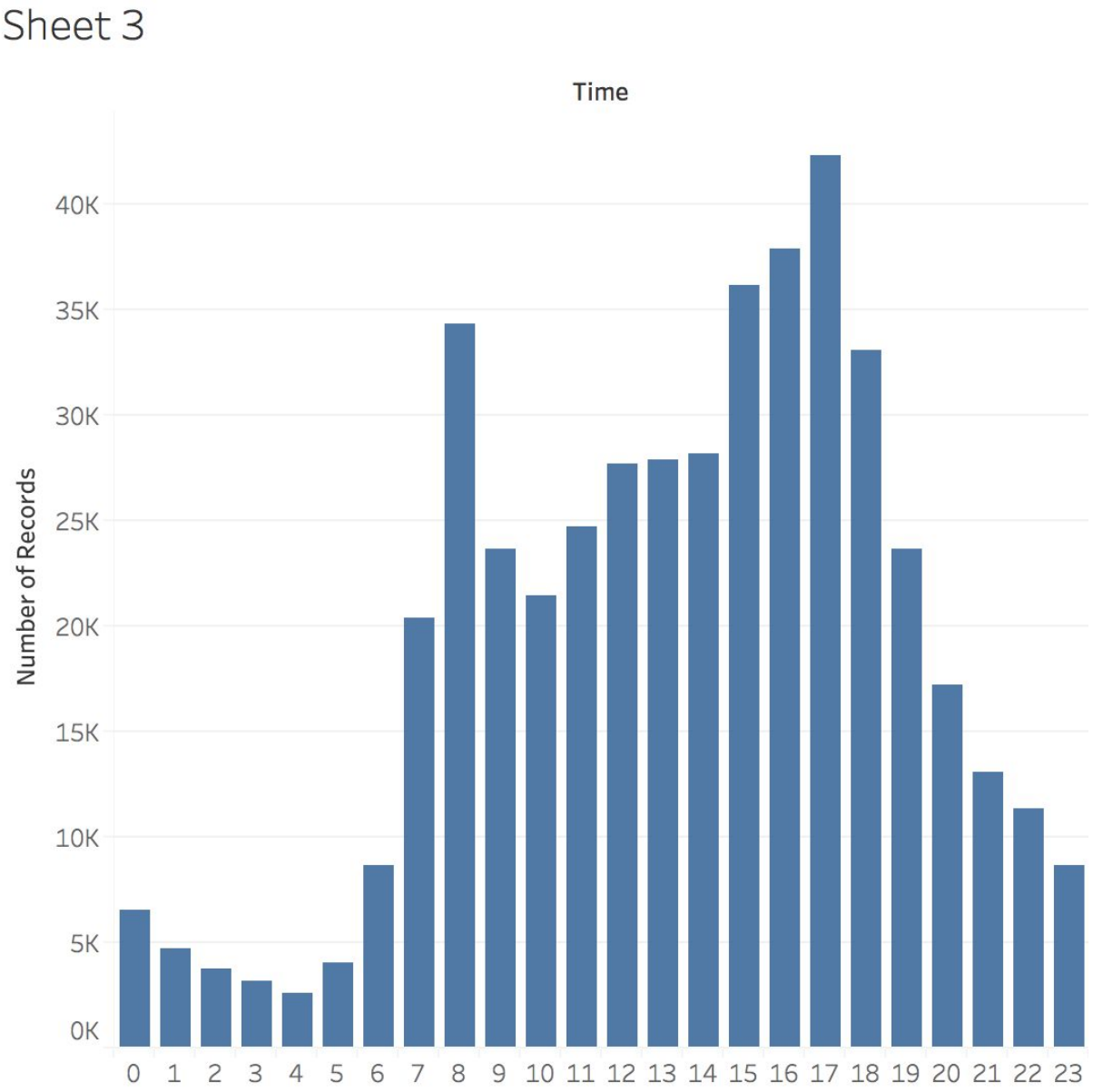
Traffic accidents cause a huge number of deaths and injuries. According to the Infographics report produced WHO [1], there are 1.35 million road traffic deaths every year, making trafic injury the 8th leading cause of death, and 1st among people aged from 5 to 29. Thus I am interested in the relationship between conditions and traffic accidents. So my initial question is, **In which conditions, there are more likely to be traffic accidents. Or say, how people can change some conditions, to reduce traffic accidents.**

Datasets: Through online searching, I found a suitable dataset contains traffic accidents in the United Kingdom from the year 2005 to 2014, which is collected and published on Kaggle [2], and owned and originally published by UK Department for Transport [3].
This dataset contains traffic accidents in year 2005 to 2014 and every 3 years are grouped in a CSV file. As the number of records is huge (1.6 million in total), I picked only the most recent data from 2012 to 2014 (0.46 million) for analysis.
In this dataset, each car accident is recorded as one entry. The location/coordiante of the accident, the number of police force used for that accident, a severity category from 1 (most severe) to 3 (least severe), number of vehicles, number of casualties, date and time, local authority information, road type, road number and class, rural or urban area, speed limit and junction control is included for each accident (with some missing data).
Before I start plotting the data, I checked that there are no missing values for severity, and only date/time information has 13 missing values, compared to the total number of 464697 entries, I would say they can be safely ignored.

