Status Report #2

2025-09-30

Tasks: Create EDR, correlations, initial visualizations

```
# Load necessary libraries
library(readr)
library(DiagrammeR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load data and view it (make sure everything looks right)
HeartAttack <- read.csv("HeartAttack.csv")
head(HeartAttack)
```

```
##   Age Gender Cholesterol BloodPressure HeartRate  BMI Smoker Diabetes
## 1  31   Male         194           162        71 22.9      0        1
## 2  69   Male         208           148        93 33.9      1        1
## 3  34 Female         132           161        94 34.0      0        0
## 4  53   Male         268           134        91 35.0      0        1
## 5  57 Female         203           140        75 30.1      0        1
## 6  41   Male         158           154        72 38.7      0        1
##   Hypertension FamilyHistory PhysicalActivity AlcoholConsumption      Diet
## 1            0             0                6                  0 Unhealthy
## 2            0             0                1                  2 Unhealthy
## 3            1             1                1                  3   Healthy
## 4            1             0                6                  0   Healthy
## 5            0             0                4                  1  Moderate
## 6            0             1                3                  2  Moderate
##   StressLevel Ethnicity Income EducationLevel Medication ChestPainType
## 1           1  Hispanic  64510    High School        Yes       Typical
## 2           6     Asian  91773        College         No      Atypical
## 3           3     Black 173550        College         No   Non-anginal
## 4           3  Hispanic  43861    High School        Yes      Atypical
## 5           1  Hispanic  83404    High School        Yes       Typical
## 6           4     Other 113011    High School        Yes   Non-anginal
##         ECGResults MaxHeartRate ST_Depression ExerciseInducedAngina       Slope
## 1 ST-T abnormality          173          0.52                   Yes Downsloping
## 2   LV hypertrophy          189          3.79                   Yes   Upsloping
## 3           Normal          122          0.17                   Yes   Upsloping
```

```
