Traffic Accident Analysis in Washington D.C.

Joanne Choi, Sam Clark, Ranjan Jaiswal, Peter Kirk and Sachin Jayaraman

DATA 601: Introduction to Data Science

Dr. Huthaifa Ashqar

May 17th 2021

Abstract

Previous studies have been done in various cities to advance our understanding of crashes in traffic, relating to traffic volume, weather, and crash data. This paper aims to do the same in Washington, D.C. The paper attempts to determine links between weather and crash data, creating a model to predict crashes by time, using a negative binomial regression after rejecting a poisson regression, and additionally explores the validity of a random forest regression. Our main consideration for an eventual application of the work done here is a reduction in crashes, using a tool to give people better options on when to commute and when to telework, when available. Although more research into additional explanatory variables is necessary to draw more applicable conclusions, our results suggest that morning times before 4am, Fridays, and October are the most unsafe commuting times. On the other hand, the time period between 6am and 1pm, February, and Thursday (excluding Sat and Sun, as they are not workdays) are the safest. Given that the data is available, we believe that a model like ours could be applied to good use in lowering the number of crashes in D.C.

Introduction

Washington, District of Columbia (D.C.) is a major metropolitan area and the capital of the United States of America. The National Capital Region, colloquially known as the "DMV" (D.C., Maryland, Virginia) is known to be a very transient area, due to how many workers reside within the district and the two states. The National Capital Region Transportation Planning Board (2019) reported that nearly two-thirds of commuters drove alone within the Washington metropolitan area, as opposed to other means of commuting, such as Metrorail or teleworking. Additionally, the average American's commute time is 27 minutes, while D.C. commuters' average is 43 minutes (Berkon, 2020). The combination of high traffic volume and extensive commute times could possibly be linked to the frequency of vehicle crashes.

COVID-19 has reshaped the workforce and the commuting patterns. Before COVID-19, about 20% of the American workforce worked from home. Currently, 71% of the same workforce are now working from home and 54% would like to continue to work from home after restrictions are lifted (Parker et al., 2021). Due to the shift from commuting to telecommuting, we believe that many employers will be more flexible with their employees, allowing them to create their own commuting schedule rather than commute to work on a daily basis.

Given the massive number of daily commuters that D.C. received prior to COVID-19, we propose developing a general regression model to predict the likelihood of vehicle collision based on various risk factors. We will analyze crash data from within D.C. from 2016 to 2019 with weather data to determine which factors contribute most to crashes. Our analysis could be used to inform commuters on the likelihood of vehicle collision on any given day, possibly impacting the decision on whether to commute into the office. This model is intended to help reduce the number of commuters on days that are predicted to be more likely to have vehicle collisions.

Literature Review

Road safety regarding traffic accidents has been an issue for as long as traffic has existed. Many studies have analyzed vehicle collisions to determine the risk factors and how they may contribute to the likelihood and severity of car crashes. Some of these risk factors include weather conditions and time of day. By determining the likelihood and the severity of the impact each factor has on car collisions, society can adapt and improve current conditions and policies to lessen the risk of car accidents.

Weather is a major risk factor that could lead to vehicle collisions. As of 2016, adverse weather conditions at the time of crash accounted for an annual average of 16% of all car crash deaths (Saha et al, 2016). Drivers were found to get into more car accidents during weather conditions involving precipitation than during extreme heat conditions (Liu et al. 2017). The accident probability was approximately 5 times higher in negative temperature than in positive temperatures (Becker et al., 2020). Precipitation conditions include rain and snow depending on the season. Of the two variables, snow was found to have more impact on the likelihood of car crashes than rainfall (El-Basyounny et al, 2014). The number of vehicle collisions was positively correlated to increase in snowfall intensity.

However, some studies have concluded that less traffic accidents occur in presence of precipitation because of the change in traffic variation (Theofilatos, 2016). During rainfall or snow, it is speculated that drivers drive more carefully during wet conditions and there are less motorcyclists or pedestrians interrupting the traffic. In the D.C. area, commuters drive through varying weather conditions throughout the year therefore understanding the impact of weather on car crashes will be critical to commuters.

Another key risk factor that could potentially impact car crashes is time of day. A linear regression analysis done in New York City concluded that 4pm is the most likely time for a

driver to get into an accident in New York City (Hopping, 2019). 4pm is the start of New York City's rush hour and the start of the busiest and heavy traffic times. Similarly, another study based on traffic in Connecticut has predicted that most crashes occurred in the afternoon between 4pm and 5pm (Mondal et al., 2020). Based on these previous studies, afternoon rush hour times after work tend to be when car crashes occur the most. Car crashes that occur during commute from work to home were found to be one of the major causes of injury and deaths among drivers (Bener et al., 2017). The traffic congestion occurring at these hours could be positively related to an increase in frequency of car accidents (Retallack et al., 2020). Another possible explanation for mid-afternoon car accidents is post-lunch sleepiness, which led to slower reaction times for drivers (Hao et al., 2016). By determining the most dangerous time of day in D.C. traffic could help commuters adjust their work schedule to avoid times when car crashes are most likely to occur.

