MIDUS Project

Overview:

MIDUS (Midlife in the United States) is a national longitudinal study of health and well-being (https://www.google.com/url?q=https%3A%2F%2Fwww.icpsr.umich.edu%2Fweb%2FICPSR%2Fsearch%2Fstudies%3Fq%3DMIDUS). It collects data on a wide range of physical, psychological, and social factors. The study has been ongoing since the 1990s, with multiple waves of data collection.

Purpose:

To explore the relationships between physical activity, social/psychological variables, and various health outcomes using the MIDUS datasets.

Data Sources:

- MIDUS 1: Baseline survey conducted in the 1990s
- MIDUS 2: Follow-up survey conducted in the 2000s
- MIDUS 3: Most recent follow-up survey conducted in the 2010s (https://www.google.com/url?q=https%3A%2F%2Fwww.icpsr.umich.edu%2Fweb%2FICPSR%2Fstudies%2F36346)
- Biomarker 1: Collected physical biomarker data from a subset of MIDUS participants (https://www.google.com/url?q=https%3A%2F%2Fwww.icpsr.umich.edu%2Fweb%2FICPSR%2Fstudies%2F38837%2Fpublications)
- Biomarker 2: Collected additional biomarker data from a different subset of MIDUS participants

Data Importing and Exploration:

The datasets were downloaded as tsv from the Inter-university Consortium for Political and Social Research (ICPSR) website. The specific URLs for the MIDUS 3 and Biomarker 2 datasets were provided. Tsvs were converted into csv and were read in pandas dataframe.

```
]  import pandas as pd

   # Step 1: Read the TSV files into DataFrames
   midus1_df = pd.read_csv('/content/36346-0001-Data.tsv', sep='\t', low_memory=False)
   midus1_df.to_csv('midus1_df.csv', index=False)

   midus2_df = pd.read_csv('/content/36346-0002-Data.tsv', sep='\t', low_memory=False)
   midus2_df.to_csv('midus2_df.csv', index=False)

   midus3_df = pd.read_csv('/content/36346-0003-Data.tsv', sep='\t', low_memory=False)
   midus3_df.to_csv('midus3_df.csv', index=False)

   biomarker1_df = pd.read_csv('/content/38837-0001-Data.tsv', sep='\t', low_memory=False)
   biomarker1_df.to_csv('biomarker1_df.csv', index=False)

   biomarker2_df = pd.read_csv('/content/38837-0002-Data.tsv', sep='\t', low_memory=False)
   biomarker2_df.to_csv('biomarker2_df.csv', index=False)
```

Figure 1: converting tsv files into csv

```
[ ]  midus1_df.head()
```

	SAMPLMAJ	C1STATUS	M3RE_FILTER	C1PRAGE	C1PBYEAR	C1PRSEX	C1PIDATE_MO	C1PIDATE_YR	...
	2	4	0	69	1943	1	7	2013	...
	1	1	0	78	1935	1	6	2013	...
	2	4	0	61	1952	2	6	2013	...
	3	4	0	63	1950	2	11	2013	...
	1	4	0	60	1952	1	6	2013	...

Figure 2: MIDUS1 Dataframe

```
[ ]  midus2_df.head()
```

	M2ID	M2FAMNUM	SAMPLMAJ	B1PGENDER	M3P1_DIS	M3FILTER
0	10001	110498	2	1	1	1
1	10002	100001	1	1	1	1
2	10004	100002	1	1	8	0
3	10005	120803	3	2	7	1
4	10006	120772	3	2	3	1

Figure 3: MIDUS2 Dataframe

```
[ ]  midus3_df.head()
```

	M2ID	SAMPLMAJ	C1PRAGE	C1PRSEX	C1PAA1A	C1PAA1B	C1PAA1C	C1PAA1D	C1PAA1E	C1PAA2LBBA	...
0	10001	2	69	1	9	99	99	99	99	99	..
1	10002	1	78	1	8	99	99	99	99	99	..
2	10011	2	61	2	99	99	99	99	99	99	..
3	10015	3	63	2	5	99	99	99	99	99	..
4	10019	1	60	1	99	99	99	99	99	99	..

5 rows × 183 columns

Figure 4: MIDUS3 Dataframe

```
# Handle missing values
merged_data.fillna(method='ffill', inplace=True)  # Forward fill missing values
merged_data.dropna(inplace=True)  # Drop rows with any missing values

# Remove duplicates
merged_data.drop_duplicates(inplace=True)

# Export the cleaned dataset to CSV
merged_data.to_csv('cleaned_dataset.csv', index=False)
print("Cleaned dataset has been created and exported as 'cleaned_dataset.csv'.")
```

```
Cleaned dataset has been created and exported as 'cleaned_dataset.csv'.
```

```
biomarker1_df.head()
```

	M2ID	M2FAMNUM	SAMPLMAJ	C1PRAGE	C1PRSEX	C4ZSITE	C4ZCOMPM	C4ZCOMPY	C4Z
0	10019	100009	1	60	1	2	1	2018	
1	10036	120944	3	64	1	2	2	2022	
2	10040	100018	1	58	1	2	5	2019	
3	10047	100022	1	54	2	1	7	2018	
4	10060	100028	1	67	1	3	11	2018	

5 rows × 3205 columns

Figure 5: Biomarker1 dataframe

```
biomarker2_df.head()
```

	M2ID	M2FAMNUM	SAMPLMAJ	C1PRSEX	C4ZAGE	C4XINDEX	C4XTYPE	C4XPM	C4XOM	C4XAM	...	C4XPC_1665	C4XPC_1710	C4XPC_1750	C4XPC_1752	C4XPC_1753	C4XPC_1756	C4XPC
0	10019	100009	1	1	65	C4XAN1	3	2	6	1	...	2	2	2	2	2	2	
1	10019	100009	1	1	65	C4XON1	2	2	6	1	...	2	2	1	2	2	2	
2	10019	100009	1	1	65	C4XON3	2	2	6	1	...	2	2	2	2	2	2	
3	10019	100009	1	1	65	C4XON4	2	2	6	1	...	1	2	2	2	2	2	
4	10019	100009	1	1	65	C4XON5	2	2	6	1	...	2	2	2	2	2	2	

5 rows × 116 columns

Figure 6: Biomarker2 dataframe

Data Merging:

Individual datasets were merged into merged_data.csv based on the column ‘M2ID’.

```
[ ]  import pandas as pd


     # Merge datasets based on 'M2ID'
     merged_data = midus1_selected.merge(midus2_selected, on='M2ID', how='outer')
     merged_data = merged_data.merge(midus3_selected, on='M2ID', how='outer')
     merged_data = merged_data.merge(biomarker1_selected, on='M2ID', how='outer')
     merged_data = merged_data.merge(biomarker2_selected, on='M2ID', how='outer', suffixes=('_1', '_2'))

     # Combine sex columns into a single column
     sex_cols = [col for col in merged_data.columns if col.startswith(('B1PRSEX', 'B4PRSEX', 'C1PRSEX'))]
     merged_data['SEX'] = merged_data[sex_cols].bfill(axis=1).iloc[:, 0]
     merged_data.drop(columns=sex_cols, inplace=True)

     # Combine age columns into a single column
     age_cols = [col for col in merged_data.columns if col.startswith(('B1PAGE_M2', 'B4PAGE_M2', 'C1PRAGE'))]
     merged_data['AGE'] = merged_data[age_cols].bfill(axis=1).iloc[:, 0]
     merged_data.drop(columns=age_cols, inplace=True)

     # Export the merged dataset to CSV
     merged_data.to_csv('merged_dataset.csv', index=False)
     print("Merged dataset has been created and exported as 'merged_dataset.csv'.")
```

Figure 7: Python code for merging of the datasets

Preprocessing of the merged dataset from null and duplicate values:

Merged_data.csv was cleaned from missing values and duplicate rows and the output was saved to ‘cleaned_dataset.csv’ and ‘cleaned_dataset_no_duplicates.csv’.

```
# Handle missing values
merged_data.fillna(method='ffill', inplace=True)  # Forward fill missing values
merged_data.dropna(inplace=True)  # Drop rows with any missing values

# Remove duplicates
merged_data.drop_duplicates(inplace=True)

# Export the cleaned dataset to CSV
merged_data.to_csv('cleaned_dataset.csv', index=False)
print("Cleaned dataset has been created and exported as 'cleaned_dataset.csv'.")


Cleaned dataset has been created and exported as 'cleaned_dataset.csv'.
```

Figure 8: Python code for data cleaning from null values

```
import pandas as pd

# Load the cleaned dataset
merged_data = pd.read_csv('cleaned_dataset.csv')

# Remove duplicate columns
merged_data = merged_data.loc[:, ~merged_data.columns.duplicated()]

# Export the updated dataset without duplicate columns
merged_data.to_csv('cleaned_dataset_no_duplicates.csv', index=False)
print("Cleaned dataset without duplicate columns has been created and exported as 'cleaned_dataset_no_duplicates.csv'.")

Cleaned dataset without duplicate columns has been created and exported as 'cleaned_dataset_no_duplicates.csv'.
```

Figure 9: Python code for dataseet cleaning from duplicate values

Renaming of the variables from codes to descriptions:

Variables were renamed from codes to their descriptions for clear understanding.

```python
import pandas as pd

# Load the cleaned dataset
merged_data = pd.read_csv('cleaned_dataset_no_duplicates.csv')
column_mapping = {
    'M2ID': 'M2ID=MIDUS 2 ID number',
    'M2FAMNUM_x': 'M2FAMNUM_x',
    'C1PRAGE': 'C1PA39 = Age began to smoke regularly',
    'C1PB1': 'C1PB1=Highest level of education completed',
    'C1PB2': 'C1PB2=Age first worked for pay for 6 or more months',
    'C1PB2A1': 'C1PB2A1 = Employment 1/2008 - Working',
    'C1PB2A2': 'C1PB2A2 = Employment 1/2008 - Self-employed',
    'C1PB2A3': 'C1PB2A3 = Employment 1/2008 - Unemployed',
    'C1PB2A4': 'C1PB2A4 = Employment 1/2008 - Temporarily laid off',
    'C1PB2A5': 'C1PB2A5 = Employment 1/2008 - Retired',
    'C1PB2A6': 'C1PB2A6 = Employment 1/2008 - Homemaker',
    'C1PB2A7': 'C1PB2A7 = Employment 1/2008 - Full-time student',
    'C1PB2A8': 'C1PB2A8 = Employment 1/2008 - Part-time student',
    'C1PB2A9': 'C1PB2A9 = Employment 1/2008 - Maternity or sick leave',
    'C1PB2A10': 'C1PB2A10 = Employment 1/2008 - Permanently disabled',
```

Figure 10: Python code for renaming the variables

Extraction of variables of interest from the merged and cleaned dataset:

From the merged and cleaned dataset which has renamed columns 'cleaned_dataset_renamed.csv', variables of interest were extracted which included variables related to physical activity, psychological variables (eg. Stress, depression, loneliness) and health outcome variables related to blood pressure, chronic kidney disease, breast cancer etc. (these list of these variables is present in the file 'Variables_final dataset_highlighted'). The dataset 'final_dataset.csv' was created with extracted relevant variables from the previous extracted dataset.

```
extracted_data = merged_data
[['M2ID=MIDUS 2 ID number',
'M2ID=MIDUS 2 ID number',
 'M2FAMNUM_x',
 'C1PA39 = Age began to smoke regularly',
 'C1PB1=Highest level of education completed',
 'C1PB2=Age first worked for pay for 6 or more months
 'C1PA6C = History of Parkinson\s disease',
 'C1PA6D = History of other neurological disorder',
 'C1PA7 = Heart trouble suspect/confirmed by n',
 'C1PA7A = Age doctor told you have heart problem',
 'C1PA7BA = Diagnosis - Heart attack'
```

```
'C4BLDL = Blood LDL Cholesterol (mg/dL)',
'C4BRTHDL = Blood Ratio Total / HDL Cholesterol',
'C4BSBAP = Blood Bone Specific Alkaline Phospatase (UL)',
'C1PB1=Highest level of education completed',
'C1PB2A1 = Employment 1/2008 - Working',
'C1PB2A2 = Employment 1/2008 - Self-employed',
'C1PB2A3 = Employment 1/2008 - Unemployed',
'C1PB2A4 = Employment 1/2008 - Temporarily laid off',
'C1PB2A5 = Employment 1/2008 - Retired',
'C1PB2A6 = Employment 1/2008 - Homemaker',
'C1PB2A7 = Employment 1/2008 - Full-time student',
'C1PB2A8 = Employment 1/2008 - Part-time student',
'C1PB2A9 = Employment 1/2008 - Maternity or sick leave',
'C1PB2A10 = Employment 1/2008 - Permanently disabled'
]
final_df = df[df.columns.intersection(variables)]
final_df.to_csv('final_dataset.csv', index=False)
```

Figure 11: final dataset created with extracted variables using python

final_df

	M2ID=MIDUS 2 ID number	M2FAMNUM_x	C1PB1=Highest level of education completed	C1PB2A1 = Employment 1/2008 - Working	C1PB2A2 = Employment 1/2008 - Self-employed	C1PB2A3 = Employment 1/2008 - Unemployed	C1PB2A4 = Employment 1/2008 - Temporarily laid off	C1PB2A5 = Employment 1/2008 - Retired	C1PB2A6 = Employment 1/2008 - Homemaker	C1PB2A7 = Employment 1/2008 - Full-time student	...	C4AD520 = Sun AM Overall Quality of Sleep
0	10019	100009.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	...	3.0
1	10019	100009.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	...	3.0
2	10019	100009.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	...	3.0
3	10019	100009.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	...	3.0
4	10019	100009.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	...	3.0

