

# **PHYSICAL AND SOCIAL PREDICTORS OF HEALTH OUTCOMES**

BY ELIZABETH VANDERGRIFT, JUAN  
CAMILO BOHORQUEZ ROZO, AND ERIC  
CROSTON

# PHYSICAL AND SOCIAL PREDICTORS OF HEALTH OUTCOMES

Data Review

Correlations

Machine Learning

Visualizations

# SOCIAL DETERMINANTS OF HEALTH

Data Source:  
US Department of Health and Human  
Services and CDC



PLACES: County Data 2023 release  
[https://data.cdc.gov/500-Cities-Places/PLACES-County-Data-GIS-Friendly-Format-2023-releas/7cmc-7y5g/about\\_data](https://data.cdc.gov/500-Cities-Places/PLACES-County-Data-GIS-Friendly-Format-2023-releas/7cmc-7y5g/about_data)

## INFORMATION ABOUT DATA USED IN THIS PROJECT

**PLACES:** County Data 2023 release :

**Covers the entire United States, by state and county level**

- Centers for Disease Control and Prevention (CDC)- Division of Population Health, Epidemiology and Surveillance Branch.

**Data sources:**

- **System (BRFSS)** 2021 or 2020 data  
**Behavioral Risk Factor Surveillance**
- **Census Bureau** 2021 or 2020 county population estimates
- **American Community Survey (ACS)** 2017–2021 or 2016–2020 estimates

<http://www.cdc.gov/nccdphp/dph/>

## Social Determinants of Health



Social Determinants of Health  
Copyright-free

 Healthy People 2030

**Healthy People 2030, U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. Retrieved [date graphic was accessed], from <https://health.gov/healthypeople/objectives-and-data/social-determinants-health>**

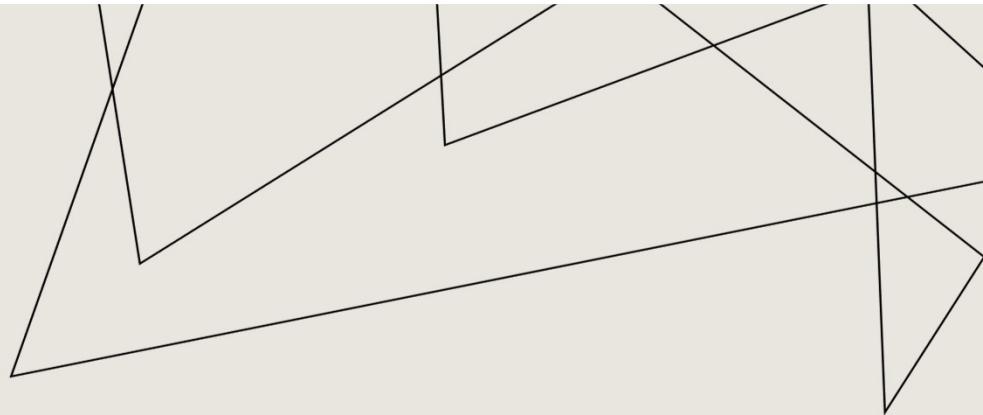


WHAT ARE FACTORS  
THAT COULD LEAD  
TO CHRONIC  
DISEASE AND  
DISABILITY?

HOW CAN WE  
IMPROVE OUR  
HEALTH OUTCOMES  
IN THE UNITED  
STATES?

# SOCIAL DETERMINANTS OF HEALTH

**The nonmedical factors that influence health outcomes**



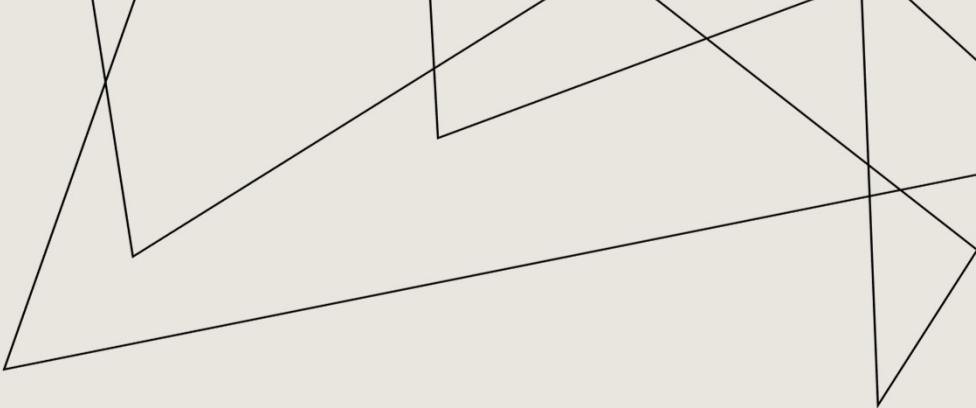
Measure_ID	Measure_Short_Name	Measure
HCOST	Housing cost burden	Housing cost burden among households
POV150	Poverty	Persons living below 150% of the poverty level
CROWD	Crowding	Crowding among housing units
AGE65	Aged 65 years or older	Persons aged 65 years or older
SNGPNT	Single-parent households	Single-parent households
BROAD	No broadband	No broadband internet subscription among households
NOHSDP	No high school diploma	No high school diploma among adults aged 25 years or older
UNEMP	Unemployment	Unemployment among people 16 years and older in the labor force
REMNRITY	Racial or ethnic minority status	Persons of racial or ethnic minority status

**Housing, Income, Education and overall conditions of a person's environment can impact their functioning, quality of life and health outcomes**

## FOOD ACCESS

- Food Insecurity and Malnutrition
- Impact of Food Deserts
- Obesity and Chronic Disease
- Cognitive Impacts and Mental Health

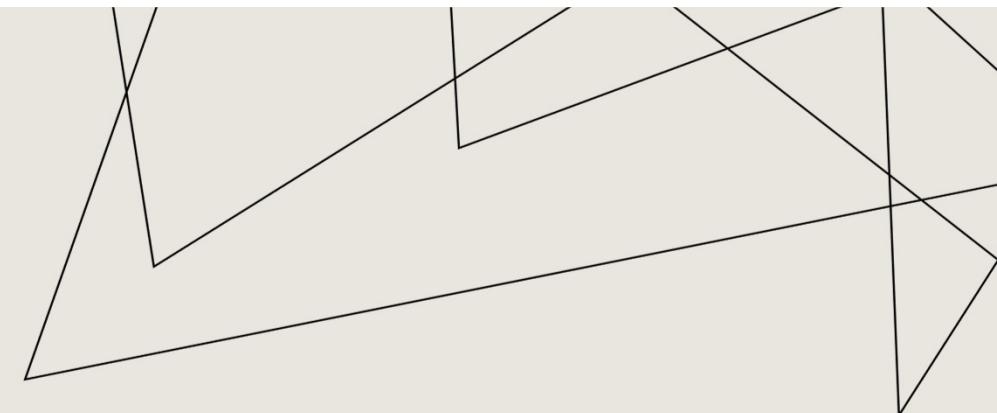
Food Access Data
No_Vehicle_Half_Mile
No_Vehicle_1_Mile
No_Vehicle_10_Miles
No_Vehicle_20_Miles
Children_Half_Mile
Children_1_Mile
Children_10_Miles
Children_20_Miles
Low_Income_Half_Mile
Low_Income_1_Mile
Low_Income_10_Miles
Low_Income_20_Miles
Population_Half_Mile
Population_1_Mile
Population_10_Miles
Population_20_Miles
Seniors_Half_Mile
Seniors_1_Mile
Seniors_10_Miles
Seniors_20_Miles



**Access to Food and  
Proper Nutrition Impacts  
Health Outcomes**

## WHAT ARE CHRONIC HEALTH CONDITIONS, DISEASES AND DISABILITIES THAT COULD BE INFLUENCED BY SDOH

<b>CODE</b>	<b>Health Measure- Chronic Disease, Disability, Negative Health Behavior</b>
ARTHRITIS	Prevalence Arthritis
BINGE	Prevalence Binge Drinking
BPHIGH	Prevalence High Blood Pressure
CANCER	Prevalence Cancer
CASTHMA	Prevalence Asthma
CHD	Prevalence Coronary Heart Disease
COPD	Prevalence COPD- Chronic Obstructive Coronary Disease
CSMOKING	Prevalence Current Smoking
DEPRESSION	Prevalence Depression
DIABETES	Prevalence Diabetes
GHLTH	Prevalence Fair or Poor Health
HIGHCHOL	Prevalence High Cholesterol
KIDNEY	Prevalence Chronic Kidney Disease
LPA	Prevalence No Leisure- Time Physical Activity
MHLTH	Prevalence Mental Health not good for 14 or more days
OBESITY	Prevalence Obesity
PHLTH	Prevalence Physical Health not good for 14 or more days
SLEEP	Prevalence Sleeping Less than 7 hours
STROKE	Prevalence Stroke
TEETHLOST	Prevalence All Teeth Lost
HEARING	Prevalence Hearing Disability
VISION	Prevalence Vision Disability
COGNITION	Prevalence Cognitive Disability
MOBILITY	Prevalence Mobility Disability
SELF CARE	Prevalence Selfcare Disability
INDEPLIVE	Prevalence Independent Living Disability
DISABILITY	Prevalence Any Disability



