Summary of the Final Project
Yishan Lin, Yimin Ding, Ruiqing Cai, Yingqi Wang

Research Question:

This project investigates the relationship between demographic, dietary, and physiological variables and blood pressure (BP) categories, using data from the National Health and Nutrition Examination Survey (NHANES). By leveraging Random Forest machine learning models, we aimed to predict systolic and diastolic BP categories and provide meaningful visualizations to explore data patterns. The project integrates functional programming, object-oriented programming (OOP), and data visualization paradigms to create a comprehensive and reproducible data analytic product.

Goals and Technical Challenges

Original Goals:

1. Develop machine learning models to predict BP categories based on NHANES data.
2. Create visualizations to explore relationships between predictors and BP.
3. Implement a reproducible workflow using advanced programming paradigms.

Accomplishments:

1. Built Random Forest models to classify BP categories.
2. Conducted multiple imputation for missing data using Predictive Mean Matching (PMM).
3. Created visualizations, including violin plots and scatter plots, to explore data patterns.
4. Integrated functional programming and OOP paradigms to streamline the analysis pipeline.

Technical Challenges:

1. **Handling Missing Data:** A significant portion of dietary variables, such as `DRQPOTA`, had missing values. We addressed this by imputing missing values and removing problematic variables.
2. **Class Imbalance:** The dataset had imbalanced BP categories, complicating model performance. This issue was mitigated using Random Forest's inherent balancing capabilities.
3. **Variable Selection:** Ensuring only meaningful predictors were included required careful preprocessing to avoid data leakage.

Previous Work:

Previous studies have explored hypertension risk factors using traditional statistical methods[1]. This project builds on prior work by integrating machine learning approaches with comprehensive preprocessing and advanced programming paradigms to improve prediction accuracy and reproducibility[2-4].

Results:

Given the structure of NHANES data, which consists of multiple subset datasets, we adopted functional programming paradigms to create a streamlined, efficient, and reproducible workflow. We developed a series of reusable functions to handle key tasks such as downloading, cleaning, stratifying, and generating a table. We used the map function to apply our custom functions to a list of datasets, ensuring consistency across operations and the reduce function was used to merge datasets with left_join. In Figure 1, we can see that the distribution of systolic blood pressure is positively skewed, but there is little difference among education levels.

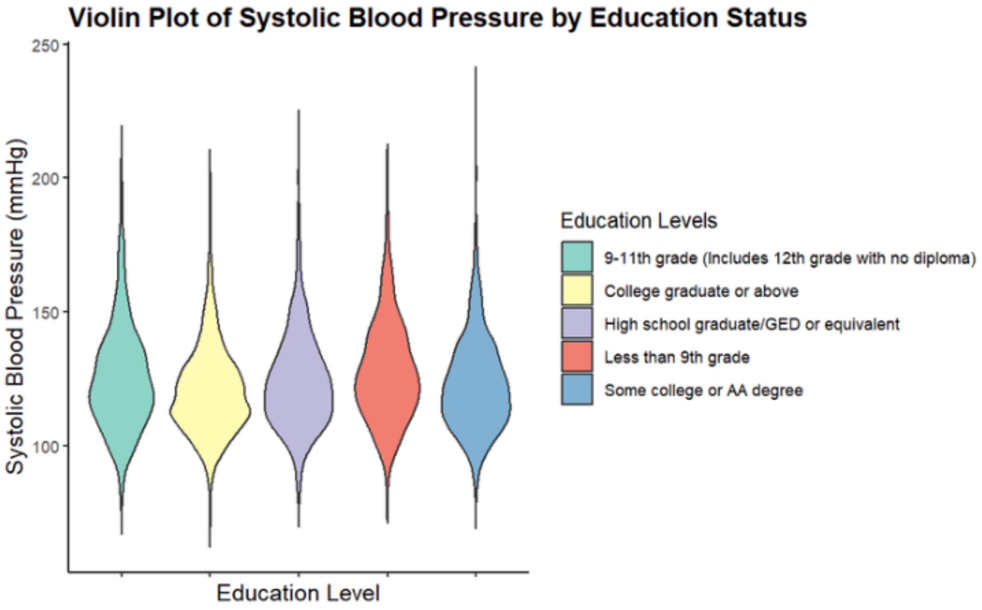
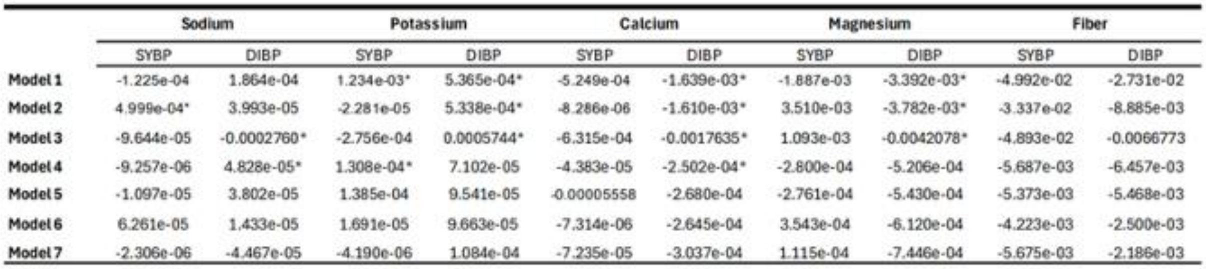


