**INTRO**

Traffic is an issue familiar to pretty much everyone. As a 7-year Los Angeles resident, I’ve sat in more than my fair share of gridlock, seemingly regardless of time of day or day of week. That’s why I was so interested when I stumbled upon a [traffic collision data set](https://data.lacity.org/A-Safe-City/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w) maintained by the city of Los Angeles. This data is cool for several reasons. While it doesn’t directly measure traffic, it measures a closely-related proxy. It’s not a stretch to hypothesize that more traffic correlates with more collisions which directly cause more traffic. I am hopeful that data sets like this one can be used to create safer and more efficient communities for everyone. In that spirit, this data set (and a [bunch of others](https://data.lacity.org/browse)) is actively maintained by the city of Los Angeles and is available to be used by anyone.

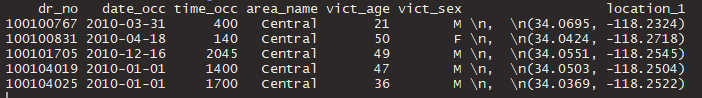
After browsing the data, I settled on 3 overarching questions I wanted to attempt to answer:

* Collisions Patterns by Time: How do traffic collision patterns vary by time of day, day of week, and time of year?
* Collision Patterns by Geography: How are traffic collisions distributed geographically? Is it possible to identify high-risk intersections or areas?
* Collision Prediction: Is it possible to predict the number of collisions in a given time frame?

Before getting into those, let’s learn a bit more about the data set.

**DATA:**

The data begins in January 2010 and is updated weekly. In my particular case, I had data from January 2010 – July 2019, which ended up being ~500K rows of data. Each row corresponds to a collision. This data is transcribed from original paper traffic reports, so it’s very likely there are some errors. Below is a sample of some of the key fields:



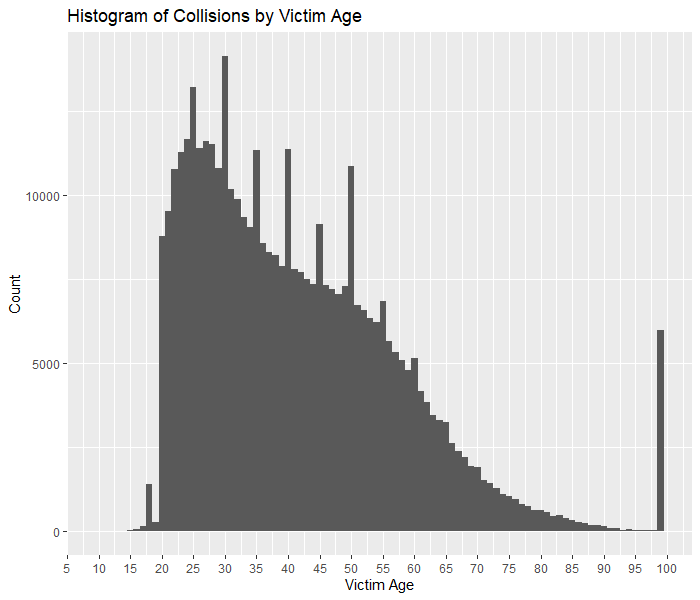
The availability of these fields is what inspired the key questions I listed above.

As with any data set, before starting analysis there was some cleaning to be done. There are a few fields with only one value, reflecting the fact that all of the rows in this data corresponds to traffic collisions. There are also multiple fields with the approximate street names of collisions (not shown above). These text fields needed cleaning, specifically removing extra spaces. Similarly, in the image above we can see the latitude/longitude coordinates contained in a string. I extracted these coordinates in separate fields for later use.

The next step I took was to check for null or missing values. ~16% of collisions (~78K) don’t have an associated victim age. There are also a small subset (~400) collisions that do not have valid latitude/longitude coordinates and are excluded from the mapping section below.

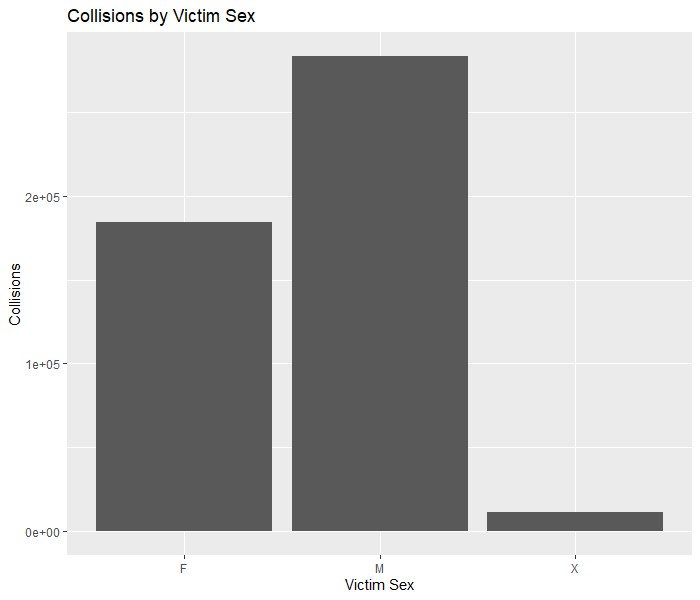
**DATA EXPLORATION**

After cleaning, but before diving into the key questions above, I wanted to do some general data exploration. I started by plotting out a few interesting looking variables.



* There are not many collision victims below age 15.
* Most collision victims are in their 20s. The number of collisions victims per age generally decreases after age 30.
* Note the spikes at most multiples of 5 (25, 30, 35, etc). This suggests some ages are estimated and that identification isn’t always used in collision reports.
* 99 seems to be a catch-all age, perhaps because it’s the maximum age recorded.
* Questions
  + How are collisions with multiple victims dealt with?
  + Is there something else going on with the age 99 bucket?
  + Emailed data owner 2019-08-30…waiting to hear back.

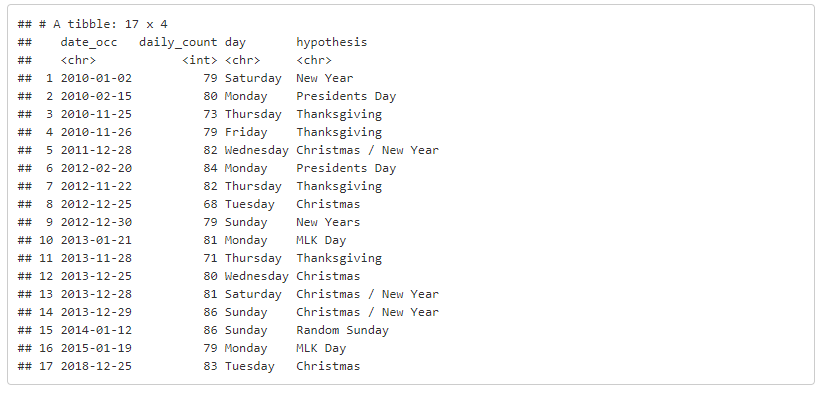
I typically find this sort of analysis useful. Even though I don’t have a specific goal in mind, I often find useful trends or insights. Let’s now look at collisions by gender.



* “X” represents unknown gender.
* Given that a collision occurred, the victim is much more likely to be male than female.

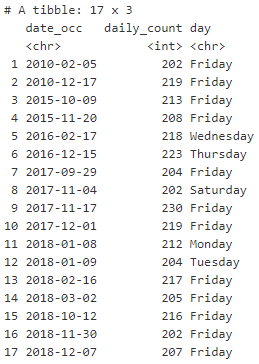
This plot would be more interesting if I had the total number of drivers by gender, allowing for a collisions per capita measure. This will be a recurring theme with this data and would be one of my main extensions of this analysis.

