1-Exploratory Data Analysis (EDA):

```
# Install the necessary libraries
library(dplyr)

library(ggplot2)

# Loading data

cat_data <-

read.csv("C:/Users/ASUS/Downloads/Cat_personality_data.csv",

header=TRUE, sep=";",encoding="UTF-8", check.names = FALSE,

stringsAsFactors = FALSE) # Read .csv file

# Display the first few lines
head(cat_data)

# Display the last few lines
tail(cat_data)

# Missing value check
missing_values <- sum(is.na(cat_data))
missing_values
[1] 0


summary(cats_data)
```

```
Personality40_Aggressive_other_cats Personality41_Excitable Personality42_Friendly_to_people Personality43_Playful
Min.   :1.000                       Min.   :1.000          Min.   :1.000                     Min.   :1.000
1st Qu.:2.000                       1st Qu.:2.000          1st Qu.:4.000                     1st Qu.:5.000
Median :3.000                       Median :3.000          Median :5.000                     Median :5.000
Mean   :3.387                       Mean   :3.304          Mean   :5.092                     Mean   :5.189
3rd Qu.:5.000                       3rd Qu.:5.000          3rd Qu.:7.000                     3rd Qu.:6.000
Max.   :7.000                       Max.   :7.000          Max.   :7.000                     Max.   :7.000
Personality44_Vocal Personality45_Decisive Personality46_Self_assured Personality47_Anxious Personality48_Trusting
Min.   :1.00        Min.   :1.00           Min.   :1.000              Min.   :1.000         Min.   :1.000
1st Qu.:4.00        1st Qu.:4.00           1st Qu.:5.000              1st Qu.:2.000         1st Qu.:3.000
Median :6.00        Median :5.00           Median :6.000              Median :3.000         Median :4.000
Mean   :5.25        Mean   :5.16           Mean   :5.363              Mean   :3.087         Mean   :4.254
3rd Qu.:7.00        3rd Qu.:6.00           3rd Qu.:6.000              3rd Qu.:4.000         3rd Qu.:6.000
Max.   :7.00        Max.   :7.00           Max.   :7.000              Max.   :7.000         Max.   :7.000
Personality49_Active Personality50_Cooperative Personality51_Shy Personality52_Eccentric   Country            Cat_sex
Min.   :1.00         Min.   :1.000             Min.   :1.00      Min.   :1.000           Length:2764        Min.   :0.0000
1st Qu.:4.00         1st Qu.:3.000             1st Qu.:2.00      1st Qu.:2.000           Class :character   1st Qu.:0.0000
Median :6.00         Median :5.000             Median :3.00      Median :4.000           Mode  :character   Median :1.0000
Mean   :5.17         Mean   :4.426             Mean   :3.43      Mean   :3.389                              Mean   :0.5018
3rd Qu.:7.00         3rd Qu.:6.000             3rd Qu.:5.00      3rd Qu.:4.000                              3rd Qu.:1.0000
Max.   :7.00         Max.   :7.000             Max.   :7.00      Max.   :7.000                              Max.   :1.0000
```

Categorical variables such as Cat_sex and Country, Considering the overall data, it can be said that the majority of personality traits do not deviate from the normal distribution and generally show a spread between 1 and 7.

```
# Visualize the number of cats by gender
ggplot(cat_data, aes(x = Cat_sex)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Number of Cats by Gender", x = "Gender", y =
"Frequency")
```

