Homework 6 - Drew Davison

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
New names:
Rows: 147 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (8): ...1, female, age, highstatus, yrsmoke, cigsday, bird, cancer

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data: Association Between Bird-Keeping and Risk of Lung Cancer

A 1972-1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague. They identified 49 cases of lung cancer among patients who were registered with a general practice, who were age 65 or younger, and who had resided in the city since 1965. Each patient (case) with cancer was matched with two control subjects (without cancer) by age and sex. Further details can be found in Holst, Kromhout, and Brand (1988).

Age, sex, and smoking history are all known to be associated with lung cancer incidence. Thus, researchers wished to determine after age, sex, socioeconomic status, and smoking have

been controlled for, is an additional risk associated with birdkeeping? The data (Ramsey and Schafer 2002) is found in [`birdkeeping.csv`(data/birdkeeping.csv)].[1]

The paper that this exercise is based upon can be found here and please read it before completing the assignment. (https://www.bmj.com/content/bmj/297/6659/1319.full.pdf)

Exercise 1

Part a

Create a segmented bar chart and appropriate table of proportions showing the relationship between birdkeeping and cancer diagnosis. Summarize the relationship in 1 - 2 sentences.

Note

```
# Create a table of counts
count_table <- table(birds$bird, birds$cancer)

# Convert to proportions by row (birdkeeping)
prop_table <- prop.table(count_table, margin = 1)
print("Proportion Table:")
```

```
[1] "Proportion Table:"
```

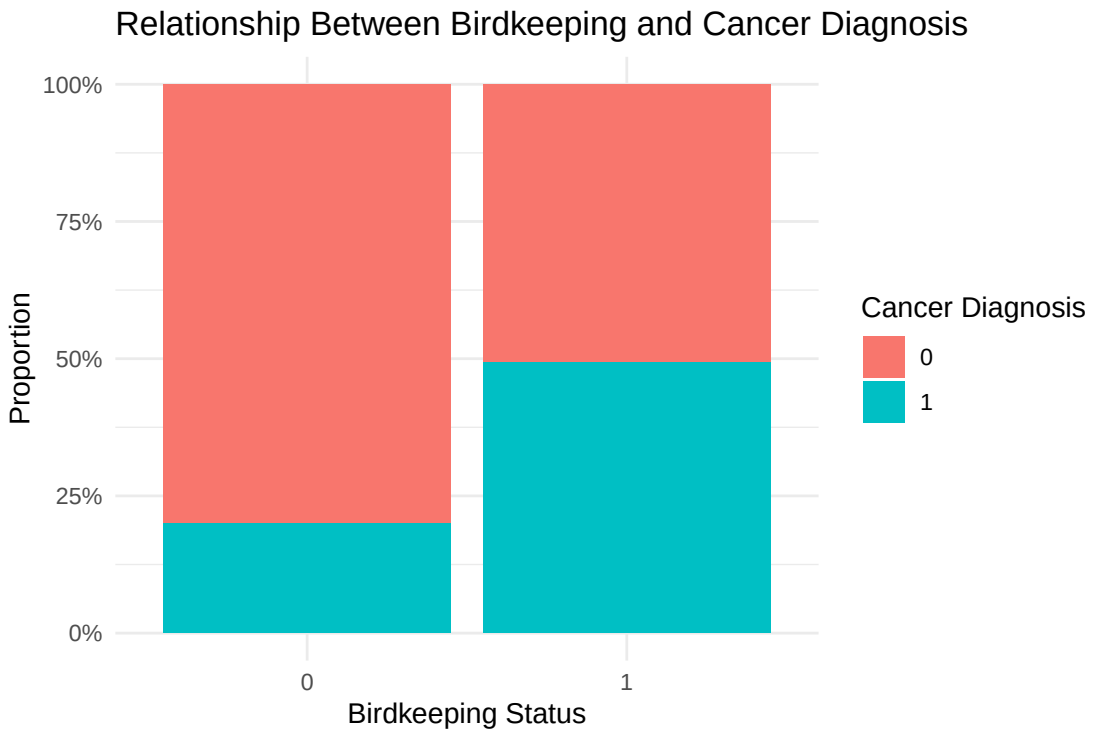
```
print(round(prop_table, 2))
```

```
     0    1
0 0.80 0.20
1 0.51 0.49
```

