Individual Project

2025-05-15

R Markdown

```
raw.data <- read.csv("/Users/minaheelkhan/Desktop/gambling2.csv", header = TRUE)
```

Inspecting the data initially

```
str(raw.data)
```

```
## 'data.frame':    13106 obs. of  36 variables:
##  $ HHSize   : int  2 2 4 5 5 4 4 4 4 2 ...
##  $ Sex      : int  2 1 1 1 2 1 2 1 1 1 ...
##  $ age      : int  58 47 39 41 37 51 48 19 16 56 ...
##  $ maritalg : int  1 1 1 1 1 1 1 3 3 1 ...
##  $ totinc   : int  26 26 97 21 21 -1 -1 -1 -1 26 ...
##  $ hhdtypb  : int  2 2 3 4 4 5 5 5 5 2 ...
##  $ OwnRnt08 : int  2 2 1 1 1 1 1 1 1 4 ...
##  $ numcars  : int  2 2 2 2 2 2 2 2 2 3 ...
##  $ SXORIEN  : int  1 1 1 1 -9 1 1 1 1 1 ...
##  $ Religsc  : int  2 1 1 1 1 1 1 1 1 2 ...
##  $ ethnicC  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ SrcInc7  : int  0 0 0 0 0 -1 -1 -1 -1 0 ...
##  $ SrcInc15 : int  0 0 0 0 0 -1 -1 -1 -1 0 ...
##  $ eqvinc   : num  105000 105000 -1 35669 35669 ...
##  $ eqv5     : int  1 1 -1 2 2 -1 -1 -1 -1 1 ...
##  $ Econact_2: int  3 1 1 1 4 1 5 4 2 1 ...
##  $ EducEnd  : int  8 7 5 7 8 8 7 1 1 5 ...
##  $ HighQual : int  1 1 6 1 1 2 3 3 6 4 ...
##  $ hpnssec5 : int  1 1 3 1 1 1 1 1 1 1 ...
##  $ RG15a    : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ docinfo1 : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ compm3   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ compm7   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ compm8   : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ compm9   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ genhelf2 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ longill12: int  1 1 1 2 2 2 1 2 1 2 ...
##  $ bmival   : num  36.1 51 -1 26.4 23.7 ...
##  $ ghq12scr : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ wemwbs   : int  -1 -1 -1 -1 -1 49 52 59 58 70 ...
##  $ cigst1   : int  1 1 1 3 3 3 3 1 1 1 ...
##  $ drating  : num  35 10.7 44 11.7 11 ...
##  $ Active   : int  2 4 2 2 -1 3 -1 2 -1 3 ...
##  $ ActPhy   : int  1 1 2 1 1 1 2 1 1 1 ...
##  $ country  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PROBGAM  : int  0 0 0 0 0 0 0 0 1 1 ...
```

```
head(raw.data)
```

```
##   HHSize Sex age maritalg totinc hhdtypb OwnRnt08 numcars SXORIEN Religsc
## 1      2   2  58        1     26       2        2       2       1       2
## 2      2   1  47        1     26       2        2       2       1       1
## 3      4   1  39        1     97       3        1       2       1       1
## 4      5   1  41        1     21       4        1       2       1       1
## 5      5   2  37        1     21       4        1       2      -9       1
## 6      4   1  51        1     -1       5        1       2       1       1
##   ethnicC SrcInc7 SrcInc15    eqvinc eqv5 Econact_2 EducEnd HighQual hpnssec5
## 1       1       0        0 105000.00    1         3       8        1        1
## 2       1       0        0 105000.00    1         1       7        1        1
## 3       1       0        0     -1.00   -1         1       5        6        3
## 4       1       0        0  35668.79    2         1       7        1        1
## 5       1       0        0  35668.79    2         4       8        1        1
## 6       1      -1       -1     -1.00   -1         1       8        2        1
##   RG15a docinfo1 compm3 compm7 compm8 compm9 genhelf2 longill12   bmival
## 1     1       -1      0      0      0      0        1         1 36.09377
## 2     2       -1      0      0      1      0        1         1 51.04027
## 3     2       -1      0      0      1      0        1         1 -1.00000
## 4     2       -1      0      0      0      0        1         2 26.42559
## 5     2       -1      0      0      0      0        1         2 23.70948
## 6     2       -1      0      0      0      0        1         2 28.04325
##   ghq12scr wemwbs cigst1 drating Active ActPhy country PROBGAM
## 1        0     -1      1  35.000      2      1       1       0
## 2        0     -1      1  10.674      4      1       1       0
## 3        0     -1      1  44.000      2      2       1       0
## 4        0     -1      3  11.710      2      1       1       0
## 5        0     -1      3  10.960     -1      1       1       0
## 6        0     49      3  41.500      3      1       1       0
```

```
colnames(raw.data)
```

```
##  [1] "HHSize"    "Sex"       "age"       "maritalg"  "totinc"    "hhdtypb"
##  [7] "OwnRnt08"  "numcars"   "SXORIEN"   "Religsc"   "ethnicC"   "SrcInc7"
## [13] "SrcInc15"  "eqvinc"    "eqv5"      "Econact_2" "EducEnd"   "HighQual"
## [19] "hpnssec5"  "RG15a"     "docinfo1"  "compm3"    "compm7"    "compm8"
## [25] "compm9"    "genhelf2"  "longill12" "bmival"    "ghq12scr"  "wemwbs"
## [31] "cigst1"    "drating"   "Active"    "ActPhy"    "country"   "PROBGAM"
```

```
table(raw.data$PROBGAM)
```

```
##
##   -9   -8   -6   -1    0    1
##  320  238  372  767 9710 1699
```

```
# keeping only the raws with values for probgam
data <- raw.data %>%
  filter(PROBGAM %in% c(0, 1))

data[data < 0] <- NA
```

```r
# the initial socio-economic predictors
socioecon_vars <- c("EducEnd", "numcars", "HHSize", "eqv5", "HighQual", "Econact_2", "hhdtypb", "OwnRnt08", "hpns
sec5", "SrcInc7", "SrcInc15", "eqvinc")

# chosen confounder variables
confounder_vars <- c("Sex", "age", "ethnicC", "Religsc")

# converting to factors
data <- data %>%
  mutate(eqv5 = factor(eqv5)) %>%
  mutate(HighQual = factor(HighQual)) %>%
  mutate(Econact_2 = factor(Econact_2)) %>%
  mutate(hhdtypb = factor(hhdtypb)) %>%
  mutate(OwnRnt08 = factor(OwnRnt08)) %>%
  mutate(hpnssec5 = factor(hpnssec5)) %>%
  mutate(SrcInc7 = factor(SrcInc7)) %>%
  mutate(SrcInc15 = factor(SrcInc15)) %>%
  mutate(sex = factor(Sex)) %>%
  mutate(ethnicC = factor(ethnicC)) %>%
  mutate(Religsc = factor(Religsc)) %>%
  mutate(numcars = factor(numcars)) %>%
  mutate(HHSize = factor(HHSize)) %>%
  mutate(totinc = factor(totinc))

summary(raw.data)
```

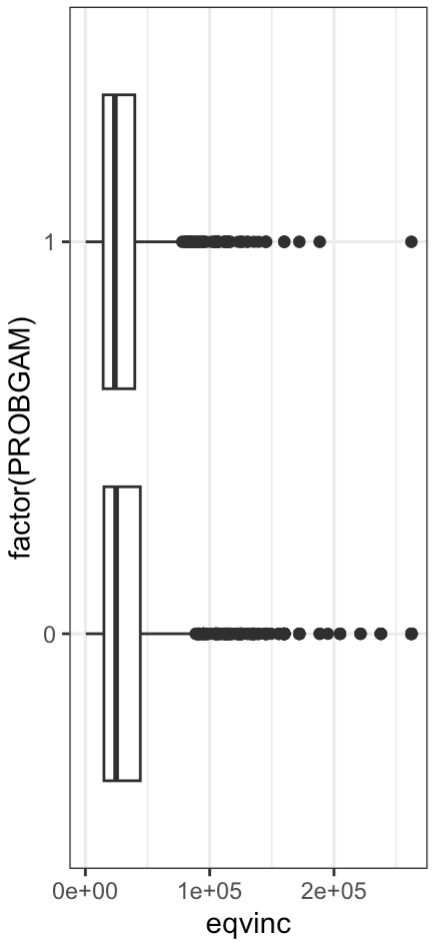
```
##      HHSize            Sex             age           maritalg
##  Min.   : 1.000   Min.   :1.000   Min.   :16.00   Min.   :-9.000
##  1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:36.00   1st Qu.: 1.000
##  Median : 2.000   Median :2.000   Median :50.00   Median : 1.000
##  Mean   : 2.581   Mean   :1.557   Mean   :50.62   Mean   : 2.243
##  3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.:65.00   3rd Qu.: 3.000
##  Max.   :11.000   Max.   :2.000   Max.   :99.00   Max.   : 6.000
##      totinc         hhdtypb          OwnRnt08         numcars
##  Min.   :-1.00   Min.   :-9.000   Min.   :-9.000   Min.   :-1.000
##  1st Qu.:10.00   1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median :16.00   Median : 5.000   Median : 2.000   Median : 1.000
##  Mean   :28.85   Mean   : 4.261   Mean   : 2.311   Mean   : 1.036
##  3rd Qu.:24.00   3rd Qu.: 6.000   3rd Qu.: 4.000   3rd Qu.: 2.000
##  Max.   :97.00   Max.   : 7.000   Max.   : 5.000   Max.   : 3.000
##     SXORIEN           Religsc          ethnicC         SrcInc7
##  Min.   :-9.0000   Min.   :-9.000   Min.   :-9.00   Min.   :-9.00000
##  1st Qu.: 1.0000   1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 0.00000
##  Median : 1.0000   Median : 2.000   Median : 1.00   Median : 0.00000
##  Mean   : 0.3667   Mean   : 2.084   Mean   : 1.26   Mean   :-0.04776
##  3rd Qu.: 1.0000   3rd Qu.: 3.000   3rd Qu.: 1.00   3rd Qu.: 0.00000
##  Max.   : 4.0000   Max.   : 9.000   Max.   : 6.00   Max.   : 1.00000
##     SrcInc15            eqvinc            eqv5          Econact_2
##  Min.   :-9.00000   Min.   :   -90   Min.   :-1.000   Min.   :-9.000
##  1st Qu.: 0.00000   1st Qu.:  8553   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 0.00000   Median : 19500   Median : 2.000   Median : 1.000
##  Mean   :-0.08797   Mean   : 27555   Mean   : 2.283   Mean   : 2.195
##  3rd Qu.: 0.00000   3rd Qu.: 36517   3rd Qu.: 4.000   3rd Qu.: 3.000
##  Max.   : 1.00000   Max.   :262295   Max.   : 5.000   Max.   : 5.000
##     EducEnd          HighQual         hpnssec5          RG15a
##  Min.   :-9.000   Min.   :-9.000   Min.   :-9.000   Min.   :-9.000
##  1st Qu.: 4.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 5.000   Median : 3.000   Median : 2.000   Median : 2.000
##  Mean   : 5.549   Mean   : 3.298   Mean   : 2.504   Mean   : 1.821
##  3rd Qu.: 8.000   3rd Qu.: 4.000   3rd Qu.: 5.000   3rd Qu.: 2.000
##  Max.   : 8.000   Max.   : 6.000   Max.   : 5.000   Max.   : 2.000
##     docinfo1           compm3            compm7            compm8
##  Min.   :-8.0000   Min.   :-9.0000   Min.   :-9.0000   Min.   :-9.00000
##  1st Qu.:-1.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000
```