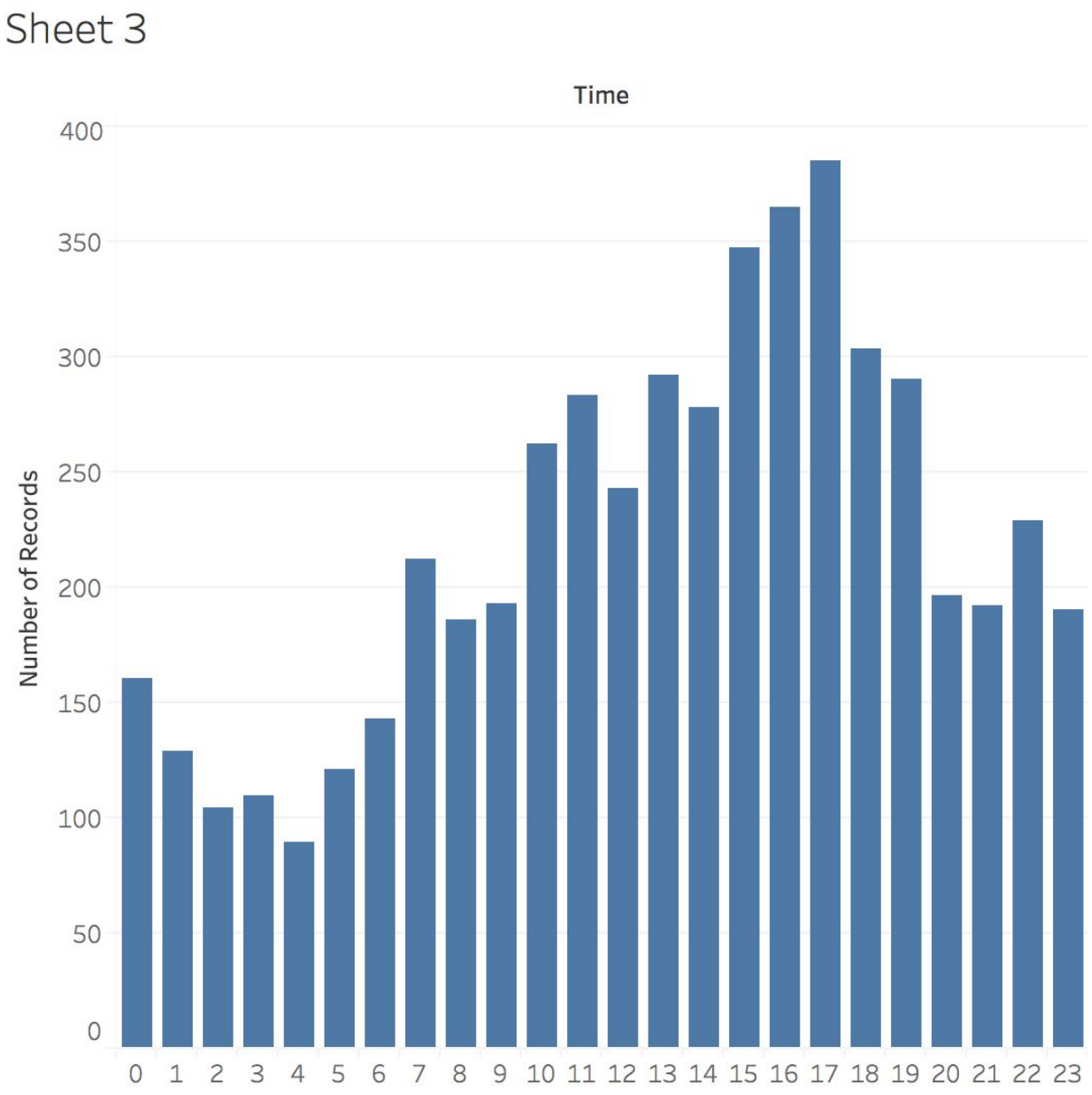
Explore the data:
To start answering the question, I started with plotting a histogram graph which the x-axis is the hour the accident happened.



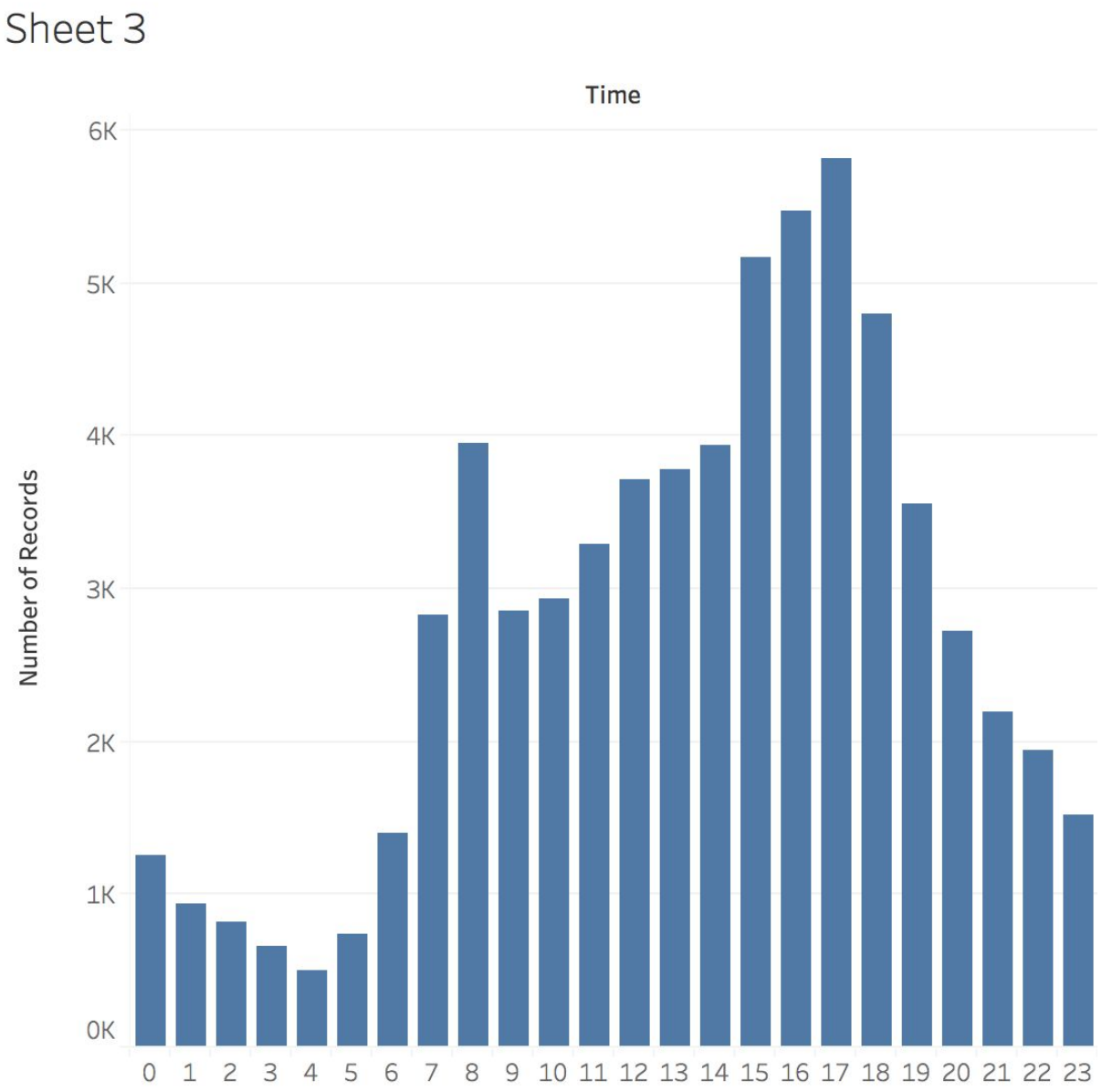
We can observe the trend that the number of accidents is strongly associated with time, wherein the day time there are more accidents and in the night there are fewer accidents. This should be due to the reason that there is more traffic in the day time than at night, thus there are more traffic accidents in the day time. I cannot find hourly traffic information online, thus I moved to explore the severity.

I plotted a histogram for time\severity this time. As there are too many level 3 (least severe) severity accidents, I separated into 3 plot by the level of severity.

