Team Lookout!

2016 Crash Analysis
when combined with
Weather and Traffic Data

Daniel Brewer
Orion Crocker
Ebelechukwu Esimai
Ernesto Martinez
Seth Seeman

Table of Contents

Project Overview

Project Goals

Data Sources

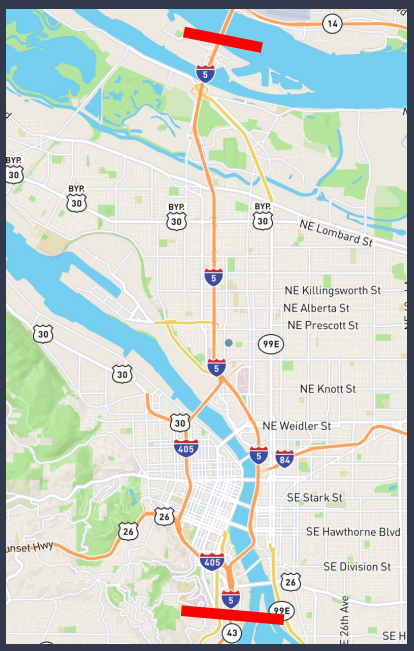
Seth

What we looked at...

Location and Timeframe:

Our team chose to focus on a subset of crash data in the Portland, OR region to analyze alongside local weather and traffic information. The main goal was to join and aggregate the datasets in an effort to pinpoint common crash locations and factors. Using these aggregates, we hoped to prove or disprove our assumptions about how time, weather, speed and traffic volume might affect the total number or severity of crashes. Depending on the degree of our success, the results could help inform the public, both organizations (i.e. Oregon Department of Transportation) and individuals (i.e. drivers) about the major factors contributing/related to vehicle collisions in the Portland Metro area.

The subset we chose was largely based on the ease of finding complete datasets for a particular timeframe and continuous section of road. We settled on a **10 mile stretch of Interstate-5** in between downtown Portland (more specifically the South Waterfront/Ross Island Bridge) and the Columbia River (state border between Oregon and Washington). We also decided to start with data from **2016** as it was the most recent, complete set we could find for **crashes, traffic and weather**. Our process and analysis could fairly easily be applied to any stretch of highway with similar data attributes.



Seth

What we looked at...

Data Source 1: ODOT Crash Data

In 2016, there were nearly 700 crashes recorded on this particular stretch of I-5 in Portland, Oregon. They involved over 1600 vehicles, caused 827 injuries and resulted in 3 fatalities. That's roughly a 50% chance of injury and 0.5% chance of death.

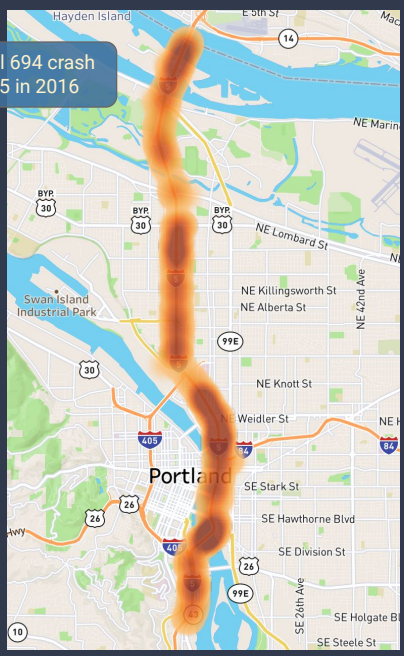
Search Parameters:

- **Interstate 5**
  - North & Southbound
  - Mileposts 299 - 308
  - Jan 1 - Dec 31, 2016

Query Statistics:

- **Northbound**
  - 297 Crashes
  - 710 Vehicles Involved
- **Southbound**
  - 397 Crashes
  - 936 Vehicles Involved

Heatmap of all 694 crash locations on I-5 in 2016



Seth

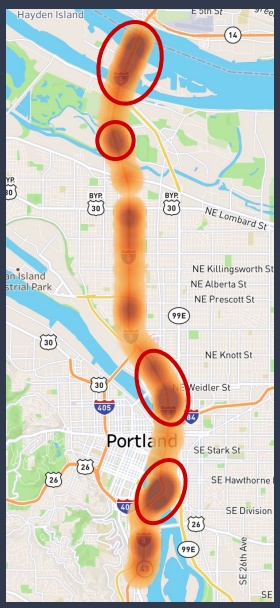
What we looked at...

Data Source 1: ODOT Crash Data

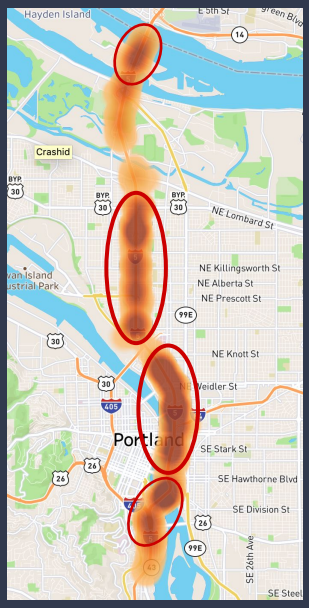
The geographical heat maps help accentuate crash location density that allows us to target common collision areas. These sorts of starting statistics are essential to keep in mind while analyzing other factors.

- **Northbound**
  - 297 Crashes
  - 710 Vehicles Involved
- **Southbound**
  - 397 Crashes
  - 936 Vehicles Involved

Northbound Crashes



Southbound Crashes



Seth

What we looked at...

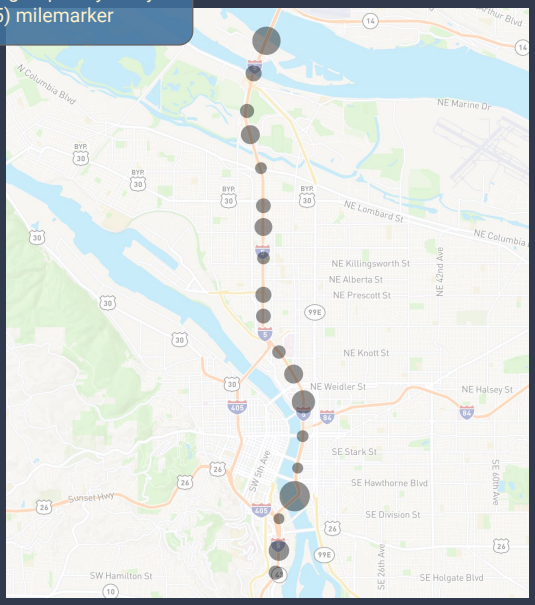
Data Source 2: PORTAL Traffic Detectors

The Traffic Detection data helps us understand what role traffic volume and speed play in vehicle collisions. By rounding the crash data mile marker locations to every 0.5 miles, we were able to isolate buckets of locations that have the most crashes and then compare that to the traffic speed and volume at those locations and even times.

