
Group C

Pei-Yu Jheng, Yifan Wang, Jiahui Yang

ALY6040: Data Mining Applications

Winter Sec 01

Instructor: Justin Grosz

Module 2 Final Project — EDA

# Data Mining Analysis: Unveiling Insights into Cardiovascular Disease Risk Factors

03/03/24

## Introduction

For this project, we have selected the "Cardiovascular Diseases Risk Prediction Dataset" obtained from the 2021 Behavioral Risk Factor Surveillance System (BRFSS), administered by the Centers for Disease Control and Prevention (CDC). Preprocessed by Kaggle poster @ALPHIREE, he hand-picked 308,854 records and 19 variables that relate to lifestyle factors of a person that can contribute to being at risk with any form of Cardiovascular Disease. Central to our analysis is the variable "Heart_Disease," which serves as a binary indicator of respondents reporting coronary heart disease or myocardial infarction. By predicting this variable, we seek to gain insights into the underlying determinants of heart disease, enabling the identification of high-risk individuals and the formulation of targeted interventions to mitigate cardiovascular risk factors.

## Data Cleaning

During the data cleaning process, we carefully examined the dataset to ensure its accuracy and reliability for the subsequent analysis including replacing missing values, checking duplicate values, handling outliers, and encoding categorical variables. The dataset comprises 19 variables, including 12 numerical and 7 categorical variables, the description of each variable is presented in Table 1. There are a total of 74,122 values missing, 49 values are duplicated, and 7 columns have outliers. We will introduce our data-cleaning process in the following.

***Missing Value***

The dataset contains missing values in 6 columns: Checkup (20.00% missing), Heart_Disease (1.00% missing), Other_Cancer (5.00% missing), Age_Category (1.00% missing), Weight_(kg) (6.00% missing), and Green_Vegetables_Consumption (11.00% missing). The percentage of missing values in each column relative to the total number of rows ranges from 0.00% to 20.00%. The columns with the highest percentage of missing values are Checkup (20.00%), Green_Vegetables_Consumption (11.00%), and Other_Cancer (5.00%).

The "Checkup" column, with 20% null values, is best handled by replacing them using the mode due to its categorical nature. Dropping null values is not ideal as it would result in significant data loss, and using the mode maintains data consistency and original distribution characteristics. For the "Heart_Disease" column, with nearly 1% null values and being the target variable, it was decided to drop the null values to maintain prediction accuracy. Since the null values are relatively few, this approach is suitable to avoid compromising the integrity of the data. In the case of the "Other_Cancer" column, with 5% null values, the mode was chosen for replacement. This decision was based on the nature of the feature as an estimate, where assuming cancer without information is inappropriate. Additionally, the 90% of results being "NO" supported the use of mode filling for consistency.

The decision to drop null values in the “age category” was made due to the low

percentage of null values (1%) and the even distribution of age data in Figure B1, making averaging or using the median inaccurate. Additionally, Using the mode would not accurately represent the age distribution, especially considering that the mode, representing "65 - 69," accounts for only 10% of the total values. In the "Weight_(kg)" column, which exhibits positive skewness and high kurtosis, we opt to use the median for filling missing values. This choice ensures that the imputed values are less influenced by outliers and maintain the overall distribution characteristics. Similarly, for the "Green_Vegetables_Consumption" column, which also shows significant skewness and kurtosis in Figure B2, mean imputation is deemed unsuitable. Instead, we choose the median for filling in missing values to preserve the central trend of the data and mitigate the impact of extreme values.

***Duplicate Values***

The presence of very few duplicate values (84 values), we have decided to keep them in the dataset. This choice aims to preserve the data as close to the original records as possible, ensuring that the integrity and authenticity of the data are maintained.

***Handle Outliners***

After using the Interquartile Range (IQR) method to identify and replace outliers with NaN values for each numeric column, we have chosen to populate these NaN values with the median of each column. This decision was made because the median is less affected by outliers and can better preserve the overall distribution of the data.

