

Exploring the effects of lifestyle habits such as smoking, physical activity, sleep duration, and alcohol consumption on HDL Cholesterol levels.

Project Proposal - STA302 Fall 2025

Group 146: Vedika Patil, Mariana Garcia Mejia and Bhavikaa Goenka

Contributions:

- **Vedika:** Did part of the Literature Review, the Introduction, edited the Data Descriptions section, and wrote the Ethics Section and the Bibliography.
- **Bhavikaa:** Did the text for the Data Descriptions section.
- **Mariana:** Did part of the Literature Review, the plots for the Data Descriptions section, the plots and text for the Preliminary Analysis section, and the Gantt chart for the team plan. Moved all the information to the Rmd file and uploaded the dataset to OneDrive.

1. Introduction:

HDL cholesterol, often referred to as “good” cholesterol, is a component of blood lipids that plays a crucial role in cardiovascular health. Multiple studies show that even a slight increase in HDL can reduce the risk of heart disease by 2–3% (Mahdy et al., 2012). Due to its importance in improving population health, this project focuses on whether lifestyle choices—specifically smoking, physical activity, sleep duration, and alcohol consumption—significantly affect HDL cholesterol levels in U.S. adults when controlling for demographic characteristics such as sex, race and age. Since heart-related conditions remain the leading cause of death in the United States, claiming one life every 34 seconds (CDC, 2021), understanding the relationship between HDL levels and modifiable habits can help policymakers, and health educators design targeted prevention programs to improve cardiovascular outcomes through changing behaviors.

Several studies examining the relationship between HDL cholesterol levels, and lifestyle behaviours and demographic factors. Smoking cessation significantly improved HDL cholesterol levels, especially in women (Gepner et al., 2011), and moderate alcohol consumption has been linked to an increase in HDL as a percentage of total cholesterol (Thornton, Symes & Heaton, 1983). Additionally, lower HDL levels are associated with more sleep problems (Cai, Zhou & Zeng, 2024), while physical activity has been shown to increase HDL cholesterol levels (Palazón-Bru, Hernández-Lozano, Gil-Guillén, 2021). Furthermore, it has been observed that biomedical risk factors for cholesterol are highest in middle adulthood (Dash et al., 2019) and that women are predisposed to higher HDL levels than men (Kim et al., 2011). Overall, we see that HDL tends to rise with lifestyle behaviours such as smoking cessation, moderate increases in alcohol consumption, exercise, and sleep, and varies with demographic factors like age and sex.

Our study will employ a multiple linear regression model because it quantifies the relationship between a continuous response variable (HDL cholesterol) and several predictors simultaneously, allowing for the isolation of each lifestyle factor’s effect on HDL, while holding others constant (e.g. sex or age). The resulting coefficients will capture the change in average HDL cholesterol per unit increase in a predictor, uncovering individual effects. Furthermore, hypothesis testing will help determine which habits explain most of the variability in HDL cholesterol (t-tests) or to compare models with different sets of predictors (F-test). Identifying the predictors that explain most of the variability in HDL cholesterol will help guide policy toward programs that target the most important lifestyle habits, while allowing for more reliable confidence and prediction intervals.

2. Data Description:

2.1. Description of data source:

We use the 2009–2010 NHANES data provided by the U.S. Centers for Disease Control and Prevention (CDC, 2015) in the R Package NHANES. The National Health and Nutrition Examination Survey is collected annually to monitor national health and nutrition trends. The data combines household interviews, standardized physical examinations, and laboratory tests of voluntary survey takers.

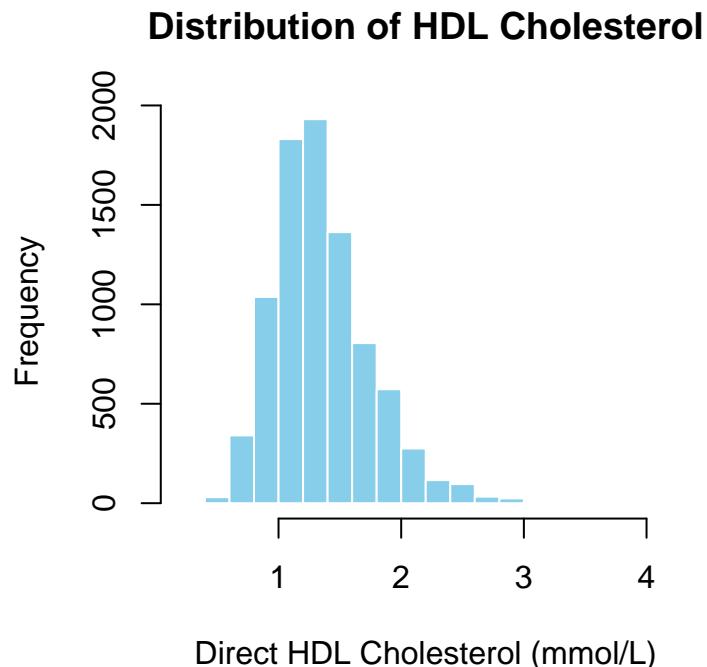
2.2. Data Description and Summary Statistics

The data contained 1,000+ observations after filtering adult respondents in 2009-10. It includes one response variable (Direct HDL Cholesterol), and nine predictor variables: some numerical, such as Income or Alcohol Days, and some categorical, such as Gender or Diabetes.

