# 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)

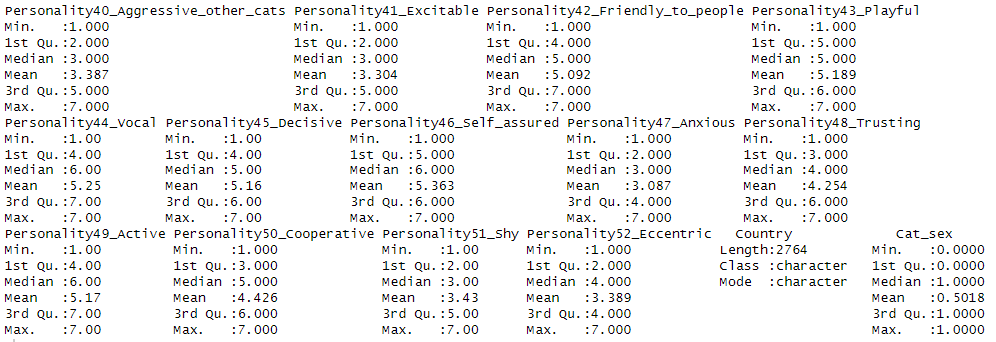
# Missing value check

missing\_values <- sum(is.na(cat\_data))

missing\_values

[1] 0

summary(cats\_data)



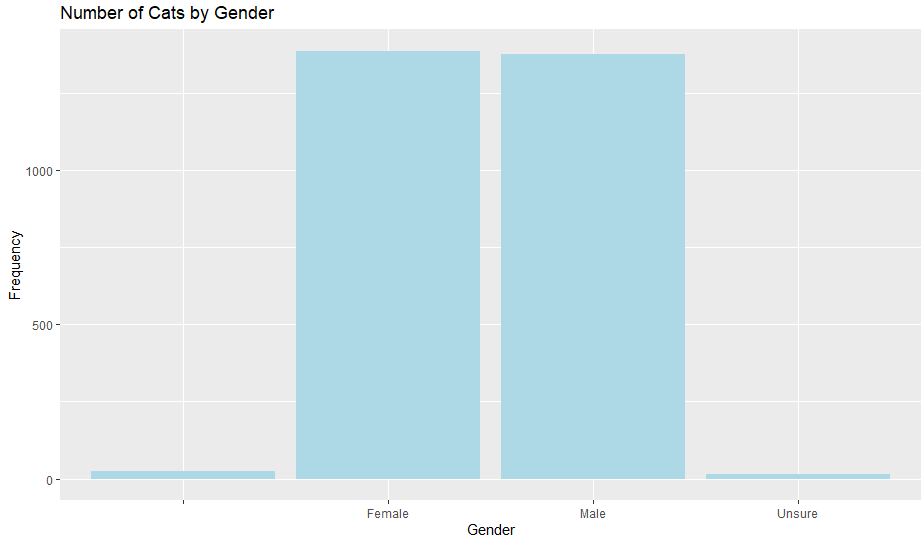
Categorical variables such as Cat\_sex and Country,Considering the overall data, it can be said that the majority of personality traits do not deviate from the normal distribution and generally show a spread between 1 and 7.

# 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")

1. 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

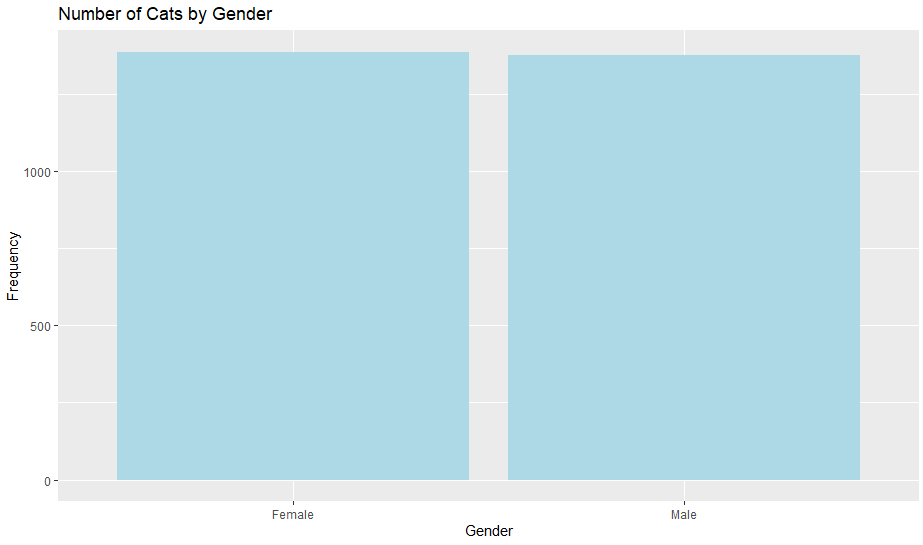
1387 1377

# Visualize the number of cats by gender

ggplot(cats\_data, aes(x = Cat\_sex)) +

geom\_bar(fill = "lightblue") +

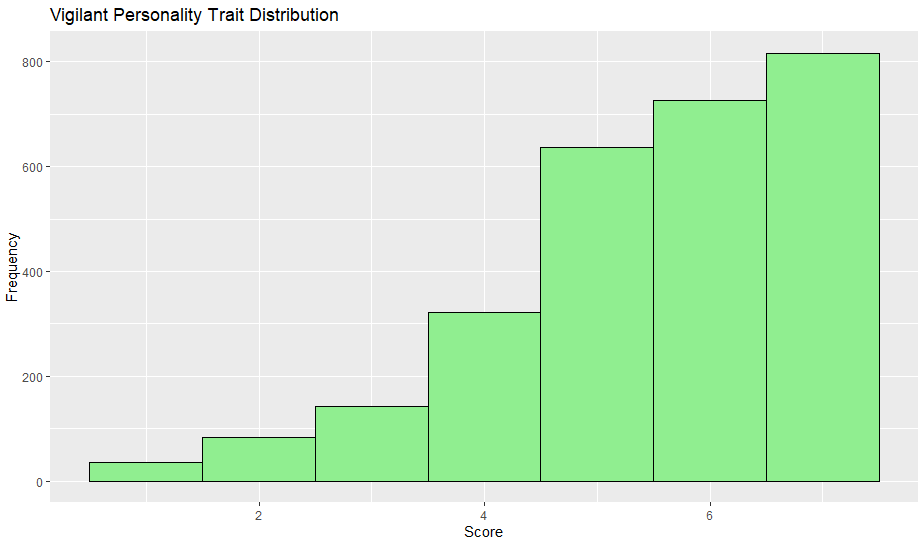
labs(title = "Number of Cats by Gender", x = "Gender", y = "Frequency")

1.  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")

1. 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 ≈ 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.