

Peligro Canino: Análisis de las Mordidas de Perro en Nueva York

Carmen Plata Fernández

Noviembre de 2024

Introducción

Las mordidas de perro en una ciudad tan diversa y poblada como Nueva York representan no solo un problema de salud pública, sino también una ventana para explorar patrones de comportamiento canino, factores de riesgo y la influencia de las políticas locales. En este análisis, investigamos la incidencia de mordidas de perro en diferentes distritos de Nueva York, considerando variables como raza, género y ubicación para identificar tendencias y posibles causas subyacentes. Con el aumento de la población de mascotas y la convivencia en espacios urbanos, comprender estos incidentes resulta esencial para prevenir conflictos y promover una coexistencia segura entre humanos y perros. Este estudio busca no solo examinar los datos de mordidas de perro en Nueva York, sino también proporcionar recomendaciones informadas para mejorar la seguridad y fomentar la tenencia responsable de mascotas.

Base de datos

Este conjunto de datos obtenido de la página: [*Mordidas de perro en NYC*](#) contiene registros de incidentes de mordeduras de perro en New York, con información detallada sobre el perro, el lugar y las circunstancias de cada mordedura. A continuación se ofrece una visión general de la estructura del conjunto de datos y los atributos clave:

“UniqueID”: Identificador único para cada registro de mordeduras de perro.

“DateOfBite”: Fecha en la que se produjo la mordedura de perro.

“Species”: La especie implicada en la mordedura.

“Breed”: La raza del perro implicada en la mordedura.

“Age”: Edad del perro implicado en la mordedura (años).

“Gender”: El sexo del perro implicado.

“SpayNeuter”: Indica si el perro estaba esterilizado o no en el momento de la mordedura.

“Borough”: Distrito de Nueva York en el que se produjo la mordedura del perro.

“ZipCode”: Código postal del lugar donde se produjo la mordedura de perro.

Análisis

Presentación de los datos

Abrimos las librerías que usaremos:

```

library(readr) #Para abrir los datos.

library(dplyr) #Proporciona herramientas para manipular los datos.

## 
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

library(ggplot2) #Creación de gráficos.

library(aod) #Herramientas para la regresión logística.

library(car) #Análisis de datos y modelado estadístico (estudio de la multicolinealidad).

## Cargando paquete requerido: carData

## 
## Adjuntando el paquete: 'car'

## The following object is masked from 'package:dplyr':
## 
##     recode

library(ResourceSelection) #Pruebas de bondad de ajuste en regresión logística.

## Warning: package 'ResourceSelection' was built under R
## version 4.4.2

## ResourceSelection 0.3-6 2023-06-27

```

Importamos los datos:

```

dog.bites <- read.csv("Dog_Bites_Data.csv")
dog.bites$SpayNeuter<-as.factor(dog.bites$SpayNeuter)
dog.bites$Gender<-as.factor(dog.bites$Gender)
dog.bites$Breed<-as.factor(dog.bites$Breed)
dog.bites$Borough<-as.factor(dog.bites$Borough)
summary(dog.bites)

##      UniqueID      DateOfBite      Species
##  Min.    : 1  Length:22663      Length:22663
##  1st Qu.: 2834 Class :character  Class :character

```

```

## Median : 5666 Mode :character Mode :character
## Mean   : 5715
## 3rd Qu.: 8499
## Max.   :12383
##
##          Breed      Age      Gender
## Pit Bull     : 4004 Length:22663    F: 3389
## UNKNOWN     : 2349 Class :character M: 8739
##           : 2217 Mode  :character U:10535
## Shih Tzu     :  731
## Chihuahua    :  646
## German Shepherd: 622
## (Other)      :12094
## SpayNeuter    Borough ZipCode
## false:16787 Bronx     :3782 Length:22663
## true  : 5876 Brooklyn :4985 Class :character
##           Manhattan :5270 Mode  :character
##           Other     : 981
##           Queens    :5773
##           Staten Island:1872
##

```

```
head(dog.bites)
```

```

## UniqueID      DateOfBite Species      Breed Age Gender
## 1            1 January 01 2018     DOG UNKNOWN      U
## 2            2 January 04 2018     DOG UNKNOWN      U
## 3            3 January 06 2018     DOG Pit Bull      U
## 4            4 January 08 2018     DOG Mixed/Other   4     M
## 5            5 January 09 2018     DOG Pit Bull      U
## 6            6 January 03 2018     DOG BASENJI     4Y     M
## SpayNeuter  Borough ZipCode
## 1       false Brooklyn 11220
## 2       false Brooklyn 11224
## 3       false Brooklyn 11231
## 4       false Brooklyn 11224
## 5       false Brooklyn 11231
## 6       false Brooklyn 11231

```

Vamos a filtrar los datos eliminando las razas desconocidas para poder trabajar con ellos más fácilmente:

```

if ("Breed" %in% colnames(dog.bites)) {
  dog.bites <- dog.bites %>%
    mutate(Breed = tolower(Breed)) %>%
    filter(!is.na(Breed) & Breed != "unknown" & Breed != "")
}

head(dog.bites)
} else {
  cat("La columna 'Breed' no se encuentra en el dataframe.\n")
}

```

```

## UniqueID      DateOfBite Species
## 1            3 January 06 2018     DOG

```

```

## 2      4 January 08 2018      DOG
## 3      5 January 09 2018      DOG
## 4      6 January 03 2018      DOG
## 5      8 January 03 2018      DOG
## 6      9 January 04 2018      DOG
##                                Breed Age Gender
## 1          pit bull        U
## 2      mixed/other        4      M
## 3          pit bull        U
## 4          basenji     4Y      M
## 5          pit bull        U
## 6 american pit bull mix / pit bull mix 5Y      M
##   SpayNeuter Borough ZipCode
## 1    false Brooklyn 11224
## 2    false Brooklyn 11231
## 3    false Brooklyn 11224
## 4    false Brooklyn 11231
## 5    false Brooklyn 11233
## 6    false Brooklyn 11235

```

Es una base demasiado grande para trabajar con ella.

Para solucionarlo vamos a disminuir nuestra muestra eligiendo las 5 razas con mayor porcentaje de mordida:

```

top_breeds <- dog.bites %>%
  count(Breed) %>%
  mutate(percent_bites = (n / sum(n)) * 100) %>%
  arrange(desc(percent_bites)) %>%
  slice(1:5)
top_breeds

```

```

##           Breed   n percent_bites
## 1      pit bull 4013    22.174946
## 2      shih tzu  732     4.044869
## 3      chihuahua 647     3.575178
## 4 german shepherd 624     3.448085
## 5      mixed/other 559     3.088910

```

Así nuestra base de datos quedaría tal que:

```

filtered_dog_bites <- dog.bites %>%
  filter(
    Breed %in% c("pit bull", "shih tzu", "american pit bull mix / pit bull mix",
    "german shepherd", "mixed/other"))

head(filtered_dog_bites)

```

```

##   UniqueID      DateOfBite Species
## 1      3 January 06 2018      DOG
## 2      4 January 08 2018      DOG
## 3      5 January 09 2018      DOG
## 4      8 January 03 2018      DOG
## 5      9 January 04 2018      DOG

```

```

## 6      14 January 01 2018      DOG
##                                Breed Age Gender
## 1                      pit bull      U
## 2                  mixed/other     4      M
## 3                      pit bull      U
## 4                      pit bull      U
## 5 american pit bull mix / pit bull mix  5Y      M
## 6                      pit bull      U
##   SpayNeuter Borough ZipCode
## 1    false Brooklyn 11224
## 2    false Brooklyn 11231
## 3    false Brooklyn 11224
## 4    false Brooklyn 11233
## 5    false Brooklyn 11235
## 6    false Brooklyn 11220