2.2.1. Predictor variable: Direct HDL Cholesterol.

Table 1: Summary statistics of the response variable: DirectChol (Direct HDL cholesterol in mmol/L.)

Group	n	Mean	Median	SD	Min	Max
Overall	1268	1.44	1.37	0.45	0.54	3.72



Values for Direct HDL cholesterol (mmol/L), the response variable, span about 0.54 to 3.72 mmol/L and centre near 1.4 mmol/L. The histogram is unimodal with a slight right skew, showing frequency of lower values. HDL is continuous and thus suitable for linear regression. Observations can be treated as independent since they are individual and de-identified. If mild skews affect diagnostics, we will consider a transformation.

2.2.2. Predictor variables.

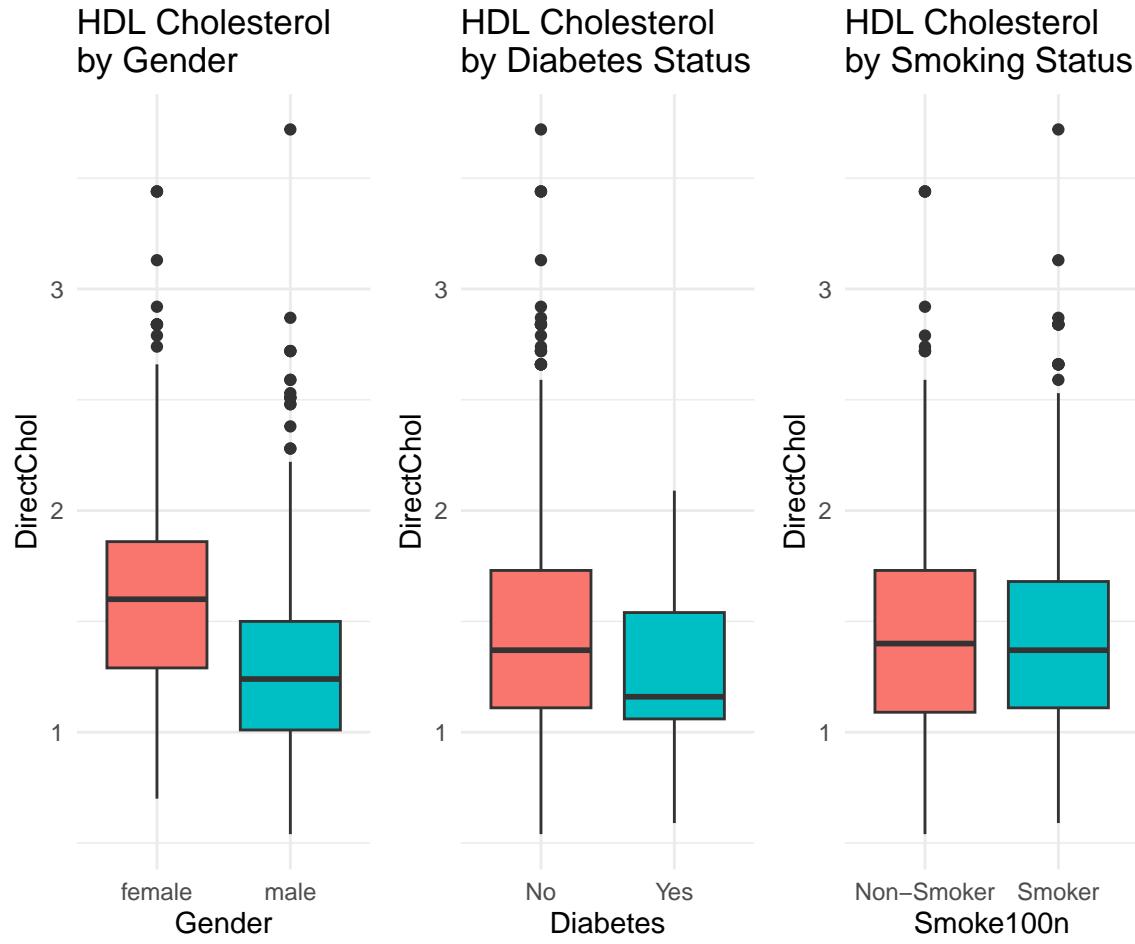
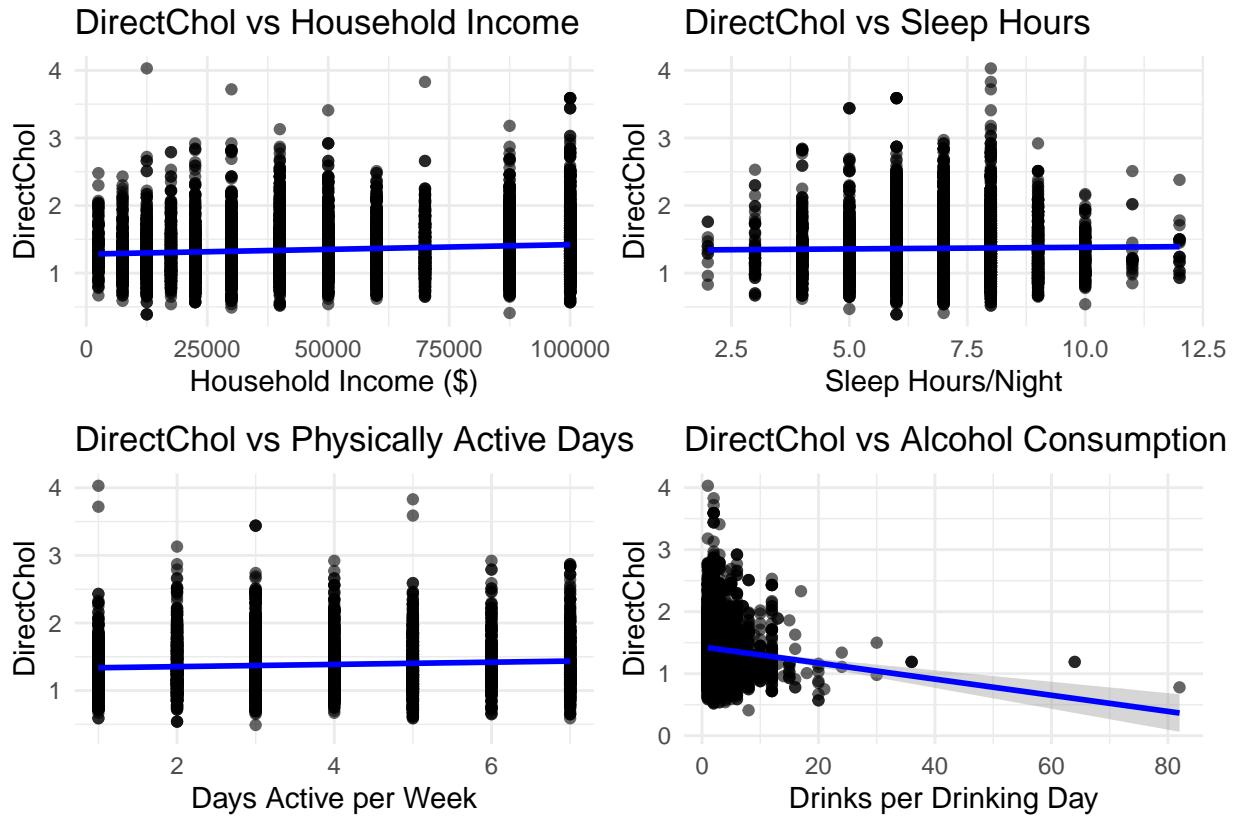


