Age Gauge:

A linear regression model to predict the age given certain health characteristics

Samantha Doyle and Brianna Noel

December 20, 2024

Abstract

There are many factors that impact one's health. Things like smoking, drinking alcohol, family medical history, gender and even education level can impact someone's health. Doctors typically have access to a patient's age, information about their past health issues, and family history. Since doctor's usually have access to a patient's age, it might seem redundant to predict age. However, predicting biological age can help determine if the patient's biological age matches their chronological age. The goal of this statistical analysis is to predict a patient's biological age, which can be used to see if they have health factors that are skewing their biological age away from their chronological age.

Introduction 5
Data Overview & Cleaning 5
Variable Selection 5
Exploratory Data Analysis 8
CORRELATION COEFFICIENTS FOR ALL NUMERIC VARIABLES 8
HISTOGRAMS FOR WEAK CORRELATION COEFFICIENTS 10
HISTOGRAMS FOR STRONGER CORRELATION COEFFICIENTS 11
HISTOGRAM FOR TARGET VARIABLE 12
DESCRIPTIVE STATISTICS FOR STRONG CORRELATION COEFFICIENT VARIABLES 12
SCATTERPLOTS FOR VARIABLES WITH >|0.5| CORRELATION COEFFICIENT 13
FREQUENCY TABLES FOR CATEGORICAL VARIABLES 14
BAR CHARTS FOR CATEGORICAL VARIABLES 16
Statistical Methods 18
ANOVA TEST FOR ACTIVITY LEVEL 19
ANOVA TEST FOR SMOKING STATUS 19
ANOVA TEST FOR ALCOHOL CONSUMPTION 19
ANOVA TEST FOR DIET 20
ANOVA TEST FOR CHRONIC DISEASES 20
ANOVA TEST FOR MEDICATION USE 21
ANOVA TEST FOR FAMILY HISTORY 21
ANOVA TEST FOR MENTAL HEALTH STATUS 21
ANOVA TEST FOR SLEEP PATTERNS 22
ANOVA TEST FOR EDUCATION LEVEL 22
ANOVA TEST FOR INCOME LEVEL 23
QUANTATIVITE T-TESTS 23
CATEGORICAL DATA TRANSFORMATION 23
BASELINE LINEAR REGRESSION 24
MULTIPLE LINEAR REGRESSION MODEL 25
REFINED MULTIPLE LINEAR REGRESSION MODEL 27
Interpretation of Results 28
Conclusion 29

References 29

Introduction

In an ever-changing world, it's important to know what factors impact health. These factors can range from self-imposed, like smoking, to inheritable, like family history. Chronological age is the age one is given based on how long they have been alive since birth. This age is calculated in years, months, days, etc. from your birth date. Chronological age is the most common way that age is reported. However, biological age differs from chronological age because it accounts for a variety of different factors more than just your birth date. It consists of genetics or family medical history, physical lifestyle and nutrition, diseases, vital signs, and more.

This statistical report aims to create a linear regression model that can predict age. With the goal of predicting age, we aim to determine which factors have major impact, to help determine if their biological age matches their chronological age.

Data Overview & Cleaning

The dataset used in this report is from Kaggle. The dataset contains many different variables that represent health factors, like cognitive function and bone density (see Data Description File for full variable list). While in Excel, an ID variable was added to create ease of reference. After this, the data was imported to SAS. Some variables still had names that were hard to reference so 9 variables were renamed.

Since the dataset is synthetic, there were no missing values. A Proc Means statement was used to determine that there were no missing values for any of the 3000 observations. Some variables were difficult to work with in their current state, so the following changes were made: blood pressure was separated into systolic and diastolic values instead of a combined value, and stress levels, sun exposure and pollution exposure were rounded to the nearest whole number. The final dataset after the cleaning process has 29 variables. There is 1 ID variable, 12 categorical variables, and 16 numeric variables. Age is the target variable.

Variable Selection

Variables:

1. **Individual ID:** The individual used as identifier variable.
2. **Gender:** The biological gender of the individual.
3. **Height:** The height of the individual in cm.
4. **Weight:** The weight of the individual in kg.
5. **Blood Pressure:** The systolic and diastolic blood pressure of the individual in mmHg.
6. **Cholesterol Level:** The cholesterol level of the individual in mg/dL.
7. **BMI:** Body Mass Index, calculated from height and weight in kg/cm.
8. **Blood Glucose Level:** The blood glucose level of the individual in mg/dL.
9. **Bone Density:** The bone density of the individual in g/cm^2.

10. **Vision Sharpness:** The vision sharpness of the individual.
11. **Hearing Ability:** The hearing ability of the individual in dB.
12. **Physical Activity Level:** The physical activity level of the individual.
13. **Smoking Status:** The smoking status of the individual.
14. **Alcohol Consumption:** The frequency of alcohol consumption.
15. **Diet:** The nutritional diet of the individual.
16. **Chronic Diseases:** The presence of chronic diseases.
17. **Medication Use:** The usage of medication.
18. **Family History:** The presence of family history of age-related conditions.
19. **Cognitive Function:** Self-reported cognitive function.
20. **Mental Health Status:** Self-reported mental health status
21. **Sleep Patterns:** The sleep patterns of the individual.
22. **Stress Levels:** Self-reported stress levels.
23. **Pollution Exposure:** Exposure to pollution.
24. **Sun Exposure:** Average sun exposure.
25. **Education Level:** Highest level of education attained.
26. **Income Level:** Annual income.
27. **Age:** The age of the individual.

Data Dictionary for Age Analysis			
Variable Name	General Type	Specific Type	Variable Dependence
Individual ID	Identifier	Identifier	NA
Gender	Categorical	Nominal	Independent
Height	Quantitative	Continuous	Independent
Weight	Quantitative	Continuous	Independent
Blood Pressure	Quantitative	Continuous	Independent
Cholesterol Level	Quantitative	Continuous	Independent
BMI	Quantitative	Continuous	Independent
Blood Glucose Level	Quantitative	Continuous	Independent
Bone Density	Quantitative	Continuous	Independent
Vision Sharpness	Quantitative	Continuous	Independent
Hearing Ability	Quantitative	Continuous	Independent
Physical Activity Level	Categorical	Ordinal	Independent
Smoking Status	Categorical	Nominal	Independent
Alcohol Consumption	Categorical	Nominal	Independent
Diet	Categorical	Nominal	Independent
Chronic Diseases	Categorical	Nominal	Independent
Medication Use	Categorical	Nominal	Independent
Family History	Categorical	Nominal	Independent
Cognitive Function	Quantitative	Continuous	Independent
Mental Health Status	Categorical	Ordinal	Independent
Sleep Patterns	Categorical	Ordinal	Independent
Stress Levels	Quantitative	Continuous	Independent
Pollution Exposure	Quantitative	Continuous	Independent
Sun Exposure	Quantitative	Continuous	Independent
Education Level	Categorical	Ordinal	Independent
Income Level	Categorical	Ordinal	Independent
Age	Quantitative	Discrete	Dependent

Exploratory Data Analysis

To begin the EDA process, a correlation coefficient matrix was created. For formatting purposes, this is just a portion of the matrix. The target variable is age, so when considering correlation coefficient going forward, the age column is the main column of importance. Please see GitHub document for the full correlation coefficient matrix.

