STA304

Yuping Hao

04/02/2022

Abstract

Open and transparent data can be the key for people to understand a certain field, data leads people to the distant future. The data speaks for itself, and the results from the data can be good or bad, but the data is never deceptive. Reasonable use of data can bring people an unexpected harvest.

```r
library(opendatatoronto)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# get package
package <- show_package("9d11c7aa-7613-4d3e-95f3-a02e2b1aa2d7")
package
```

```
## # A tibble: 1 x 11
##   title         id        topics civic_issues publisher excerpt dataset_category
##   <chr>         <chr>     <chr>  <chr>        <chr>     <chr>   <chr>
## 1 Police Annua~ 9d11c7aa~ <NA>   <NA>         <NA>      <NA>    <NA>
## # ... with 4 more variables: num_resources <int>, formats <chr>,
## #   refresh_rate <chr>, last_refreshed <date>
```

```r
# get all resources for this package
resources <- list_package_resources("9d11c7aa-7613-4d3e-95f3-a02e2b1aa2d7")

# identify datastore resources; by default, Toronto Open Data sets datastore resource format to CSV for
datastore_resources <- filter(resources, tolower(format) %in% c('csv', 'geojson'))

# load the first datastore resource as a sample
data <- filter(datastore_resources, row_number()==1) %>% get_resource()
data
```

```
## # A tibble: 2,369 x 10
##    `_id` Index_ ReportedYear GeoDivision Category   Subtype  Count_ CountCleared
##    <int> <lgl>         <int> <chr>       <chr>      <chr>     <int>        <int>
##  1     1 NA             2014 D11         Controlle~ Other       201          195
##  2     2 NA             2014 D11         Crimes Ag~ Auto Th~    119           42
##  3     3 NA             2014 D11         Crimes Ag~ Break &~     85           37
##  4     4 NA             2014 D11         Crimes Ag~ Break &~     58            18
##  5     5 NA             2014 D11         Crimes Ag~ Break &~     89            34
##  6     6 NA             2014 D11         Crimes Ag~ Break &~     23             7
##  7     7 NA             2014 D11         Crimes Ag~ Fraud       232            83
##  8     8 NA             2014 D11         Crimes Ag~ Other       628           230
##  9     9 NA             2014 D11         Crimes Ag~ Theft O~     36            12
## 10    10 NA             2014 D11         Crimes Ag~ Theft U~   1774           790
## # ... with 2,359 more rows, and 2 more variables: ObjectId <int>,
## #   geometry <chr>
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.0
```

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

```
data1 = subset(data,select = -c(Index_, ObjectId, geometry)) %>%
  filter(2017<=ReportedYear & ReportedYear<=2020) %>%
  mutate(Totalcount = Count_ - CountCleared) %>%
  filter(0<Totalcount)
```

