# ttc-subway-delay in September,2021*

Vincent Yu

02/02/2021

**Abstract**

Toronto subway is one of the greatest subway systems in the world. However, due to its age and other reasons the amount of time it delays is dramatic. TTC subway delay data is a public data posted by the Toronto Transit Commission on The City of Toronto's Open Data Portal for everyone to consume and analyze from 2017. From this report, I generate the highest delay frequency that happened on Thursday day, Vaughan mc station, southbound and line1. "No Operator Immediately Available" is the highest reason why it was delayed provided by TTC.

## Introduction

As we all know, Toronto has one of the oldest subway systems in the world. It is run by the Toronto Transit Commission (TTC). TTC is the public transport agency that operates most of the public transit in the Toronto, Peel and York region and it was established 101 years ago. It is the largest urban mass transit system in Canada and the third-largest in North America. The top two are in New York City and Mexico. In 1954, the TTC opened Canada's first underground rail line, Line1. At that moment, the total length of the route was only 7.4 kilometers. The line is also called Line-Yonge–University because it was under Yonge Street and the line was southwards to Union station under University Avenue. As of 2018, the whole network has 4 lines, 75 stations and 76.9 kilometers of route. Among them, Line 1 is still the longest route. After the latest extension from Sheppard West to Vaughan Metropolitan Center, the length of it became38.8km. In general, Line 1 has two northern terminals with a U shape. Line 2 is nearly a straight line going east-west. The other 2 are extended lines connected on these 2 main routes. (Wikimedia_Foundation (2022)).
However, as Toronto citizens, we all know that tons of questions happen every day on the subway. Although the staff tries to solve them as quickly as possible, the subway still can not arrive on time and ruins daily life. Because of the vital role of public transit, the public transit service delay could cause serious results on the urban mass transit system and citizens' life. For example, on 17th January, a storm of epic proportions arrived in Toronto and barely destroyed TTC service. The subway was delayed and huge crowds were waiting for Line 1 at victory park station for a long time that morning. Countless ongoing issues across the subway network.
For this report, I will use open-access data from the Toronto Transit Commission to analyze data of delay and get some more information about why and when the TTC Subway delay happened. There will be 2 kinds of summaries later: Numerical and Graphical. The dataset will be processed and analyzed in R The tidyverse (Hadley Wickham (2019)) dplyr (Wickham et al. (2021)) and Car (Fox and Weisberg (2019)) packages are used to generate data. Figures and tables will be created by ggplot2 (Wickham (2016)) and kableextra (Zhu (2021)). The packages knitr (Xie 2021), and tinytex (Xie 2021) and rmarkdown(R Core Team 2020) are used to generate the R markdown report.

*Code and data are available at:https://github.com/YN7666/sta304-

# Data

## Data Collection Process

The City of Toronto's Open Data Portal is an open-source delivery tool to provide and collect Toronto's data with people's lives together. This is proposed and maintained by the City of Toronto government.The TTC had posted some data on the City of Toronto's Open Data site for anyone to consume and analyze from 2017. To gain a better understanding of TTC Subway & SRT Train Service Delay is the main purpose of this data analysis. The report was using the R package opendatatoronto (Gelfand (2020)) and the dataset was last updated on October 26, 2021 and openly available to the public at any time.
The dataset contains information of TTC Subway & SRT Train Service Delay in September-2021. The counts of delay were calculated by reviewing all of the TTC eAlerts for service interruptions on the four subways. A TTC eAlert is the information about the current status of TTC service including Line & route alerts, Accessibility alerts and General alerts. In this case, Line & route alerts are counted for the event of a TTC Subway service delay. There is an article that pointed out the reason why the delay happened in Hongkong may be the following factors: power cable failure, signal cable failure, turnout communication disruption and crashes involving a casualty. Also, a longer subway operation incident delay is higher. (JinxianWeng (2014)). However, since we are far more north than Hongkong the reason may be different and here we only consider the most simple ones. Besides, the CBC has done an analysis of major reasons cause TTC subway delay. It conclude that the most common reason is caused by passengers. In April, 2017, 66 percent delay were caused by emergencies and misbehaviour triggered by passengers(CBC_News (2017)). Last but not least, due to the covid, the day has some difference compared to the previous day.
But, We all know some delays never make it to an alert. One of the problems is: studies of transit service disruptions duration are also less than highway delay concerning inspecting the effects of non-causal variables on the delay duration. (JacobLouie (2017)). Also, some drivers may lie about the delay reason since they could get fired. Besides, due to the technology issue, some period of the data is missing or miss counted for example delay in 1 min is hard to detect. This is not some issue we can fix. This problem will make the average length of delay underfitted So the delay referred by this data is not accurate enough. All in all, this is the best we have for published information right now.

