

Exploring Missing Data Imputation with ChatGPT

Table of Contents

1 Introduction	1
2 Data	2
2.1 Exploratory Analysis	3
3 Imputation Methods	6
3.1 K-Nearest Neighbor	6
3.2 Probabilistic Principal Component Analysis	6
4 Results	7
4.1 BMI - Continuous	7
4.2 BMICatUnder20yrs - Categorical	9
5 Conclusion	13
Sources	14

1 Introduction

The purpose of this project was to investigate the understanding of AI ChatGPT to both apply and explain imputing missing data, primarily in a healthcare setting. Data was utilized from the National Health and Nutrition Examination Survey (NHANES) with the NHANES library in R (1999 - 2004). This dataset contains a large amount of physical, medical, demographic, and lifestyle variables. Four imputation methods were compared; Mean Imputation, K-Nearest Neighbor (KNN), Probabilistic Principal Component Analysis (PPCA), and MICE.

Previously, I worked as a GA in the GVSU Statistical Consulting Center from 2023-2024 on a wide variety of projects. For the entire school year, I had the opportunity to work on a research study with Dr. Nicholas Lerma whose main question to answer was “can children ‘earn’ their screen time?” This research was never published, partially because there were many issues drawing inferences from the analyses attempted, which has bothered me for a while. I am using this past knowledge of the data to further attempt an analysis to answer the question while also using ChatGPT to assist.

Another reason for choosing this goal of analysis is to avoid the variables and models we previously did in class, but I felt it was interesting enough to note that the NHANES survey *does* record adolescent alcohol use, smoking, sexual behavior, and reproductive health, even though it is not in the dataset. This is because of confidentiality concerns, and those files containing the full information can be requested from the NCHS Research Data Center. This made me question how we can still assume any sort of missingness in the data, if it really isn’t missing at the end of

the day, but it would be fun to check the imputed values for these variables with the actual, true observations. Perhaps it can be done at another time!

2 Data

After data was read into R and cleaned, a subset was created only for those respondents aged 12-19. Originally, the dataset was going to include those ages 2-19, but several of the variables of interest only collected data for respondents aged 12 years and older. This resulted in 1,127 observations.

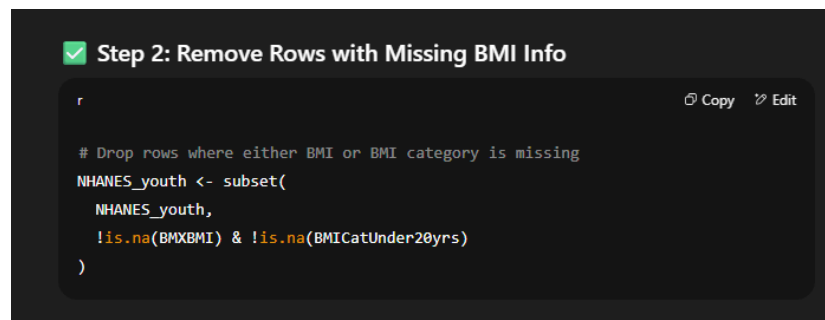
BMI was to be modeled as both a numerical and categorical response using linear regression (BMI) and ordinal logistic regression (BMICatUnder20yrs). Due to time constraints, only eight variables were considered for modeling, which were chosen based on previous knowledge of the relationships between predictors and response variables.

Variable	Type	Description
BMI	Numeric	Body mass index (weight/height ² in kg/m ²). Reported for participants aged 2 years or older
BMICatUnder20yrs	Ordered factor	Body mass index category. Reported for participants aged 2 to 19 years. One of UnderWeight (BMI < 5th percentile) NormWeight (BMI 5th to < 85th percentile), OverWeight (BMI 85th to < 95th percentile), Obese (BMI >= 95th percentile)
Age	Numeric	Age in years at screening of study participant
Gender	Categorical	Gender (sex) of study participant coded as male or female
Race1	Categorical	Reported race of study participant: Mexican, Hispanic, White, Black, or Other
TVHrsDay	Ordered factor	Number of hours per day on average participant watched TV over the past 30 days. Reported for participants 2 years or older. One of 0_to_1hr, 1_hr, 2_hr, 3_hr, 4_hr, More_4_hr. Not available 2009-2010
CompHrsDay	Ordered Factor	Number of hours per day on average participant used a computer or gaming device over the past 30 days. Reported for participants 2 years or older. One of 0_hrs, 0_to_1hr, 1_hr, 2_hr, 3_hr, 4_hr, More_4_hr. Not available 2009-2010.
PhysActive	Categorical	Participant does moderate or vigorous-intensity

