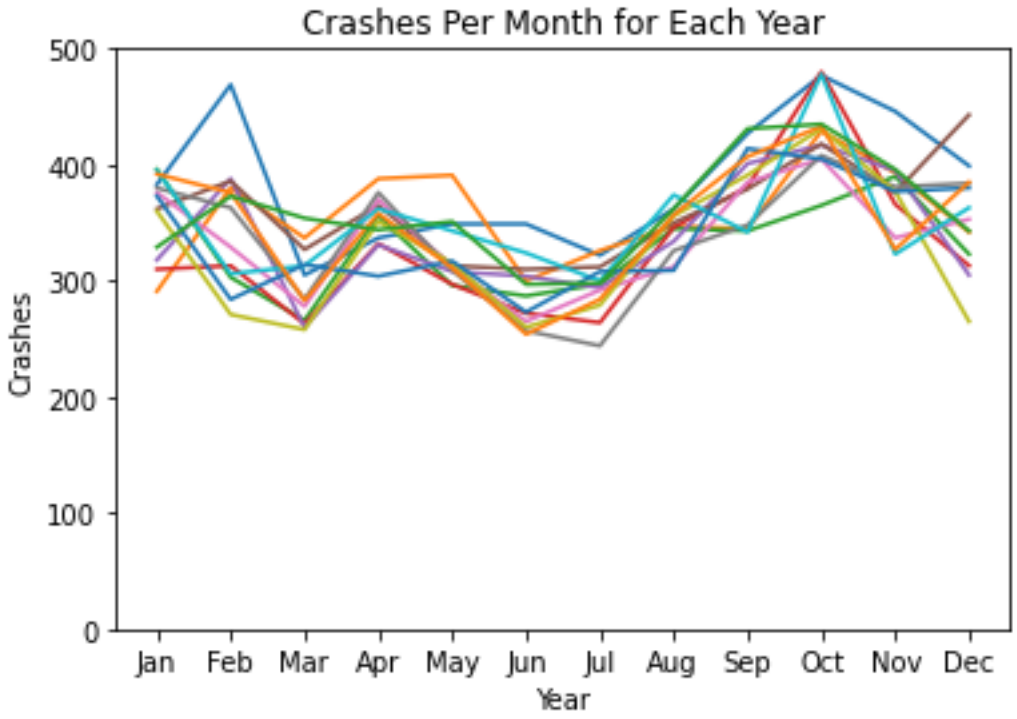
A Basic Analysis of Crash Data from Bloomington, IN from the years 2003 – 2015

