

A Study on the San Francisco Incidents Dataset Utilizing Visualization Tool

The data I explored is from the San Francisco dataset. I used Tableau for the visualization work and R for some calculations. Here are some research questions and my responses.

For either city, how do incidents vary by time of day? Which incidents are most common in the evening? During what periods of the day are robberies most common?

For San Francisco, the line chart (Figure 1) below shows the way incidents vary by the time of the day. We can see the number of records for all types of incidents peaks at 17:00-18:00 while reaches the bottom at 4:00. The number of incidents increases steadily from 4:00 to 18:00 and decreases from 18:00 to 4:00, except that at 12:00 the number of records reaches its second peak.

According to Figure 2, we can clearly tell from the visualization that Larceny/Theft is most common in the evening. Though it is a little subtle, Vehicle Theft is also more common in the evening than in other hours.

From Figure 3, we can tell that the number of incidents from robbery has three peaks. The first one is around 12:00, the second is around 23:00 and the last one is near 1:00. However, at hours near 7:00, 10:00 and 15:00, the rates are very low.

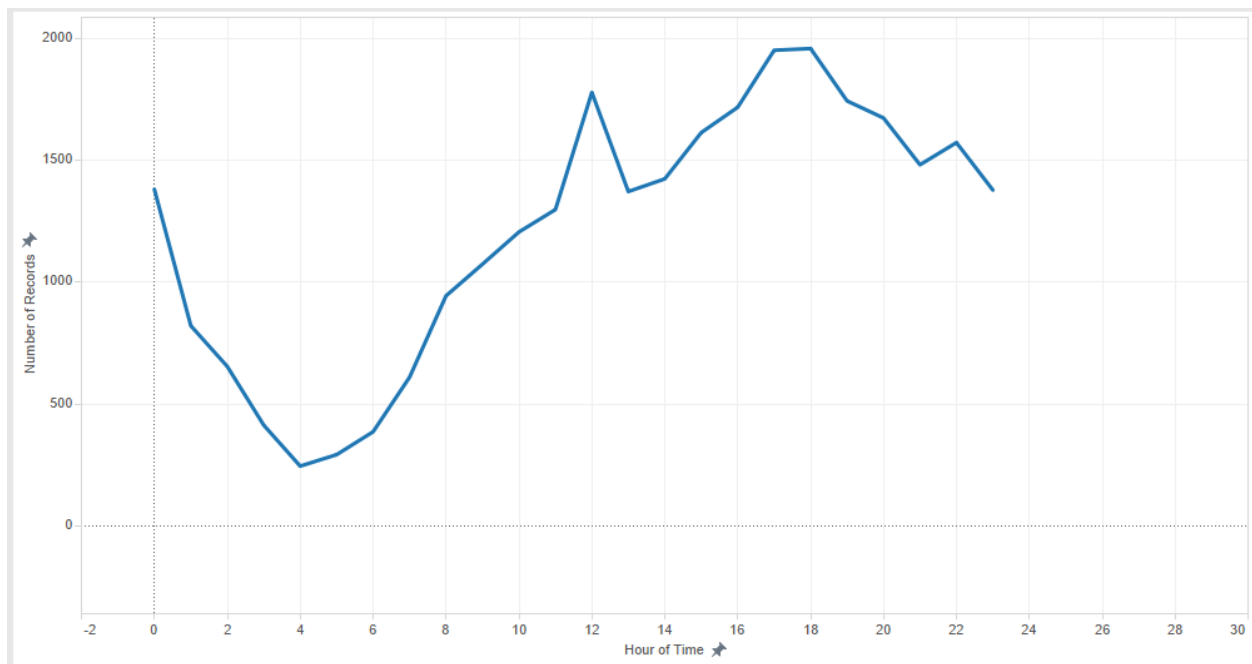


Figure 1

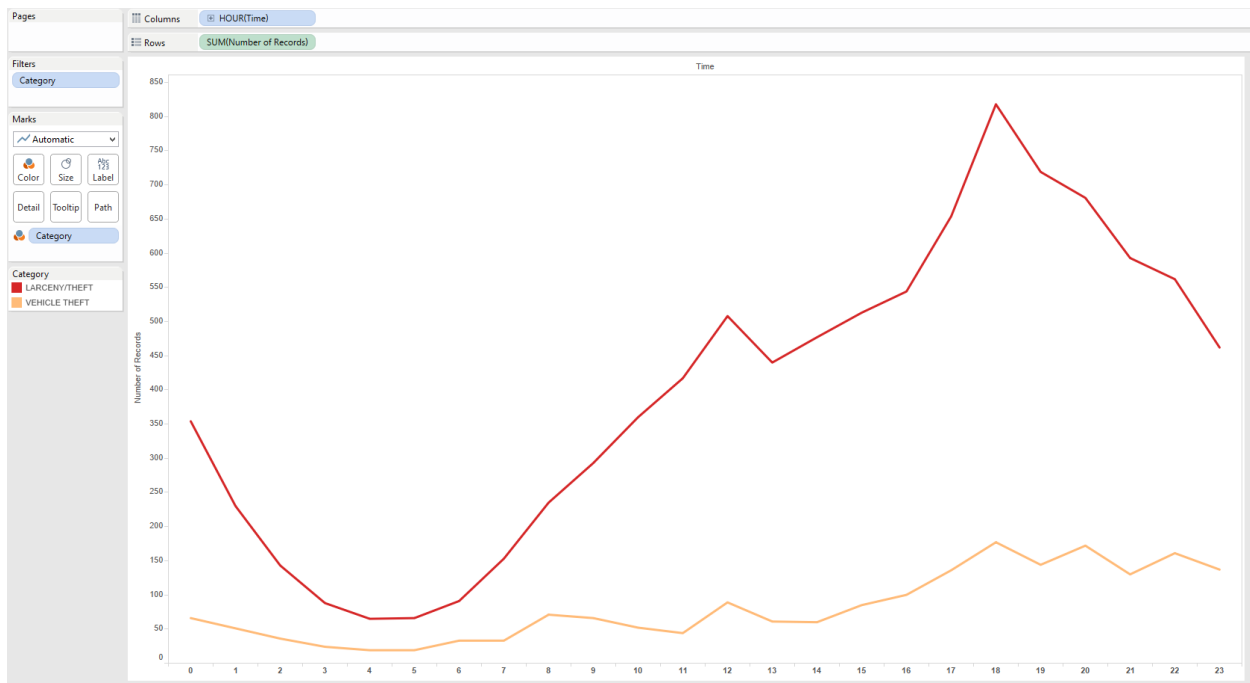


Figure 2

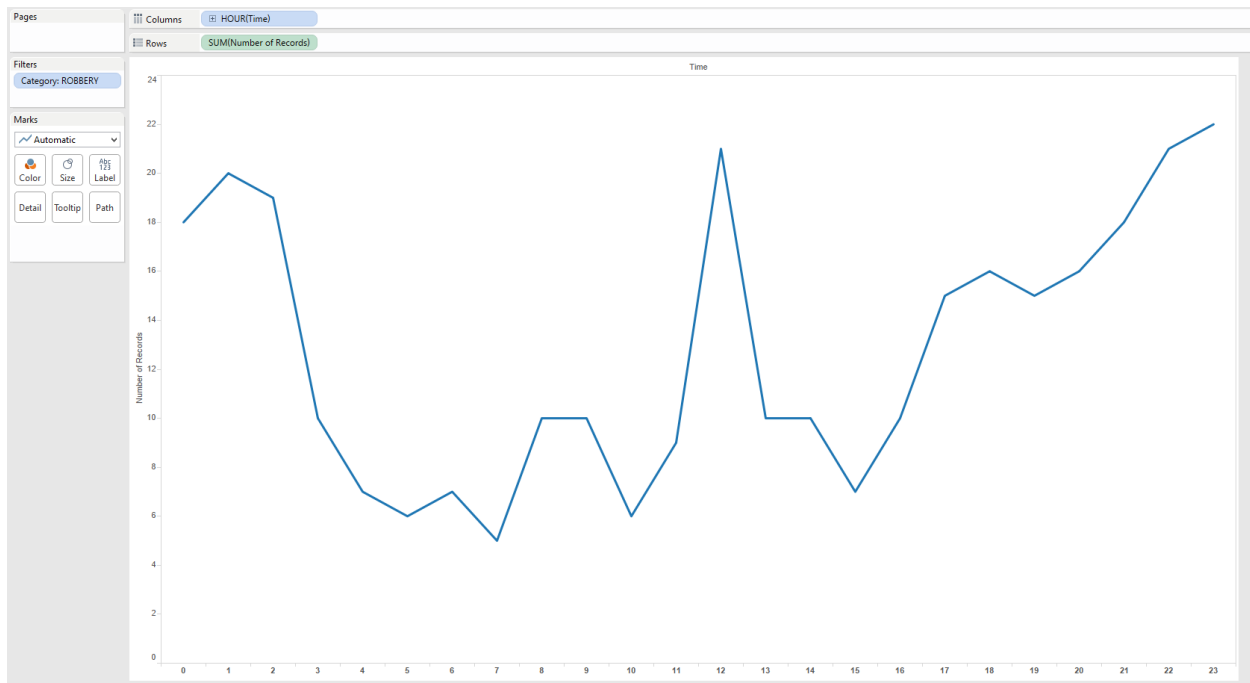


Figure 3

For either city, how do incidents vary by neighborhood? Which incidents are most common in the city center? In what areas or neighborhoods are robberies or thefts most common?

According to Figure 4, we can clearly see that the Southern district has the largest number of incidents while Park and Richmond are the two districts which have the very low rate.

From Figure 5, we can see that in the city center, the most common incident is Larceny/Theft followed by Non-criminal, Other offenses, and Assault. Since the categories of Non-criminal and Other offenses are too ambiguous, I consider Assault should be considered as the second most common incident.

According to Figure 6, robberies or thefts are most common in the neighborhood Southern followed by Central and Northern.