The Washington D.C. commuters must be prepared to drive in every weather condition as the season changes. Especially during typical rush hour times, when drivers are more prone to car accidents, drivers must be more alert and careful during the commute. Understanding the impact of these risk factors on driving conditions and likelihood of car crashes will allow drivers in D.C. to make informed decisions on when or how they will commute to work. Additionally, evaluating these risk factors will help policy makers implement safety measures to reduce possibility of car accidents as well as severity of accidents.

Data Sources and Descriptions

For our analysis car crash data was obtained from Open Data DC's "Crashes in DC" (hereafter, "Crash Data"), a free online repository of data gathered by the Washington, District of Columbia government. The Crash Data is maintained by the District Department of Transportation (DDOT). The DDOT processes crash records reported by the Metropolitan Police

Department's (MDP) crash data management system nightly. Each record indicates an individual crash reported by the MDP, and for the purpose of our analysis we considered each record as one crash. The Crash Data spans from 2015 to current time. Not all car crashes are recorded within Crash Data due to poor car crash coordinate reporting and poor address or roadway information. Data between years 2016 to 2019 was used for our analysis. The 2015 data was excluded because it contained inaccuracies from the original data migration from the historic records management system to the current database used by MDP. The 2020 data was not yet complete so we did not consider those records.

After restricting the year from 2016 to 2019, the Crash Data contains 245,136 records with no duplicate record found. There were 60 columns within the data, however, we only utilized the "REPORTDATE" column for our analysis, with the derived variables of Hour, Day, and Year. The 59 columns that were not used were mostly associated with geo-location and crash severity. Below are histograms demonstrating the distribution of the Hour, Day and Month variables:

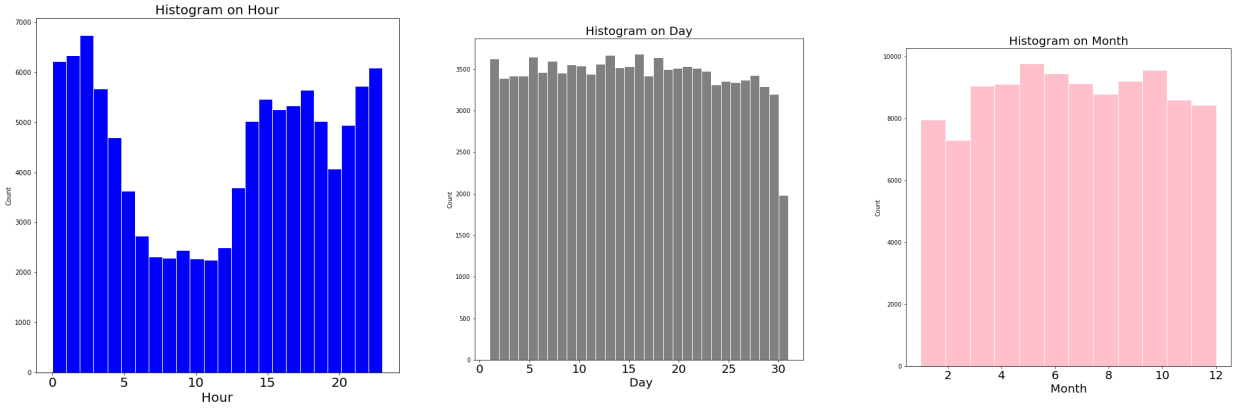


Figure 1 - Histograms of car crash distributions by Hour, Day and Month.

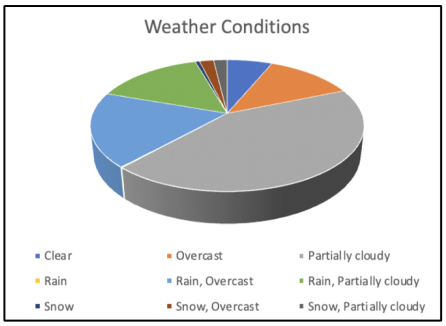
For the analysis, daily weather data was collected from the Visual Crossing website for the years between 2016 and 2019. In the initial dataset there were some duplicates, created on the end date of Daylight saving for each year. The duplicated records with minimum values were

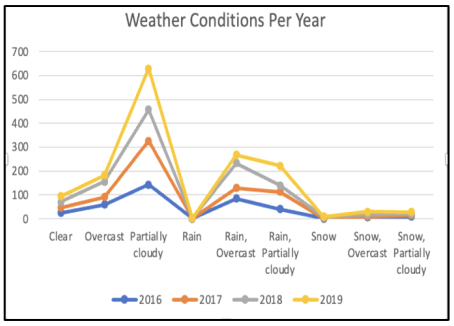
removed to match with crash data. After cleansing the weather data, there were 1461 unique records. There were 17 columns including maximum, minimum and daily average temperature (in Fahrenheit); wind speed and wind gust (in mph); precipitation, snow and snow depth (in inches); and weather conditions.