Figure 12: Dataframe 'final_dataset' with extracted variables

Feature Engineering (added new columns):

Identified the relevant columns related to family history of Alzheimer's, breast cancer, stroke, circulation problems, high blood pressure, and heart disease in the final dataset and new columns were created as 'Family history of Alzheimer's', 'Family history of breast cancer', 'Family history of stroke', 'Family history of circulation problems', 'Family history of high blood pressure', 'Family history of heart disease', with the binary value sets of (1 for presence, 0 for absence) based on the value sets of columns in the original dataset (If any of the relevant columns had a value of 1 (indicating the presence of the condition), the new variable was set to

1If all the relevant columns had values other than 1 (missing, unknown, no, etc.), the new variable was set to 0).

```
import pandas as pd

# Read the extracted dataset CSV file
df = pd.read_csv('extracted_dataset.csv')

# Create the new variables based on the conditions
df['Family history of Alzheimer\'s'] = df[['C4H75T = Blood relatives alzheimers/dementia', 'C4H75T1 = Mother alzheimers/dementia', 'C4H75T2 = Father alzheimers/dementia', 'C4H75T3
df['Family history of breast cancer'] = df[['C4H75L = Have any of your blood relatives had breast cancer?', 'C4H75L1 = Has your mother had breast cancer?', 'C4H75L2 = Has your fath
df['Family history of stroke'] = df[['C4H75E = Have any of your blood relatives had a stroke?', 'C4H75E1 = Has your mother had a stroke?', 'C4H75E2 = Has your father had a stroke?'
df['Family history of circulation problems'] = df[['C4H75D = Have any of your blood relatives had circulation problems?', 'C4H75D1 = Has your mother had circulation problems?', 'C4
df['Family history of high blood pressure'] = df[['C4H75B = Have any of your blood relatives had high blood pressure?', 'C4H75B1 = Has your mother had high blood pressure?', 'C4H75
df['Family history of heart disease'] = df[['C4H75A = Have any of your blood relatives had heart disease?', 'C4H75A1 = Has your mother had heart disease?', 'C4H75A2 = Has your fath

# Add the new variables to the final dataset
final_df = pd.concat([final_df, df[['Family history of Alzheimer\'s', 'Family history of breast cancer', 'Family history of stroke', 'Family history of circulation problems', 'Fami

# Save the updated final dataset to a new CSV file
final_df.to_csv('final_dataset.csv', index=False)
```

Figure 13: Additional Variable Creation

Analysis of correlation between AHA Life essential 8 variables:

Identified variables that could be coded as one of the 8 variables from AHA life essential 8 (https://www.google.com/url?q=https%3A%2F%2Fwww.heart.org%2Fen%2Fhealthy-living%2Fhealthy-lifestyle%2Flifes-essential-8) as –

- eat better
- be more active
- quit tobacco
- get healthy sleep
- manage weight
- control cholesterol
- manage blood sugar
- manage blood pressure

List of Identified AHA Life essential 8 Variables:

physical activity/exercise related variables, either "breast cancer" or general (other non specific cancer questions) can be found within the file ‘list of AHA life essential variables’.

- C4H73AC = Exercise/Activity A – Coded',
- 'C4H73AS = Seasonal exercise/activity A?',
- 'C4H73AFD = How many times per day do you do exercise/activity A?',
- 'C4H73AFW = How many days per week do you do exercise/activity A?',
- 'C4H73AM = What is the average number of minutes/session that you do exercise/activity A?',
- 'C4H73AI = What is the intensity of exercise/activity A?',
- 'C4H73BC = Exercise/Activity B – Coded',
- 'C4H73BS = Seasonal exercise/activity B?',
- 'C4H73BFD = How many times per day do you do exercise/activity B?',
- 'C4H73BFW = How many days per week do you do exercise/activity B?',
- 'C4H73BM = What is the average number of minutes/session that you do exercise/activity B?',
- 'C4H73BI = What is the intensity of exercise/activity B?',
- 'C4H73CC = Exercise/Activity C – Coded',
- 'C4H73CS = Seasonal exercise/activity C?',

- 'C4H73CFD = How many times per day do you do exercise/activity C?',
- 'C4H73CFW = How many days per week do you do exercise/activity C?',
- 'C4H73CM = What is the average number of minutes/session that you do exercise/activity C?',
- 'C4H73CI = What is the intensity of exercise/activity C?',
- 'C4H73DC = Exercise/Activity D – Coded',
- 'C4H73DS = Seasonal exercise/activity D?',
- 'C4H73DFD = How many times per day do you do exercise/activity D?',
- 'C4H73DFW = How many days per week do you do exercise/activity D?',
- 'C4H73DM = What is the average number of minutes/session that you do exercise/activity D?',
- 'C4H73DI = What is the intensity of exercise/activity D?',
- 'C4H73EC = Exercise/Activity E – Coded',
- 'C4H73ES = Seasonal exercise/activity E?',
- 'C4H73EFD = How many times per day do you do exercise/activity E?',
- 'C4H73EFW = How many days per week do you do exercise/activity E?',
- 'C4H73EM = What is the average number of minutes/session that you do exercise/activity E?',
- 'C4H73EI = What is the intensity of exercise/activity E?',
- 'C4H73FC = Exercise/Activity F – Coded',
- 'C4H73FS = Seasonal exercise/activity F?',
- 'C4H73FFD = How many times per day do you do exercise/activity F?',
- 'C4H73FFW = How many days per week do you do exercise/activity F?',
- 'C4H73FM = What is the average number of minutes/session that you do exercise/activity F?',
- 'C4H73FI = What is the intensity of exercise/activity F?',
- 'C4H73GC = Exercise/Activity G – Coded',
- 'C4H73GS = Seasonal exercise/activity G?',
- 'C4H73GFD = How many times per day do you do exercise/activity G?',
- 'C4H73GFW = How many days per week do you do exercise/activity G?',
- 'C4H73GM = What is the average number of minutes/session that you do exercise/activity G?',
- 'C4H73GI = What is the intensity of exercise/activity G?',
- 'C4H60 = Have you now or in the past used tobacco regularly?',
- 'C4H61 = Have you ever smoked cigarettes regularly-that is, at least a few cigarettes every day?',
- 'C4H61A = Do you currently smoke cigarettes regularly?',
- 'C4H62 = Since we last interviewed you, have you tried to quit smoking?',
- 'C4H63 = For how many years did you smoke regularly?',
- 'C4H64 = During this period, how many cigarettes did you smoke per day, on average?',
- 'C4H65 = How old were you the last time you smoked regularly?',
- 'C4H66A = Do you currently smoke a pipe or cigars, or use snuff or chewing tobacco regularly?',
- 'C4H67 = For how many years did you regularly smoke a pipe or cigars, or use snuff or chewing tobacco?',

- 'C4H68 = In the past, did anyone in your household smoke tobacco inside your home regularly?',
- 'C4H69 = Currently, does anyone regularly smoke cigarettes or other tobacco products inside your home?',
- 'C4SSQ_S1 = SLEEP Component 1 - Subjective Sleep Quality',
- 'C4SSQ_S2 = SLEEP Component 2 - Sleep Latency',
- 'C4SSQ_S3 = SLEEP Component 3 - Sleep Duration',
- 'C4SSQ_S4 = SLEEP Component 4 - Habitual Sleep Efficiency',
- 'C4AD59 = Sun AM Minutes To Fall Asleep',
- 'C4AD510 = Sun AM How Difficult To Fall Asleep',
- 'C4AD511 = Sun AM How Many Times Did You Wake',
- 'C4AD512 = Sun AM Wake Because of Noise or Activity',
- 'C4AD513 = Sun AM If Woke, Difficulty Getting To Sleep',
- 'C4AD514 = Sun AM If Woke, Number of Times Left Bed',
- 'C4AD515 = Sun AM Time Wake Up and Not Return To Sleep',
- 'C4AD516 = Sun AM Time Get Out of Bed',
- 'C4AD517 = Sun AM How Deeply Did You Sleep',
- 'C4AD518 = Sun AM How Well Rested Do You Feel',
- 'C4AD519 = Sun AM How Alert Do You Feel',
- 'C4AD520 = Sun AM Overall Quality of Sleep',
- 'C4WA1SLT = Active 1 Sleep Time',
- 'C4WA1PSL = Active 1 % Sleep Time',
- 'C4WA1SLB = Active 1 # Sleep Bouts',
- 'C4WA1ASB = Active 1 Avg Sleep Bouts',
- 'C4WS2SLT = Sleep 2 Sleep Time',
- 'C4WS2PSL = Sleep 2 % Sleep Time',
- 'C4WS2SLB = Sleep 2 # Sleep Bouts',
- 'C4WS2ASB = Sleep 2 Avg Sleep Bouts',
- 'C4BCHOL = Blood Total Cholesterol (mg/dL)',
- 'C4BHDL = Blood HDL Cholesterol (mg/dL)',
- 'C4BLDL = Blood LDL Cholesterol (mg/dL)',
- 'C4BRTHDL = Blood Ratio Total / HDL Cholesterol',
- 'C4H1I = Have you ever had diabetes?',
- 'C4BHA1C = Blood Hemoglobin A1c %',
- 'C4H1B = Have you ever had high blood pressure?',
- 'C4P1GS = Average BP (sitting) systolic',
- 'C4P1GD = Average BP (sitting) diastolic',
- 'C1PA26 = Ever had cancer',
- 'C1PA28AA = Age breast cancer diagnosed',
- 'C4H1P = Have you ever had cancer?',
- 'Family history of breast cancer'

Pearson and Spearman Bivariate correlations were analysed of the above variables selecting only women/ females and the result was saved in the file ‘bivariate_correlations_women_pearson.csv’ and the significant correlations was saved in the file ‘significance_matrix_woemn_pearson.csv’, ‘spearman_correlation_results.csv’.

Upon Pearson and Spearman Bivariate correlation analysis in females, the most frequently correlated physical activity variables with other variables are identified as -

- 'C4H73AFD = How many times per day do you do exercise/activity A',
- 'C4H73CC = Exercise/Activity C – Coded' and
- 'C4H73DFW = How many days per week do you do exercise/activity D?'

	C4H73AC = Exercise/Activity A – Coded	C4H73AS = Seasonal exercise/activity A?	C4H73AFD = How many times per day do you do exercise/activity A?	C4H73AFW = How many days per week do you do exercise/activity A?	the average number of minutes/session that you do exercise/activity A?	C4H73AI = What is the intensity of exercise/activity A?	C4H73BC = Exercise/Activity B – Coded	C4H73BS = Seasonal exercise/activity B?	C4H73BFD = How many ti day do exercise/a
C4H73AC = Exercise/Activity A – Coded	NaN	0.986 (p=0.000)	0.985 (p=0.000)	0.987 (p=0.000)	0.973 (p=0.000)	0.977 (p=0.000)	0.646 (p=0.000)	0.649 (p=0.000)	0.646 (p
C4H73AS = Seasonal exercise/activity A?	0.986 (p=0.000)	NaN	0.998 (p=0.000)	0.999 (p=0.000)	0.987 (p=0.000)	0.987 (p=0.000)	0.668 (p=0.000)	0.671 (p=0.000)	0.671 (p
C4H73AFD = How many times per day do you do exercise/activity A?	0.985 (p=0.000)	0.998 (p=0.000)	NaN	0.998 (p=0.000)	0.988 (p=0.000)	0.987 (p=0.000)	0.666 (p=0.000)	0.668 (p=0.000)	0.668 (p
C4H73AFW = How many days per week do you do exercise/activity A?	0.987 (p=0.000)	0.999 (p=0.000)	0.998 (p=0.000)	NaN	0.987 (p=0.000)	0.988 (p=0.000)	0.667 (p=0.000)	0.669 (p=0.000)	0.669 (p
C4H73AM =									

Figure 14: Pearson bivariate correlation with AHA variables in females

```
Top 3 Physical Activity/Exercise Variables:
C4H73AFD = How many times per day do you do exercise/activity A?
C4H73CC = Exercise/Activity C – Coded
C4H73DFW = How many days per week do you do exercise/activity D?

Total sample size (N): 6990
<ipython-input-46-22cedf4b8cca>:58: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  pa_corr_count['Variable'] = pa_corr_count.apply(lambda row: row['Variable 1'] if row['Variable 1'] in pa_vars else row['Variable 2'], axis=1)
```

Figure 15: Top 3 Physical activity/ Exercise variables which are most frequently correlated with AHA Life essential variables

Identification and understanding of physical activity variables in the final dataset:

The below are the list of physical activity/ exercise related variables in the final dataset –

1. C4H73AC = Exercise/Activity A – Coded
2. C4H73AS = Seasonal exercise/activity A?
3. C4H73AFD = How many times per day do you do exercise/activity A?
4. C4H73AFW = How many days per week do you do exercise/activity A?
5. C4H73AM = What is the average number of minutes/session that you do exercise/activity A?
6. C4H73AI = What is the intensity of exercise/activity A?
7. C4H73BC = Exercise/Activity B – Coded
8. C4H73BS = Seasonal exercise/activity B?
9. C4H73BFD = How many times per day do you do exercise/activity B?
10. C4H73BFW = How many days per week do you do exercise/activity B?
11. C4H73BM = What is the average number of minutes/session that you do exercise/activity B?
12. C4H73BI = What is the intensity of exercise/activity B?

