D.C. Car Crash Analysis

CSE 6242 Progress Report

1. Introduction and Motivation

Every year, car accidents are among the top causes of death in Washington, D.C. In response, the city has launched its Vision Zero Initiative, an effort to reduce vehicle-related crashes to zero by 2024. The team will determine the root causes of car accidents in Washington, D.C. by analyzing patterns, seasonality, and trends in the data collected from the Vision Zero website and other sources. **Our goal is to create an interactive and robust visualization dashboard to communicate the true impact of traffic-related crashes in Washington, D.C. to stakeholders who would be most interested in D.C.'s Vision Zero plan, such as policymakers, police, and local residents.**

1.1 Problem Definition

The current Vision Zero website is sufficient for its stated purpose, but there is substantial opportunity to improve upon it. The dataset provided is large, robust, and ripe for application of analytical techniques, especially geographical clustering, network analysis, and probability-based risk analysis. Results can be incorporated into an improved dashboard to provide further insights to DC citizens and key stakeholders.

1.2 Survey

On top of D.C.'s Vision Zero plan, many traffic safety researchers have approached the issue of reducing vehicle-related crashes using analytical approaches. Researchers have identified common Vision Zero measures and analyzed their effectiveness across countries and cultures (Kim, et al. 2017). Crashes in Washington D.C. were predicted using two separate models: ARIMA and Heston. ARIMA assumes that any volatility in the data is constant, while Heston assumes that the volatility is arbitrary. The study shows that Heston has better accuracy than ARIMA (Shannon & Fountas, 2022). Crashes were evaluated based on different collision types, such as rear-end crash, sideswipe crash, and angle crash. Bayesian analytics was used to perform the evaluation. The result shows that each collision type has different rates and risk factors (Guo et. al, 2019). Different factors affect pedestrian crashes in Texas' county-level areas using OLS Regression. The result suggests that homelessness, median household income, and poverty positively correlate with pedestrian crashes (Bernhardt & Kockelman, 2021). The results and assumptions used in the said studies will be leveraged in our analysis and model. The major shortcoming of the studies is that they do not have an interactive dashboard that can better communicate their results to non-data stakeholders.

One component of our crash-related analysis is on bicycle safety. Some of the most common forms of vehicle-related crashes are between cars and bicycles (Daraei et. al, 2021). Many studies often cite the presence of cycling lanes as a causal link to reducing crashes for cyclists; however, some of these studies only look at the absolute presence of a cycling lane (i.e., does one exist or not) as a factor for reducing crashes (Marshall et. al, 2019). In actuality, different types of cycling lanes exist and have different effects on road safety. Our research will look at these different types of road-calming and cycling infrastructure measures, trying to enhance some of the work done in papers that only use observational methods to analyze the impact on vehicle-bicycle crashes between different cycling infrastructures (Jensen, 2007).

Geographic analysis is another crucial component of analyses of pedestrian safety. Numerous studies have incorporated spatial analysis into crash analyses in order to understand how factors specific to certain geographic units influence crash frequency. Researchers in North Carolina (Pulugurtha et. al, 2010) and China (Wang et. al, 2016) analyzed the factors specific to signalized intersections and traffic analysis zones, respectively, which predicted crashes in the relevant zones. Another set of researchers in Florida (Lee et. al, 2017) went even farther, comparing the performance of crash prediction models at different levels of geographic aggregation.

Finally, an interactive and robust visualization communicates our findings to stakeholders in D.C. Researchers have superimposed photo enforcement citation data on crash data to illustrate the effectiveness of automated enforcement (Rogers et. al, 2016). Additionally, other researchers have also used transportation, mobile, and demographic data to determine safety risks in the state of Maryland (Xiong et. al, 2021). Potential shortcomings of these visualizations include their limited scopes.

2. Proposed Method

To approach this problem, the team will:

1. Pull the data using the OpenDataDC API.
2. Join in related datasets to enrich analysis, detailed further below.
3. Restructure the data for visualization purposes.
4. Conduct further analysis and clustering in Python, detailed further below.
5. Visualize results in an interactive Tableau dashboard.

2.1 The Data

The team will pull the 270,000+ records of car crashes in D.C. using the OpenDataDC API.[1] However, this method has provided some early challenges to the group. For example, the API limits pulls to just 1000 records at a time. To circumvent this, the team pulled records in small batches until all records were collected.[2] Although this took significant time (each pull took 2-3 seconds and the script pauses for 1 second in between each API call), it did provide us an avenue to get the most up-to-date crash data.

