Nicole Carter
STAT 183

Individual Project Written Report

Project Title and Goals

Cardiovascular Disease Risk Prediction

Goals / Research Questions:

- What factors are the best predictors for Cardiovascular Disease? How can we best identify Cardiovascular Disease with survey data?
- Which variables have the most significance to predict the risk of cardiovascular diseases.
- With a person's health-related behaviors and chronic health conditions, what factors would be used to best identify a person who is at risk of having cardiovascular diseases?

Data Description

Uses data from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) Survey Data and Documentation.

- Using the preprocessed Kaggle Dataset:
    - The Dataset contains 19 variables. 12 are numerical and 7 are categorical variables

Initial Number of Observations: 308,854

- No missing values
- 80 duplicate observations
    - Final data set dimensions: 308,774 observations x 19 variables

Interesting Variables:

- Respondent self-assessed general health
    - Poor
    - Very Good
    - Good
    - Fair
    - Excellent
- How long it has been since last doctor visit (checkup)
    - Within the past 2 years
    - Within the past year
    - 5 or more years ago
    - Within the past 5 years

    - Never
- During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
    - Numerical
- Respondents that reported having coronary heart disease or myocardial infarction (**Response Variable**)
    - Yes/No
- Respondents that reported having skin cancer
    - Yes/No
- Respondents that reported having any other types of cancer
    - Yes/No
- Respondents that reported having a depressive disorder (including depression, major depression, dysthymia, or minor depression)
    - Yes/No
- Respondents that reported having diabetes. If yes, what type of diabetes it is/was.
    - Yes / No
    - No, pre-diabetes or borderline diabetes
    - Yes, but female told only during pregnancy
- Respondents that reported having arthritis
    - Yes/No
- Gender
    - Female/Male
- Age Category
    - 18-24
    - 25-29
    - 30-34
    - 35-39
    - 40-44
    - 45-49
    - 50-54
    - 55-59
    - 60-64
    - 65-69
    - 70-74
    - 75-79
    - 80+
- Height (cm)
    - Numerical
- Weight (kg)

    - Numerical
- BMI
    - Numerical
- Smoking history
    - Numerical
- Alcohol Consumption (days within the last month)
    - Numerical
- Fruit Consumption (days within the last month)
    - Numerical
- Green Vegetable Consumption (days within the last month)
    - Numerical
- Fried Potato Consumption (days within the last month)
    - Numerical

Glimpse of the Data (First 7 Observations):

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_.cm.
1	Poor	Within the past 2 years	No	No	No	No	No	No	Yes	Female	70-74	150
2	Very Good	Within the past year	No	Yes	No	No	No	Yes	No	Female	70-74	165
3	Very Good	Within the past year	Yes	No	No	No	No	Yes	No	Female	60-64	163
4	Poor	Within the past year	Yes	Yes	No	No	No	Yes	No	Male	75-79	180
5	Good	Within the past year	No	No	No	No	No	No	No	Male	80+	191
6	Good	Within the past year	No	No	No	No	Yes	No	Yes	Male	60-64	183
7	Fair	Within the past year	Yes	Yes	No	No	No	No	Yes	Male	60-64	175

Weight_.kg.	BMI	Smoking_History	Alcohol_Consumption	Fruit_Consumption	Green_Vegetables_Consumption	FriedPotato_Consumption
32.66	14.54	Yes	0	30	16	12
77.11	28.29	No	0	30	0	4
88.45	33.47	No	4	12	3	16
93.44	28.73	No	0	30	30	8
88.45	24.37	Yes	0	8	4	0
154.22	46.11	No	0	12	12	12
69.85	22.74	Yes	0	16	8	0

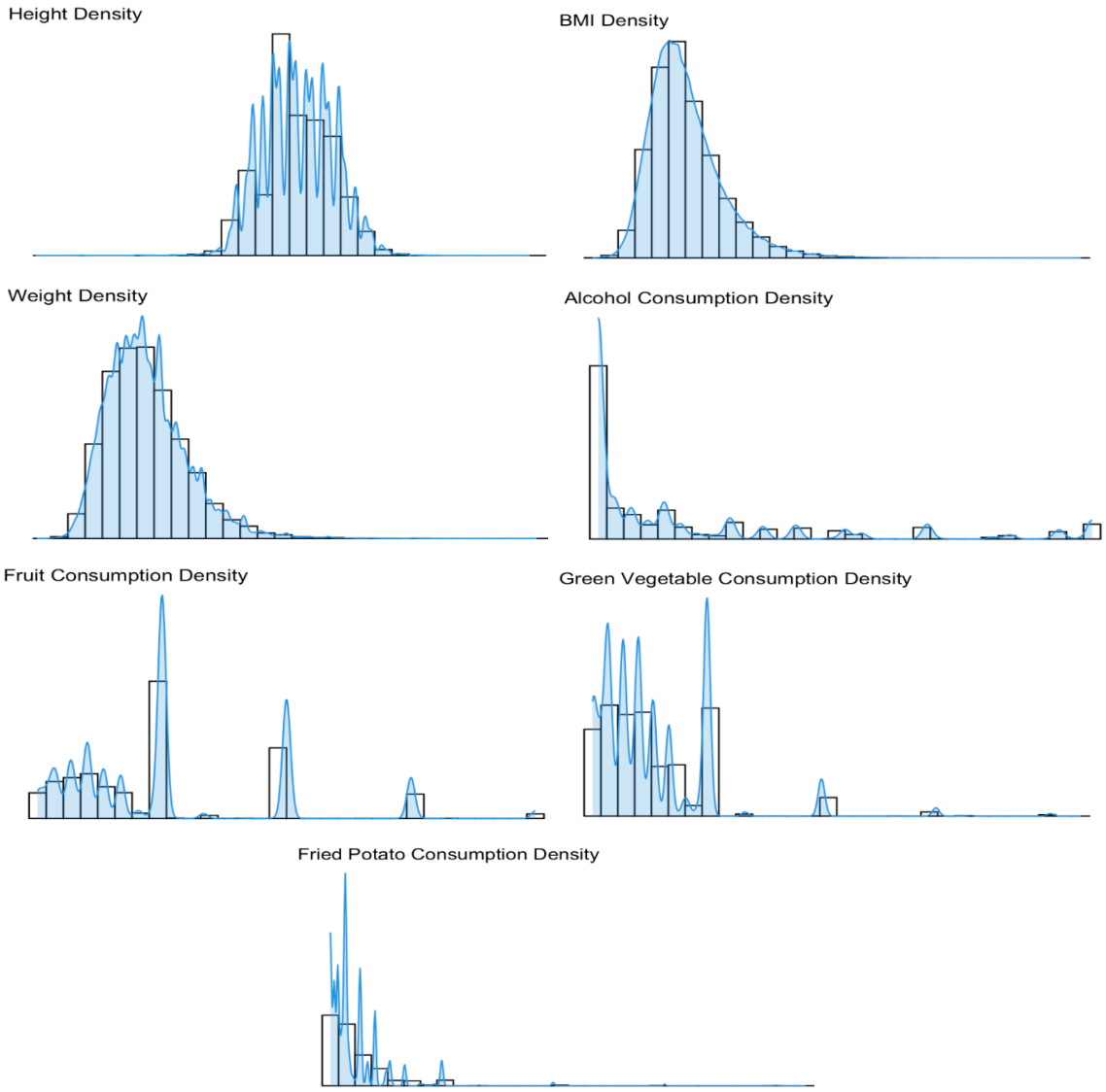
Exploratory Data Analysis

Descriptive Statistics of Variables

(After removal of 80 duplicates)

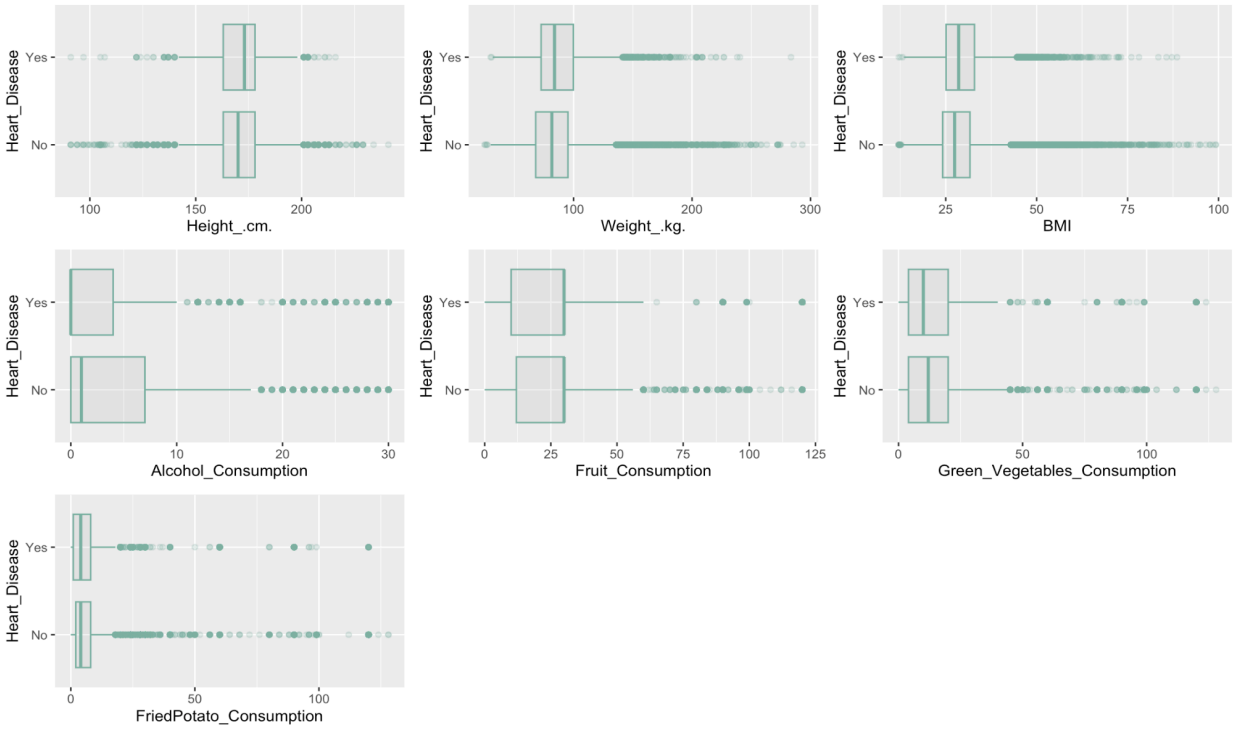
```
 General_Health       Checkup            Exercise         Heart_Disease
 Length:308774      Length:308774      Length:308774      Length:308774
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Skin_Cancer        Other_Cancer        Depression          Diabetes
 Length:308774      Length:308774      Length:308774      Length:308774
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


  Arthritis             Sex            Age_Category        Height_.cm.
 Length:308774      Length:308774      Length:308774      Min.   : 91.0
 Class :character   Class :character   Class :character   1st Qu.:163.0
 Mode  :character   Mode  :character   Mode  :character   Median :170.0
                                                          Mean   :170.6
                                                          3rd Qu.:178.0
                                                          Max.   :241.0
  Weight_.kg.          BMI        Smoking_History    Alcohol_Consumption
 Min.   : 24.95   Min.   :12.02   Length:308774      Min.   : 0.000
 1st Qu.: 68.04   1st Qu.:24.21   Class :character   1st Qu.: 0.000
 Median : 81.65   Median :27.44   Mode  :character   Median : 1.000
 Mean   : 83.59   Mean   :28.63                      Mean   : 5.098
 3rd Qu.: 95.25   3rd Qu.:31.85                      3rd Qu.: 6.000
 Max.   :293.02   Max.   :99.33                      Max.   :30.000
 Fruit_Consumption Green_Vegetables_Consumption FriedPotato_Consumption
 Min.   :  0.00    Min.   :  0.00               Min.   :  0.000
 1st Qu.: 12.00    1st Qu.:  4.00               1st Qu.:  2.000
 Median : 30.00    Median : 12.00               Median :  4.000
 Mean   : 29.83    Mean   : 15.11               Mean   :  6.297
 3rd Qu.: 30.00    3rd Qu.: 20.00               3rd Qu.:  8.000
 Max.   :120.00    Max.   :128.00               Max.   :128.000
```

