NSDUH_Drug_Analysis

Katelin Bauer

2023-12-20

Contents

Question 1

Utilize multiple regression methods to determine if there is a relationship between the age of first cocaine use during adolescence and the following predictors: demographic variables, perceived risk of cocaine use, availability of cocaine, danger seeking, age of first alcohol use, and age of first cigarette use.

Import NSDUH data

```r
library(tidyverse)
library(data.table)
library(ggplot2)
library(plyr)
library(car)
library(leaps)
library(boot)
library(glmnet)

NSDUH_2020 <- read_csv("NSDUH_2020.csv")
NSDUH_2021 <- read_csv("NSDUH_2021.csv")

df_20.21 <- rbind.fill(NSDUH_2020, NSDUH_2021)
```

Data cleaning

```r
df_20.21 |>
  select(DIFGETCOC,RSKYFQDGR,RSKYFQTES,CATAGE,IRALCAGE,IRCIGAGE,YODPREV,COCEVER,YEPRTDNG,COCAGE,RSKCOCM

df1 <- subset(df1, df1$COCAGE < 18)
df1 <- subset(df1, !(CATAGE %in% c(3, 4))) # Exclude respondents who are 26 or older.
df1 <- subset(df1, !(COCEVER %in% c(991))) # Exclude respondents who never used cocaine.
df1 <- subset(df1, !(DIFGETCOC %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKYFQDGR %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKYFQTES %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(IRALCAGE %in% c(991))) # Exclude respondents who never used alcohol.
df1 <- subset(df1, !(RSKCOCWK %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKCOCMON %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(IRCIGAGE %in% c(991))) # Exclude respondents who never used cigarettes.
df1 <- subset(df1, !(YEPRTDNG %in% c(97,99)))
df1 <- subset(df1, !(YODPREV %in% c(97,99)))

head(df1)
```

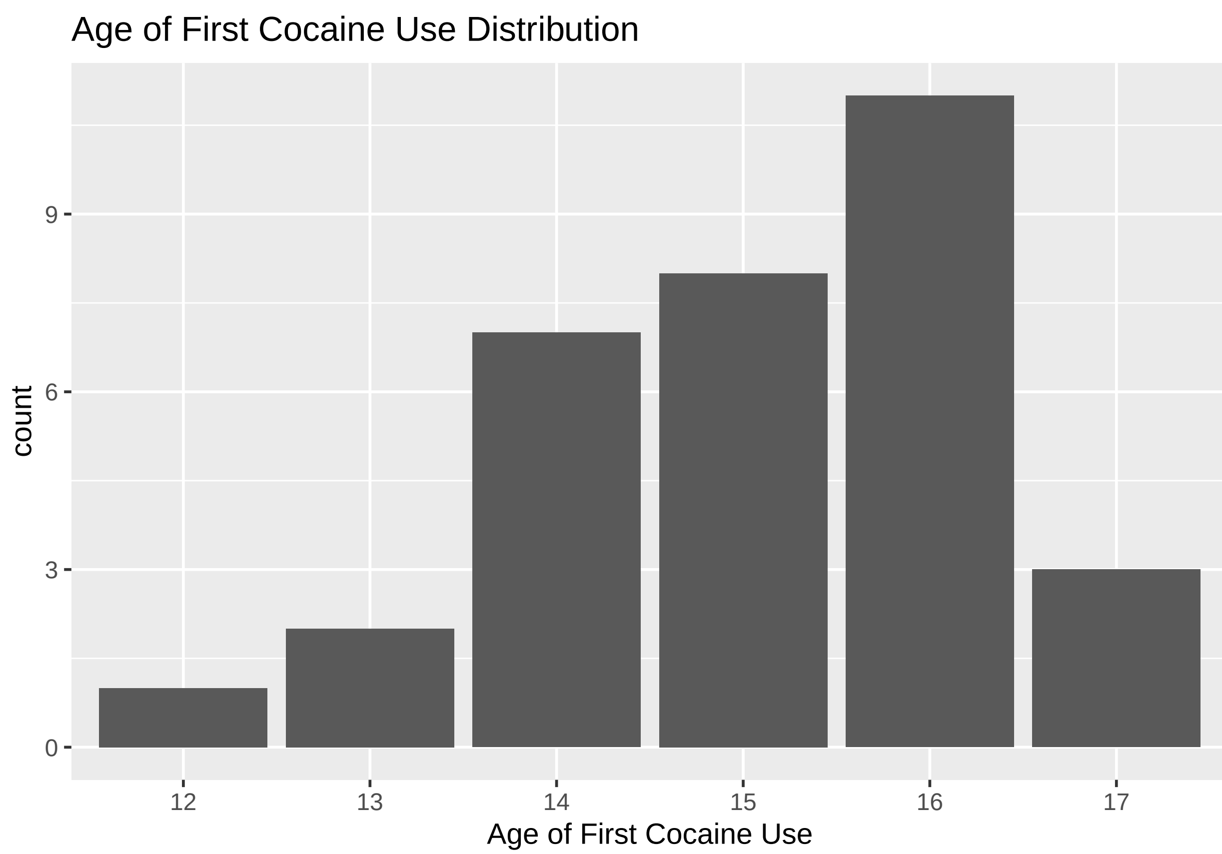
```
##      DIFGETCOC RSKYFQDGR RSKYFQTES CATAGE IRALCAGE IRCIGAGE YODPREV COCEVER
## 431          4         3         3      1        8       13       1       1
## 1730         4         2         2      1       13       15       2       1
## 2530         5         4         3      1       15       12       1       1
## 2825         2         3         3      1       15       15       1       1
## 6965         5         1         3      1       15       15       2       1
## 7862         3         3         3      1        9        9       2       1
##      YEPRTDNG COCAGE RSKCOCMON RSKCOCWK NEWRACE2 IRSEX
## 431         1     15         1        1        7     1
## 1730        2     16         3        3        7     2
## 2530        1     16         4        4        1     2
## 2825        1     16         1        2        1     2
## 6965        1     15         4        4        7     1
## 7862        1     14         2        2        1     2
```

Convert categorical variables into factors

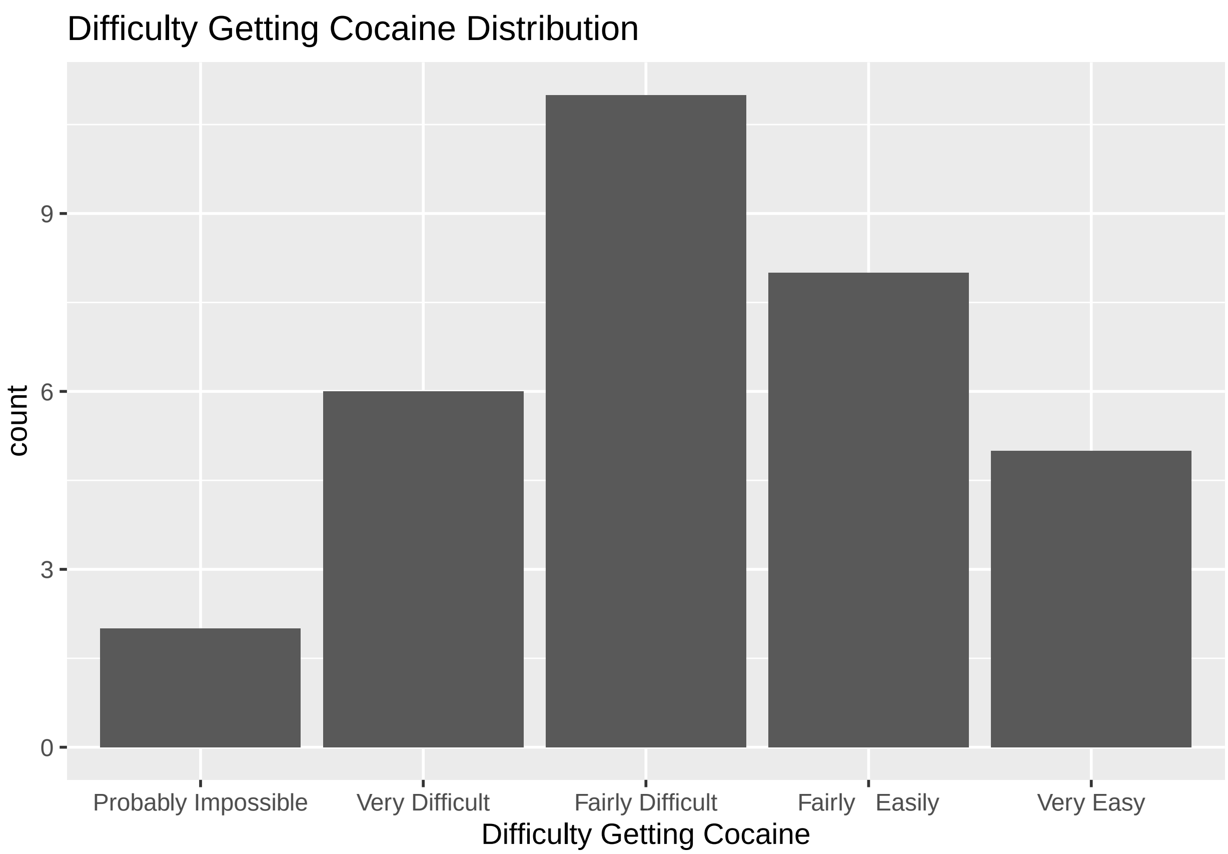
```r
df1$COCEVER <- as.factor(df1$COCEVER)
df1$DIFGETCOC<- as.factor(df1$DIFGETCOC)
df1$RSKCOCMON <- as.factor(df1$RSKCOCMON)
df1$RSKCOCWK <- as.factor(df1$RSKCOCWK)
df1$RSKYFQDGR <- as.factor(df1$RSKYFQDGR)
df1$RSKYFQTES <- as.factor(df1$RSKYFQTES)
df1$YEPRTDNG <- as.factor(df1$YEPRTDNG)
df1$YODPREV <- as.factor(df1$YODPREV)
df1$IRSEX <- factor(df1$IRSEX, labels = c("Male", "Female"))
df1$NEWRACE2 <- factor(df1$NEWRACE2)
```

Plots/ exploratory data analysis

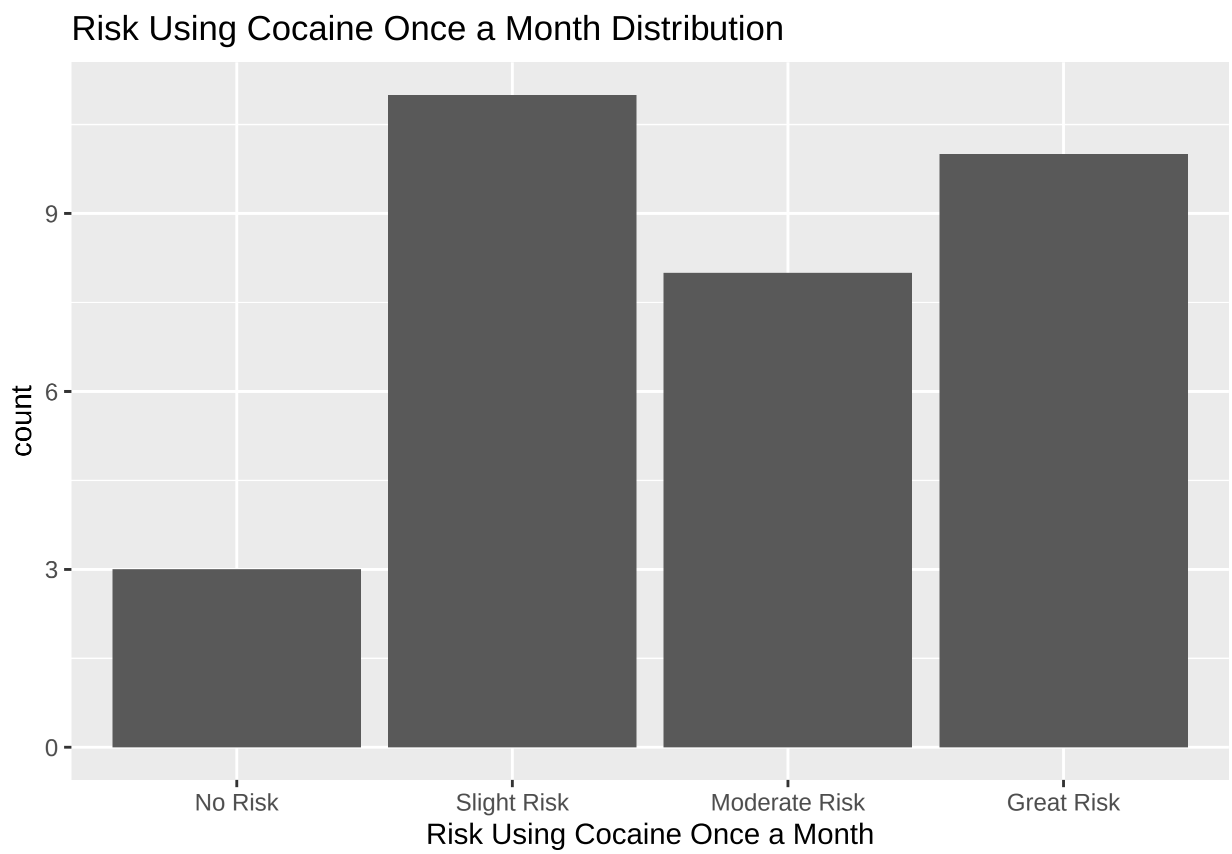
```r
ggplot(df1, aes(x = as.factor(COCAGE))) +
  geom_bar() +
  xlab("Age of First Cocaine Use") +
  ggtitle("Age of First Cocaine Use Distribution")
```



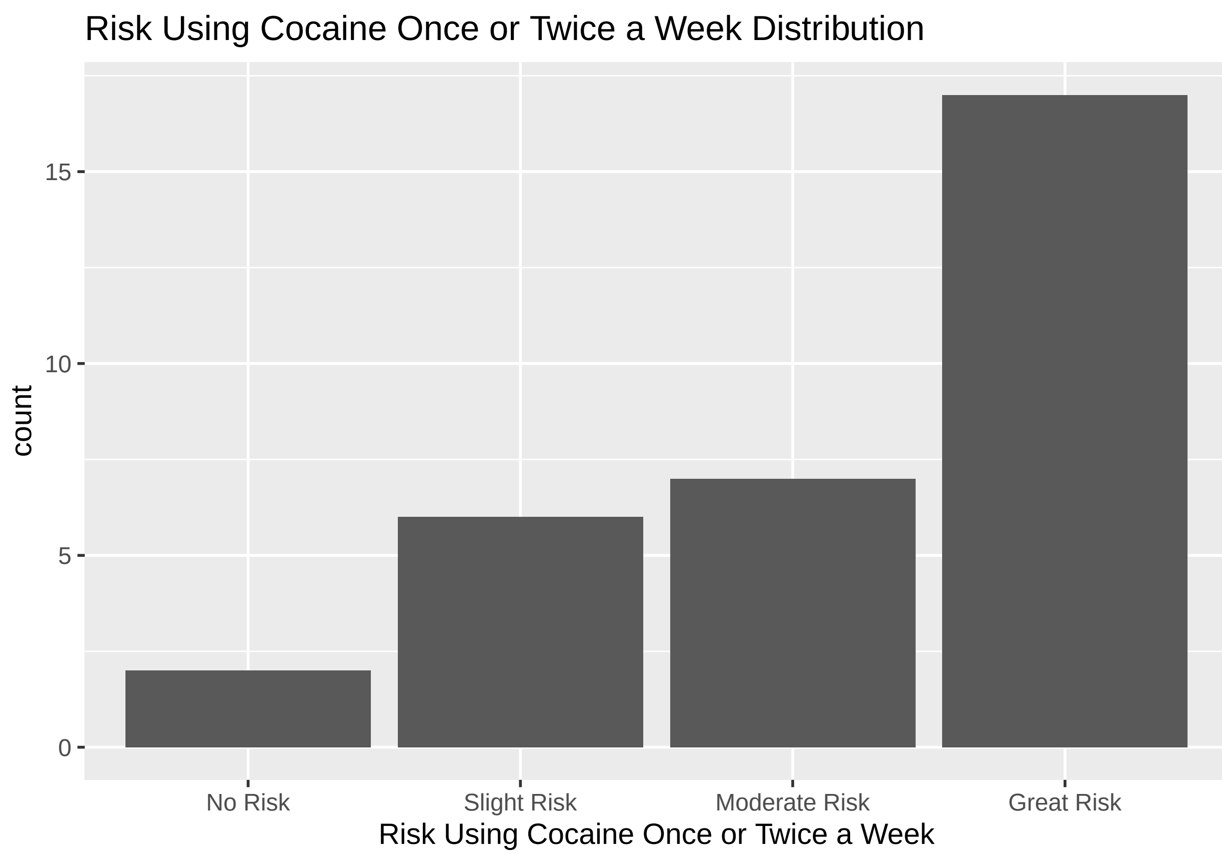

```r
ggplot(df1, aes(x = DIFGETCOC)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "Probably Impossible", "2" = "Very Difficult", "3" = "Fairly Difficul
  xlab("Difficulty Getting Cocaine") +
  ggtitle("Difficulty Getting Cocaine Distribution")
```



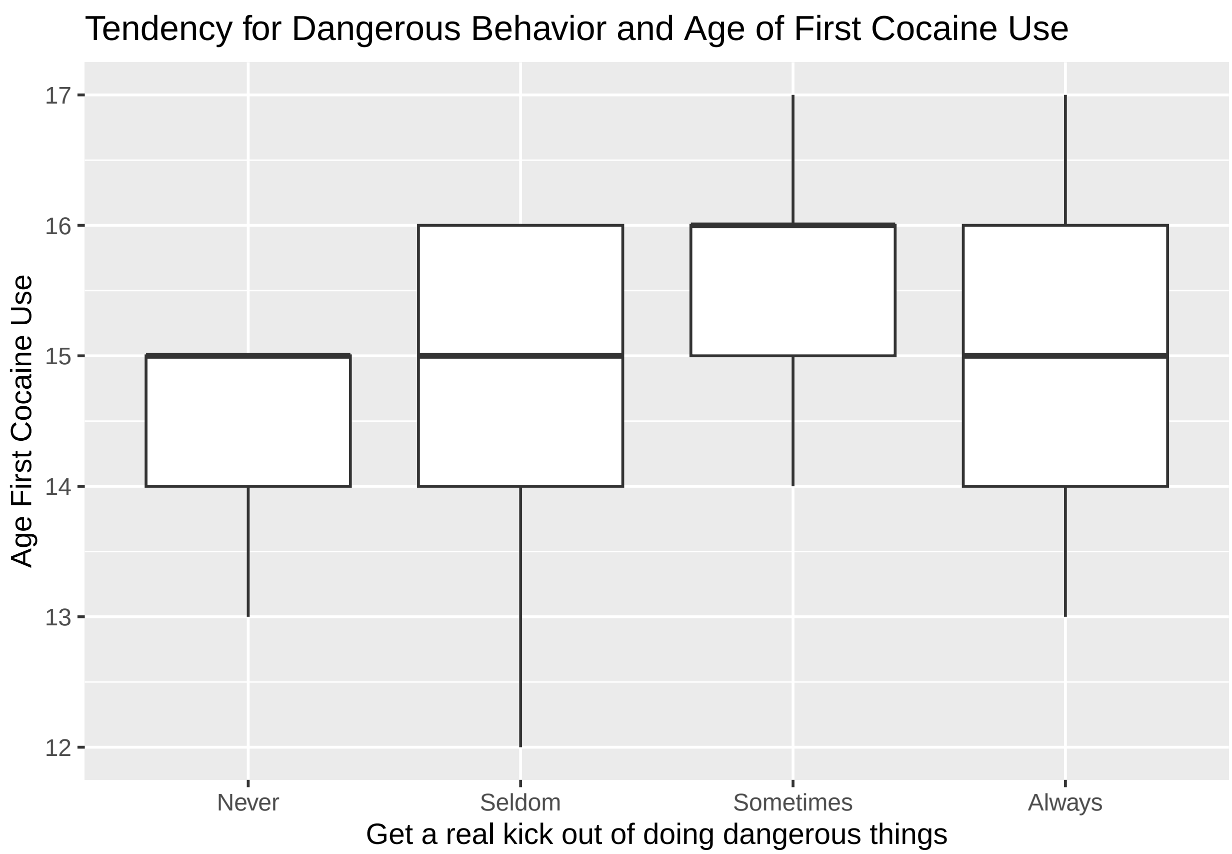

```r
ggplot(df1, aes(x = RSKCOCMON)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "No Risk", "2" = "Slight Risk", "3" = "Moderate Risk","4" = "Great Ri
  xlab("Risk Using Cocaine Once a Month") +
  ggtitle("Risk Using Cocaine Once a Month Distribution")
```