The next thing I wanted to do was look at the lowest and highest collision days in the data. I limited to the top and bottom 0.5% so I could review the results manually. Here are the lowest collision days:



* Most low-collision days occur before early 2014. We’ll see later that monthly collisions start rising after 2014.
* Most low-collision days are around holidays.

And here are the highest collision days:

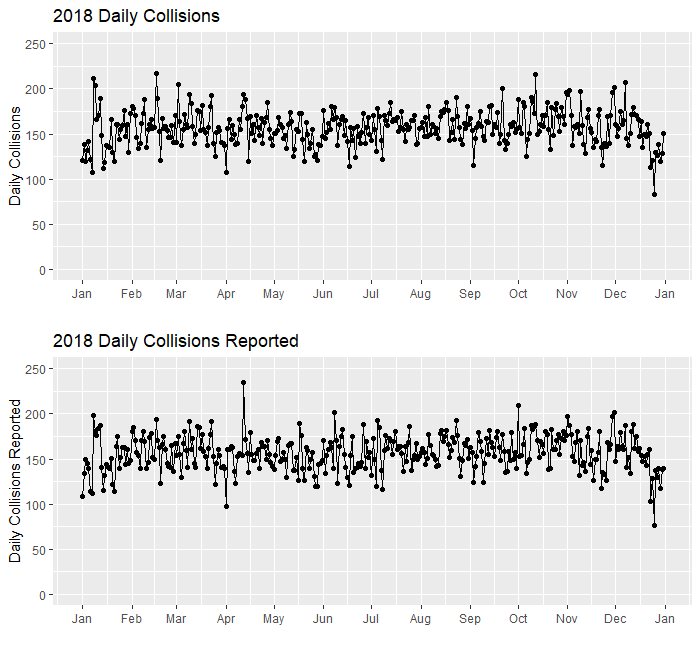


* Most high-collision days are Fridays occurring after 2015.
* We’ll see later that Fridays typically have the highest number of collisions of any day of the week.
* Questions and Thoughts
  + Why are only some holidays associated with low-collision days? For example, MLK Day often has a low number of collisions but Independence Day never does.
  + Why don’t holidays like 4th of July, Memorial Day, or Labor Day show up as low or high collision days?
  + Daylight Savings Time does **not** show up as any kind of outlier in any year. That surprised me.
  + Would it be interesting to look at weather for high-collision days?

**COLLISIONS BY TIME**

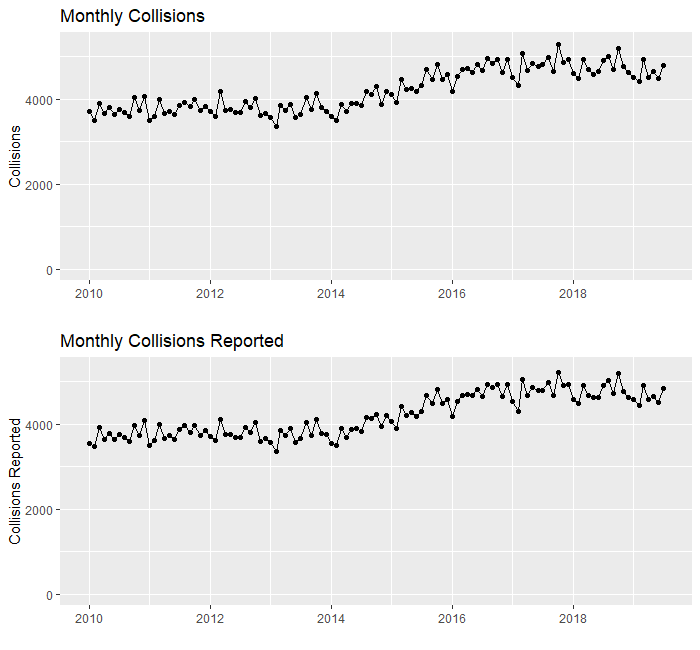
Now I’m ready to look into the main questions I identified above. In this section, I analyze how collision patterns vary by time of day, day of week, and time of year.

First, I look at a plot of daily collisions for 1 year. In addition to collision date, the data has a field for the reporting date. This is the date the collision was actually reported to police. In most cases, the reporting date is the same day or 1 day after the collision date, but not always.

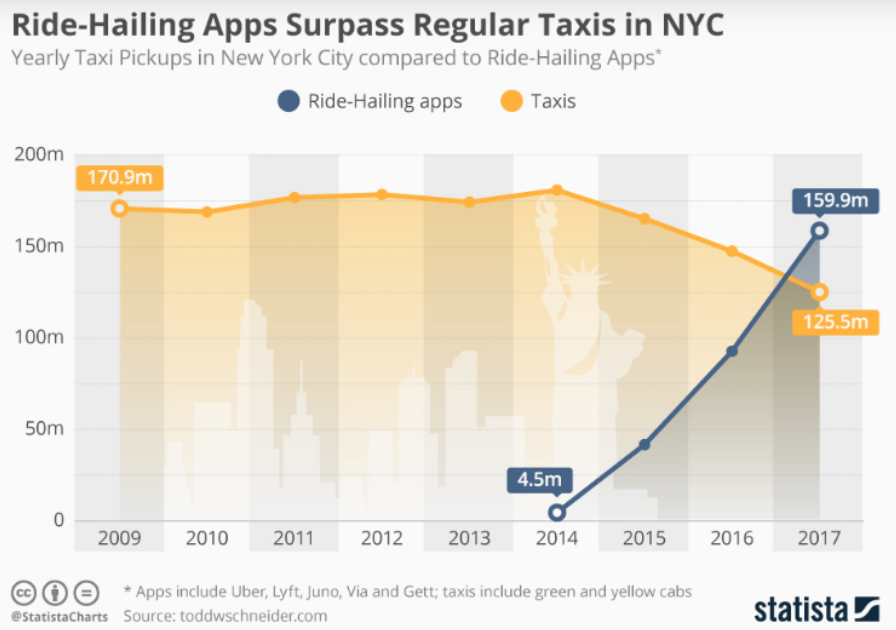


* There’s substantial variation in collisions and collisions reported per day.
* The outliers in the data don’t have an obvious pattern.
* At the daily level, collisions and collisions reported have noticeable differences. Look at the mid-April spike in collisions reported…there’s nothing similar in actual collisions!
* There may be administrative dynamics at play with when collisions are reported or processed.

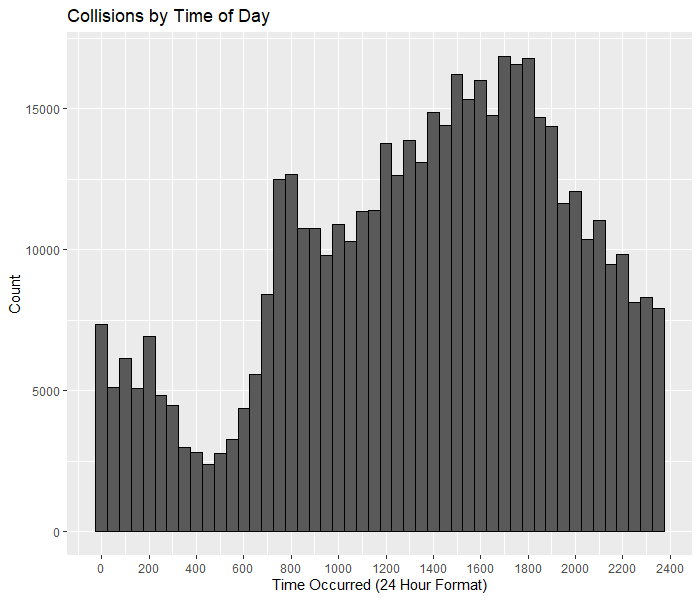
Let’s look at a similar plot aggregated by month. At this level, I can include the entire time frame from 2010-2019.