However, it's important to note that after applying the IQR method, the data distribution of 'Alcohol_Consumption' and 'Fruit_Consumption' has been significantly altered. To minimize the impact of the IQR method on the original data in these columns, we have decided to retain the original values in 'Alcohol_Consumption' and 'Fruit_Consumption'.

For the other five features ('Height_(cm)', 'Weight_(kg)', 'BMI', 'Green_Vegetables_Consumption', 'FriedPotato_Consumption'), the IQR method has been successfully implemented to handle outliers, and the NaN values have been replaced with the median of each column.

**Variable Selection**

To choose predictor variables for predicting whether a person will have heart disease or not, we should consider variables with relatively strong correlations with the "Heart Disease" variable. Based on the correlation matrix Figure C1, key predictor variables include:

1. Age Category (0.23): Age is often a significant factor in heart disease risk.

2. Diabetes (0.17): Diabetes is a known risk factor for heart disease.

3. Arthritis (0.15): While the correlation is not very strong, arthritis may be associated with lifestyle factors that could affect heart disease risk.

4. Smoking History (0.11): Smoking is a major risk factor for heart disease.

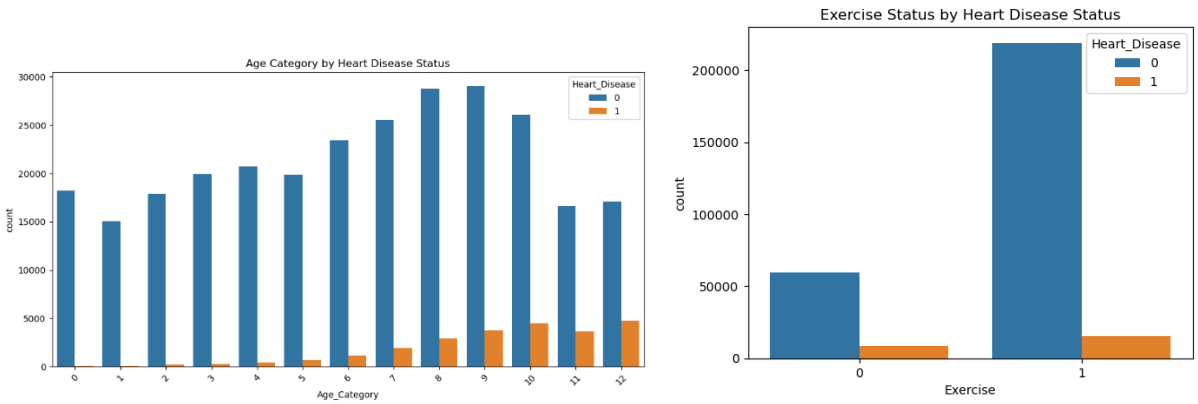
5. Skin Cancer (0.091): This correlation is relatively weak but might indicate some shared risk factors.

6. Other Cancer (0.088): Similar to skin cancer, this correlation is weak but might suggest shared risk factors.

7. Exercise (-0.097): While negatively correlated, exercise is an important factor in heart health and could be a relevant predictor.

Here is the exploratory data analysis of our key predictor variables:

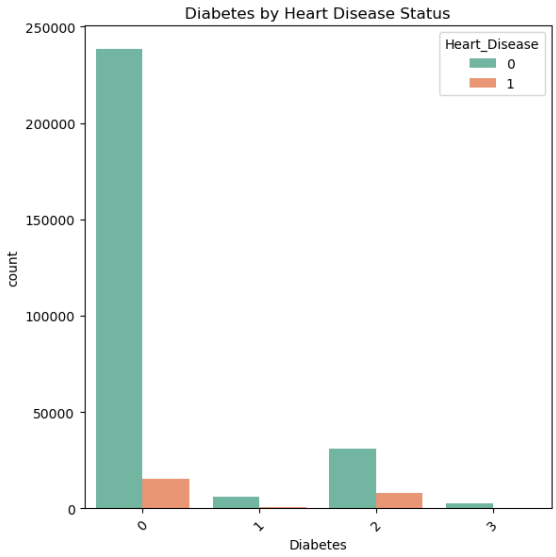
**Figure C2 & 3**




Age Group and Heart Disease: The distribution of heart disease may vary significantly among different age groups, and the incidence of heart disease may be higher in older age groups.

