Emergency Department Records Data Mining and Machine Learning from the Rapid Health Information Network (RHINO)

Report Prepared by:

Cody Carmichael
Lareina La Flair
Amanda Morse
Tom Hulse

Washington State Department of Health

WTSC Grant Number 2021-TR-4098

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
09/29/2021	Grant Deliverable - Report	Oct 1, 2020 - Sep 30, 2021

4. TITLE AND SUBTITLE
Emergency Department Records Data Mining and Machine Learning from the Rapid Health Information Network (RHINO)

5a. CONTRACT NUMBER
2021-TR-4098

5b. GRANT NUMBER
2021-TR-4098

5c. PROGRAM ELEMENT NUMBER
n/a

6. AUTHOR(S)
Cody Carmichael
Lareina La Flair
Amanda Morse
Tom Hulse

5d. PROJECT NUMBER
n/a

5e. TASK NUMBER
n/a

5f. WORK UNIT NUMBER
n/a

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Rapid Health Information Network (RHINO) Program
Office of Public Health Outbreak Coordination, Informatics, and Surveillance
Washington State Department of Health

8. PERFORMING ORGANIZATION REPORT NUMBER
n/a

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Washington Traffic Safety Commission
PO Box 40944
Olympia, Washington 98504-0944

10. SPONSOR/MONITOR'S ACRONYM(S)
WTSC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
2021-TR-4098

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release.

13. SUPPLEMENTARY NOTES
The contents of the manuscript are solely the responsibility of the authors and do not necessarily reflect the official views of the funding agency.

14. ABSTRACT
The Washington State Department of Health Rapid Health Information Network (RHINO) program staff, with support of a grant from the Washington Traffic Records Governance Council, conducted a pilot evaluation of big data text mining, machine learning, and spatiotemporal cluster detection methods for analysis of traffic injury. This report provides insights into the application of machine learning algorithms applied to emergency department records generated from January 1, 2021 - August 31, 2021.

15. SUBJECT TERMS
text mining, machine learning, syndromic surveillance, emergency department records, traffic records, motor vehicle crash,

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Peter Corier, Traffic Records Program Manager
					19b. TELEPHONE NUMBER *(Include area code)*
None	None	None	None	13	(360) 725-9879

WTSC Data Mining and ML Report

RHINO Team

9/29/2021

Summary

To further the accuracy of insights gained from ED records, a pair of Machine Learning algorithms were applied to 3,247 records gathered from the ESSENCE All Traffic Query for the purposes of validation and custom categorization as directed by WTSC staff. In validation, it was shown that the samples garnered an average of 98.21% accuracy across three key fields. In Categorization efforts, results were extremely mixed and point to a need of further examination and discussion of classification criteria for a range of injury severities involving Motor Vehicle Collisions (MVCs).

The Machine Learning algorithms, once trained, were applied to 36,744 records, from January 1, 2021 to August 31, 2021 , to allow for text mining and categorical insights. From the 34,172 (93%) confirmed records, text mining of key fields was performed to determine similarties and differences between custom criteria as informed by WTSC staff. Results of this text mining included insights that inferred that the 4 categories as described (1 = "suspected injury", 2 = "minor injury", 3 = "major injury", 4 = "death") may need some exclusion or additional criteria to perform as well as other categorizations such as scooter related collisions or pedestrian involved collisions.

Lastly, given the limitations of ED data (which is to say, current unavailability of crash or injury site information), it can be assumed that ED data alone may not be sufficient for spatiotemporal cluster detection. However, some insights into time of day and distance from home patterns may be uncovered with some limitations. Further synthesis with other data sources, name EMS data, may provide further insights.

Machine Learning and Text Mining Results

As stated within the introduction, the application of Machine Learning on traffic records was performed with two primary goals in mind. The first of these can be defined as a "Binary Classification" problem, where we wished to test if the records being pulled from ESSENCE's All Traffic Related query were truly traffic related. The second goal was to examine the possibility of using a custom criteria to assess severity of ED visits related to Traffic Injuries, which is defined as a "Multi-Classification" problem.

To solve the Binary Classification problem, it was determined that a Deep Learning Binary Classification algorithm would be the most successful across multiple fields found in records (namely, Discharge Diagnosis, Triage Notes, and Chief Complaint fields). After training and testing on a sample set of 3,427 manually tagged records, it was found that Chief Complaint based selection retained a 97.85% selection accuracy, Discharge Diagnosis 98.82%, and Triage Notes, 97.96% after 24 epochs (training periods). Graphs of both accuracy and loss (how much information is "lost" while the neural layers adapt and train) are displayed below in the graphics pages at the end of this report.

The resulting Deep Learning models were used to filter the original manually tagged samples and later a new sample set of 36,744 visits to ensure the best possible chance of success with Multi-Classification modeling.

Mutli-Classification Modeling was performed after initial filtering by a Support Vector Machine algorithm. This algorithm "clusters" visits together by most likely category and creates "borders" by which one can determine most likely grouping. In this instance, "borders" were produced based on the following criteria:

- Class 1: Individuals with no apparent visible injuries, but come to the ED seeking care.
- Class 2: Individuals with minor apparent visible injuries (scrapes, cuts, bruises, etc.).
- Class 3: Individuals with major injuries or injuries in particularly critical parts of the body (fractures, concussions, etc).
- Class 4: Individuals who died as a result of injuries sustained.