[1] This problem is adapted from Section 6.8.1, Ex 4.

```r
# Convert to a data frame for ggplot
df_plot <- as.data.frame(count_table)
colnames(df_plot) <- c("Birdkeeping", "Cancer", "Count")

# Plot segmented bar chart
ggplot(df_plot, aes(x = Birdkeeping, y = Count, fill = Cancer)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Relationship Between Birdkeeping and Cancer Diagnosis",
    y = "Proportion",
    x = "Birdkeeping Status",
    fill = "Cancer Diagnosis"
  ) +
  theme_minimal()
```



The proportion table reveals:
0-0: 80% of participants without birdkeeping did not have lung cancer. 0-1: 20% of participants without birdkeeping did have lung cancer. 1-0: 51% of participants with birdkeeping did not have lung cancer. 1-1: 49% of participants with birdkeeping did have lung cancer.

It appears that the relationship between birdkeeping and lung cancer is, for the given sample, that there is a higher percentage of people with lung cancer that keep birds than for those who do not keep birds.

Part b

Calculate the unadjusted odds ratio of a lung cancer diagnosis comparing birdkeepers to non-birdkeepers. Interpret this odds ratio in context. (Note: an unadjusted odds ratio is found by not controlling for any other variables.)

Note

```
# Create a 2x2 table for birdkeeping and lung cancer
table_lung <- table(birds$bird, birds$cancer)

# View the table
print(table_lung)
```

```
   0  1
0 64 16
1 34 33
```

```
# Calculate the odds ratio manually
# OR = (a/c) / (b/d) = (exposed cases / exposed non-cases) / (unexposed cases / unexposed n

# Extract counts
a <- table_lung["1", "1"]        # birdkeepers with lung cancer
b <- table_lung["0", "1"]         # non-birdkeepers with lung cancer
c <- table_lung["1", "0"]    # birdkeepers without lung cancer
d <- table_lung["0", "0"]     # non-birdkeepers without lung cancer

# Calculate odds ratio
odds_ratio <- (a / c) / (b / d)
print(paste("Unadjusted Odds Ratio:", round(odds_ratio, 2)))
```

```
[1] "Unadjusted Odds Ratio: 3.88"
```

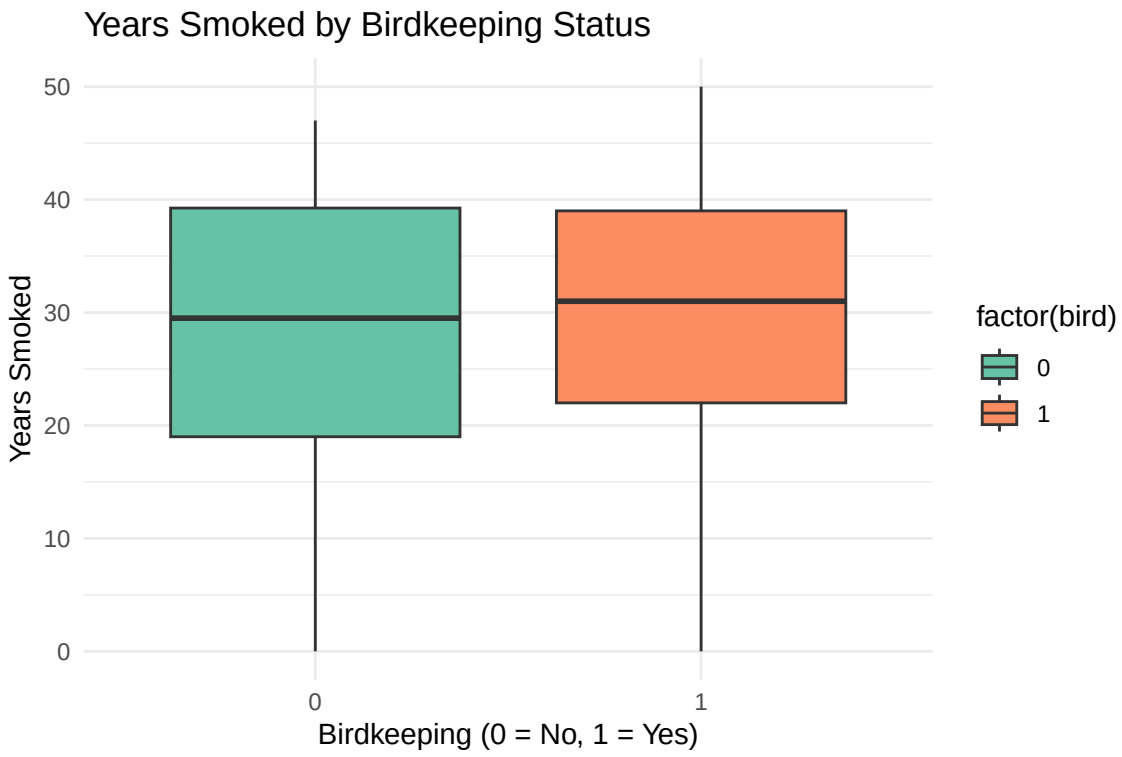
The unadjusted odds ratio of 3.88 indicates that birdkeepers had 3.88 times the odds of being diagnosed with lung cancer compared to non-birdkeepers. If the OR is greater than 1, this suggests a positive association between birdkeeping and lung cancer. If less

