Exploring factors associated with bone mineral density in adult residents of the USA

Gómez Pérez Álvaro
Lim Eduardo Luigi Miguel
Smans Agnes
Nhim Malai
Hussen Kedir Adem

December 23, 2020

Contents

Abstract

Osteoporosis, along with other age related diseases, is becoming a popular field of research as worldwide life expectancy is gradually increasing. Osteoporosis increases the probability of bone fractures and is associated with low bone mineral density (BMD, g/cm^2). In this study, the aim is to find lifestyle, demographic and genetic factors associated with BMD in adult residents of the USA. Additionally, the difference of these factors between genders is also researched. To do so, observational data from the National Health and Examination Survey (Centre for Disease Control and prevention) were analysed, containing information about 1814 USA residents. For the general population, as well as for the two different genders, separate multiple regression models were fitted using variables selected by stepwise AIC based selection. For the whole population, several factors proved to be significantly associated with BMD. Black skin colour, being overweight/obese, height and male gender are factors associated with higher BMD, whereas being underweight, low physical activity, higher age and high school graduates (in comparison to college graduates) accounts for lower BMD. For males, age does not prove to be of significant negative effect on BMD. For females, age is significantly associated with BMD, but being underweight or physical activity is not significant. The interaction effect between pack years of smoking and being overweight does prove to be of significance. Females that have ever used cocaine are also associated with higher BMD in comparison to other females.

1 Introduction

Worldwide, life expectancy has been gradually increasing for at least the past 200 years (World Health Organisation, 2019). According to the most recent data, life expectancy for people born in the United States is around 77 years for men and 82 years for women. With increasing life expectancy, the study of diseases related with old age is becoming increasingly important(Belikov A.V. et al., 2019). One disease related to old age is osteoporosis. Osteoporosis is clinically defined by the World Health Organization as areal bone mineral density (BMD, g/cm^2) more than 2.5 standard deviations below the young adult average. For adults, men generally have higher BMD than women, and BMD decreases with age. People with black skin colour also tend to have higher BMD in comparison to white skin colour(Araujo A.B. et al., 2007). Apart from gender, age & race, bone mineral density in adults is also associated with many lifestyle habits from both childhood and adulthood(Welten D.C. et al. (1994), Segheto K.J. et al., 2020). Lower bone mineral density could be associated with smoking(Rudäng R. et al., 2012), Ward K.D. et al., 2001), low physical activity(Langsetmo L. et al., 2012)) and unhealthy, unbalanced diet(Muraki S. et al., 2007). In particular, a diet lacking sufficient calcium intake during childhood and adolescence results in lower BMD later in life(Kalkwarf H.J et al., 2003). Drug use is considered to be only of minor effect directly on BMD, but can indirectly be of significance because it affects BMI(Sophocleous A. et al., 2017). Being underweight can alter adipose tissue surrounding bones, reducing its protective capacities(Palermo A. et al., 2016). On the other hand, being overweight or obese can lead to higher BMD because of larger strain posed on bones, causing bone cells to adapt(Felson et al., 1993). Understanding the factors associated with the decline of BMD in the adult phase is important, as lower BMD increases the probability of bone fractures(Kanis J.A., 2002). In this research, data from over 1800 USA residents is analysed to find associations between bone mineral density and several lifestyle, demographic and genetic factors. Additionally, dependency of these factors on gender are also researched.

2 Data and Method

2.1 Data Description

The data used in this study comes from a secondary observational data source collected in the years 2017-2018 for adults aged 20 - 59 years living in the United States(NHANES,2020). It contains information about 2258 residents of the USA who participated in the National Health and Examination Survey (NHANES) study and had a record for total bone mineral density.

The different datasets from NHANES were individually examined and studied in order to identify variables of interest. First, an overview of each of them allowed for the removal of unnecessary variables, such as those pertaining to individuals under 20 years of age. Afterwards, each of the data sets underwent a process of variable selection based on literature, and, in some cases, the construction of new, more informative variables, that will be detailed below.

From the demographics data set, age, gender and race were immediately selected, as both the literature and a quick study of their correlation with the total BMD pointed towards them significantly affecting the outcome variable. Apart from these, marital status, educational level and ratio of family income to poverty were also kept in the study in account of them possibly helping explain the variation in total bone mineral density. Ratio of family income to poverty was categorized into four levels; as lower class (ratio ≤ 2), middle class (ratio $2-3$), upper-middle class (ratio $3-5$) and upper class (ratio ≥ 5) ("Poverty Income Ratio", 2020).

The physical activity dataset questionnaire was found to be based on the Global Physical Activity Questionnaire (GPAQ) by the World Health Organization. Guidelines from the GPAQ were followed in order to calculate a single variable designated as Physical Activity Level (PAL). MET-minutes/week were calculated as a sum-product of duration (in minutes) and frequency (in days), scaled by intensity based on suggested MET (metabolic equivalent of task) scores. The individual was then classified as either low, moderate, or high based on cut-offs from GPAQ documentation.

The whole data set for the smoking of cigarettes was summarized in a single variable known as Pack Years, which was found to be a commonly used clinical measure to quantify an individual's exposure to tobacco. It was calculated as the product of the number of daily cigarette packs smoked by each individual, by the number of years the smoking habit was kept, all of which was calculated from the original variables in the data set.

Regular milk consumption status and healthy diet were selected from the diet behavior and nutrition dataset. However, 27% of Milk consumption during the age of 5-12, 13-17, and 18-35 were missing as shown in Table 10, which forced them to be excluded from the analysis despite strong evidence from the literature, in favor of keeping more observations in the dataset. From the drug use data set, ever use of hard drugs cocaine, heroin, methamphetamine and regular use of soft drug marijuana were selected to identify individual effects on the response variables. Regular use status of marijuana/weed was preferred over ever use based on the literature.

From the weight history dataset, self reported height and new computed variable BMI, which is a measure of body fat based on height and weight that applies to adult men and women, were selected. Also BMI was preferred to be used since it contains information about both weight and height (CDC, 2020). BMI was calculated as in equation 1 and categorized into four standard categories, namely; underweight (below 18.5), normal (18.5−24.9), overweight (25.0−29.9) and obese (above & 30.0) (CDC, 2020). Different guidelines were imposed for Asians as found in the literature(Lim J.U. et al. (2017)).

$$BMI = \frac{weight\ (in\ kg)}{(height\ (m))^2}$$

The final dataset contained 17 variables: total bone mineral density (TOBMD), age, gender, educational level, marital status, race, ratio of family income to poverty, self-reported healthy diet index, milk consumption, pack years, ever consumption of cocaine, heroin and methamphetamine, regular use of cannabis, BMI, and physical activity level variables which were selected based on literature. In this case, only complete cases were considered in order to facilitate the analyses which means that observations with missing values were dropped from our dataset. Only 1814 observations were used for the analysis. Throughout this study, total bone mineral density is considered the response variable. Variables age, pack years and height are continuous regressors. All remaining variables are categorical with 2-6 levels. Additional information about the structure of the separate variables is provided in Table 9.

2.2 Preliminary analysis

After understanding the datasets and all variables that are contained in them, exploratory analysis was performed in order to gain more insight into each of them. In all cases, the Total Bone Mineral Density was considered as the response variable. Summary statistics for continuous variables were computed in order to identify the summary measures for each variable such as mean, maximum and minimum. Next, scatter plots were produced for continuous variables to see how they were related to the response variable. From those plots, the relationship between some predictors were explored. For categorical variables, box plots were created in order to check how different categories of each variable differ in the amount of observations they contain and across gender by looking at the variability of the response and the median. Interaction plots also were plotted to check some variables with interaction effect on the response variable.

