NYS Traffic Tickets Issued: Four Year Window

Abstract (max 200 words)

What is the question or story you are trying to tell?

Traffic Violations are one of the major reasons for road accidents, these can even lead to one losing his/her life. We will be Analysing and Visualizing NYS Traffic Tickets Issued dataset to find some useful information out of it by finding relationships between Violation Description, Gender, Age, Day, Month, Year of Violation. We will be trying to find out what's the distribution and volume of traffic tickets issued according to age and gender of a motorist in NYC? Which were the most issued Traffic tickets in NYC month-wise for 4 consecutive years from 2014 to 2017? Which were the most active days of the week where tickets were issued and what were they? What is the gender of the motorist to which the violation ticket was issued? Whether the number of tickets issued increase with the year?

What is the conclusion that you reached?

The data set with its columns like ViolationDescription, Gender, Age can be used to plot a Violin chart to depict the relation between all 3 columns. Further, the data set can be converted into a time series data to create an animated bar chart race showing a relationship between count of violation and violation description. Also, the data set can be used to create a relationship between day of the week, year, violation description, month, age and gender. By this, we will be extracting some useful information out of the Dataset and finding some trends out of it and answer some questions.

1. Dataset [½ page]

Where/how did you retrieve it or them

- Data was taken from Kaggle Titled: **NYS Traffic Tickets Issued: Four Year Window.** Link: https://www.kaggle.com/new-york-state/nys-traffic-tickets-issued-four-year-window
- Data is extracted by Kaggle from records of tickets on file with NYS DMV. The tickets were issued to motorists for violations of : NYS Vehicle & Traffic Law (VTL), Thruway Rules and Regulations, Tax Law, Transportation Law, Parks and Recreation Regulations, Local New York City Traffic Ordinances, and NYS Penal Law.

Describe the data - size (GB or attributes), number of rows, attributes, data types present
What aspects (if any) of big data (volume, variety, velocity) are present in your data

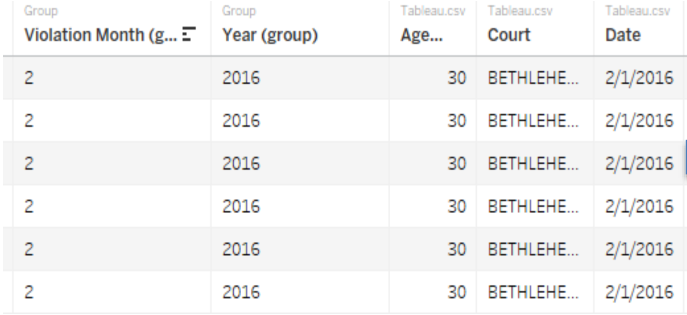
- The size of the dataset is 1.6GB and it is available in CSV format, it has 14.4 Million rows. It has 11 columns as followed
  - Violation Charged Code: Law code of the violation: Nominal data
  - Violation Description: Description of the violation: Nominal data
  - Violation Year: Year of Violation: Temporal data
  - Violation Month: Month of Violation: Temporal data
  - Violation Day of Week: Day of the week of Violation: Ordinal data
  - Age at Violation: Age in years of the person to whom the ticket was issued: Ratio
  - Gender: Gender of the person to whom the ticket was issued: Nominal Data
  - State of License: Licence State of driver license presented by the motorist: Nominal data
  - Police Agency: Police Agency that issued the ticket: Nominal data
  - Source: Processing system used for the ticket to record and track: Nominal Data
- The Dataset is huge in size and is generated on a daily basis, Kaggle has taken the data from New York State Open Data and is maintained by Kaggle making the data trustworthy. Therefore, it has **Volume, Velocity and Veracity** aspects of BigData.

2. Data Exploration, Processing, Cleaning and/or Integration [½ page]

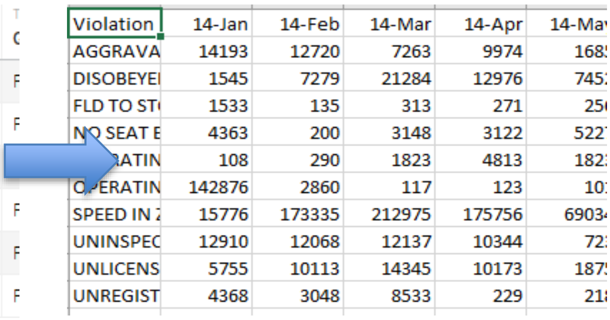
What did you need to do to prepare the dataset(s) to create your graph/chart?

