LOS ANGELES TRAFFIC COLLISION ANALYSIS

IDENTIFYING PATTERNS AND CORRELATIONS IN ORDER TO HELP REDUCE CRASHES AND POTENTIALLY SAVE LIVES

Prasad Kulkarni, Sathish Kumar Rajendiran
IST 652 SCRIPTING FOR DATA ANALYSIS
Professor Dr. Landowski
Syracuse University | September 2020

Table of Contents

INTRODUCTION

As the second largest in the United States, Los Angeles has traffic challenges due to a large and growing population and an increase in the number of cars. A better understanding of the factors that contribute to accidents can help government officials, companies, citizens and other interested parties to understand how to make the city safer and more drivable.

The goal is to explore the trends and correlations between the data to provide useful information that can help answer our proposed analysis questions:

What are the most dangerous intersections?

What are the most common collision areas in Los Angeles?

What are the best/worst times of the day for accidents? Best/worst month?

What is the demographic makeup of victims in collisions?

What is the relationship between income and collision victims? Do certain temperatures or weather play a factor?

The goal of making Los Angeles traffic safer will not only help save lives and money, but it can potentially be a translatable example to other cities around the world and inspire others. In 2018, at least 240 people were killed in Los Angeles traffic collisions. The issue is of such importance to Los Angeles that by 2025, the goal is to have zero traffic deaths. Despite programs designed to help reduce these collisions, fatal car crashes have increased 32% in Los Angeles since 2015 and more people have died in car crashes than shootings in that same timeframe. Many layers and factors exist for these traffic collisions. The objective of this report is to highlight noticeable trends and patterns that can possibly lead to solutions in the future for this important crisis in Los Angeles and abroad.

About The Data

Source Data

Original Dataset

The Los Angeles Traffic Collision Data is publicly available from Kaggle.com is owned by the City of Los Angeles. The contains 481,568 incidents from 2010 to 2019.

Source: https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data

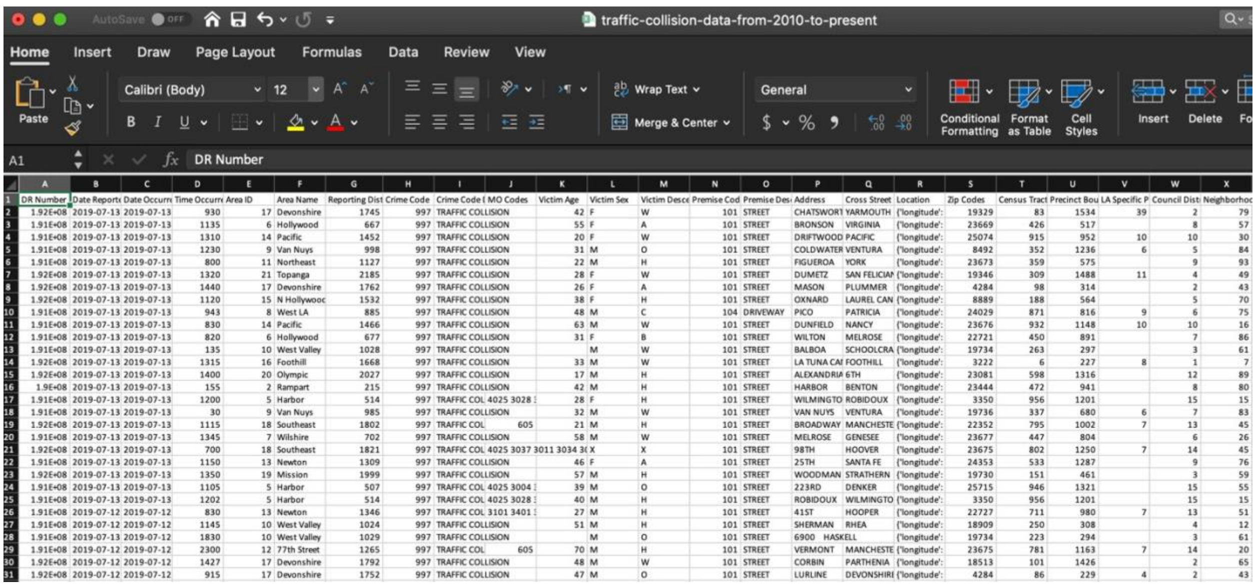


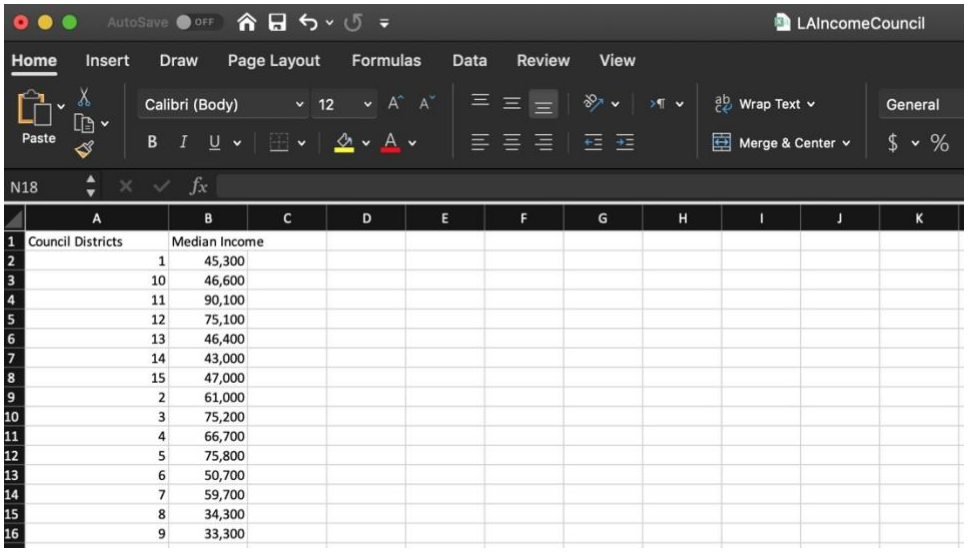
FIGURE 1 - LOS ANGELES TRAFFIC DATA 2010-2019 FROM KAGGLE.COM

Multiple data sets were incorporated into analysis beyond the original Kaggle dataset in order to get more data that was not included in this original dataset such as weather and income data.

Median Income Dataset

To answer the analysis question about income, outside data was needed since the original dataset did not have income information. For Median Income, incomes were pulled from the LA Chamber of Commerce website. They were then inputted into a CSV and merged into the original data frame. The incomes are for Council Districts in Los Angeles and are from 2016. Other ways were examined to link income to our dataset such as by Area Names and Zip Code, but in both attempts at doing that, there were not enough matches to the original dataset. Council District was found to be the most effective way to merge income with the rest of the data.

Source: https://lachamber.com/clientuploads/pdf/2018/18_BeaconReport_LR.pdf



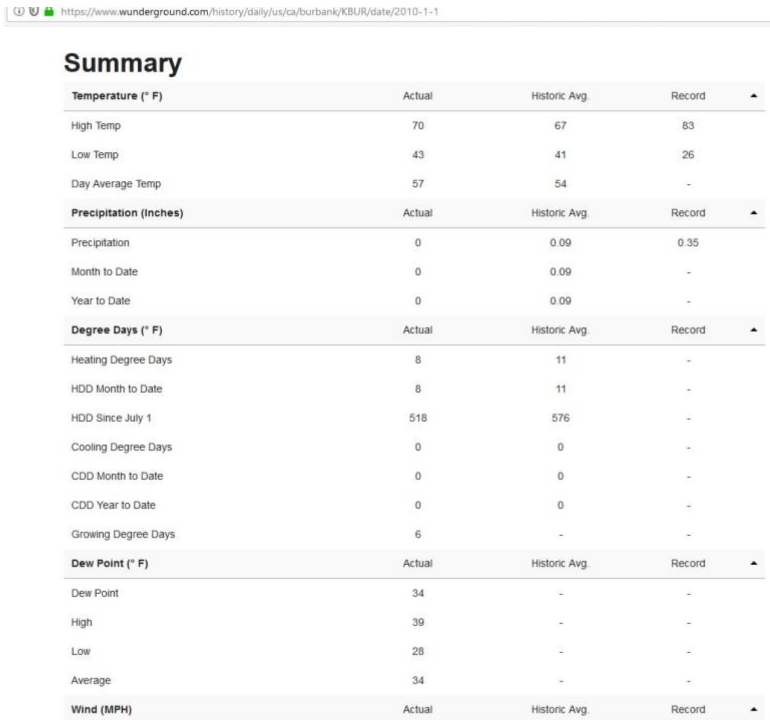
Council Districts	Median Income
1	45,300
10	46,600
11	90,100
12	75,100
13	46,400
14	43,000
15	47,000
2	61,000
3	75,200
4	66,700
5	75,800
6	50,700
7	59,700
8	34,300
9	33,300

FIGURE 2 - MEDIAN INCOME DATA FROM LA CHAMBER OF COMMERCE

WEATHER DATASET

The weather data was scraped from the website Wunderground.com. Once in CSV form, each weather CSV contained 359 rows and 7 columns (*Date*, *HighTemp*, *LowTemp*, *AverageTemp*, *Precipitation*, *NauticalTwilight*, *NauticalTwilightSet*).

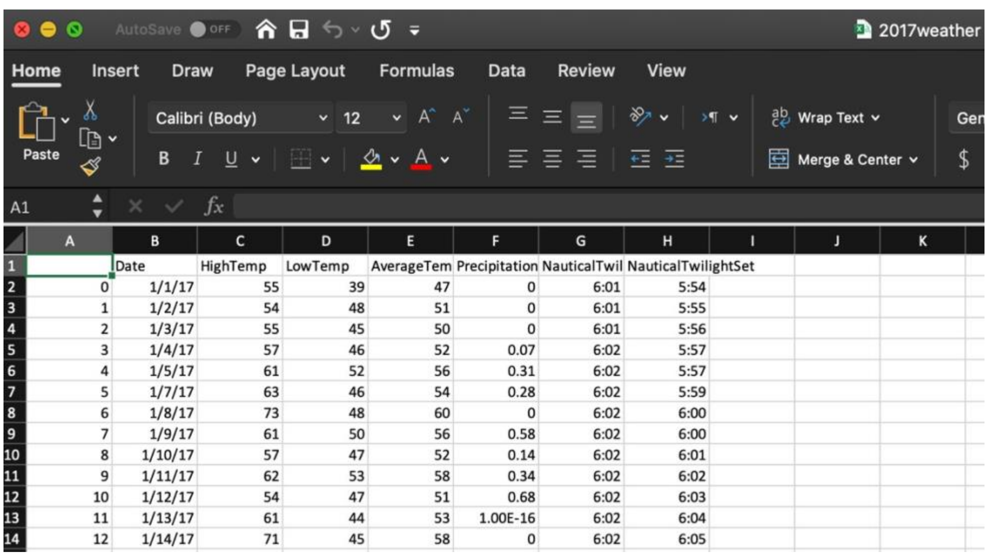
Source: www. Wunderground.com



Summary

Temperature (° F)	Actual	Historic Avg.	Record
High Temp	70	67	83
Low Temp	43	41	26
Day Average Temp	57	54	-
Precipitation (Inches)	Actual	Historic Avg.	Record
Precipitation	0	0.09	0.35
Month to Date	0	0.09	-
Year to Date	0	0.09	-
Degree Days (° F)	Actual	Historic Avg.	Record
Heating Degree Days	8	11	-
HDD Month to Date	8	11	-
HDD Since July 1	518	576	-
Cooling Degree Days	0	0	-
CDD Month to Date	0	0	-
CDD Year to Date	0	0	-
Growing Degree Days	6	-	-
Dew Point (° F)	Actual	Historic Avg.	Record
Dew Point	34	-	-
High	39	-	-
Low	28	-	-
Average	34	-	-
Wind (MPH)	Actual	Historic Avg.	Record

FIGURE 3, UNSTRUCTURED DATA IN JAVASCRIPT FORM SCRAPED

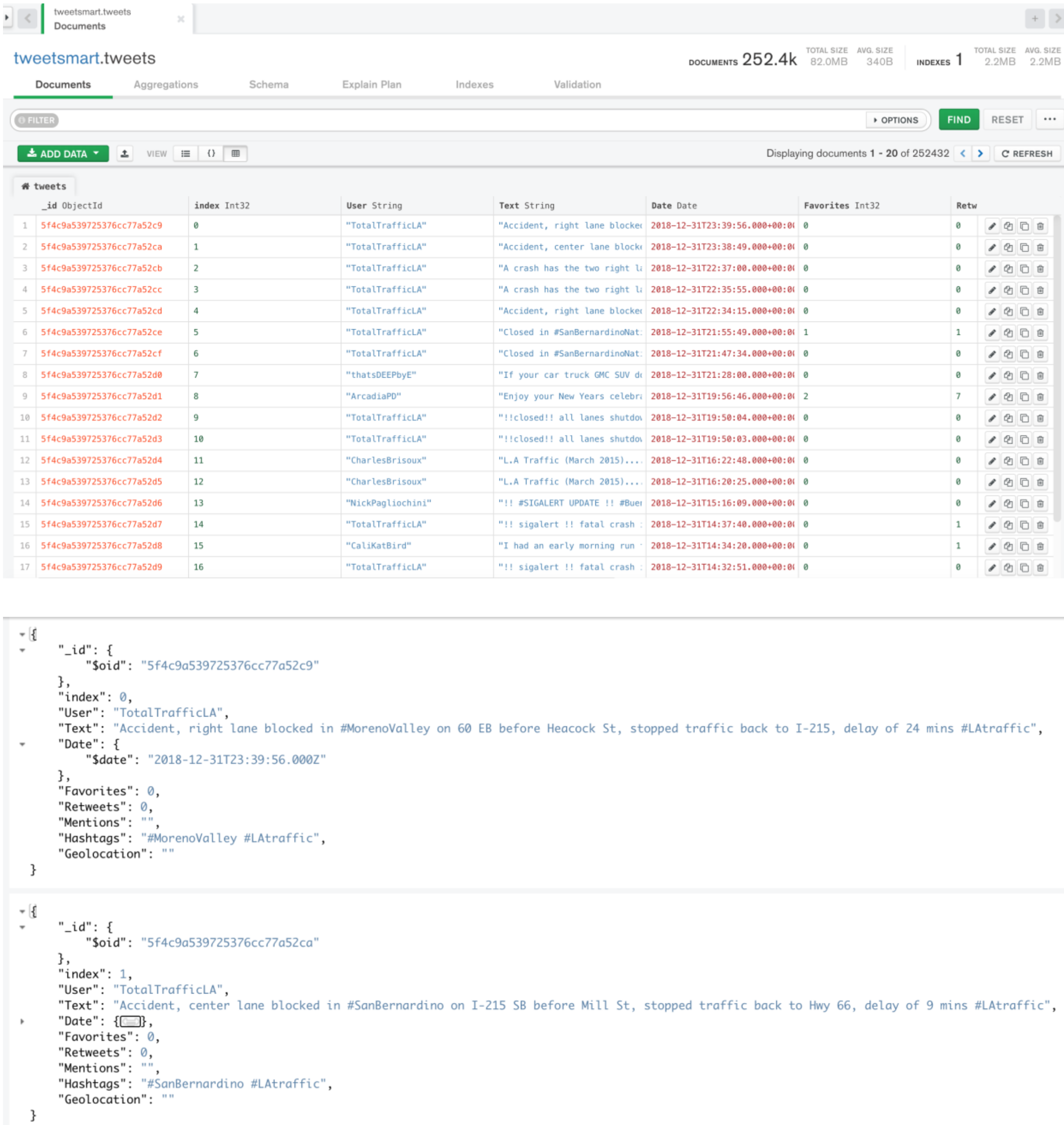


	A	B	C	D	E	F	G	H
1		Date	HighTemp	LowTemp	AverageTem	Precipitation	NauticalTwil	NauticalTwilightSet
2	0	1/1/17	55	39	47	0	6:01	5:54
3	1	1/2/17	54	48	51	0	6:01	5:55
4	2	1/3/17	55	45	50	0	6:01	5:56
5	3	1/4/17	57	46	52	0.07	6:02	5:57
6	4	1/5/17	61	52	56	0.31	6:02	5:57
7	5	1/7/17	63	46	54	0.28	6:02	5:59
8	6	1/8/17	73	48	60	0	6:02	6:00
9	7	1/9/17	61	50	56	0.58	6:02	6:00
10	8	1/10/17	57	47	52	0.14	6:02	6:01
11	9	1/11/17	62	53	58	0.34	6:02	6:02
12	10	1/12/17	54	47	51	0.68	6:02	6:03
13	11	1/13/17	61	44	53	1.00E-16	6:02	6:04
14	12	1/14/17	71	45	58	0	6:02	6:05