13. C4H73CC = Exercise/Activity C – Coded C4H73CS = Seasonal exercise/activity C?
14. C4H73CFD = How many times per day do you do exercise/activity C? C4H73CFW = How many days per week do you do exercise/activity C?
15. C4H73CM = What is the average number of minutes/session that you do exercise/activity C? C4H73CI = What is the intensity of exercise/activity C?
16. C4H73DC = Exercise/Activity D – Coded
17. C4H73DS = Seasonal exercise/activity D?
18. C4H73DFD = How many times per day do you do exercise/activity D?
19. C4H73DFW = How many days per week do you do exercise/activity D?
20. C4H73DM = What is the average number of minutes/session that you do exercise/activity D?
21. C4H73DI = What is the intensity of exercise/activity D?
22. C4H73EC = Exercise/Activity E – Coded
23. C4H73ES = Seasonal exercise/activity E?
24. C4H73EFD = How many times per day do you do exercise/activity E?
25. C4H73EFW = How many days per week do you do exercise/activity E?
26. C4H73EM = What is the average number of minutes/session that you do exercise/activity E?
27. C4H73EI = What is the intensity of exercise/activity E?
28. C4H73FC = Exercise/Activity F – Coded
29. C4H73FS = Seasonal exercise/activity F?
30. C4H73FFD = How many times per day do you do exercise/activity F?
31. C4H73FFW = How many days per week do you do exercise/activity F?
32. C4H73FM = What is the average number of minutes/session that you do exercise/activity F?
33. C4H73FI = What is the intensity of exercise/activity F?
34. C4H73GC = Exercise/Activity G – Coded
35. C4H73GS = Seasonal exercise/activity G?
36. C4H73GFD = How many times per day do you do exercise/activity G?
37. C4H73GFW = How many days per week do you do exercise/activity G?
38. C4H73GM = What is the average number of minutes/session that you do exercise/activity G?
39. C4H73GI = What is the intensity of exercise/activity G?

Based on the documentation in the codebooks of the dataset, the letters (A, B, C, D, E, F, G) represent different types of exercises or physical activities reported by the respondents. From the variable names and labels, we can infer that:

- C4H73AC, C4H73BC, C4H73CC, etc. represent the coded type of activity for each category (A, B, C, etc.).
- C4H73AS, C4H73BS, C4H73CS, etc. indicate whether the activity is seasonal or not.
- C4H73AFD, C4H73BFD, C4H73CFD, etc. represent the frequency per day for each activity.
- C4H73AFW, C4H73BFW, C4H73CFW, etc. represent the frequency per week for each activity.

- C4H73AM, C4H73BM, C4H73CM, etc. represent the average minutes per session for each activity.
- C4H73AI, C4H73BI, C4H73CI, etc. represent the intensity of each activity.

Looking at the coded values for each activity (e.g., C4H73AC, C4H73BC, etc.), we see categories like:

- Bicycling
- Conditioning exercise
- Dancing
- Fishing and hunting
- Home activities
- Home repair
- Lawn and garden
- Miscellaneous
- Music playing
- Occupation
- Religious activities
- Running
- Sports
- Treadmill
- Volunteer activities
- Walking
- Water activities
- Winter activities
- Combined activities
- Inapp

Among these, the categories that likely represent leisure-time physical activities are identified as:

- Bicycling
- Conditioning exercise
- Dancing
- Sports
- Walking
- Water activities
- Winter activities
- Treadmill

Creating refined total physical activity variable:

The physical activity variables are combined together based on the most frequently correlated physical activity variables (coded activity type variable (e.g., C4H73AC, C4H73BC, etc.), and the corresponding variables for frequency per week (C4H73AFW), minutes per session (C4H73BM), and intensity (C4H73BI)) into one "total physical activity" variable which can tell us how many minutes of physical activity per week.

Upon frequency distribution and cross tabulation of coded activity type variable (e.g., C4H73AC, C4H73BC, etc.), and the corresponding variables for frequency per week

(C4H73AFW), minutes per session (C4H73BM), and intensity (C4H73BI), here are the summary of the findings:

1. Exercise/Activity A:
    - Walking (16.0) and conditioning exercise (2.0) are the most common activities in category A.
    - Most activities in category A are done 3-5 days per week, with an average session duration of 30-60 minutes.
    - The intensity of activities in category A is mostly moderate (2.0) or light (3.0).
2. Exercise/Activity B:
    - Sports (13.0), walking (16.0), and conditioning exercise (2.0) are the most common activities in category B.
    - Most activities in category B are done 1-3 days per week, with an average session duration of 30-60 minutes.
    - The intensity of activities in category B is mostly moderate (2.0) or light (3.0).
3. Exercise/Activity C:
    - Lawn and garden (7.0) is the most common activity in category C.
    - Most activities in category C are done 3 days per week, with an average session duration of 30-180 minutes.
    - The intensity of activities in category C is mostly moderate (2.0) or light (3.0).
4. Exercise/Activity D:
    - Conditioning exercise (2.0), walking (16.0), and home activities (5.0) are the most common activities in category D.
    - Most activities in category D are done 1-2 days per week, with an average session duration of 30-120 minutes.
    - The intensity of activities in category D is mostly light (3.0) or moderate (2.0).
5. Exercise/Activity E:
    - Lawn and garden (7.0), home activities (5.0), and fishing and hunting (4.0) are the most common activities in category E.
    - Most activities in category E are done 1-3 days per week, with an average session duration of 30-60 minutes.
    - The intensity of activities in category E is mostly light (3.0) or moderate (2.0).
6. Exercise/Activity F:
    - Home activities (5.0) and conditioning exercise (2.0) are the most common activities in category F.
    - Most activities in category F are done 1-3 days per week, with an average session duration of 20-30 minutes.
    - The intensity of activities in category F is mostly light (3.0) or moderate (2.0).
7. Exercise/Activity G:
    - Conditioning exercise (2.0) and lawn and garden (7.0) are the only reported activities in category G.
    - Most activities in category G are done 1-7 days per week, with an average session duration of 60-90 minutes.
    - The intensity of activities in category G is mostly moderate (2.0) or light (3.0).

Analysis of correlation between total_physical_activity variable and other variables in the final dataset:

The correlation results are saved into the file ‘correlation of total physical activity with other variablescsv’ and ‘spearman_correlation_results.csv’ and correlations for total_physical_activity is saved to ‘physical_activity_correlations.xlsx’.

```
import pandas as pd

# Step 1: Load the data into a pandas DataFrame
data = pd.read_csv('final_dataset.csv')

# Step 2: Identify the relevant activity categories
activity_categories = ['A', 'B', 'C', 'D', 'E', 'F', 'G']
selected_activities = [1, 2, 3, 13, 14, 16, 17, 18]  # Codes for the selected activities

# Step 3: Calculate the total minutes of physical activity per week for each participant
data['total_physical_activity'] = 0

for category in activity_categories:
    coded_var = f'C4H73{category}C = Exercise/Activity {category} - Coded'
    freq_var = f'C4H73{category}FW = How many days per week do you do exercise/activity {category}?'
    min_var = f'C4H73{category}M = What is the average number of minutes/session that you do exercise/activity {category}?
    int_var = f'C4H73{category}I = What is the intensity of exercise/activity {category}?'

    if coded_var in data.columns and freq_var in data.columns and min_var in data.columns:
        activity_mask = data[coded_var].isin(selected_activities)
        data.loc[activity_mask, 'total_physical_activity'] += data[freq_var] * data[min_var]

# Step 4: Test correlations with the most recent correlation matrix
correlation_matrix = data.corr(method='spearman')

# Print the correlations of the new variable with other variables
print("Correlations of 'total_physical_activity' with other variables:")
print(correlation_matrix['total_physical_activity'].sort_values(ascending=False))
```

```
Correlations of 'total_physical_activity' with other variables:
total_physical_activity                                 1.000000
C4Q1AA = MASQ Felt like I had a lot of energy           0.239904
C4Q16C = GDLIFE The conditions of my life are excellent.  0.155025
C4QSW_SL = Subjective WellBeing - Satisfaction with Life Scale  0.150644
C4Q1FFF = MASQ Seemed to move quickly and easily        0.150458
```

Figure 16: correlation analysis of total physical activity variable and other variables in the final dataset

```
Correlations for total_physical_activity:
                                        Variable 1                Variable 2  \
265                          M2ID=MIDUS 2 ID number  total_physical_activity
530                                       M2FAMNUM_x  total_physical_activity
794     C1PB1=Highest level of education completed  total_physical_activity
1057         C1PB2A1 = Employment 1/2008 - Working  total_physical_activity
1319   C1PB2A2 = Employment 1/2008 - Self-employed  total_physical_activity
...                                             ...                       ...
35500              Family history of breast cancer  total_physical_activity
35504                     Family history of stroke  total_physical_activity
35507       Family history of circulation problems  total_physical_activity
35509        Family history of high blood pressure  total_physical_activity
35510              Family history of heart disease  total_physical_activity

       Correlation       P-value Significance      N
265       0.013568  1.332072e-01                12249
530      -0.039483  1.236555e-05           **   12249
794       0.034251  1.497910e-04           **   12249
1057      0.026656  3.174334e-03           **   12249
1319      0.010720  2.355048e-01                12249
...            ...           ...          ...    ...
35500     0.047292  1.636622e-07           **   12249
35504     0.024753  6.150607e-03           **   12249
35507    -0.035540  8.346383e-05           **   12249
35509    -0.027548  2.295311e-03           **   12249
35510     0.066496  1.747114e-13           **   12249

[266 rows x 6 columns]
```

Analysis of time-points of MIDUS and Biomarker datasets to establish temporal relationship between physical activity and biomarkers:

MIDUS 3 (2013-2014):

- Data collection period: May 2013 - November 2014
- This is the third wave of the longitudinal study
- Includes many of the same variables as previous waves, plus some new questions on topics like economic recession experiences
- Can be linked to MIDUS 1 and MIDUS 2 data using the M2ID variable

MIDUS 2 (including Biomarker Project) (2004-2009):

- Data collection period: 2004-2009
- This is the second wave of the longitudinal study
- The Biomarker Project was conducted from July 30, 2004, to May 31, 2009
- It includes biological and physiological measures not present in the first wave
- Can be linked to other MIDUS datasets using the M2ID variable

MIDUS 1 (1995-1996):

- Data collection period: 1995-1996
- This was the original baseline survey
- It established the foundation for the longitudinal study

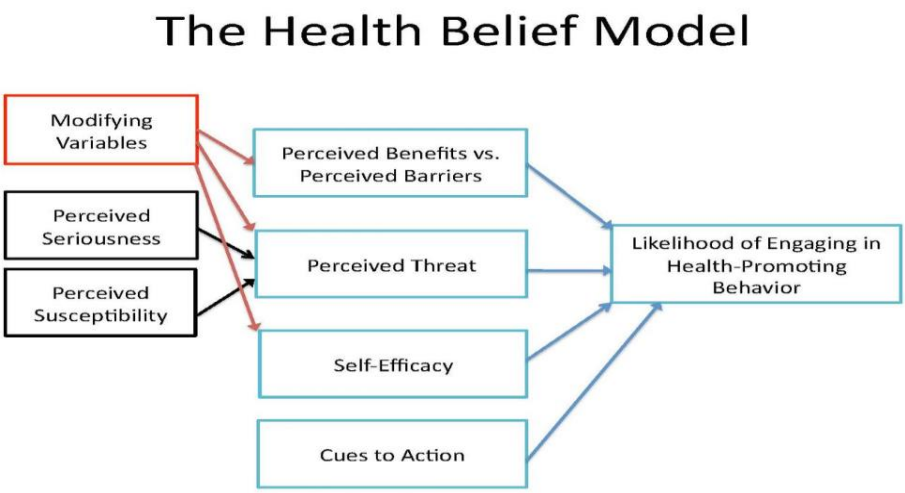
Key points:

- All waves can be linked using the M2ID variable, allowing for longitudinal analysis
- Many core variables are repeated across waves, with some additions in later waves (e.g., biomarkers in MIDUS 2, recession questions in MIDUS 3)
- MIDUS is designed as a longitudinal study following the same participants over time
- Only living respondents who completed the MIDUS 2 phone interview were eligible for MIDUS 3
- The MIDUS 2 Biomarker Project data can be particularly valuable for comparing biological measures with the MIDUS 3 Biomarker data

Correlation analysis between total physical activity variables and motivation related variables (in the behaviour models):

Health Belief Model-

https://psu.pb.unizin.org/app/uploads/sites/16/2017/08/The_Health_Belief_Model.pdf.jpg



We have identified relevant variables that fit into the above model using the keywords as below:

- Physical Activity: total_physical_activity

- Depression: 'depress', 'sad', 'blue'
- Motivation: ‘motiv’, ‘energy’, ‘vigor’
- Attitudes: attitude, fee about, think about
- Self-efficiency: confiden, able to, can do
- Intention: intend, plan to, will
- Perceived severity: serious, severe, consequen
- Perceived benefits: benefit, good for, help
- Perceived barriers: difficult, hard to, prevent
- Subjective norms: others think, expect me to, should
- Perceived behavioural control: control, up to me, decide

We have identified below relevant variables based on the above key words as per the model -

Physical Activity variables:

- total_physical_activity

Depression variables:

- C4Q1A = MASQ Felt sad
- C4Q1L = MASQ Felt depressed
- C4Q3C = CESD I felt that I could not shake off the blues even with the help of my family and friends
- C4Q3F = CESD I felt depressed
- C4Q3R = CESD I felt sad
- C4QCESDDA = CESD Depressive Affect Subscale (C,F,I,J,N,Q,R)
- C4H1V = Have you ever had depression?

