**b)** Brain Stroke Dataset Link:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

---

**c)**

Objective: Determine if there is an association between some of a person's physical characteristics and whether they've had a stroke. Specifically, if a person's age, gender, BMI, average glucose level, their history of smoking, and whether they have hypertension or a heart disease are associated with whether they've had a stroke.

Study Design: Observational. The data was taken from the Electronic Health Records of the Center for Medicare and Medicade Services. The data originally contained 5110 randomly selected patients. Entries containing "Unknown" smoking status, missing BMI, and "Other" for gender were filtered out, leaving 3425 patients. The data includes information on the patients' gender, age, body mass index (BMI) value, average glucose level in blood, history of hypertension, heart disease, smoking status, whether they've had a stroke, and a few other categories that will not be used.

Sampling Model: Multinomial sample

Analysis Unit: One patient

Variables (and measurement type):

- Response:
    - Stroke Status (Categorical/Binomial; 0 if the patient has not had a stroke, 1 if the patient has had a stroke)
- Explanatory:
    - Gender (Categorical/Binomial; "Male" or "Female")
    - Hypertension Status (Categorical/Binomial; 0 if the patient doesn't have hypertension, 1 if the patient has it)
    - Heart Disease Status (Categorical/Binomial; 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease)

- Smoking Status (Categorical/Nominal; “Never Smoked”, “Formerly Smoked”, “Smokes”)
- Age (Numerical/Continuous; in years)
- BMI (Numerical/Continuous)
- Average Glucose Level (Numerical/Continuous; in milligrams per deciliter (mg/dL))

---

**d)**

Separate marginal 2x2 tables were formed using the categorical explanatory variables (Gender, Hypertension, Heart Disease, and Smoking Status) and stroke status to determine if there was any marginal association between individual factors and stroke status (i.e. testing for statistical independence). Row proportions and mosaic plots were then taken for visual comparison.

```
        Stroke
Gender      0    1
  Female 1981  105
  Male   1264   75
```

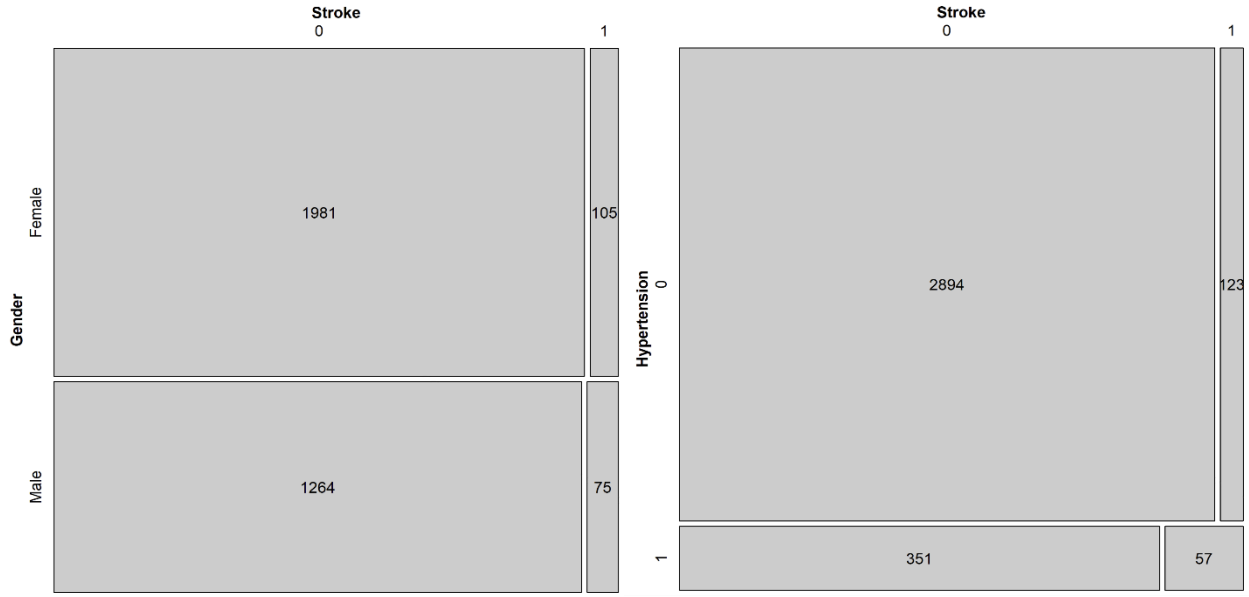
Gender

```
        Stroke
Gender       0      1
  Female 0.9497 0.0503
  Male   0.9440 0.0560
```

