Speed limit: more emphasis should be put on*

Wenxuan Li

06 February 2022

Abstract

Speed limit is important to the security of drivers and passengers in the road. However, recent drivers do not pay enough attention to the speed limit. A data set, 'About Mobile Watch Your Speed Program – Speed Summary' is adopted to see the status that the drivers care about the limit. The result is not positive and we appeal for more emphasis on the importance of speed limit.

1 Introduction

Speed limit is widely adopted in every corner of the world because it can greatly reduce the accidents in the road. However, the awareness of the importance for the speed limit may not catch enough attention from the drivers, which may bring seriously results. In this report, we will utilize a data set which records the vehicles' speed in the road of the City of Toronto Services (2022) to see if the government should emphasize more on the speed limit's importance. We focus on the median (i.e., 50%) of the speed since it represents the common speed. In the meantime, we are also curious whether the volume will influence the speed or not. Summary statistics and histograms are adopted as tools to analyze the data set.

In the following context, Section 2.1 provides the source of data, including the collection procedure. Section 2.2 talks about the possible bias issues in the process of data collection. Section 2.3 presents the numerical summary of the data set, and section 2.4 investigates the data set via graphs.

2 Data

In this section, we will summarize the data set we adopt thouroughly, including the source, collection methods, bias, numerical summary, graphical summary, and observations. Note that, we run our analysis in `R` (R Core Team 2020) and generates the report in `Rstudio`(RSudioTeam 2021). We also use the `tidyverse` and `dplyr` for data management, which was written by Wickham et al. (2019) and *A Grammar of Data Manipulation* (2021); `opendatatoronto` for data retrieval directly from the Internet, which was written by Gelf and Toronto (2020); and `knitr` for table generation, which was written by Xie (2021).

2.1 Data source and collection

The data set we adopt in this report is "About Mobile Watch Your Speed Program – Speed Summary," which is retrieved from Open Data Toronto Services (2022). As its name indicates, this data set is a summary of the data from a special program, called as "About Mobile Watch Your Speed Program." This program sets up devices (i.e., a radar device for speed measurement and a LED display for speed displaying) on the hydro pole or the streetlight to measure the speed of the passing vehicles to remind the drivers that they may exceed the speed limit. Though this program is supported by the government, the number of the device is still limited. Thus, they need to rotate their devices from spots to spots each month. Another thing we need to emphasize is that for the radar does not work 24 hours a day and 7 days a week, but only weekdays from 7 am to 9 pm.

*Code and data are available at: https://github.com/leoli2022/sta304.

According to the data collection procedure, we can see that the population is all roads in the city of Toronto; the frame population is the roads that have a hydro pole or a streetlight that can hold the radar and the LED display; and the sample population is the road that being selected to install the device set. The website does not mention how they select the spot to install the device set, but they are in a rotating scheme, which indicates that they are trying to install the device on each possible spot to collect data. Since people can install the device anywhere they want if applicable, there is no “non-response” problem.

As for the feature of the data set, the strength of the data set centers on its detailed summary on the speed data they collected. They provide the summary from two aspects – 1) the quantile of speeds; and 2) the number of vehicles recorded between a given small intervals (i.e., interval width = 5km/h). The weakness centers on the bias discussed as below.

2.2 Data bias

The bias of the data set is summarized as below. The first is the bias on the time and space. As indicated in the collection process that for each time period, only a proportion of the roads have records, which means that the record is not comprehensive from time and space, which may cause bias because the sampled raods and the sampled time periods, might not be representation. For example, the device only works on weekdays from 7 am to 8 pm, the remaining time periods have no records and they cannot be represented by the given time periods. The second thing is that when drivers see such an LED display far away (i.e., out of radar measuring range), they might be reminded and slow their speed if they exceed the limit speed, which may cause the recorded speed consistently smaller than the actual one.