Figure 1. Violin Plot of Systolic Blood Pressure by Education Status

In this project, the results aligned with some of these findings, though the relationships were weaker than anticipated after adjusting for confounding variables. Sodium intake showed a negative association with blood pressure, while nutrients such as potassium and calcium demonstrated minimal impact after accounting for age, BMI, and other factors (Table 1). These outcomes underscore the complex interplay between diet and blood pressure regulation. By integrating machine learning methods with statistical analysis, the study provided insights into

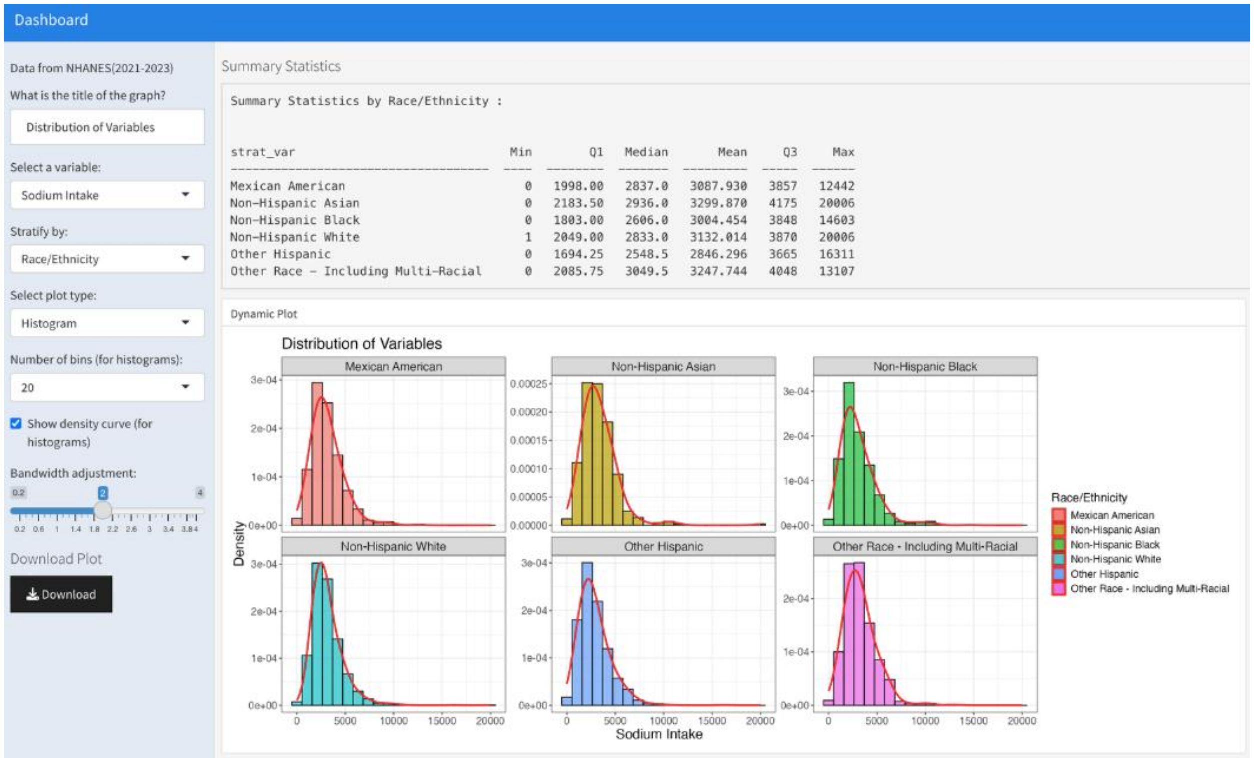
Table 1 Regression Models Results by Nutrients