Hypertension

```
              Stroke
Hypertension     0    1
           0  2894  123
           1   351   57
```

```
              Stroke
Hypertension      0      1
           0 0.9592 0.0408
           1 0.8603 0.1397
```



```
               Stroke
Heart Disease     0    1
            0  3075  144
            1   170   36
```

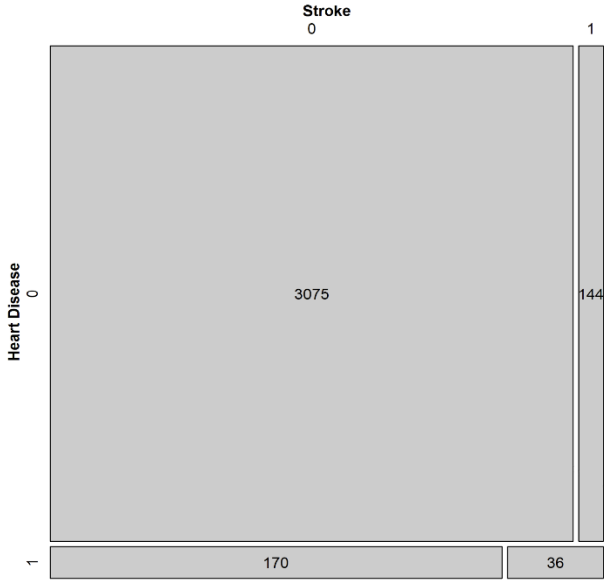
Heart Disease

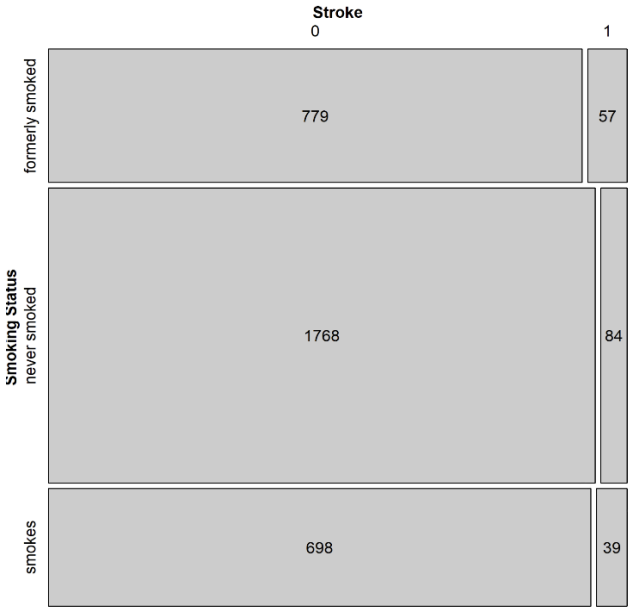
```
               Stroke
Heart Disease      0      1
            0 0.9553 0.0447
            1 0.8252 0.1748
```

Smoking Status

```
                    Stroke
Smoking Status         0    1
  formerly smoked    779   57
  never smoked      1768   84
  smokes             698   39
```

```
                    Stroke
Smoking Status         0      1
  formerly smoked 0.9318 0.0682
  never smoked    0.9546 0.0454
  smokes          0.9471 0.0529
```




Fisher's exact test was then used on the Gender, Hypertension, and Heart Disease tables, whereas both Pearson and LR tests were used on the Smoking Status table.

```
        Fisher's Exact Test for Count Data

data:  gstr.table
p-value = 0.4806
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8138991 1.5335715
sample estimates:
odds ratio
  1.119422

> mosaic(hypstr.table, labeling = labeling_values)
> fisher.test(hypstr.table)

        Fisher's Exact Test for Count Data

data:  hypstr.table
p-value = 3.135e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.685802 5.381413
sample estimates:
odds ratio
  3.819012
```

