Longevity & Behavioral Factors by State

Kandace Korver
Roberta Lee

January 11, 2020

U.S. Longevity & Behavioral Factors

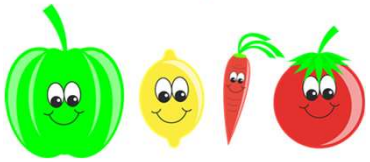
Alcohol

Physical Activity

Cigarettes

Obesity

Fruits/Vegetables

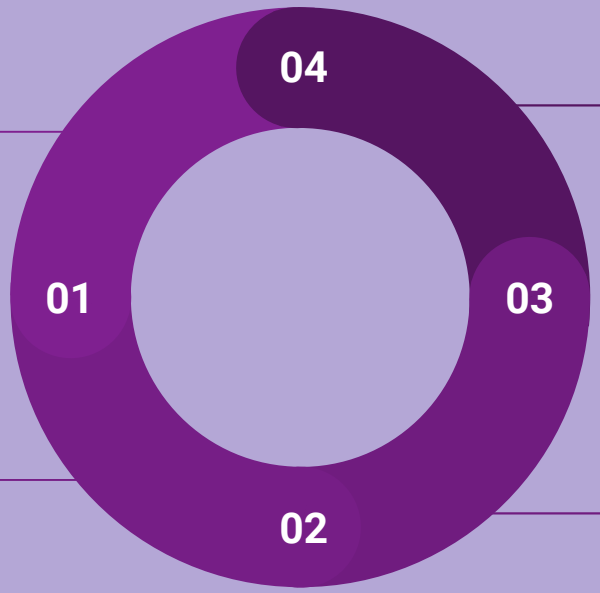


?

The Process

1. Find Datasets

2. Data Cleaning



4. Findings

3. Analysis

Data Sources

KFF Filling the need for trusted information on national health issues

Life expectancy at birth https://www.kff.org/other/stateindicator/lifeexpectancy/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

World Population Review

Alcohol consumption by state http://worldpopulationreview.com/states/alcohol-consumption-by-state/

Obesity by state http://worldpopulationreview.com/states/most-obese-states/

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Cigarette use by state https://www.cdc.gov/statesystem/cigaretteuseadult.html

Nutrition, physical activity by state https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7

Simple Data Cleaning 😊

Longevity

	Location	Life Expectancy at Birth (years)	Footnotes
0	United States	78.7	NaN
1	Alabama	75.5	NaN
2	Alaska	78.8	NaN
3	Arizona	79.9	NaN
4	Arkansas	76.0	NaN
...	...	...	...
57	Arias E, Escobedo LA, Kennedy J, Fu C, Cisewsk...	NaN	NaN
58	NaN	NaN	NaN
59	NaN	NaN	NaN
60	Footnotes	NaN	NaN
61	1. Data for Maine and Wisconsin are not availa...	NaN	NaN

Obesity

	State	obesityRank	obesityPercentage	Pop
0	West Virginia	1	0.381	1791951
1	Mississippi	2	0.373	2987895
2	Oklahoma	3	0.365	3948950
3	Iowa	4	0.364	3167997

✔ Read file:

```
life_expectancy = "./life_expectancy_by_state.csv"
```

```
life_expectancy_pd = pd.read_csv(life_expectancy)
life_expectancy_pd.head()
```

✔ Remove unnecessary columns:

```
obesity = obesity.drop(['obesityRank', 'Pop'], axis = 1)
```

✔ Rename columns so tables match

```
#rename columns
cigarette_pd.rename(columns = {'Location Desc':'State',
                               'Data Value':'cigarettePercent'}, inplace = True)
```

✔ Sort by state, Reset index:

```
alcohol_pd = alcohol_pd.sort_values(by = 'State')
alcohol_pd = alcohol_pd.reset_index(drop = True)
alcohol_pd
```

✔ Repeat process: Longevity, Obesity, Alcohol, Cigarettes

Complex Data Cleaning

Nutrition and Exercise (63028 x 33)

Multiple years per state

Isolate by class: (physical activity, fruits/vegetables)