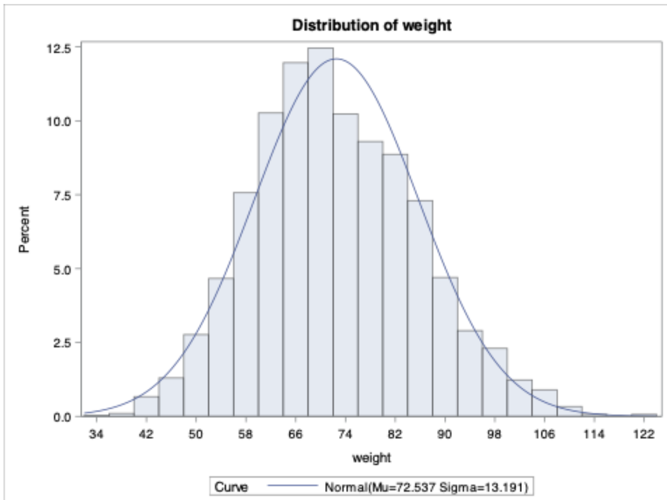
CORRELATION COEFFICIENTS FOR ALL NUMERIC VARIABLES

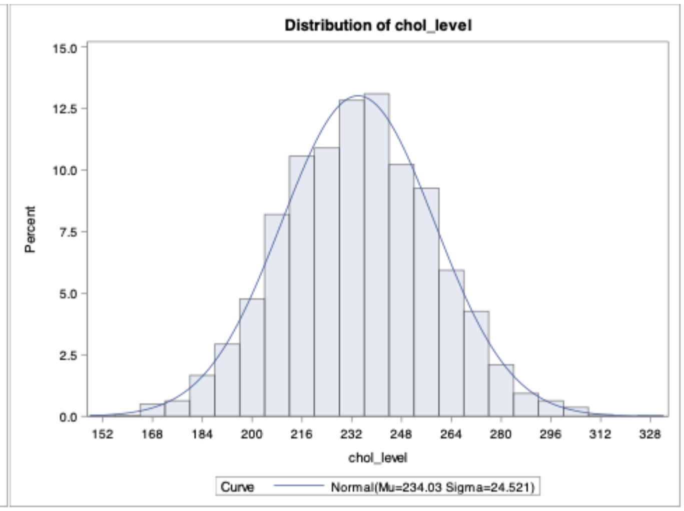
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	age	**height**	**weight**	**systolic_bp**	**diastolic_bp**	**chol_level**	**BMI**
age	1.0000 3000	0.0203 0.2658 3000	0.0025 0.8902 3000	0.6459 <.0001 2998	0.6004 <.0001 3000	0.4324 <.0001 3000	-0.0080 0.6597 3000
height	0.0203 0.2658 3000	1.00000 3000	0.3984 <.0001 3000	-0.0107 0.5572 2998	0.0224 0.2207 3000	-0.0272 0.1359 3000	-0.2228 <.0001 3000
weight	0.00252 0.8902 3000	0.3984 <.0001 3000	1.00000 3000	-0.0147 0.4202 2998	0.0069 0.7066 3000	0.0419 0.0219 3000	0.8002 <.0001 3000
systolic_bp	0.6459 <.0001 2998	-0.0107 0.5572 2998	-0.0147 0.4202 2998	1.00000 2998	0.39508 <.0001 2998	0.26888 <.0001 2998	-0.0029 0.8726 2998
diastolic_bp	0.6004 <.0001 3000	0.0224 0.2207 3000	0.0069 0.7066 3000	0.3951 <.0001 2998	1.00000 3000	0.25855 <.0001 3000	-0.0057 0.7568 3000
chol_level	0.4324 <.0001 3000	-0.0272 0.1359 3000	0.0419 0.0219 3000	0.26888 <.0001 2998	0.25855 <.0001 3000	1.0000 3000	0.0003 3000
BMI	-0.0080 0.6597 3000	-0.2228 <.0001 3000	0.8002 <.0001 3000	-0.00293 0.8726 2998	-0.00566 0.7568 3000	0.06553 0.0003 3000	1.00000 3000
blood_glucose	0.4286 <.0001 3000	0.01199 0.5115 3000	0.0160 0.3822 3000	0.26607 <.0001 2998	0.2402 <.0001 3000	0.1896 <.0001 3000	0.0127 0.4868 3000
stress_lvl	0	0	0	0	0	0	0
sun_expo	0	0	0	0	0	0	0
pollution_expo	0	0	0	0	0	0	0
Cognitive_Function	-0.5081 <.0001 3000	0.0182 0.3182 3000	0.0062 0.7334 3000	-0.3158 <.0001 2998	-0.3063 <.0001 3000	-0.2302 <.0001 3000	-0.0066 0.7199 3000
bone_density	-0.9377 <.0001 3000	-0.0236 0.1971 3000	-0.0081 0.6558 3000	-0.60823 <.0001 2998	-0.5590 <.0001 3000	-0.4012 <.0001 3000	0.0038 0.8335 3000
Vision_Sharpness	-0.8997 <.0001 3000	-0.0093 0.6117 3000	0.0004 0.9809 3000	-0.5769 <.0001 2998	-0.53710 <.0001 3000	-0.38217 <.0001 3000	0.0037 0.8402 3000
hearing_ability	0.7124 <.0001 3000	0.0095 0.6045 3000	0.0089 0.6244 3000	0.45135 <.0001 2998	0.42718 <.0001 3000	0.32847 <.0001 3000	0.0059 0.7476 3000

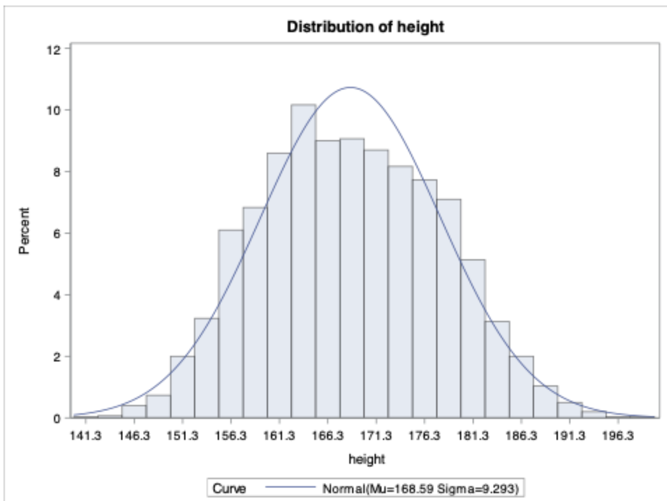
Using this matrix, there are a few variables that have a strong correlation with age. Hearing ability, vision sharpness, and bone density are among the highest correlated variables, while other variables, like systolic and diastolic blood pressure measures, have a more moderate correlation.

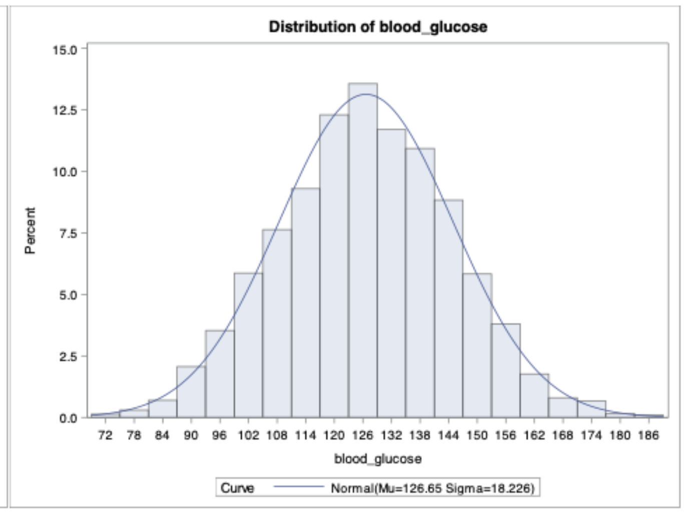
To get a better visual sense of the data, histograms were created for the variables with a weaker correlation coefficient. The distributions for the variables with weaker correlation all seem to have a mostly symmetric distribution. When the histograms for the variables with stronger correlation coefficients were created, there was more variation, with some skewed and some uniform distributions. The target variable has a very uniform distribution.