In the examination of these records based on the criteria, it was discovered there was fairly extreme and common cross-over of Triage Notes details, Discharge Diagnosis codes, and Chief Complaint details.

For Triage Notes, cross-group accuracy was noted at 42%, Discharge Diagnosis maintained a cross-group accuracy of 48.9%, and Chief Complaint cross-accuracy of 39.2%.

This leads one to the conclusion that further refinement of definitions of these categories is needed to provide the model with a successful framework with which to train another model to allow for high-volume classification and further insights.

At time of writing, text analysis that differentiates based on these results provided no meaningful results. However, frequency of Discharge Diagnosis codes in singles and pairs are reported as general trends, as well as Triage Notes and Chief Complaint unigram and bigrams based on ED visit without admission (Not Severe), ED visit with inpatient admission (Severe), and those who died while in the ED or during inpatient status (Died). These can be found in the Graphics pages at the end of this report.

Spatiotemporal Characteristics in Traffic Related Injuries

Spatial Characteristics of Traffic Related Injuries

While we do not get frequent information about where a traffic related injury may have occured, RHINO does receive information about location at which the patient sought care, and the patients home zip code. For the map displayed in the graphics pages, patient zip code was used to map out the Per 10,000 rate of Traffic Injuries as it relates to total ED visits. It's worth noting that most of the zip codes that have higher rates of injury are either very rural (low overall visit count), or near major interstate systems (I-5 and I-90).

Time of day and day of week trends in Traffic Related Injuries

It should be noted that a significant number of visits (notably, those who fall under a 2 or 1 in the custom categories) would come into an ED hours or even days after an event. As such, clustering may not be seen as the most conducive action at this time. Furthermore, this issue points to a need to integrate with near-real-time EMS data into ED analysis to provide a more comprehensive view of traffic related events within Washington State.

However, there are some insights that still may be gleaned from our data, namely that over the time period examined (January 1, 2021 to August 30, 2021), Motor Vehicle Collisions and other related injuries were most common on Thursday evenings between 6 P.M. and 8 P.M., with Wednesdays at the same time being the second most prevalent block of time. This roughly corresponds to Twilight Hours during a majority of the months examined, in which commuting may be considered more difficult due to variance in light conditions. A heatmap of these results are presented in the graphics pages.