## 4 ST-T abnormality          104          0.67                   Yes        Flat
## 5 ST-T abnormality          126          5.00                   Yes        Flat
## 6   LV hypertrophy          155          4.30                    No Downsloping
##   NumberOfMajorVessels       Thalassemia PreviousHeartAttack StrokeHistory
## 1                    1            Normal                   0             0
## 2                    2            Normal                   0             0
## 3                    0            Normal                   1             0
## 4                    0 Reversible defect                   1             0
## 5                    0      Fixed defect                   1             0
## 6                    2 Reversible defect                   0             0
##   Residence EmploymentStatus MaritalStatus         Outcome
## 1  Suburban          Retired        Single No Heart Attack
## 2  Suburban       Unemployed       Married No Heart Attack
## 3     Rural          Retired        Single    Heart Attack
## 4  Suburban          Retired       Widowed No Heart Attack
## 5     Rural          Retired       Married    Heart Attack
## 6     Rural       Unemployed       Married    Heart Attack
```

```
# Load cleaned data from Status Report #1 and view it (make sure it looks right)
HeartAttackClean <- readRDS("HeartAttackClean.rds")
head(HeartAttackClean)
```

```
##   Age Gender Cholesterol BloodPressure HeartRate  BMI Smoker Diabetes
## 1  31      0         194           162        71 22.9      0        1
## 2  69      0         208           148        93 33.9      1        1
## 3  34      1         132           161        94 34.0      0        0
## 4  53      0         268           134        91 35.0      0        1
## 5  57      1         203           140        75 30.1      0        1
## 6  41      0         158           154        72 38.7      0        1
##   Hypertension FamilyHistory PhysicalActivity AlcoholConsumption Diet
## 1            0             0                6                   0    0
## 2            0             0                1                   2    0
## 3            1             1                1                   3    1
## 4            1             0                6                   0    1
## 5            0             0                4                   1    2
## 6            0             1                3                   2    2
##   StressLevel Ethnicity Income EducationLevel Medication ChestPainType
## 1           1         0  64510              0          0             0
## 2           6         1  91773              1          1             1
## 3           3         2 173550              1          1             2
## 4           3         0  43861              0          0             1
## 5           1         0  83404              0          0             0
## 6           4         3 113011              