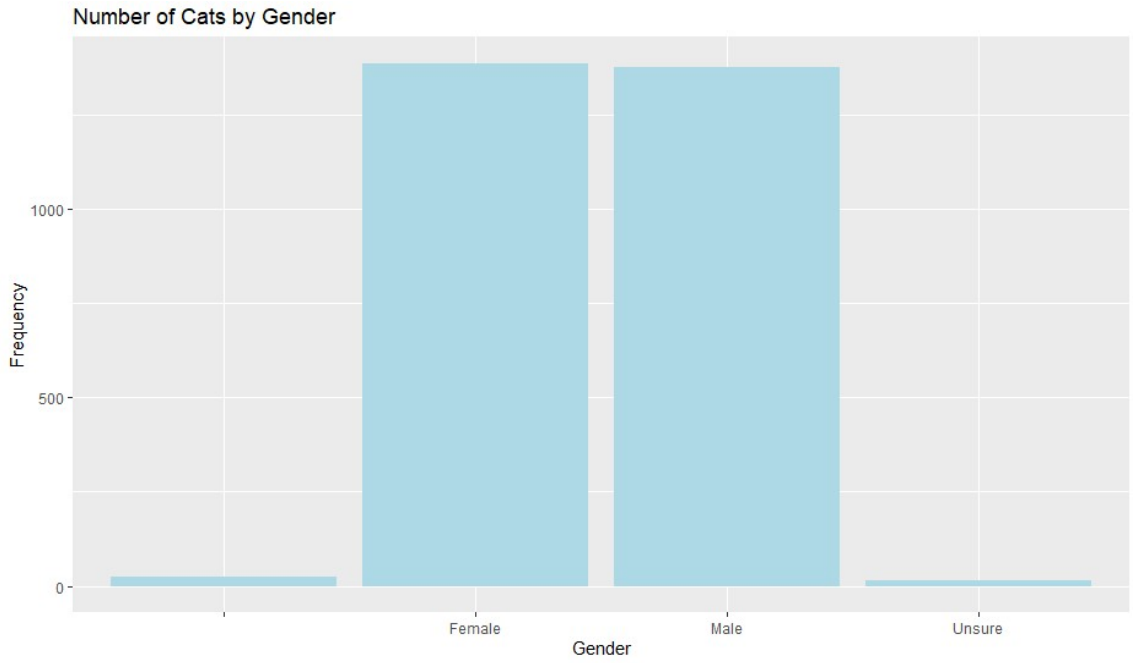

FIG1. Visualize the number of cats by gender

```
# Number of NA values
na_count <- sum((cat_data$Cat_sex != "Female" & cat_data$Cat_sex !=
"Male") & cat_data$Cat_sex != "Unsure" )
cat("Number of NA values: ", na_count, "\n")

Number of NA values:  23

# number of "Unsure" values
unsure_count <- sum(cat_data$Cat_sex == "Unsure", na.rm = TRUE)
cat("Unsure number of values: ", unsure_count, "\n")

Unsure number of values:  15

# Remove empty and "unsure" values
cleaned_data <- cat_data %>%
+   filter(cat_data$Cat_sex == "Female" | cat_data$Cat_sex ==
"Male")

# Checking the remaining data
cat("Number of rows in the cleaned dataset: ", nrow(cleaned_data),
"\n")
Number of rows in the cleaned dataset:  2764
# The number of cats by gender
table(cat_data$"Cat_sex")
       Female   Male Unsure
    23   1387   1377     15
table(cleaned_data$"Cat_sex")
     Female   Male
       1387   1377
# Visualize the number of cats by gender
ggplot(cats_data, aes(x = Cat_sex)) +
  geom_bar(fill = "lightblue") +
```

```
  labs(title = "Number of Cats by Gender", x = "Gender", y =
"Frequency")
```