		sports, fitness or recreational activities (Yes or No). Reported for participants 12 years or older.
PhysActiveDays	Numeric	Number of days in a typical week that participant does moderate or vigorous-intensity activity. Reported for participants 12 years or older
Poverty	Numeric	A ratio of family income to poverty guidelines. Smaller numbers indicate more poverty

2.1 Exploratory Analysis

A preliminary investigation of missingness was done using the VIM package in R. This was done to provide evidence as to the plausibility of the MAR assumption before doing formal testing and was not something that ChatGPT had suggested; in fact, it did not suggest I check what type of missingness the data was before starting imputation at all. It even suggested that I delete all rows of missing data even though it knew I was going to be comparing imputation methods.

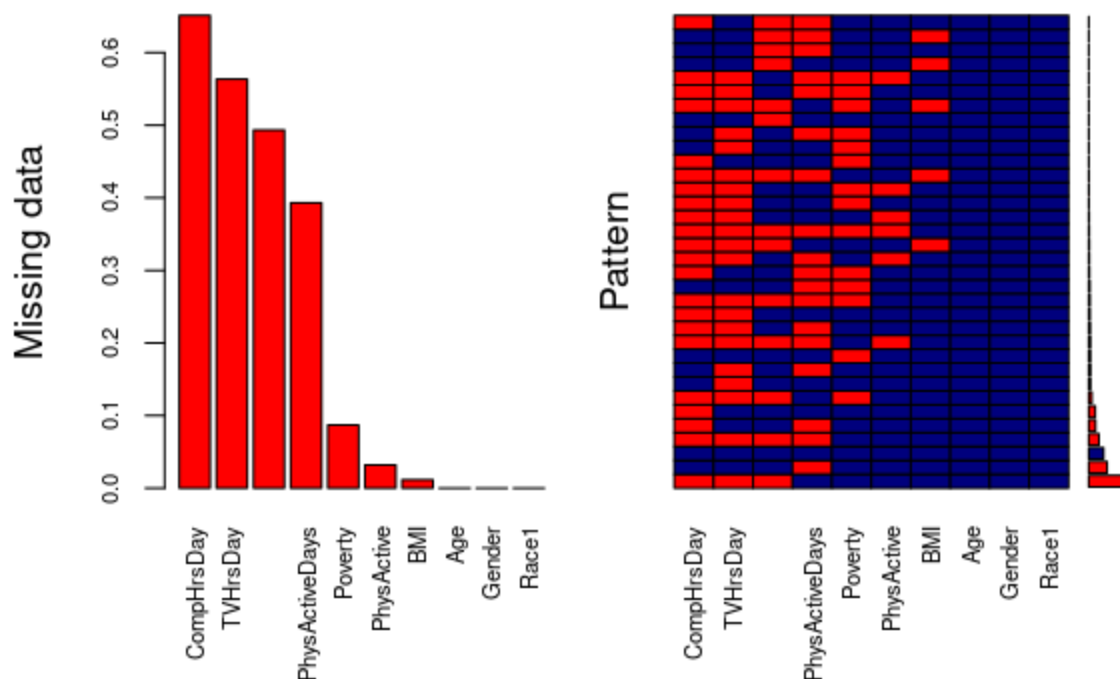


```

r                                                                    Copy Edit
# Drop rows where either BMI or BMI category is missing
NHANES_youth <- subset(
  NHANES_youth,
  !is.na(BMXBMI) & !is.na(BMICatUnder20yrs)
)

