Individual Project Documentation
Ruizhe Fu(Ritchie)
March 14th
CSC-324

Purpose

Traffic Violation is one of the significant and hottest topic in the USA. Over years, people form stereotypes about traffic violations, like the black people are more likely to violate traffic rules. This project analyze and visualize various aspects related to traffic violations and try to answer questions of what's the distribution or likelihood of traffic violation and different groups in races and ages; what are the most common traffic violations; and are peoples' stereotypes about related to traffic violation correct.

Data Description

This project utilize three data listed below:

1. Traffic and Drugs Related Violations Dataset
Link: https://www.kaggle.com/shubamsumbria/traffic-violations-dataset

Dataset:

stop_date	stop_time	country_n...	driver_gen...	driver_age...	driver_age	driver_race	violation_r...	violation	search_co...	search_type	stop_outc...
1/2/2005	1:55		M	1985	20	White	Speeding	Speeding	FALSE		Citation
1/18/2005	8:15		M	1965	40	White	Speeding	Speeding	FALSE		Citation
1/23/2005	23:15		M	1972	33	White	Speeding	Speeding	FALSE		Citation
2/20/2005	17:15		M	1986	19	White	Call for Service	Other	FALSE		Arrest Driver
3/14/2005	10:00		F	1984	21	White	Speeding	Speeding	FALSE		Citation
3/23/2005	9:45		M	1982	23	Black	Equipment/Inspection Violation	Equipment	FALSE		Citation
4/1/2005	17:30		M	1969	36	White	Speeding	Speeding	FALSE		Citation
6/6/2005	13:20		F	1986	19	White	Speeding	Speeding	FALSE		Citation
7/13/2005	10:15		M	1970	35	Black	Speeding	Speeding	FALSE		Citation
7/13/2005	15:45		M	1970	35	White	Speeding	Speeding	FALSE		Citation
7/13/2005	16:20		M	1979	26	Asian	Speeding	Speeding	FALSE		Citation
7/13/2005	19:00		F	1966	39	White	Speeding	Speeding	FALSE		Citation
7/14/2005	19:55		M	1979	26	White	Speeding	Speeding	FALSE		Citation
7/18/2005	19:30		F	1984	21	White	Speeding	Speeding	FALSE		Citation

Description:
This dataset contains more than 65000 traffic-related violation records from 2005 to 2010 in the USA. The attribute of dataset are date, time and category of violation; gender, race and age of violators; and other information including whether search is conducted, whether the driver takes drugs and the result of violation. There are over 65000 observations included in this dataset.

2. Distribution of Licensed Drivers – 2010 by Sex and Percentage in Each Age Group and

Relation to Population

Link: https://www.fhwa.dot.gov/policyinformation/statistics/2010/dl20.cfm

Data:

AGE	MALE DRIVERS			FEMALE DRIVERS			TOTAL DRIVERS		
	NUMBER	PERCENT OF TOTAL DRIVERS	DRIVERS AS PERCENT OF AGE GROUP 1/	NUMBER	PERCENT OF TOTAL DRIVERS	DRIVERS AS PERCENT OF AGE GROUP 1/	NUMBER	PERCENT OF TOTAL DRIVERS	DRIVERS AS PERCENT OF AGE GROUP 1/
UNDER 16	199,269	0.2	9.4	198,272	0.2	9.8	397,541	0.2	9.6
16	607,987	0.6	28.0	604,584	0.6	29.4	1,212,571	0.6	28.7
17	1,024,767	1.0	46.4	1,003,673	0.9	47.8	2,028,440	1.0	47.1
18	1,407,573	1.3	62.5	1,323,264	1.3	62.0	2,730,837	1.3	62.2
19	1,640,724	1.6	71.2	1,546,127	1.5	71.0	3,186,851	1.5	71.1
(19 AND UNDER)	4,880,320	4.7	44.2	4,675,920	4.4	44.6	9,556,240	4.5	44.4
20	1,743,925	1.7	78.1	1,681,843	1.6	79.8	3,425,768	1.6	78.9
21	1,756,443	1.7	79.5	1,717,300	1.6	82.4	3,473,743	1.7	80.9
22	1,757,099	1.7	80.0	1,725,417	1.6	83.4	3,482,516	1.7	81.6
23	1,767,027	1.7	79.6	1,748,038	1.7	83.8	3,515,065	1.7	81.6
24	1,792,212	1.7	80.1	1,778,871	1.7	84.7	3,571,083	1.7	82.4
(20-24)	8,816,706	8.4	79.5	8,651,469	8.2	82.8	17,468,175	8.3	81.1
25-29	9,178,507	8.8	82.6	9,252,767	8.8	87.6	18,431,274	8.8	85.0
30-34	8,934,026	8.6	88.4	8,915,067	8.4	91.2	17,849,093	8.5	89.7
35-39	9,079,149	8.7	87.7	9,082,236	8.6	89.2	18,161,385	8.6	88.4
40-44	9,612,893	9.2	91.5	9,564,857	9.0	91.2	19,177,750	9.1	91.4
45-49	10,380,717	9.9	91.9	10,433,487	9.9	90.4	20,814,204	9.9	91.2
50-54	10,240,523	9.8	95.9	10,387,582	9.8	93.7	20,628,105	9.8	94.8
55-59	9,126,397	8.7	99.1	9,313,113	8.8	95.3	18,439,510	8.8	97.2
60-64	7,846,635	7.5	103.6	8,010,950	7.6	97.3	15,857,585	7.5	100.3
65-69	5,652,418	5.4	102.6	5,815,585	5.5	92.7	11,468,003	5.5	97.3
70-74	4,028,977	3.9	98.7	4,201,935	4.0	85.3	8,230,912	3.9	91.4
75-79	2,966,194	2.8	94.2	3,191,705	3.0	76.4	6,157,899	2.9	84.1
80-84	2,090,118	2.0	90.9	2,373,492	2.2	67.4	4,463,610	2.1	76.7
85 AND OVER	1,540,916	1.5	86.4	1,870,278	1.8	48.6	3,411,194	1.6	60.6
TOTAL	104,374,496	100.0	87.1	105,740,443	100.0	84.4	210,114,939	100.0	85.7