Motivation variables:

- C4Q1AA = MASQ Felt like I had a lot of energy

Attitudes variables:

NA

Self-efficacy variables:

- C4Q1FF = MASQ Was unable to relax
- C4Q4B = PSS Unable to control the important things in your life
- C4Q4D = PSS Confident about your ability to handle your personal problems
- C4Q4G = PSS Able to control irritations in your life

intention variables:

- Perceived Susceptibility variables:
- Perceived Severity variables:

Perceived Benefits variables:

- C4Q3C = CESD I felt that I could not shake off the blues even with the help of my family and friends

Perceived Barriers variables:

- C4Q4J = PSS Difficulties were piling up so high that you couldnt overcome them
- C4AD510 = Sun AM How Difficult To Fall Asleep
- C4AD513 = Sun AM If Woke, Difficulty Getting To Sleep

Subjective Norms variables:

NA

Perceived Behavioral Control variables:

- C4Q4B = PSS Unable to control the important things in your life
- C4Q4G = PSS Able to control irritations in your life
- C4Q4I = PSS Angered because of things that were outside of your control

The correlation analysis between the above variables and total_physical_activity variable was performed and the result was sabed to 'total Physical activity_motivation_correlations_MIDUS3_Health Belief Model.csv'

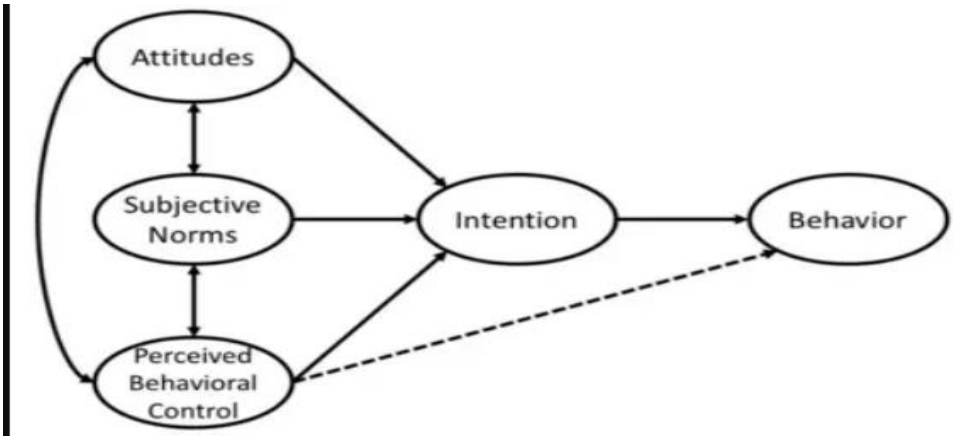
```
import pandas as pd
import numpy as np
from scipy import stats

# Load the MIDUS 3 dataset
midus3_data = pd.read_csv('final_dataset_updated.csv')

# Define categories of variables we're looking for
categories = {
    'Physical Activity': ['total_physical_activity'],
    'Depression': ['depress', 'sad', 'blue'],
    'Motivation': ['motiv', 'energy', 'vigor'],
    'Attitudes': ['attitude', 'feel about', 'think about'],
    'Self-efficacy': ['confiden', 'able to', 'can do'],
    'Intention': ['intend', 'plan to', 'will'],
    'Perceived Susceptibility': ['risk', 'chance', 'likely'],
    'Perceived Severity': ['serious', 'severe', 'consequen'],
    'Perceived Benefits': ['benefit', 'good for', 'help'],
    'Perceived Barriers': ['difficult', 'hard to', 'prevent'],
    'Subjective Norms': ['others think', 'expect me to', 'should'],
    'Perceived Behavioral Control': ['control', 'up to me', 'decide']
}

# Function to search for relevant variables
def find_relevant_variables(data, categories):
    relevant_vars = {}
    for category, keywords in categories.items():
        relevant_vars[category] = [col for col in data.columns
                                   if any(keyword.lower() in col.lower() for keyword in keywords)]
    return relevant_vars

# Find relevant variables
relevant_variables = find_relevant_variables(midus3_data, categories)

# Print relevant variables
for category, vars in relevant_variables.items():
    print(f"\n{category} variables:")
    for var in vars:
        print(f"- {var}")

# Analyze correlations between physical activity and motivation-related variables
pa_vars = relevant_variables['Physical Activity']
motivation_vars = sum([relevant_variables[cat] for cat in categories.keys() if cat != 'Physical Activity'], [])

correlation_results = []
```

The top correlations between total_physical_activity variable and health belief model based variables are identified as -

Physical Activity Variable	Motivation Variable	Correlation	P-value
total_physical_activity	C4Q1AA = MASQ Felt like I had a lot of energy	0.239904	6.87E-160
total_physical_activity	C4AD510 = Sun AM How Difficult To Fall Asleep	-0.121652	1.32E-41
total_physical_activity	C4AD513 = Sun AM If Woke, Difficulty Getting T...	-0.105802	7.82E-32
total_physical_activity	C4Q4G = PSS Able to control irritations in you...	0.049153	5.25E-08
total_physical_activity	C4Q4G = PSS Able to control irritations in you...	0.049153	5.25E-08
total_physical_activity	C4Q4J = PSS Difficulties were piling up so hig...	-0.046543	2.56E-07
total_physical_activity	C4Q4I = PSS Angered because of things that wer...	-0.045355	5.12E-07
total_physical_activity	C4Q4D = PSS Confident about your ability to ha...	0.042646	2.34E-06
total_physical_activity	C4Q3C = CESD I felt that I could not shake off...	-0.037613	3.13E-05
total_physical_activity	C4Q3C = CESD I felt that I could not shake off...	-0.03761	3.13E-05

Theory Planned Behavior model:

https://cdn.serc.carleton.edu/images/ASCN/change_theories/collection/figure_1._theory_planned_behavior_model_adapted_from_ajzen_2005._456.webp



We have identified relevant variables that fit into the above model using the keywords as below:

- 'Physical Activity': 'total_physical_activity'
- 'Attitude': 'attitude', 'feel about', 'think about', 'opinion', 'belief', 'view', 'perception', 'like', 'dislike', 'enjoy', 'value', 'importance'
- 'Subjective Norm': 'others think', 'expect me to', 'should', 'social pressure', 'norm', 'peer influence', 'family influence', 'social expectation', 'approval', 'disapproval',
- 'Perceived Behavioral Control': 'control', 'up to me', 'decide', 'able to', 'can do', 'confidence', 'self-efficacy', 'capability', 'ease', 'difficulty',
- 'Intention': 'intend', 'plan to', 'will', 'aim to', 'goal', 'objective', 'purpose', 'determination', 'motivation', 'commitment',
- 'Behavioral Beliefs': 'outcome', 'result in', 'lead to', 'consequence', 'effect', 'impact', 'benefit', 'drawback', 'advantage', 'disadvantage',
- 'Normative Beliefs': 'important others', 'approve', 'disapprove', 'social support', 'encouragement', 'discouragement', 'social influence', 'role model',

- 'Control Beliefs': 'factors', 'facilitate', 'impede', 'barrier', 'obstacle', 'enabler', 'resource', 'opportunity', 'constraint', 'limitation'

The below list of variables were identified relevant to the above behaviour models:

1. Physical Activity variables:
   total_physical_activity

2. Attitude variables:
   C4Q1U = MASQ Felt like a failure
   C4Q1V = MASQ Felt like I was having a lot of fun
   C4Q1AA = MASQ Felt like I had a lot of energy
   C4Q1EE = MASQ Felt like crying
   C4Q1PP = MASQ Felt like I was choking
   C4Q1QQ = MASQ Looked forward to things with enjoyment
   C4Q1UU = MASQ Felt like I had a lot of interesting things to do
   C4Q1WW = MASQ Felt like I had accomplished a lot
   C4Q1XX = - MASQ Felt like it took extra effort get started
   C4Q1YY = MASQ Felt like nothing was very enjoyable
   C4Q1AAA = MASQ Felt like I had a lot to look forward to
   C4Q1EEE = MASQ Felt like there wasn't anything interesting or fun to do
   C4Q1KKK = MASQ Felt like I am a good person
   C4Q3B = CESD I did not feel like eating; my appetite was poor
   C4Q3P = CESD I enjoyed life
   C4Q3S = CESD I felt that people dislike me
   C4H62 = Since we last interviewed you, have you tried to quit smoking?

3. Subjective Norm variables:
   NA

4. Perceived Behavioural Control Variables:
   C4Q1FF - MASQ Was unable to relax
   C4Q4B = PSS Unable to control the important things in your life
   C4Q4G = PSS Able to control irritations in your life
   C4Q4I = PSS Angered because of things that were outside of your control
   C4H1A = Have you ever had heart disease?
   C4H1G = Have you ever had anemia or other blood disease?
   C4AD513 = Sun AM If Woke, Difficulty Getting To Sleep
   Family history of heart disease

5. Intention Variables:
   NA
6. Behavioral Beliefs variables:
   NA
7. Normative Belief variables:
   NA
8. Control Beliefs variables:
   NA

The correlation between the above variables are saved to 'total_Physical_activity_TPB_correlations_MIDUS3.csv'

```python
# Define categories of variables based on the Theory of Planned Behavior
categories = {
    'Physical Activity': ['total_physical_activity'],
    'Attitude': ['attitude', 'feel about', 'think about', 'opinion', 'belief', 'view', 'perception', 'like', 'dislike', 'enjoy', 'value', 'importance'],
    'Subjective Norm': ['others think', 'expect me to', 'should', 'social pressure', 'norm', 'peer influence', 'family influence', 'social expectation', 'approval', 'disapproval'],
    'Perceived Behavioral Control': ['control', 'up to me', 'decide', 'able to', 'can do', 'confidence', 'self-efficacy', 'capability', 'ease', 'difficulty'],
    'Intention': ['intend', 'plan to', 'will', 'aim to', 'goal', 'objective', 'purpose', 'determination', 'motivation', 'commitment'],
    'Behavioral Beliefs': ['outcome', 'result in', 'lead to', 'consequence', 'effect', 'impact', 'benefit', 'drawback', 'advantage', 'disadvantage'],
    'Normative Beliefs': ['important others', 'approve', 'disapprove', 'social support', 'encouragement', 'discouragement', 'social influence', 'role model'],
    'Control Beliefs': ['factors', 'facilitate', 'impede', 'barrier', 'obstacle', 'enabler', 'resource', 'opportunity', 'constraint', 'limitation']
}

# Function to search for relevant variables
def find_relevant_variables(data, categories):
    relevant_vars = {}
    for category, keywords in categories.items():
        relevant_vars[category] = [col for col in data.columns
                                   if any(keyword.lower() in col.lower() for keyword in keywords)]
    return relevant_vars

# Find relevant variables
relevant_variables = find_relevant_variables(midus3_data, categories)

# Print relevant variables
for category, vars in relevant_variables.items():
    print(f"\n{category} variables:")
    for var in vars:
        print(f"- {var}")

# Analyze correlations between physical activity and TPB variables
pa_vars = relevant_variables['Physical Activity']
tpb_vars = sum([relevant_variables[cat] for cat in categories.keys() if cat != 'Physical Activity'], [])

correlation_results = []

for pa_var in pa_vars:
    for tpb_var in tpb_vars:
        corr, p_value = stats.spearmanr(midus3_data[pa_var], midus3_data[tpb_var], nan_policy='omit')
        correlation_results.append({
            'Physical Activity Variable': pa_var,
            'TPB Variable': tpb_var,
            'Correlation': corr,
            'P-value': p_value
        })

correlation_df = pd.DataFrame(correlation_results)
```

Figure 17: Python code generating correlation between behaviour model variables and physical activity variable

Top correlations between physicial activity and theory of planned behavior variables:

Physical Activity Variable	TPB Variable	Correlation	P-value
total_physical_activity	C4Q1AA = MASQ Felt like I had a lot of energy	0.239904	6.87E-160
total_physical_activity	C4Q1WW = MASQ Felt like I had accomplished a lot	0.145632	4.83E-59
total_physical_activity	C4Q1UU = MASQ Felt like I had a lot of interes...	0.134419	1.72E-50
total_physical_activity	C4H62 = Since we last interviewed you, have yo...	0.133989	3.55E-50
total_physical_activity	C4Q1XX = - MASQ Felt like it took extra effort...	-0.116404	3.22E-38
total_physical_activity	C4Q1PP = MASQ Felt like I was choking	-0.11448	5.17E-37
total_physical_activity	C4AD513 = Sun AM If Woke, Difficulty Getting T...	-0.105802	7.82E-32
total_physical_activity	C4H1A = Have you ever had heart disease?	0.103726	1.18E-30
total_physical_activity	C4Q3B = CESD I did not feel like eating; my ap...	-0.095665	2.65E-26
total_physical_activity	C4Q1V = MASQ Felt like I was having a lot of fun	0.095215	4.52E-26

Identifying the variables related to physical activity (at least two timepoints), cancer and cardiovascular related variables along with demographics (age, BMI, sex education, income, etc) and finding correlations between them using z-scores

Physical Activity Variables:

