Exploratory Data Analysis of 2003 - 2015 Car Crash Data from Monroe County, Indiana

By: Marta Baker, Kenny Berry, Zach Dallas, and Ivan Valverde

Background

We used a car crash dataset from Kaggle. The dataset contained data from Monroe County, Indiana for the years 2003 to 2015. Monroe County, Indiana is home to University of Indiana-Bloomington. We chose this dataset because of its size and inclusion of latitude and longitude that allowed us to make maps. This dataset helped us better understand the type of injuries coming from crashes and which months had the most crashes. We believe this type data could be useful for various public institutions including departments of transportation and both city and state governments

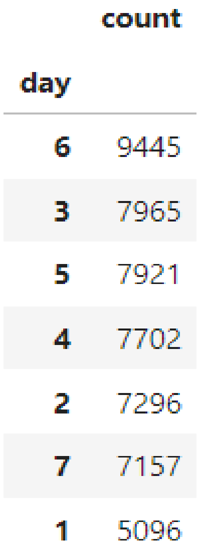
Data Cleaning

The dataset initially had 53,943 rows, each representing a crash, and eleven columns/variables. We first removed 1,361 null values. The “Primary Factor” column contained the main reported reason each crash occurred. It originally contained 52 different values, but we narrowed it down to 23. To do this, we consolidated several similar values into larger ones. For example, we created a single “Vehicle Defect” value to replace several car related values including, but not limited to, Brake Failure, Tow Hitch Failure, and Other Lights Failure.

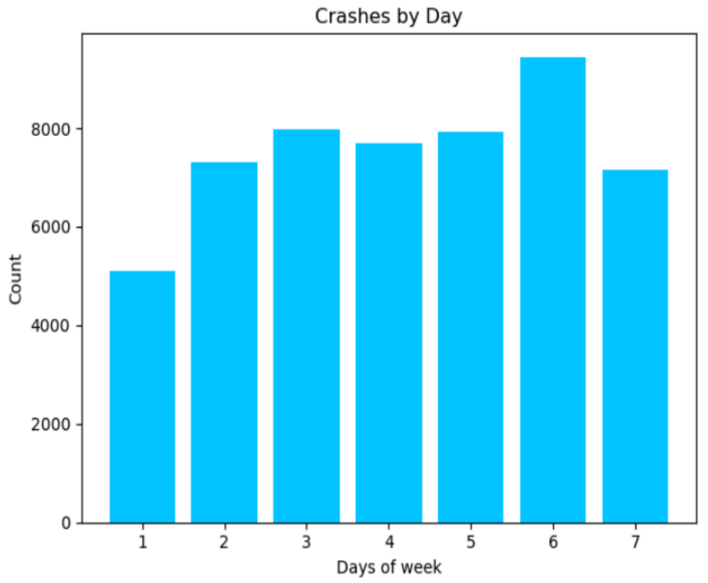
Question 1: What days of the week are people more likely to get into accidents? Does this change based on month?

We started by making a leaderboard containing the “day” column’s value counts to show how many accidents happened on each day. In the dataset, 1 represented Sunday and 7 represented Saturday with the remaining days of the week being sequential. Using the leaderboard shown below, we were able to create the following bar chart showing that most crashes happened on Fridays.

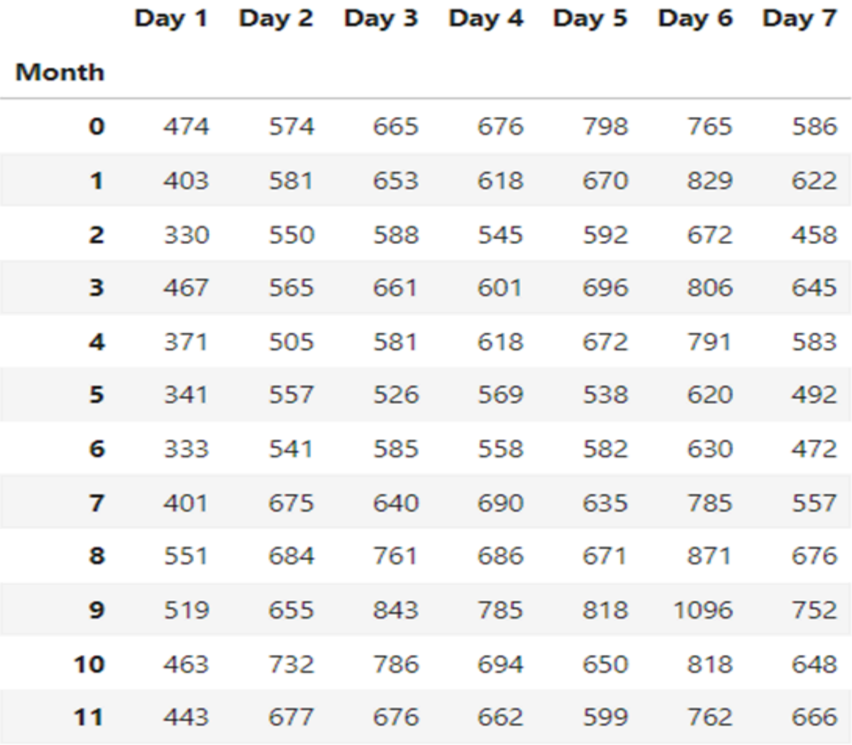
Crashes by Day charts:



	count
day	
6	9445
3	7965
5	7921
4	7702
2	7296
7	7157
1	5096



To determine if these patterns changed by month, we located the day value in the cleaned data frame, then grouped it by month. Next, we extracted the values and placed them into a dataframe which produced the below leaderboard.