than 1, it suggests a negative association This is a crude measure and does not account for potential confounders such as smoking or age.

Part c

Does there appear to be an interaction between number of years smoked and whether the subject keeps a bird? Demonstrate with an appropriate plot and briefly explain your response.

```
ggplot(birds, aes(x = factor(bird), y = yrsmoke, fill = factor(bird))) +
  geom_boxplot() +
  labs(
    title = "Years Smoked by Birdkeeping Status",
    x = "Birdkeeping (0 = No, 1 = Yes)",
    y = "Years Smoked"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#66c2a5", "#fc8d62"))
```



The median smoking years for those who keep birds is slightly higher than those who do not,

however, the IQR is smaller for those who keep birds. The 25th percentile for those who keep birds is a greater number of years than those who do not keep birds. The 75th percentile for those who keep birds is a fewer number of years than those who do not keep birds.

In conclusion, it does not appear that there is a meaningful difference in years smoked based on if someone keeps birds.

Before answering the next questions, fit logistic regression models in R with cancer as the response and the following sets of explanatory variables:

- model1 = age, yrsmoke, cigsday, female, highstatus, bird
- model2 = yrsmoke, cigsday, highstatus, bird
- model3 = yrsmoke, bird
- model4 = yrsmoke, bird, yrsmoke:bird

```
# Model 1: age, yrsmoke, cigsday, female, highstatus, bird
model1 <- glm(cancer ~ age + yrsmoke + cigsday + female + highstatus + bird,
              data = birds,
              family = binomial)

# Model 2: yrsmoke, cigsday, highstatus, bird
model2 <- glm(cancer ~ yrsmoke + cigsday + highstatus + bird,
              data = birds,
              family = binomial)

# Model 3: yrsmoke, bird
model3 <- glm(cancer ~ yrsmoke + bird,
              data = birds,
              family = binomial)

# Model 4: yrsmoke, bird, and interaction between yrsmoke and bird
model4 <- glm(cancer ~ yrsmoke * bird,
              data = birds,
              family = binomial)

# Summary of all models
tidy(model1)
```

```
# A tibble: 7 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -1.94      1.80      -1.07  0.283
2 age          -0.0398    0.0355    -1.12  0.263
```

```
3 yrsmoke       0.0729    0.0265     2.75  0.00594
4 cigsday       0.0260    0.0255     1.02  0.308
5 female        0.561     0.531      1.06  0.291
6 highstatus    0.105     0.469      0.225 0.822
7 bird          1.36      0.411      3.31  0.000923
```

```
tidy(model2)
```

```
# A tibble: 5 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)  -3.38      0.708     -4.78  0.00000177
2 yrsmoke       0.0491    0.0188     2.62  0.00891
3 cigsday       0.0286    0.0244     1.17  0.241
4 highstatus   -0.0689    0.453     -0.152 0.879
5 bird          1.49      0.403      3.69  0.000223
```

```
tidy(model3)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)  -3.18      0.636     -5.00 0.000000582
2 yrsmoke       0.0582    0.0168     3.46 0.000544
3 bird          1.48      0.396      3.73 0.000194
```

```
tidy(model4)
```

```
# A tibble: 4 x 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -3.00       0.898     -3.34  0.000844
2 yrsmoke       0.0528     0.0256     2.06  0.0394
3 bird          1.18       1.15       1.03  0.304
4 yrsmoke:bird  0.00930    0.0340     0.274 0.784
```