```

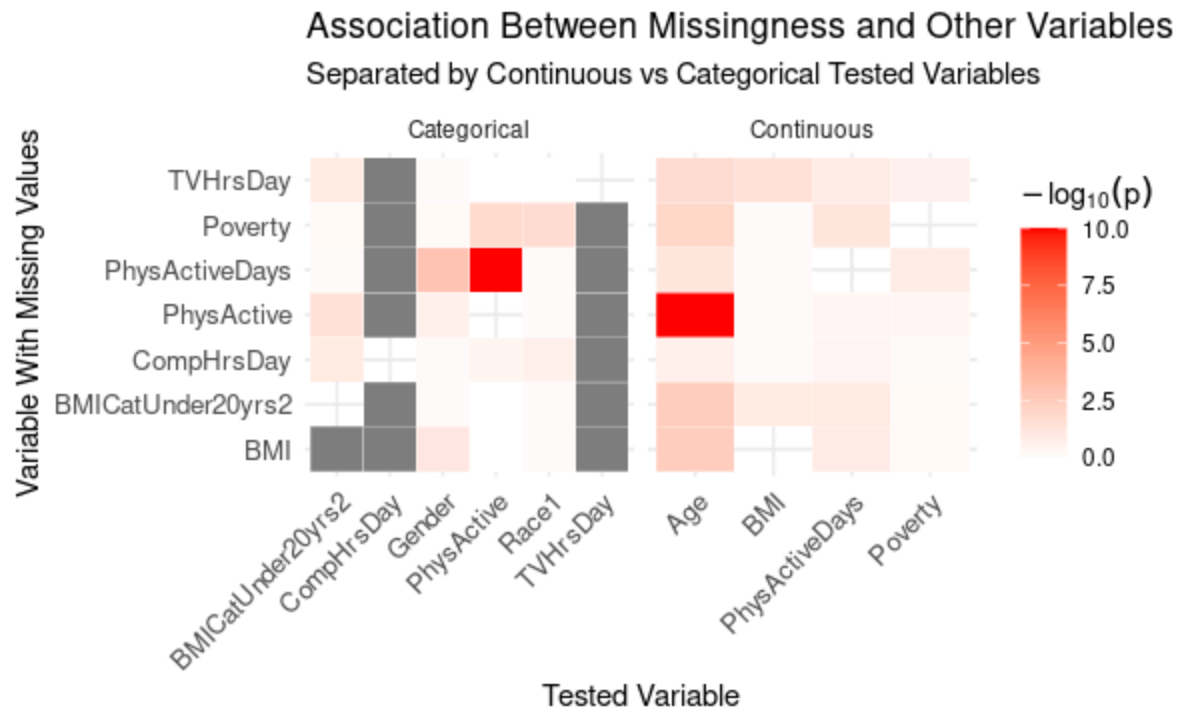
In the figure below, there are a large amount of missing observations for many of the variables of interest.



Further, a function was built in R to compare (test) missing with non-missing data using Chi-Square for categorical variables and t-test for numeric variables. If those with and without missing data differ on many observed variables, there is a possibility that they also differ on unobserved variables.

After p-values were obtained, a heat map was then generated for further visualization and quick identification of variables likely to be missing at random (MAR) since they correlate with other observed data. It's worth noting that insignificant results do not provide proof of data being MCAR or MAR.

In the following figure, rows represent variables that have missing values and columns represent variables they were tested against. The color represents the strength of the association with red meaning a very strong (significant) relationship (low p-value); the darker the red, the stronger the relationship.



A formal testing to see if the data was MCAR was then performed with the assistance of ChatGPT, which suggested doing Hawkin's Test of Normality as well as a nonparametric test. Through verification through various papers I looked at on my own, Hawkin's Test is equivalent to Little's MCAR, and the nonparametric test is legitimate as well. The results can be viewed below.

Test of normality and Homoscedasticity:

Hawkins Test:

P-value for the Hawkins test of normality and homoscedasticity:
3.166248e-13

Either the test of multivariate normality or homoscedasticity (or both) is rejected.
Provided that normality can be assumed, the hypothesis of MCAR is rejected at 0.05 significance level.

Non-Parametric Test:

P-value for the non-parametric test of homoscedasticity:
0.8186782

Reject Normality at 0.05 significance level.
There is not sufficient evidence to reject MCAR at 0.05 significance level.