Table 2: Summary statistics of Direct HDL cholesterol in mmol/L. for each Sex.

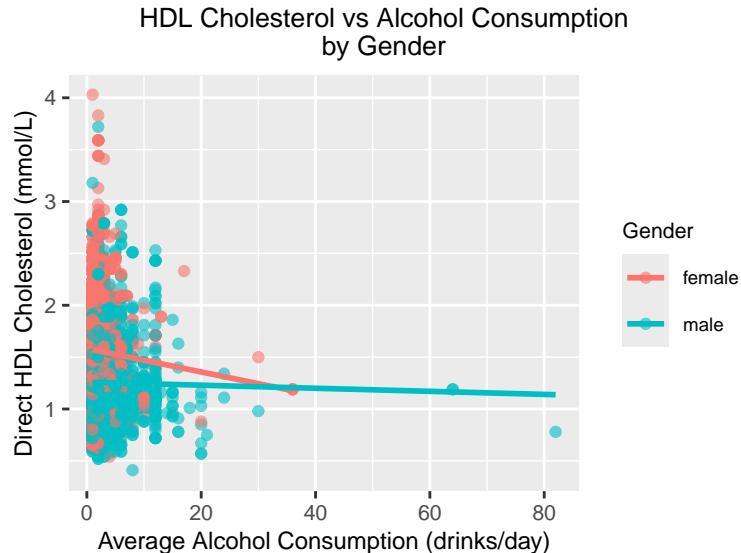
Sex	n	Mean	Median	SD	Min	Max
Female	570	1.61	1.60	0.455	0.70	3.44
Male	698	1.29	1.24	0.400	0.54	3.72

Table 3: Summary statistics of the Direct HDL cholesterol in mmol/L. for each Race.