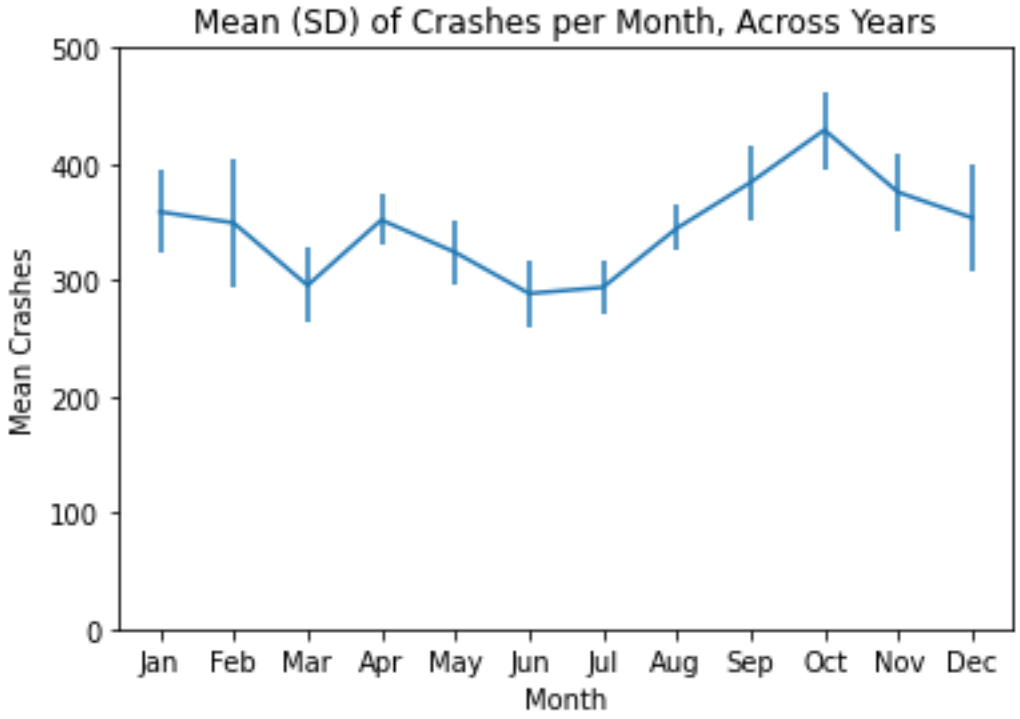
Here, I was interested in learning something important about the town I've been living in for the past several years. I was curious to see any interesting pattern in car crash counts in a small college town that works on a seasonal schedule, following the college semesters. In particular, I was expecting to see spikes in crashes in the months of August and January, where new students typically move to the town and are slowly becoming accustomed to the roads here. But, as newer students get used to the area, especially all the one-way roads, the number of crashes should decline, and this general pattern should be seen across the years measured.

But, to my surprise, this is not what was observed. Indeed, here is a graph showing the number of crashes for each month across the years the town measured:



In the Spring semester (January - May), we actually see spikes in February and April, while in the Fall semester, we see a gradual *increase* in crashes as the semester continues, with a peak in October, followed by a decline in the months of November and December. Each line here represents a different year; what struck me was how consistent the years are in terms of crashes.

To remove some clutter in the graph, here is a graph showing the average with standard deviations as error bars:



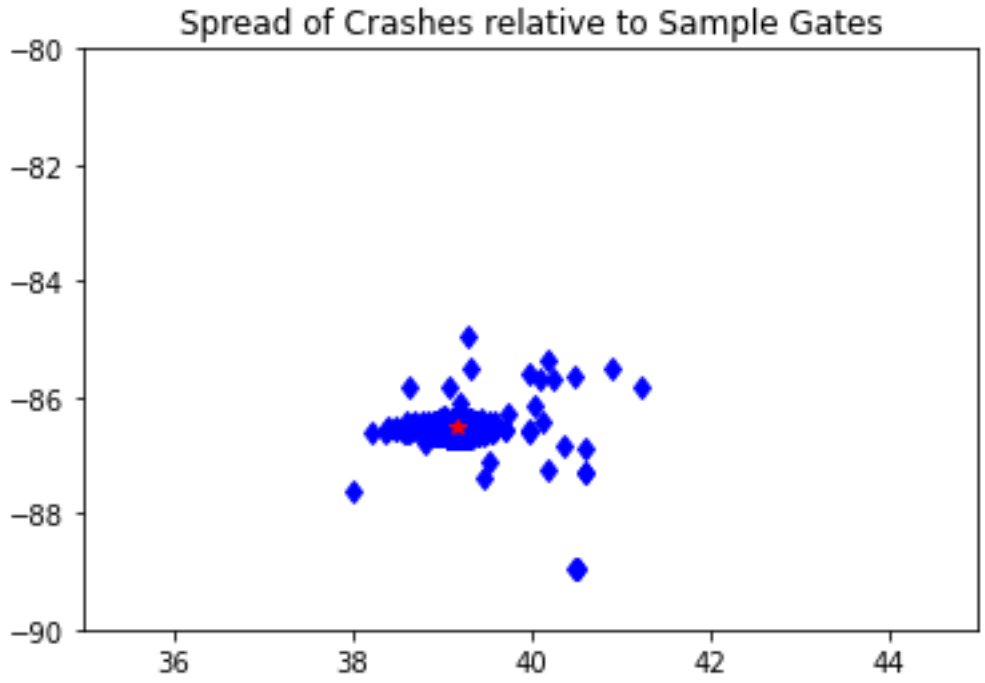
The sharp declines in March, June, July, and December can be easily explained by the mass of students who leave Bloomington for vacation (spring break in March). But it seems odd that the patterns are different across the two main semesters. Additionally, neither semester is "stable" in their crash numbers, or, the number of crashes is not constant across the months in each semester. I tested this in the corresponding Python code by looking at ANOVAs for each semester, with each month being a factor and the number of crashes for each year the dependent variable; both ANOVAs came out significant, so that there are indeed statistically significant differences between the months in each semester. In the case of the spring, the consistent drop in March may explain the significant ANOVA, while the consistent and steady increase from August to October every fall would explain the other significant ANOVA.

