D.C. Car Crash Analysis

CSE 6242 Final 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.**

2. 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, time series analysis and probability-based risk analysis. Results can be incorporated into an improved dashboard to provide further insights to DC citizens and key stakeholders.

3. 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 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. The team's research looks 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 the team's 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.

4. Proposed Method

4.1 Intuition

The DC Vision Zero website has visualizations available, but they are not interactive and do not contain statistical and machine learning models. In our method, we provide interactive visualization capability such that users can analyze specific locations, dates, individuals injured, injury type, etc. We also conduct statistical and machine learning models such as Time Series Analysis, Density Based Spatial Clustering (DBSCAN) and K-means Algorithm to enhance understanding of the data. Lastly, we have created our own website, *Vision Zero DC Analytics*, which contains all of our final outputs.

4.2 Approaches

4.2.1 Data

The team pulled 270,000+ records of car crashes in D.C. using the OpenDataDC API. Data cleaning was performed using the *pandas* and *numpy* libraries. The data was pivoted into a more effective visualization structure, strings were reformatted, dates corrected into the correct date-time format, and duplicates were removed.

4.2.2 Vision Zero DC Analytics Website

The team has cataloged its results in the following website, *Vision Zero DC Analytics*. The site includes:

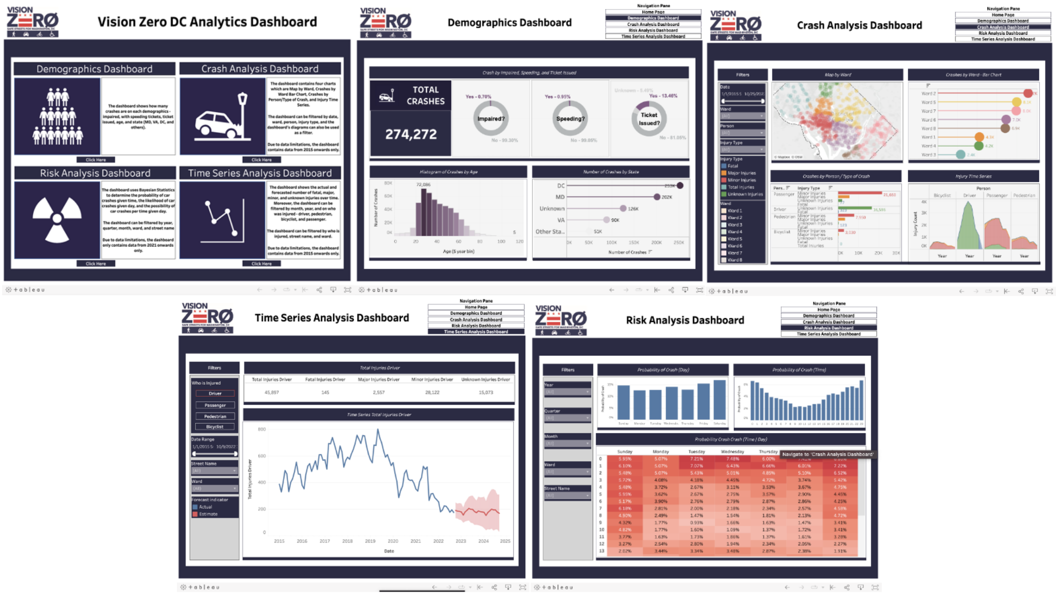
- **Home Page** | The group mission, which is to support D.C. Vision Zero initiative by creating interactive visualizations and machine learning models.
- **Dashboard** | All 4 of the team's Tableau dashboards.
- **Machine Learning** | Machine learning blogs and full write-ups.
- **About Us** | The team member bios and team contact information.