## Data Summary

This dataset is called TTC Subway & SRT Train Service Delay in September-2021 There were 1433 observations in the dataset which represent the overall Subway delay in the last September. There will be 10 variables Date, Time, Day, Station, Code, Min Delay, Min Gap, Bound, Line, vehicle. The first variable is the time variable and the other 2 are categorical variables that indicate when the subway is late. Station means the subway is delayed to arrive at this station. Code is the reason why the subway is delayed. By matching the codes to a TTC document "ttc-subway-delay_codes," we can translate it into sentences. For example MUSC - Miscellaneous Speed Control, MUPAA - Passenger Assistance Alarm Activated - No Trouble Found, TUSC - Operator Overspeeding and SUDP - Disorderly Patron. min_delay and min_gap are the duration and the frequency of delay count by minutes. Bound and Line are 2 common indicators of the subway. the vehicle is the number representing each train.
The very first step here is data cleaning. It is necessary for us to clean the data in order to make it cleaner, more accurate and more precise. Filter and rename are the 2 steps here. Filter every missing value(NA) in every column in the data. Rename `Min Delay` into min_Delay and `Min Gap` into min_Gap to avoid the ugly comma.
After these steps, there are 1104 data left.
A sample view of the dataset is displayed below.

```
## # A tibble: 6 x 10
##   Date                Time  Day    Station   Code  min_delay min_gap Bound Line
##   <dttm>              <chr> <chr>  <chr>     <chr>     <dbl>   <dbl> <chr> <chr>
## 1 2021-09-01 00:00:00 06:32 Wedne~ NORTH YO~ EUSC          0       0 S     YU
## 2 2021-09-01 00:00:00 08:08 Wedne~ ROSEDALE~ MUSC          3       6 N     YU
## 3 2021-09-01 00:00:00 09:26 Wedne~ GREENWOO~ MUD           3       7 W     BD
## 4 2021-09-01 00:00:00 11:10 Wedne~ SPADINA ~ MUDD          3       7 W     BD
## 5 2021-09-01 00:00:00 11:28 Wedne~ DUNDAS S~ MUPAA         0       0 S     YU
## 6 2021-09-01 00:00:00 12:17 Wedne~ ROSEDALE~ MUSC          0       0 N     YU
## # ... with 1 more variable: Vehicle <dbl>
```

## Numerical Summaries

Here I will analyze some interesting variables with numerical summaries from the data. This could give a clear view of how each variable is constituted. In the rest of the data, there will be three important and interesting variables that I will mainly focus on. The most important variable is Min_Delay. This is used to measure the duration of each delay. Min_Delay. It is a numerical variable. min_gap is the time gap between two delay events. This is also a numerical variable that could represent the min_delay from the side. Table 1 demonstrates the data summaries of the Min_Delay.