0          0             2
##   ECGResults MaxHeartRate ST_Depression ExerciseInducedAngina Slope
## 1          0          173          0.52                     0     0
## 2          1          189          3.79                     0     1
## 3          2          122          0.17                     0     1
## 4          0          104          0.67                     0     2
## 5          0          126          5.00                     0     2
## 6          1          155          4.30                     1     0
##   NumberOfMajorVessels Thalassemia PreviousHeartAttack StrokeHistory Residence
## 1                    1           0                   0             0         0
## 2                    2           0                   0             0         0
## 3                    0           0                   1             0         1
## 4                    0           1                   1             0         0
```

```
## 5                    0           2                    1             0         1
## 6                    2           1                    0             0         1
##   EmploymentStatus MaritalStatus Outcome
## 1                0             0       0
## 2                1             1       0
## 3                0             0       1
## 4                0             2       0
## 5                0             1       1
## 6                1             1       1
```

```
# Create EDA
# Shape & columns
dim(HeartAttackClean)           # rows, cols
```

```
## [1] 372974     32
```

```
names(HeartAttackClean)         # column names
```

```
##  [1] "Age"                   "Gender"                "Cholesterol"
##  [4] "BloodPressure"         "HeartRate"             "BMI"
##  [7] "Smoker"                "Diabetes"              "Hypertension"
## [10] "FamilyHistory"         "PhysicalActivity"      "AlcoholConsumption"
## [13] "Diet"                  "StressLevel"           "Ethnicity"
## [16] "Income"                "EducationLevel"        "Medication"
## [19] "ChestPainType"         "ECGResults"            "MaxHeartRate"
## [22] "ST_Depression"         "ExerciseInducedAngina" "Slope"
## [25] "NumberOfMajorVessels"  "Thalassemia"           "PreviousHeartAttack"
## [28] "StrokeHistory"         "Residence"             "EmploymentStatus"
## [31] "MaritalStatus"         "Outcome"
```

```
# Missing values (per column)
sort(colSums(is.na(HeartAttackClean)), decreasing = TRUE)
```

```