* Monthly collisions were roughly constant from 2010-2014, rose from 2014 to 2017, and have been roughly constant since.
* The overall trend is the same for collisions vs. collisions reported.
* Questions
  + Why the rise?
    - Population growth in LA?
    - Ridesharing service growth? I couldn’t find data for ridesharing in Los Angeles, but found the following plot for Uber in New York City showing that ridesharing grew drastically from 2014-2017. The full report link is [here](https://buildfire.com/uber-statistics/).

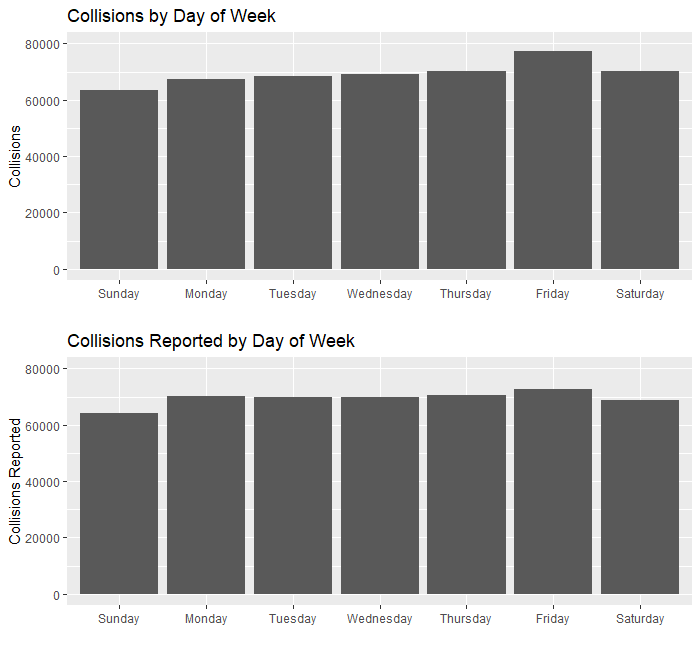


Now, I’ll look at the distribution of collisions throughout the day.



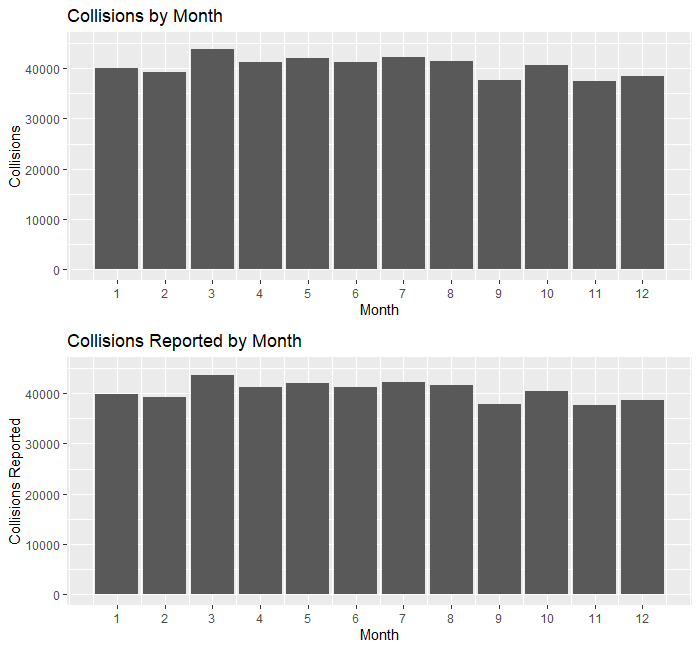
* Collisions are:
  + sharply increasing from 4/5am to 730/8am
  + decreasing from 730/8am to 830/9am
  + generally increasing from 830/9am to 6pm
  + sharply decreasing from 8pm to 4/5am
  + at their daily minimum at 4/5am
  + at their daily maximum at 5/6pm
* These results likely mirror the number of vehicles on the road.
* As mentioned previously, it would be more interesting to have a measure of total vehicles on the road per time to get a measure of collisions per capita.
* No hourly timestamps available for when collisions are reported.

Next up is looking at collisions by day of week.



* Collisions are:
  + increasing from Sun to Fri, with a sharp increase from Thu to Fri
  + at their weekly minimum on Sundays
  + at their weekly maximum on Fridays
* The end of the week is the obvious hypothesis for the high number of Friday collisions. I haven’t come up with any others so far!
* For collisions reported:
  + Sun / Fri is still the lowest / highest
  + the weekdays are pretty even

Finally, I look at collisions by month.



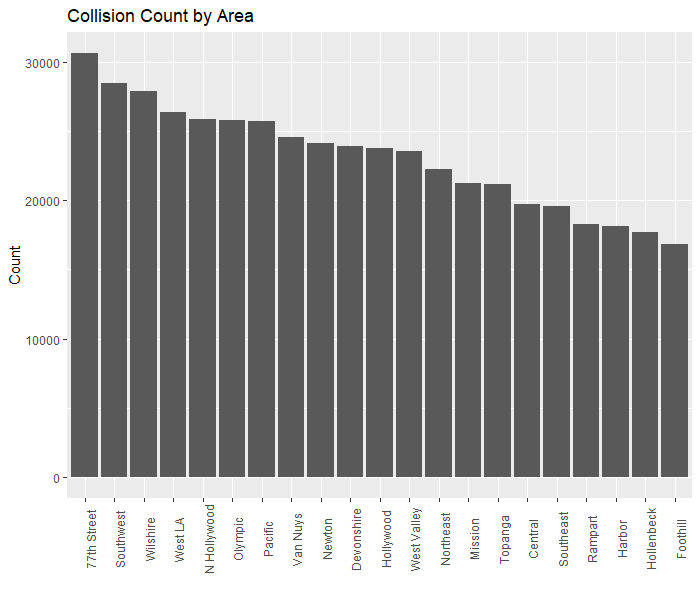
* Collisions are:
  + generally constant from April to August
  + generally lower from September to February
  + highest in March
  + my first guess was Daylight Savings Time. However, none of the high-collision outliers occurred on this date.
  + Spring break tourists?
* Collisions are lower in the colder months.
  + Less tourists?
* Collisions and collisions are very similar at the monthly level.

Given all of these plots, here are my takeaways for the temporal pattern of collisions:

* Collisions and collisions reported vary substantially at the daily level, but not at the monthly level.
* Monthly collisions were roughly constant from 2010-2014, rose from 2014 to 2017, and have been roughly constant since.
* The number of collisions is lowest at 4/5 AM and highest at 5/6 PM.
* The number of collisions is lowest on Sundays and highest on Fridays.
* The number of collisions is highest in March and lowest from September to December.

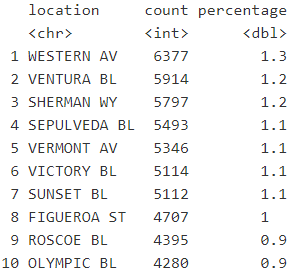
**COLLISIONS BY GEOGRAPHY**

The next topic I want to cover is analyzing collisions geographically. Before getting into mapping, I start by looking at the distribution of collisions by “area”, a field provided in the data set.