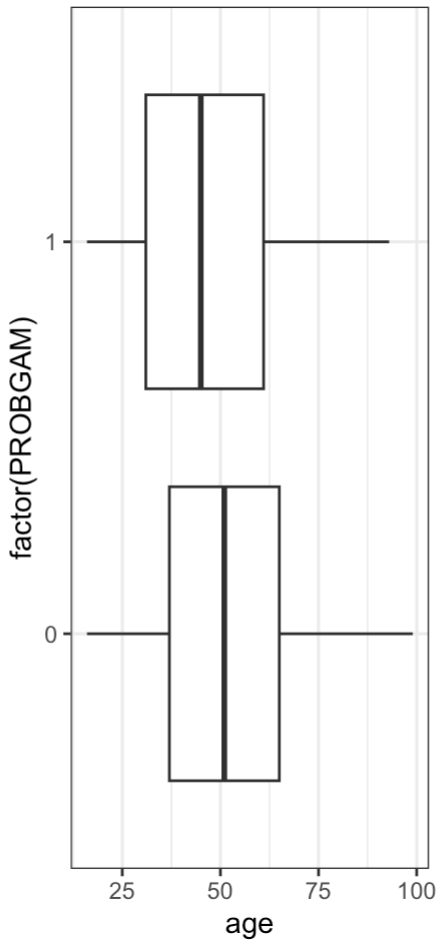
```
##  Median :-1.0000   Median : 0.0000   Median : 0.0000   Median : 0.00000
##  Mean   :-0.8555   Mean   : 0.0531   Mean   : 0.1262   Mean   : 0.06989
##  3rd Qu.:-1.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   : 2.0000   Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.00000
##      compm9           genhelf2        longill12          bmival
##  Min.   :-9.00000   Min.   :-8.000   Min.   :-9.000   Min.   :-1.00
##  1st Qu.: 0.00000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:21.45
##  Median : 0.00000   Median : 1.000   Median : 2.000   Median :25.70
##  Mean   : 0.04433   Mean   : 1.349   Mean   : 1.543   Mean   :22.82
##  3rd Qu.: 0.00000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:29.55
##  Max.   : 1.00000   Max.   : 3.000   Max.   : 2.000   Max.   :62.85
##     ghq12scr          wemwbs          cigst1          drating
##  Min.   :-9.0000   Min.   :-9.00   Min.   :-9.000   Min.   : -9.000
##  1st Qu.: 0.0000   1st Qu.:-1.00   1st Qu.: 1.000   1st Qu.:  0.116
##  Median : 0.0000   Median :47.00   Median : 2.000   Median :  3.591
##  Mean   : 0.8086   Mean   :35.45   Mean   : 2.125   Mean   : 10.473
##  3rd Qu.: 1.0000   3rd Qu.:55.00   3rd Qu.: 3.000   3rd Qu.: 14.000
##  Max.   :12.0000   Max.   :70.00   Max.   : 4.000   Max.   :595.000
##      Active           ActPhy          country         PROBGAM
##  Min.   :-8.0000   Min.   :-8.000   Min.   :1.000   Min.   :-9.0000
##  1st Qu.:-1.0000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.: 0.0000
##  Median : 1.0000   Median : 2.000   Median :1.000   Median : 0.0000
##  Mean   : 0.7783   Mean   : 1.564   Mean   :1.367   Mean   :-0.4642
##  3rd Qu.: 2.0000   3rd Qu.: 2.000   3rd Qu.:2.000   3rd Qu.: 0.0000
##  Max.   : 4.0000   Max.   : 2.000   Max.   :2.000   Max.   : 1.0000
```

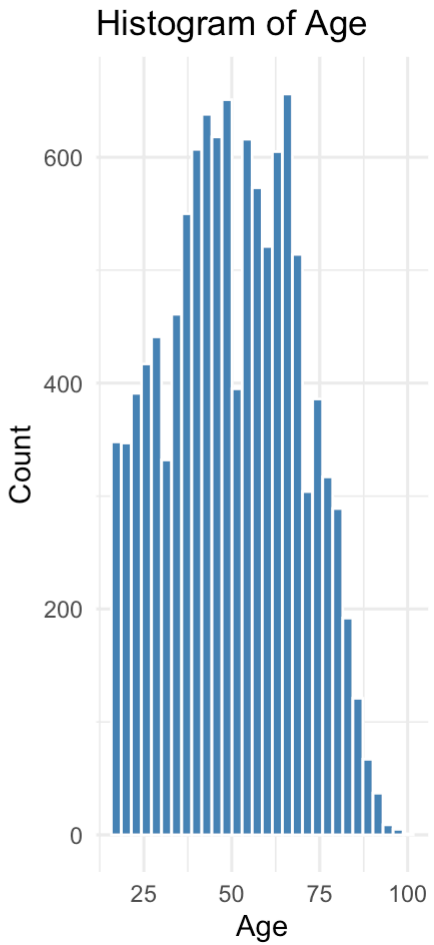
GRAPHS

```r
# boxplots of the two continous variables
p1 <- ggplot(data, aes(x = factor(PROBGAM), y = eqvinc)) +
  geom_boxplot() + coord_flip() +
  theme_bw() + theme(legend.position = "none") + scale_fill_grey()

p2 <- ggplot(data, aes(x = factor(PROBGAM), y = age)) +
  geom_boxplot() + coord_flip() +
  theme_bw() + theme(legend.position = "none") + scale_fill_grey()

# histogram of age
p3 <- ggplot(data, aes(x = age)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  labs(title = "Histogram of Age", x = "Age", y = "Count") +
  theme_minimal()

grid.arrange(p1, p2, p3, nrow = 1)
```

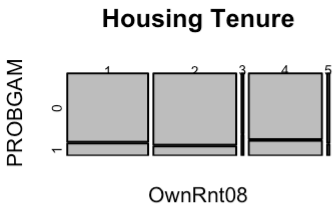
```
## Warning: Removed 1789 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

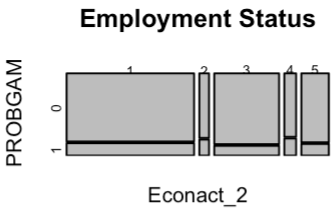

factor(PROBGAM)
1
0
0e+00
1e+05
2e+05
eqvinc

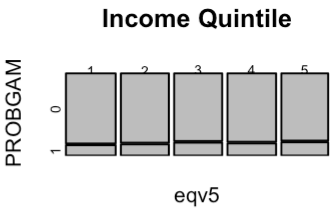

factor(PROBGAM)
1
0
25
50
75
100
age

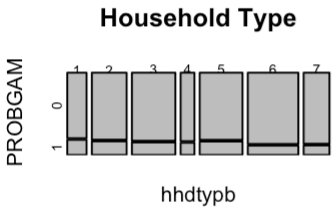

Histogram of Age
Count
600
400
200
0
25
50
75
100
Age

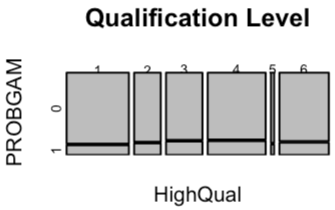
```
# mosaic plots of categorical predictors
# seperated into 2 windows as they wouldn't all fit in one

par(mfrow = c(3, 3))