2.3 Data summary

In the original data set, it contains 7528 pieces of samples and 52 variables for each sample. These 52 variables can be roughly divided into four categories.

- Identifiers. In this category, we have the sample id, the road id, ward number, locations, road names, direction, install data, removal date, recording time period, etc. These identifiers can help us locate the sample from both time and space aspects. They are all category variables.
- Speed variables. In this category, for each given recording spot and time period, the data set given the speeds of different percentile (i.e., 5%,10%,. . . ,95%) and the number of vehicles recorded in each speed interval (i.e., 0-5 km/h,5-10 km/h,. . . ,95-100 km/h, 100+km/h). Variables in this category are numerical variables.
- Volume. This volume does not represent the volume of the given road, but the total number of vehicles recorded in the given period. However, since this the recording time period is similar across different spots, we can still treat it as a “volume.” This is a numerical variable.

We will not use all of these variables in the further analysis. Instead, to check the vehicles’ speed in the city of Toronto, we clean the data set and select & create several variables of interest. These variables include the median speed, which is the 50% percentile of the recorded speed values and can be directly retrieved from the original data set; the proportion of the vehicles exceeds 40 km/h (i.e., the most common speed limit), which is summation of the number of vehicles in the speed intervals from 40-45 km/h to over 100 km/h divided by the volume; the proportion of the vehicles exceeding 100 km/h, which is computed by the number of vehicles exceeding 100 km/h divided by the volume; and the volume, which is changed to a categorical variable with five levels (i.e., 0~10000, 10001~30000, 30001~60000, 60001~80000, >80001).

Table 1 shows the summary of statistics for the variable of interests. From the table, we can observe that the speed limit is high violated by most vehicle drivers. From the median of vehicle speeds, the minimum is only 5 km/h, but the maximum can achieve 80 km/h, which is much faster than the speed limit. Fortunately, till the third quantile, the median speed is only 30 km/h. For the proportion of vehicles whose speeds are over 40 km/h, the median value is around 20%, but the maximum value is 1, which is out of expectation. For the proportion of vehicles whose speeds are over 100 km/h, the median value is 0 (which is perfect), but the

Table 1: Summary statistics for variable of interest

	Median	Over 40 km/h	Over 100 km/h
Minimum	5.00000	0.0000000	0.0000000
1st quantile	25.00000	0.0486272	0.0000000
Median	31.00000	0.1902335	0.0000000
Mean	32.43371	0.2975140	0.0003284
3rd quantile	39.00000	0.4909873	0.0000363
Maximum	80.00000	1.0000000	0.4074074

maximum value is around 40%. According to these observations, we can see that the speed limit, on some roads, are fully ignored.

2.4 Observations and discussion

In this section, we visualize the data to provide a straightforward understanding of vehicles' speeds. Figure 1 is the histogram for the median of speed. From the figure, we can see that most medians are all smaller than 40 km/h, only around 20% medians are larger than 40 km/h. This again indicates that on around 20% roads that drivers do not care about speed limits.