Graphics

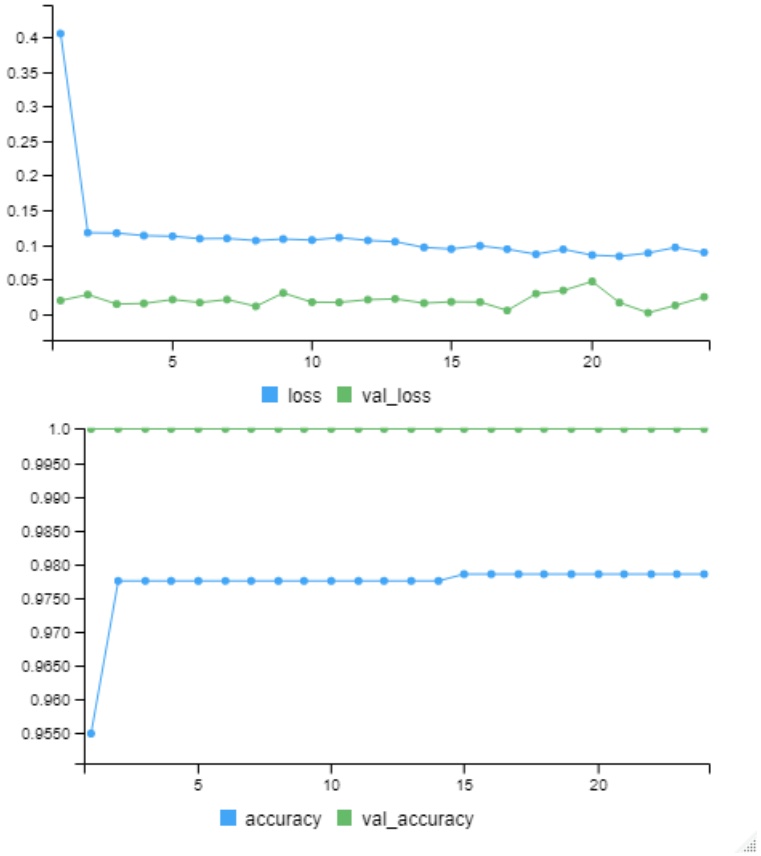


Figure 1: Graph of Loss and Accuracy for Chief Complaint

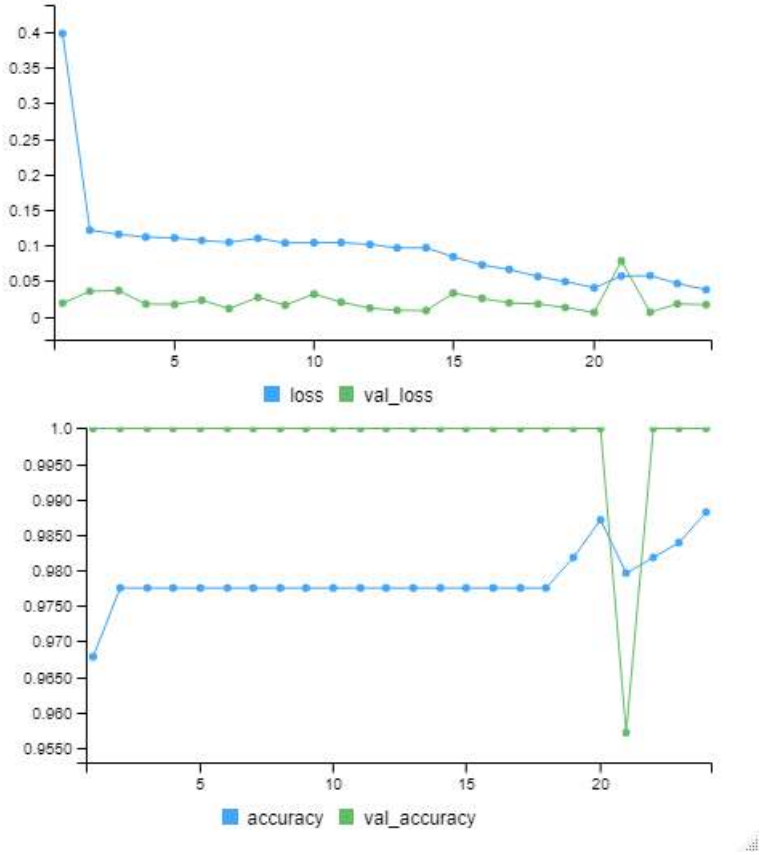


Figure 2: Graph of Loss and Accuracy for Discharge Diagnosis

