Final Project

Max Enderlein

2025-04-02

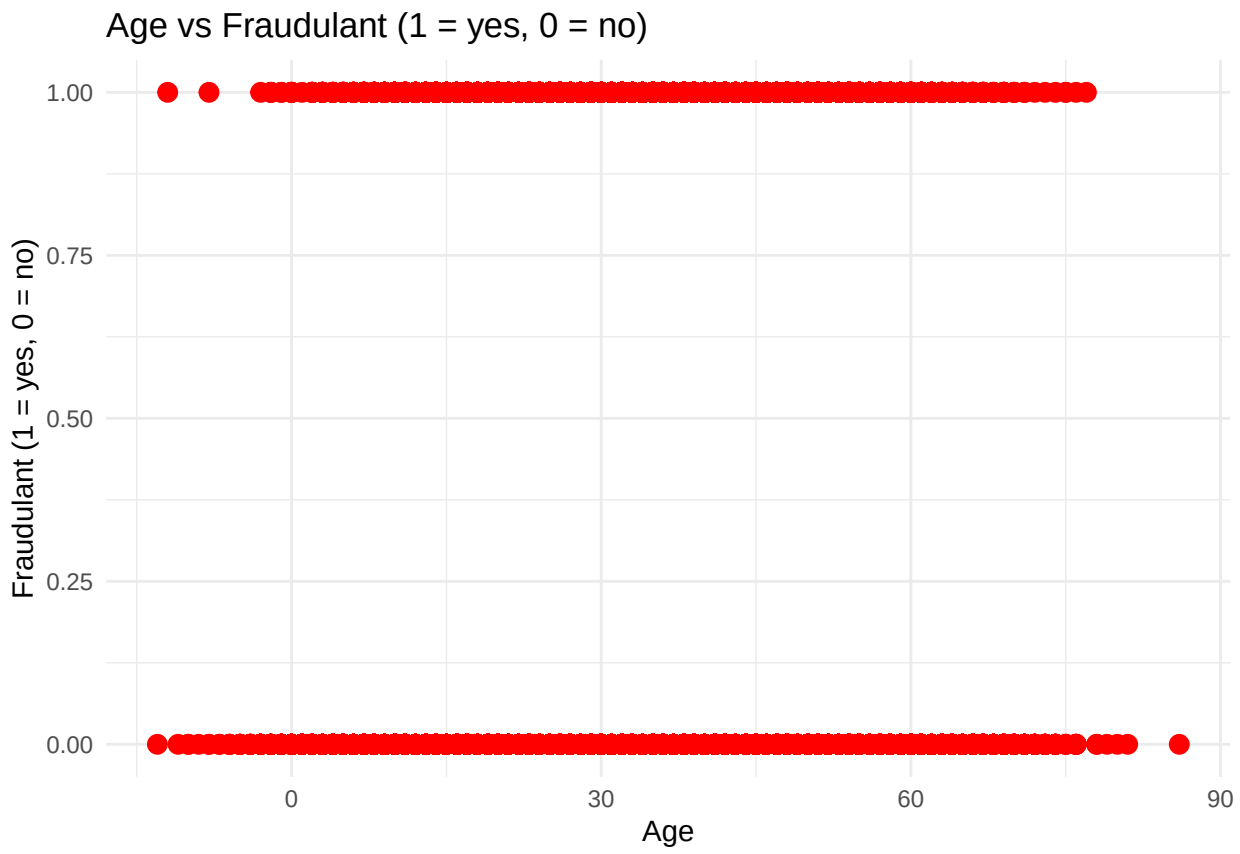
```
set.seed(1)
library(ggplot2)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
str(Fraud)
```