Data Summary

- Tracks avg speed and total volume every hour
- 192 Detectors
  - Each have 8,760 records/year (hourly)

Crashes grouped by every half (0.5) milemarker

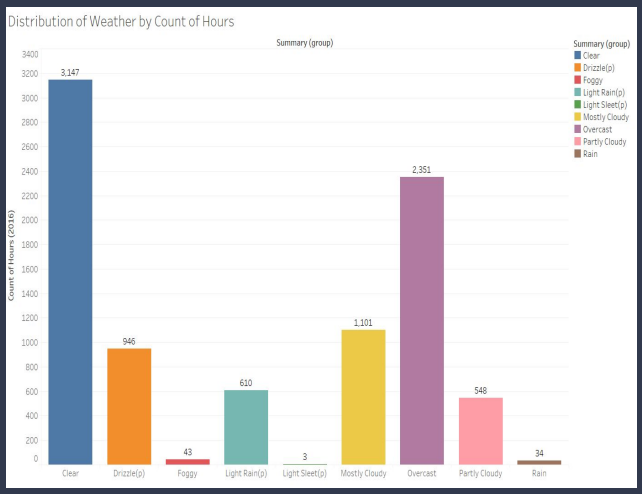


Seth

What we looked at...

Data Source 3: Weather (DarkSky API)

Weather we know is a major factor when it comes to impacting driver safety. The goal is to find correlations in days or even hours with significantly adverse weather and see if it results in a higher volume of crashes relative to better conditions. It will also be interesting to see whether traffic speeds and volume are affected in a predictable way by poor weather conditions.



Data Summary

- Single point detection (approx. milepost 302)
  - 8,760 records/year (hourly)

Seth

What does it all mean?

Traffic

Crash

Weather

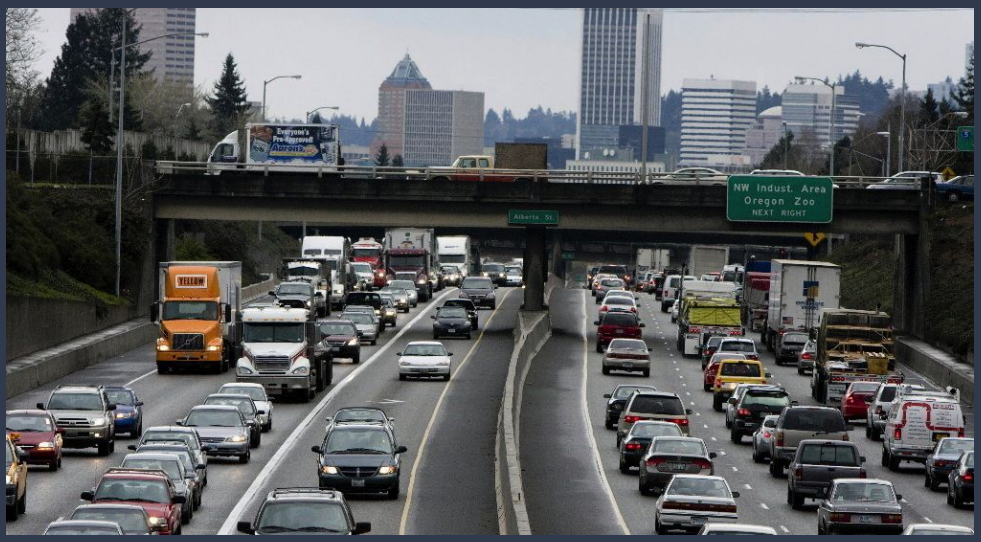
Seth

Putting it all together

The next 3 sections will provide a deeper exploration on various aspects of our datasets by analyzing the crash data with one other major contributing factor. The ultimate goal will be to draw conclusions from combining the visualization highlights and results from each section.

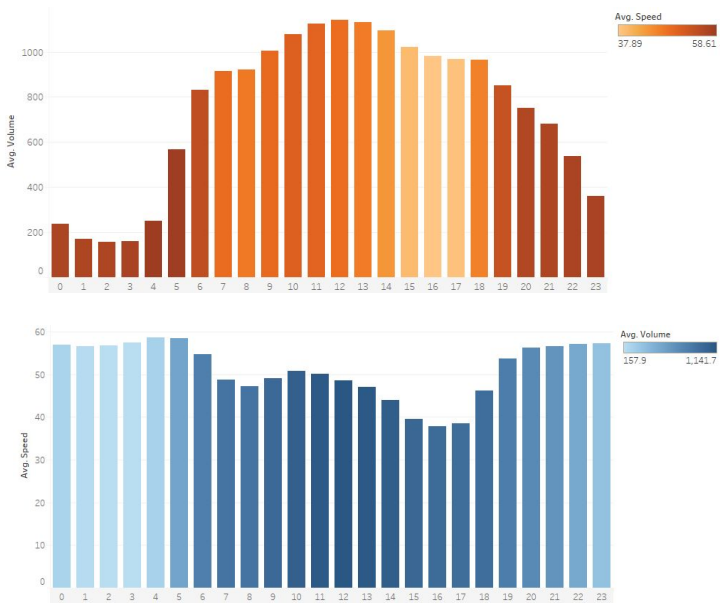
1) **Traffic Speed and Volume** as they relate to crash times and locations.
2) **Weather and Road Conditions** as they relate to crash frequencies.
3) **Time** as it relates to crashes in general (hour, day, week, month, season)

Traffic Analysis



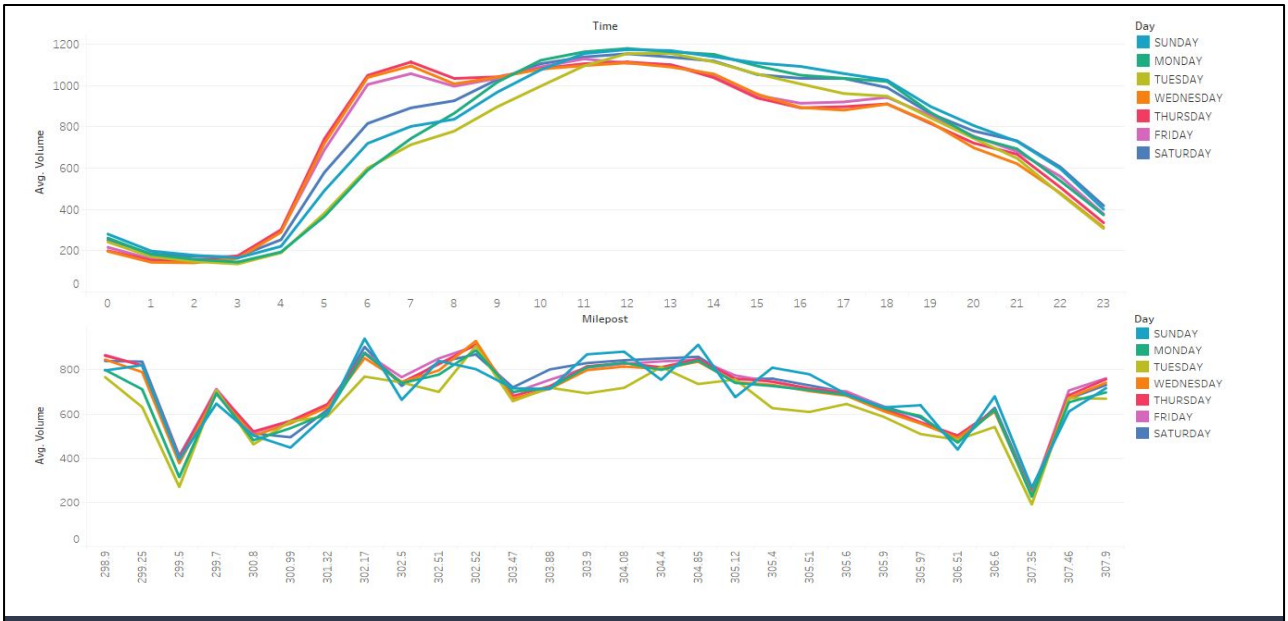
Speed and Volume

The combined data from all weekdays clearly defines a trend that describes rush hour traffic patterns between mile markers 300 and 308. The two visualizations consist of both northbound and southbound data.



