# Traffic Collisions

Exploratory Data Analysis

Md Abrar Nasir

January 31, 2021

## Data

Sourced from the City of Toronto's Open Data Portal (City of Toronto, 2020), an exploratory data analysis on traffic collisions is pursued as the literature follows. The data set includes records of all collision occurrences which took place in Toronto from 2014 to 2019 and which have been documented by the Toronto Police Service or the Collision Reporting Center.

The data set was checked for nullity and rows containing any null entry were erased. Moreover, besides the `Category` column, which essentially had the same value (*Traffic Collisions*) for all the records, insignificant variables such as `_id` and `Index_` were also removed (Bhalla), for the simplification of the analysis.

The raw data included the collision counts - grouped by years when the incidents took place, geographic subdivisions where the incidents took place and even the subcategories of collisions how the incidents took place. For the purpose of this EDA, only counts of collisions, and categorical variables such as year, geographic divisions and subcategory of collisions are to be used.

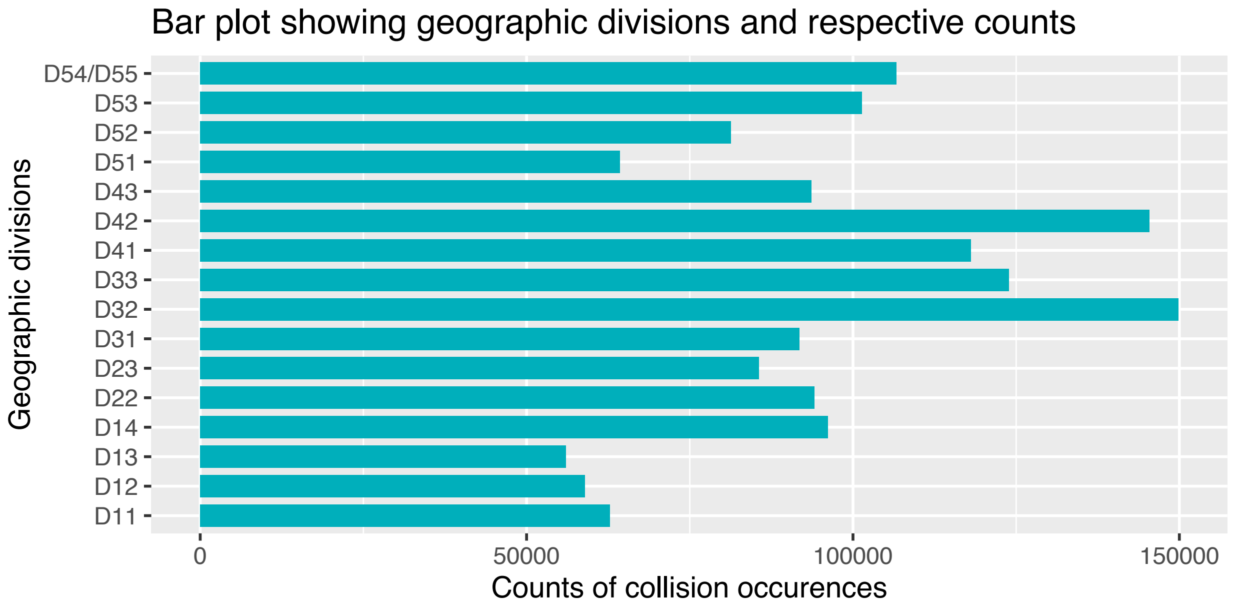
## Numerical analysis

In efforts to extract statistical measures (centres and spreads) from the sorted counts, the raw counts for each group have been aggregated (Kun, 2018) and summed up, and then categorised into the subparts of the category. This means, for example, for the *year* group, irrespective of other factors such as geographic division or subcategory of collision, figures for each year between 2014 and 2019 are to be yielded. These sorted values can then be used to compute numerical summaries such as the table (R Markdown :: Cheat Sheet) below:

| | Min | 1st Quartile | Median | Mean | Std. Deviation | 3rd Quartile | Max |
|---|---|---|---|---|---|---|---|
| By Geo. Division | 56046 | 77062 | 93877 | 95616 | 28488 | 109530 | 149864 |
| By Year | 259294 | 274154 | 290840 | 295126 | 29981 | 315310 | 337724 |
| By Subcategory | 774 | 61837 | 122900 | 247587 | 327473 | 370994 | 619088 |