```
## 'data.frame':    1048575 obs. of  16 variables:
##  $ Transaction_ID    : chr  "15d2e414-8735-46fc-9e02-80b472b2580f" "0bfee1a0-6d5e-40da-a446-d04e73b1
##  $ Customer_ID       : chr  "d1b87f62-51b2-493b-ad6a-77e0fe13e785" "37de64d5-e901-4a56-9ea0-af0c24c0
##  $ Transaction_Amount: num  58.1 390 134.2 226.2 121.5 ...
##  $ Transaction_Date  : chr  "2/20/2024 5:58" "2/25/2024 8:09" "3/18/2024 3:42" "3/16/2024 20:41" ...
##  $ Payment_Method    : chr  "bank transfer" "debit card" "PayPal" "bank transfer" ...
##  $ Product_Category  : chr  "electronics" "electronics" "home & garden" "clothing" ...
##  $ Quantity          : int  1 2 2 5 2 2 2 4 4 4 ...
##  $ Customer_Age      : int  17 40 22 31 51 34 14 42 38 39 ...
##  $ Customer_Location : chr  "Amandaborough" "East Timothy" "Davismouth" "Lynnberg" ...
##  $ Device_Used       : chr  "tablet" "desktop" "tablet" "desktop" ...
##  $ IP_Address        : chr  "212.195.49.198" "208.106.249.121" "76.63.88.212" "207.208.171.73" ...
##  $ Shipping_Address  : chr  "Unit 8934 Box 0058\nDPO AA 05437" "634 May Keys\nPort Cherylview, NV 75
##  $ Billing_Address   : chr  "Unit 8934 Box 0058\nDPO AA 05437" "634 May Keys\nPort Cherylview, NV 75
##  $ Is_Fraudulent     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Account_Age_Days  : int  30 72 63 124 158 38 119 251 190 343 ...
##  $ Transaction_Hour  : int  5 8 3 20 5 10 19 13 19 21 ...
```

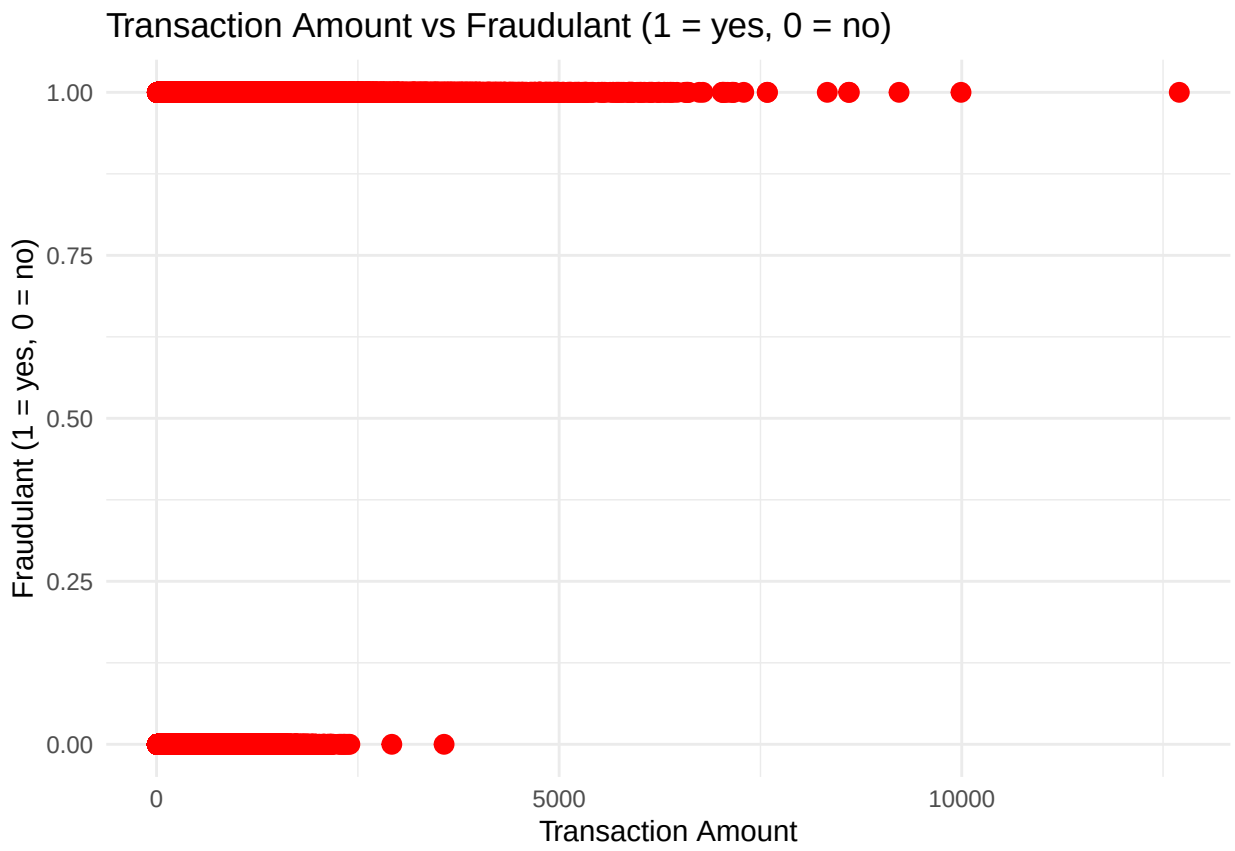
```
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
```

Testing plots below

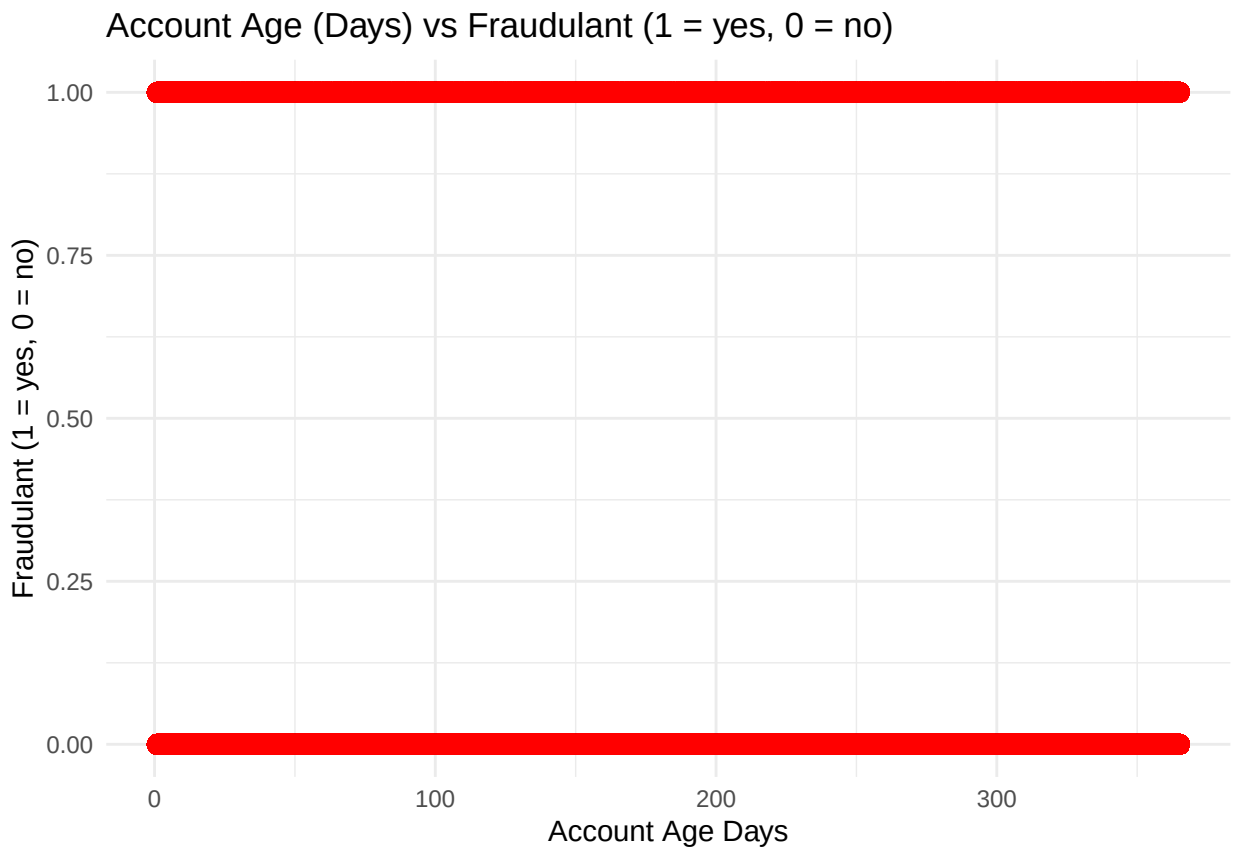
```
plot = ggplot(training, aes(x = Customer_Age, y = Is_Fraudulent)) +
          geom_point(color = "red", size = 3) +
          labs(title = "Age vs Fraudulant (1 = yes, 0 = no)",
             x = "Age",
             y = "Fraudulant (1 = yes, 0 = no)") +
           theme_minimal()
print(plot)
```



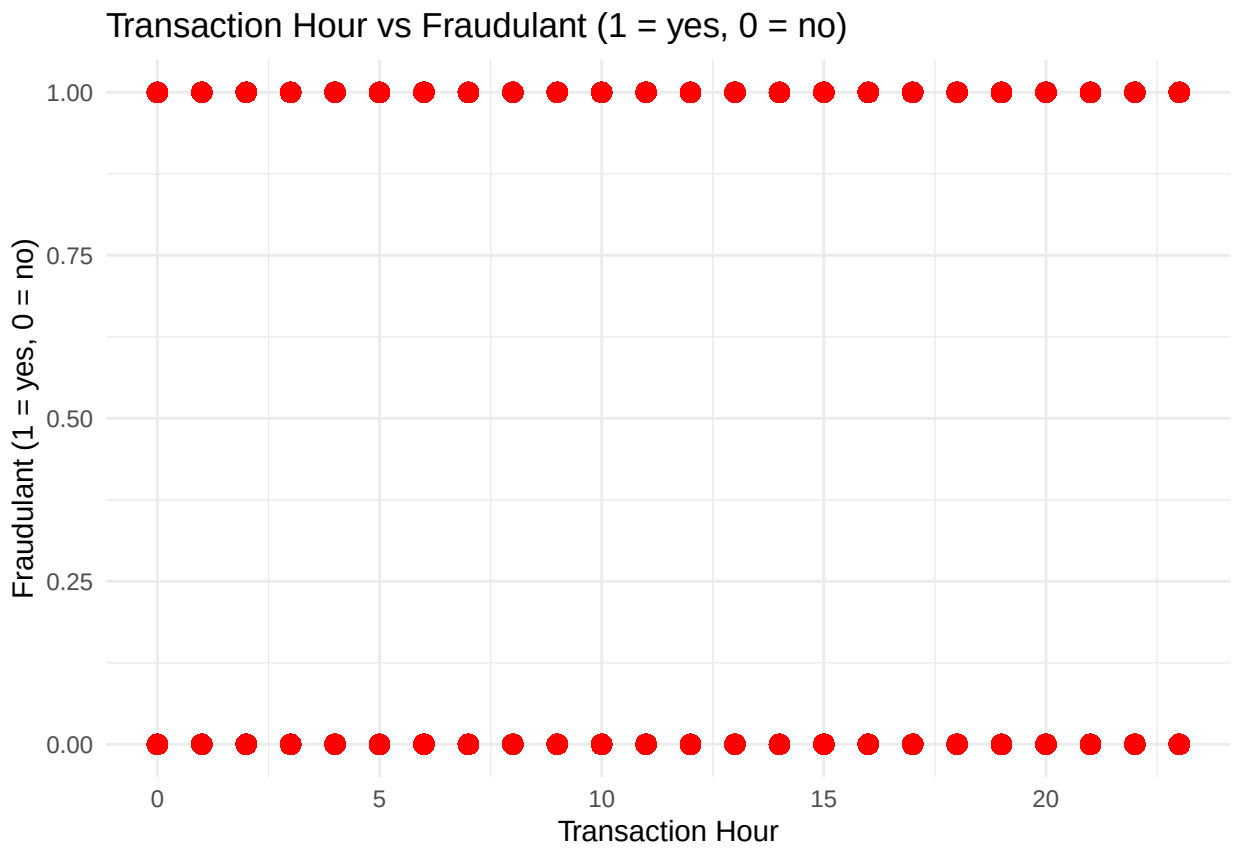

```
plot = ggplot(training, aes(x = Transaction_Amount, y = Is_Fraudulent)) +