These two visualizations, one describing speed and one volume, show a clear trend that describes the rush hour traffic patterns that all commuters know. Between the hours of 8pm to 5am, traffic has the least amount of cars on the road (volume) and highest average speed. When looking at the speed and volume of the commuter hours, 6am to 7pm respectively, the data shows a fluctuation between the the correlation of average speed and volume. While volume certainly is an important factor in determining the average speed of traffic, it is not the only one.

One example of this is clear when examining the range between 3pm and 5pm. While there is still a relatively large volume of cars on the road at this time, the total average volume is trending downwards from the earlier peak of 12pm to 3pm. However, even though the volume is steadily decreasing, the speed also declines in the same time range.

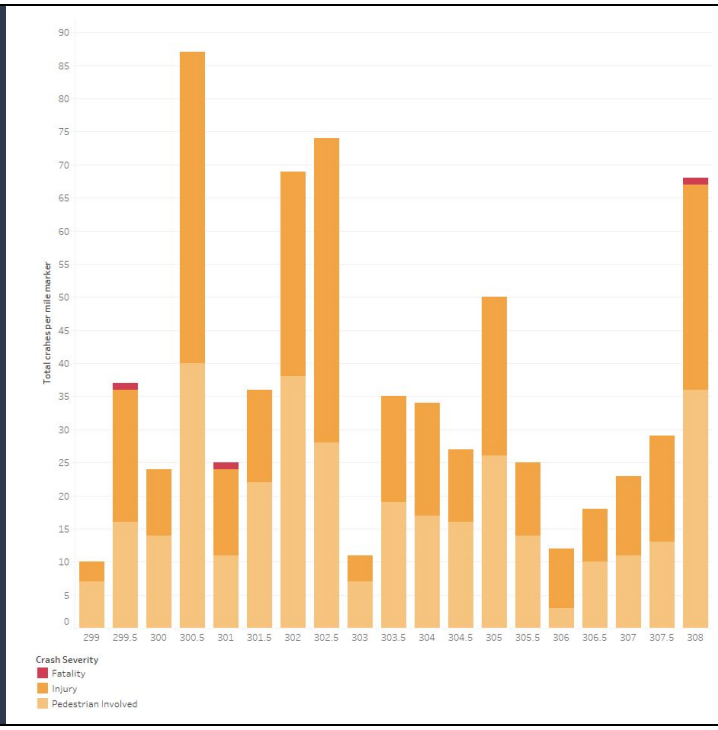


Total average volume for all days of the week, shown by hour of day and specific mile marker.

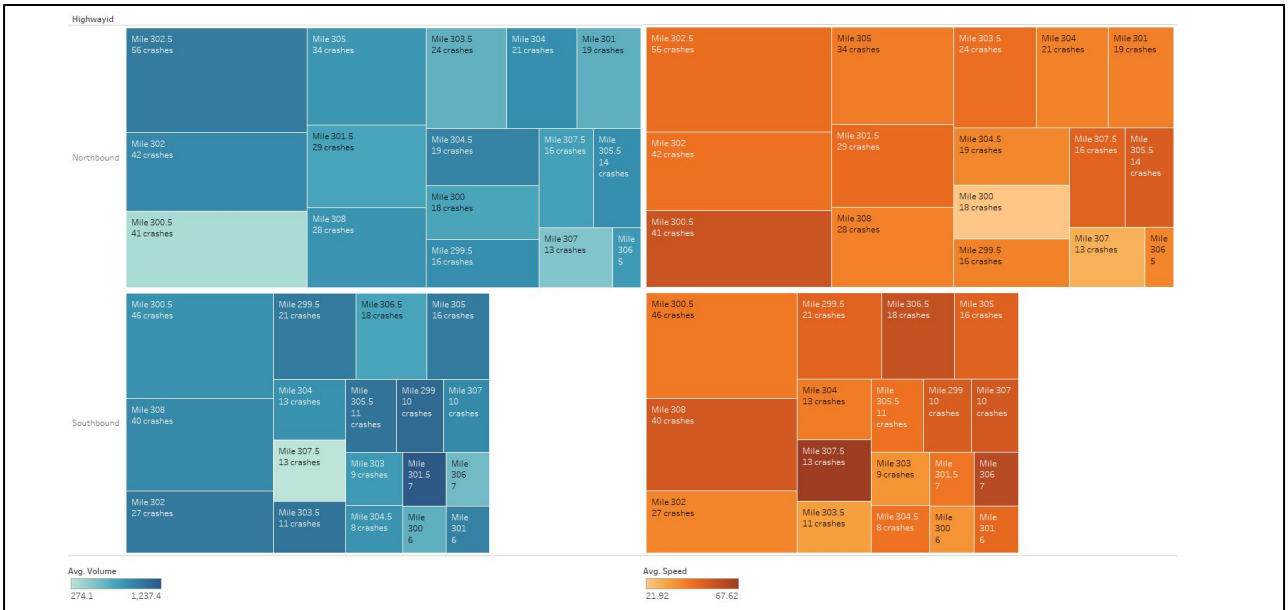
When looking at the average volume with respect to hour all of the days of the week independently, individual daily patterns can be recognized. For example, while traffic averages at around 1100 cars per hour on all days between 11am and 12pm, days Sunday, Monday, Tuesday, and Saturday have significantly less cars on the road in the hours leading up to 11am. Additionally, it can be said that commuters on Wednesday, Thursday, and Friday are on the road earlier in the day, peaking at 1100 cars per hour at 7am.

While examining volume at individual mile markers along our designated stretch of I5, it becomes immediately clear that some parts of the road are more utilized than others. This fluctuation in volume between mileposts may be due to a number of reasons, but highly probably tied to the fact that our portion of road crosses the border to Washington. This drop can be seen in the few miles leading up to the bridge crossing (306 - 308).

