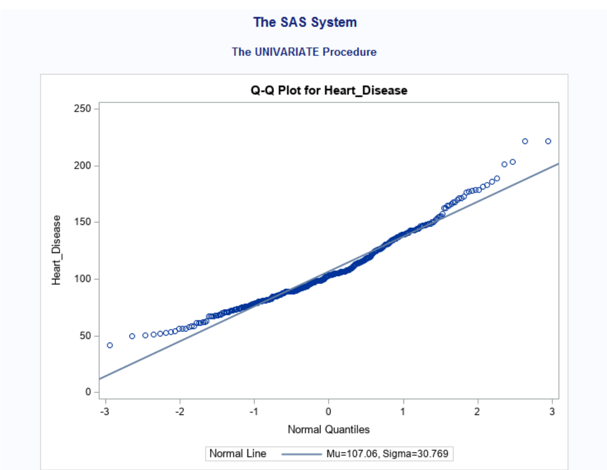
**Oscar H. Mata.**

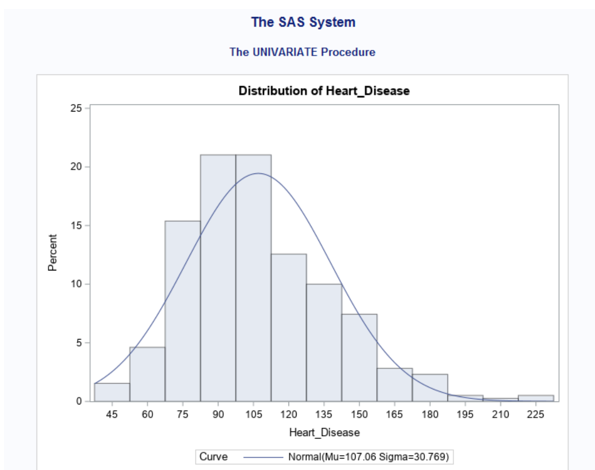
I this project analyzed the Alzheimer Disease dataset. Yet I focused on the variables Heart_Disease and physical_inactivity in the states of Michigan, Wisconsin, Minnesota, North Dakota, Montana, and Washington.

## TEST OF NORMALITY

First, I checked whether **Heart_Disease** was normally distributed.

```
proc univariate data=Alzheimer_filtered;
   var Heart_Disease;
   histogram / normal;
   qqplot / normal(mu=est sigma=est);
run;
```



The QQ plot reveals a data that does not follow a straight line. On the other hand, the histogram shows some skewness to the right. These outcomes indicate that Heart_Disease might not be normally distributed. However, this conclusion needs to be supported by further tests.
Since the QQ plot and histogram are not a 100% reliable tool to check normality of the variables, I used Shapiro Wilks test to confirm my assumptions.

Hypothesis:

H0: The data are normally distributed.

H1: The data are not normally distributed.

We reject H0 if $p<0.05$.

```
proc univariate data=Alzheimer_filtered normal;
   var Heart_Disease;
run;
```

**The SAS System**

The UNIVARIATE Procedure
Variable: Heart_Disease

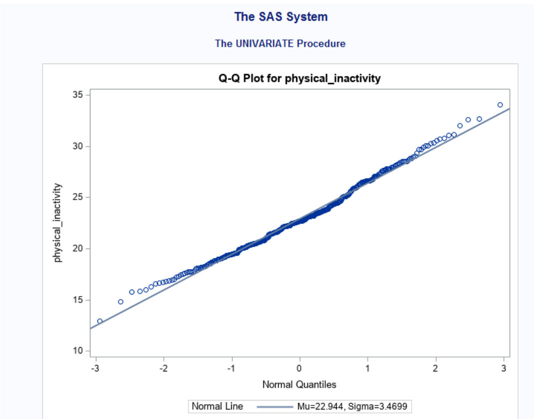
| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.964855 | Pr < W | <0.0001 |
| Kolmogorov-Smirnov | D | 0.09598 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 0.672434 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 3.720663 | Pr > A-Sq | <0.0050 |

