Does Gender Affect Preference for Dogs or Cats?

An Analysis of Pet Preferences Among College Students

Ed Dungo

December 21, 2023

Overview

The survey aimed to determine whether a person's gender affects their preference for dogs or cats. The gender of the participants was determined based on their Bitmoji and Snapchat name. The original survey had 225 votes for dogs and 154 votes for cats. However, Snapchat limited the display of the results, so only 121 votes for dogs and 79 votes for cats were visible. Note that the respondents could have been from any college graduating class between 2021 and 2028.
Note: If you'd like more information on the original data, please see the link below under "Data Source" and download the Excel file.

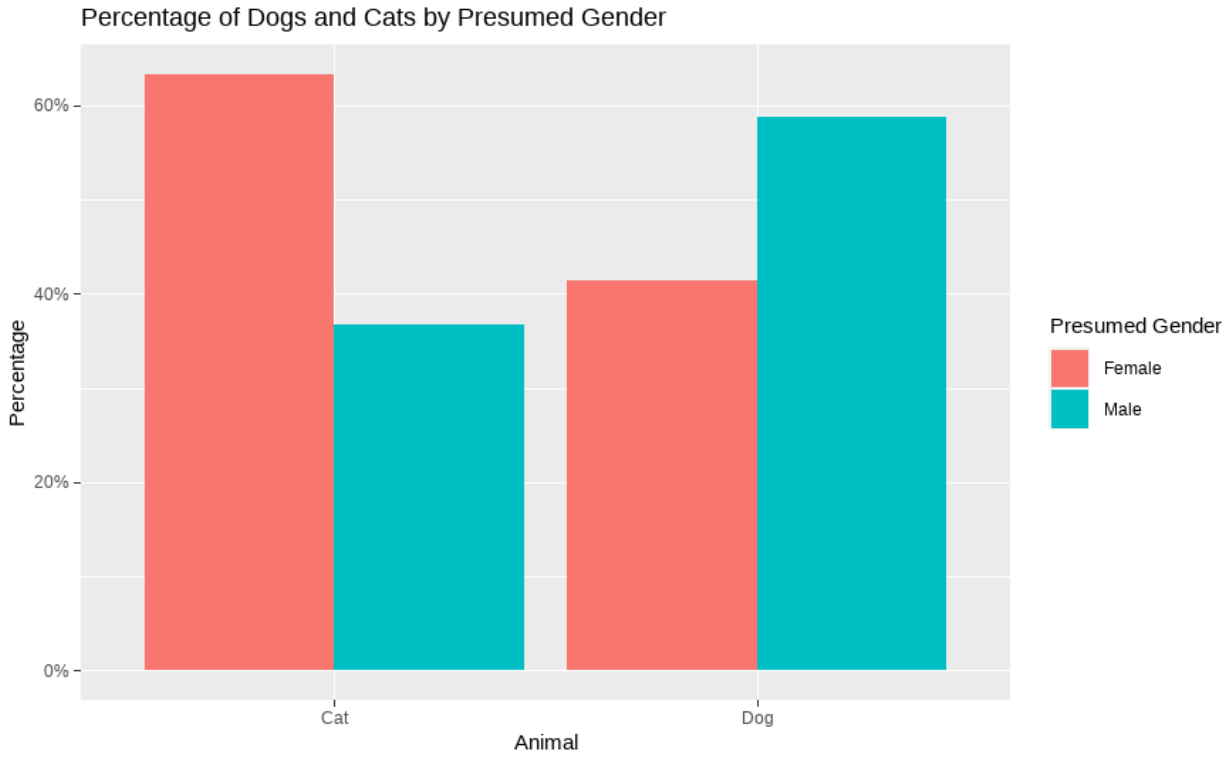
Data Source

Survey Results from Snapchat Class of 2024 Story
For a link of the detailed dataset, see link below:
https://github.com/nagnarg/wsu_animal_survey/blob/main/WSU%20Class%20Survey.xlsx

Data Description

Female: presumed WSU female undergraduate/graduate student
Male: presumed WSU female undergraduate/graduate student
Dog: a domesticated carnivorous mammal that typically has a long snout, an acute sense of smell, nonretractable claws, and a barking, howling, or whining voice.
Cat: a small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.

Exploratory Analysis



The bar chart above shows that 63.29% of females and 36.71% of males chose cats, while 41.32% of females and 58.68% of males chose dogs. These results suggest that there's a strong preference for cats among females and dogs among males.

Hypothesis Testing

```
        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
```

Using Chi-square Test of Independence, this determines if there's a significant association between the two categorical variables. As you can see, the preference for dog or car isn't independent of gender. Rather, there seems to be a relationship between these two variables. This suggests that there's a meaningful relationship between a person's gender and their preference for dogs or cats. However, it's important to note that while this test indicates a significant association, it doesn't imply causation. Moreover, the p-value of 0.003817 indicates that there's only about a 0.38% chance that the observed association is due to random variation in the sample.

Logistic Regression

```
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
```

The code

```
glm(formula = Animal ~ `Presumed Gender`, family = binomial(link = "logit"),
    data = data)
```

stands for

$$\text{Animal} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \times \text{Presumed Gender})}$$

where β_0 is the intercept, β_1 is the coefficient for the independent variable (Presumed Gender), and exp is the exponential function.