```

Estudio descriptivo

Vamos a hacer primero un estudio descriptivo de las variables para ver qué destacamos entre los distintos sucesos:

Razas

```

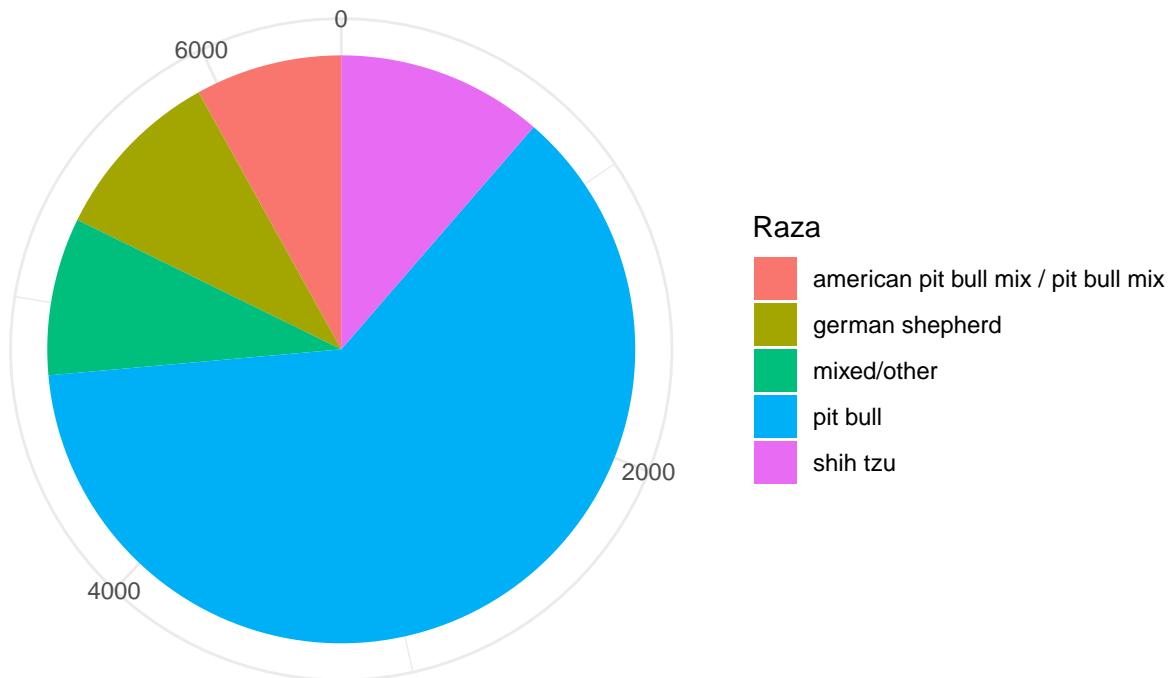
top_breeds <- names(sort(table(filtered_dog_bites$Breed), decreasing = TRUE))
top_breeds_data <- filtered_dog_bites[filtered_dog_bites$Breed %in% top_breeds, ]

top_breeds_counts <- as.data.frame(table(top_breeds_data$Breed))
colnames(top_breeds_counts) <- c("Breed", "Count")

ggplot(top_breeds_counts, aes(x = "", y = Count, fill = Breed)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_minimal() +
  labs(
    title = "Las 5 razas con más incidentes de mordeduras",
    fill = "Raza",
    x = NULL,
    y = NULL
  ) +
  theme(plot.title = element_text(hjust = 0.5))

```

Las 5 razas con más incidentes de mordeduras

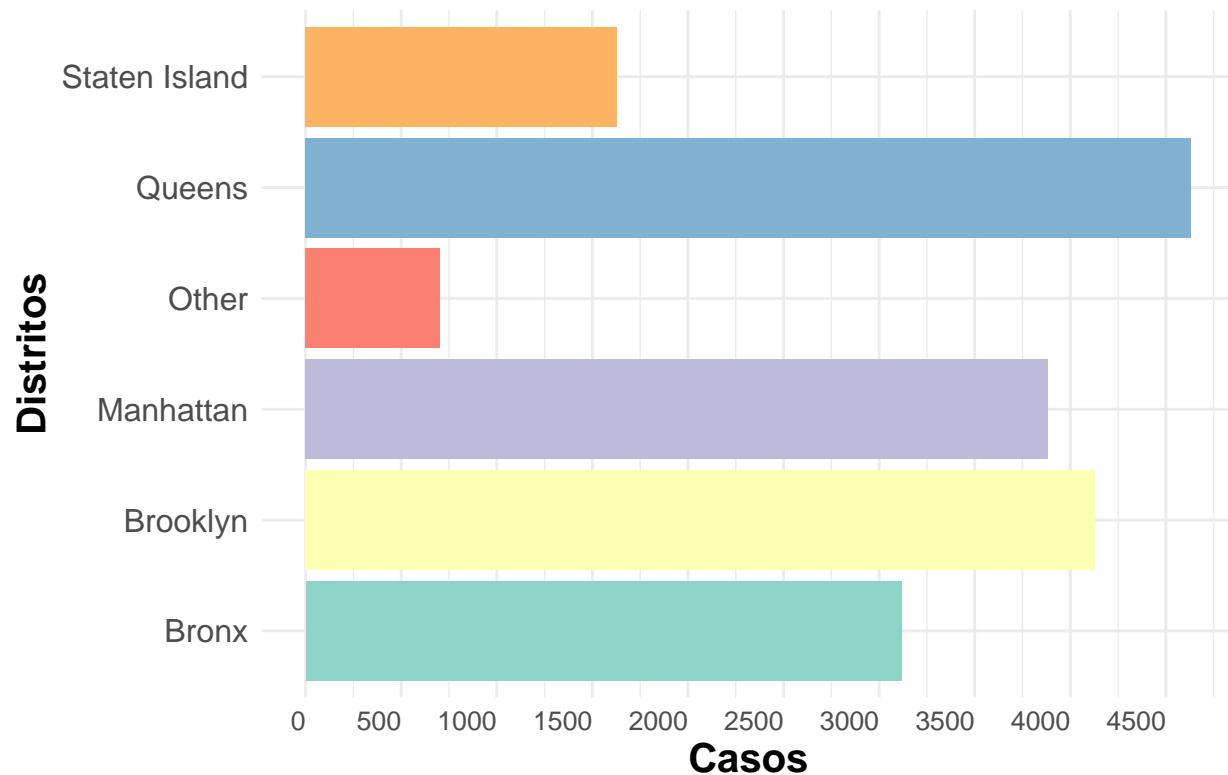


Vemos que claramente la raza más abundante es la de pit bull.