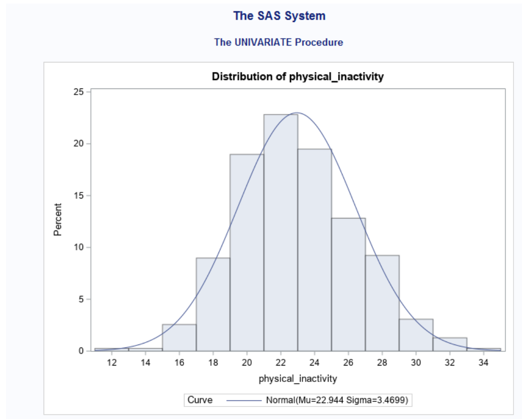
I calculated p-value=0.0001 for the variable Heart_Disease, which is less than alpha=0.05. Thus, I can reject H0. I can conclude that Heart_Disease is not normally distributed.

Secondly, I examined whether the variable **physical_inactivity** follows a normal distribution.

```
proc univariate data=Alzheimer_filtered;
   var physical_inactivity;
```

```
    histogram / normal;

    qqplot / normal(mu=est sigma=est);

run;
```




The QQ plot reveals that the data points are largely aligned with the straight line, with deviations primarily at the lower end. Additionally, the histogram displays a bell-shaped curve. Collectively, these graphical representations suggest that the variable physical_inactivity may approximate a normal distribution.

As before, next I used the Shapiro Wilks test to confirm my assumptions.

```
proc univariate data=Alzheimer_filtered normal;

    var physical_inactivity;

run;
```

The SAS System

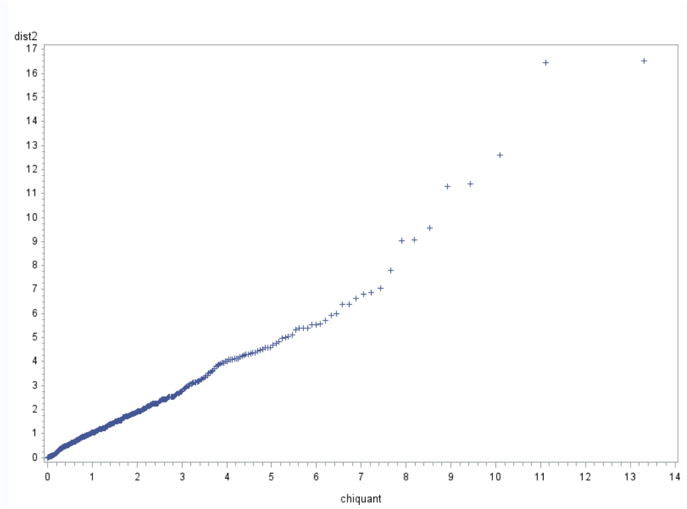
The UNIVARIATE Procedure
Variable: physical_inactivity

| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.991223 | Pr < W | 0.0205 |
| Kolmogorov-Smirnov | D | 0.057069 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 0.20887 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 1.162052 | Pr > A-Sq | <0.0050 |

I calculated p-value=0.0205 for the variable physical_inactivity, which is less than alpha=0.05. Thus, I can reject H0. I can conclude that physical_inactivity is not normally distributed.

After checking for individual normality, I checked for **bivariate normality**.

```
proc princomp std out=pcresult;

var Heart_Disease physical_inactivity;

run;

data mahal;set pcresult;dist2=uss(of prin1-prin2);

run;

proc sort;by dist2;run;data Alzheimer_filtered;set mahal;

prb=(_n_ -.5)/390;chiquant=cinv(prb,2);

run;

proc gplot;plot dist2*chiquant;

run;
```