Description:
This data summarize information about percent and number of people with car access by ages and sex in the United States in2010. Among the 20-85 year old, the data gives the number and percentage of people(divided into male, female and total) who have car access per group of five-years range.

3. Percent of households without a vehicle by race/ethnicity: United States
Link:
https://docs.google.com/spreadsheets/d/1OiTD9Xt0Fmjd79P3Vi2szT67WI_x9MsY/edit?usp=sharing&ouid=105901885807046626254&rtpof=true&sd=true

Data:

Data Dic https://docs.google.com/spreadsheets/d/1OiTD9Xt0Fmjd79P3Vi2szT67WI_x9MsY/edit?usp=sharing&ou

geo_code	GEO_NAME	year	raceth	racethd	immig	sex	racethdir	PVEHICHH	nVEHICHH_ZERO_ipums
01000000	United S	1990	All	All	All	All peop	All	1125	10324384
01000000	United S	1990	White	White	All	All peop	White	0817	6010359
01000000	United S	1990	Black	Black	All	All peop	Black	2942	2862337
01000000	United S	1990	Latino	Latino	All	All peop	Latino	1869	1094533
01000000	United S	1990	Asian or	Asian or	All	All peop	Asian or	1272	248897
01000000	United S	1990	Native Ar	Native Ar	All	All peop	Native Ar	1655	96476
01000000	United S	1990	Mixed/otl	Mixed/otl	All	All peop	Mixed/otl	2248	11782
01000000	United S	1990	People o	People o	All	All peop	People o	2373	4314025
01000000	United S	2000	All	All	All	All peop	All	1018	10744855
01000000	United S	2000	White	White	All	All peop	White	0719	5682328
01000000	United S	2000	Black	Black	All	All peop	Black	2325	2753311
01000000	United S	2000	Latino	Latino	All	All peop	Latino	1703	1579494
01000000	United S	2000	Asian or	Asian or	All	All peop	Asian or	1257	401673
01000000	United S	2000	Native Ar	Native Ar	All	All peop	Native Ar	1434	97467
01000000	United S	2000	Mixed/otl	Mixed/otl	All	All peop	Mixed/otl	1470	230582
01000000	United S	2000	People o	People o	All	All peop	People o	1906	5062527
01000000	United S	2010	All	All	All	All peop	All	0885	10113167
01000000	United S	2010	White	White	All	All peop	White	0626	5084181
01000000	United S	2010	Black	Black	All	All peop	Black	1960	2623012
01000000	United S	2010	Latino	Latino	All	All peop	Latino	1267	1631381
01000000	United S	2010	Asian or	Asian or	All	All peop	Asian or	1115	510797
01000000	United S	2010	Native Ar	Native Ar	All	All peop	Native Ar	1311	90579
01000000	United S	2010	Mixed/otl	Mixed/otl	All	All peop	Mixed/otl	1164	173217
01000000	United S	2010	People o	People o	All	All peop	People o	1523	5028986
01000000	United S	2019	All	All	All	All peop	All	0864	10433170
01000000	United S	2019	White	White	All	All peop	White	0623	5077963
01000000	United S	2019	Black	Black	All	All peop	Black	1829	2664270
01000000	United S	2019	Latino	Latino	All	All peop	Latino	1072	1707655
01000000	United S	2019	Asian or	Asian or	All	All peop	Asian or	1099	639160
01000000	United S	2019	Native Ar	Native Ar	All	All peop	Native Ar	1305	94612
01000000	United S	2019	Mixed/otl	Mixed/otl	All	All peop	Mixed/otl	1129	249510
01000000	United S	2019	People o	People o	All	All peop	People o	1364	5355207