4.2.3 Dashboard

The team's main Tableau workbook includes 4 dashboard views:

- **Demographics Dashboard** | Breaks down crashes by rider impairment, driver speeding, ticket issuance, individual's age, and license plate state (MD, VA, DC, and others).
- **Crash Analysis Dashboard** | Visualizes crashes by ward, person, and type of injury sustained. The dashboard can be filtered by date, ward, person, injury type, and is fully interactive. Due to data limitations, this dashboard contains data from 2015 onwards only.
- **Risk Analysis Dashboard** | Uses Bayesian Statistics to determine the probability of car crashes given time, the likelihood of car crashes given day, and the possibility of car crashes per time given day. The dashboard can be filtered by year, quarter, month, ward, and street name. Due to data limitations, the dashboard only contains data from 2021 onwards.
- **Time Series Analysis Dashboard** | Shows the actual and forecasted number of fatal, major, minor, and unknown injuries over time. Moreover, the dashboard can be filtered by month, year, and on who was injured - driver, pedestrian, bicyclist, and passenger. It can be filtered by who is

injured, street name, and ward. Due to data limitations, the dashboard only contains data from 2015 onwards.



4.2.4 Machine Learning

K-Means Clustering

The first approach used is a classical K-Means clustering algorithm. In this approach, the number of clusters is pre-set the number of clusters to see how cleanly the car crashes separate into respective clusters. The rationale is that DC has 8 pre-defined wards, which are similar to congressional districts in that the boundaries are assigned during each 10-year redistricting cycle such that each ward has roughly the same population. Setting a 'K' value of 8 illuminates the extent to which crash clusters correspond to existing ward boundaries. Since there can be multiple crashes at the same location, weighting locations by the number of crashes occurring in a given location provides a more meaningful understanding of crash concentration. Cluster separability is measured using an elbow plot of the sum of both weighted and unweighted squared differences between clusters.

Density Based Spatial Clustering (DBSCAN)

The second approach used is DBSCAN, a density-based clustering approach. In this approach, the maximum distance that points can be set apart in order to be clustered together is pre-set. The benefit is that the number of clusters does not need to be pre-set. This allows the number of clusters to be assigned from model parameters subject to the structure of the data itself.

5. Analytic Experiments

Risk Analysis Experiment

This experiment answers which time, day, and time of day are considered the most likely for a car crash to occur. When the team researched the most dangerous time to drive, it found multiple, conflicting answers. This inspired the team to conduct its own assessment using a Bayesian Probability method on the D.C. car crash dataset, given by following formula in which *A = Time* and *B = Day*:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Results indicate that between Sunday and Thursday the probability of a car crash is less than 15%. On the other hand, between Friday to Saturday there is a likelihood greater than 15%. Saturday is the peak, with the probability of a car crash occurring equal to 17.1%. By hour of day, car crashes are most likely to occur between 8 pm and 2 am, with a probability of more than 5% during each of those hours.

Time Series Experiment

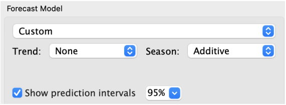
The team used time series forecasting to make scientific predictions of the number of injuries by type of person in 2024 and beyond to help the D.C. government determine if it will successfully achieve its Vision Zero Initiative of 0 crash-related fatalities. The team constructed the predictions in Tableau, which provides a built-in exponential smoothing forecasting model, by experimenting with several combinations of parameters including additive and multiplicative time series models, as well as components for trend and seasonality.

To evaluate the accuracy of the produced models, the team used the Mean Absolute Scaled Error (MASE), given by the following equation:

$$MASE = \frac{1}{n} \sum_{t=1}^{n} |q(t)|$$

where *q(t)* represents the relative absolute forecast error (forecast error divided by the mean absolute error of the forecasting method on prior data). Thus, a MASE ≥ 1 implies that the actual forecast performs worse than a naïve benchmark forecasting method calculated in-sample; a MASE < 1 means the forecast performs well. The lower the MASE value, the lower the relative absolute forecast error, the better the method. Based on the experiment, the optimal model is exponential smoothing with additive seasonality.

Below is a screenshot of the Tableau forecast model interface, a sample KPI output, and a sample of the visualization provided by Tableau, available on the team's aforementioned website.

