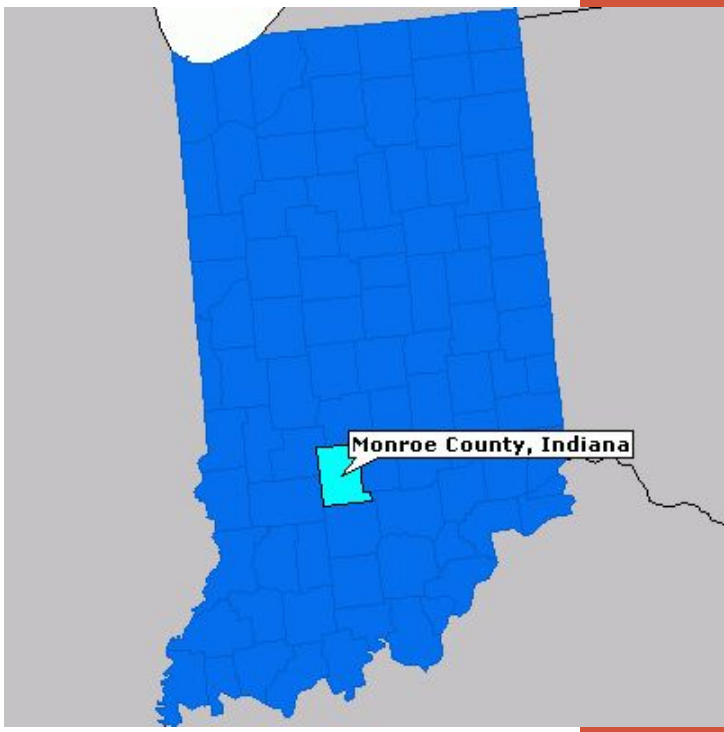
Car Crashes in Monroe County 2003-2015

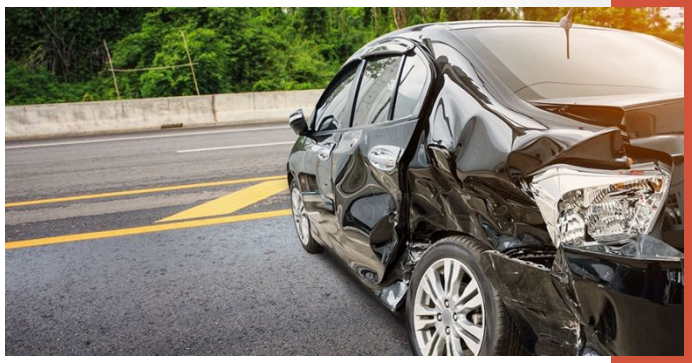
Marta Baker
Kenny Berry
Zach Dallas
Ivan Valverde

Monroe County, Indiana

- Population: 139,718 (2020)
- County seat: Bloomington
- Home to Indiana University Bloomington

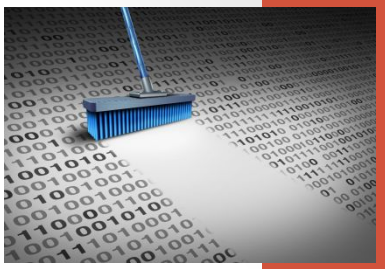


Data Set Details



- Twelve-year period (2003-2015)
- Eleven columns/variables including:
  - Accident location
  - Year, month, day of the week, and time of accident
  - Primary cause of accident
  - Type of injury sustained
- 53,943 cases

Data Set Cleaning



- Removed 1,361 null values
- Narrowed values in Accident Factor variable from 52 to 23
  - Examples:

Obstruction Not Marked
Roadway Surface Condition
Hole/Ruts in Surface
Shoulder Defective