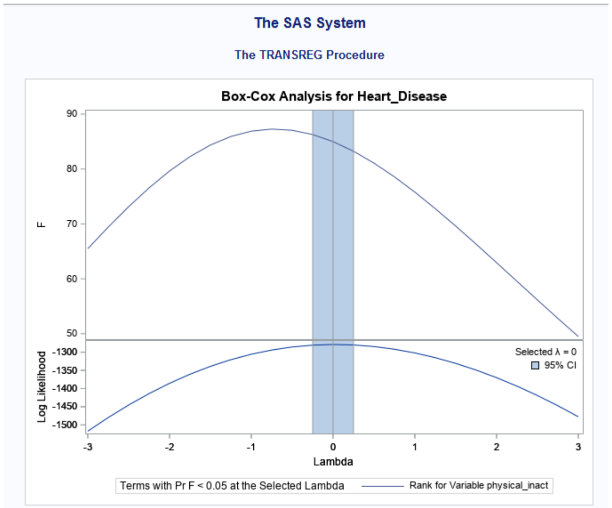


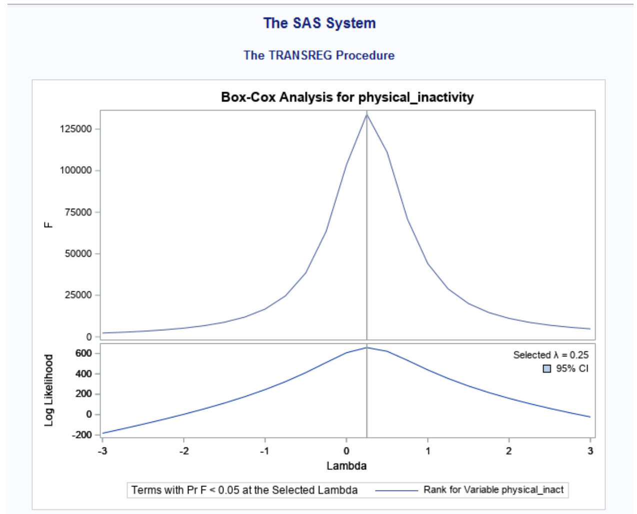

The graphic displays the data in a predominantly straight line, suggesting an apparent normality in the data. It is also noteworthy to observe the presence of outliers at the upper right extreme.

## DATA TRANSFORMATION

Given that the variables are not normally distributed, a **Box-Cox** transformation is deemed appropriate. By utilizing the **proc transreg** code, I was able to determine the power or powers (lambda values) that rendered the two variables approximately normal.

```
proc transreg;
model boxcox(Heart_Disease) = identity(q);
run;

proc transreg;
model boxcox(physical_inactivity) = identity(q);
run;
```




For the variable **Heart_Disease**, the optimal lambda is 0. Thus, a log tranformation is approapriate. Furthermore, for **physical_inactivity** the optimal lambda is 0.25. In this case the approapriate equation to transform the data is x**0.25−1/0.25

```
data Alzheimer_filtered;
   set Alzheimer_filtered;
      Heart_Disease = log(Heart_Disease);
run;


data Alzheimer_filtered;
   set Alzheimer_filtered;
    physical_inactivity = (physical_inactivity**(0.25)-1)/(0.25);
 run;
```

**The SAS System**

The UNIVARIATE Procedure
Variable: Heart_Disease

| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.996449 | Pr < W | 0.5428 |
| Kolmogorov-Smirnov | D | 0.039912 | Pr > D | 0.1335 |
| Cramer-von Mises | W-Sq | 0.069867 | Pr > W-Sq | >0.2500 |
| Anderson-Darling | A-Sq | 0.427939 | Pr > A-Sq | >0.2500 |

**The SAS System**

The UNIVARIATE Procedure
Variable: physical_inactivity

| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.997275 | Pr < W | 0.7695 |
| Kolmogorov-Smirnov | D | 0.036777 | Pr > D | >0.1500 |
| Cramer-von Mises | W-Sq | 0.069926 | Pr > W-Sq | >0.2500 |
| Anderson-Darling | A-Sq | 0.378017 | Pr > A-Sq | >0.2500 |

Following the transformation, I assessed the normality of the variables using the Shapiro-Wilk test. I calculated p-value=0.5428 for the variable Heart_Disease, which is greater than alpha=0.05. Thus, I cannot reject H0. I can conclude that Heart_Disease is normally distributed after the transformation. Additionally, I calculated p-value=0.7695 for the variable physical_inactivity, which is greater than alpha=0.05. Thus, I cannot reject H0. I can conclude that physical_inactivity is normally distributed after the transformation.

