Homework 6

```
-- 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
```

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)

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

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.

```
library(ggplot2)
library(dplyr)
library(scales)
```

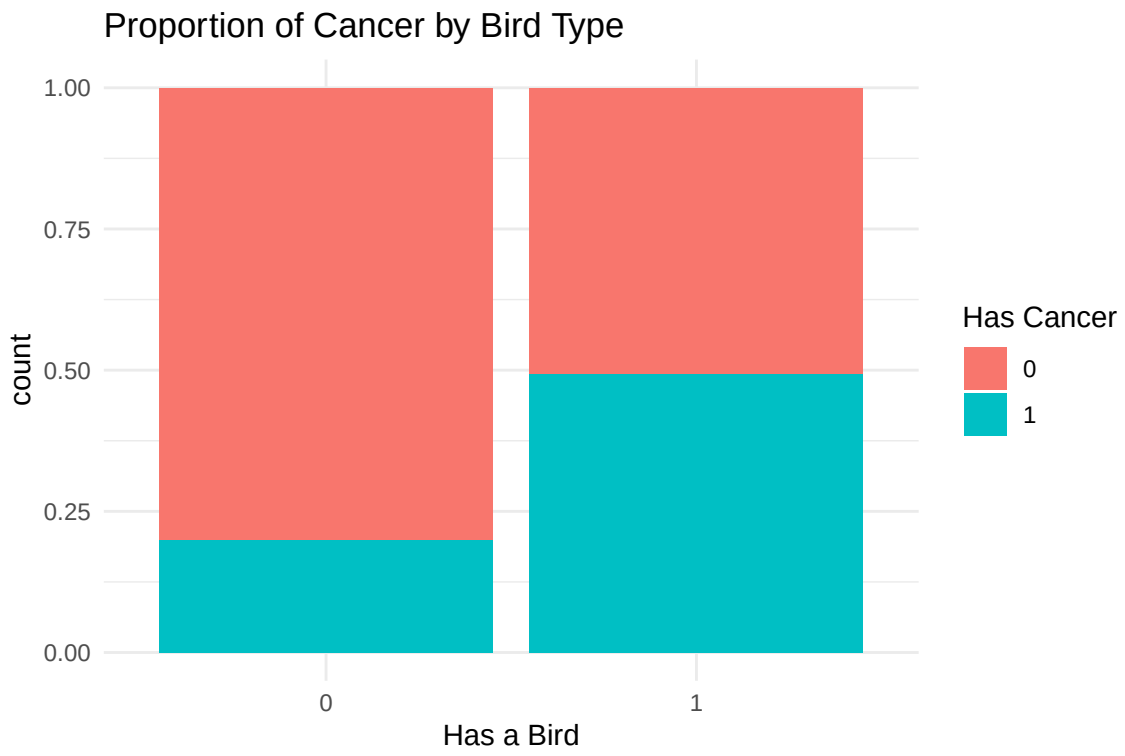
```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
birdkeeping <- read_csv("data/birdkeeping.csv")
```

```
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.
* `` -> `...1`
```

```
ggplot(birdkeeping, aes(x = factor(bird), fill = factor(cancer))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Cancer by Bird Type") +
  scale_y_continuous() +
  theme_minimal() +
```

```
  labs(
    x = "Has a Bird",
    fill = "Has Cancer"
  )
```




```
birdkeeping |>
  count(bird = factor(bird), cancer) |>
  group_by(bird) |>
  mutate(prop = n / sum(n))
```

```
# A tibble: 4 x 4
# Groups:   bird [2]
  bird  cancer     n  prop
  <fct>  <dbl> <int> <dbl>
1 0          0    64 0.8
2 0          1    16 0.2
3 1          0    34 0.507
4 1          1    33 0.493
```

From the bar chart it appears that a greater proportion of the people who have birds have lung cancer that those who do not have birds. The summary table suggests the same since 20% of those without birds had lung cancer while about 50% of those with birds had lung cancer.

chatGPT (https://chatgpt.com/c/67f4594e-1984-8008-9493-9cc0937a510f)

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.)

```
birdkeeping <- birdkeeping |>
  mutate(cancer = factor(cancer),
         bird = factor(bird))
```

```
birdkeeping |>
  group_by(cancer, bird) |>
  count()
```

```
# A tibble: 4 x 3
# Groups:   cancer, bird [4]
  cancer bird      n
  <fct>  <fct> <int>
1 0      0        64
2 0      1        34
3 1      0        16
4 1      1        33
```

```
model_og <- glm(cancer ~ bird, family = binomial, data = birdkeeping)

model_og|>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.386	0.280	-4.960	0
bird1	1.356	0.371	3.654	0