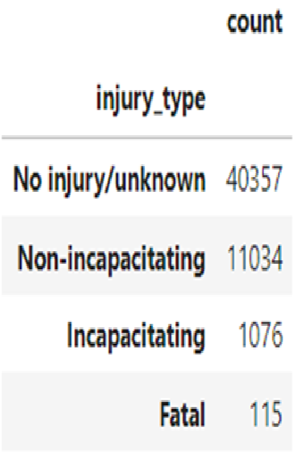
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Month							
0	474	574	665	676	798	765	586
1	403	581	653	618	670	829	622
2	330	550	588	545	592	672	458
3	467	565	661	601	696	806	645
4	371	505	581	618	672	791	583
5	341	557	526	569	538	620	492
6	333	541	585	558	582	630	472
7	401	675	640	690	635	785	557
8	551	684	761	686	671	871	676
9	519	655	843	785	818	1096	752
10	463	732	786	694	650	818	648
11	443	677	676	662	599	762	666

Friday remains the most common day to be in a crash regardless of the month. However, Fridays in October had the most crashes out of the entire dataset with a total of 1096 crashes.

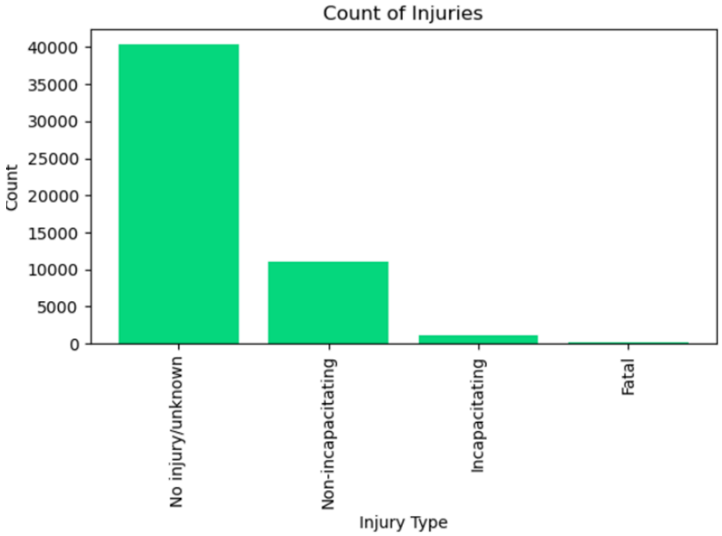
Question 2: What type of injuries are most likely to occur? Are more severe injuries likely to occur in one car accidents or multi-car accidents?

In order to find out what type of injuries were most likely to occur, we used the same logic that we used to answer question one. Instead of grouping by “day” for the value counts, we grouped by “injury type” to generate both our leaderboard and bar chart. From this data we created the below charts.

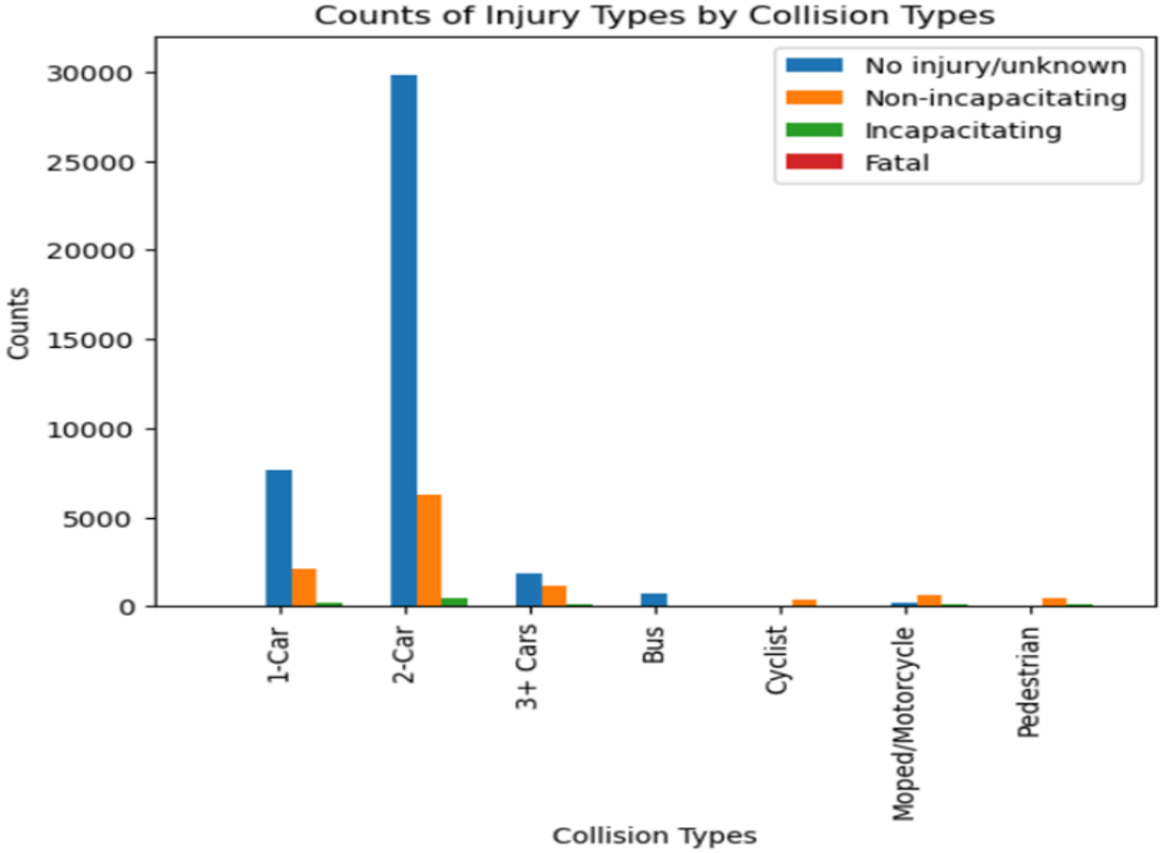
Injury type by crashes



	count
injury_type	
No injury/unknown	40357
Non-incapacitating	11034
Incapacitating	1076
Fatal	115