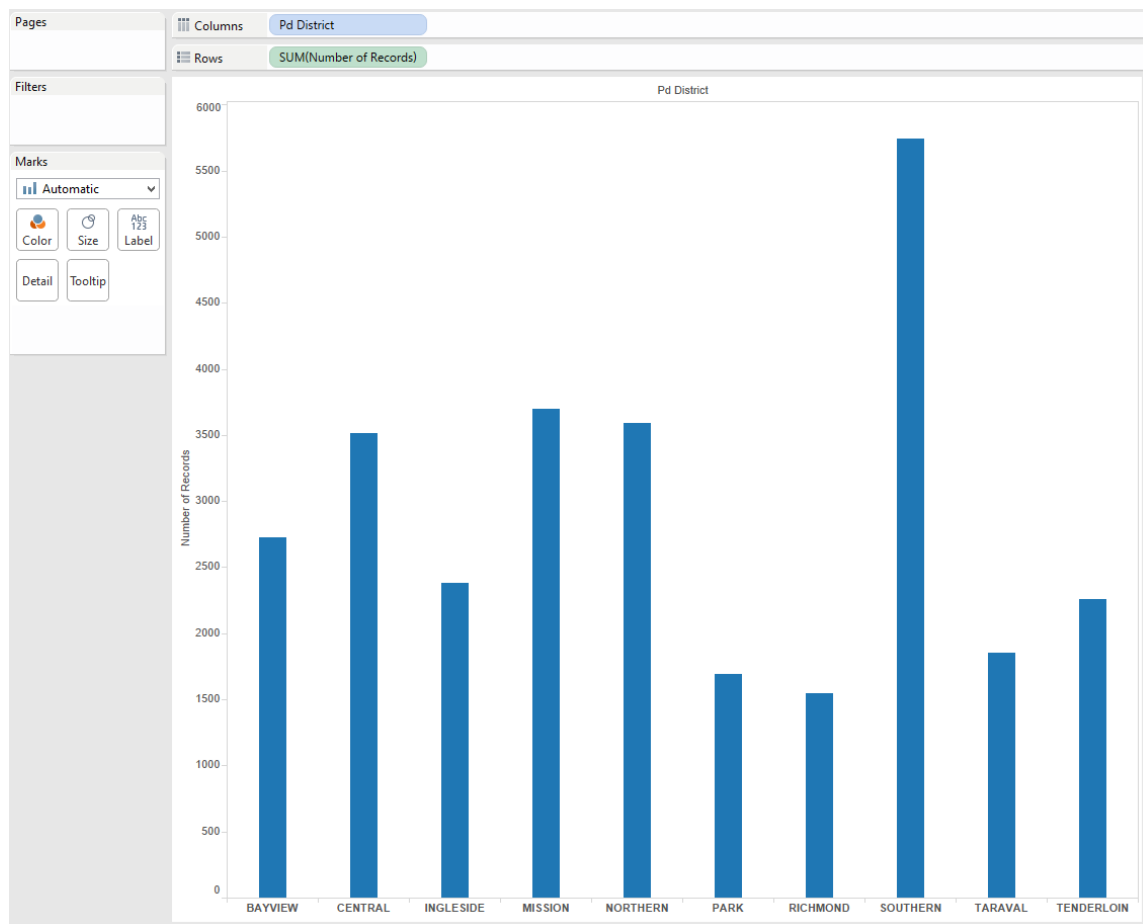


Figure 4

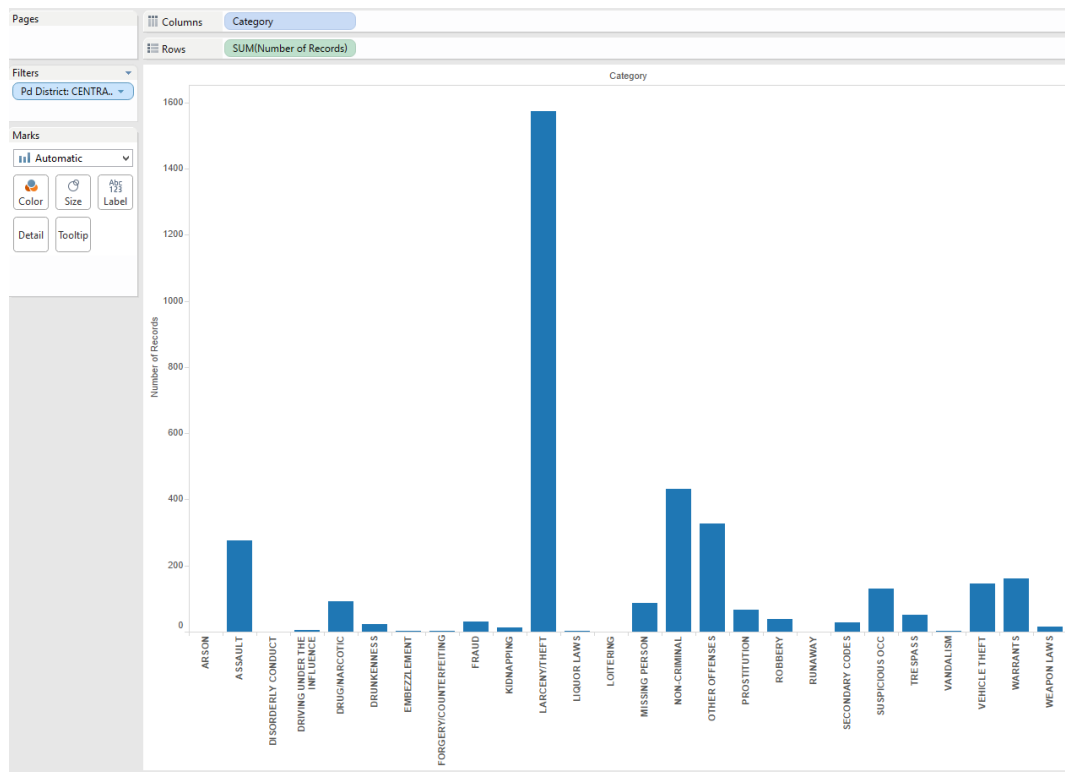


Figure 5

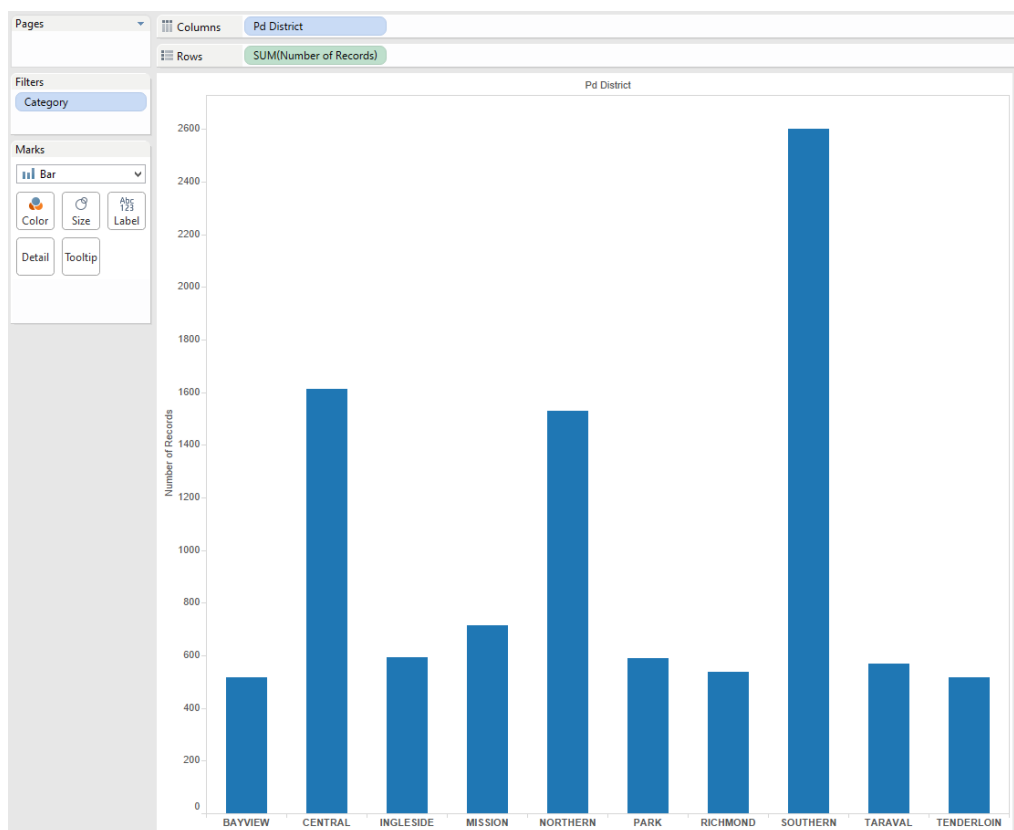


Figure 6

For either city, how do incidents vary month to month in the Summer 2014 dataset?

From Figure 7, we can tell that the overall incidents number steadily increases from around 9K to over 10K between June and August.