Gender

Hypertension

```
        Fisher's Exact Test for Count
        Data

data:  hdstr.table
p-value = 3.324e-11
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.949069 6.789554
sample estimates:
odds ratio
  4.518546

> chisq.test(smstr.table) #No Yates continuity correction needed

        Pearson's Chi-squared test

data:  smstr.table
X-squared = 6.0293, df = 2,
p-value = 0.04906

> GTest(smstr.table)

        Log likelihood ratio (G-test)
        test of independence without
        correction

data:  smstr.table
G = 5.7835, X-squared df = 2,
p-value = 0.05548
```

Heart Disease

Smoking Status

According to these tests, there is no evidence of marginal association between gender and stroke status (p-val > 0.1), extremely strong evidence of marginal association between hypertension and stroke status as well as heart disease and stroke status (p-val < 10e-10)), and moderate evidence of marginal association between smoking status and stroke status (p-val close to 0.05 for both tests).

---

For the continuous variables Age, BMI, and Average Glucose Level (abbr. AGL), a logistic regression model was created for each of them as x and the probability of having a stroke as pi(x).

```
> #Age and Stroke Status
> astr.fit <- glm(stroke~age, family = binomial, data = stroke)
> summary(astr.fit)

Call:
glm(formula = stroke ~ age, family = binomial, data = stroke)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.386079   0.413419  -17.87   <2e-16 ***
age          0.076109   0.006051   12.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1180.3  on 3423  degrees of freedom
AIC: 1184.3

Number of Fisher Scoring iterations: 7
```

$$logit\left(\hat{\pi}(x)\right) = -7.3860 + 0.0761x$$

```
> #BMI and Stroke Status
> bmistr.fit <- glm(stroke~bmi, family = binomial, data = stroke)
> summary(bmistr.fit)

Call:
glm(formula = stroke ~ bmi, family = binomial, data = stroke)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.103070   0.321963  -9.638   <2e-16 ***
bmi          0.006932   0.010210   0.679    0.497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1410.4  on 3423  degrees of freedom
AIC: 1414.4

Number of Fisher Scoring iterations: 5
```

$$logit\left(\hat{\pi}(x)\right) = -3.1031 + 0.0069x$$

```
> #Avg Glucose Level and Stroke Statuss
> aglstr.fit <- glm(stroke~avg_glucose_level, family = binomial, data = stroke)
> summary(aglstr.fit)

Call:
glm(formula = stroke ~ avg_glucose_level, family = binomial,
    data = stroke)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.109552   0.188746 -21.773  < 2e-16 ***
avg_glucose_level  0.010116   0.001286   7.868  3.6e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1355.3  on 3423  degrees of freedom
AIC: 1359.3

Number of Fisher Scoring iterations: 6
```

$$logit(\hat{\pi}(x)) = -4.1096 + 0.0101x$$

For each of these, a Wald (above) and LR (below) test was used to check for association.

```
> drop1(astr.fit, test = "LRT")
Single term deletions

Model:
stroke ~ age
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      1180.3 1184.3
age     1   1410.9 1412.9 230.58 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(bmistr.fit, test = "LRT")
Single term deletions

Model:
stroke ~ bmi
       Df Deviance    AIC     LRT Pr(>Chi)
<none>      1410.4 1414.4
bmi     1   1410.9 1412.9 0.45251   0.5011
> drop1(aglstr.fit, test = "LRT")
Single term deletions

Model:
stroke ~ avg_glucose_level
                 Df Deviance    AIC    LRT Pr(>Chi)
<none>                1355.3 1359.3
avg_glucose_level  1   1410.9 1412.9 55.623 8.78e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Age

BMI

AGL

According to these tests, there is extremely strong evidence of an association between Age and Stroke Status as well as AGL and Stroke Status (p-val < 10e-13), and no evidence of an association between BMI and Stroke Status (p-val > 0.1).

---

**e)**

Since the response variable was binomial and multiple variables both categorical and continuous are involved in the objective, a linear regression (logit) model was used to plot the relationship between the given explanatory variables and stroke status. The sample size of 3425 is more than enough to fit all 7 predictors plus any 2 factor interaction terms. Below (next page) is the chosen model:

```
Call:  glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
    family = binomial, data = stroke)

Coefficients:
      (Intercept)               age      hypertension     heart_disease
        -7.632239          0.067765          0.568359          0.453716
avg_glucose_level
         0.004701

Degrees of Freedom: 3424 Total (i.e. Null);  3420 Residual
Null Deviance:      1411
Residual Deviance: 1150         AIC: 1160
```

$logit\big(\hat{\pi}(x)\big) = -7.6322 + 0.0678x_1 + 0.5684x_2 + 0.4537x_3 + 0.0047x_4$, where:

$\hat{\pi}(x) = \ the\ probability\ of\ having\ a\ stroke$

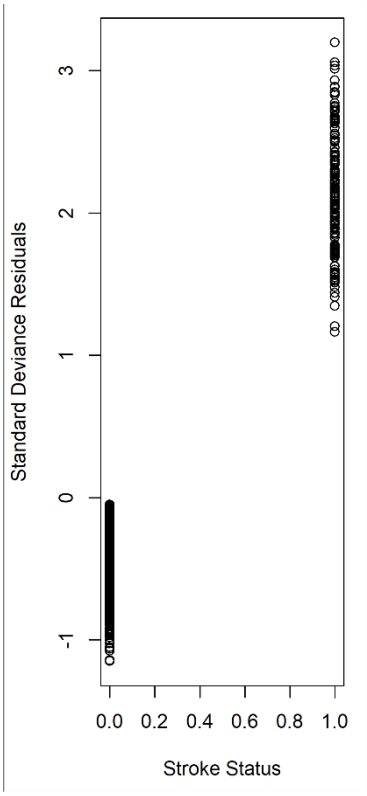
$x_1 = \ age\ (in\ years)$

$x_2 = hypertension\ (0\ if\ doesn't\ have\ hypertension, 1\ if\ has\ hypertension)$

$x_3 = heart\ disease\ (0\ if\ doesn't\ have\ heart\ diseases, 1\ if\ has\ a\ heart\ disease)$

$x_4 = average\ glucose\ level\ (in\ mg/dL)$

Also included is a scatterplot of standardized deviance residuals.




The majority of the residuals are between 0 and -1, with only 95 (less than 3% of the sample size of 3425) being greater than 2. Between this and a lack of pattern (given that Stroke Status can only be 0 or 1), the residuals indicate that the model is a strong fit for the data.

```
> sum(rstandard(fit.2) >= 2)
[1] 95
> sum(rstandard(fit.2) >= 0)
[1] 180
> sum(rstandard(fit.2) <= 0)
[1] 3245
> sum(rstandard(fit.2) <= -1)
[1] 7
> sum(rstandard(fit.2) <= -2)
[1] 0
```

This model ultimately suggests the odds of having a stroke increase if one has hypertension (increases by a factor of exp(0.5684)) or a heart disease (increases by a factor of exp(0.4537)). The odds of having a stroke also increase as one gets older (exp(0.0678)x per year) and if one has a higher average glucose level in blood (exp(0.0047)x per mg/dL). This model also suggests that gender, BMI, and smoking status have no effect on the odds of having a stroke.

**f)**

For all models, a Wald and LR GOF test are used to determine significance.

Model 1. Start with the main effects model.

```
Call:
glm(formula = stroke ~ gender + age + hypertension + heart_disease +
    smoking_status + avg_glucose_level + bmi, family = binomial,
    data = stroke)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -7.989417   0.667525 -11.969  < 2e-16 ***
genderMale                -0.074500   0.167675  -0.444 0.656817
age                        0.070205   0.006711  10.461  < 2e-16 ***
hypertension               0.566780   0.182530   3.105 0.001902 **
heart_disease              0.426566   0.219958   1.939 0.052464 .
smoking_statusnever smoked -0.064192   0.188791  -0.340 0.733843
smoking_statussmokes        0.327941   0.229990   1.426 0.153900
avg_glucose_level           0.004663   0.001373   3.396 0.000683 ***
bmi                         0.006619   0.012870   0.514 0.607041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1146.3  on 3416  degrees of freedom
AIC: 1164.3