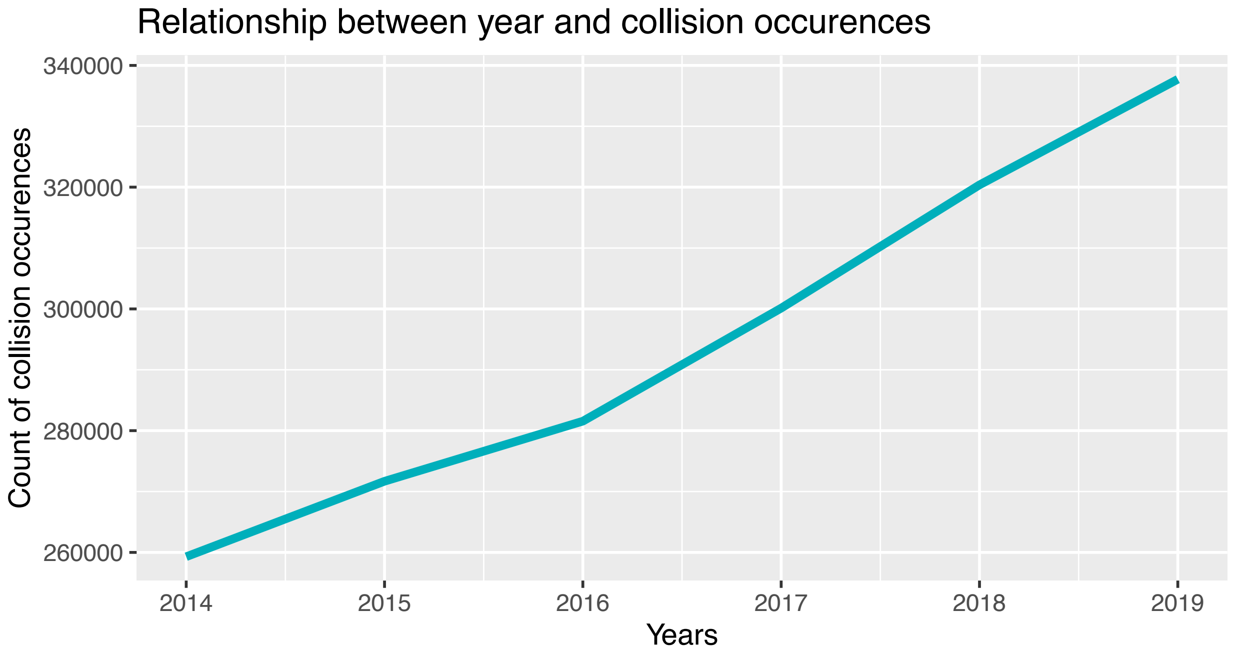
## Visualisations

In addition, the sorted data can be used to visualise and compare the counts of collisions based on geographic divisions, by means of a barplot (Holtz). In this case, `NAS` values or numbers of accidents that took place outside Toronto or those whose locations were not tracked properly, have been removed (Siddiqui, 2020) for a cleaner diagram without any sign of outlier.



Observing the plot above, it can be easily concluded *D32* and *D42* are the top two most dangerous regions with the highest numbers of collision counts, while D13 is the region with the least number of collisions that took place between 2014 and 2019.

Furthermore, a plot between years and accident counts can be presented, using our previously sorted values.



A clear positve relationship between years and count is observed from this illustration, which bears the suggestion that, as time passes, the probability of collisions occurring also increases, irrespective of any other factor.

## Conclusion

To reiterate, this EDA contains the export of numerical summaries based of traffic collisions data as well as shows how there is a supposed positive relationship between time in year and collision counts. It has also been advised that *D32* and *D42* have the highest numbers of collisions, D13 has the lowest.

## Technologies

All analysis for this report was programmed using `R version 4.3`, nicknamed as Bunny-Wunnies Freak Out, on a system running MacOS Big Sur. With `tidyverse` and `ggplot2`, the `opendatatoronto` package was used too, to import the data (Gelfand); and version control was done through GitHub.

## Bibliography

1. City of Toronto, O. (2020, November 18). Police Annual Statistical Report - Traffic Collisions. Retrieved January 31, 2021, from https://open.toronto.ca/dataset/police-annual-statistical-report-traffic-collisions/)

2. Kun, D. (2018, May 12). Aggregate – A Powerful Tool for Data Frame in R. Retrieved January 31, 2021, from https://datascienceplus.com/aggregate-data-frame-r/

3. Bhalla, D. (n.d.). R : Keep / Drop Columns from Data Frame. Retrieved January 31, 2021, from https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html

4. R Markdown :: Cheat Sheet. (n.d.). Retrieved January 31, 2021, from https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

5. Holtz, Y. (n.d.). Basic barplot with ggplot2. Retrieved January 31, 2021, from https://www.r-graph-gallery.com/218-basic-barplots-with-ggplot2.html

6. Siddiqui, N. (2020, August 24). How to remove last few rows from an R data frame? Retrieved January 31, 2021, from https://www.tutorialspoint.com/how-to-remove-last-few-rows-from-an-r-data-frame

7. Gelfand, S. (n.d.). Access the City of Toronto Open Data Portal. Retrieved January 31, 2021, from https://sharlagelfand.github.io/opendatatoronto/