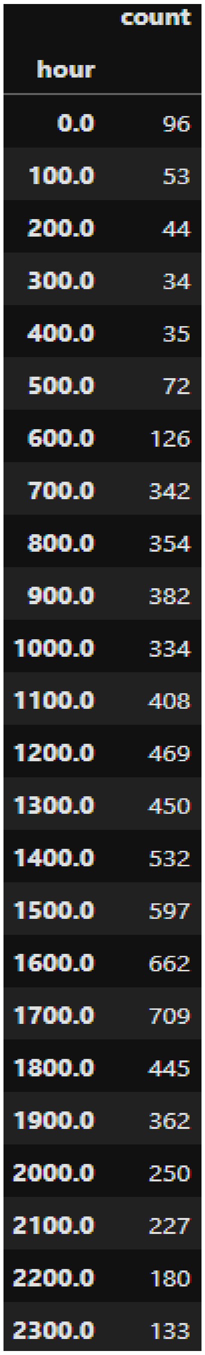


Then, we grouped by collision type and took the values count to determine what types of injuries occurred per collision types. With the help from the Xpert tool, we created the following graph.



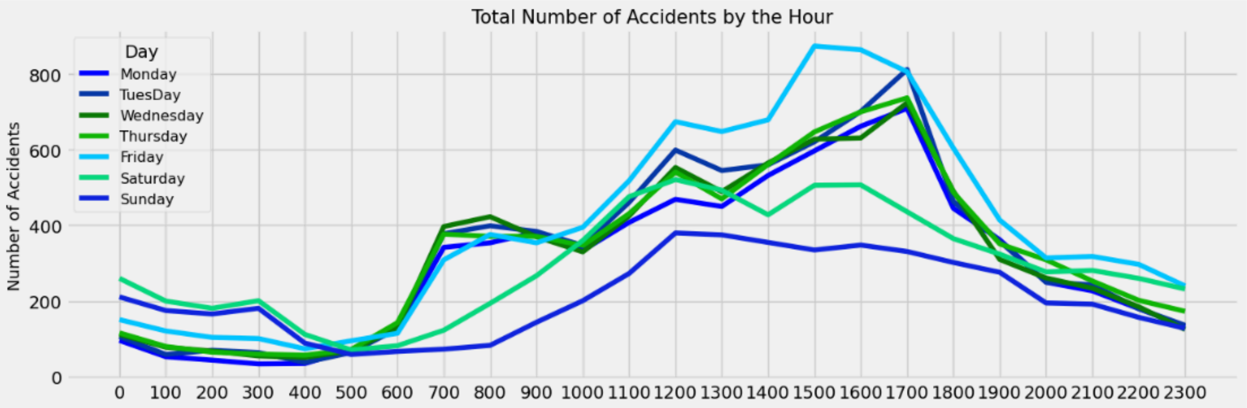
Question 3: What time of day are people most likely to get into accidents?

In order to create the line chart that would best visualize crashes over time on each of the week, we had to create 7 tables for each day of the week. These tables include the value counts of crashes from each hour of day for each day. The following table is one example of the tables that were created:



	count
hour	
0.0	96
100.0	53
200.0	44
300.0	34
400.0	35
500.0	72
600.0	126
700.0	342
800.0	354
900.0	382
1000.0	334
1100.0	408
1200.0	469
1300.0	450
1400.0	532
1500.0	597
1600.0	662
1700.0	709
1800.0	445
1900.0	362
2000.0	250
2100.0	227
2200.0	180
2300.0	133

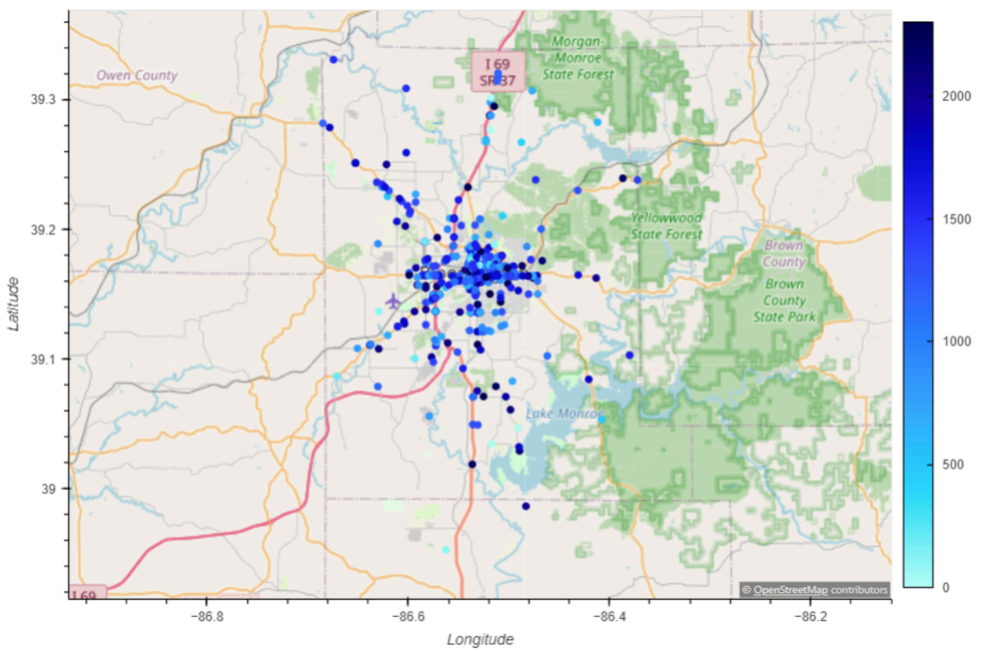
Once all the tables were created, we plotted all the tables in one line graph to make the information easier to read as shown below.