Race	n	Mean	Median	SD	Min	Max
Black	108	1.39	1.34	0.41	0.59	2.53
Hispanic	54	1.37	1.32	0.40	0.59	2.22
Mexican	72	1.34	1.27	0.50	0.65	3.72
White	955	1.46	1.40	0.46	0.54	3.44
Other	74	1.27	1.16	0.38	0.72	2.38



The boxplots align with the literature: women have higher HDL, and smokers and people with diabetes have lower HDL, with similar spreads across groups. Among the continuous predictors, income is strongly right-skewed with the widest spread; sleep hours cluster around six to eight and relate only weakly to HDL; activity days range from 0–7 and look banded but remain numeric (ordered data rather than categories), displaying a slight upward trend. Alcohol has a long right tail with a few heavy-drinker outliers, suggesting a funnel pattern and possible heteroskedasticity. Missingness is highest for income and sleep, and alcohol is unrecorded for non-drinkers.



We include a gender \times alcohol interaction because both prior work and our data suggest sex differences in HDL responses. de Oliveira e Silva (2000) and Suzuki (2025) report that HDL increases with moderate alcohol intake, with differing effects in women and men. In the plot above, the fitted lines for men and women have different slopes, suggesting an interaction term. Most points lie at low to moderate intake levels, with a few heavy-drinker outliers, creating a funnel pattern that we will examine for constant-variance issues.

3. Ethics Discussion

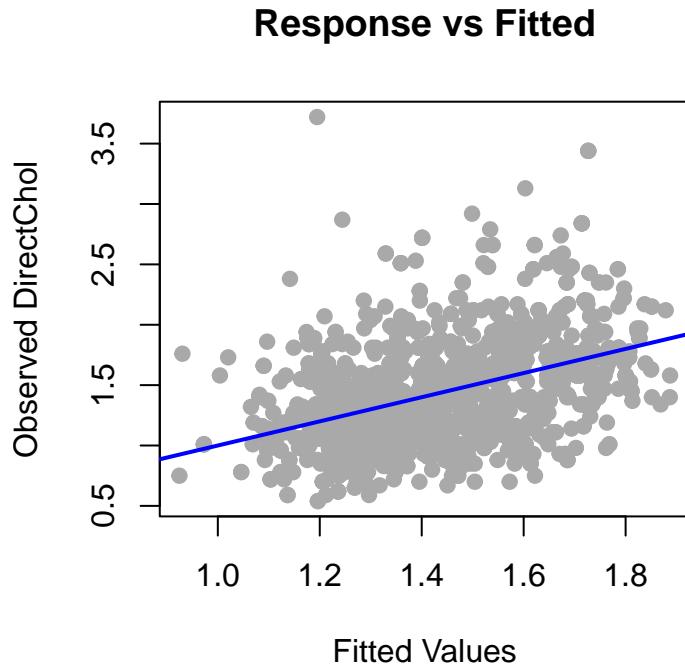
Considering ethics is central to the data used in research. The data used for our research comes from the National Health And Nutrition Examination Survey (NHANES) conducted by the US National Center for Health Statistics (NCHS). This sample specifically includes information on variables collected between 2009-10 and 2011-12. The dataset includes comprehensive metadata, including information on the sampling design, variable descriptions, and warnings and disclaimers to consider when working with the dataset, to improve transparency.

From an ethics standpoint, the data collected ensures consent, privacy and ownership. Furthermore, all variables are collected using a standardised procedure. First and foremost, even though the data uses random sampling, participation is voluntary, as participants can call in; thus, we most likely have consent. Each participant undergoes a health interview, followed by a blood test and dental exam, along with any other required procedures. Lastly, they receive a report of their medical findings. Throughout this entire process, their identity is kept confidential, and participant privacy is protected by law. The NHANES study is overseen by the Ethics Review Board (ERB), which ensures participant rights are protected. Lastly, since the data collected lies with the government, rather than third-party individuals, there's a lower risk of potential data misuse.