## MANOVA TEST

Since the variables of interest are normally distributed after the transformation, I will compute a MANOVA test to check if there are differences between states regarding heart disease and physical inactivity.

Hypotheses:

H0 : $\mu1 = \mu2 = \mu3 = \mu4$

H1 : at least one is different

We reject H0 if $p<0.05$

**Proc glm data=Alzheimer_filtered ;**

**class State;**

**model physical_inactivity Heart_Disease=State ;**

```
means State/tukey;

MANOVA h=state/printh ;

Run;
```

The SAS System

The GLM Procedure

Dependent Variable: Heart_Disease

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 10.79223537 | 2.15844707 | 39.97 | <.0001 |
| Error | 384 | 20.73816720 | 0.05400564 | | |
| Corrected Total | 389 | 31.53040257 | | | |

| R-Square | Coeff Var | Root MSE | Heart_Disease Mean |
|---|---|---|---|
| 0.342280 | 5.015798 | 0.232391 | 4.633184 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| State | 5 | 10.79223537 | 2.15844707 | 39.97 | <.0001 |

The SAS System

The GLM Procedure

Dependent Variable: physical_inactivity

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 19.20381651 | 3.84076330 | 63.15 | <.0001 |
| Error | 384 | 23.35548249 | 0.06082157 | | |
| Corrected Total | 389 | 42.55929900 | | | |

| R-Square | Coeff Var | Root MSE | physical_inactivity Mean |
|---|---|---|---|
| 0.451225 | 5.207674 | 0.246620 | 4.735709 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| State | 5 | 19.20381651 | 3.84076330 | 63.15 | <.0001 |

I calculated p-value=0.0001 which is less than $\alpha = 0.05$, so I can reject H0. I conclude that there is sufficient evidence to establish there is a difference in the means of Heart_Disease and physical inactivity in at least one of the states.

The next question is, which states are different?

PCA Scores for Alzheimer

The GLM Procedure

Tukey's Studentized Range (HSD) Test for Heart_Disease

| Comparisons significant at the 0.05 level are indicated by ***. | | | | |
|---|---|---|---|---|
| State Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
| MI - ND | 0.16932 | 0.05228 | 0.28635 | *** |
| MI - WI | 0.24886 | 0.14166 | 0.35605 | *** |
| MI - WA | 0.27311 | 0.14390 | 0.40233 | *** |
| MI - MT | 0.32421 | 0.20911 | 0.43932 | *** |
| MI - MN | 0.48808 | 0.38596 | 0.59021 | *** |
| ND - MI | -0.16932 | -0.28635 | -0.05228 | *** |
| ND - WI | 0.07954 | -0.04093 | 0.20001 | |
| ND - WA | 0.10380 | -0.03663 | 0.24422 | |
| ND - MT | 0.15490 | 0.02734 | 0.28245 | *** |
| ND - MN | 0.31877 | 0.20279 | 0.43475 | *** |
| WI - MI | -0.24886 | -0.35605 | -0.14166 | *** |
| WI - ND | -0.07954 | -0.20001 | 0.04093 | |
| WI - WA | 0.02426 | -0.10808 | 0.15660 | |
| WI - MT | 0.07536 | -0.04324 | 0.19395 | |
| WI - MN | 0.23923 | 0.13318 | 0.34527 | *** |
| WA - MI | -0.27311 | -0.40233 | -0.14390 | *** |
| WA - ND | -0.10380 | -0.24422 | 0.03663 | |
| WA - WI | -0.02426 | -0.15660 | 0.10808 | |
| WA - MT | 0.05110 | -0.08772 | 0.18992 | |
| WA - MN | 0.21497 | 0.08670 | 0.34324 | *** |
| MT - MI | -0.32421 | -0.43932 | -0.20911 | *** |
| MT - ND | -0.15490 | -0.28245 | -0.02734 | *** |
| MT - WI | -0.07536 | -0.19395 | 0.04324 | |
| MT - WA | -0.05110 | -0.18992 | 0.08772 | |
| MT - MN | 0.16387 | 0.04984 | 0.27790 | *** |
| MN - MI | -0.48808 | -0.59021 | -0.38596 | *** |
| MN - ND | -0.31877 | -0.43475 | -0.20279 | *** |
| MN - WI | -0.23923 | -0.34527 | -0.13318 | *** |
| MN - WA | -0.21497 | -0.34324 | -0.08670 | *** |
| MN - MT | -0.16387 | -0.27790 | -0.04984 | *** |