- Pandas
  - The data set had NA values, a missing column name and many extra columns that were not needed for further visualization, so all the NA values were removed and the columns were renamed and some columns were dropped from the dataset using Pandas to carry on with the Visualization.
  - Out of the 513 violations. Top 8 most occurring violations were selected.
- Tableau DataSource:
  - For Flourish Bar Chart Race the data was needed to be converted entirely into a format with columns as date and rows as traffic violation with a total count of violations per month in each cell. For this, as the month and year were in separate columns a calculated date field was created and then this csv was used in pandas for further processing. In Pandas top 10 violations were selected and the dataset

was grouped by Violation Description and newly calculated date field giving a count for each violation according to date. This newly created csv was exported and transposed and was given as an input to Flourish to create a bar chart race.



Group Violation Month (g...	Group Year (group)	Tableau.csv Age...	Tableau.csv Court	Tableau.csv Date	T C
2	2016	30	BETHLEHE...	2/1/2016	F
2	2016	30	BETHLEHE...	2/1/2016	F
2	2016	30	BETHLEHE...	2/1/2016	F
2	2016	30	BETHLEHE...	2/1/2016	F
2	2016	30	BETHLEHE...	2/1/2016	F
2	2016	30	BETHLEHE...	2/1/2016	F



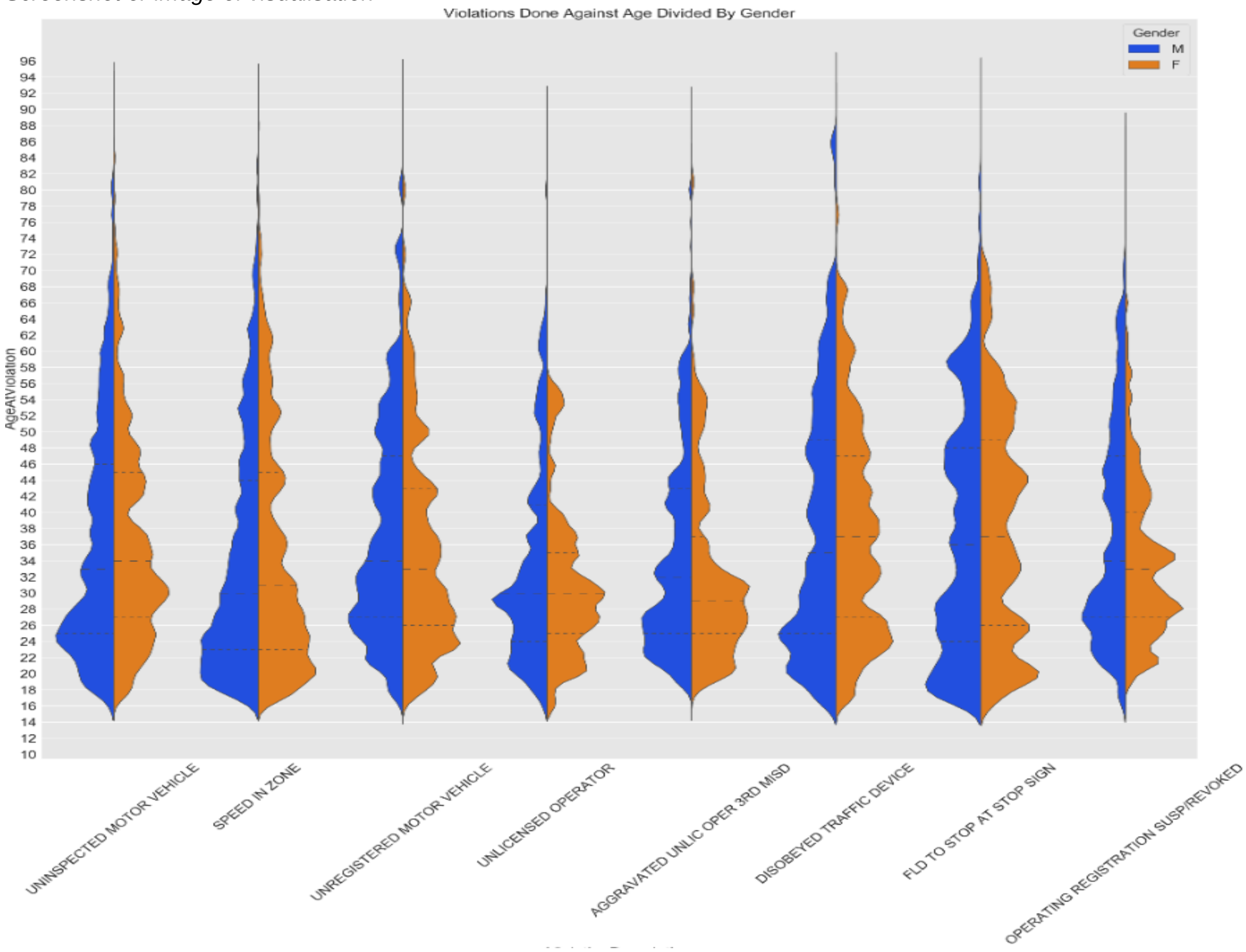
Violation	14-Jan	14-Feb	14-Mar	14-Apr	14-Ma
AGGRAVA	14193	12720	7263	9974	168
DISOBEYEI	1545	7279	21284	12976	745
FLD TO ST	1533	135	313	271	25
NO SEAT E	4363	200	3148	3122	522
RATIN	108	290	1823	4813	182
OPERATIN	142876	2860	117	123	10
SPEED IN Z	15776	173335	212975	175756	6903
UNINSPEC	12910	12068	12137	10344	72
UNLICENS	5755	10113	14345	10173	187
UNREGIST	4368	3048	8533	229	21