Part d

Is there evidence that we can remove age and female from our model? Perform an appropriate test comparing model1 to model2; give a test statistic and p-value, and state a conclusion in context.

Note

```
# Perform likelihood ratio test comparing Model 1 and Model 2
anova(model2, model1, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: cancer ~ yrsmoke + cigsday + highstatus + bird
Model 2: cancer ~ age + yrsmoke + cigsday + female + highstatus + bird
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       142     156.72
2       140     154.20  2   2.5257   0.2828
```

The test statistic, being the Deviance Difference, is 2.53.
The p-value of 0.2828 is greater than 0.05, which means we fail to reject the null hypothesis.
This suggests that removing age and female from the model does not significantly worsen the model.

Part e

Carefully interpret each of the four model coefficients (including the intercept) in model4 in context.

Note

Intercept: -2.999 This is the log-odds of cancer diagnosis for someone who does not keep birds (bird = 0) and has smoked for 0 years (yrsmoke = 0).
exp(-2.999) 0.05
So, a non-birdkeeper who has never smoked has about a 5% chance of being diagnosed with cancer.
yrsmoke: 0.0528 Among non-birdkeepers, each additional year of smoking is associated with a 5.4% increase in the odds of being diagnosed with cancer: exp(0.0528) 1.054 So, for non-birdkeepers, smoking longer slightly increases cancer risk.
bird: 1.179 For individuals who keep birds and have smoked 0 years, their log-odds of cancer are 1.179 higher than non-birdkeepers who also never smoked. That translates to:

exp(1.179) 3.25 So, birdkeepers who never smoked have 3.25 times higher odds of being diagnosed with cancer compared to non-birdkeepers who never smoked.
yrsmoke:bird interaction: 0.0093 This term tells us how the effect of smoking duration on cancer risk differs for birdkeepers. Among birdkeepers, each additional year of smoking increases the log-odds of cancer by: 0.0528 (main effect) + 0.0093 (interaction) = 0.0621 So the combined effect of smoking and birdkeeping is slightly more harmful, but the interaction is not statistically significant (p = 0.78). ## Part f If you replaced yrsmoke everywhere it appears in model4 with a mean-centered version of yrsmoke, tell what would change among these elements: the 4 coefficients, the 4 p-values for coefficients, and the residual deviance.

Note

```
# Create centered variable
birds$yrsmoke_c <- birds$yrsmoke - mean(birds$yrsmoke)

# New model with centered predictor
model5 <- glm(cancer ~ yrsmoke_c + bird + yrsmoke_c:bird,
    data = birds,
    family = binomial)

tidy(model5)
```

```
# A tibble: 4 x 5
  term           estimate std.error statistic    p.value
  <chr>             <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)    -1.53       0.319     -4.79  0.00000170
2 yrsmoke_c       0.0528     0.0256     2.06  0.0394
3 bird            1.44       0.416      3.46  0.000537
4 yrsmoke_c:bird  0.00930    0.0340     0.274 0.784
```

Changed: Intercept, coefficient for bird, p-values for intercept and bird Unchanged: coefficients for yrsmoke and interaction term, p-values for yrsmoke and interaction term, residual deviance

Part g

Observe that model3 is a potential final model based on this set of predictor variables. How does the adjusted odds ratio for birdkeeping from model3 compare with the unadjusted odds ratio you found in (b)? Is birdkeeping associated with a significant increase in the odds of developing lung cancer, even after adjusting for other factors?