Exercise and Heart Disease: People who exercise regularly may have lower rates of heart disease. A bar chart shows the relationship between exercise habits and heart disease status.

**Figure C4**




Diabetes and Heart Disease: 0 is 'No', 1 is 'Yes', 2 is 'No, pre-diabetes or borderline diabetes', 3 is 'Yes, but female told only during pregnancy'. There is a clear pattern where the number of individuals with heart disease increases in the diabetes category compared to the non-diabetic category. The data suggests a link between diabetes status and the presence of heart disease, with diabetes (category 1) showing a higher proportion of individuals with heart disease.

Key findings or correlations from the matrix and EDA:

- There is a moderate positive correlation between age category and heart disease.

- Diabetes and arthritis also show positive correlations with heart disease, although not as strong as age.

- Smoking history and exercise show some correlation with heart disease.

- General health, checkup, depression, sex, height, weight, BMI, alcohol consumption, fruit consumption, green vegetables consumption, and fried potato consumption do not show

strong correlations with heart disease based on this matrix.

**Recap Readiness of Data Set**

From our analysis, we gained several key findings and learnings emerge:

1. Age is a significant factor: The data indicates a moderate positive correlation between age category and heart disease, suggesting that older individuals may be at a higher risk.
2. Diabetes and arthritis are relevant: Both diabetes and arthritis show positive correlations with heart disease, though not as strong as age. This suggests that managing these conditions may be important in reducing cardiovascular risk.
3. Exercise is beneficial: While negatively correlated, indicating that regular exercise might reduce the risk of heart disease.
4. Limited impact of certain factors: Factors such as general health, checkup, depression, sex, height, weight, BMI, alcohol consumption, fruit consumption, green vegetables consumption, and fried potato consumption do not show strong correlations with heart disease in this dataset.
5. Data reliability and value: The dataset was obtained from the 2021 Behavioral Risk Factor Surveillance System (BRFSS), administered by the CDC, and preprocessed by a Kaggle user. This dataset includes a large number of records (308,854) and 19 variables related to lifestyle factors and heart disease, making it valuable for analyzing cardiovascular risk factors.

Stakeholders will be excited about the potential insights this dataset can provide into the determinants of heart disease. By analyzing this data, stakeholders can identify high-risk individuals and develop targeted interventions to reduce cardiovascular risk factors. The dataset's large size and detailed variables make it a valuable resource for understanding and addressing heart disease.

## References

*Cardiovascular Diseases Risk Prediction Dataset*. (2023, July 3). Kaggle. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data

*CDC - 2021 BRFSS Survey Data and Documentation*. (n.d.). https://www.cdc.gov/brfss/annual_data/annual_2021.html

# Appendix A

**Table A1**

*Description of Variables*

| | Variable | Description | Type |
|---|---|---|---|
| 1 | General Health | Respondents considered their health conditions in general | object |
| 2 | Checkup | how long has it been since respondents last visited a doctor for a routine checkup? | object |
| 3 | Exercise [Yes or No] | Any participation in physical activities or exercise? | object |
| 4 | Heart Disease [Yes or No] | Respondents reported having coronary heart disease or mycardialinfarction. | object |
| 5 | Skin Cancer [Yes or No] | Respondents reported having skin cancer | object |
| 6 | Other Cancer [Yes or No] | Respondents reported having any other types of cancer | object |
| 7 | Depression [Yes or No] | Respondents reported having a depressive disorder (including depression, major depression, dysthymia, or minor depression) | object |
| 8 | Diabetes [Yes or No] | Respondents reported having a diabetes. If yes, what type of diabetes it is/was. | object |
| 9 | Arthritis [Yes or No] | Respondents reported having an Arthritis. | object |
| 10 | Sex | Gender of Respondents | object |
| 11 | Age Category | Age group of Respondents | object |
| 12 | Height (cm) | The height of the patients seems to follow a normal distribution, with the majority of patients having heights around 160 to 180 cm. | int64 |
| 13 | Weight (kg) | The weight of the patients also appears to be normally distributed, with most patients weighing between approximately 60 and 100 kg. | float64 |
| 14 | BMI | The distribution of Body Mass Index is somewhat right-skewed. A large number | float64 |