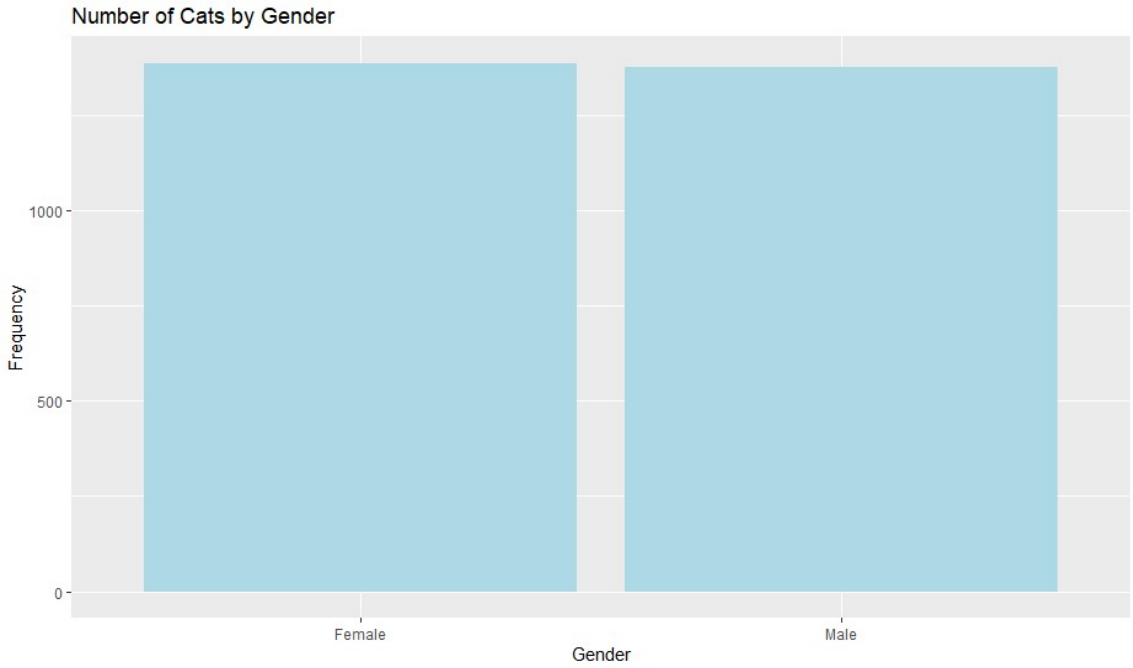

FIG2. Visualize the number of cats by gender

```
# Histogram of the vigilant personality score
ggplot(cats_data, aes(x = Personality1_Vigilant)) +
  geom_histogram(binwidth = 1, fill = "lightgreen", color = "black")
+
  labs(title = "Vigilant Personality Trait Distribution", x =
"Score", y = "Frequency")
```

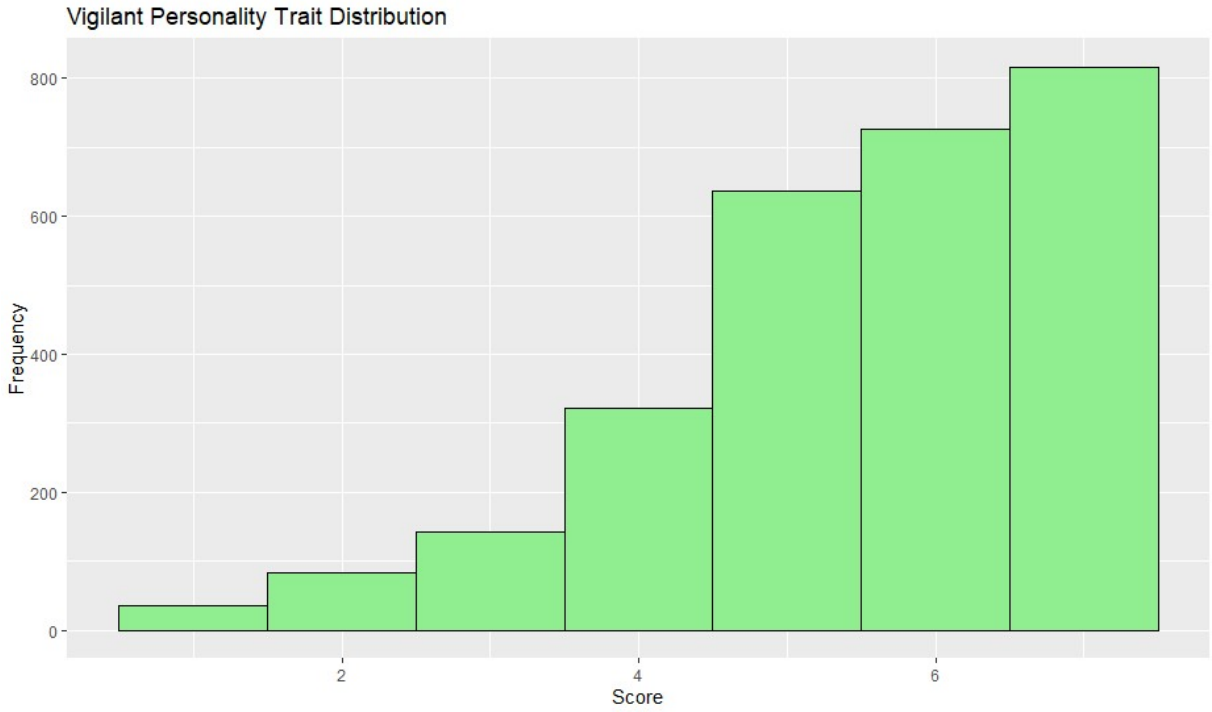

FIG3. Histogram of the vigilant personality score

2-Model Selection:

1. GLM (Generalized Linear Model)

```
cats_data$Cat_sex <- ifelse(cats_data$Cat_sex == "Female", 1, 0)
# GLM: Logistic Regression - modeling cat_sex with the whole
Personality
glm_logit <- glm(Cat_sex ~ .,
                 family = binomial(link = "logit"),
                 data = cats_data)
summary(glm_logit)

# GLM: Probit Regression - modeling cat_sex with the whole
Personality
glm_probit <- glm(Cat_sex ~ .,
                  family = binomial(link = "probit"),
                  data = cats_data)
summary(glm_probit)
```