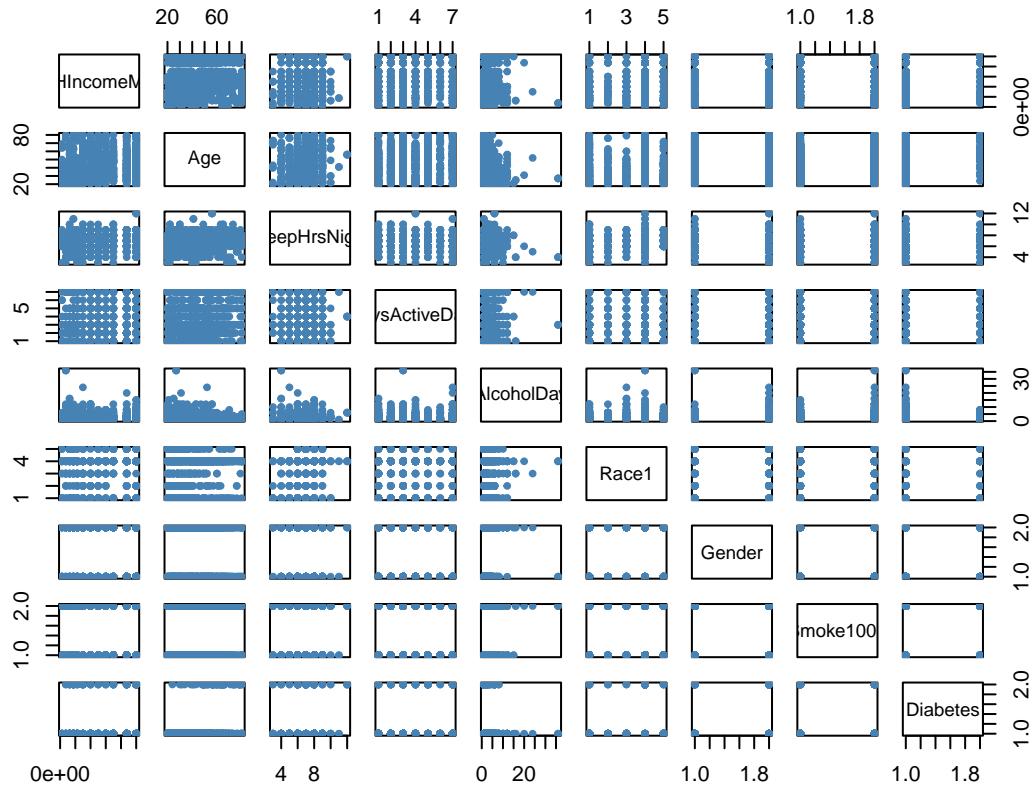
While NHANES does an excellent job of ensuring participant privacy and consent, it is also essential to consider whether the dataset is trustworthy. In this case, the dataset can be deemed trustworthy because it is widely used and approved by credible third parties. Federal agencies use the data to develop policies, and academics and students at academic institutions use NHANES data for public health and nutrition research papers. Thus, evaluating our dataset against source transparency, metadata completeness, collection method, and third-party approval is both ethical and reliable for our multiple linear regression model.

4. Preliminary Results

4.1. Residual analysis of preliminary model.

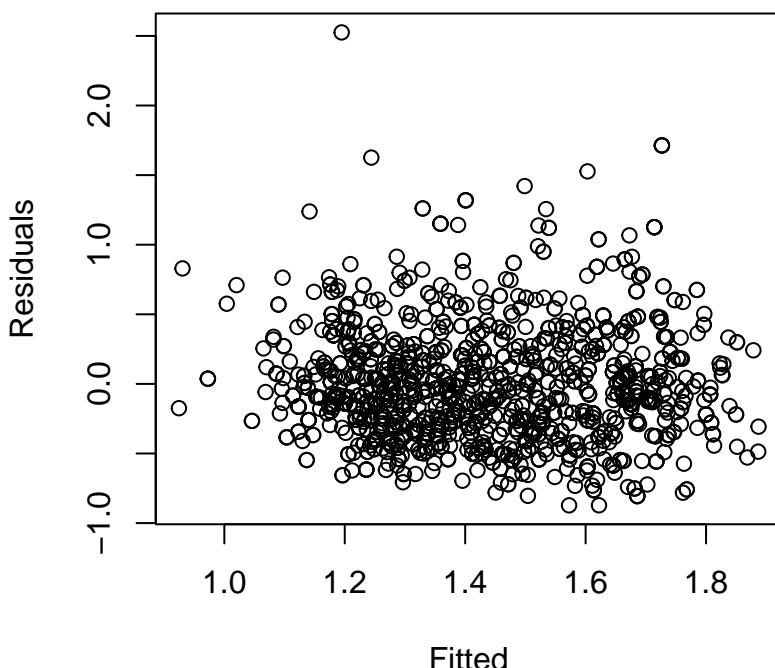


Scatterplot Matrix of Predictor Variables



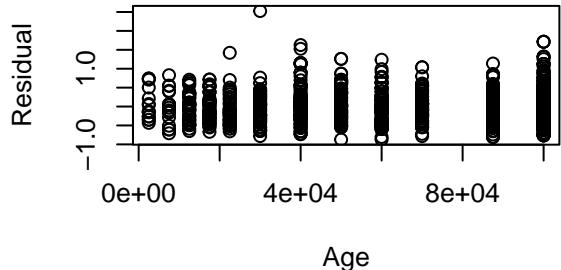
While residual plots are useful for diagnosing assumption violations, they can be misleading when the relationships among predictors or between predictors and the response are complicated. Therefore, conditional mean response and predictor conditions were first assessed. The Response vs. Fitted Values plot suggests an approximately linear relationship with large spread indicating possible linearity improvement, while pairwise plots show predictors are roughly linearly and randomly related. Given that both conditions reasonably hold, we proceed to the residual diagnostics.

Residual vs Fitted

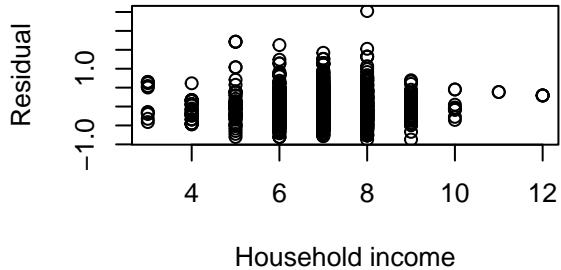


Residuals vs. Numerical Variables.