Remove unnecessary columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	YearStart	YearEnd	LocationA	LocationD	atasource	Class	Topic	Question	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Low_Conf	High_Con	Sample_S	Total	Age(years	Education	Gender	Income	Race/Eth	GeoLocat
2	20	201	WY	Wyoming	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	48.5	48.5			32.3	64.9	69						American Inc	(43.2355413
3	20	201	DC	District of Co	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	31.6	31.6			24	40.4	243			Less than high school				(38.8903713
4	20	201	AL	Alabama	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	35.2	35.2			30.7	40	598		25 - 34					(32.8405711
5	20	201	US	National	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	27.9	27.9			27.6	28.3	266452				Female			
6	20	201	US	National	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	16.9	16.9			16	17.8	20923		18 - 24					
7	20	201	US	National	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	22.1	22.1			21.4	22.8	45883		25 - 34					
8	20	201	US	National	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	28.1	28.1			27.5	28.6	109526		55 - 64					
9	20	201	RI	Rhode Island	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	40.2	40.2			33.3	47.4	354						Hispanic	(41.7082801
10	20	201	WY	Wyoming	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	32.3	32.3			25.6	39.8	484					Less than $15,000		(43.2355413
11	20	201	MN	Minnesota	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o achieve at	Value	52.8	52.8			49	56.6	3680		65 or older					(46.3556487
12	20	201	WA	Washington	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	39.5	39.5			37.5	41.5	5835				Male			(47.5222786
13	20	201	GA	Georgia	ehavioral Risk Fa	Obesity / W	Obesity / We	Percent of adults ag	d 18 years a	Value	24.6	24.6			21.3	28.2	1041					$75,000 or greater		(32.8396810
14	20	201	WI	Wisconsin	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o achieve at	Value	19.1	19.1			13.3	26.6	412					Less than $15,000		(44.3931911
15	20	201	DE	Delaware	ehavioral Risk Fa	Physical Act	Physical Acti	Percent of adults w	o engage in i	Value	24.4	24.4			20.2	29	707					$50,000 - $74,999		(39.0088306

- ✔ Read file, remove unnecessary columns
- ✔ Rename columns, reset index, sort by state
- ✔ Remove unnecessary rows (National, classes)
- ✔ Isolate by physical activity, fruits/vegetables
- ✔ Calculate weighted averages: multiple years per state with varying sample sizes

Complex Data Cleaning 😭

Nutrition and Exercise

Multiple values for every state

Different Sample Sizes

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
YearStart	YearEnd	LocationA	LocationD	Datasource	Class	Topic	Question	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Data_Vali	Low_Conf	High_Con	Sample_S
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		19.5	19.5		✖	18.2	20.8	6099
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		32.1	32.1			29	35.4	1319
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		17.7	17.7			15.8	19.9	1979
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		19.2	19.2			15.9	22.9	982
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		10.4	10.4			9.1	11.8	3962
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		32.4	32.4			29.4	35.6	1551
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		25.3	25.3			23	27.8	2297
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		20	20			17.8	22.4	2005
2018	2018	CA	California	Behavioral Risk Factc	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physica	Value		18.5	18.5			16.8	20.4	2921

Weighted Average

- To calculate a single summary value for each state, calculated the weighted average over all rows with varying sample sizes for each state
  - e.g. if 2 people have avg age 60 and 3 people have avg age 40, the avg age of all 5 people is not 50 ((60+40) / 2)
  - Instead: it is ((60*2+40*3) / (2+3)) = 48, i.e. the weighted average.
- In data set:
  - (K2*Q2 + K3*Q3…..) / sum of sample size (Q2+Q3+Q4…)

Calculating Weighted Average

Start with Isolated Class and Year
- ➔ multiple values per state
- ➔ differing sample sizes per value

	YearEnd	State	Class	Question	Data_Value	Sample_Size
727	2018	California	Physical Activity	Percent of adults who engage in no leisure-tim...	19.5	6099.0
8379	2018	California	Physical Activity	Percent of adults who engage in no leisure-tim...	32.1	1319.0
8900	2018	California	Physical Activity	Percent of adults who engage in no leisure-tim...	17.7	1979.0
9194	2018	California	Physical Activity	Percent of adults who engage in no leisure-tim...	19.2	982.0
9458	2018	California	Physical Activity	Percent of adults who engage in no leisure-tim...	10.4	3962.0