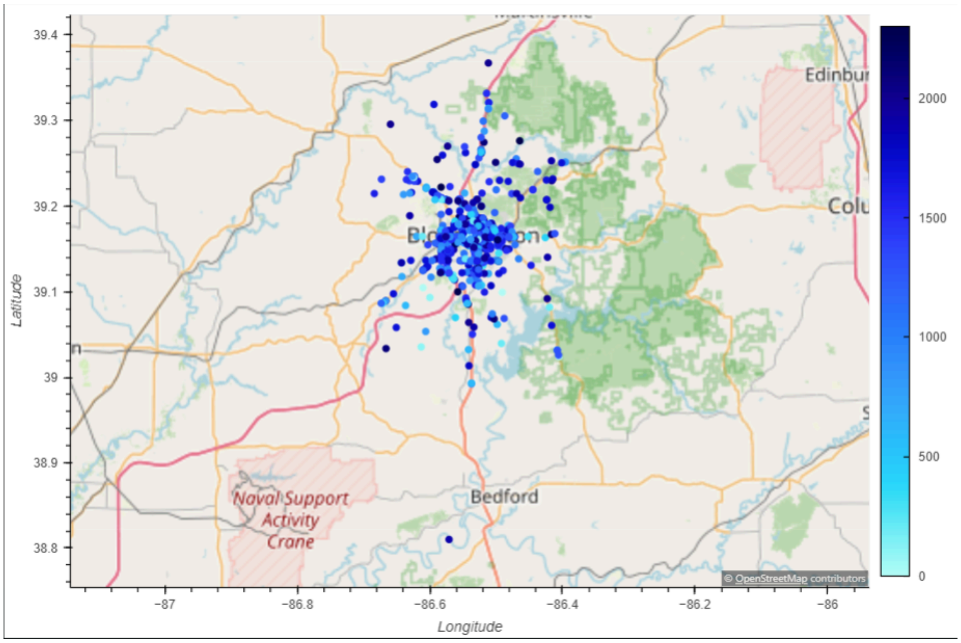
Maps: Where are accidents most likely to occur

Our final question was to determine where accidents were most likely to occur within the county. To create the maps, we generated 2 random samples of crashes in Monroe County, Indiana. Prior to generating the maps with hvplot, we first had to remove any data that had incorrect values. For example, some early data had missing Latitude and Longitude values that were input as 0.0 in the DataFrame. Once this was removed, we were able to generate the following maps.

Sample of 500 crash locations:



Sample of 1000 crash locations:



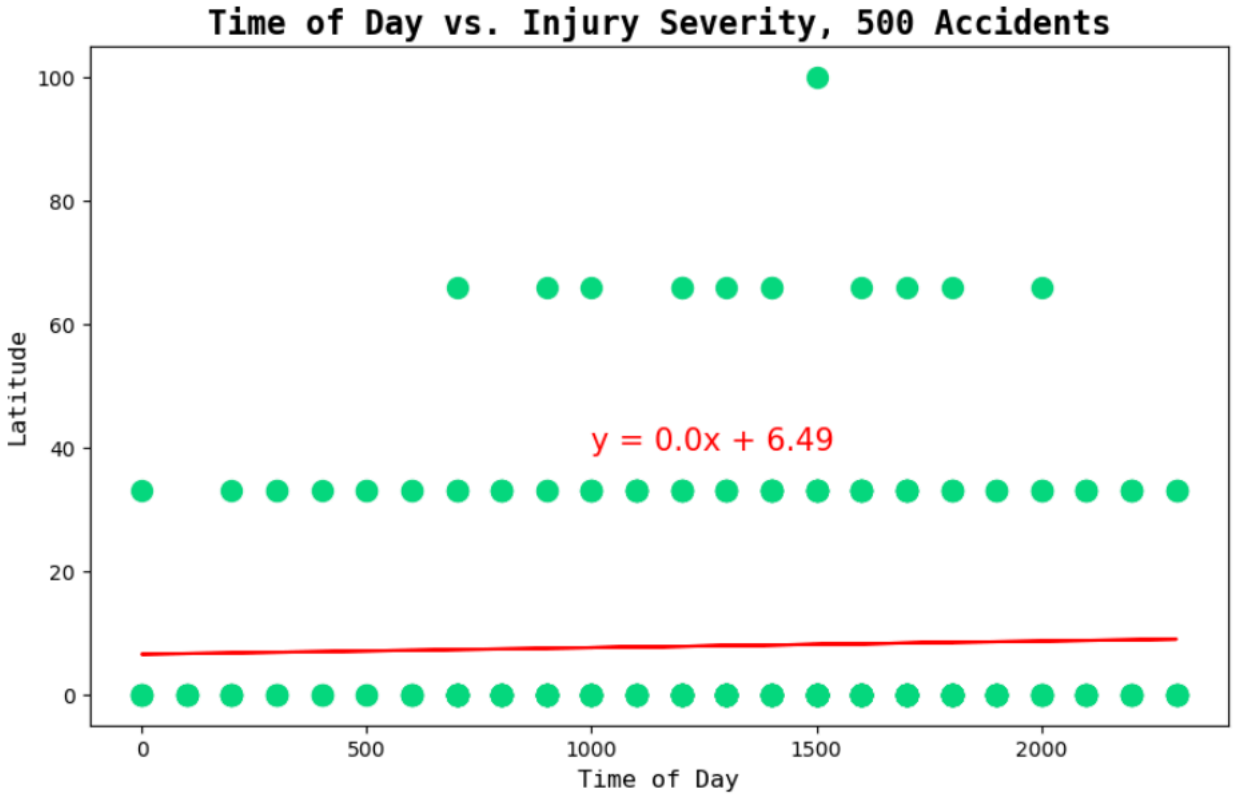
As would be expected, most accidents look to occur in the city of Bloomington and the surrounding throughways. As people move outside of the city center, accidents look to be less frequent.