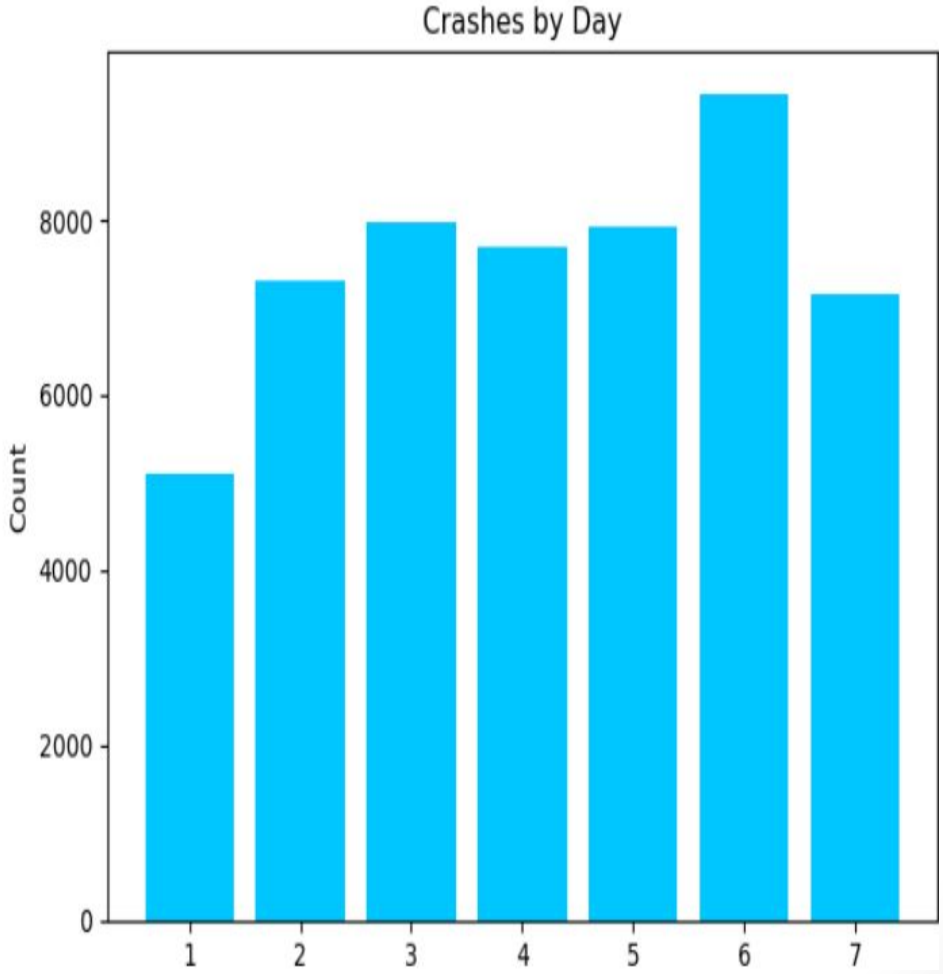
→ Road Conditions

Steering Failure
Brake Failure
Tow Hitch Failure
Other Lights Failure

→ Vehicle Defect

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

- TGIF!!
- Drive safe!

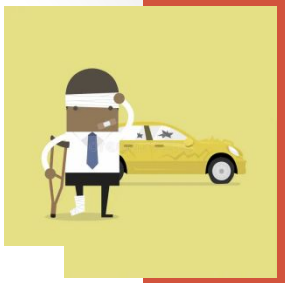


Question 1: What days of the week are accidents more likely? Does this change based on month?