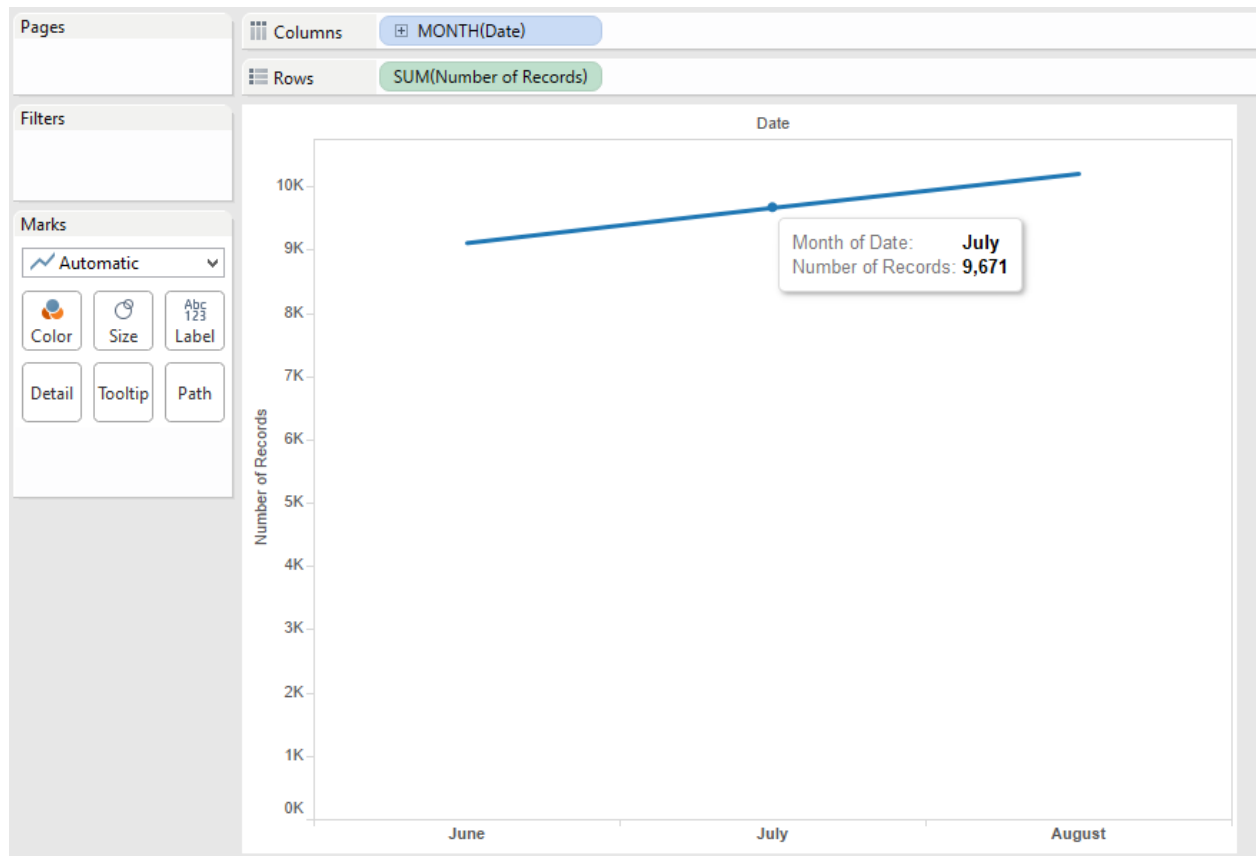


Figure 7

For either city, which incident types tend to correlate with each other on a day-by-day basis?

For San Francisco, I discover that Missing Person and Drunkenness are the two types of incidents which closely correlated (Figure 8 and Figure 9). This fact is actually very intuitive and interesting! When people get drunk, there will be a higher possibility that they will not be able to return home, so their family / friends may call the police for missing person.

The first step I did was to calculate all the pairs of correlation for each type of incident. Then I picked up the ones with absolute values over 0.5 and visualized those incidents. Among these incidents, I finally decided the visualization of the pair (Missing Person and Drunkenness) with the highest correlation (0.6543) can best show the correlated pattern on a day-by-day basis.

