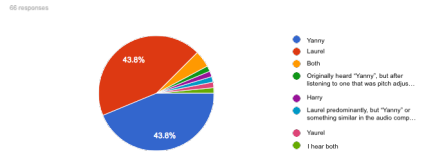
Extra tasks

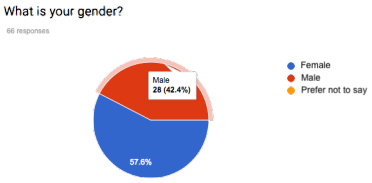
1 Case 1: Yanny or Laurel?

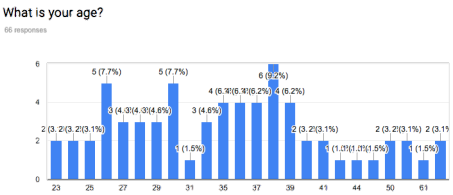
This auditory illusion first appeared on the internet in May 2018. An explanation of why people hear different things can be found in the following short video, just one of many internet sources discussing the phenomenon.

https://www.youtube.com/watch?v=yDiXQl7grPQ

The main reason behind the difference appears to be that as we age we lose the ability to hear certain sounds. To see if we could find evidence of such an age effect, we asked students and staff at the School of Mathematics and Statistics at the University of Glasgow to fill out a survey on what they hear. Below you can see summaries of the responses.







The proportions hearing `Yanny` and `Laurel` are very similar to each other, and there are some respondents who hear both or even something completely different. This may be because people do not listen to the audio file using the same device, something we couldn't control for in the survey. Ignoring the responses other than `Yanny` or `Laurel`, we have 53 observations.

> Task 1
>
> Download the `yanny.csv` data and fit a logistic regression model with `hear` as the binary response variable, and `age` and `gender` as the explanatory variables. What are your findings?
> See Solution
> Load the data:

```
yanny <- read.csv("yanny.csv",stringsAsFactors = T)
yanny <- yanny %>%
          select(hear, gender, age)
```

Exploratory Plots:

```
ggplot(data = yanny, aes(x = hear, y = age, fill = hear)) +
  geom_boxplot() +
  labs(x = "What do you hear?", y = "Age") +
  theme(legend.position = "none")
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```

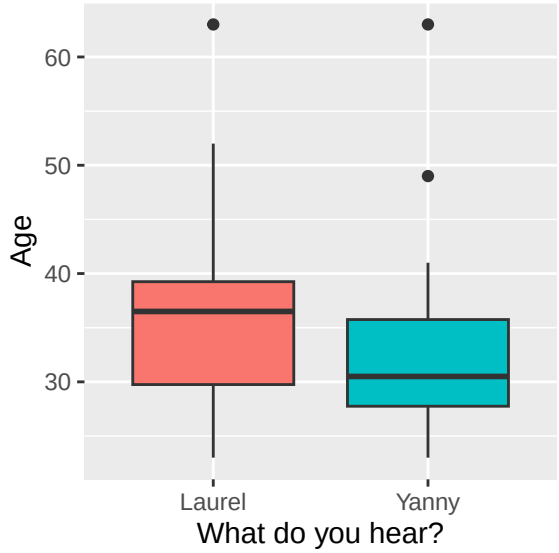


Figure 1: Hearing Yanny/Laurel by age.

We see in the boxplot that the people who hear Yanny are, on average, younger, however there is some overlap in the IQR's.