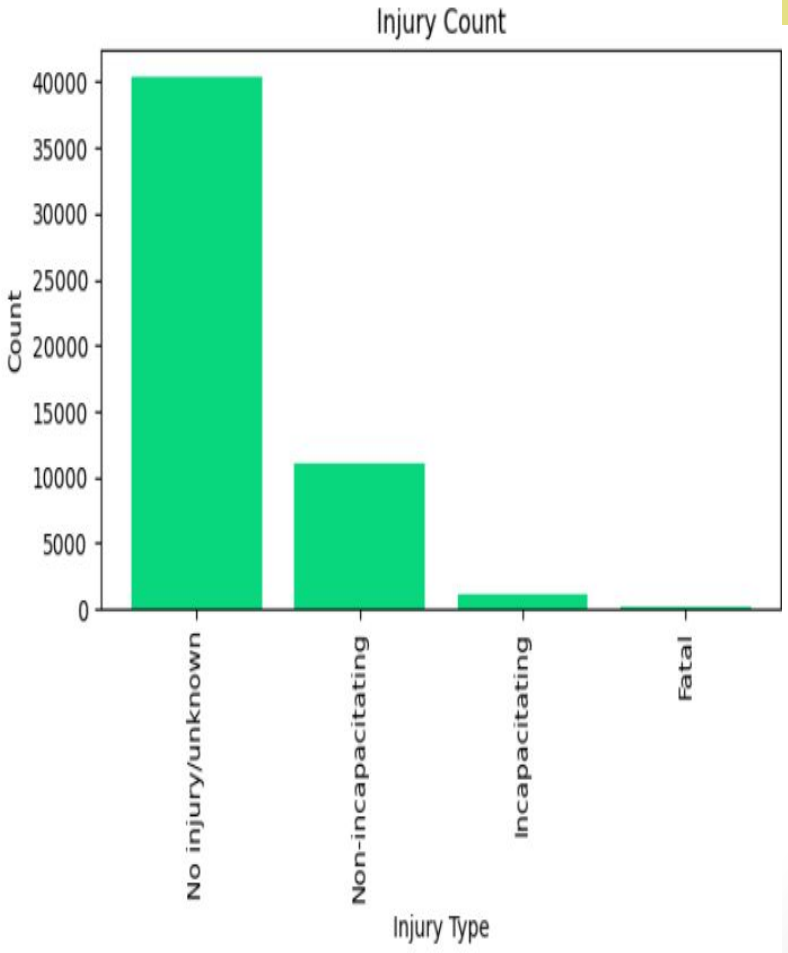
- Friday remains the most popular day to be in a crash no matter the month.
- Friday's in October hold the highest amount of crashes out of entire dataset totaling 1096.

Month	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
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

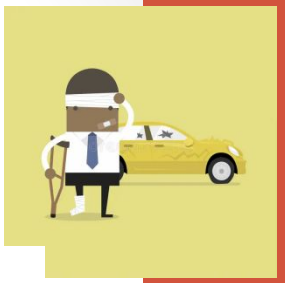
Question 2: What type of injuries are most common? Are injuries more common in single or multi-car collisions?



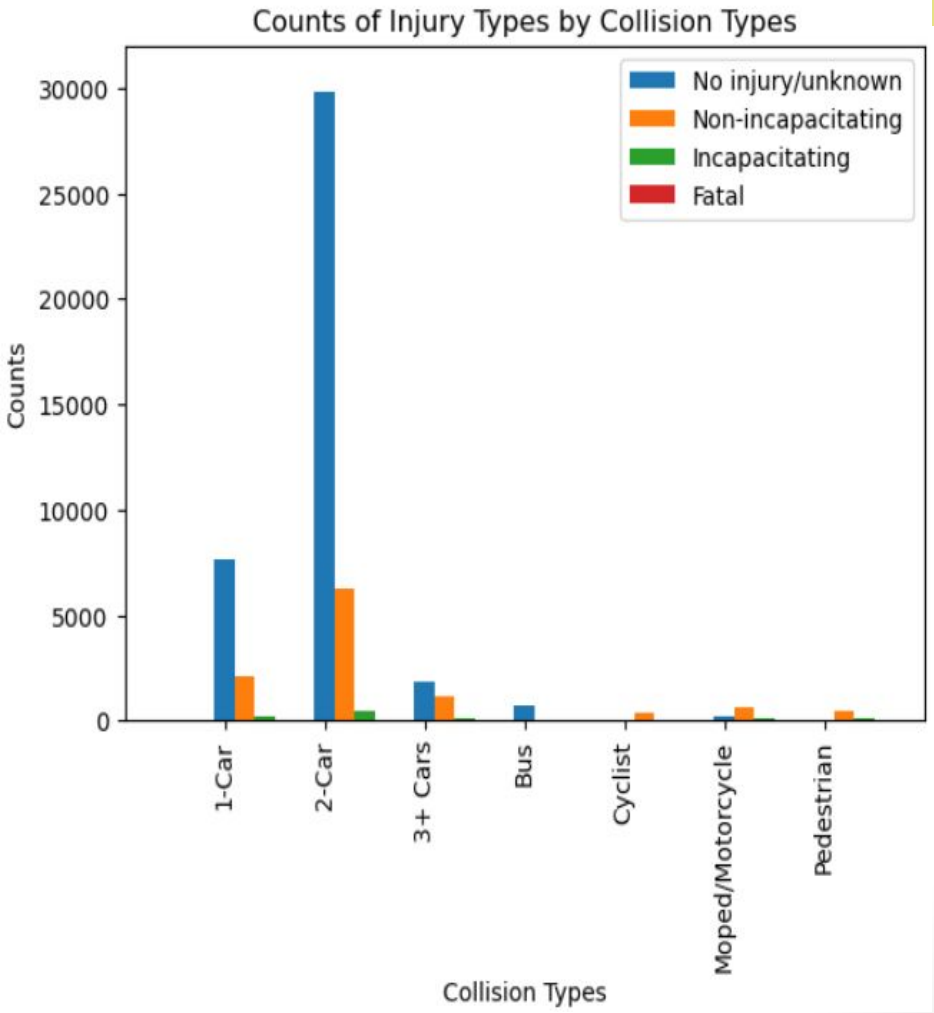
- Good news is, majority of crashes come out as No Injury/unknown.