| | | | |
|---|---|---|---|
| | | of patients have a BMI between 20 and 30, which falls within the normal to overweight range. However, there are also a significant number of patients with a BMI in the obese range (>30). | |
| 15 | Smoking History | | object |
| 16 | Alcohol Consumption | This feature is heavily right-skewed. Most patients have low alcohol consumption, but there are a few patients with high consumption. | int64 |
| 17 | Fruit Consumption | This feature is also right-skewed. A lot of patients consume fruits regularly, but a significant number consume them less frequently. | int64 |
| 18 | Green Vegetables Consumption | This feature appears to be normally distributed, with most patients consuming green vegetables moderately. | float64 |
| 19 | FriedPotato Consumption | This feature is right-skewed. Many patients consume fried potatoes less frequently, while a few consume them more often. | int64 |

# Appendix B

## Data Cleaning Process

**Figure B1**

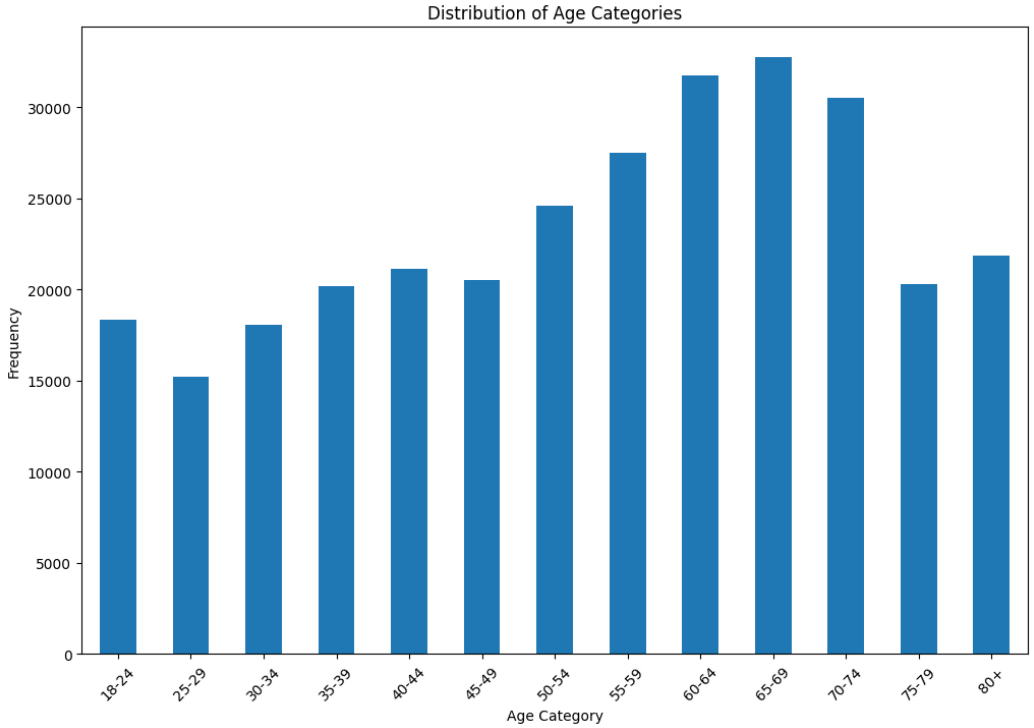
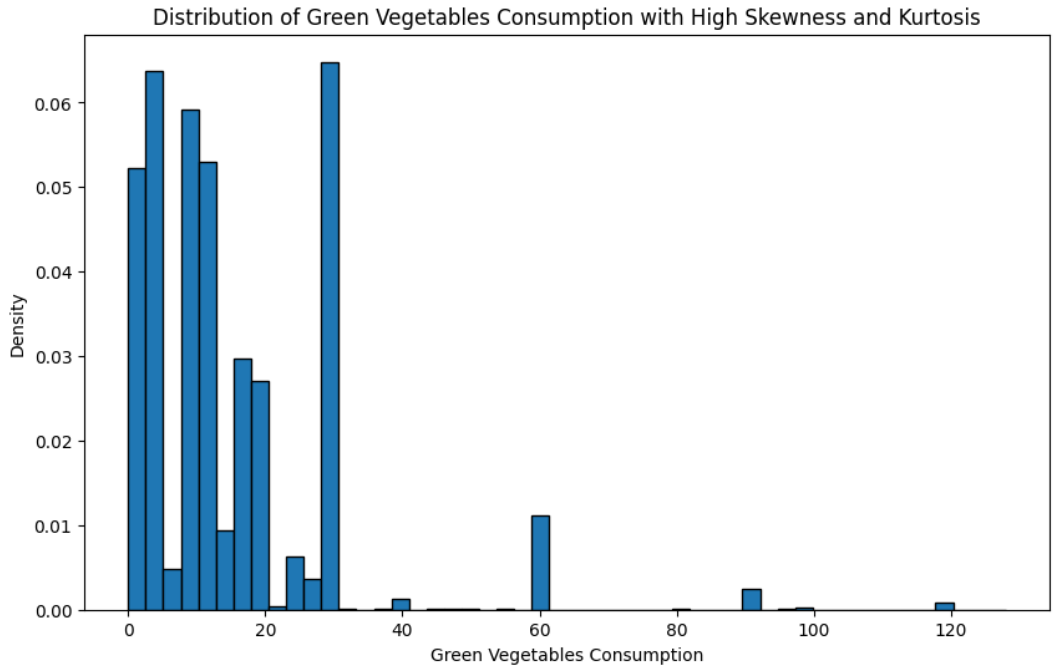