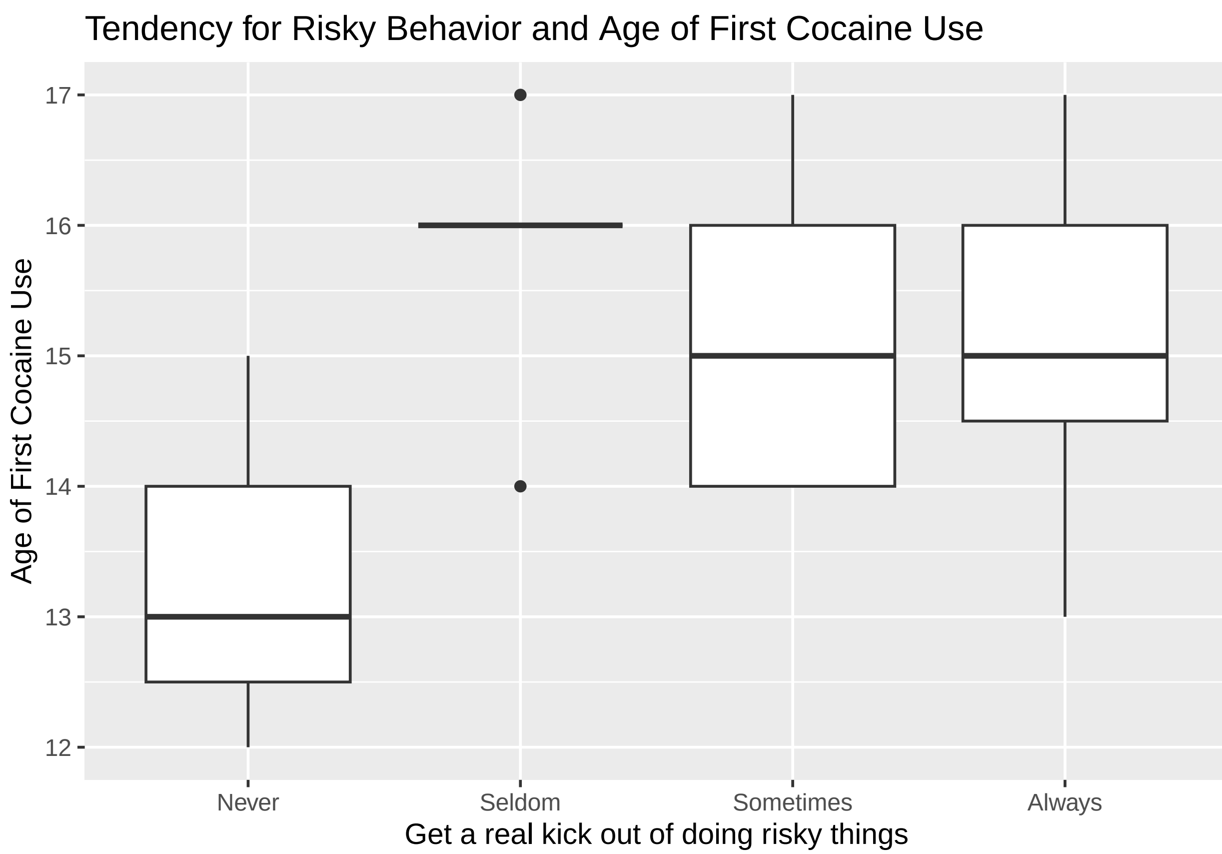

```r
ggplot(df1, aes(x = RSKCOCWK)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "No Risk", "2" = "Slight Risk", "3" = "Moderate Risk","4" = "Great Ris
  xlab("Risk Using Cocaine Once or Twice a Week") +
  ggtitle("Risk Using Cocaine Once or Twice a Week Distribution")
```



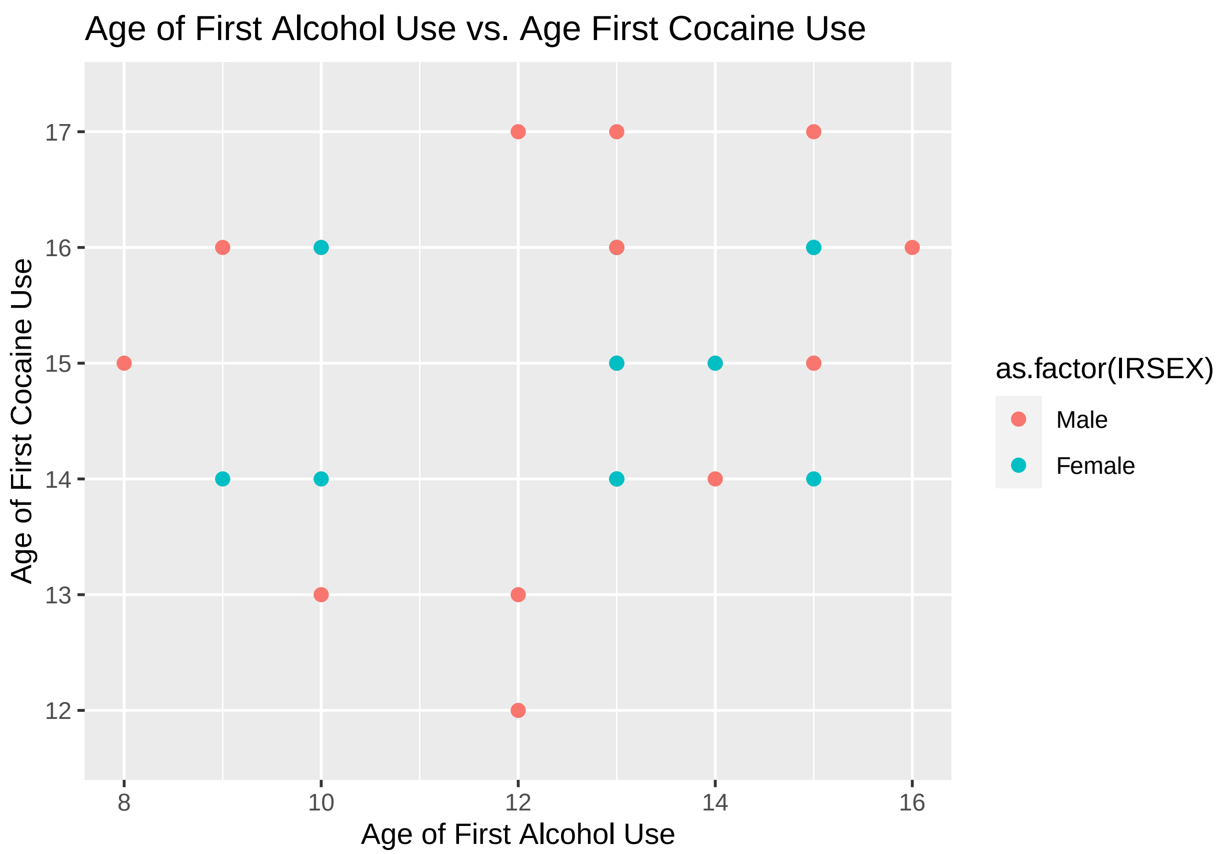

```
ggplot(df1, aes(x = RSKYFQDGR, y = COCAGE)) +
  geom_boxplot() +
  xlab("Get a real kick out of doing dangerous things") +
  ylab("Age First Cocaine Use") +
  scale_x_discrete(labels=c("1" = "Never", "2" = "Seldom", "3" = "Sometimes","4" = "Always")) +
  ggtitle("Tendency for Dangerous Behavior and Age of First Cocaine Use")
```



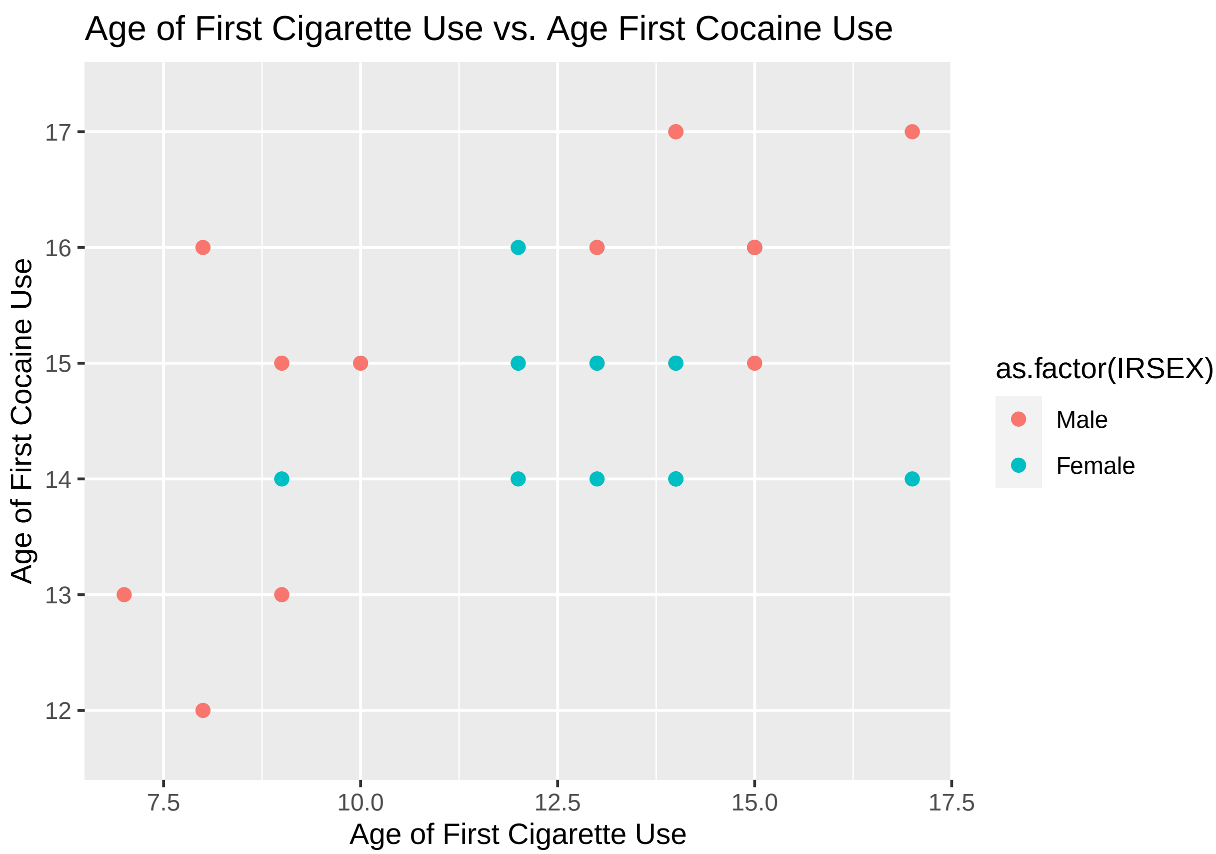

```
ggplot(df1, aes(x = RSKYFQTES, y = COCAGE)) +
  geom_boxplot() +
  xlab("Get a real kick out of doing risky things") +
  ylab("Age of First Cocaine Use") +
  scale_x_discrete(labels=c("1" = "Never", "2" = "Seldom", "3" = "Sometimes","4" = "Always")) +
  ggtitle("Tendency for Risky Behavior and Age of First Cocaine Use")
```




```r
ggplot(df1, aes(x = IRALCAGE, y = as.factor(COCAGE), color = as.factor(IRSEX))) +
  geom_point(size = 2) +
  ggtitle("Age of First Alcohol Use vs. Age First Cocaine Use") +
  xlab("Age of First Alcohol Use") +
  ylab("Age of First Cocaine Use")
```




```
ggplot(df1, aes(x = IRCIGAGE, y = as.factor(COCAGE), color = as.factor(IRSEX))) +
  geom_point(size = 2) +
  ggtitle("Age of First Cigarette Use vs. Age First Cocaine Use") +
  xlab("Age of First Cigarette Use") +
  ylab("Age of First Cocaine Use")
```



Model selection

Fitting a full multiple regression model

```r
# Adjusted R-squared:  0.2723
# p-value: 0.2833
reg_co_full <- lm(data = df1, COCAGE ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOC
summary(reg_co_full)
```

```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE +
##     IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSEX, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46551 -0.32305  0.00611  0.36226  0.90023
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.50493    1.98408   4.791  0.00137 **
## DIFGETCOC2   1.48984    1.83068   0.814  0.43930
## DIFGETCOC3   3.01646    1.92827   1.564  0.15637
```

```
## DIFGETCOC4   2.17965    1.95586   1.114  0.29746
## DIFGETCOC5   2.39416    2.11061   1.134  0.28949
## RSKYFQDGR2   0.88864    1.25423   0.709  0.49875
## RSKYFQDGR3   1.51677    1.17159   1.295  0.23156
## RSKYFQDGR4   1.91260    1.46416   1.306  0.22776
## RSKYFQTES2   0.03756    1.85704   0.020  0.98436
## RSKYFQTES3  -1.04514    1.70078  -0.615  0.55595
## RSKYFQTES4  -1.32619    1.98897  -0.667  0.52368
## IRALCAGE     0.09010    0.16388   0.550  0.59749
## IRCIGAGE     0.16736    0.15598   1.073  0.31459
## RSKCOCMON2  -1.96843    1.69911  -1.159  0.28007
## RSKCOCMON3  -1.13448    2.16176  -0.525  0.61395
## RSKCOCMON4   0.31032    2.13747   0.145  0.88816
## RSKCOCWK2    1.67421    2.23996   0.747  0.47620
## RSKCOCWK3    0.89497    2.50553   0.357  0.73018
## RSKCOCWK4    0.07795    2.77135   0.028  0.97825
## NEWRACE23    1.39942    1.08902   1.285  0.23473
## NEWRACE25    2.11820    1.75309   1.208  0.26144
## NEWRACE26   -0.18293    0.87329  -0.209  0.83931
## NEWRACE27    0.41532    0.87927   0.472  0.64930
## IRSEXFemale -1.00247    0.66808  -1.501  0.17187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.047 on 8 degrees of freedom
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.2723
## F-statistic: 1.504 on 23 and 8 DF,  p-value: 0.2833
```

Fit reduced multiple regression models

```r
# Adjusted R-squared: 0.1917
# p-value: 0.02982
reg_co_1 <- lm(data = df1, COCAGE ~ RSKYFQTES)
summary(reg_co_1)
```

```
##
## Call:
## lm(formula = COCAGE ~ RSKYFQTES, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.143 -1.125 -0.125  0.875  1.875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.3333     0.6372  20.925  < 2e-16 ***
## RSKYFQTES2    2.5000     0.7804   3.203  0.00338 **
## RSKYFQTES3    1.7917     0.6944   2.580  0.01541 *
## RSKYFQTES4    1.8095     0.7616   2.376  0.02459 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.104 on 28 degrees of freedom
## Multiple R-squared:  0.2699, Adjusted R-squared:  0.1917
## F-statistic: 3.451 on 3 and 28 DF,  p-value: 0.02982
```

```r
# Adjusted R-squared: 0.2406
# p-value: 0.02102
reg_co_2 <- lm(data = df1, COCAGE ~ DIFGETCOC)
summary(reg_co_2)
```

```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0000 -0.6591 -0.0625  0.8187  1.8750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5000     0.7565  16.524 1.21e-15 ***
## DIFGETCOC2    2.5000     0.8735   2.862 0.008035 **
## DIFGETCOC3    3.0455     0.8224   3.703 0.000965 ***
## DIFGETCOC4    2.6250     0.8458   3.104 0.004449 **
## DIFGETCOC5    2.7000     0.8951   3.016 0.005518 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 27 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2406
## F-statistic: 3.455 on 4 and 27 DF,  p-value: 0.02102
```

```r
# Adjusted R-squared: 0.2395
# p-value: 0.05224
reg_co_3 <- lm(data = df1, COCAGE ~ DIFGETCOC + RSKYFQTES)
summary(reg_co_3)
```

```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC + RSKYFQTES, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92895 -0.52089 -0.06803  0.72185  1.71001
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5000     0.7570  16.512 1.32e-14 ***
## DIFGETCOC2    2.9520     1.4832   1.990   0.0581 .
## DIFGETCOC3    3.3130     1.4317   2.314   0.0295 *
## DIFGETCOC4    2.5000     1.3112   1.907   0.0686 .
## DIFGETCOC5    3.1591     1.4991   2.107   0.0457 *
## RSKYFQTES2    0.5623     1.1721   0.480   0.6357
## RSKYFQTES3   -0.4164     1.1932  -0.349   0.7301
```

```
## RSKYFQTES4   -0.5230     1.2749  -0.410   0.6853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.071 on 24 degrees of freedom
## Multiple R-squared:  0.4112, Adjusted R-squared:  0.2395
## F-statistic: 2.395 on 7 and 24 DF,  p-value: 0.05224
```