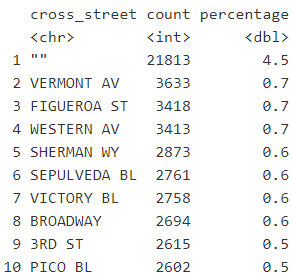


* Some areas obviously have more collisions than others.
* But this plot would be more informative with a measure of area size or traffic density.

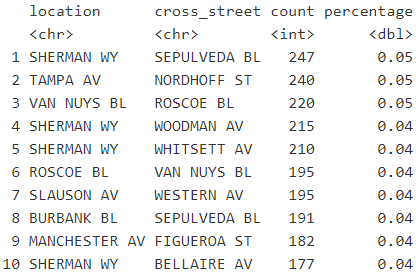
The data also includes fields called “location” and “cross\_street”. “location” is the main street a collision occurred on, while “cross\_street” is the nearest cross street. I’ll look at the 10 most common values for these fields and their combination.



* The 10 most common “location” are some of the longest and most used roads in LA.
* These top 10 streets account for >10% of total collisions. There are >25K total “location”, so there’s a very long tail.

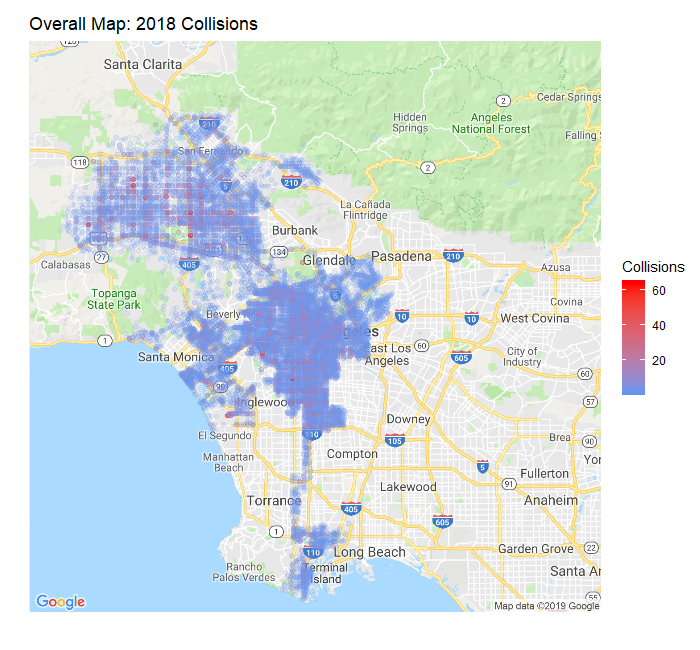


* 5% of collisions have no associated “cross\_street”.
* Otherwise, this list has a lot of overlap with the previous list.



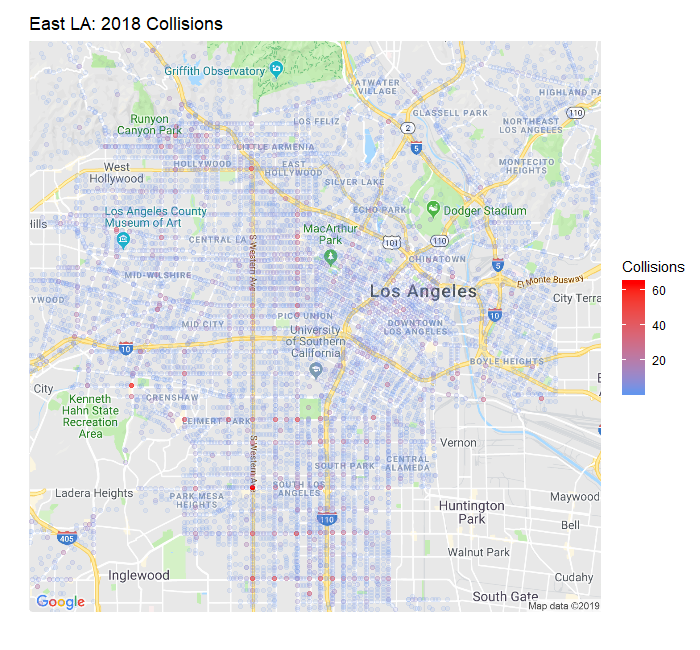
* The most common “location” / “cross\_street” combinations contains many of the streets we saw in the previous 2 lists.
* However, there are exceptions: the components of row 2 (Tampa Ave. and Nordhoff St.) don’t appear in either the most “location” or “cross\_street”.
* Even the most collision-prone intersections account for a small proportion of overall collisions.

Now I’ll move on to mapping.



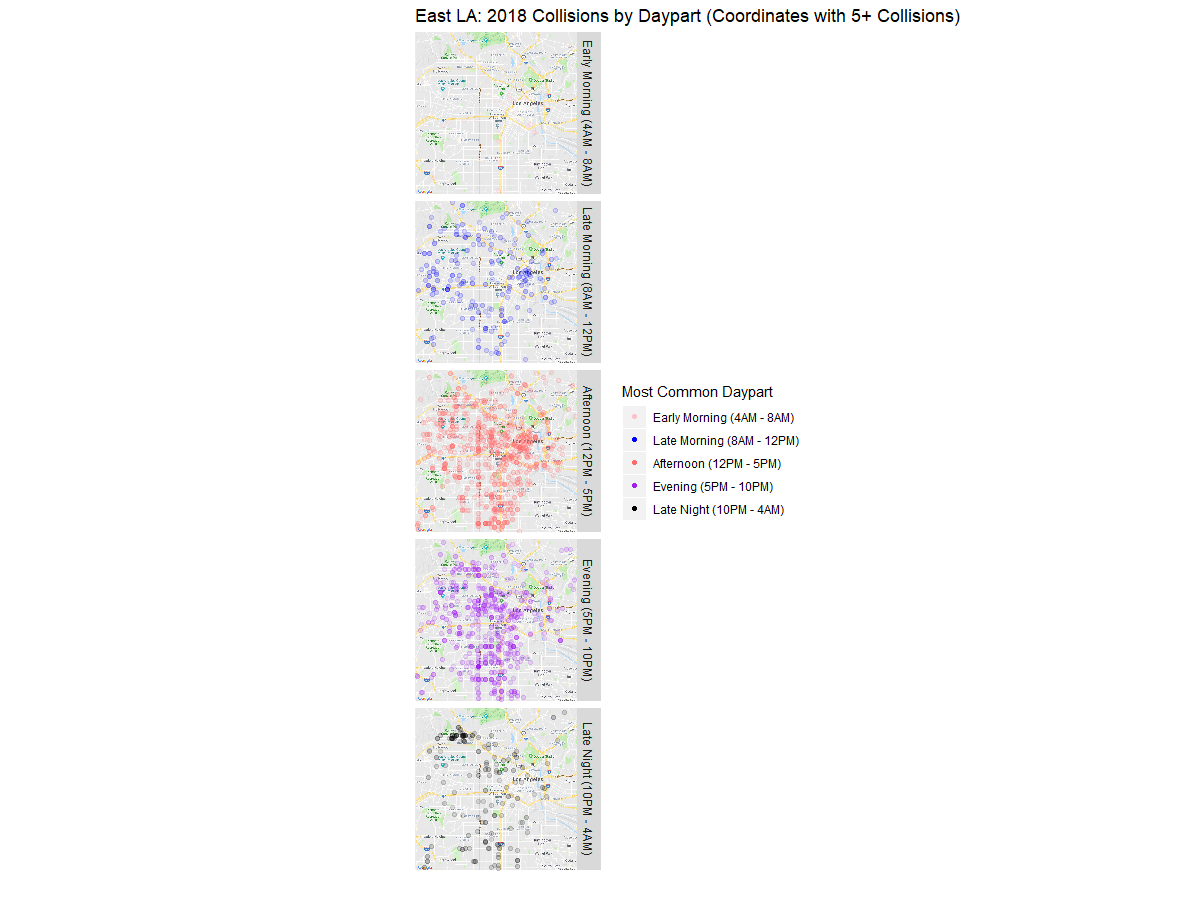
* This plot shows the interesting shape of Los Angeles.
* Blue / Red points indicate coordinates with a low / high number of collisions.
* Coordinates with a low number of collisions (blue points) are rendered with low intensity and look faded.
* Even on this zoomed-out map, I can see high collision coordinates in the Valley and East LA.