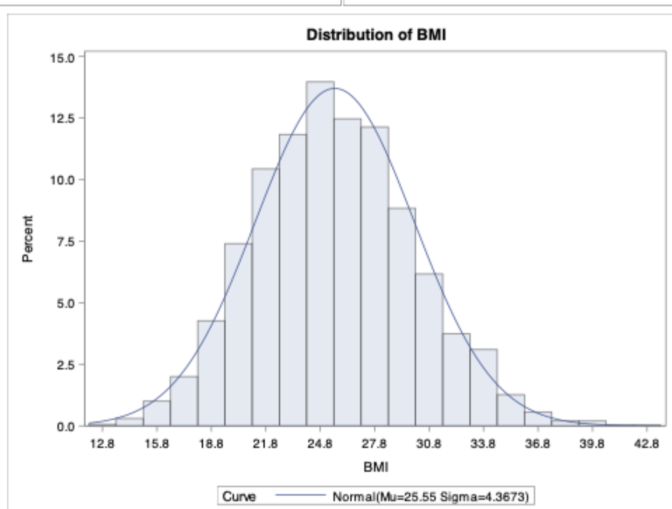
HISTOGRAMS FOR WEAK CORRELATION COEFFICIENTS



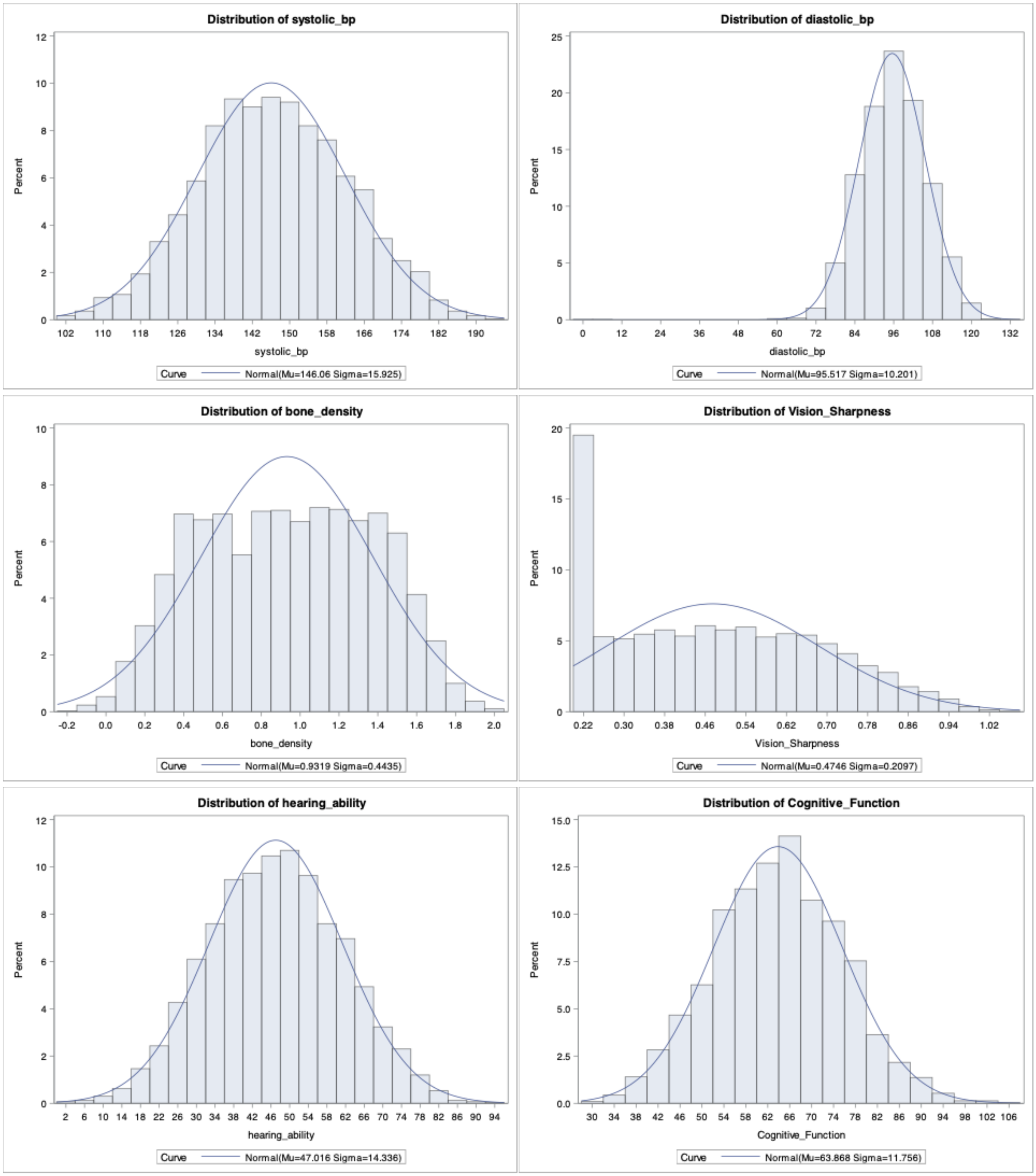




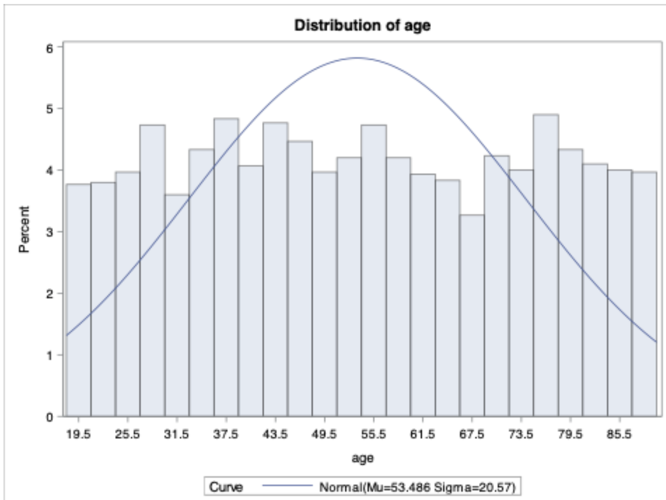




HISTOGRAMS FOR STRONGER CORRELATION COEFFICIENTS



HISTOGRAM FOR TARGET VARIABLE



To better understand the numeric variables, descriptive statistics were generated for all the variables with a moderate to strong correlation coefficient.