with(data, mosaicplot(table(OwnRnt08, PROBGAM), main = "Housing Tenure"))
with(data, mosaicplot(table(Econact_2, PROBGAM), main = "Employment Status"))
with(data, mosaicplot(table(eqv5, PROBGAM), main = "Income Quintile"))
with(data, mosaicplot(table(hhdtypb, PROBGAM), main = "Household Type"))
with(data, mosaicplot(table(HighQual, PROBGAM), main = "Qualification Level"))
with(data, mosaicplot(table(SrcInc7, PROBGAM), main = "Income Support"))
with(data, mosaicplot(table(SrcInc15, PROBGAM), main = "No Income Source"))
with(data, mosaicplot(table(hpnssec5, PROBGAM), main = "NS-SEC (5-category)"))
with(data, mosaicplot(table(numcars, PROBGAM), main = "Number of Cars"))
```

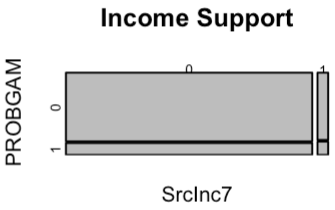


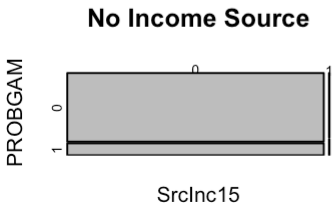


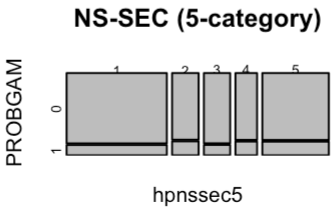


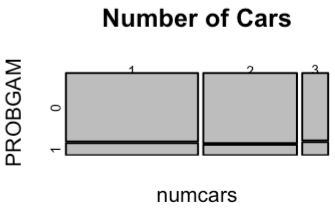




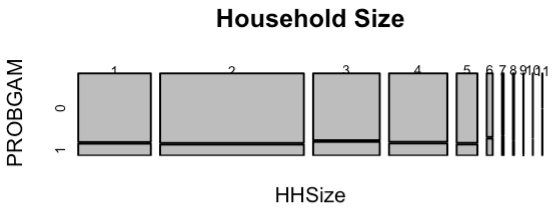


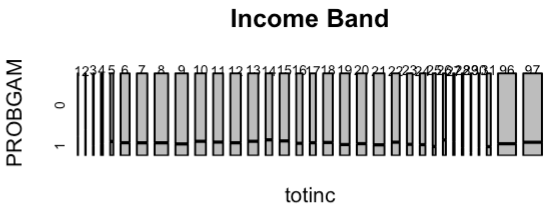


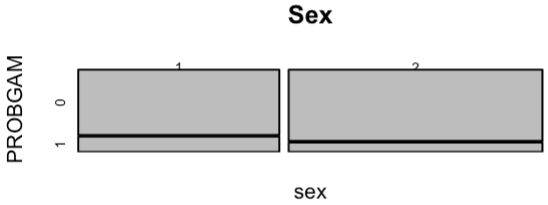


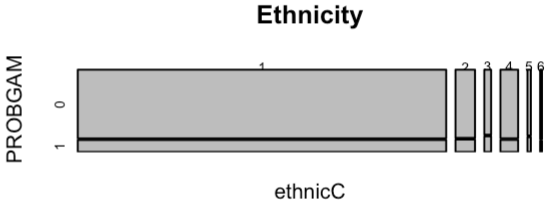


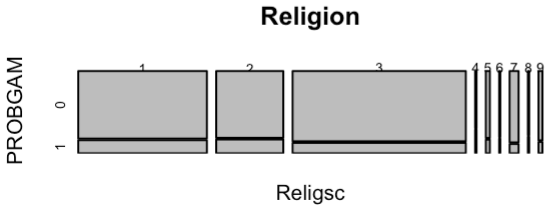

```
par(mfrow = c(3, 2))

with(data, mosaicplot(table(HHSize, PROBGAM), main = "Household Size"))
with(data, mosaicplot(table(totinc, PROBGAM), main = "Income Band"))
with(data, mosaicplot(table(sex, PROBGAM), main = "Sex"))
with(data, mosaicplot(table(ethnicC, PROBGAM), main = "Ethnicity"))
with(data, mosaicplot(table(Religsc, PROBGAM), main = "Religion"))
```











EDA

EducEnd

```
table(data$EducEnd)
```

```
##
##    1    2    3    4    5    6    7    8
##  486   17  581 2137 2977 1098 1124 2986
```

```
# contingency table converted to show percent calculations
educ_table <- table(data$EducEnd, data$PROBGAM)
educ_prop <- prop.table(educ_table, margin = 1)
educ_percent <- round(educ_prop * 100, 1)

educ_percent
```

```
##
##        0    1
##   1 80.0 20.0
##   2 88.2 11.8
##   3 87.1 12.9
##   4 84.6 15.4
##   5 83.8 16.2
##   6 84.8 15.2
##   7 84.9 15.1
##   8 87.4 12.6
```

```
# grouping and relabelling
data$EducEnd_group <- factor(
  ifelse(data$EducEnd %in% c(1, 2, 3), "Under 14",
  ifelse(data$EducEnd %in% c(4, 5, 6, 7), "Under 18",
  ifelse(data$EducEnd == 8, "19 or over", NA)))
)

# relevelling
data$EducEnd_group <- relevel(data$EducEnd_group, ref = "Under 18")

# regression to check significance
glm_educ_group <- glm(PROBGAM ~ EducEnd_group, data = data, family = binomial(link = "logit"))
display(glm_educ_group)
```

```
## glm(formula = PROBGAM ~ EducEnd_group, family = binomial(link = "logit"),
##     data = data)
##                           coef.est coef.se
## (Intercept)               -1.68     0.03
## EducEnd_group19 or over -0.25     0.06
## EducEnd_groupUnder 14    0.03     0.09
## ---
##   n = 11406, k = 3
##   residual deviance = 9583.9, null deviance = 9601.5 (difference = 17.6)
```

```
anova(glm_educ_group, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           11405     9601.5
## EducEnd_group  2   17.553     11403     9583.9 0.0001543 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

numcars

```
table(data$numcards)
```

```
## < table of extent 0 >
```

```
numcars_table <- table(data$numcars, data$PROBGAM)
numcars_prop <- prop.table(numcars_table, margin = 1)
numcars_percent <- round(numcars_prop * 100, 1)

numcars_percent
```

```
##
##        0    1
##   1 85.0 15.0
##   2 87.0 13.0
##   3 83.7 16.3
```

```
data$numcars_group <- factor(data$numcars, levels = c(1, 2, 3), labels = c("1 car", "2 cars", "3 or more"))

data$numcars_group <- relevel(data$numcars_group, ref = "1 car")

glm_numcars <- glm(PROBGAM ~ numcars_group, data = data, family = binomial(link = "logit"))
display(glm_numcars)
```

```
## glm(formula = PROBGAM ~ numcars_group, family = binomial(link = "logit"),
##     data = data)
##                       coef.est coef.se
## (Intercept)           -1.73     0.04
## numcars_group2 cars   -0.17     0.06
## numcars_group3 or more  0.10     0.10
## ---
##   n = 9152, k = 3
##   residual deviance = 7542.2, null deviance = 7551.7 (difference = 9.5)
```

```
anova(glm_numcars, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           9151     7551.7
## numcars_group  2   9.4955      9149     7542.2 0.008671 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#HHSize

```
table(data$HHSize)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11
## 2231 4416 2046 1791  639  185   50   35    4    5    7
```

```
HHSize_table <- table(data$HHSize, data$PROBGAM)
HHSize_prop <- prop.table(HHSize_table, margin = 1)
HHSize_percent <- round(HHSize_prop * 100, 1)
HHSize_percent
```

```
##
##          0     1
##   1   85.3  14.7
##   2   86.4  13.6
##   3   83.0  17.0
##   4   84.9  15.1
##   5   86.2  13.8
##   6   78.9  21.1
##   7   76.0  24.0
##   8   82.9  17.1
##   9  100.0   0.0
##   10  80.0  20.0
##   11  42.9  57.1
```

```
data$HHSize_group <- factor(
  ifelse(data$HHSize %in% c(1, 2), "1–2",
  ifelse(data$HHSize %in% c(3, 4, 5), "3–5",
  ifelse(data$HHSize %in% c(6, 7, 8, 9, 10, 11), "6+", NA)))
)

data$HHSize_group <- relevel(data$HHSize_group, ref = "1–2")

glm_HHSize <- glm(PROBGAM ~ HHSize_group, data = data, family = binomial(link = "logit"))
display(glm_HHSize)
```

```
## glm(formula = PROBGAM ~ HHSize_group, family = binomial(link = "logit"),
##     data = data)
##                 coef.est coef.se
## (Intercept)     -1.81     0.04
## HHSize_group3–5  0.14     0.05
## HHSize_group6+   0.53     0.15
## ---
##   n = 11409, k = 3
##   residual deviance = 9586.2, null deviance = 9602.4 (difference = 16.2)
```

```
anova(glm_HHSize, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          11408     9602.4
## HHSize_group  2   16.215     11406     9586.2 0.0003013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#eqv5

```
table(data$eqv5)
```

```
##
##    1    2    3    4    5
## 1982 1914 1916 1941 1867
```

```
eqv5_table <- table(data$eqv5, data$PROBGAM)
eqv5_prop <- prop.table(eqv5_table, margin = 1)
eqv5_percent <- round(eqv5_prop * 100, 1)

eqv5_percent
```

```
##
##        0    1
##   1 87.0 13.0
##   2 86.1 13.9
##   3 84.0 16.0
##   4 84.9 15.1
##   5 83.4 16.6
```

```
data$eqv5_group <- factor(
  ifelse(data$eqv5 %in% c(1, 2), "High Income",
  ifelse(data$eqv5 == 3, "Middle Income",
  ifelse(data$eqv5 %in% c(4, 5), "Low Income", NA)))
)

data$eqv5_group <- relevel(data$eqv5_group, ref = "Middle Income")

glm_eqv5_group <- glm(PROBGAM ~ eqv5_group, data = data, family = binomial(link = "logit"))
display(glm_eqv5_group)
```

```
## glm(formula = PROBGAM ~ eqv5_group, family = binomial(link = "logit"),
##     data = data)
##                       coef.est coef.se
## (Intercept)           -1.66     0.06
## eqv5_groupHigh Income -0.21     0.08
## eqv5_groupLow Income  -0.02     0.08
## ---
##   n = 9620, k = 3
##   residual deviance = 8087.2, null deviance = 8098.1 (difference = 11.0)
```

```
anova(glm_eqv5_group, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         9619     8098.1
## eqv5_group  2   10.953      9617     8087.2 0.004185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#HighQual