##         ChestPainType                   Age                Gender
##                 93609                     0                     0
##           Cholesterol         BloodPressure             HeartRate
##                     0                     0                     0
##                   BMI               Smoker              Diabetes
##                     0                     0                     0
##          Hypertension         FamilyHistory      PhysicalActivity
##                     0                     0                     0
##    AlcoholConsumption                  Diet           StressLevel
##                     0                     0                     0
##             Ethnicity                Income         EducationLevel
##                     0                     0                     0
##            Medication            ECGResults          MaxHeartRate
##                     0                     0                     0
##         ST_Depression  ExerciseInducedAngina                 Slope
##                     0                     0                     0
##  NumberOfMajorVessels           Thalassemia   PreviousHeartAttack
##                     0                     0                     0
##         StrokeHistory             Residence      EmploymentStatus
##                     0                     0                     0
##         MaritalStatus               Outcome
##                     0                     0
```

```r
# Statistics (works for numeric & 0/1)
summary(HeartAttackClean)
```

```
##       Age            Gender        Cholesterol    BloodPressure
##  Min.   :30.00   Min.   :0.0000   Min.   :100.0   Min.   : 90.0
##  1st Qu.:43.00   1st Qu.:0.0000   1st Qu.:149.0   1st Qu.:112.0
##  Median :57.00   Median :0.0000   Median :199.0   Median :134.0
##  Mean   :56.98   Mean   :0.4992   Mean   :199.5   Mean   :134.5
##  3rd Qu.:71.00   3rd Qu.:1.0000   3rd Qu.:249.0   3rd Qu.:157.0
##  Max.   :84.00   Max.   :1.0000   Max.   :299.0   Max.   :179.0
##
##    HeartRate          BMI            Smoker          Diabetes
##  Min.   : 60.0   Min.   :18.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.: 74.0   1st Qu.:23.50   1st Qu.:0.0000   1st Qu.:0.000
##  Median : 89.0   Median :29.00   Median :0.0000   Median :1.000
##  Mean   : 89.5   Mean   :29.01   Mean   :0.4992   Mean   :0.501
##  3rd Qu.:105.0   3rd Qu.:34.50   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :119.0   Max.   :40.00   Max.   :1.0000   Max.   :1.000
##
##   Hypertension   FamilyHistory    PhysicalActivity AlcoholConsumption
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000    Min.   :0
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.000    1st Qu.:1