FIGURE 4 - LA WEATHER DATA FROM WUNDERGROUND.COM

TWITTER DATA

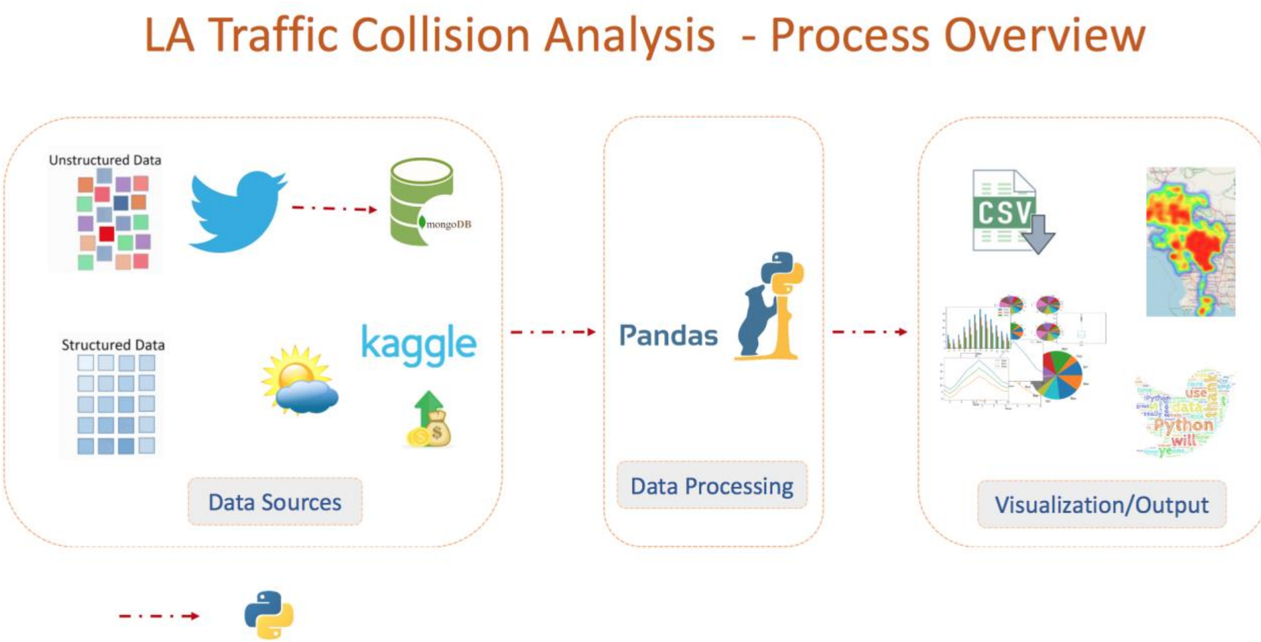
Using GetOldTweets3 (python library) import tweets between 2010 to 2019 with search keywords including #latraffic, #losangeles, #lapd to correlate LA road traffic collisions and their trend analysis from social media tweets. Here, trend in tweets are correlated directly to the number of accidents/collisions reported on Kaggle's collision dataset.



After pre-processing tweets collection has 252,432 rows and 18 columns (including derived columns such as date, year, hour, minute, month/monthname, weekday etc.)

PRE-PROCESSING

PROCESS OVERVIEW



CLEANING THE DATA

- The Median Income dataset was cleaned by doing string replace to remove the commas in the numbers and to take away the $ characters. It was also converted to float data type. After it was cleaned, it was ‘inner’ merged on the column *Council Districts* with the LA dataset.
- The Weather datasets for 2017 and 2018 were concatenated first to make a combined dataset. The weather data was converted to DateTime format, and then ‘inner’ merged with the LA dataset on the column *Date*.
- The different data types of each column were evaluated and converted to its desired type
- Columns that were not needed were then removed:
    - *DR Number*
      Area ID
      Crime Code
      Crime Code Description
      Premise Code
      Precinct Boundaries
      Date Reported

Neighborhood Councils (Certified)
Census Tracts
MO Codes
LA Specific Plans
Reporting District

- Blank values and NAs were removed with the dropna() function.
- *Time Occurred* column was broken up into hours into a *hours* column
- *Date* was converted to DateTime and broken up into *months*, *weekdays*, and *year* columns.
- Year subsets were created in order to give flexibility to analyze any given year (la_2017 and la_2018 were concatenated and used to filter main dataset to show only data from 2017 and 2018)
- *Location* was broken up into *longitude* and *latitude* columns to make it easier to analyze with map visualizations
- *Date Occurred* was dropped as well
- LA weather data from 2017-2019 was then merged with the traffic data set in a new laWeather dataset
- For the laWeather data set, the columns *Unnamed: 0* and *Location* were dropped since they were not needed anymore

STRUCTURED DATA DICTIONARY

A data dictionary with column names, description, data types, and processing steps is below. After everything was merged and cleaned, the final LA collision dataset for analysis had **90,855 rows** and **19 columns.**

Column	Description	Data Type	Range of Values

Area Name	The 21 geographic areas or Patrol Divisions given a name based on landmark or surrounding community it is responsible for	Object	'Devonshire', 'West Valley', 'Topanga', 'Mission', 'Hollywood', 'Olympic', 'Northeast', 'Rampart', 'Wilshire', 'West LA', 'Pacific', 'N Hollywood', 'Van Nuys', 'Foothill', 'Central', 'Hollenbeck', 'Newton', 'Southwest', 'Southeast', 'Harbor', '77th Street'
Time Occurred	Time of collision	Integer	Time values
Victim Age	Age of victim of car collision	Integer	Age values from 0-99
Victim Sex	Sex of the victims	Object	F - female M - male

	Genders called “H” and “N” were ignored in analysis since no indication what they represented from Kaggle website and also represented a very small amount		X - unknown
Victim Descent	Ethnicity of victim of collision	Object	A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
Premise Description	Indicates type of location where collision occurred	Object	42 unique values such as ‘STREET’, ‘PARKING LOT’, ‘FREEWAY’.
Address	Street address of collision	Object	Streets
Cross Street	Nearest intersection street to Address	Object	Cross street
Location	GPS coordinates of collision with longitude and latitude	Object	Latitude and Longitude coordinates

Zip Codes	Zip code of collision	Object (Converted from float to integer and then to string)	5 digit number
Council Districts	Council District number of collision	Integer	Values from 1-15
Median Income	Median Household Income associated with Council District,	Float	Dollar value
Date	Date of collision	DateTime	
year	Year of collision	Integer	Values from 2010 – 2019
month	Month of collision	Integer	Values from 1-12
weekday	Day of the week of collision	String	Monday to Sunday
hours	Hour in day of collision	Integer	Values from 1-23 (military time)
longitude	Longitude of location	Float	
latitude	Latitude of location	Float	
HighTemp	Highest observed temperature in the day	Integer	Temperature in F
LowTemp	Lowest observed temperature in the day	Integer	Temperature in F
AverageTemp	Average observed temperature in the day	Integer	Temperature in F
Precipitation	Amount of precipitation observed in inches	Float	Rainfall in inches
NauticalTwilightRise	Occurs when the center of the suns is between 6 - 12 degrees above the horizon. At this point artificial light is starting to not be needed for outdoor activities.	Time	

NauticalTwilightSet	Occurs when the center of the Suns between 6 12 degrees above the horizon. At this time artificial light is usually needed for outdoor activities	Time	
rain	Column to indicate if day had rain or not	Boolean	Yes or No

TABLE 1 - DATA VARIABLES INCLUDED IN ANALYSIS

TWITTER DATA DICTIONARY

Column	Description	Data Type	Range of Values
Date	UTC time when this tweet was created	String	Time values. ex. "2018-12-31T23:39:56.000Z"
_id	Unique Identifier for this tweet	Integer	Ex. "id:105011842332"
User	Actual UTF-8 Twitter text	String	Ex. "User": "TotalTrafficLA"
Text	Actual UTF-8 Twitter text	String	Ex. "Text": "Closed in #SanBernardinoNationalForest on Hwy 18 EB between Baldwin Lk Rd and Camp Rock Rd #LAtraffic http://bit.ly/10F395r"
Geolocation	Geographic location of this tweet as reported	Coordinates	"Geolocation": { " Geolocation ": [-75.14310264,40.05701649],"type":"Point"}
Retweets	Number of times this tweet has been retweeted	Integer	Ex."Retweets": 10
Mentions	What the tweet has mentioned about	String	Ex. "Mentions": "accident"
Hashtags	Key hashtags used in the tweet	String	Ex. "Hashtags": "#SanBernardinoNationalForest #LAtraffic"

METHODS OF ANALYSIS

Python software will be utilized to import the data, cleanse, develop models and create interesting visualizations to help understand the data. Analysis was broken up into four categories along with some questions to help guide the process. Each group member worked on at least one section:

1) Location

2) Demographics

3) Time/Day

4) Weather

By breaking up the analysis into four distinct categories it allowed for a more structured way to analyze the data. In order to guide analysis, a hypothesis was created for each category. This hypothesis was used as a way to find evidence to prove or disprove the statement.

Exploratory analysis was first done on all the data in order to understand the different kinds of distributions across attributes. Then multi-variable plots and visuals were created to see the types of relationships different variables. Analysis methods such as word clouds, maps, line charts, boxplots, bar graphs, heatmaps, and other types of visuals were used.

ANALYSIS: LOCATION

Hypothesis:

Certain areas of Los Angeles increase the likelihood of car collisions.

Fields:

Area Name, *Zip Codes*, *Council Districts*, *Cross Streets*, *latitude* and *longitude.*

The location was analyzed in many ways. Heat maps of the collisions were made to show patterns of where the concentration of collisions in Los Angeles occur.

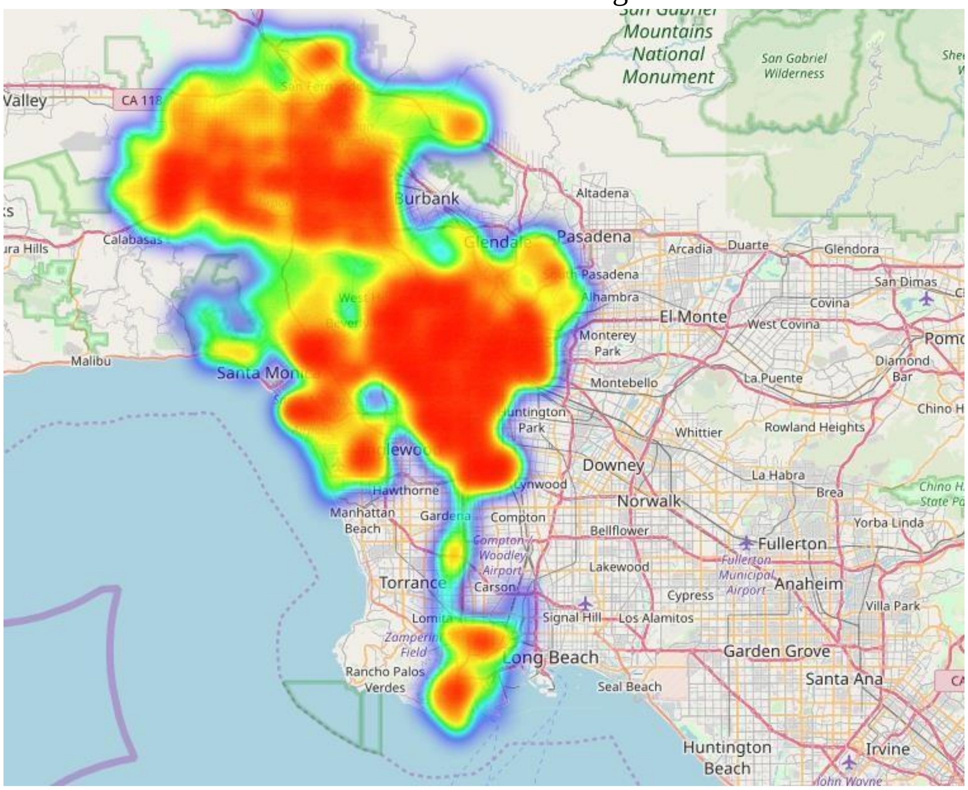


FIGURE 5, HEAT MAP OF ALL THE COLLISIONS FROM 2017 TO 2018

The map shows a concentration of collisions, the redder the color the higher concentration, in the middle of LA, towards the bottom, and farther north. These areas were also found to be concentrated during different times of the day.

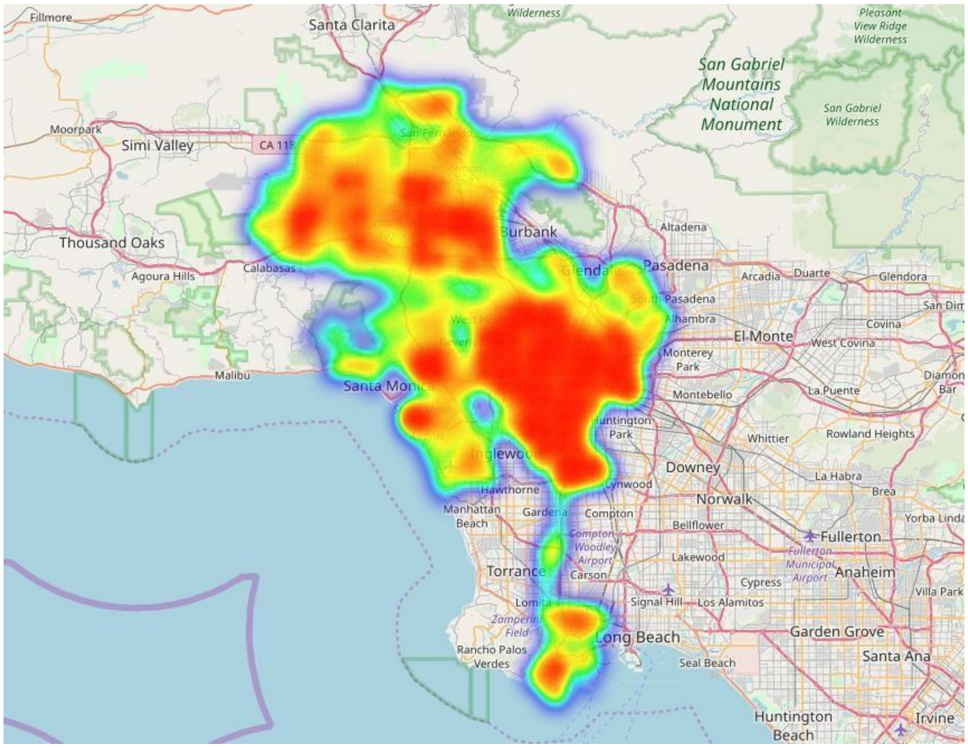


FIGURE 6, HEAT MAP OF CRASHES BETWEEN SUNRISE AND SUNSET

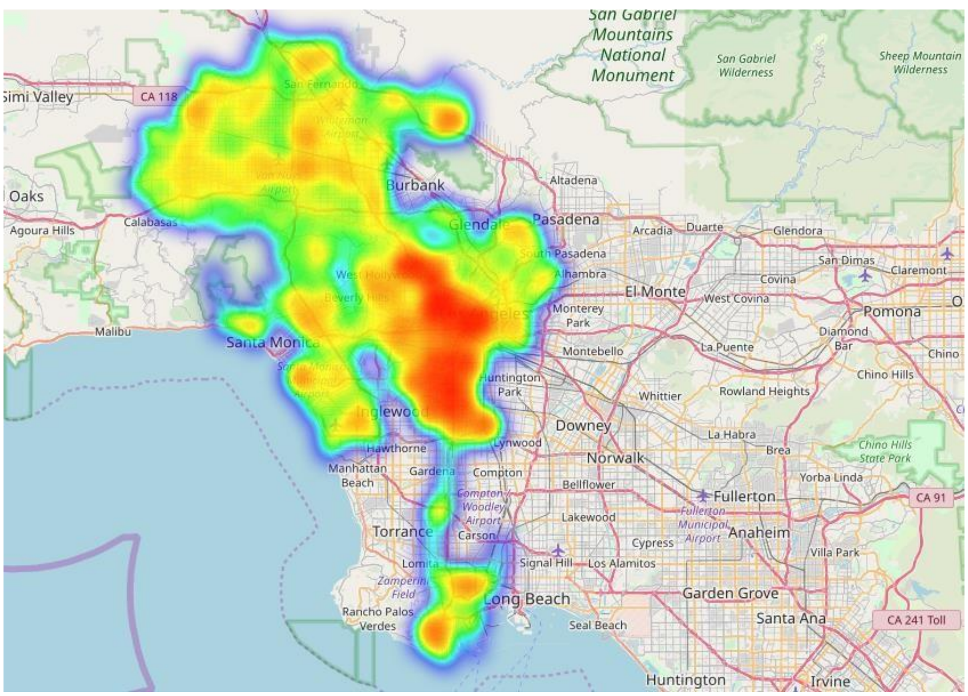


FIGURE 7, HEAT MAP OF CRASHES BEFORE SUNRISE AND AFTER SUNSET

These three heatmaps highlight how the center of LA has the most crashes at all times of the day. The map of crashes when it is visibly dark shows the northern part of LA has less accidents at night.