Weighted Average Function

```
# Lets calculate a weighted average bec

# Weighted average function
def wavg(group, data_value_col, sample_size_col):
    data_values = group[data_value_col]
    sample_sizes = group[sample_size_col]
    return (data_values * sample_sizes).sum() / sample_sizes.sum()


# Calculate weighted average for physical activity
physical_wa_by_state = nutrition_physical.groupby("State").apply(wavg, "Data_Value", "Sample_Size")
```

```
# Convert series to dataframe
physical_wa_df = pd.DataFrame({'State':physical_wa_by_state.index,
                               'infrequentPhysicalActivityPercent':physical_wa_by_state.values})
physical_wa_df = physical_wa_df.round(2) # round to 2 decimal places
physical_wa_df
```

Convert Series to Data Frame

	State	physicalActivityPercent
0	Alabama	69.12
1	Alaska	80.33
2	Arizona	77.67
3	Arkansas	68.28
4	California	79.22

✔ Repeated process twice:
- Physical Activity, Fruits/Vegetables

Analyses Steps

1. Merge Tables: pd.merge(life_expectancy_pd, factor, how='outer', on = 'State', indicator = True)
   - Merged table: factor values and longevity by state

2. Remove NaN values: .dropna()
   - states with missing data (e.g. Maine, Wisconsin)
   - non-states included in data (e.g. Guam, Puerto Rico)

```
merged_obesity_table = pd.merge(life_expectancy_pd, obesity, how='outer', on='State', indicator=True)
merged_obesity_table = merged_obesity_table.dropna() # drop rows with NaN values
merged_obesity_table
```

3. Scatterplot

4. Summary Table

Merged Table

	State	Longevity	obesityPercentage	_merge
0	Alabama	75.5	0.363	both
1	Alaska	78.8	0.342	both
2	Arizona	79.9	0.295	both
3	Arkansas	76.0	0.350	both
4	California	81.3	0.251	both
5	Colorado	80.5	0.226	both
6	Connecticut	80.9	0.269	both
7	Delaware	78.7	0.318	both
9	Florida	80.1	0.284	both
10	Georgia	77.7	0.316	both
11	Hawaii	82.0	0.238	both
12	Idaho	79.4	0.293	both
13	Illinois	79.3	0.311	both

Findings Along the Way

1. Create functions for repeated processes

Scatter Plot:
Creating graphs

```
def scatterPlotAndLine(x, y, title, y_label, x_label): # defining a function to allow code re-use
    stats = linregress(x, y)

    m = stats.slope
    b = stats.intercept
    y_intercept_sign = "+" if b >= 0 else "-"
    line_eq = "y = " + str(round(stats.slope, 2)) + "x " + y_intercept_sign + " " + str(abs(round(stats.intercept, 2))

    plt.scatter(x, y)
    plt.plot(x, m * x + b, color="red")

    plt.title(title)
    plt.ylabel(y_label)
    plt.xlabel(x_label)
    plt.annotate(line_eq, fontsize=12, color="red", xy=(0.5, 0.8), xycoords='axes fraction')

scatterPlotAndLine(x=merged_table.Longevity, y=merged_table.alcoholConsumptionGallons,
                   title="Alcohol and Life Expectancy by State", y_label="Alcohol Consumption (gallons)",
                   x_label="Life Expectancy at Birth (years)")
```

Summary Table:
Compute statistical values

```
def computeModelSummary(x, y):
    # Note the difference in argument order
    model = sm.OLS(y, x).fit()
    predictions = model.predict(x) # make the predictions by the model

    # Print out the statistics
    return model.summary()

computeModelSummary(x=merged_table.Longevity, y=merged_table.alcoholConsumptionGallons)
```