2.3 Multiple Linear Regression

In statistics, regression analysis consists of techniques for modeling the relationship between a dependent variable and one or variables. In this study the response variable is continuous and the predictor variables consist of both continuous and categorical variables, the regression type used in this study is multiple linear regressions. Three models were fitted to address the study objective. In general, a multiple linear regression model can be written as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i; i = 1, 2, 3..n$$

where

- Y_i is the i^i observation of response variable.
- β_0, β_1, β_1,...,β_p are parameters of regression coefficients.
- ϵ_i are random errors, normally distributed with mean=0 and constant variance σ^2.

2.3.1 Model building

A multiple linear regression model was fitted for the general population; separate models were fitted for males and females to address the research question of the study. In order to identify a small group of regressors from a large set of regressors, model selection was done using stepwise selection based on AIC (Akaike's Information Criterion). Additionally, for all three models, Age and Gender are kept in the model because they are considered the most important confounders. Stepwise selection is an automated search procedure designed to sequentially evaluate the pool of potential variables to arrive at a "best" subset (Kutner et al., 2005)

We started with a full model that contains all predictors and two-way interaction terms which were identified by the literature and data exploratory data analysis. A list of all interaction effects considered can be found in the appendix. Interaction effects higher than two-way were not considered as they complicate model interpretation(Neter J. et al. (1996)) Stepwise regression procedures were applied for determining the best models with reduced predictors where the decision rule is based on AIC. The model with the smallest AIC is considered as 'best'.

2.3.2 Model Diagnostics

The basic results that are used for making inferences about multiple linear regression models depend on whether the assumptions are met or not. That means, the distribution theory, confidence intervals, and tests of hypotheses are valid and have meaning only if the standard regression assumptions are satisfied. Therefore, the regression assumptions should be checked before drawing statistical conclusions from the analysis by visual inspection through the following diagnostic plots.

To investigate the normality assumption of error terms, normal probability plots (Q-Q plot) of the residuals were obtained. Normality of error terms is indicated by all the points lying fairly on a straight line. Next, a scatter plot of the square root of standardized residuals and fitted values was produced to examine constancy of error variance. The constancy of error variance was indicated by no systematic pattern of points in the plot. And then highly leveraged observations were identified by the leverage plot based on the model for further inspection. Lastly, variance inflation factor (VIF) was computed to detect multicollinearity in the final regression models.

3 Results

3.1 Exploratory data analysis

The summary statistics of the continuous variables are presented in Table 1. These summary statistics are also presented for both males and females in Table 2 and Table 3, respectively. On the average, males have higher total bone mineral density than females. Majority of the observations have 0 pack years (i.e., non-smokers) in both the general dataset (64%) and in males (57%) or females (71%). Due to the eligibility criteria for dual-energy X-ray absorptiometry and the research objectives, the range of values for age is as expected.

Table 1: Summary statistics of continuous variables

Variable	Mean	Std Dev	Minimum	Maximum
Total Bone Mineral Density (in g/cm^2)	1.12	0.109	0.697	1.73
Age (in Years)	39.9	11.7	20	59
Height (in Inches)	66.2	3.87	53	78
Pack Years	4.27	9.29	0	70

Table 2: Summary statistics of continuous variables for males

Variable	Mean	Std Dev	Minimum	Maximum
Total Bone Mineral Density (in g/cm^2)	1.16	0.108	0.876	1.73
Age (in Years)	39.7	11.7	20	59
Height (in Inches)	68.8	2.92	59	78
Pack Years	5.37	10.4	0	70

Table 3: Summary statistics of continuous variables for females

Variable	Mean	Std Dev	Minimum	Maximum
Total Bone Mineral Density (in g/cm^2)	1.08	0.0979	0.697	1.4
Age (in Years)	40.1	11.7	20	59
Height (in Inches)	63.7	2.87	53	72
Pack Years	3.26	7.98	0	58

With respect to categorical variables, most of them had a relatively uniform distribution of the observations between their different categories, but some particular ones contained categories which were less represented within the data. As this might affect the results of the analysis, the graphs for the distribution of such variables are reported below. This was the case of the variables relative to drug use, especially for cocaine, methamphetamine and heroin, where the amount of individuals consuming drugs was as low as 5% to 10%

of the total population. The only other two variables which presented this situation were BMI, where those classified as underweight represent only 2% of the total, and marital status, where there are only a 1% of widowed individuals (see Appendix).

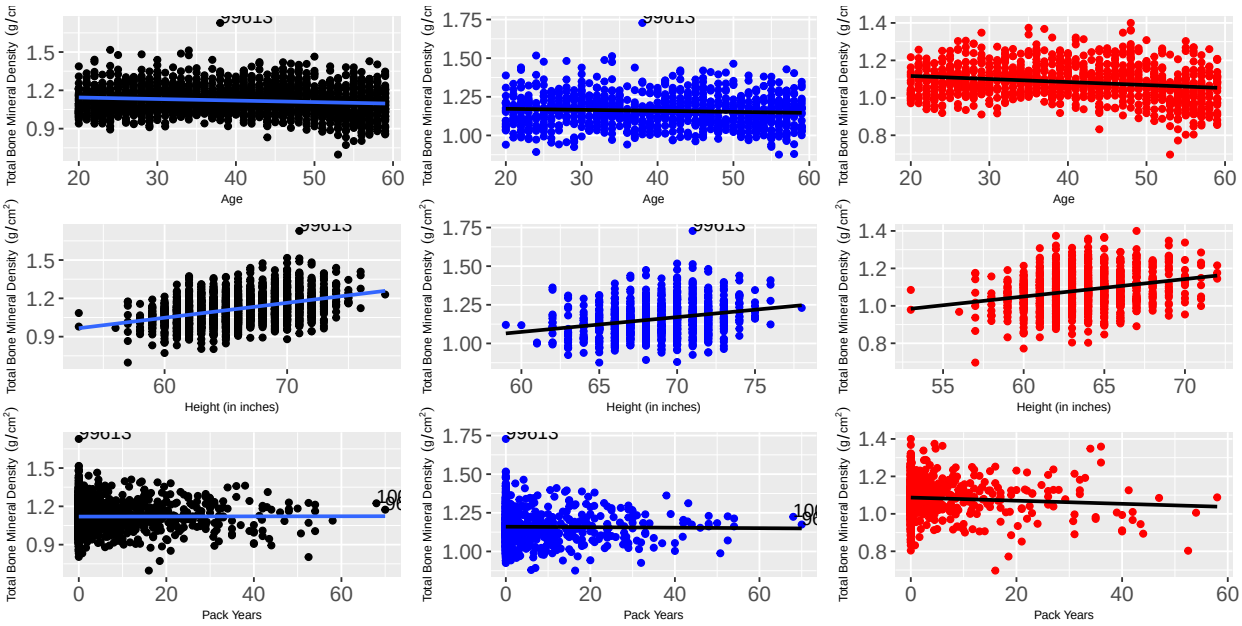

Figure 1: Scatter plot for continuous variables

3.2 Model Fitting

Evaluation of model assumptions is done solely through inspection of the residual plots obtained from each model. For the initial model of female respondents, no gross deviations from normality can be observed from the Normal Q-Q plot. The residuals vs fitted values plot do not show any discernible pattern about zero which implies that the linearity assumption has not been violated. The Scale-Location plot also shows random dispersion of the fitted values and an approximately horizontal line, evidence in support of homoscedasticity. (See Appendix) Stepwise selection is then used to choose a subset of the regressors for each model.

3.3 Model diagnostics

Diagnostic plots are presented for all models but are evaluated together to facilitate the discussion. Observations specific to a plot in a model are found underneath each specific figure.