Distritos

```
ggplot(dog.bites, aes(x = Borough, fill = Borough)) +  
  geom_bar() +  
  coord_flip() +  
  ylab("Casos") +  
  xlab("Distritos") +  
  ggtitle("Mordeduras de perro por distritos") +  
  scale_y_continuous(breaks = seq(0, 5000, by = 500)) +  
  scale_fill_brewer(palette = "Set3") +  
  theme_minimal() +  
  theme(axis.text = element_text(size = 12),  
        axis.title = element_text(size = 15, face = "bold"),  
        plot.title = element_text(size = 20, hjust = 0.5, face = "bold"),  
        legend.position = "none",  
        axis.text.x = element_text(size = 10, hjust = 1, vjust = 1))
```

Mordeduras de perro por distritos



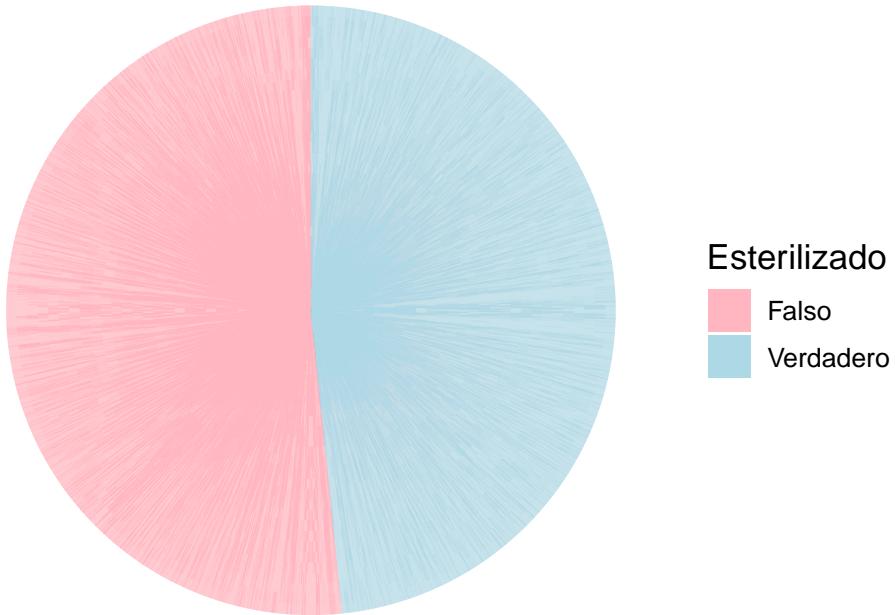
La mayoría de los hechos ocurren en Queens. Brooklyn está bastante cerca.

Esterilización

```
dog.bites$SpayNeuter_ohe = ifelse(dog.bites$SpayNeuter == TRUE, 1, 0)

ggplot(dog.bites, aes(x="", y=factor(SpayNeuter), fill=factor(SpayNeuter))) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  ggtitle("Perros esterilizados") +
  theme(plot.title = element_text(hjust = 0.5, size = 20, face = "bold"),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 13)) +
  scale_fill_manual(values=c("true" = "#ADD8E6", "false" = "#FFB6C1"),
                    labels = c("Falso", "Verdadero")) +
  labs(fill = "Esterilizado")
```

Perros esterilizados



Podemos ver que la mayoría de las mordeduras de perro provienen de perros que no están castrados. Esto también podría ver con que suponemos que los perros que están castrados en su mayoría tienen dueño.

Relacionamos las tres variables

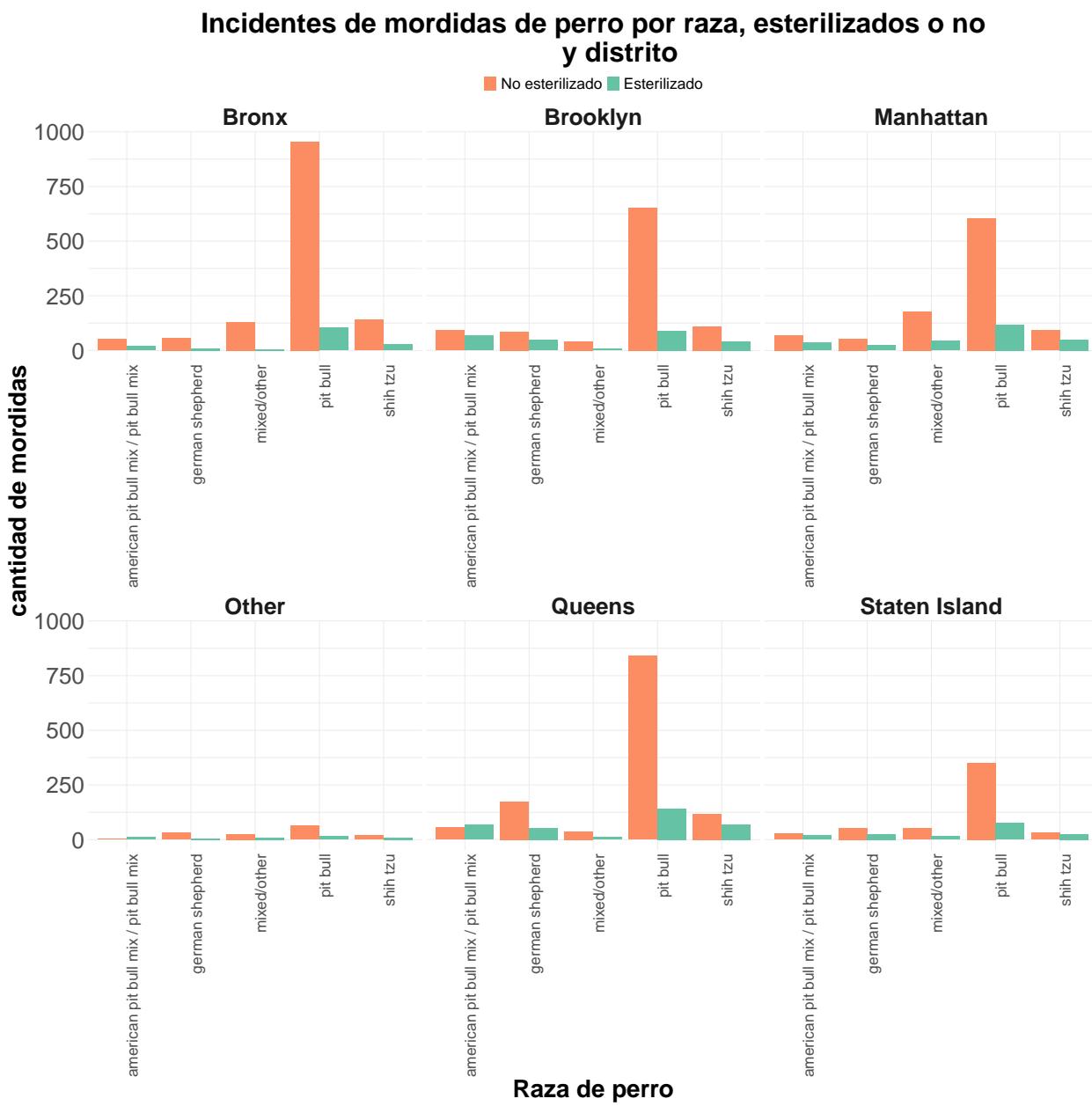
```
filtered_dog_bites$SpayNeuter <- as.factor(filtered_dog_bites$SpayNeuter)

ggplot(filtered_dog_bites, aes(x = Breed, fill = SpayNeuter)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ Borough, scales = "free_x") +
  labs(x = "Raza de perro", y = "cantidad de mordidas",
       title = "Incidentes de mordidas de perro por raza, esterilizados o no
y distrito") + scale_fill_manual(values = c("true" = "#66c2a5",
                                              "false" = "#fc8d62"),
                                    labels = c("true" = "Esterilizado", "false" = "No esterilizado")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 20),
        axis.text.y = element_text(size = 30),
        axis.title = element_text(size = 32, face = "bold"),
        plot.title = element_text(size = 36, hjust = 0.5, face = "bold"))
```

```

legend.title = element_blank(),
legend.text = element_text(size = 20),
legend.position = "top",
strip.text = element_text(size = 30, face = "bold") )

```



Podemos ver cosas muy interesantes al comparar estas tres variables:

En todos los distritos la raza que realiza más mordeduras son los pitbulls. Esto puede deberse a que hay más propietarios de perros pitbull en Nueva York.