Distributions of Numerical Variables:

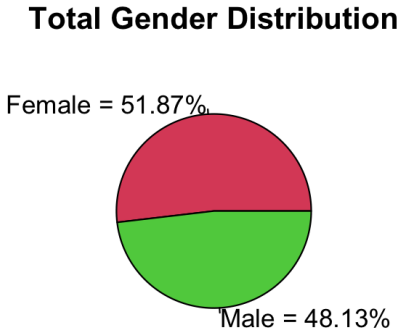


We can see that Height has a pretty normal distribution. Height, BMI, and Weight have a clearly visible bell shaped curve. Alcohol Consumption, Fruit Consumption, Green Vegetable Consumption, and Fried Potato Consumption have clear right skews. Fruit Consumption is pretty bimodal with 2 significant spikes in the plot.

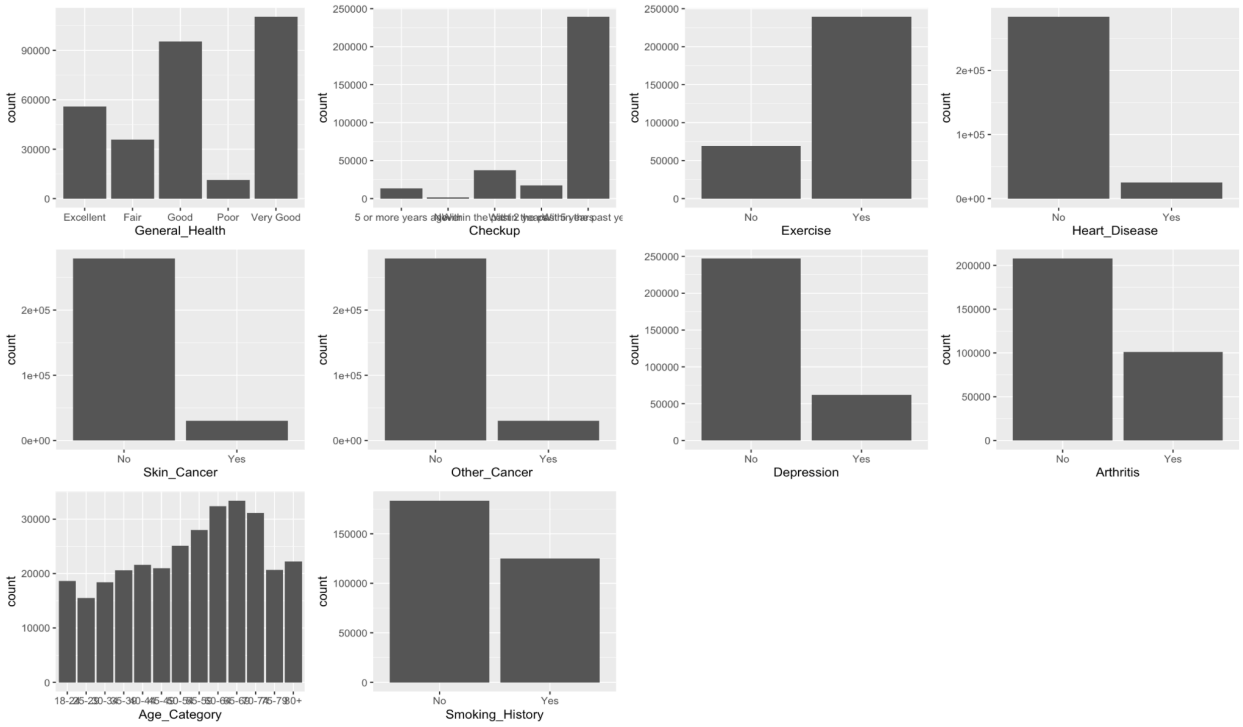
Numerical Variable Distribution in Box Plot Format:



Gender Distribution

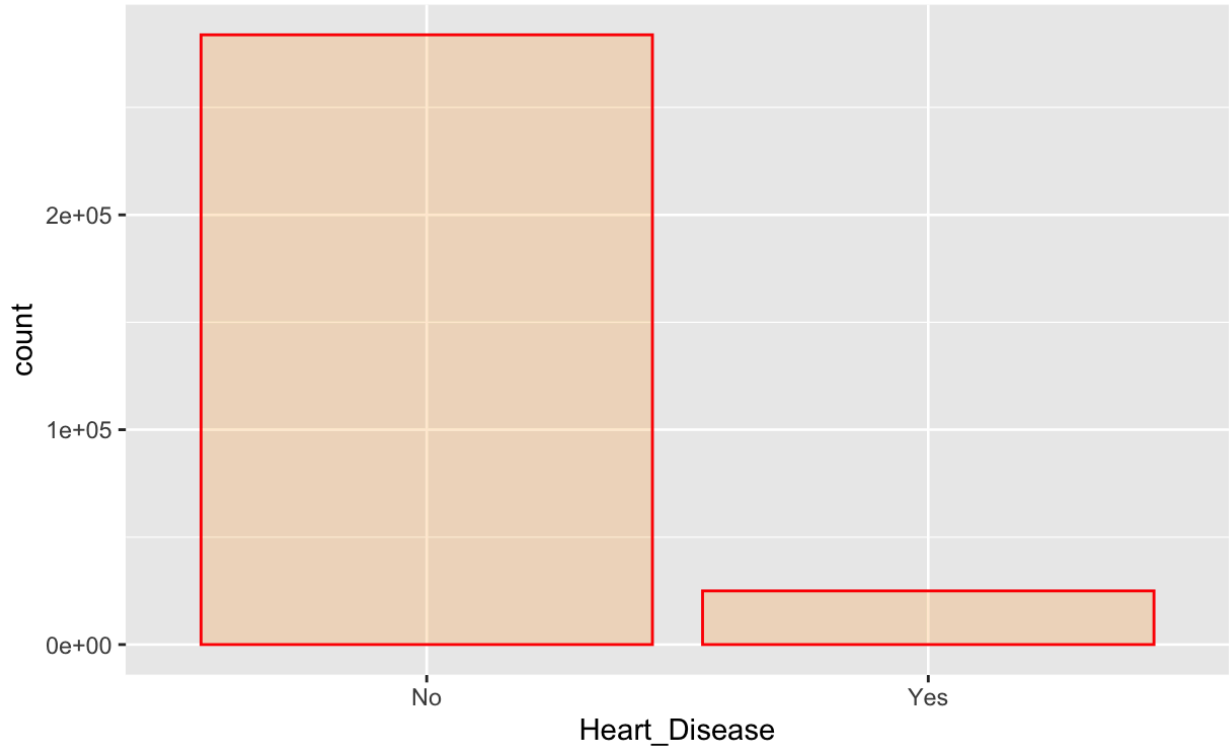


Distribution of Categorical Variables in the Dataset (Used counts of the variables):



Distribution of Response Variable

For my response variable, I calculated the distribution by count of who claims they have a history of heart disease, and who claims they don’t have a heart disease history.



As we can see from this plot, a lot more people do not have a history of heart disease compared to those who do.

Statistical Models Used

Since my response variable is binary, I used a logistic regression model to perform statistical significance.