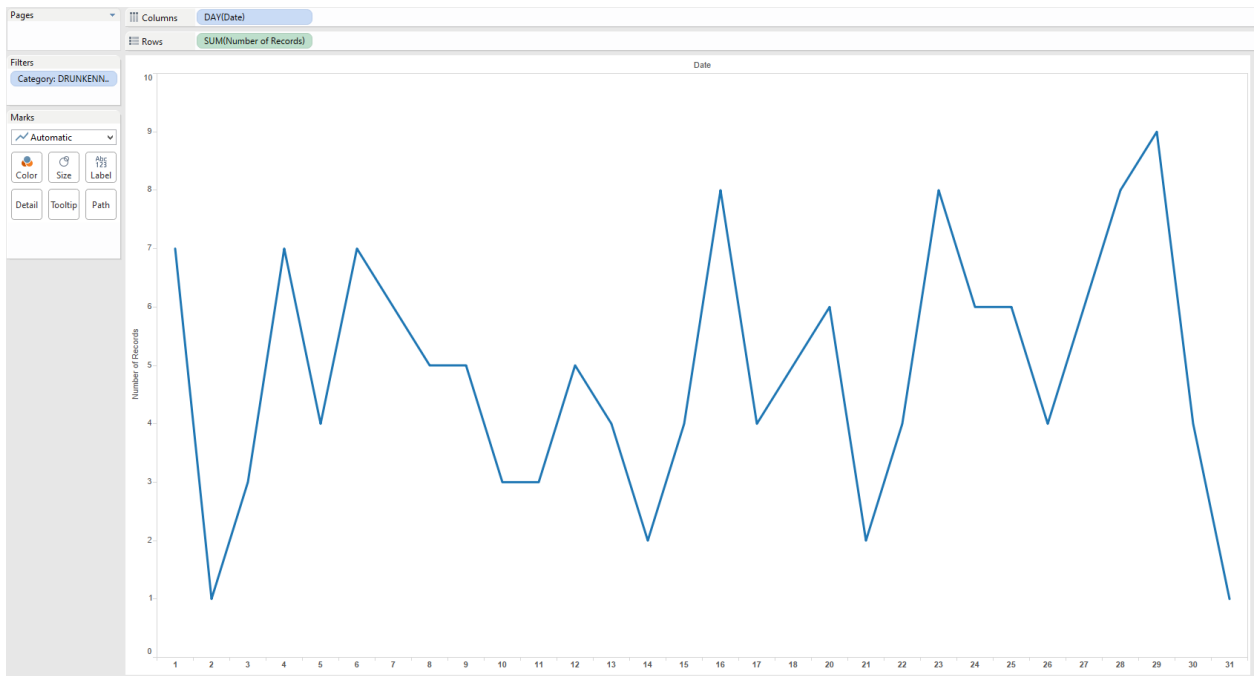


Figure 8 Drunkenness

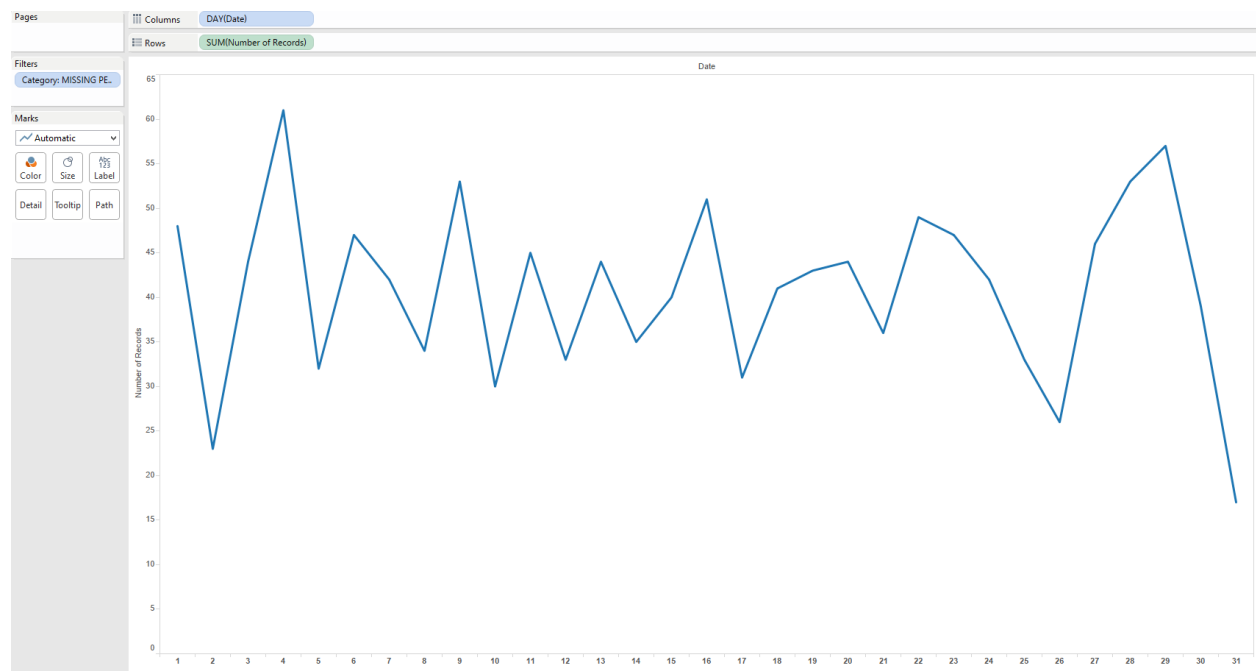


Figure 9 Missing Person