Note

```r
# Fit Model 3: Adjusted for yrsmoke
model3 <- glm(cancer ~ yrsmoke + bird, family = binomial, data = birds)
summary(model3)
```

```
Call:
glm(formula = cancer ~ yrsmoke + bird, family = binomial, data = birds)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.18016    0.63640  -4.997 5.82e-07 ***
yrsmoke      0.05825    0.01685   3.458 0.000544 ***
bird         1.47555    0.39588   3.727 0.000194 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 158.11  on 144  degrees of freedom
AIC: 164.11

Number of Fisher Scoring iterations: 4
```

```r
# Extract the bird coefficient and calculate adjusted OR
coef_bird <- coef(model3)["bird"]
adjusted_or <- exp(coef_bird)
adjusted_or
```

```
    bird
4.373447
```

The unadjusted odds ratio calculated earlier was 3.88, meaning that birdkeepers had nearly four times the odds of being diagnosed with lung cancer compared to non-birdkeepers, without accounting for any other variables.
In Model 3, which adjusts for years smoked, the adjusted odds ratio for birdkeeping is slightly higher—around 4.37. This suggests that even after accounting for smoking duration, birdkeeping is still associated with a significant increase in the odds of developing lung cancer.
Because the adjusted odds ratio remains high and statistically significant, there is strong evidence that birdkeeping is independently associated with increased lung cancer risk.

Part h

Discuss the scope of inference in this study. Can we generalize our findings beyond the subjects in this study? Can we conclude that birdkeeping causes increased odds of developing lung cancer? Do you have other concerns with this study design or the analysis you carried out?

Can we generalize the findings beyond the subjects in this study? The findings from this study are likely to be specific to this population and location, meaning that generalizing the results beyond this group may not be appropriate. The analysis adjusted for several factors like smoking history and socioeconomic status, which may not be representative of other populations. Any inference to different age groups, or individuals from outside The Hague, should be made cautiously. Therefore, the scope of inference is limited to individuals similar to those studied — specifically, young adults and middle-aged individuals in The Hague with a smoking history.

Can we conclude that birdkeeping causes increased odds of developing lung cancer? This study was observational and case-control in nature, which means it identifies associations but does not provide evidence of causal relationships. While the study adjusts for potential confounders (age, sex, smoking, and socioeconomic status), there could still be other unmeasured confounders that may affect the relationship between birdkeeping and lung cancer. Moreover, the case-control design means the study is retrospective, relying on existing data, and we cannot infer causality without a more controlled experimental design (such as a randomized controlled trial or cohort study).

Concerns with the study design or the analysis: Potential Confounders: Even though the study adjusted for smoking, age, and socioeconomic status, other confounders may exist, such as exposure to secondhand smoke, occupational exposures, or other environmental risk factors related to living in The Hague during this time period.

No Longitudinal Data: This study design does not track participants over time, so it cannot capture temporal relationships between birdkeeping and lung cancer development. Cohort studies that follow participants for extended periods of time can better establish the timeline of risk exposure and disease onset.

Data Limitations: The study is based on self-reported data, which may not be fully accurate, especially when it comes to smoking habits

Exercise 2

(Ataman and Sarıyer 2021) use ordinal logistic regression to predict patient wait and treatment times in an emergency department (ED). The goal is to identify relevant factors that can be

used to inform recommendations for reducing wait and treatment times, thus improving the quality of care in the ED.

