# MODELLING GDPR VIOLATIONS WITH TIDY MODELS

christopher okoth

4/30/2020

## EXPLORE THE DATA

```
# gdpr_violations <- readr::read_tsv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mas
gdpr_violations <-read.csv("gdpr_fines.csv")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'tibble' was built under R version 3.6.3

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

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

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

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
# gdpr_violations %<>% rename(country=name)
# gdpr_violations %<>% mutate(date=mdy(date))
# gdpr_violations %<>%mutate(date=na_if(date,"1970-01-01"))#possibly will leave these ones
```

### some notes

**Article 5:** principles for processing personal data (legitimate purpose ) **Article 6:** lawful processing of personal data ie consent etc **Article 13** inform subject if personal data is collected **Article 15:** right of access of data by subject **Article 32:** security of data processing (breach) - you have to process people's data securely

```
gdpr_violations %>% count(article_violated,sort = T) %>% top_n(10) %>% knitr::kable(align = "c")#the mo
```

```
## Selecting by n
```

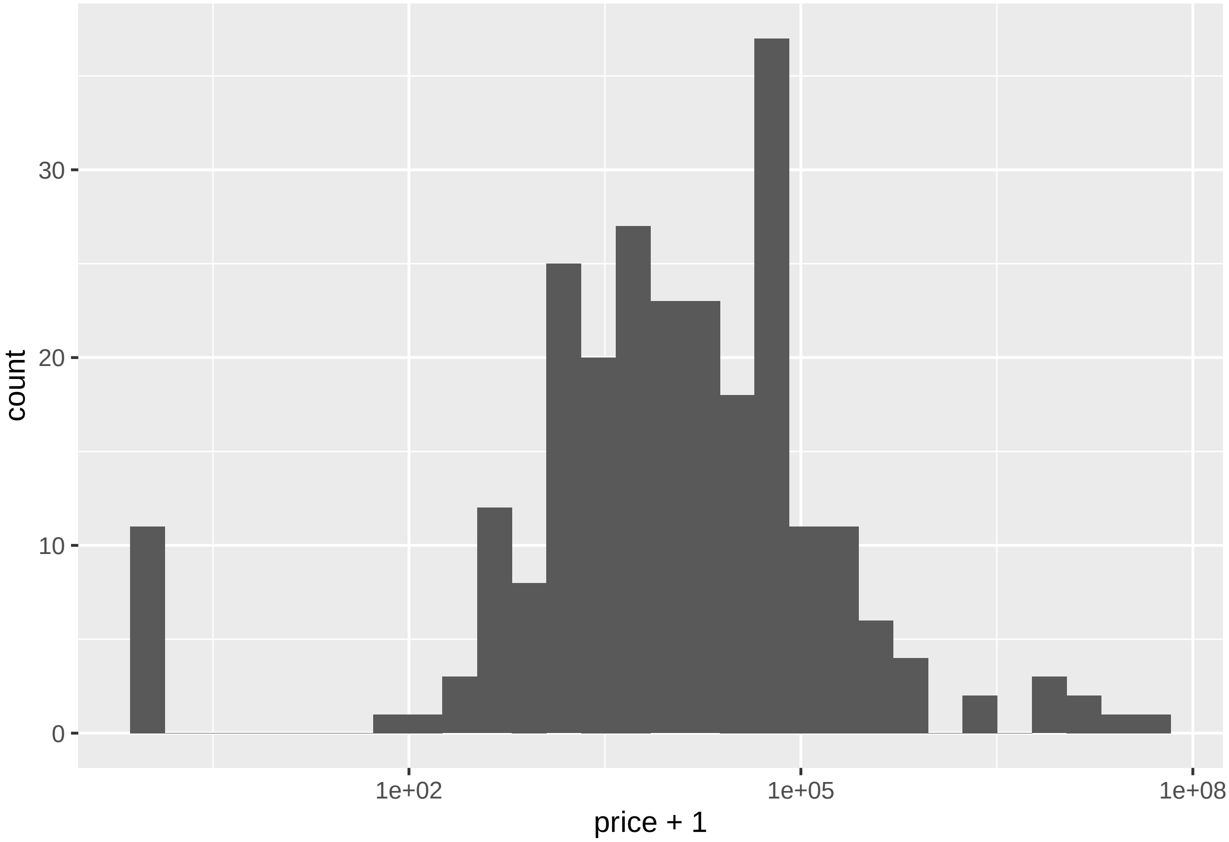
| article_violated | n |
|:---:|:---:|
| Art. 32 GDPR | 41 |
| Art. 6 GDPR | 33 |
| Art. 5 GDPR\|Art. 6 GDPR | 20 |
| Art. 15 GDPR | 10 |
| Art. 5 (1) f) GDPR\|Art. 32 GDPR | 10 |
| Art. 5 GDPR | 10 |
| Art. 13 GDPR | 7 |
| Art. 5 (1) f) GDPR | 7 |
| Art. 5 (1) a) GDPR\|Art. 6 GDPR | 6 |
| Art. 5 (1) c) GDPR | 6 |

```
gdpr_violations %>% separate_rows(article_violated,sep = "\\|") %>% count(article_violated,sort = T)#th
```

```
## # A tibble: 65 x 2
##    article_violated       n
##    <chr>              <int>
##  1 Art. 6 GDPR           82
##  2 Art. 32 GDPR          60
##  3 Art. 5 