Full Model

First, I performed logistic regression on the full data set with all variables taken into account.

```
glm(formula = HD_Int ~ . - Heart_Disease, family = "binomial",
    data = cvd)

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                  -6.272e+00  4.939e-01 -12.698
General_HealthFair                            1.707e+00  3.551e-02  48.059
General_HealthGood                            1.069e+00  3.362e-02  31.811
General_HealthPoor                            2.250e+00  3.995e-02  56.319
General_HealthVery Good                       5.127e-01  3.443e-02  14.890
CheckupNever                                  2.612e-01  1.531e-01   1.706
CheckupWithin the past 2 years                2.262e-01  6.353e-02   3.560
CheckupWithin the past 5 years                1.297e-01  7.485e-02   1.732
CheckupWithin the past year                   5.235e-01  5.787e-02   9.047
ExerciseYes                                  -1.805e-02  1.643e-02  -1.099
Skin_CancerYes                                1.213e-01  1.975e-02   6.142
Other_CancerYes                               4.802e-02  1.945e-02   2.468
DepressionYes                                 2.448e-01  1.812e-02  13.509
DiabetesNo, pre-diabetes or borderline diabetes  1.462e-01  4.120e-02   3.548
DiabetesYes                                   5.333e-01  1.695e-02  31.461
DiabetesYes, but female told only during pregnancy  1.292e-01  1.091e-01   1.184
ArthritisYes                                  2.606e-01  1.534e-02  16.986
SexMale                                       8.324e-01  2.098e-02  39.672
Age_Category25-29                             3.063e-01  1.404e-01   2.182
Age_Category30-34                             6.373e-01  1.259e-01   5.063
Age_Category35-39                             7.547e-01  1.206e-01   6.260
Age_Category40-44                             1.085e+00  1.149e-01   9.441
Age_Category45-49                             1.463e+00  1.114e-01  13.139
Age_Category50-54                             1.777e+00  1.085e-01  16.381
Age_Category55-59                             2.116e+00  1.069e-01  19.789
Age_Category60-64                             2.356e+00  1.062e-01  22.184
Age_Category65-69                             2.598e+00  1.059e-01  24.523
Age_Category70-74                             2.858e+00  1.059e-01  26.991
Age_Category75-79                             3.097e+00  1.064e-01  29.104
Age_Category80+                               3.373e+00  1.062e-01  31.753
Height_.cm.                                  -5.191e-03  2.827e-03  -1.836
Weight_.kg.                                   3.153e-04  2.564e-03   0.123
BMI                                           8.352e-04  7.401e-03   0.113
Smoking_HistoryYes                            3.958e-01  1.481e-02  26.735
Alcohol_Consumption                          -9.856e-03  9.141e-04 -10.782
Fruit_Consumption                             3.171e-05  3.105e-04   0.102
Green_Vegetables_Consumption                  7.983e-04  5.295e-04   1.508
FriedPotato_Consumption                      -6.098e-04  8.664e-04  -0.704

                                             Pr(>|z|)
(Intercept)                                   < 2e-16 ***
General_HealthFair                            < 2e-16 ***
General_HealthGood                            < 2e-16 ***
General_HealthPoor                            < 2e-16 ***
General_HealthVery Good                       < 2e-16 ***
CheckupNever                                 0.087919 .
CheckupWithin the past 2 years               0.000371 ***
CheckupWithin the past 5 years               0.083194 .
CheckupWithin the past year                   < 2e-16 ***
ExerciseYes                                  0.271973
Skin_CancerYes                               8.15e-10 ***
Other_CancerYes                              0.013570 *
DepressionYes                                 < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes  0.000388 ***
DiabetesYes                                   < 2e-16 ***
DiabetesYes, but female told only during pregnancy 0.236584
ArthritisYes                                  < 2e-16 ***
SexMale                                       < 2e-16 ***
Age_Category25-29                            0.029134 *
Age_Category30-34                            4.13e-07 ***
Age_Category35-39                            3.85e-10 ***
Age_Category40-44                             < 2e-16 ***
Age_Category45-49                             < 2e-16 ***
Age_Category50-54                             < 2e-16 ***
Age_Category55-59                             < 2e-16 ***
Age_Category60-64                             < 2e-16 ***
Age_Category65-69                             < 2e-16 ***
Age_Category70-74                             < 2e-16 ***
Age_Category75-79                             < 2e-16 ***
Age_Category80+                               < 2e-16 ***
Height_.cm.                                  0.066368 .
Weight_.kg.                                  0.902127
BMI                                          0.910157
Smoking_HistoryYes                            < 2e-16 ***
Alcohol_Consumption                           < 2e-16 ***
Fruit_Consumption                            0.918677
Green_Vegetables_Consumption                 0.131636
FriedPotato_Consumption                      0.481547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136957  on 308736  degrees of freedom
AIC: 137033

Number of Fisher Scoring iterations: 7
```

Checkup and Diabetes are selected to test overall because the model outputs they have more than 2 levels.

From the model we can see that Checkup and Diabetes have at least one factor level that has a high p-value. For this, I'm going to conduct a test on the overall variables to determine whether or not they're significant to predicting heart disease.

```
Analysis of Deviance Table (Type II tests)

Response: HD_Int
        Df Chisq Pr(>Chisq)
Checkup  4  2460  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Deviance Table (Type II tests)

Response: HD_Int
         Df  Chisq Pr(>Chisq)
Diabetes  3 9187.5  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both Checkup and Diabetes p-values are small, indicating significance to Heart Disease. Based on this information, certain levels of these variables can still be labeled as non-significant, but the overall variables should not be removed from the model.

Note: Took out Heart_Disease because I changed that from character to integer to make the response variable HD_Int for modeling.

Next, I performed Multicollinearity Analysis on the model

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.293889	4	1.032731
Checkup	1.063265	4	1.007698
Exercise	1.171483	1	1.082351
Skin_Cancer	1.085002	1	1.041634
Other_Cancer	1.059698	1	1.029416
Depression	1.147095	1	1.071025
Diabetes	1.155402	3	1.024367
Arthritis	1.137940	1	1.066743
Sex	2.095054	1	1.447430
Age_Category	1.366381	12	1.013092
Height_.cm.	17.958276	1	4.237721
Weight_.kg.	61.822837	1	7.862750
BMI	48.305413	1	6.950210
Smoking_History	1.054379	1	1.026830
Alcohol_Consumption	1.079800	1	1.039134
Fruit_Consumption	1.104466	1	1.050936
Green_Vegetables_Consumption	1.100718	1	1.049151
FriedPotato_Consumption	1.036166	1	1.017922

Since there are 3 high VIF values, I chose to remove Height, Weight, and BMI from the model.

Reduced Model 1 (Removing Height, Weight, and BMI)

```
glm(formula = HD_Int ~ . - Heart_Disease - Height_.cm. - Weight_.kg. -
    BMI, family = "binomial", data = cvd)

Coefficients:
                                             Estimate Std. Error z value
(Intercept)                                -7.079e+00  1.227e-01 -57.689
General_HealthFair                          1.715e+00  3.540e-02  48.457
General_HealthGood                          1.076e+00  3.353e-02  32.083
General_HealthPoor                          2.258e+00  3.989e-02  56.592
General_HealthVery Good                     5.149e-01  3.440e-02  14.969
CheckupNever                                2.702e-01  1.530e-01   1.766
CheckupWithin the past 2 years              2.282e-01  6.352e-02   3.593
CheckupWithin the past 5 years              1.308e-01  7.485e-02   1.748
CheckupWithin the past year                 5.265e-01  5.784e-02   9.103
ExerciseYes                                -2.223e-02  1.635e-02  -1.360
Skin_CancerYes                              1.167e-01  1.973e-02   5.915
Other_CancerYes                             4.609e-02  1.945e-02   2.370
DepressionYes                               2.458e-01  1.811e-02  13.576
DiabetesNo, pre-diabetes or borderline diabetes     1.535e-01  4.111e-02   3.734
DiabetesYes                                 5.380e-01  1.664e-02  32.325
DiabetesYes, but female told only during pregnancy  1.343e-01  1.091e-01   1.231
ArthritisYes                                2.626e-01  1.523e-02  17.244
SexMale                                     7.601e-01  1.541e-02  49.309
Age_Category25-29                           3.113e-01  1.404e-01   2.218
Age_Category30-34                           6.429e-01  1.258e-01   5.109
Age_Category35-39                           7.612e-01  1.205e-01   6.318
Age_Category40-44                           1.092e+00  1.148e-01   9.505
Age_Category45-49                           1.470e+00  1.113e-01  13.207
Age_Category50-54                           1.785e+00  1.084e-01  16.461
Age_Category55-59                           2.124e+00  1.069e-01  19.866
Age_Category60-64                           2.363e+00  1.062e-01  22.259
Age_Category65-69                           2.606e+00  1.059e-01  24.606
Age_Category70-74                           2.868e+00  1.059e-01  27.091
Age_Category75-79                           3.109e+00  1.064e-01  29.223
Age_Category80+                             3.386e+00  1.061e-01  31.903
Smoking_HistoryYes                          3.931e-01  1.478e-02  26.593
Alcohol_Consumption                        -1.006e-02  9.129e-04 -11.022
Fruit_Consumption                           7.103e-06  3.105e-04   0.023
Green_Vegetables_Consumption                7.560e-04  5.297e-04   1.427
FriedPotato_Consumption                    -6.783e-04  8.667e-04  -0.783

                                           Pr(>|z|)
(Intercept)                                 < 2e-16 ***
General_HealthFair                          < 2e-16 ***
General_HealthGood                          < 2e-16 ***
General_HealthPoor                          < 2e-16 ***
General_HealthVery Good                     < 2e-16 ***
CheckupNever                               0.077397 .
CheckupWithin the past 2 years             0.000327 ***
CheckupWithin the past 5 years             0.080453 .
CheckupWithin the past year                 < 2e-16 ***
ExerciseYes                                0.173913
Skin_CancerYes                             3.31e-09 ***
Other_CancerYes                            0.017778 *
DepressionYes                               < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes    0.000188 ***
DiabetesYes                                 < 2e-16 ***
DiabetesYes, but female told only during pregnancy 0.218496
ArthritisYes                                < 2e-16 ***
SexMale                                     < 2e-16 ***
Age_Category25-29                          0.026573 *
Age_Category30-34                          3.24e-07 ***
Age_Category35-39                          2.65e-10 ***
Age_Category40-44                           < 2e-16 ***
Age_Category45-49                           < 2e-16 ***
Age_Category50-54                           < 2e-16 ***
Age_Category55-59                           < 2e-16 ***
Age_Category60-64                           < 2e-16 ***
Age_Category65-69                           < 2e-16 ***
Age_Category70-74                           < 2e-16 ***
Age_Category75-79                           < 2e-16 ***
Age_Category80+                             < 2e-16 ***
Smoking_HistoryYes                          < 2e-16 ***
Alcohol_Consumption                         < 2e-16 ***
Fruit_Consumption                          0.981750
Green_Vegetables_Consumption               0.153487
FriedPotato_Consumption                    0.433863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136986  on 308739  degrees of freedom
AIC: 137056

Number of Fisher Scoring iterations: 7
```

Non-Significant Factors:

- Checkup (Never & Past 5 yrs.)
- Exercise
- Diabetes (Pregnancy)
- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

The AIC increased compared to the full model, meaning that the full model is less complex than this reduced model without Height, Weight, and BMI.

Conducting multicollinearity analysis, we can conclude that there is no sign of multicollinearity in this reduced model.

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.281180	4	1.031457
Checkup	1.060951	4	1.007423
Exercise	1.159146	1	1.076637
Skin_Cancer	1.082643	1	1.040501
Other_Cancer	1.058826	1	1.028993
Depression	1.145372	1	1.070221
Diabetes	1.111833	3	1.017825
Arthritis	1.121720	1	1.059113
Sex	1.130849	1	1.063414
Age_Category	1.269022	12	1.009976
Smoking_History	1.051306	1	1.025332
Alcohol_Consumption	1.076278	1	1.037438
Fruit_Consumption	1.104214	1	1.050816
Green_Vegetables_Consumption	1.100307	1	1.048955
FriedPotato_Consumption	1.035179	1	1.017437

Stepwise Model (STEP AIC on Full Model)

I decided to use STEP AIC on the full model to observe which variables would be removed using that method.