Table 1: delay in September

| total_delay | min1_delay | min1_gap | sd_min1_delay | Q1 | median | Q3 | Large_Outliers | max |
|---|---|---|---|---|---|---|---|---|
| 1104 | 3.575181 | 6.008152 | 6.243311 | 0 | 3 | 5 | 39 | 89 |

First of all, the total number of observations in this database is 1104. On the basis of this summary, the average time of delayed subways in September 2020 is 3.6 minutes, and the average time gap between two delay events in September 2020 is 6 minutes. Since we know the Q1 is 0 and time could not be less than 0. We get the lowest amount of delayed time is 0 and there is no small outlier. The longest delay is 89 minutes. Thus, the range of this data is from 0 minutes to 89 minutes. The median is 3 minutes which is a little smaller than the mean value. This means most of the delay data are gathered together. Besides, Q3 is 5 minutes long. So 75 % of the data is under 5 minutes. The standard deviation of the time subway delay is about 6.2 and it is a moderate number. 39 outliers on the right tails from the total 1104 variables.

Here are several other variables in table 2. This dataset captures mostly categorical values such as station, bound, vehicle number and delay code. So here I list a table below which contains all of them with the highest frequency of occurrence in the data. Bound is a categorical variable that indicates the direction of each subway. "SWEN" represent south, north, east, and west. However, instead of recording the direction of the end of the subway passing through this station, they consider the direction of this station only. Here, a southbound station could appear on an east-west line. B is the bound at Spadina which is a special case. There are 4 lines in this city as well. Among all 5 lines listed here, line Yu and line Yus are parts of line 1. So we do have 4 lines in total instead of 5. Combining these 2 data we can reach many conclusions. First We can see that the southbound has the highest frequency: 443 and it is the direction to downtown. And, the northbound is considered the opposite direction which has the second-highest frequency, 267. We also know that Line1 has 696 + 4, which is 700 times of delays which is almost the same as the sum north-south direction. This is reasonable since line 1 is in a north-south direction and almost none of the others have a north-south bound. Line 1 is also the longest line, which is sensible to have the highest times of delays (JinxianWeng (2014)). In the station part, I list the 5 popular stations which have the highest delay times. The highest one is Vaughan mc station.

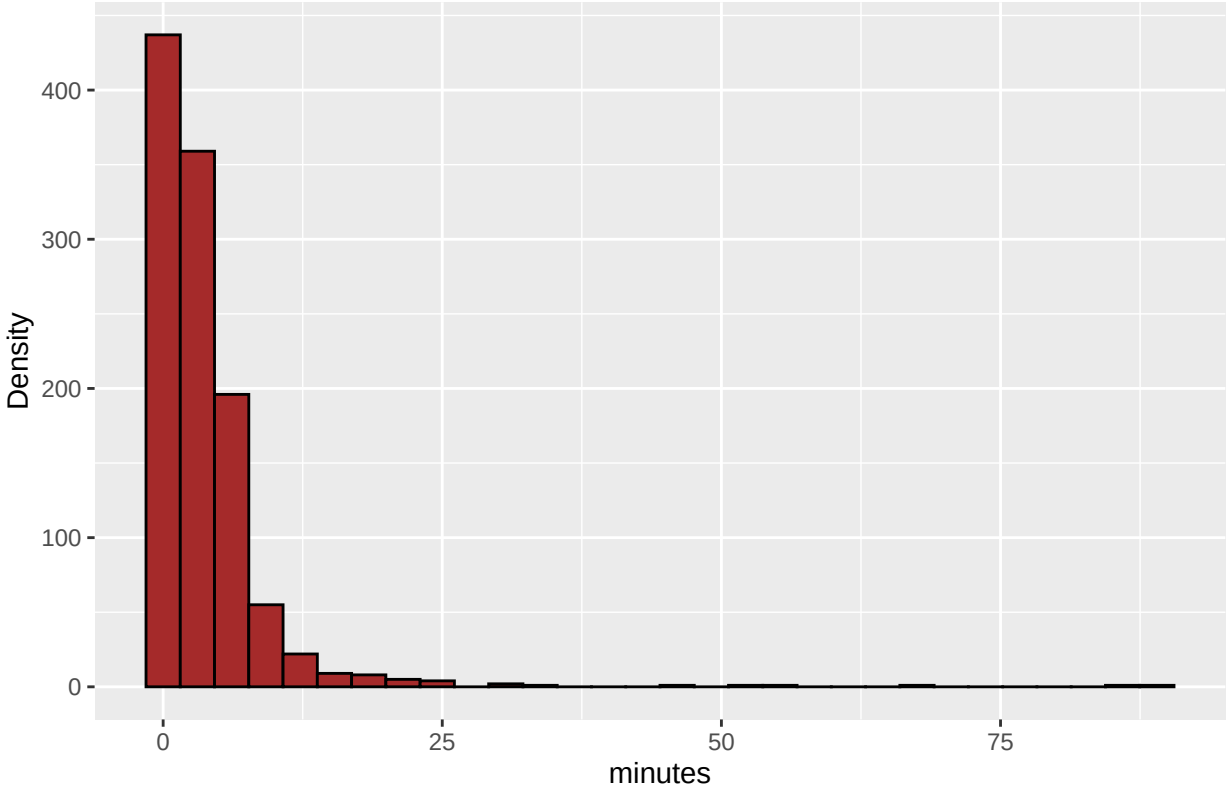
171 times the Subway delayed to arrive at this station. The other thing is, all 5 stations are from line 1 as well.