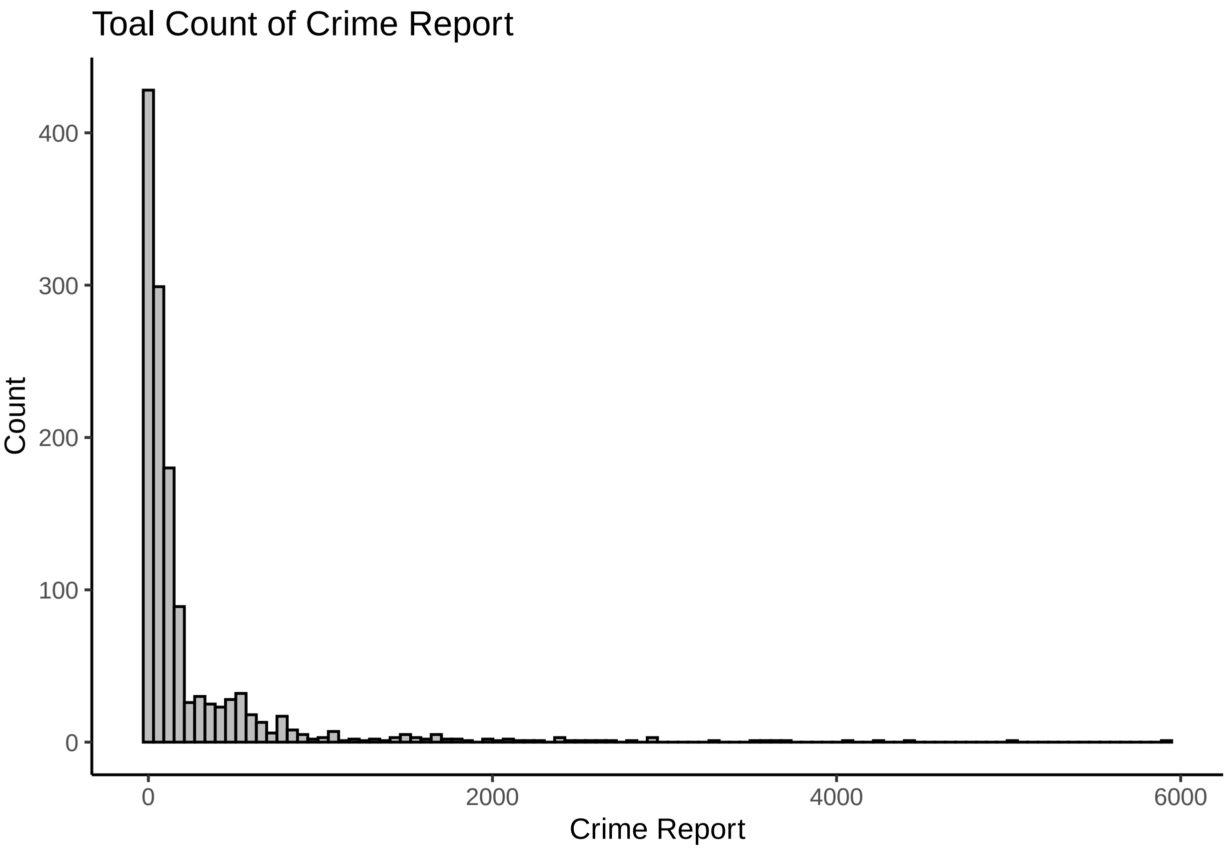
```
summary(data1)
```

```
##       _id        ReportedYear  GeoDivision          Category
##  Min.   : 801   Min.   :2017   Length:1299        Length:1299
##  1st Qu.:1356   1st Qu.:2017   Class :character   Class :character
##  Median :1690   Median :2018   Mode  :character   Mode  :character
##  Mean   :1677   Mean   :2018
##  3rd Qu.:2028   3rd Qu.:2019
##  Max.   :2369   Max.   :2020
##    Subtype              Count_        CountCleared      Totalcount
##  Length:1299        Min.   :   1.0   Min.   :   0.0   Min.   :   1.0
##  Class :character   1st Qu.:  67.0   1st Qu.:  18.0   1st Qu.:  12.0
##  Mode  :character   Median : 149.0   Median :  60.0   Median :  75.0
##                     Mean   : 412.8   Mean   : 167.6   Mean   : 245.2
##                     3rd Qu.: 414.5   3rd Qu.: 151.5   3rd Qu.: 194.0
##                     Max.   :7256.0   Max.   :2161.0   Max.   :5919.0
```