Claramente, la mayoría de perros que muerden no están castrados.

Conclusión del análisis descriptivo

Viendo que los perros que menos muerden son aquellos castrados, observaremos qué variables de las que tenemos pueden ser influyentes para que el perro sea esterilizado.

Esto podría ser muy interesante para poder evitar mordidas en un futuro.

Lo estudiaremos a través de una regresión logística.

Regresión logística

Qué es el análisis de regresión logística

El análisis de regresión logística es una técnica estadística utilizada para modelar la relación entre una variable dependiente categórica y una o más variables independientes. Este modelo es particularmente útil en situaciones donde la variable de respuesta es binaria, es decir, toma solo dos valores, como éxito/fallo, sí/no, o presencia/ausencia. La regresión logística permite predecir la probabilidad de que ocurra uno de estos eventos en función de las variables explicativas.

El modelo de regresión logística se basa en la función logística, que tiene la forma:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

donde p representa la probabilidad de que ocurra el evento de interés, e es la base del logaritmo natural, β_0 es el intercepto del modelo y $\beta_1, \beta_2, \dots, \beta_k$ son los coeficientes asociados a las variables independientes X_1, X_2, \dots, X_k .

Los supuestos fundamentales de la regresión logística incluyen:

- **Relación lineal entre las variables independientes y la log-odds:** Aunque no se requiere que las variables independientes tengan una relación lineal con la variable dependiente, sí se necesita que exista una relación lineal entre las variables independientes y el logaritmo de las probabilidades (log-odds) de los resultados.
- **Independencia de las observaciones:** Se asume que las observaciones son independientes entre sí. Esto es crítico para la validez del modelo.
- **No multicolinealidad:** Las variables independientes no deben estar altamente correlacionadas entre sí, ya que esto puede afectar la estimación de los coeficientes y la interpretación del modelo.

La regresión logística se aplica en diversos campos, como la medicina, la economía y las ciencias sociales, donde se busca entender factores que influyen en la ocurrencia de un evento. Por ejemplo, en medicina, puede usarse para predecir la probabilidad de que un paciente desarrolle una enfermedad en función de factores de riesgo como la edad, el índice de masa corporal y los hábitos de vida. En el ámbito empresarial, se utiliza para evaluar la probabilidad de que un cliente realice una compra basada en sus características demográficas y comportamentales.

En este trabajo, realizaremos un análisis exhaustivo de la regresión logística, discutiendo su aplicación práctica, el proceso de ajuste del modelo, la evaluación de los resultados y las interpretaciones de los coeficientes.

Aplicación

Función glm

Probamos ahora a realizar nuestra regresión logística usando la función glm con la familia binomial.

```
logistica<-glm(filtered_dog_bites$SpayNeuter~+
                  filtered_dog_bites$Breed+
                  filtered_dog_bites$Gender+
                  filtered_dog_bites$Borough, data = filtered_dog_bites, family=binomial)
summary(logistica)

##
## Call:
## glm(formula = filtered_dog_bites$SpayNeuter ~ +filtered_dog_bites$Breed +
##      filtered_dog_bites$Gender + filtered_dog_bites$Borough, family = binomial,
##      data = filtered_dog_bites)
##
## Coefficients:
##                               Estimate Std. Error
## (Intercept)             -0.62694   0.13828
## filtered_dog_bites$Breedgerman shepherd -0.26803   0.13965
## filtered_dog_bites$Breedmixed/other     -0.96692   0.15777
## filtered_dog_bites$Breedpit bull       -0.52508   0.10931
## filtered_dog_bites$Breedshih tzu       -0.07842   0.13124
## filtered_dog_bites$GenderM            -0.26881   0.08029
## filtered_dog_bites$GenderU            -4.59435   0.22989
## filtered_dog_bites$BoroughBrooklyn    0.81975   0.12329
## filtered_dog_bites$BoroughManhattan   0.88309   0.12035
## filtered_dog_bites$BoroughOther       0.90896   0.20644
## filtered_dog_bites$BoroughQueens     0.75053   0.11426
## filtered_dog_bites$BoroughStaten Island 0.96530   0.13780
##                               z value Pr(>|z|)
## (Intercept)             -4.534 5.79e-06
## filtered_dog_bites$Breedgerman shepherd -1.919 0.054946
## filtered_dog_bites$Breedmixed/other     -6.129 8.86e-10
## filtered_dog_bites$Breedpit bull       -4.804 1.56e-06
## filtered_dog_bites$Breedshih tzu       -0.598 0.550170
## filtered_dog_bites$GenderM            -3.348 0.000814
## filtered_dog_bites$GenderU            -19.985 < 2e-16
## filtered_dog_bites$BoroughBrooklyn    6.649 2.95e-11
## filtered_dog_bites$BoroughManhattan   7.337 2.18e-13
## filtered_dog_bites$BoroughOther       4.403 1.07e-05
## filtered_dog_bites$BoroughQueens     6.569 5.07e-11
## filtered_dog_bites$BoroughStaten Island 7.005 2.47e-12
##
## (Intercept)                 ***
## filtered_dog_bites$Breedgerman shepherd .
## filtered_dog_bites$Breedmixed/other      ***
## filtered_dog_bites$Breedpit bull        ***
## filtered_dog_bites$Breedshih tzu        .
## filtered_dog_bites$GenderM              ***
## filtered_dog_bites$GenderU              ***
## filtered_dog_bites$BoroughBrooklyn    ***
```

```

## filtered_dog_bites$BoroughManhattan      ***
## filtered_dog_bites$BoroughOther          ***
## filtered_dog_bites$BoroughQueens        ***
## filtered_dog_bites$BoroughStaten Island ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 6319.0  on 6447  degrees of freedom
## Residual deviance: 4434.6  on 6436  degrees of freedom
## AIC: 4458.6
##
## Number of Fisher Scoring iterations: 7

```

Tras haber obtenido los resultados preliminares del modelo de regresión logística, es fundamental realizar una evaluación rigurosa de la validez del modelo. Esto incluye verificar si se cumplen las hipótesis del modelo, la presencia de anomalías, y la posibilidad de simplificar el modelo.

Diagnóstico de los resultados y cumplimiento de hipótesis.

```
vif(logistica)
```

Estudio de la multicolinealidad

```

##                                     GVIF Df GVIF^(1/(2*Df))
## filtered_dog_bites$Breed    1.108649  4       1.012976
## filtered_dog_bites$Gender   1.013430  2       1.003341
## filtered_dog_bites$Borough  1.099675  5       1.009547

```

Los valores de $GVIF^{(1/(2*Df))}$ son todos cercanos a 1, lo que indica que no hay una multicolinealidad fuerte entre las variables. Es decir, las variables Breed, Gender, y Borough no están altamente correlacionadas entre sí, por lo que no hay preocupación por colinealidad en el modelo.