How did you choose the attributes to visualise?

- ViolationDescription: It has the name of the violation ticket that was issued. Count of ViolationDescription can be used with other columns and itself for visualization.
- Year: Year column gives us the year of violation and is used to divide ViolationDescription, ViolationMonth, Gender, ViolationDayOfWeek year-wise.
- ViolationMonth: ViolationMonth and Year was used to create a custom date field in Tableau for Bar and used with Violation Description Count to create a bar chart race.
- ViolationDayOfWeek: It is used to other columns like Gender, ViolationDescription, Year, etc. in terms of its occurrence in a day of week.
- AgeAtViolation: It is used to visualize at what age group violation was done.
- Gender: It is used to divide other columns as per the Gender of the motorist.

3. Visualisation [½-1 page]

Screenshot or image of visualisation



Explain your choice of chart or graph type - what relationship or data type are you showing?

- Given chart is a violin plot, they show the probability density of the data at different values. Here the volume of count of a violation done for specific violations is shown with respect to the gender on the x-axis and the count is distributed according to the age of the motorist on the y-axis. The quartiles are also presented representing the first 25%, 50%, 75% of the total violations done. It clearly represents a distribution about at what age a violation ticket was issued the most and what was the gender of the motorist.

Design choices - justify your use of colour, shapes, marks, layout, structure, font, labels

- The Pallette colour used is bright, as the inner quartiles weren't properly visible with other palette choices like dark. The x label is rotated by 45 degrees as they were overlapping with each other. The figure size has been set to (60,60) also the default Seaborn fonts are used, and the background has been set to Darkgrid. The legend is kept in the top right corner, to avoid any overlapping with the graphs.