Table 4: Test of Multicollinearity Test

	GVIF	DF	GVIF^2(1/(2*DF))
Age	2.017893	1	1.420526
Gender	19.648005	1	4.432607
Race	37.145466	5	1.435458
Height	2.127023	1	1.458432
BMI_class	1.228017	3	1.034826
PAL_GPAQ	1.176661	2	1.041508
Education	1.436776	3	1.062262
Self_Diet	1.204396	1	1.097450
Gender:Race	135.689053	5	1.634009
Age:Gender	13.753045	1	3.708510

Table 5: Test of Multicollinearity Test

	GVIF	DF	GVIF^2(1/(2*DF))
Age	1.092116	1	1.045043
Gender	2.036804	1	1.427167
Race	1.592676	5	1.047642
Height	2.122867	1	1.457006
BMI_class	1.209207	3	1.032167
PAL_GPAQ	1.169778	2	1.039982
Education	1.402089	3	1.057944
Self_Diet	1.199658	1	1.095289

Linearity of the regression model is proven by plotting the residuals against the fitted values for the outcome. This graph proves the lack of any systematic pattern in the residuals as a function of the fitted values, which therefore allows us to confirm the linear relationship between the regressors and the outcome variable as shown in Figure .

Normality of the error term is also checked by plotting the residual values, in this case in a QQ-plot, which allows us to perform a quick visual inspection of how close their distribution is to being perfectly normal. As seen below, these residuals are very close to fitting the QQ-line perfectly, except for one outlier and a small right tail. However, taking into account the number of observations in the population, normality can be assumed without any concern as shown in Figure 1(b).

Homoscedasticity is evaluated by visual inspection of the Scale-Location plot. This plot shows that residuals are homogeneously distributed, that is, their spread is roughly equal for all fitted values, and no particular pattern can be detected for the general model in Figure 1(c).

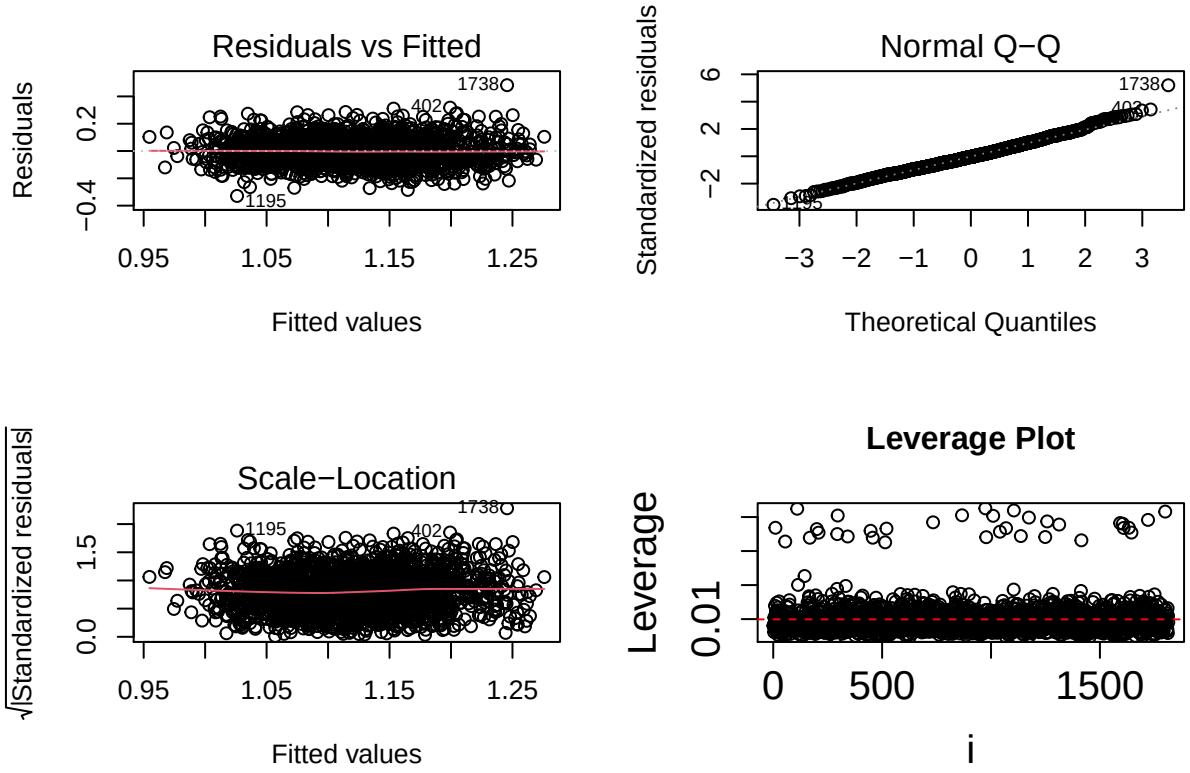

Figure 2: Model diagnostic plot for general model

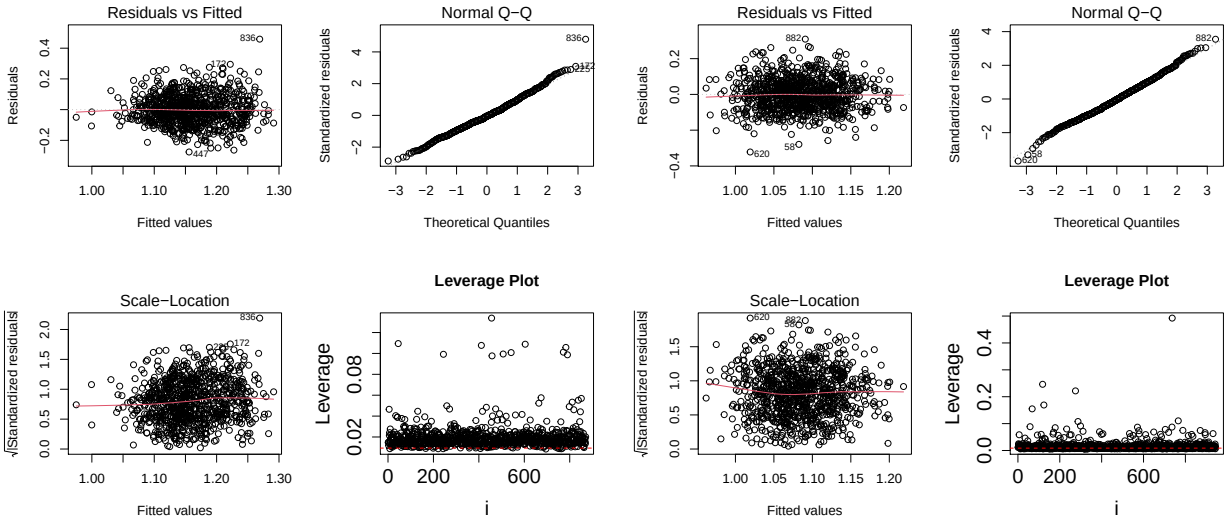

Figure 3: Model diagnostic plot for Male (Left) and Female (Right) model

3.4 Model Interpretation

From our evaluation analyses, our final model respects all the assumptions for a multiple linear regression model. Hence, our final model was adequate and fits the data very well. The detail for the parameter estimates with corresponding standard error, t-value, p-value and 95% confidence interval were presented in Table 6.

From Table 6, one can note that:

- Age shows a significant effect on BMD, suggesting an average decrease of $0.00103 g/cm^2$per every year increased in age. Gender also affects BMD, as our parameter estimate indicates a larger BMD average for men than for women, by $0.030299 g/cm^2$.
- Race is a significant factor in defining BMD as well, as in average, the total bone mineral density of Non-Hispanic Black - adults is $0.063666 g/cm^2$ larger than Mexican Americans. At the same time, other races not explicitly mentioned in the study, have a higher average bone mineral density than Mexican Americans, in this case by $0.032129 g/cm^2$.
- On average, every inch taller increases total bone mineral density by $0.007524 g/cm^2$.
- BMI also significantly affects BMD, as both obese and overweight people show averagely larger BMD values than people in a healthy weight (by $0.034559 g/cm^2$ and $0.019765 g/cm^2$ respectively), while in the case of underweight individuals, BMD is lower in average with respect to the same reference group, by $0.05146 g/cm^2$.
- People with low physical activity levels, on the average, have $0.01498 g/cm^2$ lower total bone mineral density compared to people with high physical activity levels.
- The average total bone mineral density of people with a maximum educational level of high school graduate was $0.0142 g/cm^2$ lower than those who have graduated from college or higher education.