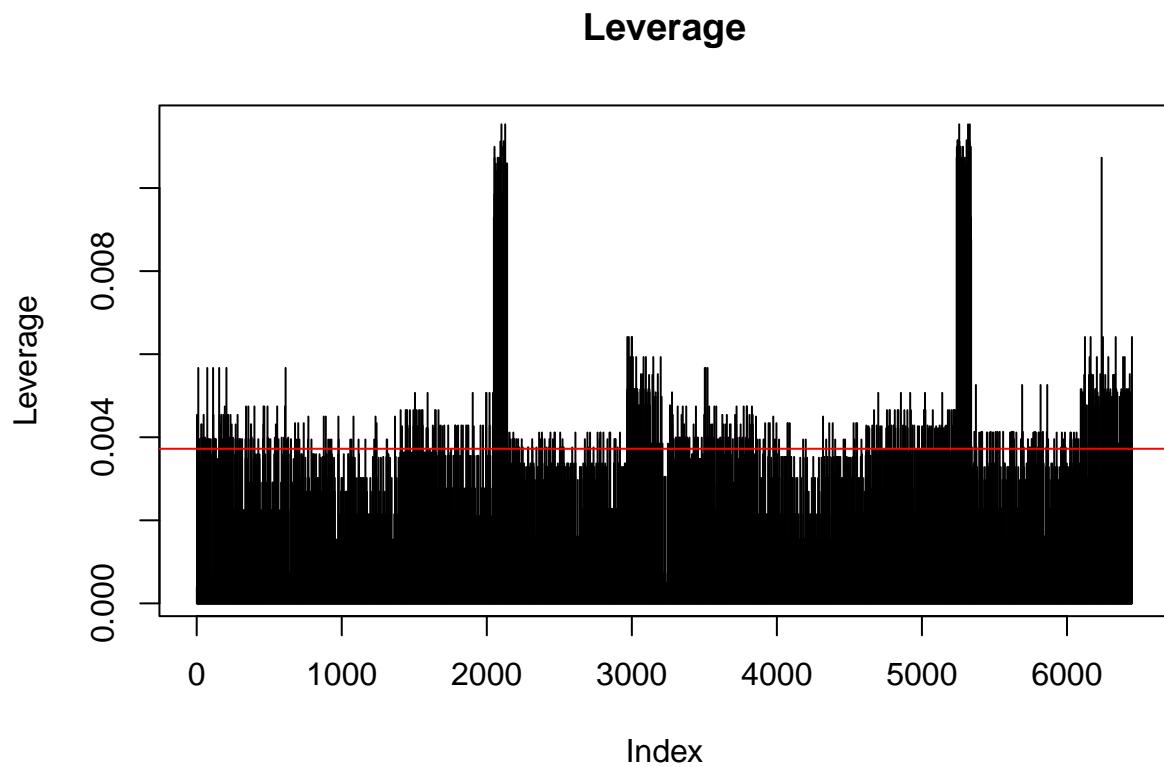
```

leverage <- hatvalues(logistica)

plot(leverage, type="h", main="Leverage", ylab="Leverage")
abline(h = 2 * (length(coefficients(logistica)) / nrow(filtered_dog_bites)), col="red")

```

Diagnóstico de influencia

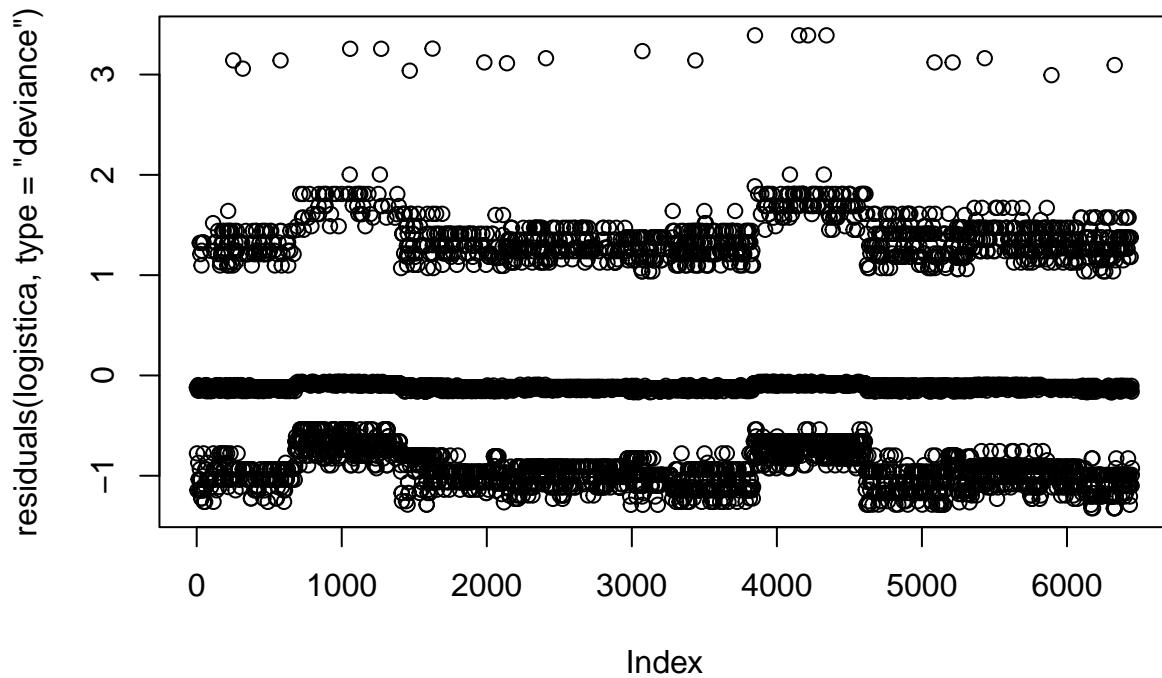


El Leverage mide qué tan lejos están las observaciones de la media de las variables predictoras. Puntos con un leverage alto podrían ser outliers. Vemos que hay algunas observaciones por encima de la linea roja que podrían tener un efecto desproporcionado en el modelo, pero no demasiadas como para producir grandes problemas.

```
plot(residuals(logistica, type = "deviance"), main = "Residuos de Deviance")
```

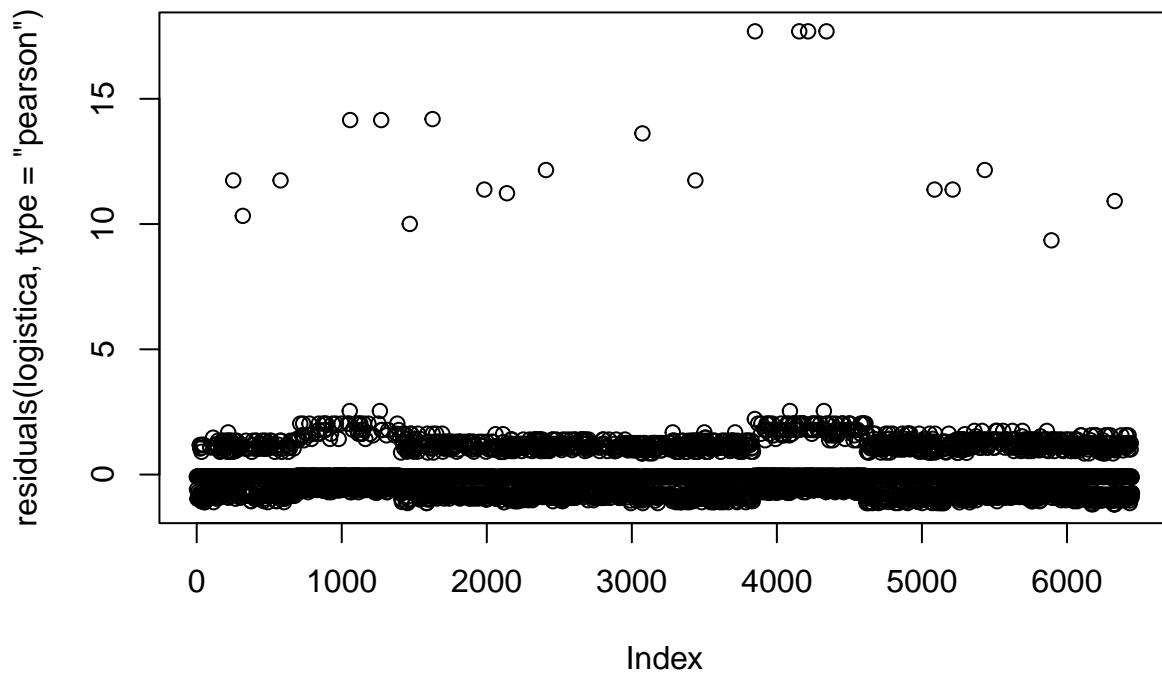
Diagnóstico de residuos.