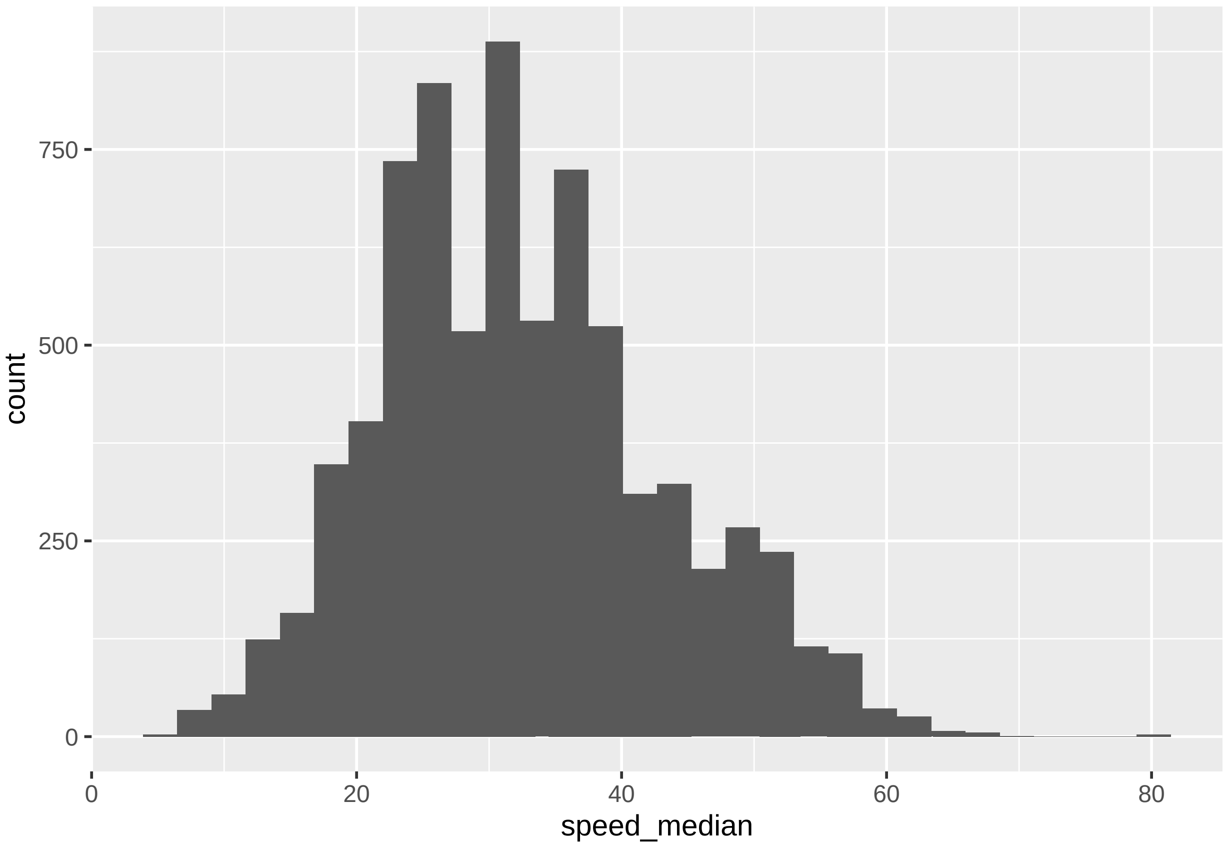

Figure 1: Medians of vehicles limits

Figure 2 shows the histogram of the proportion of vehicles that exceeds 40 km/h for different volume categories. Different colors represent different volume level. The red one yields for the smallest volume (i.e., 0~10000), and the pink one represents the highest volume (i.e., >80000). From the figure, we observe that the first three volume levels (i.e., 0~10000, 10001~30000, and 30001~60000) has an obvious decreasing trend on the number of spots whose vehicles exceed 40 km/h as the proportion increases. But for the remaining volume levels, the count of spots does not greatly vary with proportion of vehicles that exceed 40 km/h. This observation indicates that the busier the road is, the less the driver will pay attention to the speed limit. This is quite out of my expectation.