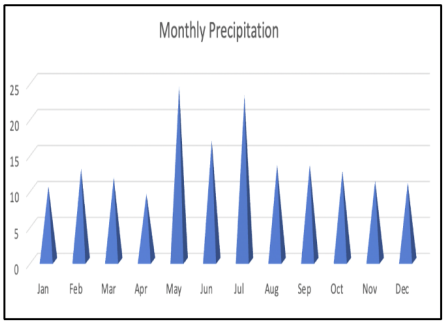
Date time	Maximum Temperature	Minimum Temperature	Temperature	Wind Chill	Heat Index	Precipitation	Snow	Snow Depth	Wind Speed	Wind Direction	Wind Gust	Visibility	Cloud Cover	Relative Humidity	Conditions
2015-01-01	45.8	26.5	36.1	19.1	0.0	0.00	0.00	0.00	14.5	211.75	29.5	9.9	39.2	45.25	Partially cloudy
2015-01-02	48.5	36.3	41.0	32.8	0.0	0.00	0.00	0.00	9.0	249.29	18.3	9.9	81.7	52.35	Overcast
2015-01-03	41.9	34.2	39.0	30.7	0.0	0.43	0.00	0.00	6.4	64.79	0.0	6.2	94.6	77.60	Rain, Overcast
2015-01-04	65.8	41.9	53.1	36.1	0.0	0.21	0.00	0.00	20.5	216.38	32.3	9.0	96.3	77.52	Rain, Overcast
2015-01-05	51.8	29.7	40.0	19.7	0.0	0.00	0.74	0.74	22.3	309.54	35.2	9.9	62.9	33.94	Snow, Partially cloudy

Figure 2 - Sample of weather data

Various studies show different weather parameters have significant impact on road accidents, but precipitation is considered to be the most important one. Thus, for our study we only utilized “Precipitation”, “Weather Conditions” along with the “Datetime” column. Note, data used here is daily average and does not indicate instant weather conditions.







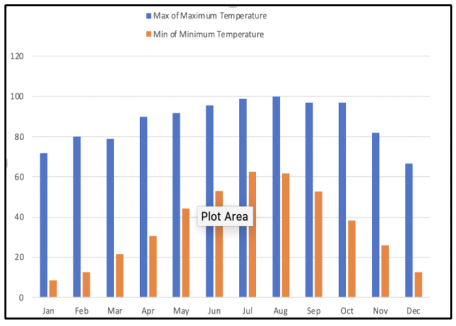


Figure 3 - Visualizations of weather data by weather conditions and monthly precipitation

Combining Data and Creating Aggregates

The Crash Data and Weather Data were merged on the date columns of each data set then grouped by Hour, Day and Month. The precipitation variable was originally a discrete value that we then converted into a binary indicator as to whether there was precipitation that day or not. The “CrashCount" and “Precipitation indicator” variables were aggregated by sum giving the total number of crashes and the total number of days there was precipitation for the grouped rows. The Hour and Month variables were converted into 24, and 12 dummy variables, respectively. The Day variable was converted into a weekday variable ranging from Monday to Sunday, and then converted into dummy variables for the 7 weekdays. The Resulting data set contained 8,756 rows and 53 columns. Below is a heatmap demonstrating the correlation of the final variables along with a histogram showing the distribution of the Crash Counts after grouping the data.