To get a better view, I’ll zoom in on one area.



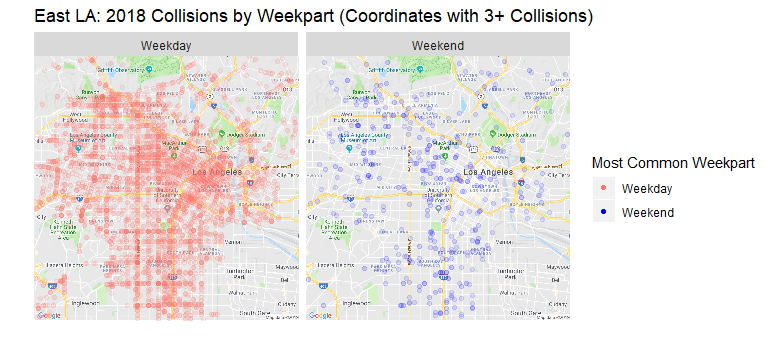
* This plot shows much of central and downtown LA.
* Many high and medium collision coordinates are clearly visible.
* Some main streets seem to have high or medium collision intersections occurring quite often.

Next, I layer time of day onto the map.

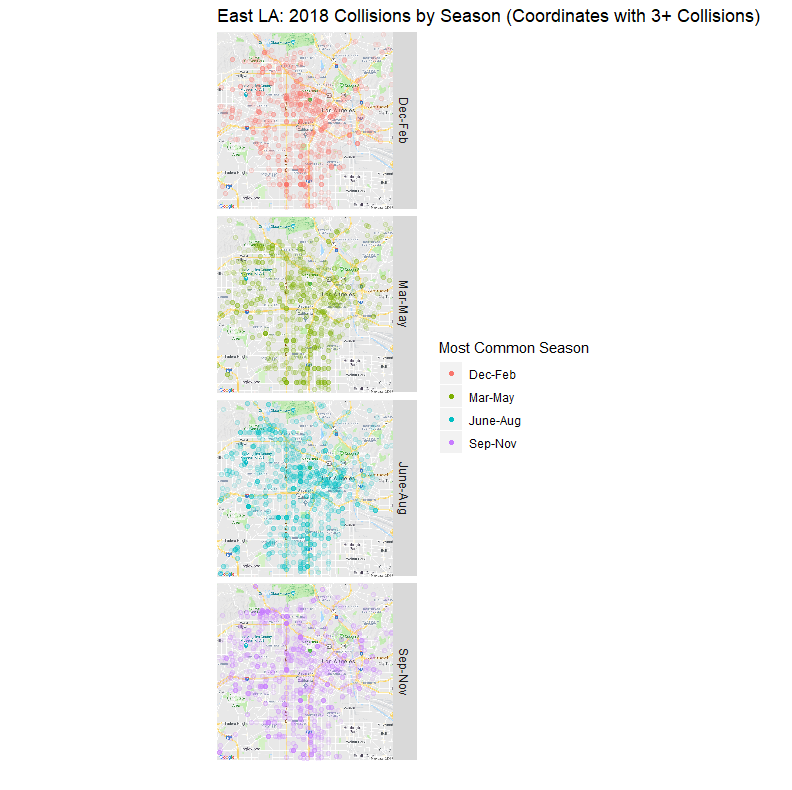


* This plot only includes coordinates with 5+ collisions. Each coordinate is assigned to the daypart where most of its collisions occur in. Coordinates with ties are thrown out.
* There are no coordinates with a majority of collisions occurring in the early morning. This makes sense given the Collisions by Time analysis above.
* The afternoon and evening are when many collisions happen. We also saw this in the Collisions by Time analysis
* Interestingly, we can see a cluster of coordinates where late night collisions are common.

Let’s look at a similar map broken out by weekday/weekend.



* There are obviously more days and collisions in the “Weekday” bucket.
* This map can be used to identify areas with many weekend collisions.
* It might also be interesting to include part or all of Friday in the “Weekend” bucket.



* It looks like more collisions occur in the summer months (Mar-Aug). This would line up with the results in the Collisions by Time section.
* Given that Los Angeles doesn’t have distinct seasons like fall or winter, there may be other ways to split up the year for a plot like this.

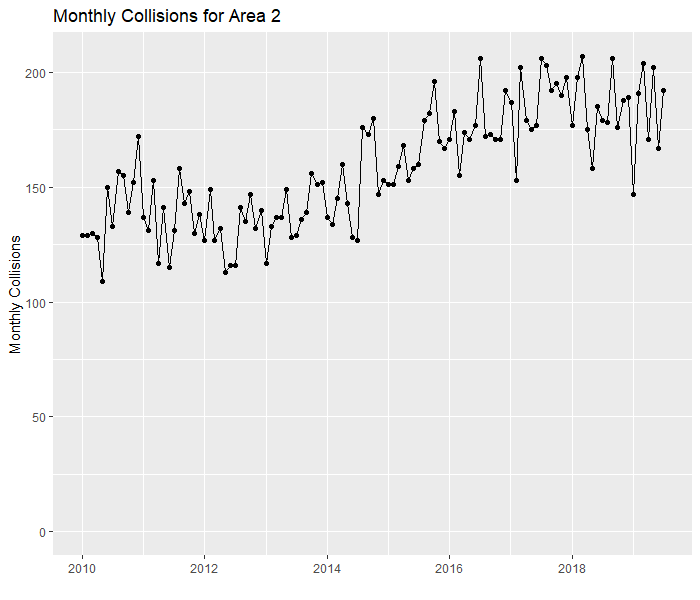
Here are my conclusions for the Collisions by Geography section:

* Through the *location* field, *cross\_street* field, and mapping it’s possible to identify the most accident-prone coordinates in Los Angeles.
* The most common daypart for collisions is the afternoon or evening. Mapping collisions shows areas where particular dayparts are most common.
* Many more collisions occur during the weekend than the weekend. Mapping collisions by weekpart shows areas where weekend collisions are more common.
* More collisions happen in the summer than winter months.

**COLLISION PREDICTION**

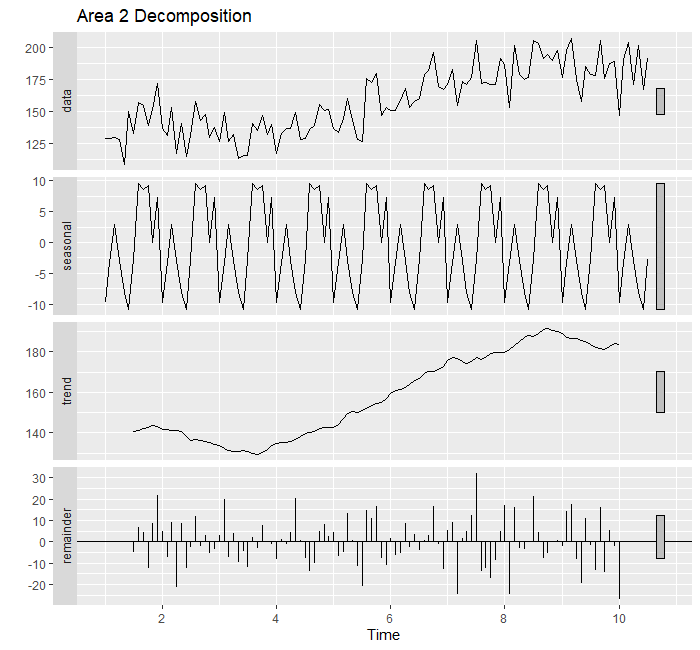
The final section deals with trying to predict collisions. I specifically try to predict the number of collisions that will occur per month and area.