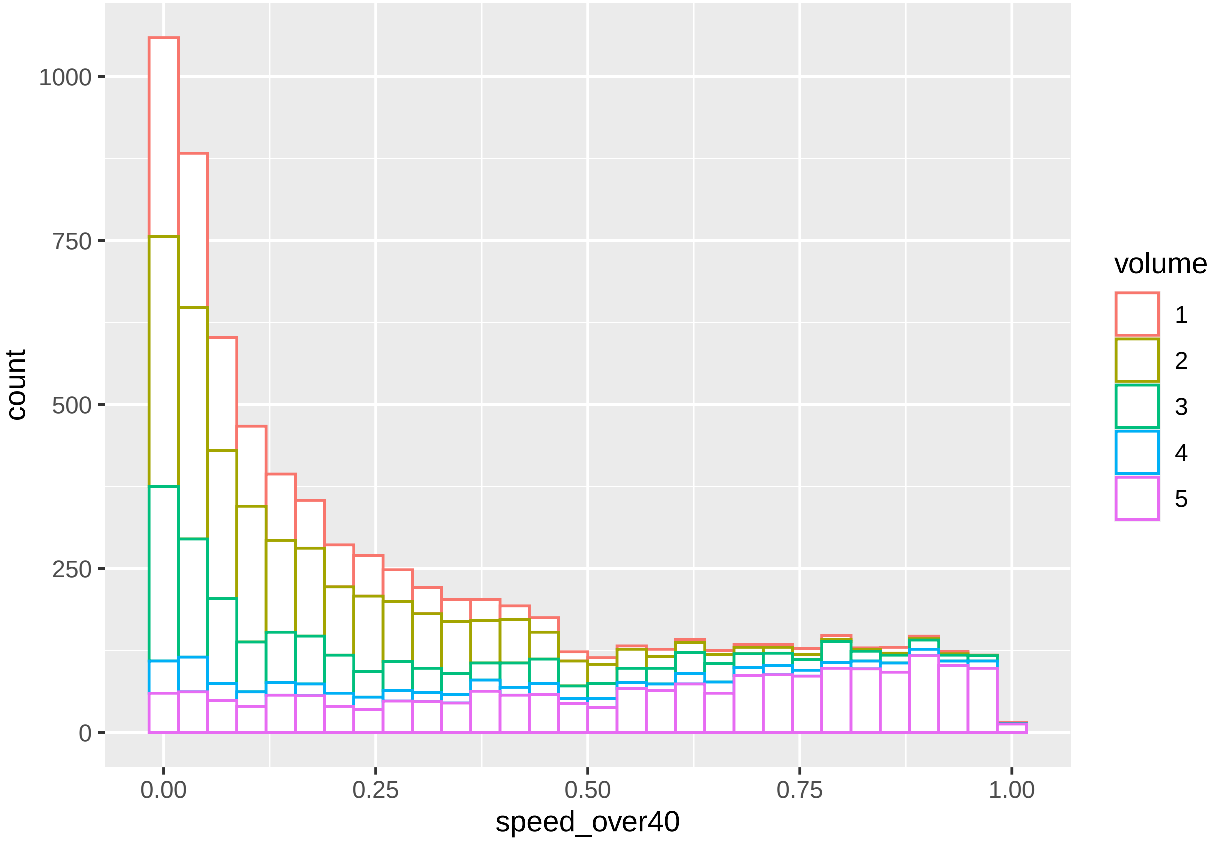

Figure 2: Proportion of vehicles that exceed 40 km/h

Figure 3 shows the histogram of the proportion of vehicles that exceeds 100 km/h for different volume levels. Unlike 2, this figure shows that only the latter two volume levels have proportion obviously larger than 0. Especially for the level 5, whose proportion can even exceed 0.4.

From these figures, we observed two things: 1) in some roads, no vehicle pays attention to the speed limit, which means we may need to emphasize more on the important of the speed limit; 2) vehicles in roads with higher volume tend to pay less attention on the speed limit and tend to drive on the highest possible speed. The latter observation is quite interesting, and we may need to explore more to find the answer of this question.