```
table(data$HighQual)
```

```
##
##    1    2    3    4    5    6
## 3007 1286 1759 2800  154 2387
```

```
HighQual_table <- table(data$HighQual, data$PROBGAM)
HighQual_prop <- prop.table(HighQual_table, margin = 1)
HighQual_percent <- round(HighQual_prop * 100, 1)
HighQual_percent
```

```
##
##        0    1
##   1 88.0 12.0
##   2 85.7 14.3
##   3 83.3 16.7
##   4 82.9 17.1
##   5 87.0 13.0
##   6 84.8 15.2
```

```
data$HighQual_group <- factor(
  ifelse(data$HighQual %in% c(1, 2), "Higher Education",
  ifelse(data$HighQual %in% c(3, 4), "Secondary/A-Levels",
  ifelse(data$HighQual %in% c(5, 6), "Low/No Quals", NA)))
)

data$HighQual_group <- relevel(data$HighQual_group, ref = "Secondary/A-Levels")

glm_HighQual <- glm(PROBGAM ~ HighQual_group, data = data, family = binomial(link = "logit"))
display(glm_HighQual)
```

```
## glm(formula = PROBGAM ~ HighQual_group, family = binomial(link = "logit"),
##     data = data)
##                               coef.est coef.se
## (Intercept)                   -1.59     0.04
## HighQual_groupHigher Education -0.34     0.06
## HighQual_groupLow/No Quals     -0.14     0.07
## ---
##   n = 11393, k = 3
##   residual deviance = 9558.7, null deviance = 9590.3 (difference = 31.6)
```

```
anova(glm_HighQual, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           11392     9590.3
## HighQual_group  2   31.639     11390     9558.7 1.348e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Econact_2

```
table(data$Econact_2)
```

```
##
##    1    2    3    4    5
## 6022  458 3061  557 1298
```

```
econ_table <- table(data$Econact_2, data$PROBGAM)
econ_prop <- prop.table(econ_table, margin = 1)
econ_percent <- round(econ_prop * 100, 1)
econ_percent
```

```
##
##        0    1
##   1 84.6 15.4
##   2 80.1 19.9
##   3 87.9 12.1
##   4 78.8 21.2
##   5 85.6 14.4
```

```
data$Econact_2 <- factor(data$Econact_2,
                         levels = c(1, 2, 3, 4, 5),
                         labels = c("Employed/Training",
                                    "Education",
                                    "Retired",
                                    "Unemployed",
                                    "Other"))

data$Econact_2 <- relevel(data$Econact_2, ref = "Employed/Training")

glm_Econact <- glm(PROBGAM ~ Econact_2, data = data, family = binomial(link = "logit"))
display(glm_Econact)
```

```
## glm(formula = PROBGAM ~ Econact_2, family = binomial(link = "logit"),
##     data = data)
##                     coef.est coef.se
## (Intercept)         -1.70     0.04
## Econact_2Education   0.31     0.12
## Econact_2Retired    -0.28     0.07
## Econact_2Unemployed  0.39     0.11
## Econact_2Other      -0.08     0.09
## ---
##   n = 11396, k = 5
##   residual deviance = 9538.6, null deviance = 9584.3 (difference = 45.7)
```

```
anova(glm_Econact, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      11395     9584.3
## Econact_2  4   45.657     11391     9538.6 2.902e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

hhdtypb

```
table(data$hhdtypb)
```

```
##
##    1    2    3    4    5    6    7
##  956 1721 2149  684 2117 2501 1278
```

```
hhd_table <- table(data$hhdtypb, data$PROBGAM)
hhd_prop <- prop.table(hhd_table, margin = 1)
hhd_percent <- round(hhd_prop * 100, 1)
hhd_percent
```

```
##
##        0    1
##   1 81.3 18.7
##   2 83.1 16.9
##   3 84.5 15.5
##   4 84.9 15.1
##   5 83.3 16.7
##   6 88.4 11.6
##   7 88.2 11.8
```

```
data$hhdtypb_group <- factor(
  ifelse(data$hhdtypb == 1, "Single Adult",
  ifelse(data$hhdtypb %in% c(2, 5), "Multiple Adults",
  ifelse(data$hhdtypb %in% c(3, 4), "Family Household",
  ifelse(data$hhdtypb %in% c(6, 7), "Senior Household", NA))))
)

data$hhdtypb_group <- relevel(data$hhdtypb_group, ref = "Family Household")

glm_hhdtypb_group <- glm(PROBGAM ~ hhdtypb_group, data = data, family = binomial(link = "logit"))
display(glm_hhdtypb_group)
```

```
## glm(formula = PROBGAM ~ hhdtypb_group, family = binomial(link = "logit"),
##     data = data)
##                               coef.est coef.se
## (Intercept)                   -1.70     0.05
## hhdtypb_groupMultiple Adults   0.10     0.07
## hhdtypb_groupSenior Household -0.33     0.07
## hhdtypb_groupSingle Adult      0.23     0.10
## ---
##   n = 11406, k = 4
##   residual deviance = 9546.7, null deviance = 9601.5 (difference = 54.8)
```

```
anova(glm_hhdtypb_group, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           11405     9601.5
## hhdtypb_group  3   54.769     11402     9546.7 7.692e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#OwnRnt08

```
table(data$OwnRnt08)
```

```
##
##    1    2    3    4    5
## 3820 3912   98 3454  108
```

```
own_table <- table(data$OwnRnt08, data$PROBGAM)
own_prop <- prop.table(own_table, margin = 1)
own_percent <- round(own_prop * 100, 1)
own_percent
```

```
##
##        0    1
##   1 84.9 15.1
##   2 88.6 11.4
##   3 74.5 25.5
##   4 81.6 18.4
##   5 86.1 13.9
```

```
data$OwnRnt08 <- factor(data$OwnRnt08,
  levels = c(1, 2, 3, 4, 5),
  labels = c("Owner",
             "Shared Ownership",
             "Social Rent",
             "Private Rent",
             "Other"))

data$OwnRnt08 <- relevel(data$OwnRnt08, ref = "Owner")

glm_own <- glm(PROBGAM ~ OwnRnt08, data = data, family = binomial(link = "logit"))
display(glm_own)
```

```
## glm(formula = PROBGAM ~ OwnRnt08, family = binomial(link = "logit"),
##     data = data)
##                           coef.est coef.se
## (Intercept)               -1.73     0.05
## OwnRnt08Shared Ownership -0.32     0.07
## OwnRnt08Social Rent       0.65     0.24
## OwnRnt08Private Rent      0.23     0.06
## OwnRnt08Other            -0.10     0.28
## ---
##   n = 11392, k = 5
##   residual deviance = 9515.2, null deviance = 9593.5 (difference = 78.3)
```

```
anova(glm_own, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     11391     9593.5
## OwnRnt08  4   78.278     11387     9515.2 4.033e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#hppnsec

```
table(data$hpnssec5)
```

```
##
##    1    2    3    4    5
## 4631 1227 1215 1000 3073
```

```
hpnssec_table <- table(data$hpnssec5, data$PROBGAM)
hpnssec_prop <- prop.table(hpnssec_table, margin = 1)
hpnssec_percent <- round(hpnssec_prop * 100, 1)
hpnssec_percent
```

```
##
##        0    1
##   1 87.3 12.7
##   2 82.7 17.3
##   3 87.2 12.8
##   4 83.0 17.0
##   5 83.1 16.9
```

```
data$hpnssec5 <- factor(data$hpnssec5,
  levels = c(1, 2, 3, 4, 5, 99),
  labels = c("Managerial/Professional",
             "Intermediate",
             "Self-Employed",
             "Technical/Supervisory",
             "Semi-Routine",
             "Other")
)

data$hpnssec5 <- relevel(data$hpnssec5, ref = "Managerial/Professional")

glm_hpnssec <- glm(PROBGAM ~ hpnssec5, data = data, family = binomial(link = "logit"))
display(glm_hpnssec)
```

```
## glm(formula = PROBGAM ~ hpnssec5, family = binomial(link = "logit"),
##     data = data)
##                               coef.est coef.se
## (Intercept)                   -1.92     0.04
## hpnssec5Intermediate           0.36     0.09
## hpnssec5Self-Employed          0.01     0.10
## hpnssec5Technical/Supervisory  0.34     0.10
## hpnssec5Semi-Routine           0.33     0.07
## ---
##   n = 11146, k = 5
##   residual deviance = 9299.7, null deviance = 9339.7 (difference = 40.1)
```

```
anova(glm_hpnssec, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     11145     9339.7
## hpnssec5  4   40.052     11141     9299.7 4.223e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SrcInc7

```
table(data$SrcInc7)
```

```
##
##     0     1
## 10578   459
```

```
src7_table <- table(data$SrcInc7, data$PROBGAM)
src7_prop <- prop.table(src7_table, margin = 1)
src7_percent <- round(src7_prop * 100, 1)
src7_percent
```

```
##
##        0    1
##   0 85.2 14.8
##   1 83.4 16.6
```

```
data$SrcInc7 <- factor(data$SrcInc7,
  levels = c(0, 1),
  labels = c("No", "Yes")
)

data$SrcInc7 <- relevel(data$SrcInc7, ref = "No")

glm_SrcInc7 <- glm(PROBGAM ~ SrcInc7, data = data, family = binomial(link = "logit"))
display(glm_SrcInc7)
```

```
## glm(formula = PROBGAM ~ SrcInc7, family = binomial(link = "logit"),
##     data = data)
##             coef.est coef.se
## (Intercept) -1.75     0.03
## SrcInc7Yes   0.14     0.13
## ---
##   n = 11037, k = 2
##   residual deviance = 9265.2, null deviance = 9266.3 (difference = 1.1)
```

```
anova(glm_SrcInc7, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    11036     9266.3
## SrcInc7  1   1.0965     11035     9265.2    0.295
```

