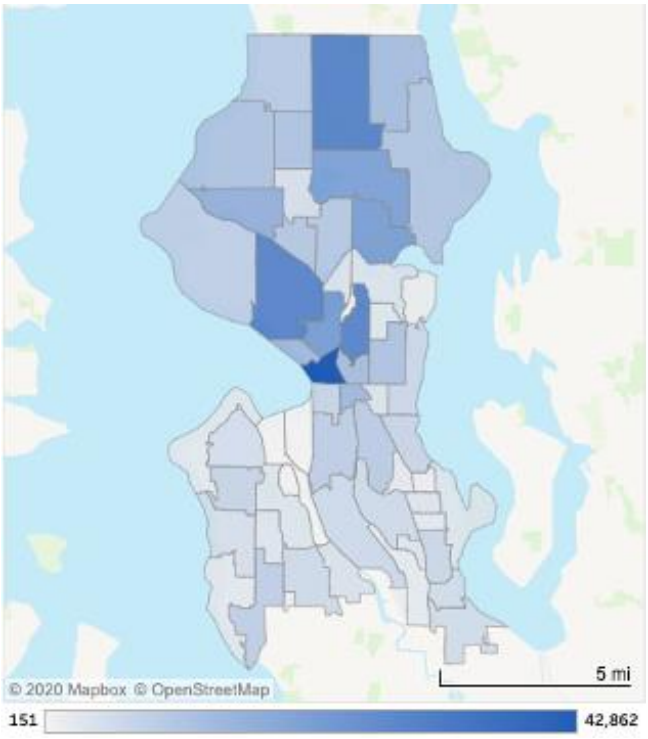


Seattle Crime Analysis

- Seattle Crime from 2008 to 2020

ABSTRACT

This paper studied on the dataset of crime in Seattle from 2008 to 2020 to seek crime patterns in Seattle and give recommendations for residents on how to avoid crimes and stay safe.

- Iggy Zhao

Section 1 Introduction

Seattle is the largest city and business center in Washington State. Its economy, population, and culture are rich in diversity. Because of this, the crime rate is also different in all areas of Seattle. This article analyzes the crime data of Seattle since 2008, explores the types of crime, time, and region, and attempts to mine potential information to provide safety recommendations for residents. Specifically, this article explores the crimes of the three major categories and 30 parent categories, ranks the crimes in different regions, and compares the same time in different years to get the patterns. In addition, this article focuses on the relationship between crime in the downtown area and the categories and start time of crimes. This project is accomplished by Spark powered by Databricks.

Section 2 Data Overview and Preprocessing

Section 2.1 The dataset

The dataset contains the crime record for the city of Seattle from January 1, 2008 to the end of April 2020 with 824220 rows and 17 columns. The dataset describes the start and end time of the offenses, the category (including three levels of classification including the overall category, the parent category, etc.), the offense name, the police jurisdiction, the district, address, and the latitude and longitude information of the place where the crime occurred.

Section 2.2 Data Dictionary

The data dictionary of the original dataset is shown below.

Column Name	Description	Type
Report Number	Primary key/UID for the overall report. One report can con...	Plain Text
Offense ID	Distinct identifier to denote when there are multiple offens...	Plain Text
Offense Start DateTime	Start date and time the offense(s) occurred.	Plain Text
Offense End DateTime	End date and time the offense(s) occurred, when applicable.	Plain Text
Report DateTime	Date and time the offense(s) was reported. (Can differ fro...	Plain Text
Group A B	Corresponding offense group.	Plain Text
Crime Against Category	Corresponding offense crime against category.	Plain Text
Offense Parent Group	Offense_Parent_Group	Plain Text
Offense	Corresponding offense.	Plain Text
Offense Code	Corresponding offense code.	Plain Text
Precinct	Designated police precinct boundary where offense(s) occ...	Plain Text
Sector	Designated police sector boundary where offense(s) occurr...	Plain Text
Beat	Designated police sector boundary where offense(s) occurr...	Plain Text
MCPP	Designated Micro-Community Policing Plans (MCPP) bound...	Plain Text
100 Block Address	Offense(s) address location blurred to the one hundred blo...	Plain Text
Longitude	Offense(s) spatial coordinate blurred to the one hundred b...	Number
Latitude	Offense(s) spatial coordinate blurred to the one hundred b...	Number

Section 2.3 Data Sampling

For simplicity and without loss of generality, I random sampled 5% from the original dataset when finding the data patterns. The sample data consists of 41211 rows. Knowing features such as Precinct, Sector, Beat and _c0 are less necessary for the data exploration, at this stage, they were dropped. Then the columns were renamed into more commonly understandable terms.

Section 2.4 Missing Data

There are a few features including missing values. Among 41211 attributes, 19866 of the Offense End Datetime was missing, and I identify the Offense Start Datetime as the actual crime time when analyzing, so the end time column was dropped. Moreover, 1519 address records were missed, but generally, the place of occurrence of crimes can be displayed by the District. At the same time, the proportion of missing value for this feature was 3.6%, which would be fine if I ignore the value. In the original dataset, the only integer features are the _c0, which is the number of rows and dropped before, and Offense ID, which is not useful in modeling as well. All other features are in string or double type, so at this stage, it's less likely to have outlier problems. The missing values and outliers will be further processed whenever I update the dataset or conduct modeling with specific objectives.