          geom_point(color = "red", size = 3) +
          labs(title = "Transaction Amount vs Fraudulant (1 = yes, 0 = no)",
             x = "Transaction Amount",
             y = "Fraudulant (1 = yes, 0 = no)") +
           theme_minimal()
print(plot)
```




```
plot = ggplot(training, aes(x = Account_Age_Days, y = Is_Fraudulent)) +
          geom_point(color = "red", size = 3) +
          labs(title = "Account Age (Days) vs Fraudulant (1 = yes, 0 = no)",
             x = "Account Age Days",
             y = "Fraudulant (1 = yes, 0 = no)") +
           theme_minimal()
print(plot)
```




```
plot = ggplot(training, aes(x = Transaction_Hour, y = Is_Fraudulent)) +
          geom_point(color = "red", size = 3) +
          labs(title = "Transaction Hour vs Fraudulant (1 = yes, 0 = no)",
             x = "Transaction Hour",
             y = "Fraudulant (1 = yes, 0 = no)") +
           theme_minimal()
print(plot)
```




```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccAge = glm(Is_Fraudulent ~ Account_Age_Days, data = training, family=binomial(link='logit'))
summary(SumAccAge)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Account_Age_Days, family = binomial(link = "logit"),
##     data = training)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.004e+00  8.401e-03  -238.6   <2e-16 ***
## Account_Age_Days -6.326e-03  5.495e-05  -115.1   <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: 312422  on 786430  degrees of freedom
## Residual deviance: 297406  on 786429  degrees of freedom
```

```
## AIC: 297410
##
## Number of Fisher Scoring iterations: 6
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccTM = glm(Is_Fraudulent ~ Transaction_Amount, data = training, family=binomial(link='logit'))
summary(SumAccTM)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Transaction_Amount, family = binomial(link = "logit"),