The data include daily records for ED arrivals in August 2018 at a public hospital in Izmir, Turkey. The response variable is Wait time, a categorical variable with three levels:

- Patients who wait less than 10 minutes
- Patients whose waiting time is in the range of 10-60 minutes
- Patients who wait more than 60 minutes

Part a

Compare and contrast the proportional odds model with the multinomial logistic regression model. Write your response using 3 - 5 sentences. You can find a brief review of the proportional odds model here: https://library.virginia.edu/data/articles/fitting-and-interpreting-a-proportional-odds-model and https://online.stat.psu.edu/stat504/lesson/8/8.4

Note

The proportional odds model and the multinomial logistic regression model are both used for analyzing categorical outcome variables, but they differ in their assumptions and the types of relationships they model. The proportional odds model is used for ordinal outcomes, assuming that the relationship between each pair of outcome categories is the same (i.e., the odds ratios are constant across thresholds). In contrast, the multinomial logistic regression model is suited for nominal outcomes and does not assume any inherent ordering between categories, modeling each category's probability relative to a reference category. While the proportional odds model is more restrictive, it is simpler and easier to interpret when the assumption of proportionality holds, whereas the multinomial logistic model allows for more flexibility but requires estimating more parameters. ## Part b Table 5 in the paper contains the output for the wait time and treatment time models. Consider only the model for wait time. Describe the effect of arrival mode (ambulance, walk-in) on the waiting time. Note: walk-in is the baseline in the model. (A link to the paper can be found in the slides).

Note

The coefficient for arrival mode is -3.398, with a p-value of 0.000, and a 95% confidence interval of $[-3.616, -3.180]$. This negative and highly significant coefficient indicates that patients arriving by ambulance are much less likely to experience longer wait times compared to walk-in patients. In other words, ambulance arrivals are prioritized and tend to be seen more quickly, making them more likely to fall into the "less than 10 minutes"

wait category. The strong significance and large magnitude of the effect underscore how crucial mode of arrival is in determining how quickly a patient is attended to in the ED. ## Part c Consider output from both the wait time and treatment time models. Use the results from both models to describe the effect of triage level (red = urgent, green = non-urgent) on the wait and treatment times in the ED. Note: red is the baseline level.

Wait time model: The coefficient for triage level is 0.016, with a p-value of 0.153 and a 95% confidence interval of $[-0.006, 0.037]$. This suggests that, when compared to red (urgent) cases, green (non-urgent) cases tend to wait slightly longer, but this difference is not statistically significant. We cannot confidently conclude that triage level affects wait time.

Treatment time model: The coefficient for triage level is -0.950, with a p-value of 0.000 and a 95% confidence interval of $[-0.973, -0.926]$. This means that green (non-urgent) cases receive significantly shorter treatment times than red (urgent) cases. The negative coefficient indicates that patients with less urgent conditions tend to be treated faster once seen.

Triage level has a strong and significant effect on treatment time, with non-urgent patients receiving shorter treatments. However, its effect on wait time is minimal and not statistically significant, suggesting that urgent and non-urgent patients wait similar amounts of time before being treated.

Exercise 3

Ibanez and Roussel (2022) conducted an experiment to understand the impact of watching a nature documentary on pro-environmental behavior. The researchers randomly assigned the 113 participants to watch an video about architecture in NYC (control) or a video about Yellowstone National Park (treatment). As part of the experiment, participants played a game in which they had an opportunity to donate to an environmental organization. The data set is available in nature.csv in the data folder. We will use the following variables:

- donation_binary: 1 - participant donated to environmental organization versus 0 - participant did not donate.
- Age: age in years
- Gender: Participant's reported gender
- Treatment: Urban (T1)" - the control group versus "Nature (T2)" - the treatment group.
- NEP_high: 1 - score of 4 or higher on the New Ecological Paradigm (NEP) versus 0 - score of less than 4.