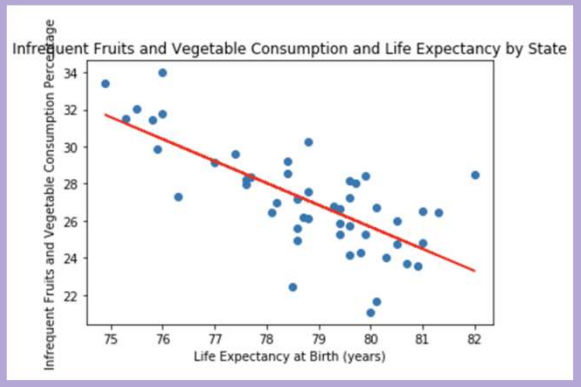
✓ Repeated for the 5 factors

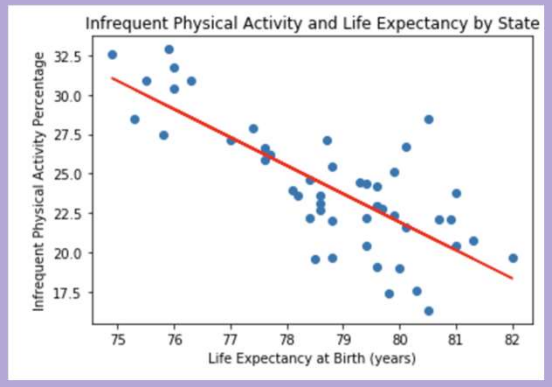
Findings Along the Way

2. Inverted Values for Physical Activity & Nutrition data

At first, calculated negative correlations for physical activity and fruits/veg consumption

% of adults who report consuming vegetables less than one time daily





% of adults who engage in no leisure-time physical activity

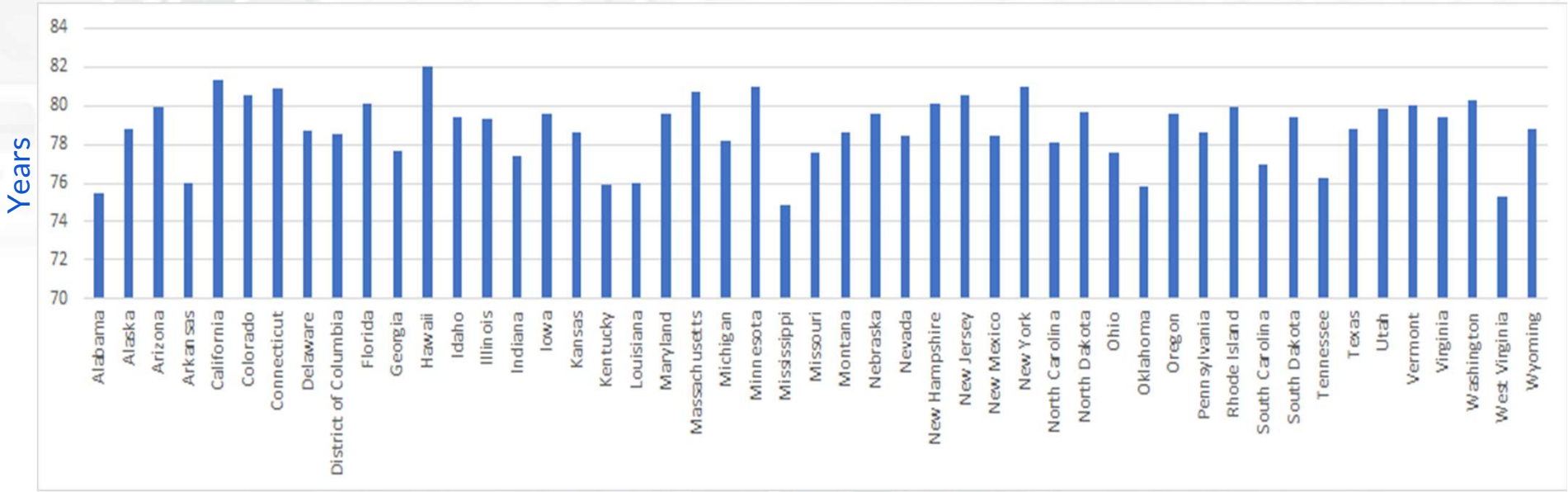
✓ Invert values to show frequency percentages: .map(lambda x: 100 -x)

```
# Convert the inverted percent (people who don't exercise frequently) to the regular percent (people who do exercise
physical_wa_df.rename(columns = {'infrequentPhysicalActivityPercent':'physicalActivityPercent'}, inplace = True)
physical_wa_df['physicalActivityPercent'] = physical_wa_df['physicalActivityPercent'].map(lambda x: 100-x)
physical_wa_df
```