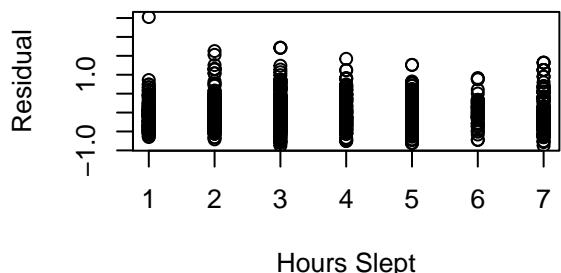
Residual vs Age



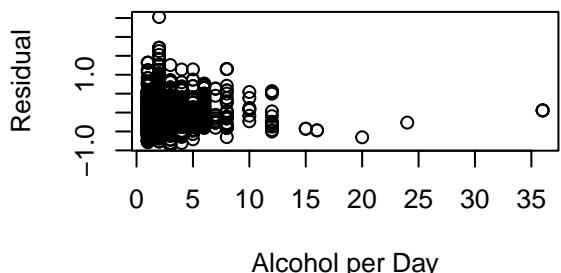
Residual vs Household Income



Residual vs Hours Slept

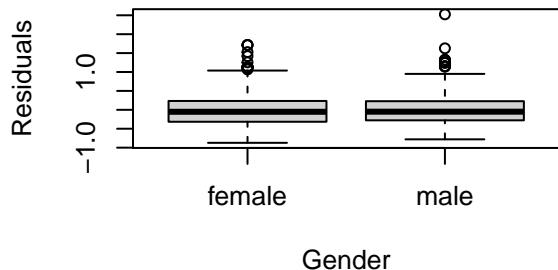


Residual vs Alcohol per Day

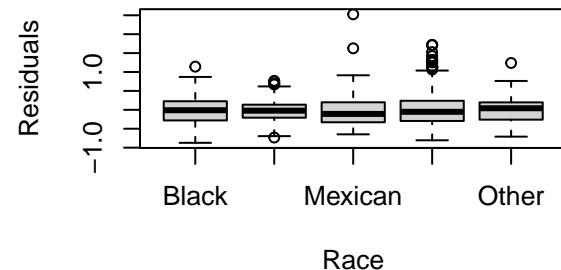


Residuals vs. Categorical Variables.

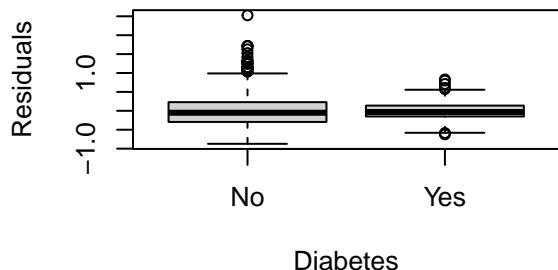
Residual vs Gender



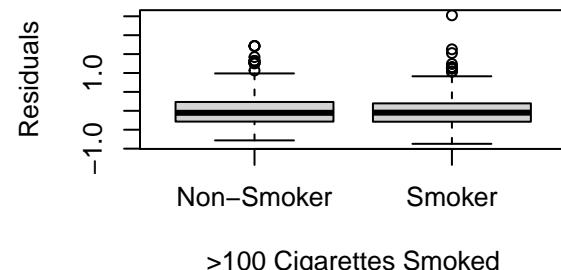
Residual vs Race



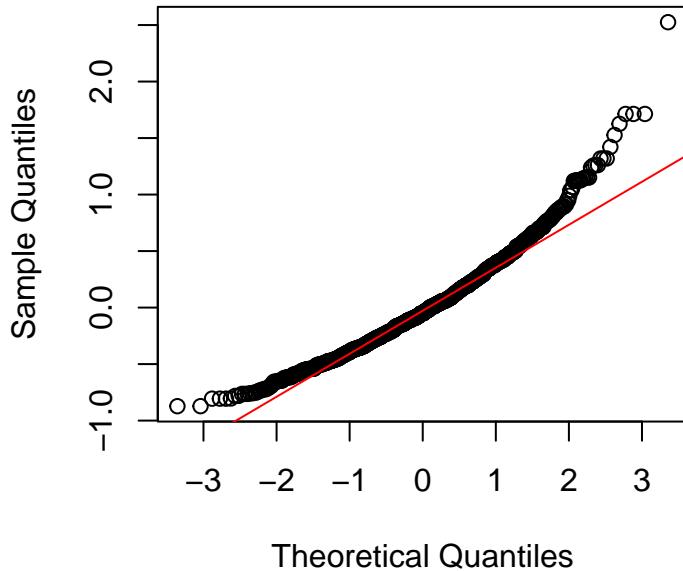
Residuals vs Diabetes



Residuals vs Smoke 100 cigs



Normal Q–Q Plot



The Residuals vs. Fitted plot shows residuals centered around zero with no curvature or clusters, supporting linearity and uncorrelated errors. Variance appears mostly constant here, as well as in the Residual vs. Numerical predictors plots. However, the Residuals vs. Alcohol per Day plot suggests non-constant variance, as residuals narrow as alcohol consumption increases, possibly due to outliers. For categorical predictors, residuals are centered and randomly spread across groups, indicating no major violations aside from mild outliers among males, Mexicans, and smokers. Lastly, the Normal Q–Q plot shows deviations in the upper tail, confirming non-normality and the skewed distribution in Section 2. Overall, residual analysis indicates minor non-linearity and violations of normality and constant variance.