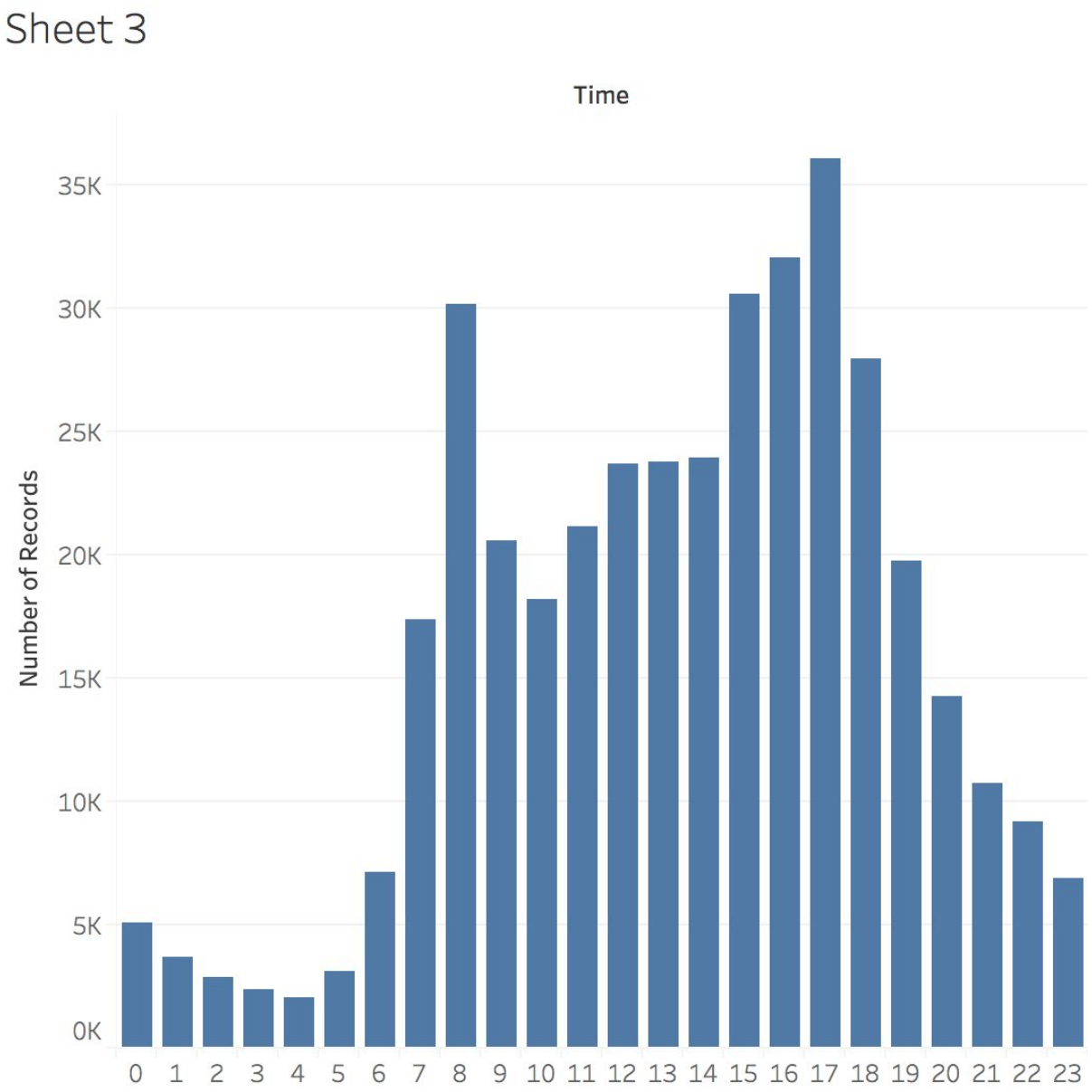
The first one is level 1 severity:



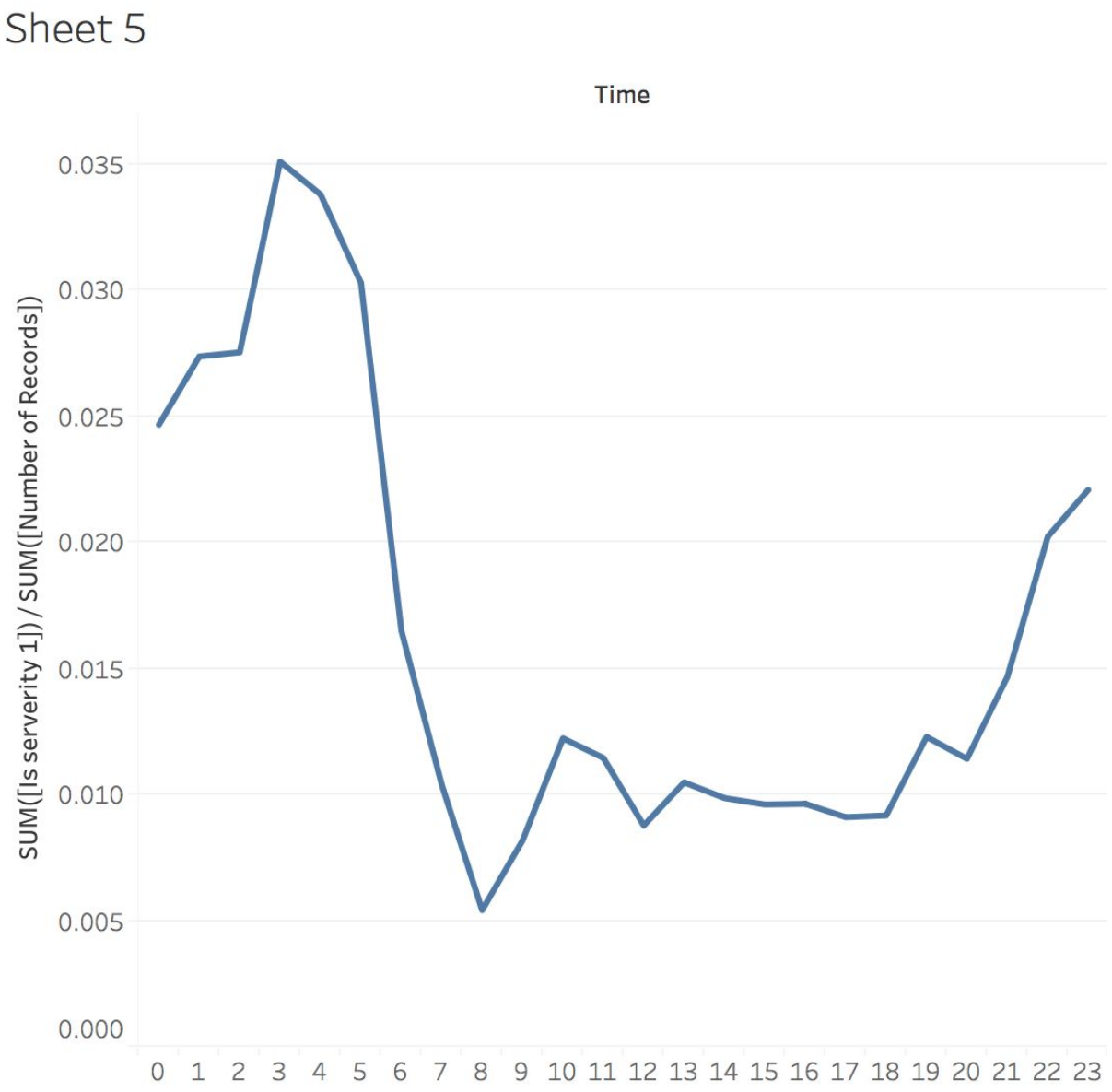
Level 2 severity:



Level 3 severity:



Compare the 3 graphs of different severity level, I found a very interesting observation is that, the histogram for level 1 severity by hour is more “flat” than the other severity levels. Does that mean there is a higher percentage of level 1 severity accidents in late night?
To confirm this, I come up with this plot:



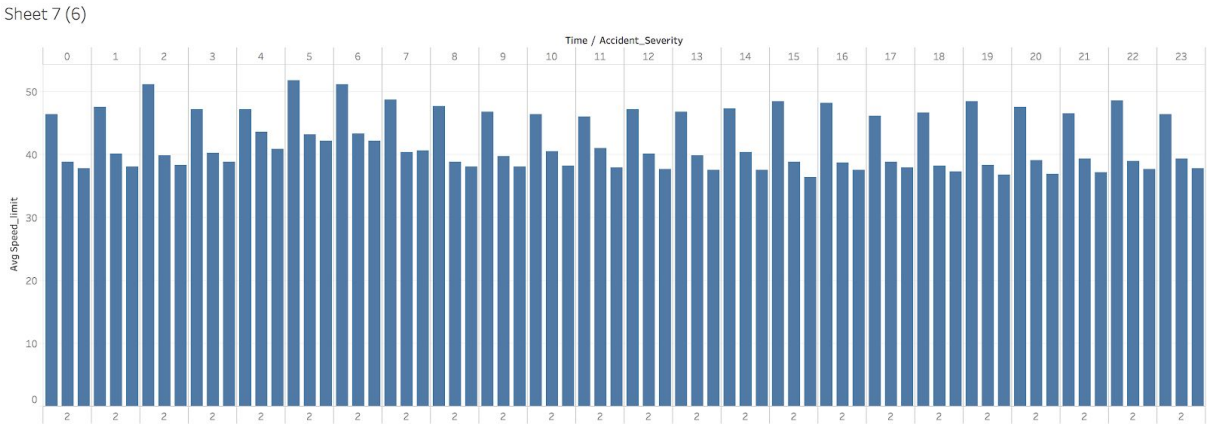
This is the percentage of level 1 accidents among all accidents at that hour. A higher rate means there are more percentage of severe accidents in that hour. The peak is at 3 am, where 3.5% of accidents are level 1 accidents, and it is almost 7 times of the rate at 8 am. We can see that the percentage keeps generally the same from 9 am to 7 pm, and then increase from 8 pm to the peak 3 am, and then decrease from 3 am to 8 am.

The graph clearly shows that severe accidents are strongly associated with time. However, it is impossible for people to choose what time to travel. If they have a need, they need to travel, no matter the time.

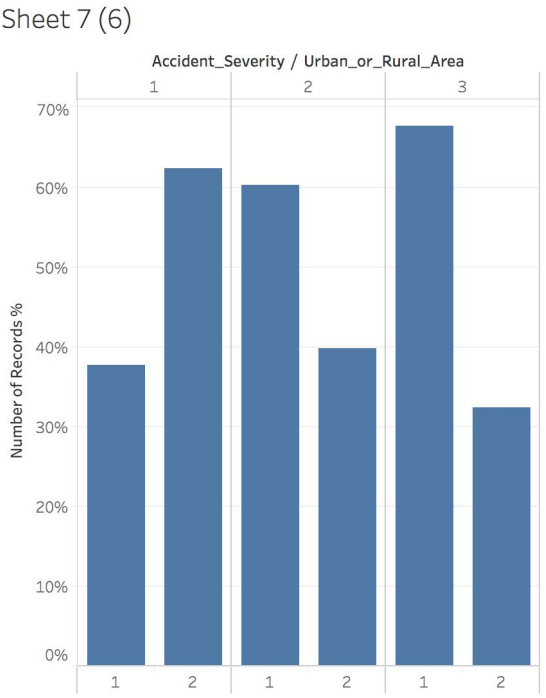
However, the cause of high severe accidents rate at late night is still unclear. Is that because of the speed limit? Is that because the rural-urban area they are driving? Is that because of the road type people traveling with? Is that because of the weather condition? Is that because of no lights?

So now I want to rephrase the question and to be more specific, **What is the cause of high severe accident rate in late night?**

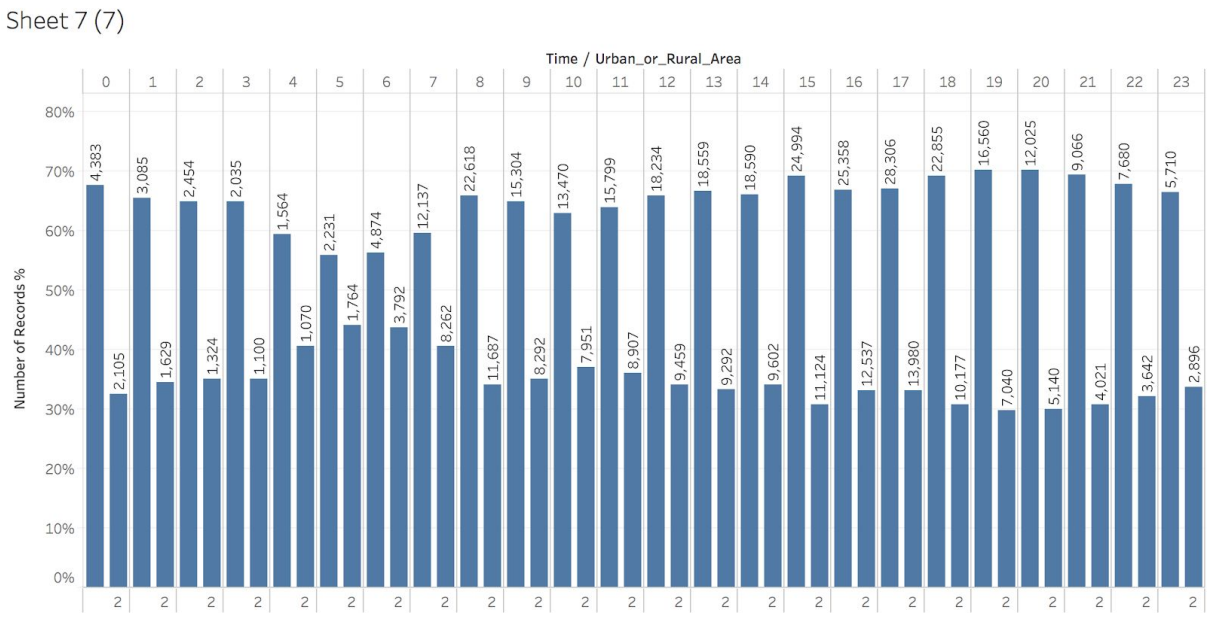
I started with doubting that a higher speed limit at night might be the cause of a higher severe accident rate. So I come up with this plot of the average speed limit by hour and level of severity.