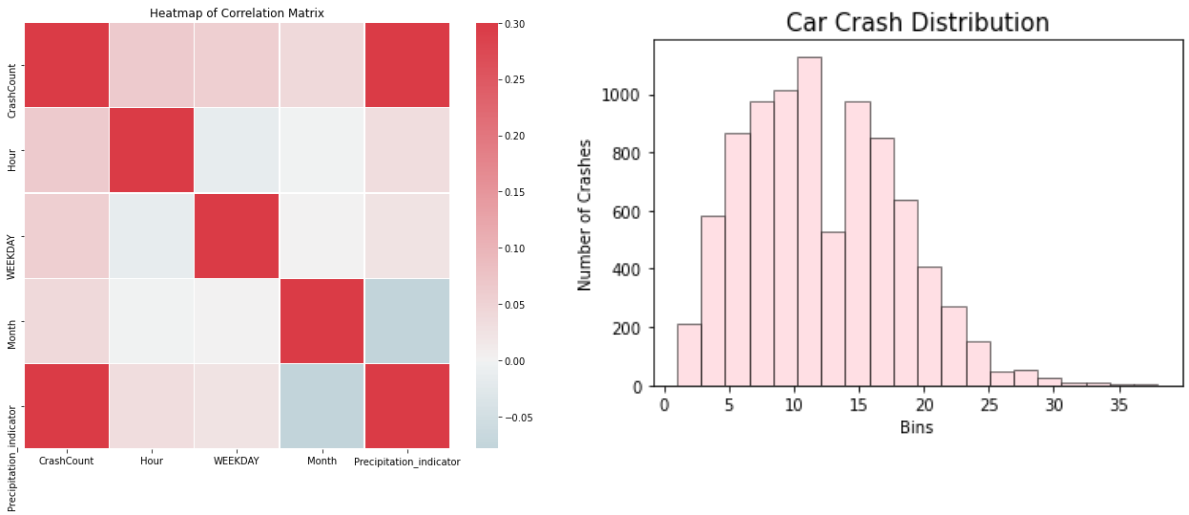


Figure 4 - Heatmap of final variables and bar chart of car crash distributions

Analysis and Results

Ordinary Least Square (OLS) regression model was used to understand the significance of each variable on car crashes with statsmodels. Initially, bivariate models were run to examine the individual R^2 values of Month, Hour and Precipitation separately. The R^2 value refers to how closely the data fits the regression line. Higher R^2 value indicates higher significance of the variable. Month and Hour had R^2 value of less than 0.005. The Precipitation had the highest R^2 of 0.080, which is still insignificant. The multivariate model with all three variables ran against the CrashCount data, but the model returned an R^2 value of 0.085, which does not indicate overall significance either.

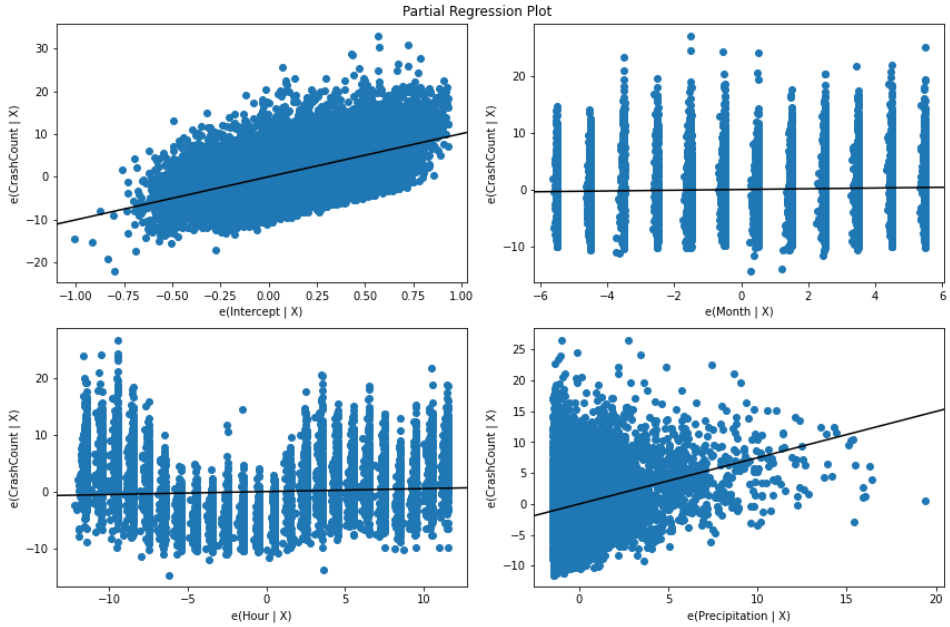


Figure 5 - Results of the Ordinary Least Square Analysis by Month, Hour and Precipitation

OLS regression requires a linear relationship between dependent and independent relationship, which is not the case for the variables in our analysis. The month of the year and hour of the day or level of precipitation does not always cause vehicle collisions. Therefore, we explored other models for our analysis.

Random Forest

Random Forest (RF) Regression is one of the widely used machine learning algorithms, which involves constructing a large number of trees from the training datasets. Prediction on RF Regression model is calculated from the average prediction across the decision trees. In the study, crash count is considered as a response variable whereas 69 other variables (precipitation, snow, 24 dummy variables for hour, 31 dummy variables for day and 12 dummy variables for month) are used as explanatory variables. Dataset was then split into 75% for training and 25% for testing the model. In order to get the best accuracy, we ran the model on a constant dataset for different values of 'n_estimators' i.e. number of trees used in the model and then observed the Mean absolute error and Accuracy.

Initially, the model experienced a steady decline in Mean Absolute error with increase in number of trees but after some point, a flat line was observed with maximum accuracy of ~60%.

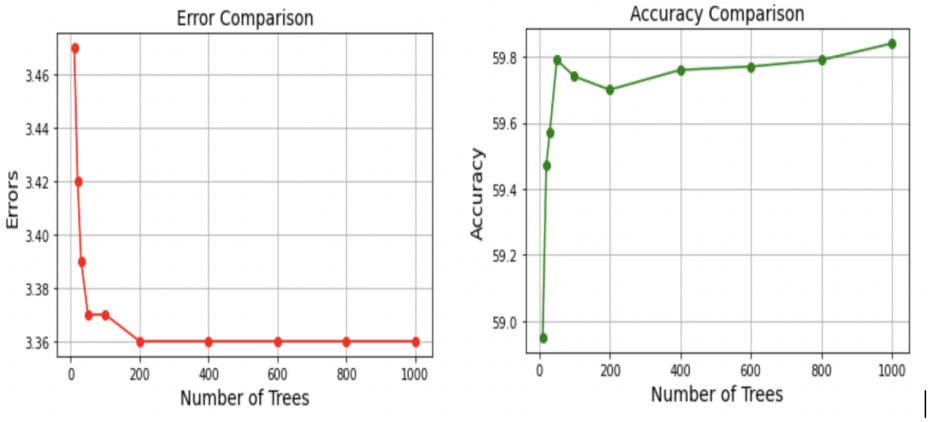
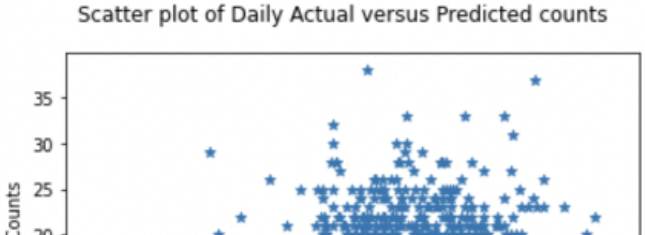