## Heart Disease differences

Michigan (MI) shows a statistically significant higher mean compared to North Dakota (ND), Wisconsin (WI), Washington (WA), Montana (MT), and Minnesota (MN). The differences and their respective confidence intervals all exclude zero and are marked with "***", indicating significant differences.

North Dakota's (ND) mean is different than Michigan (MI) and Minnesota (MN)

Wisconsin (WI) is significantly different than MI and higher than MN.

Washington (WA) is significantly lower than MI and higher than MN, with no significant differences when compared to ND, WI, and MT.

Montana (MT) is significantly lower than MI and higher than MN.

Minnesota (MN) has a significantly lower mean compared to all other states mentioned (MI, ND, WI, WA, MT).

**PCA Scores for Alzheimer**

**The GLM Procedure**

**Tukey's Studentized Range (HSD) Test for physical_inactivity**

Comparisons significant at the 0.05 level are indicated by ***.

| State Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| ND - MT | 0.25797 | 0.12260 | 0.39333 | *** |
| ND - MI | 0.38259 | 0.25839 | 0.50679 | *** |
| ND - WI | 0.52233 | 0.39448 | 0.65017 | *** |
| ND - MN | 0.63810 | 0.51502 | 0.76118 | *** |
| ND - WA | 0.71553 | 0.56650 | 0.86455 | *** |
| MT - ND | -0.25797 | -0.39333 | -0.12260 | *** |
| MT - MI | 0.12462 | 0.00247 | 0.24678 | *** |
| MT - WI | 0.26436 | 0.13850 | 0.39022 | *** |
| MT - MN | 0.38013 | 0.25912 | 0.50115 | *** |
| MT - WA | 0.45756 | 0.31024 | 0.60488 | *** |
| MI - ND | -0.38259 | -0.50679 | -0.25839 | *** |
| MI - MT | -0.12462 | -0.24678 | -0.00247 | *** |
| MI - WI | 0.13974 | 0.02598 | 0.25350 | *** |
| MI - MN | 0.25551 | 0.14713 | 0.36389 | *** |
| MI - WA | 0.33294 | 0.19580 | 0.47007 | *** |
| WI - ND | -0.52233 | -0.65017 | -0.39448 | *** |
| WI - MT | -0.26436 | -0.39022 | -0.13850 | *** |
| WI - MI | -0.13974 | -0.25350 | -0.02598 | *** |
| WI - MN | 0.11577 | 0.00324 | 0.22831 | *** |
| WI - WA | 0.19320 | 0.05276 | 0.33364 | *** |
| MN - ND | -0.63810 | -0.76118 | -0.51502 | *** |
| MN - MT | -0.38013 | -0.50115 | -0.25912 | *** |
| MN - MI | -0.25551 | -0.36389 | -0.14713 | *** |
| MN - WI | -0.11577 | -0.22831 | -0.00324 | *** |
| MN - WA | 0.07743 | -0.05869 | 0.21355 | |
| WA - ND | -0.71553 | -0.86455 | -0.56650 | *** |
| WA - MT | -0.45756 | -0.60488 | -0.31024 | *** |
| WA - MI | -0.33294 | -0.47007 | -0.19580 | *** |
| WA - WI | -0.19320 | -0.33364 | -0.05276 | *** |
| WA - MN | -0.07743 | -0.21355 | 0.05869 | |

## Physical Inactivity Differences

ND shows a significantly higher mean compared to MT, MI, WI, MN, and WA. The positive differences range from 0.25797 (ND vs. MT) to 0.71553 (ND vs. WA).