SrcInc15

it only has two categories and second category only has a handful of data points so excluding it from anaylsis.

```
table(data$SrcInc15)
```

```
##
##     0     1
## 11017    20
```

#totinc # initially considered including totinc which represents the total income band, however evinc and eqv5 are already similar.

```
table(data$totinc)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##  13  19  36  95 150 431 502 648 597 516 473 500 492 322 451 299 292 460 459 452
##  21  22  23  24  25  26  27  28  29  30  31  96  97
## 522 381 267 268 144 141  72  69  35  43 195 860 894
```

```
totinc_table <- table(data$totinc, data$PROBGAM)
totinc_prop <- prop.table(totinc_table, margin = 1)
totinc_percent <- round(totinc_prop * 100, 1)
totinc_percent
```

```
##
##         0    1
##   1  76.9 23.1
##   2  68.4 31.6
##   3  88.9 11.1
##   4  86.3 13.7
##   5  83.3 16.7
##   6  84.9 15.1
##   7  85.1 14.9
##   8  85.0 15.0
##   9  86.4 13.6
##   10 83.1 16.9
##   11 84.1 15.9
##   12 85.0 15.0
##   13 83.1 16.9
##   14 81.4 18.6
##   15 82.7 17.3
##   16 85.6 14.4
##   17 84.9 15.1
##   18 84.6 15.4
##   19 87.1 12.9
##   20 86.1 13.9
##   21 87.5 12.5
##   22 84.3 15.7
##   23 86.9 13.1
##   24 87.3 12.7
##   25 89.6 10.4
##   26 81.6 18.4
##   27 83.3 16.7
##   28 91.3  8.7
##   29 97.1  2.9
##   30 90.7  9.3
##   31 89.7 10.3
##   96 86.2 13.8
##   97 84.3 15.7
```

Sex

```
table(data$Sex)
```

```
##
##    1    2
## 5044 6365
```

```
sex_table <- table(data$Sex, data$PROBGAM)
sex_prop <- prop.table(sex_table, margin = 1)
sex_percent <- round(sex_prop * 100, 1)
sex_percent
```

```
##
##        0    1
##   1 81.0 19.0
##   2 88.3 11.7
```

```
data$Sex <- factor(
  ifelse(data$Sex == 1, "Male",
  ifelse(data$Sex == 2, "Female", NA))
)

data$Sex <- relevel(data$Sex, ref = "Male")

glm_sex <- glm(PROBGAM ~ Sex, data = data, family = binomial(link = "logit"))
display(glm_sex)
```

```
## glm(formula = PROBGAM ~ Sex, family = binomial(link = "logit"),
##     data = data)
##             coef.est coef.se
## (Intercept) -1.45     0.04
## SexFemale   -0.57     0.05
## ---
##   n = 11409, k = 2
##   residual deviance = 9484.4, null deviance = 9602.4 (difference = 118.0)
```

```
anova(glm_sex, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                 11408     9602.4
## Sex   1   117.98     11407     9484.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ethnicC

```
table(data$ethnicC)
```

```
##
##     1     2     3     4     5     6
## 10052   533   186   484    93    50
```

```
ethnic_table <- table(data$ethnicC, data$PROBGAM)
ethnic_prop <- prop.table(ethnic_table, margin = 1)
ethnic_percent <- round(ethnic_prop * 100, 1)
ethnic_percent
```

```
##
##        0    1
##   1 85.2 14.8
##   2 84.6 15.4
##   3 80.6 19.4
##   4 85.1 14.9
##   5 81.7 18.3
##   6 86.0 14.0
```

```
data$ethnicC_group <- factor(
  ifelse(data$ethnicC %in% c(1, 2), "White",
  ifelse(data$ethnicC == 3, "Black",
  ifelse(data$ethnicC == 4, "Asian",
  ifelse(data$ethnicC == 5, "Mixed",
  ifelse(data$ethnicC == 6, "Other", NA)))))
)

data$ethnicC_group <- relevel(data$ethnicC_group, ref = "White")

glm_ethnic_group <- glm(PROBGAM ~ ethnicC_group, data = data, family = binomial(link = "logit"))
display(glm_ethnic_group)
```

```
## glm(formula = PROBGAM ~ ethnicC_group, family = binomial(link = "logit"),
##     data = data)
##                    coef.est coef.se
## (Intercept)        -1.75     0.03
## ethnicC_groupAsian  0.01     0.13
## ethnicC_groupBlack  0.32     0.19
## ethnicC_groupMixed  0.25     0.27
## ethnicC_groupOther -0.07     0.41
## ---
##   n = 11398, k = 5
##   residual deviance = 9595.3, null deviance = 9598.9 (difference = 3.6)
```

```
anova(glm_ethnic_group, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          11397     9598.9
## ethnicC_group  4   3.6038     11393     9595.3   0.4623
```

Religsc

```
table(data$Religsc)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 3732 1947 5025   41  118   33  261   39  114
```

```
relig_table <- table(data$Religsc, data$PROBGAM)
relig_prop <- prop.table(relig_table, margin = 1)
relig_percent <- round(relig_prop * 100, 1)
relig_percent
```

```
##
##        0    1
##   1 83.8 16.2
##   2 82.9 17.1
##   3 87.1 12.9
##   4 70.7 29.3
##   5 83.1 16.9
##   6 81.8 18.2
##   7 88.5 11.5
##   8 79.5 20.5
##   9 86.0 14.0
```

```
data$Religsc_group <- factor(
  ifelse(data$Religsc == 1, "No Religion",
  ifelse(data$Religsc == 2, "Catholic Christian",
  ifelse(data$Religsc == 3, "Non-Catholic Crhistian",
  ifelse(data$Religsc %in% 4:9, "Other Religion", NA))))
)

data$Religsc_group <- relevel(data$Religsc_group, ref = "Catholic Christian")

glm_relig <- glm(PROBGAM ~ Religsc_group, data = data, family = binomial(link = "logit"))
display(glm_relig)
```

```
## glm(formula = PROBGAM ~ Religsc_group, family = binomial(link = "logit"),
##     data = data)
##                                   coef.est coef.se
## (Intercept)                       -1.58     0.06
## Religsc_groupNo Religion          -0.06     0.07
## Religsc_groupNon-Catholic Crhistian -0.33     0.07
## Religsc_groupOther Religion       -0.14     0.13
## ---
##   n = 11310, k = 4
##   residual deviance = 9471.9, null deviance = 9500.8 (difference = 28.9)
```

```
anova(glm_relig, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PROBGAM
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          11309     9500.8
## Religsc_group  3   28.884     11306     9471.9 2.369e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#VIF # All the adjusted values are under 5 so keeping everything

```
cleaned_data <- data[, c("PROBGAM", "age", "Sex", "ethnicC_group", "Religsc_group",
                         "HHSize_group", "numcars_group", "eqvinc", "hhdtypb_group",
                         "eqv5_group", "HighQual_group", "Econact_2", "OwnRnt08",
                         "hpnssec5", "SrcInc7")]


vif_model <- lm(as.numeric(PROBGAM) ~ ., data = cleaned_data)
vif(vif_model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## age             3.350658  1        1.830480
## Sex             1.045124  1        1.022313
## ethnicC_group   1.886897  4        1.082601
## Religsc_group   2.033457  3        1.125570
## HHSize_group    3.020606  2        1.318328
## numcars_group   1.563476  2        1.118208
## eqvinc          1.924212  1        1.387160
## hhdtypb_group   7.150598  3        1.388003
## eqv5_group      2.389286  2        1.243275
## HighQual_group 1.510521  2        1.108617
## Econact_2       3.684533  4        1.177058
## OwnRnt08        1.992435  4        1.089991
## hpnssec5        1.519373  4        1.053678
## SrcInc7         1.133183  1        1.064511
```

MODEL

forward selection. After adding my chosen predictors of interest, I removed the least significant variable ethnicity, which I had initially added as a confounder variable Multiple socio-economic variables also had non-