Any interactivity or animation and how it helps answer your question

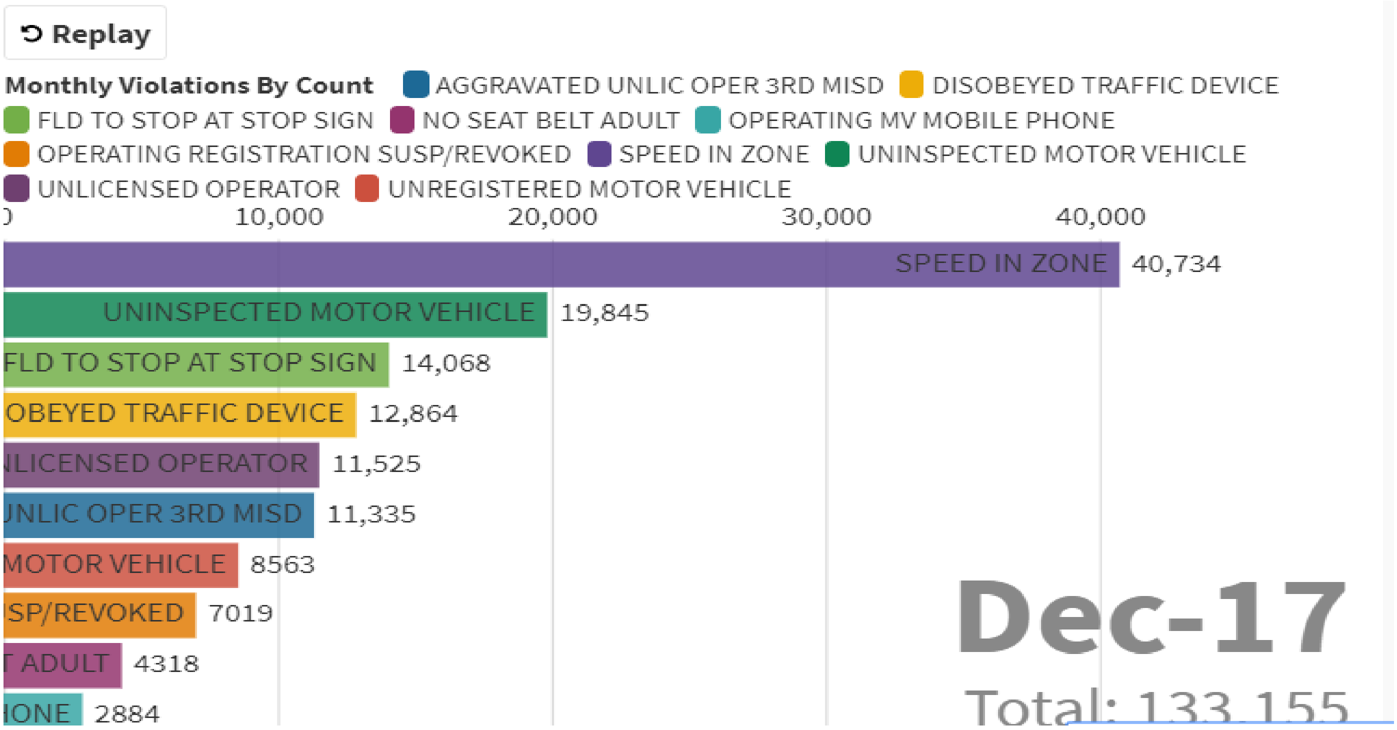
- The first chart is a simple noninteractive chart.

List of tools or libraries used

- Seaborn, Pandas, MatPlotlib, Python

4. Visualisation [½-1 page]

Screenshot or image of visualisation



Explain your choice of chart or graph type - what relationship or data type are you showing?

- The Given chart is an animated bar chart race, it needs data as time series, and it changes accordingly with values. Here the violations were grouped with their count every month. For this, the dataset was transformed into a time series data with columns as individual months with year and rows with violations and its count for that particular time.

Design choices - justify your use of colour, shapes, marks, layout, structure, font, labels

- The colour for the bars used is Flourish 1 as all the bars can be easily distinguished, the changing size of the bar easily represents violations count per month. The font used was Source Sans Pro and labels were added on each bar apart from the legend to make the bars more distinguished. The animation scale was set to 2 in flourish, so the animation and major changes were clearly visible.