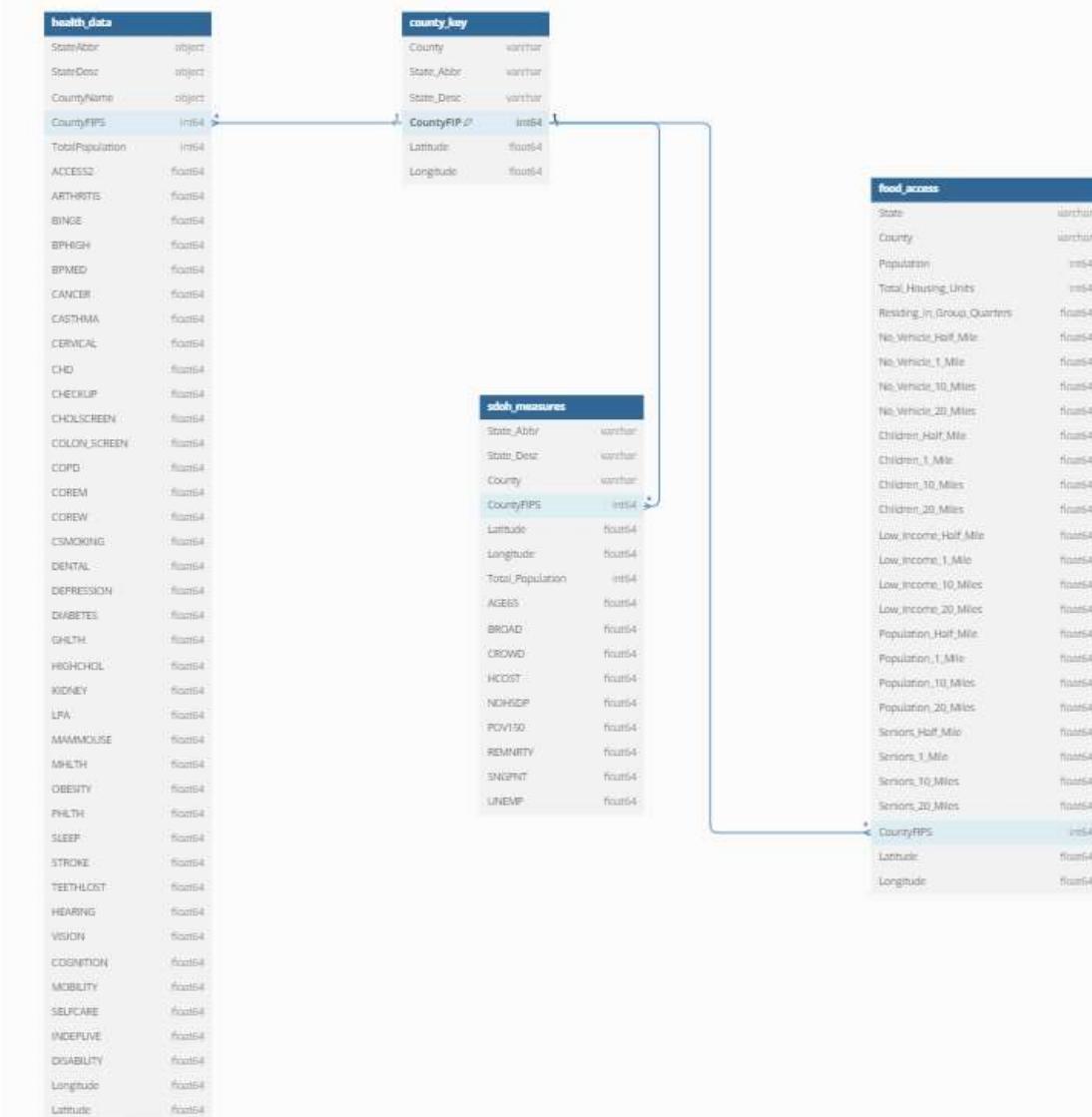
## Is Health Screening and Prevention Impacted by SDOH?

<b>CODE</b>	<b>Health Measure- Screening and Prevention</b>
ACCESS2	Prevalence of Adults with Health Insurance 18-64
BPMED	Prevalence of Taking Medication for High Blood Pressure
CERVICAL	Prevalence Cervical Cancer Screening
CHECKUP	Prevalence Routine Checkup within Past Year
CHOLSCREEN	Prevalence Cholesterol Screening
COLON_SCREEN	Prevalence Colon Screening
COREM	Prevalence Older Men Core set of Clinical Preventative Service
COREW	Prevalence Older Women Core set of Clinical Preventative Service
DENTAL	Prevalence Dental Visits
MAMMOUSE	Prevalence Mammography women within 50-74 Years

# CONNECTING THE DATA IN POSTGRES SQL DATABASE

## ENTITY RELATIONSHIP DIAGRAM

County FIP- A specific ID for each State and County used in government datasets



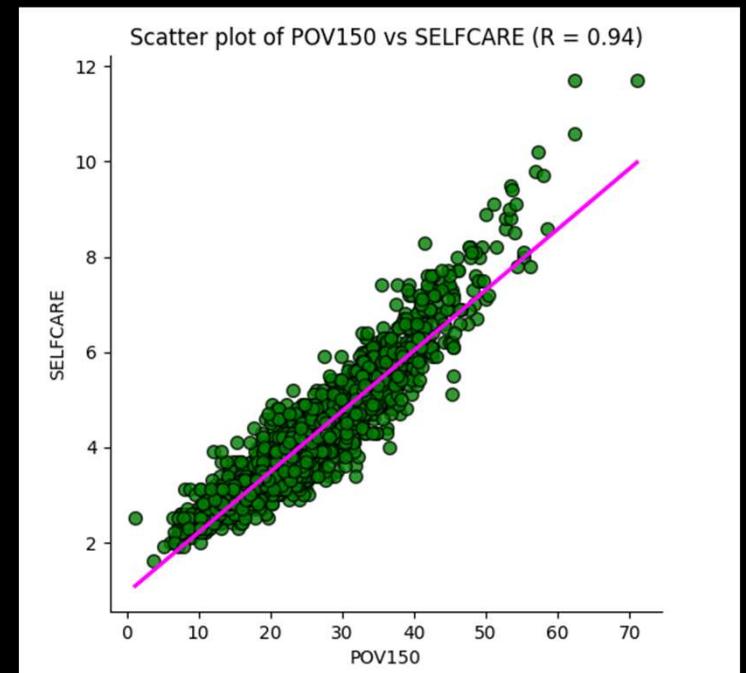
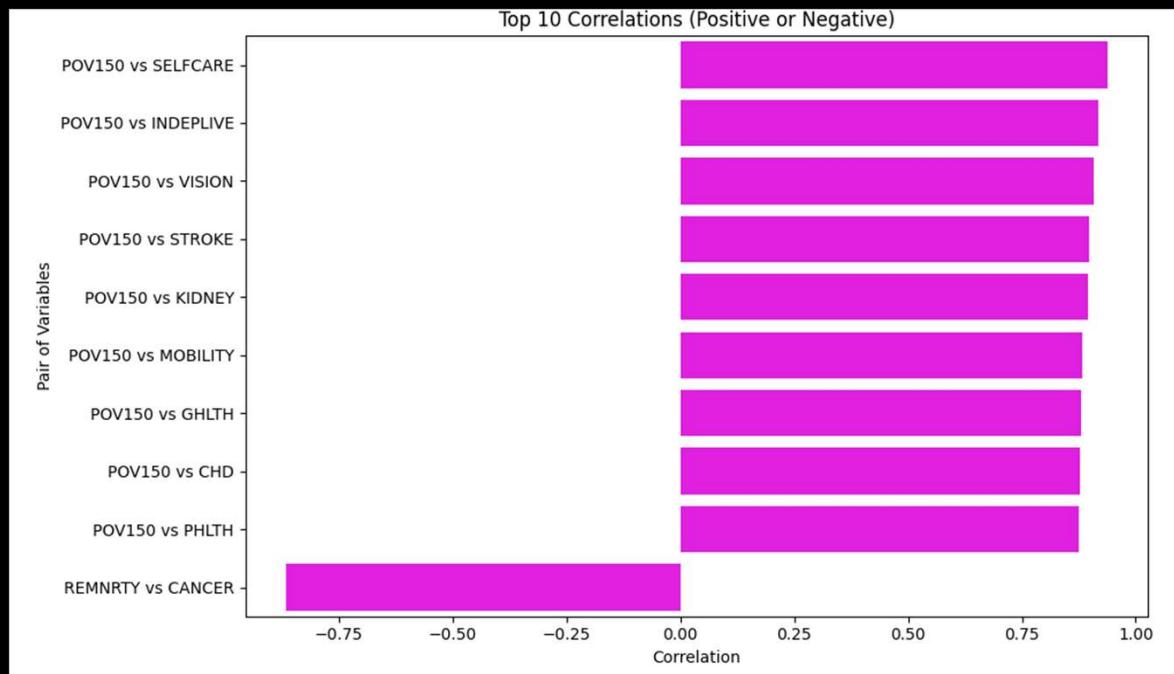
# CORRELATIONS SOCIAL DETERMINANTS OF HEALTH OUTCOMES

## 10 Strongest and Weakest Correlations

10 Strongest Correlations (Positive or Negative):		
	Pair	Correlation
152	POV150 vs SELFCARE	0.937454
153	POV150 vs INDEPLIVE	0.918315
149	POV150 vs VISION	0.908080
146	POV150 vs STROKE	0.896523
140	POV150 vs KIDNEY	0.895394
151	POV150 vs MOBILITY	0.883747
138	POV150 vs GHLTH	0.881256
130	POV150 vs CHD	0.876814
144	POV150 vs PHLTH	0.875616
159	REMNRTY vs CANCER	-0.865038

10 Weakest Correlations:		
	Pair	Correlation
68	HCOST vs CHD	0.001094
297	Residing_in_Group_Quarters vs MHLTH	-0.005350
74	HCOST vs DEPRESSION	0.006388
165	REMNRTY vs CSMOKING	0.006587
162	REMNRTY vs COPD	-0.013162
69	HCOST vs COPD	-0.013740
79	HCOST vs LPA	-0.015794
279	Residing_in_Group_Quarters vs ARTHRITIS	0.018980
280	Residing_in_Group_Quarters vs BINGE	-0.019216
198	SNGPNT vs DEPRESSION	-0.020138

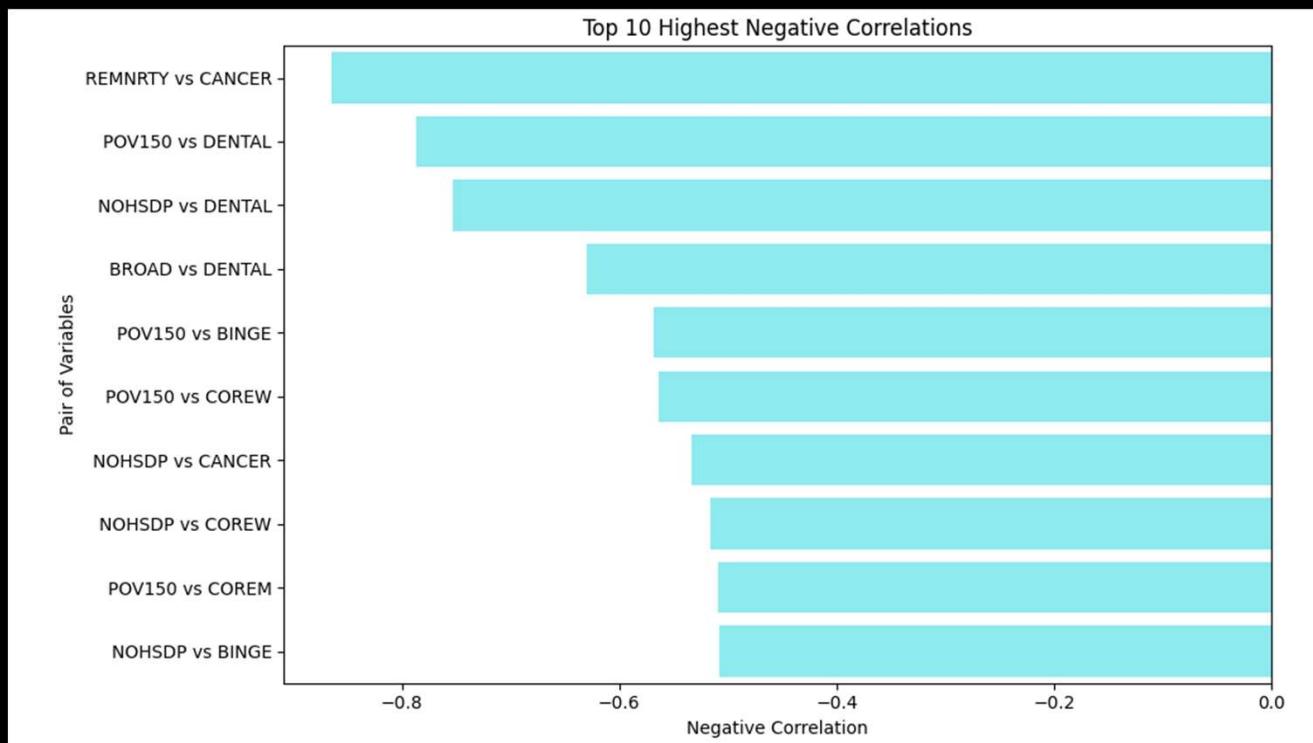
For a single person in 2024, 150% below FPL is \$14,580 and 150% of that would be \$21,870