```
glm(formula = HD_Int ~ General_Health + Checkup + Skin_Cancer +
    Other_Cancer + Depression + Diabetes + Arthritis + Sex +
    Age_Category + Height_.cm. + Weight_.kg. + Smoking_History +
    Alcohol_Consumption + Green_Vegetables_Consumption, family = "binomial",
    data = cvd)

Coefficients:
                                             Estimate Std. Error z value
(Intercept)                                -6.2311969  0.2005056 -31.077
General_HealthFair                          1.7096980  0.0352880  48.450
General_HealthGood                          1.0702314  0.0335476  31.902
General_HealthPoor                          2.2560762  0.0394223  57.228
General_HealthVery Good                     0.5123969  0.0344188  14.887
CheckupNever                                0.2623874  0.1530742   1.714
CheckupWithin the past 2 years              0.2260931  0.0635142   3.560
CheckupWithin the past 5 years              0.1293801  0.0748481   1.729
CheckupWithin the past year                 0.5234075  0.0578510   9.048
Skin_CancerYes                              0.1207921  0.0197426   6.118
Other_CancerYes                             0.0479837  0.0194509   2.467
DepressionYes                               0.2456192  0.0181038  13.567
DiabetesNo, pre-diabetes or borderline diabetes     0.1462278  0.0412005   3.549
DiabetesYes                                 0.5340382  0.0169280  31.548
DiabetesYes, but female told only during pregnancy  0.1283153  0.1091454   1.176
ArthritisYes                                0.2607693  0.0153372  17.002
SexMale                                     0.8303160  0.0208292  39.863
Age_Category25-29                           0.3065668  0.1404123   2.183
Age_Category30-34                           0.6379956  0.1258801   5.068
Age_Category35-39                           0.7555672  0.1205478   6.268
Age_Category40-44                           1.0867726  0.1149207   9.457
Age_Category45-49                           1.4656747  0.1113442  13.163
Age_Category50-54                           1.7798942  0.1084610  16.410
Age_Category55-59                           2.1195917  0.1068943  19.829
Age_Category60-64                           2.3592314  0.1061230  22.231
Age_Category65-69                           2.6013223  0.1058589  24.573
Age_Category70-74                           2.8618893  0.1058114  27.047
Age_Category75-79                           3.1015121  0.1063255  29.170
Age_Category80+                             3.3787104  0.1060891  31.848
Height_.cm.                                -0.0055565  0.0010177  -5.460
Weight_.kg.                                 0.0006379  0.0003995   1.597
Smoking_HistoryYes                          0.3962494  0.0147515  26.862
Alcohol_Consumption                        -0.0098946  0.0009135 -10.832
Green_Vegetables_Consumption                0.0007387  0.0005111   1.445

                                           Pr(>|z|)
(Intercept)                                 < 2e-16 ***
General_HealthFair                          < 2e-16 ***
General_HealthGood                          < 2e-16 ***
General_HealthPoor                          < 2e-16 ***
General_HealthVery Good                     < 2e-16 ***
CheckupNever                               0.086507 .
CheckupWithin the past 2 years             0.000371 ***
CheckupWithin the past 5 years             0.083886 .
CheckupWithin the past year                 < 2e-16 ***
Skin_CancerYes                             9.46e-10 ***
Other_CancerYes                            0.013628 *
DepressionYes                               < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes    0.000386 ***
DiabetesYes                                 < 2e-16 ***
DiabetesYes, but female told only during pregnancy 0.239740
ArthritisYes                                < 2e-16 ***
SexMale                                     < 2e-16 ***
Age_Category25-29                          0.029011 *
Age_Category30-34                          4.01e-07 ***
Age_Category35-39                          3.66e-10 ***
Age_Category40-44                           < 2e-16 ***
Age_Category45-49                           < 2e-16 ***
Age_Category50-54                           < 2e-16 ***
Age_Category55-59                           < 2e-16 ***
Age_Category60-64                           < 2e-16 ***
Age_Category65-69                           < 2e-16 ***
Age_Category70-74                           < 2e-16 ***
Age_Category75-79                           < 2e-16 ***
Age_Category80+                             < 2e-16 ***
Height_.cm.                                4.77e-08 ***
Weight_.kg.                                0.110351
Smoking_HistoryYes                          < 2e-16 ***
Alcohol_Consumption                         < 2e-16 ***
Green_Vegetables_Consumption               0.148370
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136959  on 308740  degrees of freedom
AIC: 137027

Number of Fisher Scoring iterations: 7
```

Removed Variables

- Exercise
- BMI

- Fruit Consumption
- Fried Potato Consumption

Non-Significant Factors
- Checkup (Never & Past 5 yrs.)
- Diabetes (Pregnancy)
- Weight
- Green Vegetables Consumption

Seeing as the AIC decreased from the full model, we can conclude that this model is less complex than the full model.

Results of Data Analysis

ANOVA on Full Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
General_Health	4	1429	357.3	5497.475	< 2e-16	***
Checkup	4	119	29.7	456.776	< 2e-16	***
Exercise	1	14	14.0	215.936	< 2e-16	***
Skin_Cancer	1	128	127.9	1968.513	< 2e-16	***
Other_Cancer	1	36	36.3	558.612	< 2e-16	***
Depression	1	8	8.2	126.209	< 2e-16	***
Diabetes	3	268	89.4	1375.067	< 2e-16	***
Arthritis	1	106	105.7	1627.085	< 2e-16	***
Sex	1	164	163.7	2519.636	< 2e-16	***
Age_Category	12	559	46.6	716.544	< 2e-16	***
Height_.cm.	1	1	0.9	14.016	0.000181	***
Weight_.kg.	1	1	1.5	23.032	1.59e-06	***
BMI	1	0	0.1	0.877	0.348892	
Smoking_History	1	44	44.0	677.229	< 2e-16	***
Alcohol_Consumption	1	10	9.8	150.987	< 2e-16	***
Fruit_Consumption	1	0	0.0	0.653	0.418894	
Green_Vegetables_Consumption	1	0	0.2	2.889	0.089176	.
FriedPotato_Consumption	1	0	0.1	0.841	0.359101	
Residuals	308736	20064	0.1			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Non-Significant Factors
- BMI
- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

ANOVA on Reduced Model 1

```
                             Df Sum Sq Mean Sq  F value Pr(>F)
General_Health                4   1429   357.3 5496.895 <2e-16 ***
Checkup                       4    119    29.7  456.728 <2e-16 ***
Exercise                      1     14    14.0  215.914 <2e-16 ***
Skin_Cancer                   1    128   127.9 1968.306 <2e-16 ***
Other_Cancer                  1     36    36.3  558.553 <2e-16 ***
Depression                    1      8     8.2  126.195 <2e-16 ***
Diabetes                      3    268    89.4 1374.922 <2e-16 ***
Arthritis                     1    106   105.7 1626.913 <2e-16 ***
Sex                           1    164   163.7 2519.370 <2e-16 ***
Age_Category                 12    559    46.6  716.469 <2e-16 ***
Smoking_History               1     44    44.3  681.413 <2e-16 ***
Alcohol_Consumption           1     10     9.6  148.111 <2e-16 ***
Fruit_Consumption             1      0     0.1    0.847 0.3573
Green_Vegetables_Consumption  1      0     0.2    3.163 0.0753 .
FriedPotato_Consumption       1      0     0.1    1.326 0.2495
Residuals                308739  20067     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Non-Significant Factors

- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

ANOVA on Stepwise Model

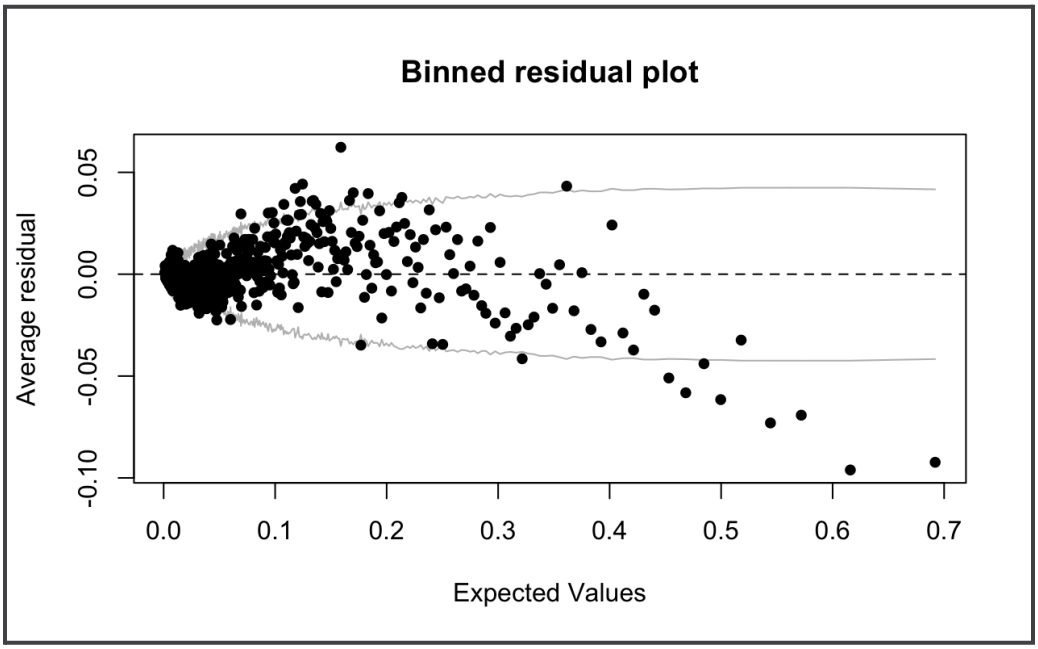
```
                             Df Sum Sq Mean Sq  F value   Pr(>F)
General_Health                4   1429   357.3 5497.361  < 2e-16 ***
Checkup                       4    119    29.7  456.767  < 2e-16 ***
Skin_Cancer                   1    127   127.1 1955.380  < 2e-16 ***
Other_Cancer                  1     37    36.9  567.510  < 2e-16 ***
Depression                    1      8     7.6  117.540  < 2e-16 ***
Diabetes                      3    276    91.9 1413.357  < 2e-16 ***
Arthritis                     1    108   108.2 1664.915  < 2e-16 ***
Sex                           1    160   160.3 2466.710  < 2e-16 ***
Age_Category                 12    566    47.2  725.694  < 2e-16 ***
Height_.cm.                   1      1     1.0   15.338 8.99e-05 ***
Weight_.kg.                   1      1     1.2   19.167 1.20e-05 ***
Smoking_History               1     45    44.7  687.230  < 2e-16 ***
Alcohol_Consumption           1     10    10.1  154.808  < 2e-16 ***
Green_Vegetables_Consumption  1      0     0.2    2.468    0.116
Residuals                308740  20065     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Non-Significant Factor

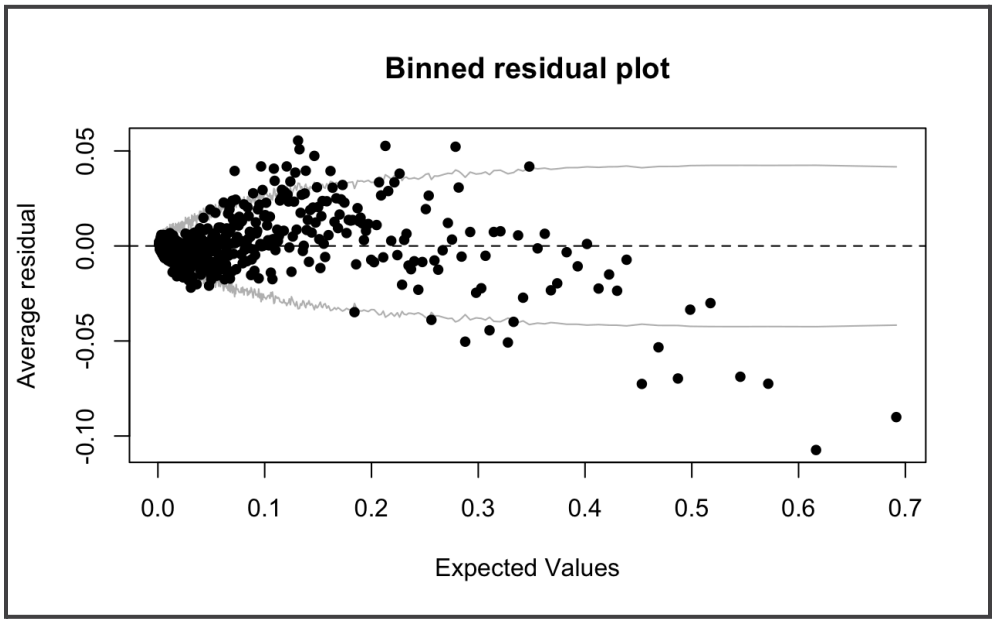
- Green Vegetables Consumption

In order to evaluate the normality and variance of our model, we needed to observe the residuals. Since I chose a logistic regression model, a normal residuals vs. fitted plot was not going to be useful, therefore we chose to create a Binned Residuals Plot in order to observe the residuals' behavior

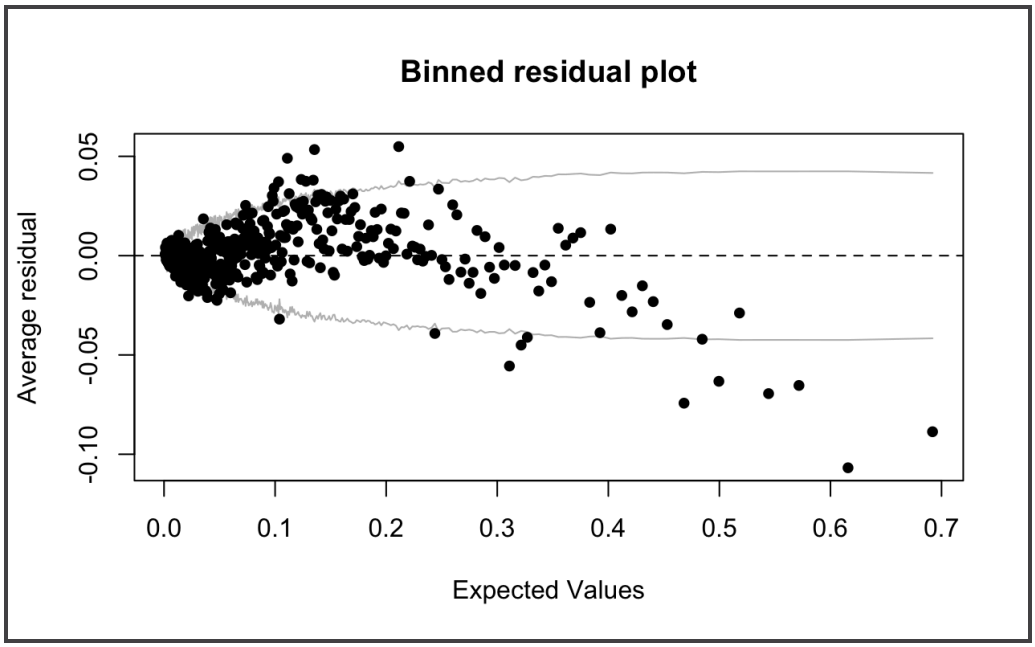
Full Model - Binned Residuals Plot



Reduced Model 1 - Binned Residuals Plot



Stepwise Model - Binned Residuals Plot



We can identify many outliers in all of the residual plots. Leading to the conclusion that there could be unexpected factors that are influencing the adequacy of these models.

Rationale of Fitted Model

Model Accuracy

Model Accuracy (Full Model)

- Accuracy: 0.9193261
- Area Under Curve: 0.8351

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,311	1,492
Heart Disease	23,418	1,553

Model Accuracy (Reduced Model 1)

- Accuracy: 0.9193876
- Area Under Curve: 0.835

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,350	1,453
Heart Disease	23,438	1,533

Model Accuracy (Stepwise Model)

- Accuracy: 0.9193488
- Area Under Curve: 0.8351

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,317	1,486
Heart Disease	23,417	1,554

The model accuracies and areas under the curve I achieved were significantly high, leading me to conclude that the models are all pretty useful for predicting the effects of heart disease.

Conclusions and Limitations

In terms of model accuracy, the best-to-worst of the 3 models are:

1. Reduced Model 1
2. Stepwise Model
3. Full Model

For future use, it would be possible to obtain better performing results with other techniques like Ridge or Lasso Regression. Since some of the variables did not seem to have a perfect normal distribution, another method would be to transform the data to have a more standardized pattern. Seeing as there were many outliers in the residuals, I could also suggest that to go more in depth of the data, further methods could be to target more specific characteristics to look into, such as combinations of factors (interaction terms).

At the end of my code, I did try some initial log transformations on the models, but I did not include this in the final presentation as I got pretty similar results from all my previous models, so I am not fully confident in that regard that I found the best possible model.

Appendix (R Code)

Loading data

```{r}
cvd <- read.csv("~/Desktop/UC_Riverside/2023-2024/Spring/STAT_183/Individual Project/CVD_cleaned.csv")
```

Loading Libraries

```{r}
library(ggplot2)
library(dplyr)
# install.packages(gridExtra)
library(gridExtra) # Plots
library(tidyr)
library(MASS) # Stepwise Regression
library(car) # Multicollinearity (vif)
library(arm) # Binned Plotted Residuals
library(pROC) # Correlation Matrix
library(reshape2) # Correlation
```

Glimpse of the data:

```{r}
glimpse(cvd)
```

Summary Statistics

```{r}
summary(cvd)
```

Variable classifications

```{r}
str(cvd)
```

Data Cleaning:

* Remove any NA or duplicate observations

```{r}
# Checking for NA or Duplicates
sum(is.na(cvd))
sum(duplicated(cvd))
```

* There are 80 duplicates, need to remove these observations

```{r}
cvd <- unique(cvd)
sum(duplicated(cvd))
```

```{r}
summary(cvd) # Summary Statistics
```

```{r}
glimpse(cvd)
```

* Changing Categorical variables to binary 0/1 integers
* Making factor levels for the predictor variables

```{r}
# Changing Character variables to Factors
# General Health
cvd$General_Health <- as.factor(cvd$General_Health)

# Checkup
cvd$Checkup <- as.factor(cvd$Checkup)

# Exercise
```

```
# n_distinct(cvd$Exercise) # 2
cvd$Exercise <- as.factor(cvd$Exercise)

# Heart Disease
n_distinct(cvd$Heart_Disease) # 2
cvd$HD_Int <- ifelse(cvd$Heart_Disease == "Yes", 1, 0)

# # Skin Cancer
# n_distinct(cvd$Skin_Cancer) # 2
cvd$Skin_Cancer <- as.factor(cvd$Skin_Cancer)
#
# # Other Cancer
# n_distinct(cvd$Other_Cancer) # 2
cvd$Other_Cancer <- as.factor(cvd$Other_Cancer)
#
# # Depression
# n_distinct(cvd$Depression) # 2
cvd$Depression <- as.factor(cvd$Depression)
#
# # Diabetes
# n_distinct(cvd$Diabetes) # 4
cvd$Diabetes <- as.factor(cvd$Diabetes)
#
# # Arthritis
# n_distinct(cvd$Arthritis)
cvd$Arthritis <- as.factor(cvd$Arthritis)
#
# # Sex
# n_distinct(cvd$Sex)
cvd$Sex <- as.factor(cvd$Sex)