It seems there is no significant difference between the average speed limit of 3 am and the day time. Although the speed limit at 2 am, 5 am and 6 am are indeed higher than other times, the difference is not significant, compared to the difference in the server accident rate. Then I moved to discover about the rural/urban area. First I plotted this graph about the percentage in urban (encoded as 1) and rural (encoded as 2) area for different severity:

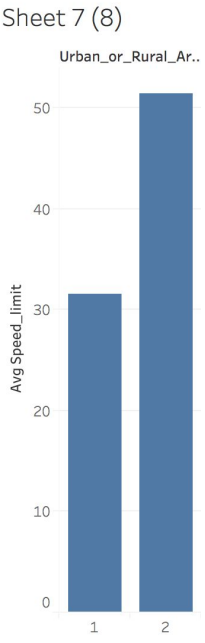


It is true that level 1 accidents are more likely to happen in rural areas. Is that the true cause of why severe accidents rate is higher in late night? I plotted the percentage of rural/urban driving for hours:

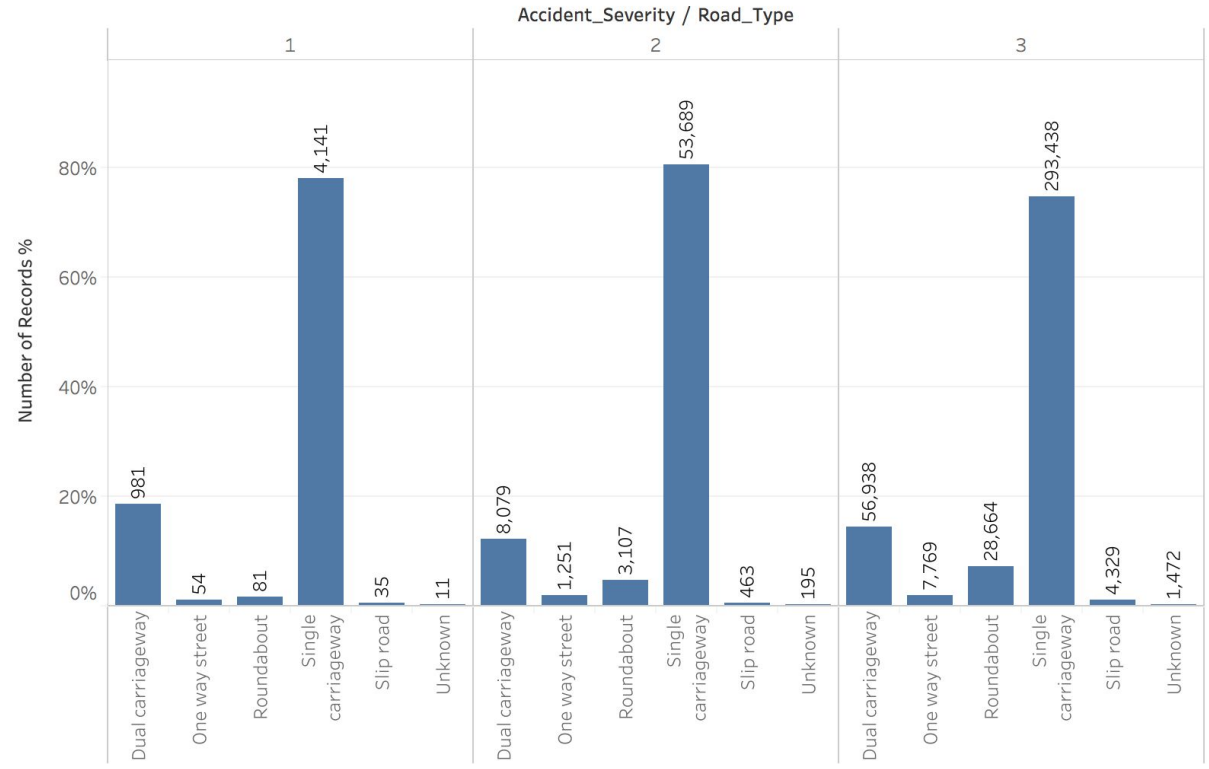


It seems the percentage of rural driving is indeed higher in late night, but the peak is around 5 am instead of 3 am. I would say it would be a possible factor causing the higher severe accident rate, but it is hard to tell.

A possible explanation of why severe accidents are likely to happen in rural areas is that there is less traffic in rural areas, making the driver less cautious about other traffics, and the speed limit in rural areas is higher than in urban areas. The following plot of average speed limit in urban and rural areas would support that “the speed limit in rural areas is higher than in urban areas”:

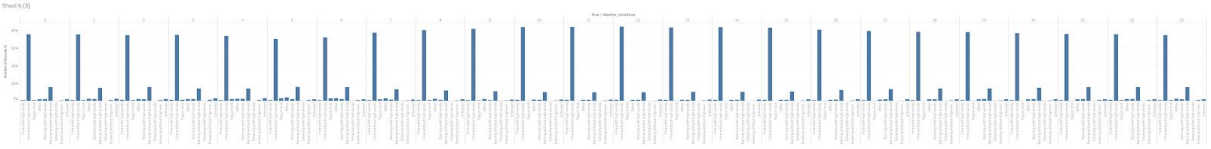


Then I moved to road type. Does road type have a correlation with severity? I made the following plot of percentage of road type for different severity level(the label is the absolute number of entries):

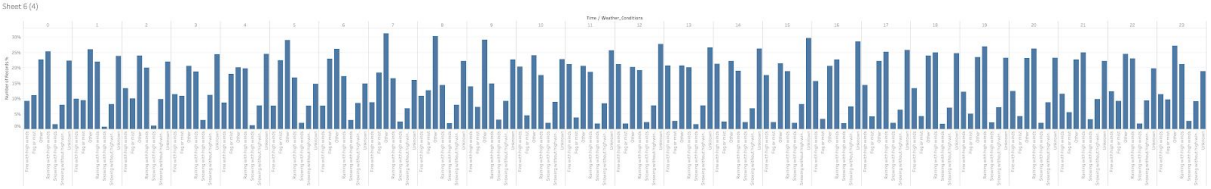


There is no significant difference in road types of different severity levels. Only 1 side observation is that for the roundabout, it is more likely to be not so severe accidents, maybe because they are mostly in urban areas or the speed is low.