Step forward variable selection

```r
reg_co_1_null <- lm(COCAGE ~1, data = df1)
reg_co_1_stepout <- step(reg_co_1_null,
                      scope = list(lwer = reg_co_1_null, upper = reg_co_full),
                      method = "forward")
```

```
## Start:  AIC=14.11
## COCAGE ~ 1
##
##              Df Sum of Sq    RSS     AIC
## + IRCIGAGE   1   13.0994 33.619  5.5797
## + DIFGETCOC  4   15.8165 30.902  8.8830
## + RSKYFQTES  3   12.6116 34.107 10.0407
## <none>                   46.719 14.1091
## + IRALCAGE   1    2.0938 44.625 14.6419
## + IRSEX      1    0.2482 46.471 15.9387
## + RSKCOCMON  3    5.0733 41.645 16.4306
## + RSKYFQDGR  3    4.5039 42.215 16.8652
## + RSKCOCWK   3    4.1599 42.559 17.1248
## + NEWRACE2   4    2.1207 44.598 20.6225
##
## Step:  AIC=5.58
## COCAGE ~ IRCIGAGE
##
##              Df Sum of Sq    RSS     AIC
## + DIFGETCOC  4    7.6807 25.939  5.2799
## + RSKYFQTES  3    5.7698 27.849  5.5545
## <none>                   33.619  5.5797
## + IRSEX      1    1.8292 31.790  5.7894
## + IRALCAGE   1    0.1763 33.443  7.4114
## + RSKYFQDGR  3    3.0709 30.548  8.5144
## + RSKCOCMON  3    2.4772 31.142  9.1304
## + RSKCOCWK   3    1.9492 31.670  9.6684
## + NEWRACE2   4    1.7110 31.908 11.9082
## - IRCIGAGE   1   13.0994 46.719 14.1091
##
## Step:  AIC=5.28
## COCAGE ~ IRCIGAGE + DIFGETCOC
##
##              Df Sum of Sq    RSS     AIC
## + IRSEX      1    3.4036 22.535  2.7788
## <none>                   25.939  5.2799
## - DIFGETCOC  4    7.6807 33.619  5.5797
```

```
## + RSKCOCMON  3    3.2912 22.647  6.9379
## + IRALCAGE   1    0.0000 25.939  7.2799
## + RSKYFQTES  3    2.2981 23.641  8.3113
## - IRCIGAGE   1    4.9636 30.902  8.8830
## + RSKCOCWK   3    1.8353 24.103  8.9317
## + RSKYFQDGR  3    0.5309 25.408 10.6182
## + NEWRACE2   4    1.9090 24.030 10.8337
##
## Step:  AIC=2.78
## COCAGE ~ IRCIGAGE + DIFGETCOC + IRSEX
##
##              Df Sum of Sq    RSS    AIC
## <none>                    22.535 2.7788
## + IRALCAGE   1    0.0405 22.495 4.7212
## + RSKCOCMON  3    2.3771 20.158 5.2117
## - IRSEX      1    3.4036 25.939 5.2799
## - DIFGETCOC  4    9.2550 31.790 5.7894
## + NEWRACE2   4    2.7533 19.782 6.6089
## + RSKYFQTES  3    1.4485 21.087 6.6529
## + RSKCOCWK   3    0.7554 21.780 7.6878
## - IRCIGAGE   1    5.7118 28.247 8.0079
## + RSKYFQDGR  3    0.3412 22.194 8.2906
```

```r
summary(reg_co_1_stepout)
```

```
##
## Call:
## lm(formula = COCAGE ~ IRCIGAGE + DIFGETCOC + IRSEX, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82239 -0.77183  0.07097  0.61314  1.27618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.86768    0.93338  11.643 1.37e-11 ***
## IRCIGAGE     0.19204    0.07629   2.517  0.01861 *
## DIFGETCOC2   1.82327    0.83800   2.176  0.03924 *
## DIFGETCOC3   2.43060    0.83267   2.919  0.00733 **
## DIFGETCOC4   2.10019    0.89799   2.339  0.02764 *
## DIFGETCOC5   2.51060    0.85786   2.927  0.00720 **
## IRSEXFemale -0.74054    0.38110  -1.943  0.06334 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9494 on 25 degrees of freedom
## Multiple R-squared:  0.5176, Adjusted R-squared:  0.4019
## F-statistic: 4.471 on 6 and 25 DF,  p-value: 0.003314
```

```r
# Chosen model: lm(formula = COCAGE ~ IRCIGAGE + DIFGETCOC + IRSEX, data = df1)

# Adjusted R-squared: 0.4019
# p-value: 0.003314
```

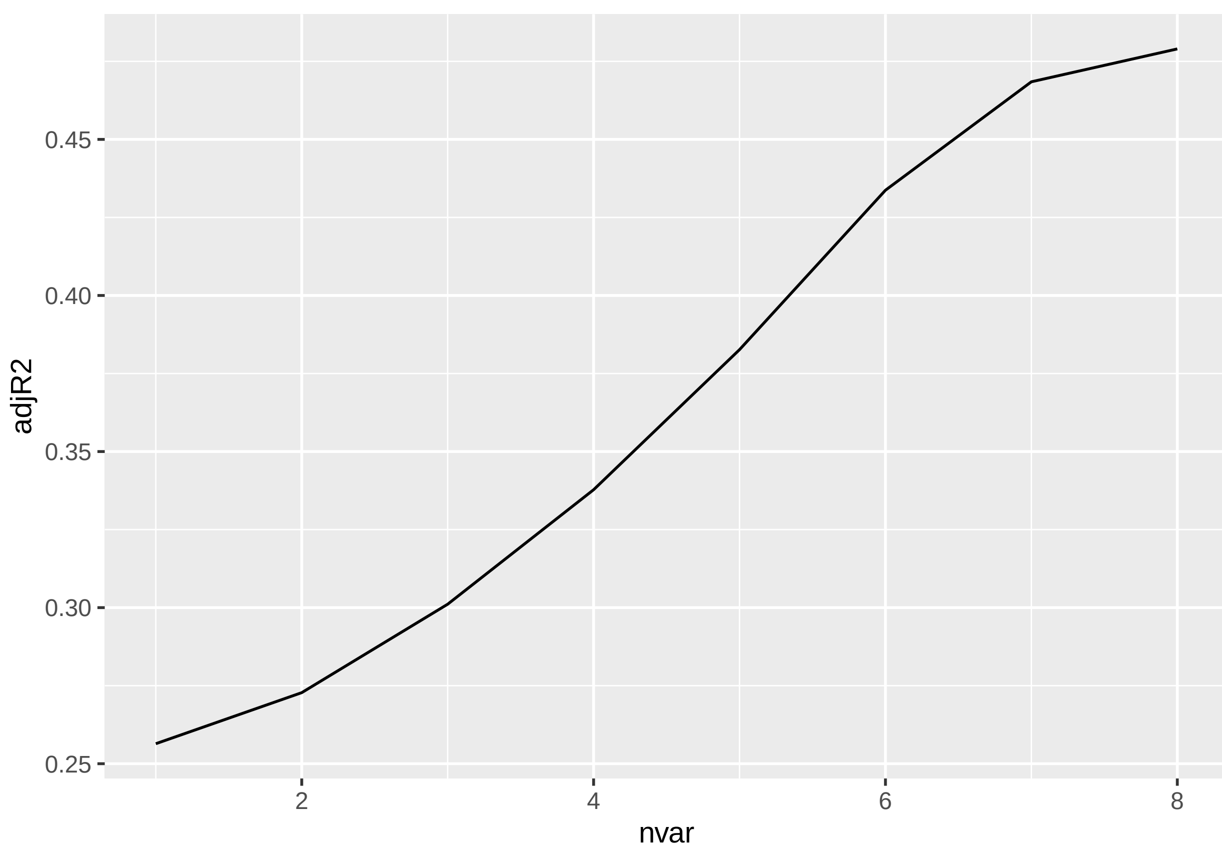
Exhaustive search variable selection

```
reg_co_1_ex <- regsubsets(data = df1, COCAGE ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE +
reg_co_1_summary <- summary(reg_co_1_ex)
reg_co_1_summary
```

```
## Subset selection object
## Call: regsubsets.formula(data = df1, COCAGE ~ DIFGETCOC + RSKYFQDGR +
##     RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK +
##     NEWRACE2 + IRSEX)
## 23 Variables  (and intercept)
##              Forced in Forced out
## DIFGETCOC2       FALSE      FALSE
## DIFGETCOC3       FALSE      FALSE
## DIFGETCOC4       FALSE      FALSE
## DIFGETCOC5       FALSE      FALSE
## RSKYFQDGR2       FALSE      FALSE
## RSKYFQDGR3       FALSE      FALSE
## RSKYFQDGR4       FALSE      FALSE
## RSKYFQTES2       FALSE      FALSE
## RSKYFQTES3       FALSE      FALSE
## RSKYFQTES4       FALSE      FALSE
## IRALCAGE         FALSE      FALSE
## IRCIGAGE         FALSE      FALSE
## RSKCOCMON2       FALSE      FALSE
## RSKCOCMON3       FALSE      FALSE
## RSKCOCMON4       FALSE      FALSE
## RSKCOCWK2        FALSE      FALSE
## RSKCOCWK3        FALSE      FALSE
## RSKCOCWK4        FALSE      FALSE
## NEWRACE23        FALSE      FALSE
## NEWRACE25        FALSE      FALSE
## NEWRACE26        FALSE      FALSE
## NEWRACE27        FALSE      FALSE
## IRSEXFemale      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          DIFGETCOC2 DIFGETCOC3 DIFGETCOC4 DIFGETCOC5 RSKYFQDGR2 RSKYFQDGR3
## 1  ( 1 ) " "        " "        " "        " "        " "        " "
## 2  ( 1 ) " "        " "        " "        " "        " "        "*"
## 3  ( 1 ) " "        "*"        " "        " "        " "        " "
## 4  ( 1 ) " "        "*"        " "        " "        " "        " "
## 5  ( 1 ) " "        " "        " "        " "        " "        "*"
## 6  ( 1 ) " "        "*"        " "        " "        " "        " "
## 7  ( 1 ) " "        "*"        " "        " "        " "        " "
## 8  ( 1 ) "*"        "*"        "*"        "*"        " "        " "
##          RSKYFQDGR4 RSKYFQTES2 RSKYFQTES3 RSKYFQTES4 IRALCAGE IRCIGAGE
## 1  ( 1 ) " "        " "        " "        " "        " "      "*"
## 2  ( 1 ) " "        " "        " "        " "        " "      "*"
## 3  ( 1 ) " "        " "        " "        " "        " "      "*"
## 4  ( 1 ) " "        " "        " "        " "        " "      "*"
## 5  ( 1 ) " "        " "        " "        "*"        " "      "*"
## 6  ( 1 ) " "        "*"        "*"        "*"        " "      " "
```

```
## 7  ( 1 ) " "        "*"        "*"        "*"         " "       "*"
## 8  ( 1 ) " "        "*"        " "        " "         "*"       " "
##          RSKCOCMON2 RSKCOCMON3 RSKCOCMON4 RSKCOCWK2 RSKCOCWK3 RSKCOCWK4
## 1  ( 1 ) " "        " "        " "        " "       " "       " "
## 2  ( 1 ) " "        " "        " "        " "       " "       " "
## 3  ( 1 ) "*"        " "        " "        " "       " "       " "
## 4  ( 1 ) "*"        " "        " "        " "       " "       " "
## 5  ( 1 ) "*"        " "        " "        " "       " "       " "
## 6  ( 1 ) "*"        "*"        " "        " "       " "       " "
## 7  ( 1 ) "*"        "*"        " "        " "       " "       " "
## 8  ( 1 ) "*"        "*"        " "        " "       " "       " "
##          NEWRACE23 NEWRACE25 NEWRACE26 NEWRACE27 IRSEXFemale
## 1  ( 1 ) " "       " "       " "       " "       " "
## 2  ( 1 ) " "       " "       " "       " "       " "
## 3  ( 1 ) " "       " "       " "       " "       " "
## 4  ( 1 ) " "       "*"       " "       " "       " "
## 5  ( 1 ) " "       "*"       " "       " "       " "
## 6  ( 1 ) " "       " "       " "       " "       " "
## 7  ( 1 ) " "       " "       " "       " "       " "
## 8  ( 1 ) " "       " "       " "       " "       " "
```

```r
df_exh <- data.frame(adjR2 = reg_co_1_summary$adjr2, nvar = 1:length(reg_co_1_summary$adjr2))
ggplot(df_exh, (aes(nvar, adjR2))) +
  geom_line()
```



```r
which.max(reg_co_1_summary$adjr2)
```

```
## [1] 8
```

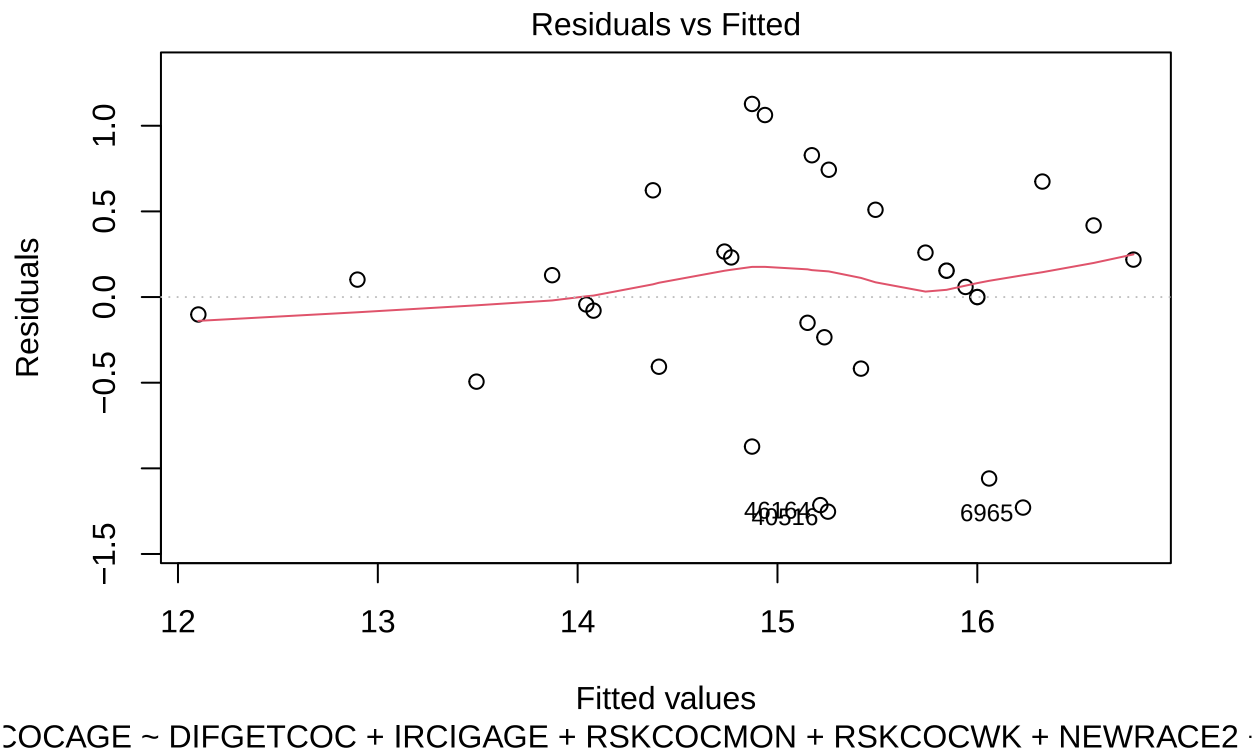
```r
# Exhaustive search recommends 8 variables, however, since there are factors, each level counts as 1 va

reg_ex_model <- lm(data = df1, COCAGE ~ DIFGETCOC + RSKYFQTES + IRALCAGE + RSKCOCMON)
summary(reg_ex_model)
```

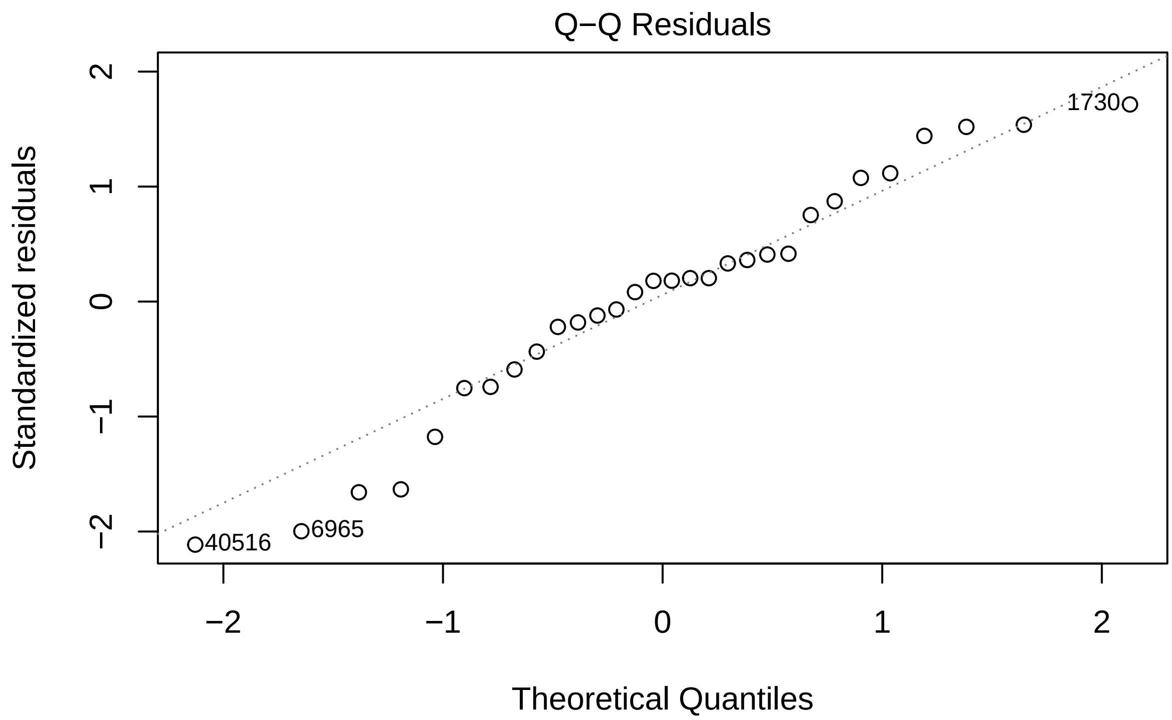
```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC + RSKYFQTES + IRALCAGE + RSKCOCMON,
##     data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49567 -0.31874 -0.02383  0.36230  1.92808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.15927    1.28355   9.473  7.8e-09 ***
## DIFGETCOC2   1.78908    1.38353   1.293   0.2107
## DIFGETCOC3   2.81240    1.27395   2.208   0.0391 *
## DIFGETCOC4   1.88125    1.16201   1.619   0.1211
## DIFGETCOC5   2.10079    1.45577   1.443   0.1645
## RSKYFQTES2   1.69070    1.16423   1.452   0.1620
## RSKYFQTES3   0.23714    1.14709   0.207   0.8383
## RSKYFQTES4   0.21419    1.24442   0.172   0.8651
## IRALCAGE     0.12431    0.08739   1.422   0.1703
## RSKCOCMON2  -1.64548    0.63206  -2.603   0.0170 *
## RSKCOCMON3  -1.85164    0.71618  -2.585   0.0177 *
## RSKCOCMON4  -0.65661    0.73699  -0.891   0.3836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9252 on 20 degrees of freedom
## Multiple R-squared:  0.6335, Adjusted R-squared:  0.432
## F-statistic: 3.143 on 11 and 20 DF,  p-value: 0.01265
```

```r
# Adjusted R-squared: 0.432
# p-value: 0.01265
```

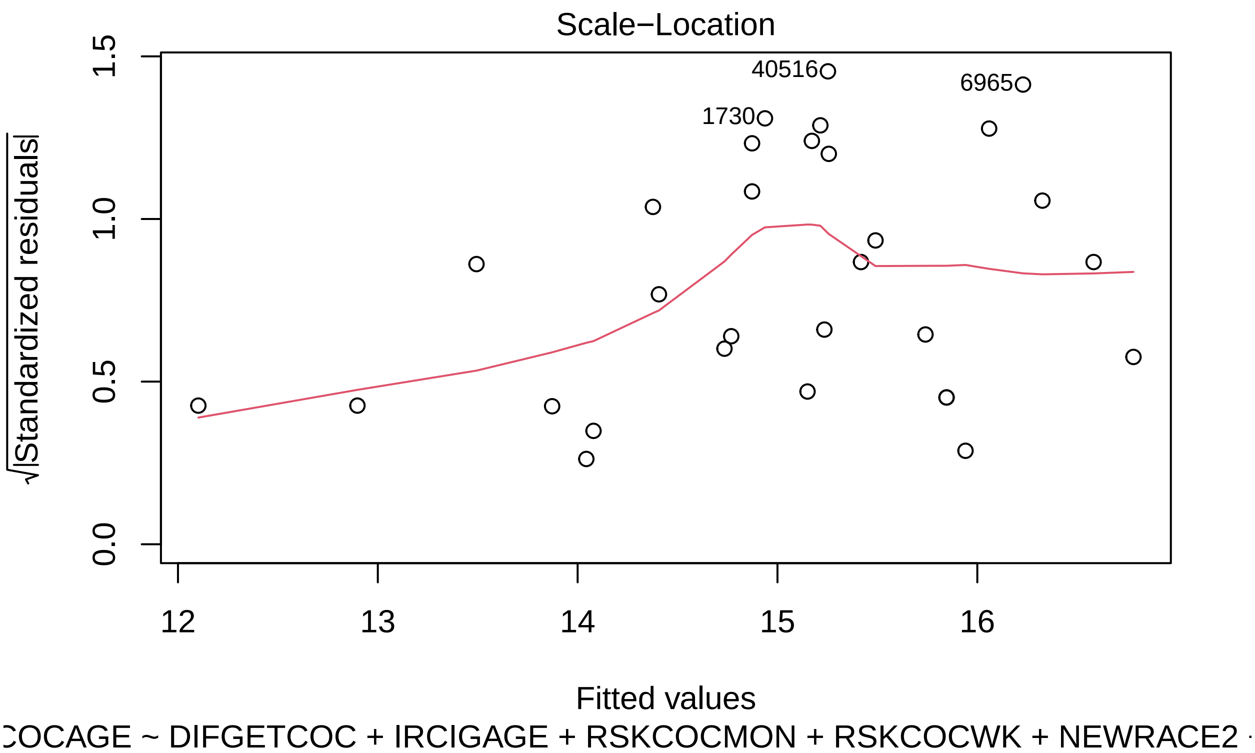
Multiple regression model with the highest adjusted r-squared, thus far:

```r
model_best_pt1 <- lm(data = df1, COCAGE ~ DIFGETCOC + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSE
plot(model_best_pt1)
```