# Age Category
cvd$Age_Category <- as.factor(cvd$Age_Category)

# Smoking History
cvd$Smoking_History <- as.factor(cvd$Smoking_History)

# head(cvd, 5)
str(cvd)
```

Exploratory Data Analysis (EDA)

```{r}
# Gender Pie Chart for distribution
gender <- table(cvd$Sex)
gender_labels <- paste0(rownames(gender), " = ", round(100 * gender/sum(gender), 2), "%")

pie(gender, col = 2:3, labels = gender_labels, main = "Total Gender Distribution")
```

Gender vs. Target

```{r}
ggplot(data = cvd, aes(x = Heart_Disease), fill = Sex) +
  geom_bar(color="red", fill="orange", alpha=0.2)
```

Based on this plot, we can see that a majority of the dataset contains people who do not have any history of heart disease.

* Distribution of Numerical Variables

```{r}
pl1 <- ggplot(data = cvd, aes(x=Height_.cm., y=Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl2 <- ggplot(data = cvd, aes(Weight_.kg., Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl3 <- ggplot(data = cvd, aes(BMI, Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl4 <- ggplot(data = cvd, aes(Alcohol_Consumption, Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl5 <- ggplot(data = cvd, aes(Fruit_Consumption, Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl6 <- ggplot(data = cvd, aes(Green_Vegetables_Consumption, Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

pl7 <- ggplot(data = cvd, aes(FriedPotato_Consumption, Heart_Disease)) +
  geom_boxplot(color="#69b3a2", fill="gray", alpha=0.2)

grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,pl7)
```

```

* Distributions of Numerical Variables (Density Histograms)

```{r}
pl1 <- ggplot(data = cvd, aes(x = Height_.cm., y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Height Density") +
  xlab("Height (cm)") + ylab("Density") +
  theme_void()

pl2 <- ggplot(data = cvd, aes(x = Weight_.kg., y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Weight Density") +
  xlab("Weight (kg)") + ylab("Density") +
  theme_void()

pl3 <- ggplot(data = cvd, aes(x = BMI, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("BMI Density") +
  xlab("BMI") + ylab("Density") +
  theme_void()

pl4 <- ggplot(data = cvd, aes(x = Alcohol_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Alcohol Consumption Density") +
  xlab("Number of Days Alcohol was Consumed within Last 30 Days") + ylab("Density") +
  theme_void()

pl5 <- ggplot(data = cvd, aes(x = Fruit_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Fruit Consumption Density") +
  xlab("Fruit Consumption in last 30 days") + ylab("Density") +
  theme_void()

pl6 <- ggplot(data = cvd, aes(x = Green_Vegetables_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
```
```

```
  ggtitle("Green Vegetable Consumption Density") +
  xlab("Green Vegetable Consumption in last 30 days") + ylab("Density") +
  theme_void()

pl7 <- ggplot(data = cvd, aes(x = FriedPotato_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Fried Potato Consumption Density") +
  xlab("Fried Potato Consumption in last 30 days") + ylab("Density") +
  theme_void()

grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,pl7)
```

* Distributions of Categorical Variables

```{r}
pl1 <- ggplot(data = cvd, aes(x = General_Health)) +
  geom_bar()

pl2 <- ggplot(data = cvd, aes(x = Checkup)) +
  geom_bar()

pl3 <- ggplot(data = cvd, aes(x = Exercise)) +
  geom_bar()

pl4 <- ggplot(data = cvd, aes(x = Heart_Disease)) +
  geom_bar()

pl5 <- ggplot(data = cvd, aes(x = Skin_Cancer)) +
  geom_bar()

pl6 <- ggplot(data = cvd, aes(x = Other_Cancer)) +
  geom_bar()

pl7 <- ggplot(data = cvd, aes(x = Depression)) +
  geom_bar()

pl8 <- ggplot(data = cvd, aes(x = Arthritis)) +
  geom_bar()

pl9 <- ggplot(data = cvd, aes(x = Age_Category)) +
  geom_bar()
```

```
pl10 <- ggplot(data = cvd, aes(x = Smoking_History)) +
  geom_bar()

grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,pl7,pl8,pl9,pl10, ncol = 4)
```

```{r}
table(cvd$Heart_Disease)
```

Significantly more people without Heart Disease than with. Comparing 24,000s to 280,000s.

Beginning Logistic Regression Modeling

```{r}
# Factor variables need to have at least 2 levels

model1 <- glm(formula = HD_Int ~ . - Heart_Disease, data = cvd, family = "binomial")
summary(model1)
```

* Contains all 18 predictor variables.

Wald test on predictors that have at least one level where p-value > alpha

```{r, warning = F}
# Wald Test on Checkup
model <- glm(HD_Int ~ Checkup, data = cvd, family = binomial)
Anova(model, type = "II", test = "Wald")
cat("--------------------------------------------------------------------------------", "\n")

# Wald Test on Diabetes
model <- glm(HD_Int ~ Diabetes, data = cvd, family = binomial)
Anova(model, type = "II", test = "Wald")
```

* Overall P-Values for both Checkup and Diabetes are small, so we can conclude they are significant to the model.

Wald test on Exercise

```{r}
# Wald Test on Diabetes
model <- glm(HD_Int ~ Exercise, data = cvd, family = binomial)
Anova(model, type = "II", test = "Wald")
```

ANOVA

```{r}
anova <- aov(model1)
summary(anova)
```

Binned Residuals Plot

```{r}
binnedplot(fitted(model1),
        residuals(model1, type = "response"),
        nclass = NULL,
        xlab = "Expected Values",
        ylab = "Average residual",
        main = "Binned residual plot",
        cex.pts = 0.8,
        col.pts = 1,
        col.int = "gray")
```

Accuracy of Full Model

```{r}
prediction <- ifelse(model1$fitted.values > 0.5, "pos", "neg")

confusion_matrix <- table(cvd$HD_Int, prediction)
rownames(confusion_matrix) <- c("No Heart Disease", "Heart Disease")
colnames(confusion_matrix) <- c("Predicted No Heart Disease", "Predicted Heart Disease")
confusion_matrix

accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
```

accuracy

# Area under curve, 1 indicates perfect predictive model
auc(cvd$HD_Int~model1$fitted.values, data = cvd)
```

## Multicollinearity Analysis

```{r}
vif(model1)
```

* By squaring the GVIF^(1/(2*Df)) values, we can conclude that there is a high multicollinearity problem regarding Height, Weight, and BMI.
* Reduced_model1 is removing Heart_Disease, Height, Weight, and BMI

```{r}
# Reduced model after VIF

reduced_model1 <- glm(formula = HD_Int ~ . - Heart_Disease - Height_.cm. - Weight_.kg. - BMI, data = cvd, family = "binomial")
summary(reduced_model1)
```

* Increased AIC from 137038 to 137061, so the full model is better than without 3 variables.

* Confirm low VIF values

```{r}
vif(reduced_model1)
```

* No sign of multicollinearity.
* Results in 15 variables.

Binned Residuals Plot

```{r}
binnedplot(fitted(reduced_model1),
        residuals(reduced_model1, type = "response"),
        nclass = NULL,
        xlab = "Expected Values",
        ylab = "Average residual",
        main = "Binned residual plot",

```
        cex.pts = 0.8,
        col.pts = 1,
        col.int = "gray")
```

* Visually, lots of outliers exist.
* Outliers can be caused by some unknown outside factors.

ANOVA

```{r}
anova <- aov(reduced_model1)
summary(anova)
```

Accuracy of Reduced Model 1

```{r}
prediction <- ifelse(reduced_model1$fitted.values > 0.5, "pos", "neg")
#
confusion_matrix <- table(cvd$HD_Int, prediction)
rownames(confusion_matrix) <- c("No Heart Disease", "Heart Disease")
colnames(confusion_matrix) <- c("Predicted No Heart Disease", "Predicted Heart Disease")
confusion_matrix

accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
accuracy

# Area under curve, 1 indicates perfect predictive model
auc(cvd$HD_Int~reduced_model1$fitted.values, data = cvd)
```

Stepwise Regression Model

* Performs stepwise regression on the full model

```{r warning=F}
model2 <- stepAIC(model1, direction = "both", trace = F) # On full model
summary(model2)
```

* Lowered AIC from full model 137033 to 137027.
* This model only contains 14 variables: General_Health, Checkup, Skin_Cancer, Other_Cancer, Depression, Diabetes, Arthritis, Sex, Age_Category, Height_.cm., Weight_.kg., Smoking_History, Alcohol_Consumption, Green_Vegetables_Consumption

ANOVA of Stepwise Model

```{r}
anova <- aov(model2)
summary(anova)
```

Binned Residuals Plot

```{r}
binnedplot(fitted(model2),
        residuals(model2, type = "response"),
        nclass = NULL,
        xlab = "Expected Values",
        ylab = "Average residual",
        main = "Binned residual plot",
        cex.pts = 0.8,
        col.pts = 1,
        col.int = "gray")
```

* Visually, lots of outliers exist.
* Outliers can be caused by some unknown outside factors.

Accuracy of Stepwise Model

```{r}
prediction <- ifelse(model2$fitted.values > 0.5, "pos", "neg")

confusion_matrix <- table(cvd$HD_Int, prediction)
rownames(confusion_matrix) <- c("No Heart Disease", "Heart Disease")
colnames(confusion_matrix) <- c("Predicted No Heart Disease", "Predicted Heart Disease")
confusion_matrix

accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
accuracy
```

```
# Area under curve, 1 indicates perfect predictive model
auc(cvd$HD_Int~model2$fitted.values, data = cvd)
```