##  Median :1.000   Median :0.0000   Median :3.000    Median :2
##  Mean   :0.501   Mean   :0.4999   Mean   :3.003    Mean   :2
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:5.000    3rd Qu.:3
##  Max.   :1.000   Max.   :1.0000   Max.   :6.000    Max.   :4
##
##       Diet        StressLevel      Ethnicity         Income
##  Min.   :0.000   Min.   :1.000   Min.   :0.000   Min.   : 20000
##  1st Qu.:0.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.: 64957
##  Median :1.000   Median :5.000   Median :2.000   Median :110111
##  Mean   :1.003   Mean   :5.002   Mean   :2.001   Mean   :110033
##  3rd Qu.:2.000   3rd Qu.:7.000   3rd Qu.:3.000   3rd Qu.:155012
##  Max.   :2.000   Max.   :9.000   Max.   :4.000   Max.   :199999
##
##  EducationLevel     Medication     ChestPainType     ECGResults
##  Min.   :0.0000   Min.   :0.0000   Min.   :0       Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0       1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1       Median :1.0000
##  Mean   :0.9976   Mean   :0.5001   Mean   :1       Mean   :0.9994
##  3rd Qu.:2.0000   3rd Qu.:1.0000   3rd Qu.:2       3rd Qu.:2.0000
##  Max.   :2.0000   Max.   :1.0000   Max.   :2       Max.   :2.0000
##                                    NA's   :93609
##   MaxHeartRate   ST_Depression   ExerciseInducedAngina     Slope
##  Min.   :100.0   Min.   :0.000   Min.   :0.0           Min.   :0.000
##  1st Qu.:124.0   1st Qu.:1.250   1st Qu.:0.0           1st Qu.:0.000
##  Median :149.0   Median :2.500   Median :1.0           Median :1.000
##  Mean   :149.5   Mean   :2.502   Mean   :0.5           Mean   :1.001
##  3rd Qu.:174.0   3rd Qu.:3.750   3rd Qu.:1.0           3rd Qu.:2.000
##  Max.   :199.0   Max.   :5.000   Max.   :1.0           Max.   :2.000
##
##  NumberOfMajorVessels  Thalassemia     PreviousHeartAttack StrokeHistory
##  Min.   :0.000         Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.000         1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000
```

```
##  Median :1.000        Median :1.0000   Median :0.0000      Median :1.0000
##  Mean   :1.499        Mean   :0.9965   Mean   :0.4969      Mean   :0.5008
##  3rd Qu.:2.000        3rd Qu.:2.0000   3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :3.000        Max.   :2.0000   Max.   :1.0000      Max.   :1.0000
##
##    Residence      EmploymentStatus MaritalStatus      Outcome
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :2.000   Median :0.0000