RESULTS

Life Expectancy at Birth By State

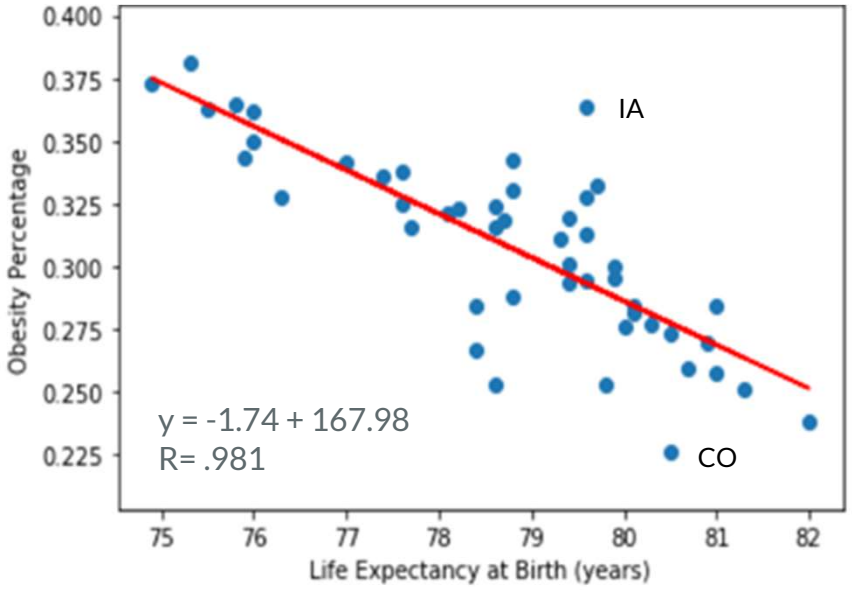
	State	Longevity
0	Hawaii	82.0
1	California	81.3
2	Minnesota	81.0
3	New York	81.0
4	Connecticut	80.9



Obesity and Longevity

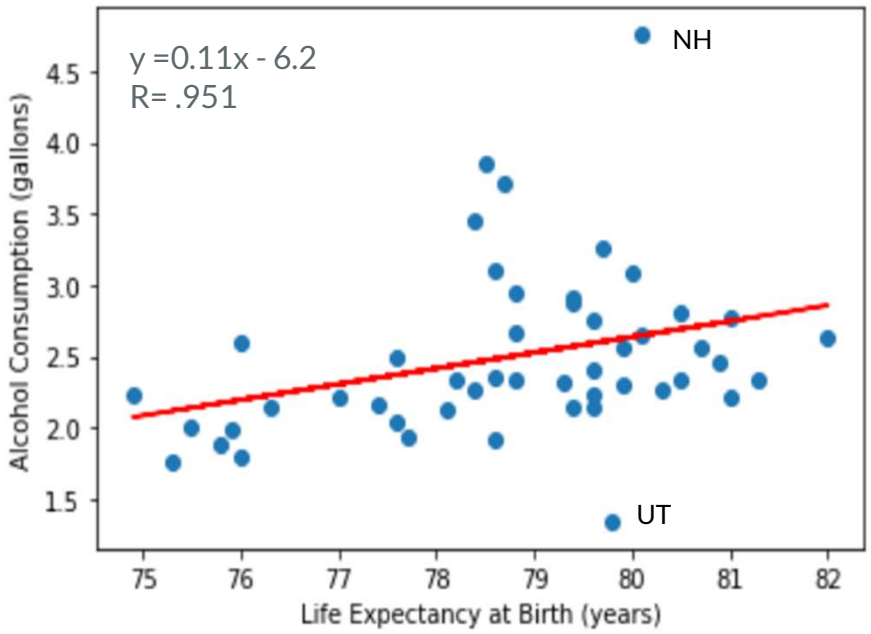
Body Mass Index >= 30 (normal 18.5 - 24.9)

