Final Report

Group 6

Lucas Anderton, Hannah Brown, Lucas Gorak, Edward Mikkelson

1/10/2020

1 Research Question

⁀In our analysis, we were interested in understanding the relationship, if any, between various factors, like population increase and road conditions, and the crashes that occurred in the District of Columbia. Previous research has indicated that D.C. has the third worst traffic congestion in the United States. However, their ranking has improved since 2011, which D.C. had the worst congestion of all 50 states. In that same period of time, the number of crashes in D.C. has increased significantly. Intuitively, we credited the increase to more traffic from the constantly growing population in the Washington metro area. D.C.'s public datasets on traffic are convoluted and difficult to manipulate for analysis. It is recorded in 15,000+ rows from various traffic-volume meters throughout the city on an annual basis. However, every year, the number of meters and other variables in the data change. Alternately, we chose to compare the number of crashes to the 311 traffic service requests. 311 traffic service requests offer information regarding the number of commuters out on the roads and the state of the city's vehicle infrastructure. In our research, we wanted to understand what the relationship between these 311 traffic requests and the frequency of crashes in D.C. Additionally, these data include timestamp information, like the date they were submitted, the date they were due (most likely based on a city algorithm for how long certain requests should take), and the date they were resolved. We wanted to see if a relationship existed between rising latency in resolution times and the frequency of crashes. The reasoning behind our hypothesis was that if the city was taking longer to respond to poor road and navigation conditions, more incidents may occur. We also wanted to investigate if there was a geographical relationship between crashes and 311 requests. And lastly, we were interested in analyzing what factors best estimate risk of faility in a crash.

2 Data Collection Procedure

In our preliminary research, we found three comprehensive datasets from D.C.'s open data site, each offering salient variables pertaining to vehicle crashes in D.C. over time. The first set we explored was mostly categorical and qualitative, which left much to be desired in terms of quantitative analysis. The remaining sets include a combined 63 variables, both qualitative and quantitative in nature.

Next, we happened upon the city's 311 service request data portal, on the same OpenData library. The portal offered custom data downloaded, allowing the user to select a date range or data for one specific type of request. The full data set included more than 1.5 million rows with approximately ten types of requests. We were able to import our data directly from D.C.'s OpenData portal.

```
Details <- read.csv("https://opendata.arcgis.com/datasets/70248b73c20f46b0a5ee895fc91d6222_25.csv")
crashes <- read_csv("https://opendata.arcgis.com/datasets/70392a096a8e431381f1f692aaa06afd_24.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   CCN = col_character(),
##   REPORTDATE = col_datetime(format = ""),
##   ROUTEID = col_character(),
##   FROMDATE = col_datetime(format = ""),
##   TODATE = col_logical(),
##   ADDRESS = col_character(),
##   WARD = col_character(),
##   EVENTID = col_character(),
##   MAR_ADDRESS = col_character(),
##   NEARESTINTROUTEID = col_character(),
##   NEARESTINTSTREETNAME = col_character(),
##   INTAPPROACHDIRECTION = col_character(),
##   LOCATIONERROR = col_character(),
##   LASTUPDATEDATE = col_datetime(format = ""),
##   BLOCKKEY = col_character(),
##   SUBBLOCKKEY = col_character(),
##   FATALPASSENGER = col_logical(),
##   MAJORINJURIESPASSENGER = col_logical(),
##   MINORINJURIESPASSENGER = col_logical(),
##   UNKNOWNINJURIESPASSENGER = col_logical()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 11 parsing failures.
##  row                    col           expected actual
## 1061 MINORINJURIESPASSENGER 1/0/T/F/TRUE/FALSE      3 'https://opendata.arcgis.com/datasets/70392a09
## 1098 MINORINJURIESPASSENGER 1/0/T/F/TRUE/FALSE      2 'https://opendata.arcgis.com/datasets/70392a09
## 1106 MINORINJURIESPASSENGER 1/0/T/F/TRUE/FALSE      4 'https://opendata.arcgis.com/datasets/70392a09
## 1201 MINORINJURIESPASSENGER 1/0/T/F/TRUE/FALSE      2 'https://opendata.arcgis.com/datasets/70392a09
## 1215 MAJORINJURIESPASSENGER 1/0/T/F/TRUE/FALSE      2 'https://opendata.arcgis.com/datasets/70392a09
## .... ...................... .................. ...... .............................................
## See problems(...) for more details.
```

```
threeoneone <- read_csv("https://datagate.dc.gov/search/open/311requests?daterange=8years&details=true&
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   XCOORD = col_double(),
##   LONGITUDE = col_double(),
##   RESOLUTIONDATE = col_datetime(format = ""),
##   INSPECTIONDATE = col_datetime(format = ""),
##   SERVICEDUEDATE = col_datetime(format = ""),
##   YEAR = col_double(),
##   SERVICECALLCOUNT = col_double(),
##   MARADDRESSREPOSITORYID = col_double(),
##   ZIPCODE = col_double(),
##   YCOORD = col_double(),
##   ADDDATE = col_datetime(format = ""),
##   SERVICEORDERDATE = col_datetime(format = ""),
##   LATITUDE = col_double()
## )
## See spec(...) for full column specifications.
```

3 Important/Interesting Facets of Data Processing

4 Statistical Methods

5 Analysis of Results

6 Implications