MT has a higher mean compared to MI, WI, MN, and WA, with differences ranging from 0.12462 (MT vs. MI) to 0.45756 (MT vs. WA).

MI has a lower mean compared to ND and a slightly higher mean compared to MT, WI, MN, and WA. The smallest significant difference is against MT (0.12462) and the largest against WA (0.33294).

WI shows lower means compared to ND, MT, and MI, and a higher mean compared to MN and WA.

MN has significantly lower means compared to ND, MT, MI, and WI, with differences indicating that MN is at the lower end for physical inactivity being compared.

WA consistently shows lower means compared to ND, MT, MI, WI, and is insignificantly higher than MN, suggesting WA generally ranks lowest for physical inactivity among these states.

The following **MANOVA (Multivariate Analysis of Variance)** table considers the combined effect of Heart_Disease and physical_inanctivity. MANOVA is specifically designed to test differences across states on more than one dependent variable simultaneously, unlike univariate ANOVA which tests for differences on a single dependent variable.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall State Effect
H = Type III SSCP Matrix for State
E = Error SSCP Matrix

S=2 M=1 N=190.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.36971201 | 49.38 | 10 | 766 | <.0001 |
| Pillai's Trace | 0.77741308 | 48.84 | 10 | 768 | <.0001 |
| Hotelling-Lawley Trace | 1.30686290 | 49.97 | 10 | 571.76 | <.0001 |
| Roy's Greatest Root | 0.82380618 | 63.27 | 5 | 384 | <.0001 |

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Based on the one-way MANOVA, we can reject the null hypothesis that there is not a difference between means for Heart_Disease and physical_inactivity across the chosen states. I calculated the p-values=.0001, which is less than alpha=0.05.

## PRINCIPAL FACTOR ANALYSIS (FA)

Next, I computed a FA for data reduction and to identify the underlying factors or hidden variables that explain the correlations among our variables. I searched for variables with high loadings, where typically absolute values greater than 0.5 are considered significant.

```
proc factor data=Alzheimer_filtered method=principal priors=smc n=2 rotate=promax score out=FA_Alzheimer_filtered ;

   var physical_inactivity Heart_Disease sixtyfiveandup Smoking_Rate Diabetes Cancer Mercury_TPY Lead_TPY Glyphosates NATA_Cancer_11;

   where State = 'MN';

run;
```

**PCA Scores for Alzheimer**

**The FACTOR Procedure**
**Initial Factor Method: Principal Factors**

| Eigenvalues of the Reduced Correlation Matrix: Total = 4.94398983 Average = 0.49439898 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 3.31140317 | 2.08639853 | 0.6698 | 0.6698 |
| 2 | 1.22500464 | 0.51692743 | 0.2478 | 0.9176 |
| 3 | 0.70807721 | 0.33085773 | 0.1432 | 1.0608 |
| 4 | 0.37721949 | 0.36408784 | 0.0763 | 1.1371 |
| 5 | 0.01313165 | 0.04802185 | 0.0027 | 1.1397 |
| 6 | -.03489020 | 0.05252051 | -0.0071 | 1.1327 |
| 7 | -.08741071 | 0.05348177 | -0.0177 | 1.1150 |
| 8 | -.14089248 | 0.05498552 | -0.0285 | 1.0865 |
| 9 | -.19587800 | 0.03589694 | -0.0396 | 1.0469 |
| 10 | -.23177494 | | -0.0469 | 1.0000 |

| Rotated Factor Pattern | | |
|---|---|---|
| | Factor1 | Factor2 |
| physical_inactivity | -0.64614 | 0.40497 |
| Heart_Disease | -0.54488 | 0.24920 |
| sixtyfiveandup | -0.68518 | 0.17996 |
| Smoking_Rate | -0.37048 | 0.59512 |
| Diabetes | -0.03892 | 0.54828 |
| Cancer | 0.07337 | 0.49360 |
| Mercury_TPY | 0.55134 | 0.34024 |
| Lead_TPY | 0.79187 | 0.10296 |
| Glyphosates | -0.44910 | -0.21605 |
| NATA_Cancer_11 | 0.83137 | -0.23350 |