Table 6: Parameter Estimates for the general model

term	estimate	std.error	statistic	p.value	lower CI	upper CI
(Intercept)	0.6413820	0.0540999	11.8555123	0.0000000	0.5352766	0.7474874
Age	-0.0010305	0.0001958	-5.2625757	0.0000002	-0.0014145	-0.0006464
GenderMale	0.0302992	0.0062516	4.8466305	0.0000014	0.0180380	0.0425604
RaceNon-Hispanic Asian	-0.0111575	0.0088591	-1.2594389	0.2080355	-0.0285327	0.0062177
RaceNon-Hispanic Black	0.0636655	0.0080643	7.8947173	0.0000000	0.0478491	0.0794819
RaceNon-Hispanic White	0.0052429	0.0074988	0.6991657	0.4845390	-0.0094643	0.0199501
RaceOther	0.0321289	0.0105983	3.0315074	0.0024682	0.0113426	0.0529153
RaceOther Hispanic	0.0054983	0.0095197	0.5775719	0.5636256	-0.0131725	0.0241691
Height	0.0075239	0.0008238	9.1334025	0.0000000	0.0059082	0.0091395
BMI_classObese	0.0345590	0.0057322	6.0288979	0.0000000	0.0233164	0.0458015
BMI_classOverweight	0.0197645	0.0056889	3.4742355	0.0005245	0.0086070	0.0309220
BMI_classUnderweight	-0.0514601	0.0165435	-3.1105958	0.0018965	-0.0839065	-0.0190136
PAL_GPAQLow	-0.0149848	0.0051889	-2.8878361	0.0039255	-0.0251617	-0.0048078
PAL_GPAQModerate	-0.0102444	0.0062910	-1.6284100	0.1036134	-0.0225829	0.0020941
EducationHigh school Graduate	-0.0141999	0.0067440	-2.1055626	0.0353809	-0.0274268	-0.0009730
EducationSome high school	0.0081200	0.0079670	1.0192066	0.3082422	-0.0075055	0.0237454
EducationUndergraduate or AA Degree	-0.0052933	0.0059017	-0.8969100	0.3698872	-0.0168681	0.0062816
Self_Diet	-0.0045172	0.0024272	-1.8611063	0.0628924	-0.0092776	0.0002432

From Table 7, the final model for males contains eight factors such as age, race, height, BMI classification, milk consumption, physical activity, ever used heroin, and educational level. Among men, holding all other factor constant:

- On average, the total bone mineral density of Non-Hispanic Black men is $0.0726 g/cm^2$ higher than Mexican American men. Men from other races, on average, have a total bone mineral density of $0.0622 g/cm^2$ higher compared to Mexican American men.
- An inch increase in height in men, on average, increases their total bone mineral density by $0.0070 g/cm^2$.
- On average, underweight men have $0.0980 g/cm^2$ less total bone mineral density compared to healthy weight men.
- Compared to healthy weight men, on average, overweight men have more total bone mineral density by $0.0255 g/cm^2$.
- Obese men have higher total bone mineral density than healthy weight men by $0.0325 g/cm^2$ on average.
- On average men who have low physical activity have less total bone mineral density than men who have high physical activity by $0.0209 g/cm^2$.

Table 7: Parameter Estimates for Male's model

term	estimate	std.error	statistic	p.value	lower CI	upper CI
(Intercept)	0.6775715	0.0846443	8.0049260	0.0000000	0.5114360	0.8437070
Age	-0.0004187	0.0002970	-1.4096836	0.1589976	-0.0010016	0.0001643
RaceNon-Hispanic Asian	-0.0174528	0.0123892	-1.4087172	0.1592832	-0.0417696	0.0068640
RaceNon-Hispanic Black	0.0726323	0.0120409	6.0321449	0.0000000	0.0489991	0.0962656
RaceNon-Hispanic White	0.0011269	0.0113337	0.0994304	0.9208199	-0.0211182	0.0233721
RaceOther	0.0622331	0.0155817	3.9939982	0.0000706	0.0316502	0.0928160
RaceOther Hispanic	0.0213015	0.0143540	1.4840041	0.1381772	-0.0068719	0.0494748
Height	0.0070200	0.0012226	5.7418039	0.0000000	0.0046203	0.0094197
BMI_classObese	0.0324600	0.0089625	3.6217493	0.0003099	0.0148688	0.0500511
BMI_classOverweight	0.0254509	0.0084645	3.0067720	0.0027178	0.0088372	0.0420646
BMI_classUnderweight	-0.0980622	0.0299654	-3.2725099	0.0011088	-0.1568768	-0.0392476
Milk_RegUseRegular Milk Drinker	0.0160745	0.0092871	1.7308473	0.0838406	-0.0021537	0.0343027
Milk_RegUseVaried	-0.0068381	0.0084382	-0.8103820	0.4179467	-0.0234001	0.0097239
PAL_GPAQLow	-0.0209046	0.0078831	-2.6518138	0.0081547	-0.0363773	-0.0054320
PAL_GPAQModerate	-0.0101518	0.0095043	-1.0681321	0.2857631	-0.0288064	0.0085027
Self_Diet	-0.0051989	0.0034308	-1.5153627	0.1300510	-0.0119326	0.0015349
EverHeroinUseYes	-0.0236463	0.0168573	-1.4027339	0.1610600	-0.0567329	0.0094403

From Table 8, the final model consists of the following variables: race, height, age, BMI classification, pack years, cocaine use, and the interaction term between BMI classification and pack years. Table **??** shows the parameter estimates from the fitted model. Based on the model, the following conclusions can be made. Among women, holding all other factors constant:

- On average, the total bone mineral density of Non-Hispanic Black women is $0.0491 g/cm^2$ higher than Mexican American women.
- On average, an inch increase in a woman's height increases total bone mineral density by $0.0082 g/cm^2$
- On average, a woman's total bone mineral density decreases by $0.0015 g/cm^2$ every year she gets older
- On average, an additional pack year decreases overweight women's total bone mineral density by $0.00029 g/cm^2$
- On average, the total bone mineral density of women who have ever used cocaine is $0.0189 g/cm^2$ higher than women who have never used the substance.

Table 8: Parameter Estimates for Female's model

term	estimate	std.error	statistic	p.value	lower CI	upper CI
(Intercept)	0.5979167	0.0694912	8.6042022	0.0000000	0.4615385	0.7342949
Age	-0.0014869	0.0002567	-5.7925447	0.0000000	-0.0019907	-0.0009831
RaceNon-Hispanic Asian	-0.0064837	0.0108629	-0.5968668	0.5507419	-0.0278023	0.0148349
RaceNon-Hispanic Black	0.0491428	0.0103471	4.7494191	0.0000024	0.0288363	0.0694493
RaceNon-Hispanic White	0.0038398	0.0094397	0.4067739	0.6842679	-0.0146859	0.0223656
RaceOther	-0.0039873	0.0142651	-0.2795160	0.7799112	-0.0319829	0.0240083
RaceOther Hispanic	-0.0147508	0.0123184	-1.1974606	0.2314328	-0.0389259	0.0094243
Height	0.0081920	0.0010826	7.5666045	0.0000000	0.0060673	0.0103167
BMI_classObese	0.0336638	0.0077166	4.3625337	0.0000143	0.0185199	0.0488078
BMI_classOverweight	0.0193678	0.0081717	2.3701195	0.0179859	0.0033307	0.0354049
BMI_classUnderweight	-0.0302828	0.0210253	-1.4403034	0.1501189	-0.0715454	0.0109798
Pack_Years	-0.0003659	0.0006594	-0.5548111	0.5791576	-0.0016600	0.0009283
EverCocaineUseYes	0.0189094	0.0091612	2.0640728	0.0392883	0.0009303	0.0368885
BMI_classObese:Pack_Years	0.0000925	0.0008346	0.1108650	0.9117474	-0.0015454	0.0017305
BMI_classOverweight:Pack_Years	-0.0025219	0.0010730	-2.3504301	0.0189603	-0.0046277	-0.0004162
BMI_classUnderweight:Pack_Years	-0.0013608	0.0022779	-0.5973795	0.5503998	-0.0058313	0.0031097