The result shows the p-value for male is 0.00262, which is less than 0.05, indicating that the difference in pet preference between males and females is statistically significant.

Probabilities

Using the coefficients from the logistic regression model, I can calculate the probabilities for pet preferences. The model suggests that females have an equal probability (50%) of preferring cats or dogs. For males, the model predicts a higher probability of preferring cats (approximately 71%), which implies about a 29% probability of preferring dogs. Note that this is based on the sample collected from the survey, not the entire population of WSU students.

Conclusion

Using hypothesis testing and logistic regression regression, I can conclude that preference for dogs or cats matters when it comes to gender. Although, this has been proven with just a few observable data, adding more variables into play such as socioeconomic background could drastically provide more insights to the results. This can alter the results for the probabilities as well.

Codes

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
```

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

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

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

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

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

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

Clean Data

```
# 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

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

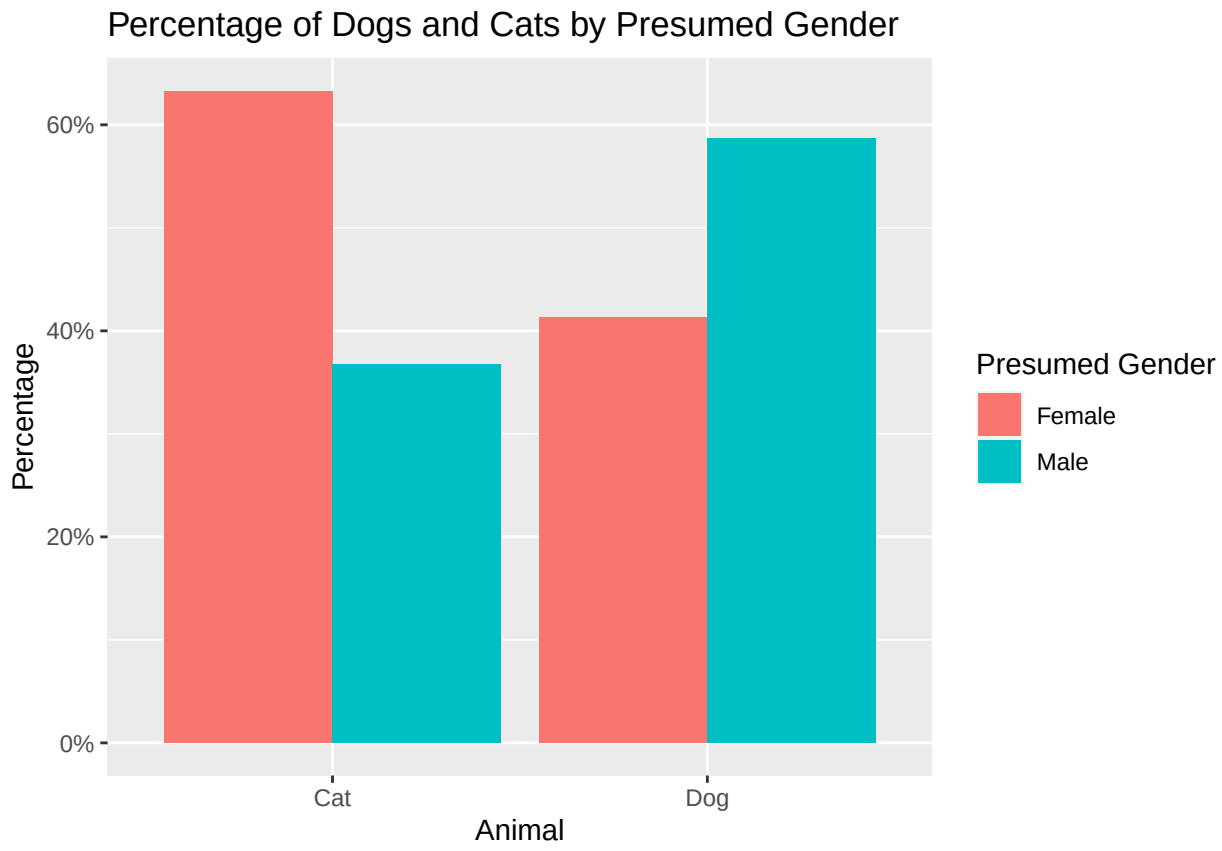
```
## `summarise()` has grouped output by 'Animal'. You can override using the
## `.groups` argument.
```

```
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
```

```
# coefficients from the logistic regression model
intercept <- 0
coef_male <- 0.8954

# calculate the probability of preferring cats for females and males
# For females (GenderMale = 0)
prob_female_cat <- 1 / (1 + exp(-(intercept + coef_male * 0)))
# For males (GenderMale = 1)
prob_male_cat <- 1 / (1 + exp(-(intercept + coef_male * 1)))
print(prob_female_cat)
```

```
## [1] 0.5
```

```
print(prob_male_cat)
```

```
## [1] 0.7100033
```

```
# calculate the probability of preferring dog for females and males
prob_female_cat <- 1 / (1 + exp(-(intercept + coef_male * 0)))
prob_female_dog <- 1 - prob_female_cat

prob_male_cat <- 1 / (1 + exp(-(intercept + coef_male * 1)))
prob_male_dog <- 1 - prob_male_cat

print(prob_female_dog)
```

```
## [1] 0.5
```

```r
print(prob_male_dog)
```

```
## [1] 0.2899967
```

Export Cleaned Data

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