**Short paper**

***Introduction***

Police and government in the United States have made efforts to reduce crimes continuously in the past decades. Compared with 17 other politically stable, economically affluent democracies, America is distinctive. It not only has a much higher homicide rate and deaths from guns, but also has higher imprisonment rates, and is one of two countries that retains capital punishment (Rodney Tiffen, 2019). In this study, we analyzed the crime patterns in San Francisco to provide data-supported facts for both the police system and public for their own safety. Data-driven analysis that is much more effective to combat crime. We want to examine the relationship of how factors like crime occurred time and demographic patterns influence the crime rate. By recognizing crime-frequently time and places, resources will be more focused and effective to solve the crime events. Then we can optimize resources utilization and improve the police reaction system through understanding the type and pattern of crime and provide timely reaction.

San Francisco, as the cultural, commercial, and financial center for Northern California, has 883,305 residents as of 2018. It was the seventh-highest income county in the United States, with a per capita personal income of $130,696. As of the 2010 census, the ethnic makeup and population of San Francisco included: 48% whites, 33% Asiana, 6% African Americans, and 15% Hispanics or Latinos of any race (Gibson, 2010). It is notable that there are several street gangs operating in the city, including African-American and Chinese gangs. The city has a 2.0% unemployment rate and 11.7% of poverty rate, both lower than the national rates by 1 to 2 percent.

The aim of the present study is to address the incongruous findings of the time-crime relationship in San Francisco, along with crime types and resolution results. We want our analysis to offer a helpful guide for people who live in San Francisco to reduce the ‘avoidable’ incidents and crimes. We will address the following questions:

* Find out the most frequent time of crimes occurred
* Display the distribution of crimes categories and resolution rates of each
* Crimes rates in the past 15 years and trends for different crime types

***Related Work***

The study (Lu & Tang, 2011) used geographically weighted regression (GER model) to analyze the cause of crime rate spatial distribution and also indicated social factors which are significantly influenced crime rate. The study indicates that spatial distribution of theft crime rate is positive related to road network, police intensity. In addition, the crime rate is negatively correlated to population density, but this negative correlation is not significant. There is also an insignificant negative relationship between crime rate and land price on average. On the social and economical aspects, the crime rate is influenced by family income, education level and unemployment rate. Compared with our study, we both have the similarity in choosing the big city as an example to start with. Besides, we also have two same measurements, population density and policy density, as potential factors may influence the crime rates. As for the differences between two studies, first of all we used different analysis models from them. Lu and Tang’s study used a local model, which able to consider the differences of spatial location and the spatial correlation, to analyze the local parameters estimations. Secondly, two studies focused on two different countries. Lu and Tang study a big city in China and our study focuses on San Francisco, a big modern city on the west coast of the U.S.. Crime rate is a social study which is highly related to the local culture behaviors and general mode of society. In this case, both of us need to customize the influence factors corresponding to the city we study in. For example, Lu’s study chose average land price as one factor, because land price is one obvious indicator of the neighborhood development in China. In contrast, we focused on analyzing the frequency of crime that occurred in each period, given the fact that SF is a city known for its difference of day and night time.

As we went through all the databases, we found that there was a significant drop for “Vehicle Theft” from 2005 to 2006, reducing from 18000 cases per year to 7000 cases per year and remaining around 6000 cases per year to now. In order to figure out the reason, we compared our research with a previous study, which focused on car theft in Mexico City from spatial patterns and time series perspectives (Carlos J., 2011). Based on Carlos’s research, the annual historical data showed that the total frequency of crimes had a high correlation with the crime of vehicle theft (r = 0.852). Additionally, it was also detectable that vehicle theft is a very self-correlated crime, and which is predictable over time. The study also showed there was a distinctive increasing trend of car theft in the 4 years after the economic crisis and notable reduction presented in 5th year. Although Mexico City is in a different country than San Francisco, both studies used the same data type and recorded by the police sector.

Study of crime rates and medical marijuana dispensaries in the city Los Angeles is another study we found related to our research. Christopher Contreras found that liberalization of marijuana may have implications for neighborhood crime, specifically in the distribution of marijuana through a dispensary system potentially providing additional opportunities for criminal behavior to take place (2016). It also suggested that marijuana dispensaries may increase crime rates on socially organized blocks, which experiencing a slight perturbation in their ecological continuity from a dispensary’s establishment. Same as LA, San Francisco experienced the same process of legalizing use of marijuana. According to Christopher research, he found that average household income for blocks without dispensaries is nearly double than those blocks with dispensaries, but only have half the amount of liquor stores inversely. Relating to our study, we also found the increasing trend of crime rates since 2011, similar to the City of Los Angeles. The difference between two studies reflects on the emphasis. Our study focused on the crime types and the prevention of crimes in the future.