GDPR           46
##  4 Art. 13 GDPR          17
##  5 Art. 5 (1) f) GDPR    17
##  6 Art. 5 (1) a) GDPR    16
##  7 Art. 5 (1) c) GDPR    16
##  8 Art. 15 GDPR          15
##  9 Art. 21 GDPR           8
## 10 Art. 6 (1) GDPR        8
## # ... with 55 more rows
```

```
gdpr_violations %>% ggplot(aes(price+1))+geom_histogram()+scale_x_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



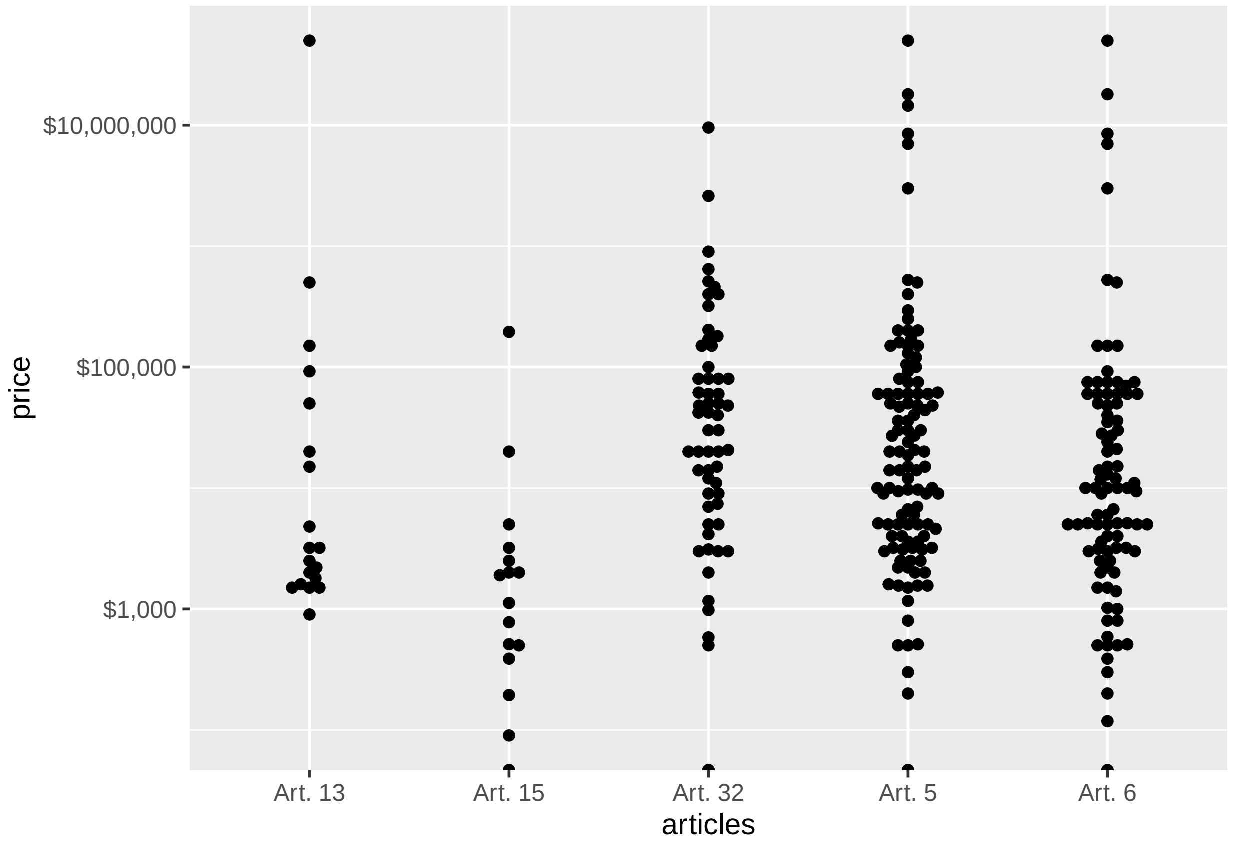
```
gdpr_cleaned <- gdpr_violations %>% transmute(id,country,price,
                              article_violated,
                              articles=str_extract_all(article_violated,pattern = "Art. \\d+|Art.\\d+"))
  mutate(total_articles =map_int(articles,length)) %>% #parse the column to map and return an integer l
  unnest(articles) %>%add_count(articles) %>% filter(n>10) %>%
  select(-n)#basically just remove that new column
#the data is now not in one violation per row but in article per row
```