Due to BMI not being a normally distributed variable anyways (i.e. it's normally right skewed since there are more overweight/obese people in America than there are underweight,

which is also true in this dataset), the non-parametric test seemed more legitimate. Non-parametric forms of imputation include K-Nearest Neighbors (KNN), Random Forests, and Kernel methods. For the sake of my own curiosity, as well as to have something to discuss for the presentation, I went ahead with all four imputation methods mentioned previously to compare.

3 Imputation Methods

3.1 K-Nearest Neighbor

KNN imputation is a non-parametric method for handling missing data based on the idea of measuring similarity with some distance metric between observations. It replaces a missing value with one computed from the k most similar (nearest) observations in the dataset that do not have missing values for that feature. The similarity between observations using the VIM package uses Gower Distance, which is used for mixed data types.

Gower Distance is between 0 and 1, inclusive. For numeric variables, it's computed using the absolute difference between observations and dividing by the range for that specific feature. For categorical variables, 0 is assigned if the observations are the same (i.e. comparing Female to Female as observations 1 and 2), and 1 is assigned if the observations are as far apart as possible relative to the data (i.e. Female compared to Male). In general, Gower Distance that is closer to 0 means the observations are more similar (nearer); Gower Distance closer to 1 means the observations are as different (far) as possible.

Some better properties of KNN imputation is that it has no assumptions about the underlying data and uses the local structure, and it can naturally adapt to nonlinear relationships. Several drawbacks include the choice of k (number of neighbors) and distance metric making a difference in your outcomes, may perform poorly if there are too many missing observations, doesn't model uncertainty, and is sensitive to scaling.

3.2 Probabilistic Principal Component Analysis

PCA in general is commonly used for dimension reduction. Its goal is to reduce dimensionality of a multivariate dataset while still accounting for as much variation in the original data as possible. A new set of variables, principal components, are formed, which are linear combinations of the original variables in the dataset. These are uncorrelated variables in order such that the first principal component accounts for as much variation in the data as possible, the second accounts for as much of the leftover variation as possible, etc.

When I asked ChatGPT to explain how PCA is used for missing data imputation, it neglected to mention anything about preserving variation. It also neglected to mention what specific form of PCA imputation it suggested for me to try (there are several), and it failed to mention that PCA is for numerical data only. I am partially to blame for this, since I know what PCA is, but it didn't occur to me until coding the data analysis and receiving several errors due to the factors included in my model. ChatGPT and several other sources stated that factors can be

converted to numerical variables for the purpose of PCA, but there is something about this that doesn't seem ethical to me, and I'm not sure how it would be interpreted if the factors were not in order.

Moving on, the form of PCA imputation that was suggested turned out to be PPCA (refer to the title of the section for the full name, I am so tired of typing it), which is sometimes referred to as EM-PCA. It assumes the data can be generated by projecting latent variables into a high-dimensional space, which are estimated using MLE and EM-Algorithm. It assumes that data is either MCAR or Mar.

A few benefits of using PPCA is that it handles multivariate correlation well, improves accuracy over simple imputations, and is useful when data lies in a low-dimensional space. Some drawbacks include it not handling nonlinear structures unless extended (ex. Kernel methods), not suitable if data is NMAR, is sensitive to the number of principal components, and is less effective if there is lots of missingness.

4 Results

4.1 BMI - Continuous

After completing all four types of imputation and fitting models for each, the MICE method showed to be the best out of all four methods when comparing Adjusted R-Squared and RMSE for each (ChatGPT offered code for that), though they did not differ by a dramatic amount. To me, this is surprising considering KNN is the only nonparametric method involved.

Method <chr>	Adj_R2 <dbl>	RMSE <dbl>
Mean	0.07600384	5.735644
KNN	0.08414312	5.848150
PCA	0.08563711	5.737052
MICE	0.09435668	5.702046

When comparing the output for each model, *age* and *poverty* were consistently significant amongst all imputation methods. At times, *RaceOther* was significant at a 0.05 level, as well as various types of *CompHrsDay*. Full output for each is shown below.