Number of Fisher Scoring iterations: 7
```

```
Single term deletions

Model:
stroke ~ gender + age + hypertension + heart_disease + smoking_status +
    avg_glucose_level + bmi
                  Df Deviance    AIC     LRT  Pr(>Chi)
<none>                 1146.2 1164.2
gender             1   1146.5 1162.5   0.198 0.6562906
age                1   1291.8 1307.8 145.540 < 2.2e-16 ***
hypertension       1   1155.4 1171.4   9.171 0.0024593 **
heart_disease      1   1149.8 1165.8   3.586 0.0582814 .
smoking_status     2   1149.5 1163.5   3.303 0.1917533
avg_glucose_level  1   1157.5 1173.5  11.208 0.0008143 ***
bmi                1   1146.5 1162.5   0.262 0.6088411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to both tests, the parameters associated with gender and BMI are not significant (p-val > 0.1; backed up by initial testing in part d). Thus, Gender and BMI are dropped (Smoking Status is kept for now due to previous evidence of marginal association in part d).

Model 2: Main effects model (without Gender/BMI).

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + smoking_status +
    avg_glucose_level, family = binomial, data = stroke)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -7.793238   0.480222 -16.228  < 2e-16 ***
age                         0.069521   0.006516  10.669  < 2e-16 ***
hypertension                0.576742   0.181475   3.178 0.001483 **
heart_disease               0.410991   0.218359   1.882 0.059812 .
smoking_statusnever smoked -0.054520   0.186935  -0.292 0.770555
smoking_statussmokes        0.327136   0.229921   1.423 0.154789
avg_glucose_level           0.004785   0.001335   3.586 0.000336 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1146.7  on 3418  degrees of freedom
AIC: 1160.7

Number of Fisher Scoring iterations: 7
```

```
Single term deletions

Model:
stroke ~ age + hypertension + heart_disease + smoking_status +
    avg_glucose_level
                  Df Deviance    AIC     LRT  Pr(>Chi)
<none>                 1146.7 1160.7
age                1   1293.6 1305.6 146.904 < 2.2e-16 ***
hypertension       1   1156.3 1168.3   9.577 0.0019703 **
heart_disease      1   1150.1 1162.1   3.377 0.0661011 .
smoking_status     2   1149.9 1159.9   3.178 0.2041742
avg_glucose_level  1   1159.1 1171.1  12.402 0.0004288 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to both tests, the parameters associated with smoking status are not significant (p-val > 0.1). Thus, Smoking Status is dropped.

Model 3: Main effects model (without Gender/BMI/Smoking Status).

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
    family = binomial, data = stroke)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.632239   0.439507 -17.365  < 2e-16 ***
age                0.067765   0.006359  10.656  < 2e-16 ***
hypertension       0.568359   0.181384   3.133 0.001728 **
heart_disease      0.453716   0.216659   2.094 0.036247 *
avg_glucose_level  0.004701   0.001334   3.524 0.000425 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1149.9  on 3420  degrees of freedom
AIC: 1159.9

Number of Fisher Scoring iterations: 7
```

```
Single term deletions

Model:
stroke ~ age + hypertension + heart_disease + avg_glucose_level
                  Df Deviance    AIC     LRT  Pr(>Chi)
<none>                 1149.9 1159.9
age                1   1296.0 1304.0 146.089 < 2.2e-16 ***
hypertension       1   1159.2 1167.2   9.316 0.0022713 **
heart_disease      1   1154.1 1162.1   4.155 0.0415074 *
avg_glucose_level  1   1161.9 1169.9  11.983 0.0005368 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to both tests, all parameters have somewhat strong significance (p-val < 0.05).