```
exp(1.356)
```

```
[1] 3.88064
```

The unadjusted odds ratio is 3.88, meaning that we expect the odds of having a cancer to have a multiplicative change of 3.88 when a person has a bird. This does appear to be significant (very small p-value).

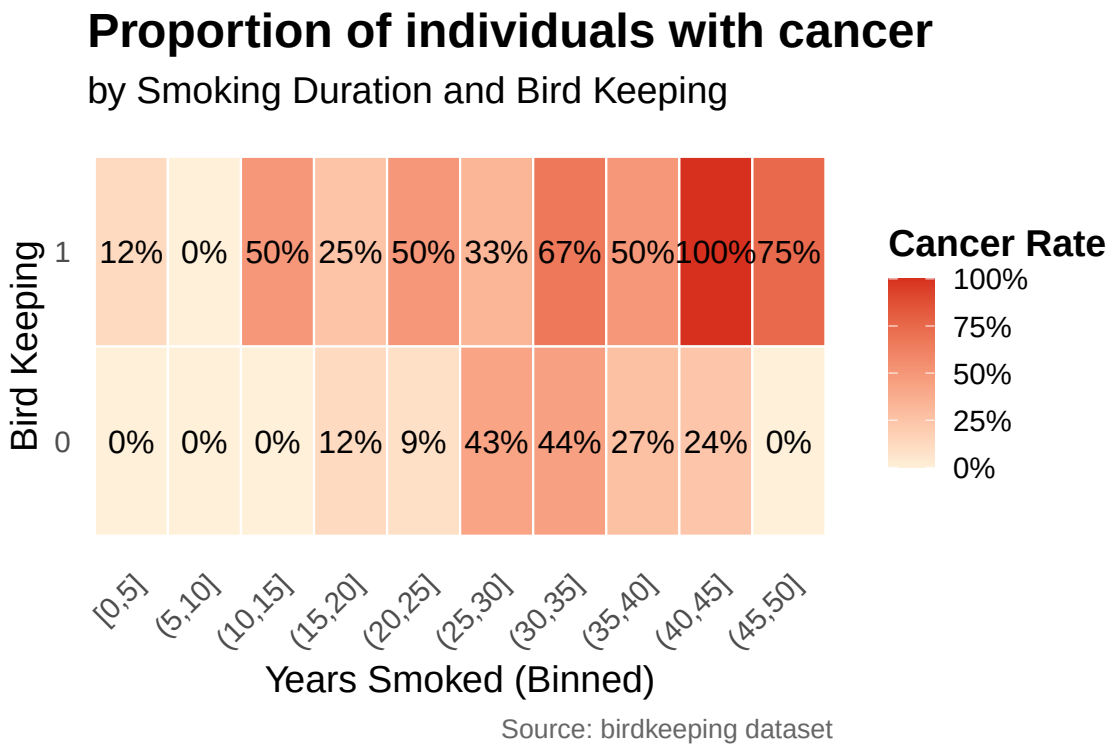
Part c

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.

There does appear to be an interaction between number of years smoked and whether the subject keeps a bird. I examined this by creating a heatmap, binning by years smoked. It seems like there's a higher proportion of individuals with cancer who both have birds and have high smoking duration than just individuals with either birds standalone or have been smoking a long time. This makes me believe that the extent to which these variables influence cancer risk depends on the other variable.

```
  # for percent formatting

# Step 1: Ensure cancer is numeric 0/1
birdkeeping <- birdkeeping |>
  mutate(
    cancer_num = case_when(
      cancer %in% c("1", 1, TRUE, "yes", "Yes") ~ 1,
      cancer %in% c("0", 0, FALSE, "no", "No") ~ 0,
      TRUE ~ NA_real_
    )
  )

# Step 2: Bin years smoked
birdkeeping <- birdkeeping |>
  mutate(
    yrsmoke_bin = cut(yrsmoke, breaks = seq(0, max(yrsmoke, na.rm = TRUE),
                                            by = 5), include.lowest = TRUE)
  )

# Step 3: Summarize data
heatmap_df <- birdkeeping |>
  group_by(yrsmoke_bin, bird) |>
  summarise(
    cancer_rate = mean(cancer_num, na.rm = TRUE),
    count = n(),
    .groups = "drop"
  )