	State	Longevity	obesityPercentage	_merge
0	Alabama	75.5	0.363	both
1	Alaska	78.8	0.342	both
2	Arizona	79.9	0.295	both
3	Arkansas	76.0	0.350	both
4	California	81.3	0.251	both
5	Colorado	80.5	0.226	both
6	Connecticut	80.9	0.269	both
7	Delaware	78.7	0.318	both
9	Florida	80.1	0.284	both
10	Georgia	77.7	0.316	both



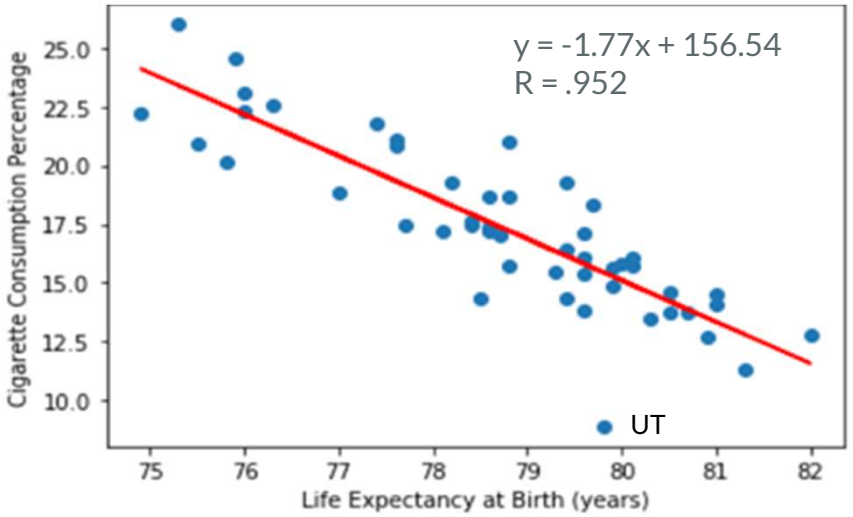
Alcohol and Longevity

	State	Longevity	alcoholConsumptionGallons	_merge
0	Alabama	75.5	2.01	both
1	Alaska	78.8	2.94	both
2	Arizona	79.9	2.31	both
3	Arkansas	76.0	1.80	both
4	California	81.3	2.33	both
5	Colorado	80.5	2.81	both
6	Connecticut	80.9	2.45	both
7	Delaware	78.7	3.72	both
8	District of Columbia	78.5	3.85	both
9	Florida	80.1	2.65	both
10	Georgia	77.7	1.94	both



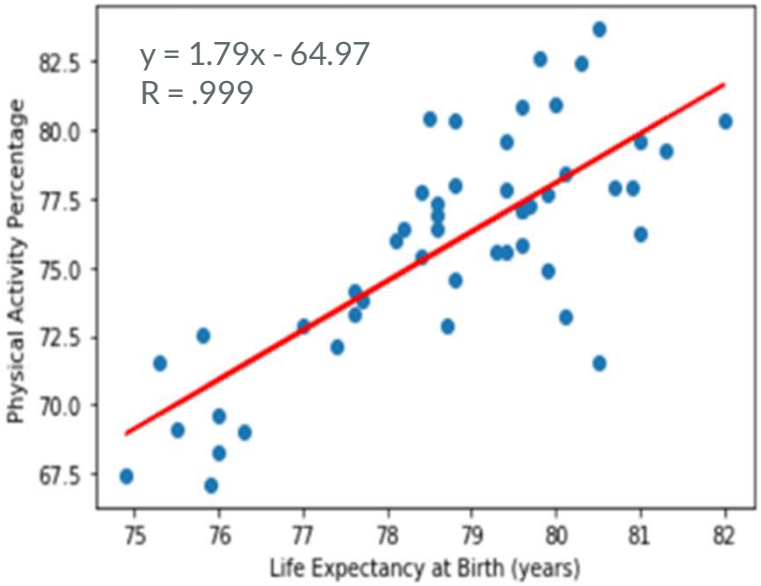
Cigarette Use and Longevity

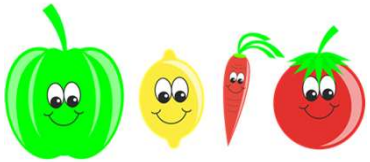
	State	Longevity	cigarettePercent	_merge
0	Alabama	75.5	20.9	both
1	Alaska	78.8	21.0	both
2	Arizona	79.9	15.6	both
3	Arkansas	76.0	22.3	both
4	California	81.3	11.3	both
5	Colorado	80.5	14.6	both
6	Connecticut	80.9	12.7	both
7	Delaware	78.7	17.0	both
8	District of Columbia	78.5	14.3	both
9	Florida	80.1	16.1	both
10	Georgia	77.7	17.5	both