4.2. Preliminary model discussion

Table 4: Model Estimates with 95% Confidence Intervals

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.348	0.091	14.812	0.000	1.170	1.527
Gendermale	-0.356	0.033	-10.742	0.000	-0.421	-0.291
HHIncomeMid	0.000	0.000	2.052	0.040	0.000	0.000
Race1Hispanic	-0.070	0.069	-1.009	0.313	-0.206	0.066
Race1Mexican	-0.033	0.064	-0.512	0.609	-0.157	0.092
Race1White	-0.022	0.044	-0.490	0.624	-0.108	0.065
Race1Other	-0.134	0.062	-2.161	0.031	-0.255	-0.012
Age	0.005	0.001	6.287	0.000	0.004	0.007
Smoke100nSmoker	-0.017	0.024	-0.705	0.481	-0.064	0.030
SleepHrsNight	-0.005	0.010	-0.523	0.601	-0.024	0.014
DiabetesYes	-0.212	0.050	-4.244	0.000	-0.310	-0.114
PhysActiveDays	0.020	0.007	3.056	0.002	0.007	0.033
AlcoholDay	-0.010	0.006	-1.627	0.104	-0.023	0.002
Gendermale:AlcoholDay	0.020	0.008	2.418	0.016	0.004	0.037

The preliminary regression results identify several significant predictors of HDL cholesterol, aligning partly with prior research but with unexpected findings. Aligning with the literature, gender, income, and age were strong predictors: males had significantly lower HDL than females, while HDL increased modestly with age. Additionally, physical activity was positively associated with HDL and significant.

However, smoking, alcohol consumption and sleep were not significant once other predictors were included. Surprisingly, the interaction between gender and alcohol was significant, suggesting that males experience a higher increase

in HDL levels per additional drink consumed than women. Finally, confidence intervals for income and some race predictors were relatively wide, indicating possible multicollinearity. This means that demographic variables may already capture much of the explained variation in HDL, reducing the additional explanatory power of lifestyle predictors.

5. Plan

5.1. Analysis Plan

Transformations:

The preliminary model showed strong violations of normality, which can lead to biased inference due to inaccurate critical values. Therefore, our first step will be to apply a Box–Cox transformation to the response variable (DirectChol), as this method is commonly used to correct non-normality and improve linearity. We will apply a simplified Box–Cox transformation on the DirectChol to maintain interpretability and re-examine the residual plots to determine whether the normality, linearity, and constant variance assumptions improve. If the Box–Cox transformation is insufficient, we will test alternative variance-stabilizing transformations on the response, such as the natural logarithm or square root, and select the one that most effectively corrects the violations. We will not initially transform predictors, as transforming the response often improves both normality and constant variance. Model assumptions will be reassessed in the residual plots when using a new transformation to confirm improvement and ensure the model remains interpretable.

Predictor selection:

Since the literature mostly considers lifestyle factors separately, and this study examines them jointly, this may explain the differences in results and limited significance observed. To address this, additional literature will be reviewed to identify potential relationships between variables (e.g., individuals who exercise frequently may be less likely to drink). Variance Inflation Factors (VIFs) will then be calculated to detect potential multicollinearity, and predictors with high VIFs may be removed in the reduced model. Model assumptions will be reassessed, and if satisfied, a partial F-test will compare the reduced and full models. This process will continue until we arrive at a model with significant improvement, where coefficient variances are small and confidence intervals are narrow.