See the Introduction and Methods sections of Ibanez and Roussel (2022) for more details about the variables and see the class slides regarding the url for the paper.

```r
nature <- read_csv("~/STA310/HW-6/STA310/Data/nature.csv", show_col_types = FALSE)
# https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275806
nature = nature %>% select(c("donation_binary", "Age", "Gender", "Treatment", "nep_high", "D
# summary(nature)
```

Part a

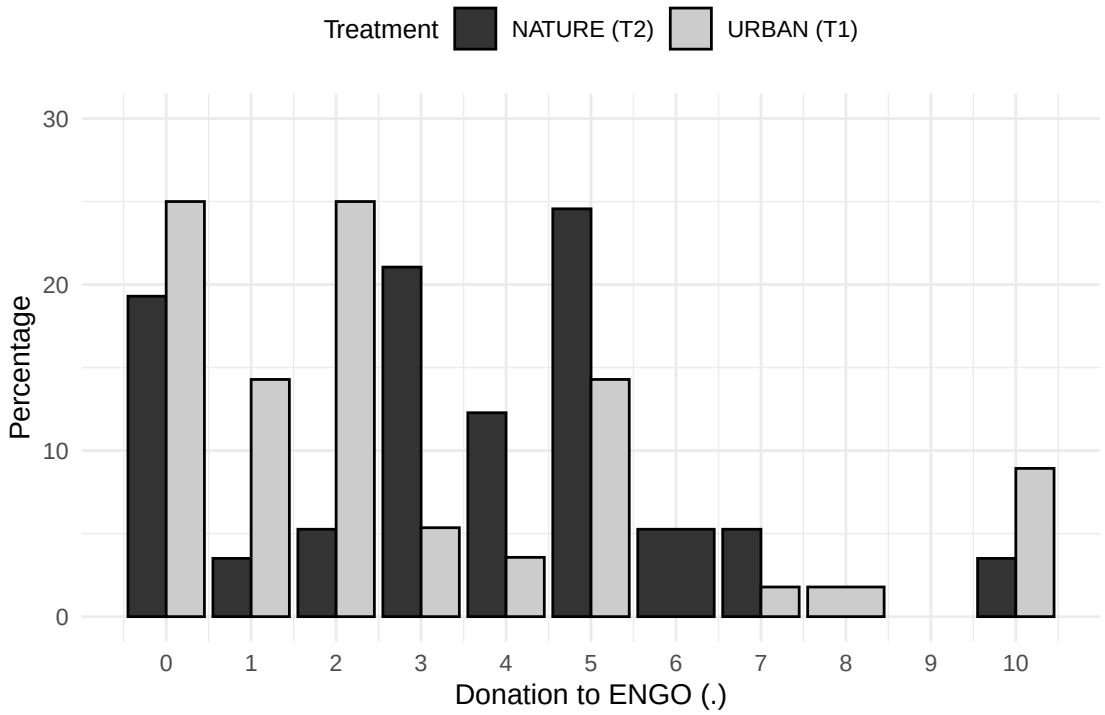
Figure 2 on pg. 9 of the article visualizes the relationship between donation amount and treatment. Recreate this visualization using your own code. Use the visualization to describe the relationship between donating and the treatment.

Note

```r
plot_data <- nature %>%
  group_by(Treatment, `Donation (level)`) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(Treatment) %>%
  mutate(percentage = 100 * count / sum(count))

plot_data$Treatment <- factor(plot_data$Treatment, levels = c("NATURE (T2)", "URBAN (T1)")

ggplot(plot_data, aes(x = `Donation (level)`, y = percentage, fill = Treatment)) +
  geom_bar(stat = "identity", position = position_dodge(), color = "black") +
  scale_fill_manual(values = c( "NATURE (T2)" = "gray20", "URBAN (T1)" = "gray80")) +
  scale_x_continuous(breaks = 0:10) +
  scale_y_continuous(limit = c(0, 30)) +
  labs(x = "Donation to ENGO (€)", y = "Percentage", fill = "Treatment") +
  theme_minimal() +
  theme(
    legend.position = "top",
    legend.justification = "center",
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 9)
  )
```



Part b

Fit a probit regression model using age, gender, treatment, nep_high and the interaction between nep_high and treatment predict the likelihood of donating. (Note: Your model will be similar (but not exactly the same) as the "Likelihood" model in Table 5 on pg. 11.) Display the model.