	Sodium		Potassium		Calcium		Magnesium		Fiber	
	SYBP	DIBP	SYBP	DIBP	SYBP	DIBP	SYBP	DIBP	SYBP	DIBP
Model 1	-1.225e-04	1.864e-04	1.234e-03*	5.365e-04*	-5.249e-04	-1.639e-03*	-1.887e-03	-3.392e-03*	-4.992e-02	-2.731e-02
Model 2	4.999e-04*	3.993e-05	-2.281e-05	5.338e-04*	-8.286e-06	-1.610e-03*	3.510e-03	-3.782e-03*	-3.337e-02	-8.885e-03
Model 3	-9.644e-05	-0.0002760*	-2.756e-04	0.0005744*	-6.315e-04	-0.0017635*	1.093e-03	-0.0042078*	-4.893e-02	-0.0066773
Model 4	-9.257e-06	4.828e-05*	1.308e-04*	7.102e-05	-4.383e-05	-2.502e-04*	-2.800e-04	-5.206e-04	-5.687e-03	-6.457e-03
Model 5	-1.097e-05	3.802e-05	1.385e-04	9.541e-05	-0.00005558	-2.680e-04	-2.761e-04	-5.430e-04	-5.373e-03	-5.468e-03
Model 6	6.261e-05	1.433e-05	1.691e-05	9.663e-05	-7.314e-06	-2.645e-04	3.543e-04	-6.120e-04	-4.223e-03	-2.500e-03
Model 7	-2.306e-06	-4.467e-05	-4.190e-06	1.084e-04	-7.235e-05	-3.037e-04	1.115e-04	-7.446e-04	-5.675e-03	-2.186e-03

hypertension risk prediction while reinforcing the need for targeted nutritional strategies to reduce cardiovascular risks.

Our dashboard allows us to visualize relationships between key variables and systolic blood pressure, such as sodium, potassium, calcium and so on (Figure 2). Then we can also select the variables of interest to see stratified distribution. Here we can choose to stratify by gender, race/ethnicity, or education level for tailored subgroup analyses. The dynamic plot section includes histograms, boxplots, and scatterplots. For histograms, we could also customize the number of bins and overlay a density curve to better understand the data distribution. Above the plots, summary statistics section dynamically updates to reflect the chosen stratification, providing distribution, means, and quartiles for each group. Additionally, we can easily download the generated plots for further reporting or presentations.



Dashboard

Data from NHANES(2021-2023)

What is the title of the graph? Distribution of Variables

Select a variable: Sodium Intake

Stratify by: Race/Ethnicity

Select plot type: Histogram

Number of bins (for histograms): 20

☑ Show density curve (for histograms)

Bandwidth adjustment: 2

Summary Statistics

```
Summary Statistics by Race/Ethnicity :


strat_var                            Min        Q1   Median       Mean     Q3     Max
------------------------------------ ----  --------  -------  ---------  -----  ------
Mexican American                        0   1998.00   2837.0   3087.930   3857   12442
Non-Hispanic Asian                      0   2183.50   2936.0   3299.870   4175   20006
Non-Hispanic Black                      0   1803.00   2606.0   3004.454   3848   14603
Non-Hispanic White                      1   2049.00   2833.0   3132.014   3870   20006
Other Hispanic                          0   1694.25   2548.5   2846.296   3665   16311
Other Race - Including Multi-Racial     0   2085.75   3049.5   3247.744   4048   13107
```

Figure 2. Dashboard Display with Diagrams

In order to conduct stratification and regression model, we used S3 based analysis. First, we defined S3 class of blood pressure analysis. Then we created a `model.BPAnalysis` and `summary.BPAnalysis` for the regression model and output. Finally run it with `run_model.BPAnalysis. We also created a machine learning model to see if we have these kinds of such data in hands in the future, how the predicted possible blood pressure index would be. Before that, we imputed the missing value in the dataset by predictive mean matching. We had tried to use numeric blood pressure variables to produce the model, but only 10 % of the data were predicted. Then, we switched to binary outcomes, either normal or high blood pressure and defined hypertension based on the current handbook, which is either systolic blood pressure or diastolic blood pressure is in high level we defined. All the variables in the NHANES database, especially demographic data and nutrition data were included. The random forest method was used and trained with 250 trees.