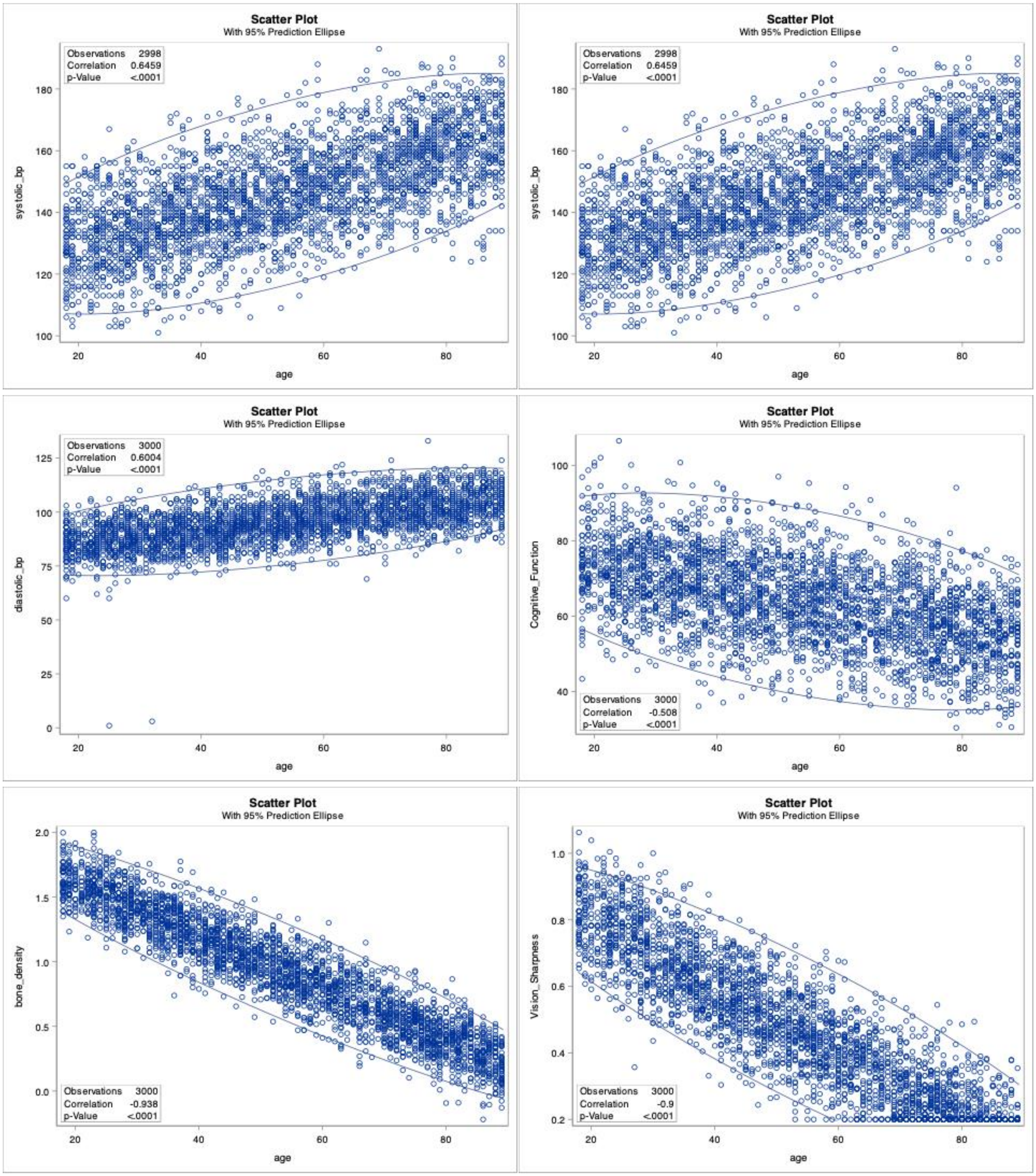
DESCRIPTIVE STATISTICS FOR STRONG CORRELATION COEFFICIENT VARIABLES

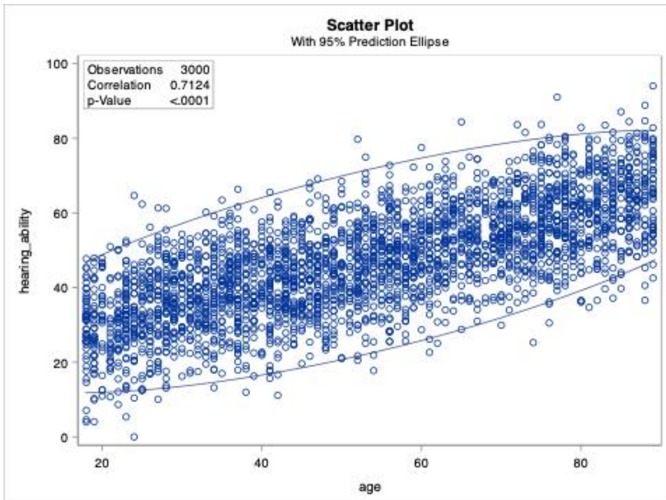
Variable	N	Mean	Std Dev	Median	Lower Quartile	Upper Quartile
systolic_bp	2998	146.0593729	15.9250982	146.0000000	135.0000000	157.0000000
diastolic_bp	3000	95.5170000	10.2006730	95.0000000	89.0000000	103.0000000
Vision_Sharpness	3000	0.4745905	0.2097256	0.4620701	0.2816224	0.6395051
hearing_ability	3000	47.0162137	14.3364639	46.9637327	36.7250629	56.8354916
Cognitive_Function	3000	63.8683760	11.7557384	64.0146519	55.6473284	72.0947020
bone_density	3000	0.9318990	0.4435497	0.9395855	0.5608143	1.2945963

Variable	Quartile Range	Minimum	Maximum	Range
systolic_bp	22.0000000	101.0000000	193.0000000	92.0000000
diastolic_bp	14.0000000	1.0000000	133.0000000	132.0000000
Vision_Sharpness	0.3578828	0.2000000	1.0625375	0.8625375
hearing_ability	20.1104287	0	94.0038243	94.0038243
Cognitive_Function	16.4473736	30.3820982	106.4798308	76.0977326
bone_density	0.7337820	-0.2197872	1.9998289	2.2196161

To visualize these variables against age, scatterplots were created, again using numeric variables with a correlation coefficient of greater than |0.5|.

SCATTERPLOTS FOR VARIABLES WITH >|0.5| CORRELATION COEFFICIENT





From these scatterplots, it can be seen that a linear regression model is appropriate, as all the scatterplots have a linear tendency. Now that there is a better understanding of the quantitative variables, the categorical variables will be looked at, starting with frequency tables.

FREQUENCY TABLES FOR CATEGORICAL VARIABLES

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	1511	50.37	1511	50.37
Male	1489	49.63	3000	100.00

activity_level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	691	23.03	691	23.03
Low	902	30.07	1593	53.10
Moderate	1407	46.90	3000	100.00

Smoking_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Current	793	26.43	793	26.43
Former	1181	39.37	1974	65.80
Never	1026	34.20	3000	100.00

Alcohol_Consumption	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Frequent	742	24.73	742	24.73
None	1201	40.03	1943	64.77
Occasional	1057	35.23	3000	100.00

Diet	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Balanced	1183	39.43	1183	39.43
High-fat	662	22.07	1845	61.50
Low-carb	605	20.17	2450	81.67
Vegetarian	550	18.33	3000	100.00

Chronic_Diseases	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Diabetes	532	17.73	532	17.73
Heart Disease	493	16.43	1025	34.17
Hypertension	676	22.53	1701	56.70
None	1299	43.30	3000	100.00

Medication_Use	Frequency	Percent	Cumulative Frequency	Cumulative Percent
None	1198	39.93	1198	39.93
Occasional	739	24.63	1937	64.57
Regular	1063	35.43	3000	100.00

Family_History	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Diabetes	645	21.50	645	21.50
Heart Disease	453	15.10	1098	36.60
Hypertension	451	15.03	1549	51.63
None	1451	48.37	3000	100.00

Mental_Health_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Excellent	439	14.63	439	14.63
Fair	1009	33.63	1448	48.27
Good	1073	35.77	2521	84.03
Poor	479	15.97	3000	100.00