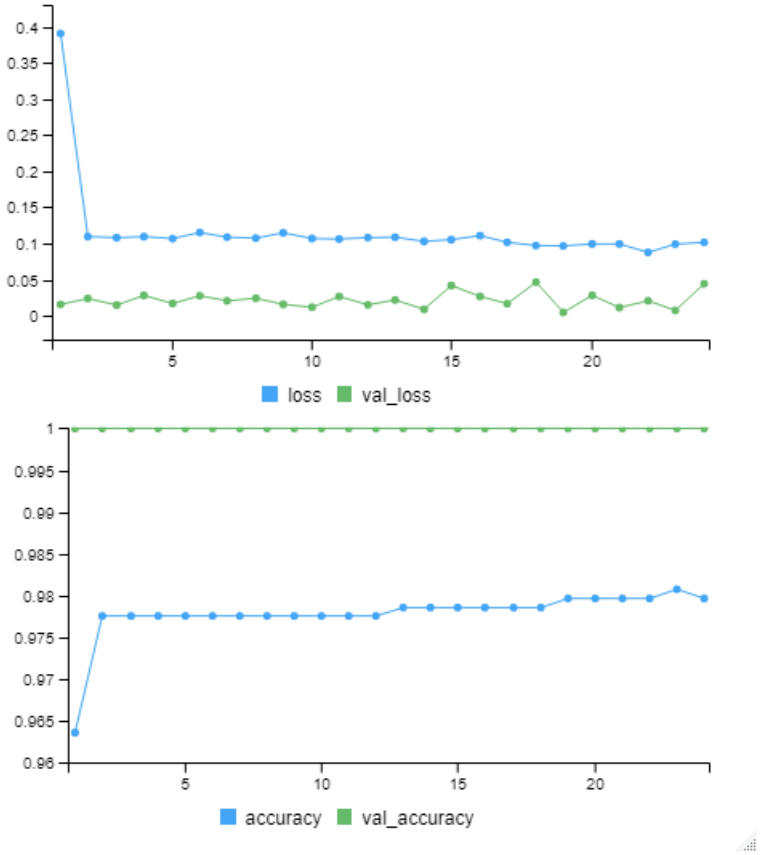


Figure 3: Graph of Loss and Accuracy for Triage Notes

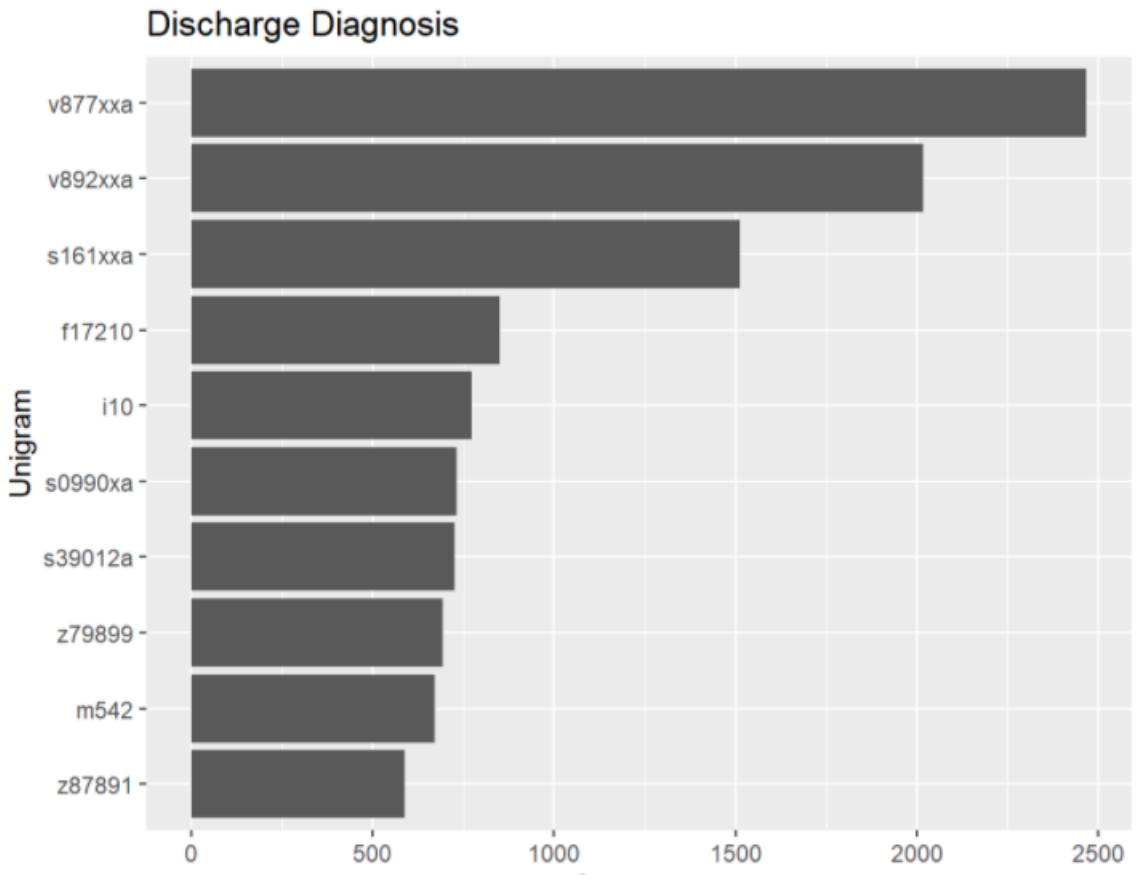


Figure 4: Discharge Diagnosis Frequency - Singular