Table 2: general delay by line or bound or station

| Bound | n | Line | n | Station | n |
|---|---|---|---|---|---|
| S | 443 | YU | 696 | VAUGHAN MC STATION | 171 |
| N | 267 | BD | 328 | ROSEDALE STATION | 44 |
| E | 222 | SHP | 50 | FINCH STATION | 42 |
| W | 159 | SRT | 26 | BLOOR STATION | 34 |
| B | 13 | YUS | 4 | WILSON STATION | 29 |

## Graphical Summaries

Fig.1: Number of TTC Subway Delays by minutes in September 2021




I also make a histogram to show the general distribution of the delay. Based on figure 1, we can see the aspect of the distribution is right-skewed. Most of the data are assembled between 0 -25mins. There is a long tail on the right of the figure. Most of the bins are 0. Only a few of them have values like 1or 2. The highest bin is at 0 minutes. This is because every delay under 3 minutes is considered as 0 minutes. This indicates over 400 delays are under 3 minutes. There are approximately 350 delays of around 3 minutes. Only 0 min to 12 mins have more than 50 times delays, which is significant. This graph proved the 3 minutes median and 3.6 minutes mean we get from table 1 is reasonable.

On the other hand, I would like to study the relationship between the reason why the subway is delayed(Code) and how delays it costs. Due to the code is a categorical variable, in figure 2 we use a bar plot

to show the top 10 reasons TTC Subway Delays in September 2021. After matching the codes to the TTC document "ttc-subway-delay_codes"I generate the top 5 reasons and the number occurs here:

| Code | reason | frequency |
|---|---|---|
| TUNOA | No Operator Immediately Available | numbers of occurs: 132 |
| SUDP | Disorderly Patron | numbers of occurs: 125 |
| PUOPO | OPTO (COMMS) Train Door Monitoring | numbers of occurs: 90 |
| MUPAA | Passenger Assistance Alarm Activated - No Trouble Found | numbers of occurs: 88 |
| MUSC | Miscellaneous Speed Control | numbers of occurs: 47 |

We could learn that "No Operator Immediately Available" is the most common error here. The top 4 reasons have a frequency of around 100 times. Meanwhile, the others are under 50. The graph shows a clear view that they have a huge gap with the 5th one: "Miscellaneous Speed Control" and all other reasons.