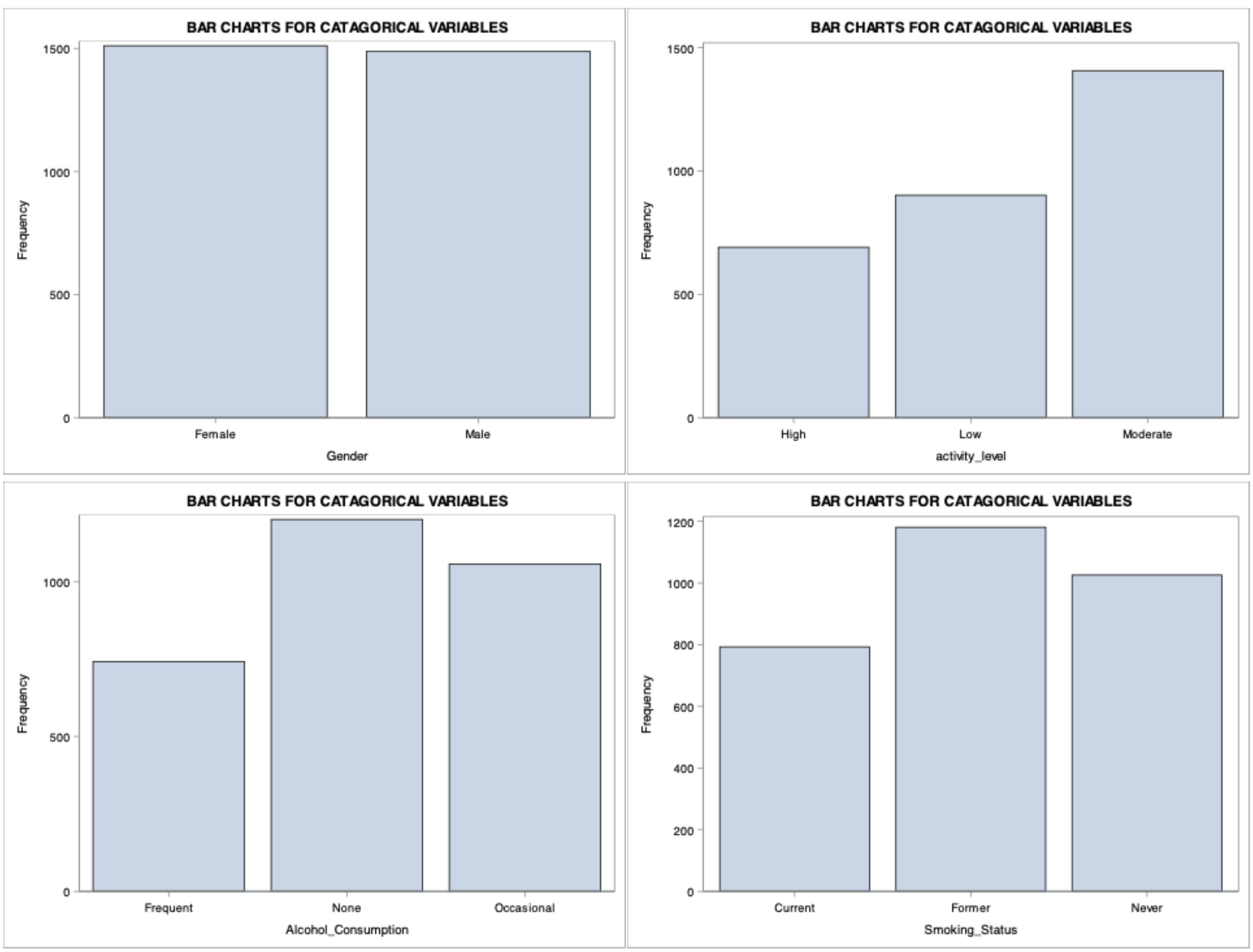
Sleep_Patterns	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Excessive	428	14.27	428	14.27
Insomnia	1053	35.10	1481	49.37
Normal	1519	50.63	3000	100.00

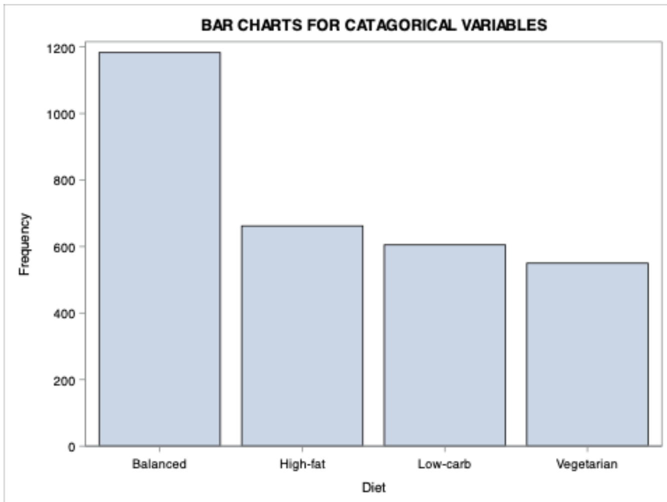
Income_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	861	28.70	861	28.70
Low	916	30.53	1777	59.23
Medium	1223	40.77	3000	100.00

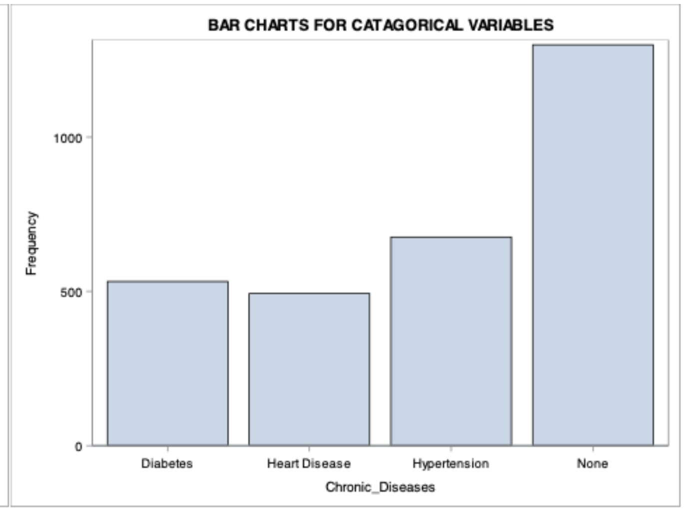
Education_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High School	883	29.43	883	29.43
None	627	20.90	1510	50.33
Postgraduate	606	20.20	2116	70.53
Undergraduate	884	29.47	3000	100.00

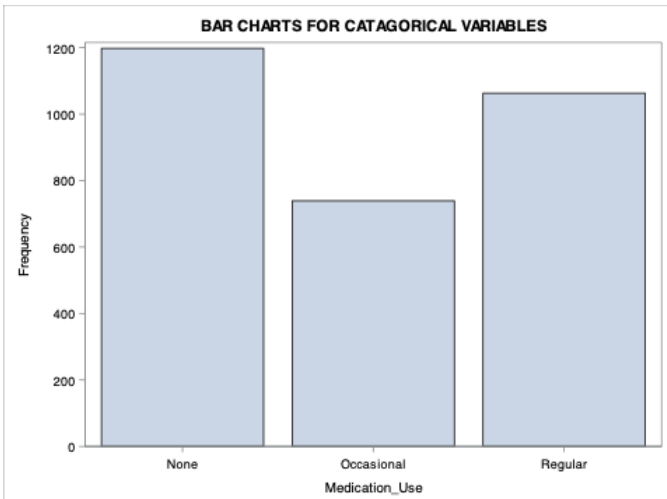
To visualize these tables, bar charts were created for each categorical variable. Below are those visuals.