- C4H73AC = Exercise/Activity A – Coded

- C4H73AS = Seasonal exercise/activity A?
- C4H73AFD = How many times per day do you do exercise/activity A?
- C4H73AFW = How many days per week do you do exercise/activity A?
- C4H73AM = What is the average number of minutes/session that you do exercise/activity A?
- C4H73AI = What is the intensity of exercise/activity A?
- C4H73BC = Exercise/Activity B – Coded
- C4H73BS = Seasonal exercise/activity B?
- C4H73BFD = How many times per day do you do exercise/activity B?
- C4H73BFW = How many days per week do you do exercise/activity B?
- C4H73BM = What is the average number of minutes/session that you do exercise/activity B?
- C4H73BI = What is the intensity of exercise/activity B?
- C4H73CC = Exercise/Activity C – Coded
- C4H73CS = Seasonal exercise/activity C?
- C4H73CFD = How many times per day do you do exercise/activity C?
- C4H73CFW = How many days per week do you do exercise/activity C?
- C4H73CM = What is the average number of minutes/session that you do exercise/activity C?
- C4H73CI = What is the intensity of exercise/activity C?
- C4H73DC = Exercise/Activity D – Coded
- C4H73DS = Seasonal exercise/activity D?
- C4H73DFD = How many times per day do you do exercise/activity D?
- C4H73DFW = How many days per week do you do exercise/activity D?
- C4H73DM = What is the average number of minutes/session that you do exercise/activity D? C4H73DI = What is the intensity of exercise/activity D?
- C4H73EC = Exercise/Activity E – Coded
- C4H73ES = Seasonal exercise/activity E?
- C4H73EFD = How many times per day do you do exercise/activity E?
- C4H73EFW = How many days per week do you do exercise/activity E?
- C4H73EM = What is the average number of minutes/session that you do exercise/activity E?
- C4H73EI = What is the intensity of exercise/activity E?
- C4H73FC = Exercise/Activity F – Coded
- C4H73FS = Seasonal exercise/activity F?
- C4H73FFD = How many times per day do you do exercise/activity F?
- C4H73FFW = How many days per week do you do exercise/activity F?
- C4H73FM = What is the average number of minutes/session that you do exercise/activity F?
- C4H73FI = What is the intensity of exercise/activity F?
- C4H73GC = Exercise/Activity G – Coded
- C4H73GS = Seasonal exercise/activity G?
- C4H73GFD = How many times per day do you do exercise/activity G?
- C4H73GFW = How many days per week do you do exercise/activity G?
- C4H73GM = What is the average number of minutes/session that you do exercise/activity G?

- C4H73GI = What is the intensity of exercise/activity G?
- total_physical_activity

Cancer variables:

- 'C1PA26 = Ever had cancer',
- 'C1PA28AA = Age breast cancer diagnosed',
- 'C4H1P = Have you ever had cancer?',
- 'Family history of breast cancer'

Cardiovascular Variables:

- 'C1PA24 = High blood pressure ever diagnosed',
- 'C1PA24A = Number years ago told high blood pressure',
- 'C1PA24B = Ever taken high blood pressure medicine',
- 'C1PA24EC = High blood pressure therapy - Exercise',
- 'C4H1A = Have you ever had heart disease?',
- 'C4H1B = Have you ever had high blood pressure?',
- 'C4H1C = Have you ever had circulation problems?',
- 'C4H1D = Have you ever had blood clots?',
- 'C4H1E = Have you ever had a heart murmur?',
- 'C4H1F = Have you ever had a TIA (mini-stroke) or stroke?',
- 'C4H1H = Have you ever had cholesterol problems?',
- 'C4P1F1D = BP (sitting) 1 diastolic',
- 'C4P1F2S = BP (sitting) 2 systolic',
- 'C4P1F2D = BP (sitting) 2 diastolic',
- 'C4P1F3S = BP (sitting) 3 systolic',
- 'C4P1F3D = BP (sitting) 3 diastolic',
- 'C4P1GS = Average BP (sitting) systolic',
- 'C4P1GD = Average BP (sitting) diastolic',
- 'C4P1GS23 = Average of 2nd and 3rd systolic BPs',
- 'C4P1GD23 = Average of 2nd and 3rd diastolic BPs',
- 'C4BCHOL = Blood Total Cholesterol (mg/dL)',
- 'C4BTRIGL = Blood Triglycerides (mg/dL)',
- 'C4BHDL = Blood HDL Cholesterol (mg/dL)',
- 'C4BLDL = Blood LDL Cholesterol (mg/dL)',
- 'C4BRTHDL = Blood Ratio Total / HDL Cholesterol',
- 'Family history of stroke',
- 'Family history of circulation problems',
- 'Family history of high blood pressure',
- 'Family history of heart disease'

Demographics variables:

- C1PB1 = Highest level of education completed
- B1PGENDER = Gender

The correlations between demographic, physical activity, cancer and cardiovascular variables was calculated and saved as 'significant_correlations_PA_cancer_CVD_demographics.csv' and

their z scores were saved as ‘significant_correlations_PA_cancer_CVD_demographics_relevant_variables_zscores.csv’ and ‘Z-scores for relevant variables saved to 'significant_correlations_PA_cancer_CVD_demographics_relevant_variables_zscores.csv’.

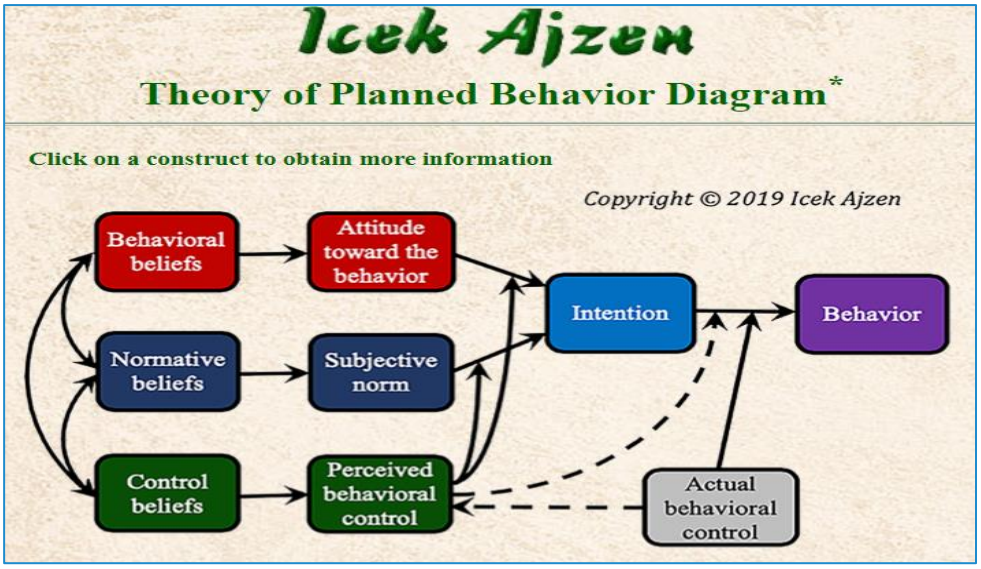
```
# Calculate p-values
def calculate_pvalues(df):
    df = df.dropna()._get_numeric_data()
    dfcols = pd.DataFrame(columns=df.columns)
    pvalues = dfcols.transpose().join(dfcols, how='outer')
    for r in df.columns:
        for c in df.columns:
            pvalues[r][c] = round(stats.spearmanr(df[r], df[c])[1], 4)
    return pvalues

pvalues = calculate_pvalues(relevant_data_z)

# Identify significant correlations
alpha = 0.05  # Significance level
significant_corr_list = []

for var1 in all_relevant_vars:
    for var2 in all_relevant_vars:
        if var1 != var2:
            corr = correlation_matrix.loc[var1, var2]
            pval = pvalues.loc[var1, var2]
            if pval < alpha:
                significant_corr_list.append({
                    'Variable 1': var1,
                    'Variable 2': var2,
                    'Correlation': corr,
                    'P-value': pval
                })

significant_corr = pd.DataFrame(significant_corr_list)

# Sort by absolute correlation value
significant_corr['Abs_Correlation'] = abs(significant_corr['Correlation'])
significant_corr = significant_corr.sort_values('Abs_Correlation', ascending=False)
significant_corr = significant_corr.drop('Abs_Correlation', axis=1)

# Print results
print("\nSignificant correlations with total physical activity:")
print(significant_corr[significant_corr['Variable 1'] == 'total_physical_activity'])

# Save results to CSV
significant_corr.to_csv('significant_correlations_PA_cancer_CVD_demographics_zscores.csv', index=False)
print("\nFull results saved to 'significant_correlations_PA_cancer_CVD_demographics_zscores.csv'")
```

Theory Determinants:

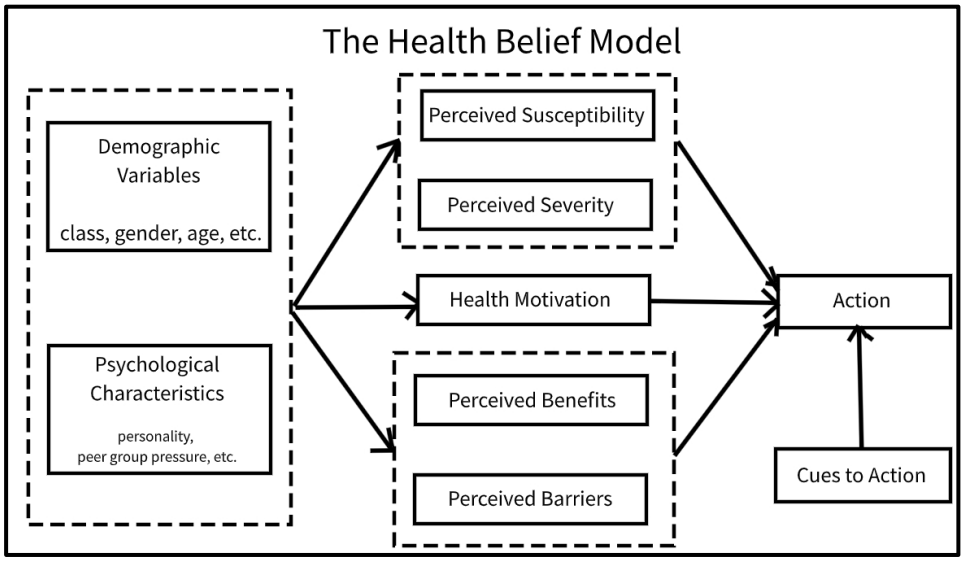
Identifying behaviour theory variables and assessing their correlation with the total physical activity variables using z-scores:

Behaviour theory variables - these are psychological or social correlates to exercise/physical activity (behaviour). Behaviour theories as below -

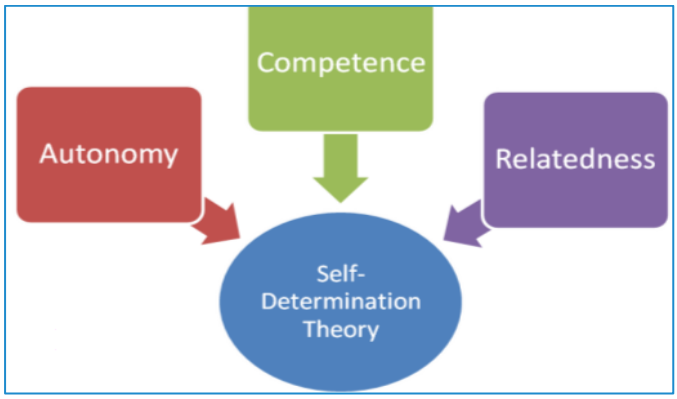
- https://people.umass.edu/aizen/tpb.diag.html



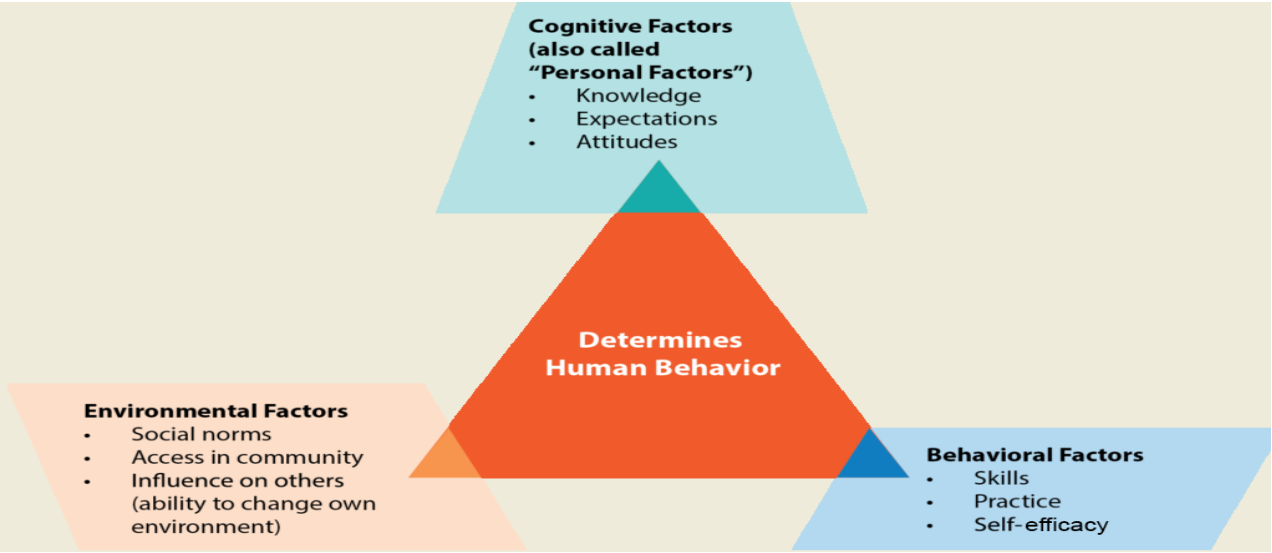
- https://en.wikipedia.org/wiki/Health_belief_model



- https://en.wikipedia.org/wiki/Self-determination_theory



- https://sbccimplementationkits.org/sbcc-in-emergencies/social-cognitive-learning-theory/



Based on the above theory models, the below relevant variables are identified from the dataset 'final_dataset_updated' -

'Theory of Planned Behavior':

'Attitudes':

- 'C4Q16A = GDLIFE Compared to most of my peers, I consider myself to be more happy.',
- 'C4Q16B = GDLIFE In most ways my life is close to my ideal.',
- 'C4Q16C = GDLIFE The conditions of my life are excellent.',
- 'C4Q16D = GDLIFE I am satisfied with my life.',
- 'C4Q16E = GDLIFE So far I have gotten the important things I want in life',
- 'C4Q16F = GDLIFE I have so much in life to be thankful for',
- 'C4Q16G = GDLIFE I am grateful to a wide variety of people.'