significant values but I chose to keep them. Next, I tested out all other variables I had initially foregone and checked for significance by gradually adding them to the final model.

```
glm1 <- glm(PROBGAM ~ ., data = cleaned_data, family = binomial(link = "logit"))
display(glm1, detail = TRUE, digits = 3)
```

```
## glm(formula = PROBGAM ~ ., family = binomial(link = "logit"),
##     data = cleaned_data)
##                                   coef.est coef.se z value Pr(>|z|)
## (Intercept)                       -0.425    0.232  -1.836   0.066
## age                               -0.018    0.003  -5.077   0.000
## SexFemale                         -0.625    0.069  -9.101   0.000
## ethnicC_groupAsian                -0.197    0.251  -0.785   0.433
## ethnicC_groupBlack                 0.275    0.256   1.074   0.283
## ethnicC_groupMixed                 0.120    0.347   0.345   0.730
## ethnicC_groupOther                 0.032    0.547   0.059   0.953
## Religsc_groupNo Religion          -0.159    0.098  -1.633   0.102
## Religsc_groupNon-Catholic Crhistian -0.166  0.095  -1.745   0.081
## Religsc_groupOther Religion       -0.095    0.223  -0.425   0.671
## HHSize_group3-5                    0.012    0.109   0.109   0.913
## HHSize_group6+                     0.404    0.248   1.628   0.103
## numcars_group2 cars               -0.162    0.081  -2.000   0.046
## numcars_group3 or more             0.028    0.129   0.217   0.828
## eqvinc                             0.000    0.000  -0.883   0.377
## hhdtypb_groupMultiple Adults       0.152    0.102   1.490   0.136
## hhdtypb_groupSenior Household     -0.014    0.187  -0.076   0.940
## hhdtypb_groupSingle Adult          0.302    0.170   1.778   0.075
## eqv5_groupHigh Income             -0.113    0.102  -1.108   0.268
## eqv5_groupLow Income              -0.160    0.097  -1.648   0.099
## HighQual_groupHigher Education    -0.221    0.079  -2.804   0.005
## HighQual_groupLow/No Quals        -0.067    0.107  -0.629   0.529
## Econact_2Education                -0.314    0.195  -1.604   0.109
## Econact_2Retired                   0.314    0.140   2.241   0.025
## Econact_2Unemployed                0.222    0.163   1.360   0.174
## Econact_2Other                     0.042    0.135   0.307   0.759
## OwnRnt08Shared Ownership          -0.164    0.099  -1.660   0.097
## OwnRnt08Social Rent                0.316    0.347   0.911   0.363
## OwnRnt08Private Rent               0.147    0.090   1.629   0.103
## OwnRnt08Other                      0.130    0.331   0.392   0.695
## hpnssec5Intermediate               0.222    0.115   1.924   0.054
## hpnssec5Self-Employed             -0.085    0.123  -0.689   0.491
## hpnssec5Technical/Supervisory      0.182    0.123   1.477   0.140
## hpnssec5Semi-Routine               0.223    0.098   2.279   0.023
## SrcInc7Yes                        -0.395    0.263  -1.501   0.133
```

```
## ---
##   n = 7666, k = 35
##   residual deviance = 6037.2, null deviance = 6280.2 (difference = 243.1)
```

```
Anova(glm1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              25.771  1  3.844e-07 ***
## Sex              84.399  1  < 2.2e-16 ***
## ethnicC_group     1.966  4    0.74204
## Religsc_group     3.418  3    0.33156
## HHSize_group      2.968  2    0.22671
## numcars_group     5.146  2    0.07630 .
## eqvinc            0.802  1    0.37059
## hhdtypb_group     6.181  3    0.10313
## eqv5_group        3.068  2    0.21564
## HighQual_group    7.893  2    0.01932 *
## Econact_2        10.110  4    0.03862 *
## OwnRnt08          9.070  4    0.05936 .
## hpnssec5         10.585  4    0.03165 *
## SrcInc7           2.406  1    0.12090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# dropping ethnicity but keeping the rest because they're in my set of interest even if they're not significant
glmtemp <- glm(PROBGAM ~ . - ethnicC_group, data = cleaned_data, family = binomial(link = "logit"))
Anova(glmtemp)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              25.630  1  4.136e-07 ***
## Sex              83.689  1  < 2.2e-16 ***
## Religsc_group     3.688  3    0.29724
## HHSize_group      2.784  2    0.24862
## numcars_group     5.242  2    0.07275 .
## eqvinc            0.811  1    0.36792
## hhdtypb_group     6.205  3    0.10204
## eqv5_group        3.014  2    0.22160
## HighQual_group    7.769  2    0.02056 *
## Econact_2        10.035  4    0.03985 *
## OwnRnt08          9.725  4    0.04532 *
## hpnssec5         10.635  4    0.03099 *
## SrcInc7           2.335  1    0.12646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
display(glmtemp, detail = TRUE, digits = 3)
```

```
## glm(formula = PROBGAM ~ . - ethnicC_group, family = binomial(link = "logit"),
##     data = cleaned_data)
##                                   coef.est coef.se z value Pr(>|z|)
## (Intercept)                       -0.422    0.231  -1.827   0.068
## age                               -0.018    0.003  -5.063   0.000
## SexFemale                         -0.622    0.069  -9.064   0.000
## Religsc_groupNo Religion          -0.164    0.097  -1.693   0.090
## Religsc_groupNon-Catholic Crhistian -0.165  0.095  -1.736   0.083
## Religsc_groupOther Religion       -0.203    0.180  -1.131   0.258
## HHSize_group3-5                    0.012    0.109   0.107   0.915
## HHSize_group6+                     0.389    0.247   1.576   0.115
## numcars_group2 cars               -0.165    0.081  -2.041   0.041
## numcars_group3 or more             0.022    0.129   0.175   0.861
## eqvinc                             0.000    0.000  -0.888   0.375
## hhdtypb_groupMultiple Adults       0.151    0.102   1.484   0.138
## hhdtypb_groupSenior Household     -0.018    0.187  -0.094   0.925
## hhdtypb_groupSingle Adult          0.300    0.170   1.770   0.077
## eqv5_groupHigh Income             -0.112    0.102  -1.098   0.272
## eqv5_groupLow Income              -0.158    0.097  -1.634   0.102
## HighQual_groupHigher Education    -0.219    0.079  -2.784   0.005
## HighQual_groupLow/No Quals        -0.070    0.107  -0.658   0.511
## Econact_2Education                -0.304    0.195  -1.555   0.120
## Econact_2Retired                   0.313    0.140   2.233   0.026
## Econact_2Unemployed                0.228    0.163   1.402   0.161
## Econact_2Other                     0.036    0.135   0.263   0.792
## OwnRnt08Shared Ownership          -0.166    0.099  -1.682   0.093
## OwnRnt08Social Rent                0.336    0.346   0.971   0.331
## OwnRnt08Private Rent               0.154    0.090   1.713   0.087
## OwnRnt08Other                      0.121    0.331   0.367   0.714
## hpnssec5Intermediate               0.224    0.116   1.936   0.053
## hpnssec5Self-Employed             -0.089    0.123  -0.727   0.467
## hpnssec5Technical/Supervisory      0.179    0.123   1.451   0.147
## hpnssec5Semi-Routine               0.221    0.098   2.255   0.024
## SrcInc7Yes                        -0.389    0.263  -1.480   0.139
## ---
##   n = 7666, k = 31
##   residual deviance = 6039.1, null deviance = 6280.2 (difference = 241.1)
```

```r
# removed ethnicity and added first test variable, drating
model_data <- cleaned_data[, !(names(cleaned_data) %in% "ethnicC_group")]
model_data$drating <- data$drating

glm_updated <- glm(PROBGAM ~ ., data = model_data, family = binomial(link = "logit"))
Anova(glm_updated)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              27.807  1  1.340e-07 ***
## Sex              70.887  1  < 2.2e-16 ***
## Religsc_group     4.097  3    0.25121
## HHSize_group      2.526  2    0.28285
## numcars_group     5.610  2    0.06051 .
## eqvinc            0.987  1    0.32057
## hhdtypb_group     5.945  3    0.11433
## eqv5_group        2.683  2    0.26150
## HighQual_group    7.757  2    0.02068 *
## Econact_2         8.021  4    0.09082 .
## OwnRnt08          8.833  4    0.06542 .
## hpnssec5         10.761  4    0.02939 *
## SrcInc7           2.119  1    0.14551
## drating          16.038  1  6.209e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# adding wemwbs
model_data <- model_data %>%
  mutate(wemwbs = data$wemwbs)
glm_updated <- glm(PROBGAM ~ ., data = model_data, family = binomial(link = "logit"))
Anova(glm_updated)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              17.480  1  2.904e-05 ***
## Sex              69.257  1  < 2.2e-16 ***
## Religsc_group     1.738  3  0.6284591
## HHSize_group      4.722  2  0.0943334 .
## numcars_group     5.514  2  0.0634711 .
## eqvinc            2.192  1  0.1387155
## hhdtypb_group     4.575  3  0.2056909
## eqv5_group        1.903  2  0.3860665
## HighQual_group    5.468  2  0.0649558 .
## Econact_2         8.115  4  0.0874411 .
## OwnRnt08          9.078  4  0.0591758 .
## hpnssec5          5.104  4  0.2768145
## SrcInc7           1.567  1  0.2106994
## drating          14.367  1  0.0001504 ***
## wemwbs            8.770  1  0.0030616 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# adding cigst1
model_data <- model_data %>%
  mutate(cigst1 = data$cigst1)
glm_updated <- glm(PROBGAM ~ ., data = model_data, family = binomial(link = "logit"))
Anova(glm_updated)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              18.622  1  1.594e-05 ***
## Sex              69.055  1  < 2.2e-16 ***
## Religsc_group     2.091  3  0.5536880
## HHSize_group      4.630  2  0.0987461 .
## numcars_group     5.411  2  0.0668530 .
## eqvinc            2.064  1  0.1507737
## hhdtypb_group     4.543  3  0.2084717
## eqv5_group        1.968  2  0.3738531
## HighQual_group    4.993  2  0.0823874 .
## Econact_2         7.903  4  0.0952012 .
## OwnRnt08          7.972  4  0.0926271 .
## hpnssec5          4.495  4  0.3431254
## SrcInc7           1.647  1  0.1994197
## drating          11.872  1  0.0005699 ***
## wemwbs            7.875  1  0.0050114 **
## cigst1            3.548  1  0.0596078 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# adding country
model_data <- model_data %>%
  mutate(country = data$country)
glm_updated <- glm(PROBGAM ~ ., data = model_data, family = binomial(link = "logit"))
Anova(glm_updated)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##                LR Chisq Df Pr(>Chisq)
## age              18.395  1  1.795e-05 ***
## Sex              69.104  1  < 2.2e-16 ***
## Religsc_group     1.490  3  0.6845706
## HHSize_group      4.559  2  0.1023301
## numcars_group     5.605  2  0.0606453 .
## eqvinc            2.346  1  0.1255790
## hhdtypb_group     4.748  3  0.1912194
## eqv5_group        2.211  2  0.3310181
## HighQual_group    4.691  2  0.0957937 .
## Econact_2         7.880  4  0.0960765 .
## OwnRnt08          7.857  4  0.0969545 .
## hpnssec5          4.776  4  0.3110122
## SrcInc7           1.637  1  0.2007135
## drating          11.317  1  0.0007682 ***
## wemwbs            9.342  1  0.0022398 **
## cigst1            3.926  1  0.0475510 *
## country           5.226  1  0.0222539 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

factoring the newly added variables

drating (total units of alchol a week) is a scale so leave as is # wemwbs (well-being scale) also continuous, so leave as is # cigst1, kept the four categories as is, relevelled to never smoked