# Step 4: Aesthetic plot
ggplot(heatmap_df, aes(x = yrsmoke_bin, y = factor(bird),
```

```
                      fill = cancer_rate)) +
geom_tile(color = "white", linewidth = 0.4) +
geom_text(aes(label = percent(cancer_rate, accuracy = 1)), color = "black",
          size = 4.2) +
scale_fill_gradient(
  low = "#fef0d9", high = "#d7301f",
  limits = c(0, 1),
  name = "Cancer Rate",
  labels = percent_format(accuracy = 1)
) +
labs(
  x = "Years Smoked (Binned)",
  y = "Bird Keeping",
  title = "Proportion of individuals with cancer",
  subtitle = "by Smoking Duration and Bird Keeping",
  caption = "Source: birdkeeping dataset"
) +
theme_minimal(base_size = 14) +
theme(
  plot.title = element_text(face = "bold", size = 18, family = "sans"),
  plot.subtitle = element_text(size = 14, margin = margin(b = 10)),
  plot.caption = element_text(size = 10, color = "gray40", hjust = 1),
  axis.text.x = element_text(angle = 45, hjust = 1),
  legend.title = element_text(face = "bold"),
  legend.key.height = unit(0.5, "cm"),
  panel.grid = element_blank(),
  panel.border = element_blank()
)
```



citation: https://chatgpt.com/c/67f55cb3-4250-8008-97ba-7d1c56b541fe

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

```
model1 <- glm(cancer ~ age + yrsmoke + cigsday + female + highstatus + bird,
              family = binomial, data = birdkeeping)

model1 |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.937	1.804	-1.074	0.283
age	-0.040	0.035	-1.120	0.263

term	estimate	std.error	statistic	p.value
yrsmoke	0.073	0.026	2.751	0.006
cigsday	0.026	0.026	1.019	0.308
female	0.561	0.531	1.057	0.291
highstatus	0.105	0.469	0.225	0.822
bird1	1.363	0.411	3.313	0.001

```
model2 <- glm(cancer ~ yrsmoke + bird + cigsday + highstatus, family = binomial, data = birdl

model2 |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.381	0.708	-4.778	0.000
yrsmoke	0.049	0.019	2.616	0.009
bird1	1.487	0.403	3.692	0.000
cigsday	0.029	0.024	1.174	0.241
highstatus	-0.069	0.453	-0.152	0.879

```
model3 <- glm(cancer ~ yrsmoke + bird, family = binomial, data = birdkeeping)

model3 |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.180	0.636	-4.997	0.000
yrsmoke	0.058	0.017	3.458	0.001
bird1	1.476	0.396	3.727	0.000

```
model4 <- glm(cancer ~ yrsmoke + bird + yrsmoke:bird,
              family = binomial, data = birdkeeping)

model4 |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.999	0.898	-3.338	0.001
yrsmoke	0.053	0.026	2.060	0.039
bird1	1.179	1.147	1.028	0.304
yrsmoke:bird1	0.009	0.034	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.

```
anova(model1, model2)
```

```
Analysis of Deviance Table

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

The test statistic is -2.5257.

Since the p-value (0.2828), there is sufficient evidence in favor of the null hypothesis, or the simpler model. The null hypothesis that the slope coefficients of the sex and age terms equals 0 can't be rejected. This means, we can likely remove age and female and have a decent model.

Part e

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

```
model4 |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.999	0.898	-3.338	0.001
yrsmoke	0.053	0.026	2.060	0.039
bird1	1.179	1.147	1.028	0.304
yrsmoke:bird1	0.009	0.034	0.274	0.784

Intercept: The p-value is small (0.001) so we say we expect an individual who has never smoked and doesn't own a bird to have 0.0498369 odds of having cancer.

yrsmoke: The p-value is small (0.039) so we can say for each additional year an individual smokes, we expect the odds of having cancer to multiply by a factor of 1.0544296, holding whether an individuals owns a bird constant.

bird1: The p-value (0.304) doesn't indicate significance of this estimate. If the p-value was below 0.05, we'd expect that owning a bird changes odds of having cancer by a multiplicative factor of 3.2511215, holding years smoked constant. However, we aren't able to conclude that here.

yrsmoke:bird1: The p-value (0.784) doesn't indicate significance of this esimate. If the p-value was small, we'd be able to say, for each additional year an individual smokes, we expect the odds of cancer for individuals who have birds to multiply by an additional 1.0090406. However, we aren't able to conclude that here.

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.

The intercept would change so that the baseline refers to a person who had smoked whatever the average number of years for years smoked was. Also, the coefficient would change for variable that indicated whether a person had a bird. As a consequence the p-values for both thte intercept and bird variable should change.