Regression:

Time of Day vs. Injury Severity Regression

Given the nature of the data, there weren't many linear relationships. One correlation we examined was whether there was a relationship between the time of day an accident occurred and the severity of any injuries from the accident. This was accomplished by assigning different numbers to the categorical injury data provided. We assigned values as follows: 0 for "No injury/unknown", 33 for "Non-incapacitating" injuries, 66 for "Incapacitating" injuries, and 100 for "Fatal" accidents. This scale was decided upon because "No injury/unknown" and "Fatal" injuries represented the two extremes of injury types. The last two injury types, "Non-incapacitating" and "Incapacitating" didn't seem to be strongly favoring either "No injury/unknown" or "Fatal." As such, they were given values roughly 1/3 and 2/3 between "No injury/unknown" and "Fatal".

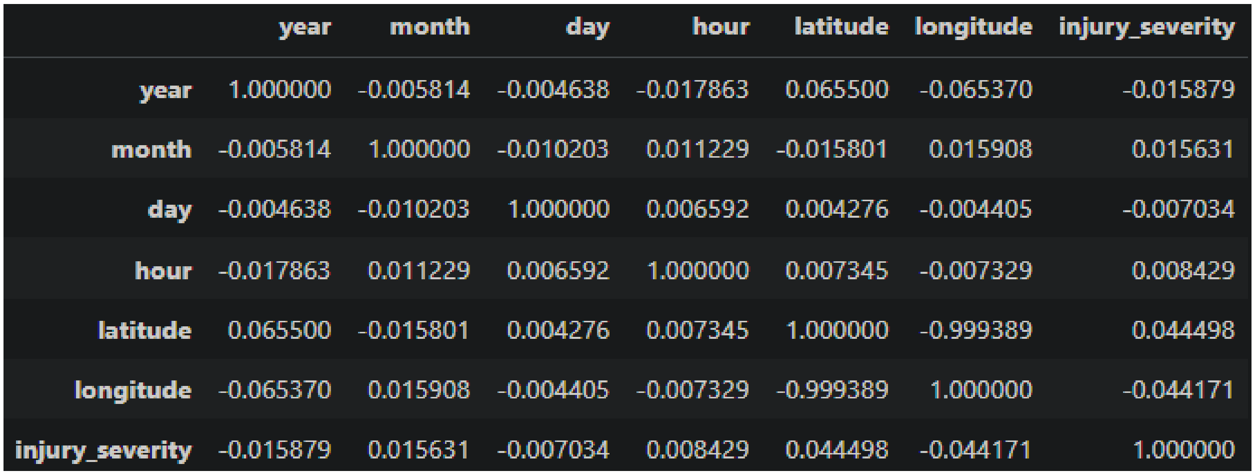
However, when a sample of 500 accidents were plotted, time of day an accident occurred and injury severity were independent of one another. The r2 value was 0.0011 which is effectively 0. The regression plot can be seen below.