Description:
This data summarize information about percent and number of households without a vehicle by race/ethnicity in the United States in four different years(1990, 2000, 2010 and 2019). For each of the different year, it provides numbers and percentage of people that don't have car access among their races.

How was Data Collected

For the first data(Traffic and Drugs Related Violations Dataset), I searched directly on the Kaggle website, a website providing many datasets in various areas. Since my topic is to explore the general pattern of traffic violations according to different groups in the USA, I searched "traffic violations in USA" directly on Kaggle website. For the second data(Distribution of Licensed Drivers), I need to find the car access pattern by ages in the USA, so I go to the website of U.S. Department of Transportation -Federal Highway Administration, and find out there exists this data about the relationship between car access and ages in 2010 in the USA. For the third data(Percent of households without a vehicle by race/ethnicity), I need to find car access pattern by races in the USA. I searched it on website of National Equity Atlas and find out there is no direct data about it. Instead, there exists percentage and number of people don't have car access by races. Thus, I will use this data and tidy it to make it usable(described in later section).

Potential Users

The potential users for this visualization are people who are interested in exploring the various aspects related to traffic violations in the USA. Potential users also includes people who wants to explore the distribution and likelihood of traffic violation in different groups in races, ages and traffic violations in general in the USA.

Questions and What Works

The topic of this project is trying to explore various aspects related to traffic violations in the USA. Questions include what's the distribution or likelihood of traffic violation and different groups in races and ages; What are the most common traffic violations; and are peoples' stereotypes about traffic violation correct(the black people are more likely to involve in traffic violation).

The insight of my data

Only from the first dataset(Traffic and Drugs Related Violations Dataset), we can notice that for the

exploration of age, age range of 20-24 constitute the highest number of violation, and decrease as age range increases, but age below 20 constitutes very tiny number of traffic violation. However, if we combine this dataset with the car access pattern in the second data(Distribution of Licensed Drivers), we actually see that driver ages below 20 year old have the highest percentage of violation rate. This is because even if the absolute value of traffic violations for people under the age of 20 is very small, it is still a lot in percentage since there are not many drivers under the age of 20. In addition, only from the first dataset, we can also notice white constitute the highest violation number. However, again, if we combine this dataset with the car access pattern in the third data(Percent of households without a vehicle by race), we can actually see that different race group share similar violation rate because there are many more white drivers in the USA. Lastly, we can notice from the first dataset that speeding is the most common violation type in the USA.

Potential Improvement

Improvement includes two parts. Firstly, there are more areas I can explore. For instance, the first dataset(Traffic and Drugs Related Violations Dataset) also includes attributes like stop time, whether search is conducted and the reason of the stop. I can also research, work and visualize these areas to make my topic, various aspects related to traffic violations, more comprehensive. Secondly, the dataset keeps track of traffic violations from 2005 – 2010, which is quite a lot of time from current. Thus, the dataset can only represent the traffic violation situations from 2005 to 2010 instead of representing the current situation. Thus, it can be improved by finding and combining more current dataset about traffic violations.

Source of Reference

1. Traffic and Drugs Related Violations Dataset
https://www.kaggle.com/shubamsumbria/traffic-violations-dataset

2. Distribution of Licensed Drivers – 2010 by Sex and Percentage in Each Age Group and Relation to Population
https://www.fhwa.dot.gov/policyinformation/statistics/2010/dl20.cfm

3. Percent of households without a vehicle by race/ethnicity: United States
https://docs.google.com/spreadsheets/d/1OiTD9Xt0Fmjd79P3Vi2szT67WI_x9MsY/edit?usp=sharing&ouid=105901885807046626254&rtpof=true&sd=true

Process and Development

Tibbles:
The table of dataset is directly downloaded from the three links mentioned above, including Traffic and Drugs Related Violations Dataset, Distribution of Licensed Drivers – 2010 by Sex and Percentage in Each Age Group and Relation to Population, Percent of households without a vehicle by race/ethnicity: United States

Data import:
This part is direct and straightforward. I use read.csv function to import my data to the R file.