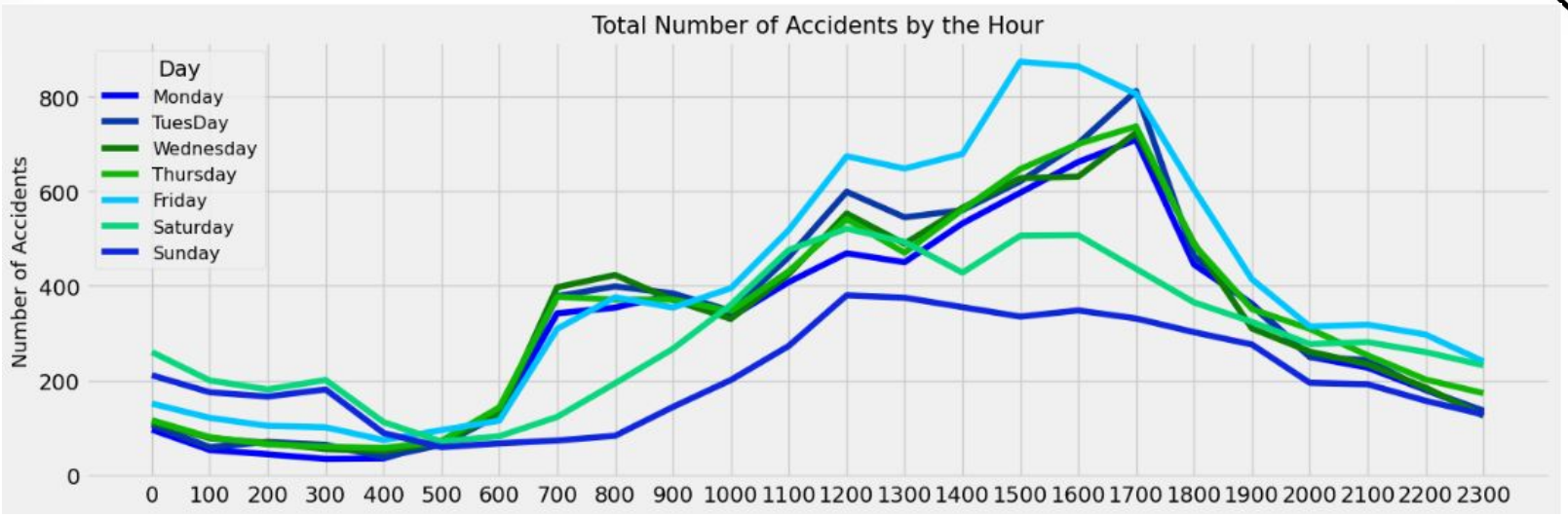
Question 2: What type of injuries are most common? Are injuries more common in single or multi-car collisions?



- 2-car collisions are by far the most common type of collisions.
- More people were incapacitated during a 2-car collision, however 1-car collisions held a higher fatality rate.



Question 3: What time of day are accidents most common?