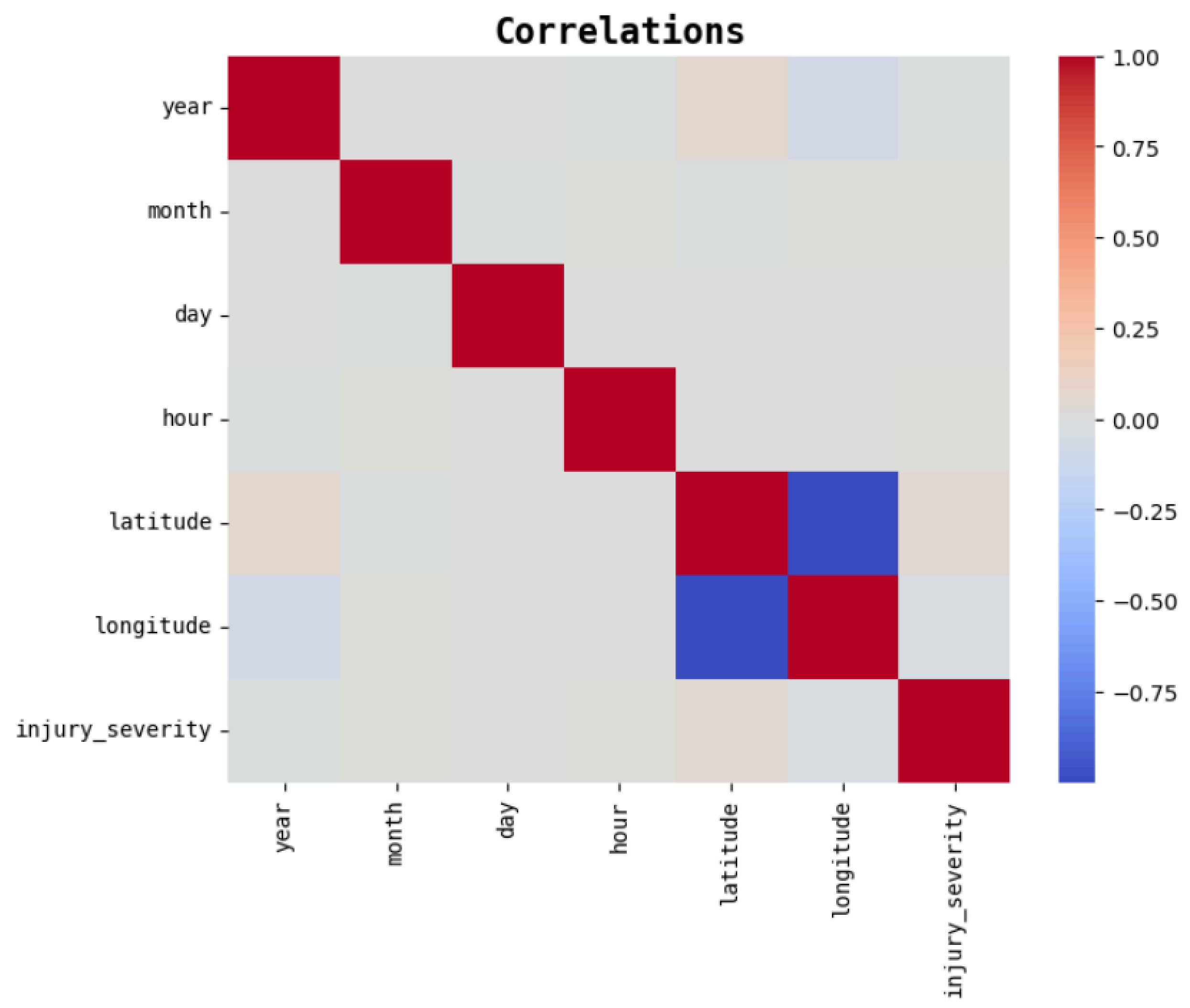
Correlation between all data

After we got the low r2 value, we decided to create both a leaderboard and heatmap to demonstrate the correlation between the various data to show the low correlations between the various numeric columns.

.

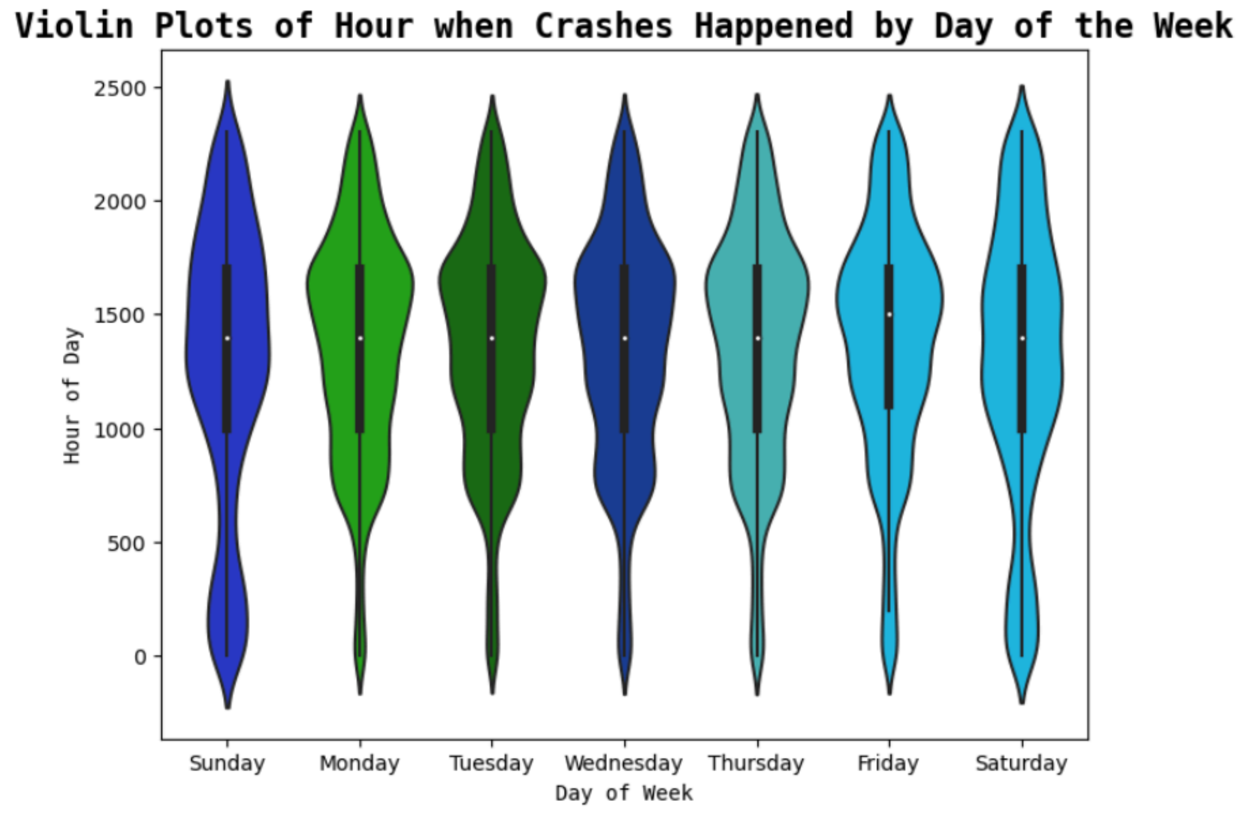


	year	month	day	hour	latitude	longitude	injury_severity
year	1.000000	-0.005814	-0.004638	-0.017863	0.065500	-0.065370	-0.015879
month	-0.005814	1.000000	-0.010203	0.011229	-0.015801	0.015908	0.015631
day	-0.004638	-0.010203	1.000000	0.006592	0.004276	-0.004405	-0.007034
hour	-0.017863	0.011229	0.006592	1.000000	0.007345	-0.007329	0.008429
latitude	0.065500	-0.015801	0.004276	0.007345	1.000000	-0.999389	0.044498
longitude	-0.065370	0.015908	-0.004405	-0.007329	-0.999389	1.000000	-0.044171
injury_severity	-0.015879	0.015631	-0.007034	0.008429	0.044498	-0.044171	1.000000



