predictions for classes 1, 4 or 5 were made as a result of those classes being predicted 0 times. as features increase the spread of predictions widens somewhat closer towards the actual spread of y however the predictions consistently cluster towards classes 2 and 3.


Ordinal Logistic Regression Performance During Feature Reduction

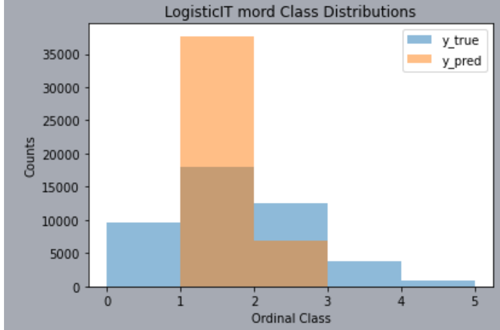
All Significant Features


LogisticAT mord Class Distributions

Confusion Matrix

44	8337	687	1	0
30	14680	2401	4	0
7	8063	3473	3	0
4	1702	1597	8	0
0	312	445	0	0

Only Most Significant Feature, Income


LogisticIT mord Class Distributions

Confusion Matrix

0	8694	847	0	0
0	16140	1876	0	0
0	10096	2346	0	0
0	2357	1342	0	0
0	415	447	0	0

Future Analysis

- Try to improve models able to predict edge cases (classes 1, 4 and 5)
    - Try oversampling to create simulated data points for edge case classifications to see if this improves the models ability to predict such cases.
- Investigate relationship between income and health more closely, look for mechanisms behind the relationship to find actionable steps one could take to improve health that are more accessible than increasing their income or paths for government organizations to increase access to these sources of improved health (e.g., does access to health-care or time for a second source of physical activity increase with income and account for some of the variable of general health?)
- Investigate BMI independently from general health. Use binary logistic regression to identify the factors at play with increased BMI (e.g., does increase physical activity or nutrition increase probability of being obese more? What are the relationships between income and nutrition and exercise?)
- Investigate impact of age on relationships to see if samples older demographic has an impact on findings that doesn't relate to general U.S. population.

What I would have done differently looking back

- Tried out tableau for a more convenient dashboard tool
- Carried out less analysis on individual features until after predictor analysis so I could focus my energy on the features that played the greatest role in predicting general health (for example, I would have focused more on relationships between income and other features and focused more on the secondary source of exercise than the primary source, which I had anticipated being more important).
- Used R instead of python as the documentation for MASS is much more informative than that for mord and it appears to have a more active user-base.