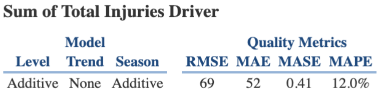


Forecast Model

Custom

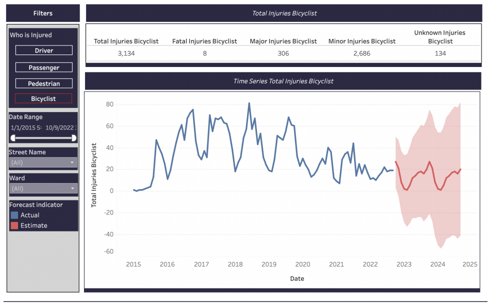
Trend: None Season: Additive

☑ Show prediction intervals 95%



Sum of Total Injuries Driver

Model			Quality Metrics			
Level	Trend	Season	RMSE	MAE	MASE	MAPE
Additive	None	Additive	69	52	0.41	12.0%



Total Injuries Bicyclist

Total Injuries Bicyclist	Fatal Injuries Bicyclist	Major Injuries Bicyclist	Minor Injuries Bicyclist	Unknown Injuries Bicyclist
3,134	8	306	2,686	134

Date Range: 1/1/2015 5: 10/9/2022 :

Street Name: (All)

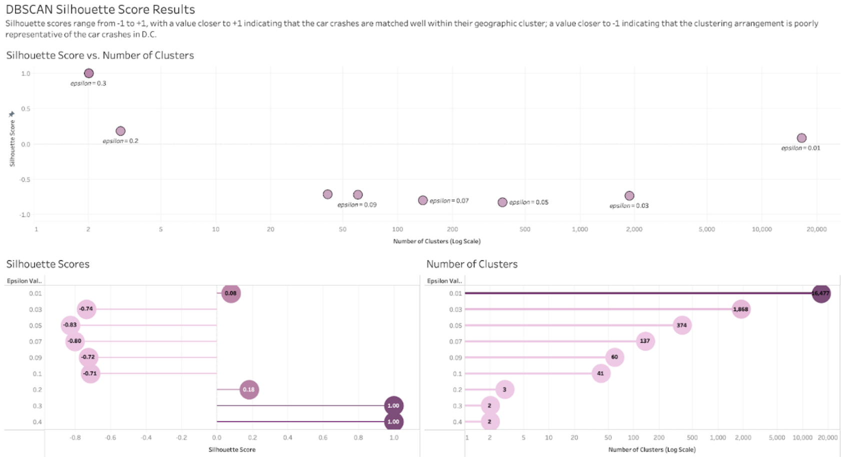
Ward: (All)

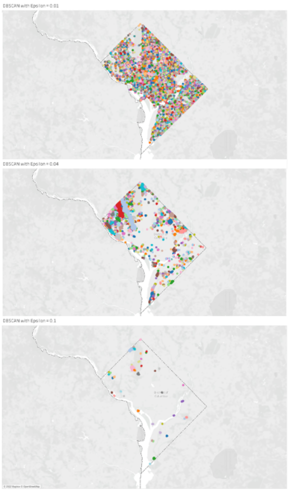
DBSCAN Results

To use the correct implementation of DBSCAN, the team followed 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 properly incorporate Latitude and Longitude data. To evaluate the results of DBSCAN, the team measured the silhouette score for various values of the main parameter, *epsilon*, one of the two main model parameters. 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:

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

According to 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 effectively illustrates how well-separated the clusters are for various values of *epsilon*. Here are the initial results from the analysis of various epsilon values on clustering the car crash data:

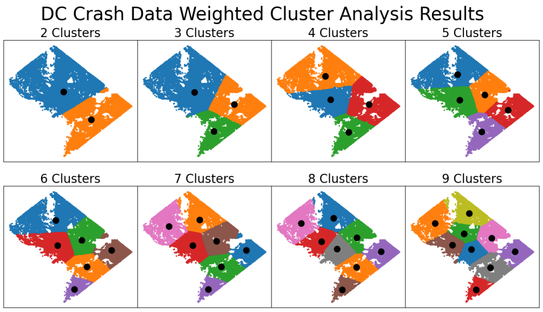


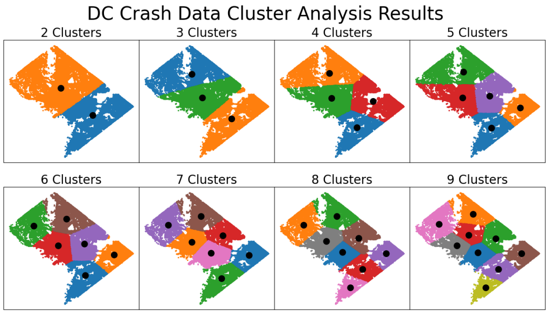

DBSCAN with Epsilon = 0.01
DBSCAN with Epsilon = 0.04
DBSCAN with Epsilon = 0.1

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). Lower values of *epsilon* (i.e. ≤ .1) yield more meaningful clusters, but the algorithm still struggles to deal with the density of points. For example, the following three maps represent DBSCAN with *epsilon* equal to .01, .04, and .1, respectively, where points attributed to noise are filtered out. Overall, the results are less than meaningful.

