Preference of Dog or Cat based on Gender

Ed Dungo

December 21, 2023

Packages

```
if (!require("readxl")) install.packages("readxl")
if (!require("dplyr")) install.packages("dplyr")
if (!require("ggplot2")) install.packages("ggplot2")
```

Library

```
library(writexl)
library(dplyr)
library(ggplot2)
```

Import Data

```
data <- read_excel("WSU Class Survey.xlsx")
```

Check Data

```
# get a preview of the data
head(data)
```

```
## # A tibble: 6 x 2
##   ‘presumed gender‘ ‘dog or cat?‘
##   <chr>             <chr>
## 1 m                 dog
## 2 m                 dog
## 3 f                 dog
## 4 m                 dog
## 5 m                 dog
## 6 m                 dog
```

```r
# get the dimensions of the data
dim(data)
```

```
## [1] 201   2
```

```r
# check the values, which is cat and dog, in the data
table(data$`presumed gender`)
```

```
## 
##              f              m section_cutoff 
##            100            100              1
```

```r
table(data$`dog or cat?`)
```

```
## 
##            cat            dog section_cutoff 
##             79            121              1
```

Clean Data

```r
# remove "section_cutoff" from both variables (columns)
data <- data %>% filter_all(all_vars(. != "section_cutoff"))

# rename values
data <-
  data %>% mutate(`presumed gender` =
                    recode(`presumed gender`, "m" = "Male", "f" = "Female"))
data <-
  data %>% mutate(`dog or cat?` =
                    recode(`dog or cat?`, "dog" = "Dog", "cat" = "Cat"))

# rename columns
colnames(data) <- c("Presumed Gender", "Animal")
```

Analyze Data

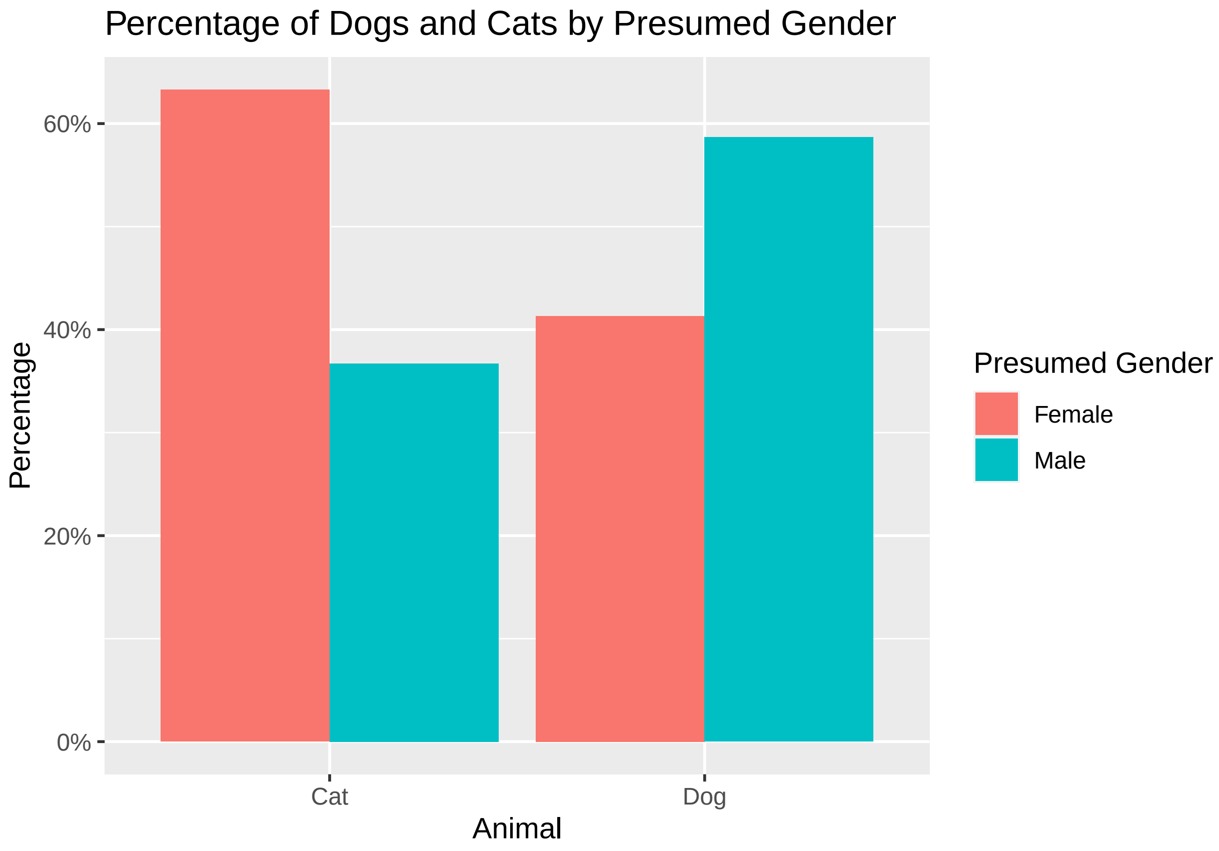
```r
# calculate the percentages
percentages <- data %>%
  group_by(Animal, `Presumed Gender`) %>%
  summarise(n = n()) %>%
  mutate(percent = n / sum(n))
```

```
## ‘summarise()‘ has grouped output by ’Animal’. You can override using the
## ‘.groups‘ argument.
```

```
print(percentages)
```

```
## # A tibble: 4 x 4
## # Groups:   Animal [2]
##   Animal ‘Presumed Gender‘     n percent
##   <chr>  <chr>             <int>   <dbl>
## 1 Cat    Female               50   0.633
## 2 Cat    Male                 29   0.367
## 3 Dog    Female               50   0.413
## 4 Dog    Male                 71   0.587
```

```
# exploratory data analysis with visualization
ggplot(percentages, aes(x = Animal, y = percent, fill = `Presumed Gender`)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Percentage of Dogs and Cats by Presumed Gender",
       x = "Animal",
       y = "Percentage")
```



```
# perform the Chi-Square Test of Independence
chi_square_result <- chisq.test(table(data$`Presumed Gender`, data$Animal))
print(chi_square_result)
```

```
##
```

```
##  Pearson’s Chi-squared test with Yates’ continuity correction
##
## data:  table(data$‘Presumed Gender‘, data$Animal)
## X-squared = 8.3691, df = 1, p-value = 0.003817
```

```
# convert categorical variables to factors
data$`Presumed Gender` <- as.factor(data$`Presumed Gender`)
data$Animal <- as.factor(data$Animal)

# Logistic Regression Model
# 'Animal' is the dependent/response variable
# 'Presumed Gender' is the independent/predictor variable
model <-
  glm(Animal ~ `Presumed Gender`,
      family = binomial(link = "logit"), data = data)
print(summary(model))
```

```
##
## Call:
## glm(formula = Animal ~ ‘Presumed Gender‘, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.697e-15  2.000e-01   0.000  1.00000
## ‘Presumed Gender‘Male 8.954e-01  2.976e-01   3.009  0.00262 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.37  on 199  degrees of freedom
## Residual deviance: 259.06  on 198  degrees of freedom
## AIC: 263.06
##
## Number of Fisher Scoring iterations: 4
```

Export Cleaned Data

```
# export the data frame to a CSV file
write.csv(data, "wsu_class_survey.csv", row.names = FALSE)
```