We also conduct parallel processing to boost the process. To evaluate the model, confusion matrices were used. About 60% of the data can be predicted, which is the best we can do due to the small sample size of our datasets. Interestingly, we also conduct machine learning on only \`sybp\` or \`dibp\`, the model can have an almost 100% sensitivity to predict normal person, but only 10% specificity to predict hypertension.

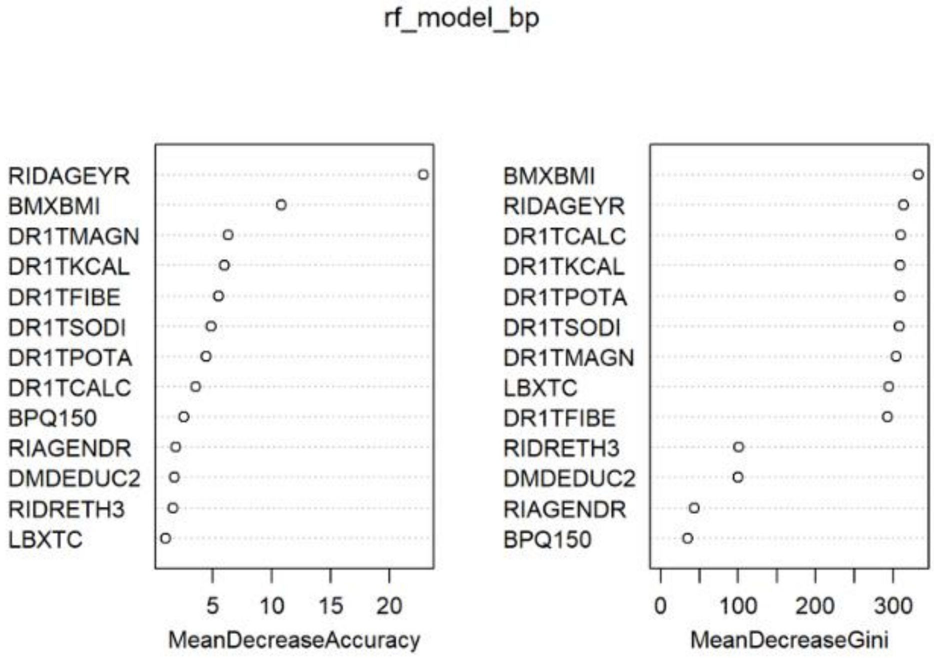


Figure 3. Importance of Variables in the Random Forest Model

In all our model results (Figure 3), we can see the most important variable to predict the blood pressure would be the age and BMI, which would be reasonable since we also see the significant during our stratified analysis. But five of the nutrients did not have a tremendous impact on the model. This project offered valuable insights into both scientific and technical domains. Scientifically, it underscored the complexities of isolating direct effects of dietary nutrients due to confounding and reverse causation. It also highlighted the limitations of cross-sectional data for causal inference. Technically, the project demonstrated the power of programming paradigms like functional and object-oriented approaches. A web-based dashboard further enhanced usability and accessibility.

Reference:

1. Appel LJ, Moore TJ, Obarzanek E, et al. A clinical trial of the effects of dietary patterns on blood pressure. *N Engl J Med*. 1997;336(16):1117-1124. doi:10.1056/NEJM199704173361601.
2. Turin TC, Saeed A, Shahana N, et al. Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis. PLoS One. 2022;17(4):e0266334. doi:10.1371/journal.pone.0266334.
3. Zhao H, Zhang X, Xu Y, et al. Predicting the risk of hypertension based on several easy-to-collect risk factors: A machine learning method. *Front Public Health*. 2021;9:619429. doi:10.3389/fpubh.2021.619429.
4. Krittanawong C, Zhang H, Wang Z, et al. Machine learning prediction in cardiovascular diseases: A meta-analysis. *Sci Rep*. 2020;10(1):16057. doi:10.1038/s41598-020-72685-1.