Dataset	Description
Vision Zero Data	The official website, Vision Zero. The site contains the data related to Vision Zero.
Vision Zero Safety	This dataset supports the Vision Zero Initiative and comes from a web-based application developed to allow the public to communicate the real and perceived dangers along the roadway from the perspective of either a pedestrian, bicyclist or motorists.
Crashes in D.C.	This dataset represents the crash locations associated along the District of Columbia roadway blocks network. A companion crash details related table also exists for download.
Crash Table Data	This table is a companion to the Crashes in DC layer. It is a related table containing details for each crash such as methods of transportation, some demographics for persons and injury types.

2.2 Innovations

To innovate upon current methods, the team is working to incorporate a large number of datasets and analyze them thoroughly in Python. For example, most current research on D.C. car crash data focuses on wards – the political voting boundaries pre-designated by the city – to analyze the data. However, the team plans to see if there are any other logical, geospatial clusters that better portray patterns within the data. To do so, we plan on comparing the results of a few unsupervised learning methods like K-Means Clustering and DBSCAN. The team will also perform probabilistic clustering to understand the most likely days/times for crashes to occur. The results will be displayed in a Tableau dashboard, creating user-friendly and interactive views to the end users.

2.2.1 K-Means Clustering (Under Development)

The first approach we will use is a classical K-Means clustering algorithm. In this approach, we pre-set the number of clusters we would expect, and see how cleanly the car crashes separate into respective clusters. The rationale here is that DC has 8 pre-defined wards, so we can see if a ‘K’ value of 8 provides optimal separability in the data, or if a different value of ‘K’ does. We will measure separability of the clusters by looking at an elbow plot of the sum of squared differences between clusters.

2.2.2 DBSCAN

The second approach we will use is DBSCAN, a density-based clustering approach. In this approach, we pre-set the maximum distance that points can be set apart in order to be clustered together. The benefit

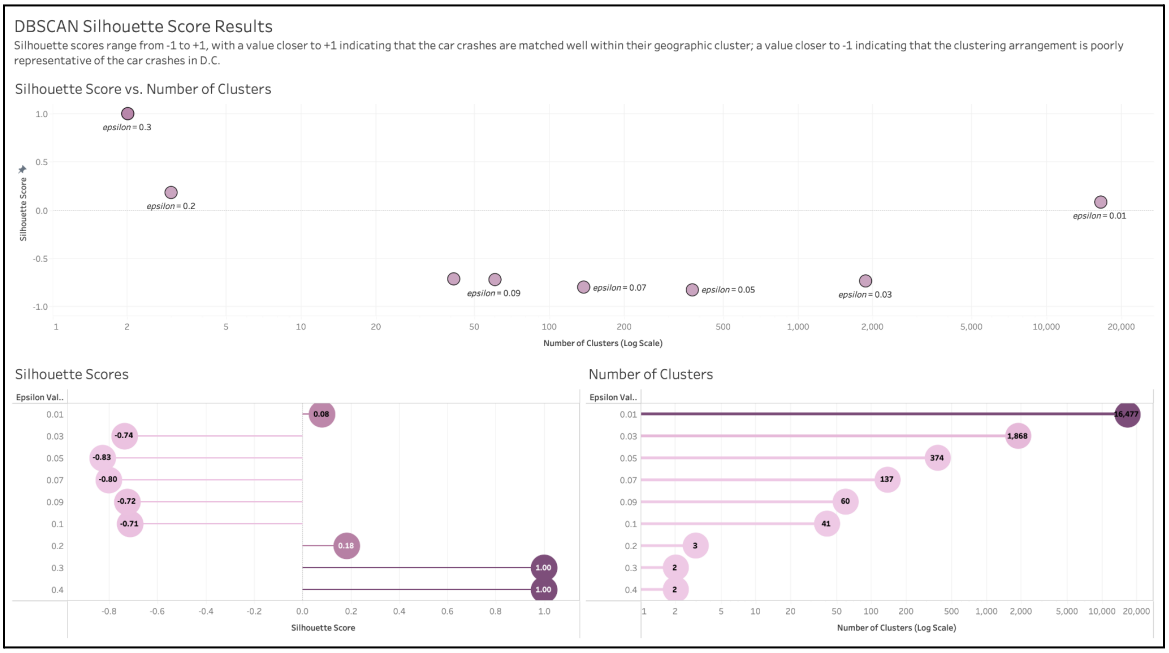
[1] https://opendata.dc.gov/datasets/DCGIS::crashes-in-dc/api
[2] https://github.com/jschulberg/DC-Transportation-Crashes/issues/3