##  Mean   :0.9978   Mean   :0.9987   Mean   :1.499   Mean   :0.4995
##  3rd Qu.:2.0000   3rd Qu.:2.0000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :2.0000   Max.   :3.000   Max.   :1.0000
##
```

```r
# Discrete columns, show counts + %
uniq_n <- sapply(HeartAttackClean, function(x) length(unique(na.omit(x))))
disc_cols <- names(uniq_n[uniq_n >= 2 & uniq_n <= 10])

freq <- lapply(disc_cols, function(v) {
  tab <- table(HeartAttackClean[[v]], useNA = "no")
  data.frame(level = names(tab),
             count = as.integer(tab),
             pct = round(100 * prop.table(tab), 1),
             row.names = NULL)
})
names(freq) <- disc_cols
freq   # prints a small table for each discrete column
```

```
## $Gender
##   level  count pct.Var1 pct.Freq
## 1     0 186770        0     50.1
## 2     1 186204        1     49.9
##
## $Smoker
##   level  count pct.Var1 pct.Freq
## 1     0 186776        0     50.1
## 2     1 186198        1     49.9
##
## $Diabetes
##   level  count pct.Var1 pct.Freq
## 1     0 186119        0     49.9
## 2     1 186855        1     50.1
##
## $Hypertension
##   level  count pct.Var1 pct.Freq
## 1     0 186124        0     49.9
## 2     1 186850        1     50.1
##
## $FamilyHistory
##   level  count pct.Var1 pct.Freq
## 1     0 186524        0       50
## 2     1 186450        1       50
##
## $PhysicalActivity
##   level count pct.Var1 pct.Freq
```

```
## 1     0 53125        0     14.2
## 2     1 53268        1     14.3
## 3     2 53155        2     14.3
## 4     3 53290        3     14.3
## 5     4 53360        4     14.3
## 6     5 53326        5     14.3
## 7     6 53450        6     14.3
##
## $AlcoholConsumption
##   level count pct.Var1 pct.Freq
## 1     0 74524        0     20.0
## 2     1 74893        1     20.1
## 3     2 74286        2     19.9
## 4     3 74583        3     20.0
## 5     4 74688        4     20.0
##
## $Diet
##   level  count pct.Var1 pct.Freq
## 1     0 123804        0     33.2
## 2     1 124091        1     33.3
## 3     2 125079        2     33.5
##
## $StressLevel
##   level count pct.Var1 pct.Freq
## 1     1 41245        1     11.1
## 2     2 41895        2     11.2
## 3     3 41008        3     11.0
## 4     4 41460        4     11.1
## 5     5 41346        5     11.1
## 6     6 41586        6     11.1
## 7     7 41453        7     11.1
## 8     8 41572        8     11.1
## 9     9 41409        9     11.1
##
## $Ethnicity
##   level count pct.Var1 pct.Freq
## 1     0 74350        0     19.9
## 2     1 74942        1     20.1
## 3     2 74532        2     20.0
## 4     3 74181        3     19.9
## 5     4 74969        4     20.1
##
## $EducationLevel
##   level  count pct.Var1 pct.Freq
## 1     0 124912        0     33.5
## 2     1 124040        1     33.3
## 3     2 124022        2     33.3
##
## $Medication
##   level  count pct.Var1 pct.Freq
## 1     0 186448        0       50
## 2     1 186526        1       50
##
## $ChestPainType
```