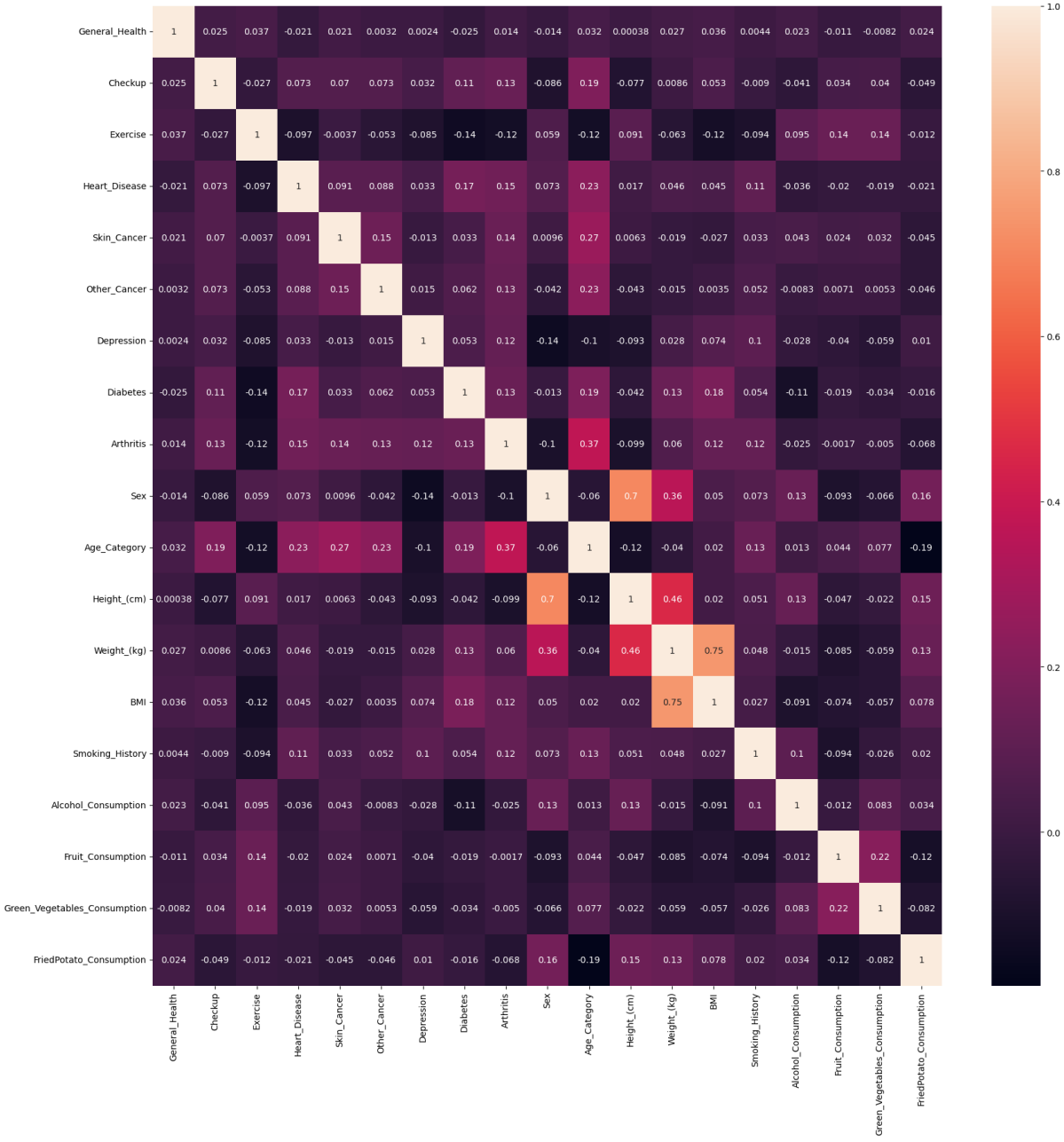

**Figure B2**



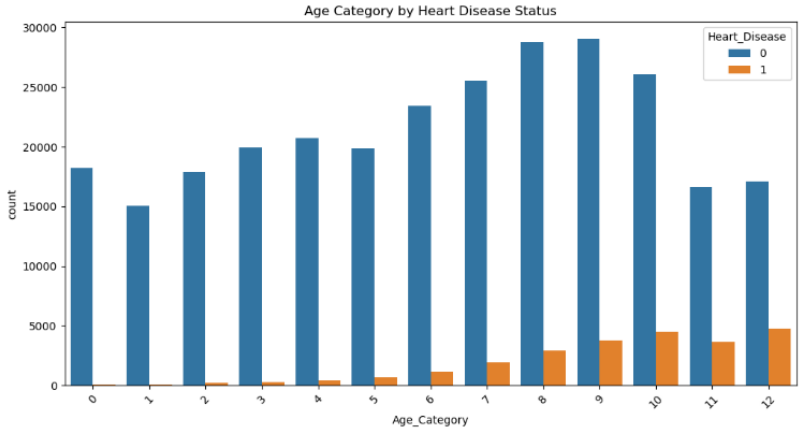
# Appendix C

## Predictor Selection

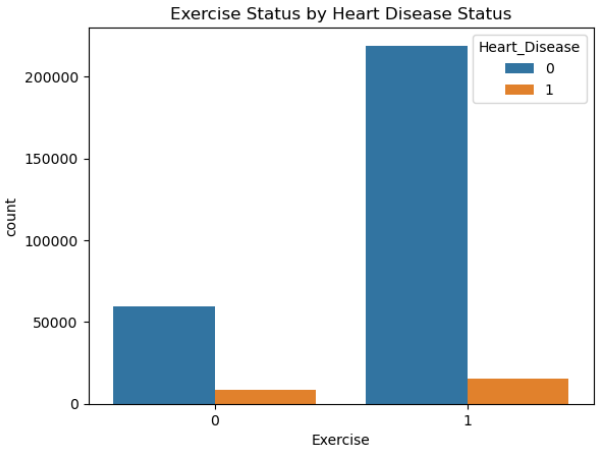
**Figure C1 – Correlation Metrix**



## Figure C2




## Figure C3




## Figure C4