4 Discussion, Conclusions and Recommendation

4.1 Discussion

To determine sociodemographic and behavioral factors associated with total bone mineral density, exploratory data analysis was first performed on the reduced dataset with complete information. This was done to identify any trends between the regressors and total bone mineral density prior to modelling, to flag extreme values, and to explore possible interactions between regressors. Height exhibits an upward trend with BMD. Slight downward trends are observed for age and pack years. SEQN ID's 100521 and 96372 have been identified as observations with extreme values for pack years. Some categorical variables had levels with few observations, as stated in the data exploration. This might cause the interpretation of these categories to be biased. Interactions identified in the literature were also observed in the interaction plots.

Race, gender, height, physical activity, body mass index and educational attainment were found to significantly determine bone mineral density in the general model. This generally agrees with previously published literature, as gender has been postulated as one of the main factors affecting BMD, with its values being generally higher for men than for women, especially when age increases (Seeman, 2001). Similarly, in the case of race, previous studies have shown that people of color show a generally higher BMD than White or Asian individuals (Nam et al., 2013, Araujo et al., 2007). Physical activity as well has been confirmed by other authors as an important factor in increasing BMD (Langsetmo et al., 2012), although the relationship between its frequency and intensity and its effects on BMD are still not well known. In our case, the general model suggested that any level of physical activity below "high" (in the WHO classification) relates to a decrease in BMD with respect to individuals that do achieve this level. Lower educational attainment may indicate less knowledge of or a lack of access to a healthy and balanced diet, therefore contributing to lower total bone mineral density relative to those with higher educational attainment. Similar results have been found before, although in studies including only women (Sham et al., 2006).

With respect to height, our model pointed to it being a significant factor in defining BMD, slightly increasing it with every inch. This might be due to the fact that the bones of taller people are exposed to larger strains, especially during childhood and adolescence (when peak bone mass is developed), which results in a higher BMD later in life (Zemel, 2013). This is intimately related to BMI as well, as our general model implies a larger BMD for obese and overweight people with respect to people in a healthy weight, and this could also be explained by the larger strain posed on bones having an effect on bone cells that results in increased BMD

(Felson et al., 1993). On the other hand, underweight people are shown to tend to a lower BMD than people with a healthy weight, which is justifiable as well because a general lack of nutrients could be assumed in such a situation.

Subgroup analysis was approached in a similar fashion. Separate datasets were made for women (n=944) and men (n=870) from the reduced dataset. Exploratory data analysis of height, age, and pack years showed the same trends for both genders, albeit being slightly stronger in females across all variables.

In both separate models for each gender, race, height and BMI still proved to be significant in affecting BMI. However, in the model for men, milk consumption was also significant, suggesting that both men who never consumed milk regularly and those who did it only for some periods of time generally have a lower BMD than men who have always regularly consumed milk. On the other hand, the model for women showed an important effect of increasing age on BMD for them, as has been already widely presented by many researchers, successfully linking osteoporosis and general BMD loss to ageing and specially to post-menopausal age in women (Finkelstein et al., 2008). Lastly, a significant interaction was found between the BMI and pack years (smoking of cigarettes) in the case of women.

It must be noted that, reducing the size of our data by discarding observations with missing values, some bias might be introduced in the parameter estimates. Since we have not learned yet how to handle missing data, discarding them and considering only the complete cases was the solution. Possible other important factors on bone mineral density that are not in our dataset. Beside the selected 16 predictor variables in the dataset, there are other variables, represented for other elements still not considered. As all predictor variables in the final model are categorical variables, the three-factor interactions between more than two variables are not covered due to their complexity.

4.2 Conclusion

Various demographic and behavioral factors have been identified to significantly determine total bone mineral density in adults. Important factors were found to differ between males and females. Programs for improving bone health can use the results of the study to target specific lifestyles and identify individuals whose worsening conditions could be a precursor to fractures, osteoporosis, or other related diseases.

4.3 Recommendations

For future studies, the researchers recommend including genetic factors as a control as it was seen to affect total bone mineral density (Deng et al., 2000). Several factors in early life were found to be important based on the literature, such as milk consumption, which was indeed included in the study, but also others like peak bone mass. This value for bone mass is usually achieved during early adulthood, and affects the BMD of individuals during the rest of their lives, as BMD at any time passed this point is defined as the peak bone mass minus the bone loss happened afterwards (Kalkwarf et al., 2003; Rizzoli et al., 2001). Therefore, it would be of great help if such a value were to be included in the study. Sex-specific and sex-related factors such as menopausal state, pregnancy, and estrogen use should be included in the analyses to control for differences between genders.

References

Araujo, A. B., Travison, T. G., Harris, S. S., Holick, M. F., Turner, A. K., & McKinlay, J. B. (2007). Race/ethnic differences in bone mineral density in men. Osteoporosis International, 18(7), 943-953.

Belikov, A. V. (2019). Age-related diseases as vicious cycles. Ageing Research Reviews, 49, 11-26.

Bourne, D., Plinke, W., Hooker, E. R., & Nielson, C. M. (2017). Cannabis use and bone mineral density: NHANES 2007–2010. Archives of osteoporosis, 12(1), 29.

CDC. (2020, December 12). Defining Adult Overweight and Obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/adult/defining.html

Chastin, S. F., Mandrichenko, O., Helbostadt, J. L., & Skelton, D. A. (2014). Associations between objectively-measured sedentary behaviour and physical activity with bone mineral density in adults and older adults, the NHANES study. Bone, 64, 254-262.

Demeter, S., Leslie, W. D., Lix, L., MacWilliam, L., Finlayson, G. S., & Reed, M. (2007). The effect of socioeconomic status on bone density testing in a public health-care system. Osteoporosis international, 18(2), 153-158.

Deng, H., Chen, W., Conway, T., Zhou, Y., Davies, K. M., Stegman, M. R., . . . Recker, R. R. (2000). Determination of bone mineral density of the hip and spine in human pedigrees by genetic and life-style factors. Genetic Epidemiology, 19(2), 160-177. doi:10.1002/1098-2272(200009)19:23.0.co;2-h

Du, Y.; Zhao, L.-J.; Xu, Q.; Wu, K-H.; Deng, H.-W. (2017). Socioeconomic status and bone mineral density in adults by race/ethnicity and gender: the Louisiana osteoporosis study. Osteoporosis International, 28(5), 1699–1709. doi:10.1007/s00198-017-3951-1.

Eleftheriou, K. I., Rawal, J. S., James, L. E., Payne, J. R., Loosemore, M., Pennell, D. J., … & Sanders, J. (2013). Bone structure and geometry in young men: the influence of smoking, alcohol intake and physical activity. Bone, 52(1), 17-26. Felson, D. T., Zhang, Y., Hannan, M. T., & Anderson, J. J. (1993). Effects of weight and body mass index on bone mineral density in men and women: the Framingham study. Journal of Bone and Mineral Research, 8(5), 567-573.