```
yanny %>%
  tabyl(gender, hear) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
gender     Laurel      Yanny
Female 50.0% (14) 50.0 (14)
  Male 56.0% (14) 44.0 (11)
```

```
ggplot(data = yanny, aes(x = hear, group = gender)) +
  geom_bar(aes(y = ..prop.., fill = gender),
           stat = "count", position = "dodge") +
  labs(x = "What do you hear?", y = "Proportion")
```

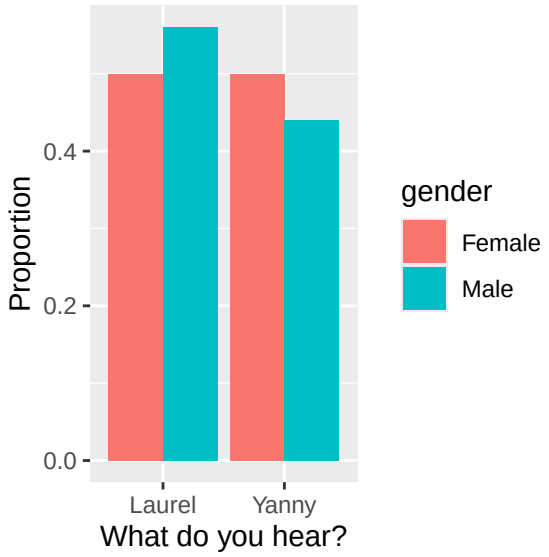


Figure 2: Barplot of what participants heard by gender.

There is a slightly smaller proportion of men hearing `Yanny`, but the proportions are very similar overall.

```
mod.yanny <- glm(hear ~ age + gender, data = yanny,family="binomial")

mod.yanny %>%
  tab_model(transform = NULL)
```

		hear	
Predictors	Log-Odds	CI	p
(Intercept)	1.63	-0.72 – 4.24	0.191
age	-0.05	-0.12 – 0.01	0.155
gender [Male]	-0.21	-1.34 – 0.91	0.717
Observations	52		
R^2 Tjur	0.045		

Notice that the coefficient of age is negative, suggesting that older people are less likely to hear Yanny. However, the coefficient of age is not significant (p-value of 0.16). Still, if we wanted to use the estimated coefficient to quantify the effect of age, we would need to look at exp(-0.05) = 0.95. This suggests that for two people who differ by one year in age, the older person's odds of hearing Yanny are 0.95 times those of the younger person. If we want to look at a ten-year age difference then the odds multiplier becomes exp(-0.05 * 10) = 0.6. Hence, for two people who differ by 10 years in age, the older person's odds of hearing Yanny are 0.6 times those of the younger person. Conversely, the odds for a younger person of hearing Yanny would be `exp(-0.05 * 10)^-1 = 1.65` times the odds (64% higher odds) of hearing Yanny than 10 year older person.

```
plot_model(mod.yanny, show.values = TRUE,
           title = "Odds (Age)", show.p = TRUE)
```

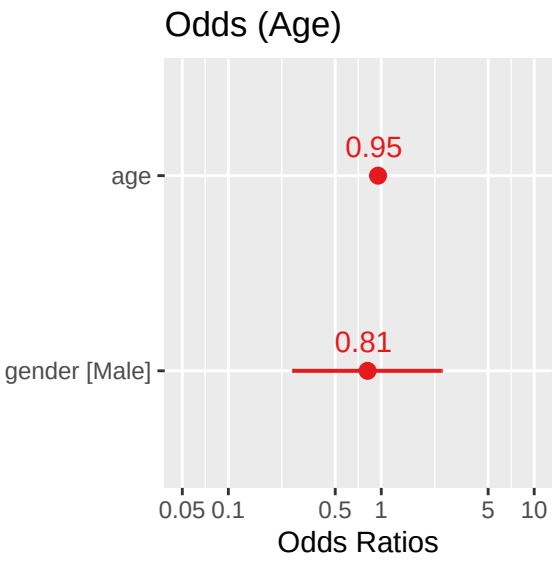


Figure 3: Odds of hearing yanny with age.