Any interactivity or animation and how it helps answer your question

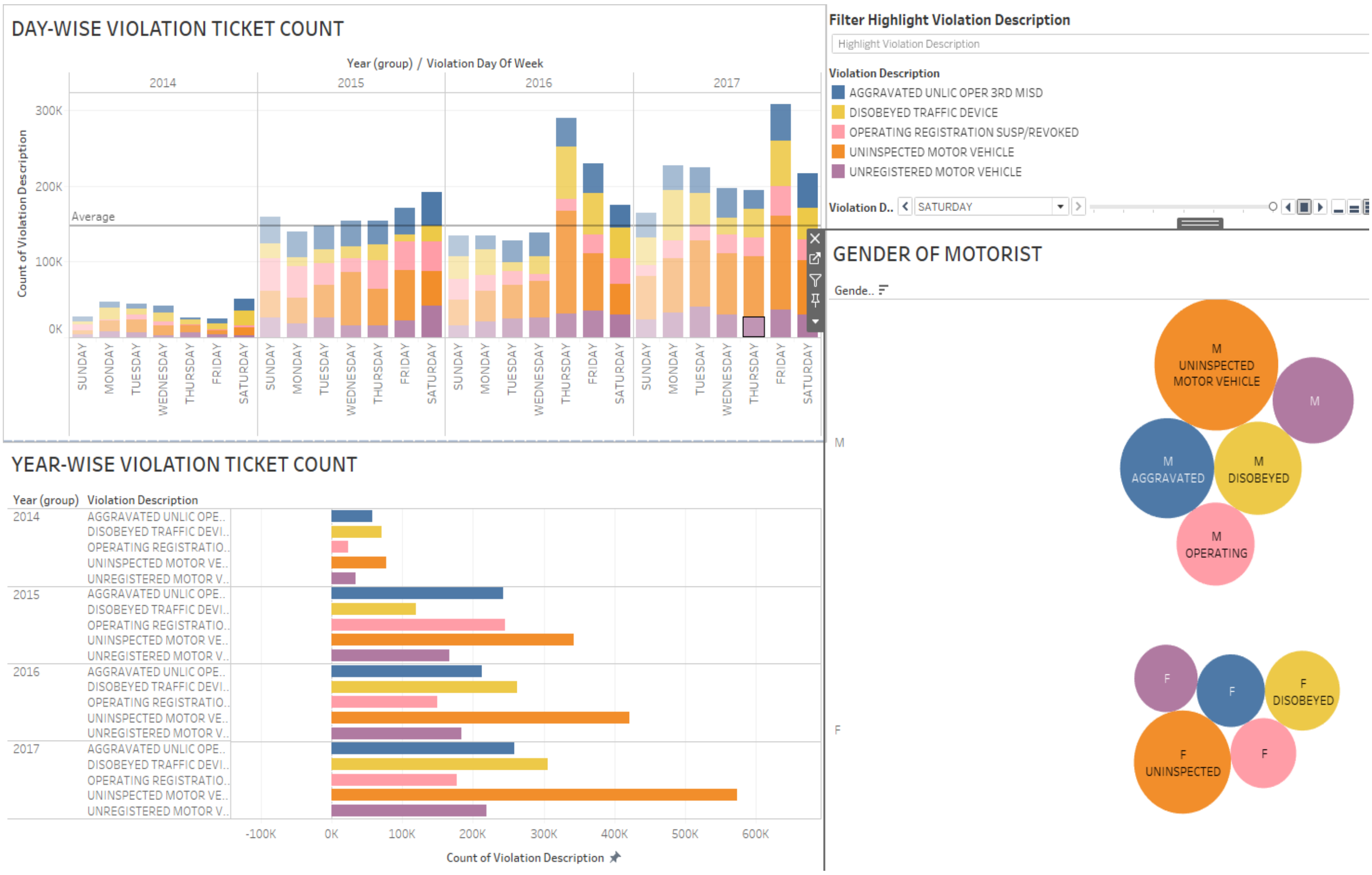
- All the bars represent ViolationDescription and change its length depending upon its count every month. Therefore, the viewer easily knows the count and description of tickets in descending order every month. If the count of a particular violation bar is increased if compared to other it changes its position accordingly.

List of tools or libraries used:

- Pandas, Tableau Data Source, Excel for creating a customized column, cleaning and transposing the data into a time series form. Flourish for Visualization.

5. Visualisation [½-1 page]

Screenshot or image of visualisation



Explain your choice of chart or graph type - what relationship or data type are you showing?