I’ll start by looking at an example of the collision time series for one area:



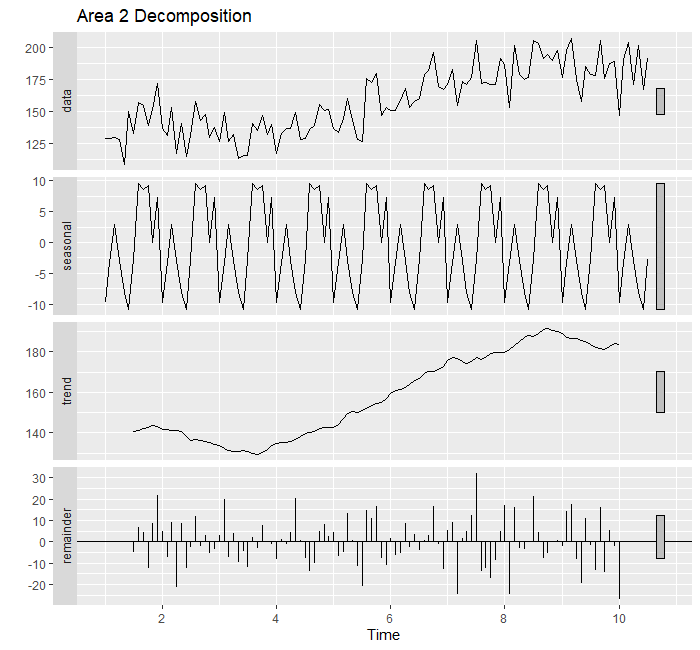
* The trend for area 2 generally matches the trend of overall monthly collisions

Let’s decompose this time series into trend, seasonality, and remainder components.



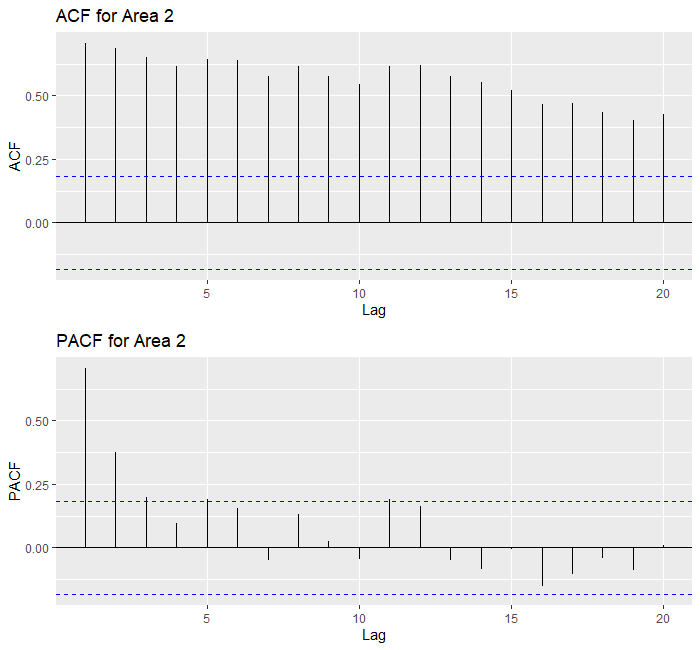
* Keep the shape of the seasonality curve in mind. I’ll compare it against overall monthly collisions next.
* The trend looks generally as expected.

Let’s compare the area 2 decomposition to the overall monthly collisions decomposition.

****

* The seasonality curve for the overall data looks very different! This is my first indication that different areas can have different dynamics.

Next, I look at more exploratory plots for area 2.

****

* The ACF analyzes how correlated lagged values of area 2 collisions are to the current value.
* The PACF analyzes how much previously unexplained variance each lag explains.
* These plots indicate that any model predicting area 2 collisions should include at least 2 lagged terms.

I try a few different model types:

* 3 and 6 month moving average models (MA)
* ARIMA
* Prophet

MA models average past values to generate predictions and are the simplest time series model. ARIMA models can use past values, differencing, and previous errors. Prophet is an additive forecasting model where non-linear trends are fit with yearly and monthly seasonality. I evaluate each model on the final 12 months of data (August 2018 to July 2019) with the following metrics:

* MAPE: average absolute percentage difference between prediction and actual
* Bias: average percentage difference between prediction and actual

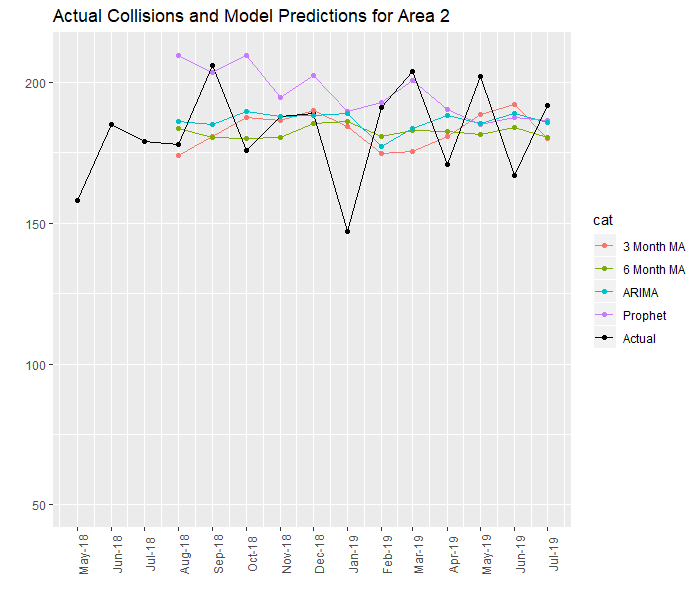
The MAPE gives me an idea of how far my predictions are from actual collision values while the bias lets me know if I am systematically over- or under-predicting the data.

Here are the overall model results:



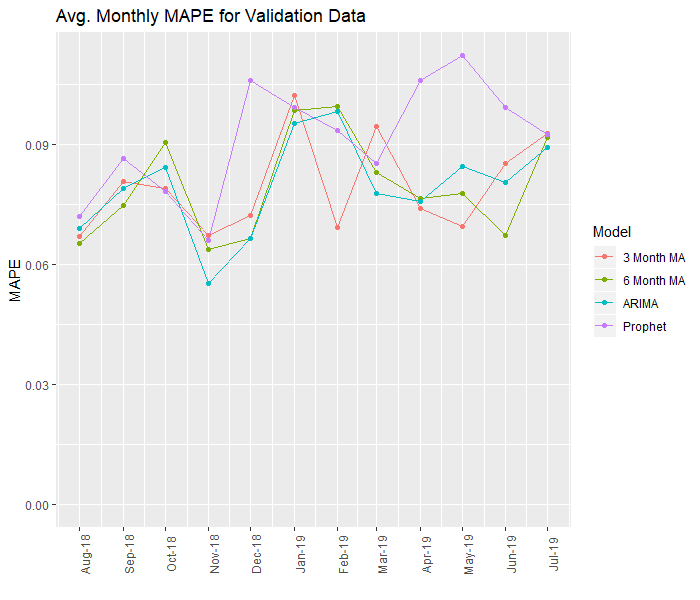
* This table reflects:
  + Auto-fit ARIMA models for each area. So each area can have a different p, d, and q.
  + The best result of multiple Prophet model specifications.
* The 3 and 6 month MA models have very similar results.
* The ARIMA model has a similar MAPE to the MA models, but worse bias.
* The Prophet model has the worse results by far (even after tuning).
* It’s surprising that the MA models have the best performance!