```
table(data$cigst1)
```

```
##
##    1    2    3    4
## 5512  621 2930 2310
```

```
cigst_table <- table(data$cigst1, data$PROBGAM)
cigst_prop <- prop.table(cigst_table, margin = 1)
cigst_percent <- round(cigst_prop * 100, 1)
print(cigst_percent)
```

```
##
##        0    1
##   1 86.3 13.7
##   2 84.5 15.5
##   3 86.5 13.5
##   4 80.5 19.5
```

```
model_data$cigst1 <- factor(
  ifelse(model_data$cigst1 == 1, "Never smoked",
  ifelse(model_data$cigst1 == 2, "Occasional smoker",
  ifelse(model_data$cigst1 == 3, "Regular smoker",
  ifelse(model_data$cigst1 == 4, "Current smoker", NA))))
)

model_data$cigst1 <- relevel(model_data$cigst1, ref = "Never smoked")

glm_cigst <- glm(PROBGAM ~ cigst1, data = model_data, family = binomial(link = "logit"))

library(car)
Anova(glm_cigst)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##        LR Chisq Df Pr(>Chisq)
## cigst1    46.99  3  3.493e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

country, kept the two categories, relevelled to England

```
model_data$country <- factor(
  ifelse(model_data$country == 1, "England",
  ifelse(model_data$country == 2, "Scotland", NA)),
  levels = c("England", "Scotland")
)

model_data$country <- relevel(model_data$country, ref = "England")

glm_country <- glm(PROBGAM ~ country, data = model_data, family = binomial(link = "logit"))
display(glm_country)
```

```
## glm(formula = PROBGAM ~ country, family = binomial(link = "logit"),
##     data = model_data)
##                 coef.est coef.se
## (Intercept)     -1.69     0.03
## countryScotland -0.16     0.06
## ---
##   n = 11409, k = 2
##   residual deviance = 9594.2, null deviance = 9602.4 (difference = 8.2)
```

```
Anova(glm_country)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PROBGAM
##         LR Chisq Df Pr(>Chisq)
## country   8.2466  1   0.004083 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# cleaning model_data
vars <- names(model_data)
model_data_complete <- model_data[complete.cases(model_data[, vars]), ]

glmfinal <- glm(PROBGAM ~ ., data = model_data, family = binomial(link = "logit"))
display(glmfinal, detail=T, digits=3)
```

```
## glm(formula = PROBGAM ~ ., family = binomial(link = "logit"),
##     data = model_data)
##                                   coef.est coef.se z value Pr(>|z|)
## (Intercept)                        0.303    0.386   0.784   0.433
## age                               -0.017    0.004  -4.040   0.000
## SexFemale                         -0.677    0.081  -8.335   0.000
## Religsc_groupNo Religion          -0.119    0.119  -1.000   0.317
## Religsc_groupNon-Catholic Crhistian -0.039  0.116  -0.337   0.736
## Religsc_groupOther Religion        0.040    0.222   0.180   0.857
## HHSize_group3-5                   -0.122    0.128  -0.951   0.342
## HHSize_group6+                     0.367    0.286   1.283   0.199
## numcars_group2 cars               -0.217    0.094  -2.314   0.021
## numcars_group3 or more            -0.068    0.151  -0.448   0.654
## eqvinc                             0.000    0.000  -1.502   0.133
## hhdtypb_groupMultiple Adults       0.045    0.123   0.362   0.717
## hhdtypb_groupSenior Household     -0.292    0.217  -1.341   0.180
## hhdtypb_groupSingle Adult          0.017    0.200   0.086   0.931
## eqv5_groupHigh Income              0.019    0.119   0.158   0.875
## eqv5_groupLow Income              -0.157    0.114  -1.375   0.169
## HighQual_groupHigher Education    -0.194    0.092  -2.101   0.036
## HighQual_groupLow/No Quals        -0.022    0.123  -0.181   0.857
## Econact_2Education                -0.043    0.233  -0.183   0.855
## Econact_2Retired                   0.439    0.159   2.756   0.006
## Econact_2Unemployed               -0.002    0.209  -0.009   0.993
## Econact_2Other                    -0.049    0.164  -0.301   0.764
## OwnRnt08Shared Ownership          -0.172    0.116  -1.483   0.138
## OwnRnt08Social Rent                0.271    0.402   0.673   0.501
## OwnRnt08Private Rent               0.125    0.107   1.167   0.243
## OwnRnt08Other                      0.512    0.357   1.433   0.152
## hpnssec5Intermediate               0.237    0.137   1.735   0.083
## hpnssec5Self-Employed              0.043    0.141   0.301   0.764
## hpnssec5Technical/Supervisory      0.202    0.145   1.395   0.163
## hpnssec5Semi-Routine               0.193    0.116   1.664   0.096
## SrcInc7Yes                        -0.419    0.317  -1.320   0.187
## drating                            0.006    0.002   3.431   0.001
## wemwbs                            -0.014    0.005  -2.954   0.003
## cigst1Current smoker               0.304    0.108   2.804   0.005
## cigst1Occasional smoker            0.178    0.164   1.081   0.280
```

```
## cigst1Regular smoker                -0.030    0.099  -0.306   0.760
## countryScotland                     -0.195    0.080  -2.431   0.015
## ---
##   n = 5994, k = 37
##   residual deviance = 4489.2, null deviance = 4721.6 (difference = 232.4)
```

```
# null model
null.glm <- glm(PROBGAM ~ 1, data = model_data_complete, family = binomial(link = "logit"))
anova(null.glm, glmfinal, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: PROBGAM ~ 1
## Model 2: PROBGAM ~ age + Sex + Religsc_group + HHSize_group + numcars_group +
##     eqvinc + hhdtypb_group + eqv5_group + HighQual_group + Econact_2 +
##     OwnRnt08 + hpnssec5 + SrcInc7 + drating + wemwbs + cigst1 +
##     country
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5993     4721.6
## 2      5957     4489.2 36   232.38 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# function from project guide code
ct.op<-function(predicted,observed){ #arguments
#create the data frame
df.op<-data.frame(predicted=predicted,observed=observed)
#create a table
op.tab<-table(df.op)
#use the prop.table function to obtain the rows we need and stack them on top of each other with rbind
op.tab<-rbind(op.tab,c(round(prop.table(op.tab,2)[1,1],2),
                       round((prop.table(op.tab,2)[2,2]),2)))
#name the rows
rownames(op.tab)<-c("pred=0","pred=1","%corr")
#name the columns
colnames(op.tab)<-c("obs=0","obs=1")
#return the table
op.tab
}


# checking for initial model, as there is an error for inconsistent rows, changing the data to only include rows
with all values (which isn't much)

rows1 <- as.numeric(names(glm1$fitted.values))
final_data1 <- data$PROBGAM[rows1]
rowsfinal <- as.numeric(names(glmfinal$fitted.values))
final_datafinal <- data$PROBGAM[rowsfinal]

pred.glm1 <- as.numeric(glm1$fitted.values>0.2)
ct.op(pred.glm1, final_data1)
```

```
##          obs=0 obs=1
## pred=0 5624.00 760.0
## pred=1  949.00 333.0
## %corr     0.86   0.3
```

```
pred.glmfinal <- as.numeric(glmfinal$fitted.values>0.2)
ct.op(pred.glmfinal, final_datafinal)
```

```
##          obs=0  obs=1
## pred=0 4515.00 554.00
## pred=1  676.00 249.00
## %corr     0.87   0.31
```

```
summary(glmfinal)
```

```
##
## Call:
## glm(formula = PROBGAM ~ ., family = binomial(link = "logit"),
##     data = model_data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     3.027e-01  3.860e-01   0.784  0.43301
## age                            -1.688e-02  4.178e-03  -4.040 5.34e-05 ***
## SexFemale                      -6.774e-01  8.126e-02  -8.335  < 2e-16 ***
## Religsc_groupNo Religion       -1.191e-01  1.190e-01  -1.000  0.31719
## Religsc_groupNon-Catholic Crhistian -3.908e-02  1.159e-01  -0.337  0.73594
## Religsc_groupOther Religion     3.997e-02  2.217e-01   0.180  0.85692
## HHSize_group3-5                -1.215e-01  1.278e-01  -0.951  0.34185
## HHSize_group6+                  3.669e-01  2.860e-01   1.283  0.19946
## numcars_group2 cars            -2.172e-01  9.387e-02  -2.314  0.02069 *
## numcars_group3 or more         -6.793e-02  1.515e-01  -0.448  0.65385
## eqvinc                         -2.933e-06  1.953e-06  -1.502  0.13321
## hhdtypb_groupMultiple Adults    4.451e-02  1.229e-01   0.362  0.71724
## hhdtypb_groupSenior Household  -2.915e-01  2.173e-01  -1.341  0.17981
## hhdtypb_groupSingle Adult       1.723e-02  2.002e-01   0.086  0.93141
## eqv5_groupHigh Income           1.876e-02  1.188e-01   0.158  0.87452
## eqv5_groupLow Income           -1.569e-01  1.141e-01  -1.375  0.16922
## HighQual_groupHigher Education -1.943e-01  9.246e-02  -2.101  0.03565 *
## HighQual_groupLow/No Quals     -2.228e-02  1.234e-01  -0.181  0.85667
## Econact_2Education             -4.255e-02  2.326e-01  -0.183  0.85489
## Econact_2Retired                4.392e-01  1.594e-01   2.756  0.00586 **
## Econact_2Unemployed            -1.802e-03  2.086e-01  -0.009  0.99311
## Econact_2Other                 -4.922e-02  1.637e-01  -0.301  0.76375
## OwnRnt08Shared Ownership       -1.715e-01  1.157e-01  -1.483  0.13821
## OwnRnt08Social Rent             2.705e-01  4.017e-01   0.673  0.50068
## OwnRnt08Private Rent            1.245e-01  1.068e-01   1.167  0.24338
## OwnRnt08Other                   5.116e-01  3.570e-01   1.433  0.15182
## hpnssec5Intermediate            2.370e-01  1.366e-01   1.735  0.08266 .
## hpnssec5Self-Employed           4.255e-02  1.414e-01   0.301  0.76353
## hpnssec5Technical/Supervisory   2.019e-01  1.447e-01   1.395  0.16314
## hpnssec5Semi-Routine            1.932e-01  1.162e-01   1.664  0.09618 .
## SrcInc7Yes                     -4.187e-01  3.172e-01  -1.320  0.18686
```

```
## drating                              6.058e-03  1.765e-03   3.431  0.00060 ***
## wemwbs                              -1.428e-02  4.833e-03  -2.954  0.00314 **
## cigst1Current smoker                 3.040e-01  1.084e-01   2.804  0.00505 **
## cigst1Occasional smoker              1.776e-01  1.642e-01   1.081  0.27955
## cigst1Regular smoker                -3.033e-02  9.911e-02  -0.306  0.75962
## countryScotland                     -1.949e-01  8.017e-02  -2.431  0.01504 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4721.6  on 5993  degrees of freedom
## Residual deviance: 4489.2  on 5957  degrees of freedom
##   (5415 observations deleted due to missingness)
## AIC: 4563.2
##
## Number of Fisher Scoring iterations: 5
```