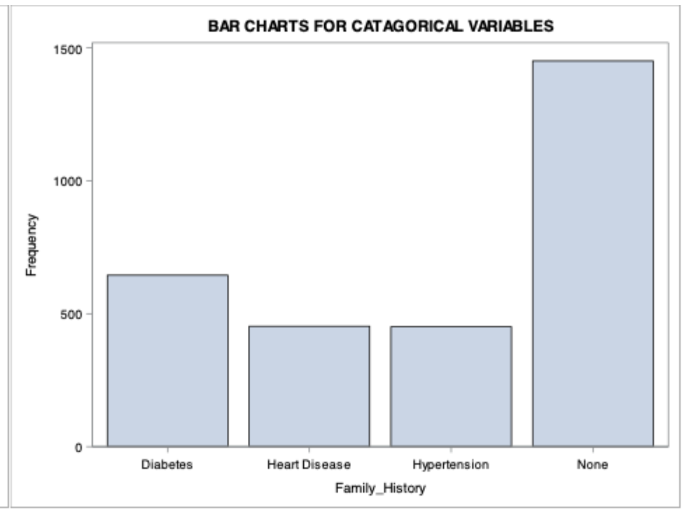
BAR CHARTS FOR CATEGORICAL VARIABLES

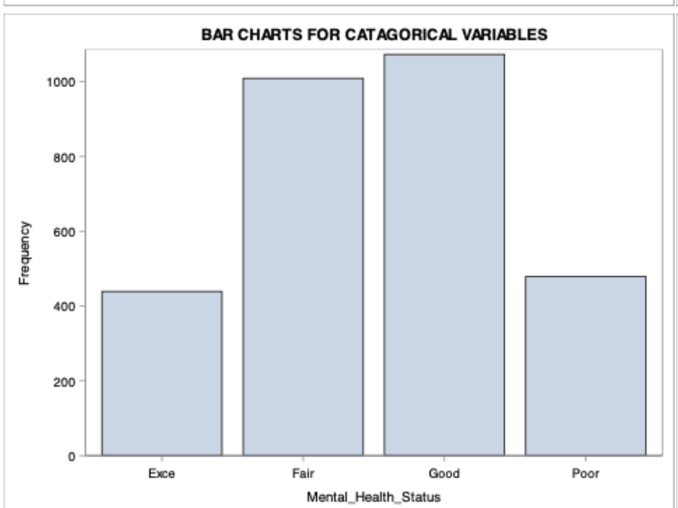


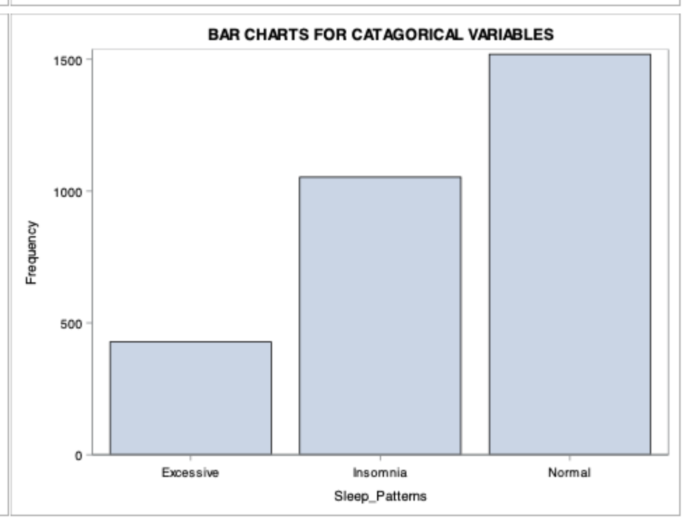

BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
200
400
600
800
1000
1200
Balanced
High-fat
Low-carb
Vegetarian
Diet

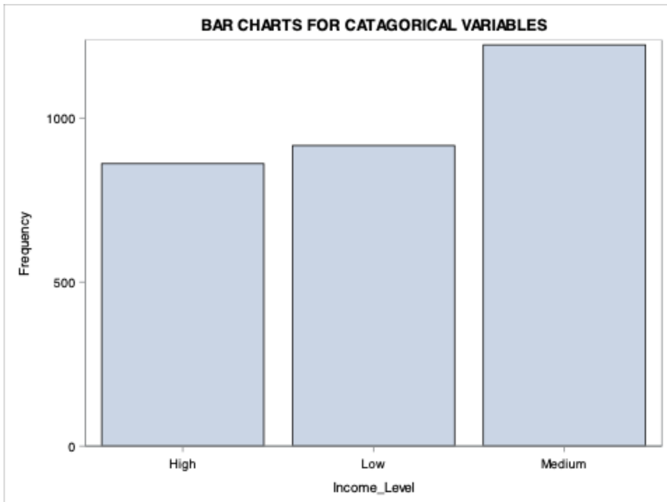

BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
500
1000
Diabetes
Heart Disease
Hypertension
None
Chronic_Diseases

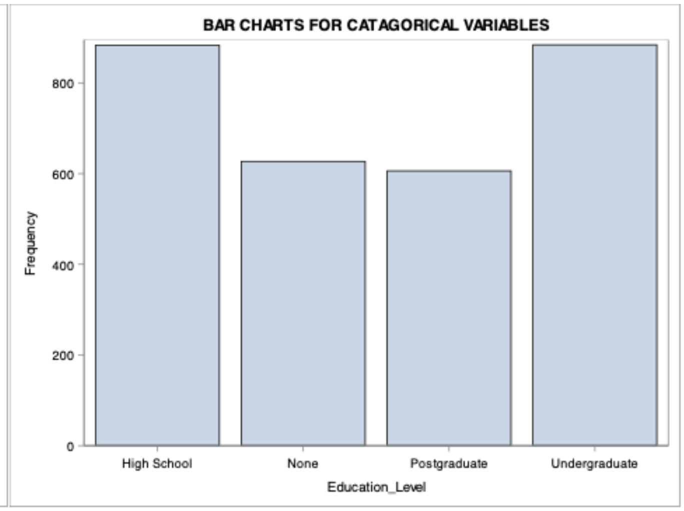

BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
200
400
600
800
1000
1200
None
Occasional
Regular
Medication_Use


BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
500
1000
1500
Diabetes
Heart Disease
Hypertension
None
Family_History


BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
200
400
600
800
1000
Exce
Fair
Good
Poor
Mental_Health_Status


BAR CHARTS FOR CATAGORICAL VARIABLES
Frequency
0
500
1000
1500
Excessive
Insomnia
Normal
Sleep_Patterns





Statistical Methods

To begin the linear regression model, ANOVA tests were run on all categorical variables to see if there is a statistically significant difference between each category in the variables by age. All the ANOVA tests were run with an alpha value of 0.05.

ANOVA TEST FOR GENDER

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	598.284	598.284	1.41	0.2345
Error	2998	1268315.100	423.054		
Corrected Total	2999	1268913.384			

$$H_0: \mu_F = \mu_M$$

$$H_a: \mu_F \neq \mu_M$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for females and males.

ANOVA TEST FOR ACTIVITY LEVEL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	341.491	170.745	0.40	0.6681
Error	2997	1268571.893	423.281		
Corrected Total	2999	1268913.384			

$$H_0:\ \mu_{Low} = \mu_{Moderate} = \mu_{High}$$

$$H_a:\ At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for low, moderate and high activity levels.

ANOVA TEST FOR SMOKING STATUS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	47106.831	23553.415	57.77	<.0001
Error	2997	1221806.553	407.677		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{Never} = \mu_{Former} = \mu_{Current}$$

$$H_a:\ At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is less than the alpha value of 0.05, the null hypothesis is rejected. There is a statistically significant difference between the means of age for never, former and current smoker status. Since there is a significant difference, the variable smoking_status was dummy coded to numeric values to be included in the linear regression model.

ANOVA TEST FOR ALCOHOL CONSUMPTION

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1969.776	984.888	2.33	0.0975
Error	2997	1266943.607	422.737		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{None} = \mu_{Occasional} = \mu_{Frequent}$$

$$H_a:\ At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for none, occasional and frequent alcohol consumption.

ANOVA TEST FOR DIET

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	570.101	190.034	0.45	0.7181
Error	2996	1268343.282	423.346		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{Low-carb} = \mu_{Balanced} = \mu_{Vegetarian} = \mu_{High-fat}$$

$$H_a: At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for low-carb, balanced, vegetarian and high-fat diets.

ANOVA TEST FOR CHRONIC DISEASES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	739.571	246.524	0.58	0.6265
Error	2996	1268173.813	423.289		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{None} = \mu_{Hypertension} = \mu_{Diabetes} = \mu_{HeartDisease}$$

$$H_a: At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for none, hypertension, diabetes or heart disease in chronic diseases.

ANOVA TEST FOR MEDICATION USE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	446.089	223.044	0.53	0.5904
Error	2997	1268467.295	423.246		
Corrected Total	2999	1268913.384			

$$H_0\colon \mu_{None} = \mu_{Occassional} = \mu_{Regular}$$

$$H_a\colon At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for none, occasional and regular medication use.