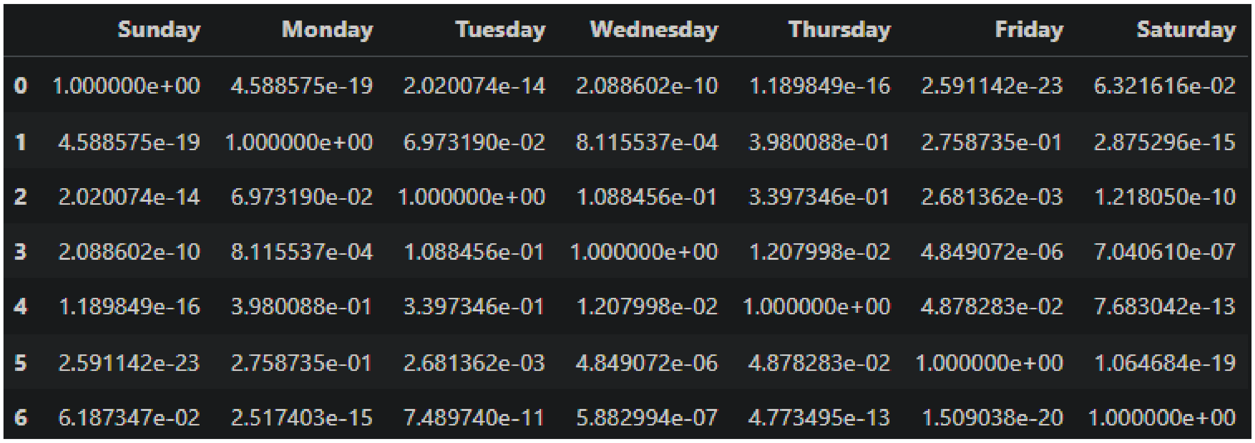
Statistical Tests

After we discovered there weren't any strong relationships between the data we were provided, we decided to see if there were any relationships between the time of day an accident occurred and the day of the week accidents occurred. Our first step to explore this relationship was to create a violin plot showing the distribution of accidents per day of the week by the time they occurred.

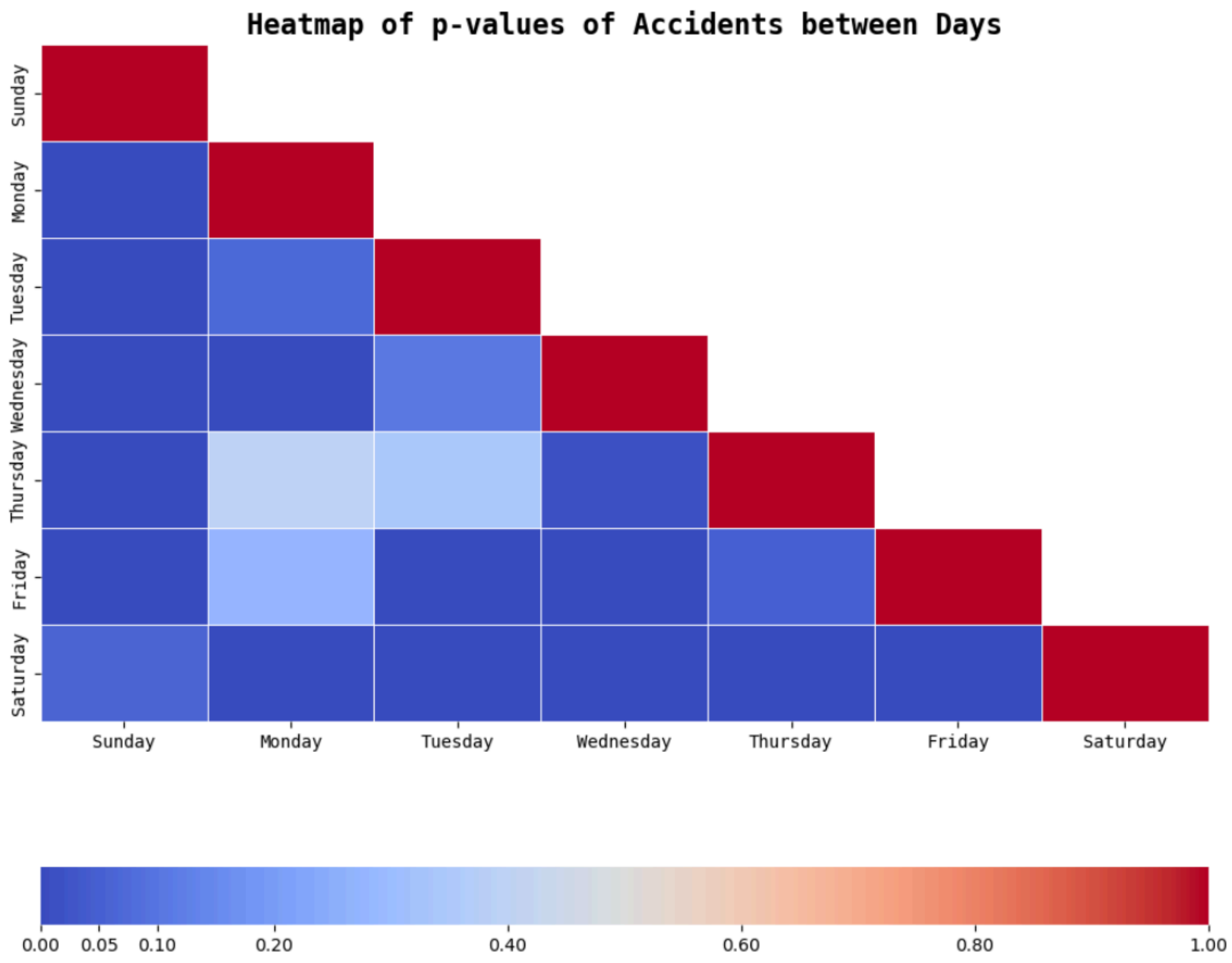


Looking at the violin plots, while the means look relatively similar (with the exception of Friday), the distribution of accidents definitely differs. For example, accidents on Sunday and Saturday look to be bimodal with a mode around 2 or 3 am and another towards the middle of the day. In comparison, the modes of accidents for Monday through Friday appear to be unimodal around 4 pm. Because of this, we ran an ANOVA to determine if there were any significant differences in accidents between any days. The ANOVA test returned a p-value of 1.87e-42 indicating that there were significant differences between any of the days.