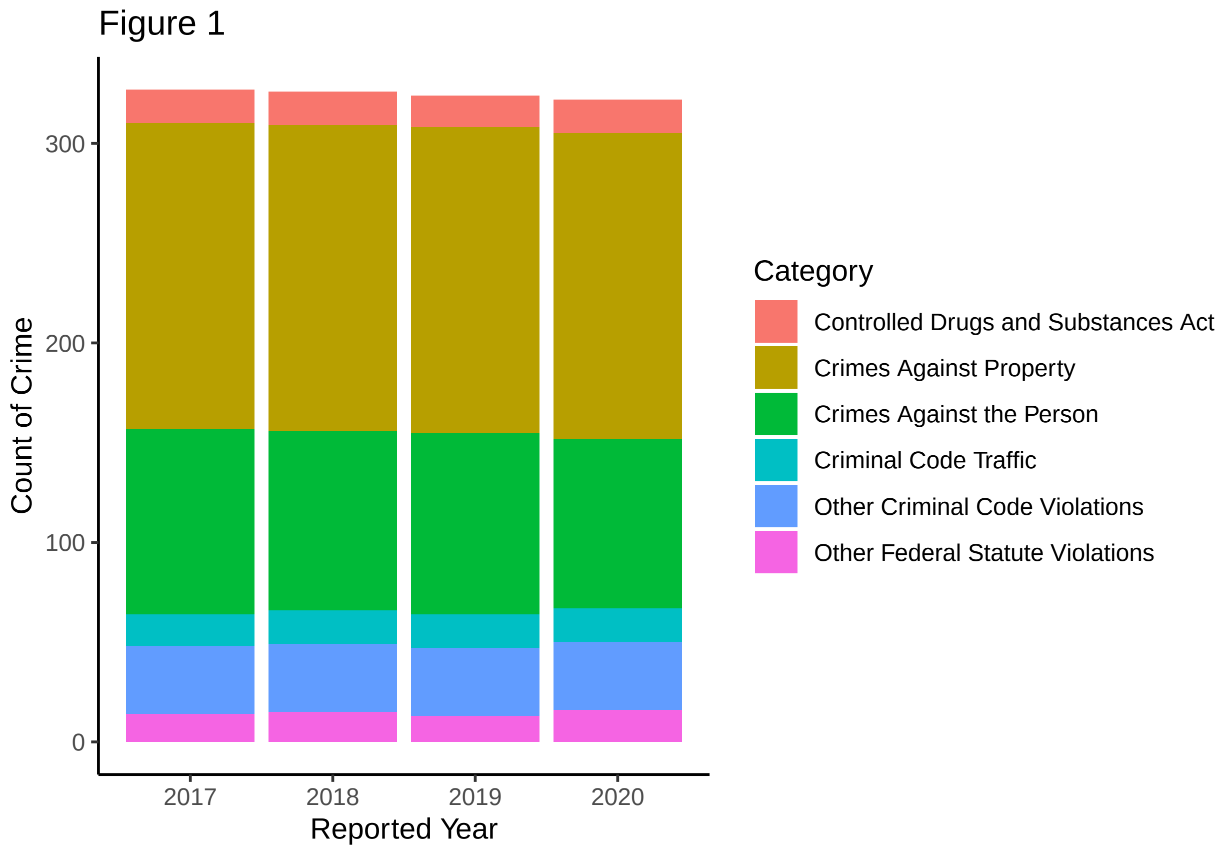
```
glimpse(data1)
```

```
## Rows: 1,299
## Columns: 8
## $ `_id`        <int> 801, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 81~
## $ ReportedYear <int> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 201~
## $ GeoDivision  <chr> "D22", "D22", "D22", "D22", "D22", "D22", "D22", "D22", "~
## $ Category     <chr> "Crimes Against Property", "Crimes Against Property", "Cr~
## $ Subtype      <chr> "Fraud", "Other", "Theft Over $5000", "Theft Under $5000"~
## $ Count_       <int> 728, 827, 85, 2434, 980, 7, 280, 19, 175, 128, 138, 1177,~
## $ CountCleared <int> 154, 267, 15, 672, 571, 5, 157, 15, 77, 66, 135, 1050, 66~
## $ Totalcount   <int> 574, 560, 70, 1762, 409, 2, 123, 4, 98, 62, 3, 127, 5, 4,~
```

```
data1 %>% ggplot(aes(x=Totalcount))+geom_histogram(fill="grey",color="black",
                                                    bins = 100) +
  theme_classic() + labs(x="Crime Report", y="Count",
                         title="Toal Count of Crime Report ")
```

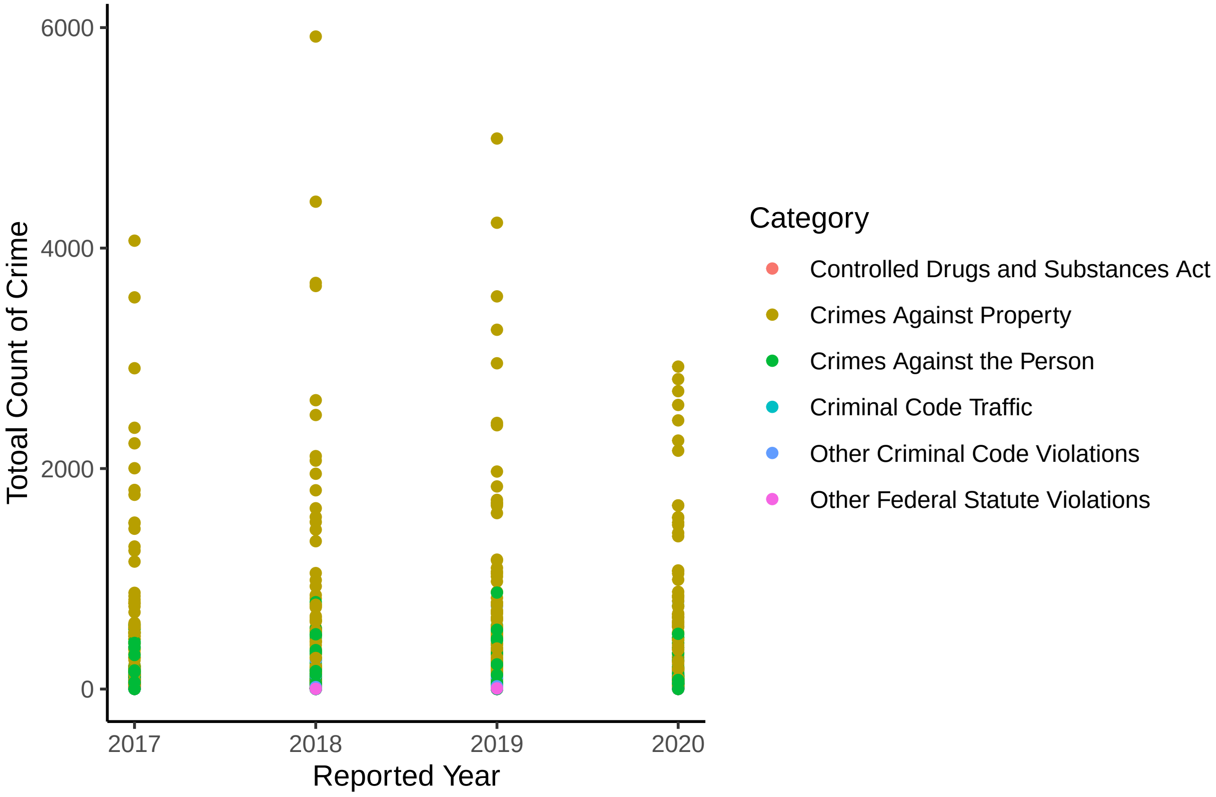


```
data1 %>% ggplot(aes(x=ReportedYear, fill=Category)) +
  geom_bar()+theme_classic()  +
  labs(x="Reported Year", y="Count of Crime", title="Figure 1")
```



```
data1 %>% ggplot(aes(x=ReportedYear, y=Totalcount, color=Category)) + geom_point() +
  theme_classic() + labs(x="Reported Year", y="Totoal Count of Crime", title =
                                          "Figure 2")
```