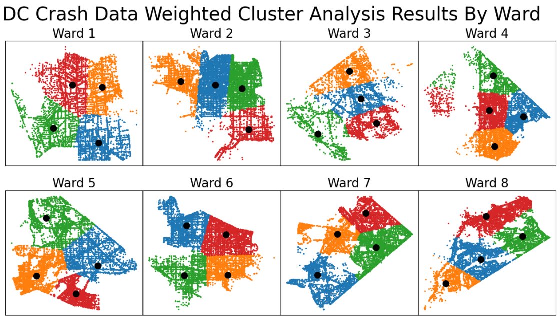
K-Means Results

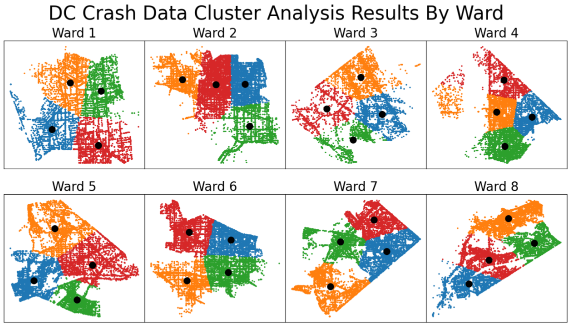
Results indicate that the K-Means algorithm, using both weighted and unweighted data, provides little value to the analysis of crash data. Optimal separability occurs at about 4 clusters, with limited improvement in performance beyond that, meaning that separability is only slightly better for 8 clusters than for 4. The crash data is so dense relative to the size of each ward, that the results are rarely meaningful outside of specific subgroups exhibiting sufficient sparsity for meaningful patterns to emerge.





Weighting by the number of crashes occurring at a location does meaningfully change the cluster centers in situations where a disproportionate share of crashes in a region occur in a specific area. In ward 3 for example, in the far northwest of the city, a heavy concentration of crashes occur along key road arteries. Weighting significantly changes the location of cluster centers, denoted by black dots.





5. Team Member Contribution

All team members have contributed to the project equally across all facet areas. The team met bi-weekly to tackle the weekly project goals laid out in the project proposal.

Distribution of Work

Tasks	Adam Pier	Justin Schulberg	Maynard Miranda	Ryan Doogan
Proposal	✓	✓	✓	✓
Proposal Presentation	✓	✓	✓	✓
Data Gathering and Cleaning	✓	✓	✓	✓
Analysis	✓	✓	✓	✓
Progress Report	✓	✓	✓	✓
Final Report	✓	✓	✓	✓
Final Poster Presentation	✓	✓	✓	✓

6. Conclusion and Discussion

This study evaluated the D.C. car crash dataset, and used interactive dashboards, machine learning models, and an easy-to-use website to help the D.C. government achieve its Vision Zero initiative (Website Link). Users are able to utilize the resources the team has created directly from the website to analyze common demographics and risk factors associated with increased risks of car crashes. From our analysis, we determined that some of these primary risk factors include, but are not limited to: those with ticket violation(s) are 13.45% more likely to get involved in a car crash; Saturday evening is the most dangerous time to drive; people aged 25-30 are more predisposed to being involved in a crash; drivers from Maryland contributed to more crashes than drivers from any other state (including Washington, D.C.); and crashes are most prevalent in Wards 2, 5, and 7. Lastly, despite the team's best efforts to use unsupervised clustering algorithms to look for geographic trends in the data, the data is unfortunately too dense to get meaningful results.

7. 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., &amp; 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 &amp; 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., &amp; 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., &amp; 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