# *Spatial Distribution and Analysis of Cyclist Crashes in Washington, DC*

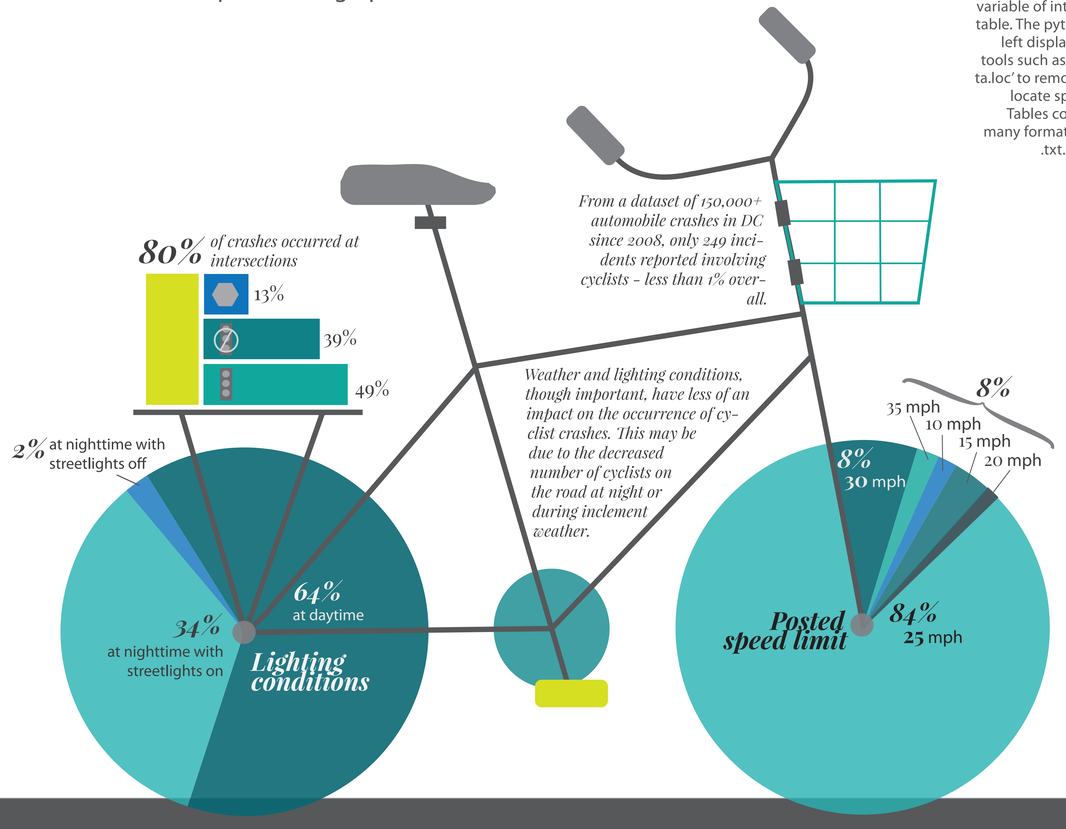
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Kelsey Taylor
Master's Candidate, Department of Geography
ktaylor_@gwu.edu

## *Abstract*

As part of its Open Data initiative, the government of the District of Columbia offers hundreds of geospatial datasets for public use in their online database. While many of the datasets are well curated, consistency in open-source data is not guaranteed. For this study, a 150,000+ feature dataset of vehicle crashes in Washington, DC was pre-processed and cleaned in Python using the pandas library and ArcPy toolset. The resulting shapefile contained 250 crash incidents involving cyclists.

Each record includes attributes for its corresponding crash that were used for analysis of the data, including lighting and weather conditions, intersection type, and speed limit at the crash site. For each factor analyzed, the standard distance ellipse was calculated and mapped. Additionally, the point data was aggregated to the neighborhood level and "problem" areas were identified through optimized hot-spot analysis. Spatial and attribute data was visualized with a series of maps and infographics.




## *Methodology*

The dataset used in this analysis, though extensive, was not consistently kept. Several attributes were left blank for each point, in addition to many free-form response attributes within the data. Python scripting using the pandas library was used to clean the data for processing and analyses. With the number of records exceeding 150,000, .csv and .txt files allowed for quicker data processing than reading .csv files into GIS software or converting to .shp or other spatial data formats.

Once cleaned, the data were processed using the ArcPy toolset in the ArcMap Python console. For the primary point dataset, standard tools such as TableToTableConversion and MakeXYEventLayer allowed for visualization of the crash incidents. For further analysis, point data were aggregated to the neighborhood level using a spatial join. Once aggregated, an Optimized Hotspot Analysis illustrated the concentration of crashes in each neighborhood within Washington, DC, as outlined in a polygon shapefile from the same Open Data initative.

