Analysis FAA Wildlife strike

NewMirai / @alangel12407606

2019-07-24

Tidy Tuesday

For this 30th edition in 2019 of **Tidy Tuesday** we have a dataset coming from the FAA Wildlife Strike database to look at. A report on the full dataset can be found **here**.

Source

Full data dictionary

EDA

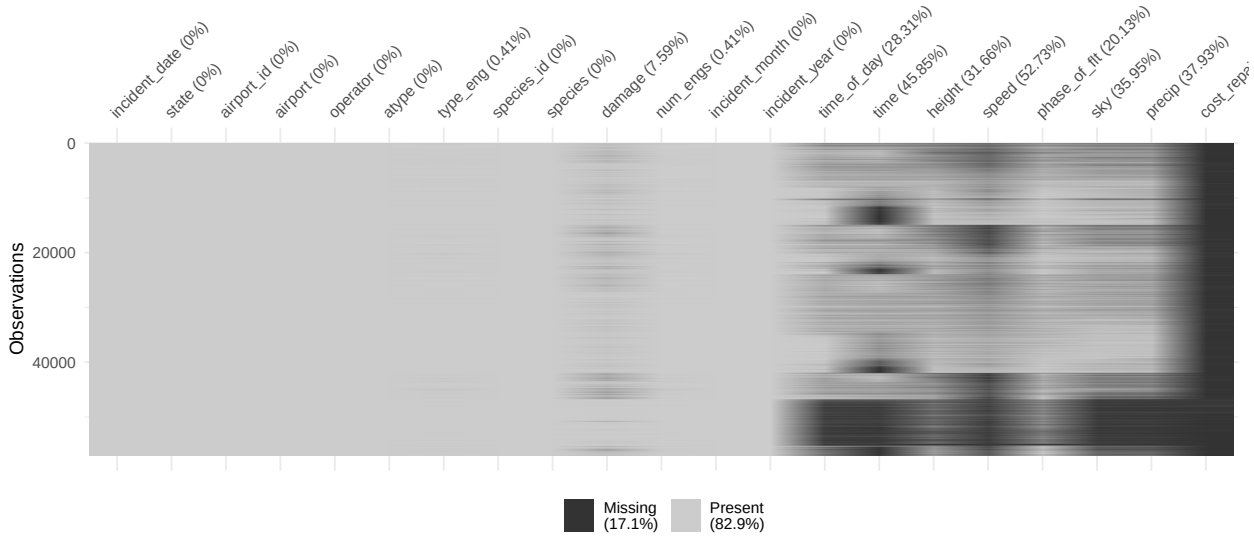
First we need to import some basics libraries in order to observe the dataset.

```
library(tidyverse)
library(visdat)
library(skimr)
library(RColorBrewer)
library(glue)
```

Then, we get the data from this **link**

```
wildlife_impacts <- readr::read_csv(link)
```

We can first, check if we have missing values



Interesting. It seems that most of the dataset is complete but **17,1%** of missing values is not negligible! We have to be careful about the choices of our variables.

Let's see a different reprenetation of our data using the `skirmr` library.

```
skim_to_list(wildlife_impacts)
```

```
## $character
## # A tibble: 13 x 8
##    variable     missing complete n     min   max   empty n_unique
##  * <chr>        <chr>   <chr>    <chr> <chr> <chr> <chr> <chr>
##  1 airport      0       56978    56978 4     53    0     423
##  2 airport_id   0       56978    56978 3     5     0     423
##  3 atype        0       56978    56978 4     18    0     91
##  4 damage       4324    52654    56978 1     2     0     4
##  5 operator     0       56978    56978 15    18    0     4
##  6 phase_of_flt 11469   45509    56978 4     12    0     24
##  7 precip       21614   35364    56978 3     10    0     7
##  8 sky          20485   36493    56978 8     10    0     3
##  9 species      0       56978    56978 4     50    0     527
## 10 species_id   0       56978    56978 1     6     0     528
## 11 state        0       56978    56978 2     3     0     59
## 12 time_of_day  16133   40845    56978 3     5     0     4
## 13 type_eng     234     56744    56978 1     1     0     4
##
## $numeric
## # A tibble: 7 x 12
##   variable missing complete n     mean  sd    p0    p25   p50   p75   p100
## * <chr>    <chr>   <chr>    <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 cost_re~ 56363   615      56978 "242~ 9426~ 11    5128  26783 9312~ "   ~
## 2 height   18038   38940    56978 "   ~ "  2~ 0     0     50    " 10~ "250~
## 3 inciden~ 0       56978    56978 "   ~ "   ~ 1     5     8     "   ~ "   ~
## 4 inciden~ 0       56978    56978 "  2~ "   ~ 1990  2001  2009  " 20~ " 20~
## 5 num_engs 233     56745    56978 "   ~ "   ~ 1     2     2     "   ~ "   ~
## 6 speed    30046   26932    56978 "   ~ "   ~ 0     130   140   "  1~ "  3~
## 7 time     26124   30854    56978 "  1~ "   ~ -84   930   1426  " 19~ " 23~
## # ... with 1 more variable: hist <chr>
##
## $POSIXct
## # A tibble: 1 x 8
##   variable     missing complete n     min       max       median   n_unique
## * <chr>        <chr>   <chr>    <chr> <chr>     <chr>     <chr>    <chr>
## 1 incident_da~ 0       56978    56978 1990-01-~ 2018-12-~ 2009-11~ 9880
```

We have a dataset with high dimensions 56978, 21. There is time-series data and geospatial data in this dataset so there is a lot of possibilities for the choice of plot.

Data wrangling

```
proportion <- wildlife_impacts %>%
  select(incident_date,operator,state,damage,time_of_day,atype,time_of_day,phase_of_flt,species,speed) %
  mutate(time_of_day=if_else(is.na(time_of_day),"Unknown",time_of_day),
         damage=case_when(
           damage=="M"~ "Minor",
           damage=="M?"~ "Minor Uncertain",
           damage=="N"~ "None",
           damage=="S"~ "Substantial",
           damage=="D"~ "Destroyed",
           is.na(damage)~ "Unknown Damage"
```

```
        ))

nonedmg <- 1.83+2.53+19.63+24.15+37.25

p1 <- proportion %>%
  ggplot(aes(x=time_of_day,fill=time_of_day))+
  geom_bar(aes(y=(..count..)/sum(..count..)))+
  scale_y_continuous(labels=scales::percent)+
  labs(x="Period of the day",
       y="Relative frequencies",
       title = "Incident per period of day with damage distinction",
       subtitle = glue("{nonedmg} % of incident implies None damage. Good News!!!\n\n"),
       fill="",
       caption = "Author: @alangel12407606 | #TidyTuesday")+
  geom_text(aes( label = paste(round(100*(..count..)/sum(..count..),digits = 2),"%"),
                 y= (..count..)/sum(..count..) ), stat= "count", vjust = -.5,size=3)+
  scale_fill_brewer(type = "qual",palette = "Set3")+
  theme_minimal()+
  facet_wrap(~damage)+
  theme(legend.position = c(0.85,0.25),
        plot.title = element_text(size=18,hjust = .5),
        plot.subtitle = element_text(size=14,hjust = .5),
        strip.background = element_rect(fill = "#f2f2f2"))

p1
```