Figure 6 - Error and Accuracy Comparison for Random Forest Regression

Then the model was trained on the training dataset with the best possible value of 'n_estimators' and predictions were made on the testing dataset. The scatter plot chart shows the actual daily crash count versus predicted crash count by the model. Based on the RF regression, Precipitation was the most important feature in determining the car crashes.

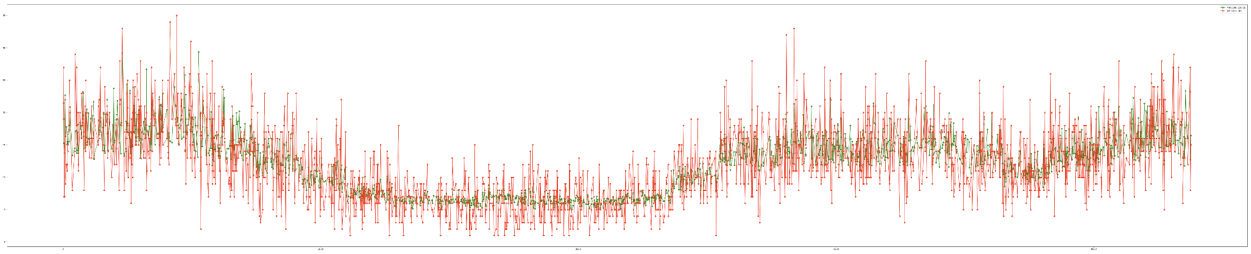


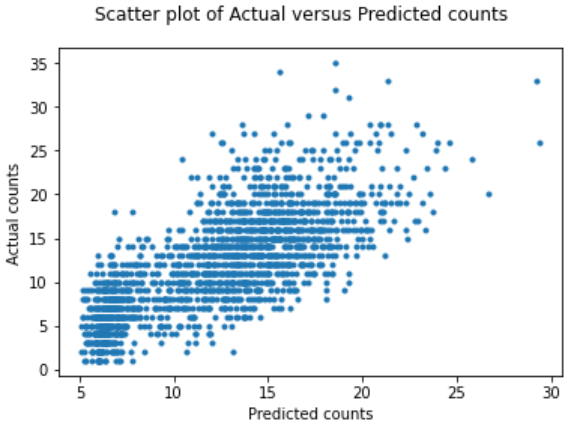
Poisson and Negative Binomial Regressions

We initially tried a Poisson regression to predict accident counts based on our designated explanatory variables. Poisson regressions are designed to model count data, however, the Poisson regression makes the assumption that the mean equals the variance. After an initial investigation we found that our data was over dispersed and thus did not follow a Poisson distribution. When count based data does not meet the criteria for a Poisson distribution, the Negative Binomial Regression (NB) model is generally recommended. The NB regression is considered a variant of the Poisson regression that includes additional parameters to account for over-dispersion.

To run the Poisson and NB regressions the data set was then split into 80% for training the model and 20% for testing. The aggregate number of crashes is the independent variable for our model. The dependent variables are identified as the 24 dummy variables for hour, 7 dummy variables for weekdays, 12 dummy variables for month, and the aggregate indicator variable for precipitation.

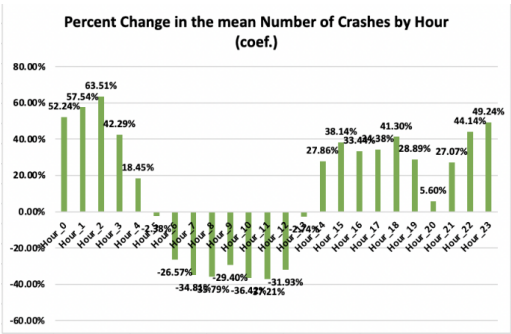
To initially analyze the results of the NB regression we looked at the fit of our predicted results versus our actual testing results to see if the model tracked the trend of the data. We used visuals of actual versus predicted counts and saw that the trend of the predicted data tracked the actual test data well. Below are the two graphs we used for the analysis. The first is a plot of actual (red) and predicted (green) with the number of crashes on the Y axis and an index of the data on the Y. The second graph is a scatterplot of actual vs predicted counts.