```
##   level count pct.Var1 pct.Freq
## 1     0 93126        0     33.3
## 2     1 93262        1     33.4
## 3     2 92977        2     33.3
##
## $ECGResults
##   level  count pct.Var1 pct.Freq
## 1     0 124227        0     33.3
## 2     1 124750        1     33.4
## 3     2 123997        2     33.2
##
## $ExerciseInducedAngina
##   level  count pct.Var1 pct.Freq
## 1     0 186473        0       50
## 2     1 186501        1       50
##
## $Slope
##   level  count pct.Var1 pct.Freq
## 1     0 124045        0     33.3
## 2     1 124383        1     33.3
## 3     2 124546        2     33.4
##
## $NumberOfMajorVessels
##   level count pct.Var1 pct.Freq
## 1     0 93346        0     25.0
## 2     1 93473        1     25.1
## 3     2 93027        2     24.9
## 4     3 93128        3     25.0
##
## $Thalassemia
##   level  count pct.Var1 pct.Freq
## 1     0 125106        0     33.5
## 2     1 124071        1     33.3
## 3     2 123797        2     33.2
##
## $PreviousHeartAttack
##   level  count pct.Var1 pct.Freq
## 1     0 187629        0     50.3
## 2     1 185345        1     49.7
##
## $StrokeHistory
##   level  count pct.Var1 pct.Freq
## 1     0 186183        0     49.9
## 2     1 186791        1     50.1
##
## $Residence
##   level  count pct.Var1 pct.Freq
## 1     0 124721        0     33.4
## 2     1 124336        1     33.3
## 3     2 123917        2     33.2
##
## $EmploymentStatus
##   level  count pct.Var1 pct.Freq
## 1     0 124373        0     33.3
```

```
## 2     1 124706        1     33.4
## 3     2 123895        2     33.2
##
## $MaritalStatus
##   level count pct.Var1 pct.Freq
## 1     0 93599        0     25.1
## 2     1 92814        1     24.9
## 3     2 93467        2     25.1
## 4     3 93094        3     25.0
##
## $Outcome
##   level  count pct.Var1 pct.Freq
## 1     0 186658        0       50
## 2     1 186316        1       50
```

```
## 5) Duplicate rows (optional)
sum(duplicated(HeartAttackClean))
```

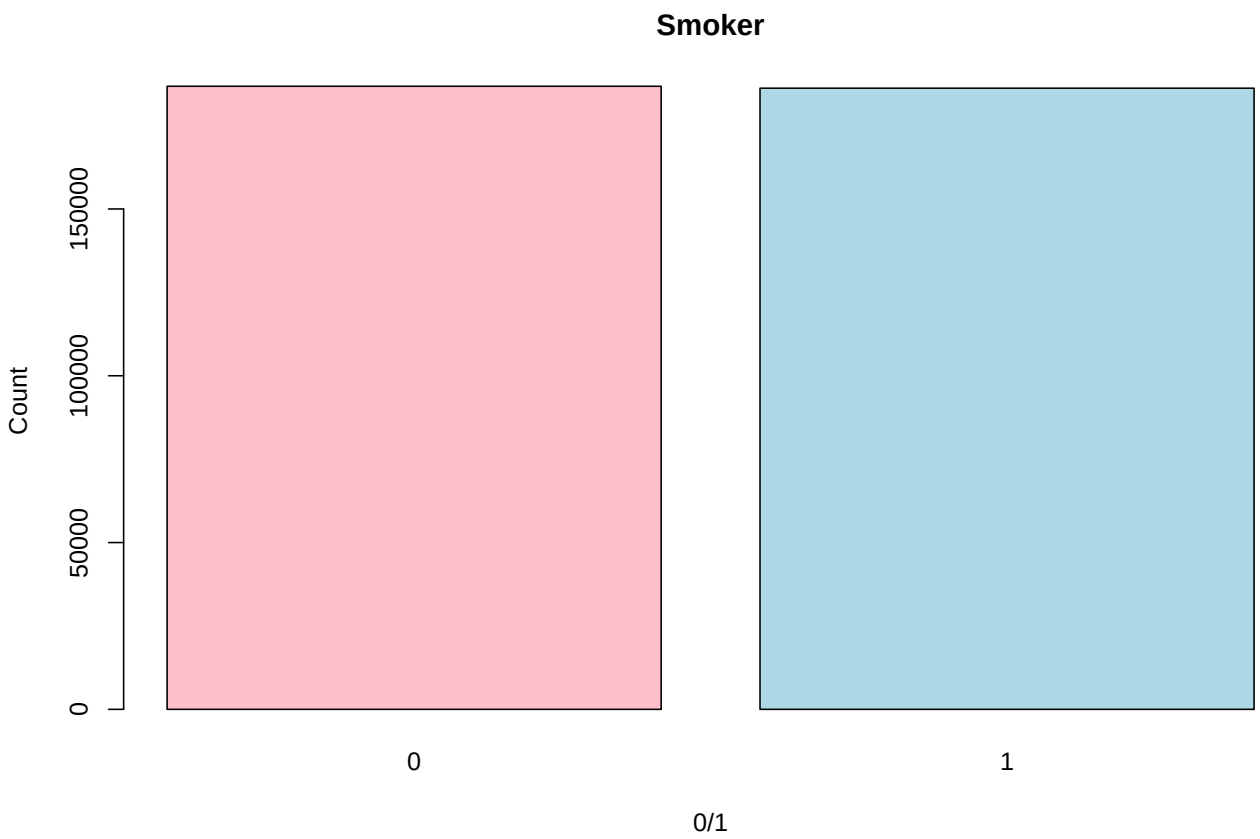
```
## [1] 0
```

```
# These are the variables I want to see a correlation with Outcome
vars <- c("Age","Gender","FamilyHistory","Smoker",
          "PhysicalActivity","AlcoholConsumption",
          "Diet","StressLevel","EducationLevel")