MICE

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	18.46493174	1.71571113	10.7622615	136.354720	6.207404e-20
2	Age	0.48306998	0.08843597	5.4623698	135.323345	2.180900e-07
3	Gendermale	-0.33760683	0.37851206	-0.8919315	188.967372	3.735640e-01
4	Race1Hispanic	-0.74835178	0.82958269	-0.9020822	960.499660	3.672394e-01
5	Race1Mexican	-0.11775381	0.71938442	-0.1636869	223.308881	8.701257e-01
6	Race1Other	-1.90116901	0.76216669	-2.4944268	1058.693018	1.276764e-02
7	Race1White	-0.50426956	0.61360209	-0.8218185	204.056804	4.121394e-01
8	TVHrsDay.L	0.62101392	1.29566811	0.4793002	9.800623	6.422401e-01
9	TVHrsDay.Q	1.66880608	1.08796891	1.5338729	12.271737	1.504296e-01
10	TVHrsDay.C	-0.10516020	0.75322303	-0.1396136	29.812252	8.899042e-01
11	TVHrsDay^4	0.88143114	0.57052797	1.5449394	66.544730	1.271036e-01
12	TVHrsDay^5	0.05080128	0.47008353	0.1080686	33.017387	9.145953e-01
13	CompHrsDay.L	1.02806050	0.91248560	1.1266594	18.793366	2.740676e-01
14	CompHrsDay.Q	1.06489944	1.04562735	1.0184311	7.302093	3.410240e-01
15	CompHrsDay.C	1.31708342	0.91160352	1.4447985	23.187648	1.618934e-01
16	CompHrsDay^4	1.03932029	0.99328768	1.0463437	16.553003	3.104419e-01
17	CompHrsDay^5	0.15174306	0.51910824	0.2923149	1006.376036	7.701061e-01
18	PhysActiveYes	0.29212273	0.48267218	0.6052197	158.651050	5.458977e-01
19	PhysActiveDays	-0.08830488	0.13838867	-0.6380933	14.373700	5.334477e-01
20	Poverty	-0.35127467	0.12297745	-2.8564154	163.718685	4.840383e-03

KNN

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	18.49451	1.67138	11.065	< 2e-16	***
Age	0.55932	0.08413	6.649	4.64e-11	***
Gendermale	-0.61458	0.38733	-1.587	0.11286	
Race1Hispanic	-1.10597	0.84566	-1.308	0.19121	
Race1Mexican	-0.72236	0.69058	-1.046	0.29578	
Race1Other	-2.32073	0.77279	-3.003	0.00273	**
Race1White	-1.14758	0.59204	-1.938	0.05283	.
TVHrsDay.L	-1.72217	1.28111	-1.344	0.17913	
TVHrsDay.Q	0.81892	1.16967	0.700	0.48399	
TVHrsDay.C	-0.72867	0.88589	-0.823	0.41095	
TVHrsDay^4	0.80112	0.61665	1.299	0.19416	
TVHrsDay^5	-0.13774	0.41022	-0.336	0.73710	
CompHrsDay.L	2.55648	0.89401	2.860	0.00432	**
CompHrsDay.Q	1.51941	0.70230	2.163	0.03072	*
CompHrsDay.C	1.62005	0.85810	1.888	0.05929	.
CompHrsDay^4	1.58359	0.83992	1.885	0.05964	.
CompHrsDay^5	0.64859	0.58606	1.107	0.26866	
PhysActiveYes	-0.06326	0.46975	-0.135	0.89290	
PhysActiveDays	-0.07114	0.10808	-0.658	0.51054	
Poverty	-0.37868	0.11644	-3.252	0.00118	**

PPCA

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12.55971	2.02158	6.213	7.33e-10	***
Age	0.50687	0.08117	6.244	6.03e-10	***
Gender	-0.20461	0.34890	-0.586	0.5577	
Race1	-0.14242	0.12681	-1.123	0.2616	
TVHrsDay	0.21769	0.21535	1.011	0.3123	
CompHrsDay	0.86971	0.19813	4.390	1.24e-05	***
PhysActive	0.24355	0.44467	0.548	0.5840	
PhysActiveDays	0.12833	0.12142	1.057	0.2908	
Poverty	-0.32681	0.11508	-2.840	0.0046	**