```
birdkeeping <- birdkeeping |>
  mutate(modYrsSmoke = yrsmoke-mean(birdkeeping$yrsmoke))
```

```
model4Mod <- glm(cancer ~ modYrsSmoke + bird + modYrsSmoke:bird,
                 family = binomial,
                 data = birdkeeping)

model4Mod |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.528	0.319	-4.786	0.000
modYrsSmoke	0.053	0.026	2.060	0.039
bird1	1.438	0.416	3.462	0.001
modYrsSmoke:bird1	0.009	0.034	0.274	0.784

```
model4 |>
  tidy()|>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.999	0.898	-3.338	0.001
yrsmoke	0.053	0.026	2.060	0.039
bird1	1.179	1.147	1.028	0.304
yrsmoke:bird1	0.009	0.034	0.274	0.784

```r
anova(model4, model4Mod)
```

```
Analysis of Deviance Table

Model 1: cancer ~ yrsmoke + bird + yrsmoke:bird
Model 2: cancer ~ modYrsSmoke + bird + modYrsSmoke:bird
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       143     158.04
2       143     158.04  0        0
```

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?

```
model3 |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.180	0.636	-4.997	0.000
yrsmoke	0.058	0.017	3.458	0.001
bird1	1.476	0.396	3.727	0.000

The adjusted odds ratio is 4.375409, meaning the odds of having cancer changes by a multiplicative factor of 4.375409 when someone has a bird, holding number of years smoked constant. The unadjusted odds ratio was 3.88, meaning that we expect the odds of having a cancer to have a multiplicative change of 3.88 when a person has a bird. Because the adjusted odds ratio is > 1, birdkeeping is associated with a significant increase in having cancer, even after adjusting for years smoked. In fact, it seems to be even more so, after adjusting for years smoked.

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?

I don't think we're really able to generalize findings beyond the subjects in this study. In fact, the patient population were those from a single general practice office in the Hague, and doesn't represent a popoulation beyond that.

I don't think we can include that birdkeeping causes increased odds of developing lung cancer.

I'm also concerned about the study design of how individuals were selected. In theory, all individuals from that general practice should have been potentially eligible for selection in this study but seems like many people with severe lung cancer were excluded. In fact, many of these patient were ineligible.

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

The multinomial logistic model and proportional odds models are similar in that they both model events that consist of more than 2 categories. They both use logit, assume multinomial response, and interpret coefficients as log odds. Both the multinomial logistic regression and proportional odds have multiple intercepts. They differ in the type of categories they model: proportional odds is used for categories that are ordinal (i.e., have a natural ordering) while MLR is used for categories that are nominal (i.e., don't have a natural ordering). For proportional odds, they each have the same coefficients for each level, but for multinomial logistic regression, we fit k-1 coefficients for each covariate term. This means multinomial logistic regression is more computationally intensive than proportional odds. We interpret prop odds in regards to either higher versus lower groups, but multinomial logistic regression interpretations are specific to a group.

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).

The effect for arrival mode was significant (p-value small). When a person enters the ER via ambulance as opposed to walking in, the odds of being in a higher waiting time category changes by a factor of 0.0334401, holding all else constant. This means the odds of being in a higher wait time category decreases, while the odds of being in a lower wait time category increases, holding all else constant.

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.

For the waiting time model, the effect of triage level was not significant since it had a p-value of 0.153. If it was significant, when the triage level is green, the odds of being in a higher waiting time category changes by a multiplicative factor of 1.0161287, holding all else constant. This would have meant that the odds of being in a higher waiting time category would be greater when the triage level is green, holding all else constant. Since the p-value is not less than 0.05, we can not make this conclusion.

When the triage level is green, the odds of being in a higher treatment time category changes by a multiplicative factor of 0.386741, holding all else constant. This means the odds of being in a higher treatment time category would decrease when the triage level is green. Here, the p-value is small, so we can make this conclusion.

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.