# CORRELATIONS SOCIAL DETERMINANTS OF HEALTH OUTCOMES

## 10 Strongest Negative Correlations

Top 10 Highest Negative Correlations:			
	Pair	Correlation	abs_corr
159	REMNRITY vs CANCER	-0.865038	0.865038
135	POV150 vs DENTAL	-0.786584	0.786584
104	NOHSDP vs DENTAL	-0.752744	0.752744
11	BROAD vs DENTAL	-0.630112	0.630112
125	POV150 vs BINGE	-0.568558	0.568558
133	POV150 vs COREW	-0.563635	0.563635
97	NOHSDP vs CANCER	-0.533278	0.533278
102	NOHSDP vs COREW	-0.515512	0.515512
132	POV150 vs COREM	-0.509307	0.509307
94	NOHSDP vs BINGE	-0.508078	0.508078



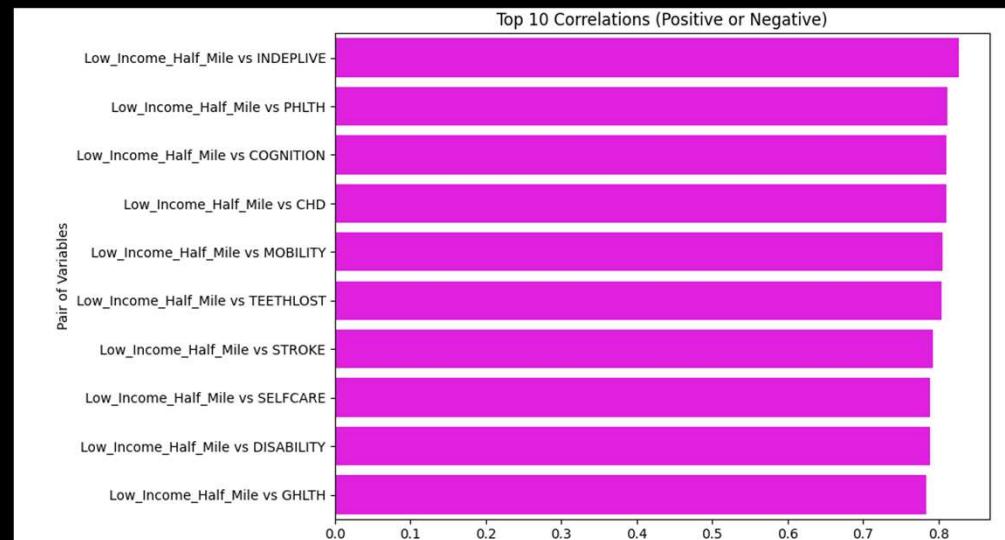
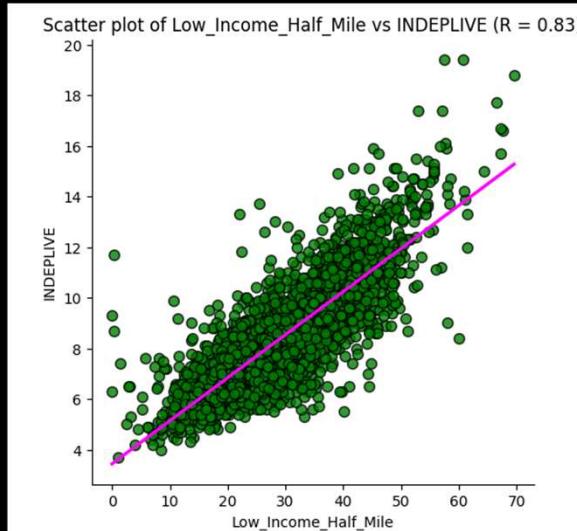
# CORRELATIONS FOOD ACCESS AND HEALTH OUTCOMES

## 10 Strongest Correlations

10 Strongest Correlations (Positive or Negative):		
	Pair	Correlation
277	Low_Income_Half_Mile vs INDEPLIVE	0.826557
268	Low_Income_Half_Mile vs PHLTH	0.812193
274	Low_Income_Half_Mile vs COGNITION	0.811168
254	Low_Income_Half_Mile vs CHD	0.810643
275	Low_Income_Half_Mile vs MOBILITY	0.805490
271	Low_Income_Half_Mile vs TEETHLOST	0.804450
270	Low_Income_Half_Mile vs STROKE	0.792466
276	Low_Income_Half_Mile vs SELFCARE	0.789198
278	Low_Income_Half_Mile vs DISABILITY	0.788572
262	Low_Income_Half_Mile vs GHLTH	0.783908

10 Weakest Correlations:		
	Pair	Correlation
315	Low_Income_10_Miles vs CASTHMA	0.000107
575	Seniors_10_Miles vs LPA	-0.000756
237	Children_20_Miles vs PHLTH	0.001250
524	Seniors_Half_Mile vs SELFCARE	0.001476
600	Seniors_20_Miles vs DENTAL	-0.002221
351	Low_Income_20_Miles vs CSMOKING	-0.002228
63	No_Vehicle_10_Miles vs BINGE	-0.002385
62	No_Vehicle_10_Miles vs ARTHRITIS	0.002415
367	Low_Income_20_Miles vs COGNITION	0.002515
614	Seniors_20_Miles vs VISION	-0.002725