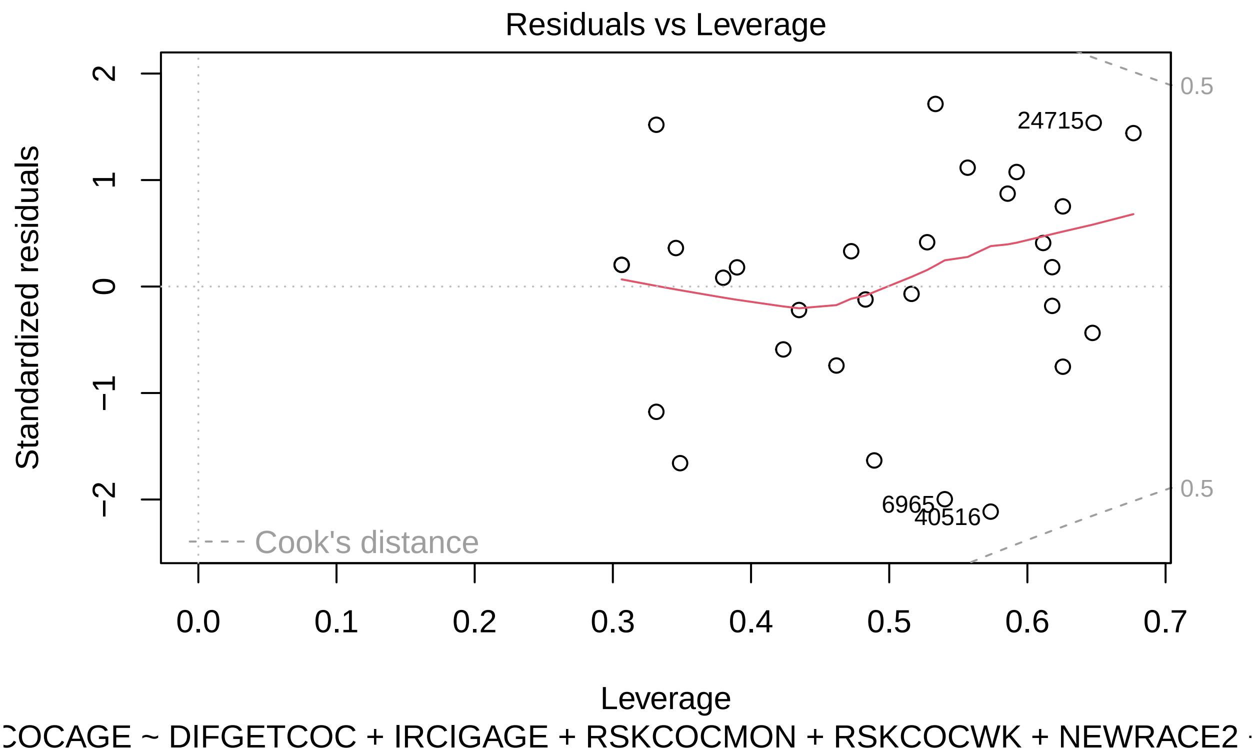

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.5
12
13
14
15
16
Fitted values
46164
40516
6965
:OCAGE ~ DIFGETCOC + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 ·



COCAGE ~ DIFGETCOC + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 ·



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
12
13
14
15
16
Fitted values
40516
6965
1730
COCAGE ~ DIFGETCOC + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 ·




```r
summary(model_best_pt1)
```

```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC + IRCIGAGE + RSKCOCMON + RSKCOCWK +
##     NEWRACE2 + IRSEX, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25267 -0.27778  0.08044  0.30349  1.12718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.26064    1.38646   7.401 2.22e-06 ***
## DIFGETCOC2   1.61424    0.81404   1.983  0.06599 .
## DIFGETCOC3   3.20939    0.99762   3.217  0.00576 **
## DIFGETCOC4   2.70502    1.18308   2.286  0.03719 *
## DIFGETCOC5   2.43451    0.92083   2.644  0.01842 *
## IRCIGAGE     0.22229    0.09781   2.273  0.03818 *
## RSKCOCMON2  -1.85454    1.31924  -1.406  0.18017
## RSKCOCMON3  -0.74528    1.62910  -0.457  0.65388
## RSKCOCMON4   0.31040    1.60564   0.193  0.84930
## RSKCOCWK2    1.91746    1.56737   1.223  0.24007
## RSKCOCWK3    0.94690    1.80278   0.525  0.60709
## RSKCOCWK4    0.32648    1.89318   0.172  0.86539
```

```
## NEWRACE23    0.72044    0.64502   1.117  0.28160
## NEWRACE25    2.17890    1.14034   1.911  0.07534 .
## NEWRACE26   -0.10414    0.70132  -0.148  0.88393
## NEWRACE27   -0.43752    0.53032  -0.825  0.42230
## IRSEXFemale -1.12669    0.47113  -2.391  0.03033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9074 on 15 degrees of freedom
## Multiple R-squared:  0.7357, Adjusted R-squared:  0.4537
## F-statistic: 2.609 on 16 and 15 DF,  p-value: 0.03522
```

```
# Adjusted R-squared: 0.4537
# p-value: 0.03522
```

There are outliers in the data which could contribute to the model only being able to explain approximately 45% of the variance in the response variable according to the adjusted r-squared value. The Q-Q plot shows a fairly normal linear distribution of the data, however, around the max/min values the regression line is displaying some curvature. It is possible that there is multicollinearity among predictors.

VIF test for multicollinearity

```
vif(reg_co_full) # full multiple regression model
```

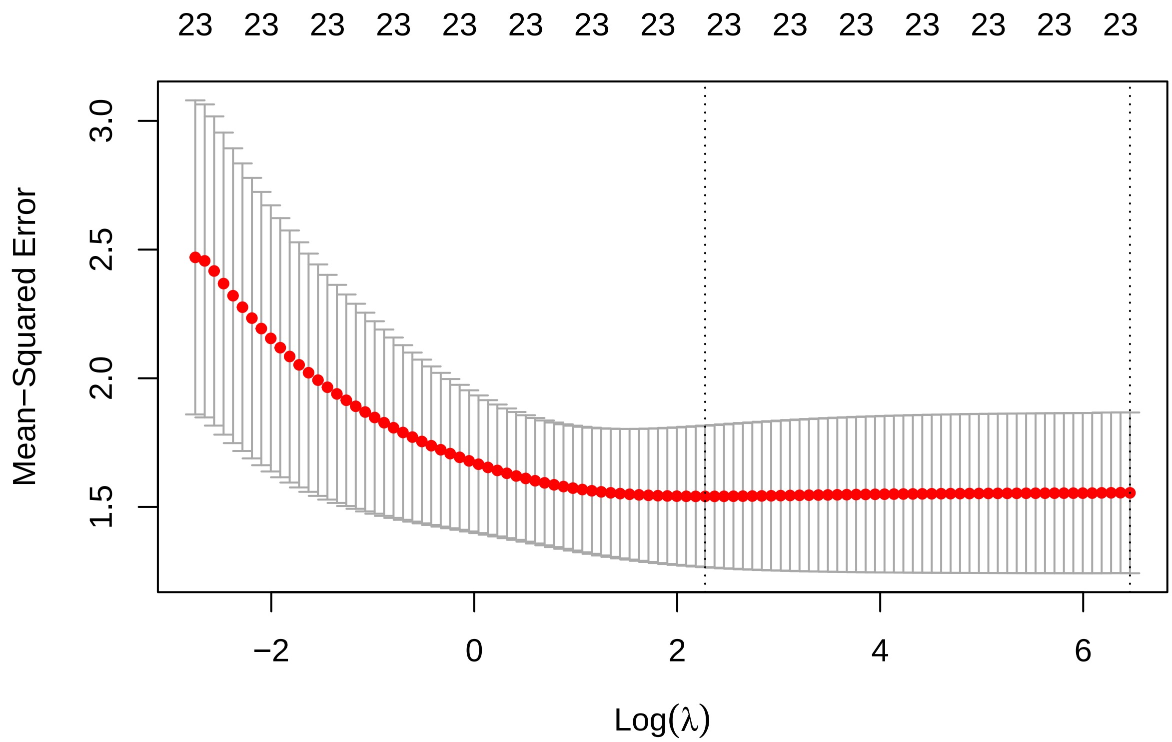
```
##                  GVIF Df GVIF^(1/(2*Df))
## DIFGETCOC  94.212852  4        1.765077
## RSKYFQDGR  49.082416  3        1.913467
## RSKYFQTES  71.583190  3        2.037676
## IRALCAGE    3.281790  1        1.811571
## IRCIGAGE    4.808258  1        2.192774
## RSKCOCMON 187.159736  3        2.391674
## RSKCOCWK  137.863677  3        2.272872
## NEWRACE2   28.594388  4        1.520670
## IRSEX       3.243370  1        1.800936
```

The variance inflation test indicates a high probability of the presence of multicollinearity among the following predictors:

DIFGETCOC, RSKYFQDGR, RSKYFQTES, RSKCOCMON, RSKCOCWK, and NEWRACE2.

Ridge regression model for the shrinkage of predictor coefficent values

```
X <- model.matrix(data = df1, COCAGE ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOC
set.seed(123)
cv.ridge = cv.glmnet(X, df1$COCAGE, alpha = 0)

plot(cv.ridge)
```




```r
cv.ridge # cross-validated MSE: 1.555
```

```
##
## Call:  cv.glmnet(x = X, y = df1$COCAGE, alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure     SE Nonzero
## min    9.7    46   1.541 0.2749      23
## 1se  639.8     1   1.555 0.3123      23
```

Adding interaction terms to the model

```r
reg_co_5 <- lm(data = df1, COCAGE ~ IRCIGAGE*DIFGETCOC + IRSEX)
summary(reg_co_5)
```

```
##
## Call:
## lm(formula = COCAGE ~ IRCIGAGE * DIFGETCOC + IRSEX, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6706 -0.3222  0.0000  0.6072  1.5411
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0000    11.0166   0.363   0.7202
## IRCIGAGE               1.0000     1.2938   0.773   0.4482
## DIFGETCOC2             7.0092    11.1295   0.630   0.5356
## DIFGETCOC3             9.9825    11.1327   0.897   0.3801
## DIFGETCOC4             5.3807    11.6387   0.462   0.6486
## DIFGETCOC5            13.2199    11.2360   1.177   0.2525
## IRSEXFemale           -0.6735     0.3778  -1.783   0.0891 .
## IRCIGAGE:DIFGETCOC2  -0.6761     1.2996  -0.520   0.6083
## IRCIGAGE:DIFGETCOC3  -0.8611     1.2992  -0.663   0.5147
## IRCIGAGE:DIFGETCOC4  -0.5576     1.3197  -0.422   0.6770
## IRCIGAGE:DIFGETCOC5  -1.1369     1.3068  -0.870   0.3941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9149 on 21 degrees of freedom
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.4446
## F-statistic: 3.482 on 10 and 21 DF,  p-value: 0.00769
```

```r
# Adjusted R-squared:  0.4446
# p-value: 0.00769

reg_co_6 <- lm(data = df1, COCAGE ~ DIFGETCOC*IRCIGAGE + NEWRACE2 + IRSEX + RSKCOCMON + RSKCOCWK)
summary(reg_co_6)
```

```
##
## Call:
## lm(formula = COCAGE ~ DIFGETCOC * IRCIGAGE + NEWRACE2 + IRSEX +
##     RSKCOCMON + RSKCOCWK, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2478 -0.2135  0.0000  0.2737  1.0647
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          11.905933  11.497359   1.036   0.3226
## DIFGETCOC2            0.468186  11.202066   0.042   0.9674
## DIFGETCOC3            2.304238  11.301114   0.204   0.8422
## DIFGETCOC4           -7.886776  12.455839  -0.633   0.5396
## DIFGETCOC5            6.544602  11.243498   0.582   0.5723
## IRCIGAGE              0.127250   1.341927   0.095   0.9262
## NEWRACE23             0.479285   0.639380   0.750   0.4692
## NEWRACE25             3.124347   1.117285   2.796   0.0174 *
## NEWRACE26             0.009009   0.637668   0.014   0.9890
## NEWRACE27             0.318757   0.567679   0.562   0.5857
## IRSEXFemale          -0.574050   0.486892  -1.179   0.2633
## RSKCOCMON2           -1.512975   1.271836  -1.190   0.2592
## RSKCOCMON3           -1.089306   1.517365  -0.718   0.4878
## RSKCOCMON4           -0.026129   1.466706  -0.018   0.9861
## RSKCOCWK2             0.589041   1.546323   0.381   0.7105
## RSKCOCWK3             0.035582   1.726412   0.021   0.9839
```

```
## RSKCOCWK4           -0.025054   1.738935  -0.014   0.9888
## DIFGETCOC2:IRCIGAGE  0.113476   1.314495   0.086   0.9328
## DIFGETCOC3:IRCIGAGE  0.054463   1.330172   0.041   0.9681
## DIFGETCOC4:IRCIGAGE  0.711822   1.374656   0.518   0.6148
## DIFGETCOC5:IRCIGAGE -0.368079   1.312840  -0.280   0.7844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8069 on 11 degrees of freedom
## Multiple R-squared:  0.8467, Adjusted R-squared:  0.5679
## F-statistic: 3.037 on 20 and 11 DF,  p-value: 0.03108
```

```
# Adjusted R-squared: 0.5679
# p-value: 0.03108
```

Cross-validation

```
# Full model:
glm_co2 <- glm(data = df1, COCAGE ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON
glm_co2_cv <- cv.glm(data = df1, glm_co2)
glm_co2_cv$delta # Prediction MSE = 3.696113 3.613343
```

```
## [1] 3.696113 3.613343
```

```
# Model with the lowest cross-validated prediction MSE:
glm_co3 <- glm(data = df1, COCAGE ~ DIFGETCOC + IRCIGAGE)
glm_co_cv3 <- cv.glm(data = df1, glm_co3)
glm_co_cv3$delta # Prediction MSE = 1.246017 1.238274
```

```
## [1] 1.246017 1.238274
```

General takeaway:

The model with the lowest cross-validated prediction MSE only included two predictor variables. However, this model has quite a low adjusted r-squared value. Therefore, I would recommend further evaluation of the data and the predictor variables with high GVIF values before recommending a model.

Question 2

Utilize classification methods to determine whether a respondent used cocaine for the first time before 18 years old (yes/no) can be effectively classified based on demographic variables, perceived risk of cocaine use, availability of cocaine, danger seeking, age of first alcohol use, and age of first cigarette use.

Data cleaning

```r
df_20.21 |>
  select(DIFGETCOC,FUCOC18,RSKYFQDGR,RSKYFQTES,IRALCAGE,IRCIGAGE,COCEVER,COCAGE,RSKCOCMON,RSKCOCWK,NEWR

df1 <- subset(df1, !(COCAGE %in% c(991, 985, 994, 997, 998)))
df1 <- subset(df1, !(COCEVER %in% c(991))) # Exclude respondents who never used cocaine.
df1 <- subset(df1, !(DIFGETCOC %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKYFQDGR %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKYFQTES %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(IRALCAGE %in% c(991))) # Exclude respondents who never used alcohol.
df1 <- subset(df1, !(RSKCOCWK %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(RSKCOCMON %in% c(85, 94, 97, 98)))
df1 <- subset(df1, !(IRCIGAGE %in% c(991))) # Exclude respondents who never used cigarettes.

head(df1)
```

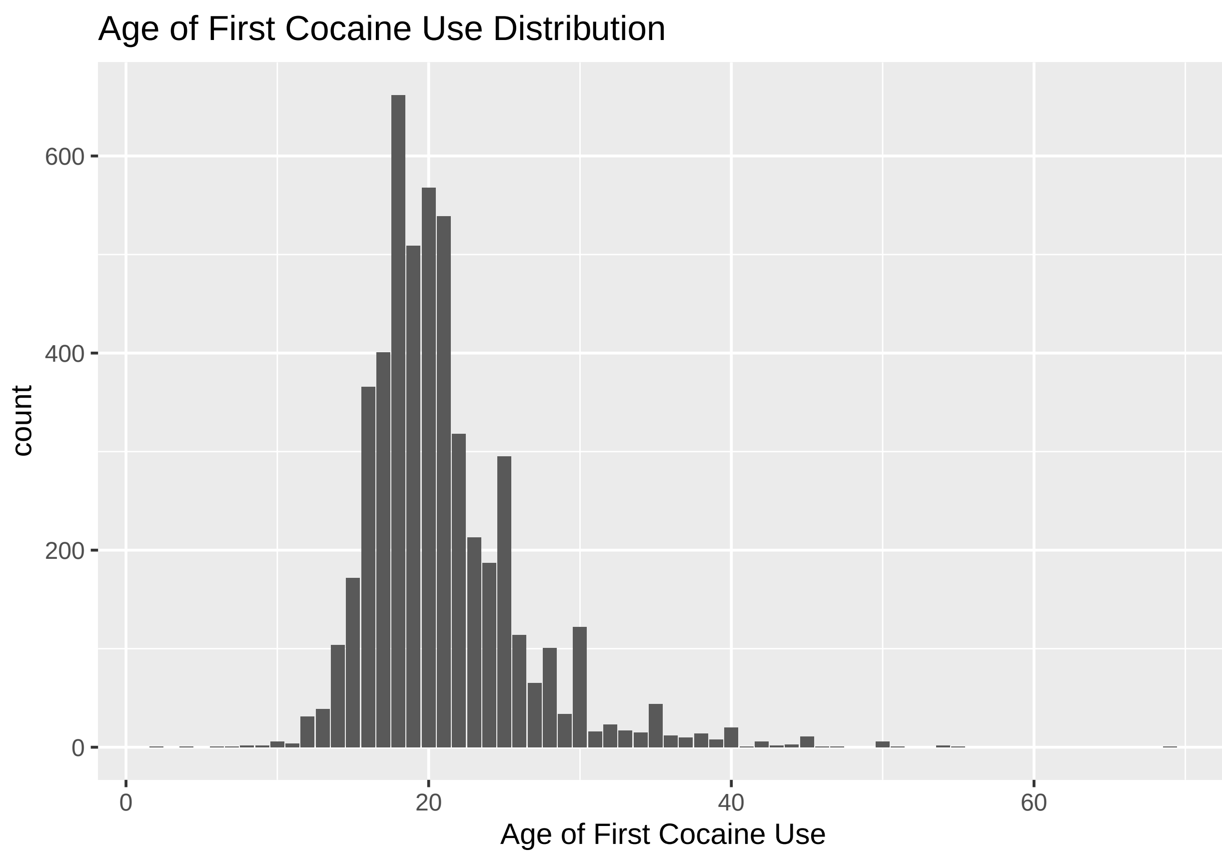
```
##    DIFGETCOC FUCOC18 RSKYFQDGR RSKYFQTES IRALCAGE IRCIGAGE COCEVER COCAGE
## 2          4       2         3         3       16        8       1     25
## 16         3       2         1         1       17       15       1     25
## 20         3       2         4         3       16       29       1     18
## 28         4       2         1         1       14       14       1     19
## 29         2       2         4         4       13       15       1     18
## 34         5       2         1         1        8       15       1     21
##    RSKCOCMON RSKCOCWK NEWRACE2 IRSEX
## 2          4        4        1     1
## 16         4        4        1     2
## 20         3        4        1     1
## 28         4        4        1     2
## 29         2        4        1     1
## 34         4        4        7     2
```