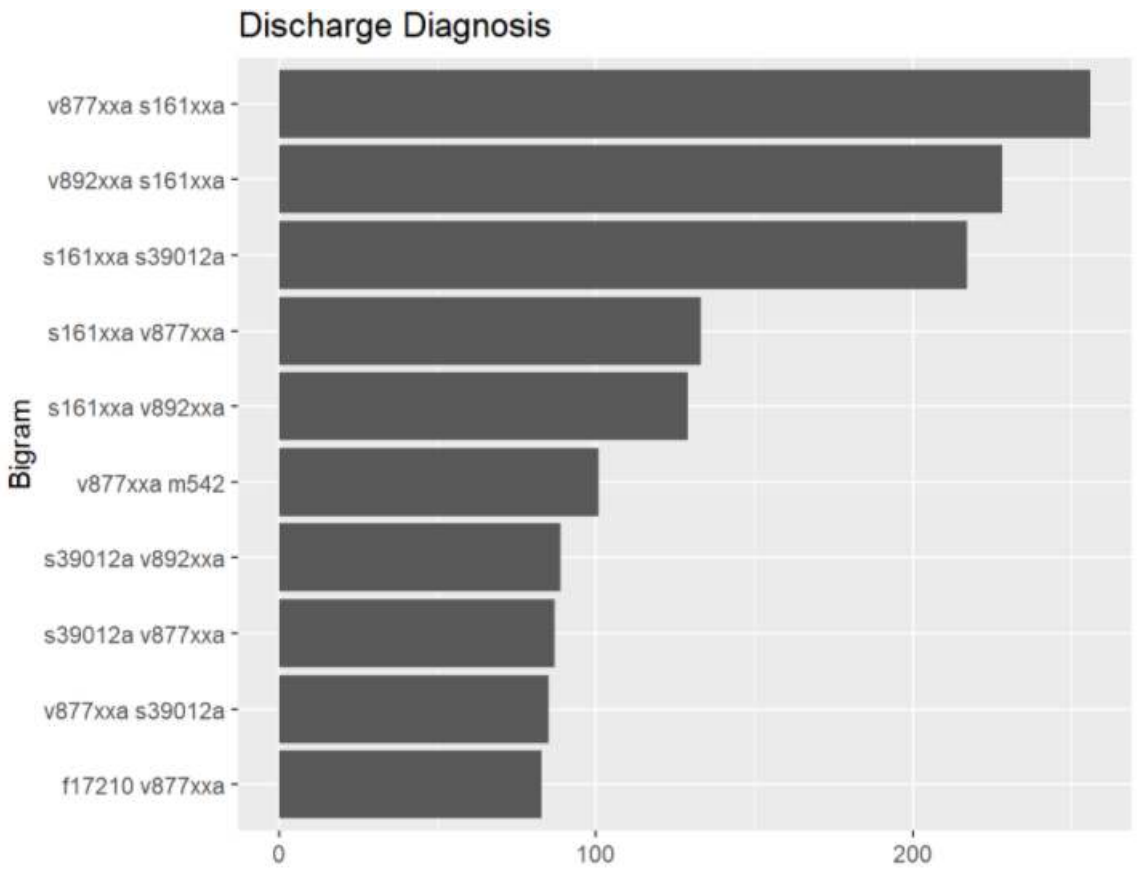


Figure 5: Discharge Diagnosis Frequency - Pair

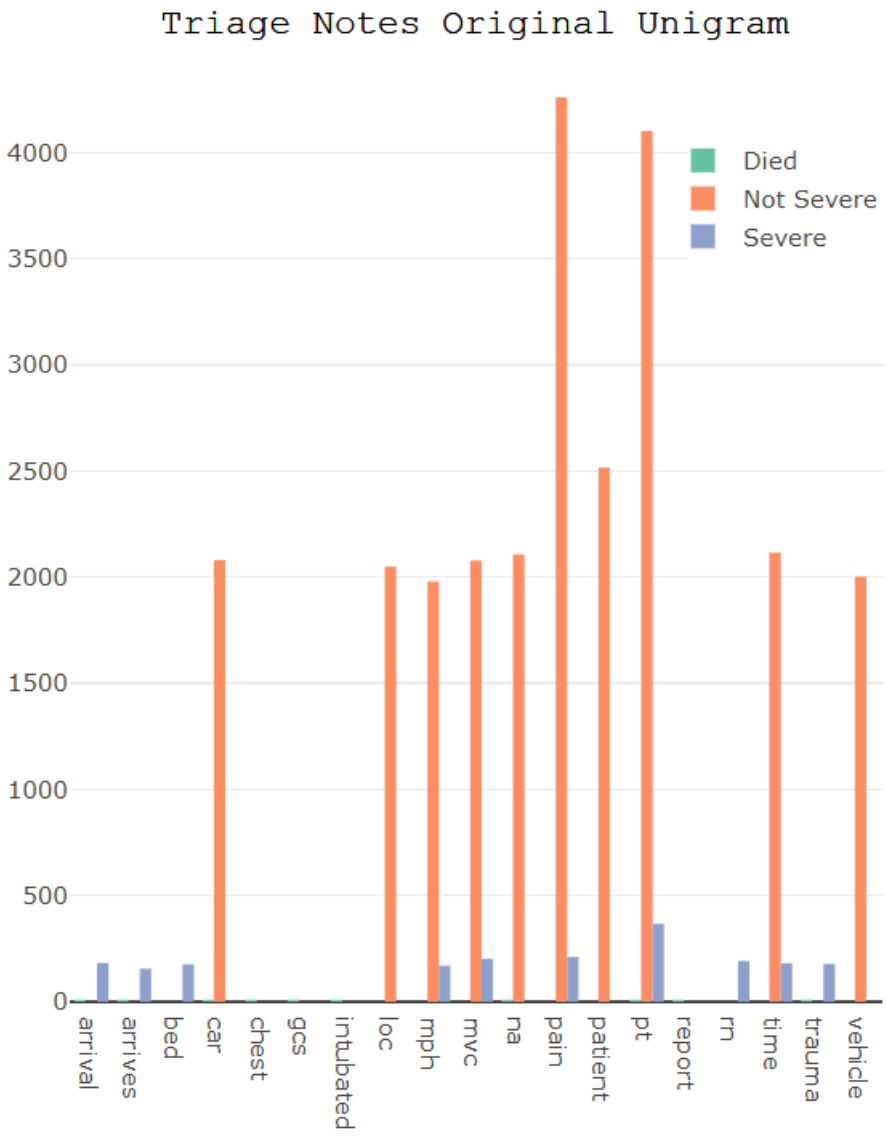


Figure 6: Triage Notes Frequency - Unigram