Note

```
# Fit the probit regression model
nature_model <- glm(donation_binary ~ Age + Gender + Treatment + nep_high + Treatment:nep_h
             data = nature, family = binomial(link = "probit"))

# Display the model summary
tidy(nature_model)
```

```
# A tibble: 6 x 5
  term                        estimate std.error statistic p.value
  <chr>                          <dbl>     <dbl>     <dbl>   <dbl>
```

```
1 (Intercept)                     0.284     0.996      0.285  0.776
2 Age                             0.0597    0.0423     1.41   0.158
3 Gender                         -0.737     0.310     -2.38   0.0173
4 TreatmentURBAN (T1)            -0.191     0.406     -0.471  0.637
5 nep_high                       -0.639     0.418     -1.53   0.127
6 TreatmentURBAN (T1):nep_high    0.183     0.568      0.323  0.747
```

Part c

Describe the effect of watching the documentary on the likelihood of donating.

Note

Since NATURE (T2) is the reference group in the model, the effect of watching the Nature documentary is captured by the Intercept term.
Intercept: 0.2835 (p-value = 0.776)
This represents the baseline probability of donating for individuals in the Nature documentary (T2) group, assuming all other variables are held constant (such as age, gender, and NEP score). The intercept value of 0.2835 suggests that for individuals in the NATURE (T2) treatment group, the likelihood of donating is approximately 28.35%.
Gender: (-0.7373, p-value = 0.0173): This coefficient suggests that, controlling for other factors, being male is associated with a lower likelihood of donating. This effect is statistically significant (p-value < 0.05), meaning that males are less likely to donate compared to females in the NATURE (T2) treatment group.
Age: (0.0597, p-value = 0.158): The positive coefficient for age suggests that older individuals are slightly more likely to donate, but this effect is not statistically significant (p-value = 0.158), meaning age does not have a strong effect on the likelihood of donating.
NEP High: (-0.6386, p-value = 0.127): The negative coefficient for nep_high indicates that individuals with a high NEP score (indicating higher environmental concern) are less likely to donate, but this effect is also not statistically significant (p-value = 0.127).
TreatmentURBAN (T1): (-0.1911, p-value = 0.637): The negative coefficient (-0.1911) suggests that individuals in the Urban documentary (T1) group are slightly less likely to donate compared to those in the Nature documentary (T2) group, but this difference is not statistically significant (p-value = 0.637). Therefore, we do not have enough evidence to conclude that the Urban documentary (T1) treatment significantly affects donation likelihood compared to the Nature documentary (T2).
TreatmentURBAN (T1):nep_high: (0.1833, p-value = 0.747): The interaction term between Treatment URBAN (T1) and nep_high is also not statistically significant (p-value = 0.747), suggesting that the combined effect of being in the Urban documentary group and having a high NEP score does not significantly affect the likelihood of donating.

Part d

Based on the model, what is the predicted probability of donating for a 20-year old female in the treatment group with a NEP score of 3?

- Intercept ((β_0)): (0.2835)
- Age coefficient ((β_1)): (0.0597)
- Gender coefficient ((β_2)): Since the individual is female, the gender coefficient for females is (0).
- NEP high coefficient ((β_3)): Since a NEP score of 3 is not considered high, we assume (nep_high = 0).
- Treatment (NATURE (T2)): Since NATURE (T2) is the reference group, the treatment coefficient is (0).
- Interaction term: Since we are in the NATURE (T2) group, the interaction term is (0).

Now, the linear predictor (LP) is calculated as:

[$LP = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{nep\_high}$]

Substitute the values into the equation:

[$LP = 0.2835 + (0.0597 \cdot 20) + (-0.7373 \cdot 0) + (-0.6386 \cdot 0)$]

Simplify the expression:

[$LP = 0.2835 + 1.194 + 0$]

[$LP = 1.4775$]

Next, we convert the linear predictor to the predicted probability using the probit link:

[$P = \Phi(LP) = \Phi(1.4775)$]

Where (Φ) is the CDF of the standard normal distribution. Using the normal CDF:

[$P \approx 0.9306$]

Thus, the predicted probability of donating for a 20-year-old female in the NATURE (T2) treatment group with a NEP score of 3 is approximately **93.06%**.