Beside above, some researchers did the crime study about the relationship between age structure and homicide rates. The results found that age structure, in which from 15 to 29 considered as the range of high possibility to make criminal choice, has positive effects on homicide rate among the group of people who are not engaged in any social institution, like school, work and military; to those people who are involved in the social institution, the age structure has negative effects on homicide rates (Mccall et al, 2013).

***Process***

We download the data set ‘*Police Department Incident Reports: Historical 2003 to May 2018*’, from [DataSF](https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry) and restore it as a CSV file. At the beginning, we used the pandas library to read the downloaded CSV file and named ‘f’. Since this is the completed data set of San Francisco from 2003 to 2018, lots of data are not very necessary in our analysis. Thus, we use the drop function to remove all the columns we will not use. To process the data more effectively, we replace the column’s name from ‘Category’ to ‘Incident’. In addition, we create a new column, named ‘Year’ in dataframe. The values of column ‘Year’ are cut out from the values of columns ‘Date’. By these steps, we are prepared to answer the first question.

We group the data by the values of column ‘Year’ and count the total number of appearances of the values of column ‘Incident’. Since the incident data in 2018 is only collected until May, the incident sum result in 2018 is not accurate and much lower than the sum in other years. To make sure the analysis is more accurate and reliable, we decide to drop the data in 2018. Then, we get the line graph about the total number of all incidents in SF from 2003 to 2017 by applying *plot* function. To analyze deeply, we wonder which type of incidents are high-frequently happened in San Francisco. So, we use *groupby* function to get the total number of each incidents over the past 15 years. Some types of incidents only occasionally happened, while other types of incidents have an extremely large number of records. Therefore, we want to be more focused on the types of incidents with high frequency. Two special things we do in the data processing are dropping two categories: ‘other offenses’ and ‘non-criminal’ since the data set dictionary doesn’t explain the specific types category ‘other offenses’ includes and the category ‘non-criminal’ is considered irrelevant to analyzing the San Francisco’ safety level. Based on the result, some types of incidents have far larger records compared with other types. Taking those ‘top’ types of incidents away, we also plot those incidents type who have a large number of records but not that extremely large.

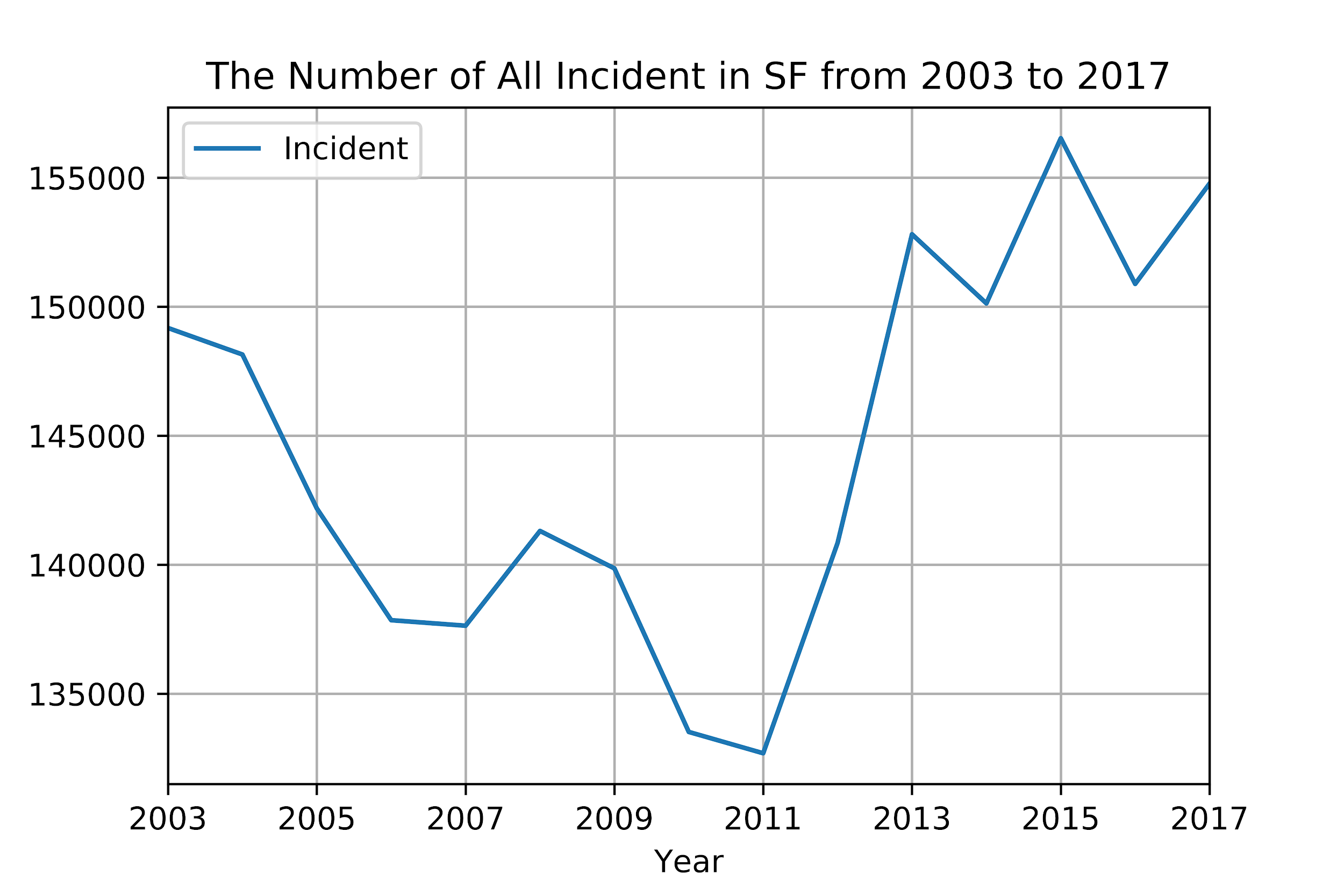
To figure out the most dangerous time in San Francisco, we decide to count the total number of incidents in each two hours. Basically, we add the value as 1 in the columns named Totoal Number of Incident and use the *resample* function to change the period into two hours and get the sum in each two hours period.