```
plot_model(mod.yanny,
           type = "pred",
           terms = "age",
           title = "",
           axis.title = c("Age", "Probability of hearing Yanny"))
```

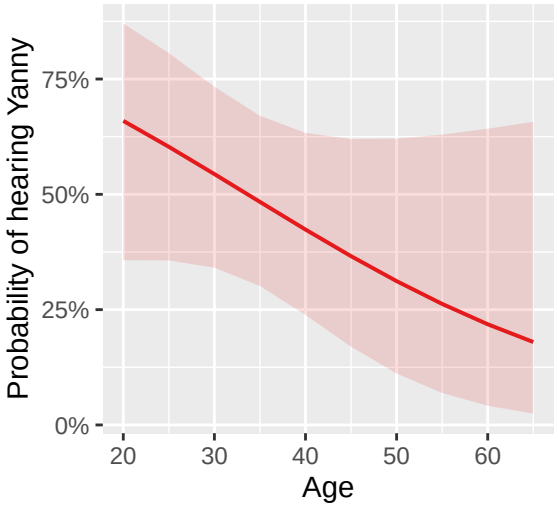


Figure 4: Probability of hearing Yanny with age.

2 Case 2: Surviving the Titanic

On 15th April 1912, during its maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Task 2

Download the data (`titanic.csv`) for $n = 891$ passengers aboard the Titanic and fit a logistic regression model with `survived` as the binary response variable, and `age`, `gender`, and `passenger.class` as the explanatory variables. What are your findings?
See Solution
Load the data

```
titanic <- read.csv("titanic.csv",stringsAsFactors = T)
titanic <- titanic %>%
          select(survived, age, gender, passenger.class) %>%
  mutate(passenger.class = as.factor(passenger.class),
         survived_fct = as.factor(survived))
levels(titanic$survived_fct) <- c("Died", "Survived")
```

```
ggplot(data = titanic, aes(x = survived_fct, y = age, fill = survived_fct)) +
  geom_boxplot() +
  labs(x = "Survived the Titanic?", y = "Age") +
  theme(legend.position = "none")
```

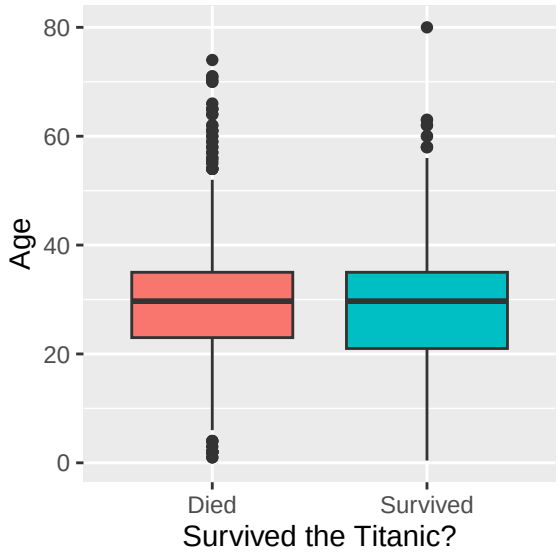


Figure 5: Titanic passenger age by survival.

We see in the boxplot that there is very little difference in the age of passengers who died or survived the sinking of the Titanic.

```
titanic %>%
  tabyl(gender, survived_fct) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
gender        Died    Survived                                    (233)
female 25.8%  (81) 74.2                                          (109)
 male 81.1% (468) 18.9
```

```
ggplot(data = titanic, aes(x = survived_fct, group = gender)) +
  geom_bar(aes(y = ..prop.., fill = gender), stat = "count", position = "dodge") +
  labs(x = "Survived the Titanic?", y = "Proportion")
```

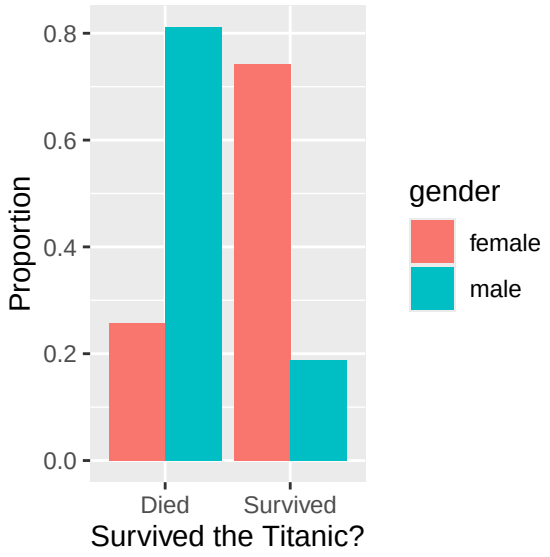


Figure 6: Barplot of passenger survival by gender.

There is a clear pattern here with the proportion surviving much higher for females than for males.