Residuos de Deviance



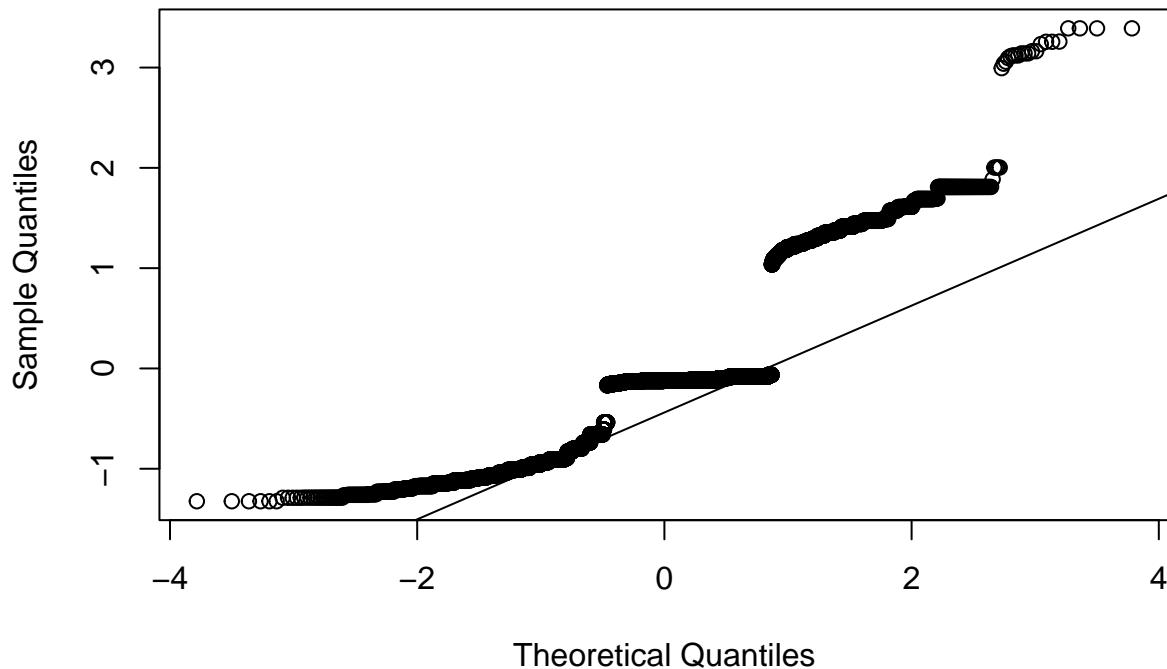
```
plot(residuals(logistica, type = "pearson"), main = "Residuos de Pearson")
```

Residuos de Pearson



```
qqnorm(resid(logistica, type = "deviance"))
qqline(resid(logistica, type = "deviance"))
```

Normal Q-Q Plot



Estos gráficos indican que no existen patrones e irregularidades que puedan darnos problemas. Advertimos algunos outliers.

```
hoslem.test(logistica$y, fitted(logistica))
```

Prueba de bondad de ajuste - Prueba de Hosmer-Lemeshow

```
##  
##  Hosmer and Lemeshow goodness of fit (GOF) test  
##  
##  data: logistica$y, fitted(logistica)  
##  X-squared = 8.569, df = 8, p-value = 0.38
```

Como el valor p es de 0.38, no hay evidencia suficiente para rechazar la hipótesis nula, lo que sugiere que el modelo de regresión logística tiene un buen ajuste a los datos.

En conclusión, el modelo de regresión logística es válido y está bien ajustado a los datos: no hay preocupaciones de multicolinealidad, los residuos no presentan patrones que sugieran incumplimiento de supuestos y la prueba de bondad del ajuste confirma que el modelo es apropiado.

Sin embargo, es recomendable revisar las observaciones y considerar si su exclusión mejora la estabilidad del modelo. También podría ser útil simplificar el modelo eliminando predictores no significativos.

Determinar si los parámetros son significativos

Establecemos los contrastes de Ward, que miden si el parámetro es estadísticamente significativo y por tanto si la variable asociada lo será o no.

Como todas nuestras variables son de varias modalidades se toma una de referencia. Por ejemplo, en la raza toma de referencia la raza pitbull americana mezclada y las compara con esta:

- Raza German Shepherd: p-valor es 0.054946, por lo que la variable es marginalmente significativa.
- Raza mezclada: p-valor es 8.86e-10, por lo que la variable es significativa.
- Raza Pitbull: p-valor es 1.56e-06, por lo que la variable es significativa.
- Raza Shih Tzu: p-valor es 0.550170, por lo que la variable no es significativa.

En el género toma de referencia el género femenino:

- Género Macho: p-valor es 0.000814, por lo que la variable es significativa.
- Género Desconocido: p-valor es <2e-16, por lo que la variable es significativa.

Y en distritos toma de referencia el distrito de Bronx:

- Distrito Brooklyn: p-valor es 2.95e-11, por lo que la variable es significativa.
- Distrito Manhattan: p-valor es 2.18e-13, por lo que la variable es significativa.
- Distrito Other: p-valor es 1.07e-05, por lo que la variable es significativa.
- Distrito Queens: p-valor es 5.07e-11, por lo que la variable es significativa.
- Distrito Staten Island: p-valor es 2.47e-12, por lo que la variable es significativa.

Es cierto que, de forma general, si hay alguna modalidad significativa suele ocurrir que la variable globalmente lo sea, pero puede que haya combinaciones que nos generen problemas.

Para ello se suele contrastar mediante la librería aod que la variable sea significativa globalmente. Hacemos el test de Wald iniciando el orden de la salida glm() cuales son las modalidades implicadas.

Esto lo hacemos también con las demás variables que tienen varias modalidades.

```
wald.test(b = coef(logistica), Sigma = vcov(logistica), Terms = 2:5)
```

```
## Wald test:  
## -----  
##  
## Chi-squared test:  
## X2 = 56.7, df = 4, P(> X2) = 1.4e-11
```

Se contrasta si la variable raza mezcla pitbull es significativa, por lo que aceptamos que es significativa.

```
wald.test(b = coef(logistica), Sigma = vcov(logistica), Terms = 6)
```

```

## Wald test:
## -----
##
## Chi-squared test:
## X2 = 11.2, df = 1, P(> X2) = 0.00081

```

Se contrasta si la variable género femenino es significativa, por lo que aceptamos que es significativa.

```
wald.test(b = coef(logistica), Sigma = vcov(logistica), Terms = 7:11)
```

```

## Wald test:
## -----
##
## Chi-squared test:
## X2 = 457.8, df = 5, P(> X2) = 0.0

```

En la variable distrito, tomamos de referencia Bronx. La variable es significativa.