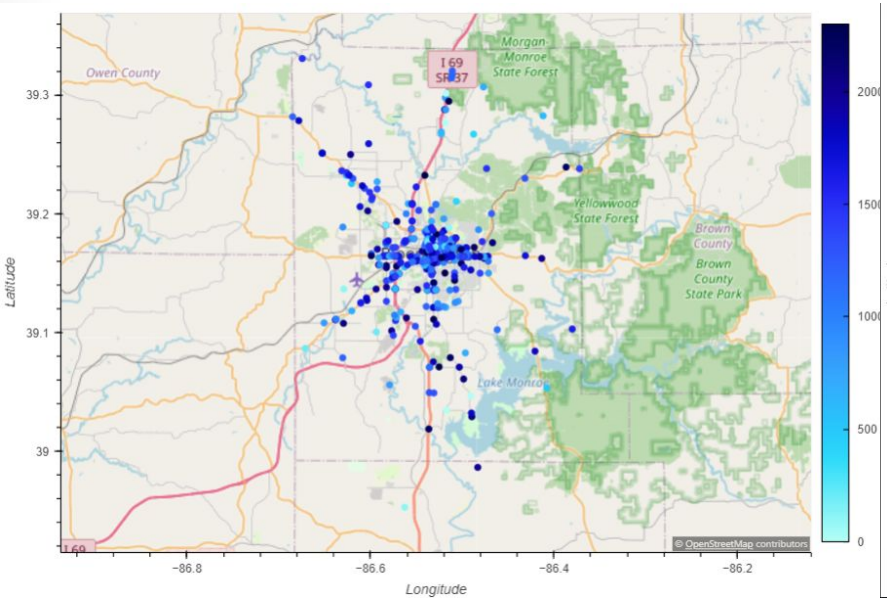


- Friday has a higher trend of accidents as well as a higher peak
- Sunday has a more shallow curve showing more consistency throughout the day
- Tuesday seems to have an above average peak for a weekday
- Saturday seems to trend low even though it seems to be a high traffic day

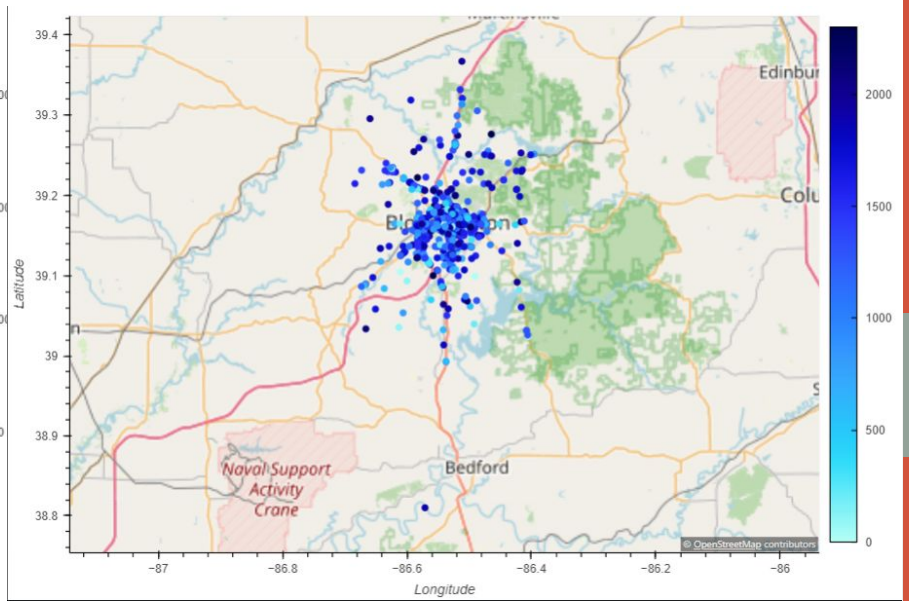
Accident locations within Monroe County

- The crashes are colored by time of day
- Accidents more likely to occur within the city and along major throughways
- Looking at colors, accidents appear to happen more between roughly 10 am and 4 pm