ANOVA TEST FOR FAMILY HISTORY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	548.124	182.708	0.43	0.7304
Error	2996	1268365.259	423.353		
Corrected Total	2999	1268913.384			

$$H_0\colon \mu_{None} = \mu_{Hypertension} = \mu_{Diabetes} = \mu_{HeartDisease}$$

$$H_a\colon At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for none, hypertension, diabetes and heart disease in family history.

ANOVA TEST FOR MENTAL HEALTH STATUS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	165.463	55.154	0.13	0.9421
Error	2996	1268747.921	423.481		
Corrected Total	2999	1268913.384			

$$H_0\colon \mu_{Poor} = \mu_{Fair} = \mu_{Good} = \mu_{Excellent}$$

$$H_a\colon At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for poor, fair, good and excellent mental health.

ANOVA TEST FOR SLEEP PATTERNS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	146.326	73.163	0.17	0.8413
Error	2997	1268767.058	423.346		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{Normal} = \mu_{Excessive} = \mu_{Insomnia}$$

$$H_a: At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for normal, excessive and insomnia sleep patterns.

ANOVA TEST FOR EDUCATION LEVEL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3306.868	1102.289	2.61	0.0499
Error	2996	1265606.516	422.432		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{None} = \mu_{HighSchool} = \mu_{Undergrad} = \mu_{PostGrad}$$

$$H_a: At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is less than the alpha value of 0.05, the null hypothesis is rejected. There is a statistically significant difference between the means of age for none, high school, undergraduate, and postgraduate education levels. Since there is a significant difference, the variable education_level was dummy coded to ordinal numeric values to be included in the linear regression model.

ANOVA TEST FOR INCOME LEVEL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	978.123	489.061	1.16	0.3149
Error	2997	1267935.261	423.068		
Corrected Total	2999	1268913.384			

$$H_0: \mu_{Low} = \mu_{Medium} = \mu_{High}$$

$$H_a: At\ least\ one\ of\ the\ means\ is\ not\ equal\ to\ the\ others.$$

Since the p-value is greater than the alpha value of 0.05, the null hypothesis fails to be rejected. There is not a statistically significant difference between the means of age for low, medium, and high income levels.

Statistical Method Test Decision	
Test	Decision
ANOVA test for Gender	Fail to Reject
ANOVA test for Activity Level	Fail to Reject
ANOVA test for Smoking Status	Reject
ANOVA test for Alcohol Consumption	Fail to Reject
ANOVA test for Diet	Fail to Reject
ANOVA test for Chronic Diseases	Fail to Reject
ANOVA test for Medication Use	Fail to Reject
ANOVA test for Family History	Fail to Reject
ANOVA test for Mental Health Status	Fail to Reject
ANOVA for Sleep Patterns	Fail to Reject
ANOVA test for Education Level	Reject
ANOVA test for Income Level	Fail to Reject

QUANTATIVITE T-TESTS

After conducting the ANOVA tests, we conducted t-tests on all of the quantitative variables. After performing these t-tests, the null hypothesis for each t-test was rejected. It was determined that all of the quantitative variables would be used in the multiple linear regression model.

CATEGORICAL DATA TRANSFORMATION

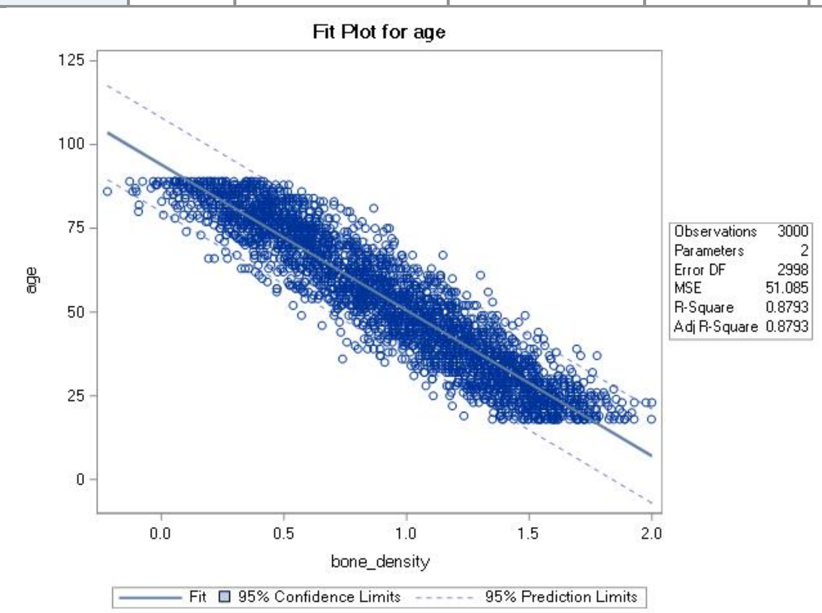
After determining that smoking status and education level are the categorical variables that have a difference in mean age, a few variables were altered. Education level was ordinal and ranked using the numbers 0-3. The new variable was called education_level_rank. Smoking status was dummy coded into two variables. Smoking_status_current, where current smoker=1 and everyone else=0, and smoking_status_former, where a former smoker=1 and everyone else=0. Then, the linear regression model was built with the transformed categorical variables and all of the quantitative variables.

BASELINE LINEAR REGRESSION

A baseline linear regression model was developed based on the variable with the strongest correlation coefficient, bone density.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	94.01074	0.30368	309.57	<.0001
bone_density	1	-43.48655	0.29425	-147.79	<.0001



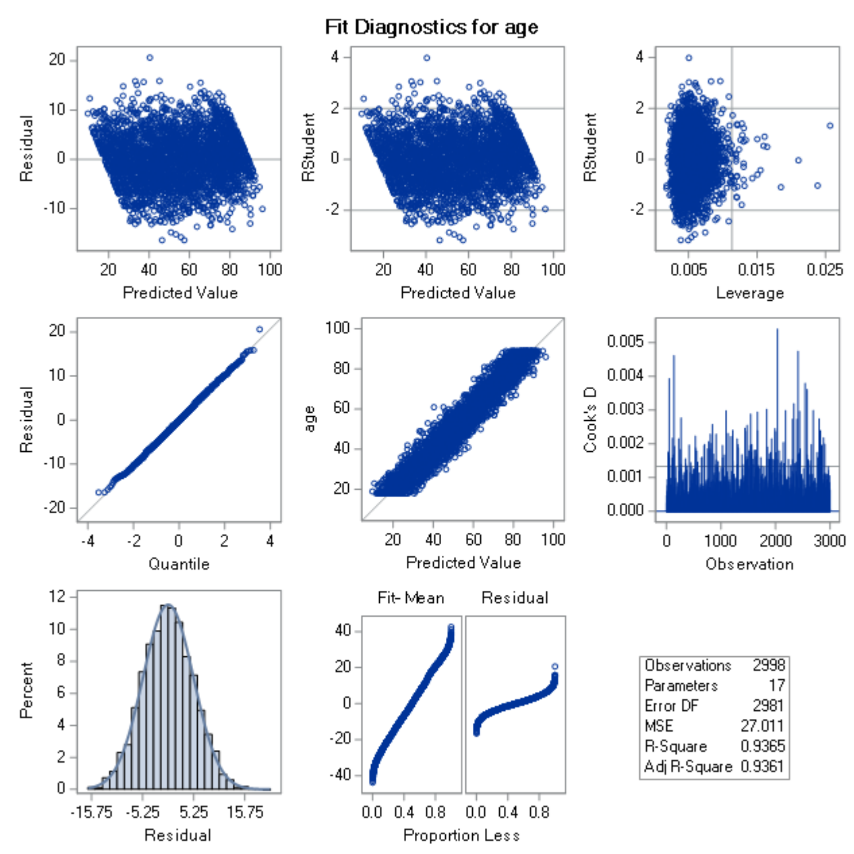
The following estimated baseline linear regression equation can be concluded from the output:

$$\hat{y} = 94.01074 - 43.48655(x_{bone\_density})$$

MULTIPLE LINEAR REGRESSION MODEL

A multiple linear regression model was developed based on the transformed categorical variables and the quantitative variables.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	61.91962	10.85676	5.70	<.0001
smoking_status_current	1	0.55645	0.24959	2.23	0.0259
smoking_status_former	1	0.57417	0.22497	2.55	0.0108
education_rank	1	0.00555	0.09197	0.06	0.9519
height	1	-0.08836	0.06337	-1.39	0.1633
weight	1	0.11288	0.07260	1.55	0.1201
chol_level	1	0.03138	0.00428	7.34	<.0001
BMI	1	-0.35273	0.20631	-1.71	0.0874
blood_glucose	1	0.04153	0.00573	7.25	<.0001
bone_density	1	-22.88466	0.45303	-50.51	<.0001
Vision_Sharpness	1	-28.95163	0.89399	-32.38	<.0001
hearing_ability	1	0.12943	0.00915	14.14	<.0001
Cognitive_Function	1	-0.06714	0.00933	-7.19	<.0001
systolic_bp	1	0.09653	0.00763	12.65	<.0001
diastolic_bp	1	0.14226	0.01183	12.02	<.0001
stress_lvl	1	-0.00787	0.03634	-0.22	0.8286
pollution_expo	1	-0.00294	0.03300	-0.09	0.9290
sun_expo	1	-0.01097	0.02723	-0.40	0.6870



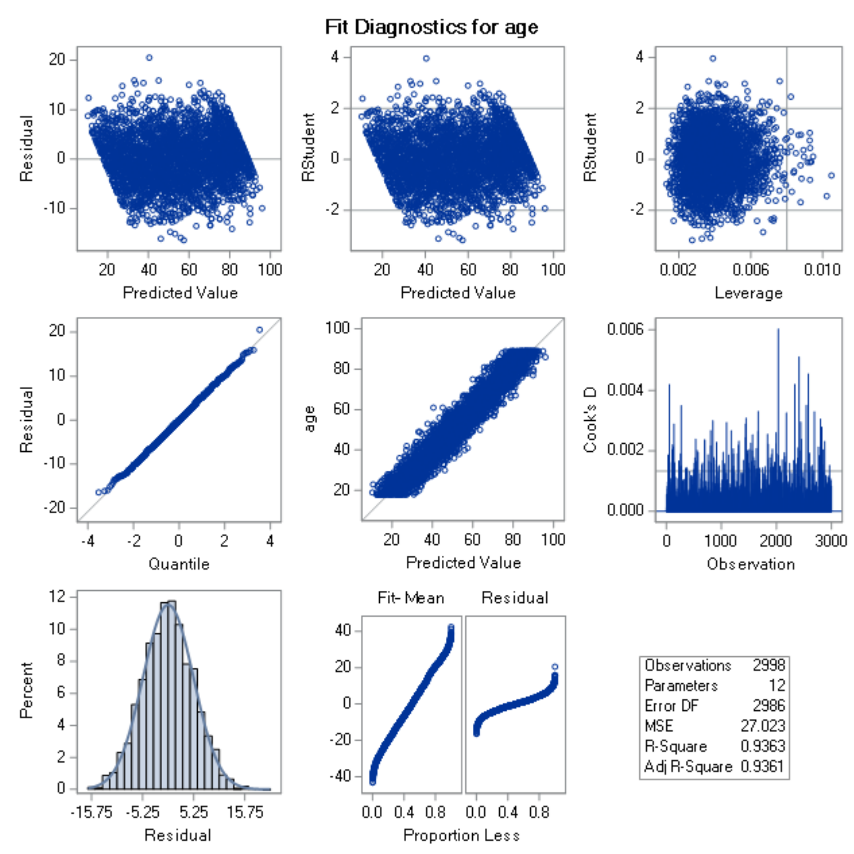
The following estimated multiple linear regression equation can be concluded from the output:

$$\begin{aligned}\hat{y} = 61.89787 &+ 0.55839(x_{smoking\_status\_current}) + 0.57583(x_{smoking\_status\_former}) \\ &+ 0.00529(x_{education\_rank}) - 0.08845(x_{height}) + 0.11300(x_{weight}) \\ &+ 0.03137(x_{chol\_level}) - 0.35292(x_{BMI}) + 0.04149(x_{blood\_glucose}) \\ &- 22.88865(x_{bone\_density}) - 28.94983(x_{vision\_sharpnness}) \\ &+ 0.12936(x_{hearing\_ability}) - 0.06716(x_{cognitive\_function}) \\ &+ 0.09653(x_{systolic\_bp}) + 0.14210(x_{diastolic\_bp}) - 0.00819(x_{stress\_lvl}) \\ &- 0.00327(x_{pollution\_expo}) - 0.01097(x_{sun\_expo})\end{aligned}$$

REFINED MULTIPLE LINEAR REGRESSION MODEL

A refined multiple linear regression model was developed based on the transformed categorical variables and the quantitative variables that had a p-value less than alpha = 0.05.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	46.37602	2.30358	20.13	<.0001
smoking_status_current	1	0.55679	0.24941	2.23	0.0257
smoking_status_former	1	0.57464	0.22483	2.56	0.0106
education_rank	1	0.00729	0.09187	0.08	0.9368
chol_level	1	0.03064	0.00426	7.19	<.0001
blood_glucose	1	0.04125	0.00573	7.20	<.0001
bone_density	1	-22.91896	0.45242	-50.66	<.0001
Vision_Sharpness	1	-28.95953	0.89375	-32.40	<.0001
hearing_ability	1	0.12922	0.00915	14.12	<.0001
Cognitive_Function	1	-0.06676	0.00932	-7.16	<.0001
systolic_bp	1	0.09578	0.00762	12.57	<.0001
diastolic_bp	1	0.14264	0.01182	12.07	<.0001



The following estimated refined multiple linear regression equation can be concluded from the output:

$$
\begin{aligned}
\hat{y} = 46.37602 &+ 0.55679(x_{smoking\_status\_current}) + 0.57464(x_{smoking\_status\_former}) \\
&+ 0.00729(x_{education\_rank}) + 0.03064(x_{chol\_level}) + 0.04125(x_{blood\_glucose}) \\
&- 22.91896(x_{bone\_density}) - 28.95953(x_{vision\_sharpness}) \\
&+ 0.12922(x_{hearing\_ability}) - 0.06676(x_{cognitive\_function}) \\
&+ 0.09578(x_{systolic\_bp}) + 0.14264(x_{diastolic\_bp})
\end{aligned}
$$

Interpretation of Results

From the statistical analysis, various health factors (smoking status, education level, height, weight, cholesterol level, body mass index, blood glucose, bone density, vision sharpness, hearing ability, cognitive function, blood pressure, stress levels, and pollution exposure) display a linear relationship with age. This provides meaningful interpretations regarding how these factors can be used to potentially predict age. This may add valuable insights into how various lifestyle factors, physical conditions, mental health and previous medical history can impact one's health conditions.

Additionally, the results suggest that this may add valuable information for patients who desire to have their biological age tested against their chronological age. This could be useful for people who are trying to see if they are aging healthily based on personal factors.

Conclusion

The results of our analysis imply that health factors, such as smoking status, education level, height, weight, cholesterol level, body mass index, blood glucose, bone density, vision sharpness, hearing ability, cognitive function, blood pressure, stress levels, and pollution exposure, may be meaningful predictors of age. In this statistical analysis, a multiple linear regression model was built based on these factors to potentially predict one's biological age in comparison to their chronological age.

References

Frothingham, S. (2023, July 24). *Chronological vs. Biological Aging: Differences & More*. Healthline. https://www.healthline.com/health/chronological-ageing#biological-aging