Let’s look at the model predictions for area 2 only.



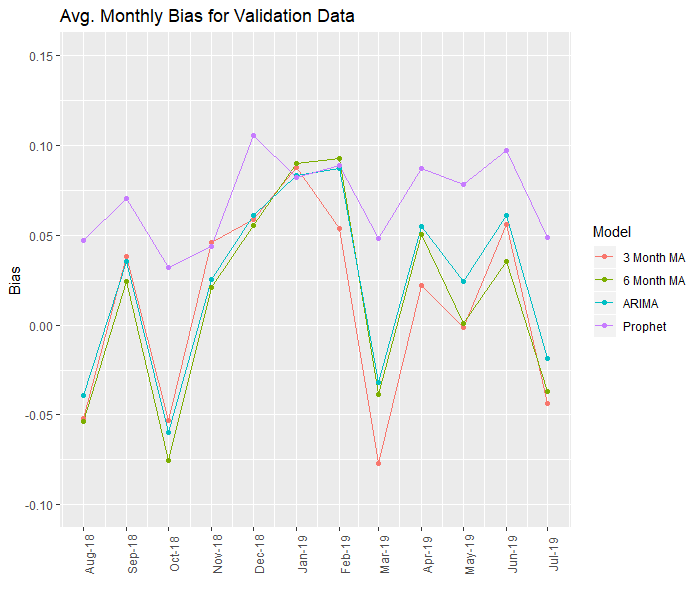
* The Prophet model has much higher predictions than other models for the first 5 months.
* All models miss the drop in Jan 2019 and the up/down pattern of April 2019-July 2019.
* The 6 month MA model predictions don’t vary that much.
* The MA and ARIMA models seem to make conservative predictions that don’t capture the fluctuating nature of the data.

Next, I’ll look at the average MAPE per month per model.



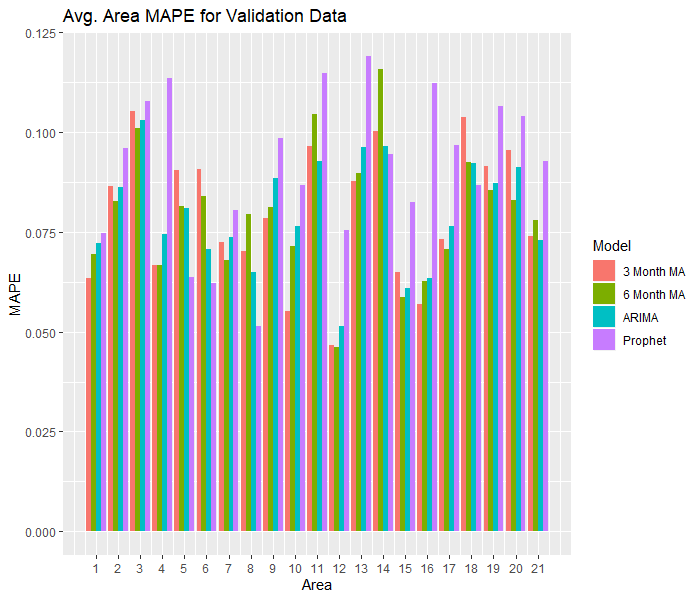
* For the first half of the validation data, the MA and ARIMA models move together.

It’s also worth looking at the average bias per month per model.

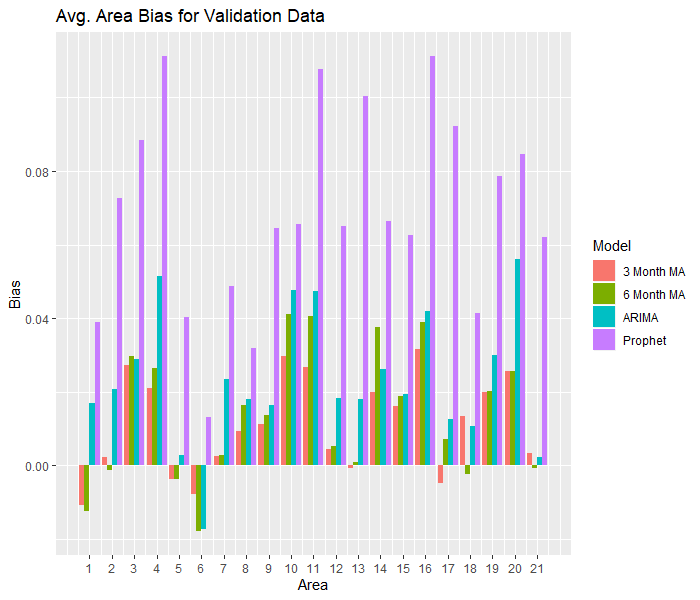


* The bias for the MA and ARIMA models moves together.
* The Prophet model has positive bias (overpredicts) for the entire validation set.

Now, I’ll look at the average MAPE and bias per model per area.