here is that we do *not* need to pre-set the number of clusters we would expect. The rationale here is that it is difficult to pre-determine the number of clusters needed; instead, looking at the density of points on a map would be useful. To use the correct implementation of DBSCAN, we plan on following the approach used by Geoff Boeing in his paper *Clustering to Reduce Spatial Data Set Size* using the built-in 'haversine' metric, which takes into account curvature of the Earth so we can properly use Latitude and Longitude points.

To evaluate the results of DBSCAN, we plan on measuring the silhouette score for various values of the main parameter, *epsilon*, one of the two main parameters we can change in the algorithm. In this case, the silhouette score measures the ratio of the average distance between points *within a cluster* (*a*) divided by the average distance between points *within that cluster to the next nearest cluster* (*b*). Thus, the formula is:

$$\frac{b-a}{max(a,b)}$$

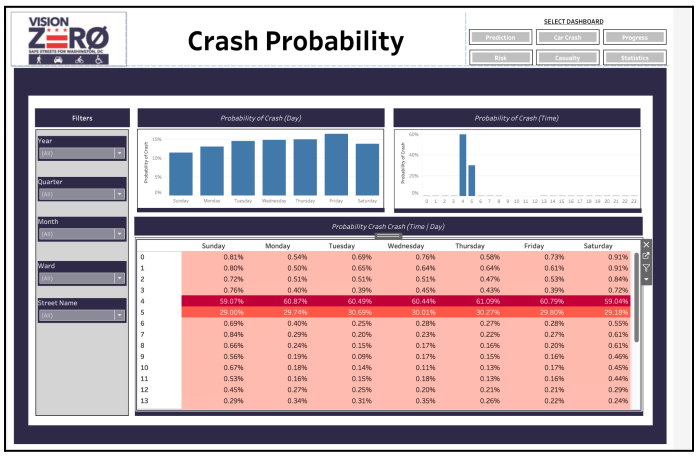
We know from Fabrice Muhlenbach's paper *A New Clustering Algorithm Based on Regions of Influence with Self-Detection of the Best Number of Clusters*, a value closer to 1 indicates that the car crashes are well-matched within their geographic clusters; a value closer to -1 indicates that the clustering arrangement is poorly representative of the car crashes in D.C. This will give us a good sense of how well-separated the clusters are for various values of *epsilon*. Here are the initial results from our analysis of various epsilon values on clustering the car crash data:



We can see that higher *epsilon* values (i.e. ≥ .2) yield a high silhouette score of almost 1; however, they come with the caveat that the number of clusters created is extremely low (just 2). We will have to do further visual analysis of this to ensure that these results make sense on a geographic map.

2.2.3 Probabilistic Risk Analysis

We will also use marginal/conditional probability of car crashes within Washington D.C. given the evidence of crashes happening derived from an intrinsic, conditional probability distribution with respect to month, week of the month, day of the week, time of day, and ward. The result will assist district police in their operations, policymakers in developing traffic laws, and will make D.C. residents more aware of local traffic risks.



The dashboard for probabilistic risk analysis is still under development, however, a mockup of the dashboard is provided below.

2.2.4 Tableau Dashboard (Under Development)

Lastly, the results of the analyses will be displayed in a tableau dashboard, creating user-friendly and interactive views for the end users. This end-product is still under development.

3. Experiments and Evaluations

The team will perform the following experiments/evaluations related to this analytical effort:

- Experiment on the optimal number of clusters for K-means clustering
- Evaluate the classification measures for K-means clustering
- Experiment to get the optimal cluster and distance for DBSCAN
- Evaluate the classification measures for DBSCAN
- Experiment on the probabilistic risk of crashes at different times, locations, days, and zones to determine high risk areas and times based on different factors.

4. Team Member Contribution

All team members have contributed to the project equally across all facet areas. The team continues to meet every Sunday and Thursday to meet the weekly project goals as discussed in the project proposal. The remaining tasks are listed to the right.