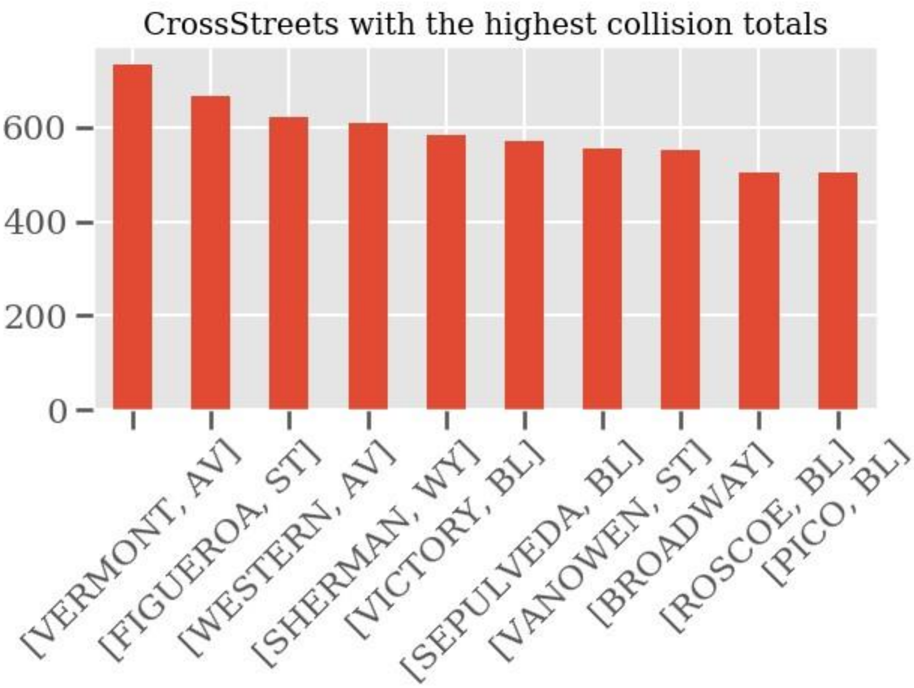


FIGURE 8, BARCHART SHOWING CROSS STREETS WITH MOST COLLISIONS

Intersections of Collisions

Cross Street	Address	Total Collisions
SEPULVEDA BL	SHERMAN WY	60
NORDHOFF ST	TAMPA AV	59
WHITSETT AV	SHERMAN WY	52
WOODMAN AV	SHERMAN WY	52
RODEO RD	BREA AV	50
SEPULVEDA BL	BURBANK BL	47
VICTORY BL	TOPANGA CANYON BL	47
PLUMMER ST	ST TAMPA AV	46

TABLE 2, MOST COLLISIONS BY SAME CROSS STREET AND ADDRESS

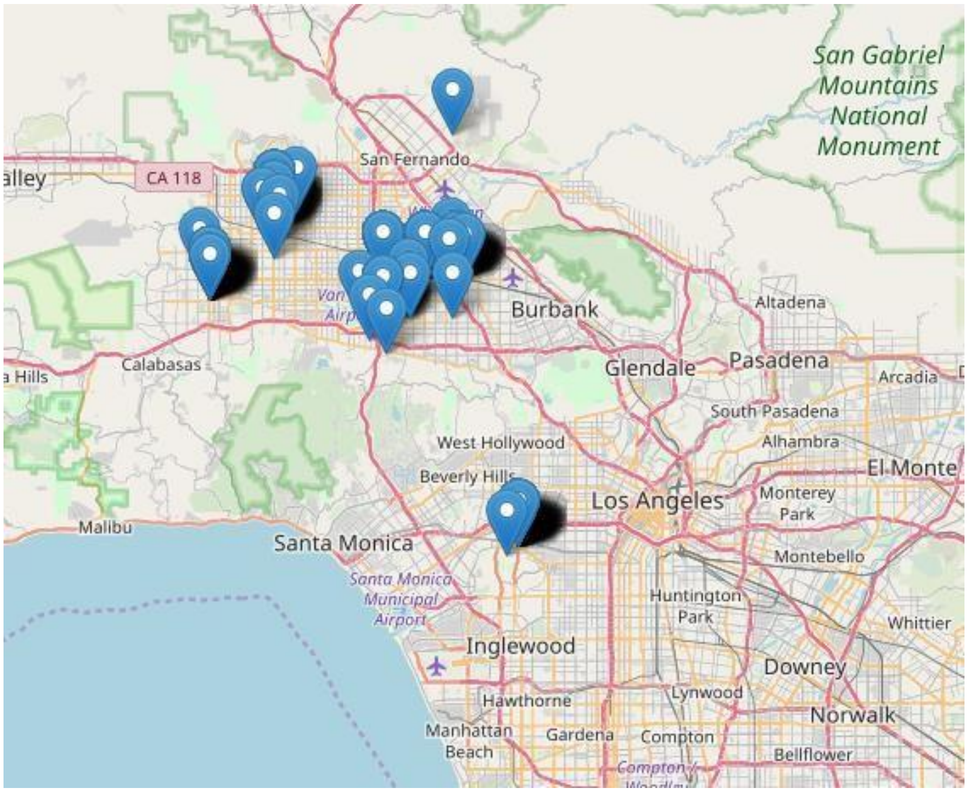


FIGURE 9, UNIQUE CROSS STREET AND ADDRESSES WITH MOST COLLISIONS (EACH POINT REPRESENTING MANY COLLISIONS)

Although the middle of Los Angeles had the most concentration of collisions, the northern parts near Burbank had unique combinations with more frequent accidents. These may want to be analyzed to see if there is an issue. The airport may be a heavy traffic area.

AREA NAME

Area Name was examined to find out which areas were the most impacted by collisions:

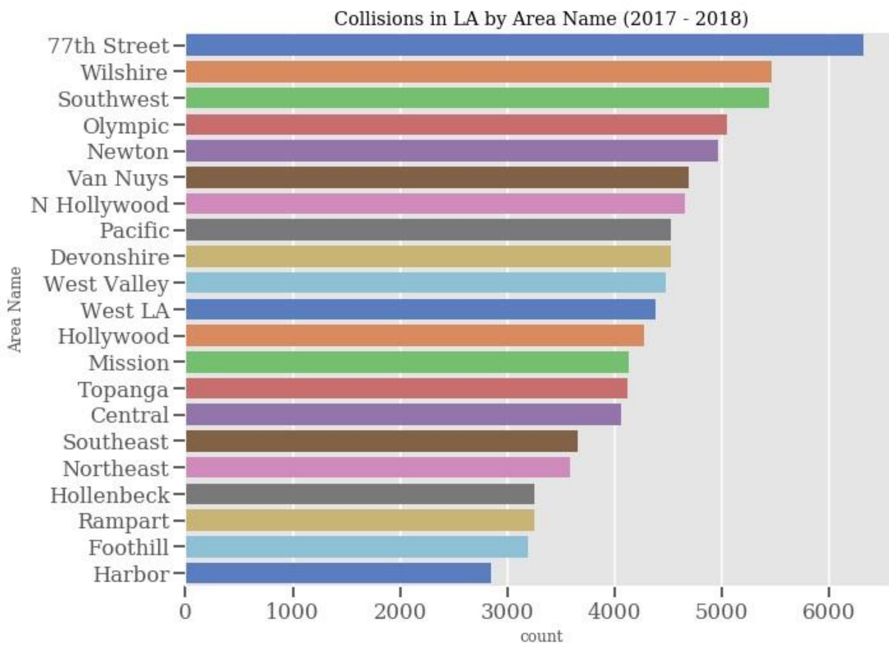


FIGURE 10 - LA COLLISIONS BY AREA NAME

COUNCIL DISTRICTS

Council Districts were also looked at:

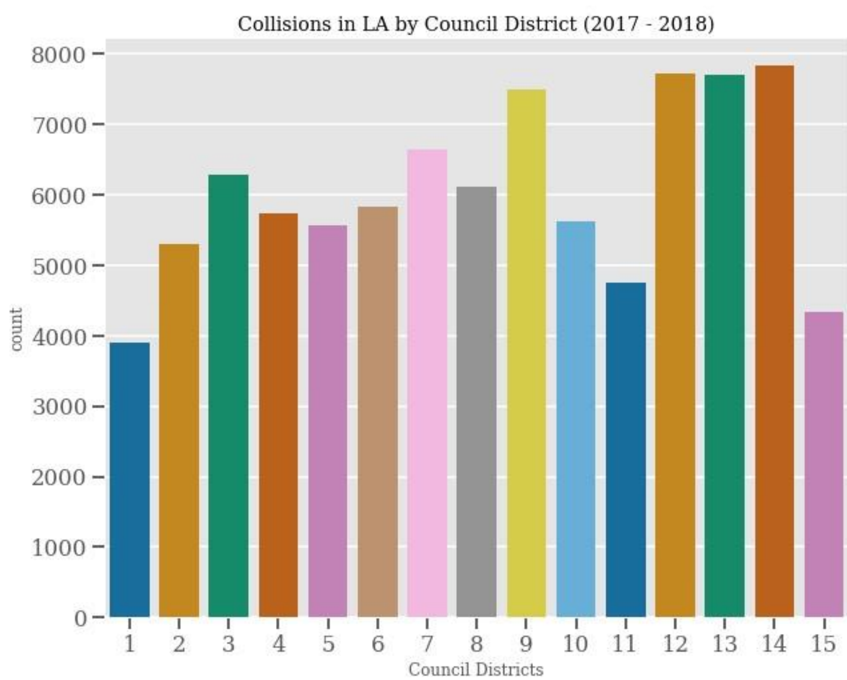


FIGURE 11 - LA COLLISIONS BY COUNCIL DISTRICT

STREETS

Streets from the *Address* field were examined to see which streets were involved in the most collisions:

```
WESTERN                      AV        1322SHERMAN
WY        1242VENTURA                      BL        1239VICTORY
BL        1129SEPULVEDA                    BL        1120VERMONT
AV        1083FIGUEROA                     ST        1053ROSCOE
BL         938VANOWEN                      ST         900OLYMPIC
BL         888VAN NUYS                     BL         876SUNSET
BL         834PICO                         BL         788BROADWAY
729
```

Addresses with the most collisions

WESTERN	AV	1322
SHERMAN	WY	1242
VENTURA	BL	1239
VICTORY	BL	1129
SEPULVEDA	BL	1120
VERMONT	AV	1083
FIGUEROA	ST	1053
ROSCOE	BL	938
VANOWEN	ST	900
OLYMPIC	BL	888
VAN NUYS	BL	876
SUNSET	BL	834
PICO	BL	788
BROADWAY		729
NORMANDIE	AV	701
WILSHIRE	BL	696
VENICE	BL	661
LAUREL CANYON	BL	634
CENTRAL	AV	622
WASHINGTON	BL	613
FLORENCE	AV	594
3RD	ST	576
LA BREA	AV	572
CRENSHAW	BL	557
TOPANGA CANYON	BL	557
RESEDA	BL	551
SATICOY	ST	542
BURBANK	BL	530
MAIN	ST	527
MANCHESTER	AV	517

FIGURE 12 - WORD CLOUD OF ADDRESS STREETS INVOLVED IN COLLISIONS

Streets from the *Cross Streets* field were examined to see which streets were involved in the most collisions.

```
Cross Streets with the most collisions

VERMONT                    AV        740
FIGUEROA                   ST        674
WESTERN                    AV        634
SHERMAN                    WY        612
VICTORY                    BL        587
SEPULVEDA                  BL        577
BROADWAY                             561
VANOWEN                    ST        558
ROSCOE                     BL        515
```

FIGURE 13 - WORD CLOUD OF CROSS STREETS INVOLVED IN COLLISIONS

When combining the cross streets with addresses, the **top 8 streets with the most collisions** was determined:

```
Cross Street                    Address
```

SEPULVEDA	BL	SHERMAN	WY	60
NORDHOFF	ST	TAMPA	AV	59
WOODMAN	AV	SHERMAN	WY	53
WHITSETT	AV	SHERMAN	WY	52
RODEO	RD	LA BREA	AV	50
SEPULVEDA	BL	BURBANK	BL	47
VICTORY	BL	TOPANGA CANYON	BL	47
PLUMMER	ST	TAMPA	AV	45

RESULTS AND FINDINGS: LOCATION

- The middle of the city has a similar density of collisions during at all times, at night, and during the day.
- When it is dark there are fewer collisions in the northern area compared to the daytime.
- The center of Los Angeles had the heaviest concentration of crashes but, the northern area had a handful of locations with more collisions than anywhere where else.
- Council Districts with the most collisions: 12/13/14.
- Cross Streets with the most collisions:
- SEPULVEDA BL & SHERMAN WY
- NORDHOFF ST & TAMPA AV
- WHITSETT AV & SHERMAN WY
- RODEO RD & LA BREA AV

ANALYSIS: DEMOGRAPHICS

Hypothesis:

Certain ages, genders, and ethnicities make an individual more susceptible to becoming a victim of a car collision.

Fields:

Victim Age, Victim Descent, Victim Sex, Median Income, Council Districts

VICTIM GENDER

Breakdown of gender counts in collisions:

M	53676
F	36914
X	231
H	18
N	6

There are substantially more men involved in collisions compared to women:

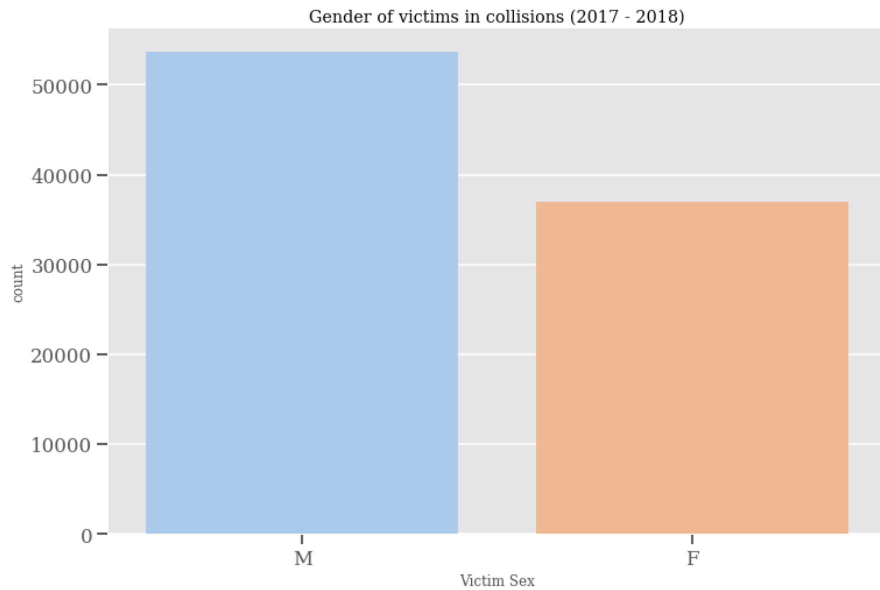


FIGURE 14

- **Men comprise 59.09% of victims In a car crash in Los Angeles.** This is larger than the 50% male population in Los Angeles. (https://censusreporter.org/profiles/16000US0644000-los-angeles-ca/)
- Many possibilities could account for this, such as how men are often more aggressive in general and this could possibly translate onto the road. Further exploration is necessary.