Convert categorical variables into factors

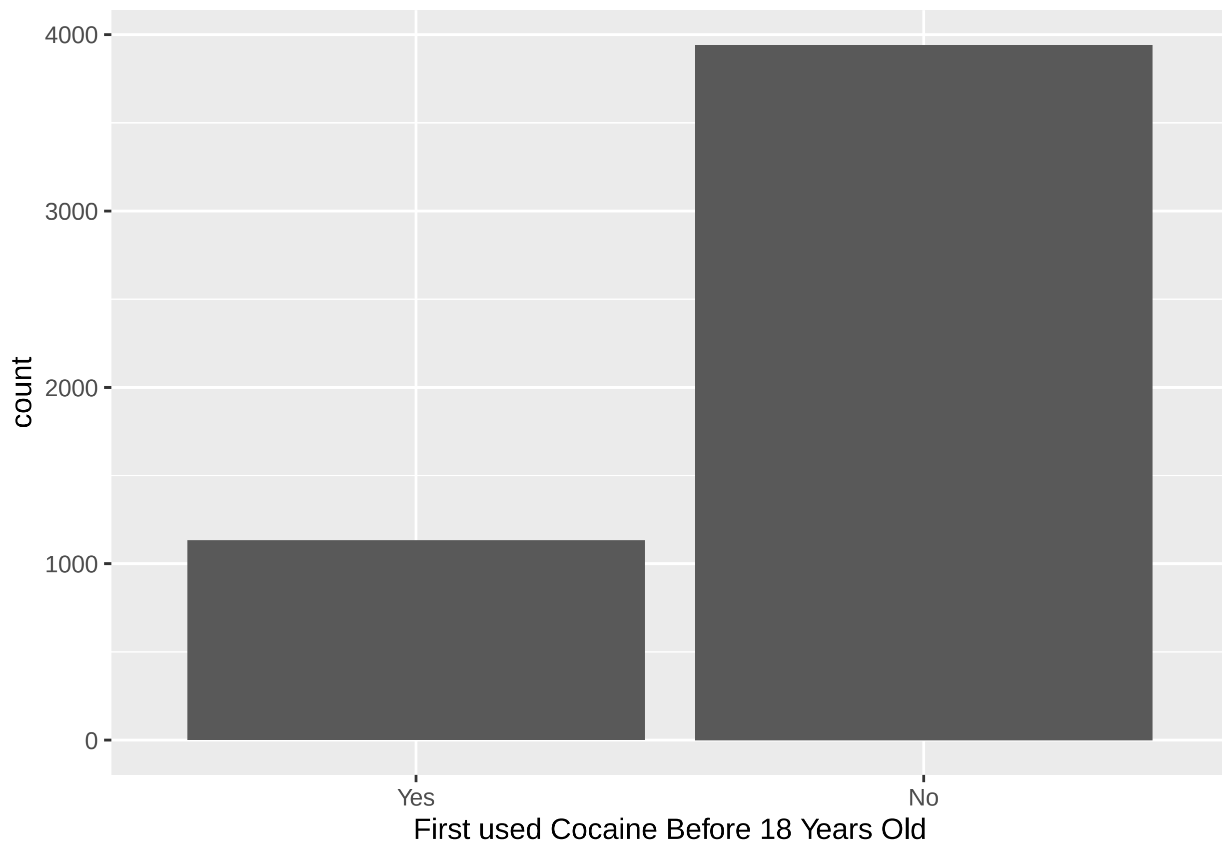
```r
df1$COCEVER <- as.factor(df1$COCEVER)
df1$DIFGETCOC<- as.factor(df1$DIFGETCOC)
df1$RSKCOCMON <- as.factor(df1$RSKCOCMON)
df1$RSKCOCWK <- as.factor(df1$RSKCOCWK)
df1$RSKYFQDGR <- as.factor(df1$RSKYFQDGR)
df1$RSKYFQTES <- as.factor(df1$RSKYFQTES)
df1$IRSEX <- factor(df1$IRSEX, labels = c("Male", "Female"))
df1$NEWRACE2 <- factor(df1$NEWRACE2)
df1$FUCOC18 <- factor(df1$FUCOC18)
```

Plots/ exploratory data analysis

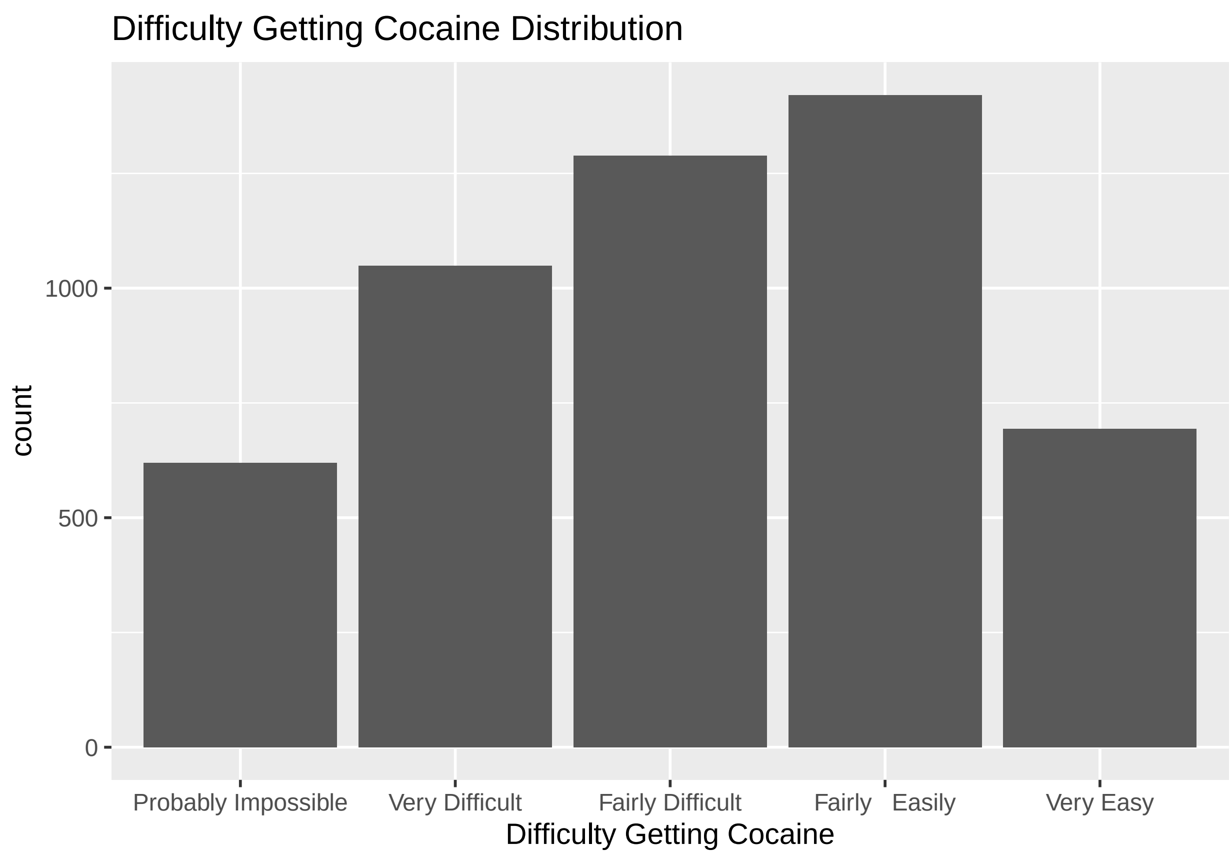
```r
ggplot(df1, aes(x = COCAGE))  +
  geom_bar() +
  xlab("Age of First Cocaine Use") +
  ggtitle("Age of First Cocaine Use Distribution")
```



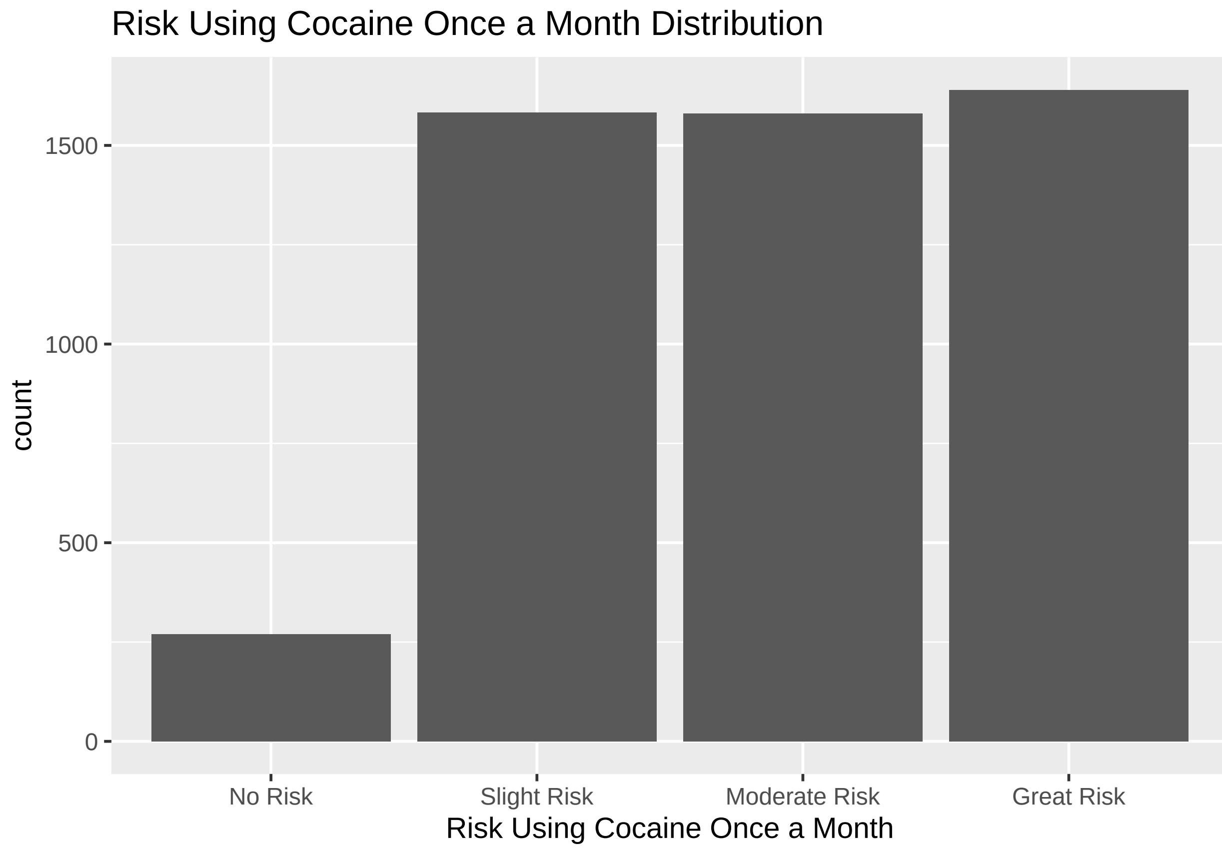

```r
ggplot(df1, aes(x = FUCOC18)) +
  geom_bar() +
  xlab("First used Cocaine Before 18 Years Old") +
  scale_x_discrete(labels=c("1" = "Yes", "2" = "No"))
```



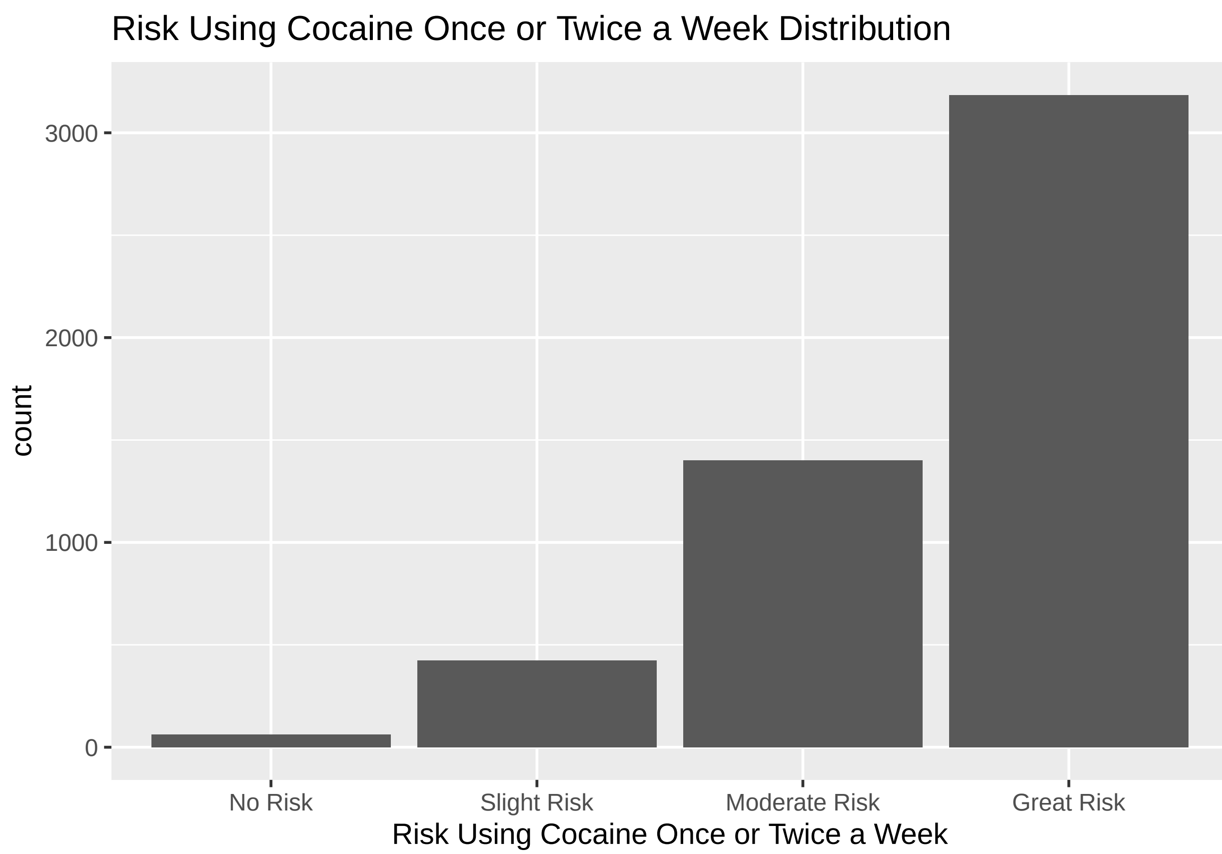

```
ggplot(df1, aes(x = DIFGETCOC)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "Probably Impossible", "2" = "Very Difficult", "3" = "Fairly Difficul
  xlab("Difficulty Getting Cocaine") +
  ggtitle("Difficulty Getting Cocaine Distribution")
```



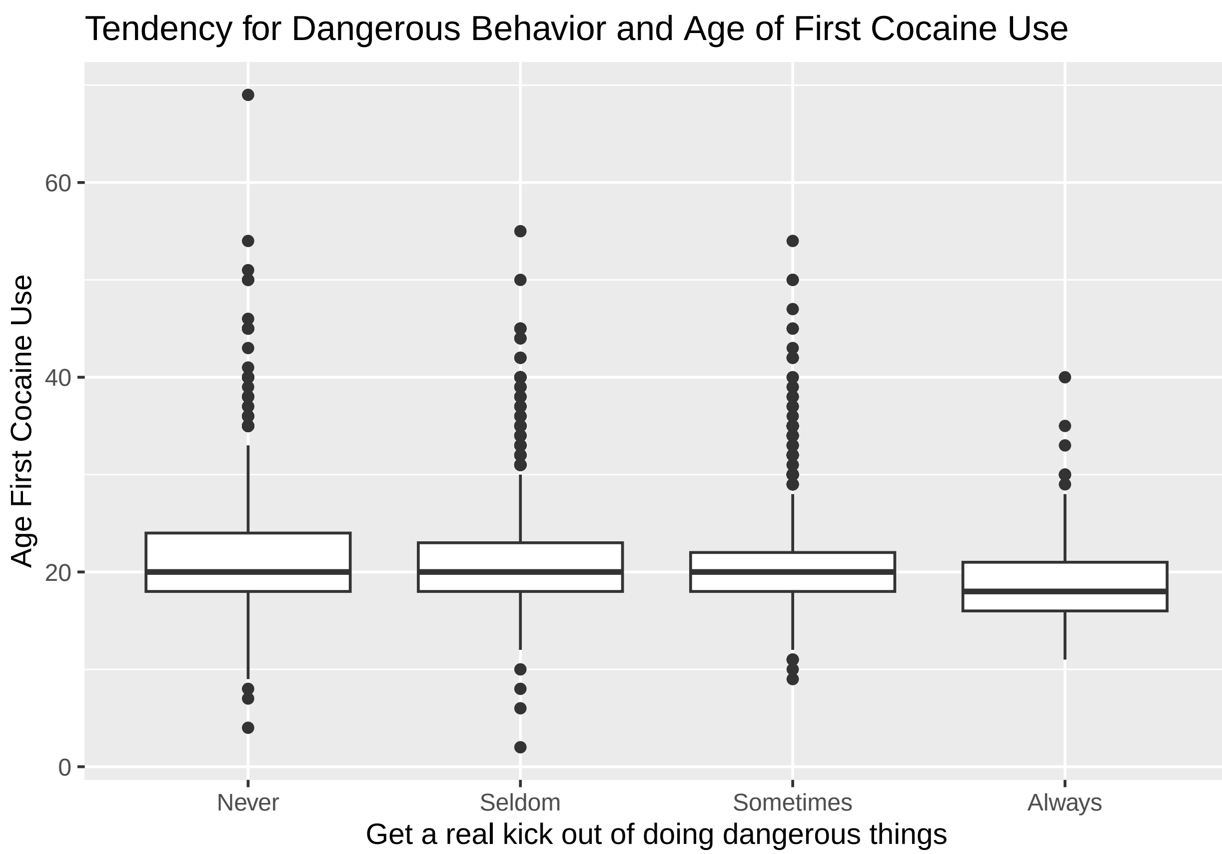

```
ggplot(df1, aes(x = RSKCOCMON)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "No Risk", "2" = "Slight Risk", "3" = "Moderate Risk","4" = "Great Ri
  xlab("Risk Using Cocaine Once a Month") +
  ggtitle("Risk Using Cocaine Once a Month Distribution")
```