5.2. Team Plan

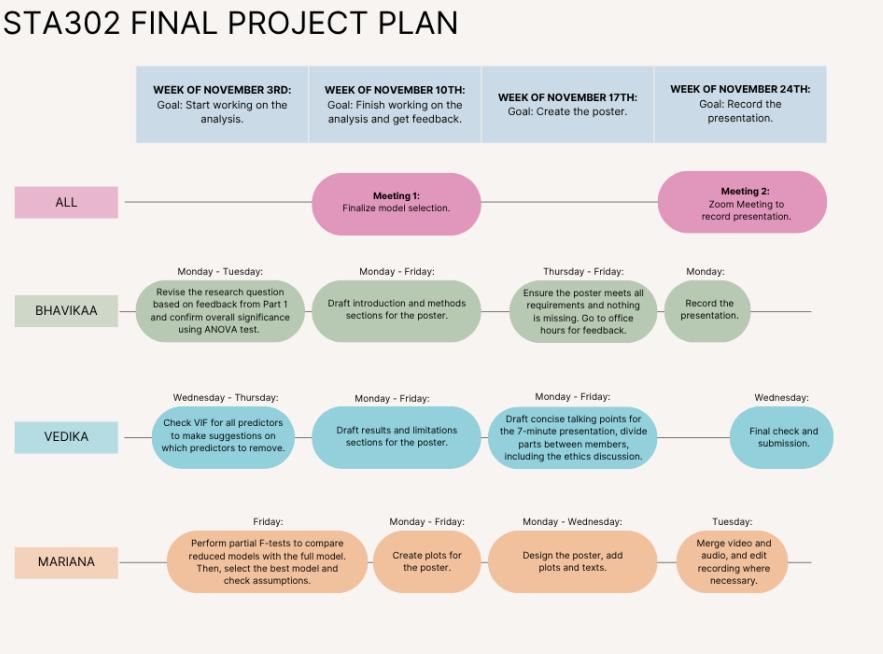


Figure 1: Gantt Chart

6. Bibliography

1. Cai, Y., Zhou, Z., & Zeng, Y. (2024). Association between non-high-density lipoprotein cholesterol to high-density lipoprotein cholesterol ratio (NHHR) and sleep disorders in US adults: NHANES 2005 to 2016. *Medicine*, 103(26), e38748. <https://doi.org/10.1097/MD.00000000000038748>
2. U.S. Centre for Disease Control and Prevention (2025, July 10). Heart disease facts. *Heart Disease*. <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html>
3. Dash, S. R., Hoare, E., Varsamis, P., Jennings, G. L. R., & Kingwell, B. A. (2019). Sex-specific lifestyle and biomedical risk factors for chronic disease among early-middle, middle and older aged Australian adults. *International Journal of Environmental Research and Public Health*, 16(2), 224. <https://doi.org/10.3390/ijerph16020224>
4. De Oliveira E Silva, E. R., Foster, D., McGee Harper, M., Seidman, C. E., Smith, J. D., Breslow, J. L., & Brinton, E. A. (2000). Alcohol consumption raises HDL cholesterol levels by increasing the transport rate of apolipoproteins A-I and A-II. *Circulation*, 102(19), 2347–2352. <https://doi.org/10.1161/01.cir.102.19.2347>
5. Gepner, A. D., Piper, M. E., Johnson, H. M., Fiore, M. C., Baker, T. B., & Stein, J. H. (2011). Effects of smoking and smoking cessation on lipids and lipoproteins: outcomes from a randomized clinical trial. *American Heart Journal*, 161(1), 145–151. <https://doi.org/10.1016/j.ahj.2010.09.023>
6. Kim, H. J., Park, H. A., Cho, Y. G., Kang, J. H., Kim, K. W., Kang, J. H., Kim, N.-R., Chung, W.-C., Kim, C. H., Whang, D. H., & Park, J. K. (2011). Gender difference in the level of HDL cholesterol in Korean adults. *Korean Journal of Family Medicine*, 32(3), 173–181. <https://doi.org/10.4082/kjfm.2011.32.3.173>
7. Lin, Y.-Y., Chen, H.-C., Lai, W.-S., Wu, L.-W., Wang, C.-H., Lee, J.-C., Kao, T.-W., & Chen, W.-L. (2017). Gender differences in the association between moderate alcohol consumption and hearing threshold shifts. *Scientific Reports*, 7(1), 2201. <https://doi.org/10.1038/s41598-017-02426-4>
8. Mahdy Ali, K., Wonnerth, A., Huber, K., & Wojta, J. (2012). Cardiovascular disease risk reduction by raising HDL cholesterol—current therapies and future opportunities: Cardiovascular risk and HDL cholesterol. *British Journal of Pharmacology*, 167(6), 1177–1194. <https://doi.org/10.1111/j.1476-5381.2012.02081.x>
9. NHANES: Data from the US National Health and Nutrition Examination Study (2015). Version 2.1.0 [R package: NHANES].
10. Palazón-Bru, A., Hernández-Lozano, D., & Gil-Guillén, V. F. (2021). Which physical exercise interventions increase HDL-cholesterol levels? A systematic review of meta-analyses of randomized controlled trials. *Sports Medicine* (Auckland, N.Z.), 51(2), 243–253. <https://doi.org/10.1007/s40279-020-01364-y>
11. Thornton, J., Symes, C., & Heaton, K. (1983). Moderate alcohol intake reduces bile cholesterol saturation and raises HDL cholesterol. *Lancet*, 2(8354), 819–822. [https://doi.org/10.1016/s0140-6736\(83\)90738-9](https://doi.org/10.1016/s0140-6736(83)90738-9)
12. Whitehead, T. P., Robinson, D., & Allaway, S. L. (1996). The effects of cigarette smoking and alcohol consumption on blood lipids: a dose-related study on men. *Annals of Clinical Biochemistry*, 33 (Pt 2)(2), 99–106. <https://doi.org/10.1177/000456329603300201>