# I want to see if the above variables have correlations with Outcome
target <- "Outcome"   # <-- change if your column is named differently

df <- HeartAttackClean %>% select(any_of(c(vars, target)))

# Compute correlation with Outcome
res <- lapply(setdiff(names(df), target), function(v) {
  x <- df[[v]]
  y <- df[[target]]
  ok <- is.finite(df[[v]]) & is.finite(y)
  r  <- suppressWarnings(cor(df[[v]][ok], y[ok], method = "pearson"))  # phi for 0/1
  data.frame(variable = v,
             r = r,
             direction = ifelse(is.na(r), NA,
                                ifelse(r > 0, "positive",
                                       ifelse(r < 0, "negative", "zero"))))
})
res <- bind_rows(res) %>% arrange(match(variable, vars))
print(res)
```

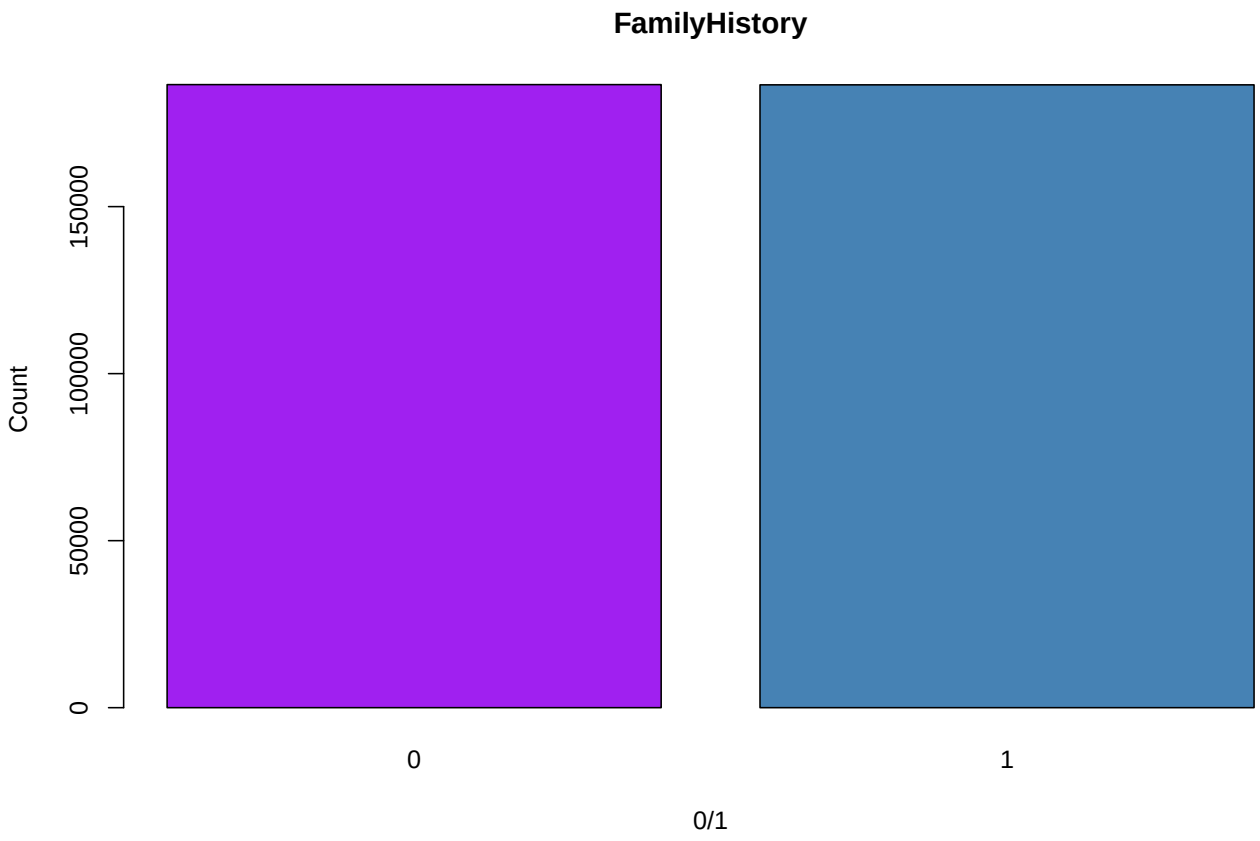
```
##             variable             r direction
## 1                Age -0.0009466919  negative
## 2             Gender  0.0018915048  positive
## 3      FamilyHistory  0.0004127157  positive
## 4             Smoker  0.0008297375  positive
## 5   PhysicalActivity  0.0003313566  positive
## 6 AlcoholConsumption  0.0016793415  positive
## 7               Diet  0.0006373049  positive
## 8        StressLevel -0.0023159857  negative
## 9     EducationLevel  0.0006799495  positive
```

- Positive: More heart attacks than no heart attacks
- Negative: Fewer heart attacks than heart attacks
- Age and stress level have less heart attack outcomes
- Gender, family history, smoker, physical activity, alcohol consumption, diet, and education level have more heart attack outcomes
- This tells us that the positive factors can increase the risk of a heart attack

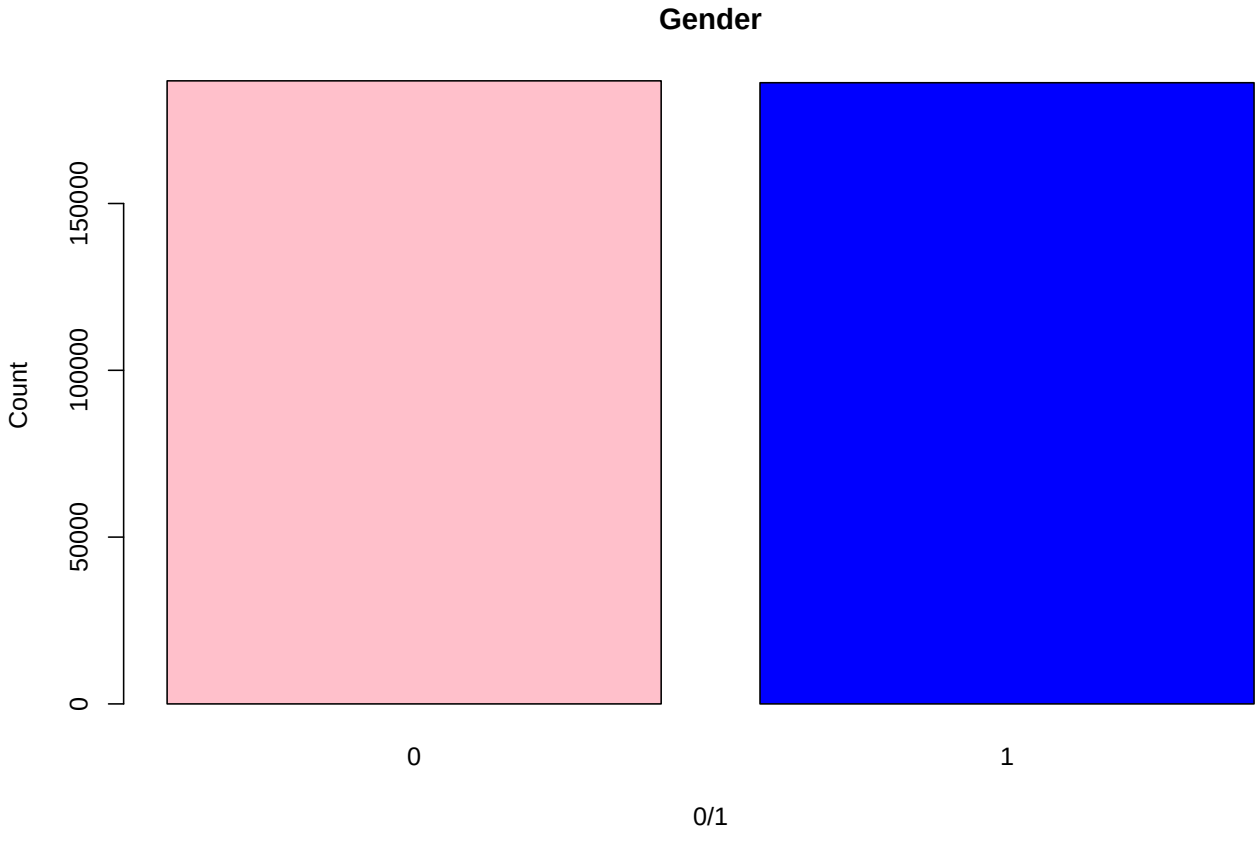
```
# Create initial plots
barplot(table(HeartAttackClean$Smoker), main="Smoker", xlab="0/1", ylab="Count", col = c("pink", "lightl
```



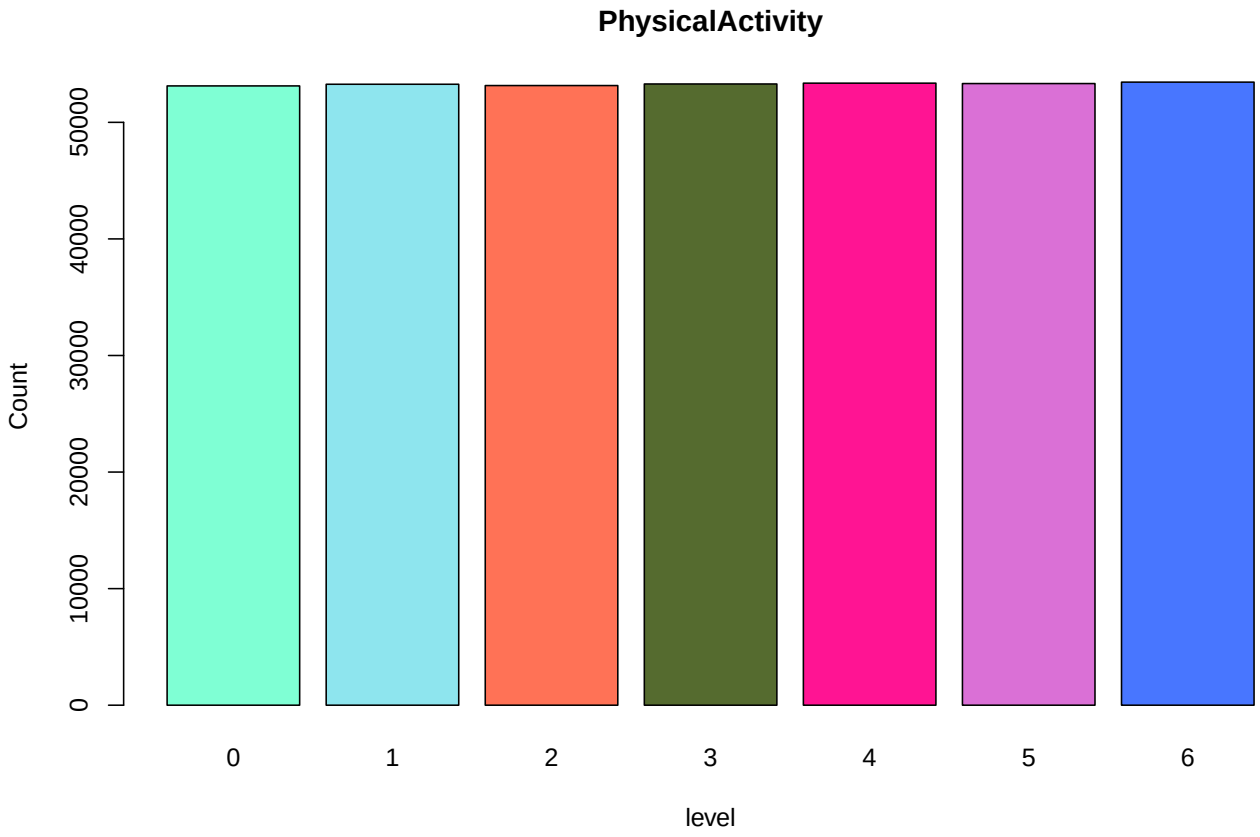

```
barplot(table(HeartAttackClean$FamilyHistory), main="FamilyHistory", xlab="0/1", ylab="Count", col = c(
```



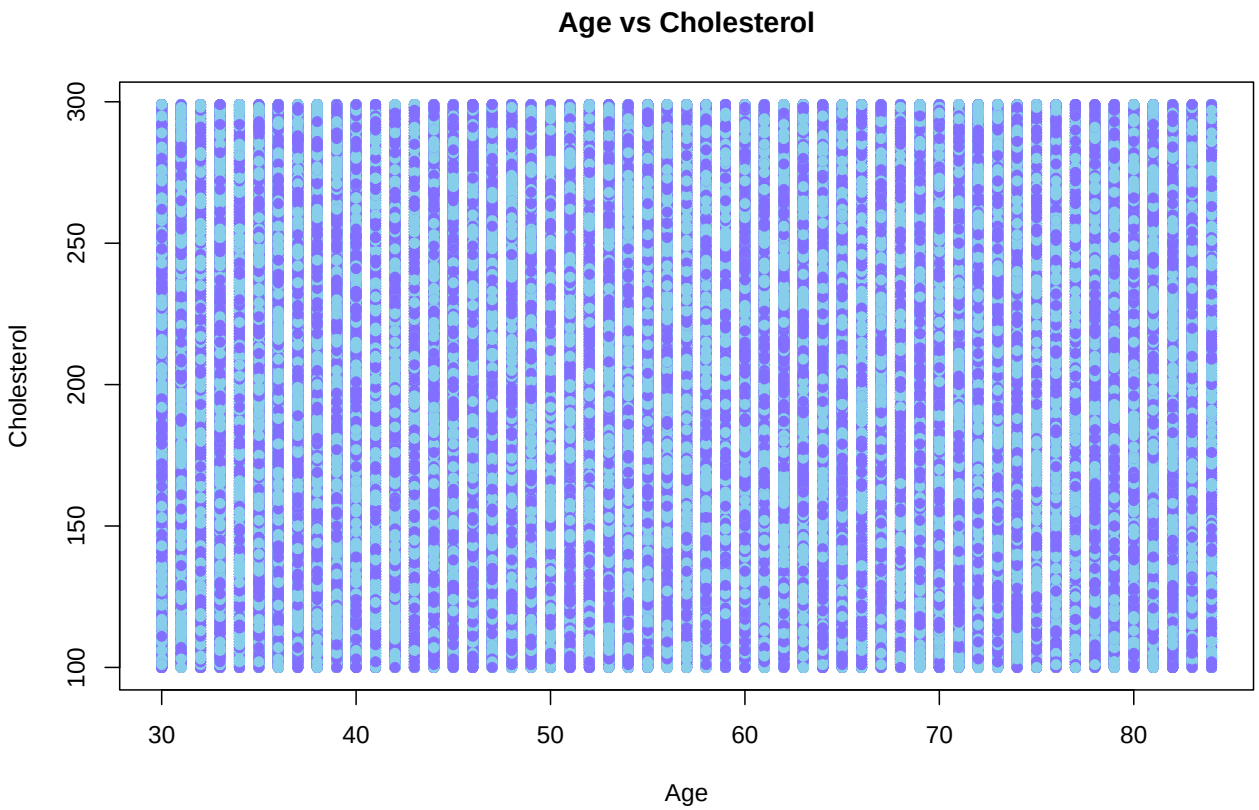

```
barplot(table(HeartAttackClean$Gender), main="Gender", xlab="0/1", ylab="Count", col = c("pink", "blue"
```



```r
barplot(table(HeartAttackClean$PhysicalActivity), main="PhysicalActivity", xlab="level", ylab="Count", c
```




```r
plot(HeartAttackClean$Age, HeartAttackClean$Cholesterol,
     xlab="Age", ylab="Cholesterol", pch=16, main="Age vs Cholesterol", col = c("skyblue", "slateblue1"
```




```
plot(HeartAttackClean$BMI, HeartAttackClean$BloodPressure,
     xlab="BMI", ylab="BloodPressure", pch=16, main="BMI vs BloodPressure", col = c("thistle1", "violet
```