

## lab7

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5    v purrr  0.3.4
## v tibble  3.1.6    v dplyr  1.0.7
## v tidyr   1.1.4    v stringr 1.4.0
## v readr   2.1.1    v forcats 0.5.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(knitr)
library(broom)
library(nnet)
```

Exercise 1:

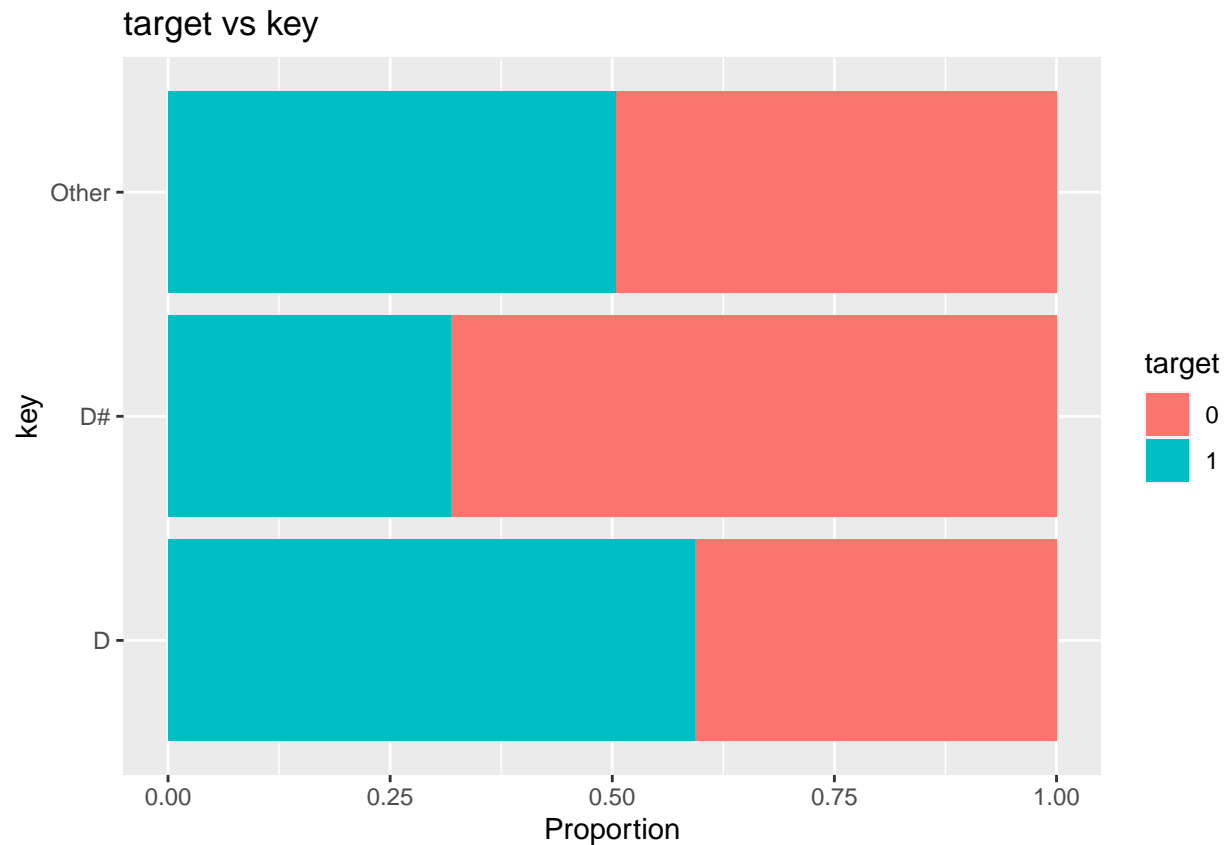
```
data <- read.csv("data.csv")
glimpse(data)
```

```
## Rows: 2,017
## Columns: 17
## $ X                <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ acousticness     <dbl> 0.010200, 0.199000, 0.034400, 0.604000, 0.180000, 0.0~
## $ danceability      <dbl> 0.833, 0.743, 0.838, 0.494, 0.678, 0.804, 0.739, 0.26~
## $ duration_ms       <int> 204600, 326933, 185707, 199413, 392893, 251333, 24140~
## $ energy            <dbl> 0.434, 0.359, 0.412, 0.338, 0.561, 0.560, 0.472, 0.34~
## $ instrumentalness  <dbl> 2.19e-02, 6.11e-03, 2.34e-04, 5.10e-01, 5.12e-01, 0.0~
## $ key               <int> 2, 1, 2, 5, 5, 8, 1, 10, 11, 7, 5, 10, 0, 0, 9, 6, 1,~
## $ liveness          <dbl> 0.1650, 0.1370, 0.1590, 0.0922, 0.4390, 0.1640, 0.207~
## $ loudness          <dbl> -8.795, -10.401, -7.148, -15.236, -11.648, -6.682, -1~
## $ mode              <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0,~
## $ speechiness       <dbl> 0.4310, 0.0794, 0.2890, 0.0261, 0.0694, 0.1850, 0.156~
## $ tempo              <dbl> 150.062, 160.083, 75.044, 86.468, 174.004, 85.023, 80~
## $ time_signature    <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4,~
## $ valence            <dbl> 0.286, 0.588, 0.173, 0.230, 0.904, 0.264, 0.308, 0.39~
## $ target            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ song_title        <chr> "Mask Off", "Redbone", "Xanny Family", "Master Of Non~
## $ artist            <chr> "Future", "Childish Gambino", "Future", "Beach House"~
```