Traffic and Incidents



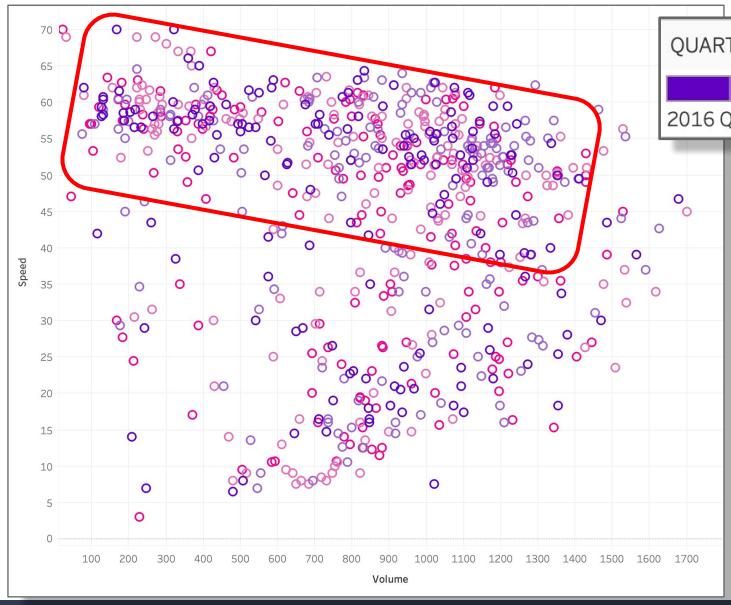
The above graph represents the number of individual crashes per half-mile for our 8 mile stretch of road. Additionally, it can be see at a glance that there were three total fatalities, and a relatively equal distribution of incidents involving pedestrians as there were vehicular injuries. The number of incidents involving pedestrians was surprisingly high, especially because the data represents freeway incidents. However, I suspect that this is a symptom of the inaccuracies that occur when individuals self-report crash data to ODOT. Only when an incident involves a fatality does a police officer file the incident report.



Volume, speed and number of Incidents for each half-mile

Each square represents a specific half-mile on our stretch of road between mile markers 300 and 308. The size of each square represents the number of crashes that took place there over the course of 2016. At a glance, it can easily be determined that there were significantly more crashes on the northbound route than the southbound. Additionally, it can be observed that the half-miles with the largest number of crashes did not have a higher than average speed or volume.

While this trend doesn't "rule out" the effects of speed or volume on the number of crashes that occur at a location, it can be determined that other factors play a part and must be analyzed accordingly.



- This plot shows that the **majority of crashes occurred at higher average speeds**.
  (shown by the cluster in the red rectangle)
- However, **volume is relatively evenly distributed** across all crashes until volume reaches 1400 vehicles/hour at which point crashes are less likely (probably because those volumes happen so infrequently over the course of a single day or the year as a whole).
- Also, important to note that there are no crashes above 65 mph as volume increases (as it is likely no drivers are able to go that fast during those higher volumes).

Crash Plot by Speed and Volume (Color represents time of year)

Weather Analysis

Weather and Crashes

Intuition:

We expected weather to have a direct correlation with crashes or, even further, to show a causality between “Bad” weather and increase in the number of crashes.

Investigation:

Despite our bias, we focused on inferences that drivers and planners care about:

- Drivers care about safety and about travel times
- Transit authorities care about planning Advisory speeds and Resource Allocation.

Plan:

- Identify “hotspots”: areas where crashes are more likely, possibly at certain times or in certain weather conditions.

Combining Data Sources

Data join is on DATE and HOUR

ODOT(primary source of crash data):

Represents weather information as a 5 category "Summary" - **CLEAR, RAIN, SNOW, CLOUDY,** and **FOG.** Since ODOT data can be self reported, its accuracy of weather conditions are yet to be determined. Also, since ODOT only provides weather at the time of crashes, there is no information for other times and locations. We therefore chose to augment this report with data from DarkSky.

Dark Sky (primary source of weather data):

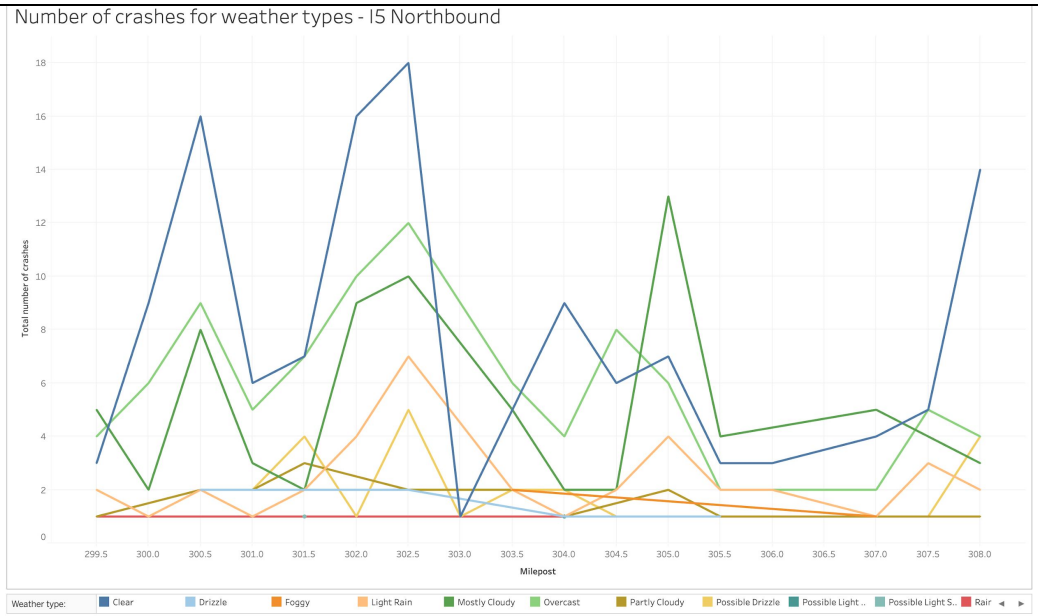
The Dark Sky weather service aggregates many sources of weather, from satellite images to terrestrial radar to weather stations, to produce detailed weather reports about any place and any past time. It provides

- Summary (10 categories)
- Visibility in miles
- and Many other fields for all hours from 01/01/2016 - 12/31/2016

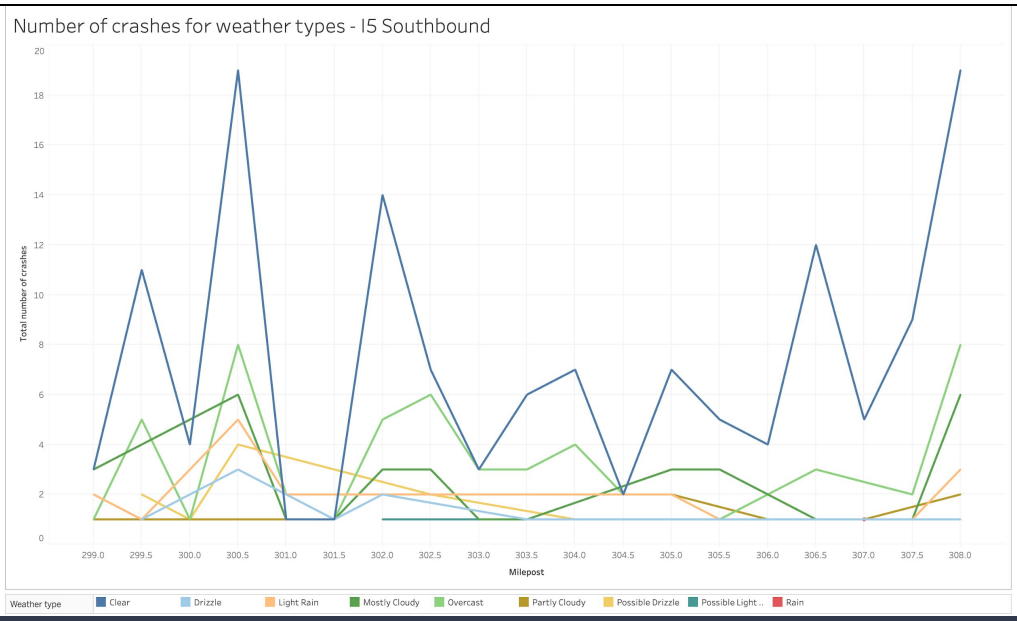
Simply comparing weather and crashes

The next two graphs show our initial, very simple analysis of weather and crashes. We plotted the **total** number of crashes that occurred under each of the listed weather types. Here the crash data comes from ODOT, but the the weather categories come from Dark Sky so that we can fairly compare it against later graphs.