Calculating Average Predictive Comparisons

```
mod_mat <- model.matrix(glmfinal)
betas <- coef(glmfinal)
```

Age (20 to 50) = -0.06636827

```
# checking from 20 to 50
lo.hi <- c(20, 50)

colnames(mod_mat)
```

```
##  [1] "(Intercept)"                          "age"
##  [3] "SexFemale"                            "Religsc_groupNo Religion"
##  [5] "Religsc_groupNon-Catholic Crhistian" "Religsc_groupOther Religion"
##  [7] "HHSize_group3-5"                      "HHSize_group6+"
##  [9] "numcars_group2 cars"                  "numcars_group3 or more"
## [11] "eqvinc"                               "hhdtypb_groupMultiple Adults"
## [13] "hhdtypb_groupSenior Household"        "hhdtypb_groupSingle Adult"
## [15] "eqv5_groupHigh Income"                "eqv5_groupLow Income"
## [17] "HighQual_groupHigher Education"       "HighQual_groupLow/No Quals"
## [19] "Econact_2Education"                   "Econact_2Retired"
## [21] "Econact_2Unemployed"                  "Econact_2Other"
## [23] "OwnRnt08Shared Ownership"             "OwnRnt08Social Rent"
## [25] "OwnRnt08Private Rent"                 "OwnRnt08Other"
## [27] "hpnssec5Intermediate"                 "hpnssec5Self-Employed"
## [29] "hpnssec5Technical/Supervisory"        "hpnssec5Semi-Routine"
## [31] "SrcInc7Yes"                           "drating"
## [33] "wemwbs"                               "cigst1Current smoker"
## [35] "cigst1Occasional smoker"              "cigst1Regular smoker"
## [37] "countryScotland"
```

```
col_age <- which(colnames(mod_mat) == "age")

mm_hi <- mod_mat
mm_hi[, col_age] <- rep(lo.hi[2], nrow(mod_mat))

mm_lo <- mod_mat
mm_lo[, col_age] <- rep(lo.hi[1], nrow(mod_mat))

delta_age <- with(model_data, (invlogit(mm_hi %*% betas) - invlogit(mm_lo %*% betas)))

mean_delta_age <- mean(delta_age)
print(mean_delta_age)
```

```
## [1] -0.06636827
```

Sex (Male to Female) = -0.07600407

```
lo.hi <- c(0, 1)

col_sex <- which(colnames(mod_mat) == "SexFemale")

mm_hi <- mod_mat
mm_hi[, col_sex] <- rep(lo.hi[2], nrow(mod_mat))

mm_lo <- mod_mat
mm_lo[, col_sex] <- rep(lo.hi[1], nrow(mod_mat))

delta_sex <- with(model_data, (invlogit(mm_hi %*% betas) - invlogit(mm_lo %*% betas)))
mean_delta_sex <- mean(delta_sex)
print(mean_delta_sex)
```

```
## [1] -0.07600407
```

Houshold Size group (1-2 to 3-5) = -0.008180366

```
                  (1–2 to 6+) = 0.0525688
```

```
col_hhsize_3_5 <- which(colnames(mod_mat) == "HHSize_group3–5")
col_hhsize_6p  <- which(colnames(mod_mat) == "HHSize_group6+")

mm_baseline <- mod_mat
mm_hhsize_3_5 <- mod_mat
mm_hhsize_6p <- mod_mat

mm_hhsize_3_5[, col_hhsize_3_5] <- 1
mm_hhsize_3_5[, col_hhsize_6p] <- 0

mm_hhsize_6p[, col_hhsize_3_5] <- 0
mm_hhsize_6p[, col_hhsize_6p] <- 1

pred_baseline <- invlogit(mm_baseline %*% betas)
pred_hhsize_3_5 <- invlogit(mm_hhsize_3_5 %*% betas)
pred_hhsize_6p <- invlogit(mm_hhsize_6p %*% betas)

delta_3_5 <- pred_hhsize_3_5 - pred_baseline
delta_6p <- pred_hhsize_6p - pred_baseline

mean_delta_3_5 <- mean(delta_3_5)
mean_delta_6p <- mean(delta_6p)

print(mean_delta_3_5)
```

```
## [1] -0.008180366
```

```
print(mean_delta_6p)
```

```
## [1] 0.0525688
```

eqvinc (20000 to 80000) = -0.01907012

```r
lo_hi_eqvinc <- c(20000, 80000)

col_eqvinc <- which(colnames(mod_mat) == "eqvinc")

mm_hi <- mod_mat
mm_hi[, col_eqvinc] <- rep(lo_hi_eqvinc[2], nrow(mod_mat))

mm_lo <- mod_mat
mm_lo[, col_eqvinc] <- rep(lo_hi_eqvinc[1], nrow(mod_mat))

pred_hi <- invlogit(mm_hi %*% betas)
pred_lo <- invlogit(mm_lo %*% betas)

delta_eqvinc <- pred_hi - pred_lo
mean_delta_eqvinc <- mean(delta_eqvinc)

print(mean_delta_eqvinc)
```

```
## [1] -0.01907012
```

HighQual_group (Low/No Qualification to Higher Education) = -0.01887061

```
col_highqual_higher <- which(colnames(mod_mat) == "HighQual_groupHigher Education")
col_highqual_lowno <- which(colnames(mod_mat) == "HighQual_groupLow/No Quals")

mm_higher <- mod_mat
mm_lowno <- mod_mat

mm_higher[, col_highqual_higher] <- 1
mm_higher[, col_highqual_lowno] <- 0

mm_lowno[, col_highqual_higher] <- 0
mm_lowno[, col_highqual_lowno] <- 1

pred_higher <- invlogit(mm_higher %*% betas)
pred_lowno <- invlogit(mm_lowno %*% betas)

delta_highqual <- pred_higher - pred_lowno
mean_delta_highqual <- mean(delta_highqual)

print(mean_delta_highqual)
```

```
## [1] -0.01887061
```

Units of alcohol a week (0 to 50) = 0.03551692

```
col_drating <- which(colnames(mod_mat) == "drating")

lo <- 0
hi <- 50

mm_lo <- mod_mat
mm_lo[, col_drating] <- lo

mm_hi <- mod_mat
mm_hi[, col_drating] <- hi

pred_lo <- invlogit(mm_lo %*% betas)
pred_hi <- invlogit(mm_hi %*% betas)

delta_drating <- pred_hi - pred_lo
mean_delta_drating <- mean(delta_drating)

print(mean_delta_drating)
```

```
## [1] 0.03551692
```

Mental Health Wellbeing Score (30 to 70) = -0.06481117

```
col_wemwbs <- which(colnames(mod_mat) == "wemwbs")

lo <- 30
hi <- 70

mm_lo <- mod_mat
mm_lo[, col_wemwbs] <- lo

mm_hi <- mod_mat
mm_hi[, col_wemwbs] <- hi

pred_lo <- invlogit(mm_lo %*% betas)
pred_hi <- invlogit(mm_hi %*% betas)

delta_wemwbs <- pred_hi - pred_lo

mean_delta_wemwbs <- mean(delta_wemwbs)

print(mean_delta_wemwbs)
```

```
## [1] -0.06481117
```

Smoking Status (Never smoked to Current smoker) = 0.03615152

```
col_regular <- which(colnames(mod_mat) == "cigst1Current smoker")

mm_lo <- mod_mat
mm_lo[, col_regular] <- 0

mm_hi <- mod_mat
mm_hi[, col_regular] <- 1

pred_lo <- invlogit(mm_lo %*% betas)
pred_hi <- invlogit(mm_hi %*% betas)

delta_cigst1 <- pred_hi - pred_lo

mean_delta_cigst1 <- mean(delta_cigst1)
print(mean_delta_cigst1)
```

```
## [1] 0.03615152
```

Country (England to Scotland) = -0.02162515

```r
col_country <- which(colnames(mod_mat) == "countryScotland")

mm_lo <- mod_mat
mm_lo[, col_country] <- 0

mm_hi <- mod_mat
mm_hi[, col_country] <- 1

pred_lo <- invlogit(mm_lo %*% betas)
pred_hi <- invlogit(mm_hi %*% betas)

delta_country <- pred_hi - pred_lo

mean_delta_country <- mean(delta_country)

print(mean_delta_country)
```

```
## [1] -0.02162515
```