```
data_lv <- data %>% mutate(target = as.factor(target), key = as.factor(ifelse(key == 2, "D", ifelse(key
glimpse(data_lv)
```

```
## Rows: 2,017
## Columns: 17
## $ X                <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ acousticness     <dbl> 0.010200, 0.199000, 0.034400, 0.604000, 0.180000, 0.0~
## $ danceability     <dbl> 0.833, 0.743, 0.838, 0.494, 0.678, 0.804, 0.739, 0.26~
## $ duration_ms      <int> 204600, 326933, 185707, 199413, 392893, 251333, 24140~
## $ energy           <dbl> 0.434, 0.359, 0.412, 0.338, 0.561, 0.560, 0.472, 0.34~
## $ instrumentalness <dbl> 2.19e-02, 6.11e-03, 2.34e-04, 5.10e-01, 5.12e-01, 0.0~
## $ key              <fct> D, Other, D, Other, Other, Other, Other, Other, Other~
## $ liveness         <dbl> 0.1650, 0.1370, 0.1590, 0.0922, 0.4390, 0.1640, 0.207~
## $ loudness         <dbl> -8.795, -10.401, -7.148, -15.236, -11.648, -6.682, -1~
## $ mode             <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0,~
## $ speechiness      <dbl> 0.4310, 0.0794, 0.2890, 0.0261, 0.0694, 0.1850, 0.156~
## $ tempo            <dbl> 150.062, 160.083, 75.044, 86.468, 174.004, 85.023, 80~
## $ time_signature   <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4,~
## $ valence          <dbl> 0.286, 0.588, 0.173, 0.230, 0.904, 0.264, 0.308, 0.39~
## $ target           <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ song_title       <chr> "Mask Off", "Redbone", "Xanny Family", "Master Of Non~
## $ artist           <chr> "Future", "Childish Gambino", "Future", "Beach House"~
```

```
p <- ggplot(data = data_lv, aes(x = key, fill = target)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion",
       title = "target vs key") +
  coord_flip()
p
```



Exercise 2:

```
target_m_red <- glm(target ~ acousticness + danceability + duration_ms + instrumentalness + loudness + speechiness + valence,
  data = data_lv, family = binomial)
tidy(target_m_red, conf.int = TRUE, exponentiate = FALSE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.955	0.276	-10.693	0	-3.504	-2.420
acousticness	-1.722	0.240	-7.182	0	-2.197	-1.257
danceability	1.630	0.344	4.737	0	0.958	2.308
duration_ms	0.000	0.000	4.225	0	0.000	0.000
instrumentalness	1.353	0.207	6.549	0	0.952	1.763
loudness	-0.087	0.017	-5.062	0	-0.122	-0.054
speechiness	4.072	0.583	6.985	0	2.947	5.234
valence	0.856	0.223	3.836	0	0.420	1.296

Exercise 3:

```
target_m_full <- glm(target ~ acousticness + danceability + duration_ms + instrumentalness + loudness + speechiness + valence,
  data = data_lv, family = binomial)
```

```
(dev_red <- glance(target_m_red)$deviance)
```

```
## [1] 2518.517
```

```
(dev_full <- glance(target_m_full)$deviance)
```

```
## [1] 2505.16
```

```
(test_stat <- dev_red - dev_full)
```

```
## [1] 13.357
```

```
1 - pchisq(test_stat, 8)
```

```
## [1] 0.1001425
```

There is not significant evidence to suggest that the response depends on the key variable. A p value of .1 is generally very high. But on top of that, the drop-in-deviance statistic suggests that even if the response depends on the key, the key will not improve the model by that much at all. Given a high p-value and low deviance statistic when comparing the model with key and the model without key, I conclude that the reduced model (without key) would be the best. approach.

Exercise 4:

```
model <- target_m_red  
tidy(model, conf.int = TRUE, exponentiate = FALSE) %>%  
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.955	0.276	-10.693	0	-3.504	-2.420
acousticness	-1.722	0.240	-7.182	0	-2.197	-1.257
danceability	1.630	0.344	4.737	0	0.958	2.308
duration_ms	0.000	0.000	4.225	0	0.000	0.000
instrumentalness	1.353	0.207	6.549	0	0.952	1.763
loudness	-0.087	0.017	-5.062	0	-0.122	-0.054
speechiness	4.072	0.583	6.985	0	2.947	5.234
valence	0.856	0.223	3.836	0	0.420	1.296

It is not appropriate to interpret key D# because it is not included in the model.