LA Traffic - Structured Data Analysis

- IST 652
- Authors: Prasad Kulkarni, Sathish Kumar Rajendiran
- Homework # 1
- Date: July 27,2020

Introduction

As the second largest city 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 such as the most dangerous intersections, best/worst times of the day for accidents and other useful information.

Analysis Questions and Sources

To guide our analysis we came up with a bunch of questions to answer in our analysis:

What are the most dangerous streets?

What are the most common collision areas in Los Angeles?

What are the best/worst days of the week for accidents? Month?

Sources

Traffic dataset

The Los Angeles Traffic Collision Data is publicly available from Kaggle.com is owned by the City of Los Angeles. The dataset contains 481,568 incidents from 2010 to 2019 https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data

Median Income dataset

This Dataset is mainly for future use during detailed project analysis. For Median Income, incomes were pulled from the into a CSV and merged into the original data
frame. https://lachamber.com/clientuploads/pdf/2018/18_BeaconReport_LR.pdf

Loading and Cleaning the Data

Data Cleaning

- 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.

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	

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 two categories along with some questions to help guide the process. Each group member worked on at least one section:

1) Location
2) Time/Day

ANALYSIS: LOCATION

Hypothesis:

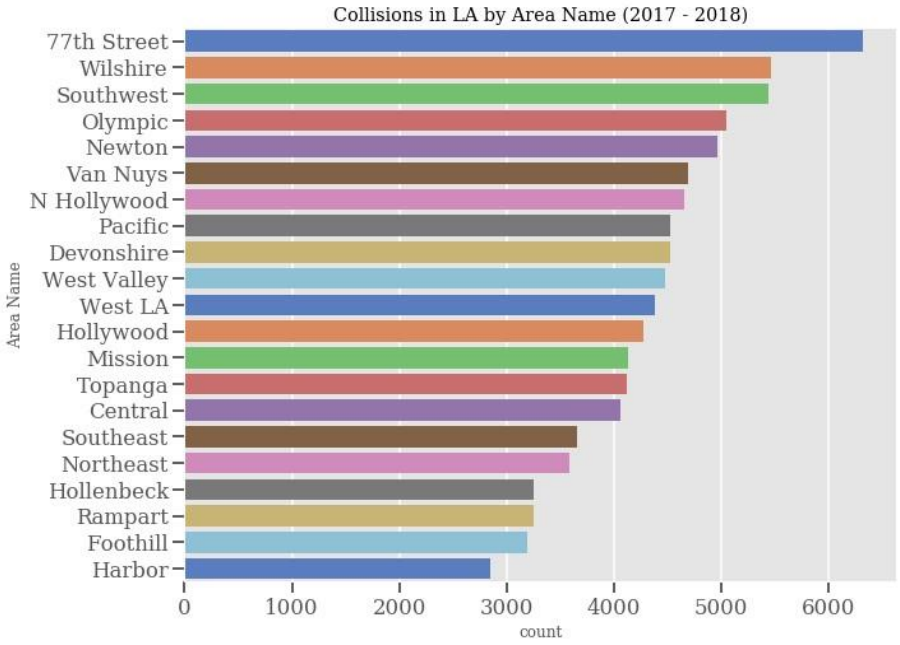
Certain areas of Los Angeles increase the likelihood of car collisions.