The Eigenvalues of the Reduced Correlation Matrix Table reveal that the first factor is very strong; its eigenvalue of approximately 3.3 accounts for a significant proportion of the variance—67%. The second factor, with an eigenvalue of approximately 1.22, is much smaller, contributing less to the variance explanation. Factor 2 adds 24%, totaling 92% of the variance explained. The results suggest that Factor 1 and Factor 2 combined summarize much of the information contained in the ten measurements.

The Rotated Factor Pattern Table of the Principal Factor reveals that Mercury_TPY (0.77), Lead_TPY (0.79), and NATA_Cancer_11(0.83) have a strong relationship with Factor 1. Conversely smoking_rate (0.595) and Diabetes (0.548) are linked to Factor 2. In summary, FA's results indicate that Factor 1 is associated with **Toxicities and their risk of getting cancer**, whereas Factor 2 relates to lifestyle **and chronic disease**.

## PRINCIPAL COMPONENT ANALYSIS (PCA)

The next step involves conducting a Principal Component Analysis (PCA) to reduce the number of variables by creating new ones that capture essential information from the dataset. I have utilized all the variables selected by our team, focusing specifically on the state of Minnesota.

```
proc factor data=Alzheimer_filtered method=prin priors=one n=3 rotate=varimax score out=PCA_Alzheimer_filtered ;
   var physical_inactivity Heart_Disease sixtyfiveandup Smoking_Rate Diabetes Cancer Mercury_TPY Lead_TPY Glyphosates NATA_Cancer_11;
   where State = 'MN';
run;
```

The SAS System

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

| Eigenvalues of the Correlation Matrix: Total = 10 Average = 1 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 3.72098745 | 1.89184945 | 0.3721 | 0.3721 |
| 2 | 1.82913800 | 0.65936586 | 0.1829 | 0.5550 |
| 3 | 1.16977213 | 0.25900890 | 0.1170 | 0.6720 |
| 4 | 0.91076323 | 0.22769780 | 0.0911 | 0.7631 |
| 5 | 0.68306543 | 0.12686107 | 0.0683 | 0.8314 |
| 6 | 0.55620436 | 0.18647154 | 0.0556 | 0.8870 |
| 7 | 0.36973283 | 0.03092597 | 0.0370 | 0.9240 |
| 8 | 0.33880685 | 0.05951434 | 0.0339 | 0.9578 |
| 9 | 0.27929251 | 0.13705531 | 0.0279 | 0.9858 |
| 10 | 0.14223720 | | 0.0142 | 1.0000 |

| Rotated Factor Pattern | | | |
|---|---|---|---|
| | Factor1 | Factor2 | Factor3 |
| physical_inactivity | 0.58522 | -0.39554 | 0.40367 |
| Heart_Disease | 0.78760 | -0.00376 | 0.05737 |
| sixtyfiveandup | 0.84689 | -0.10931 | -0.02992 |
| Smoking_Rate | 0.48181 | -0.06224 | 0.63385 |
| Diabetes | -0.05599 | -0.15174 | 0.80942 |
| Cancer | 0.04255 | 0.22238 | 0.68950 |
| Mercury_TPY | -0.01885 | 0.91509 | 0.08712 |
| Lead_TPY | -0.38630 | 0.80526 | -0.06225 |
| Glyphosates | 0.41085 | -0.32885 | -0.36031 |
| NATA_Cancer_11 | -0.83249 | 0.31962 | -0.09856 |

The Eigenvalues of the Correlation Matrix Table reveal that the first factor is strong; its eigenvalue of approximately 3.7 accounts for a 37% proportion of the variance. The second factor, with an eigenvalue of approximately 1.8, is much smaller, contributing less to the variance explanation. Factor 2 adds 18%, totaling 55% of the variance explained. Moreover, Factor 3 has an approximate eigenvalue equal to 1.2, thus contributing 12% to the variance explanation. The results suggest that Factor 1, Factor 2, and Factor 3 combined summarize much of the information contained in the ten measurements (67% of the variance explained).