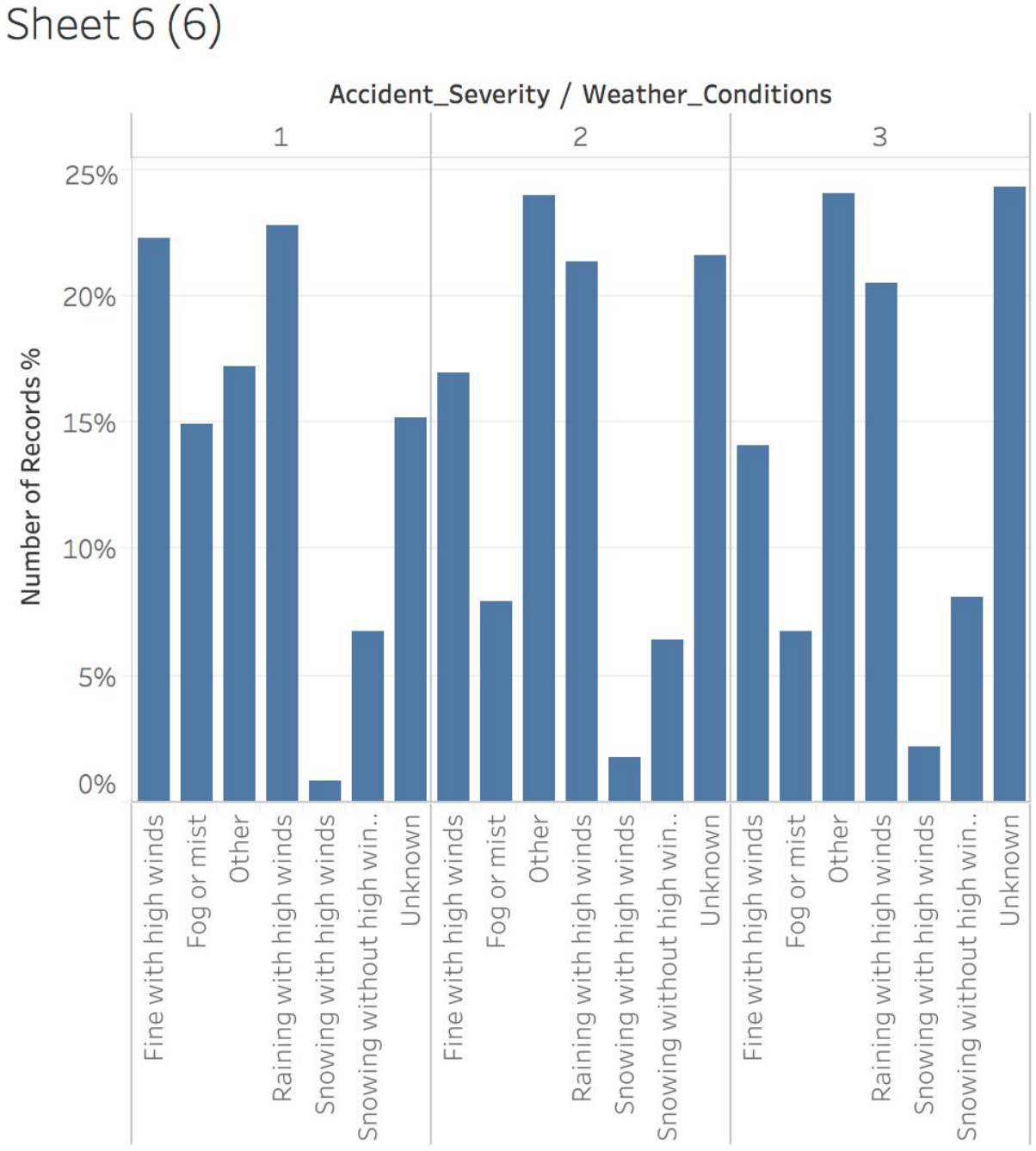
The next candidate is the weather conditions. First I tried to make a plot for percentage of weather condition of different hours:



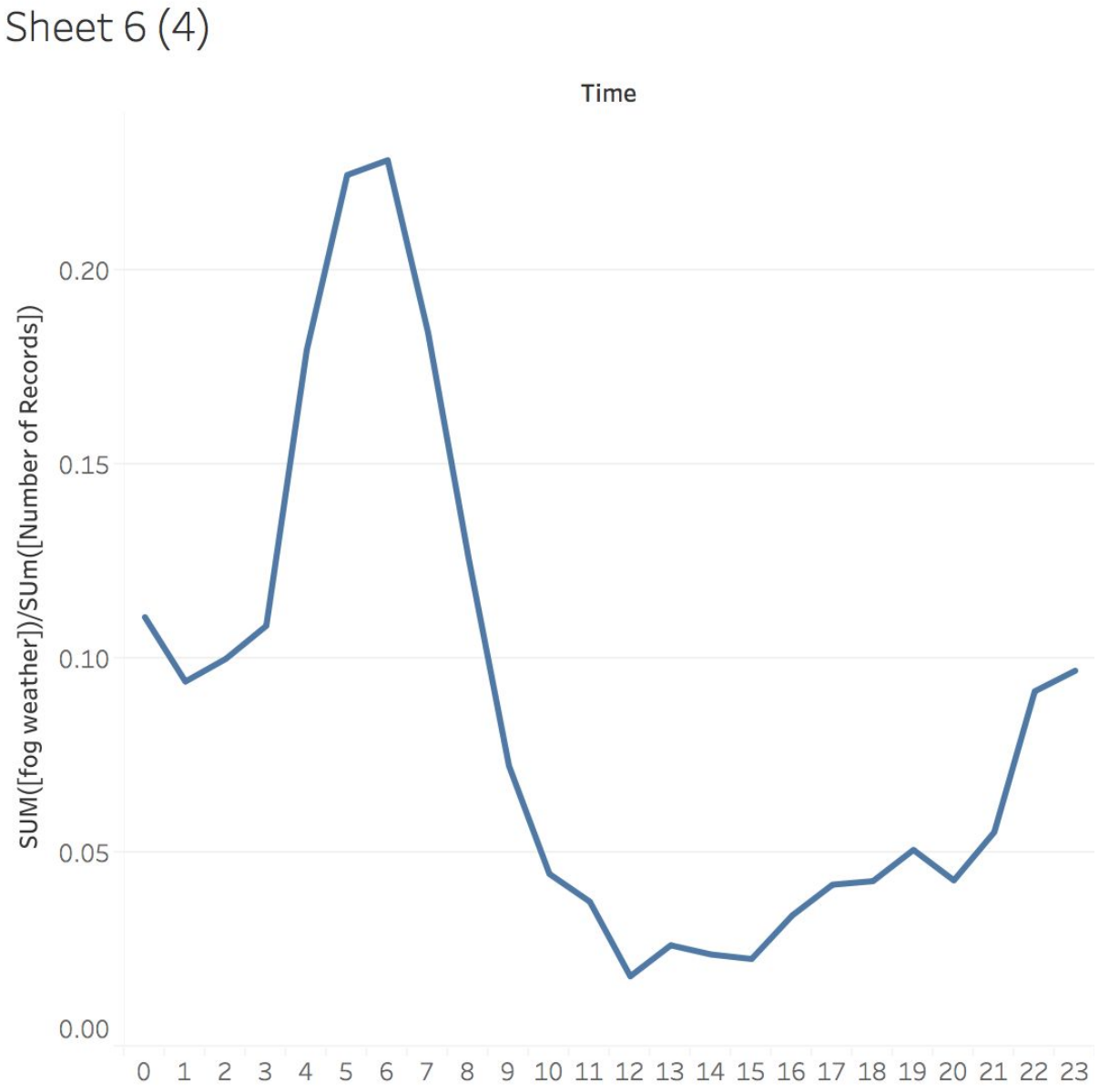
(I know it is too small to see, the longest bar is always “fine without high winds”)We can see that most of the cases are all “fine without high winds” and “rainning without high winds”, and they are all almost similar percentage, so I decided to plot it again without these two common cases:



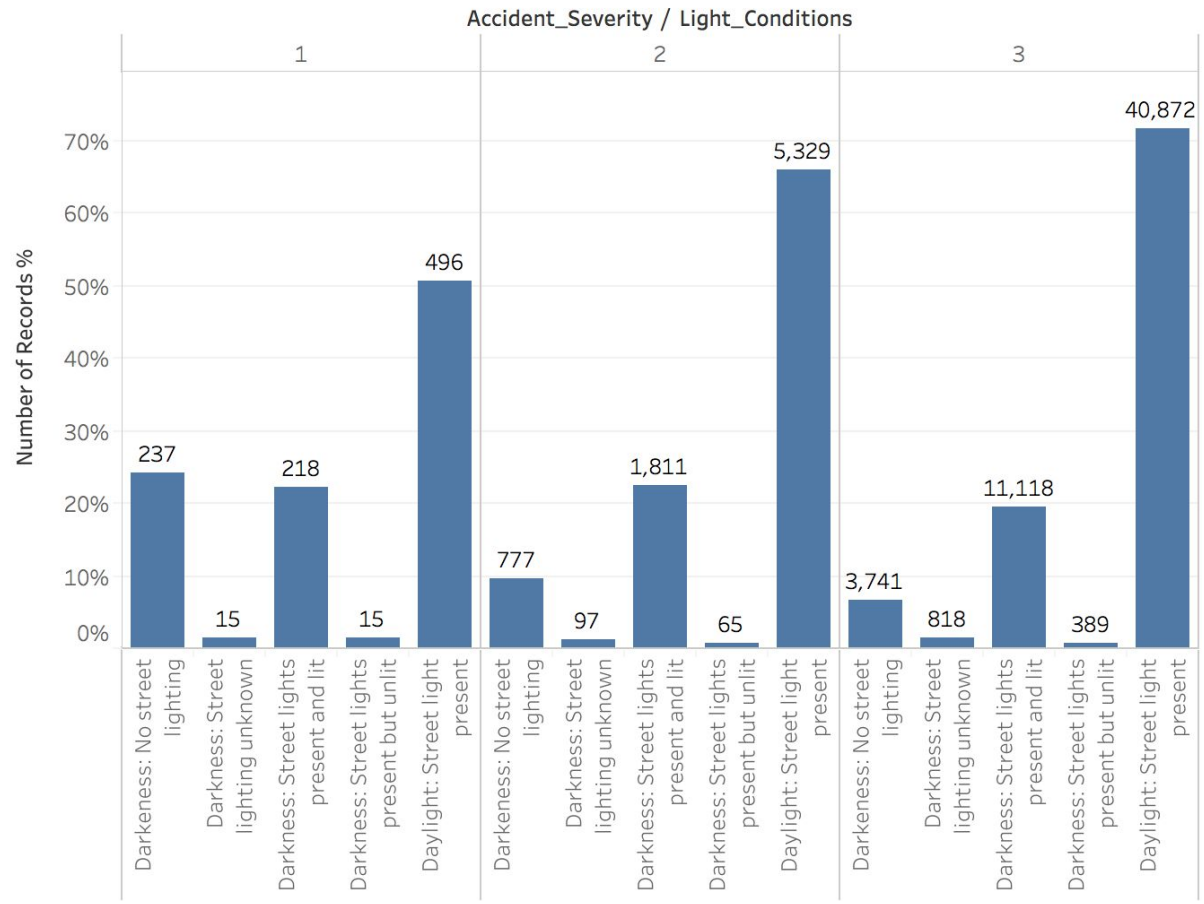
Although it is still too small to see, there are no much big variations between different hours. And to be sure, I also plotted the percentage of weather conditions for different level of severity:



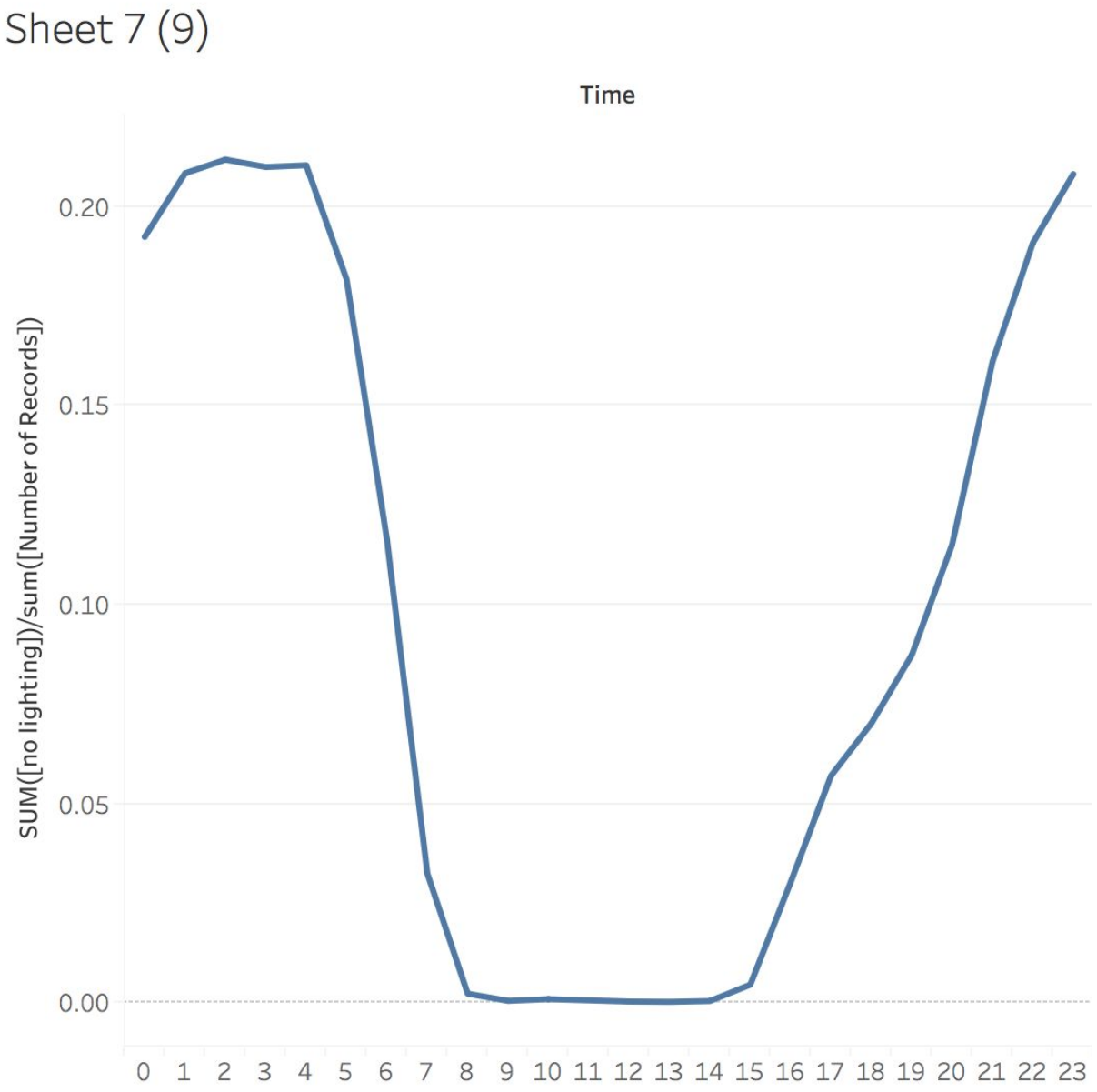
We can see there is much more fog or mist weather for severe accidents! Then I tried to plot the percentage of fog weathers for each hour:



We can see that at late night, there is indeed a higher chance to be foggy hours. However, this still cannot explain why there is most likely to have severe accidents at 3 am.
Now only the final suspect left: light conditions. I made a plot of the percentage of light conditions at different level of severities (the label is the absolute number of records):



We can see there are more “no street lighting” in level 1 severity.
To check if it is true that there is more “no street lighting” at late night, I made the following plot of the percentage of “no street lighting” in different hours:

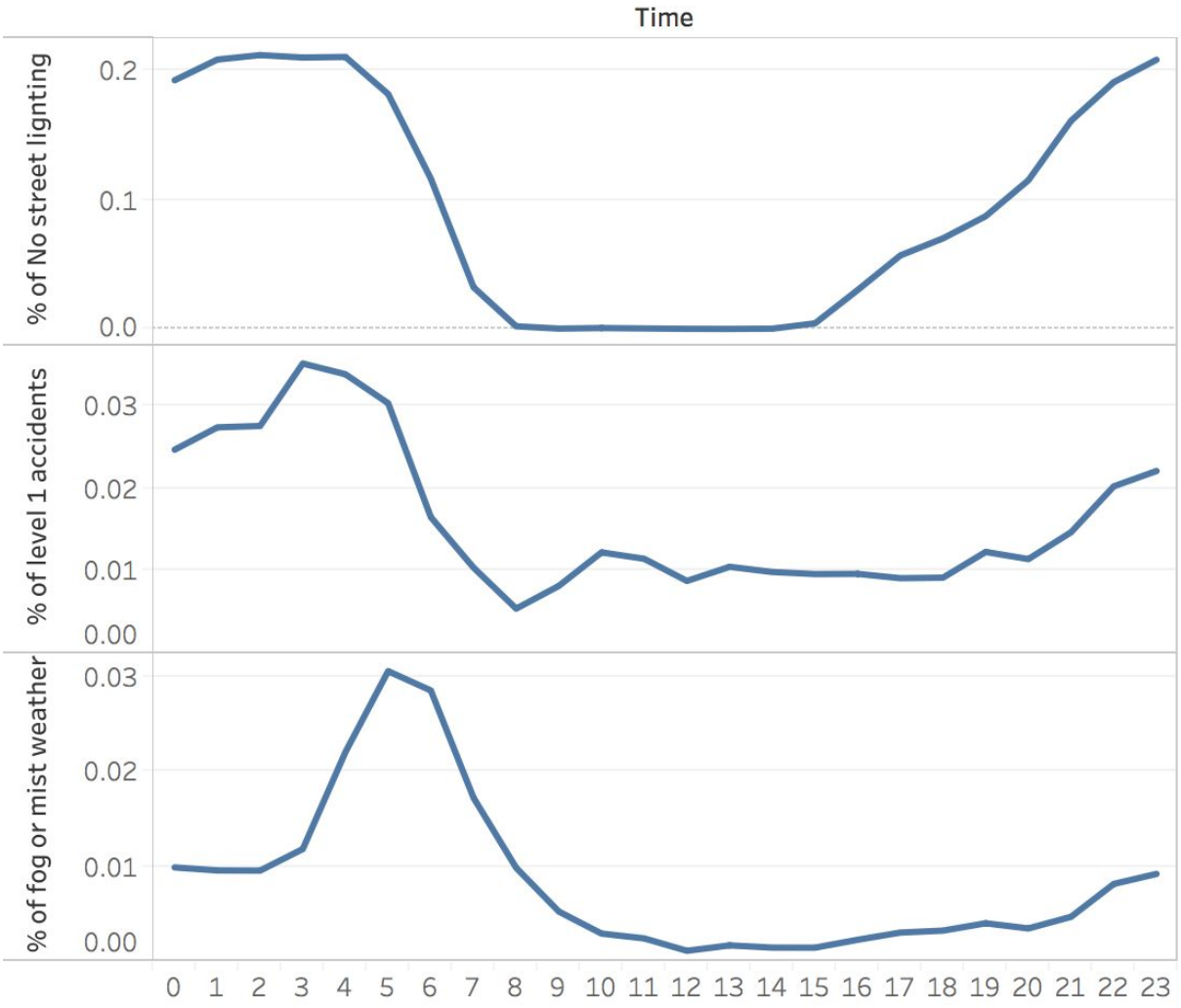


We can see the peak is around 11 pm to 4 am. So it seems true that at around late-night (2 or 3 am), there is a higher chance of having no street lighting, which could be a possible cause of higher severe accident rate.

Conclusion:

Based on the analysis of the data, I found that “fog or mist” weather and “darkness: no street lighting” light condition both have correlations with time and severe accident rate. But correlation is not causation. There might be some other variables such as tiredness of driver, which is affected by time and affect the severe accident rate. However, this part of data is not included in this dataset, thus we cannot test on all possible variables to find the actual causation, or how much each variable contributes to the causation.

Final visualization:



In the above visualization, we see that the percentage of accidents with no street lighting among all accidents in that hour, the percentage of level 1 accidents among all accidents in that hour and the percentage of accidents with fog or mist weather among all accidents in that hour. We can see there is a correlation between the percentage of fog or mist weather with the percentage of severe accidents, and also the correlation between the percentage of no street lighting and the percentage of severe accidents.

I would say light condition and weather could be (but may not be) two possible causes that make higher severe accident rate in late night, but there would be some more as well, such as tiredness of driver and so on.

Limitations:

1. The severity of accidents is only measured in 1-3 scale, which is not accurate enough.
2. Many factors are not counted in such as the reason for the accident (such as driver careless and vehicle defects)

3. This kind of data cannot deduce causation. Even the light condition is improved, the percentage of severe accidents may not improve, as they may not be a cause-effect relationship.
4. Some factors cannot be changed because people do not have control on it, such as weather, urban-rural areas.

[1] WHO, “WHO | Infographics,” *WHO*, 07-Dec-2018. [Online]. Available: http://www.who.int/violence_injury_prevention/publications/infographics/en/. [Accessed: 16-Oct-2019].
[2] D. Fisher-Hickey, “1.6 million UK traffic accidents,” *Kaggle*. [Online]. Available: https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales/download.
[3] UK Department for Transport, “Map Road traffic statistics - Road traffic statistics.” [Online]. Available: https://roadtraffic.dft.gov.uk/. [Accessed: 16-Oct-2019].