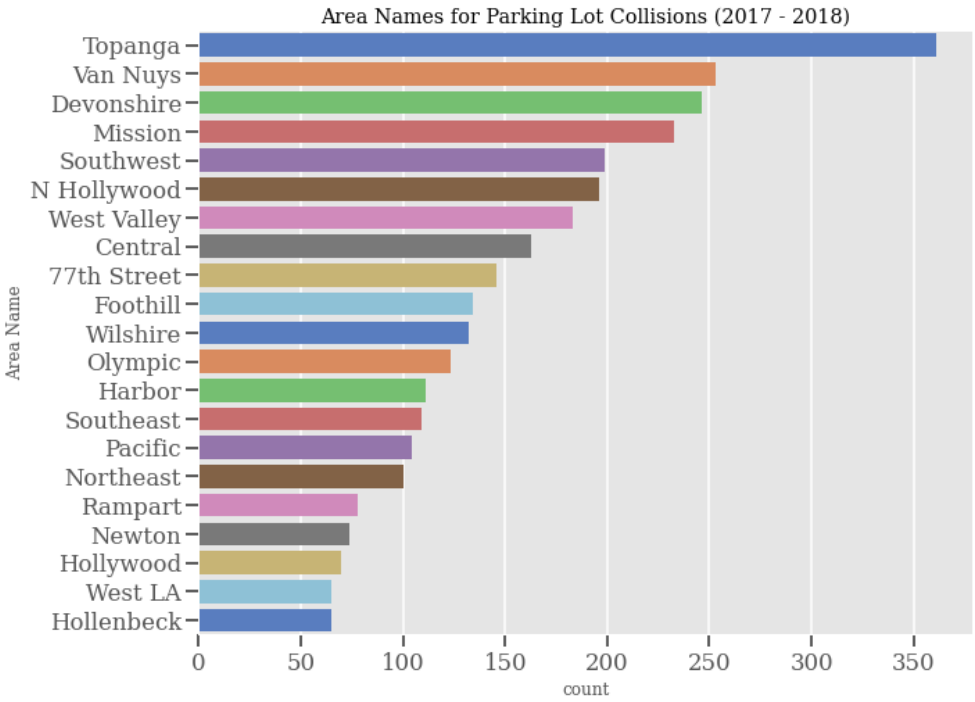
Fields:

Area Name, Zip Codes, Council Districts, Cross Streets,.

AREA NAME

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





