**Isaac Carranza Project 2: ETL**

**Zoey Charif**

Our mission is to perform data engineering on police shootings. We searched the web and collected data from the following sources:

* https://github.com/washingtonpost/data-police-shootings
* https://www.kaggle.com/brendanhasz/fatal-police-shootings?select=CityLocations.csv

EXTRACT:

* imported both files to Jupyter Notebook
* converted files to separate Pandas dataframes.

TRANSFORM:

* To clean our data, we utilized the following functions: rsplit, join, merge, drop, dropna, rename, isnull.sum
* We saved our newly structured data into CSV and JSON files.

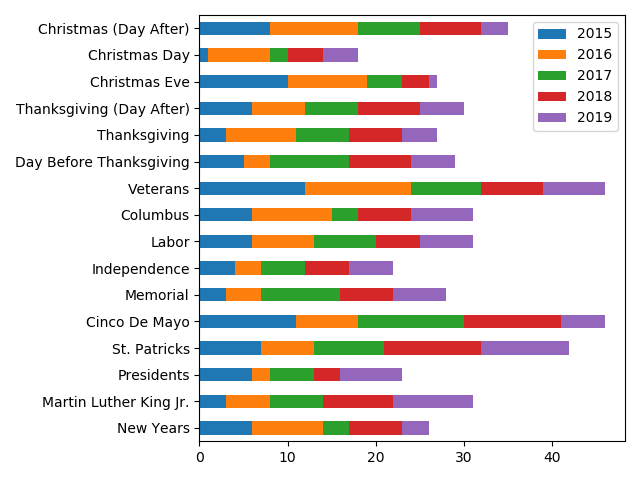
LOAD:

* We loaded our clean dataframe into Postgres SQL (relational db) and MongoDB (non-relational db)
* To load to Postgres, we first created a database inside PGAdmin called ‘police\_shootings’. We created a connection between our SQL database to Jupyter Notebook. We queried the table to verify all data was properly loaded
* To load to MongoDB, we used 2 processes. Firstly, we utilized MongoDB Compass to manually upload the new CSV file. Our second method was to use the insert\_many() function to import our dataframe directly from Pandas. Both functions adequately transferred the data to MongoDB.

Periods of Time proved difficult to code as the original dataset was not formatted to a transferable timestamp. The times of days were entered as whole numbers from 0 – 2400. The formulas to address this proved difficult to work with and time was short. More work is needed.

We successfully tested for the number of collisions that occurred on Federally recognized or/and popular holidays. We also included some holiday-adjacent days, where we expected more vehicle traffic to occur and, potentially, the collision rates to be higher.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **HOLIDAY** | **2015** | **2016** | **2017** | **2018** | **2019** |
| New Year's | 2015-01-01 | 2016-01-01 | 2017-01-02 | 2018-01-01 | 2019-01-01 |
| Martin Luther King Jr. | 2015-01-19 | 2016-01-18 | 2017-01-16 | 2018-01-15 | 2019-01-21 |
| President's | 2015-02-16 | 2016-02-15 | 2017-02-20 | 2018-02-19 | 2019-02-18 |
| St. Patrick’s\* | 2015-03-17 | 2016-03-17 | 2017-03-17 | 2018-03-17 | 2019-03-17 |
| Cinco De Mayo\* | 2015-05-05 | 2016-05-05 | 2017-05-05 | 2018-05-05 | 2019-05-05 |
| Memorial | 2015-05-25 | 2016-05-30 | 2017-05-09 | 2018-05-28 | 2019-05-27 |
| Independence | 2015-07-03 | 2016-07-04 | 2017-07-04 | 2018-07-04 | 2016-07-04 |
| Labor Day | 2015-09-07 | 2016-09-05 | 2017-09-04 | 2018-09-03 | 2019-09-02 |
| Columbus | 2015-10-12 | 2016-10-10 | 2017-10-09 | 2018-10-08 | 2019-10-14 |
| Veterans | 2015-11-11 | 2016-11-11 | 2017-11-10 | 2018-11-12 | 2019-11-11 |
| Thanksgiving | 2015-11-26 | 2016-11-24 | 2017-11-23 | 2018-11-22 | 2019-11-28 |
| Thanksgiving Day After\* | 2015-11-27 | 2016-11-25 | 2017-11-24 | 2018-11-23 | 2019-11-29 |
| Christmas Eve\* | 2015-12-24 | 2016-12-24 | 2017-12-24 | 2018-12-24 | 2019-12-24 |
| Christmas | 2015-12-25 | 2016-12-26 | 2017-12-25 | 2018-12-25 | 2019-12-25 |
| Christmas Day After\* | 2015-12-26 | 2016-12-27 | 2017-12-26 | 2018-12-26 | 2019-12-26 |
| \* not a Federally recognized holiday | | | | | |

Among the selected holidays and holiday-adjacent days, Veteran’s Day, Cinco de Mayo and St. Patrick’s Day experienced the most traffic collisions between 2015 – 2019. Although one could immediately hypothesize the potential factors surrounding these holidays, these elements are not included in this data set and further analysis is required.