Finkelstein, J. S., Brockwell, S. E., Mehta, V., Greendale, G. A., Sowers, M. R., Ettinger, B., … & Neer, R. M. (2008). Bone mineral density changes during the menopause transition in a multiethnic cohort of women. The Journal of Clinical Endocrinology & Metabolism, 93(3), 861-868.

Ghadimi, R., Hosseini, S. R., Asefi, S., Bijani, A., Heidari, B., & Babaei, M. (2018). Influence of smoking on bone mineral density in elderly men. International journal of preventive medicine, 9.

Husten, C. G. (2009). How should we define light or intermittent smoking? Does it matter?. Nicotine & Tobacco Research, 11(2), 111.

Jouanny, P., Jeandel, C., Pourel, J., Guillemin, F., & Kuntz, C. (1995). Environmental and genetic factors affecting bone mass similarity of bone density among members of healthy families. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 38(1), 61-67.

Kalkwarf, H. J., Khoury, J. C., & Lanphear, B. P. (2003). Milk intake during childhood and adolescence, adult bone density, and osteoporotic fractures in US women. The American Journal of Clinical Nutrition, 77(1), 257-265. doi:10.1093/ajcn/77.1.257

Kanis, J. A. (2002). Diagnosis of osteoporosis and assessment of fracture risk. The Lancet, 359(9321), 1929-1936. ISO 690

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. New York: McGrawHill Education.

Langsetmo, L., Hitchcock, C. L., Kingwell, E. J., Davison, K. S., Berger, C., Forsmo, S., … & Research, T. C. M. O. S. (2012). Physical activity, body mass index and bone mineral density—associations in a prospective

population-based cohort of women and men: The Canadian Multicentre Osteoporosis Study (CaMos). Bone, 50(1), 401-408.

Lim, J. U., Lee, J. H., Kim, J. S., Hwang, Y. I., Kim, T. H., Lim, S. Y., … & Rhee, C. K. (2017). Comparison of World Health Organization and Asia-Pacific body mass index classifications in COPD patients. International journal of chronic obstructive pulmonary disease, 12, 2465.

Morseth, B., Emaus, N., & Jørgensen, L. (2011). Physical activity and bone: The importance of the various mechanical stimuli for bone mineral density. A review. Norsk epidemiologi, 20(2).

Muir, J. M., Ye, C., Bhandari, M., Adachi, J. D., & Thabane, L. (2013). The effect of regular physical activity on bone mineral density in post-menopausal women aged 75 and over: a retrospective analysis from the Canadian multicentre osteoporosis study. BMC musculoskeletal disorders, 14(1), 253.

Muraki, S., Yamamoto, S., Ishibashi, H., Oka, H., Yoshimura, N., Kawaguchi, H., & Nakamura, K. (2007). Diet and lifestyle associated with increased bone mineral density: cross-sectional study of Japanese elderly women at an osteoporosis outpatient clinic. Journal of Orthopaedic Science, 12(4), 317-320.

Nam, H. S., Kweon, S. S., Choi, J. S., Zmuda, J. M., Leung, P. C., Lui, L. Y., … & Cauley, J. A. (2013). Racial/ethnic differences in bone mineral density among older women. Journal of bone and mineral metabolism, 31(2), 190-198.

Neter, J, M Kutner, C Nachtsheim, and W Wasserman.(1996). Applied Linear Statistical Models. Irwin Chicago

NHANES. (2020, December 10). NHANES Questionnaire Data. Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire.

Palermo, A., Tuccinardi, D., Defeudis, G., Watanabe, M., D'Onofrio, L., Lauria Pantano, A., … & Manfrini, S. (2016). BMI and BMD: the potential interplay between obesity and bone fragility. International journal of environmental research and public health, 13(6), 544.

Rudäng, R., Darelid, A., Nilsson, M., Nilsson, S., Mellström, D., Ohlsson, C., & Lorentzon, M. (2012). Smoking is associated with impaired bone mass development in young adult men: A 5-year longitudinal study. Journal of bone and mineral research, 27(10), 2189-2197.

Segheto, K. J., Juvanhol, L. L., Carvalho, C. J. D., Silva, D. C. G. D., Kakehasi, A. M., & Longo, G. Z. (2020). Factors associated with bone mineral content in adults: a population-based study. Einstein (São Paulo), 18.

Sophocleous, A., Robertson, R., Ferreira, N. B., McKenzie, J., Fraser, W. D., & Ralston, S. H. (2017). Heavy cannabis use is associated with low bone mineral density and an increased risk of fractures. The American journal of medicine, 130(2), 214-221. Ward, K. D., & Klesges, R. C. (2001). A meta-analysis of the effects of cigarette smoking on bone mineral density. Calcified tissue international, 68(5), 259-270.

Welten, D. C., Kemper, H. C. G., Post, G. B., Van Mechelen, W., Twisk, J., Lips, P., & Teule, G. J. (1994). Weight-bearing activity during youth is a more important factor for peak bone mass than calcium intake. Journal of Bone and Mineral Research, 9(7), 1089-1096.

World Health Organisation (2020,December 18) Life expectancy by Country and in the World. https: //www.worldometers.info/demographics/life-expectancy

Zemel, B. (2013). Bone mineral accretion and its relationship to growth, sexual maturation and body composition during childhood and adolescence. In Nutrition and Growth (Vol. 106, pp. 39-45). Karger Publishers.

Appendix

Table 9: Variable descripition.

	Variable	Codes/Value	Description
1	TOBMD	in g/cm^2	Total Bone Mineral Density
2	Age	in year	Age at the screening time
3	Gender	1=Male, 2=Female (ref)	Gender
4	Education	1=Less than Higher School (ref), 2= Higher School and some college degree, 3=College Graduate and above	Educational level of participant
5	Race	1=Mexican American, 2=Other Hispanic, 3=Non-Hispanic White, 4=Non-Hispanic Black, 5=Non-Hispanic Asian, 6=Other race or multi race	Race or ethincity
6	Marital Status	1=Married, 2=Never married, 3=Widowed, 4=Divorced	Marital Status of participant
7	Ratio	1=Lower Class (ref), 2=Middle Class, 3=Upper-Middle Class, 4=Upper Class	Ratio of family income to poverty
8	Healthy Diet Index	1=Excellent, 2=Vey Good, 3=Good, 4=Fair, 5=Poor (ref)	How healthy is the diet?
9	Regular Milk Drinking	1=Yes, 2=No (ref), 3=Sometimes(Varied)	Regular milk use 5 times per week
10	Marijuana	1=Yes, 2=No (ref)	Regular uses of marijuana or hashish
11	Heroin	1=Yes, 2=No (ref)	Have you ever, even once, used heroin?
12	Cocaine	1=Yes, 2=No(ref)	Have you ever use any form of cocaine
13	Methamphetamine	1=Yes, 2=No(ref)	Have you ever, even once, used methamphetamine?
14	PAL	1=High (ref), 2=Moderate, 3=Low	Physical activity level based on GPAQ
15	Pack Year	count	Number of pack
16	Height	in inches	Current self-reported height
17	Weight	in pound	Current self-reported weight
18	BMI	1 = Normal (ref), 2 = Obese, 3 = Overweight, 4 = Underweight	Body mass index

Table 10: Missing observation of variables

	Variable	Count	Percentage %
1	Annual Family Income	253	11.20
2	Milk consumption during 5-12	618	27.40
3	Milk consumption during 5-12	618	27.40
4	Milk consumption during 5-12	618	27.40
5	Fast food consumption during past 30 days	1	0.04
6	Marijuana	137	6.07
7	cocaine	137	6.07
8	Heroine	137	6.07
9	Methamphetamine	138	6.07
10	Years of smoking	2	0.09
11	BMI	96	3.06