Mean

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.86652	1.65350	10.805	< 2e-16	***
Age	0.55184	0.08054	6.852	1.21e-11	***
Gendermale	-0.28069	0.35360	-0.794	0.42748	
Race1Hispanic	-0.97274	0.82221	-1.183	0.23703	
Race1Mexican	-0.63909	0.66693	-0.958	0.33814	
Race1Other	-2.24434	0.75876	-2.958	0.00316	**
Race1White	-1.17774	0.55647	-2.116	0.03453	*
TVHrsDay.L	-1.42872	1.32582	-1.078	0.28144	
TVHrsDay.Q	1.01762	1.20629	0.844	0.39907	
TVHrsDay.C	-1.03766	0.98885	-1.049	0.29424	
TVHrsDay^4	0.46679	0.77527	0.602	0.54723	
TVHrsDay^5	-0.83247	0.52687	-1.580	0.11438	
CompHrsDay.L	2.61433	1.07588	2.430	0.01526	*
CompHrsDay.Q	2.19882	0.88856	2.475	0.01349	*
CompHrsDay.C	2.58230	1.11702	2.312	0.02097	*
CompHrsDay^4	1.62768	1.12566	1.446	0.14847	
CompHrsDay^5	0.99209	0.87301	1.136	0.25604	
PhysActiveYes	0.24016	0.44920	0.535	0.59301	
PhysActiveDays	-0.07351	0.12104	-0.607	0.54376	
Poverty	-0.32872	0.11653	-2.821	0.00487	**

4.2 BMICatUnder20yrs - Categorical

After completing all four types of imputation and fitting models for each, the KNN method showed to be the best out of all four methods when comparing McFadden's R-Squared and RMSE for each (ChatGPT again offered code for that). This makes sense considering we are assuming nonparametric models are more suitable for this analysis.

Method <chr>	AIC <dbl>	McFadden_R2 <dbl>
mean	2013.385	0.03489192
knn	1999.527	0.04168288
pca	2006.300	0.02758257
mice	2381.059	0.03648793

When comparing the output for each model, *age*, *gender*, *poverty*, and *CompHrsDay* were consistently significant amongst all imputation methods. Full output for each is shown below, and a discussion for interpretation follows.

MICE

	Value	Std. Error	t value	p value
Age	-0.101487744	0.02862149	-3.54585853	3.913362e-04
Gendermale	0.228374805	0.12233378	1.86681715	6.192713e-02
Race1Hispanic	0.020044479	0.27351254	0.07328541	9.415790e-01
Race1Mexican	0.146778755	0.22518788	0.65180576	5.145265e-01
Race1Other	0.055962135	0.25650349	0.21817300	8.272943e-01
Race1White	0.062755749	0.19364485	0.32407652	7.458801e-01
TVHrsDay.L	0.308225868	0.37883420	0.81361680	4.158645e-01
TVHrsDay.Q	0.467851460	0.25914358	1.80537544	7.101595e-02
TVHrsDay.C	-0.318473045	0.33336784	-0.95532023	3.394158e-01
TVHrsDay^4	0.166076287	0.35420748	0.46886725	6.391645e-01
TVHrsDay^5	-0.025406168	0.31848894	-0.07977096	9.364194e-01
CompHrsDay.L	-0.441984490	0.45658330	-0.96802596	3.330314e-01
CompHrsDay.Q	-0.001893707	0.17908242	-0.01057450	9.915629e-01
CompHrsDay.C	0.664442121	0.46986458	1.41411410	1.573284e-01
CompHrsDay^4	-0.099027493	0.37286985	-0.26558193	7.905612e-01
CompHrsDay^5	0.605370815	0.51246384	1.18129470	2.374857e-01
PhysActiveYes	-0.127240503	0.15263853	-0.83360669	4.045027e-01
PhysActiveDays	0.019838120	0.03361919	0.59008332	5.551348e-01
Poverty	-0.297572429	0.04080574	-7.29241660	3.044440e-13
UnderWeight NormWeight	-5.339732206	0.57651920	-9.26201971	2.006012e-20
NormWeight OverWeight	-1.626829757	0.54874066	-2.96466050	3.030171e-03
OverWeight Obese	-0.811067684	0.54762397	-1.48106680	1.385888e-01

