
Majd Abdo - Ghareeb Falazi - Yuliya Volga

BTW 2017 Data Science Challenge SDSC17
Goals

• Find interesting patterns and behaviors and visualize them.

• Detect *when* and *where* the number of accidents and casualties was the highest.

• Detect *when* and *where* the number of accidents showed big changes.

• Show the effect of *weather* and *traffic* on the accidents.

• Suggest *recommendations* for decreasing the number of accidents.
Major Technologies Used

Cloud Based:
- Cloud Foundry Python App
- Data Connect
- Compose for Postgresql

Python:
- Bokeh
- Flask Serve

Javascript:
- Google Maps
- LeafletJS
Data Sources

1. NYPD Motor Vehicle Collisions:  
   1 CSV files (~170MB) representing ~1M accidents (07/2012 - 02/2017)

2. Yellow Taxi Trip Data:  
   12 CSV files (~40GB) representing ~150 M taxi trips (2015)

3. Daily weather statistics of central park ground station in NY.  
   (source: www.wunderground.com)
Data Sources Integration

Visual Analytics Application

Materialized Views
Hold the results of static frequent queries

Geo-Indexing:
PostGIS column with index

Join:
With accidents source by date

Aggregation:
# of trips by date and hour-of-day

Integrated Source

Materialized Views

ETL

Weather

Accidents

Traffic

ETL

ETL

MV

queries
Interaction model - Initial Request

1. HTTP request to some route: (e.g. /vis3)

2. Database

3. Flask

4. Bokeh visualization embedded in an HTML page

User Browser
Interaction model - Subsequent Requests

1. User applies new filters,..
2. Javascript callback triggered
3. Ajax request to some route: (e.g. /histogram )
4. Data received, and Bokeh visualization populated
5. JSON file returned with the new data
6. User Browser
DEMO
Change Rate Formula

For each ZIP code, we calculate the change rate as the following:

\[
\frac{\sum^n_i |x_i - x_{i-1}|}{\sum^n_i x_i} \cdot 50
\]

Where:

\( n \): number of years/months with accidents
\( x_i \): number of accidents at the current (\( i^{th} \)) year/month
\( x_{i-1} \): number of accidents at previous (\( i-1^{th} \)) year/month
\( 50 \): to make the range between 0% and 100%
New York City on the 25th of February 2015

“Daily Telegraph” Picture of the Day
Summary of Results (1)

- Brooklyn, Manhattan and Queens are the boroughs with most accidents.
- No big difference in accidents count between the four seasons, but a slight increase during Fall.
- We can consider Manhattan Midtown and Jackson Heights as the most frequent hotspots of accidents in NYC. On the other side, Staten Island is considered as the safest place to drive.
- **W 42nd street in Manhattan** has a big number of accidents in general, and especially caused by buses (> 50) and taxis (> 80).
- **Linden Boulevard in Brooklyn** has the highest # of casualties in the whole city (183)
- **Flatbush external to Manhattan bridge** has the highest # of accidents ever 678!
- Manhattan and Brooklyn have recently a considerable decrease in accidents. However, Staten Island has the most decreasing rate of accidents in last three years.
Summary of Results (2)

- All the boroughs share the same behaviour of accidents during hours of day with a local peak at 8:00, a smooth increasing until 16:00 and then a tremendous decrease of accidents.
- Safest period to drive a vehicle is between 20:00 until 7:00 (next day).
- The rush hour for Bicycle accidents is between 16:00 until 20:00
- Bicycle accidents show strong seasonality.
- A significant increase in Bicycle accidents recently.
- Accidents affected by Glare usually take place at 8 (sunrise) and at 16 (sunset). And especially during winter. Drivers should wear sunglasses during these times.
- Number of injured persons increased recently after a small drop during 2015.
- Number of killed persons is stable during last four years with 0~2 casualties daily. Most critical hour is at 17:00.
- Speeding accidents has been going crazily up since Sept 2016. An action is needed by the city to control this issue again.
Summary of Results (3)

• The temperature of -5 °C and the visibility of 3 km are associated with the highest number of accidents.

• Extreme weather values have little number of accidents, whereas a little less extreme values have the highest rate for accidents. People should be extra cautious while certain weather conditions are present.

• Extra caution can tremendously decrease the number of accidents. Media should encourage people to take caution at certain cases.

• The reason “Pavement Slippery” appears to be one of the most popular reasons (among infrastructural, non-person behavior) for causing accidents. Suggestion: to foresee situations with cold temperatures and provide work on roads to prevent accidents (put chemicals on the road for snow and ice melting, for instance).
Summary of Results (4)

• There exist areas with high number of accidents and low change rate of accidents - which is critical, because it means that the number of accidents is always high and not decreasing. Such areas are mostly in Brooklyn.

• Three ZIPs of Staten Island borough have huge decrease of number of accidents - investigate why those happened. Provide the same politics for other boroughs to decrease number of accidents.

• The data quality dropped significantly starting from April 26\textsuperscript{th} 2016 and was partially recovered starting from September 8\textsuperscript{th} 2017 this should be closely investigated especially that the source of the data is NYPD.