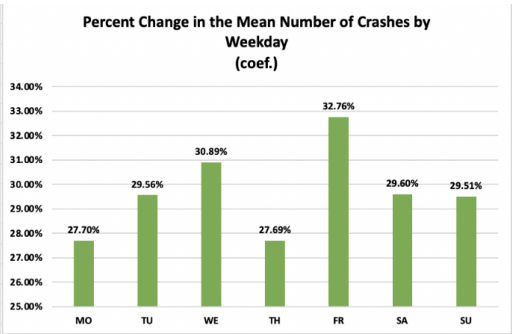


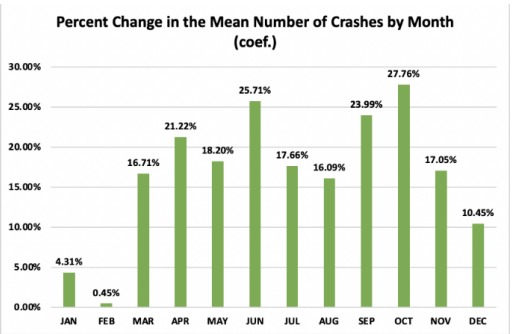


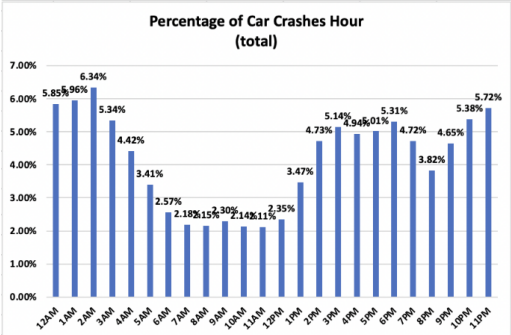
Additionally we computed a root mean square error value of 4 and compared it to the max value of 30, and minimum value of 5 within our predicted dataset.

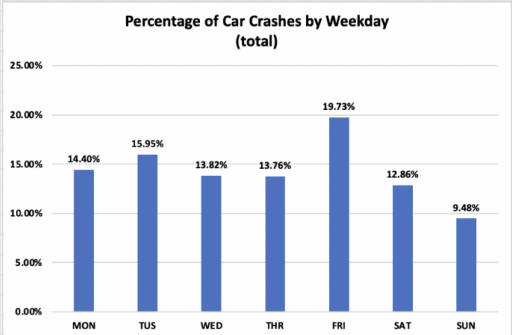
The coefficients for our variables were exponentiated multiplicative factors by which the mean count changes. We then computed the percent change that our coefficients implied. The percentage of the coefficient tells us by what percent increase that variable would impact the mean crash count by 1 unit. The resulting matrix of all the variables can be found after the conclusion. Below are plots for each variable grouping (Hour, Weekday and Month) for the predicted impact each coefficient had on the mean crash count in our model, along with a corresponding graph of the percentage of the total number of crashes each variable represented in the whole dataset.

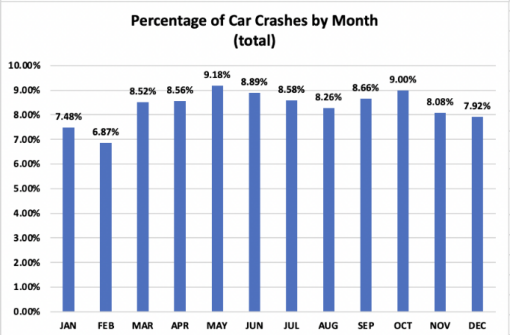












The variables found to not be statistically significant based on a Wald-type test were Hours 6 and 14, and March. They were likely found to not be statistically significant because they were closely associated with the mean.

Conclusion

We analyzed 106,422 crashes over 4 years (2016-2019). Our project intended to find a relationship between time, crash data and weather data. We created a model that predicts actual crashes to a reasonable degree of accuracy. We found that the most likely day to crash on was Friday, the most likely month was October, and the most likely hour was 2am. Conversely, the least likely day was Thursday, the least likely month was February, and the least likely hour was 11am. We would need to perform further analysis with additional explanatory factors, such as D.C. bar hours, driver's alcohol consumption level and trends in commuting patterns throughout the year to definitively explain our results.

The main application for our model we considered was as a tool to assist in commuting. We anticipated, with the ever increasing ability for people to telework, that a tool using this model or some improvement upon it could be used to determine days in which crashes are more likely to occur. Ideally, this would allow people to pick the most desirable days to choose not to commute, and even possibly have the effect of reducing the likelihood of crashes in the first place.

Ways in which our model was limited:

- Incorporation of volume data. We attempted to use data publicly available on https://opendata.dc.gov/. However, it only had average annual daily traffic (AADT) for all areas of DC, rather than a count by date, making it challenging to find a sensible purpose to merge it with our other data. As a result, we decided not to use the volume

data, but we recognize that it would be a useful addition if data was available in a more granular (at least weekly, if not daily) fashion.

- Granularity of weather data. We only had access to weather data by day. With more granular data in regards to weather data, our model could be improved to better match reality.
- Incorporation of GIS location data. GIS location data could potentially be integrated into the model we created to improve its efficacy. Additionally, it could be used in an implementation, to show potential “hotspots” where a crash is predicted more likely to occur, potentially allowing commuters to take other routes. As a longer term focus, determining the type of areas where frequent hot spots for crashes occur could assist in future city planning, or redesigns of what already exists.

