Estimation of Fatal Crashes

Lucas Anderton, Hannah Brown, Lucas Gorak, Edward Mikkelson

1/8/2020

Traffic in DC

Washington Post, 2011: "D.C. area is No. 1 nationwide in traffic congestion, study says"

Washington Post, 2019: "Washington is No. 3 in traffic congestion, study says"

Research Questions

Is there a relationship between. . .

1. 311 traffic requests and the frequency of crashes in D.C.?
2. Rising latency in resolution times and the frequency of crashes?

Do crashes occur more in geographic areas with more 311 requests?

What factors are most significant in estimating the likelihood of a fatality in a crash?

Data Collection Procedure

- Categorical, descriptive crash data had analytical shortcomings.
- OpenDataDC's `crashes` dataset was thorough and contained mostly boolean values (True/False; Yes/No).
- Explored related `crashdetails` which is a relational table.
  - Utilizing that for additional data on a subset of the DC crashes.
- Summarized 311 Requests `threeoneonerequestts` and `crashes` by count per day.
  - Eventually merged these into a single tibble, grouped by date and ward.

```
Details <- read.csv("https://opendata.arcgis.com/datasets/7
crashes <- read_csv("https://opendata.arcgis.com/datasets/7
threeoneone <- read_csv("https://datagate.dc.gov/search/ope
```

Initial Processing

- Filtering out unnecessary variables(columns) in all three frames
- Lining up start and end dates of data

```
crashes <- crashes %>% filter(as.Date(FROMDATE) >=
"2012-01-01")
```

- Filtering 311 requests down to traffic-related requests

```
trafficrequests <- filter(threeoneone,
SERVICECODEDESCRIPTION == c("roadway
signs","streetlight repair investigation","pothole"))
```

Crash Data Tidbits

- Crashes has 223990 observations and 60 columns
- Details has 443867 observations and 15 variables
- Ages beginning at -7990 and ranged up to 237
- License plate `None` produced a high number of crashes in the data set
- The full dataset reaches back until 1975

Wrangling Into Counts

Creates new variable, INDEX, and assigns every row 1 for counting purposes.

```
crashes <- crashes %>%
  mutate(INDEX = 1)
View(crashes)
```

Frequency tibble for crashes on each day

```
crashes_by_date <- crashes %>% group_by(FROMDATE,WARD) %>%
  summarise(dailycrashes = sum(INDEX))
```

Frequency tibble for 311 requests on each day

```
requests_by_date <- trafficrequests %>% group_by(ADDDATE,WA
  summarise(dailyrequests = sum(INDEX)) <-- using those inc
```

Linear Regression on Frequency

Modeling the relationship between frequency of 311 requests and crashes

Methods

- Recoding data from categorical to binary for use in regression models
- Ran a series of exploratory models with `ggplot`
- Mapped geospatial data with `ggmap` and Google API
- Used clustering algorithm on geospatial data to find groupings

Analysis

-Regression model for 311 requests and traffic incidents -Tested model on training data and found consistent results

Future Implications

- Opportunity to analyze particularly dangerous intersections
- Increase usage of 311 complaints to address traffic and transportation concerns
- Importance of responding to 311 requests
- Increase pedestrian safety infrastructure in crash-dense areas
- No snowmobiles?