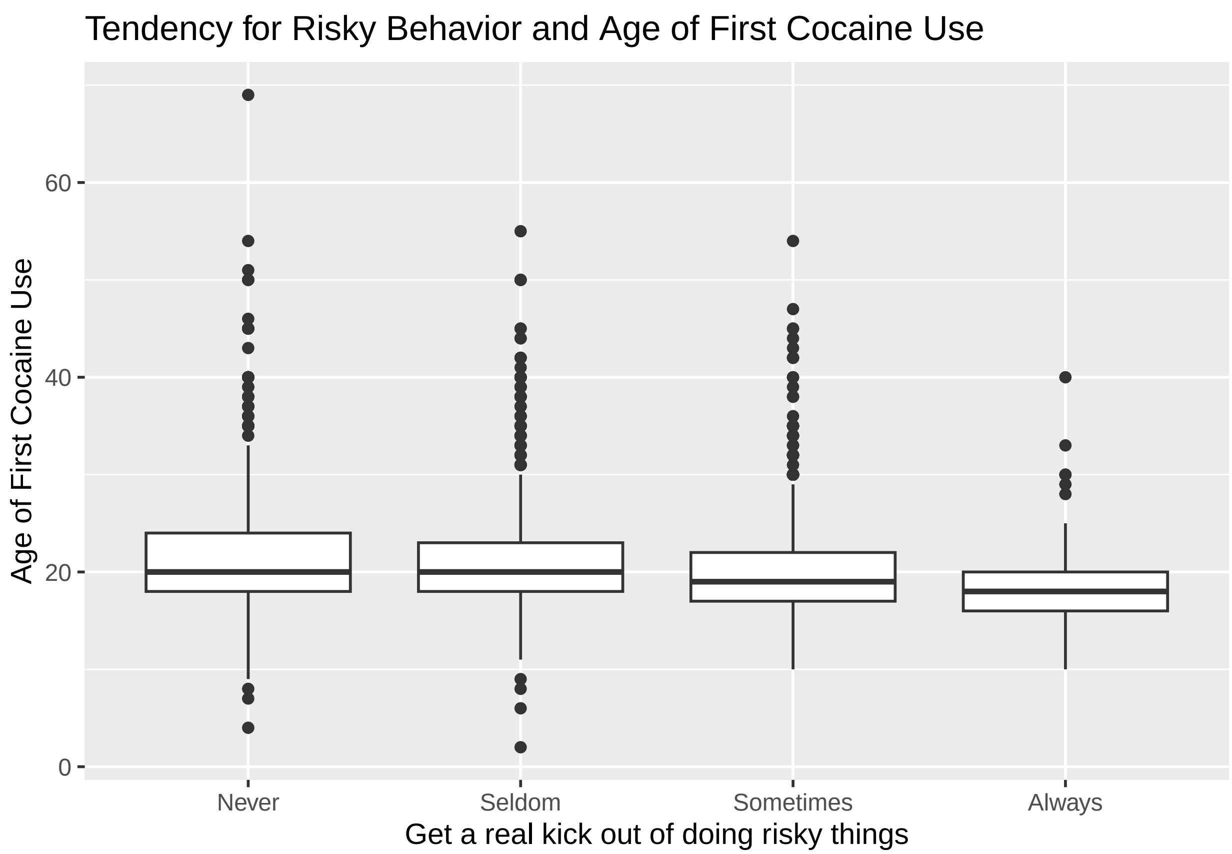

```
ggplot(df1, aes(x = RSKCOCWK)) +
  geom_bar() +
  scale_x_discrete(labels=c("1" = "No Risk", "2" = "Slight Risk", "3" = "Moderate Risk","4" = "Great Ris
  xlab("Risk Using Cocaine Once or Twice a Week") +
  ggtitle("Risk Using Cocaine Once or Twice a Week Distribution")
```



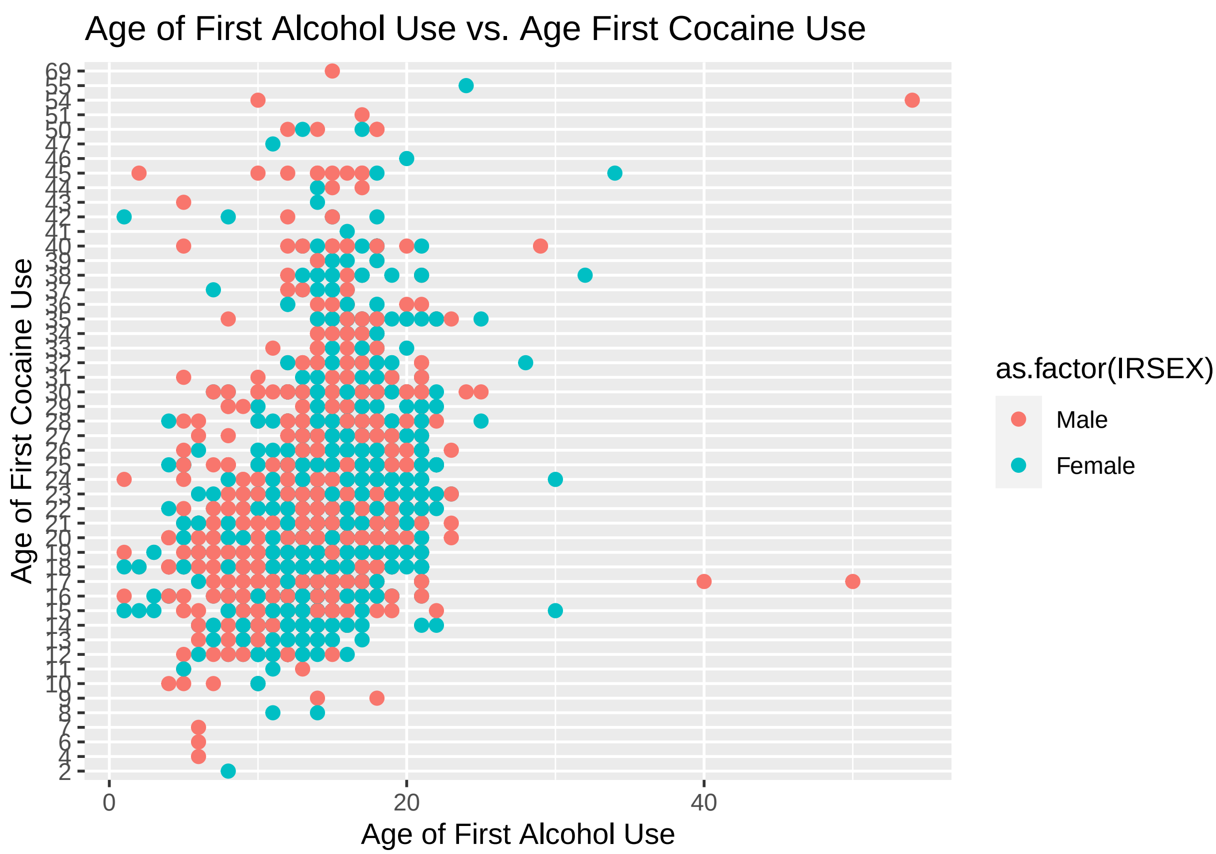

```r
ggplot(df1, aes(x = RSKYFQDGR, y = COCAGE)) +
  geom_boxplot() +
  xlab("Get a real kick out of doing dangerous things") +
  ylab("Age First Cocaine Use") +
  scale_x_discrete(labels=c("1" = "Never", "2" = "Seldom", "3" = "Sometimes","4" = "Always")) +
  ggtitle("Tendency for Dangerous Behavior and Age of First Cocaine Use")
```



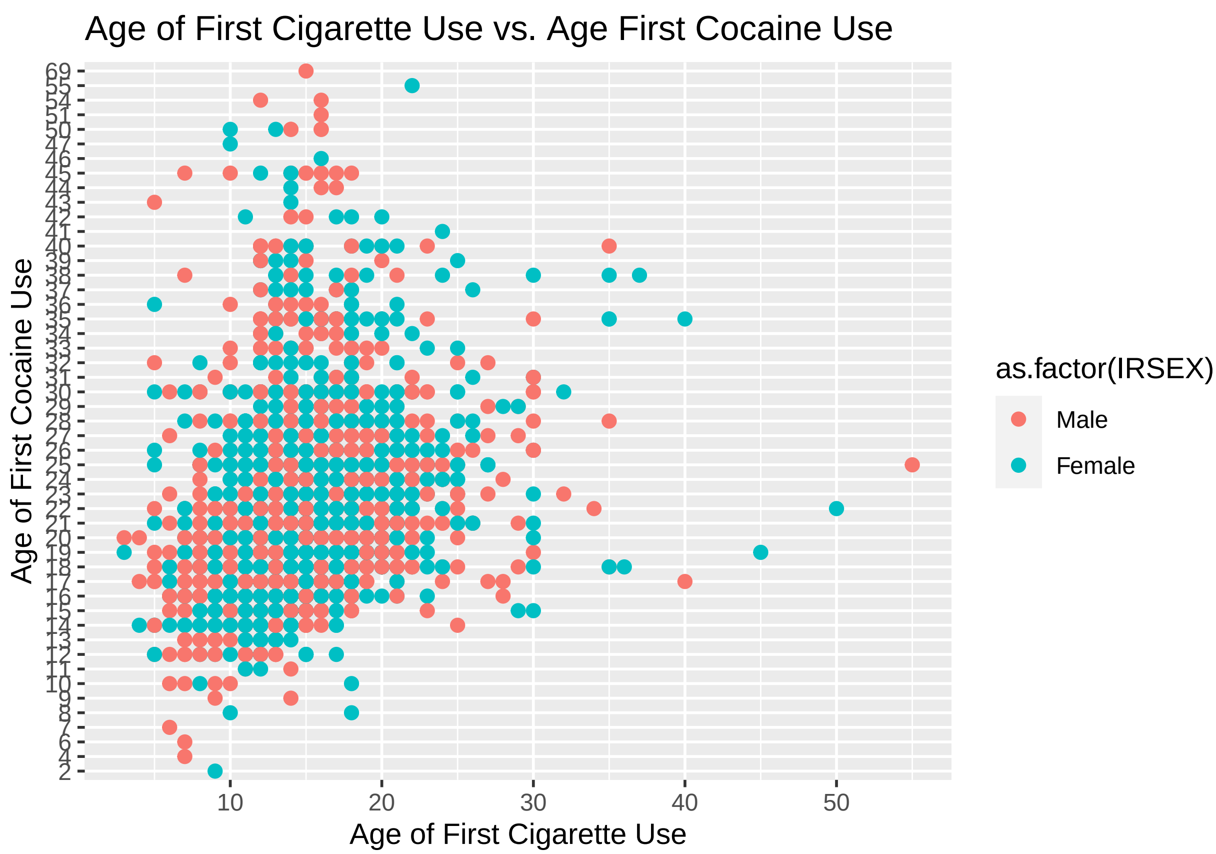

```r
ggplot(df1, aes(x = RSKYFQTES, y = COCAGE)) +
  geom_boxplot() +
  xlab("Get a real kick out of doing risky things") +
  ylab("Age of First Cocaine Use") +
  scale_x_discrete(labels=c("1" = "Never", "2" = "Seldom", "3" = "Sometimes","4" = "Always")) +
  ggtitle("Tendency for Risky Behavior and Age of First Cocaine Use")
```




```r
ggplot(df1, aes(x = IRALCAGE, y = as.factor(COCAGE), color = as.factor(IRSEX))) +
  geom_point(size = 2) +
  ggtitle("Age of First Alcohol Use vs. Age First Cocaine Use") +
  xlab("Age of First Alcohol Use") +
  ylab("Age of First Cocaine Use")
```




```
ggplot(df1, aes(x = IRCIGAGE, y = as.factor(COCAGE), color = as.factor(IRSEX))) +
  geom_point(size = 2) +
  ggtitle("Age of First Cigarette Use vs. Age First Cocaine Use") +
  xlab("Age of First Cigarette Use") +
  ylab("Age of First Cocaine Use")
```



Logistic regression

Predicting FUCOC18: First used cocaine before 18 years old. 1 = yes, 2 = no.

```
# Full model
logreg <- glm(FUCOC18 ~  DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK
summary(logreg)
```

```
##
## Call:
## glm(formula = FUCOC18 ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE +
##     IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSEX, family = binomial,
##     data = df1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.289464   0.386181  -8.518  < 2e-16 ***
## DIFGETCOC2  -0.142517   0.140519  -1.014  0.31048
## DIFGETCOC3  -0.352274   0.135498  -2.600  0.00933 **
## DIFGETCOC4  -0.234533   0.135260  -1.734  0.08293 .
## DIFGETCOC5  -0.378135   0.147796  -2.558  0.01051 *
## RSKYFQDGR2   0.099670   0.118968   0.838  0.40215
## RSKYFQDGR3  -0.126927   0.139946  -0.907  0.36442
## RSKYFQDGR4  -0.373253   0.212885  -1.753  0.07955 .
## RSKYFQTES2   0.001057   0.114458   0.009  0.99263
```

```
## RSKYFQTES3  -0.091399   0.143492  -0.637  0.52415
## RSKYFQTES4  -0.608691   0.256123  -2.377  0.01747 *
## IRALCAGE     0.138420   0.014115   9.807  < 2e-16 ***
## IRCIGAGE     0.181156   0.013345  13.574  < 2e-16 ***
## RSKCOCMON2  -0.172217   0.207011  -0.832  0.40546
## RSKCOCMON3  -0.284822   0.222996  -1.277  0.20151
## RSKCOCMON4  -0.557617   0.232801  -2.395  0.01661 *
## RSKCOCWK2    0.498379   0.354274   1.407  0.15950
## RSKCOCWK3    0.794486   0.363097   2.188  0.02866 *
## RSKCOCWK4    0.848723   0.371272   2.286  0.02226 *
## NEWRACE22    0.383105   0.214286   1.788  0.07380 .
## NEWRACE23   -0.337340   0.285515  -1.182  0.23740
## NEWRACE24   -0.186630   0.738435  -0.253  0.80047
## NEWRACE25   -0.309381   0.312654  -0.990  0.32240
## NEWRACE26   -0.049339   0.167275  -0.295  0.76803
## NEWRACE27   -0.579327   0.106765  -5.426 5.76e-08 ***
## IRSEXFemale -0.227410   0.074744  -3.043  0.00235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5383.6  on 5072  degrees of freedom
## Residual deviance: 4680.8  on 5047  degrees of freedom
## AIC: 4732.8
##
## Number of Fisher Scoring iterations: 5
```

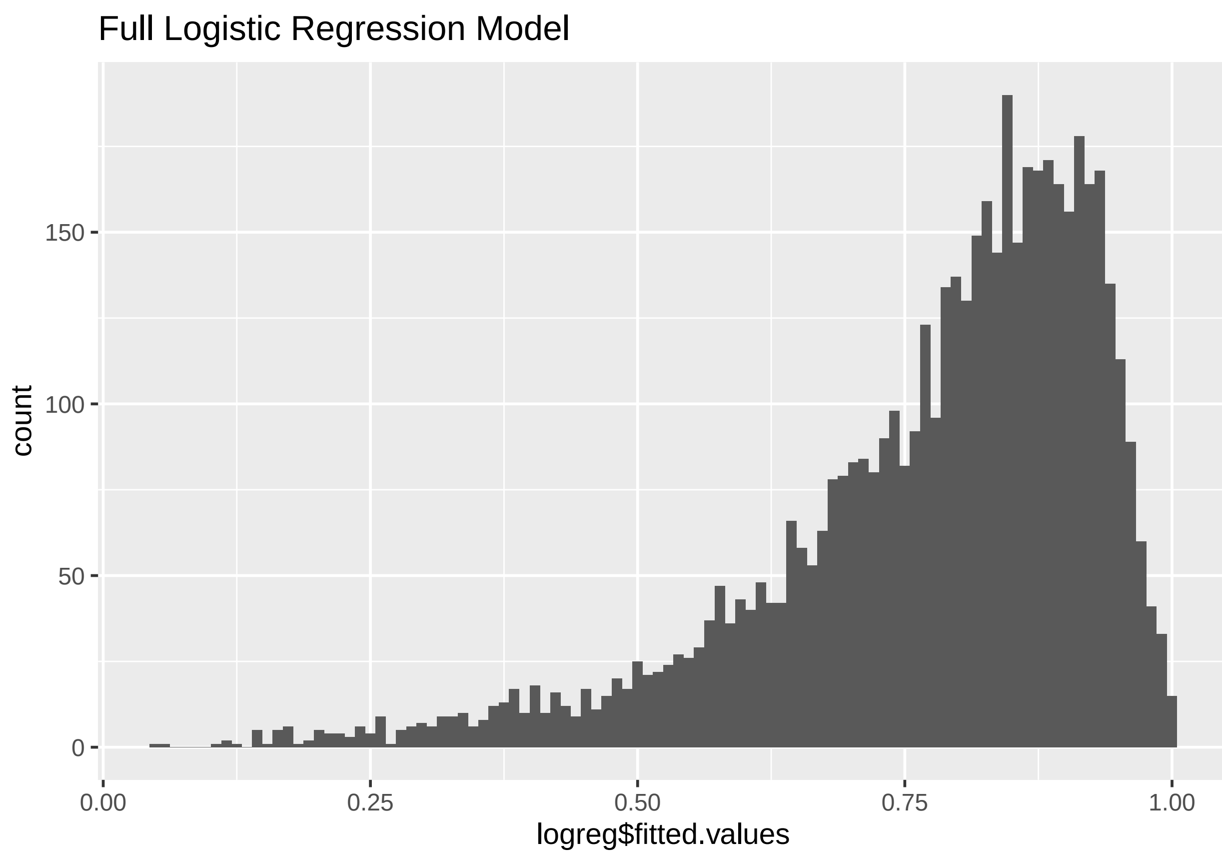
```
# Reduced model (Removed RSKYFDGR, the variable with the least significant levels overall)
logreg2 <- glm(FUCOC18 ~  DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2
summary(logreg2)
```

```
##
## Call:
## glm(formula = FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE +
##     RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSEX, family = binomial,
##     data = df1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.295490   0.383769  -8.587  < 2e-16 ***
## DIFGETCOC2  -0.144450   0.140308  -1.030  0.30323
## DIFGETCOC3  -0.358371   0.135211  -2.650  0.00804 **
## DIFGETCOC4  -0.245276   0.134762  -1.820  0.06875 .
## DIFGETCOC5  -0.401765   0.147272  -2.728  0.00637 **
## RSKYFQTES2   0.007429   0.091783   0.081  0.93549
## RSKYFQTES3  -0.227770   0.104031  -2.189  0.02857 *
## RSKYFQTES4  -0.929852   0.197891  -4.699 2.62e-06 ***
## IRALCAGE     0.139232   0.014078   9.890  < 2e-16 ***
## IRCIGAGE     0.181087   0.013324  13.591  < 2e-16 ***
## RSKCOCMON2  -0.155924   0.205966  -0.757  0.44903
## RSKCOCMON3  -0.264049   0.222158  -1.189  0.23461
## RSKCOCMON4  -0.543480   0.231984  -2.343  0.01914 *
## RSKCOCWK2    0.484703   0.352865   1.374  0.16956
```