Variables	Coefficent	exp(coef.)	Percent change	P-Score	Total Number of Crashes	Percentage of Total Number of Crashes
Hour_0	0.420	1.522	52.24%	0.000	6222	5.85%
Hour_1	0.455	1.575	57.54%	0.000	6338	5.96%
Hour_2	0.492	1.635	63.51%	0.000	6743	6.34%
Hour_3	0.353	1.423	42.29%	0.000	5681	5.34%
Hour_4	0.169	1.184	18.45%	0.000	4701	4.42%
Hour_5	-0.024	0.976	-2.38%	0.000	3627	3.41%
Hour_6	-0.309	0.734	-26.57%	0.198	2733	2.57%
Hour_7	-0.428	0.652	-34.81%	0.000	2320	2.18%
Hour_8	-0.443	0.642	-35.79%	0.000	2293	2.15%
Hour_9	-0.348	0.706	-29.40%	0.000	2446	2.30%
Hour_10	-0.453	0.636	-36.42%	0.000	2276	2.14%
Hour_11	-0.465	0.628	-37.21%	0.000	2250	2.11%
Hour_12	-0.385	0.681	-31.93%	0.000	2504	2.35%
Hour_13	-0.028	0.973	-2.74%	0.000	3694	3.47%
Hour_14	0.246	1.279	27.86%	0.122	5030	4.73%
Hour_15	0.323	1.381	38.14%	0.000	5469	5.14%
Hour_16	0.289	1.334	33.44%	0.000	5259	4.94%
Hour_17	0.296	1.344	34.38%	0.000	5336	5.01%
Hour_18	0.346	1.413	41.30%	0.000	5651	5.31%
Hour_19	0.254	1.289	28.89%	0.000	5020	4.72%
Hour_20	0.055	1.056	5.60%	0.000	4069	3.82%
Hour_21	0.240	1.271	27.07%	0.002	4950	4.65%
Hour_22	0.366	1.441	44.14%	0.000	5723	5.38%
Hour_23	0.400	1.492	49.24%	0.000	6087	5.72%
Variables	Coefficent	exp(coef.)	Percent change	P-Score	Total Number of Crashes	Percentage of Total Number of Crashes
Precipitation	0.027	1.027	2.73%	0.000		0.00%
Variables	Coefficent	exp(coef.)	Percent change	P-Score	Total Number of Crashes	Percentage of Total Number of Crashes
MO	0.245	1.277	27.70%	0.000	15321	14.40%
TU	0.259	1.296	29.56%	0.000	16975	15.95%
WE	0.269	1.309	30.89%	0.000	14707	13.82%
TH	0.244	1.277	27.69%	0.000	14644	13.76%
FR	0.283	1.328	32.76%	0.000	20999	19.73%
SA	0.259	1.296	29.60%	0.000	13683	12.86%
SU	0.259	1.295	29.51%	0.000	10093	9.48%
Variables	Coefficent	exp(coef.)	Percent change	P-Score	Total Number of Crashes	Percentage of Total Number of Crashes
JAN	0.042	1.043	4.31%	0.000	7957	7.48%
FEB	0.005	1.005	0.45%	0.000	7311	6.87%
MAR	0.155	1.167	16.71%	0.720	9066	8.52%
APR	0.192	1.212	21.22%	0.000	9115	8.56%
MAY	0.167	1.182	18.20%	0.000	9774	9.18%
JUN	0.229	1.257	25.71%	0.000	9458	8.89%
JUL	0.163	1.177	17.66%	0.000	9133	8.58%
AUG	0.149	1.161	16.09%	0.000	8793	8.26%
SEP	0.215	1.240	23.99%	0.000	9214	8.66%
OCT	0.245	1.278	27.76%	0.000	9573	9.00%
NOV	0.157	1.170	17.05%	0.000	8598	8.08%
DEC	0.099	1.105	10.45%	0.000	8430	7.92%

References

Becker, N., Rust, H. W., & Ulbrich, U. (2020, October 29). *Predictive modeling of hourly probabilities for weather-related road accidents*. Natural Hazards and Earth System Sciences. https://nhess.copernicus.org/articles/20/2857/2020/.

Bener, A., Lajunen, T., Özkan, T., Yildirim, E., & Jadaan, K. S. (2017). T*he impact of aggressive behaviour, sleeping, and fatigue on road traffic crashes as comparison between minibus/van/pick-up and commercial taxi drivers*. https://www.researchgate.net/profile/Abdulbari-Bener/publication/315595379_The_Impact_of_Aggressive_Behaviour_Sleeping_and_Fatigue_on_Road_Traffic_Crashes_as_Comparison_between_MinibusVanPick-up_and_Commercial_Taxi_Drivers/links/58d8b025aca2727e5e06e646/The-Impact-of-Aggressive-Behaviour-Sleeping-and-Fatigue-on-Road-Traffic-Crashes-as-Comparison-between-Minibus-Van-Pick-up-and-Commercial-Taxi-Drivers.pdf.