FIGURE 15 - COMPARISON OF GENDERS IN COLLISIONS

We then looked at the gender breakdown by the ethnicity of the victims in order to see if there were any interesting relationships:

Hispanic

```
M    22328 F    14662 X        4 H        3 N        1
```

Asians

```
M    1987 F    1675 N        1 X        1 H        1
```

Whites

```
M    12004 F      8122
H        2 N        1
```

Blacks

```
M    6971 F    6283 H        6 X        1 N        1
```

- For Asians and Blacks, the gender breakdown of victims is fairly even, near 50%, which is in line with the 50% gender divide in Los Angeles among all races.
- **For Whites and Hispanics, though, males make up 60% of the gender of victims within the same race.**
    - This is an interesting deviation from the standard 50/50 gender divide in the Los Angeles population. This could be due to a culture differences or some other factor. It would be interesting for further examination.

VICTIM AGE

An exploration of the distribution of victim ages was done:

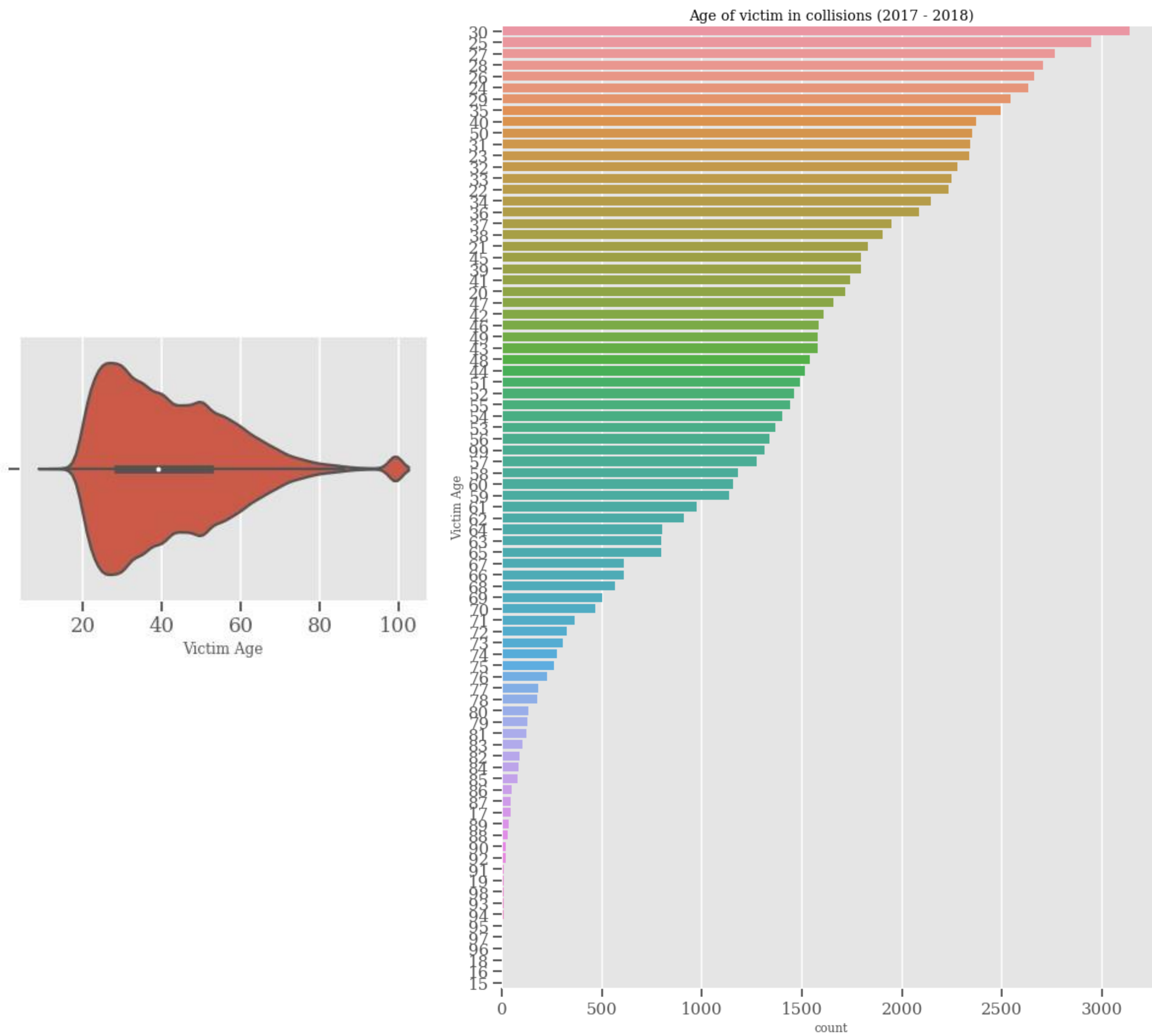


FIGURE 16 - COMPARISON OF VICTIM AGE IN COLLISIONS

- From this we can see that the **top 5 victim ages are all under 30 years old** and a younger demographic.
- **The average age of a victim was 41.81 years old** compared to the average age in Los Angeles of 35.8 years old .[1]

Victim Age distribution by race:

[1] https://censusreporter.org/profiles/16000US0644000-los-angeles-ca/

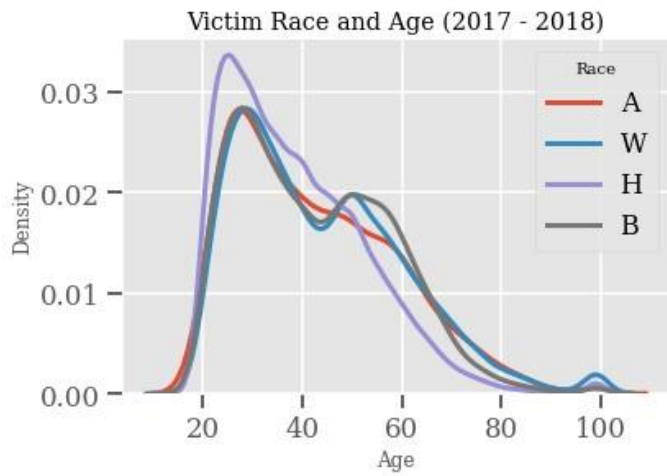


FIGURE 17 - COMPARISON OF VICTIM RACE AND AGE IN COLLISIONS

Most of the victim ages spike from the range of 20 to 40 years old for all the major ethnicities.

Victim age distribution by gender:

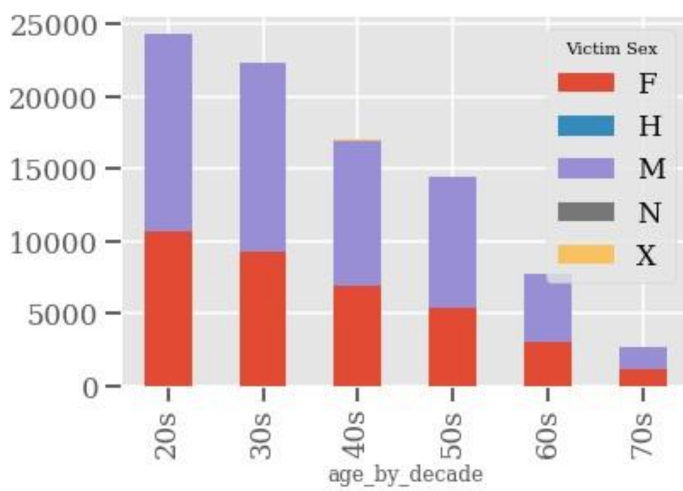


FIGURE 18, BINNED AGES AND GENDERS

Binned age groups help show the gender breakdown for each age decade. Men unsurprisingly hold a higher proportion of victims for every decade until around the 70s where it is almost 50/50. We can also see that as each decade passes, the amount of victims goes down.

VICTIM DESCENT

Breakdown of victim descents in collisions:

H	37657
W	20484
B	13529
O	12256
A	3730
X	1820
K	717
F	285
C	126
U	48
J	42
Z	35
V	35
P	30
I	29
G	12
S	5
D	4
L	1

The ethnicities of victims in the most crashes were analyzed:

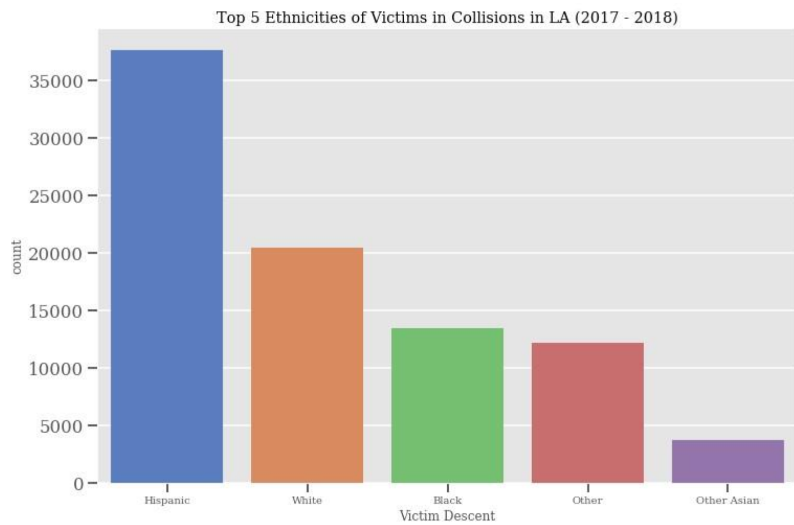


FIGURE 19, TOP 5 ETHNICITIES IN COLLISIONS

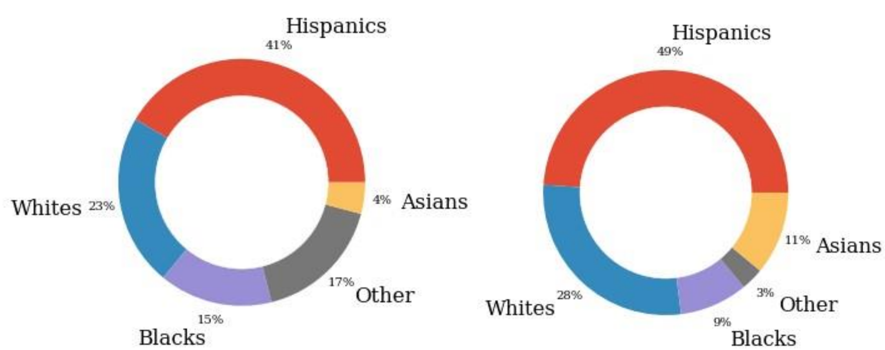


FIGURE 20, COLLISION ETHNICITIES (LEFT) AND LA POPULATION ETHNICITIES (RIGHT)

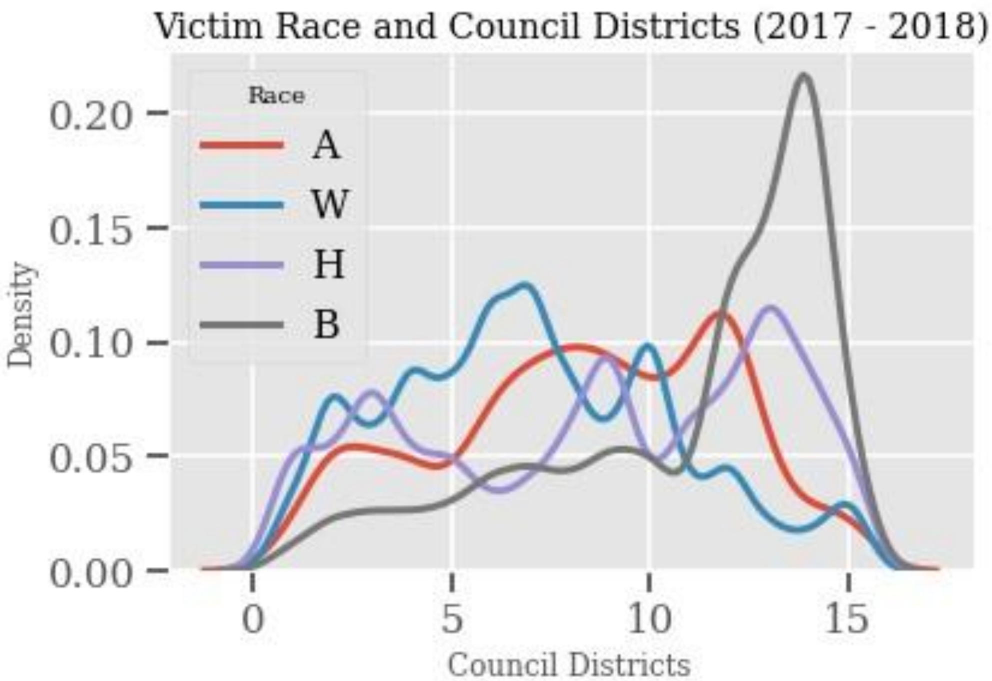
SOURCE: [HTTPS://CENSUSREPORTER.ORG/PROFILES/16000US0644000-LOS-ANGELES-CA/](https://censusreporter.org/profiles/16000US0644000-los-angeles-ca/)

FIGURE 21 - COMPARISON OF ETHNICITIES IN COLLISIONS

- Hispanics comprise the highest total for ethnicities involved in a collision (41.45%) followed by Whites (22.55%) and Blacks (14.89%). It Is not surprising that Hispanics are the highest since there is a 49% Hispanic population in Los Angeles. The high percentage is actually below the population average.
- Whites 22.55% total Is below the 28% total of Whites in Los Angeles population.

- **Blacks only compose 9% of the population yet are victims in 14.89% of car accidents**.
- **Asians make up only 4% of car victims with a 11% population in general in Los Angeles**.