Tasks	Contributors
Finalize k-means clustering	All team members
Finalize DBSCAN	All team members
Finalize Probabilistic Risk Analysis	All team members
Finalize Tableau Dashboard and Final Report	All team members

5. Conclusion and Discussion

Thus far, we have encountered the following issues, which we are currently (and optimistically) working through:

- API limits us to pulling 1k rows at a time. 😕
- Date-time data is poorly maintained, precluding meaningful analysis prior to August 2021.
- We will likely need to do some additional geocoding in order to make certain data sources usable in modeling.

6. References

Bernhardt, M., & Kockelman, K. (2021, June 3). An analysis of pedestrian crash trends and contributing factors in Texas. Journal of Transport & Health. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/pii/S2214140521001201

Boeing, G. (2018, March 22). Clustering to Reduce Spatial Data Set Size. https://doi.org/10.31235/osf.io/nzhdc

Daraei, S., Pelechrinis, K. & Quercia, D. A data-driven approach for assessing biking safety in cities. EPJ Data Sci. 10, 11 (2021). https://doi.org/10.1140/epjds/s13688-021-00265-y

Guo, Y., Li, Z., Liu, P., & Wu, Y. (2019, April 29). Modeling correlation and heterogeneity in crash rates by collision types using full Bayesian random parameters multivariate Tobit model. Accident Analysis & Prevention. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/pii/S0001457518311576

Jensen, S. U. (2007, November 7). Bicycle tracks and lanes: A before-after study - researchgate. Retrieved October 14, 2022, from https://www.researchgate.net/profile/Soren-Jensen-16/publication/237524182_Bicycle_Tracks_and_Lanes_a_Before-After_Study/links/5a548377458515e7b732688e/Bicycle-Tracks-and-Lanes-a-Before-After-Study.pdf?origin=publication_detail

Kim, E., Muennig, P., & Rosen, Z. (2017, January 9). Vision Zero: A toolkit for road safety in the modern era - injury epidemiology. SpringerLink. Retrieved October 12, 2022, from https://link.springer.com/article/10.1186/s40621-016-0098-z

Kondo MC, Morrison C, Guerra E, Kaufman EJ, Wiebe DJ. Where do bike lanes work best? A Bayesian spatial model of bicycle lanes and bicycle crashes. Saf Sci. 2018 Mar;103:225-233. doi: 10.1016/j.ssci.2017.12.002. PMID: 32713993; PMCID: PMC7380879.

Lee, J., Abdel-Aty, M., & Cai, Q. (2017, March 21). *Intersection crash prediction modeling with macro-level data from various geographic units*. Accident Analysis & Prevention. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0001457517301070

Marshall, W. E., &; Ferenchak, N. N. (2019, May 29). Why cities with high bicycling rates are safer for all road users. Journal of Transport & Health. Retrieved October 13, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S2214140518301488

Muhlenbach, F., & Lallich, S. (2009). A New Clustering Algorithm Based on Regions of Influence with Self-Detection of the Best Number of Clusters. *2009 Ninth IEEE International Conference on Data Mining*, 884-889.

Rogers, J. M., Dey, S. S., Retting, R., Jain, R., Liang, X., & Askarzadeh, N. (2016, December 5). Using automated enforcement data to achieve vision zero goals: A case study. Retrieved October 12, 2022, from https://ieeexplore.ieee.org/abstract/document/7841099/authors#authors

Shannon, D., & Fountas, G. (2022, March 3). Amending the heston stochastic volatility model to forecast local motor vehicle crash rates: A case study of washington, d.c. Transportation Research Interdisciplinary Perspectives. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/pii/S2590198222000392

Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016, July 29). *Macro-level safety analysis of pedestrian crashes in Shanghai, China*. Accident Analysis & Prevention. Retrieved October 12, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0001457516302573

Xiong, C., Mahmoudi, J., Luo, W., Yang, M., Zheng, J., & Delion, C. (2021, August 31). A data-driven safety dashboard assessing Maryland statewide density exposure of pedestrians, bicycles, and e-scooters. A Data-Driven Safety Dashboard Assessing Maryland Statewide Density Exposure of Pedestrians, Bicycles, and E-Scooters. Retrieved October 12, 2022, from https://rosap.ntl.bts.gov/view/dot/61218