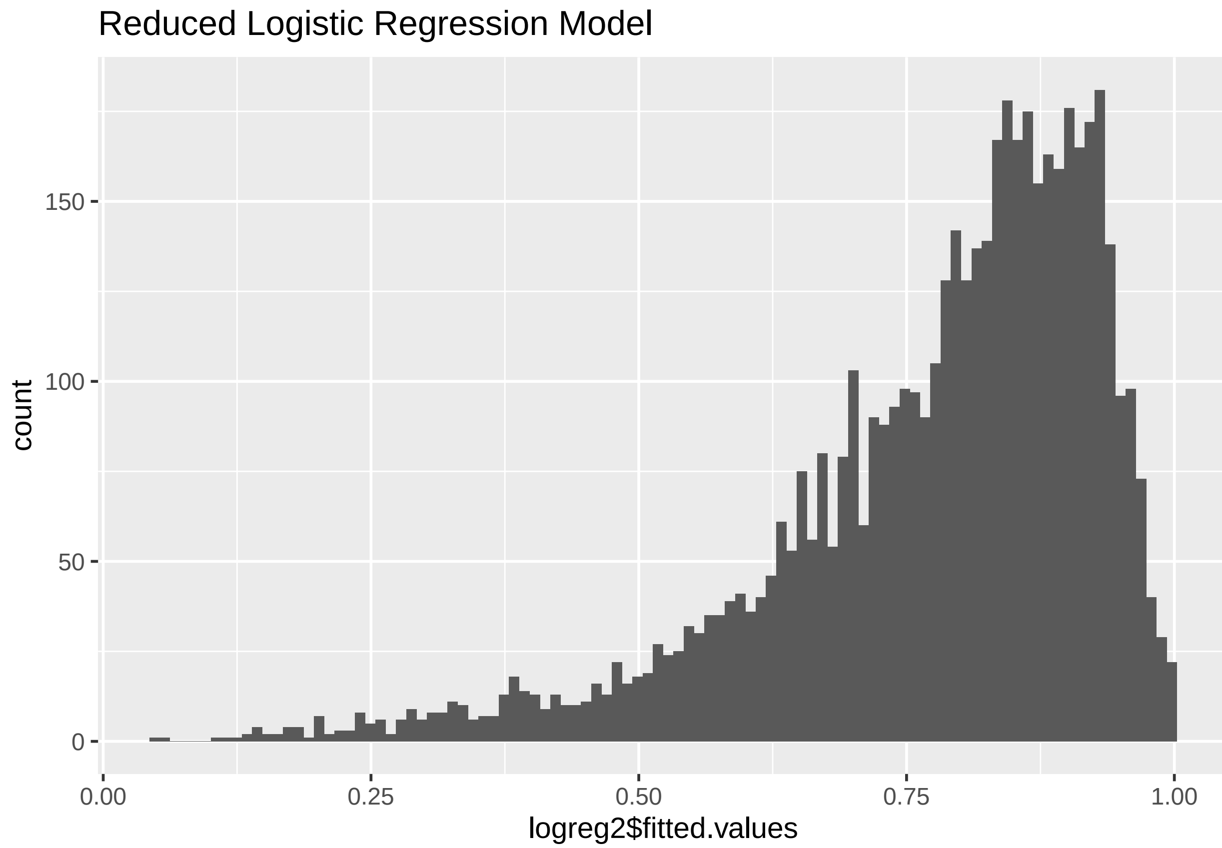
```
## RSKCOCWK3    0.788379   0.362053   2.178  0.02944 *
## RSKCOCWK4    0.851700   0.370256   2.300  0.02143 *
## NEWRACE22    0.383211   0.213620   1.794  0.07283 .
## NEWRACE23   -0.317186   0.285133  -1.112  0.26596
## NEWRACE24   -0.125825   0.736641  -0.171  0.86437
## NEWRACE25   -0.326197   0.310948  -1.049  0.29416
## NEWRACE26   -0.058411   0.166966  -0.350  0.72646
## NEWRACE27   -0.585598   0.106671  -5.490 4.02e-08 ***
## IRSEXFemale -0.219033   0.074515  -2.939  0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5383.6  on 5072  degrees of freedom
## Residual deviance: 4688.5  on 5050  degrees of freedom
## AIC: 4734.5
##
## Number of Fisher Scoring iterations: 5
```

Histogram of the fitted values and the plots of the OLS results

```r
# Full logistic regression model
ggplot(logreg$df1, aes(x = logreg$fitted.values)) +
  geom_histogram( bins = 100) +
  ggtitle("Full Logistic Regression Model")
```

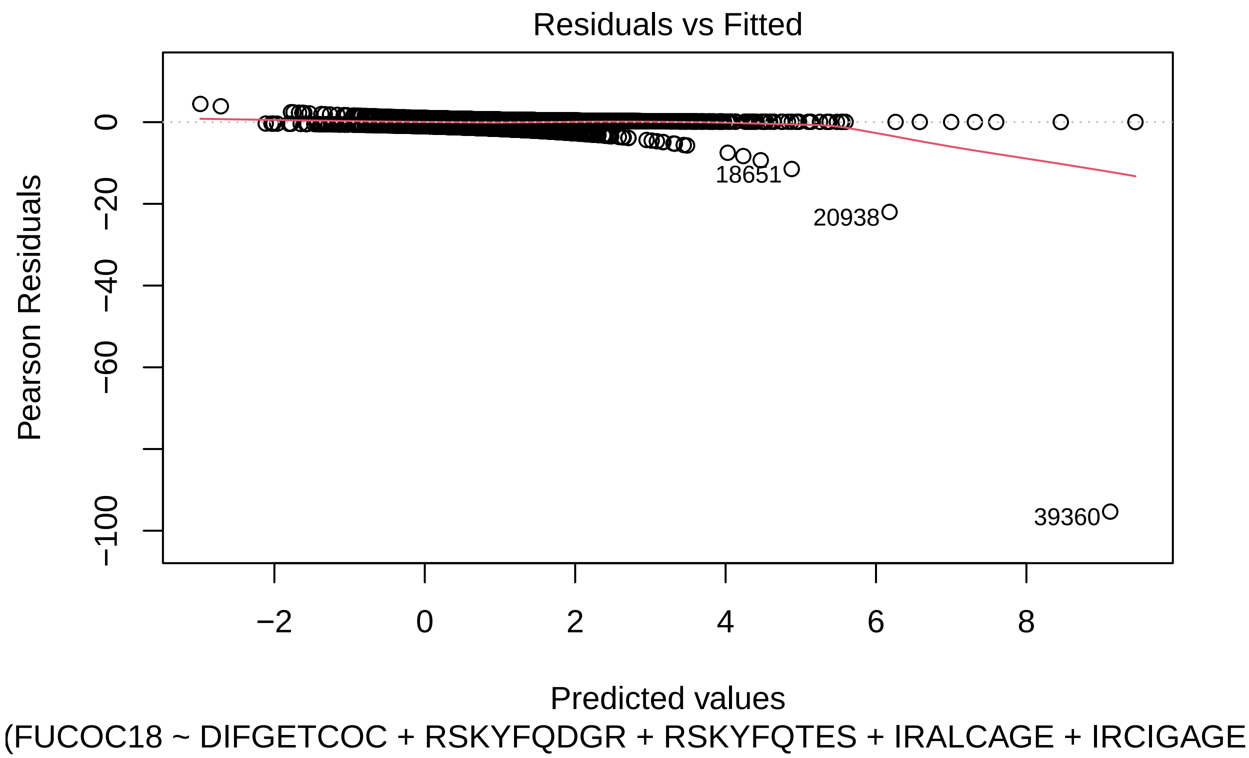


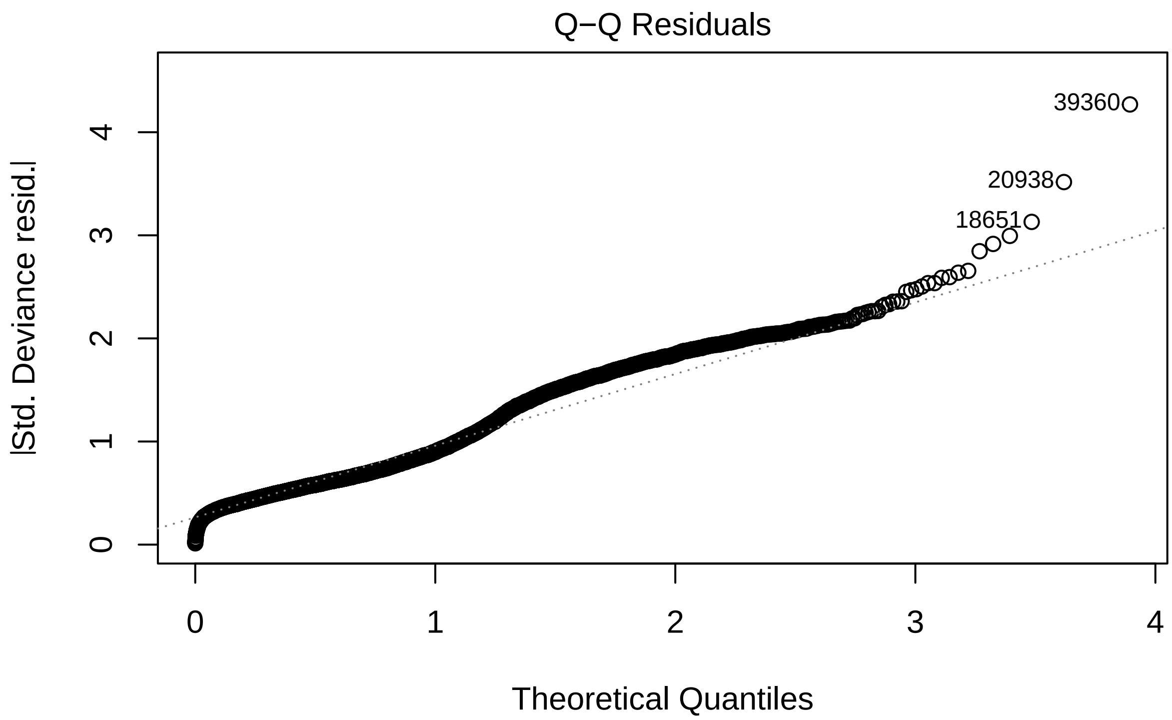

```r
# Reduced logistic regression model
ggplot(logreg2$df1, aes(x = logreg2$fitted.values)) +
  geom_histogram( bins = 100) +
  ggtitle("Reduced Logistic Regression Model")
```



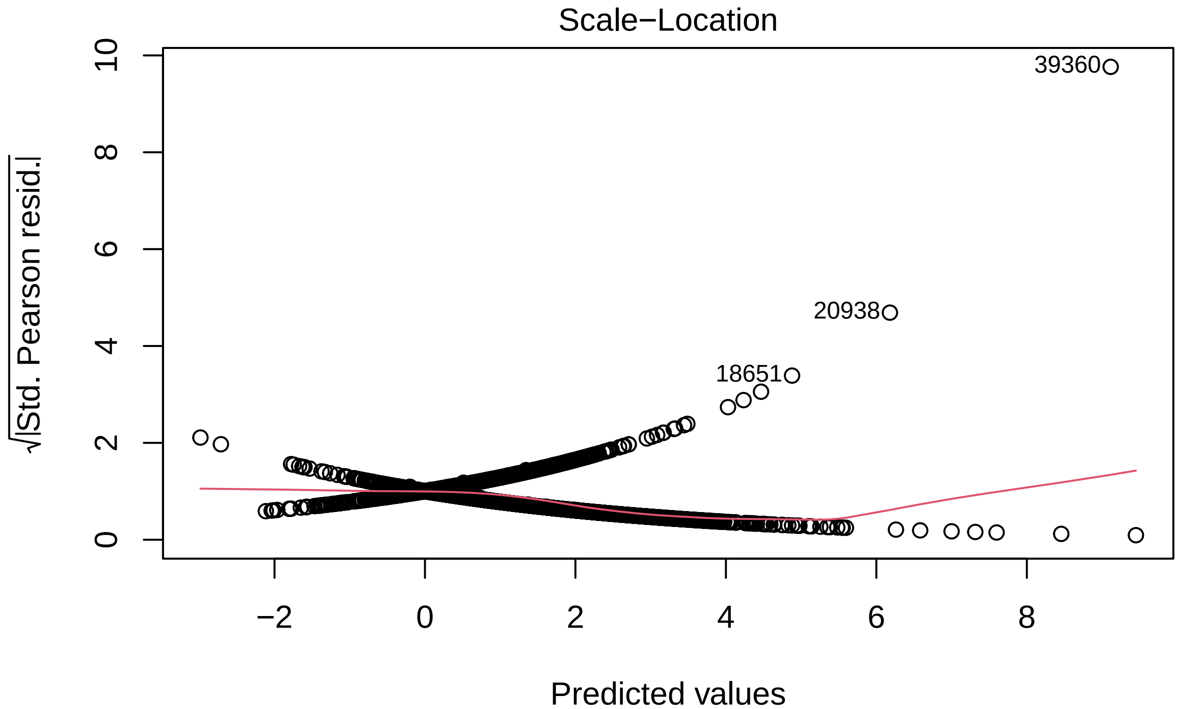
Plots of logistic regression models

```
# Full logistic regression model
plot(logreg)
```

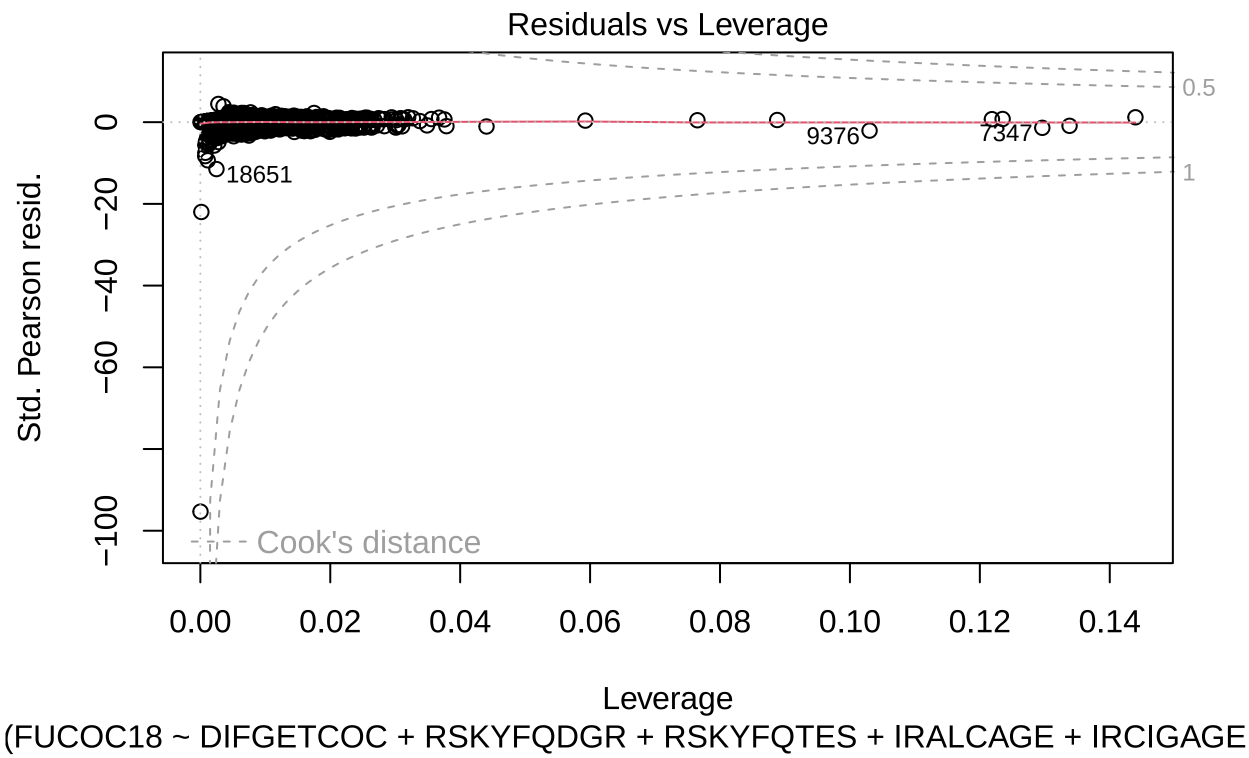

Residuals vs Fitted
Pearson Residuals
0
−20
−40
−60
−100
−2
0
2
4
6
8
18651
20938
39360
Predicted values
(FUCOC18 ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE



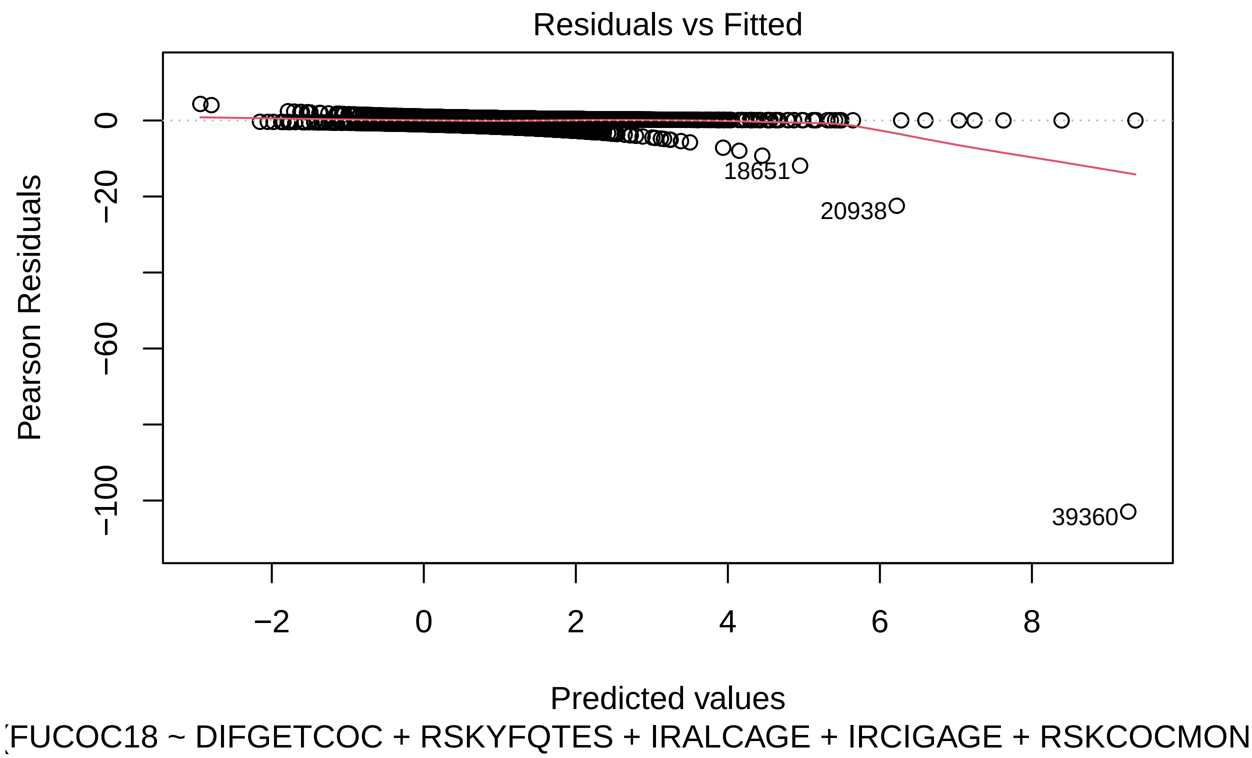
(FUCOC18 ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE

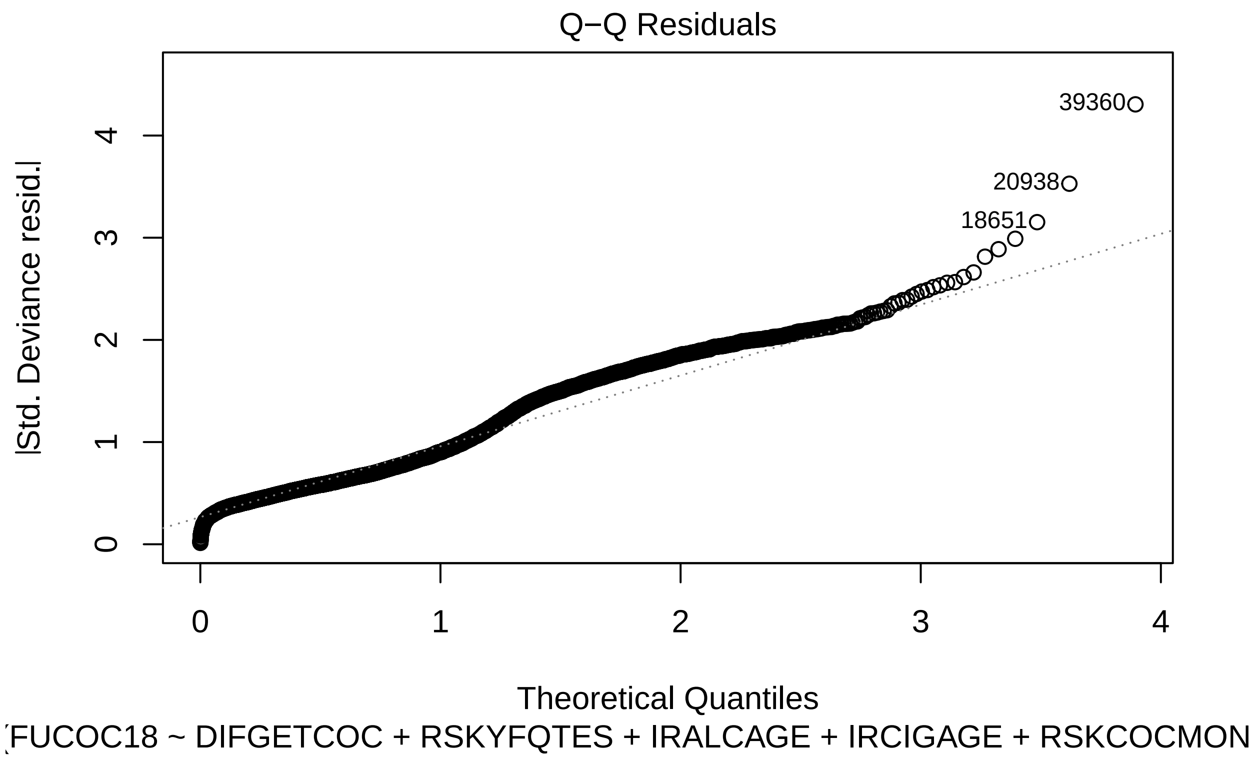


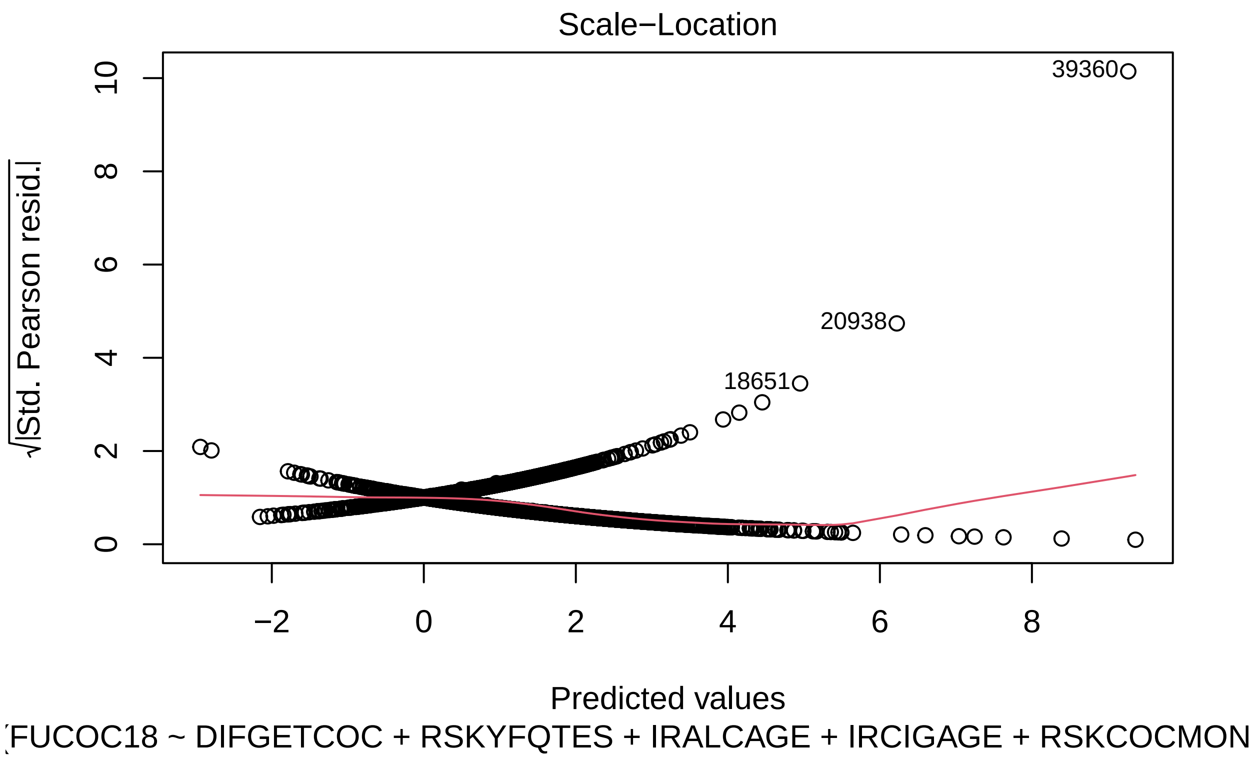
(FUCOC18 ~ DIFGETCOC + RSKYFQDGR + RSKYFQTES + IRALCAGE + IRCIGAGE

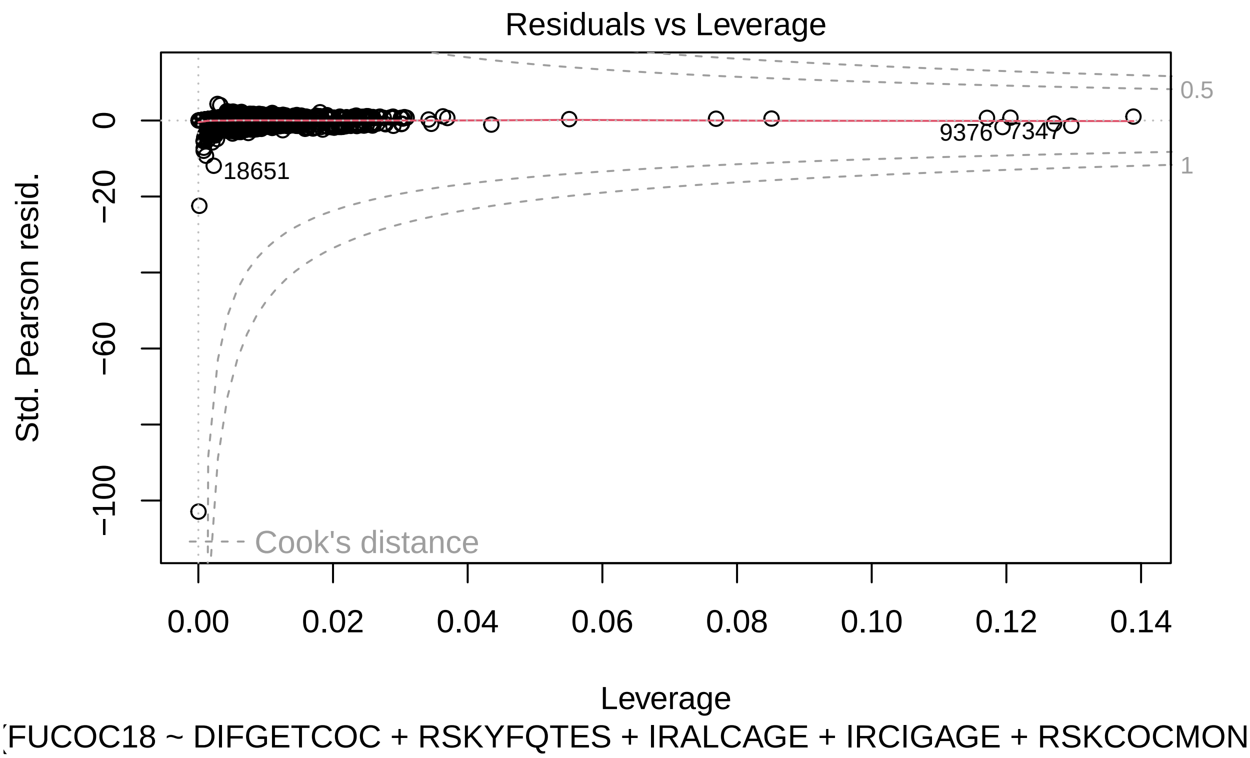



```r
# Reduced logistic regression model
plot(logreg2)
```


Residuals vs Fitted
Pearson Residuals
0
−20
−60
−100
−2
0
2
4
6
8
Predicted values
18651
20938
39360
FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON


Q−Q Residuals
39360
20938
18651
|Std. Deviance resid.|
0
1
2
3
4
Theoretical Quantiles
(FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON


Scale−Location
√|Std. Pearson resid.|
0
2
4
6
8
10
−2
0
2
4
6
8
39360
20938
18651
Predicted values
(FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON



Plots of both the full and reduced models show that there are outliers in the data. Since the points curve off at the extremities of the Q-Q plot, this would indicate the data has more extreme values than data coming from a perfectly normal distribution.

Deviance: the measure of "goodness of fit" used in general linear models

Note: The closer the p-value is to one, the closer the model corresponds to a "perfect" saturated model.

```
# Just the intercept term/ null deviance
pchisq(logreg$null.deviance, logreg$df.null, lower.tail = FALSE)
```

```
## [1] 0.001187091
```

```
# Full model
pchisq(logreg$deviance, logreg$df.residual, lower.tail = FALSE)
```

```
## [1] 0.9999042
```

```
# Reduced model
pchisq(logreg2$deviance, logreg2$df.residual, lower.tail = FALSE)
```

```
## [1] 0.9998834
```

Both of the models are very close to 1, which suggests that they fit the data well. Specifically, the full model is a better fit according to the amount of deviance.

Predicting new values

Tune the model to select a threshold

```r
df1 <- tidyr::drop_na(df1)

# Define the split between training and testing data
set.seed(1234)
training_pct <- .5
Z <- sample(nrow(df1), floor(training_pct*nrow(df1)))
log_train <- df1[Z, ]
log_test <- df1[-Z, ]

# Run the model on the training data
logreg <- glm(FUCOC18 ~  DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 +

summary(logreg)
```

```
##
## Call:
## glm(formula = FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE +
##     RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSEX, family = "binomial",
##     data = log_train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.429724   0.576343  -5.951 2.67e-09 ***
## DIFGETCOC2  -0.048536   0.202248  -0.240  0.81034
## DIFGETCOC3  -0.343506   0.190828  -1.800  0.07185 .
## DIFGETCOC4  -0.403279   0.189628  -2.127  0.03345 *
## DIFGETCOC5  -0.427156   0.211855  -2.016  0.04377 *
## RSKYFQTES2  -0.003562   0.128781  -0.028  0.97794
## RSKYFQTES3  -0.263333   0.148027  -1.779  0.07525 .
## RSKYFQTES4  -0.805029   0.299535  -2.688  0.00720 **
## IRALCAGE     0.155732   0.019702   7.905 2.69e-15 ***
## IRCIGAGE     0.163241   0.018982   8.600  < 2e-16 ***
## RSKCOCMON2   0.007186   0.274099   0.026  0.97909
## RSKCOCMON3  -0.146295   0.299351  -0.489  0.62505
## RSKCOCMON4  -0.437367   0.314575  -1.390  0.16442
## RSKCOCWK2    0.417824   0.533985   0.782  0.43394
## RSKCOCWK3    0.618447   0.549783   1.125  0.26063
## RSKCOCWK4    0.855958   0.561724   1.524  0.12756
## NEWRACE22    0.828476   0.335428   2.470  0.01351 *
## NEWRACE23   -0.259782   0.381433  -0.681  0.49583
## NEWRACE24   -0.209290   1.195018  -0.175  0.86097
## NEWRACE25   -0.282903   0.447623  -0.632  0.52738
## NEWRACE26    0.375575   0.256003   1.467  0.14236
## NEWRACE27   -0.471092   0.152918  -3.081  0.00207 **
## IRSEXFemale -0.079330   0.105267  -0.754  0.45109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2705.2  on 2535  degrees of freedom
## Residual deviance: 2360.6  on 2513  degrees of freedom
## AIC: 2406.6
##
## Number of Fisher Scoring iterations: 5
```

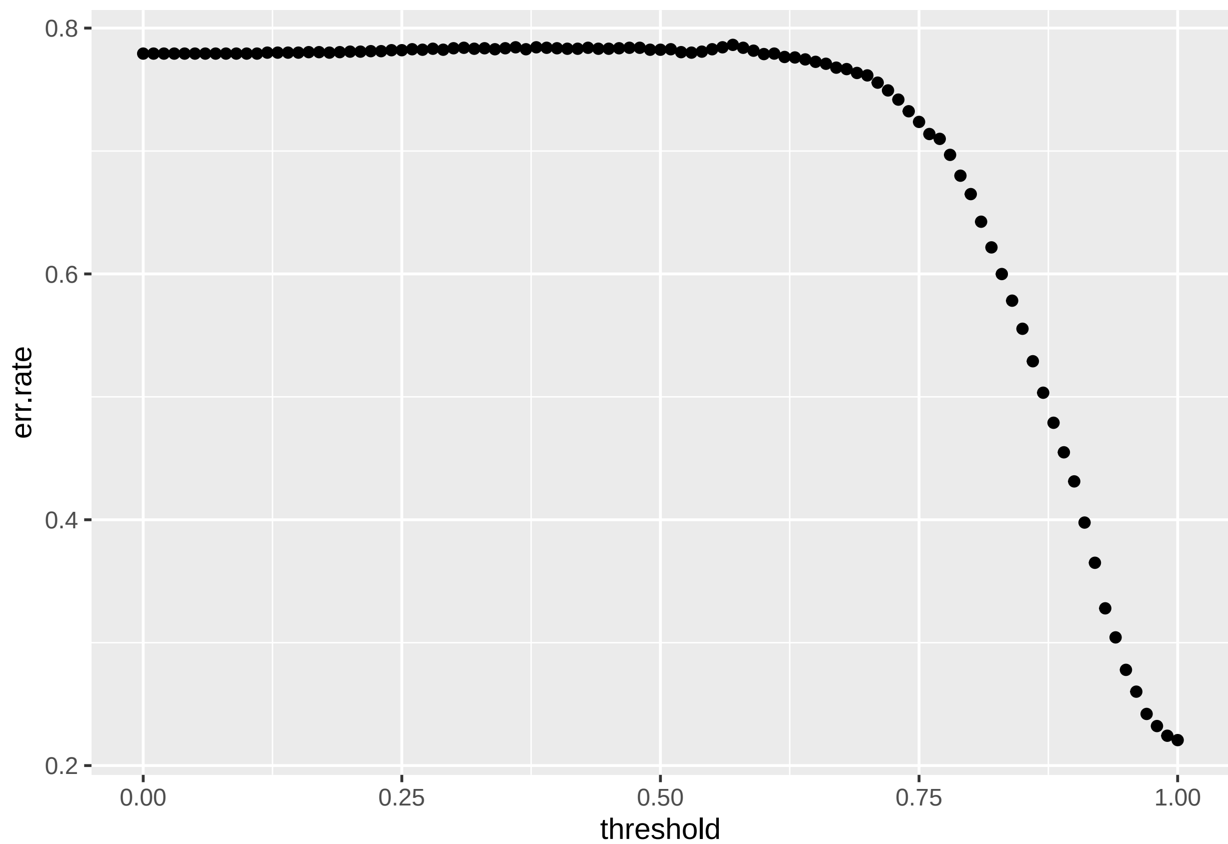
Predicting with the test data

```r
# Get predictions on the test data
Prob <- predict(logreg, type = "response", newdata = log_test)

# Set up the possible thresholds
threshold <- seq(0, 1, .01)
length(threshold)
```

```
## [1] 101
```

Test all the possible thresholds

```r
TPR <-  FPR <- err.rate <- rep(0, length(threshold))

for (i in seq_along(threshold)) {
Yhat <- rep(NA_character_, nrow(log_test))
Yhat <-  ifelse(Prob >= threshold[[i]], "1", "2")

err.rate[i] <- mean(Yhat != log_test$FUCOC18)
TPR[[i]] <- sum(Yhat == "1" & log_test$FUCOC18 == "1") /
  sum(log_test$FUCOC18 == "1")
FPR[[i]] <- sum(Yhat == "1" & log_test$FUCOC18 == "2") /
  sum(log_test$FUCOC18 == "2")
}

ggplot(tibble(threshold, err.rate),
       aes(threshold, err.rate)) +
  geom_point()
```




```r
table(log_test$FUCOC18)
```

```
##
##    1    2
##  560 1977
```

```r
# What is the minimum error rate of our model? 0.2207331
min(err.rate)
```

```
## [1] 0.2207331
```

```r
# What is the best threshold?
which.min(err.rate)
```

```
## [1] 101
```

```r
threshold[which.min(err.rate)]
```

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

```r
Yhat <- ifelse(Prob >= threshold[which.min(err.rate)], "1", "2")
table(Yhat, log_test$FUCOC18)
```

```
##
## Yhat    1    2
##    2  560 1977
```

Determine how well the logistic regression model performs

```
round(mean(log_test$FUCOC18 == Yhat), 3) # Correct classification rate
```

```
## [1] 0.779
```

Correct classification rate of 77.9%

LDA & QDA

LDA: The LDA discriminant function assumes equal variance for all classes

```
suppressMessages(library(tidyverse))
library(MASS)
library(ISLR2)

# Define the split between training and testing data
set.seed(1234)
training_pct <- .5
Z <- sample(nrow(df1), floor(training_pct*nrow(df1)))
lda_train <- df1[Z, ]
lda_test <- df1[-Z, ]

lda_out <- lda(FUCOC18 ~  DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2

Predicted.Direction_lda <- predict(lda_out, data.frame(lda_test))$class

table(lda_test$FUCOC18, Predicted.Direction_lda)
```

```
##    Predicted.Direction_lda
##        1    2
##   1   81  479
##   2   76 1901
```

How well did the LDA model perform?

```
round(mean(lda_test$FUCOC18 == Predicted.Direction_lda), 3) # Classification Rate
```

```
## [1] 0.781
```

Correct classification rate of 78.1%

QDA: The QDA discriminant function does not assume equal variance for all classes.

```r
# Define the split between training and testing data
set.seed(1234)
training_pct <- .5
Z <- sample(nrow(df1), floor(training_pct*nrow(df1)))
qda_train <- df1[Z, ]
qda_test <- df1[-Z, ]

qda_out <- qda(FUCOC18 ~  DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2

Predicted.Direction_qda <- predict(qda_out, data.frame(qda_test))$class

table(qda_test$FUCOC18, Predicted.Direction_qda)
```

```
##    Predicted.Direction_qda
##        1    2
##   1  152  408
##   2  205 1772
```

How well did the QDA model perform?

```r
round(mean(qda_test$FUCOC18 == Predicted.Direction_qda), 3)
```

```
## [1] 0.758
```

Correct classification rate of 75.8%

General takeaway:

The model with the highest correct classification rate on the testing data was the LDA model. Therefore, I would recommend the following model for classifying whether an respondent used cocaine for the first time before age 18:

lda(FUCOC18 ~ DIFGETCOC + RSKYFQTES + IRALCAGE + IRCIGAGE + RSKCOCMON + RSKCOCWK + NEWRACE2 + IRSEX)