We also looked at how Council Districts related to victim race:



a

FIGURE 22 - COMPARISON OF COUNCIL DISTRICTS IN COLLISIONS

It seems that the latter number districts are composed with more Hispanics and Blacks. For whatever reason, the districts from 10-15 seem to have a lot more collisions which are areas comprised of mostly Blacks

VICTIM INCOME

Victim Income Breakdown		
Income Range	Count	Percentage
30-39k	13410	15.86%
40-49k	28816	34.08%
50-59k	12266	14.51%
60-69k	10900	12.89%
70-79k	19166	22.66%

TABLE 3 - VICTIM INCOME BREAKDOWN

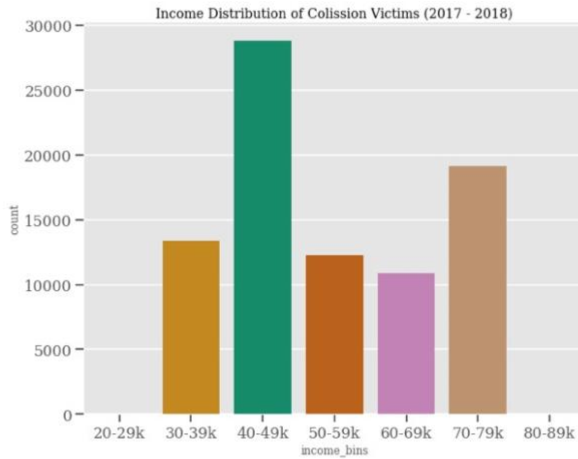


FIGURE 16 - INCOME DISTRIBUTION OF COLLISIONS

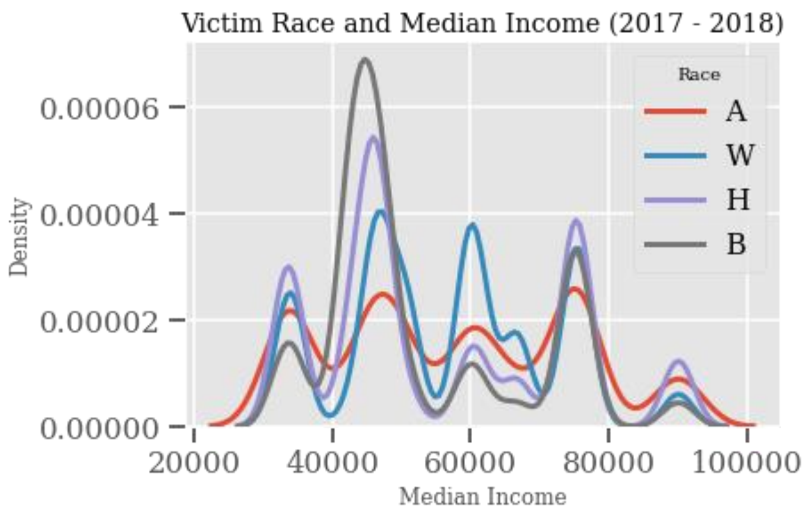


FIGURE 23 - VICTIM RACE AND MEDIAN INCOME IN COLLISIONS

- Here we see a spike in the 40k income range for the Hispanic and Black demographic. It is the highest point for both races. For Whites, the peak is at around 55k while Asians peak around 63k.
- Average income for Blacks in LA is $34,500 while its peak income for victims is around 37k
- Average income for Whites in LA is $61,100 while its peak income for victims is around 55k
- Average income for Hispanics in LA is $40,300 while its peak income for victims is near 40k
- Average income for Asians in LA is $57,800 while its peak income for victims is around 65k
    - This is an interesting finding since Asians were also much likely to be victims in a car collision compared to others, and they are also the only group that when someone does become a victim, they tend to be from the higher range of income.

Source: https://statisticalatlas.com/place/California/Los-Angeles/Household-Income

Results and Findings: Demographics

- Men are more likely to be in an accident compared to women.
- Frequency of collisions is proportional to race/ethnicity.
- Age 30 has the highest number of collisions and the top 5 ages are all below 30.

- People who self-identify as Black, compose 9% of the population but are victims in 14.89% of car accidents.
- People who self-identify as Asian make up only 4% of car victims but comprise 11% of the population.
- 34.08% of accident victims have a median income between $40-49K.

ANALYSIS: TIME/DAY

Hypothesis:

Certain times of the day and days of the week are more dangerous and result in more car collisions.

Fields:

Month, year, hours, weekday, NaughticalTwilightSet, NaughticalTwilightRise

YEARLY PATTERN

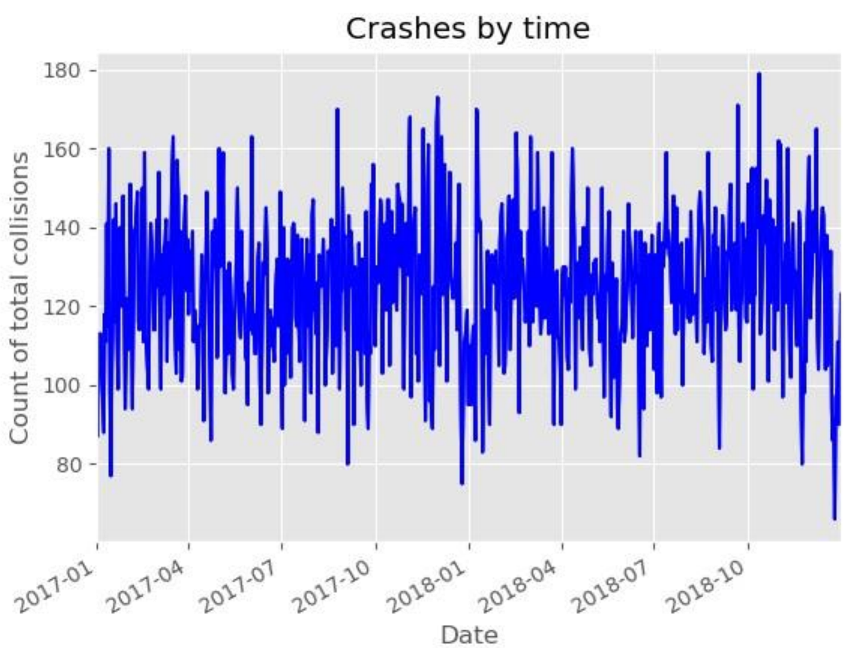


FIGURE 24, COLLISIONS IN A TWO YEAR PERIOD

This is a classic example of a time series plot. Clearly there is some variation in the bigger picture as the thickest blue area varies in height, but on a smaller level there is much more variability.

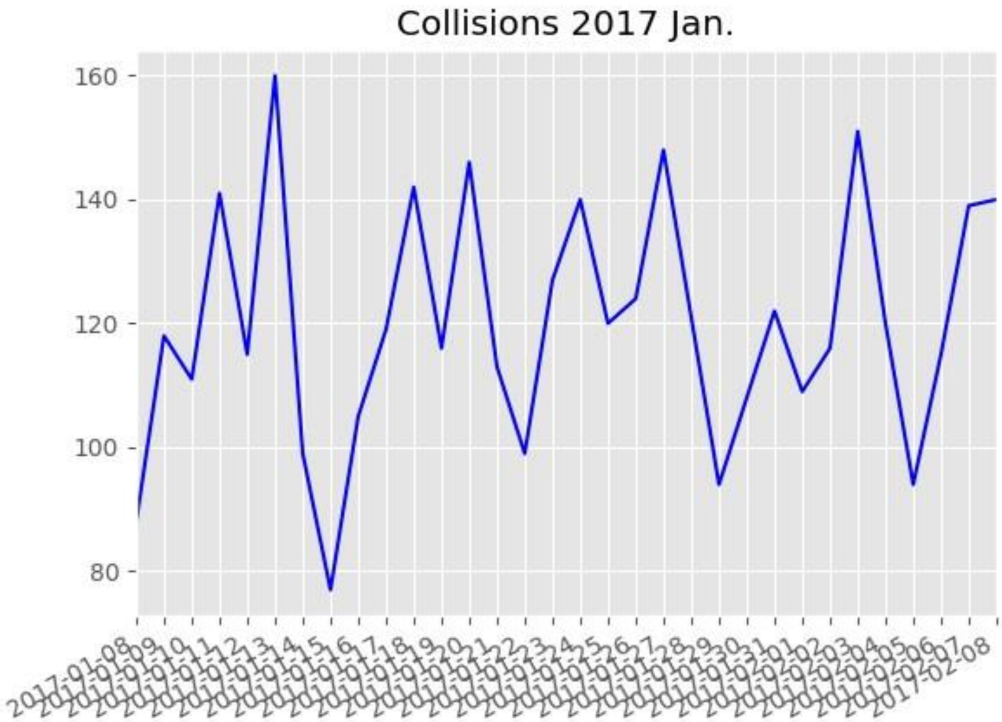


FIGURE 25, COLLISIONS BY DAY IN A ONE MONTH PERIOD

Here the variability from the graphs is more noticeable when viewed from a one-month period. The line has a pattern of building up, slightly dipping, peaking, then falling again. The lowest points are Sunday. As the week goes on Monday has more crashes, Tuesday even more, Wednesday has increased collisions. Then collisions drop off some on Thursday. **Friday is the peak collision day.** Each of the highest points are Fridays. Collisions drop off on Saturdays and Sundays.

Summary of top weekday occurrences

Friday	14313
Wednesday	13456
Thursday	13270
Tuesday	13054
Monday	12627
Saturday	12157
Sunday	11092

The time of day shows a pattern of when collisions occurred:

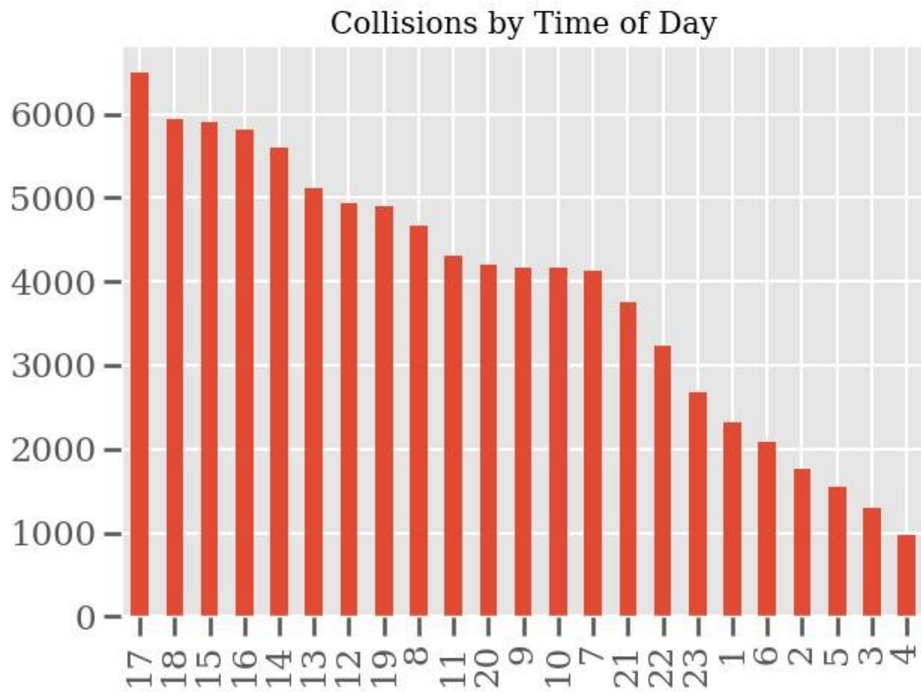


FIGURE 26, COLLISIONS BY THE HOUR IN MILITARY TIME

This graph shows that most collisions in LA occur around 12PM to 5PM. It appears that Friday and Wednesday at from 12PM to 5PM are peak times, however, this is by no means a majority of the collisions.

DAY OF WEEK

Day of week was explored to see if any particular days were troublesome:

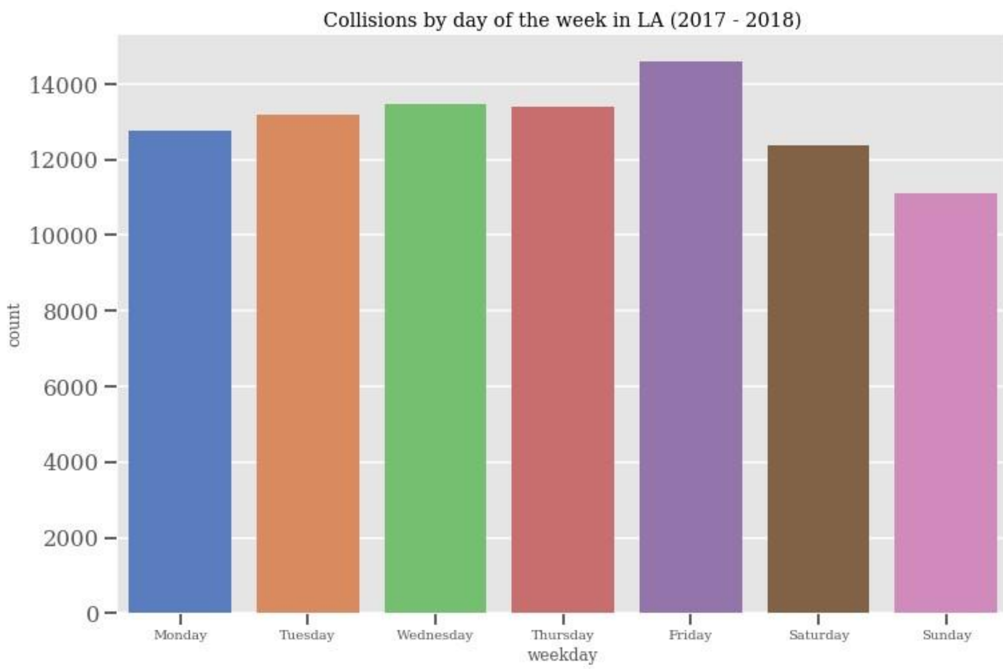


FIGURE 27 - COLLISIONS BY DAY OF THE WEEK 2017-2018

This shows the counts by day.

MONTH

Collisions by month was looked at to see which months had an increased chance of car accidents:

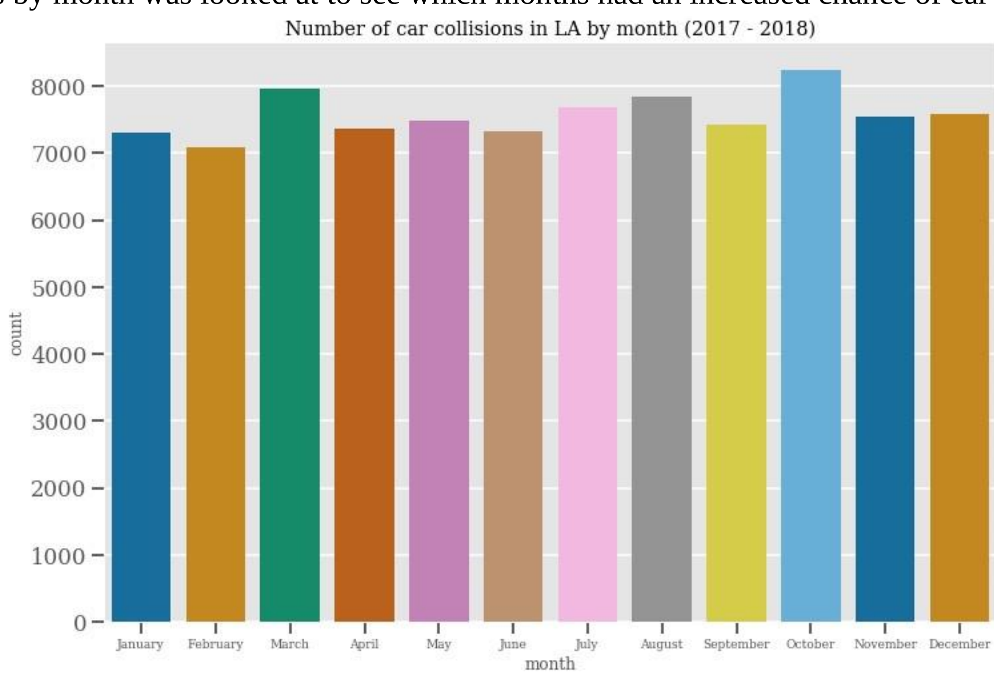


FIGURE 28 - CAR COLLISIONS BY MONTH 2017-2018

We can see that October is the most common month as well as March and August. 2017 January to April month patterns:

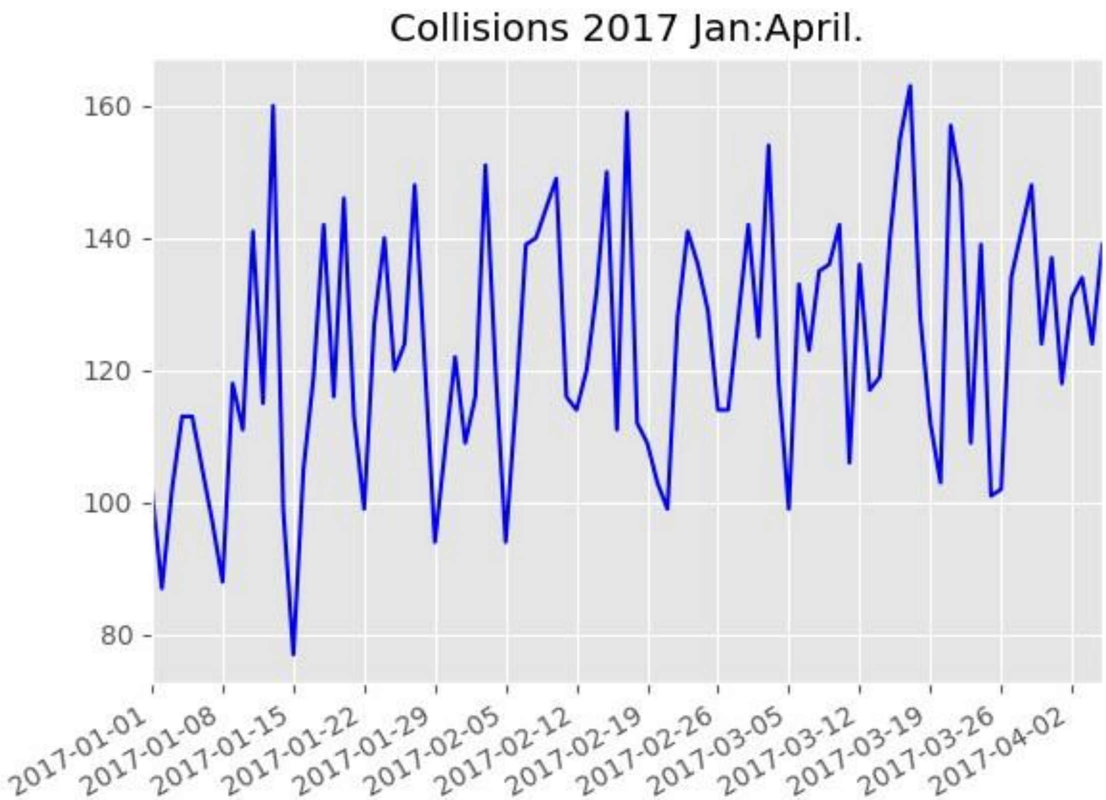


FIGURE 29

HOUR OF DAY

The hour of day was examined by day of the week:

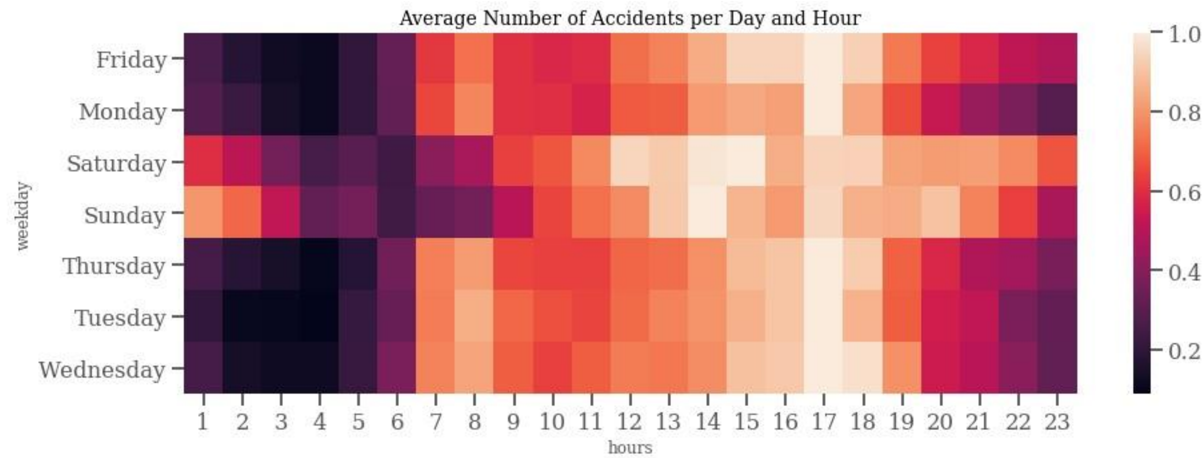


FIGURE 30 - HEATMAP OF COLLISIONS BY DAY OF WEEK AND HOURS

It is interesting to see that there are a lot of darker boxes from 1am to 6am during the weekdays, but that changes on the weekend during 1am to 3am. These brighter boxes indicate more collisions. A possible reason for this could be due to people going out during these late hours since they do not have work the next day.

Victim descent was analyzed during daylight hours and sundown hours. Crashes were higher during the day overall, and no significant patterns emerged. Victim descent in day and night collisions.

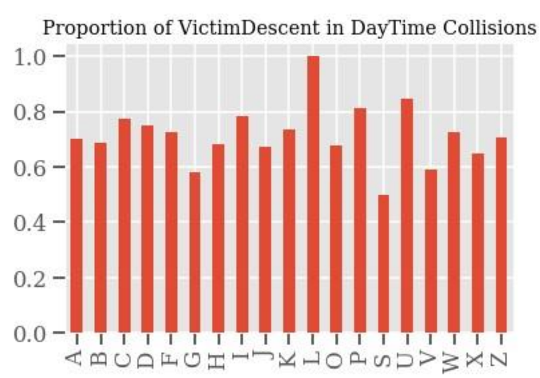


FIGURE 31

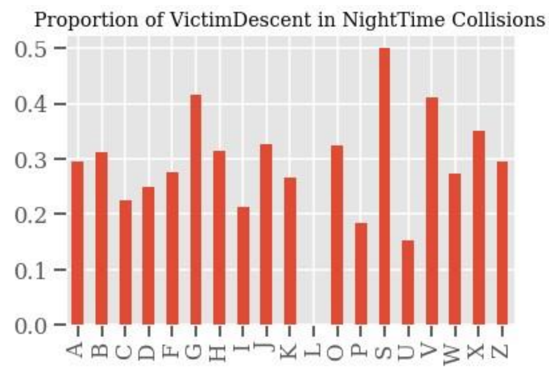


FIGURE 32

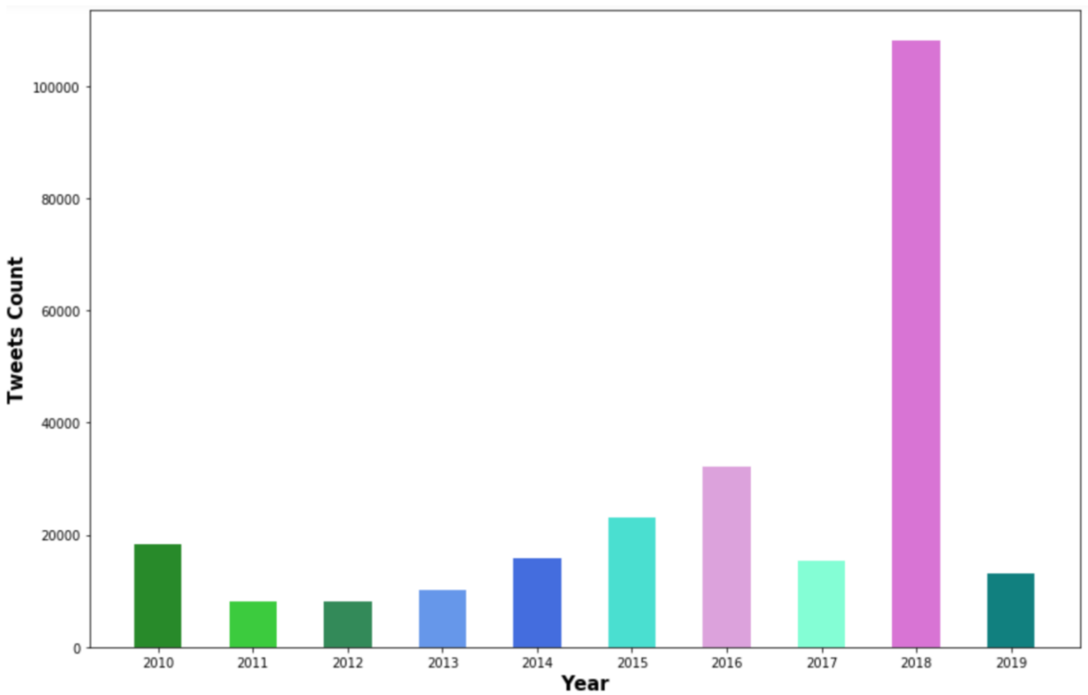
Summary of most collisions by specific dates in 2017-2018

2018-10-12	179
2017-12-01	173
2018-09-21	171
2017-08-25	170
2018-01-08	170
2018-01-09	169
2017-11-04	168
2017-11-30	166
2017-11-17	165
2018-12-07	165

These dates do not seem to be special, however there may have been large events or conventions during these dates. It was hypothesized that certain dates like New Year's eve would have the most crashes.

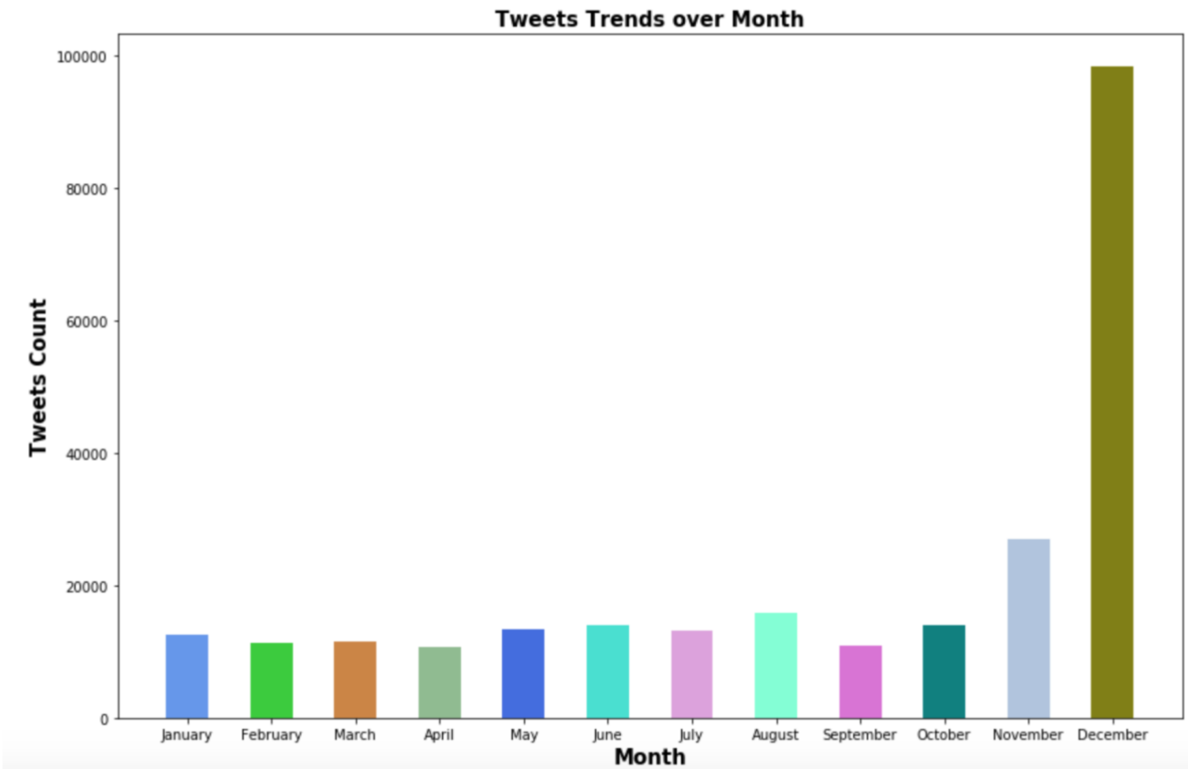
TWEETS ACROSS YEARS 2010 THROUGH 2019

Below chart shows the number of tweets mentioning latraffic spread across 2010 to 2019. Twitter API returned more data on 2018 compared to other years in scope.



TWEETS ACROSS YEARS 2010 THROUGH 2019 BY MONTH

Below chart shows the number of tweets mentioning latraffic spread by month across 2010 to 2019. Twitter API returned more data on December 2018.Hence the spike in month of December.



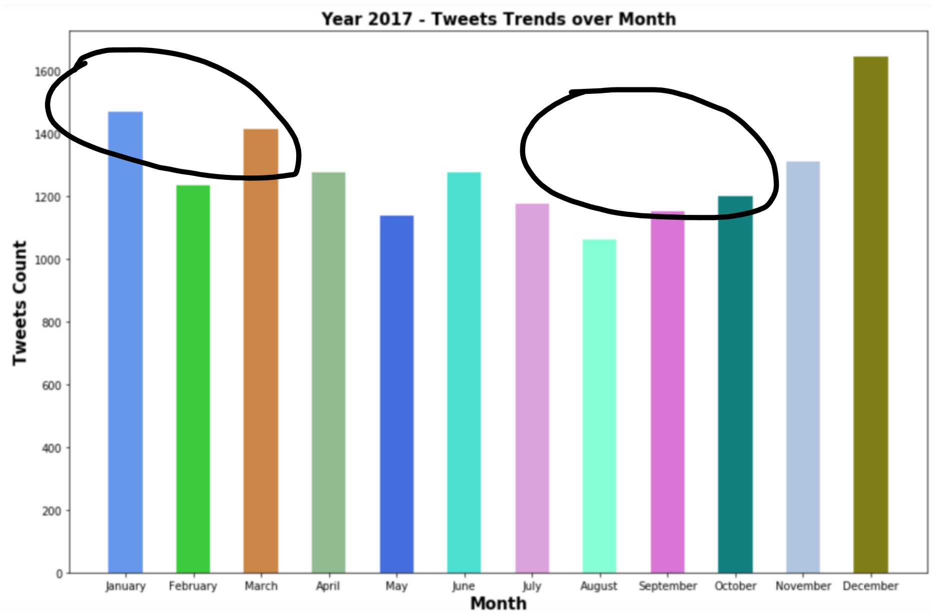
2017- 2018 TWEETS BY MONTH

In alignment with data analysis on Kaggle's collision data – twitter analysis restricted to 2017-2018 data as well. Here, trend in tweets are correlated directly to the number of accidents/collisions reported on Kaggle's dataset. Based on that, below table summarizes overall spread of number of tweets (collisions) by month.

2017 - 2018 Tweets by Month:

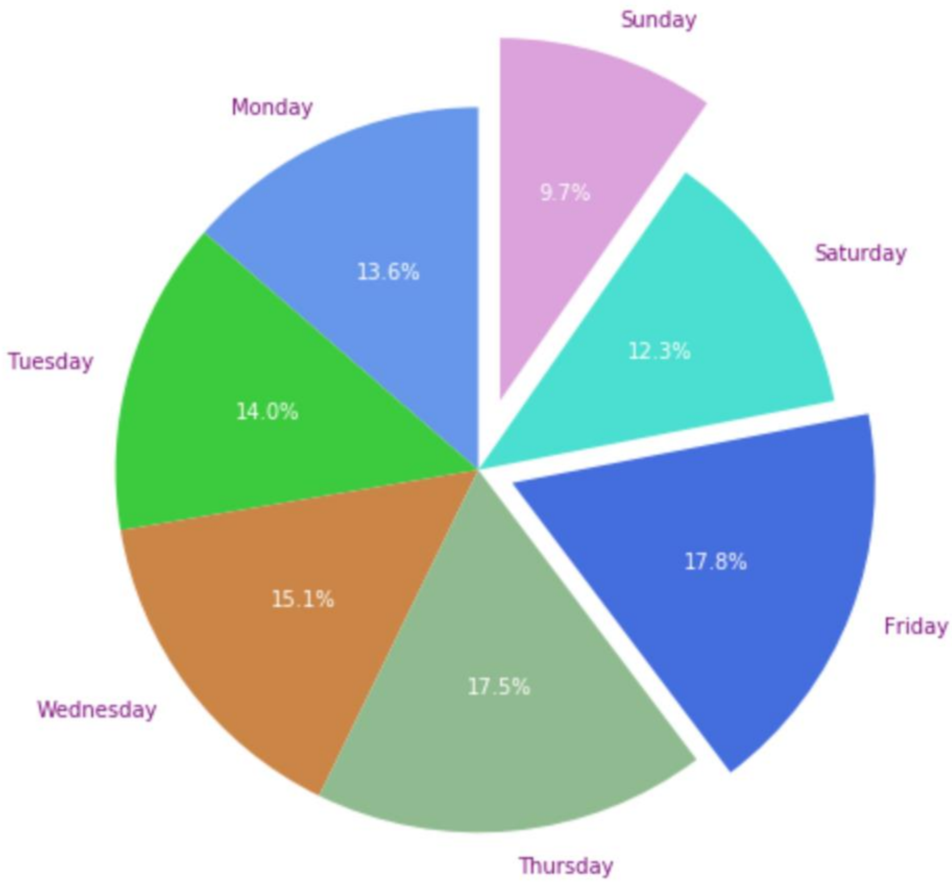
	year	month	monthname	counts
1	2017	1	January	1472
2	2017	2	February	1237
3	2017	3	March	1417
4	2017	4	April	1277
5	2017	5	May	1139
6	2017	6	June	1276
7	2017	7	July	1176
8	2017	8	August	1062
9	2017	9	September	1152
10	2017	10	October	1200
11	2017	11	November	1311
12	2017	12	December	1646
13	2018	8	August	2816
14	2018	9	September	1204
15	2018	10	October	1363
16	2018	11	November	14576
17	2018	12	December	88142

Let’s, zoom in on 2017 tweets as the data is consistent across whole year. To balance out the numbers, whenever there is a seasonal change – the number of tweets seems to increase.