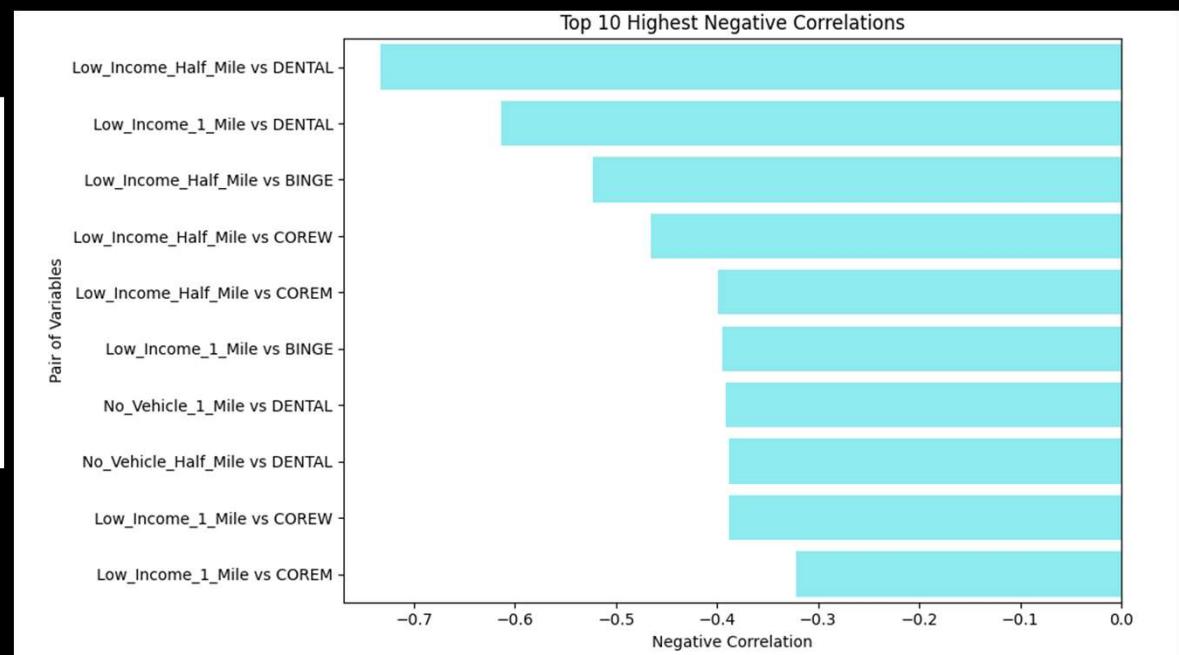


# CORRELATIONS FOOD ACCESS AND HEALTH OUTCOMES

## 10 Strongest Negative Correlations

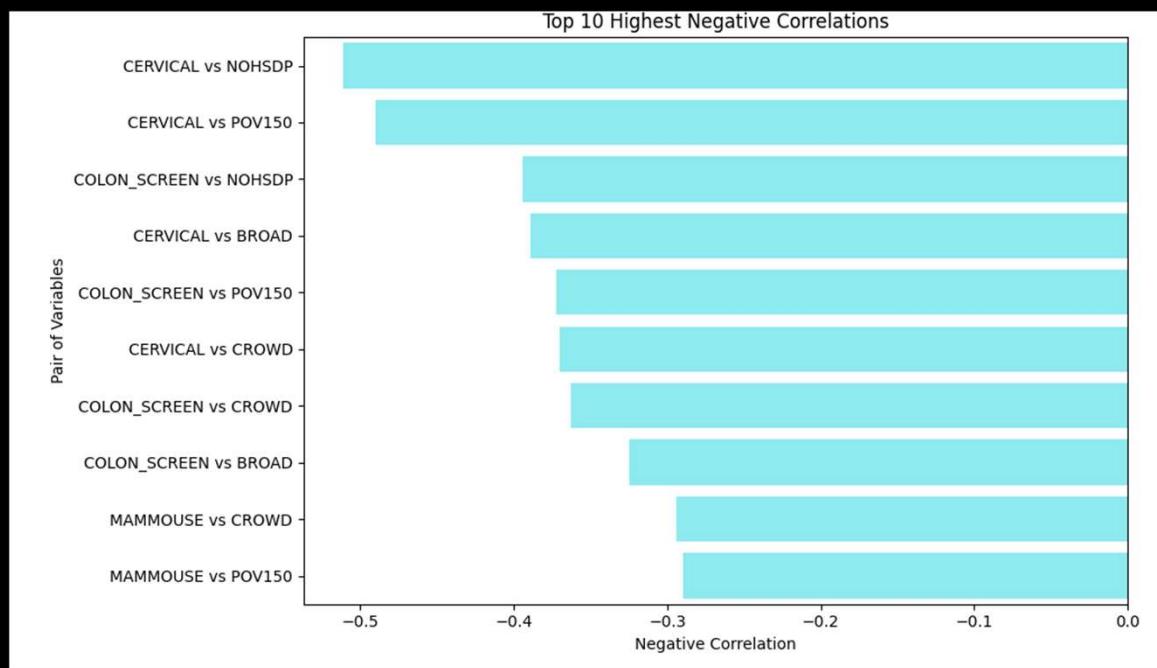
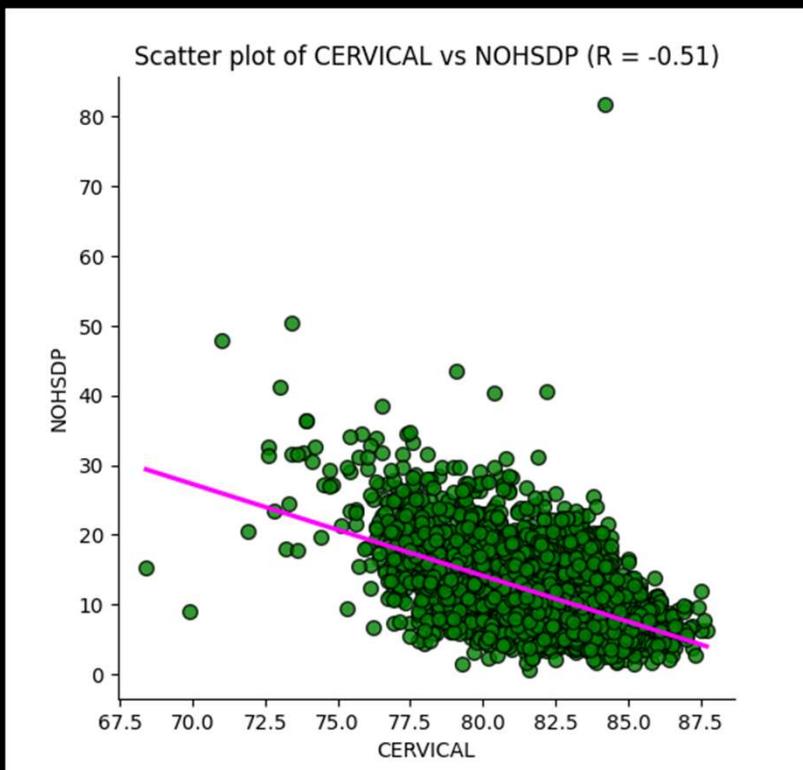
### Top 10 Highest Negative Correlations:

	Pair	Correlation	abs_corr
259	Low_Income_Half_Mile vs DENTAL	-0.732625	0.732625
290	Low_Income_1_Mile vs DENTAL	-0.613549	0.613549
249	Low_Income_Half_Mile vs BINGE	-0.523113	0.523113
257	Low_Income_Half_Mile vs COREW	-0.465940	0.465940
256	Low_Income_Half_Mile vs COREM	-0.398940	0.398940
280	Low_Income_1_Mile vs BINGE	-0.395179	0.395179
42	No_Vehicle_1_Mile vs DENTAL	-0.391396	0.391396
11	No_Vehicle_Half_Mile vs DENTAL	-0.388544	0.388544
288	Low_Income_1_Mile vs COREW	-0.388464	0.388464
287	Low_Income_1_Mile vs COREM	-0.321915	0.321915



# CORRELATIONS PREVENTATIVE SCREENINGS AND SDOH

## 10 Strongest Negative Correlations

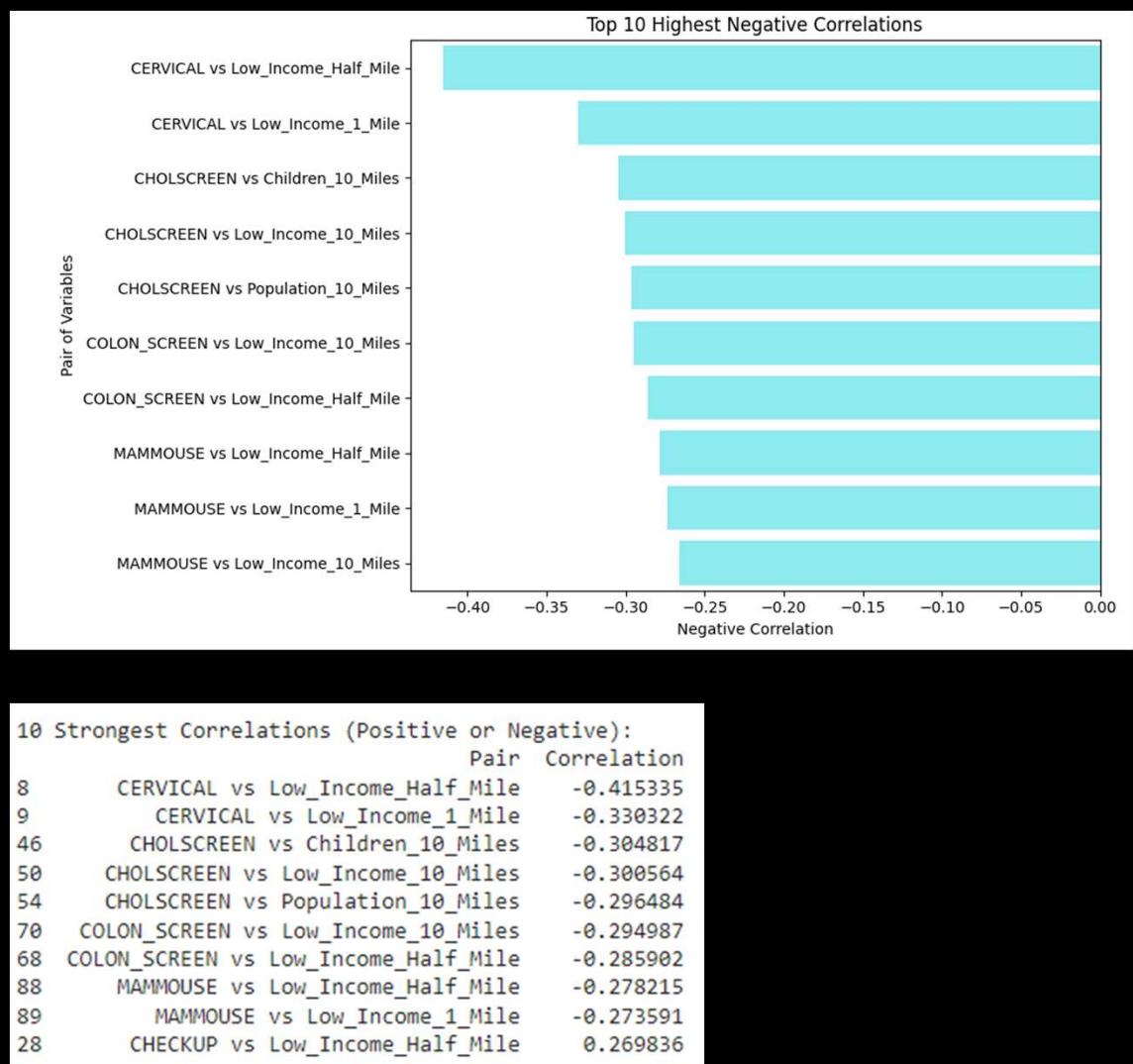
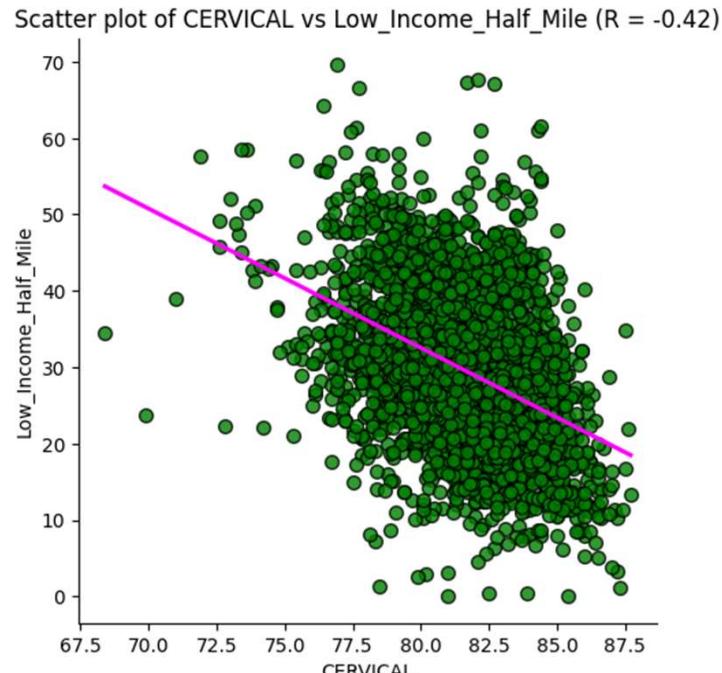


## 10 Strongest Correlations (Positive or Negative):

	Pair	Correlation
3	CERVICAL vs NOHSDP	-0.511206
4	CERVICAL vs POV150	-0.489667
33	COLON_SCREEN vs NOHSDP	-0.393859
0	CERVICAL vs BROAD	-0.389234
34	COLON_SCREEN vs POV150	-0.372206
1	CERVICAL vs CROWD	-0.370167
31	COLON_SCREEN vs CROWD	-0.362426
30	COLON_SCREEN vs BROAD	-0.324758
41	MAMMOUSE vs CROWD	-0.293665
44	MAMMOUSE vs POV150	-0.289866