I wanted to model the cat genders with all personality traits and so I was able to evaluate all personality traits and choose personality traits for 3 models

```
Call:

glm(formula = Cat_sex ~ ., family = binomial(link = "logit"),
    data = cats_data)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.5695413  0.5715345   0.997 0.319001
Personality1_Vigilant          0.0241180  0.0335574   0.719 0.472320
Personality2_Stable           -0.0247397  0.0343726  -0.720 0.471679
Personality3_Bold             -0.0105866  0.0319203  -0.332 0.740148
Personality4_Clumsy           -0.0519107  0.0272276  -1.907 0.056579 .
Personality5_Defiant           0.0257624  0.0273370   0.942 0.345987
Personality6_Gentle           -0.0170730  0.0344558  -0.496 0.620243
Personality7_Constrained      -0.0040690  0.0293265  -0.139 0.889648
Personality8_Inquisitive       0.0218850  0.0376179   0.582 0.560720
Personality9_Inventive         0.0316281  0.0302562   1.045 0.295864
Personality10_Irritable        0.0631699  0.0317329   1.991 0.046517 *
Personality11_Distractable     0.0532089  0.0296062   1.797 0.072300 .
Personality12_Erratic          0.0489592  0.0334052   1.466 0.142754
Personality13_Solitary         0.0520134  0.0259741   2.003 0.045230 *
Personality14_Impulsive        0.0203849  0.0293047   0.696 0.486669
Personality15_Quitting        -0.0443243  0.0313980  -1.412 0.158041
Personality16_Independent     -0.0207814  0.0293792  -0.707 0.479349
Personality17_Smart            0.0006066  0.0354853   0.017 0.986361
Personality18_Jealous          0.0414558  0.0256664   1.615 0.106272
Personality19_Fearful_other_cats  0.1220054  0.0271601   4.492 7.05e-06 ***
Personality20_Persevering      0.0212339  0.0309198   0.687 0.492245
Personality21_Greedy          -0.0835506  0.0237336  -3.520 0.000431 ***
Personality22_Friendly_other_cats -0.1648045  0.0299196  -5.508 3.62e-08 ***
Personality23_Submissive      -0.0434098  0.0298238  -1.456 0.145519
Personality24_Dominant        -0.1171114  0.0328668  -3.563 0.000366 ***
Personality25_Reckless        -0.0872247  0.0304614  -2.863 0.004190 **
Personality26_Predictable     -0.0121446  0.0325875  -0.373 0.709389
Personality27_Suspicious      -0.0382870  0.0319528  -1.198 0.230824
Personality28_Individualistic  0.0060801  0.0272296   0.223 0.823310
Personality29_Affectionate     0.0370873  0.0353637   1.049 0.294299
Personality30_Insecure        -0.0022996  0.0347716  -0.066 0.947270
Personality31_Bullying        -0.1295796  0.0343090  -3.777 0.000159 ***
Personality32_Curious          0.1149593  0.0420534   2.734 0.006264 **
Personality33_Aimless         -0.0130021  0.0345383  -0.376 0.706580
Personality34_Deliberate       0.0149872  0.0372488   0.402 0.687424
Personality35_Tense           -0.0312143  0.0328389  -0.951 0.341845
Personality36_Fearful_of_people  0.0007249  0.0332300   0.022 0.982596
Personality37_Cool            -0.0706304  0.0285795  -2.471 0.013460 *
```

```
Personality38_Aggressive_to_people  -0.0062616  0.0375935  -0.167 0.867716
Personality39_Calm                   0.0796564  0.0354099   2.250 0.024477 *
Personality40_Aggressive_other_cats  0.0663395  0.0323078   2.053 0.040038 *
Personality41_Excitable             -0.0055121  0.0296364  -0.186 0.852451
Personality42_Friendly_to_people     0.0278557  0.0343979   0.810 0.418051
Personality43_Playful                0.0830525  0.0339190   2.449 0.014343 *
Personality44_Vocal                 -0.0339076  0.0246453  -1.376 0.168876
Personality45_Decisive              -0.0583260  0.0435371  -1.340 0.180348
Personality46_Self_assured           0.0001124  0.0423710   0.003 0.997884
Personality47_Anxious                0.0058796  0.0371462   0.158 0.874234
Personality48_Trusting              -0.0473081  0.0355288  -1.332 0.183011
Personality49_Active                -0.0738572  0.0305919  -2.414 0.015766 *
Personality50_Cooperative           -0.0569953  0.0287983  -1.979 0.047802 *
Personality51_Shy                    0.0816809  0.0325142   2.512 0.011999 *
Personality52_Eccentric             -0.0380432  0.0252401  -1.507 0.131747
CountryNewZealand                    0.3262295  0.0854197   3.819 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3831.7  on 2763  degrees of freedom
Residual deviance: 3514.1  on 2710  degrees of freedom
AIC: 3622.1

Number of Fisher Scoring iterations: 4
```

```
log_model1 <- glm(Cat_sex ~ Personality51_Shy + Personality17_Smart
+ Personality31_Bullying,
                  data = cats_data,
                  family = "binomial")
summary(log_model1)
```

```
Call:
glm(formula = Cat_sex ~ Personality51_Shy + Personality17_Smart +
    Personality31_Bullying, family = "binomial", data = cats_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.24377    0.18441  -1.322    0.186
Personality51_Shy       0.13834    0.02141   6.462 1.04e-10 ***
Personality17_Smart     0.01173    0.02649   0.443    0.658
Personality31_Bullying -0.09279    0.02071  -4.479 7.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3831.7  on 2763  degrees of freedom
Residual deviance: 3763.5  on 2760  degrees of freedom
AIC: 3771.5

Number of Fisher Scoring iterations: 4
```