2017- 2018 Tweets by Weekday

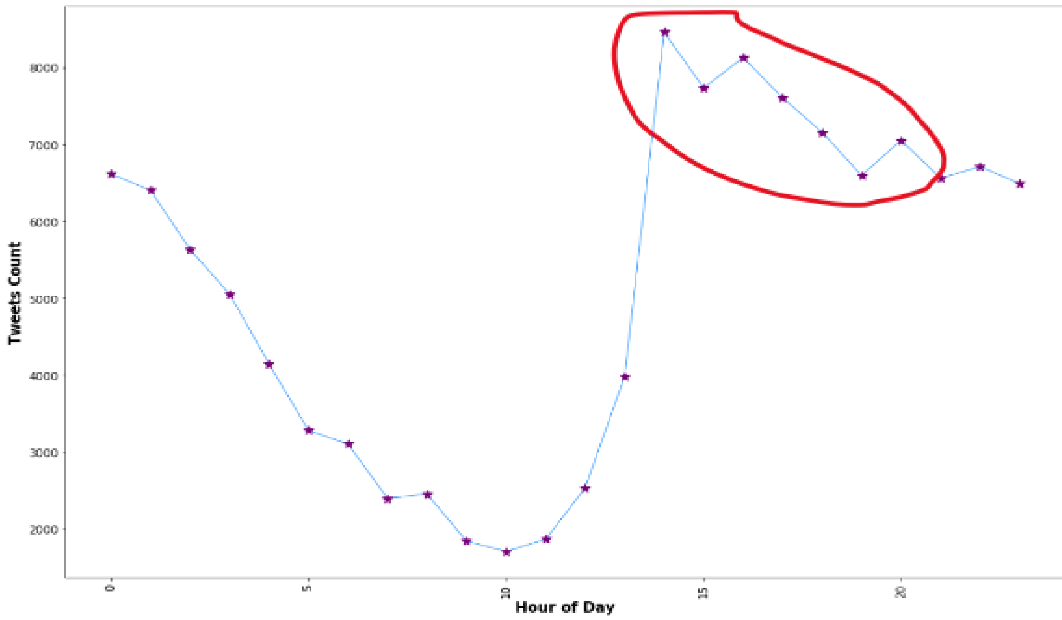
Moving onto weekday analysis. Chart below shows, higher number of tweets on Friday followed by Thursday and least on Sundays. Suggesting, more collisions/accidents towards end of the week than beginning of week. This pattern is also in alignment with Kaggle’s collision report suggesting more accidents on Fridays and least on Sundays.

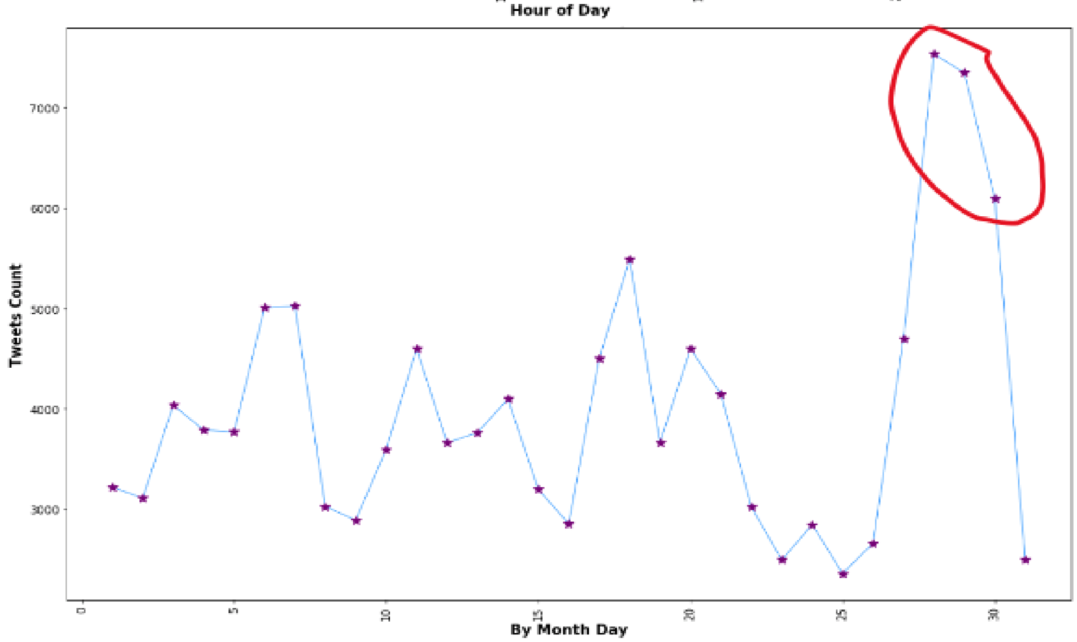


2017- 2018 Tweets by Hour of day

Further into hourly analysis - Chart below shows, higher number of tweets between 15 to 20h. Suggesting, more collisions/accidents towards end of the day as more people returning from work. This pattern is also in alignment with Kaggle's collision report suggesting more accidents during this hour range. With 10 AM being the least suggesting less to minimal traffic during this window.

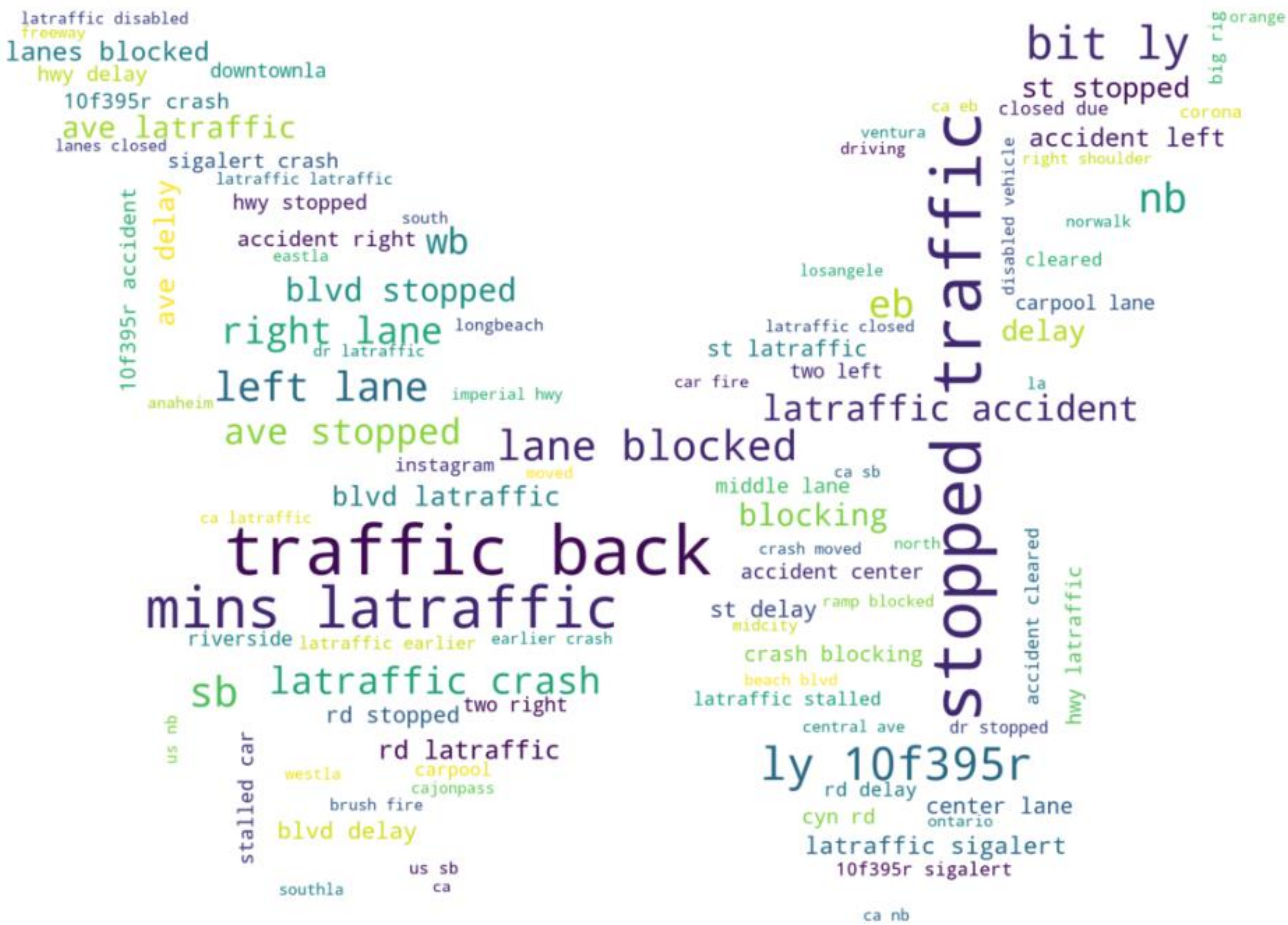
In addition, towards end of the month, there seems like more accidents/collisions based on the day of the month analysis followed.





WORDCLOUD

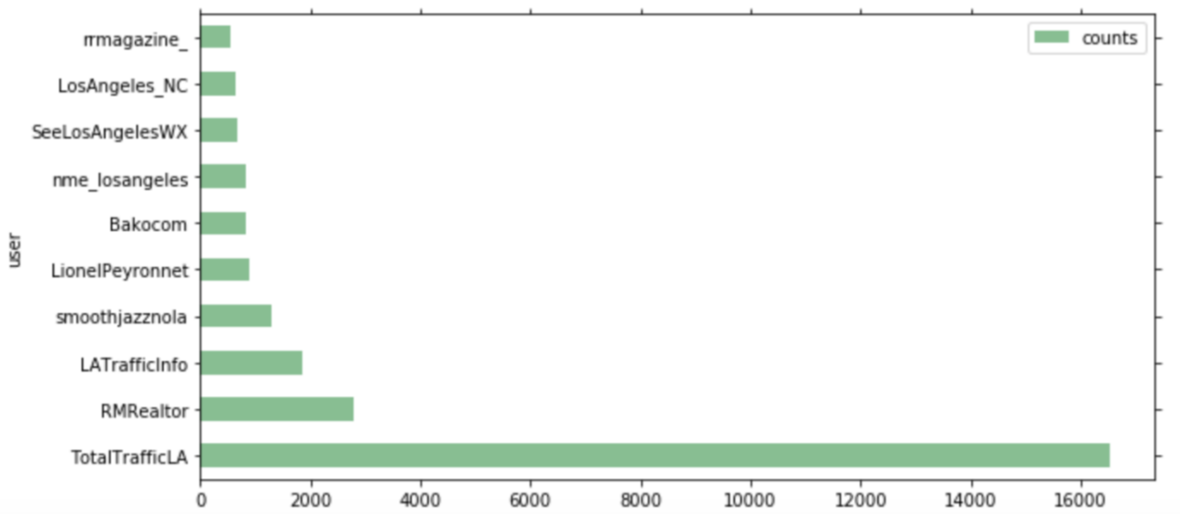
Finally, a simple Wordcloud depicting the overall talk about the words(tokens) trended in tweets are correlated directly to the number of accidents/collisions reported on Kaggle's dataset.


latraffic disabled
freeway
lanes blocked
hwy delay
downtownla
10f395r crash
ave latraffic
lanes closed
sigalert crash
latraffic latraffic
hwy stopped
south
accident right
wb
eastla
10f395r accident
ave delay
blvd stopped
right lane
longbeach
dr latraffic
left lane
anaheim
imperial hwy
ave stopped
lane blocked
instagram
moved
blvd latraffic
ca latraffic
traffic back
mins latraffic
riverside
latraffic earlier
earlier crash
sb
latraffic crash
rd stopped
two right
us nb
stalled car
rd latraffic
westla
carpool
cajonpass
brush fire
blvd delay
us sb
southla
ca
bit ly
orange
big rig
st stopped
ca eb
closed due
corona
ventura
driving
accident left
right shoulder
stopped traffic
disabled vehicle
nb
norwalk
cleared
losangele
eb
carpool lane
latraffic closed
delay
st latraffic
two left
car fire
la
latraffic accident
ca sb
middle lane
blocking
crash moved
north
accident center
accident cleared
hwy latraffic
st delay
ramp blocked
midcity
crash blocking
beach blvd
latraffic stalled
central ave
dr stopped
ly 10f395r
rd delay
center lane
ontario
cyn rd
latraffic sigalert
10f395r sigalert
ca nb

MOST TRENDING USERS

```
Top 10 | 2017 - 2018 Tweets by Users:

TotalTrafficLA :  16,513
RMRealtor :  2,793
LATrafficInfo :  1,855
smoothjazznola :  1,296
LionelPeyronnet :  891
Bakocom :  843
nme_losangeles :  818
SeeLosAngelesWX :  668
LosAngeles_NC :  643
rrmagazine_ :  540
```



RESULTS AND FINDINGS: TIME/DAY

- The months have some variability, but some patterns are more obvious collisions have a cycle on the line charts.
- Trends occur by day of the week. On Sunday they are the lowest. Then increase on Monday, Tuesday, and Wednesday. Thursday's are slightly lower on average, then Crashes Peak.
- Holidays were assumed to have the most crashes; however, the data does not support this.
- The highest frequency of accidents is on Monday-Friday between 4-5pm

ANALYSIS: WEATHER

Hypothesis:
Colder and adverse weather (ex. Rain, severe heat) result in more car collisions.