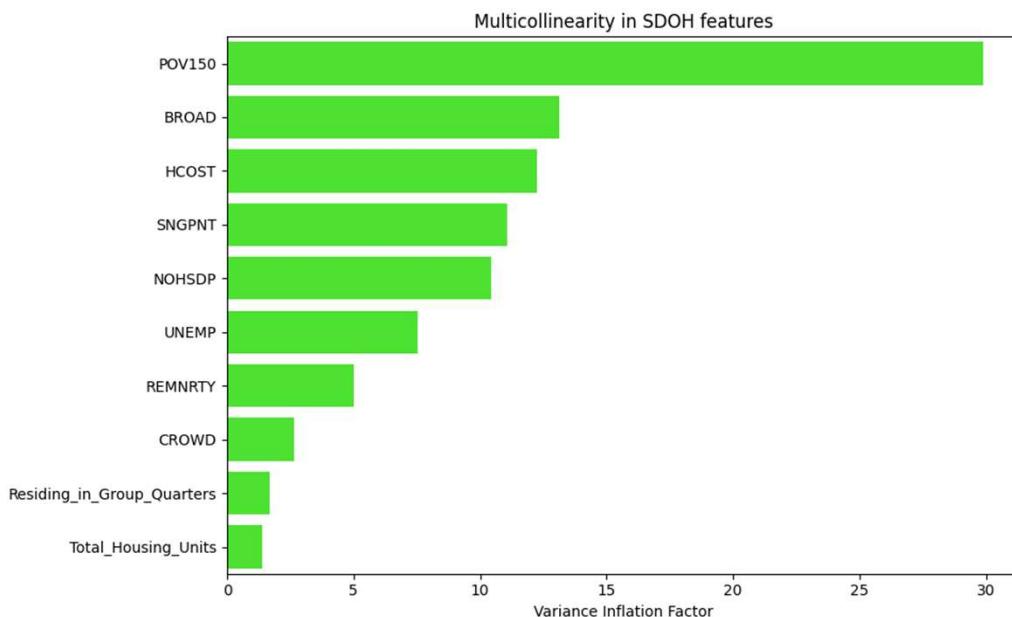
# CORRELATIONS PREVENTATIVE SCREENINGS AND FOOD ACCESS

## 10 Strongest Negative Correlations



# CHECKING FOR MULTICOLLINEARITY IN SDOH FEATURES

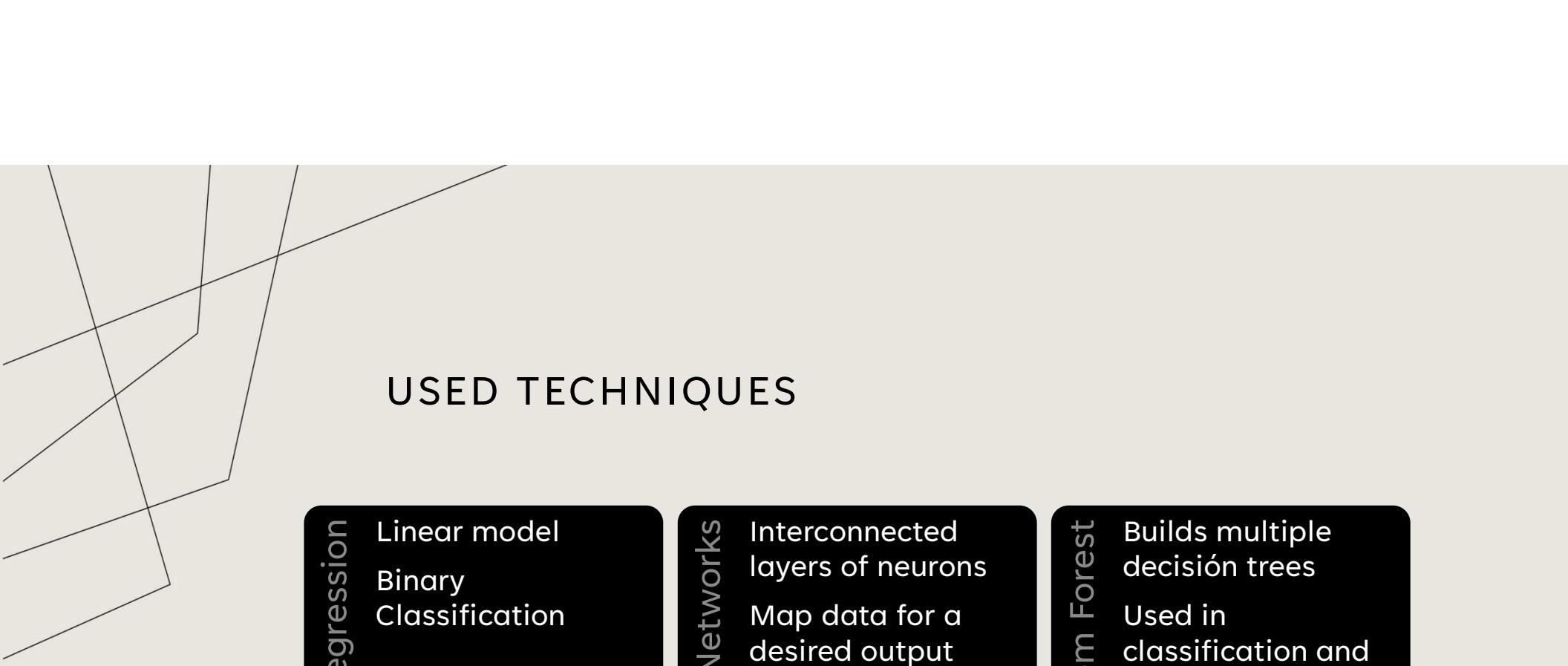
SDOH Features



	Feature	VIF
4	POV150	29.870483
0	BROAD	13.144014
2	HCOST	12.242866
6	SNGPNT	11.087698
3	NOHSDP	10.425348
7	UNEMP	7.523886
5	REMNRTY	5.004775
1	CROWD	2.671960
9	Residing_in_Group_Quarters	1.668267
8	Total_Housing_Units	1.370490



**MACHINE  
LEARNING**



## USED TECHNIQUES

### Logistic Regression

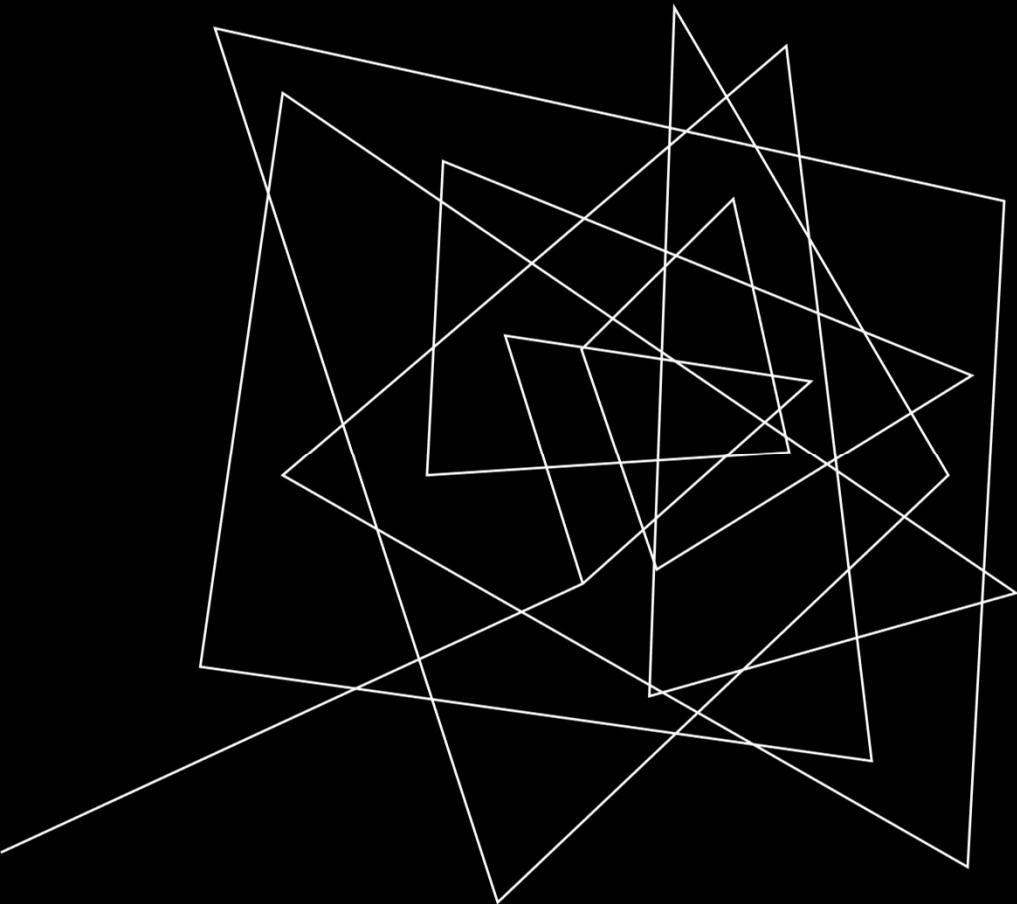
Linear model  
Binary Classification

### Neural Networks

Interconnected layers of neurons  
Map data for a desired output  
Used for binary and non-binary data  
Inputs > Act. Function > Outputs

### Random Forest

Builds multiple decision trees  
Used in classification and regression tasks



# LOGISTIC REGRESSION

\*JUPYTER NOTEBOOK

# LOGISTIC REGRESSION CONFUSION MATRIX

Lung Cancer

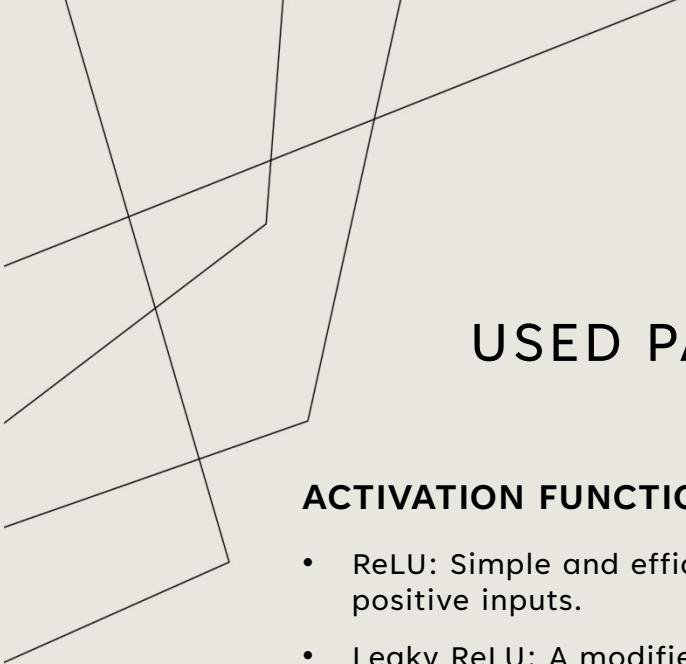
	precision	recall	f1-score	support
Yes	0.33	0.20	0.25	10
No	0.89	0.94	0.91	68
accuracy			0.85	78
macro avg	0.61	0.57	0.58	78
weighted avg	0.82	0.85	0.83	78

Heart Attack

	precision	recall	f1-score	support
Yes	0.88	0.83	0.85	35
No	0.86	0.90	0.88	41
accuracy			0.87	76
macro avg	0.87	0.87	0.87	76
weighted avg	0.87	0.87	0.87	76

- **True Positives (Yes are actual Yes):**  
2/10 correctly identified.
- **False Negatives (No are actual Yes):**  
8 lung cases missed.
- **True Negatives (No are actual No):**  
64/68 correctly identified.
- **False Positives(Yes are actual No):**  
4 cases incorrectly classified.