COUNCIL DISTRICTS

Council Districts were also looked at:

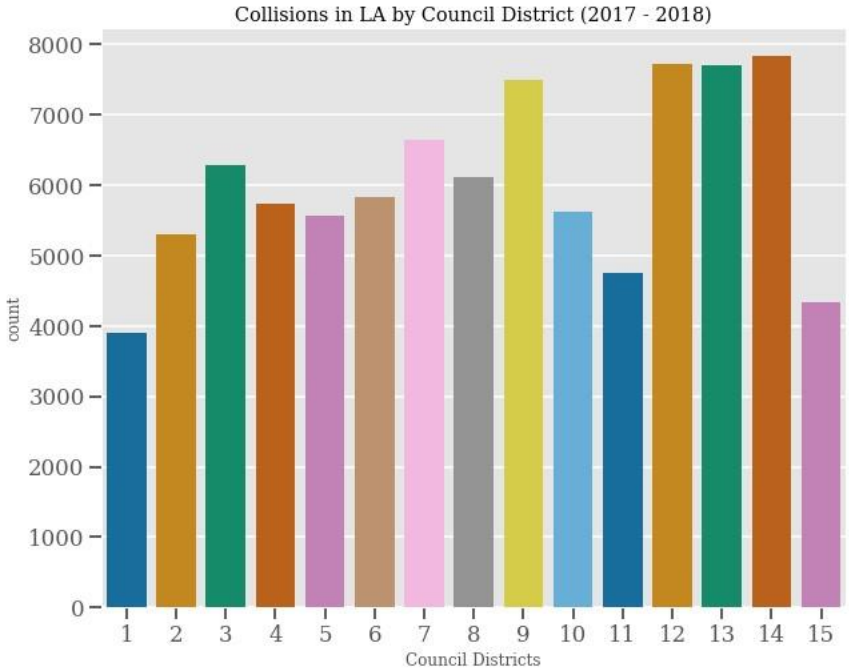
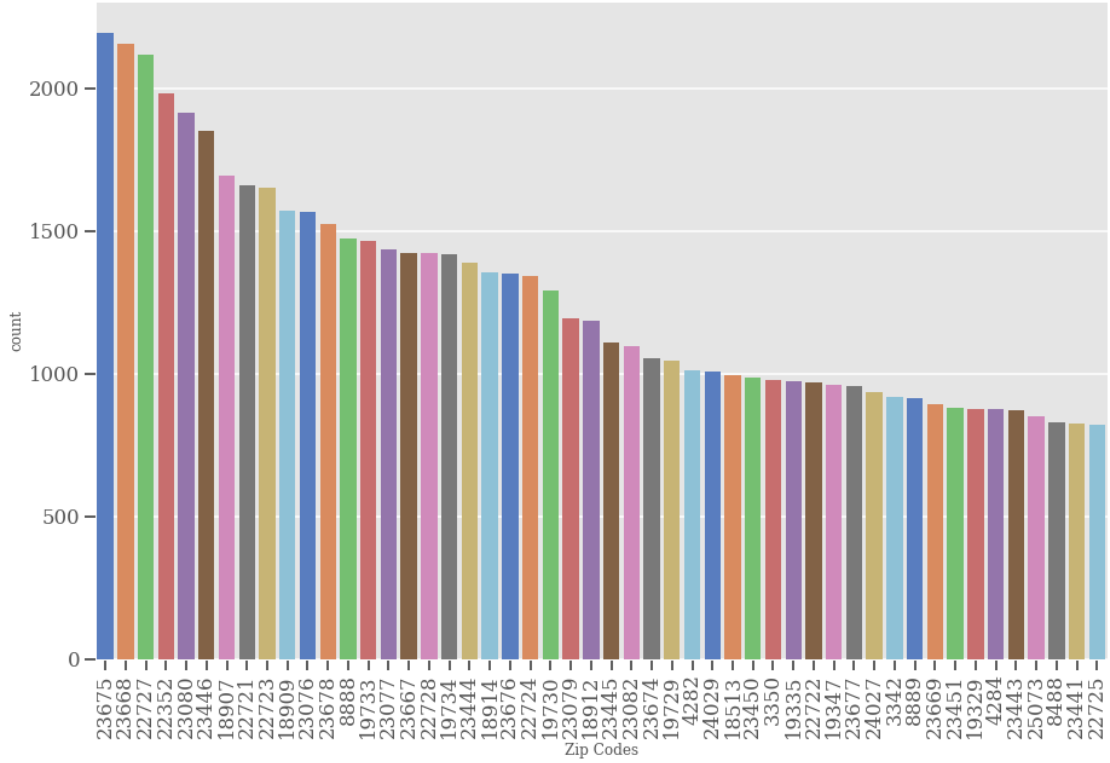