```
library(ggbeeswarm)
```

```
## Warning: package 'ggbeeswarm' was built under R version 3.6.3
```

```
gdpr_cleaned %>% ggplot(aes(articles,price))+
  geom_beeswarm(priority = "random")+scale_y_log10(labels=scales::dollar_format())
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```



```
gdpr_cleaned %>%mutate(articles=str_replace_all(articles,pattern = "Art. ","Article") ,
                        articles=fct_reorder(articles,price)) %>% #the default function that the reoder 
  ggplot(aes(articles,price,color=articles))+geom_boxplot(alpha=0.2)+
  geom_quasirandom()+scale_y_log10(labels=scales::dollar_format())+ggdark::dark_theme_gray()+theme(legen
  ggtitle("Article 32 seems to have the largest fines associated with it ")
```

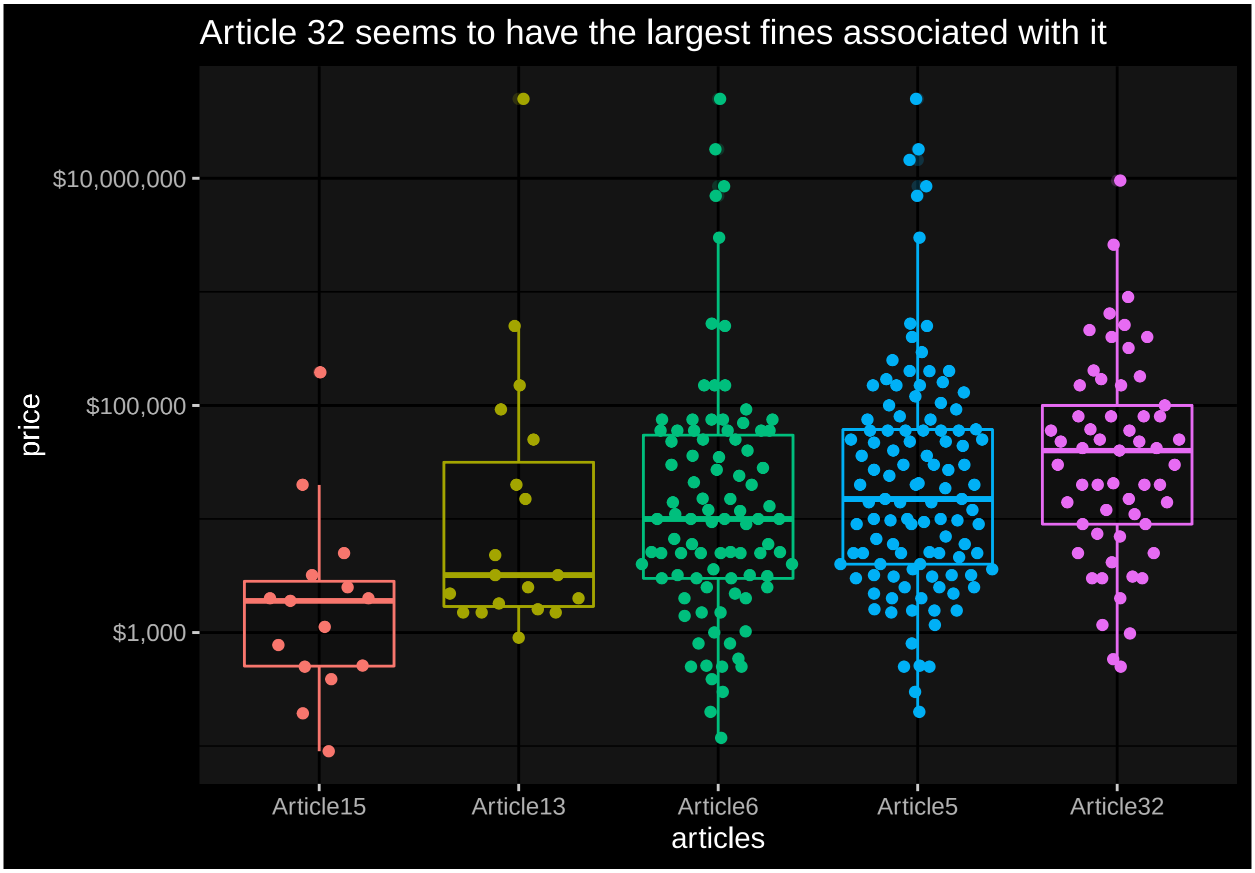
```
## Inverted geom defaults of fill and color/colour.
## To change them back, use invert_geom_defaults().

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 12 rows containing non-finite values (stat_boxplot).

## Warning: Removed 12 rows containing missing values (geom_point).
```



```r
gdpr_cleaned %>% mutate(value=1) %>% select(-article_violated) %>%
  pivot_wider(names_from = articles,values_from = value,values_fn = list(value = min),values_fill = lis
```