```
titanic %>%
  tabyl(passenger.class, survived) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
passenger.class           0           1
              1 37.0%  (80) 63.0%  (136)
              2 52.7%  (97) 47.3%  (87)
              3 75.8% (372) 24.2%  (119)
```

```
ggplot(data = titanic, aes(x = survived_fct, group = passenger.class)) +
  geom_bar(aes(y = ..prop.., fill = passenger.class),
           stat = "count", position = "dodge") +
  labs(x = "Survived the Titanic?", y = "Proportion")
```

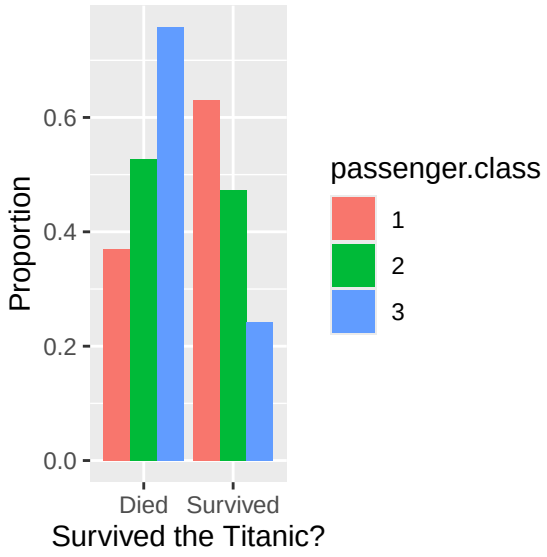


Figure 7: Barplot of passenger survival by gender.

The largest group of passengers who died were third class passengers, while among those who survived the largest group was first class passengers.
Now we focus on the event Probability$(\text{Survival} = 1)$. We will as our response the binary variable `survived` which takes the value of 1 if the i*th* passenger survived and zero othewise.

```
mod.titanic <- glm(survived ~ gender + passenger.class + age, data = titanic)
mod.titanic %>%
  tab_model(transform = NULL)
```

		survived	
Predictors	Estimates	CI	p
(Intercept)	1.10	1.00 – 1.19	**<0.001**
gender [male]	-0.50	-0.55 – -0.45	**<0.001**
passenger class [2]	-0.18	-0.26 – -0.10	**<0.001**
passenger class [3]	-0.37	-0.44 – -0.30	**<0.001**
age	-0.01	-0.01 – -0.00	**<0.001**
Observations	891		
R^2	0.383		

We see that the coefficient for males (`gendermale`) is negative, indicating a lower chance of survival for male passengers. Similarly, the coefficients for second (`passenger.class2`) and third (`passenger.class3`) class passengers are negative, with the magnitude of the third class coefficient larger than that of the second class coefficient. This suggests that second class passengers chances of survival were worse in comparison with first class passengers, and that third class passengers chances of survival were even worse. Finally the `age` coefficient is negative, suggesting that older people were less likely to survive.

```
plot_model(mod.titanic,
           transform = "exp",
           show.values = TRUE,
           title = "",
           show.p = FALSE,
           value.offset = 0.25)
```

Figure 8: Odds of surviving the sinking of the Titanic.

We interpret the odds ratios as follows: men's odds of survival were 0.61 times those of women, third class passengers' odds of survival were 0.7 times those of first class passengers, and second class passengers' odds of survival were 0.83 times those of first class passengers. Finally, for each year increase in the passenger's age, their odds of survival decrease (by a factor of 0.99).