FIGURE 11 - LA COLLISIONS BY COUNCIL DISTRICT

Zip code

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



RESULTS AND FINDINGS: LOCATION

- Council Districts with the most collisions: 12/13/14.
- Streets / Areas with the most collisions: 77th Street, Wilshire, Southwest, Olypic
- Zip codes with most collisions: 23675,23668,22727,22352

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*,

DAY OF WEEK

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

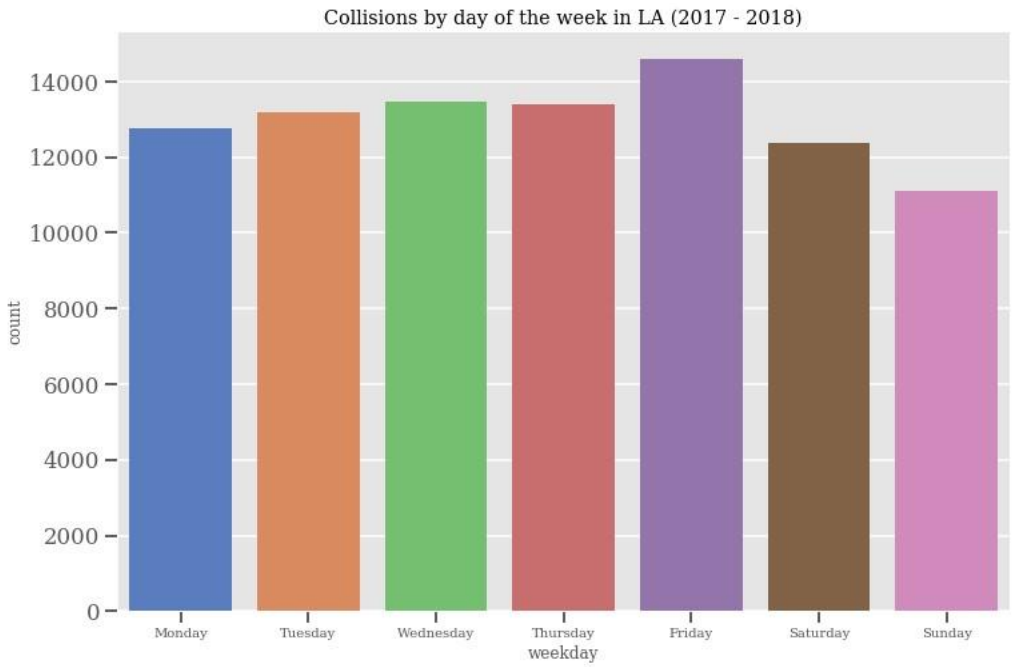
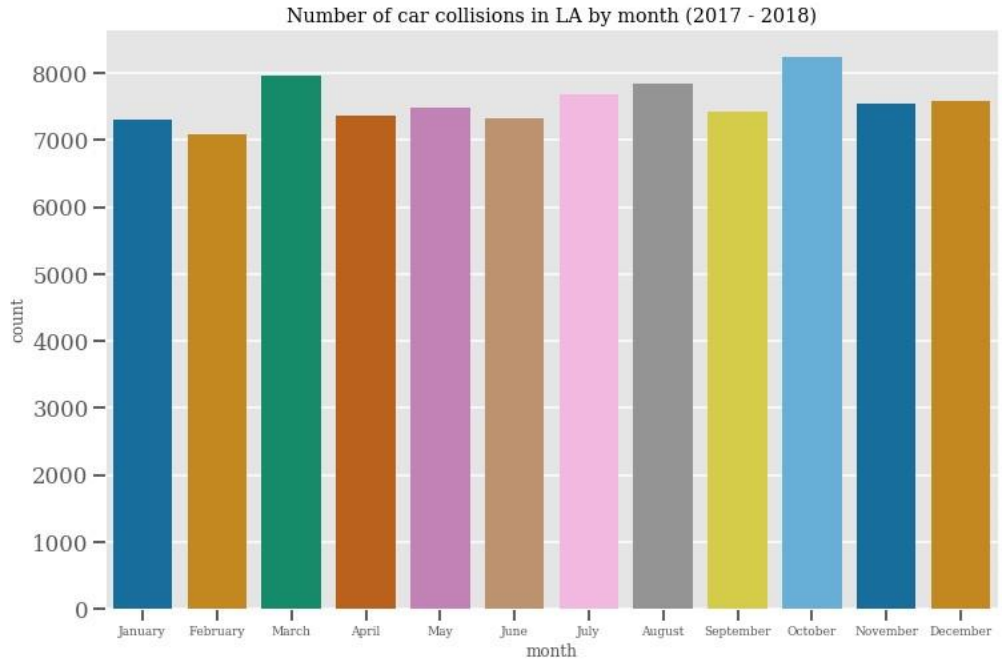


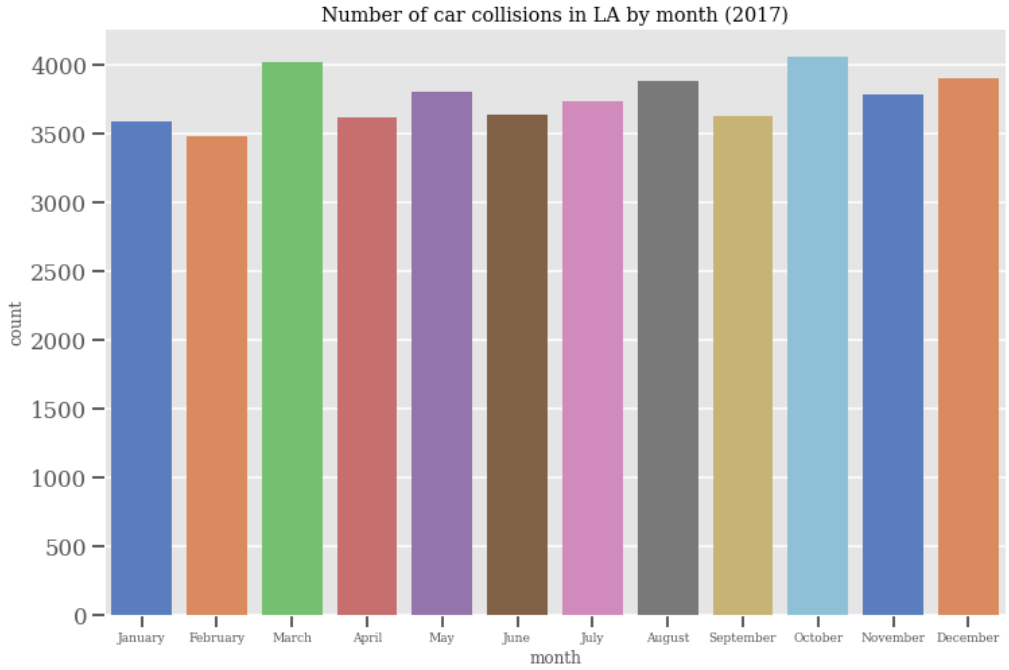
FIGURE 27 - COLLISIONS BY DAY OF THE WEEK 2017-2018

