DSCI 510 HW5 Project Description
BingQing Liu

Description: For the final project, I have created a final dataset called “final_data.csv” to do the data analysis. In this dataset, 2000 rows of data collected and generated from different sources have been included. The final dataset comes from three csv files which are “raw_data.csv”, “zip_data.csv” and “Population_by_zipcode.csv”. In this report I have analyzed traffic collisions taking place in some specific zip codes of Los Angeles using these datasets and their correlation with the population density in that particular zip code.

Motivation: The first time I arrived in Los Angeles, I could feel its prosperity. And when I drive in the freeways I can see the busy roads and crowded traffic streams. I also happened to see some traffic collisions while driving. Big trunk or small vehicle, young teenagers or old man drivers. I started to wonder what factors account for these traffic collisions. Maybe because of the weather, the age of the driver, the humidity or temperature of that day or the level of prosperity in that specific area. There are many things I can come up with. In this article I took population density into account and analyzed the relationship between the frequency of traffic collisions happening and the population density in that particular zip code. Because more population in some way indicates that there are more people and more vehicles and in common sense it will lead to more traffic collisions.

Data Sources:

Dataset1:Traffic_Collision_Data_from_2010_to_Present.csv

It is scraped from the website(https://www.splitgraph.com/lacity/traffic-collision-data-from-2010-to-present-d5tf-ez2w). I read the source code through the requests and beautifulsoup and get the csv file download link so that I can use urllib library to open the link and get the file. The data I need to use from the dataset is the victim age, the latitudes and longitudes which I can get the

address and zip codes from. After that I can form a csv file called “raw_data.csv” to collect the data. I use 500 rows of the csv because the api has rate limits every day.
The original Traffic_Collision_Data_from_2010_to_Present.csv data:

index	DR Number	Date Reported	Date Occurred	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code Description	MO Codes
0	190319651	08/24/2019	08/24/2019	450	3	Southwest	356	997	TRAFFIC COLLISION	3036 3004 3026 3101 4003
1	190319680	08/30/2019	08/30/2019	2320	3	Southwest	355	997	TRAFFIC COLLISION	3037 3006 3028 3030 3039 3101 4003
2	190413769	08/25/2019	08/25/2019	545	4	Hollenbeck	422	997	TRAFFIC COLLISION	3101 3401 3701 3006 3030
3	190127578	11/20/2019	11/20/2019	350	1	Central	128	997	TRAFFIC COLLISION	0605 3101 3401 3701 3011 3034
4	190319695	08/30/2019	08/30/2019	2100	3	Southwest	374	997	TRAFFIC COLLISION	0605 4025 3037 3004 3025 3101

Victim Age	Victim Sex	Victim Descent	Premise Code	Premise Description	Address	Cross Street	Location
22.0	M	H	101.0	STREET	JEFFERSON BL	NORMANDIE AV	(34.0255, -118.3002)
30.0	F	H	101.0	STREET	JEFFERSON BL	W WESTERN	(34.0256, -118.3089)
NaN	M	X	101.0	STREET	N BROADWAY	W EASTLAKE AV	(34.0738, -118.2078)
21.0	M	H	101.0	STREET	1ST	CENTRAL	(34.0492, -118.2391)
49.0	M	B	101.0	STREET	MARTIN LUTHER KING JR	ARLINGTON AV	(34.0108, -118.3182)

After extracting the columns i need from the dataset, i get the raw_data.csv:

	Victim Age	Time Occurred	Lat	Lng
0	22.0	2315	34.1006	−118.3387
1	30.0	1520	33.7636	−118.2961
2	21.0	1700	33.7324	−118.2857
3	49.0	1115	34.1347	−118.3427
4	60.0	1745	34.1042	−118.3383

Dataset2:Reverse Geocoding service by opencage:

It is about zipcodes use opencage to reverse geocode(https://towardsdatascience.com/reverse-geocoding-in-python-a915acf29eb6) so that I can get the address using the location including the latitudes and longitudes. Then I use the re library to search and get the zipcodes in the address. The address and corresponding zip codes are used to form a new csv called “zip_data.csv”.