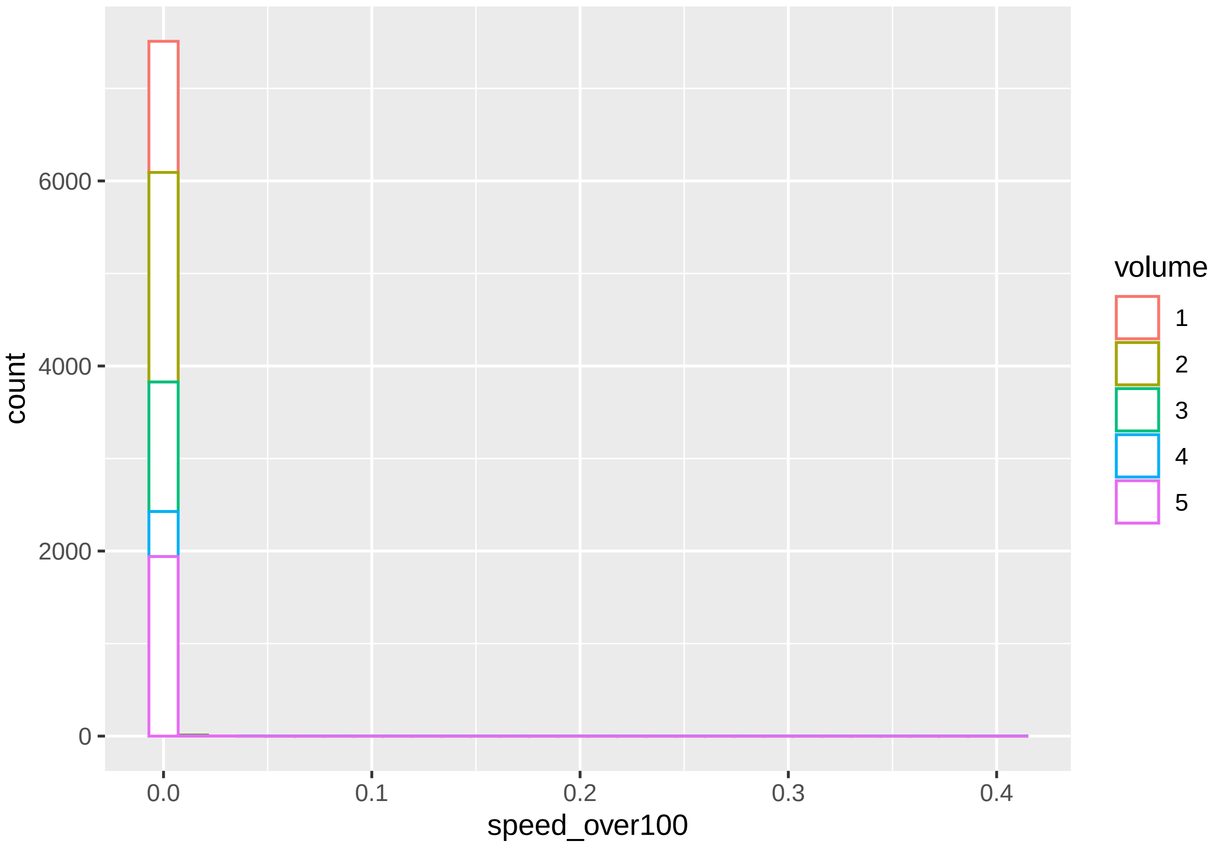

Figure 3: Proportion of vehicles that exceed 100 km/h

References

A Grammar of Data Manipulation. 2021. https://cran.r-project.org/web/packages/dplyr/dplyr.pdf.

Gelf, Sharla, and City of Toronto. 2020. *Access the City of Toronto Open Data Portal.* https://cran.r-project.org/web/packages/opendatatoronto/opendatatoronto.pdf.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RSudioTeam. 2021. *Deliver the Insights That Your Stakeholders Want with RStudio Connect.* https://www.rstudio.com/.

Services, Transportation. 2022. "About Mobile Watch Your Speed Program – Speed Summary." *City of Toronto Open Data Portal.* https://open.toronto.ca/dataset/mobile-watch-your-speed-program-speed-summary/.

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.

Xie, Yihui. 2021. *A General-Purpose Package for Dynamic Report Generation in r.* https://cran.r-project.org/web/packages/knitr/knitr.pdf.