- **True Positives (Yes are actual Yes):**  
29/35 cases correctly identified.
- **False Negatives (No are actual Yes):**  
6 cases were missed
- **True Negatives (No are actual No):**  
37/41 cases correctly identified
- **False Positives(Yes are actual No):**  
4 cases incorrectly classified



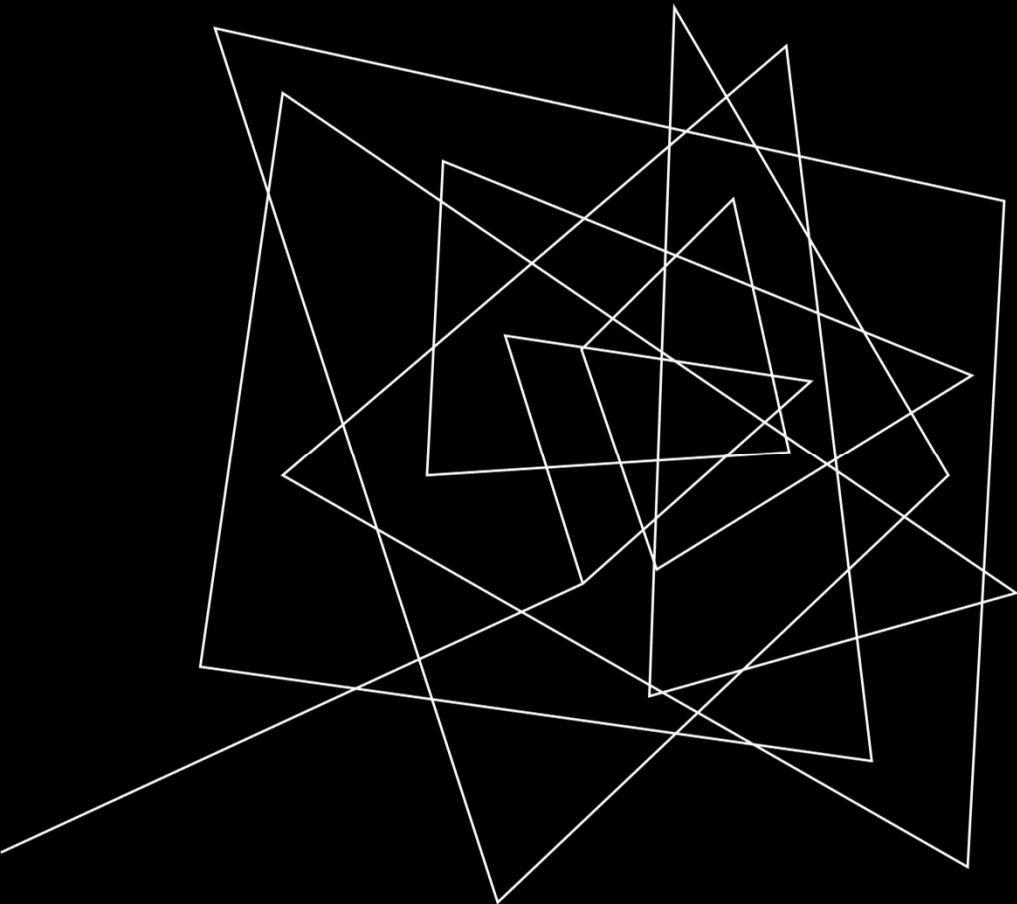
## USED PARAMETERS

### ACTIVATION FUNCTIONS

- ReLU: Simple and efficient. Activates only positive inputs.
- Leaky ReLU: A modified ReLU that allows a small, non-zero gradient for negative inputs.
- ELU: Unlike ReLU, ELU allows negative inputs, helping to make the mean activations closer to zero
- PReLU: A variant of Leaky ReLU, but with the slope of the negative part of the function being a learnable parameter
- Swish: A smooth, non-monotonic function that outperforms ReLU in certain deep networks. Swish allows for small negative inputs

### OPTIMIZERS

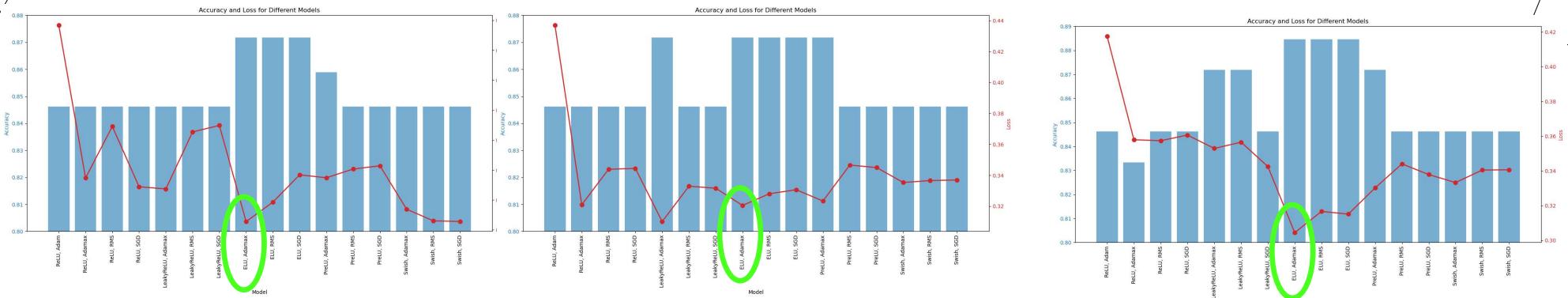
- Adam: Widely used, computes individual rates for each parameter.
- Adamax: More stable versión of Adam.
- RMSprop: handles non-stationary objectives.
- SGD: Gradient descent updating weights by the loss function.



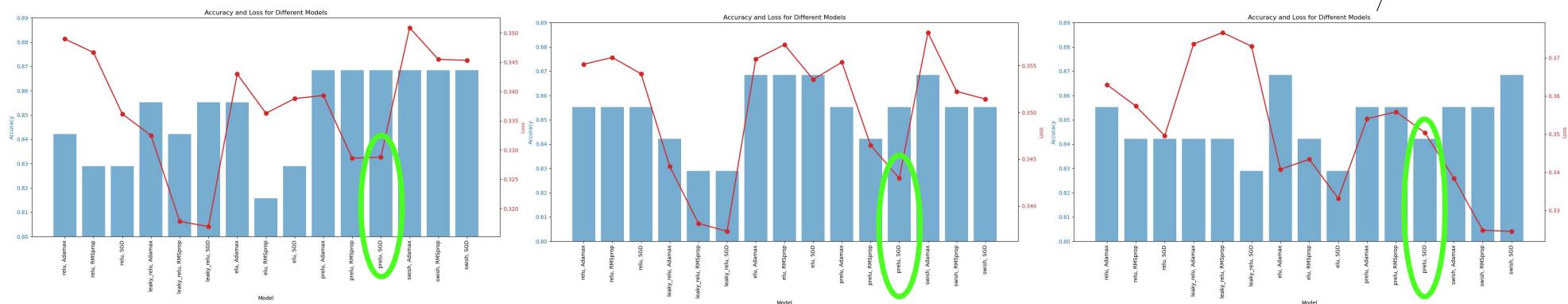
# NEURAL NETWORKS

\*JUPYTER NOTEBOOK

## LUNG CANCER (ELU, ADAMAX)



## HEART ATTACK (PRELU, SGD)\*\*\*





# Data Cleaning and Database Prep

SDOH Measures for County, ACS 2017-2021  
data.cdc.gov



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

**Geolocation**  
POINT  
(-86.2434818  
32.9314453)

Latitude	Longitude
32.931445	-86.243482

```
# Function to extract Latitude and Longitude
def extract_lat_lng(geo_point):
    # Check if the value is a string
    if isinstance(geo_point, str):
        match = re.match(r'POINT \((-?\d.+\) (\-?\d.+)*)', geo_point)
        if match:
            return pd.Series([float(match.group(2)), float(match.group(1))], index=['Latitude', 'Longitude'])
    # Return None if the value is not a valid POINT string
    return pd.Series([None, None], index=['Latitude', 'Longitude'])

# Apply the function to create Latitude and Longitude columns
sdoh_measures_df[['Latitude', 'Longitude']] = sdoh_measures_df['Geolocation'].apply(extract_lat_lng)

# Drop the original Geolocation column
sdoh_measures_df.drop(columns='Geolocation', inplace=True)
```

# Data Cleaning and Database Prep

SDOH Measures for County, ACS 2017-2021  
data.cdc.gov



```
sdoh_measures_df.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 28287 entries, 0 to 28286  
Data columns (total 12 columns):
```

```
sdoh_pivot_df.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3142 entries, 0 to 3141  
Data columns (total 16 columns):
```

Measure_ID	Measure_Short_Name	Measure
0	H_COST	Housing cost burden
1	POV150	Persons living below 150% of the poverty level
2	CROWD	Crowding among housing units
3	AGE65	Persons aged 65 years or older
4	SNGPNT	Single-parent households
5	BROAD	No broadband internet subscription among house...
6	NOHSDP	No high school diploma among adults aged 25 ye...
7	UNEMP	Unemployment among people 16 years and older i...
8	REMNRITY	Persons of racial or ethnic minority status