Berkon, E. (2020, January 24). D.C. has some of the longest commutes in the COUNTRY. what help is available? Retrieved April 03, 2021, from https://www.npr.org/local/305/2020/01/24/799292338/d-c-has-some-of-the-longest-commutes-in-the-country-what-help-is-available.

El-Basyouny, K., Barua, S., & Islam, M. T. (2014, September 3). *Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models*. Accident Analysis & Prevention. https://www.sciencedirect.com/science/article/pii/S0001457514002516?casa_token=6xj_

-e_y-TQAAAAA%3AoZBBJdYufVIgSOBhCaq-Tv-IqFt4omUAmPRr34t9S-dqTg7LmcEnMFTSZGmX6_nfwbn47z1hfQ.

Hao, W., Kamga, C., & Wan, D. (2016, January 14). *The effect of time of day on driver's injury severity at highway-rail grade crossings in the United States*. Journal of Traffic and Transportation Engineering (English Edition). https://www.sciencedirect.com/science/article/pii/S2095756415200104.

Hopping, G. (2019, December 9). *Traffic accidents in New York city - a linear Regression study*. https://towardsdatascience.com/traffic-accidents-in-new-york-city-a-linear-regression-study-3af7159ef088.

Lee, M. L., Howard, M. E., Horrey, W. J., Liang, Y., Anderson, C., Shreeve, M. S., … Czeisler, C. A. (2016, January 5). *High risk of near-crash driving events following night-shift work*. PNAS. https://www.pnas.org/content/113/1/176.full.

Liu, A., Soneja, S. I., Jiang, C., Huang, C., Kerns, T., Beck, K., … Sapkota, A. (2016, December 15). *Frequency of extreme weather events and increased risk of motor vehicle collision in Maryland*. Science of The Total Environment. https://www.sciencedirect.com/science/article/pii/S0048969716326791?casa_token=rGYBxDahW28AAAAA%3AhT7VeMFzjPPm2kWfaSva9m8PdVKmxRwLHzJ-YY03bEik3CXRBeEI1lVaQux9vsC8m6kRUYBgVw.

Malin, F., Norros, I., & Innamaa, S. (2018, October 29). *Accident risk of road and weather conditions on different road types*. Accident Analysis & Prevention. https://www.sciencedirect.com/science/article/pii/S0001457518308455.

Mondal, A. R., Bhuiyan, M. A. E., & Yang, F. (2020, July 14). *Advancement of weather-related crash prediction model using nonparametric machine learning algorithms*. SN Applied Sciences. https://link.springer.com/article/10.1007/s42452-020-03196-x.

National Capital Region Transportation Planning Board. (2019, September 17). *2019 STATE OF THE COMMUTE SURVEY Technical Survey Report.* https://www.mwcog.org/file.aspx?&A=1AAuS26tuk0qvTVF52Q7%2BD87l582VWw4yNkHhrI8JrM%3D.

Parker, K., Horowitz, J., & Minkin, R. (2021, February 09). How coronavirus has changed the way Americans work. Retrieved April 03, 2021, from https://www.pewresearch.org/social-trends/2020/12/09/how-the-coronavirus-outbreak-has-and-hasnt-changed-the-way-americans-work/.

Retallack, A. E., & Ostendorf, B. (2020, February 21). *Relationship Between Traffic Volume and Accident Frequency at Intersections*. International journal of environmental research and public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7068508/.

Saha, S., Schramm, P., Nolan, A., & Hess, J. (2016, November 8). *Adverse weather conditions and fatal motor vehicle crashes in the United States, 1994-2012*. Environmental Health. https://ehjournal.biomedcentral.com/articles/10.1186/s12940-016-0189-x.

Theofilatos, A. (2017, March 2). *Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials*. Journal of Safety Research. https://www.sciencedirect.com/science/article/pii/S0022437517301378#s0060.

Authors' Contributions

Joanne Choi

- Data discovery
- Literature review research
- Analysis
    - Ordinary Least Square Regression
- Contribution to the paper
    - Introduction
    - Literature review
    - Analysis
- Combined all notebooks (EDA, data merging, various analyses) into one notebook for submission
- Presentation preparation - contributed to the deck and rehearsed for the presentation

Sam Clark

- Data discovery
- Literature review research
- Merged and created final dataset
- EDA
    - Crash dataset
- Analysis
    - Poisson regression
    - Negative binomial regression
- Contribution to the paper
    - Crash dataset
    - Methods
    - Analysis
- Presentation preparation - contributed to the deck and rehearsed for the presentation

Ranjan Jaiswal

- Data discovery
- Literature review research

- EDA
    - Weather dataset
- Analysis
    - Random Forest
- Contribution to the paper
    - Weather dataset
    - Analysis
- Presentation preparation - contributed to the deck and rehearsed for the presentation

Peter Kirk

- Data discovery
- Literature review research
- EDA
    - Traffic volume dataset
- Contribution to the paper
    - Abstract
    - Conclusion
- Presentation preparation - contributed to the deck and rehearsed for the presentation

Sachin Jayaraman

- Data discovery
- Literature review research
- Project presentation prep