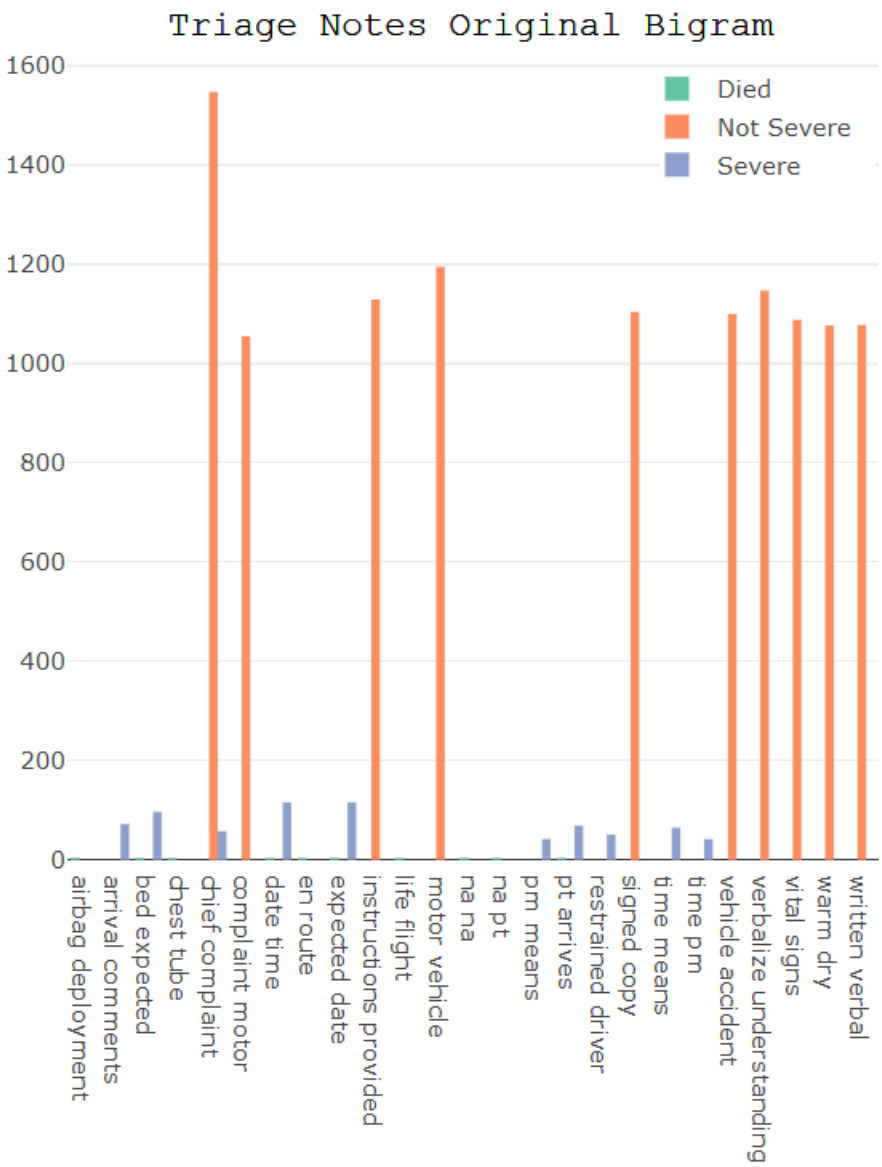


Figure 7: Triage Notes Frequency - Bigram

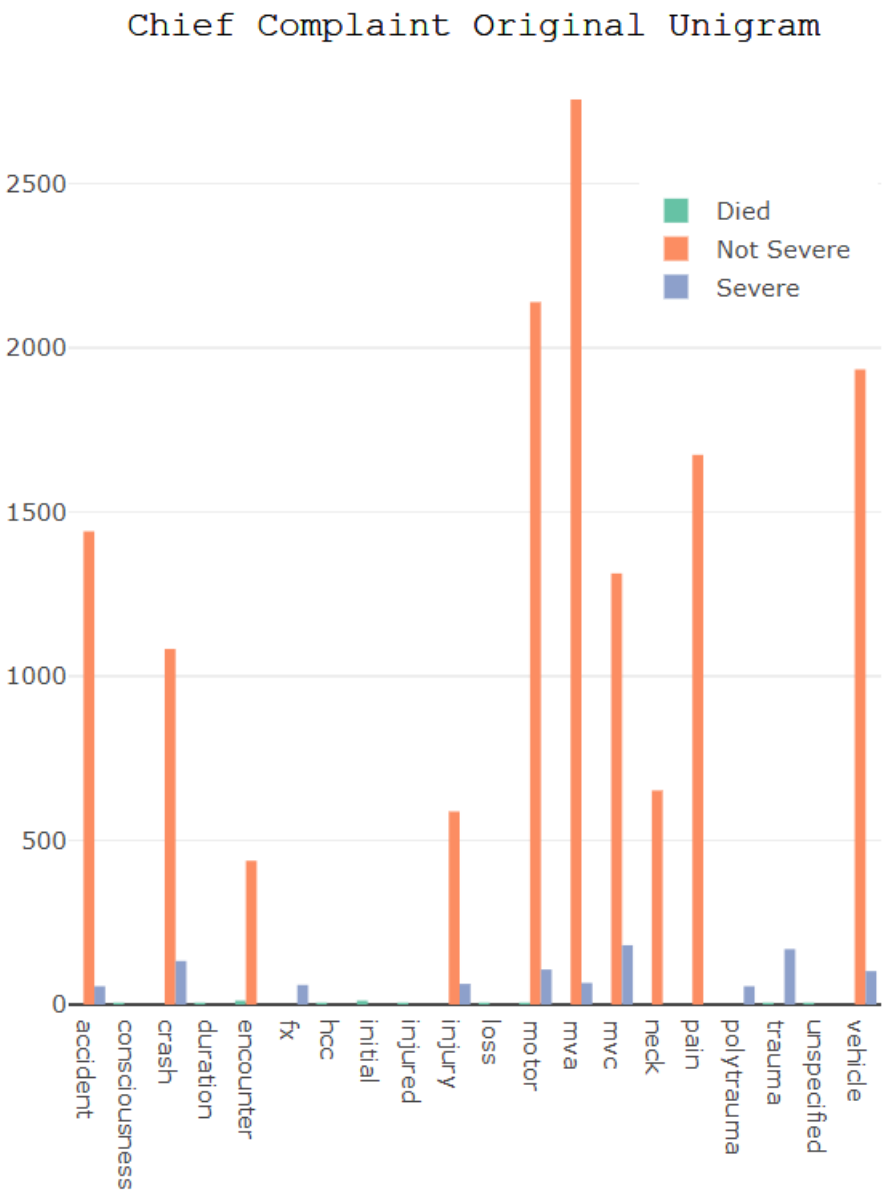


Figure 8: Chief Complaint Frequency - Unigram

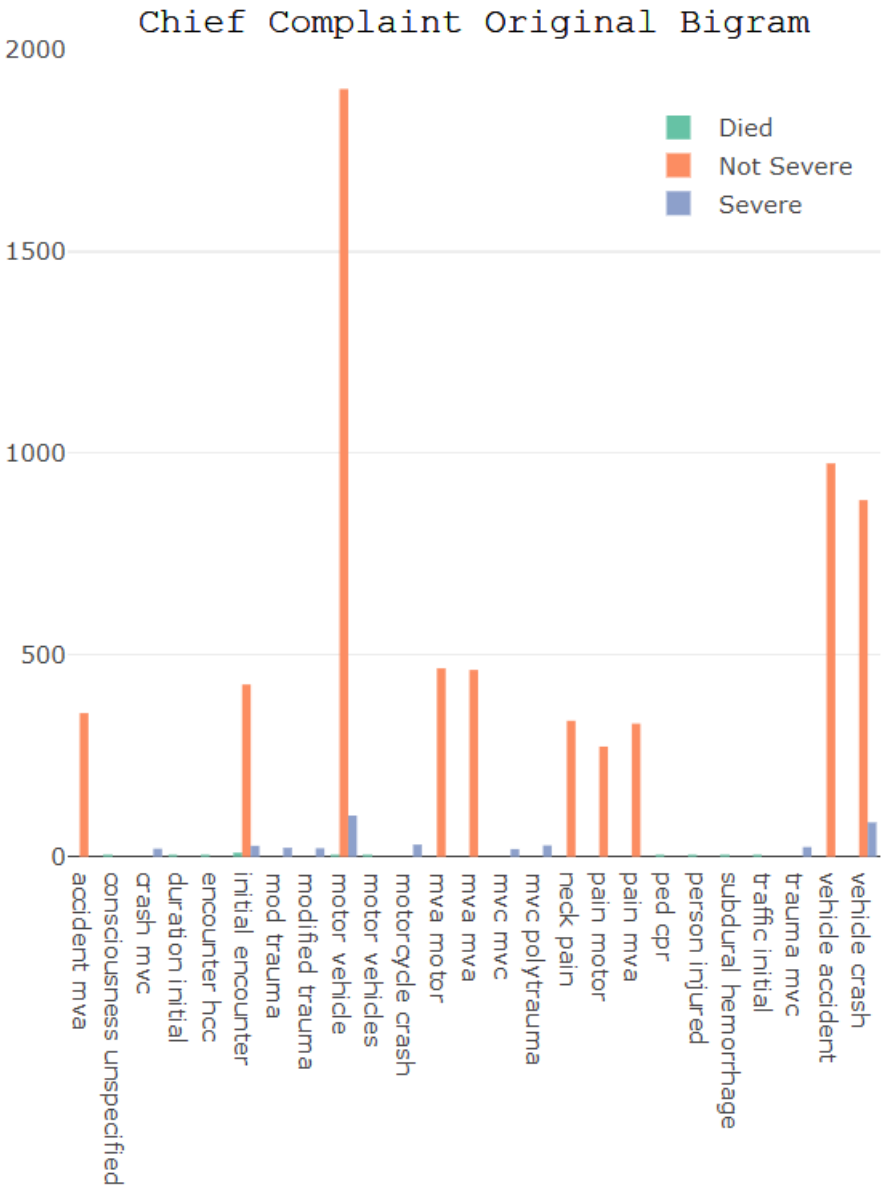