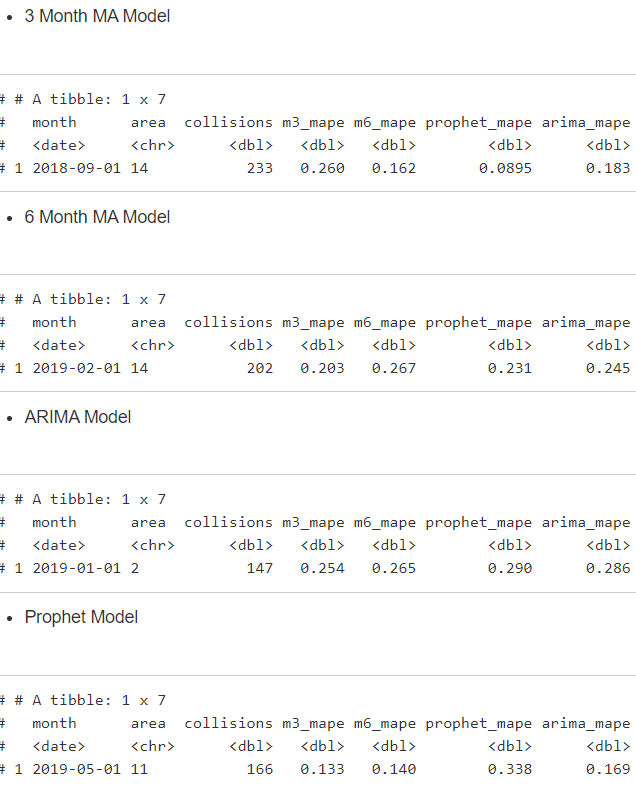


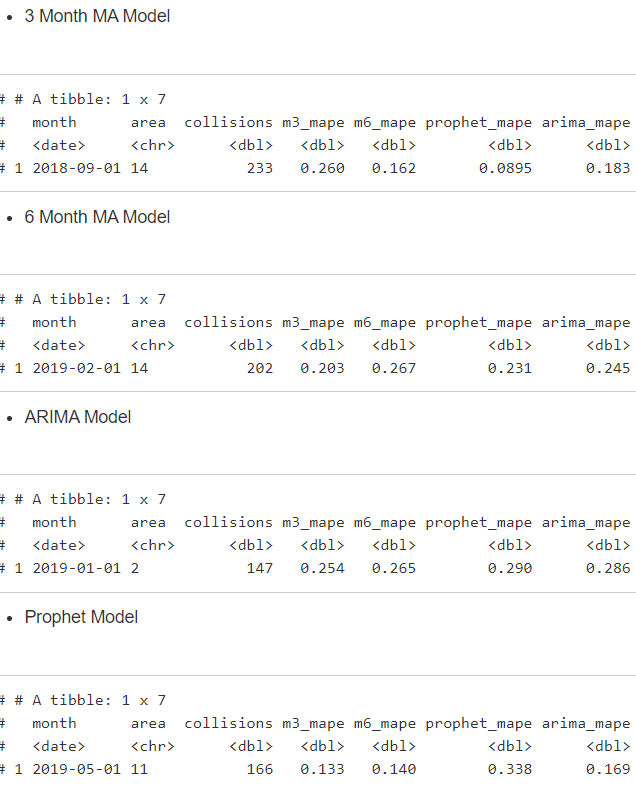
* Average MAPE performance varies substantially by area. For example, model performance on area 12 [3] is pretty good [bad].

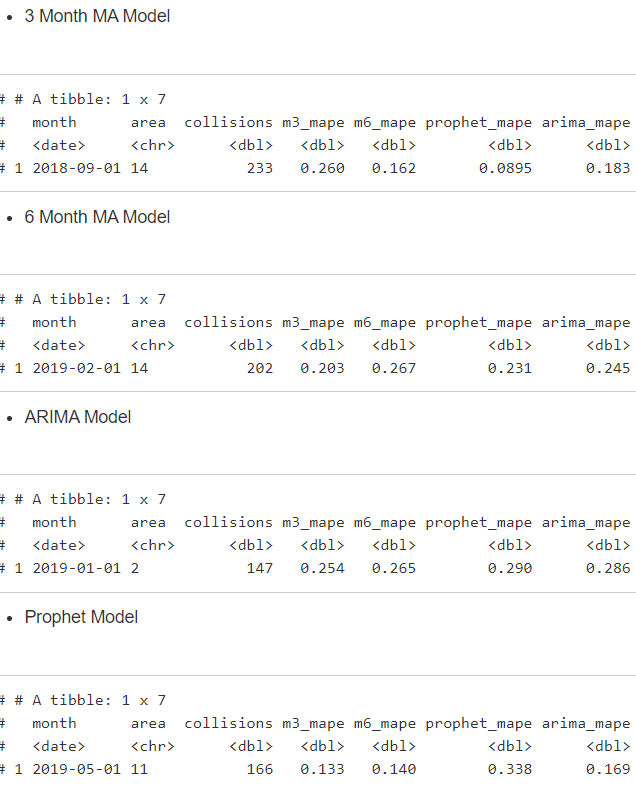


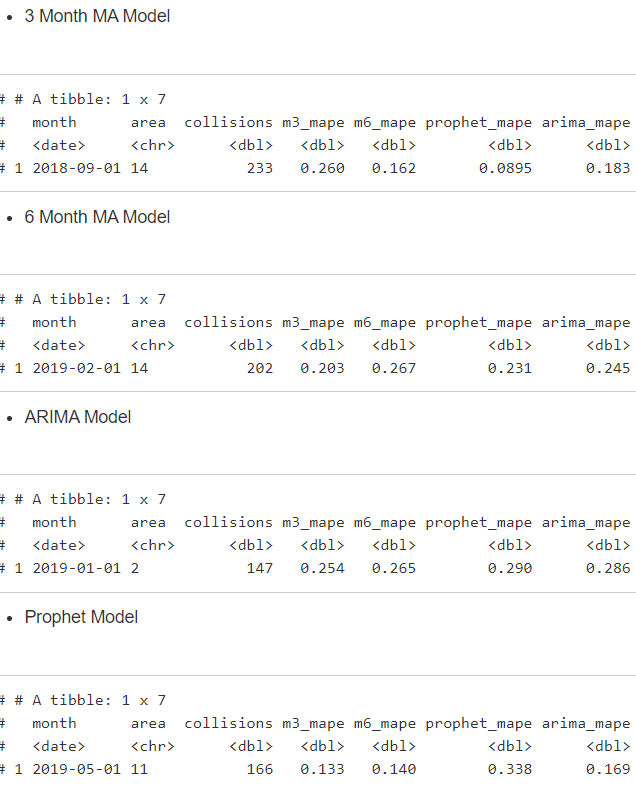
* Average bias performance also varies substantially by area.
* Surprisingly, most models have positive bias (overpredict) most areas!

Next, I’ll look at the worst area/month predictions per model.



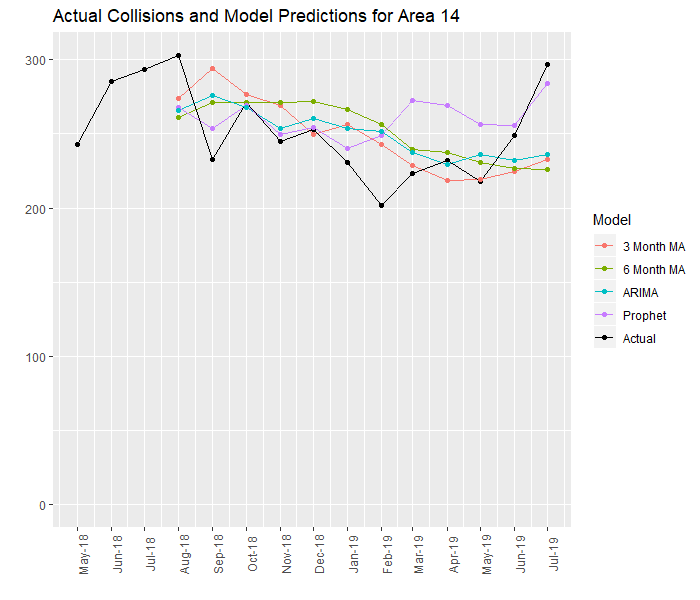






* Area 14 shows up twice
* Both Jan and Feb of 2019 show up
* It’s interesting to see cases where all models struggled (Jan 2019 in area 2) vs. cases where one model in particular struggled (Sept 2018 in area 14).

Based on the worst predictions, I’ll zoom into area 14.



* The trend for area 14 is completely different than for area 2 above.
* All models except Prophet miss the spike in July 2019.

Here are my conclusions for the Collision Prediction section:

* Overall monthly performance is not bad (<10% MAPE and bias in most cases).
* However, MA models having the best performance indicates that longer-term lagged data, differencing, and previous errors doesn’t help performance.
* This is a surprising result, but suggests that the number of collisions per month/area is largely random within a certain range.
* Trends seem to vary by area. This should in theory be addressed by the ARIMA and Prophet methods which fit a separate model per area.

**NEXT STEPS:**

Here are the ways I would extend this analysis:

* Get a measure of per capita collisions. To do this, I need to know how many vehicles are on the road by time, day, gender, etc. This would add context to much of the Collisions by Time section.
* Additional work on Collision Prediction. External data sources (such as weather) may be helpful. I would start by zooming into 1 area and trying to understand the dynamics affecting actual collision values.

Concerns

* Catch-all age
* Multiple victims
* Unreported minor incidents