# Data Cleaning and Database Prep



www.kaggle.com/datasets

```
food_access_df.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3142 entries, 0 to 3141  
Data columns (total 25 columns):  
 #   Column           Non-Null Count  Dtype     
 ---  --  
 0   State            3142 non-null    object    
 1   County           3142 non-null    object    
 2   Population       3142 non-null    int64     
 3   Total_Housing_Units 3142 non-null    int64    
 4   Residing_in_Group_Qarters 3142 non-null    float64  
 5   No_Vehicle_Half_Mile 3142 non-null    float64  
 6   No_Vehicle_1_Mile   3142 non-null    float64  
 7   No_Vehicle_10_Miles 3142 non-null    float64  
 8   No_Vehicle_20_Miles 3142 non-null    float64  
 9   Children_Half_Mile 3142 non-null    float64  
 10  Children_1_Mile    3142 non-null    float64  
 11  Children_10_Miles  3142 non-null    float64  
 12  Children_20_Miles  3142 non-null    float64  
 13  Low_Income_Half_Mile 3142 non-null    float64  
 14  Low_Income_1_Mile   3142 non-null    float64  
 15  Low_Income_10_Miles 3142 non-null    float64  
 16  Low_Income_20_Miles 3142 non-null    float64  
 17  Population_Half_Mile 3142 non-null    float64  
 18  Population_1_Mile   3142 non-null    float64  
 19  Population_10_Miles 3142 non-null    float64  
 20  Population_20_Miles 3142 non-null    float64  
 21  Seniors_Half_Mile  3142 non-null    float64  
 22  Seniors_1_Mile     3142 non-null    float64  
 23  Seniors_10_Miles   3142 non-null    float64  
 24  Seniors_20_Miles   3142 non-null    float64
```

```
# Merge the two DataFrames on County and State  
food_access_fips_df = pd.merge(  
    food_access_df,  
    county_key_df[['County', 'State_Desc', 'CountyFIPS', 'Latitude', 'Longitude']],  
    left_on=['County', 'State'],  
    right_on=['County', 'State_Desc'],  
    how='left'  
)
```

	County	State_Abbr	State_Desc	CountyFIPS	Latitude	Longitude
0	Aleutians East Borough	AK	Alaska	2013	55.245044	-161.997477
1	Aleutians West Census Area	AK	Alaska	2016	51.959447	178.338813
2	Anchorage Municipality	AK	Alaska	2020	61.174250	-149.284329
3	Bethel Census Area	AK	Alaska	2050	60.929141	-160.152625
4	Bristol Bay Borough	AK	Alaska	2060	58.741661	-156.966805

# CHANGE FUNCTION TO WORK WITH CONTINUOUS DATA

- Output Layer Activation
- Loss Function and Metrics
- Evaluate MAE
- Variables to Adjust to Different Datasets
  - input\_shape
  - hidden\_nodes\_layer1
  - hidden\_nodes\_layer2
  - hidden\_nodes\_layer3

```
# Define the model
nn = tf.keras.models.Sequential()

# Add the input layer
nn.add(tf.keras.layers.Input(shape=(input_shape,)))

# First hidden layer
if isinstance(get_activation(activation_func), str): # If it's a standard activation
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer1, activation=get_activation(activation_func)))
else: # If it's a custom activation layer
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer1))
    nn.add(get_activation(activation_func)) # Add custom activation layer

# Dropout and subsequent layers
nn.add(Dropout(0.3))
if isinstance(get_activation(activation_func), str):
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer2, activation=get_activation(activation_func)))
else:
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer2))
    nn.add(get_activation(activation_func))

nn.add(Dropout(0.3))
if isinstance(get_activation(activation_func), str):
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer3, activation=get_activation(activation_func)))
else:
    nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer3))
    nn.add(get_activation(activation_func))

# Output layer (for continuous data)
nn.add(tf.keras.layers.Dense(units=1, activation="linear"))

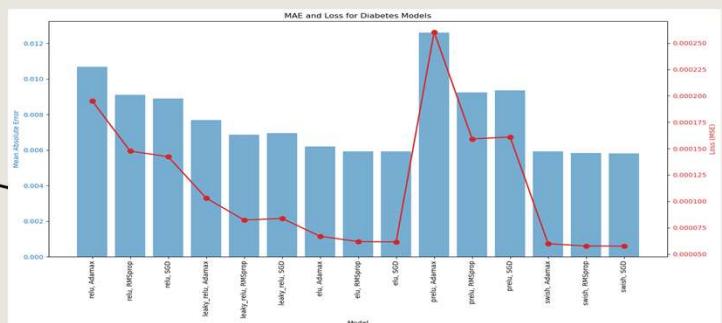
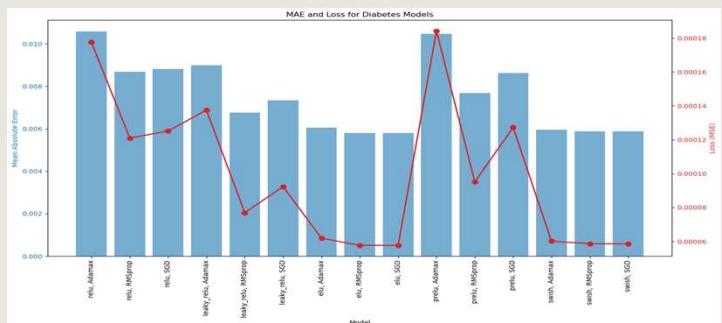
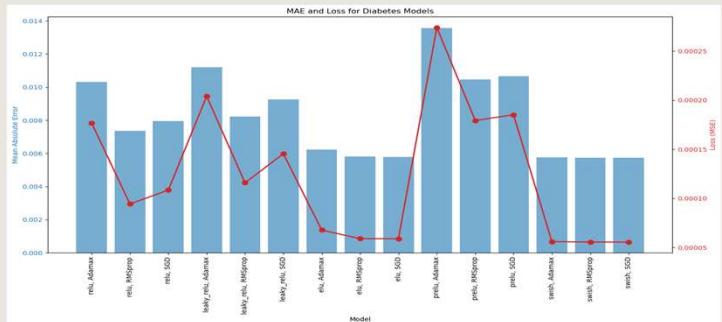
# Experiment with different optimizers
optimizers = {
    'Adamax': Adamax(learning_rate=0.002),
    'RMSprop': RMSprop(learning_rate=0.001),
    'SGD': SGD(learning_rate=0.01, momentum=0.9)
}

mae_list2 = []
loss_list2 = []

# Iterate through each optimizer
for name, optimizer in optimizers.items():
    print(f"\nTraining with {name} optimizer and {activation_func} activation function...")
    nn.compile(optimizer=optimizer, loss="mean_squared_error", metrics=["MAE"])

# Evaluate the model
loss, mae = nn.evaluate(X_test_scaled, y_test, verbose=0)
print(f'{name} Optimizer - Loss (MSE): {loss:.6f}, Mean Absolute Error: {mae:.6f}')
mae_list2.append(mae)
loss_list2.append(loss)
```

# Diabetes Optimization Function



## Model Optimization

- Activation Function: Exponential Linear Unit (elu)
- Optimizer: Stochastic Gradient Descent (SGD)

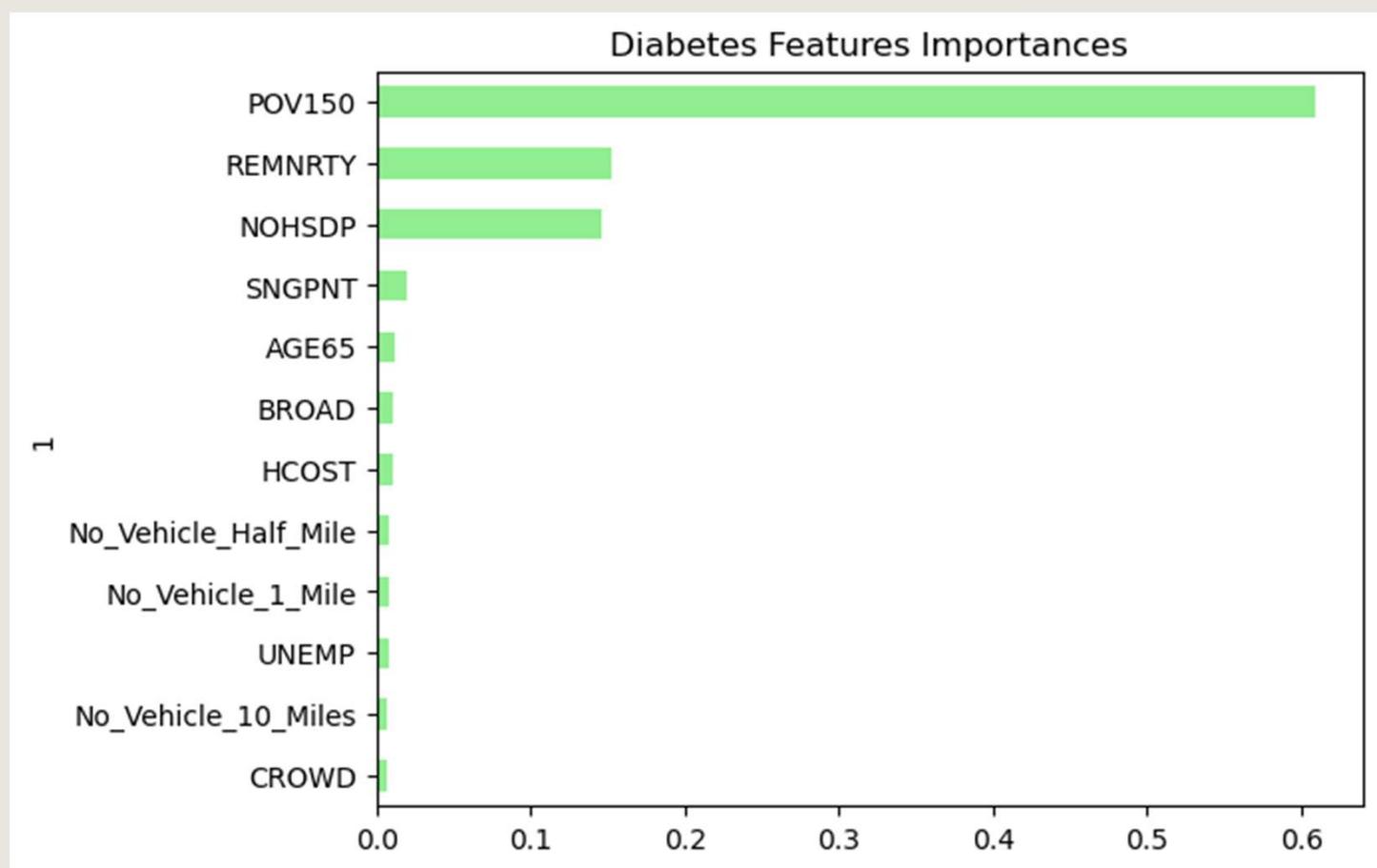
## Model Performance

- MAE 0.005796
- Loss (MSE) 0.000058

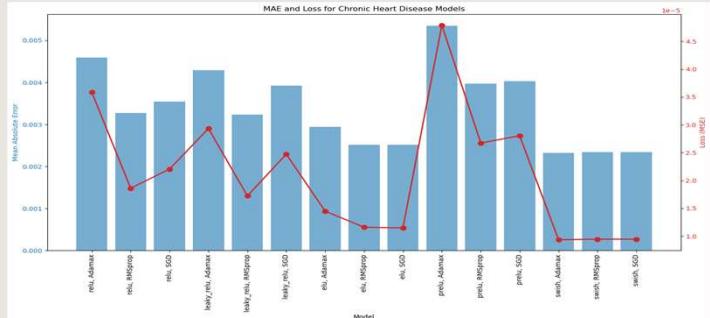
Model	elu, SGD
Activation_Function	elu
Optimizer	SGD
MAE	0.005796
Loss	0.000058
MAE_Rank	1.0
Loss_Rank	1.0
Average_Rank	1.0