```{r}
vif(model2)
```

* No sign of multicollinearity.

Log Transform Data (Did not remove duplicate observations)

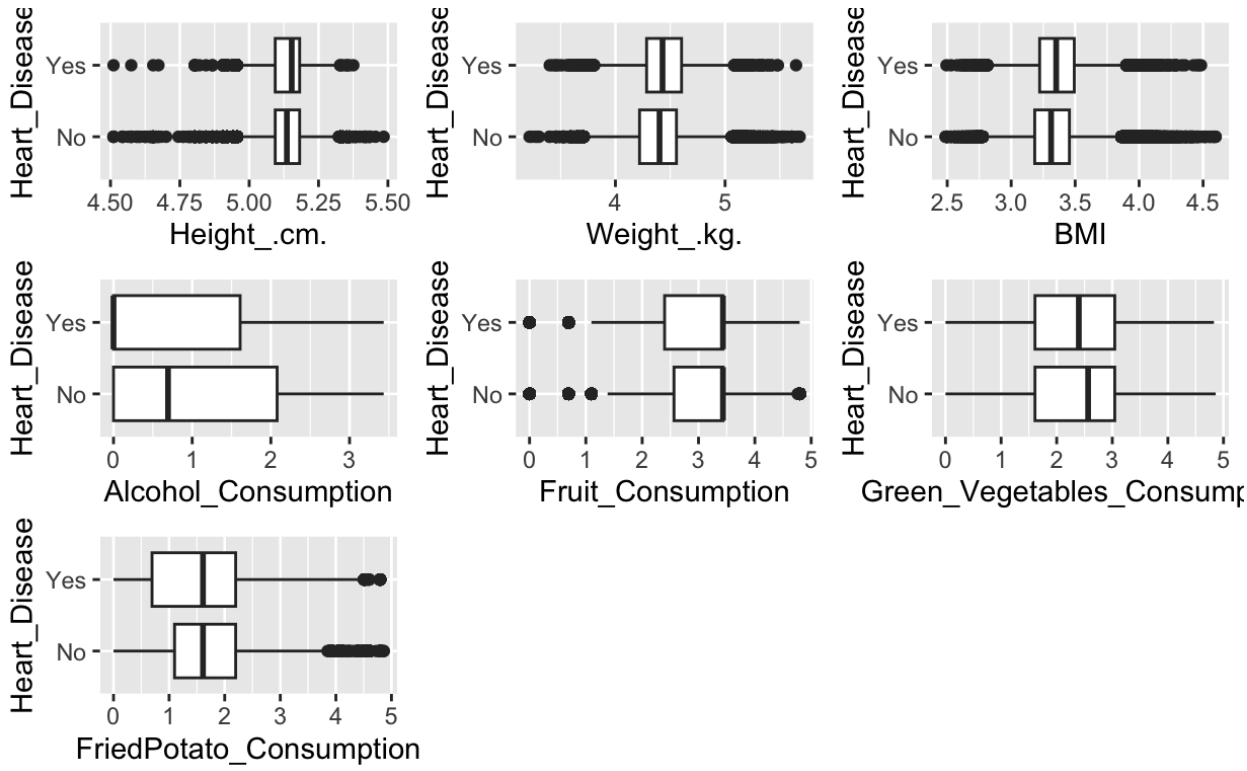
* Caution, with the log transformation, now it's possible to obtain -Infinity values because we have zero values in our numerical data
* Going to shift the data so that values can stay positive

```{r}
new_cvd <- cvd
new_cvd$Alcohol_Consumption <- new_cvd$Alcohol_Consumption + 1
new_cvd$Fruit_Consumption <- new_cvd$Fruit_Consumption + 1
new_cvd$Green_Vegetables_Consumption <- new_cvd$Green_Vegetables_Consumption + 1
new_cvd$FriedPotato_Consumption <- new_cvd$FriedPotato_Consumption + 1
new_cvd$HD_Int <- new_cvd$HD_Int + 1

new_cvd <- new_cvd %>%
  mutate_at(vars(Height_.cm.,
Weight_.kg.,BMI,Alcohol_Consumption,Fruit_Consumption,Green_Vegetables_Consumption,,Fr
iedPotato_Consumption, HD_Int), ~log(.))

glimpse(new_cvd)
```

Boxplot Distributions after log transformation

```{r}
pl1 <- ggplot(data = new_cvd, aes(Height_.cm., Heart_Disease)) +
  geom_boxplot()

pl2 <- ggplot(data = new_cvd, aes(Weight_.kg., Heart_Disease)) +
```

```
  geom_boxplot()

pl3 <- ggplot(data = new_cvd, aes(BMI, Heart_Disease)) +
  geom_boxplot()

pl4 <- ggplot(data = new_cvd, aes(Alcohol_Consumption, Heart_Disease)) +
  geom_boxplot()

pl5 <- ggplot(data = new_cvd, aes(Fruit_Consumption, Heart_Disease)) +
  geom_boxplot()

pl6 <- ggplot(data = new_cvd, aes(Green_Vegetables_Consumption, Heart_Disease)) +
  geom_boxplot()

pl7 <- ggplot(data = new_cvd, aes(FriedPotato_Consumption, Heart_Disease)) +
  geom_boxplot()

grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,pl7)
```

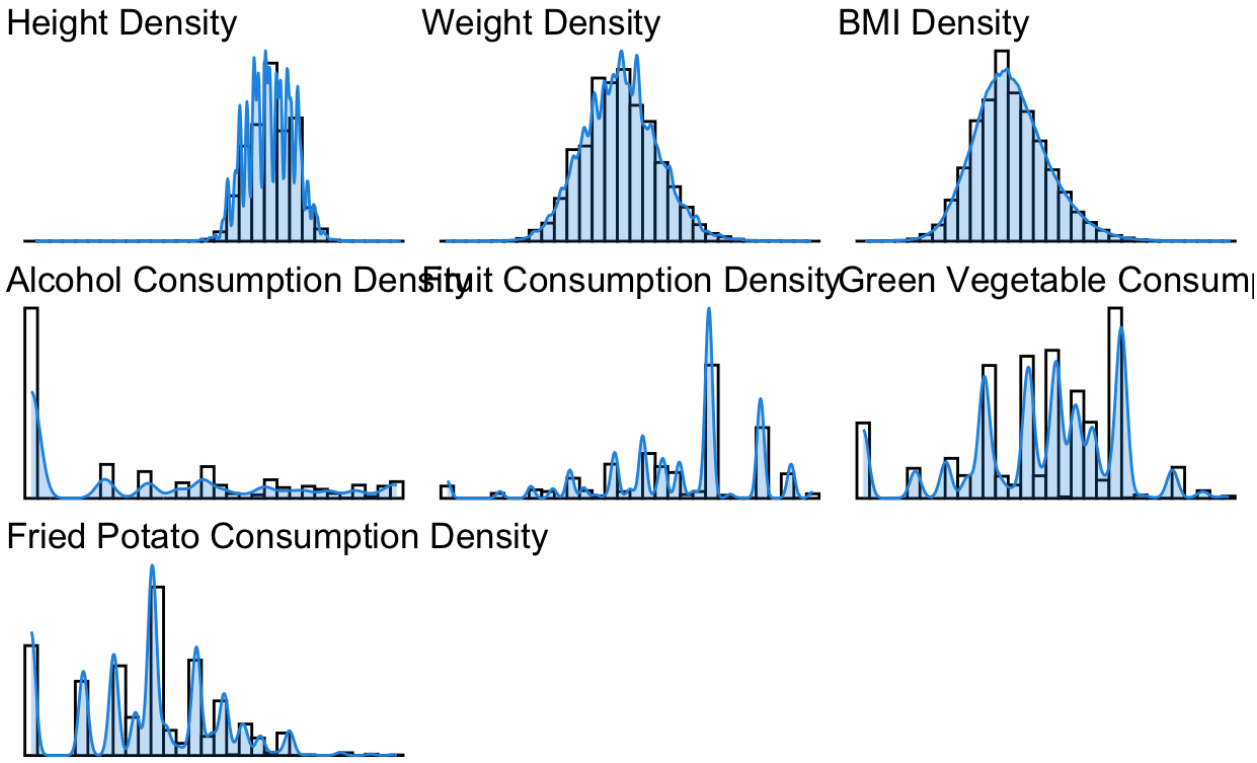


* Density Distributions of Numerical Variables (After Log-Transformation)

```{r}
pl1 <- ggplot(data = new_cvd, aes(x = Height_.cm., y = after_stat(density))) +
```

```
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Height Density") +
  xlab("Height (cm)") + ylab("Density") +
  theme_void()

pl2 <- ggplot(data = new_cvd, aes(x = Weight_.kg., y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Weight Density") +
  xlab("Weight (kg)") + ylab("Density") +
  theme_void()

pl3 <- ggplot(data = new_cvd, aes(x = BMI, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("BMI Density") +
  xlab("BMI") + ylab("Density") +
  theme_void()

pl4 <- ggplot(data = new_cvd, aes(x = Alcohol_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Alcohol Consumption Density") +
  xlab("Number of Days Alcohol was Consumed within Last 30 Days") + ylab("Density") +
  theme_void()

pl5 <- ggplot(data = new_cvd, aes(x = Fruit_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Fruit Consumption Density") +
  xlab("Fruit Consumption in last 30 days") + ylab("Density") +
  theme_void()

pl6 <- ggplot(data = new_cvd, aes(x = Green_Vegetables_Consumption, y =
after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Green Vegetable Consumption Density") +
  xlab("Green Vegetable Consumption in last 30 days") + ylab("Density") +
  theme_void()

pl7 <- ggplot(data = new_cvd, aes(x = FriedPotato_Consumption, y = after_stat(density))) +
  geom_histogram(fill = "white", colour = "black") +
```

```
  geom_density(color = 4, fill = 4, alpha = 0.25) +
  ggtitle("Fried Potato Consumption Density") +
  xlab("Fried Potato Consumption in last 30 days") + ylab("Density") +
  theme_void()

grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,pl7)
```



New model with log transformations:

```{r}
# Heart Disease to int

newmodel1 <- glm(formula = HD_Int ~ . - Heart_Disease, data = new_cvd, family = "binomial")
summary(newmodel1)
```

```
Warning: non-integer #successes in a binomial glm!
Call:
glm(formula = HD_Int ~ . - Heart_Disease, family = "binomial",
    data = new_cvd)