Fig.2: Top 10 reason TTC Subway Delays in September 2021

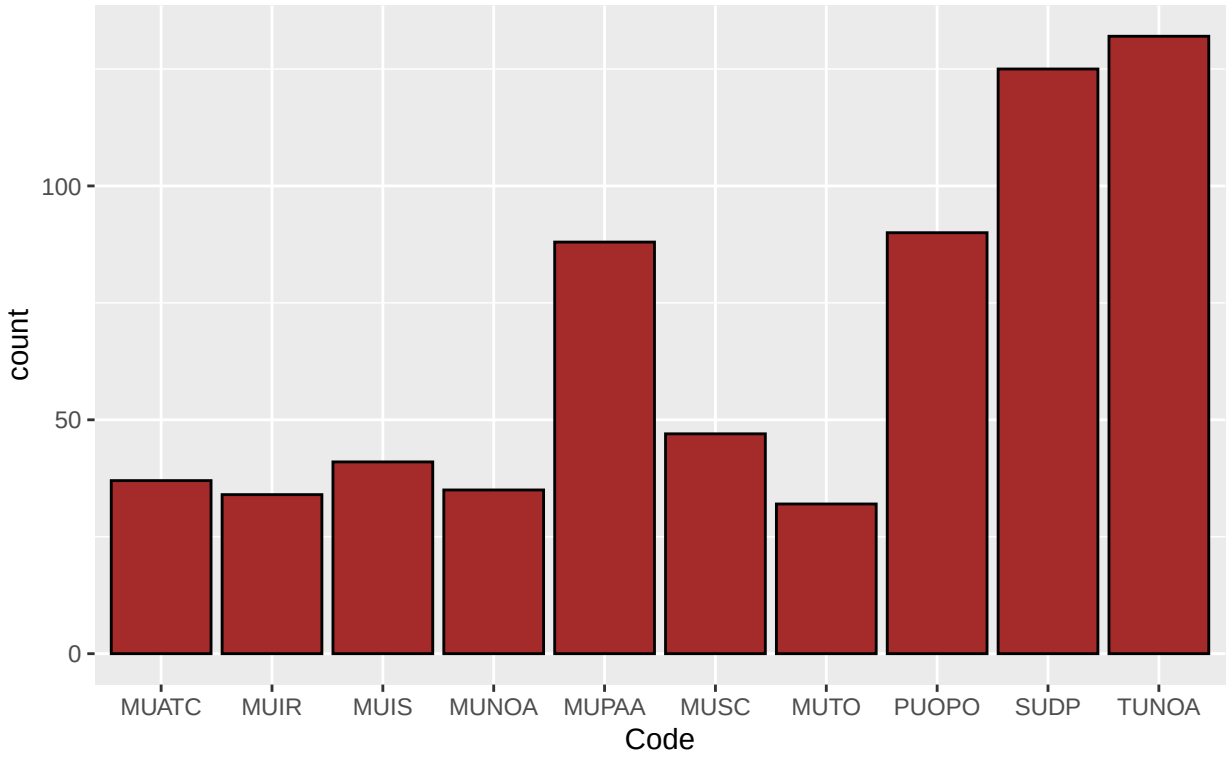
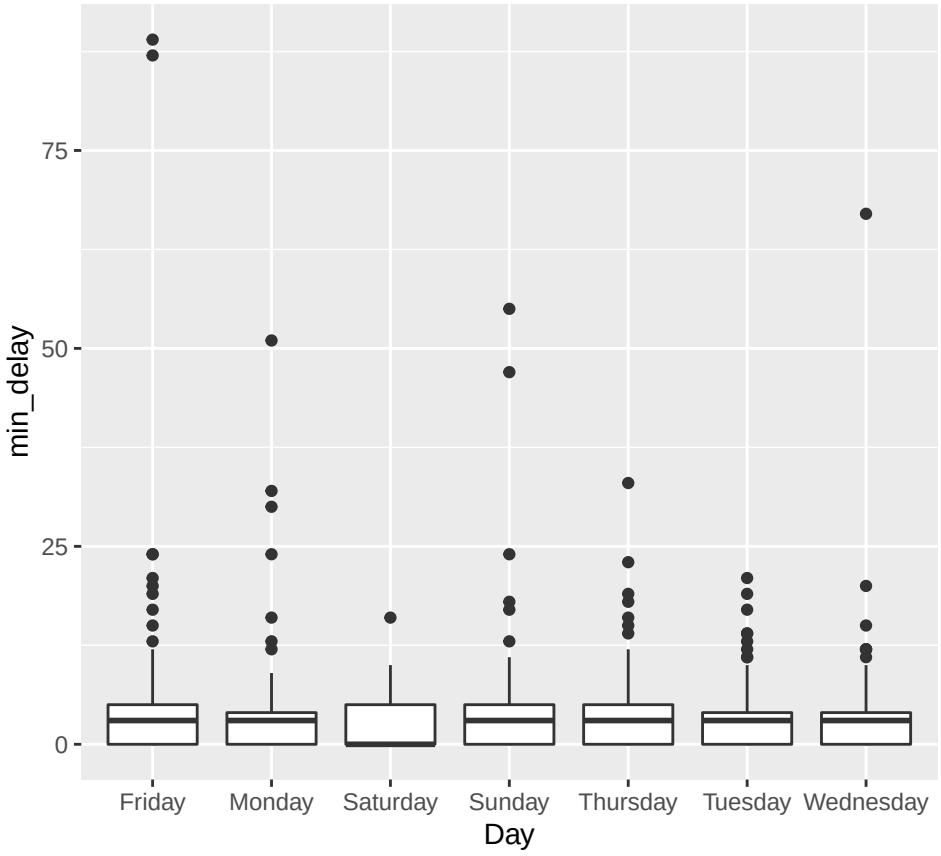