---

From here, an automated stepwise algorithm was applied involving the two-factor interaction terms for Model 3 (i.e. (age+hypertension+heart_disease+avg_glucose_lvl)^2). Forwards, backwards, and bidirectional algorithms were used, leading to the same model:

Model 4: Model 3 with interaction terms AGL:Heart Disease and Hypertension:Heart Disease

```
Call:
glm(formula = stroke ~ age + avg_glucose_level + hypertension +
    heart_disease + avg_glucose_level:heart_disease + hypertension:heart_disease,
    family = binomial, data = stroke)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -7.522081   0.444336 -16.929  < 2e-16 ***
age                           0.068182   0.006398  10.656  < 2e-16 ***
avg_glucose_level             0.003305   0.001524   2.169 0.030090 *
hypertension                  0.713674   0.199703   3.574 0.000352 ***
heart_disease                -0.305215   0.578676  -0.527 0.597890
avg_glucose_level:heart_disease  0.006588   0.003432   1.919 0.054935 .
hypertension:heart_disease   -0.699712   0.471691  -1.483 0.137965
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1144.2  on 3418  degrees of freedom
AIC: 1158.2

Number of Fisher Scoring iterations: 7
```

```
Single term deletions

Model:
stroke ~ age + avg_glucose_level + hypertension + heart_disease +
    avg_glucose_level:heart_disease + hypertension:heart_disease
                                Df Deviance    AIC     LRT Pr(>Chi)
<none>                               1144.2 1158.2
age                              1   1289.2 1301.2 145.015  < 2e-16 ***
avg_glucose_level:heart_disease  1   1148.0 1160.0   3.791  0.05154 .
hypertension:heart_disease       1   1146.5 1158.5   2.256  0.13308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to both tests, the interaction factor Hypertension:Heart Disease is not significant (p-val > 0.10). Thus, it is dropped. Note that the standard error for heart disease is far larger (0.57 > 0.21) and that it is no longer significant (according to the Wald test) in this model, indicating signs of multicollinearity.

Model 5: Model 3 with interaction term AGL:Heart Disease

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level +
    avg_glucose_level:heart_disease, family = binomial, data = stroke)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -7.529974   0.443133 -16.993   <2e-16 ***
age                          0.068596   0.006375  10.760   <2e-16 ***
hypertension                 0.582124   0.182229   3.194   0.0014 **
heart_disease               -0.481568   0.571376  -0.843   0.3993
avg_glucose_level            0.003447   0.001517   2.272   0.0231 *
heart_disease:avg_glucose_level  0.006259   0.003426   1.827   0.0678 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1410.9  on 3424  degrees of freedom
Residual deviance: 1146.5  on 3419  degrees of freedom
AIC: 1158.5

Number of Fisher Scoring iterations: 7
```

```
Single term deletions

Model:
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
    avg_glucose_level:heart_disease
                             Df Deviance    AIC     LRT  Pr(>Chi)
<none>                            1146.5 1158.5
age                           1   1295.5 1305.5 148.994 < 2.2e-16 ***
hypertension                  1   1156.2 1166.2   9.674  0.001869 **
heart_disease:avg_glucose_level  1   1149.9 1159.9   3.428  0.064119 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

While the interaction term AGL:Heart Disease has some significance (p-val < 0.1), this model also shows signs of multicollinearity regarding Heart Disease. Thus, we compare Model 3 and Model 5 by testing for goodness of fit with an Analysis of Deviance.

```
Analysis of Deviance Table

Model 1: stroke ~ age + hypertension + heart_disease + avg_glucose_level
Model 2: stroke ~ age + avg_glucose_level + hypertension + heart_disease +
    avg_glucose_level:heart_disease + hypertension:heart_disease
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3420     1149.9
2      3418     1144.2  2   5.6837  0.05832 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is slightly significant evidence ($0.05 < \text{p-val} < 0.1$) of Model 5 being a better fit than Model 3. However, due to their AIC values being close enough (1159.9 vs 1158.5), Model 5 having multicollinearity (also the Heart Disease parameter in Model 5 being negative), and to make the final model easier to understand and observe, Model 3 is ultimately chosen as the final model.

---

**g)**

In conclusion, there appears to be an association between some of the explanatory variables and whether someone has had a stroke. Preliminary testing revealed that Hypertension, Heart Disease, Smoking Status, Age, and Average Glucose Level had marginal association with Stroke Status, while Gender and BMI had no marginal association with Stroke Status (part d). According to the final model (part e), the same variables were still associated with Stroke Status except for Smoking Status, which despite being marginally associated, did not have enough evidence of association when taking the other variables into account. Overall, people with hypertension or a heart disease are more likely to have had a stroke; additionally, the older one gets, and the more glucose in their bloodstream, the more likely they are to have had a stroke.