Aryan Sharma

Logistic regression

Data Cleaning and Manipulation

Explore the variables included in the dataset.

1. Load the data into your environment and perform any necessary cleaning steps and any data preprocessing steps needed for your analysis.

```
# Load data
file_path <- ("C:/Users/91884/Desktop/BAIS/Advance data science/Assignment 4/SAHD.xlsx")
data <- read_excel(file_path, sheet = "Data")
summary(data)
str(data)
```

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
9	114	0.00	3.83	19.40	Present	49	24.86	2.49	29	0
10	132	0.00	5.80	30.96	Present	69	30.11	0.00	53	1
11	206	6.00	2.95	32.27	Absent	72	26.81	56.06	60	1
12	134	14.10	4.44	22.39	Present	65	23.09	0.00	40	1
13	118	0.00	1.88	10.05	Absent	59	21.57	0.00	17	0
14	132	0.00	1.87	17.21	Absent	49	23.63	0.97	15	0
15	112	9.65	2.29	17.20	Present	54	23.53	0.68	53	0

2. Conduct a descriptive analysis of the key variables. Note any major observations.

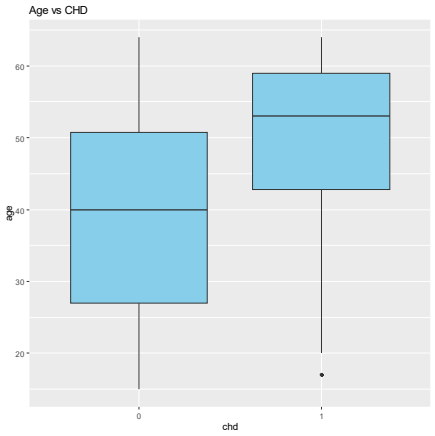
```
# 3. Descriptive Analysis of Key Variables
summary(data %>% select(age, ldl, sbp, tobacco, adiposity, obesity, alcohol, typea, famhist, chd))
table(data$chd)
```

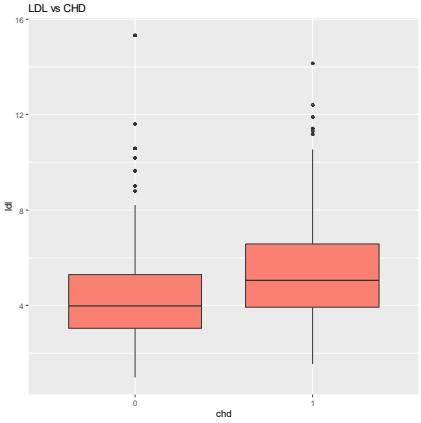
```
      age              ldl              sbp            tobacco          adiposity        obesity          alcohol          
 Min.   :15.00   Min.   : 0.980   Min.   :101.0   Min.   : 0.0000   Min.   : 6.74   Min.   :14.70   Min.   :  0.00  
 1st Qu.:31.00   1st Qu.: 3.283   1st Qu.:124.0   1st Qu.: 0.0525   1st Qu.:19.77   1st Qu.:22.98   1st Qu.:  0.51  
 Median :45.00   Median : 4.340   Median :134.0   Median : 2.0000   Median :26.11   Median :25.80   Median :  7.51  
 Mean   :42.82   Mean   : 4.740   Mean   :138.3   Mean   : 3.6356   Mean   :25.41   Mean   :26.04   Mean   : 17.04  
 3rd Qu.:55.00   3rd Qu.: 5.790   3rd Qu.:148.0   3rd Qu.: 5.5000   3rd Qu.:31.23   3rd Qu.:28.50   3rd Qu.: 23.89  
 Max.   :64.00   Max.   :15.330   Max.   :218.0   Max.   :31.2000   Max.   :42.49   Max.   :46.58   Max.   :147.19  
     typea        famhist              chd        
 Min.   :13.0   Length:462         Min.   :0.0000  
 1st Qu.:47.0   Class :character   1st Qu.:0.0000  
 Median :53.0   Mode  :character   Median :0.0000  
 Mean   :53.1                      Mean   :0.3463  
 3rd Qu.:60.0                      3rd Qu.:1.0000  
 Max.   :78.0                      Max.   :1.0000  
> table(data$chd)

  0   1 
302 160 
```