**do we have evidence that violating multiple articles is associated with higher fines**

## BUILD THE MODEL

```r
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## -- Attaching packages -----------------------------------------------------------------------
```

```
## v broom     0.5.6      v rsample   0.0.6
## v dials     0.0.6      v tune      0.1.0
## v infer     0.5.1      v workflows 0.1.1
## v parsnip   0.1.0      v yardstick 0.0.6
## v recipes   0.1.10
```

```
## Warning: package 'dials' was built under R version 3.6.3
```

```
## Warning: package 'scales' was built under R version 3.6.3
```

```
## Warning: package 'infer' was built under R version 3.6.3
```

```
## Warning: package 'parsnip' was built under R version 3.6.3
## Warning: package 'recipes' was built under R version 3.6.3
```

```
## Warning: package 'rsample' was built under R version 3.6.3
```

```
## Warning: package 'tune' was built under R version 3.6.3
```

```
## Warning: package 'workflows' was built under R version 3.6.3
```

```
## Warning: package 'yardstick' was built under R version 3.6.3
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x scales::discard()     masks purrr::discard()
## x magrittr::extract()   masks tidyr::extract()
## x dplyr::filter()       masks stats::filter()
## x recipes::fixed()      masks stringr::fixed()
## x dplyr::lag()          masks stats::lag()
## x dials::margin()       masks ggplot2::margin()
## x magrittr::set_names() masks purrr::set_names()
## x yardstick::spec()     masks readr::spec()
## x recipes::step()       masks stats::step()
```

```r
gdpr_recipe <- recipe(price~.,data=gdpr_articles) %>%
  step_other(country) %>%
  update_role(id,new_role = "id") %>%
  step_dummy(all_nominal())
```

```r
gdpr_prep <- prep(gdpr_recipe)
juice(gdpr_prep)
```

```
## # A tibble: 219 x 14
##       id total_articles art_13 art_5 art_6 art_32 art_15  price
##    <int>          <int>  <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>
##  1     2              4      1     1     1      0      0   2500
##  2     3              2      0     1     1      0      0  60000
##  3     5              1      0     0     0      1      0 150000
##  4     6              2      0     0     0      1      0  20000
##  5     7              2      0     1     0      0      0 200000
##  6     9              2      0     1     1      0      0  30000
##  7    10              2      0     1     1      0      0   9000
##  8    11              3      0     0     0      0      1 195407
##  9    12              1      0     1     0      0      0  10000
## 10    13              1      0     0     0      1      0 644780
## # ... with 209 more rows, and 6 more variables:
## #   country_Czech.Republic <dbl>, country_Germany <dbl>,
## #   country_Hungary <dbl>, country_Romania <dbl>, country_Spain <dbl>,
## #   country_other <dbl>
```

```r
gdpr_workflow <- workflow() %>% add_recipe(gdpr_recipe) %>%
  add_model(linear_reg() %>% set_engine("lm"))
```

# EXPLORE THE RESULTS

```r
#then we use the fit dunction to fit the model using the wflow
gdpr_workflow %>% fit(data=gdpr_articles)->gdpr_fit
#since the above is a workflow object we have to pull stuf out of it
gdpr_fit %>% pull_workflow_fit() %>% tidy() %>% filter(p.value<0.5)
```

```
## # A tibble: 5 x 5
##   term             estimate std.error statistic p.value
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     -1200240.  1254706.    -0.957  0.340
## 2 total_articles   1229050.   508274.     2.42   0.0165
## 3 art_15           -996666.  1427277.    -0.698  0.486
## 4 country_Germany  1247605.  1285475.     0.971  0.333
## 5 country_other     826264.  1089648.     0.758  0.449
```

The more articles one violates the higher the fines one pays And those who violate article 15 get the highest fines

### prediction on new data

```
new_data <- crossing(country="Other",
                     art_5=0:1,
                     art_15=0:1,
                     art_6=0:1,
                     art_32=0:1,
                     art_13=0:1) %>% mutate(total_articles=art_5+art_15+art_6+art_32+art_13,id=row_numbe
new_data
```

```
## # A tibble: 32 x 8
##    country art_5 art_15 art_6 art_32 art_13 total_articles    id
##    <chr>   <int>  <int> <int>  <int>  <int>          <int> <int>
##  1 Other       0      0     0      0      0              0     1
##  2 Other       0      0     0      0      1              1     2
##  3 Other       0      0     0      1      0              1     3
##  4 Other       0      0     0      1      1              2     4
##  5 Other       0      0     1      0      0              1     5
##  6 Other       0      0     1      0      1              2     6
##  7 Other       0      0     1      1      0              2     7
##  8 Other       0      0     1      1      1              3     8
##  9 Other       0      1     0      0      0              1     9
## 10 Other       0      1     0      0      1              2    10
## # ... with 22 more rows
```