Notice that weather types like “Clear”, “Mostly cloudy”, and “Overcast” dominate the chart. This is not necessarily because those weather conditions cause crashes, but possibly because they are simply more frequent than conditions like rain and sleet. Because of this, the initial analysis didn’t tell us what we wanted to know about weather and crashes.



Total number of crashes fails to account for number of hours with that weather



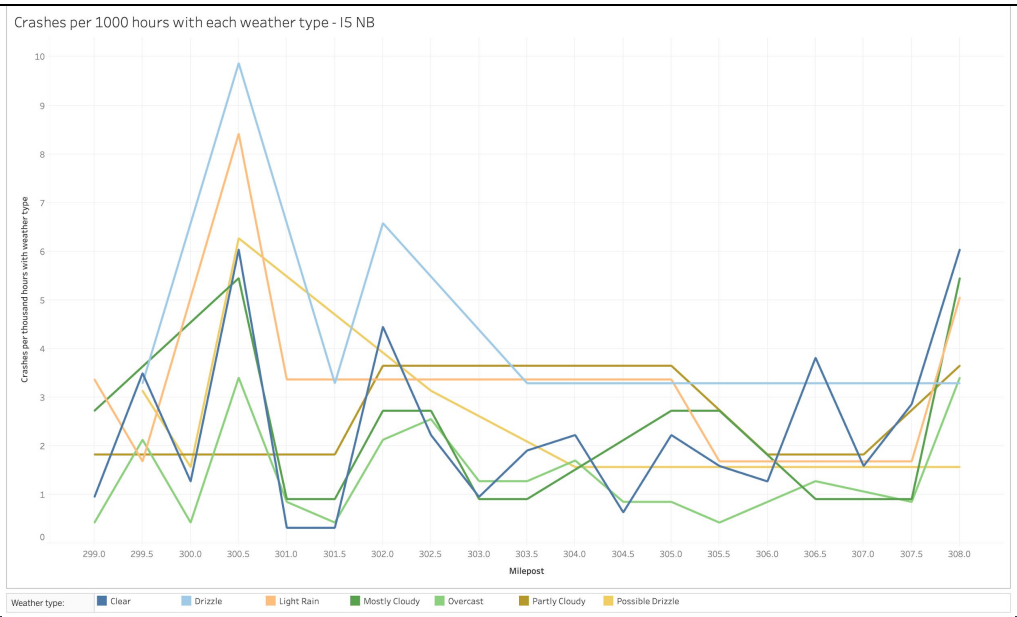
Total number of crashes fails to account for number of hours with that weather

Fairly comparing weather and crashes

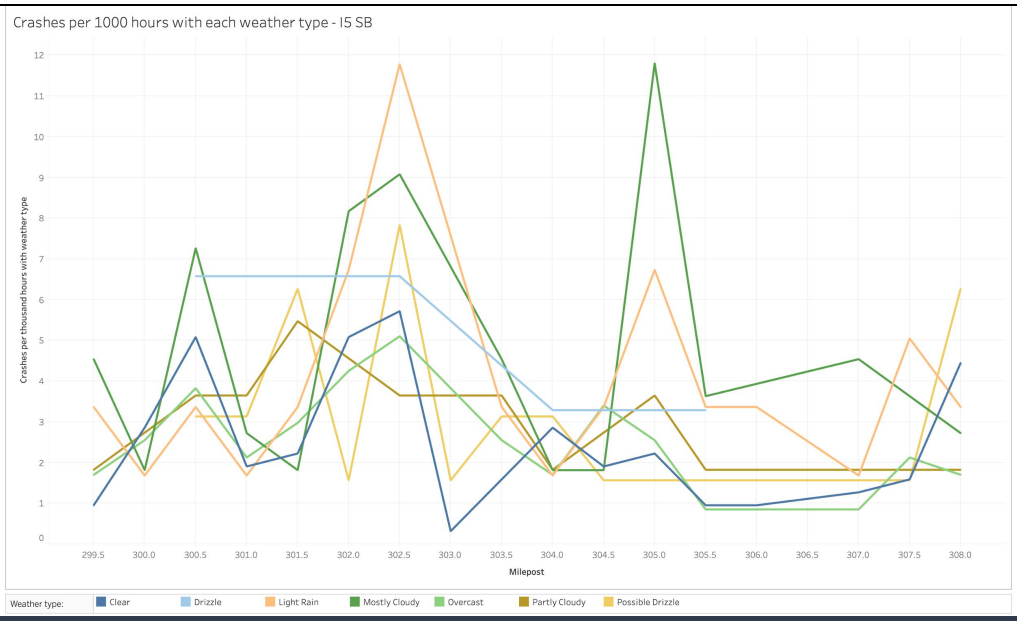
The next two graphs show a more fair comparison between crashes in different types of weather. We needed to scale the previous graphs by the overall frequency of those weather types. ODOT only reports the weather *when there are crashes*, but not overall. To find overall frequency, we used Dark Sky data to count how many total hours of each weather type happened over that year. Then, dividing the total number of crashes by the total number of hours gave us the number of crashes *per hour of that weather.*

We scaled the Y-axis up by 1,000 (crashes per 1000 hours) because the tiny numbers were harder to read and visualize.

Drizzle and Light Rain are associated with the highest number of crashes, as expected. This disproportionate risk of certain weather conditions suggests that crash reduction efforts during some weather conditions could have more impact than in other conditions.