Ajustamos el modelo

En un principio, la única variable que parece no ser significativa es la raza shih tzu (ya que tiene un p-valor mayor que 0.05 ($p=0.550$)), por lo que la eliminaremos para optimizar nuestro modelo:

```

logistica_simplified <- glm(SpayNeuter ~ Breed + Gender + Borough,
                             data = filtered_dog_bites %>% filter(Breed != "shih tzu"),
                             family = binomial)

summary(logistica_simplified)

```

```

##
## Call:
## glm(formula = SpayNeuter ~ Breed + Gender + Borough, family = binomial,
##      data = filtered_dog_bites %>% filter(Breed != "shih tzu"))
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.51749   0.14498 -3.569 0.000358
## Breedgerman shepherd -0.26173   0.13989 -1.871 0.061345
## Breedmixed/other     -0.94494   0.15833 -5.968 2.40e-09
## Breedpit bull        -0.52296   0.10958 -4.773 1.82e-06
## GenderM              -0.37027   0.08686 -4.263 2.02e-05
## GenderU              -4.64621   0.24213 -19.189 < 2e-16
## BoroughBrooklyn      0.81081   0.13487  6.012 1.83e-09
## BoroughManhattan     0.79689   0.13116  6.076 1.24e-09
## BoroughOther          0.89124   0.22577  3.948 7.89e-05
## BoroughQueens         0.70883   0.12566  5.641 1.69e-08
## BoroughStaten Island  0.86111   0.14853  5.798 6.73e-09
##
## (Intercept) ***  

## Breedgerman shepherd .  

## Breedmixed/other ***

```

```

## Breedpit bull      ***
## GenderM          ***
## GenderU          ***
## BoroughBrooklyn   ***
## BoroughManhattan  ***
## BoroughOther      ***
## BoroughQueens     ***
## BoroughStaten Island ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 5361.9  on 5715  degrees of freedom
## Residual deviance: 3741.5  on 5705  degrees of freedom
## AIC: 3763.5
##
## Number of Fisher Scoring iterations: 7

```

Viendo los resultados obtenidos, surge la duda de si optimizaría el modelo eliminar la variable raza German Shepherd ya que tiene un p-valor de 0.061345, mayor de 0.05 pero ligeramente.

Debido a que su magnitud de efecto (-0.26173) es modesto podría no haber gran variación. Vamos a comprobarlo:

```

logistica_simplified <- glm(SpayNeuter ~ Breed +
                           Gender +
                           Borough,
                           data = filtered_dog_bites %>% filter(Breed != "shih tzu" &
                           Breed != "german shepherd"),
                           family = binomial)

summary(logistica_simplified)

##
## Call:
## glm(formula = SpayNeuter ~ Breed + Gender + Borough, family = binomial,
##      data = filtered_dog_bites %>% filter(Breed != "shih tzu" &
##                                             Breed != "german shepherd"))
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.50288   0.14949 -3.364 0.000768
## Breedmixed/other     -0.95011   0.15918 -5.969 2.39e-09
## Breedpit bull        -0.53297   0.10990 -4.850 1.24e-06
## GenderM              -0.37695   0.09385 -4.016 5.91e-05
## GenderU              -4.63939   0.25641 -18.094 < 2e-16
## BoroughBrooklyn      0.73381   0.14464  5.073 3.91e-07
## BoroughManhattan     0.77020   0.13698  5.623 1.88e-08
## BoroughOther         1.08168   0.24773  4.366 1.26e-05
## BoroughQueens        0.77073   0.13357  5.770 7.91e-09
## BoroughStaten Island 0.81995   0.15851  5.173 2.31e-07
##

```

```

## (Intercept)      ***
## Breedmixed/other ***
## Breedpit bull   ***
## GenderM          ***
## GenderU          ***
## BoroughBrooklyn ***
## BoroughManhattan ***
## BoroughOther     ***
## BoroughQueens    ***
## BoroughStaten Island ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 4605.6 on 5091 degrees of freedom
## Residual deviance: 3190.3 on 5082 degrees of freedom
## AIC: 3210.3
##
## Number of Fisher Scoring iterations: 7

```

Dado que la exclusión de la raza German Shepherd no afecta negativamente el modelo y, de hecho, mejora el AIC¹ considerablemente (de 3763.5 a 3210.3), la eliminación parece merecer la pena.

Interpretación de los parámetros

Estudiaremos si las exponenciales de los parámetros a las que se denominan Odd Ratios:

```

exp(cbind(OR = coef(logistica_simplified), confint(logistica_simplified, level = 0.95)))

## Waiting for profiling to be done...

##                               OR      2.5 %     97.5 %
## (Intercept)      0.604788810 0.450351918 0.80943508
## Breedmixed/other 0.386699833 0.282222027 0.52695292
## Breedpit bull   0.586861108 0.473156161 0.72806541
## GenderM          0.685947329 0.570776334 0.82468097
## GenderU          0.009663601 0.005633568 0.01549062
## BoroughBrooklyn 2.083006138 1.569938540 2.76865292
## BoroughManhattan 2.160188153 1.653677735 2.82990707
## BoroughOther     2.949630408 1.807689903 4.78566858
## BoroughQueens    2.161348556 1.665879763 2.81292049
## BoroughStaten Island 2.270391828 1.663882869 3.09868106

```

Hemos obtenido que:

El intercepto (OR = 0.605): Representa la probabilidad base de esterilización para el grupo de referencia (pitbull mezclado, género femenino, ubicado en el Bronx). Un OR menor a 1 sugiere una menor probabilidad base de esterilización en este grupo.

¹El Akaike Information Criterion (AIC) es una medida estadística utilizada para evaluar y comparar modelos, penalizando la complejidad del modelo para evitar sobreajuste. Un AIC más bajo indica un modelo preferible, ya que sugiere un mejor balance entre ajuste y simplicidad.

Raza mixta u otra (OR = 0.387): Los perros de raza mixta u “otras” tienen una menor probabilidad de esterilización en comparación con los pitbull mezclados, con aproximadamente 0.39 veces las probabilidades de ser esterilizados. El intervalo de confianza (0.2822 - 0.5270) indica que estamos bastante seguros de que el OR es significativamente menor a 1.

Esto puede deberse a que los dueños de perros de razas mixtas u “otras” podrían tener menos interés o acceso a la esterilización, posiblemente porque estos perros son menos costosos y tienen menor demanda de adopción, lo que reduce la presión para esterilizarlos. También podría influir que algunas razas mixtas se perciben como menos propensas a problemas de comportamiento, por lo que los dueños ven la esterilización como menos urgente.

Raza Pit bull (OR = 0.587): Los pitbulls de raza pura también tienen una probabilidad de esterilización menor que los pitbull mezclados, con aproximadamente 0.59 veces las probabilidades de ser esterilizados. El intervalo de confianza (0.4732 - 0.7281) está completamente por debajo de 1, lo que respalda una menor probabilidad de esterilización para esta raza.

Suponemos que los dueños de pitbulls de raza pura podrían preferir no esterilizarlos, ya sea porque consideran que mantenerlos sin esterilizar puede preservar características de la raza o por un estigma cultural. Además, podría haber intentos de reproducción en esta raza debido a su popularidad en ciertos entornos, lo que reduciría la tasa de esterilización.

Género Masculino (OR = 0.686): Los perros machos tienen un 0.69 veces menos probabilidad de ser esterilizados en comparación con las hembras. El intervalo de confianza (0.5708 - 0.8247) indica que este efecto es estadísticamente significativo.