The incident patterns at ten police districts of San Francisco are different. We try to figure out the efficiency of each police system. The steps of choosing relevant data and dropping irrelevant data are similar to what we do in answering the former two questions. To observe the gap between the total number of different incident types, we plot all the incident types initially and remove the incident type which has the largest total number. Then we build a DataFrame that only includes the information about police districts and resolution results. If any single incident was labeled ‘solved’, then we mark it as 1; otherwise, we mark it as 0. By *groupby* function, we can get the sum of all incident records and the solved incidents in each police district.

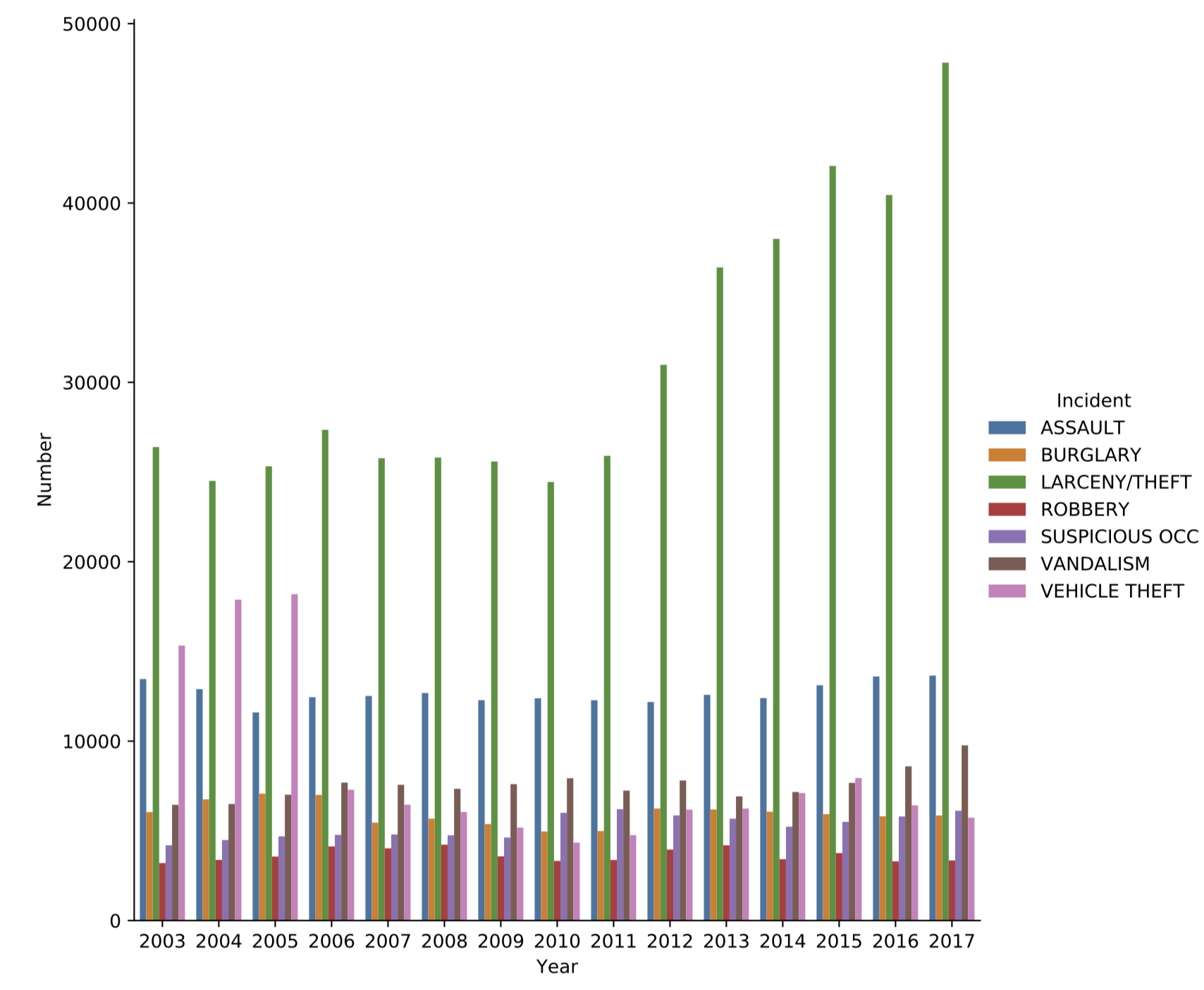
***Results***

**Question 1: Crimes rates in the past 15 years and trends for different crime types.**

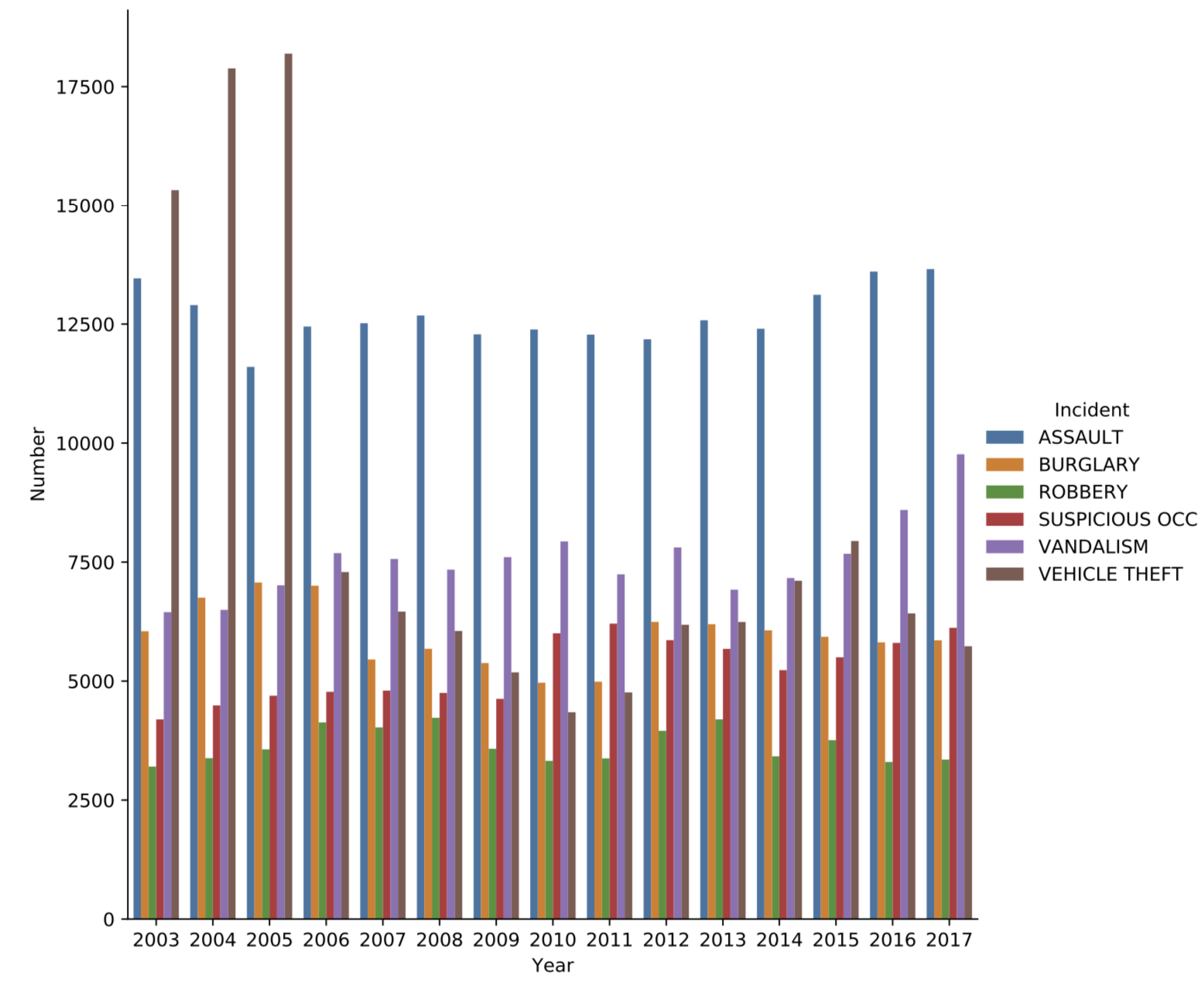
As shown in Figure1, the number of all incidents in San Francisco has decreased from 2003 to 2011, and then gradually climbed up after on. The highest number in 2015 has more than 20,000 cases occurred compared with the lowest point in 2011. Get into details of each different crime type in Figure 2, “Larceny / Theft” has always been the highest since 2003. Since “Larceny / Theft” is significantly higher than other crime types, we eliminated this in Figure 3 and kept all others. It is notable that all the crimes have a stable number of incidents except for “Vehicle Theft”, which had a huge drop from 2005 to 2006.