Dividing by the number of hours with that weather type corrects for the distribution of the weather types and shows how *driving an hour in that weather type* relates to crashes



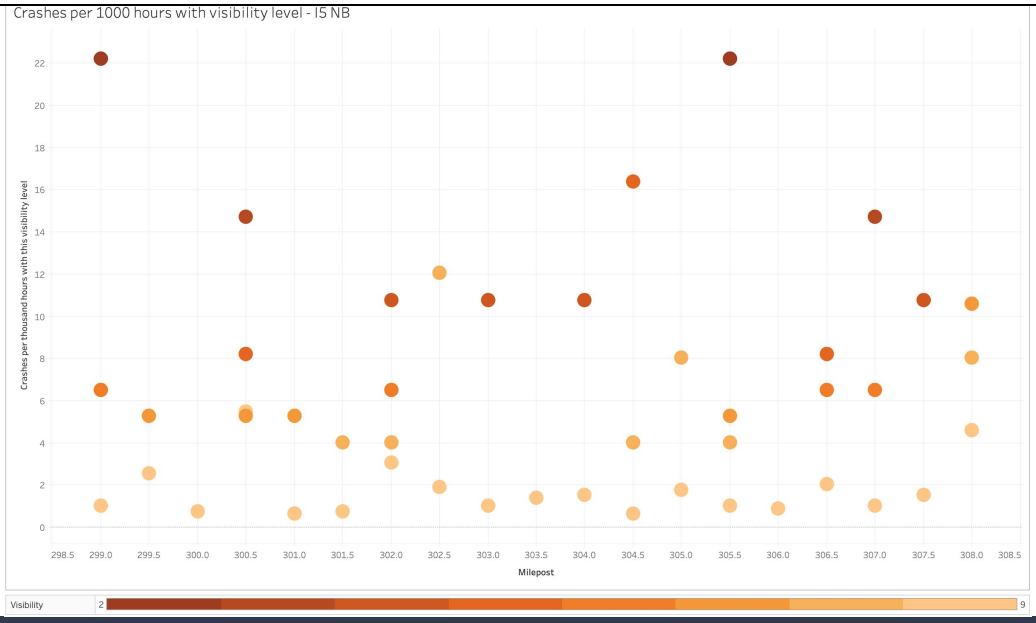
Dividing by the number of hours with that weather type corrects for the distribution of the weather types and shows how *driving an hour in that weather type* relates to crashes.

Comparing weather and visibility

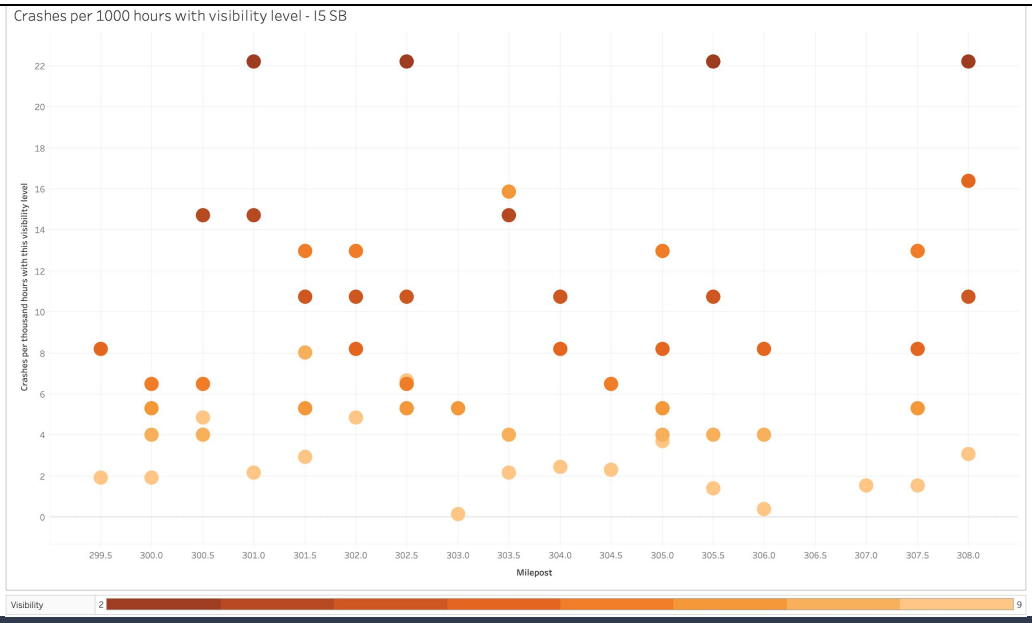
The next two graphs show the relationship between crashes and the visibility level. Dark Sky reports “visibility” level as an integer representing visibility in miles, including at nighttime. We wanted to explore visibility rather than the discrete and vague weather categories because it’s a more continuous number. Like the weather categories in the previous slides, we needed to divide each crash count by the number of hours with that visibility over the year.

By plotting number of crashes vs visibility, there’s a clear trend that crashes happen during bad visibility. We had expected less of an impact, since the difference between 9 miles of visibility and 2 miles of visibility sounds like an insignificant difference.

Like with the crashes/hour graphs, the resulting numbers were confusingly small, so we scaled them to to crashes/1000 hours.



After correcting for # of hours with each visibility level, the trend is clear: crashes happen less frequently when visibility is good.



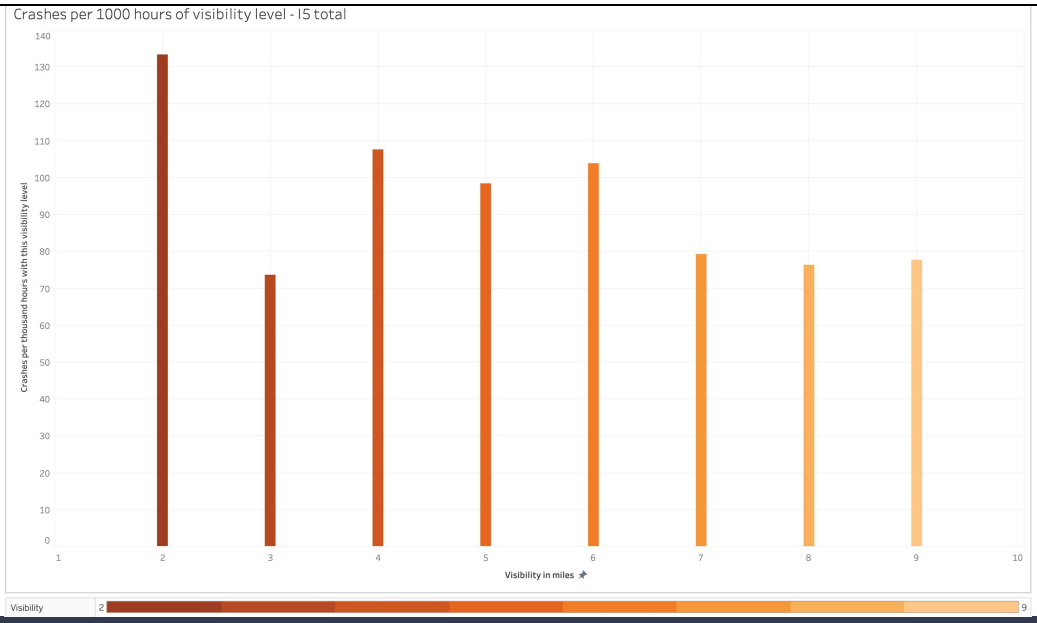
After correcting for # of hours with each visibility level, the trend is clear: crashes happen less frequently when visibility is good.