The dataset “zip_data.csv”:

	Address	Zip code
0	Highland Avenue, Los Angeles, CA 90028, United...	90028
1	2101 Barrywood Avenue, Los Angeles, CA 90731, ...	90731
2	398 West 13th Street, Los Angeles, CA 90731, U...	90731
3	3481 Barham Boulevard, Los Angeles, CA 90068, ...	90068
4	1855 Highland Avenue, Los Angeles, CA 90028, U...	90028

Dataset3:

It is about the population in Los Angeles by zip codes downloaded from the website(https://data.lacity.org/Community-Economic-Development/2010-Census-Populations-by-Zip-Code/nxs9-385f). And according to the csv formed in the dataset2 processing, I get a new column corresponding to the zipcodes where the traffic collision happened. I also counted the frequency of the zipcodes shown in the traffic collision data and zip codes’ responding population number. The new csv is called “frequency.csv". The original dataset of “population.csv”:

	Zip Code	Total Population	Median Age	Total Males	Total Females	Total Households	Average Household Size
0	90001	1	73.5	0	1	1	1.00
1	90002	57110	26.6	28468	28642	12971	4.40
2	90003	51223	25.5	24876	26347	11731	4.36
3	90004	66266	26.3	32631	33635	15642	4.22
4	90005	62180	34.8	31302	30878	22547	2.73

The population number corresponding to the zipcodes in dataset2:

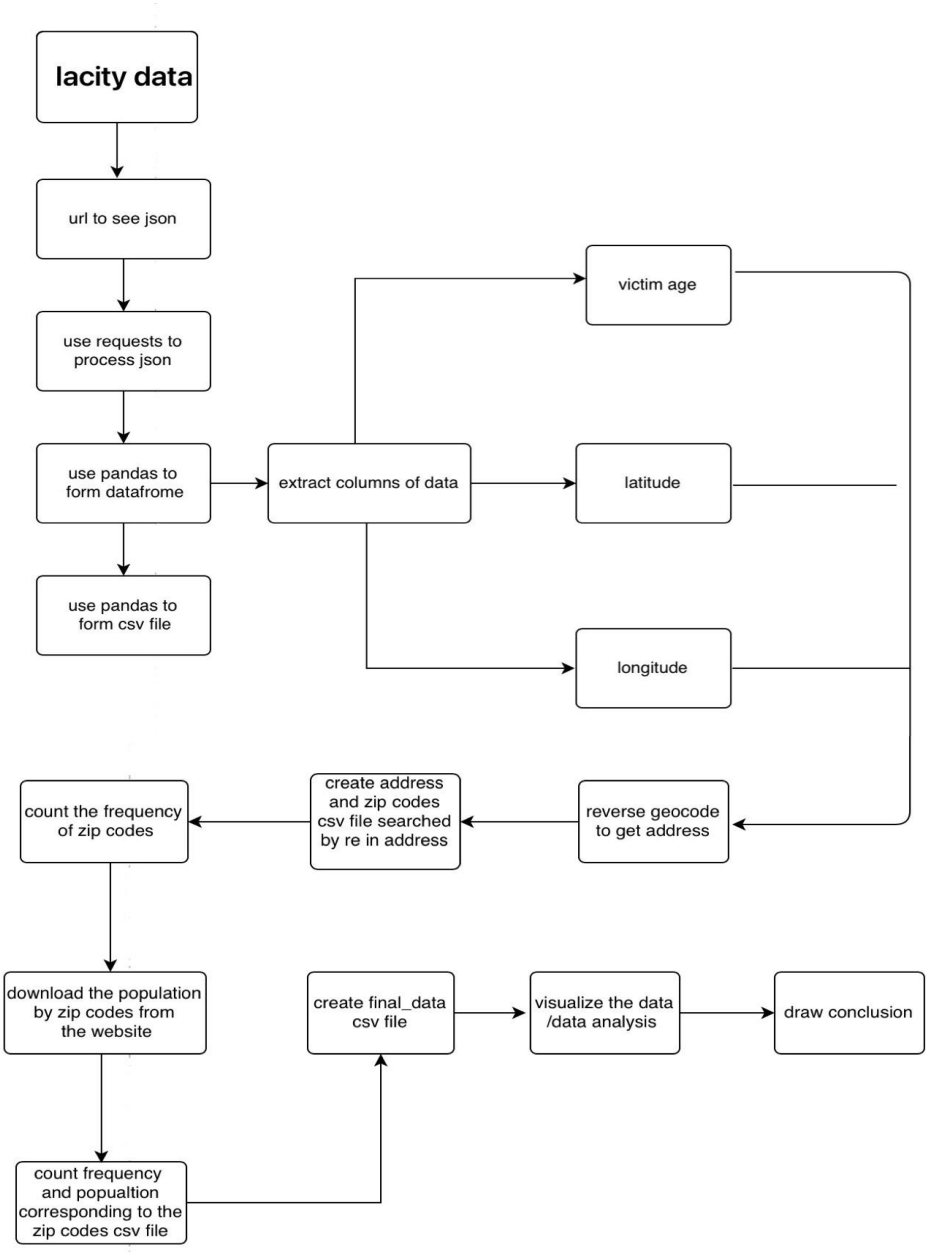
	zipcode	frequency	population
0	90028	589	45151
1	90731	328	54099
2	90068	101	2424
3	90036	32	28418
4	90029	175	28714