General Interpretation of the Model:

- Personality51_Shy : The coefficient is positive and significant ($p < 0.001$). This result indicates that the higher the level of shyness of cats, the more likely they are to belong to a particular gender.
- Personality31_Bullying: The coefficient is negative and significant ($p < 0.001$). This result indicates that as cats' bullying levels increase, the likelihood of belonging to a specific gender decreases.

- Personality17_Smart: The coefficient is negative but the significance level is just above 5% ($p \approx 0.067$). This suggests that the level of calmness has a weak effect on the sex of the cat.

```
log_model2 <- glm(Cat_sex ~ Personality19_Fearful_other_cats +
Personality22_Friendly_other_cats,
                  data = cats_data,
                  family = "binomial")
summary(log_model2)
```

```
Call:
glm(formula = Cat_sex ~ Personality19_Fearful_other_cats +
Personality22_Friendly_other_cats,
    family = "binomial", data = cats_data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -0.10451    0.12649  -0.826    0.409
Personality19_Fearful_other_cats   0.18278    0.02212   8.262  < 2e-16 ***
Personality22_Friendly_other_cats -0.13554    0.02313  -5.860 4.63e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3831.7  on 2763  degrees of freedom
Residual deviance: 3695.4  on 2761  degrees of freedom
AIC: 3701.4

Number of Fisher Scoring iterations: 4
```

General Interpretation of the Model:

- Personality19_Fearful_other_cats: The coefficient is positive (0.18278) and statistically significant ($p < 0.001$). This result indicates that the higher the level of fear of other cats, the higher the probability that the cat belongs to a particular gender.
- Personality22_Friendly_other_cats (Friendliness towards other cats): The coefficient is negative (-0.13554) and statistically significant ($p < 0.001$). This result indicates that as the level of friendliness towards other cats increases, the probability of a cat belonging to a particular gender decreases
- Intercept (Constant Term): The constant term is not significant ($p = 0.409$). This indicates that without the independent variables there is no significant difference in the baseline probability of the cats' gender.

```
log_model3 <- glm(Cat_sex ~ Personality19_Fearful_other_cats +
Personality22_Friendly_other_cats + Personality21_Greedy +
Personality24_Dominant + Country,
                  family = binomial(link = "logit"),
                  data = cats_data)
summary(log_model3)
```

```
Call:
glm(formula = Cat_sex ~ Personality19_Fearful_other_cats +
Personality22_Friendly_other_cats +
    Personality21_Greedy + Personality24_Dominant + Country,
    family = binomial(link = "logit"), data = cats_data)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.73951    0.19822   3.731 0.000191 ***
Personality19_Fearful_other_cats  0.14809    0.02343   6.320 2.62e-10 ***
Personality22_Friendly_other_cats -0.17879    0.02456  -7.280 3.35e-13 ***
Personality21_Greedy             -0.07849    0.02115  -3.710 0.000207 ***
Personality24_Dominant           -0.10414    0.02439  -4.270 1.96e-05 ***
CountryNewZealand                 0.28270    0.08076   3.501 0.000464 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3831.7  on 2763  degrees of freedom
Residual deviance: 3635.3  on 2758  degrees of freedom
AIC: 3647.3

Number of Fisher Scoring iterations: 4
```