Looking in vain for “hotspots”

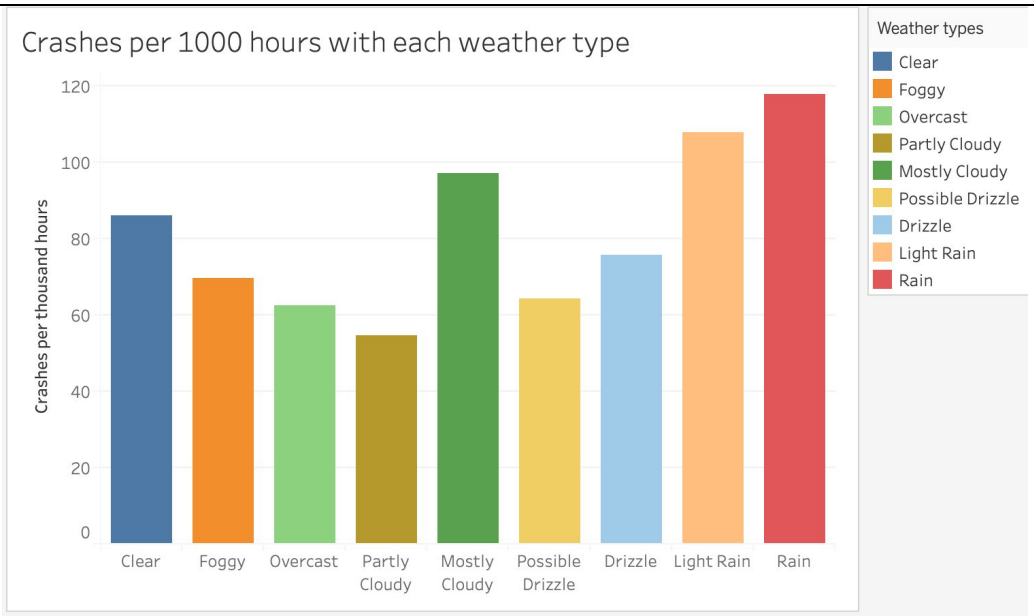
In the graphs above, we always used milepost as the X axis, even though this detracts more from the graphs than it adds to them. We tried this because we wanted to identify “hotspots”, especially dangerous locations. Planners and traffic authorities want to know that certain locations are dangerous under specific conditions. Plotting on mileposts could help reveal those hotspot locations.

Our hope was to normalize the data to eliminate the “noise” caused by some mileposts being more crash-prone than others. That should reveal the “signal” of crashes caused *disproportionately* by weather at that milepost. We tried using the means and standard deviations of crashes with regards to a few factors at each milepost, but these graphs never showed interesting or useful relationships.

Unfortunately, this kind of graph was a failed experiment.



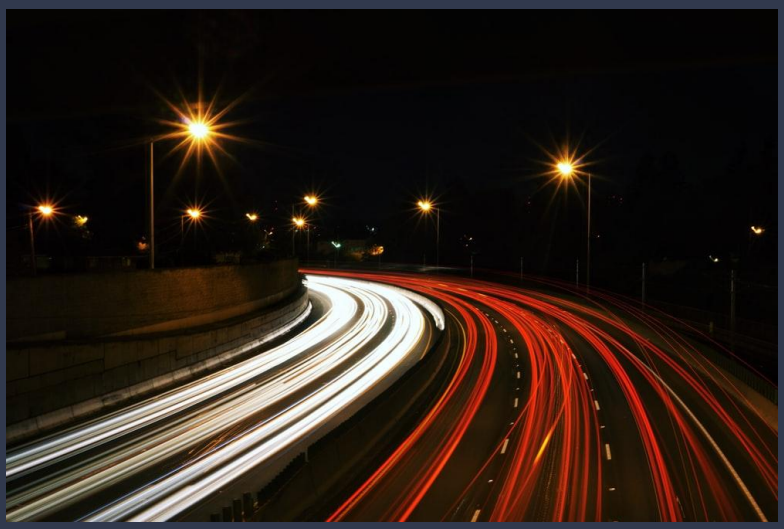
Simple conclusions for drivers: Driving in very low visibility is almost twice as dangerous as driving in great visibility.



Simple conclusions for drivers: Rain is the most dangerous weather type, and some cloudy conditions are surprisingly safe.

Simple conclusions for drivers: On the road, this safety information should be further condensed into very simple advisory messages.

Time Analysis



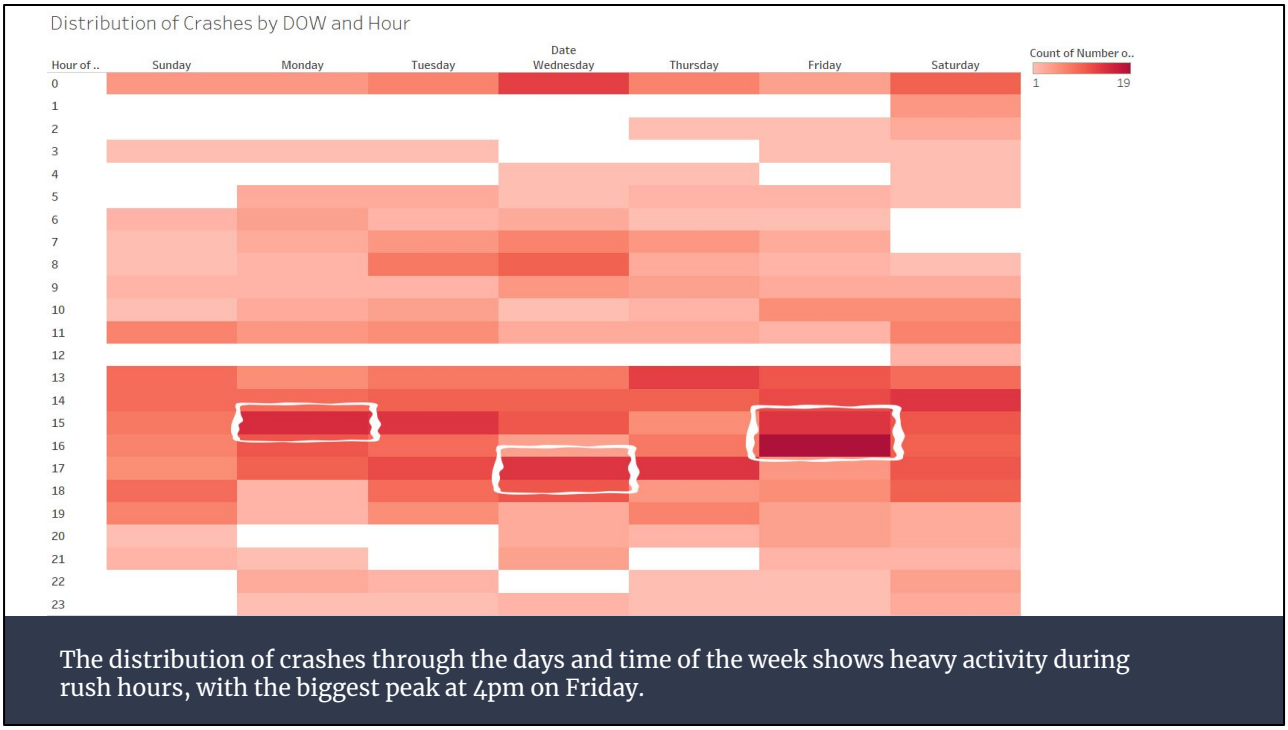
Time and crashes

Time of day could affect the number of crashes in many ways.