Next, I conducted an orthogonal rotation of the 3 factors in order to make the factor structure easier to understand. I searched for variables with high loadings, where typically absolute values greater than 0.5 are considered significant. The Rotated Factor Pattern Table of the Principal Factor reveals that physical_inactivity (0.58), Heart_Disease (0.79), age 65 and up (0.85), and NATA_Cancer (-0.83) have a strong relationship with Factor 1.

Based on the previous results Factor 1 seems to grab elements related to **General Health Conditions** impacted by lifestyle and age. It includes aspects of diseases and conditions prevalent in elderly populations and inactive lifestyle in Minnesota. On the other hand, the strong negative loading in NATA_Cancer suggests an inverse relationship with the general health conditions captured by Factor 1. The negative association infers that in counties with worse general health conditions (more heart disease and physical inactivity), the values of NATA_Cancer_11 are low. This negative relationship needs further analysis.

On the other hand, Factor 2 appears to relate to **Toxicities**, with strong loadings on mercury (0.91) and lead toxicity (0.80) variables. Factor 3 is primarily connected with health conditions directly impacted by specific behavioral factors like smoking (0.63), and **Chronic Conditions** like diabetes (0.80) and cancer (0.70), which might also correlate with lifestyle choices.

```
proc univariate data=PCA_Alzheimer_filtered;
  var Factor1;
      run;
proc univariate data=PCA_Alzheimer_filtered;
  var Factor2;
      run;
proc univariate data=PCA_Alzheimer_filtered;
  var Factor3;
      run;
```

The SAS System

The UNIVARIATE Procedure
Variable: Factor1

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| -2.37390 | 77 | 1.31629 | 9 |
| -2.16609 | 86 | 1.51825 | 17 |
| -2.16519 | 83 | 1.82034 | 87 |
| -2.02955 | 84 | 1.82738 | 20 |
| -1.97338 | 81 | 1.85286 | 68 |

The SAS System

The UNIVARIATE Procedure
Variable: Factor2

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| -1.195070 | 75 | 1.93283 | 83 |
| -1.090751 | 52 | 2.06148 | 85 |
| -0.915780 | 82 | 2.11638 | 86 |
| -0.891395 | 73 | 2.17600 | 53 |
| -0.846694 | 71 | 7.19279 | 67 |

The SAS System

The UNIVARIATE Procedure
Variable: Factor3

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| -1.64206 | 85 | 1.70591 | 12 |
| -1.46838 | 36 | 1.75464 | 18 |
| -1.46057 | 62 | 2.06288 | 38 |
| -1.44228 | 81 | 2.20085 | 16 |
| -1.41515 | 17 | 3.93288 | 45 |

After examining the three relevant factors of the PCA, I looked for unusual counties in Minnesota. For instance, Table 1 points out that observation 83, which is Ramsey County is associated with low levels of heart disease and physical inactivity (Factor 1 General Health Conditions). One of the reasons could be that its population is mainly young. Yet, considering that NATA_Cancer_11 has a negative loading on this factor, counties with extreme negative scores (-2.166) on Factor 1 might also have higher NATA_Cancer_11 values, suggesting potentially higher estimated risks of cancer from air toxins. Moreover, Ramsey County also appears on Factor 2 extreme observations associated with Toxicities. The county scored high (1.932) on levels of mercury and lead toxicity. On the other hand, Mahnomen County which is extreme observation 45 appears on the Factor 3 extreme observations table with a score of 3.93 indicating that it is linked to chronic conditions like diabetes and cancer, and in a lower-level behavioral factors like smoking.