For now, I can't quite understand why there would be such an increase, except for the following possible explanation: September and October are the months in the fall with no significant vacation days. The former gets Labor Day, and the latter gets a one-day Fall Break, with the gap between each being a bit over a month long. This prolonged gap in vacation days

may be what contributes to the higher numbers in both of these months. Indeed, by the average graph, it seems like both months are the highest for the year for crashes.

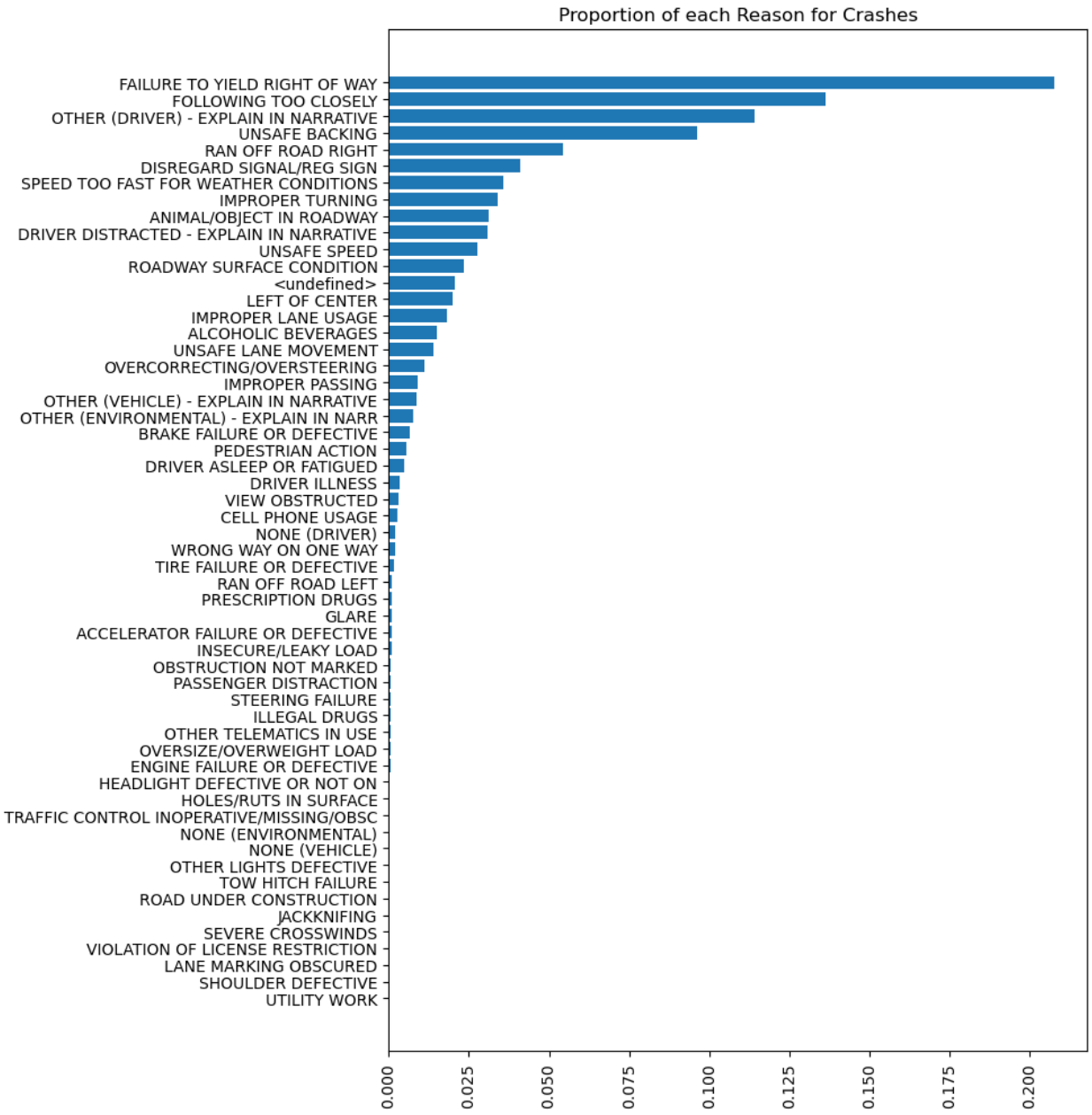
As a side note, we see more variability in the crash counts in the winter months – this, no doubt, is because of the variable snow that Bloomington sees each year. This may also contribute to the higher counts in October, as sometimes we can get snow here as early as October, and that combined with the area still being somewhat new for many students, may combine to yield the higher crash counts in this month.