- 1st graph shows the count of 5 different violations tickets issued every day of the week each year, 2nd graph shows each year count of every violation and the third graph showing gender-based division on the count of particular Violation tickets issued interactively combined with each other.

Design choices - justify your use of colour, shapes, marks, layout, structure, font, labels

- Tableau 10 colour was used for different sections of the bar and the Gender circles as all sections are clearly visible, the bar chart was used to depict different Violation ticket counts in sections of the day of week and year. Circles with specific volumes depending upon the count of violations tickets issued were used to represent the gender of the motorist for specific violations. The legend was shifted to an empty space on the graph without obstructing, the font used was Tableau book

Any interactivity or animation and how it helps answer your question

- The bars have an animation showing growing trends, each violation interacts with the violation count in a day of the week and year and gender graph. Highlighting a specific violation highlights that violation count in all the three graphs.

List of tools or libraries used

- Tableau

4. Conclusion [½ page]

Critically analyse the outcome of your visualisation.

- Speeding in zone tickets were issued the most followed by Uninspected Motor Vehicle.
- Mostly the motorist to whom the violation tickets are issued are in the age group 20-38
- Around 25 % of the tickets are issued to motorist aged around above 47 by looking at the upper quartile
- Around 25% of the tickets are issued to motorist aged below 25/26 by looking at the lower quartile
- The volume of the number of tickets issued is decreasing as the age is increasing for all violation types.
- Tickets FLD TO STOP AT STOP SIGN and SPEED IN ZONE has been issued in large number to the younger aged population. Around 50% of the tickets are issued to people aged below 30 for SPEED IN Zone
- Some tickets have been issued to people aged below 18.

- From the Bar Chart race of month-wise count of the number of violations done it is seen that SPEED IN ZONE ticket issues was leading in most of the months, there were many months where UNINSPECTED MOTOR VEHICLE was the most issued ticket. DISOBEYED TRAFFIC DEVICE and UNREGISTERED MOTOR VEHICLE followed them. NO SEAT BELT and OPERATING MV MOBILE PHONE were the least issued tickets.
- In 2014 most of the tickets issued were on Saturday, and in 2015 most tickets were issued on Saturday followed by Friday. In 2016 most tickets were issued on Thursday followed by Friday.
- In 2017 most tickets were issued on Friday followed by Monday and Saturday.
- There's a constant increase in the number of tickets issued every year, day-wise. Comparatively more number of tickets were issued on Friday and Saturday.
- A number of tickets issued to males were higher than females in all types of violation tickets issued.
- From the graphs, we can conclude that the number of tickets issued each year is increasing linearly for all types of violations.

Were there aspects that you think could be improved upon?

- This data can be added with a new column stating whether the day on which the ticket was issued was a festive day/occasion or just a normal day.
- The data could have been further divided into time of the day like Morning, Afternoon, Evening, Night. By this, we could have calculated at what time of the day the maximum number of tickets were issued.
- Another column with the amount of fine imposed on the motorist could have been added.
- Also, weather conditions could also have been added like rain, fog, etc.

Were there effects or functionality that you were technically unable to achieve?

- With this huge volume of data, having a lot of variety with around 533 types of violation and 1.4 million rows description it was difficult to visualize each aspect of violation description.

References:

Include any citation of the dataset

[1] Kaggle.com, NYC Traffic Ticket Issued: Four Year Window, Available at https://www.kaggle.com/new-york-state/nys-traffic-tickets-issued-four-year-window

Include links to any tutorial or example that contributed significantly to your work

[2] Flourish.studio for Data Visualization, Available at: https://flourish.studio/
[3] Seaborn: https://seaborn.pydata.org/
[4] Matplotlib: https://matplotlib.org/tutorials/index.html
[5] Tutorialspoint: https://www.tutorialspoint.com/seaborn/index.htm:
[6] Tableau.com: https://www.tableau.com/learn/training

Include any articles or web resources supporting your design choices

[1] Dataquest.io: https://www.dataquest.io/blog/design-tips-for-data-viz/
[2] Columnfivemedia: https://www.columnfivemedia.com/25-tips-to-upgrade-your-data-visualization-design