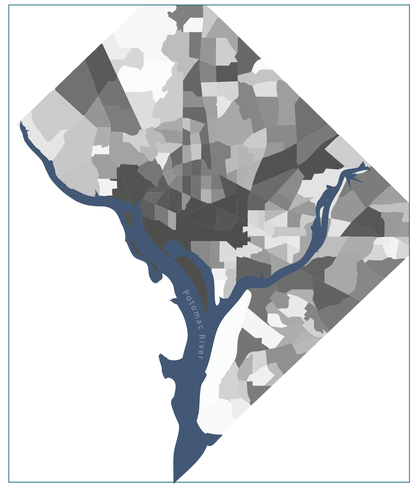
For non-spatial analysis and visualization, the pandas library allowed for quick processing of the large point dataset. Of particular use were pivot tables, which provided summary statistics for each variable of interest in a simple table. The python script to the left displays use of pandas tools such as 'dropna' and 'data.loc' to remove null data and locate specific attributes. Tables could be output in many formats, including .csv, .txt., .xml, and so on.

```
# using the pandas library and modules to read and write to .csv files
import pandas as pd

data = pd.read_csv(path + r"\Cyclist_crashes.csv")
data.head()
data.columns
# drop all records that don't have coordinates associated with them for the
# data set we are mapping since they won't show up anyway
data.dropna(subset=['Field2'])
data.fillna(value=' ',inplace=True)
data.replace('Unknown', '-9999', inplace=True)

# script doesn't work if "Null" values aren't rewritten as strings
data.loc[data['Field5'].str.contains('null'),'INTERSECTION'] = "-9999"
# converting any partial strings with "Not" to 0
data.loc[data['Field5'].str.contains('Not'),'INTERSECTION'] = "0"
# converting any remaining rows containing the word intersection to 1
data.loc[data['Field5'].str.contains('ntersection'),'INTERSECTION'] = "1"
data.loc[data['Field5'].str.contains('P'),'INTERSECTION'] = "0"
print "Intersections converted"

# streetlights on = 1, off = 0, else null
data.loc[data['Field6'].str.contains('On'),'STREETLIGH'] = "1"
data.loc[data['Field6'].str.contains('Off'),'STREETLIGH'] = "0"
data.loc[data['Field6'].str.contains('Other'),'STREETLIGH'] = "-9999"
```

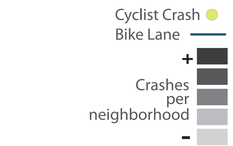
All data from DCGIS.gov
NAD 1983 State Plane Maryland FIPS 1900
Lambert Conformal Conic Projection

### *All Reported Crashes*

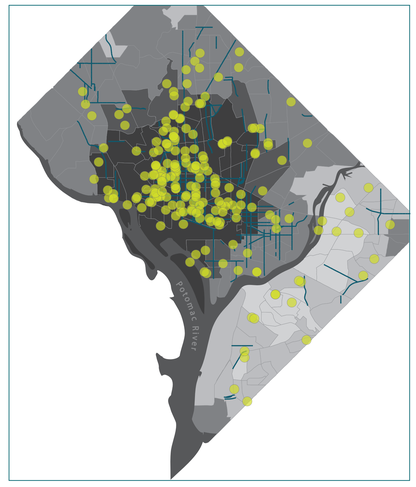

As expected, the vast majority of crashes recorded within the District of Columbia occur in the central business district and near the National Mall. However, there are other pockets of high crash counts per neighborhood in areas along the city's diagonal thoroughfares, which take commuters and tourists in and out of the downtown area.




### *Vehicle Crashes Involving Cyclists*




With a considerable portion of the population commuting to the central business district (darkest gray), it comes as little surprise that this area has the highest concentration of cyclist crashes in the District of Columbia. There does appear to be a positive correlation between crash incidents and bike lanes, with more bike lanes present in the central business district.




## *Discussion & Conclusion*

While most results of interest are featured in the infographic (left), three primary outcomes summarize the larger trends:

- vehicle crashes involving cyclists were concentrated in the downtown/central business district area
- speed limit appears to have a correlation with cyclist crash incidents in the study area, with 25mph zoned areas dominating the dataset
- the majority of cyclist crashes occur at intersections

A number of factors not considered in this analysis may have contributed to the outcomes of this study. For example, the count of crashes in 25mph areas may be skewed by the high proportion of these areas within the District of Columbia. As a high-commuter city, the percentage of crashes involving cyclists at nighttime may be lower than expected due to decreased vehicle and cyclist traffic at night. Another attribute, weather conditions, yielded a similar result, likely due to the decreased number of cyclists on the road during inclement weather events. Ideally, a more complete and systematically curated dataset would yield a more accurate reflection of each attribute's impact on the count and distribution of cyclist crashes.