****

**Figure 1**

****

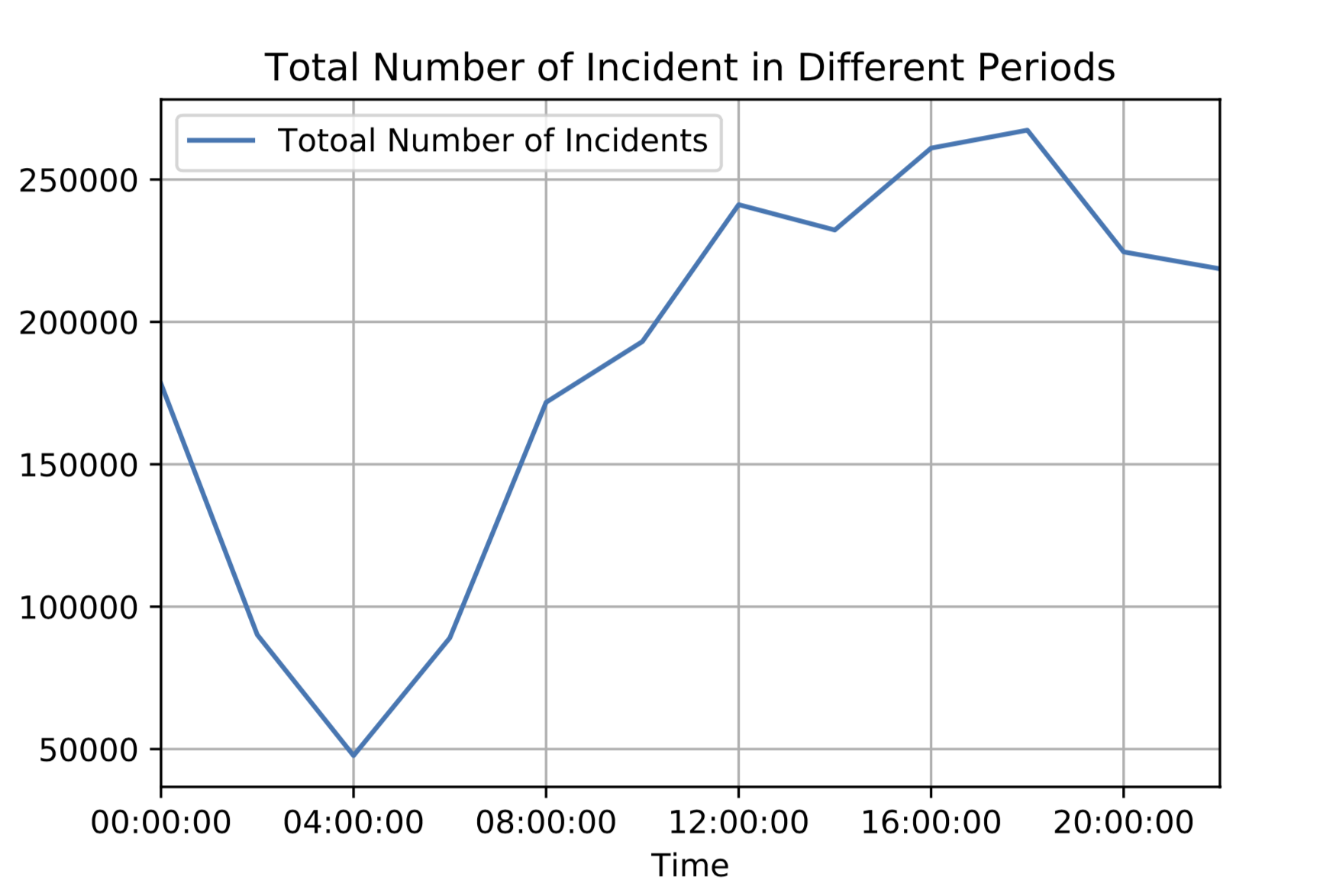
**Figure 2**



**Figure 3 (Without “LARCENY / THEFT”)**

**Question 2: Total number of incidents in every two hours**

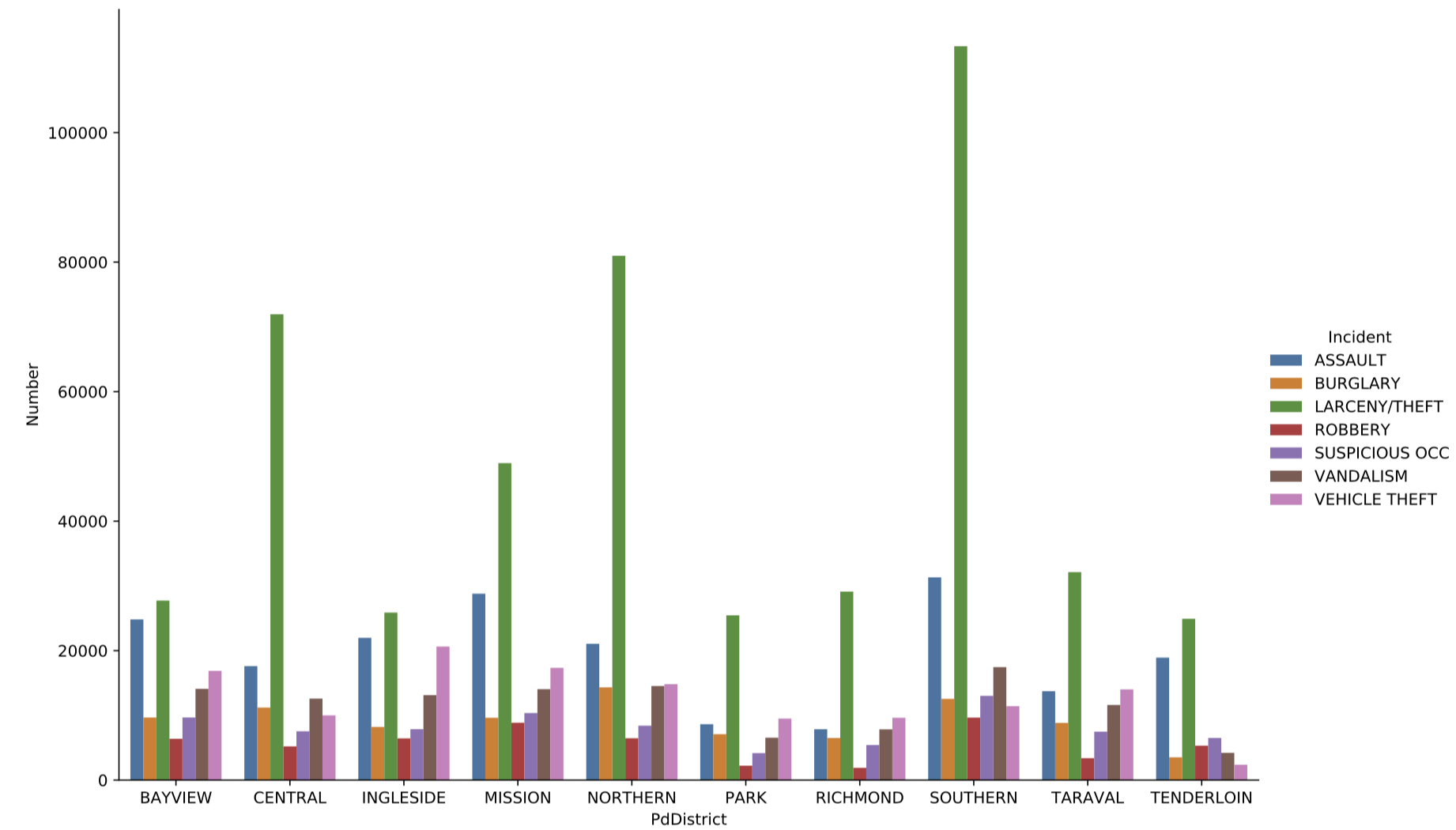
Before we do the analysis, the conjecture we made is that people are more likely to be attacked at night. Therefore, the total incident number at night is supposed to be higher in the daytime. However, the data result shows the opposite things. The result indicates that 18:00 have the highest total number of incidents (Figure 4).