from dataset3

	Address	Zip code
0	Highland Avenue, Los Angeles, CA 90028, United...	90028
1	2101 Barrywood Avenue, Los Angeles, CA 90731, ...	90731
2	398 West 13th Street, Los Angeles, CA 90731, U...	90731
3	3481 Barham Boulevard, Los Angeles, CA 90068, ...	90068
4	1855 Highland Avenue, Los Angeles, CA 90028, U...	90028

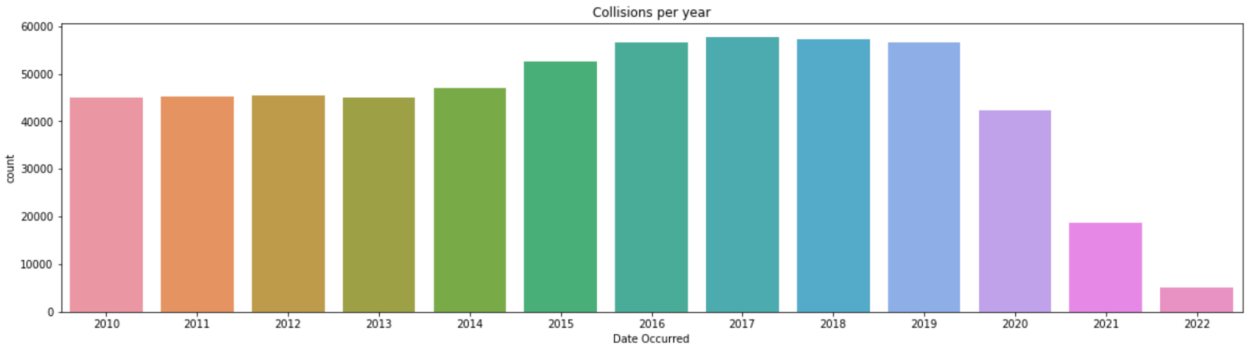
from dataset2

Flowchart:



Analysis Performed:
1.the whole original dataset is used to visualize the data

(1)Number of collisions through time



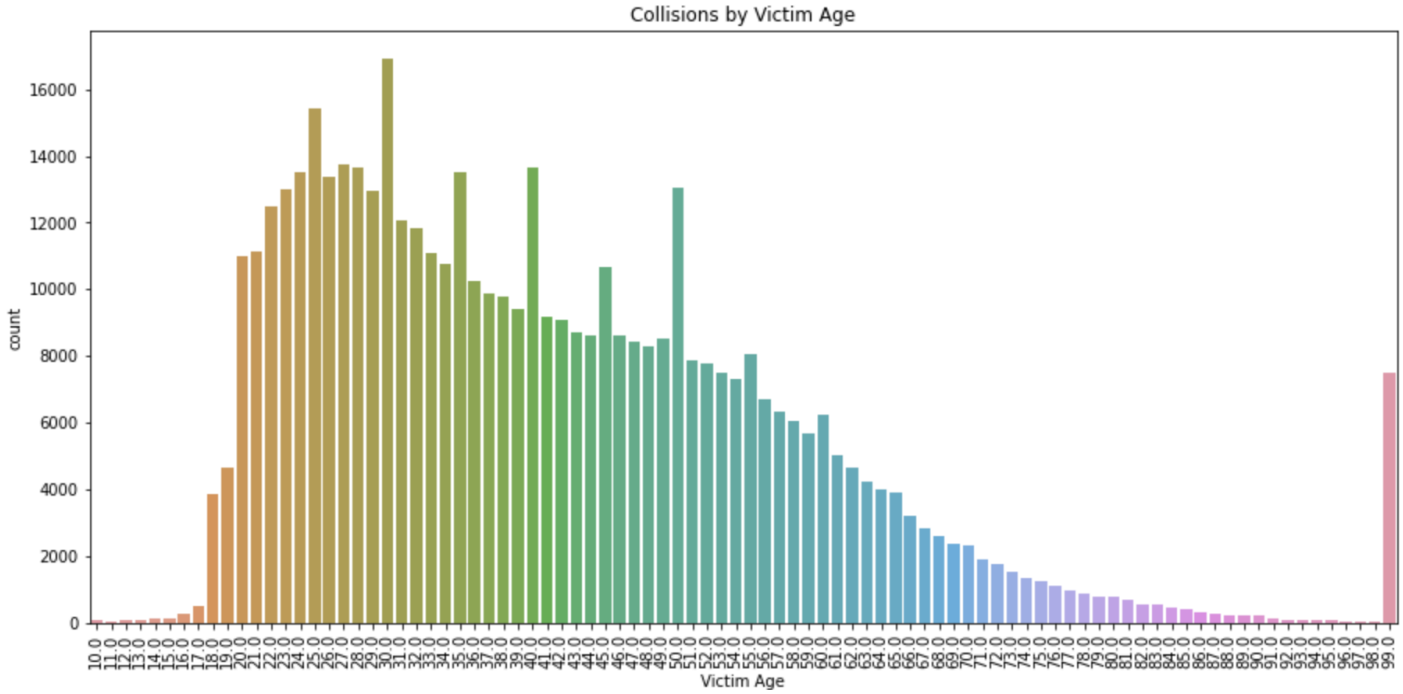
This chart summarizes the trend in collisions over the past decade. From which we can conclude that in 2017 the number of collisions is biggest and it then went down every year.

(2)Location of collisions

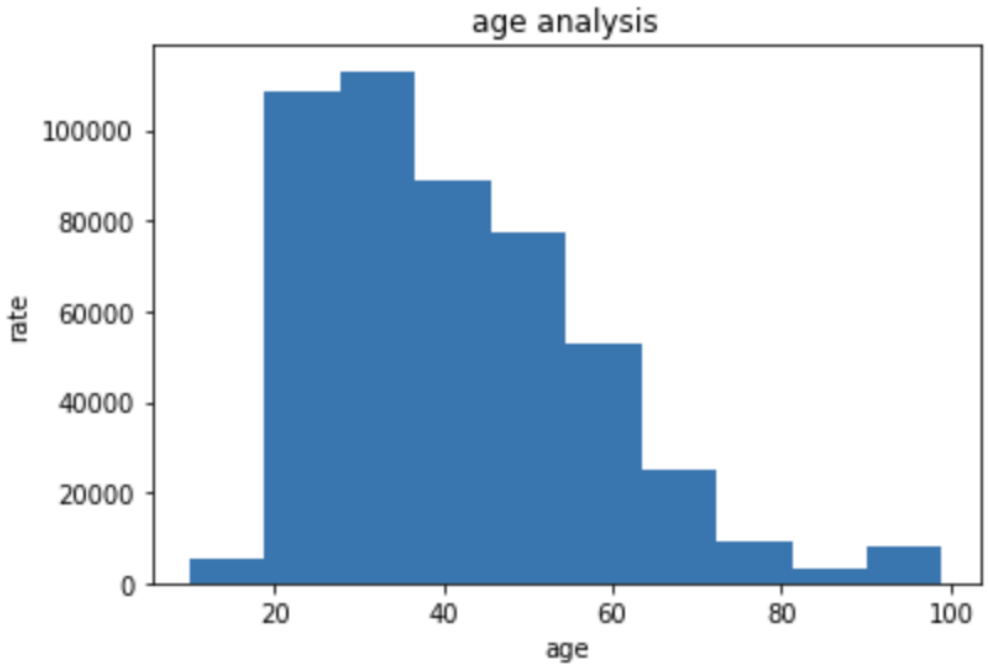
STREET	547176
PARKING LOT	18128
SIDEWALK	3174
ALLEY	1034
DRIVEWAY	991
FREEWAY	537
SINGLE FAMILY DWELLING	360
GAS STATION	345
TRANSPORTATION FACILITY (AIRPORT)	206
OTHER PREMISE	173