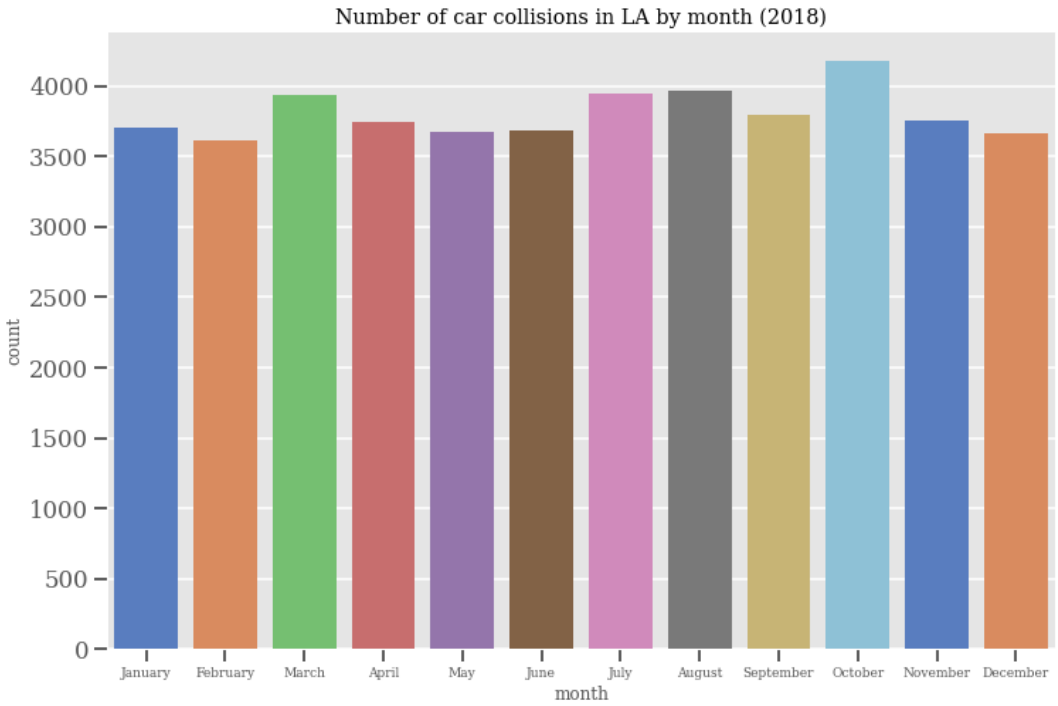
This shows the counts by day.

MONTH

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


Number of car collisions in LA by month (2017 - 2018)
count
8000
7000
6000
5000
4000
3000
2000
1000
0
January
February
March
April
May
June
July
August
September
October
November
December
month


Number of car collisions in LA by month (2017)
count
4000
3500
3000
2500
2000
1500
1000
500
0
January
February
March
April
May
June
July
August
September
October
November
December
month



We can see that October is the most common month as well as March and August.

RESULTS AND FINDINGS: TIME/DAY

- 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.
- We can see that October is the most common month as well as March and August.

CONTRIBUTIONS

Project Topic: Prasad Kulkarni
Kaggle Data Import: Prasad Kulkarni, Sathish Kumar Rajendiran
Kaggle Data Cleansing and Formatting: Prasad Kulkarni, Sathish Kumar Rajendiran
Location Analysis – Prasad Kulkarni
Day / Month Analysis - Sathish Kumar Rajendiran