After running the ANOVA, we created a DataFrame to show the p-values for all of the days compared to one another. This was accomplished by extracting the p-value from the SciPy Stats independent t-test formula. The resulting DataFrame and heatmap can be seen below.



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
0	1.000000e+00	4.588575e-19	2.020074e-14	2.088602e-10	1.189849e-16	2.591142e-23	6.321616e-02
1	4.588575e-19	1.000000e+00	6.973190e-02	8.115537e-04	3.980088e-01	2.758735e-01	2.875296e-15
2	2.020074e-14	6.973190e-02	1.000000e+00	1.088456e-01	3.397346e-01	2.681362e-03	1.218050e-10
3	2.088602e-10	8.115537e-04	1.088456e-01	1.000000e+00	1.207998e-02	4.849072e-06	7.040610e-07
4	1.189849e-16	3.980088e-01	3.397346e-01	1.207998e-02	1.000000e+00	4.878283e-02	7.683042e-13
5	2.591142e-23	2.758735e-01	2.681362e-03	4.849072e-06	4.878283e-02	1.000000e+00	1.064684e-19
6	6.187347e-02	2.517403e-15	7.489740e-11	5.882994e-07	4.773495e-13	1.509038e-20	1.000000e+00



Looking at both the DataFrame and Heatmap, there are some values that look like they could be considered significant. For example, the p-value for Sunday and Saturday is 0.062. This value is quite close to 0.05 and under normal circumstances likely would be considered a significant difference in accidents between the two days. However, a number of p-values are much smaller (i.e. 4.85 e-6, 2.02 e-14, etc.). As a result, these values are considered to not be significant due to the comparative difference in p-values.

Random Forest Regressor:

Given the lack of linear relationships between many of our variables, we decided to see if we could use a Random Forest Regressor to see if we could predict what day an accident would occur. To prepare the data for the Random Forest Regressor, we one-hot encoded the collision_type, injury_type, and primary_factor columns. We excluded the reported_location column from the one-hot encoding because the column was redundant to the latitude and longitude and because there were so many different reported locations that it would have bloated the model. Finally, we removed the reported location column, and deleted any rows with incorrect latitude and longitude values (i.e. latitude that was under 10 and positive longitude).

The Random Forest we created to predict the day an accident occurred was not accurate. Using a test size of 30% and with 100 estimators, our random forest's mean squared error is 3.78. As a result, the model currently is effectively guessing what day an accident is occurring. Simply decreasing the test size or increasing the estimators decreases the mean squared error, but they don't decrease it below 3.6. As a result, the model can be refined, however it likely would require a lot of fine-tuning to begin to accurately predict the day an accident would occur.

Conclusions:

We were able to determine that Fridays during afternoon rush hour had the most crashes both generally and across every month across the 13 year timespan of the reported accidents. Additionally, afternoon rush hours have the most crashes across any day while early morning hours while people are sleeping have the fewest. Most crashes appear to be fairly minor 2-car collisions with no injuries or no known injuries. Finally, while there were significant differences in the time accidents occurred across days, there weren't many linear relationships between the numeric columns contained within the data set.

Bias, Limitations, and Future Work:

The data had some biases including that the crashes were only reported accidents (either via police reports or insurance claims). However, the data likely doesn't include minor accidents, and 1-car collisions might be underreported if the driver never filed an insurance claim for the damage. Other collision types including car v. cyclist or car v. pedestrian might not have been reported either to the police or insurance if there were no injuries. Additionally, with the county containing a lot of college students, those individuals might be less likely to report accidents.

We had a number of limitations with the dataset. For example, the injury type data was limited to 4 rather broad categories. Additionally, the data lacked a lot of potentially influential variables including: age and gender of involved drivers, type of vehicle (including make and model), color of vehicle, etc. Another limitation of the data is that the data is almost 10 years out of date. With the pandemic, there has been a lot of anecdotal evidence indicating that accidents are more likely or that drivers are more aggressive post-pandemic. While that is anecdotal, due to the age of the data, the dataset prevents those claims from being examined further.

Given more time we would have liked to examine trends in the primary factor of accidents to see if there was any trend between the primary factor and either the collision type or the injury severity. We would also like to compare this dataset with a newer dataset for the same county that includes the pandemic years and the years following the pandemic. Additionally, it would be interesting to see if we could include a column in the data for "distance from home" for the accidents for the involved drivers. While providing the home addresses for involved drivers could be considered a privacy issue, checking how far people were from their home or temporary residence when they crashed would be useful to see if people were more likely to get into accidents closer to home.

For the regression model, while the days are provided numerically, a different model would likely have been a better fit for the data. Instead of using a Random Forest Regressor, we would try to make a predictive model with a Random Forest Classifier. While we would need to make a number of similar adjustments as were needed with the Random Forest Regressor, the Random Forest Classifier would probably provide a more accurate estimation of what day an accident would have occurred.

Sources:

Harold, Tosin. "Enhancing Correlation Matrix Heatmap Plots with p-values in Python." *Medium*, 23 Jan. 2020, tosinharold.medium.com/enhancing-correlation-matrix-heatmap-plots-with-p-values-in-python-41bac6a7fd77.

Jackson, Divakarr. "Car Crash Dataset." *Kaggle*, Kaggle, www.kaggle.com/datasets/jacksondivakarr/car-crash-dataset/data?select=monroe+county+car+crach+2003-2015.csv.

Stack Overflow. "LabelEncoder vs OneHot Encoding in Random Forest Regressor." *Stack Overflow*, 14 Jan. 2021, stackoverflow.com/questions/65749305/labelencoder-vs-onehot-encoding-in-random-forest-regressor.