Existe una percepción común de que la esterilización de machos es menos necesaria que la de hembras, debido a que no pueden quedar embarazadas ni directamente causar camadas no deseadas. Los dueños pueden también subestimar el impacto del comportamiento relacionado con la reproducción en machos, como la agresividad o el marcaje, lo cual reduce la tasa de esterilización en perros machos.

Género desconocido (OR = 0.0097): Los perros con género desconocido tienen solo un 0.97% de las probabilidades de ser esterilizados en comparación con las hembras, lo que sugiere una probabilidad extremadamente baja de esterilización para este grupo. El intervalo de confianza (0.0056 - 0.0155) es muy estrecho y bajo, lo que sugiere una alta precisión en esta estimación.

Los perros cuyo género es desconocido podrían ser animales callejeros o abandonados, sobre los que no se ha recopilado suficiente información. Es posible que estos animales tengan menos probabilidades de recibir servicios veterinarios en general, incluida la esterilización, debido a la falta de intervención humana o recursos.

Distrito Brooklyn (OR = 2.083): En Brooklyn, la probabilidad de esterilización es 2.08 veces mayor que en el Bronx. El intervalo de confianza (1.5699 - 2.7687) indica que este aumento es estadísticamente significativo.

Brooklyn podría tener programas o subvenciones específicas para la esterilización, así como un mayor acceso a servicios veterinarios y campañas de sensibilización. También es posible que la población de Brooklyn tenga mayor conciencia o disposición hacia la esterilización, debido a factores socioeconómicos o de cultura local.

Distrito Manhattan (OR = 2.160): En Manhattan, la probabilidad de esterilización es 2.16 veces mayor que en el Bronx. El intervalo (1.6537 - 2.8299) está por encima de 1, indicando significancia estadística.

Manhattan, como un área con mayores ingresos en promedio, puede tener más recursos para clínicas veterinarias y programas de control de población animal. La población de Manhattan también puede tener una mentalidad más progresista sobre el control animal, lo cual favorece la esterilización y genera tasas más altas.

Distrito Otros (OR = 2.950): En otros distritos agrupados, la probabilidad de esterilización es 2.95 veces mayor que en el Bronx, y el intervalo (1.8077 - 4.7857) sugiere una alta variabilidad, pero también una diferencia significativa.

En otros distritos fuera del Bronx y los principales (Brooklyn, Manhattan, etc.), es posible que haya iniciativas locales o regulaciones que promuevan la esterilización de animales de compañía. Además, estos distritos pueden incluir zonas rurales o suburbanas con políticas municipales específicas para el control de animales.

Distrito Queens (OR = 2.161): En Queens, la probabilidad de esterilización es 2.16 veces mayor que en el Bronx. El intervalo de confianza (1.6659 - 2.8129) indica que el OR es significativamente mayor a 1.

Queens podría beneficiarse de programas de esterilización patrocinados por organizaciones locales o una mejor infraestructura veterinaria en comparación con el Bronx. También es posible que los residentes de Queens tengan acceso a subsidios o incentivos para esterilizar a sus mascotas.

Distrito Staten Island (OR = 2.270): En Staten Island, la probabilidad de esterilización es 2.27 veces mayor que en el Bronx. El intervalo (1.6639 - 3.0987) respalda una diferencia significativa.

Staten Island tiene una comunidad más cerrada y una cultura de vecindad que puede fomentar el cumplimiento de las normas de esterilización. Esto, sumado a programas de control animal o mayor disponibilidad de clínicas veterinarias, podría explicar la mayor tasa de esterilización en comparación con el Bronx.

En resumen, los perros de algunas razas, como “mezclada/otras” y pitbulls de raza pura, tienen menos probabilidades de ser esterilizados en comparación con los pitbull mezclados. Los machos también tienen menos probabilidades de ser esterilizados que las hembras, y los perros de género desconocido tienen una probabilidad extremadamente baja de esterilización. Por otro lado, vivir en cualquier distrito fuera del Bronx está asociado con mayores probabilidades de esterilización, lo que sugiere que el Bronx tiene tasas de esterilización más bajas o menos acceso a estos servicios.

Evaluamos el modelo

La bondad del modelo se realiza contrastando si hay diferencias significativas entre el modelo sin variables y el que hemos creado, para ello:

```
dev <- logistica_simplified$deviance
nullDev <- logistica_simplified)null.deviance
modelChi <- nullDev - dev
modelChi
```

```
## [1] 1415.322
```

Como este valor es positivo se reduce la verosimilitud, ahora habrá que contrastar si esta reducción es significativa, sabiendo que este estadístico sigue una chi cuadrado con $k-1$ grados de libertad siendo k el número de parámetros calculados:

```
chidf <- logistica_simplified$df.null -logistica_simplified$df.residual
chisq.prob <- 1 - pchisq(modelChi, df=chidf)
chisq.prob
```

```
## [1] 0
```

Tendremos que el p-valor del contraste es menor que 0.05 por lo que hay una reducción verosimilitud significativa, es decir la diferencia entre ambos valores es significativa.

También se puede comparar si hay diferencias significativas entre el modelo completo y el que hemos calculado:

```
anova(logistica,logistica_simplified, test="Chisq")
```

```
## Warning in anova.glmlist(c(list(object), dotargs),
## dispersion = dispersion, : models with response
## '"SpayNeuter"' removed because response differs from model
## 1

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: filtered_dog_bites$SpayNeuter
##
## Terms added sequentially (first to last)
##
##
##                               Df Deviance Resid. Df Resid. Dev
## NULL                           6447     6319.0
## filtered_dog_bites$Breed      4    314.46    6443     6004.5
## filtered_dog_bites$Gender     2   1486.83    6441     4517.7
## filtered_dog_bites$Borough    5     83.12    6436     4434.6
##                               Pr(>Chi)
## NULL
## filtered_dog_bites$Breed < 2.2e-16 ***
## filtered_dog_bites$Gender < 2.2e-16 ***
## filtered_dog_bites$Borough < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que el modelo ha mejorado significativamente al excluir las dos razas anteriores, y que las tres variables raza, género y distrito son importantes para entender si un perro que ha mordido ha sido esterilizado o no.

Conclusión

En conclusión, el análisis de los incidentes de mordeduras de perro en Nueva York revela patrones clave en relación con la raza, el género, la esterilización y la ubicación geográfica. Los resultados sugieren que los perros no esterilizados tienen más probabilidades de morder, y las razas como los pitbulls y los perros de raza mixta están más involucradas en estos incidentes. Además, los distritos con mayores tasas de esterilización, como Brooklyn y Manhattan, muestran tasas de mordeduras relativamente más bajas. Para abordar este problema de salud pública, se recomienda la implementación de políticas de esterilización más accesibles y efectivas, especialmente en distritos con tasas más bajas de esterilización.

Es importante señalar que los resultados obtenidos deben interpretarse con cautela debido a posibles sesgos en los datos. La falta de información completa sobre el comportamiento de los perros, las circunstancias específicas de los incidentes y la representación desigual de algunas razas o áreas geográficas puede haber influido

en los patrones observados. Además, el sesgo de selección en los datos disponibles, como la subrepresentación de ciertas razas o zonas con menores tasas de reporte, podría haber afectado los resultados del modelo. Por lo tanto, es crucial que futuros estudios utilicen conjuntos de datos más completos y representativos para obtener conclusiones más robustas.

Futuras investigaciones podrían explorar otros factores, como la educación del propietario y las políticas locales sobre tenencia responsable de mascotas, para desarrollar estrategias preventivas más completas.