This table shows that traffic collisions happen most in streets.

(3)Collisions by age group

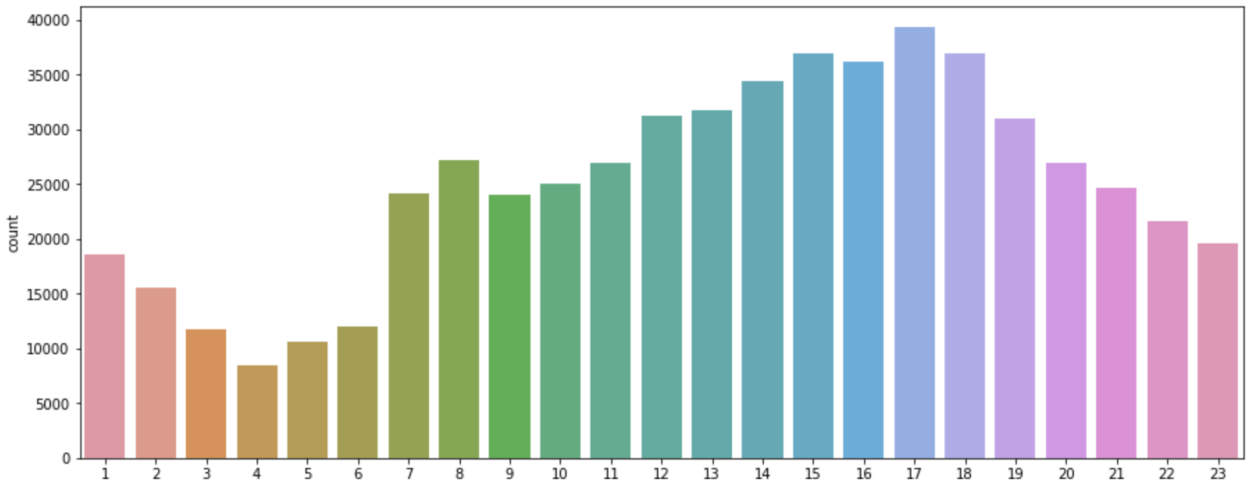


This chart shows that in every age the number of collisions happened.



The histogram shows that the people aging between 30-40 are targeted the most.

(4)Collisions by time of day



This chart shows that at 5pm, which is the time of getting off work and the rush hour, the frequency of traffic collisions is highest.

2. 2000 rows of data are used to do analysis

(1)Pearson correlation between traffic collision frequency and population density

The pearson's coefficient for correlation between collision frequency and population density is 0.10317095429617772 and the p value is 0.6837266021937809
The correlation coefficient is a positive number of (0,1], indicating that there is a linear positive correlation between x and y; the correlation coefficient is 0, indicating that there is no linear correlation between the two, but it does not rule out the existence of other nonlinear correlations; the correlation coefficient is [-1,0) is negative, indicating that there is a linear negative correlation between x and y.
Generally speaking, when the Pearson coefficient is greater than zero, the larger the correlation coefficient value, the smaller the p-value, and the greater the linear correlation. But note that p_value is not completely reliable, when the amount of data is greater than 500, it may be reasonable.

So concluded from the results got by codes, the p value is more than 0.05,which proves that the collision frequency is statistically insignificant while predicting population density corresponding to the zip codes.