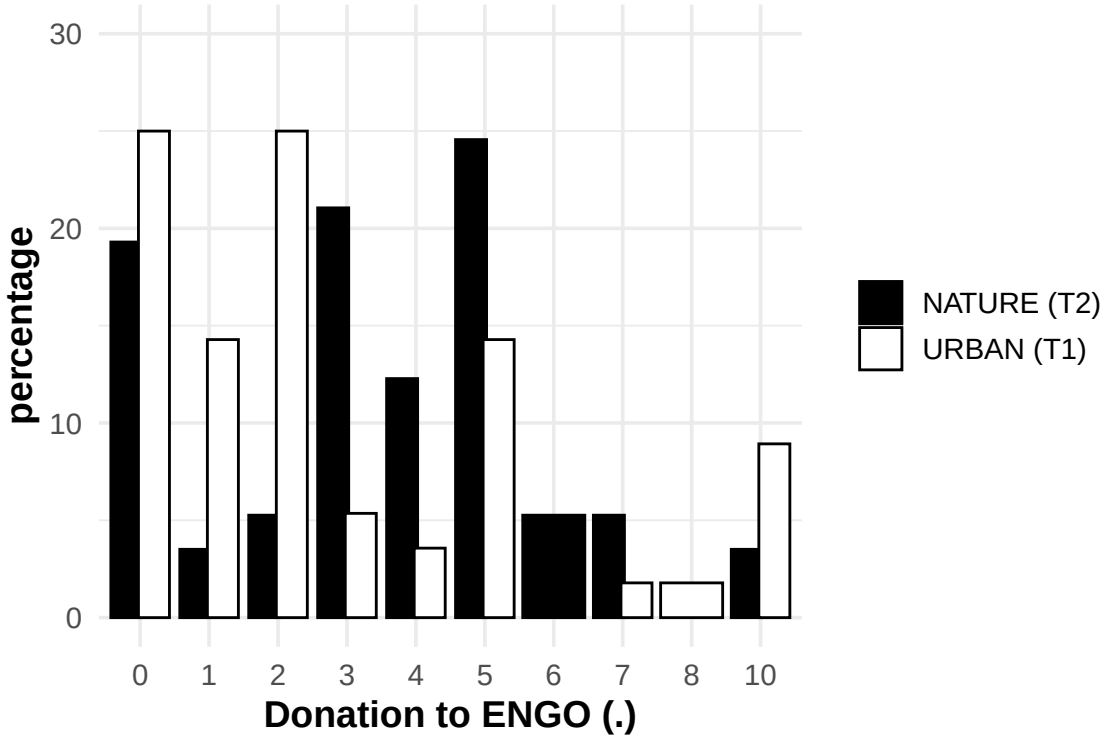
```
nature <- read_csv("data/nature.csv", show_col_types = FALSE)
# https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275806
nature = nature %>% select(c("donation_binary", "Age", "Gender", "Treatment", "nep_high", "D
# summary(nature)
```

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.

```
nature |>
  count(Treatment, `Donation (level)`) |>
  group_by(Treatment) |>
  mutate(percentage = n / sum(n) * 100) |>
  ggplot(aes(x = factor(`Donation (level)`), y = percentage,
             fill = Treatment)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
           color = "black") +
  scale_fill_manual(values = c("NATURE (T2)" = "black", "URBAN (T1)" = "white")) +
  labs(x = "Donation to ENGO (€)", y = "percentage") +
  ylim(0, 30) +
```

```
  theme_minimal(base_size = 14) +
  theme(
    axis.title = element_text(face = "bold"),
    legend.title = element_blank()
  )
```



It seems like in the higher donation groups, the percent of people receiving treatment (T2) is much less than in the other groups.

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.

```
fit <- glm(donation_binary ~ Age + Gender + Treatment + nep_high +
             Treatment*nep_high, family = binomial(link = "probit"), nature)
fit |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.284	0.996	0.285	0.776
Age	0.060	0.042	1.412	0.158
Gender	-0.737	0.310	-2.379	0.017
TreatmentURBAN (T1)	-0.191	0.406	-0.471	0.637
nep_high	-0.639	0.418	-1.527	0.127
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.

The p-value was not significant (0.637), but if it was, we'd be able to say participants who watched the documentary with urban footage (Urban T1) instead of the one with nature footage (Nature T2) were less likely to donate (-0.191). This would have meant for individuals with high NEP scores, they were still less likely to donate, but only slightly less when they say the urban documentary (0.183-0.191=-0.008). However, this can't really be said since our p-value wasn't small.

The probability of donating decreases by about 2.4% for those who watched the urban documentary.

```
margins(fit)
```

```
Average marginal effects
```

```
glm(formula = donation_binary ~ Age + Gender + Treatment + nep_high +     Treatment * nep_hig
```

```
     Age  Gender nep_high TreatmentURBAN (T1)
 0.01601 -0.1976  -0.1448            -0.02446
```

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?

```
new_data <- data.frame(
  Age = 20,
  Gender = 0,
  Treatment = "NATURE (T2)",
  nep_high = 0
)

predict(fit, newdata = new_data, type = "response")
```

```
        1
0.9303456
```

The predicted probability of donating is 93%.