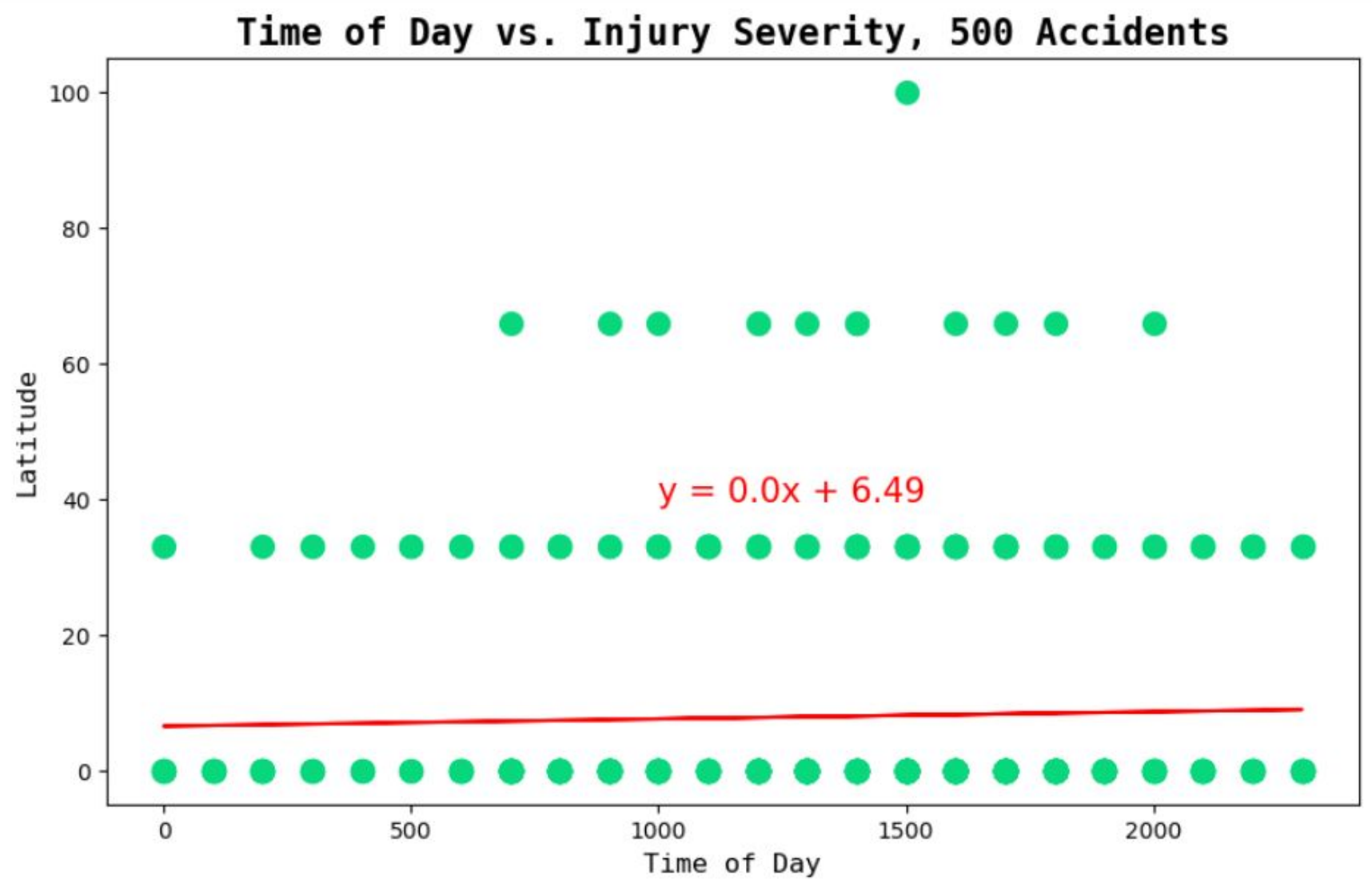
Five hundred-crash random selection



One Thousand-crash random selection

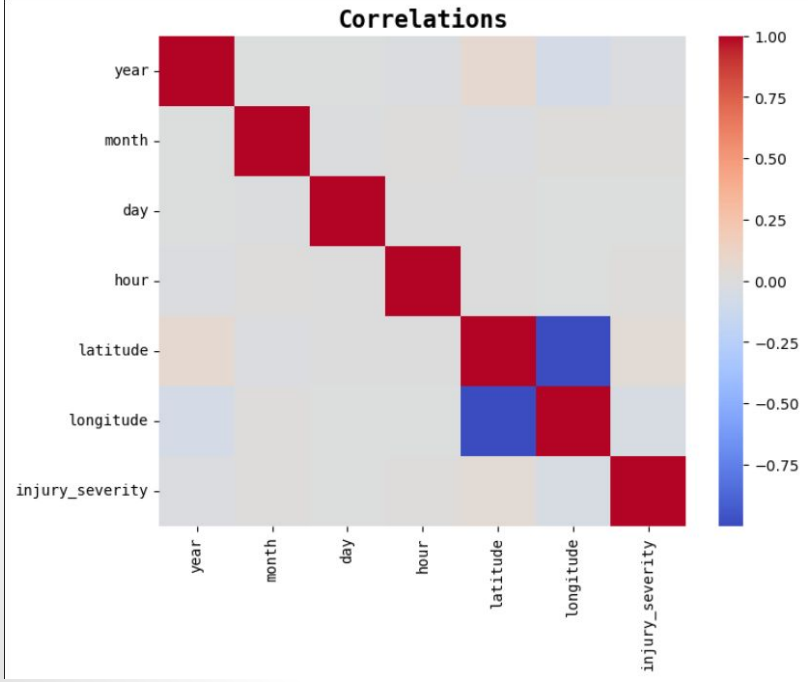


Time- injury severity relation



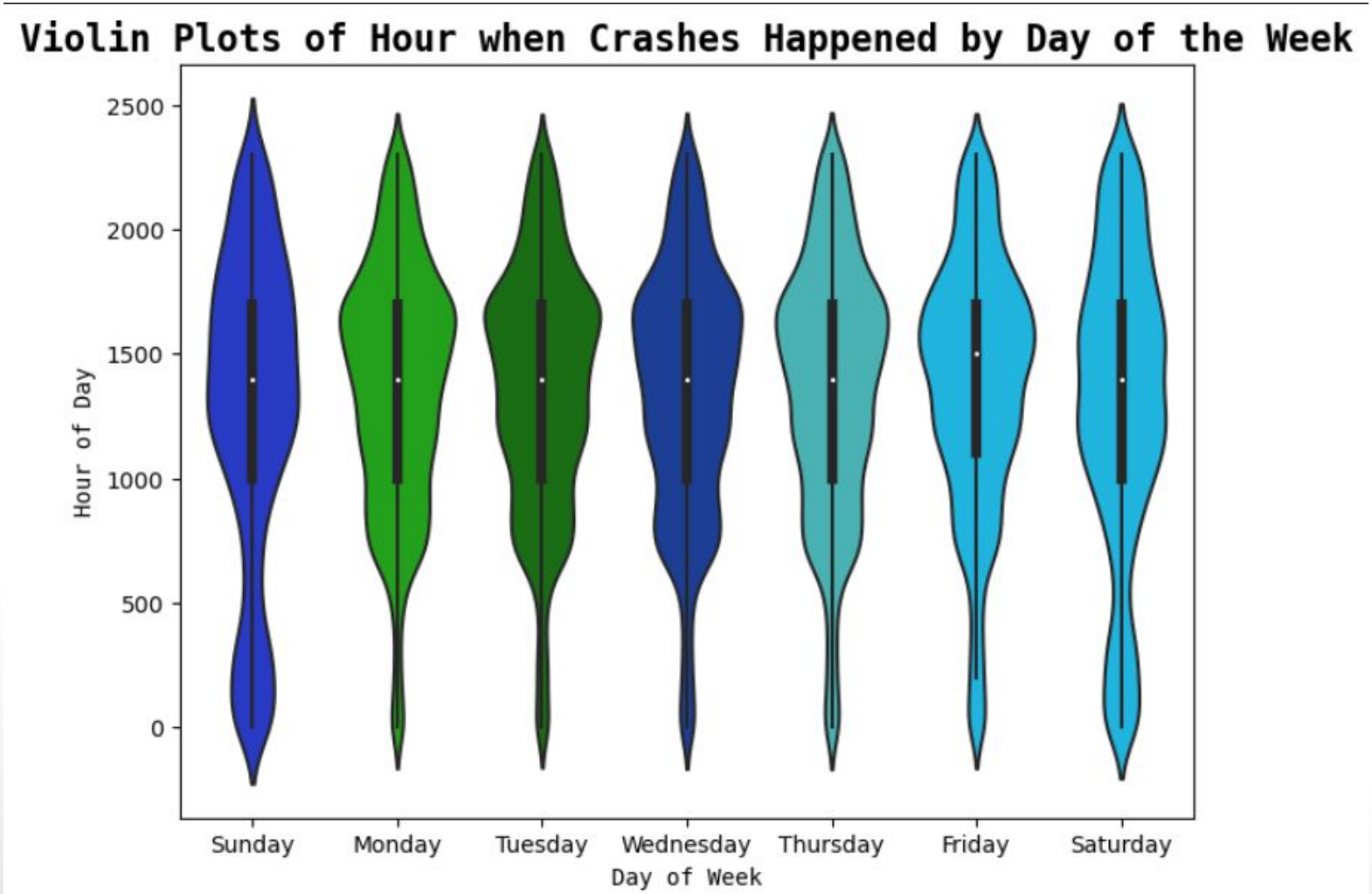
Correlations

	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



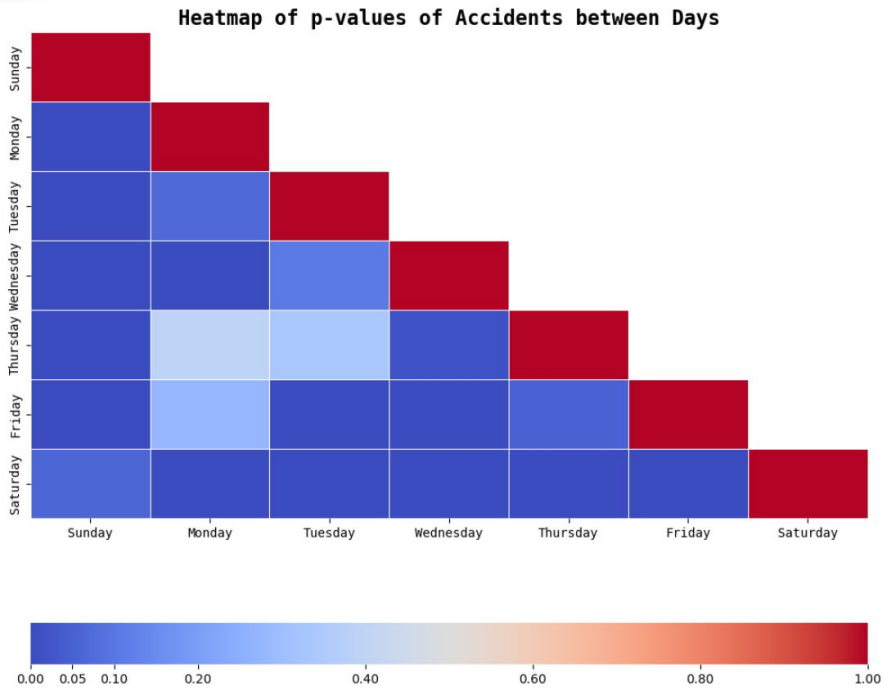
No correlations between any of the numeric columns

Hour of day- day of the week relation



Accidents between days P-value heatmap

	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



- p-values gotten from extracting the p-value from independent t-tests comparing the time accidents happened between each day
- Used a hard upper limit of 0.05 for significance
- Comparing the time accidents occur across the days of the week yields significant differences for most days