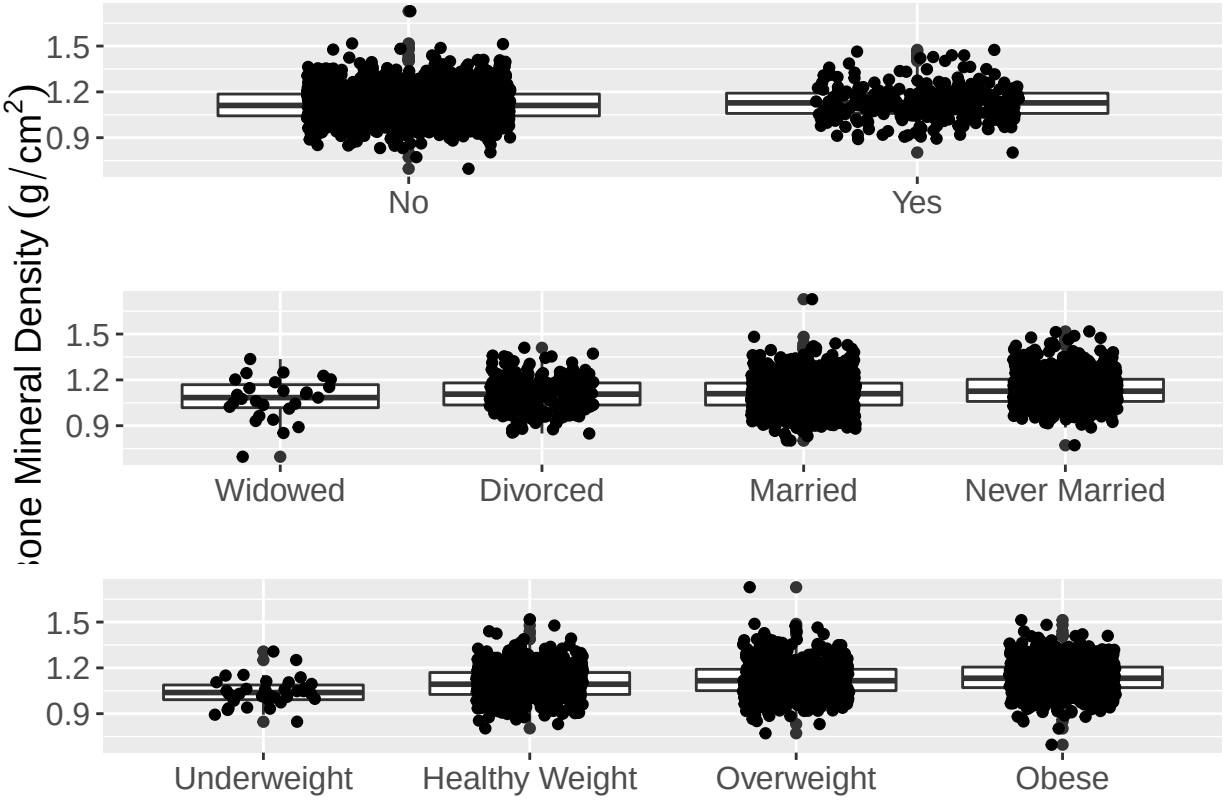

Figure 4: Boxplot for Cocaine use, Marital status and BMI

Table 11: Selected Interaction Item

BMI:Income	Drug Use:Poverty	Gender: Physical activity
Gender: Smoking	Marihuana: BMI	Marihuana: Milk Use
Marihuana: Other drug use	Smoking: Physical activity	Drug use : Education
Age: Physical activity	Gender: Age	Gender: Race
Gender: Poverty ratio	Physical activity: Milk Use	Diet: Milk Use
Race: Poverty ratio	Smoking: BMI	