Figure 9: Chief Complaint Frequency - Bigram

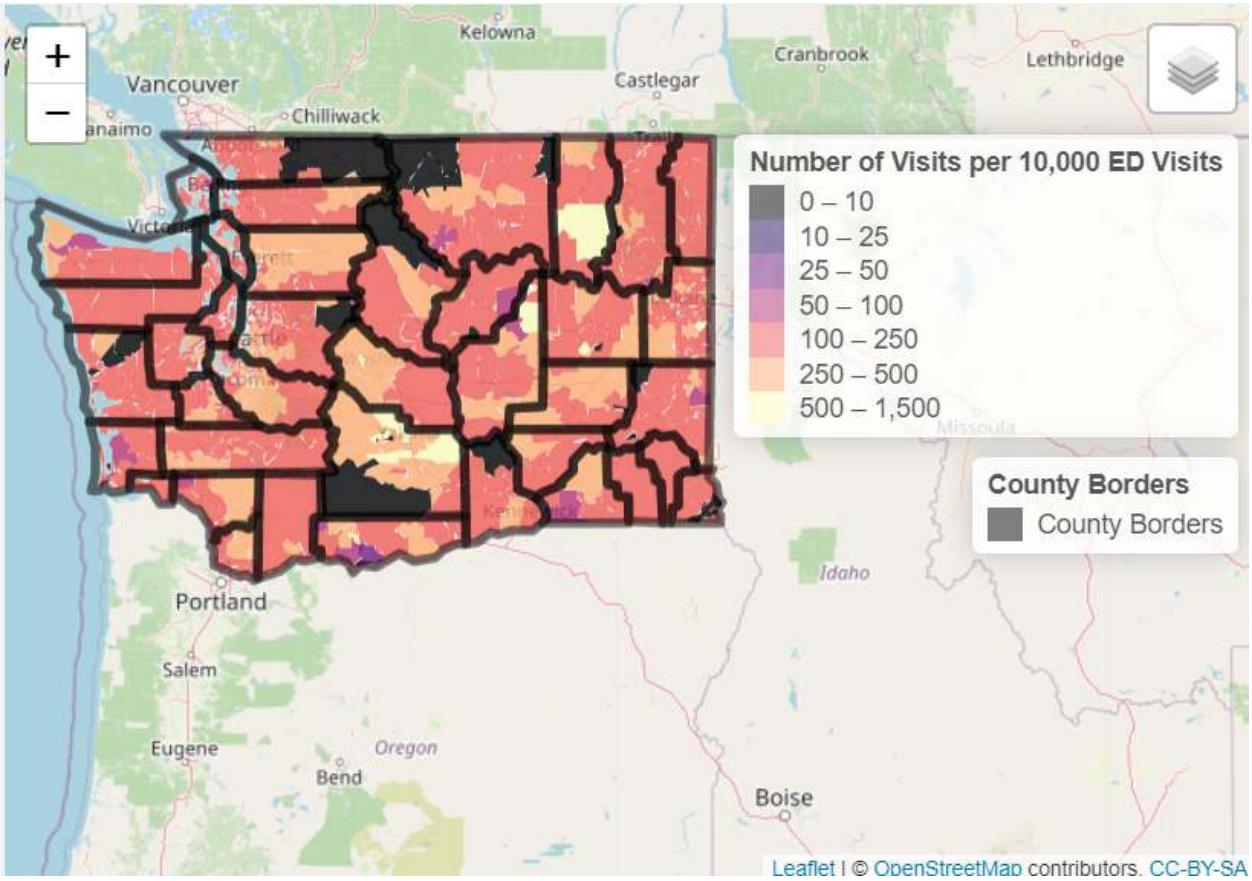
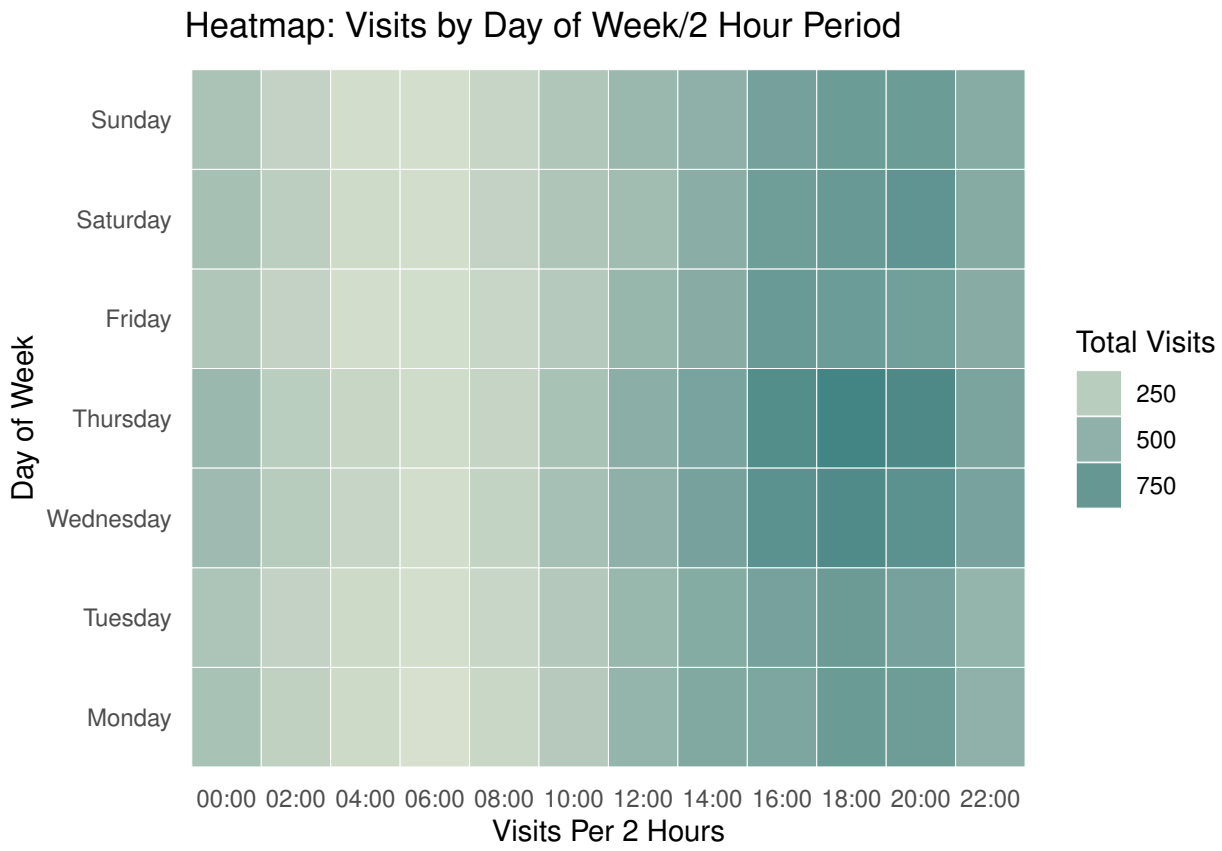


Figure 10: Map of Washington State Traffic Related Visits, Visits per 10,000 ED Visits


Heatmap: Visits by Day of Week/2 Hour Period
Day of Week
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Visits Per 2 Hours
Total Visits
250
500
750