Table 4: delay on each day

| Day | n |
|---|---|
| Thursday | 196 |
| Sunday | 195 |
| Wednesday | 168 |
| Tuesday | 160 |
| Friday | 148 |
| Monday | 133 |
| Saturday | 104 |

As I mentioned above, the last one of the three important and interesting variables is the day. Day is a variable that represents the subways have more delays on which day of the week? In this section, I will do both Graphical Summaries and Numerical Summaries to represent the number of TTC Subway Delays by minutes in September 2021 in different week of days. Of all 7 days, Thursday has the highest number of delays: 196. This could be caused by Wednesday and Thursday being repeated 5 times while others only had 4 times in September. The Mean of delay is 3.6 minutes with an IQR from 0 to 5 minutes. The longest delay is 33 minutes. Saturday has the lowest number of delays: 104. The Mean of delay is 2.6 minutes with an IQR from 0 to 5 minutes. The longest delay is 16 minutes which is also the lowest one. Besides, all of the boxes have a left bound equal to 0 and some outliers on the right tail. The biggest outlier is 89 minutes, which is on Friday. The lowest boxes(IQR) are Monday, Tuesday, and Wednesday: from 0 to 4 minutes. All the others are 0 to 5 minutes. The median of most boxes is 3 minutes, except Saturday which is 0.

Fig.3: Number of TTC Subway Delays by minutes in September 2021 in different days



# Reference


R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2021). rmarkdown: Dynamic Documents for R. R package version 2.9. URL https://rmarkdown.rstudio.com.

Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown.

Yihui Xie and Christophe Dervieux and Emily Riederer (2020). R Markdown Cookbook. Chapman and Hall/CRC. ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.

Yihui Xie (2021). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.35.

CBC_News. 2017. "66." https://www.cbc.ca/news/canada/toronto/ttc-subway-delays-1.4068358.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Hadley Wickham, Jennifer Bryan, Mara Averick. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

JacobLouie, Shalaby, M.Eng.sAmer. 2017. "Modelling the Impact of Causal and Non Causal Factors on Disruption Duration for Toronto Subway System: An Exploratory Investigation Using Hazard Modelling." https://www.sciencedirect.com/science/article/abs/pii/S0001457516303694.

JinxianWeng, Xuedong, YangZheng. 2014. "Development of a Subway Operation Incident Delay Model Using Accelerated Failure Time Approaches." https://www.sciencedirect.com/science/article/abs/pii/S0001457514002322.

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

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

Wikimedia_Foundation. 2022. "Toronto Subway." https://en.wikipedia.org/wiki/Toronto_Transit_Commission.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.