A natural question may be where most of these crashes are concentrated. The reader may not be that surprised to find out that most are concentrated near the university, the center of the town. The crash data had the longitude and latitude for each crash, when available, and so I graphed these points with a reference point in town to see the spatial distribution of crashes:



The red dot here is the reference point: the Sample Gates that form the formal entrance to the university, located at the end of the busy Kirkwood avenue, and parallel to the one-way, also busy, Indiana Avenue. The university buildings are north of the Sample Gates, and the downtown area is located below the red dot here. I should note, the other clustering of crashes, not shown here for simplicity, occurs at a highway juncture more to the north of the town, but the cluster by the university by far eclipses the other.

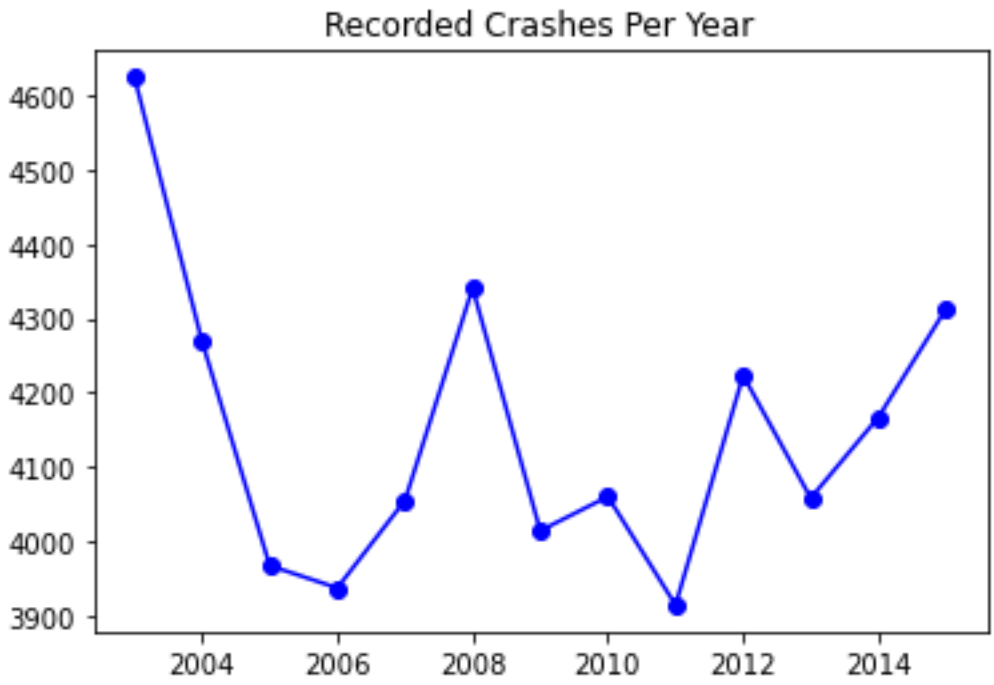
We also have the luxury of examining the reasons recorded for the crashes. Here is a graphic showing the percentages of crashes corresponding to each reason:



The two most common reasons, that account jointly for about 35% of crashes, are a failure to yield the right of way and following too closely. Note that both of these reasons can become exaggerated under snowy conditions. The attentive reader might also note a reason of

“<undefined>;” this not only indicates missing data, but can also indicate a hit-and-run. Thankfully, these only seem to account for about 3% of the crashes.

One analysis that stumped me, as I could not find a reasonable pattern, was in checking the total number of crashes across the years in the dataset. Here is a graphic:



There was a steep drop off until 2006, where there was again an increase, peaking in 2008, and then another decline from 2009 to 2011, with a general trend upwards (save in 2013) until 2015, the last year in the dataset. These data may reflect a history of this town that I’m not aware of; for instance, it may be that some winters were harsher than others, or maybe it had to do with the university’s enrollment numbers for the years. The current dataset is limited, however, in testing these ideas.

A final analysis I did with these data was to check if there was anything special, probabilistically, about October. Were the reasons for crashes different in October than in the year as a whole? One way to test this idea is to see if the proportions of crashes accounted for by each reason are similar when we look at the whole year versus October only. In other words, if we let X be a random variable measuring the number of times, say, failure to yield the right of way was the cause of the crash, we want to know if $P(X) = P(X \mid \text{October})$. As the following

table of values demonstrates, indeed October appears to be probabilistically identical to the entire year:

	PrimaryFactor_Year	PrimaryFactor_Oct
FAILURE TO YIELD RIGHT OF WAY	0.207497	0.222003
FOLLOWING TOO CLOSELY	0.136422	0.153019
OTHER (DRIVER) - EXPLAIN IN NARRATIVE	0.114158	0.101416
UNSAFE BACKING	0.096176	0.091023
RAN OFF ROAD RIGHT	0.054224	0.047303

This is the first five rows of the dataframe comparing the proportions for each reason for the year and for October only, in descending order. Indeed, the percentages are quite similar. In particular, this suggests that the main reasons for crashes for the year are the same as those for October only. (Of course, one could do a chi-squared test to formally assess the differences, but a glance shows this largely an unnecessary formality.)

Conclusion

It would seem, then, that October is the most dangerous month to drive in Bloomington, IN. In particular, drivers should check twice, maybe thrice, to make sure that it is safe to go, even if you have the right of way. Drivers should also be wary of driving too closely to each other, especially in the winter months. Finally, it seems that most of the crashes occur around the university, especially in and around the downtown area.

It does appear that the crash patterns follow the university class schedule, but in complicated ways. Of course, those months with the longest breaks are those with the lowest crash numbers, but the Spring and Fall semesters are otherwise not equivalent in their patterns of crash numbers. Indeed, the prolonged lack of a long enough break between September and October means that more students are likely to stay in town; combine this with the beginning of the snowy season in October (in some years), and this inevitably leads to an increasing in the

number of crashes. However, the probabilistic pattern of crash data does not appear to change significantly in October, the month with the consistently highest number of crashes year after year.