Physical Activity and Longevity

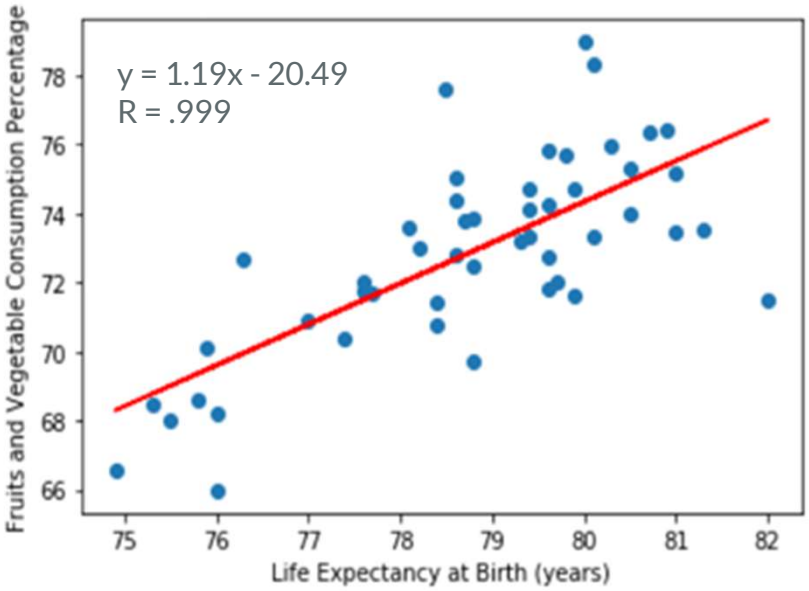
	State	Longevity	physicalActivityPercent	_merge
0	Alabama	75.5	69.12	both
1	Alaska	78.8	80.33	both
2	Arizona	79.9	77.67	both
3	Arkansas	76.0	68.28	both
4	California	81.3	79.22	both
5	Colorado	80.5	83.66	both
6	Connecticut	80.9	77.94	both
7	Delaware	78.7	72.91	both
8	District of Columbia	78.5	80.40	both
9	Florida	80.1	73.26	both
10	Georgia	77.7	73.80	both





Fruits, Vegetables and Longevity

	State	Longevity	fruitsVegetablesConsumptionPercent	_merge
0	Alabama	75.5	67.98	both
1	Alaska	78.8	72.46	both
2	Arizona	79.9	71.60	both
3	Arkansas	76.0	68.23	both
4	California	81.3	73.53	both
5	Colorado	80.5	75.28	both
6	Connecticut	80.9	76.43	both
7	Delaware	78.7	73.81	both
8	District of Columbia	78.5	77.58	both
9	Florida	80.1	73.31	both
10	Georgia	77.7	71.67	both



Findings of Regression Equations

Correlation

+

STRONG

Physical Activity

$y = 1.79x - 64.97$

Fruits, Vegetables

$y = 1.19x - 20.49$

WEAK

Alcohol

$y = 0.11x - 6.2$

−

STRONG

Cigarettes

$y = -1.77x + 156.54$

Obesity

$y = -1.74x + 167.98$

Does not consider hereditary , environmental or socioeconomic factors

Caveats in Data and Analyses

- Differing Time Periods
  - Life expectancy at birth based on pooled data from 2011 - 2015
  - Alcohol consumption 2019
  - Obesity 2019
  - Cigarettes 2019
  - Nutrition and Exercise 2019

- Surveyed Data

  Behavioral Risk Factor Surveillance System

  Annual telephone survey of 400,000 people in all 50 states