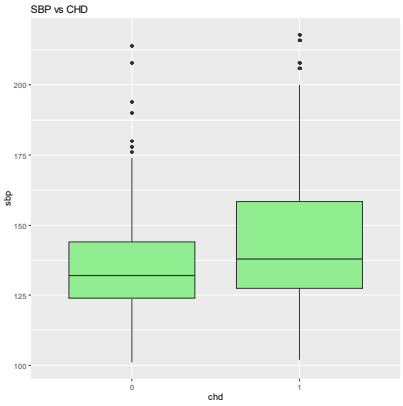
Aryan Sharma

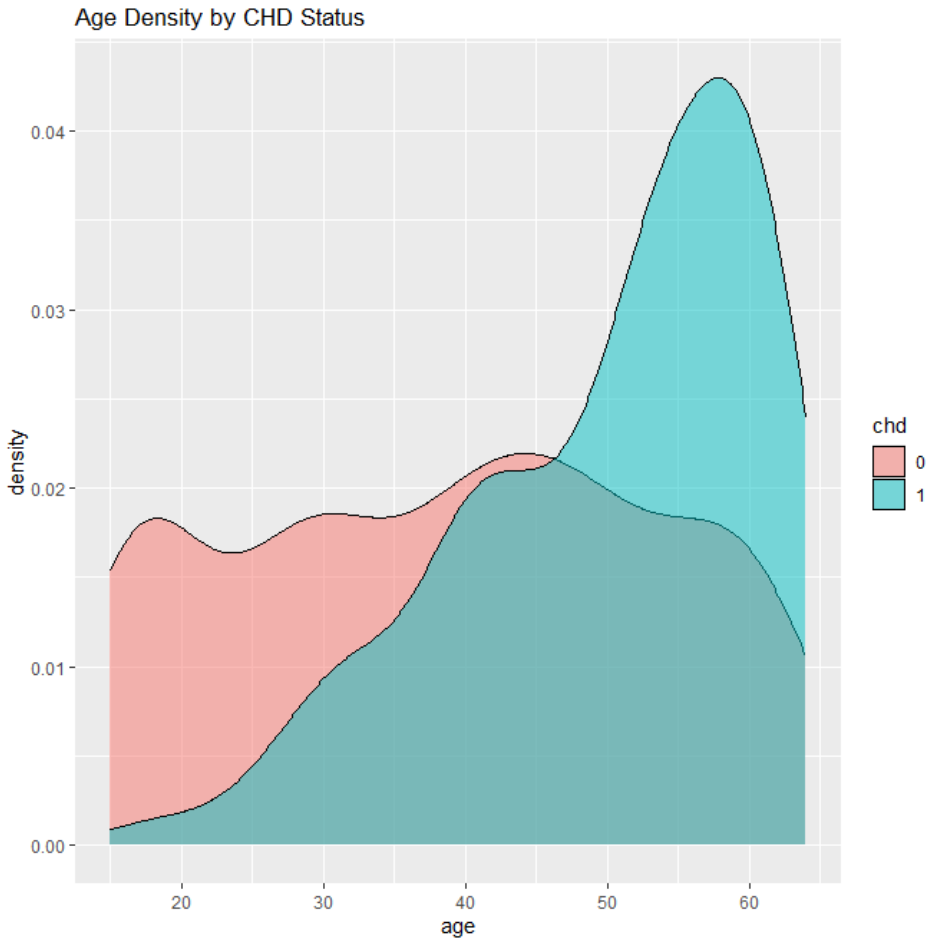
3. Use visualization techniques to explore the relationships between various predictors (such as age, ldl, blood pressure) and the binary outcome of heart disease.

```
# 4. Visualizations: Predictors vs Heart Disease
data$famhist <- as.factor(data$famhist)
data$chd <- as.factor(data$chd)

ggplot(data, aes(x = chd, y = age)) + geom_boxplot(fill = "skyblue") + ggtitle("Age vs CHD")
ggplot(data, aes(x = chd, y = ldl)) + geom_boxplot(fill = "salmon") + ggtitle("LDL vs CHD")
ggplot(data, aes(x = chd, y = sbp)) + geom_boxplot(fill = "lightgreen") + ggtitle("SBP vs CHD")

ggplot(data, aes(x = age, fill = chd)) + geom_density(alpha = 0.5) + ggtitle("Age Density by CHD Status")
```

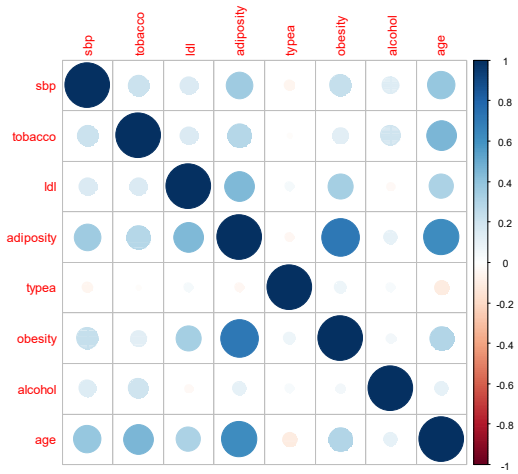








4. Identify and discuss any observable trends or patterns that could influence your logistic regression analysis.



- **Age** shows a clear upward trend in CHD risk — individuals with CHD tend to be older, confirming age as a strong predictive factor.
- **LDL levels** and **systolic blood pressure (SBP)** are generally higher in individuals with CHD, supporting their inclusion in logistic models.
- **Density and boxplots** reveal that CHD cases cluster around higher values of age, LDL, and SBP, suggesting these variables meaningfully separate the two outcome groups.
- Other variables like **tobacco use**, **alcohol**, and **obesity** show high variability and outliers, which could reduce model clarity if included without transformation or filtering.