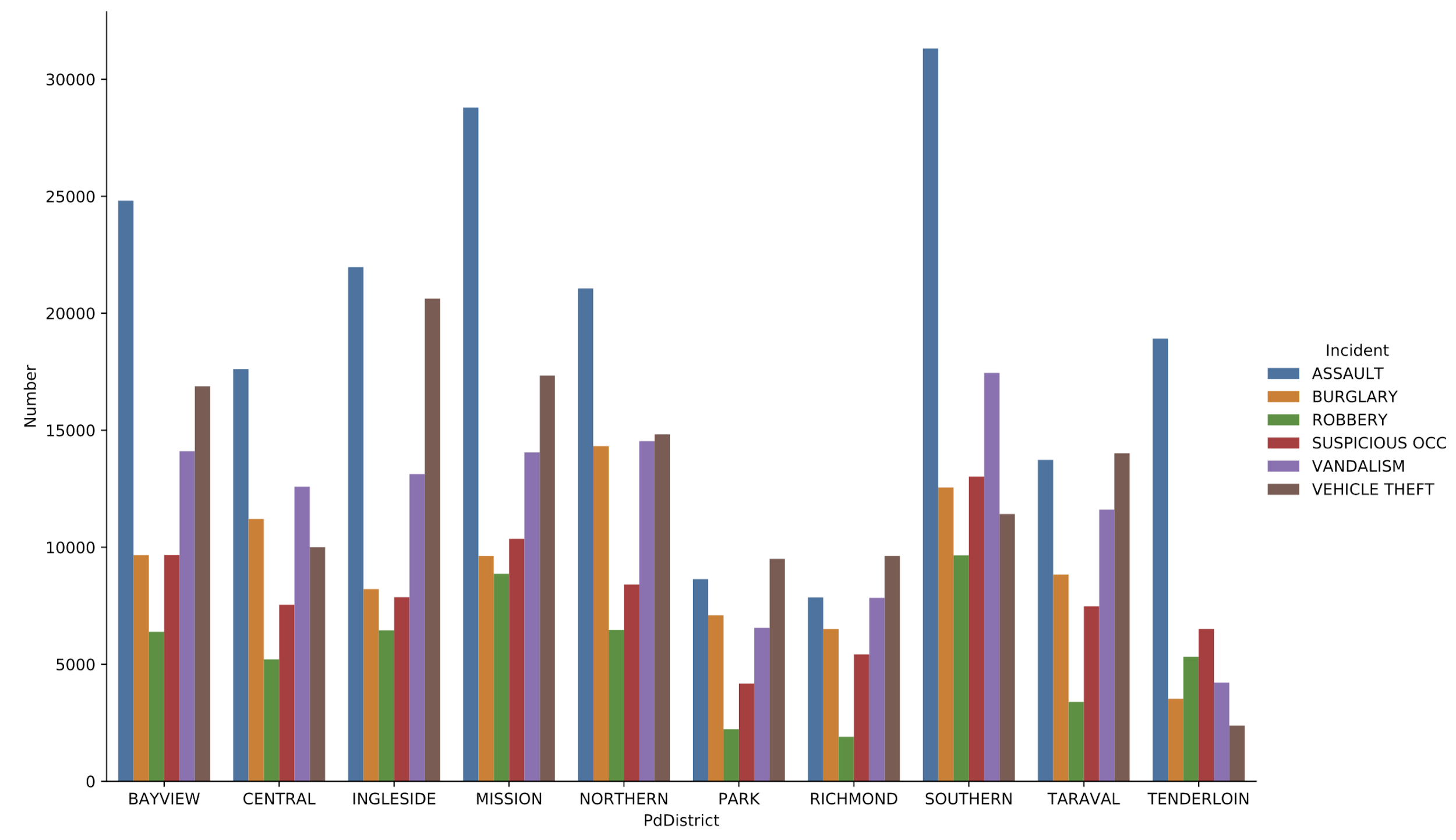
**Figure 4**

**Question 3: Display the distribution of crimes categories and resolution rates of each**

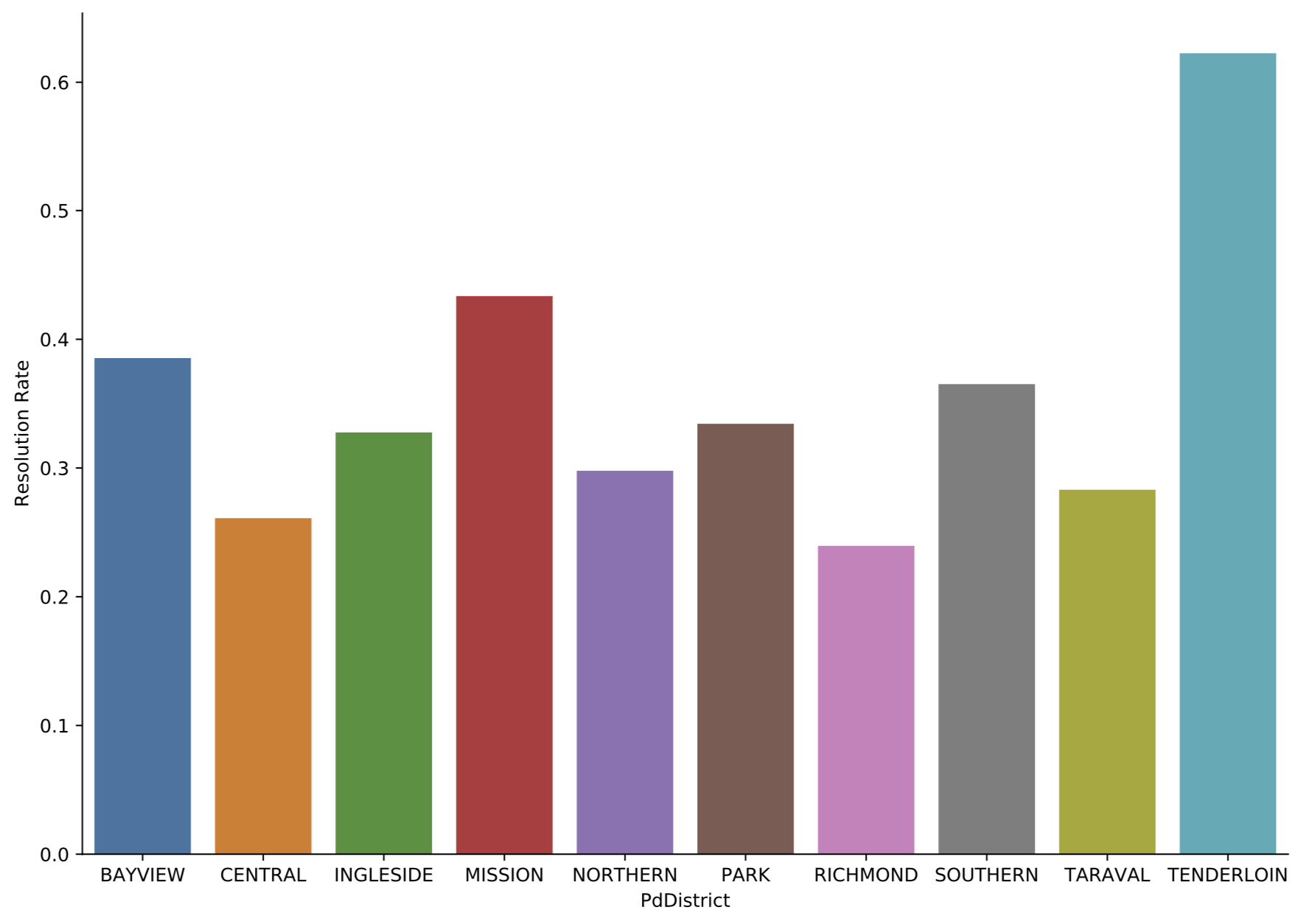
In figure 5 and 6, we displayed the total number of incidents received by PdDistricts in San Francisco from 2003 to 2018. Same trend as shown in question 1, “Larceny / Theft” has been the top 1 for all PdDistricts, following with “Assault” as the second. The distribution of the number of crime types in each PdDistrict is similar to one another. We finally calculated the resolution rate for each PdDistrict, showing that Tenderloin has 60 percent, the highest resolution rate among 10 PdDistricts in SF.



**Figure 5**



**Figure 6 (Without “LARCENY / THEFT”)**

****

**Figure 7**

***Conclusion***

By analyzing the incident data in San Francisco from 2003 to 2018, the incidents in San Francisco have had an increasing pattern since 2011. “Larceny / Theft” is the most serious problem no matter in which year. The number of those types of incidents mostly are stable or increasing. What unexpected is that the most dangerous time in San Francisco is in 18:00 to 20:00, rather than at midnight. Among ten police district, Tenderloin, Bayview and Mission are considered to have better ability to solve the incidents. To improve the analysis result of Question 2, more factors need to be considered. Since fewer people will go outside at night, it will be less likely to incidents happening. Moreover, this report mainly focuses on analyzing the incidents in San Francisco from time and district. There are some possible good options to evaluate the relationship between incident and other factors like, race, education level and policy.