Figure 2



```
summary_table <- data1 %>%
  group_by(Category) %>%
  summarise(Min = min(Totalcount),
            Q1 = quantile(Totalcount,0.25),
            Median = median(Totalcount),
            Q3 = quantile(Totalcount,0.75),
            Max = max(Totalcount),
            sd = sd(Totalcount))
knitr::kable(summary_table)
```

Category	Min	Q1	Median	Q3	Max	sd
Controlled Drugs and Substances Act	1	5	9.0	12.00	27	5.447625
Crimes Against Property	1	57	140.5	512.00	5919	722.505335
Crimes Against the Person	1	30	86.0	146.50	875	148.452451
Criminal Code Traffic	1	3	5.0	8.50	20	4.312006
Other Criminal Code Violations	2	10	30.0	81.25	796	99.798202
Other Federal Statute Violations	1	1	2.0	4.00	15	2.699510

```
summary_table <- data1 %>%
  group_by(ReportedYear) %>%
  summarise(Min = min(Totalcount),
            Q1 = quantile(Totalcount,0.25),
            Median = median(Totalcount),
            Q3 = quantile(Totalcount,0.75),
            Max = max(Totalcount),
            sd = sd(Totalcount))
knitr::kable(summary_table)
```

ReportedYear	Min	Q1	Median	Q3	Max	sd
2017	1	10.00	69	167.00	4067	465.8102
2018	1	11.00	79	193.00	5919	606.1348
2019	1	12.75	83	220.50	4994	583.8484
2020	1	15.00	68	192.75	2925	464.1884

```
citation("tidyverse")
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostin
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
```

```
citation("knitr")
```

```
##
## To cite the 'knitr' package in publications use:
##
##   Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report
##   Generation in R. R package version 1.33.
##
##   Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
##   Chapman and Hall/CRC. ISBN 978-1498716963
##
##   Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
##   Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
##   Peng, editors, Implementing Reproducible Computational Research.
##   Chapman and Hall/CRC. ISBN 978-1466561595
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

```
citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```

Wickham et al. (2019)

Wickham et al. (2021)

Wickham (2016)

Nivette et al. (2021)

Gramlich (2020)

Casey (2021)

Moreau (2021)

Street (2019)

Greg Moreau and Armstrong (2020)

Service (2019)

Reference

Casey, Liam. 2021. "'I Was Actually Really Pissed':behind the Rise Fo Car Thefts Across Canada."

Gramlich, John. 2020. "What the Data Says(and Doesn't Say)about Crime in the United States."

Greg Moreau, Brianna Jaffray, and Amelia Armstrong. 2020. "Police-Reported Crime Statistics in Canada, 2019."

Moreau, Greg. 2021. "Police-Reported Crime Statistics in Canada, 2020."

Nivette, Amy E., Renee Zahnow, Raul Aguilar, Andri Ahven, Shai Amram, Barak Ariel, María José Arosemena Burbano, et al. 2021. "A Global Analysisi of the Impact of COVID-19 Stay-at-Home Restrictions on Crime."

Service, Toronto Police. 2019. "Annual Statistical Report."

Street, Brittany. 2019. "The Impact of Economic Activity on Criminal Behavior: Evidence from the Fracking Boom."

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.