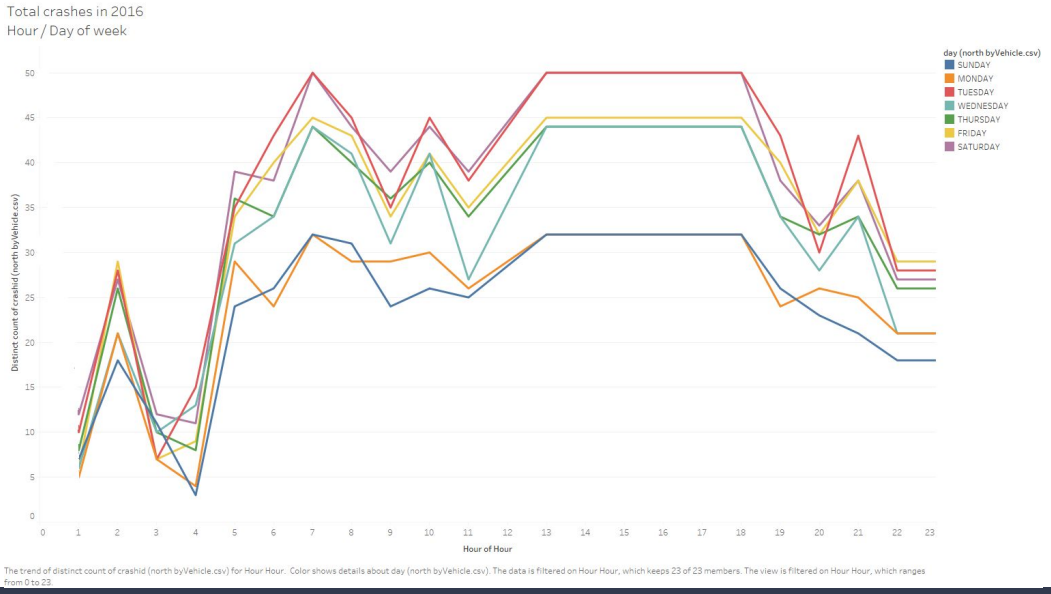
Some times have higher traffic volume than others. Mornings are usually low volume. Rush hour, when a high percentage of the population is driving out/in from work, has much higher volume. We also expected rush hour to have more crashes.

There could be other time-related factors. Night time crashes may happen due to speeding or alcohol. Day crashes may happen due to light conditions.

Looking at graphs about crashes based on the time of the day may help us discover insights into time-related factors that affect the number of crashes.



The distribution of crashes through the days and time of the week shows heavy activity during rush hours, with the biggest peak at 4pm on Friday.



This graph showcase the same information but with a different approach. We can see Sunday & Monday as a relative calm start of the week. On the other extreme, Tuesday and Saturday are more intense. On Saturdays, people go out at all times during the day, causing a high intensity of crashes.

Project Summary

Challenges

Conclusions

Challenges

- Breaking tasks between Tableau and SQL
  - Finding a good separation of tasks was difficult because we had more experience in SQL than Tableau. As such, we opted to do joins and some aggregation in SQL that were not intuitive in Tableau.
- Appropriate combination of dataset for valuable inferences
  - Combining the data was challenging in that though they are holistically related, the details did not match well. For example, the traffic data is collected by the hour and though the hour a crash occurs is recorded, we could not show the lasting effects of a crash on traffic flow since the effect may be resolved within the our that it occurs. We opted to let the trends in the summaries of the datasets guide our analyses.
- Bias versus Outcome
  - We had bias on our projections such as there will be more crashes in “bad” weather and we subconsciously preferred graphs that confirmed our bias. However, the outcomes of the analysis disproved some of our bias and showed correlation on some of our assumptions.
- Complexity
  - In our ambition to show multiple features relating, we developed complex graphs which were less informative than simple graphs. More complex “hotspot” graphs took **much** more planning and analysis than simpler graphs.

Conclusions

- Traffic
  - Most collisions occur at higher speeds (62% above 45 mph, 72% above 35 mph)
- Weather
  - Bad visibility and weather are both weakly correlated with more crashes. Lowering advisory speeds could slightly improve safety in these conditions
- Time
  - There is a weak correlation between time of day and crashes. This is evident in the increase during heavy traffic hours and also seen at midnight. We can theorize that increased traffic volume increase crash probability and ending of social activities about midnight may contribute to crashes.

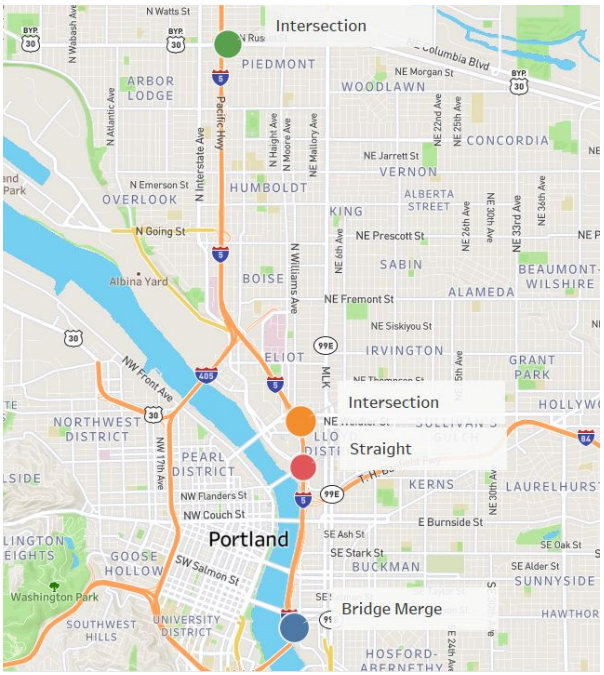
Although, we didn't consider it as the primary factor, we found that the most notable factor to crashes is **LOCATION**. Regardless of the variation in weather, visibility, speed, volume and time, some location are more prone to crash incidents than others.

- The highest crash volume areas are near bridges and major junctions (marquam bridge, columbia river i5 bridge, 405/i5/84 junction area - milepost 301-304)
- Afternoon rush hour is clearly the worst time of day everyday to travel. This might lead us to believe volume actually plays a major role in crashes despite not seeing the distinct correlation it in the traffic section.
- Also note: Saturday night and for some reason wednesday night have significantly more crashes than other nights.

Conclusion

These four spots are examples of spots with high number of crashes regardless of weather, traffic and time.

Each spot is either a busy intersection or merge from/to a bridge. These locations are hotspots and require extra resource allocation and planning.



Appendix

Appendix A: Combining Datasets

SQL Data Processing

- **JOIN:** Traffic Readings + Traffic Metadata Tables (station, highway, detector) → 2016 Traffic Dataset (combined)
- **JOIN:** 2016 Traffic + DarkSky Weather Data → Traffic/Weather Dataset (hourly)
- **Split:** Traffic/Weather (by highwayID) → TW_Southbound + TW_Northbound
- **JOIN:** TW_Southbound + 2016 Vehicle Crash Data → TWV_South (By individual vehicles, not crashes)
- **JOIN:** TW_Northbound + 2016 Vehicle Crash Data → TWV_North (By individual vehicles, not crashes)
- **Aggregate:** TWV_South *group by* CrashID → TWC_South (Final Dataset)
- **Aggregate:** TWV_North *group by* CrashID → TWC_North (Final Dataset)