KNN

	Value	Std. Error	t value	p value
Age	-0.084844804	0.03172178	-2.67465440	7.480632e-03
Gendermale	0.428432231	0.14670814	2.92030309	3.496911e-03
Race1Hispanic	-0.404544263	0.31173704	-1.29770995	1.943870e-01
Race1Mexican	-0.013908702	0.24790957	-0.05610393	9.552590e-01
Race1Other	-0.152965009	0.28027050	-0.54577634	5.852197e-01
Race1White	0.005245964	0.21401370	0.02451228	9.804440e-01
TVHrsDay.L	-1.717484231	0.53032676	-3.23853964	1.201433e-03
TVHrsDay.Q	0.741431399	0.39860151	1.86008177	6.287396e-02
TVHrsDay.C	-0.262071765	0.42351734	-0.61879820	5.360493e-01
TVHrsDay^4	-0.264520353	0.38048343	-0.69522174	4.869163e-01
TVHrsDay^5	0.487147031	0.35752001	1.36257277	1.730172e-01
CompHrsDay.L	0.633901254	0.56651660	1.11894559	2.631634e-01
CompHrsDay.Q	0.151708297	0.26378210	0.57512734	5.652052e-01
CompHrsDay.C	0.127055275	0.55728052	0.22799160	8.196528e-01
CompHrsDay^4	0.931780193	0.38656791	2.41039199	1.593539e-02
CompHrsDay^5	0.430350197	0.58591706	0.73448995	4.626502e-01
PhysActiveYes	-0.432008985	0.17723480	-2.43749524	1.478941e-02
PhysActiveDays	0.092626433	0.03999002	2.31623887	2.054523e-02
Poverty	-0.151132022	0.04350499	-3.47390077	5.129507e-04
UnderWeight NormWeight	-5.525740255	0.65193681	-8.47588324	2.333027e-17
NormWeight OverWeight	-0.800062150	0.61678229	-1.29715487	1.945779e-01
OverWeight Obese	0.331708061	0.61705187	0.53756917	5.908745e-01

PPCA

	Value	Std. Error	t value	p value
Age	-0.06741179	0.03114287	-2.1645980	3.041848e-02
Gender	0.64045730	0.13415627	4.7739647	1.806341e-06
Race1	0.03473909	0.04696339	0.7397057	4.594786e-01
TVHrsDay	-0.02258688	0.08376563	-0.2696437	7.874344e-01
CompHrsDay	0.14518196	0.07978819	1.8195920	6.882115e-02
PhysActive	-0.38278513	0.16926560	-2.2614467	2.373161e-02
PhysActiveDays	0.13236668	0.04481195	2.9538258	3.138611e-03
Poverty	-0.16239267	0.04377161	-3.7099995	2.072597e-04
UnderWeight NormWeight	-3.75130078	0.78524917	-4.7772108	1.777435e-06
NormWeight OverWeight	0.90803835	0.76665809	1.1844111	2.362504e-01
OverWeight Obese	2.01244348	0.76898473	2.6170136	8.870281e-03

Mean

	Value	Std. Error	t value	p value
Age	-0.071105696	0.03060757	-2.32314104	2.017158e-02
Gendermale	0.632999795	0.13542535	4.67416024	2.951588e-06
Race1Hispanic	-0.425396049	0.30705648	-1.38540002	1.659302e-01
Race1Mexican	-0.044890130	0.24173188	-0.18570215	8.526783e-01
Race1Other	-0.094030517	0.28009770	-0.33570614	7.370925e-01
Race1White	0.002177297	0.20151196	0.01080480	9.913792e-01
TVHrsDay.L	-1.132006540	0.59758748	-1.89429427	5.818596e-02
TVHrsDay.Q	0.790448957	0.41917829	1.88571062	5.933396e-02
TVHrsDay.C	-0.345685527	0.51245818	-0.67456339	4.999532e-01
TVHrsDay^4	-0.583006903	0.46825475	-1.24506351	2.131084e-01
TVHrsDay^5	0.340502306	0.48708852	0.69905632	4.845168e-01
CompHrsDay.L	-0.221924395	0.75513458	-0.29388721	7.688441e-01
CompHrsDay.Q	0.291238346	0.32049020	0.90872779	3.634938e-01
CompHrsDay.C	0.977981580	0.76779387	1.27375539	2.027501e-01
CompHrsDay^4	0.010573828	0.66120384	0.01599178	9.872409e-01
CompHrsDay^5	1.072993582	0.87069093	1.23234726	2.178194e-01
PhysActiveYes	-0.320517453	0.16967547	-1.88900293	5.889144e-02
PhysActiveDays	0.096071194	0.04474839	2.14691960	3.179968e-02
Poverty	-0.152321773	0.04419668	-3.44645318	5.679970e-04
UnderWeight NormWeight	-5.502900236	0.65449655	-8.40783694	4.176412e-17
NormWeight OverWeight	-0.820231208	0.62081196	-1.32122326	1.864269e-01
OverWeight Obese	0.303634593	0.62073235	0.48915542	6.247317e-01