##     data = training)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.688e+00  7.734e-03  -476.9   <2e-16 ***
## Transaction_Amount  2.406e-03  1.496e-05   160.8   <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: 312422  on 786430  degrees of freedom
## Residual deviance: 281392  on 786429  degrees of freedom
## AIC: 281396
##
## Number of Fisher Scoring iterations: 6
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccAge = glm(Is_Fraudulent ~ Customer_Age, data = training, family=binomial(link='logit'))
summary(SumAccAge)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Customer_Age, family = binomial(link = "logit"),
##     data = training)
##
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  -2.9538055  0.0186019 -158.790   <2e-16 ***
## Customer_Age  0.0002956  0.0005172    0.572    0.568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 312422  on 786430  degrees of freedom
## Residual deviance: 312422  on 786429  degrees of freedom
## AIC: 312426
##
## Number of Fisher Scoring iterations: 5
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccHour = glm(Is_Fraudulent ~ Transaction_Hour, data = training, family=binomial(link='logit'))
summary(SumAccHour)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Transaction_Hour, family = binomial(link = "logit"),
##     data = training)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.1871479  0.0083516 -261.88   <2e-16 ***
## Transaction_Hour -0.0786297  0.0008119  -96.85   <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: 312422  on 786430  degrees of freedom
## Residual deviance: 302176  on 786429  degrees of freedom
## AIC: 302180
##
## Number of Fisher Scoring iterations: 6
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccQuantity = glm(Is_Fraudulent ~ Quantity, data = training, family=binomial(link='logit'))
summary(SumAccQuantity)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Quantity, family = binomial(link = "logit"),
##     data = training)
##
## Coefficients:
```

```
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.927995   0.012092 -242.133   <2e-16 ***