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                   8.241458  22.299084   0.370
General_HealthFair                            1.620707   0.041981  38.605
General_HealthGood                            1.042361   0.039922  26.110
General_HealthPoor                            2.080274   0.046532  44.706
General_HealthVery Good                       0.509998   0.040928  12.461
CheckupNever                                  0.258636   0.179356   1.442
CheckupWithin the past 2 years                0.224127   0.074935   2.991
CheckupWithin the past 5 years                0.127670   0.088328   1.445
CheckupWithin the past year                   0.507304   0.068356   7.421
ExerciseYes                                  -0.008066   0.019007  -0.424
Skin_CancerYes                                0.107914   0.022586   4.778
Other_CancerYes                               0.044336   0.022233   1.994
DepressionYes                                 0.226784   0.020808  10.899
DiabetesNo, pre-diabetes or borderline diabetes 0.133730   0.047601   2.809
DiabetesYes                                   0.474987   0.019453  24.417
DiabetesYes, but female told only during pregnancy 0.109719   0.128442   0.854
ArthritisYes                                  0.243117   0.017752  13.695
SexMale                                       0.773716   0.024279  31.867
Age_Category25-29                             0.314635   0.168407   1.868
Age_Category30-34                             0.642072   0.150944   4.254
Age_Category35-39                             0.760509   0.144546   5.261
Age_Category40-44                             1.091785   0.137771   7.925
Age_Category45-49                             1.467499   0.133482  10.994
Age_Category50-54                             1.774008   0.130062  13.640
Age_Category55-59                             2.102377   0.128216  16.397
Age_Category60-64                             2.328093   0.127335  18.283
Age_Category65-69                             2.551564   0.127035  20.086
Age_Category70-74                             2.790487   0.126960  21.979
Age_Category75-79                             3.007278   0.127515  23.584
Age_Category80+                               3.254988   0.127293  25.571
Height_.cm.                                  -3.357538   4.843630  -0.693
Weight_.kg.                                   1.329745   2.432495   0.547
BMI                                          -1.241810   2.433065  -0.510
Smoking_HistoryYes                            0.370634   0.017161  21.598
Alcohol_Consumption                          -0.081557   0.007638 -10.678
Fruit_Consumption                            -0.013945   0.008534  -1.634
Green_Vegetables_Consumption                  0.017844   0.008555   2.086
FriedPotato_Consumption                      -0.016374   0.008874  -1.845
```

```
                                                Pr(>|z|)
(Intercept)                                      0.71169
General_HealthFair                               < 2e-16 ***
General_HealthGood                               < 2e-16 ***
General_HealthPoor                               < 2e-16 ***
General_HealthVery Good                          < 2e-16 ***
CheckupNever                                     0.14929
CheckupWithin the past 2 years                   0.00278 **
CheckupWithin the past 5 years                   0.14834
CheckupWithin the past year                     1.16e-13 ***
ExerciseYes                                      0.67129
Skin_CancerYes                                  1.77e-06 ***
Other_CancerYes                                  0.04613 *
DepressionYes                                    < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes  0.00496 **
DiabetesYes                                      < 2e-16 ***
DiabetesYes, but female told only during pregnancy  0.39298
ArthritisYes                                     < 2e-16 ***
SexMale                                          < 2e-16 ***
Age_Category25-29                                0.06172 .
Age_Category30-34                               2.10e-05 ***
Age_Category35-39                               1.43e-07 ***
Age_Category40-44                               2.29e-15 ***
Age_Category45-49                                < 2e-16 ***
Age_Category50-54                                < 2e-16 ***
Age_Category55-59                                < 2e-16 ***
Age_Category60-64                                < 2e-16 ***
Age_Category65-69                                < 2e-16 ***
Age_Category70-74                                < 2e-16 ***
Age_Category75-79                                < 2e-16 ***
Age_Category80+                                  < 2e-16 ***
Height_.cm.                                      0.48819
Weight_.kg.                                      0.58461
BMI                                              0.60978
Smoking_HistoryYes                               < 2e-16 ***
Alcohol_Consumption                              < 2e-16 ***
Fruit_Consumption                                0.10225
Green_Vegetables_Consumption                     0.03700 *
FriedPotato_Consumption                          0.06501 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102582  on 308773  degrees of freedom
Residual deviance:  78291  on 308736  degrees of freedom
AIC: 140671

Number of Fisher Scoring iterations: 8
```

```{r}
# str(new_cvd)
```

```

```{r}
anova <- aov(newmodel1)
summary(anova)
```

```
                                Df Sum Sq Mean Sq  F value   Pr(>F)    
General_Health                   4    687  171.65 5497.561  < 2e-16 ***
Checkup                          4     57   14.26  456.783  < 2e-16 ***
Exercise                         1      7    6.74  215.940  < 2e-16 ***
Skin_Cancer                      1     61   61.46 1968.544  < 2e-16 ***
Other_Cancer                     1     17   17.44  558.621  < 2e-16 ***
Depression                       1      4    3.94  126.211  < 2e-16 ***
Diabetes                         3    129   42.94 1375.089  < 2e-16 ***
Arthritis                        1     51   50.80 1627.111  < 2e-16 ***
Sex                              1     79   78.67 2519.676  < 2e-16 ***
Age_Category                    12    268   22.37  716.555  < 2e-16 ***
Height_.cm.                      1      0    0.35   11.126 0.000851 ***
Weight_.kg.                      1      0    0.35   11.189 0.000823 ***
BMI                              1      0    0.01    0.254 0.614297    
Smoking_History                  1     21   21.23  679.991  < 2e-16 ***
Alcohol_Consumption              1      5    5.05  161.713  < 2e-16 ***
Fruit_Consumption                1      0    0.00    0.004 0.947213    
Green_Vegetables_Consumption     1      0    0.34   11.013 0.000905 ***
FriedPotato_Consumption          1      0    0.00    0.107 0.743751    
Residuals                   308736   9640    0.03                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

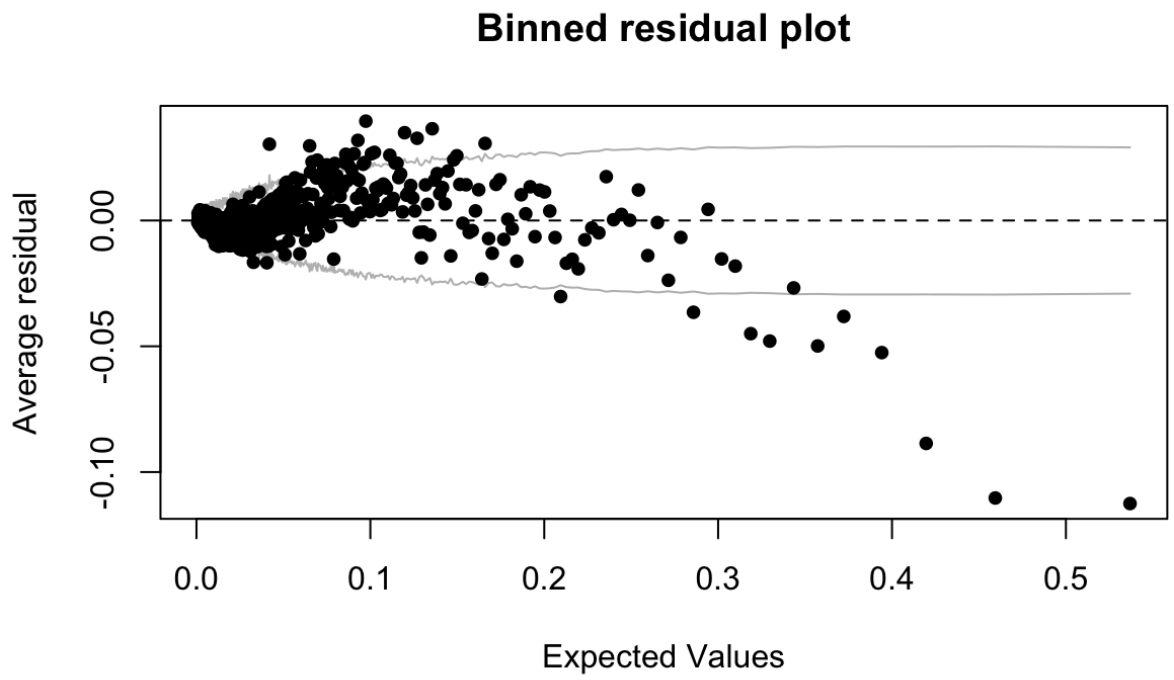
```{r}
vif(newmodel1)
```
```

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.323397	4	1.035646
Checkup	1.064685	4	1.007866
Exercise	1.191087	1	1.091369
Skin_Cancer	1.087019	1	1.042602
Other_Cancer	1.060991	1	1.030044
Depression	1.145421	1	1.070243
Diabetes	1.170209	3	1.026543
Arthritis	1.140383	1	1.067887
Sex	2.095214	1	1.447485
Age_Category	1.374702	12	1.013348
Height_.cm.	1384.279784	1	37.205911
Weight_.kg.	5276.069872	1	72.636560
BMI	3990.345739	1	63.169183
Smoking_History	1.059146	1	1.029148
Alcohol_Consumption	1.117056	1	1.056909
Fruit_Consumption	1.144861	1	1.069982
Green_Vegetables_Consumption	1.156101	1	1.075221
FriedPotato_Consumption	1.066637	1	1.032781

* By squaring the GVIF^(1/(2*Df)) values, we can conclude that there is a high multicollinearity problem regarding Height, Weight, and BMI.

Binned Residuals of Log Transformed

```{r}
binnedplot(fitted(newmodel1),
        residuals(newmodel1, type = "response"),
        nclass = NULL,
        xlab = "Expected Values",
        ylab = "Average residual",
        main = "Binned residual plot",
        cex.pts = 0.8,
        col.pts = 1,
        col.int = "gray")
```



* Still not performing great in terms of outliers.
* Outliers can be caused by some unknown outside factors.

Accuracy of Log-Transformed Model

```{r}
prediction <- ifelse(newmodel1$fitted.values > 0.5, "pos", "neg")

confusion_matrix <- table(cvd$HD_Int, prediction)
rownames(confusion_matrix) <- c("No Heart Disease", "Heart Disease")
colnames(confusion_matrix) <- c("Predicted No Heart Disease", "Predicted Heart Disease")
confusion_matrix

accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
accuracy

# Area under curve, 1 indicates perfect predictive model
auc(cvd$HD_Int~newmodel1$fitted.values, data = cvd)
```

```
                 prediction
                  Predicted No Heart Disease Predicted Heart Disease
  No Heart Disease                     283639                     164
  Heart Disease                         24712                     259
[1] 0.9194362
Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.8351
```