(2)OLS regression

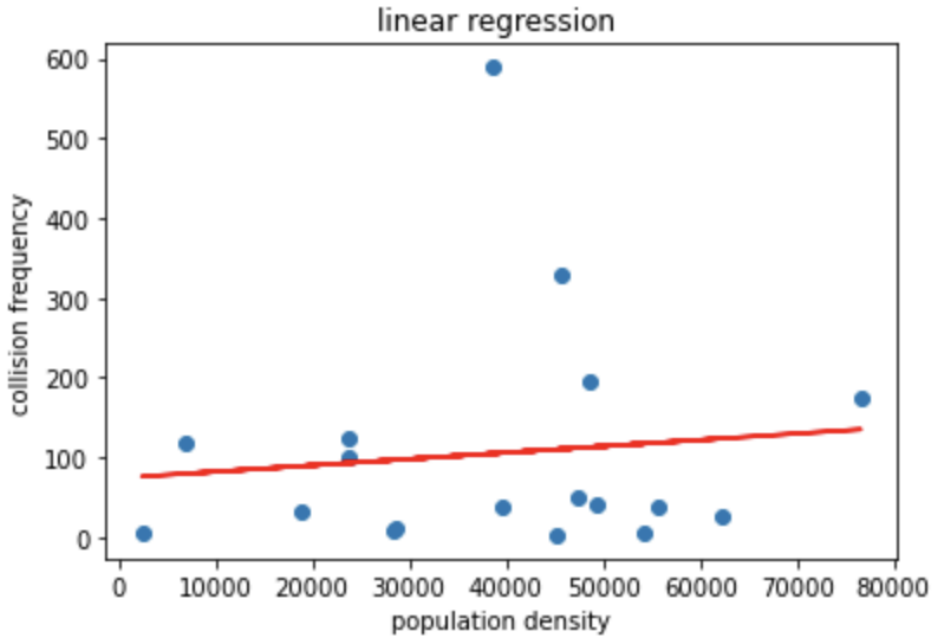
Dep. Variable:	y	**R-squared:**	0.015
Model:	OLS	**Adj. R-squared:**	-0.047
Method:	Least Squares	**F-statistic:**	0.2418
Date:	Fri, 06 May 2022	**Prob (F-statistic):**	0.630
Time:	23:11:50	**Log-Likelihood:**	-202.41
No. Observations:	18	**AIC:**	408.8
Df Residuals:	16	**BIC:**	410.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.702e+04	5730.255	6.460	0.000	2.49e+04	4.92e+04
x1	15.8129	32.155	0.492	0.630	-52.353	83.978

Omnibus:	0.038	**Durbin-Watson:**	1.866
Prob(Omnibus):	0.981	**Jarque-Bera (JB):**	0.109
Skew:	-0.052	**Prob(JB):**	0.947
Kurtosis:	2.634	**Cond. No.**	221.

After using stats model to do ols regression, we can see from the screenshot of the person output result that pearson's coefficient for correlation between collision frequency and population density is slightly positive, which is 0.10317095429617772 and indicate that population positively in some way affects frequency of collisions. But the value of 0.1 is not strong enough to support the fact.

(3)linear regression



The chart shows that the linear relationship between collison frequency and population density is subtle. The population density slightly affects the collision probability and frequency.

Conclusion:

I first used the whole original database to visualize the data so that we can see many information from the chart, including the victim age, the location of collisons, the time of collisions and the collisions per year, which can let me know better about the dataset and traffic collisions in la.
Then i runned some algrothms. We can conclude from the analysis we have applied that the correlation between the collision and population density is small and the population density slightly affects the collision frequency. The linear relationship is tiny.

The analysis is not that perfect and can be optimized in some way. Fist of all, the api has rate limits every day so it is hard for me to extract the address and get the zip codes in several days. And it will take long time to run the codes. So the database quantity is not that big. If the number of rows of data is bigger, the results will be more reasonable. Also, the data

collected from the source is not very inclusive despite big quantity of database. It does not include every zip code’s collison and every collision happened. It will in some way affect the results. More datasets can be taken into consideration.