References:

[1] Juan Lu, & Guo-an Tang. (2011). The spatial distribution cause analysis of theft crime rate based on GWR model. *2011 International Conference on Multimedia Technology,* 3761-3764.

[2] Baciu, O., & Parpucea, I. (2011). Socio-economic factors impact on crime rate. *Virgil Madgearu Review of Economic Studies and Research,* *4*(2), 5.

[3] Mccall, P., Land, K., Dollar, C., & Parker, K. (2013). The Age Structure-Crime Rate Relationship: Solving a Long-Standing Puzzle. *Journal of Quantitative Criminology,* *29*(2), 167-190.

[4] Vilalta Perdomo, CJ (2011). Car theft in Mexico City: spatial patterns and time series. *Management and Public Policy* , *20* (1), 97+. Retrieved from https://link-gale-com.ezp3.lib.umn.edu/apps/doc/A304726188/IFME?u=mnaumntwin&sid=IFME&xid=6805b5c5

[5] Contreras, C. (2017). A Block-Level Analysis of Medical Marijuana Dispensaries and Crime in the City of Los Angeles. Justice Quarterly, 34(6), 1069-1095.

[6] Tiffen, R., Gauja, A., O’Connor, B., Gittins, R., & Smith, D. (2020). Crime and Social Problems. In How America Compares (pp. 181-189). Springer, Singapore.

[7] Gibson, C., & Jung, K. (2010). Historical census statistics on population totals by race, 1790 to 1990, and by Hispanic origin, 1990 to 2010, for the United States, regions, divisions, and states. Washington, DC: US Census Bureau.