Random Forest

Goal: Predict what day an accident would occur

One-hot encoded:

- Collision Type
- Injury Type
- Primary Factor

Model:

- Test size: 0.3
- Estimators: 100

Mean-squared error: 3.78

Limitations:

- Model is essentially guessing on what day an accident would occur
- Would need to adjust the model a lot to increase accuracy
  - Adjust: Min Samples Split, Min Samples Leaf, etc.

Conclusions

- Fridays during the afternoon rush hour have the single most crashes
- Overall, afternoon rush hours have the most crashes; early morning hours, the least
- The majority of crashes involve two cars and have unknown or no injuries
- Data allows very limited predictive capabilities, if any
- Majority of crashes have an outcome of no injuries

Bias & Limitations

- Bias
  - Only includes reported accidents
- Limitations
  - Lack of continuous data
    - All data except for latitude and longitude was discrete
  - Injury Type data is limited to 4 categories
  - Data is almost 10 years out of date; unlikely to provide good predictions for current accidents

Future Work

- Adjust the Random Forest Regressor and Test-Train Split to better train the model
- Compare this dataset to one containing locations of suburbs or housing to understand if people are more likely to get into an accident close to home
- Analyze “Primary Factor” variable to determine a relationship between it and injuries sustained or type of collision
- Define Injury type category into more specific buckets

Works Cited

- 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.

Any Questions?