R code for data manipulation

```
setwd()

library("SASxport")
library("Hmisc")
library("MASS")
library("tidyverse")
library("skimr")
library("sjlabelled")
library("data.table")
library("broom")
library("car")
library("scales")
library("ggrepel")
library("cowplot")

DXX_JData <- read.xport("DXX_J.XPT")
DEMO_JData <- read.xport("DEMO_J.XPT")
DIET_JData <- read.xport("DBQ_J.XPT")
DRUG_JData <- read.xport("DUQ_J.XPT")
ACTIVITY_JData <- read.xport("PAQ_J.XPT")
SMOKE_JData <- read.xport("SMQ_J.XPT")
WEIGHT_JData <- read.xport("WHQ_J.XPT")
totaldataset <- Reduce(function(x, y) merge(x, y, all=TRUE,by="SEQN"),
                       list(DXX_JData,DEMO_JData,DIET_JData,DRUG_JData,
                            ACTIVITY_JData,SMOKE_JData,WEIGHT_JData))

prelim_data <- totaldataset[!is.na(totaldataset$DXDTOBMD), ]
prelim_findata <- remove_all_labels(prelim_data)
remove(list=ls(pattern="_JData$"),totaldataset,prelim_data)

##### Full Data Set based on Selected Variables#####

### Demographics

DEMO.data<- prelim_findata %>% filter(RIDAGEYR>=20)
%>% mutate(Gender=case_when(RIAGENDR==1~"Male",
                            RIAGENDR==2~"Female"),
           Age=RIDAGEYR,
           Race=case_when(RIDRETH3==1~"Mexican American",
 RIDRETH3==2~"Other Hispanic",
 RIDRETH3==3~"Non-Hispanic White",
```

```
 RIDRETH3==4~"Non-Hispanic Black",
 RIDRETH3==6~"Non-Hispanic Asian",
 RIDRETH3==7~"Other"),
 Education=case_when(DMDEDUC2==1~"Some high school",
 DMDEDUC2==2~"Some high school",
 DMDEDUC2==3~"High school Graduate",
 DMDEDUC2==4~"Undergraduate or AA Degree",
 DMDEDUC2==5~"College graduate and Above"),
 MaritalStatus=case_when(DMDMARTL==1~"Married",
                         DMDMARTL==2~"Widowed",
                        DMDMARTL==3~"Divorced",
                        DMDMARTL==4~"Divorced",
                        DMDMARTL==5~"Never Married",
                        DMDMARTL==6~"Never Married"),
FamilySize=DMDFMSIZ,
AnnualFamilyIncome=case_when(INDFMIN2==77|INDFMIN2==99 |INDFMIN2==12|
                             INDFMIN2==13~NA_real_,
                            INDFMIN2==14~12,
                            INDFMIN2==15~13,
                            TRUE~as.numeric(INDFMIN2)),
Poverty_Ratio_cat = case_when(INDFMPIR<1.00~"Lower Class",
          between(INDFMPIR,1.00,1.99)~"Middle Class",
          between(INDFMPIR,2.00,3.99)~"Upper Middle Class",
          between(INDFMPIR,4.00,5.00)~"Upper Class"
                                                    ))%>%
  select(SEQN,DXDTOBMD,Gender,Age,Education,MaritalStatus,Race,Poverty_Ratio_cat) %>%
  mutate_at(vars(Gender,Education,MaritalStatus,Race,Poverty_Ratio_cat),factor)

### Drug Use

DRUG.data <- prelim_findata %>% filter(RIDAGEYR>=20) %>%
  mutate(RegWeedUse = case_when(DUQ200==2&is.na(DUQ211)~2,
                                TRUE~as.numeric(DUQ211)),
         EverCocaineUse = case_when(DUQ240==2&is.na(DUQ250)~2,
                                    TRUE~as.numeric(DUQ250)),
         EverHeroinUse = case_when(DUQ240==2&is.na(DUQ290)~2,
                                   TRUE~as.numeric(DUQ290)),
         EverMethUse = case_when(DUQ240==2&is.na(DUQ330)~2,
                                 DUQ330==9~NA_real_,
                                 TRUE~as.numeric(DUQ330))) %>%
  select(SEQN,DXDTOBMD,RegWeedUse,EverCocaineUse,EverHeroinUse,EverMethUse) %>%
  mutate_at(vars(-SEQN,-DXDTOBMD),factor,levels=c(1:2),labels=c("Yes","No"))

### Physical Activity

PAQ.data <- prelim_findata %>% filter(RIDAGEYR>=20) %>%
  mutate(PAQ605=case_when(PAQ605 == 9 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ605)),
         PAQ610=case_when(PAQ605 == 2 & is.na(PAQ610)~0,
                          PAQ610 == 99 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ610)),
         PAD615=case_when(PAQ605 == 2 & is.na(PAD615) ~ 0,
                          PAD615 == 9999 ~ NA_real_,
```

```
                          TRUE ~ as.numeric(PAD615)),
         PAQ620=case_when(PAQ620 == 9 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ620)),
         PAQ625=case_when(PAQ620 == 2 & is.na(PAQ625)~0,
                          PAQ625 == 99 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ625)),
         PAD630=case_when(PAQ620 == 2 & is.na(PAD630) ~ 0,
                          PAD630 == 9999 ~ NA_real_,
                          TRUE ~ as.numeric(PAD630)),
         PAQ635=case_when(PAQ635 == 9 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ635)),
         PAQ640=case_when(PAQ635 == 2 & is.na(PAQ640)~0,
                          PAQ640 == 99 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ640)),
         PAD645=case_when(PAQ635 == 2 & is.na(PAD645) ~ 0,
                          PAD645 == 9999 ~ NA_real_,
                          TRUE ~ as.numeric(PAD645)),
         PAQ650=case_when(PAQ650 == 9 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ650)),
         PAQ655=case_when(PAQ650 == 2 & is.na(PAQ655)~0,
                          PAQ655 == 99 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ655)),
         PAD660=case_when(PAQ650 == 2 & is.na(PAD660) ~ 0,
                          PAD660 == 9999 ~ NA_real_,
                          TRUE ~ as.numeric(PAD660)),
         PAQ665=case_when(PAQ665 == 9 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ665)),
         PAQ670=case_when(PAQ665 == 2 & is.na(PAQ670)~0,
                          PAQ670 == 99 ~ NA_real_,
                          TRUE ~ as.numeric(PAQ670)),
         PAD675=case_when(PAQ665 == 2 & is.na(PAD675) ~ 0,
                          PAD675 == 9999 ~ NA_real_,
                          TRUE ~ as.numeric(PAD675)),
         MET_minutes=(PAQ610 * PAD615 * 8) + (PAQ625 * PAD630 * 4)+
                       (PAQ640 * PAD645 * 4) + (PAQ655 * PAD660 * 8) +
           (PAQ670 * PAD675 * 4),
         total_active = PAD615+PAD630+PAD645+PAD660+PAD675,
         PAL_GPAQ = case_when((MET_minutes>=1500&(PAQ610+PAQ655)>=3)|
                                MET_minutes>=3000&
                                (PAQ610+PAQ625+PAQ640+PAQ655+PAQ670)>=7~"High",
                              (PAQ610+PAQ655)>=3&(PAQ610*PAD615+PAQ655*PAD660)>=60|
                                (PAQ625+PAQ640+PAQ670)>=5&
                                (PAQ625*PAD630+PAQ640*PAD645+PAQ670*PAD675)>=150|
                                MET_minutes>=600&(PAQ610+PAQ625+PAQ640+PAQ655+PAQ670)>=5~"Moderate",
                              is.na(MET_minutes)~NA_character_,
                              TRUE~"Low")) %>%
  select(SEQN,DXDTOBMD,PAL_GPAQ,total_active) %>% mutate_at(vars(PAL_GPAQ),factor)

### Smoking - Cigarette Use

SCU.data <- prelim_findata %>% filter(RIDAGEYR>=20) %>%
  mutate(SMD030 = case_when(SMD030==999~NA_real_,
                            TRUE~as.numeric(SMD030)),
```

```
         SMQ040 = case_when(SMQ020==2&is.na(SMQ040)~3,
                            TRUE~as.numeric(SMQ040)),
         SMD057 = case_when(SMQ020 == 2 ~ 0,
                            SMD057 == 999 ~ NA_real_,
                            TRUE ~ as.numeric(SMD057)),
         Smoker = case_when(SMQ020==2~"Non-smoker",
                            SMQ020==1 & SMQ040==3~"Former Smoker",
                            SMQ020==1 & SMQ040!=3~"Current Smoker"),
         YSS = case_when(SMQ020==2~0,
                         SMQ050U==1~SMQ050Q/365,
                         SMQ050U==2~SMQ050Q/52.143,
                         SMQ050U==3~as.double(SMQ050Q/12),
                         SMQ050U==4~as.double(SMQ050Q)),
         Years_Smoking = case_when(Smoker=="Non-smoker"~as.double(0),
                                   Smoker=="Former Smoker"~as.double(RIDAGEYR-YSS-SMD030),
                                   Smoker=="Current Smoker"~as.double(RIDAGEYR-SMD030)),
         Pack_Years = case_when(SMQ020 == 2~0,
                                SMQ040 == 1 | SMQ040 == 2~Years_Smoking*SMD650/20,
                                SMQ040 == 3~Years_Smoking*SMD057/20))%>%
  select(SEQN,DXDTOBMD,Pack_Years,Years_Smoking)


### Weight History
WH.data <- prelim_findata %>% filter(RIDAGEYR>=20) %>%
  mutate(WHD010 = case_when(WHD010==9999~NA_real_,
                            TRUE~as.numeric(WHD010)),
         WHD020 = case_when(WHD020==9999~NA_real_,
                            WHD020==7777~NA_real_,
                            TRUE~as.numeric(WHD020)),
         BMI = round(((WHD020/2.205)/(WHD010/39.37)^2),1),
         BMI_class=case_when(BMI<18.5~"Underweight",
                             BMI>=18.5 & BMI<=24.9~"Healthy Weight",
                             BMI>=25.0 & BMI<=29.9~"Overweight",
                             BMI>=30.0~"Obese",
                             RIDRETH3==6&BMI<17.50~"Underweight",
                             RIDRETH3==6&between(BMI,17.50,22.99)~"Healthy Weight",
                             RIDRETH3==6&between(BMI,23.00,27.99)~"Overweight",
                             RIDRETH3==6&BMI>28.00~"Obese")) %>%
  mutate(BMI_class = factor(BMI_class,levels=c("Underweight","Healthy Weight","Overweight","Obese"))) %
  select(SEQN,DXDTOBMD,WHD010,BMI_class) %>% rename(Height=WHD010)

### Diet Behavior & Nutrition
DBN.data <- prelim_findata %>% filter(RIDAGEYR>=20) %>%
  mutate(Milk_RegUse=case_when(DBQ229==9~NA_real_,
                               TRUE~as.numeric(DBQ229))) %>%
  select(SEQN,DXDTOBMD,Milk_RegUse,DBQ700) %>% rename(Self_Diet=DBQ700) %>%
  mutate_at(vars(Milk_RegUse),factor,levels=c(1:3),
            labels=c("Regular Milk Drinker","Never Drank Milk Regularly","Varied")) %>%
  mutate_at(vars(Self_Diet),factor)


### Combining datasets
datasets<-list(DEMO.data,DRUG.data,PAQ.data,SCU.data,WH.data,DBN.data)
```

```r
data_full<-Reduce(function(x, y) merge(x, y, all=TRUE,by=c("SEQN","DXDTOBMD")), datasets) %>%
  filter(is.na(total_active)|total_active<1440) %>% filter(is.na(Years_Smoking)|Years_Smoking>=0)

### Complete Data
data_complete<-data_full %>% select(-total_active,-Years_Smoking) %>% drop_na()
remove(list=ls(pattern=".data$"))
```