## Quantity    -0.005209   0.003655   -1.425    0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 312422  on 786430  degrees of freedom
## Residual deviance: 312420  on 786429  degrees of freedom
## AIC: 312424
##
## Number of Fisher Scoring iterations: 5
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccSig = glm(Is_Fraudulent ~ Account_Age_Days + Transaction_Amount + Transaction_Hour, data = trainin
summary(SumAccSig)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Account_Age_Days + Transaction_Amount +
##     Transaction_Hour, family = binomial(link = "logit"), data = training)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.015e+00  1.246e-02 -161.78   <2e-16 ***
## Account_Age_Days   -6.163e-03  5.724e-05 -107.66   <2e-16 ***
## Transaction_Amount  2.379e-03  1.576e-05  151.01   <2e-16 ***
## Transaction_Hour   -7.712e-02  8.562e-04  -90.07   <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: 312422  on 786430  degrees of freedom
## Residual deviance: 259254  on 786427  degrees of freedom
## AIC: 259262
##
## Number of Fisher Scoring iterations: 6
```

```
set.seed(1)
Fraud = read.csv("Fraudulent_E-Commerce_Transaction_Data.csv")
smp <- floor(.75 * nrow(Fraud))
train <- sample(seq_len(nrow(Fraud)), size = smp)
training <- Fraud[train, ]
testing <- Fraud[-train, ]
SumAccSigTest = glm(Is_Fraudulent ~ Account_Age_Days + Transaction_Amount + Transaction_Hour, data = tes
summary(SumAccSigTest)
```

```
##
## Call:
## glm(formula = Is_Fraudulent ~ Account_Age_Days + Transaction_Amount +
##     Transaction_Hour, family = binomial(link = "logit"), data = testing)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.006e+00  2.147e-02  -93.44   <2e-16 ***
## Account_Age_Days   -6.034e-03  9.846e-05  -61.29   <2e-16 ***
## Transaction_Amount  2.371e-03  2.717e-05   87.28   <2e-16 ***
## Transaction_Hour   -7.918e-02  1.488e-03  -53.23   <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: 104431  on 262143  degrees of freedom
## Residual deviance:  86874  on 262140  degrees of freedom
## AIC: 86882
##
## Number of Fisher Scoring iterations: 6
```