Focusing on the KNN imputation model for ordinal logistic regression, physical activity appears to “justify” screen time based on BMI, but at the same time, it doesn’t. The results are fairly conflicting. As an example, the significant coefficient for PhysActiveYes is -0.432, meaning physically active kids have lower odds of being in a higher BMI category than non-active kids. However, the significant coefficient for PhysActiveDays is 0.0926, meaning for each extra day someone aged 12-19 is active, the higher the odds of being in a higher BMI category.

One thing that ChatGPT was very helpful with was interpreting output I had never seen before. The TVHrsDay and CompHrsDay variables have output including L, Q, C, 4, and 5, which are not included in the data. These apparently stand for linear, quadratic, cubic, etc. In the KNN imputation model, the linear relationship between TV Hours and BMI was statistically significant. Specifically, for every one-step decrease in the underlying TVHrsDay order, the log odds of being in a higher BMI category decrease by 1.717 on the logit scale.

However, again, the quadratic trend in the relationship was also statistically significant, and the coefficient is positive (0.7414), meaning the trend is a U shape; youth who watch very little or a ton of TV tend to have a higher BMI.

When prompting ChatGPT about this, I was urged to investigate the relationship between BMI category and TV Hours separately. Unfortunately, there are so few observations in the underweight category that I was not able to do any further analyses on this variable. I had attempted to collapse the 7 categories for TV Hours into 3, but it was no use.

After I had thought about this for a while, I remembered talking about natural smoothing splines in Statistical Learning last year, which are linear at boundaries but can be non-linear in other areas. This isn't something ChatGPT brought up, and is something that may be worth investigating in the future.

5 Conclusion

A previously encountered issue, which I ran into again, was incorporating sample weights and adjusting for clusters and strata from the NHANES data. In other types of analyses, mean BMI and prevalence of overweight/obesity was calculated by incorporating these, but in these studies, missing information was excluded from analysis (such as in [this](#) paper). I did not attempt to use them, but they should be accounted for in any type of serious research.

ChatGPT as a research “partner” was useful at times, especially to clarify theory around MCAR, MAR, and NMAR definitions. However, it doesn't catch a lot of smaller details that matter, such as some of the variables of interest not having data for anyone younger than 12 years old. It did help with debugging and generating R code, as well as converting some of my old SAS code into R, but it also seems to forget things after a while, despite being in the same chat thread. At the end of the day, it only does what you tell it to do, and the wording of the prompts you give matter (ex. “Can I do this,” because of course you *can*, but it doesn't mean you should).

ChatGPT helped to bridge theory and practice, even if there were some discrepancies, so fact checking is still a necessity. I've found it to be better at applying rather than understanding theory, but even the applications can be incorrect if the underlying assumptions of it are incorrect. I've also found that if it gives a response that includes something like “it's not this, it's this,” or if it starts talking like a middle schooler trying to meet the word requirements in an essay about a book they didn't read, it's usually not going to be completely correct and there is a lot of research I need to do on my own.

All in all, AI can support missing data imputation and analysis when used carefully, but this is a more complicated topic (especially to me), so it needs to be used with care.

Sources

ChatGPT

MCAR Tests: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3124223/>

PCA: <https://www.nature.com/articles/s41598-025-93333-6>

PCA: https://math.montana.edu/grad_students/writing-projects/2024/Moh2024.pdf

KNN: <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0318-z>

VIM (for knn imputation) R Package info: <https://rdr.io/cran/VIM/man/kNN.html>

Gower Distance:

<https://jamesmccaffrey.wordpress.com/2020/04/21/example-of-calculating-the-gower-distance/>

Missing data in clinical settings: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8499698/>