# Diabetes Random Forest Regressor Evaluation Metrics

MAE: 0.005177508452535746  
MSE: 4.548317280624185e-05  
RMSE: 0.006744121351684135  
R<sup>2</sup>: 0.9166733060194924

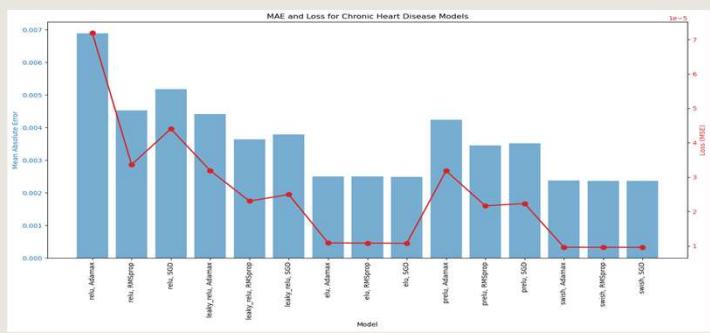


# Chronic Heart Disease Optimization Function



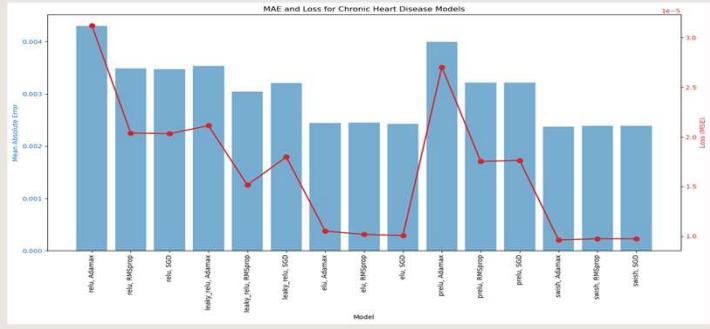
## Model Optimization

- Activation Function: Swish
- Optimizer: Adamax



## Model Performance

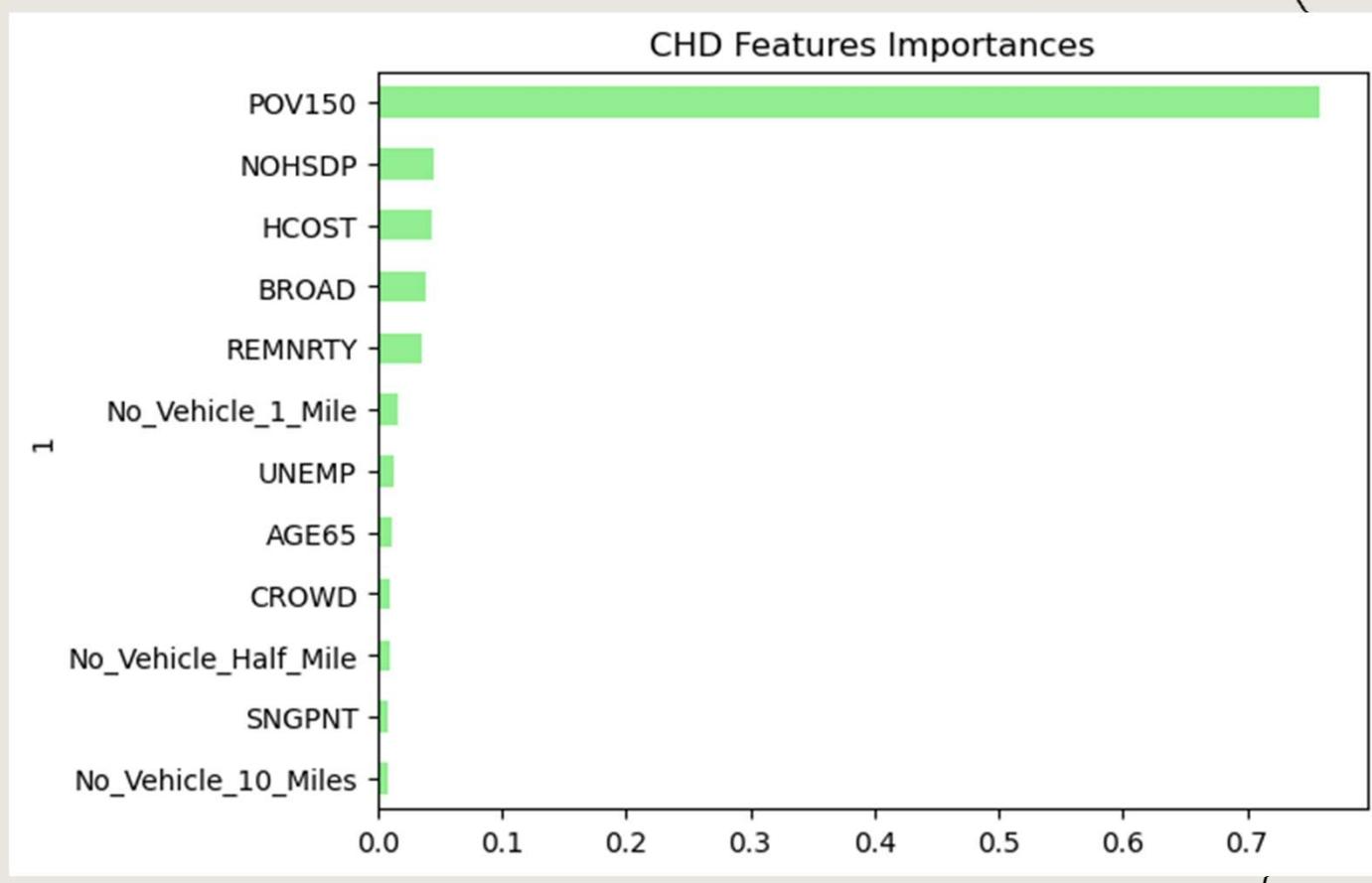
- MAE 0.002372
- Loss (MSE) 0.00001



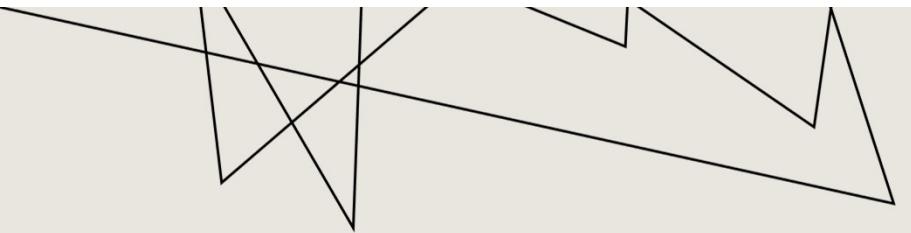
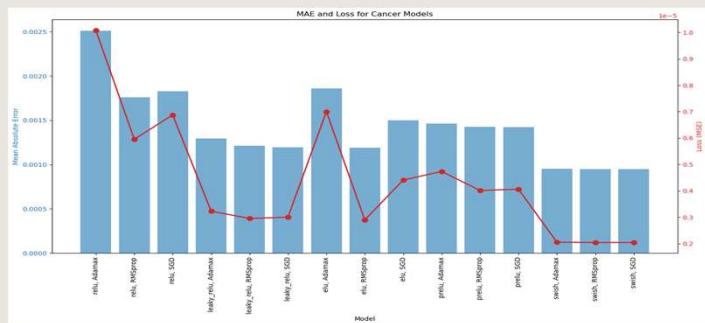
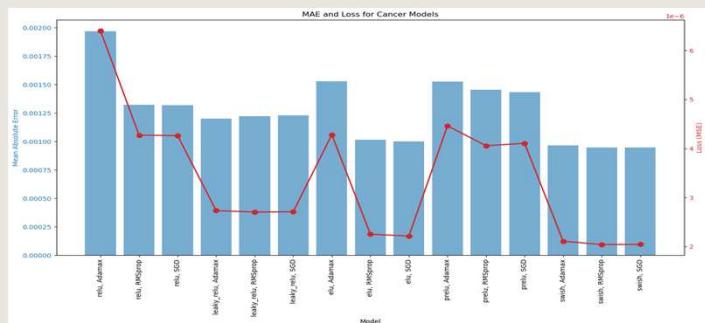
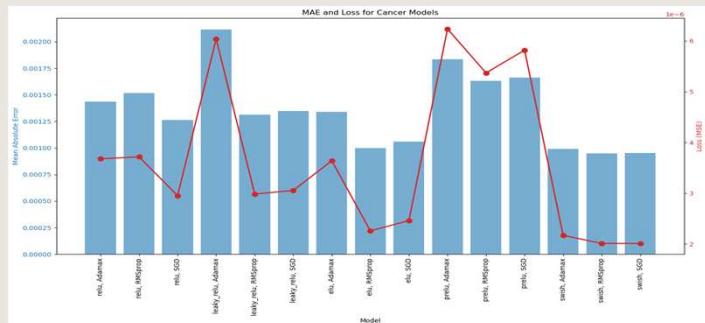
Model	swish, Adamax
Activation_Function	swish
Optimizer	Adamax
MAE	0.002372
Loss	0.00001
MAE_Rank	1.0
Loss_Rank	1.0
Average_Rank	1.0

# Chronic Heart Disease Random Forest Regressor Evaluation Metrics

MAE: 0.00226032769830949  
MSE: 9.15610218985695e-06  
RMSE: 0.003025905185206065  
 $R^2$ : 0.881870567942117



# Cancer Optimization Function



## Model Optimization

- Activation Function: Swish
- Optimizer: RMSprop

## Model Performance

- MAE 0.000948
- Loss (MSE) 0.000002

Model	swish, RMSprop
Activation_Function	swish
Optimizer	RMSprop
MAE	0.000948
Loss	0.000002
MAE_Rank	1.0
Loss_Rank	1.0
Average_Rank	1.0

# Cancer Random Forest Regressor Evaluation Metrics