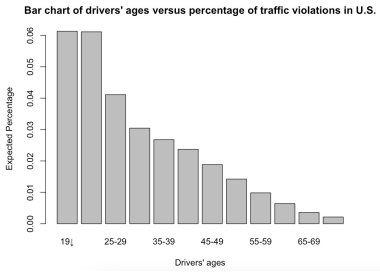
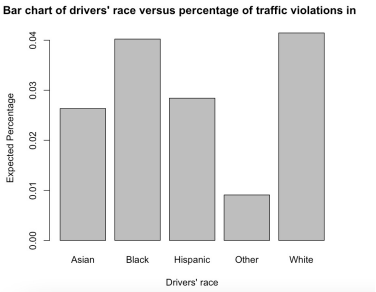
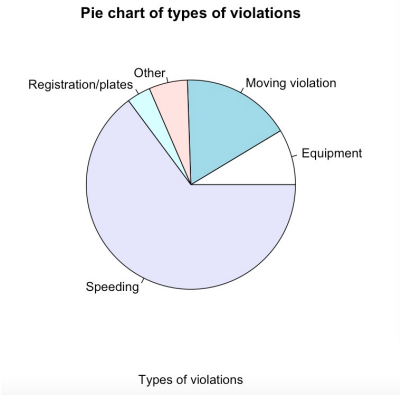
Tidy data:
This section includes three part, which are ages, races and violation type.

1. Part of Ages: Firstly, I group my the data by the attribute of age(per every 5 years from 20 years old to 85 years old), including creating the boundary and using filter(), mutate() and group_by() to assign which range each observation falls. Secondly, I use the count() to calculate the total number of observations in each group of ages. Thirdly, I combine the above tidied data with the data of Distribution of Licensed Drivers using cbind(). Lastly, I use arithmetic equation to derive the percentage of violations in each age group.
2. Part of Races: Firstly, I group my the data by the attribute of races using group_by(). Secondly, I use the count() to calculate the total number of observations in each group of races. Thirdly, I combine the above tidied data with the data of Percent of Households Without a Vehicle by Race using cbind(). Lastly, I use the arithmetic equation to derive the percentage of violations in each race group.
3. Part of Violation Type: I group my the data by the attribute of races using group_by() and then I use the count() to calculate the total number of observations in each group of races.

Visualization:
This section also includes three part, which are ages, races and violation type.

1. Part of Ages: The first part shows one bar chart about how ages are related to violations within all violations and another bar chart shows how ages are related to car access in the U.S. Then, we combine the above two to derive a bar chart showing percentage and likelihood of traffic violations in each age group.
2. Part of Races: The first part shows one bar chart about how races are related to violations within all violations and another pie chart shows how races are related to car access in the U.S. Then, we combine the above two to derive a bar chart showing percentage and likelihood of traffic violations in each race group.
3. Part of Violation Type: This section shows one bar chart and one pie chart about the relationship between the traffic violation and different violation types.

Design Decision

Chart	What?	Why?	How?
First Section: Bar Chart 	This is a bar chart showing the traffic violation rate in each drivers' ages group in the USA. Attributes are ranges of age and are in increasing order.	Users will use this bar chart to find out the likelihood(distr-ibution) of traffic violations by each age groups in the USA.	The likelihood of traffic violations is reflected in heights of the bar. Thus, we can easily see that drivers under 25 are more likely to violate traffic rules.
Second Section: Bar Chart 	This is a bar chart showing the traffic violation rate in each drivers' race group in the USA. Attributes are different races.	Users will use this bar chart to find out the likelihood(distr-ibution) of traffic violations by each races in the USA.	The likelihood of traffic violations is reflected in heights of the bar. Thus, we can easily see that black and white drivers share similar traffic violation rate.
Third Section: Pie Chart 	This is a pie chart showing the percentage of each traffic violation type from all traffic violations. Attributes are different violation types.	Users will use this pie chart to find out the most common traffic violation type, which are speeding, in the USA.	The proportions of each traffic violation types are reflected in color and area of the pie chart. We can easily see that speeding is the most common violation type in the USA.