  Summary: Age, LDL, and SBP show consistent, meaningful trends with CHD r

5. Based on your exploratory data analysis, formulate two statistical inference questions that logistic regression can address. Explain why these questions are relevant and how they relate to the dataset.
   Q1: Does age predict the likelihood of developing CHD?
   This question is relevant because **age is a well-established risk factor** for heart disease. In your dataset, age ranges from 15 to 64, with a mean of ~43 years, capturing both younger and older adults. Since coronary heart disease (CHD) risk typically increases with age, analyzing this relationship helps determine whether age alone can significantly explain variations in CHD occurrence. This aligns with clinical understanding and helps validate the dataset's quality and predictive power.
   Q2: Do LDL and SBP together predict CHD?

Aryan Sharma

This question targets two major **physiological risk indicators—LDL cholesterol** and **systolic blood pressure (SBP)**—both of which are routinely monitored in cardiovascular health assessments. The dataset includes considerable variation in both LDL (0.98–15.33) and SBP (101–218), making it suitable for evaluating their **combined effect** on heart disease. This question is relevant because it mirrors real-world diagnostics where these two factors are considered together to assess heart disease risk, thereby enhancing model accuracy beyond single-variable prediction.

6. For each question, identify the predictor variable(s) and the binary outcome variable.

Predictor Variable	Reason for Inclusion/Deletion	Used in Question
age	Included — age is a key clinical predictor of CHD risk	Q1
ldl	Included — LDL is a major cholesterol indicator for heart health	Q2
sbp	Included — SBP is directly related to hypertension and CHD risk	Q2
tobacco	Excluded — not directly relevant to Q1 or Q2; related to lifestyle	-
adiposity	Excluded — measures body fat but not directly tested in Q1 or Q2	-
obesity	Excluded — correlated with adiposity, not tested in selected questions	-
alcohol	Excluded — high variance, outliers, not part of selected questions	-
typea	Excluded — psychological measure, not used in Q1 or Q2	-
famhist	Excluded — categorical; not included in current logistic models	-

7. For each question, perform logistic regression analysis to assess the relationship between the predictor(s) and the outcome.

```
Logistic Regression Results
=================================================
                        Dependent variable:
                    -----------------------------
                               CHD
                        Age Only          LDL + SBP
-------------------------------------------------
Age                 0.064*** (0.009)
LDL                                   0.255*** (0.052)
SBP                                   0.017*** (0.005)
Constant            -3.522*** (0.416) -4.180*** (0.730)
-------------------------------------------------
Observations              462               462
Log Likelihood         -262.781          -276.497
Akaike Inf. Crit.       529.562           558.993
=================================================
Note:                       *p<0.1; **p<0.05; ***p<0.01
```

Model 1 (Age Only):
Age has a significant positive effect on CHD ($\beta = 0.064$, $p < 0.01$), meaning each additional year of age increases the log-odds of heart disease.

Model 2 (LDL + SBP):
Both LDL (β = 0.255) and SBP (β = 0.017) are statistically significant predictors of CHD ($p < 0.01$), suggesting that increases in cholesterol and blood pressure are both linked to higher CHD risk.

8. Interpret the logistic regression coefficients and discuss their implications in the context of heart disease risk.
   The logistic regression results show that **age**, **LDL**, and **systolic blood pressure (SBP)** are all significant predictors of coronary heart disease (CHD). In Model 1, each additional year of age increases the odds of CHD by approximately **6.6%**, highlighting age as a strong risk factor. In Model 2, a one-unit increase in **LDL** raises the odds by **29%**, while each mmHg increase in **SBP** raises the odds by **1.7%**. These findings confirm that both age and key physiological measures significantly contribute to heart disease risk and should be prioritized in prevention strategies.

9. Evaluate the model's fit and discuss any limitations or assumptions in your analysis.

```
#10. Model Fit and Evaluation
# Pseudo R-squared
pR2(model1)
pR2(model2)

# Hosmer-Lemeshow test
hoslem.test(data$chd, fitted(model1))
hoslem.test(data$chd, fitted(model2))
```

```
> pR2(model1)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-262.7811684 -298.0542100   70.5460833    0.1183444    0.1416104    0.1953770
> pR2(model2)
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-276.49672016 -298.05421000   43.11497967    0.07232741    0.08910028    0.12292988
```

```
> hoslem.test(data$chd, fitted(model1))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  data$chd, fitted(model1)
X-squared = NA, df = 8, p-value = NA
```

```
> hoslem.test(data$chd, fitted(model2))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  data$chd, fitted(model2)
X-squared = NA, df = 8, p-value = NA
```

The McFadden R^2 values indicate modest model fits:

- Model 1 (age): 0.118
- Model 2 (ldl + sbp): 0.072

This suggests age alone explains more variation in CHD risk than LDL and SBP combined.

However, the Hosmer-Lemeshow test failed to return a valid p-value (likely due to variable format issues or a constant predicted value bin). As a result, model fit can't be validated using this test