General Interpretation of the Model:

- Personality19_Fearful_other_cats: Significant ($p < 0.001$). The higher the trait "cowardice", the higher the probability that the cat belongs to the specific gender.
- Personality22_Friendly_other_cats Significant ($p < 0.001$). The higher the "friendly" trait, the lower the probability that the cat belongs to the specific gender.
- Personality21_Greedy: Significant ($p < 0.001$). As the trait "greed" increases, the probability of the cat belonging to the specific gender decreases.
- Personality24_Dominant: Significant ($p < 0.001$). The higher the "Dominance" trait, the lower the probability that the cat belongs to the specific gender.
- CountryNewZealand: Significant ($p < 0.001$). The sex probability of cats in New Zealand is positively influenced.
- Intercept (Constant Term): The constant term is significant. The basic probability is positively affected.

```
# GLM: Probit Regression - modeling cat_sex with
Personality19_Fearful_other_cats, Personality22_Friendly_other_cats,
Personality21_Greedy, Personality24_Dominant and Country
probit_model3 <- glm(Cat_sex ~ Personality19_Fearful_other_cats +
Personality22_Friendly_other_cats + Personality21_Greedy +
Personality24_Dominant + Country,
                  family = binomial(link = "probit"),
                  data = cats_data)

summary(probit_model3)
```

```
Call:
glm(formula = Cat_sex ~ Personality19_Fearful_other_cats + 
Personality22_Friendly_other_cats + 
    Personality21_Greedy + Personality24_Dominant + Country, 
    family = binomial(link = "probit"), data = cats_data)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      0.45287    0.12225   3.704 0.000212 ***
Personality19_Fearful_other_cats 0.09176    0.01445   6.352 2.13e-10 ***
Personality22_Friendly_other_cats -0.10964    0.01507  -7.277 3.41e-13 ***
Personality21_Greedy             -0.04871    0.01304  -3.736 0.000187 ***
Personality24_Dominant           -0.06387    0.01501  -4.256 2.08e-05 ***
CountryNewZealand                 0.17207    0.04988   3.449 0.000562 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3831.7  on 2763  degrees of freedom
Residual deviance: 3635.6  on 2758  degrees of freedom
AIC: 3647.6

Number of Fisher Scoring iterations: 4
```

The log_model can be made similar to the log_model3 model only for comparison in terms of AIC. AIC is close to log_model3 model.

General Interpretation of the Model:

- Personality19_Fearful_other_cats: The coefficient is positive (0.18278) and statistically significant ($p < 0.001$). This result indicates that the higher the level of fear of other cats, the higher the probability that the cat belongs to a particular gender.
- Personality22_Friendly_other_cats Significant ($p < 0.001$). The higher the "friendly" trait, the lower the probability that the cat belongs to the specific gender.
- Personality21_Greedy: Significant ($p < 0.001$). As the trait "greed" increases, the probability of the cat belonging to the specific gender decreases.
- Personality24_Dominant: Significant ($p < 0.001$). The higher the "Dominance" trait, the lower the probability that the cat belongs to the specific gender.
- CountryNewZealand: Significant ($p < 0.001$). The sex probability of cats in New Zealand is positively influenced.
- Intercept (Constant Term): The constant term is significant. The basic probability is positively affected.

```
> AIC(log_model1)
[1] 3771.5
> AIC(log_model2)
[1] 3701.449
> AIC(log_model3)
[1] 3647.324
> AIC(probit_model3)
[1] 3647.646
```

Model	Residual Deviance	AIC	Explanation
log_model1	3763.5	3771.5	Contains 3 variables (Shy, Smart, Bullying)
log_model2	3695.4	3701.4	Contains 3 variables (Shy, Calm, Bullying)
log_model3	3635.3	3647.3	5 variables (Fearful, Friendly, Greedy, Dominant, Country)
probit_model3	3635.6	3647.6	Same variables (using Probit model)

Best Performing Model:

- log_model3 and probit_model3 with the lowest Residual Deviance and AIC values perform better.
- The difference between log_model3 (Logit model) and probit_model3 (Probit model) is very small. However, log_model3 is preferable as its AIC is slightly lower.