'Subjective Norms':

- 'C4Q18B = GDLIFE Be needed by others',
- 'C4Q18C = GDLIFE Be in harmony with others and surrounding events',
- 'C4Q18F = GDLIFE To receive sympathy from others',
- 'C4Q18G = GDLIFE To receive respect from others'

'Perceived Behavioral Control':

- 'C4Q4D = PSS Confident about your ability to handle your personal problems',
- 'C4Q4G = PSS Able to control irritations in your life',
- 'C4Q4H = PSS Felt that you were on top of things',
- 'C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it'

'Intention': No clear measures of intention

'Behavior':

- 'total_physical_activity'

Health Belief Model:

'Perceived Susceptibility':

- 'Family history of breast cancer',
- 'Family history of stroke',
- 'Family history of circulation problems',
- 'Family history of high blood pressure',
- 'Family history of heart disease',
- "Family history of Alzheimer's"

'Perceived Severity':

- 'C4H1A = Have you ever had heart disease?',
- 'C4H1B = Have you ever had high blood pressure?',
- 'C4H1C = Have you ever had circulation problems?',
- 'C4H1D = Have you ever had blood clots?',
- 'C4H1E = Have you ever had a heart murmur?',
- 'C4H1F = Have you ever had a TIA (mini-stroke) or stroke?',
- 'C4H1G = Have you ever had anemia or other blood disease?',
- 'C4H1H = Have you ever had cholesterol problems?',
- 'C4H1I = Have you ever had diabetes?',
- 'C4H1P = Have you ever had cancer?',
- 'C4H1V = Have you ever had depression?'

'Perceived Benefits':

- 'C1PA24EC = High blood pressure therapy - Exercise'

'Perceived Barriers':

- 'C4Q4J = PSS Difficulties were piling up so high that you couldnt overcome them',
- 'C4Q3G = CESD I felt that everything I did was an effort'

'Cues to Action':

- 'C4H68 = In the past, did anyone in your household smoke tobacco inside your home regularly?',
- 'C4H69 = Currently, does anyone regularly smoke cigarettes or other tobacco products inside your home?'

'Self-Efficacy':

- 'C4Q4D = PSS Confident about your ability to handle your personal problems',
- 'C4Q4G = PSS Able to control irritations in your life',
- 'C4Q4H = PSS Felt that you were on top of things',
- 'C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it'

Self-Determination Theory:

'Autonomy':

- 'C4Q4G = PSS Able to control irritations in your life',
- 'C4Q4H = PSS Felt that you were on top of things'

'Competence':

- 'C4Q4D = PSS Confident about your ability to handle your personal problems',
- 'C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it',
- 'C4Q1O = MASQ Was proud of myself',
- 'C4Q1WW = MASQ Felt like I had accomplished a lot'

'Relatedness':

- 'C4Q17A = LONELY There is no one I can turn to',
- 'C4Q17B = LONELY No one really knows me well',
- 'C4Q17C = LONELY Feel isolated from others',
- 'C4Q17D = LONELY There are people who really understand me',
- 'C4Q17E = LONELY People are around me but not with me',
- 'C4Q17F = LONELY There are people I can talk to',
- 'C4Q17G = LONELY There are people I can turn to',
- 'C4QLONEL = UCLA Loneliness Scale'

'Intrinsic Motivation':

- 'C4Q1AA = MASQ Felt like I had a lot of energy',
- 'C4Q1MM = MASQ Felt really up or lively',
- 'C4Q1UU = MASQ Felt like I had a lot of interesting things to do',
- 'C4Q1V = MASQ Felt like I was having a lot of fun'

'Extrinsic Motivation':

- 'C4Q18F = GDLIFE To receive sympathy from others',
- 'C4Q18G = GDLIFE To receive respect from others'

'Social Cognitive Learning Theory:

'Self-efficacy':

- 'C4Q4D = PSS Confident about your ability to handle your personal problems',
- 'C4Q4G = PSS Able to control irritations in your life',
- 'C4Q4H = PSS Felt that you were on top of things',
- 'C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it

'Outcome Expectations':

- 'C4Q16B = GDLIFE In most ways my life is close to my ideal.',
- 'C4Q16C = GDLIFE The conditions of my life are excellent.',
- 'C4Q16D = GDLIFE I am satisfied with my life.',
- 'C4Q16E = GDLIFE So far I have gotten the important things I want in life

'Goals':

- 'C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it',
- 'C4Q18H = GDLIFE To give something back to society

'Sociostructural Factors':

- 'C1PB1=Highest level of education completed',
- 'B1PGENDER = Gender',
- 'C4H85 = Marital status changed - Current status',
- 'C1PB2A1 = Employment 1/2008 - Working',
- 'C1PB2A2 = Employment 1/2008 - Self-employed',
- 'C1PB2A3 = Employment 1/2008 - Unemployed',
- 'C1PB2A4 = Employment 1/2008 - Temporarily laid off',
- 'C1PB2A5 = Employment 1/2008 - Retired',
- 'C1PB2A6 = Employment 1/2008 - Homemaker',
- 'C1PB2A7 = Employment 1/2008 - Full-time student',
- 'C1PB2A8 = Employment 1/2008 - Part-time student',
- 'C1PB2A9 = Employment 1/2008 - Maternity or sick leave',
- 'C1PB2A10 = Employment 1/2008 - Permanently disabled'

Correlation analysis of Theory planned behaviour model variables:

The correlation between above TPB variables are saved in 'significant_correlations_TPB_PA_according_to_Theory_of_Planned_Behavior_model_zscores.csv' and their Z scores are saved in 'TPB_variables_PA_zscores_according_to_Theory_of_Planned_Behavior_model.csv'.

The significant Significant correlations between TPB variables and total physical activity according to Theory of Planned Behavior model:

TPB Category	Variable	Correlatio	P-value
Attitudes	C4Q16C = GDLIFE The conditions of my life are excellent	0.155025	0
Attitudes	C4Q16E = GDLIFE So far I have gotten the important things I want in	0.134265	0
Attitudes	C4Q16F = GDLIFE I have so much in life to be thankful for	0.1213	0
Attitudes	C4Q16D = GDLIFE I am satisfied with my life	0.11998	0
Subjective Norms	C4Q18G = GDLIFE To receive respect from others	-0.09696	0
Attitudes	C4Q16B = GDLIFE In most ways my life is close to my ideal.	0.087914	0
Attitudes	C4Q16A = GDLIFE Compared to most of my peers, I consider myself	0.069221	0
Subjective Norms	C4Q18F = GDLIFE To receive sympathy from others	-0.05467	0
Perceived Behaviora	C4Q4G = PSS Able to control irritations in your life	0.049153	0
Perceived Behaviora	C4Q4H = PSS Felt that you were on top of things	0.043766	0
Perceived Behaviora	C4Q4D = PSS Confident about your ability to handle your personal p	0.042646	0
Perceived Behaviora	C4Q18D = GDLIFE Have the ability to make a good effort at somethir	0.020647	0.0223

Correlation analysis of Health Belief model variables:

```
if not significant_corr.empty:
    # Sort by absolute correlation value
    significant_corr['Abs_Correlation'] = abs(significant_corr['Correlation'])
    significant_corr = significant_corr.sort_values('Abs_Correlation', ascending=False)
    significant_corr = significant_corr.drop('Abs_Correlation', axis=1)

    # Print results
    print("\nSignificant correlations between HBM variables and total physical activity according to Health Belief Model:")
    print(significant_corr.to_string(index=False))

    # Save results to CSV
    significant_corr.to_csv('significant_correlations_HBM_PA_according_to_Health_Belief_Model_zscores.csv', index=False)
    print("\nFull results saved to 'significant_correlations_HBM_PA_according_to_Health_Belief_Model_zscores.csv'")

    # Calculate average correlations for each HBM category
    avg_correlations = {}
    for category, vars in hbm_variables.items():
        if category != 'Behavior':
            category_vars = [v for v in vars if v in available_vars]
            if category_vars:
                category_corrs = correlation_matrix.loc[category_vars, 'total_physical_activity']
                avg_correlations[category] = category_corrs.mean()

    print("\nAverage correlations for each HBM category with total physical activity:")
    for category, avg_corr in avg_correlations.items():
        print(f"{category}: {avg_corr:.4f}")

    # Save z-scores to CSV
    z_scores_df = relevant_data_z.copy()
    z_scores_df.to_csv('HBM_variables_PA_zscores_according_to_Health_Belief_Model.csv', index=False)
    print("Z-scores for HBM variables and physical activity saved to 'HBM_variables_PA_zscores_according_to_Health_Belief_Model.csv'")
else:
    print("No significant correlations found.")

Available variables: ['Family history of breast cancer', 'Family history of stroke', 'Family history of circulation problems', 'Family history of high blood pressure', 'Family
Missing variables: []

Significant correlations between HBM variables and total physical activity according to Health Belief Model:
     HBM Category                                                                    Variable  Correlation  P-value
Perceived Severity                                 C4H1H = Have you ever had cholesterol problems?     0.130041   0.0000
   Cues to Action C4H69 = Currently, does anyone regularly smoke cigarettes or other tobacco products inside your home?     0.127351   0.0000
Perceived Severity                                   C4H1B = Have you ever had high blood pressure?     0.123167   0.0000
Perceived Severity                                          C4H1I = Have you ever had diabetes?     0.118892   0.0000
```

The significant correlations between HBM variables and total physical activity according to Health Belief Model are saved to 'significant_correlations_HBM_PA_according_to_Health_Belief_Model_zscores.csv' and their Z-scores are saved to 'HBM_variables_PA_zscores_according_to_Health_Belief_Model.csv'.

Significant correlations between HBM variables and total physical activity according to Health Belief Model:

HBM Category	Variable	Correlation	P-Value
Perceived Severity	C4H1H = Have you ever had cholesterol problems?	0.130041	0
Cues to Action	C4H69 = Currently, does anyone regularly smoke cigarettes or other tobacco products inside your home?	0.127351	0
Perceived Severity	C4H1B = Have you ever had high blood pressure?	0.123167	0
Perceived Severity	C4H1I = Have you ever had diabetes?	0.118892	0
Perceived Severity	C4H1A = Have you ever had heart disease?	0.103726	0
Perceived Severity	C4H1C = Have you ever had circulation problems?	0.098071	0
Perceived Barriers	C4Q3G = CESD I felt that everything I did was an effort	-0.09459	0
Perceived Severity	C4H1D = Have you ever had blood clots?	0.082521	0
Perceived Severity	C4H1F = Have you ever had a TIA (mini-stroke) or stroke?	0.078772	0
Perceived Susceptibility	Family history of Alzheimer's	0.066705	0
Perceived Susceptibility	Family history of heart disease	0.066496	0
Cues to Action	C4H68 = In the past, did anyone in your household smoke tobacco inside your home regularly?	0.053541	0

Self-Efficacy	C4Q4G = PSS Able to control irritations in your life	0.049153	0
Perceived Susceptibility	Family history of breast cancer	0.047292	0
Perceived Barriers	C4Q4J = PSS Difficulties were piling up so high that you couldnt overcome them	-0.04654	0
Perceived Severity	C4H1P = Have you ever had cancer?	0.045094	0
Self-Efficacy	C4Q4H = PSS Felt that you were on top of things	0.043766	0
Perceived Severity	C4H1G = Have you ever had anemia or other blood disease?	0.043304	0
Self-Efficacy	C4Q4D = PSS Confident about your ability to handle your personal problems	0.042646	0
Perceived Benefits	C1PA24EC = High blood pressure therapy - Exercise	-0.04119	0
Perceived Severity	C4H1E = Have you ever had a heart murmur?	-0.03592	0.0001
Perceived Susceptibility	Family history of circulation problems	-0.03554	0.0001
Perceived Susceptibility	Family history of high blood pressure	-0.02755	0.0023
Perceived Susceptibility	Family history of stroke	0.024753	0.0062
Perceived Severity	C4H1V = Have you ever had depression?	0.020929	0.0205
Self-Efficacy	C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it	0.020647	0.0223

Correlation analysis of Self Determination Theory:

The Significant correlations between SDT variables and total physical activity according to Self-Determination Theory are saved in the file ‘significant_correlations_SDT_PA_according_to_Self_Determination_Theory_zscores.csv' and their Z-scores are saved to ‘'SDT_variables_PA_zscores_according_to_Self_Determination_Theory.csv'.

```
                if pval < alpha:
                    significant_corr_list.append({
                        'SDT Category': category,
                        'Variable': var,
                        'Correlation': corr,
                        'P-value': pval
                    })

significant_corr = pd.DataFrame(significant_corr_list)

if not significant_corr.empty:
    # Sort by absolute correlation value
    significant_corr['Abs_Correlation'] = abs(significant_corr['Correlation'])
    significant_corr = significant_corr.sort_values('Abs_Correlation', ascending=False)
    significant_corr = significant_corr.drop('Abs_Correlation', axis=1)

    # Print results
    print("\nSignificant correlations between SDT variables and total physical activity according to Self-Determination Theory:")
    print(significant_corr.to_string(index=False))

    # Save results to CSV
    significant_corr.to_csv('significant_correlations_SDT_PA_according_to_Self_Determination_Theory_zscores.csv', index=False)
    print("\nFull results saved to 'significant_correlations_SDT_PA_according_to_Self_Determination_Theory_zscores.csv'")

    # Calculate average correlations for each SDT category
    avg_correlations = {}
    for category, vars in sdt_variables.items():
        if category != 'Behavior':
            category_vars = [v for v in vars if v in available_vars]
            if category_vars:
                category_corrs = correlation_matrix.loc[category_vars, 'total_physical_activity']
                avg_correlations[category] = category_corrs.mean()

    print("\nAverage correlations for each SDT category with total physical activity:")
    for category, avg_corr in avg_correlations.items():
        print(f"{category}: {avg_corr:.4f}")

    # Save z-scores to CSV
    z_scores_df = relevant_data_z.copy()
    z_scores_df.to_csv('SDT_variables_PA_zscores_according_to_Self_Determination_Theory.csv', index=False)
    print("Z-scores for SDT variables and physical activity saved to 'SDT_variables_PA_zscores_according_to_Self_Determination_Theory.csv'")
else:
    print("No significant correlations found.")
```

Significant correlations between SDT variables and the total physical activity according to self-Determination theory:

SDT Category	Variable	Correlation	P-value
Intrinsic Motivation	C4Q1AA = MASQ Felt like I had a lot of energy	0.239904	0
Competence	C4Q1WW = MASQ Felt like I had accomplished a lot	0.145632	0
Intrinsic Motivation	C4Q1MM = MASQ Felt really up or lively	0.137642	0
Intrinsic Motivation	C4Q1UU = MASQ Felt like I had a lot of interesting things to do	0.134419	0
Competence	C4Q1O = MASQ Was proud of myself	0.115801	0
Extrinsic Motivation	C4Q18G = GDLIFE To receive respect from others	-0.096957	0
Intrinsic Motivation	C4Q1V = MASQ Felt like I was having a lot of fun	0.095215	0
Relatedness	C4Q17A = LONELY There is no one I can turn to	-0.078402	0
Extrinsic Motivation	C4Q18F = GDLIFE To receive sympathy from others	-0.054665	0

Autonomy	C4Q4G = PSS Able to control irritations in your life	0.049153	0
Relatedness	C4Q17B = LONELY No one really knows me well	-0.046269	0
Autonomy	C4Q4H = PSS Felt that you were on top of things	0.043766	0
Relatedness	C4QLONEL = UCLA Loneliness Scale	-0.043435	0
Competence	C4Q4D = PSS Confident about your ability to handle your personal problems	0.042646	0
Relatedness	C4Q17G = LONELY There are people I can turn to	0.042269	0
Competence	C4Q18D = GDLIFE Have the ability to make a good effort at something and stick to it	0.020647	0.0223
Relatedness	C4Q17D = LONELY There are people who really understand me	0.019535	0.0306

Correlation analysis of social cognitive learning theory:

The Significant correlations between SCLT variables and total physical activity according to Social Cognitive Learning Theory are saved as 'significant_correlations_SCLT_PA_according_to_Social_Cognitive_Learning_Theory_zscores.csv' and their z-scores are saved as Z-scores for SCLT variables and physical activity saved to 'SCLT_variables_PA_zscores_according_to_Social_Cognitive_Learning_Theory.csv.

Significant correlations between SCLT variables and total physical activity according to Social Cognitive Learning Theory:

SCLT Category	**Variable**	**Correlation**	**P-value**
Outcome Expectations	C4Q16C = GDLIFE The conditions of my life are excellent.	0.155025	9.40E-67
Outcome Expectations	C4Q16E = GDLIFE So far I have gotten the important things I want in life	0.134265	2.23E-50
Outcome Expectations	C4Q16D = GDLIFE I am satisfied with my life.	0.11998	1.64E-40
Goals	C4Q18H = GDLIFE To give something back to society	-0.091911	2.13E-24
Outcome Expectations	C4Q16B = GDLIFE In most ways my life is close to my ideal.	0.087914	1.89E-22
Sociostructural Factors	B1PGENDER = Gender	-0.060113	2.77E-11
Self-efficacy	C4Q4G = PSS Able to control irritations in your life	0.049153	5.25E-08

Self-efficacy	C4Q4H = PSS Felt that you were on top of things	0.043766	1.26E-06
Self-efficacy	C4Q4D = PSS Confident about your ability to handle your personal problems	0.042646	2.34E-06
Sociostructural Factors	C4H85 = Marital status changed - Current status	0.038116	2.45E-05
Sociostructural Factors	C1PB1=Highest level of education completed	0.034251	1.50E-04

Interpretations of the above correlations between model variables and total physical activity:

Based on the significant correlations between the model variables and total physical activity, the below are the interpretations for each each theoretical model.

1. **Health Belief Model (HBM):**

- The HBM results suggest that perceived severity of health conditions is strongly associated with physical activity. People who have experienced various health issues (e.g., cholesterol problems, high blood pressure, diabetes) tend to be more physically active. This could indicate that experiencing health problems motivates individuals to engage in preventive behaviors like exercise.
- Cues to action, particularly exposure to secondhand smoke, are positively correlated with physical activity. This unexpected result might suggest that awareness of health risks in one's environment could motivate healthier behaviors in other areas.
- Perceived barriers, such as feeling that everything is an effort, are negatively associated with physical activity, which aligns with the model's predictions.
- Self-efficacy shows a positive correlation with physical activity, supporting the idea that confidence in one's abilities promotes healthier behaviors.
- Perceived susceptibility shows mixed results, with some family history factors positively correlated and others negatively correlated with physical activity.

2. **Theory of Planned Behavior (TPB):**

- Attitudes towards life satisfaction and well-being show the strongest positive correlations with physical activity. This suggests that individuals who have a more positive outlook on life are more likely to engage in physical activity.
- Subjective norms, particularly those related to receiving respect or sympathy from others, show negative correlations with physical activity. This unexpected result might indicate that individuals who are less concerned with others' opinions are more likely to engage in physical activity.
- Perceived behavioral control variables all show positive correlations with physical activity, supporting the theory's prediction that individuals who feel more in control of their actions are more likely to engage in planned behaviors like exercise.

3. **Self-Determination Theory (SDT):**

- Intrinsic motivation shows the strongest positive correlations with physical activity. Feeling energetic, lively, and having interesting things to do are all associated with higher levels of physical activity. This supports the SDT's emphasis on intrinsic motivation as a key factor in sustained behavior.
- Competence, another key component of SDT, also shows positive correlations with physical activity. Feeling accomplished and proud of oneself is associated with higher levels of physical activity.
- Autonomy, represented by feelings of control and being on top of things, shows positive correlations with physical activity, aligning with SDT's predictions.
- Relatedness shows mixed results, with some loneliness indicators negatively correlated and others positively correlated with physical activity.
- Extrinsic motivation shows negative correlations with physical activity, which is consistent with SDT's proposition that external motivators are less effective for sustaining behaviors than intrinsic ones.

4. **Social Cognitive Learning Theory (SCLT):**

- Outcome expectations, particularly those related to life satisfaction and achieving important life goals, show the strongest positive correlations with physical activity. This supports SCLT's emphasis on the role of expected outcomes in shaping behavior.
- Self-efficacy shows positive correlations with physical activity, aligning with SCLT's central tenet that belief in one's abilities influences behavior.
- Sociostructural factors like gender, marital status, and education level show significant correlations with physical activity, highlighting the importance of environmental and personal factors in shaping behavior.
- Goals, specifically the goal to give back to society, shows a negative correlation with physical activity. This unexpected result might warrant further investigation.
- Overall, these results provide support for aspects of each theory while also highlighting areas that may require further research or refinement in the context of physical activity behavior.

Summary table that compares the strengths of correlations across all four models (HBM, TPB, SDT, and SCLT):

Health Belief Model Summary:

HBM Category	mean	min	max
Cues to Action	0.090446	0.053541	0.127351
Perceived Barriers	-0.070569	-0.094594	-0.046543
Perceived Benefits	-0.041192	-0.041192	-0.041192
Perceived Severity	0.073509	-0.035923	0.130041
Perceived Susceptibility	0.023693	-0.035540	0.066705
Self-Efficacy	0.039053	0.020647	0.049153

Theory of Planned Behavior Summary:

TPB Category	mean	min	max
Attitudes	0.114617	0.069221	0.155025
Perceived Behavioral Control	0.039053	0.020647	0.049153
Subjective Norms	-0.075811	-0.096957	-0.054665

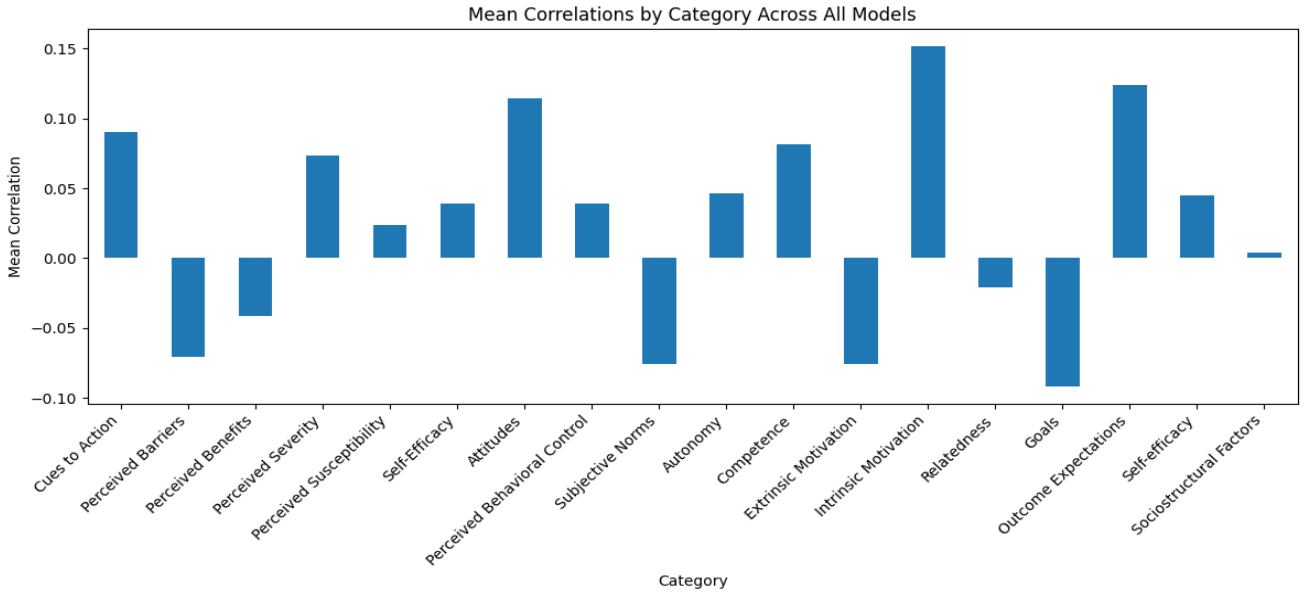
Self-Determination Theory Summary:

SDT Category	mean	min	max
Autonomy	0.046460	0.043766	0.049153
Competence	0.081182	0.020647	0.145632
Extrinsic Motivation	-0.075811	-0.096957	-0.054665
Intrinsic Motivation	0.151795	0.095215	0.239904
Relatedness	-0.021260	-0.078402	0.042269

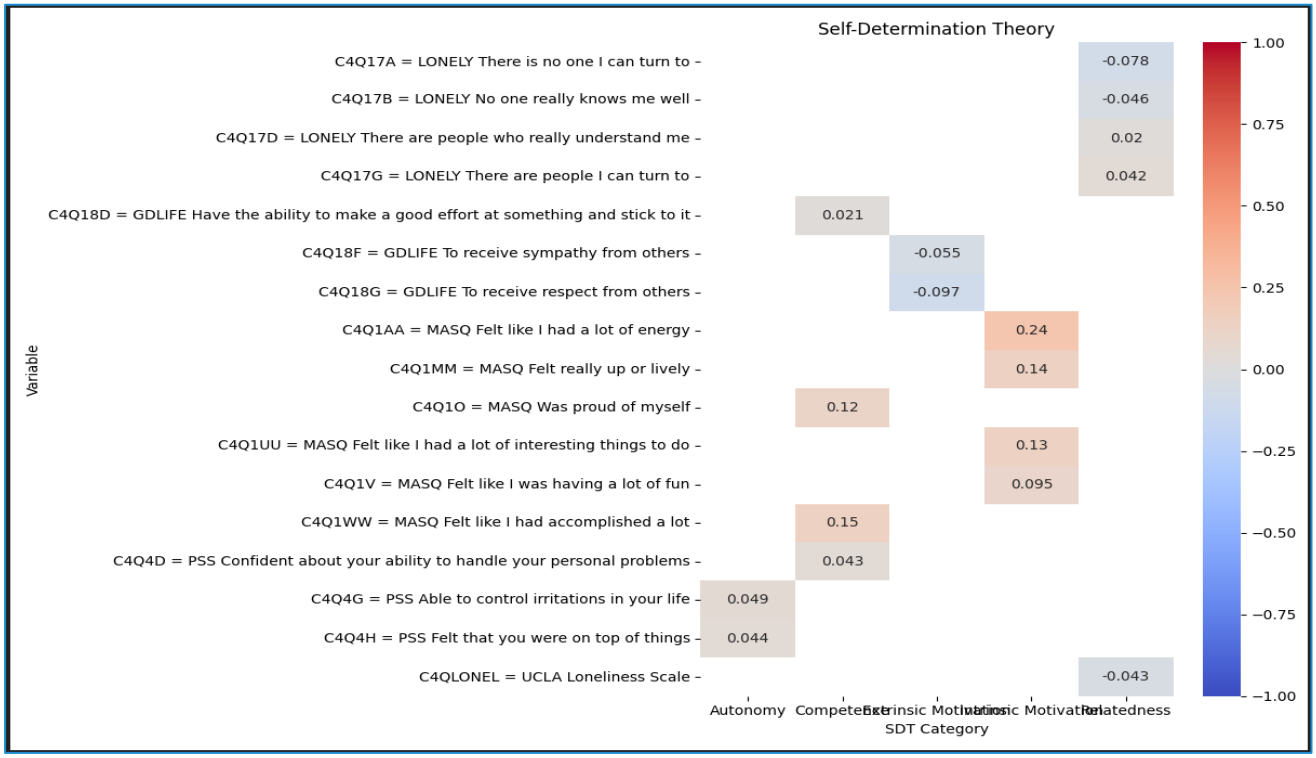
Social Cognitive Learning Theory Summary:

SCLT Category	mean	min	max
Goals	-0.091911	-0.091911	-0.091911
Outcome Expectations	0.124296	0.087914	0.155025
Self-efficacy	0.045188	0.042646	0.049153
Sociostructural Factors	0.004085	-0.060113	0.038116

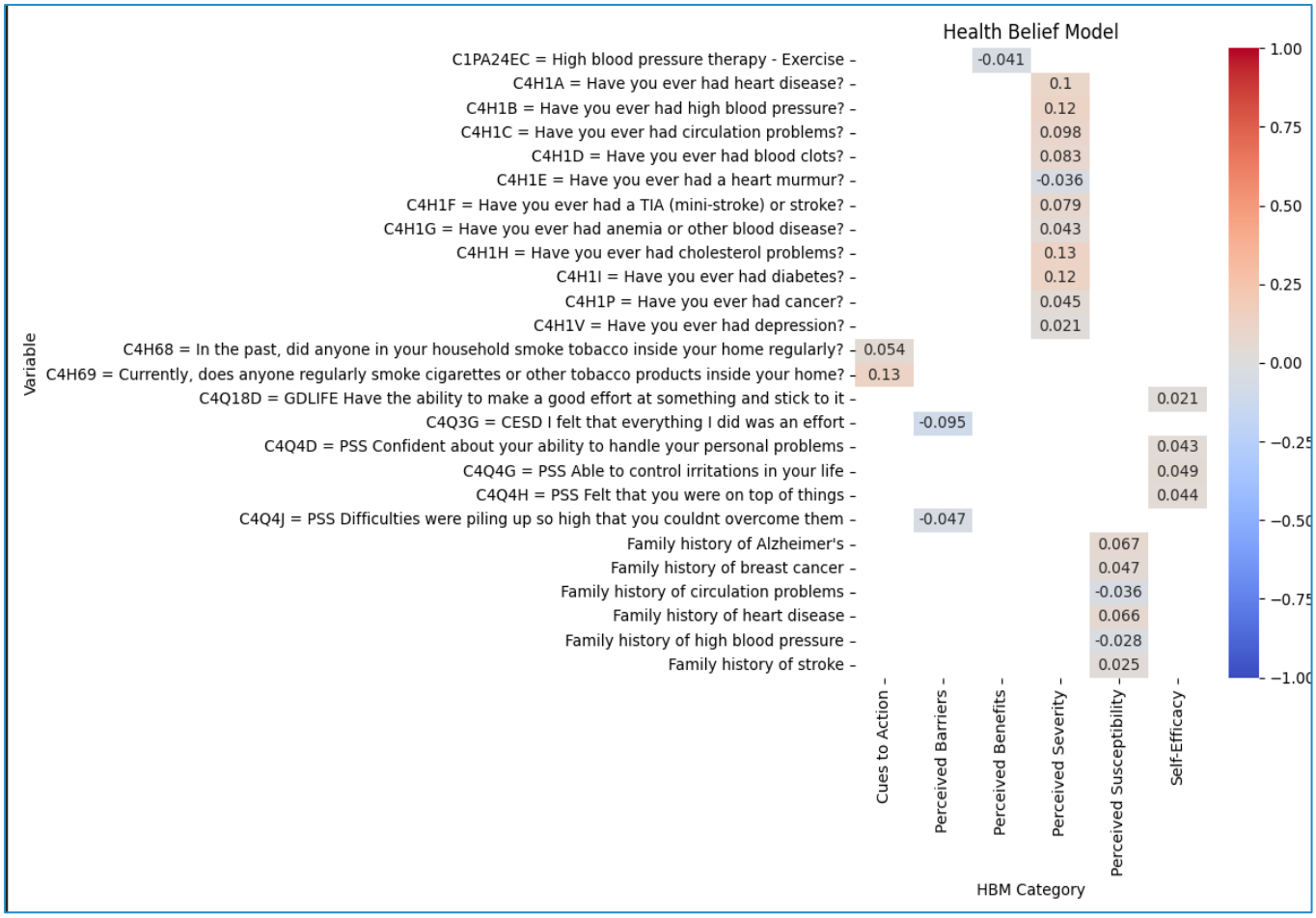
Mean correlation of variables across all the models:



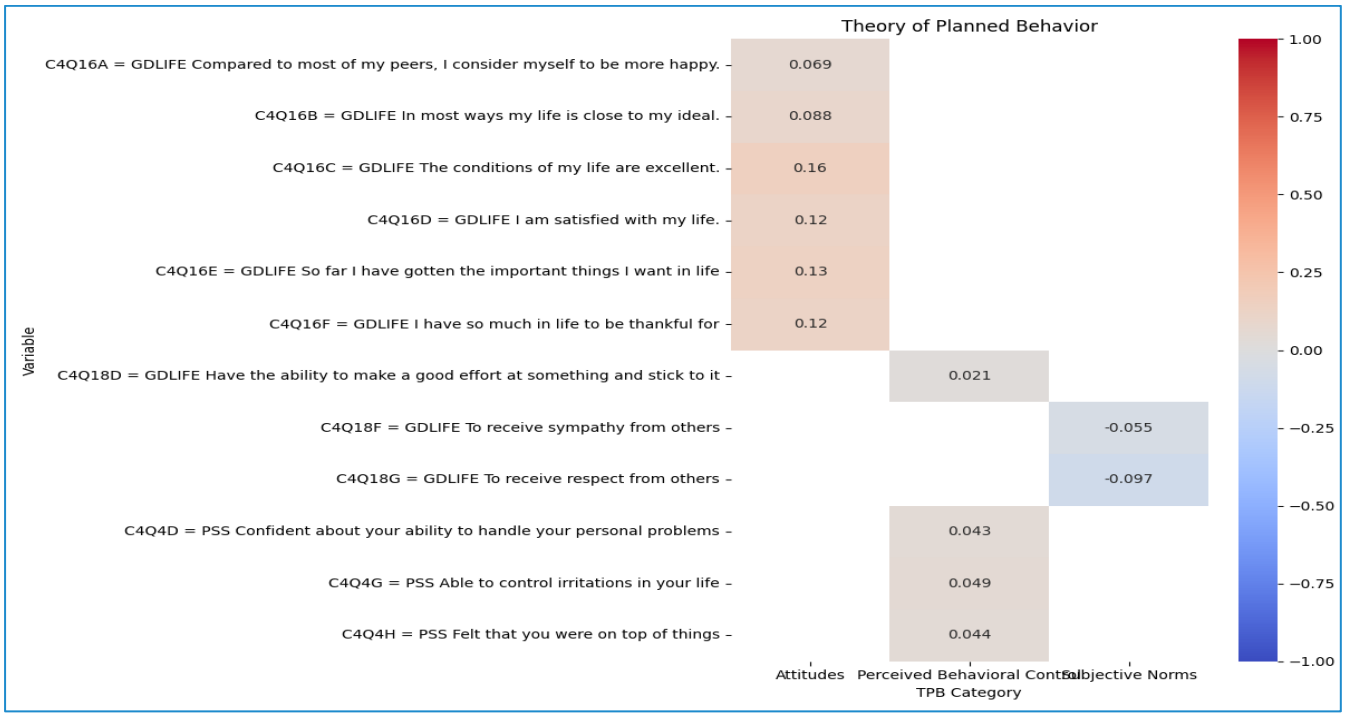
Self determination theory variables correlation heatmap:



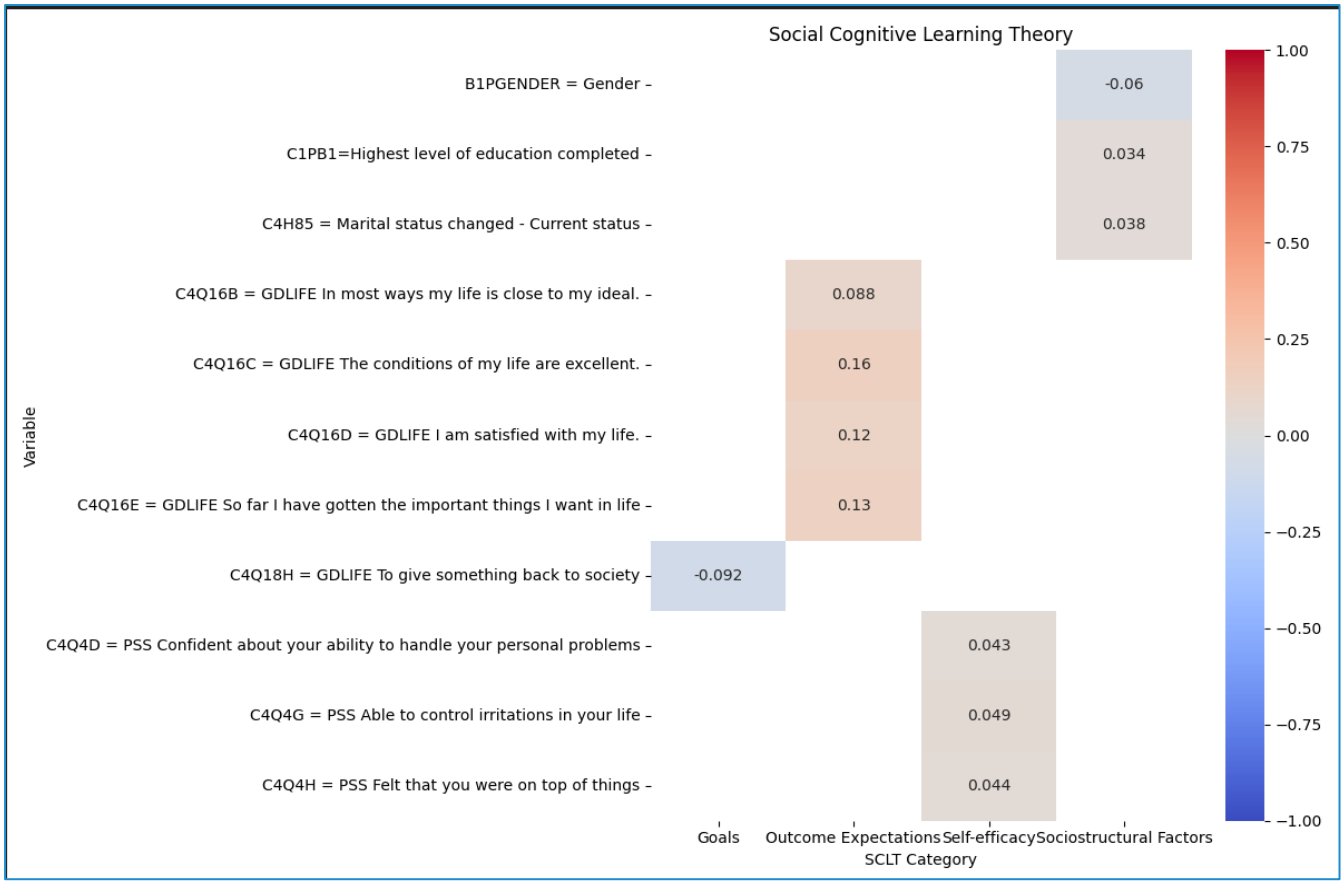
Health belief model variables correlation heatmap:



Theory of planned behavior variables correlation heatmap:



Social cognitive learning theory variables correlation heatmap:


Social Cognitive Learning Theory
B1PGENDER = Gender
C1PB1=Highest level of education completed
C4H85 = Marital status changed - Current status
C4Q16B = GDLIFE In most ways my life is close to my ideal.
C4Q16C = GDLIFE The conditions of my life are excellent.
C4Q16D = GDLIFE I am satisfied with my life.
C4Q16E = GDLIFE So far I have gotten the important things I want in life
C4Q18H = GDLIFE To give something back to society
C4Q4D = PSS Confident about your ability to handle your personal problems
C4Q4G = PSS Able to control irritations in your life
C4Q4H = PSS Felt that you were on top of things
Variable
-0.06
0.034
0.038
0.088
0.16
0.12
0.13
-0.092
0.043
0.049
0.044
Goals
Outcome Expectations
Self-efficacy
Sociostructural Factors
SCLT Category
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00