Fields:
Temperature, precipitation

TEMPERATURE

Temperature was analyzed to see what kind of factor it played in collisions:

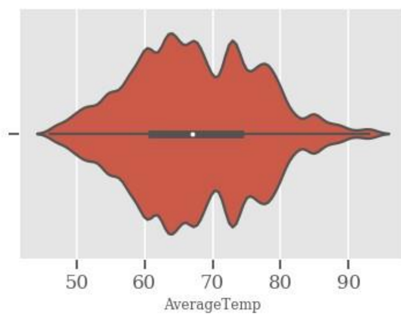


FIGURE 33, AVERAGE TEMP

The average temperature for the days of collisions is 67.46 degrees F which is higher than the average temperature in Los Angeles of 60.95 degrees F.[4]

Low temperatures below 40 degrees F and high temperatures above 100 degrees F were compared in the graph below in order to see the effect of extreme weather. There does not appear to be a correlation between low/high temperatures and more accidents. The average annual high temperature in LA is 72 degrees F and the average low temperature is 64 degrees F.[2]

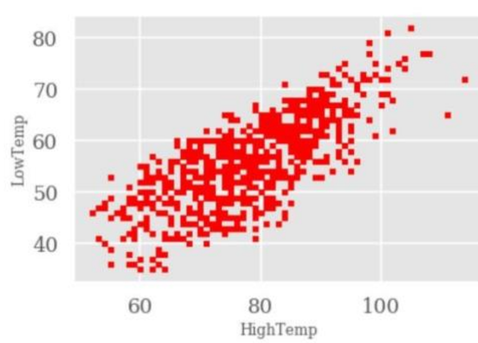


FIGURE 34 - SCATTER PLOT OF LOW/HIGH TEMPERATURES AND COLLISIONS

2 : https://www.usclimatedata.com/climate/los-angeles/california/united-states/usca1339

[4] https://www.usclimatedata.com/climate/california/united-states/3174

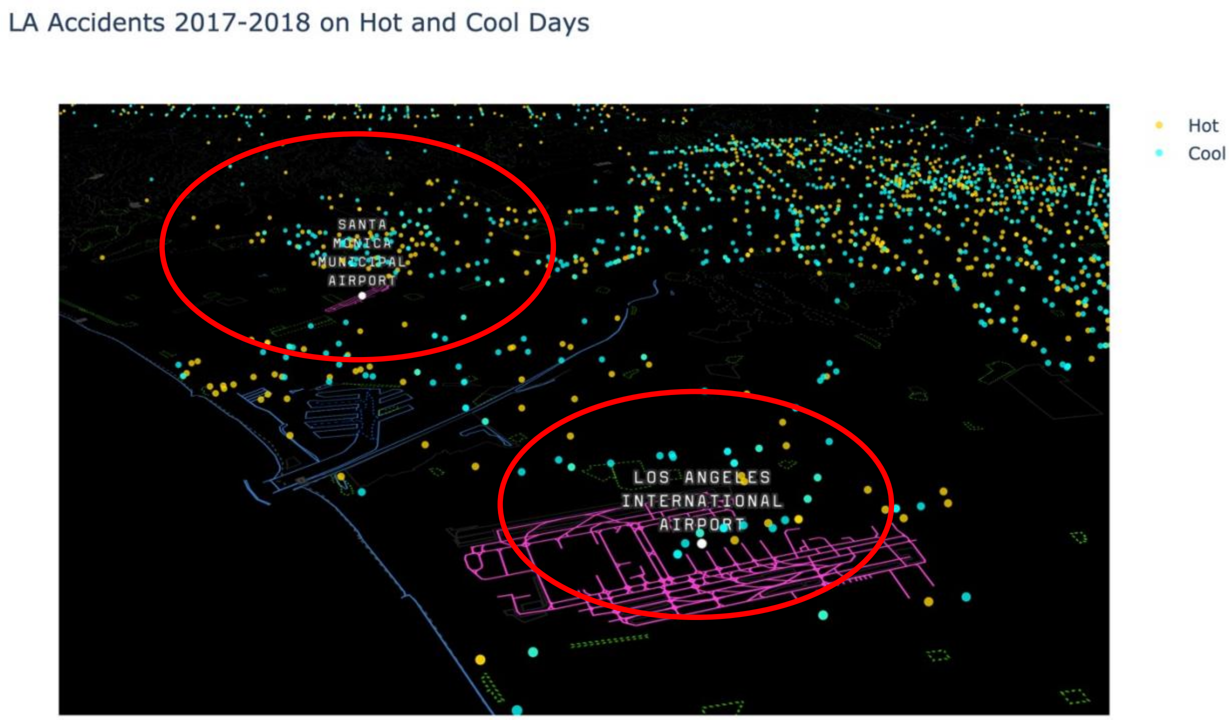


FIGURE 35 - MAP OF SANTA MONICA AND LA AIRPORTS WITH COLLISIONS ON HOT/COLD DAYS

This map of LA near the airports shows a higher concentration of accidents near the Santa Monica airport compared to the LA Airport on Hot (>100 deg. F) and Cool (<40 deg. F) days. It is also interesting to note that the Santa Monica airport is closing in 2028 and will be converted to a park. These maps of the area would be useful for city planners as they plan the area to understand how traffic collisions could be reduced in this area and if weather is a factor

Temperatures during the day of each collision were analyzed. The following tables show summaries of the Average Temperatures, daily high temperatures, and daily low temperatures.

Summary of top Average Temperature occurrences

63	4603
67	4347
73	4345
64	4058
60	3932
65	3881
72	3858

68	3641
61	3546

Summary of top High Temperature occurrences

79	3739
77	3710
75	3471
74	3369
88	3134
83	3099
86	2961
71	2836
80	2744
82	2690

Summary of top Low Temperature occurrences

53	4892
54	4873
57	4795
60	4650
56	3789
48	3641
63	3621
62	3568
58	3374
50	3290

There may be a correlation with cooler temperatures and accidents. No temperatures in the 90s or above was found in the top 10 frequently occurring temperatures. It appears that more moderate weather in the 60s to 80s has most of the accidents.

Precipitation

Since LA only averages about 15 inches of rain per year, precipitation is not as much of a factor as other large cities that are wetter. Our analysis showed different results for 2017-2018 with higher precipitation and accidents occurring around 5pm.

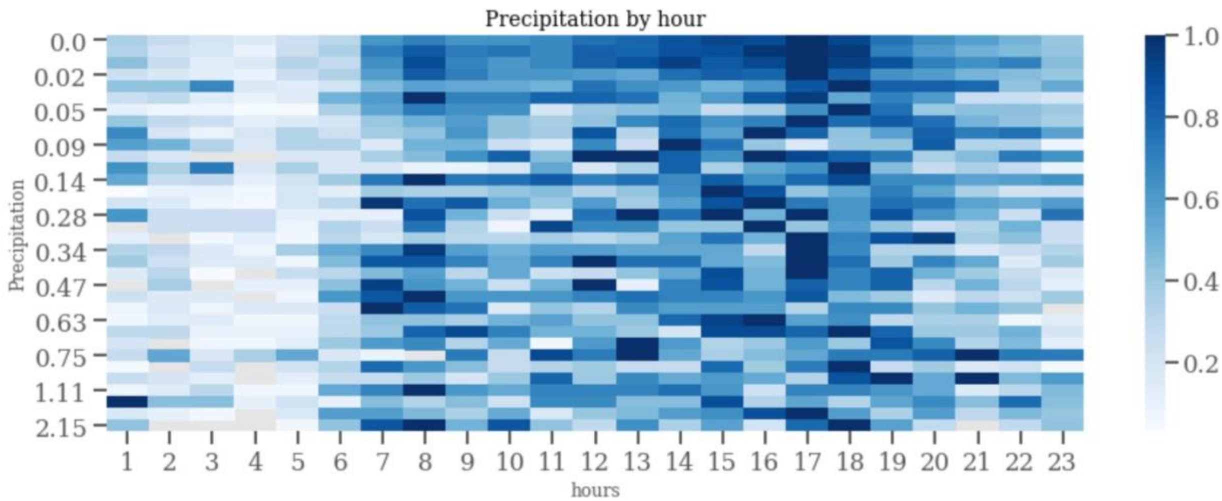


FIGURE 36 - 2017-2018 PRECIPITATION BY HOUR

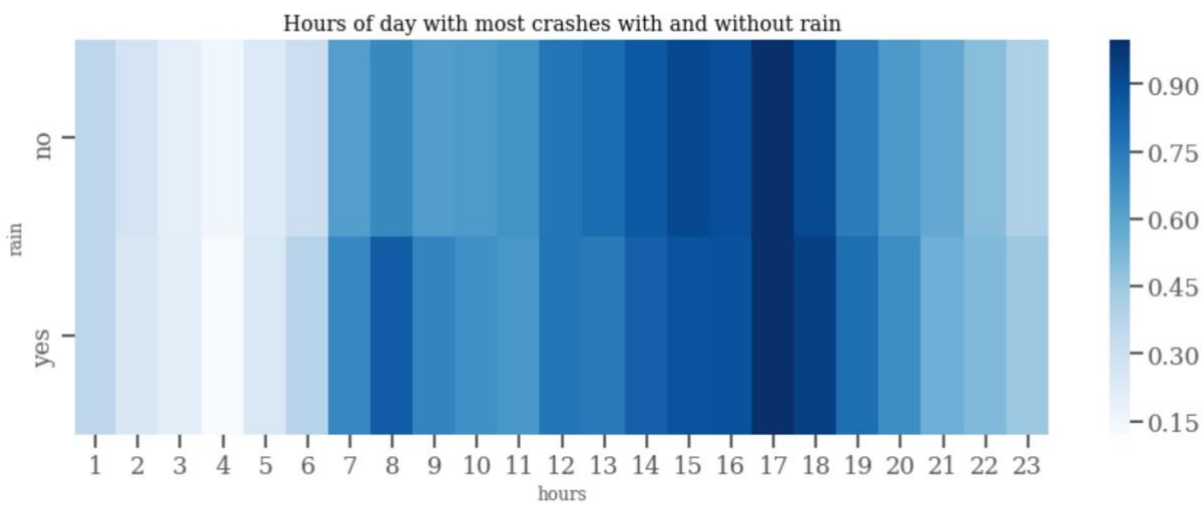


FIGURE 37 - 2017-2018 COLLISIONS BY HOUR WITH/WITHOUT PRECIPITATION

For this project, we also choose to explore the use of new 3-D maps to explore the traffic collision data. This allows for zooming/panning the data in a way where a subject matter expert (government official, city planner, citizen, etc) could observe variations in the data.

Although we did not have team members from LA, we did observe an area northeast of the Hawthorne Municipal airport where collisions appear to be more frequent (yellow dots show where a collision occurred with rain). A more detailed analysis of this area could reveal causes for these accidents at the boundary with the airport. Note: This report contains a static map of the precipitation. When running the code in Python, this map can be zoomed/panned to view the data interactively.

FIGURE 38 - MAP OF AREA NEAR HAWTHORNE MUNICIPAL AIRPORT WITH RAIN/NO RAIN COLLISIONS

Summaries of the top precipitation occurrences were analyzed for patterns.

Summary of top Precipitation occurrences

0.000000e+00	77563
1.000000e-16	4358
1.000000e-02	948
2.000000e-02	916
1.400000e-01	584
7.000000e-02	568
2.300000e-01	459
3.000000e-02	365
1.700000e-01	275
3.400000e-01	256

RESULTS AND FINDINGS: WEATHER

Los Angeles benefits from having great weather most of the time and weather is not a major factor as in other "wetter" cities. However, when it rains during evening rush hour traffic, there is an increase in the number of collisions.

CONCLUSIONS AND RECOMMENDATIONS

Los Angeles is a large and growing city which continues to attract more residents and vehicles. With the increase in the number of cars on the road comes additional traffic collisions and congestion. The analysis of the 2017-2018 data provides some interesting insights that can be used by a variety of people who are interested in LA traffic.

Below is a summary of our observations from this analysis:

1. What address/cross street combinations had the most collisions?
2. What are the most dangerous intersections?
    a. SEPULVEDA BL & SHERMAN WY
    b. NORDHOFF ST & TAMPA AV
    c. WHITSETT AV & SHERMAN WY
    d. RODEO RD & LA BREA AV
3. What are the most common collision areas in Los Angeles?
    a. 77th Street Area
    b. Council Districts 12/13/14
    c. Generally in the heart of LA
4. What are the best/worst times of the day for accidents? Best/worst month?
    a. Friday has the highest frequency of collisions.
    b. Sunday has the fewest amount of collisions.
    c. March and October have the highest number of collisions.
    d. The hours between 12PM to 5PM have the highest frequencies of collisions.
5. What patterns occur due to the amount of natural sunlight?
    a. There appears to be a significantly less concentration of accidents in Northern LA during the night time. This may be due to less traffic by the airports.
    b. The highest frequency of accidents is on Monday-Friday between 45pm.
6. What is the demographic makeup of victims in collisions?
    a. Men are more likely to be in an accident compared to women.
    b. Frequency of collisions is proportional to race/ethnicity.
    c. Age 30 has the highest number of collisions.
    d. 34.08% of accident victims have a median income between $40-49K.
7. Do certain temperatures or weather play a factor?
    a. When it rains during evening rush hour traffic, there is an increase in the number of collisions around 5pm.
    b. The area near the Hawthorne airport appears to have a higher proportion of weather-related accidents.

The results of this study provide additional insights that can be used by city planners, government officials and citizens to better understand the Los Angeles traffic conditions sin 2017 and 2018. We recommend the use of this information in the following areas:

1. Additional studies of dangerous intersections and locations.
2. Warnings on interactive road signs during rush hour when it is raining.
3. Utilize this data and interactive maps for planning sessions with subject matter experts and citizens when starting new projects which impact roadways.
4. Public service campaign to educate the public with targeted messages to men.

Limitations of Study

For this study, we were limited by the years used. Ideally, we would have liked to use all the years since 2010 but we were only able to scrape two complete years for weather, and therefore this was the reason we limited the scope to two years.

For income, we were limited to finding the income by Council Districts since we were not able to find incomes for every Zip Code in the dataset easily. This was easier since there were only 15 Council District incomes to get instead of the hundreds of Zip Codes. On the other hand, it is not as accurate an indicator of income since each Council District is spread out over a big area and not as precise as maybe a specific Zip Code.

The dataset also only included the *victims* of a car collision and not the perpetrator. It also did not indicate the severity of injuries of the victim. It would have been interesting to have more information about the types and severity of injuries since that would add another layer of depth to the analysis. It would have also been interesting to get information about speed of the vehicles in the collision and whether or not the cars were speeding.

The code to import Twitter data related to #latraffic was created but the data was not included in the analysis because the API only provides real-time information. Since historical tweets were not available, the limited number of tweets using hashtags related to #latraffic was low and not enough data was collected for analysis.

Contributions

Project Topic: Prasad Kulkarni
Team Meeting Organizer: Prasad Kulkarni, Sathish Kumar Rajendiran
Kaggle Data Import: Prasad Kulkarni
Kaggle Data Cleansing and Formatting: Prasad Kulkarni, Sathish Kumar Rajendiran
Creation of Data Dictionary: Prasad Kulkarni, Sathish Kumar Rajendiran
Weather Data Scrape: Prasad Kulkarni
Weather Data Import: Prasad Kulkarni
Twitter Data Import + API: Sathish Kumar Rajendiran
Financial Data Import: Prasad Kulkarni
Overall Analysis: Prasad Kulkarni, Sathish Kumar Rajendiran
Weather/Street/Time of Day Analysis: Prasad Kulkarni, Sathish Kumar Rajendiran
Map Visualizations: Prasad Kulkarni
Word Template Creation: Prasad Kulkarni, Sathish Kumar Rajendiran
Word Document Content and Editing: Prasad Kulkarni, Sathish Kumar Rajendiran
PowerPoint Template Creation: Prasad Kulkarni, Sathish Kumar Rajendiran
PowerPoint Content and Editing: Prasad Kulkarni, Sathish Kumar Rajendiran

REFERENCES

Data and supporting material was obtained from https://www.kaggle.com/cityofLA/los-angelestraffic-collision-data

Other sources included:

Unstructured Data
www.wunderground.com/

Color Themes https://seaborn.pydata.org/tutorial/color_palettes.html
https://medium.com/@andykashyap/top-5-tricks-to-make-plots-look-better-9f6e687c1e08
Income Data https://lachamber.com/clientuploads/pdf/2018/18_BeaconReport_LR.pdf

Twitter API & Mongo DB http://social-metrics.org/downloading-tweets-by-a-list-of-users-take3/
https://stats.seandolinar.com/collecting-twitter-data-storing-tweets-in-mongodb/
http://mrbool.com/tweepy-retrieving-and-storing-twitter-data-using-python-and-mongodb/36853
http://www.networkx.nl/data-science/twitter-data-python-store-mongodb/
https://pythondata.com/collecting-storing-tweets-with-python-and-mongodb/

Cover images: https://www.rockylawfirm.com/wp-content/uploads/2015/04/side-impact.jpg
https://ca-https://www.latimes.com/local/lanow/la-me-la-worst-traffic-20180206-story.html

Word cloud:
https://www.datacamp.com/community/tutorials/wordcloud-python