Section3 Data Exploration

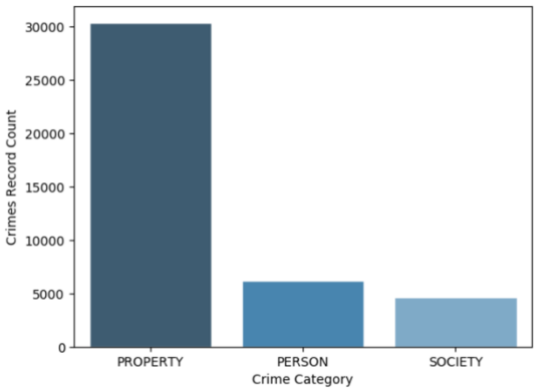
Section 3.1 Crimes Category

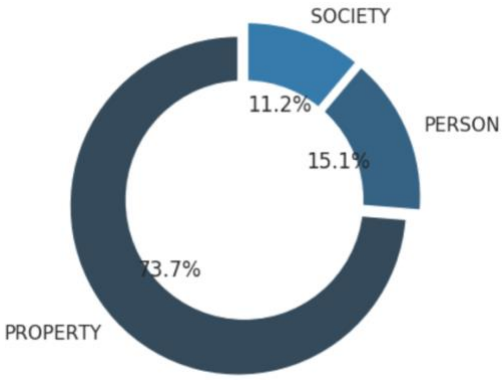
3.1.1 Main Category

In this section, I want to capture the number of crimes in each category to see which kinds of offenses are more likely to occur in Seattle. The crimes in Seattle has three main categories: Property, Person, and Society. Among them, 73.7% were property crimes, indicating that property crime accounted for the majority of all crimes, and the cases of person and society crimes were much less than those of property crimes.

Category	count
PROPERTY	30384
PERSON	6228
SOCIETY	4599

To visualize the categories:



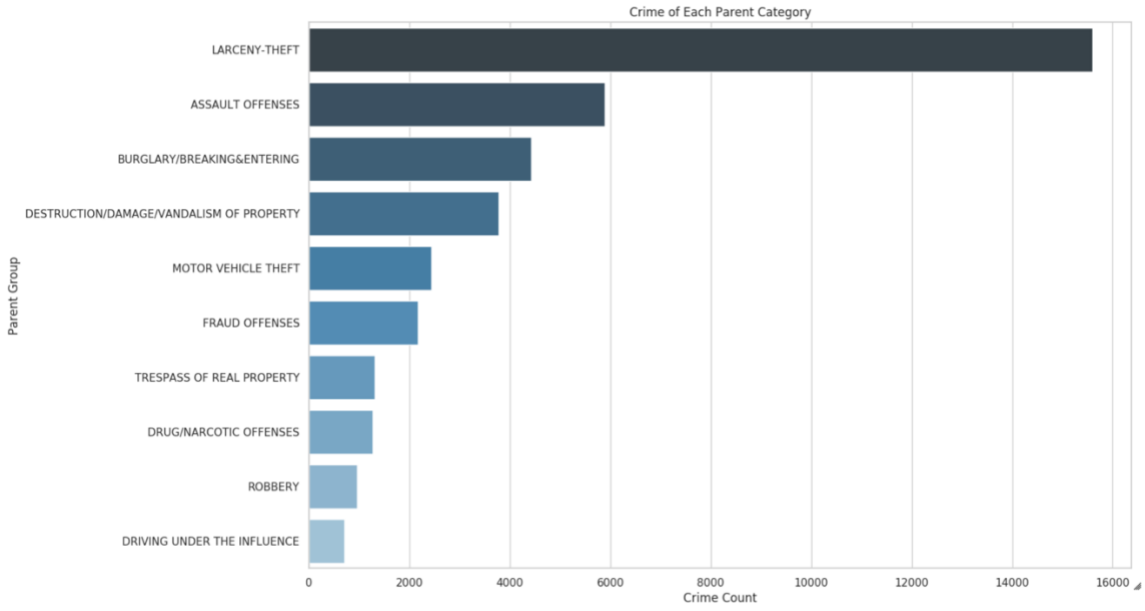


3.1.2 Parent Group

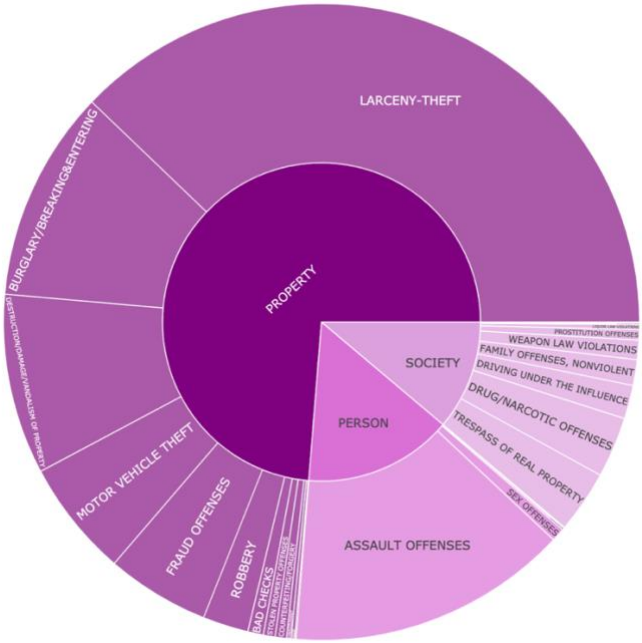
The crimes can be further categorized into 30 parent groups, here lists the main parent groups:

ParentGroup	count
LARCENY-THEFT	15600
ASSAULT OFFENSES	5888
BURGLARY/BREAKING&ENTERING	4429
DESTRUCTION/DAMAGE/VANDALISM OF PROPERTY	3780
MOTOR VEHICLE THEFT	2445
FRAUD OFFENSES	2164
TRESPASS OF REAL PROPERTY	1307
DRUG/NARCOTIC OFFENSES	1271

Looking at 10 parent groups with the highest number of crimes, among 41211 crimes, nearly 16,000 was larceny-theft.

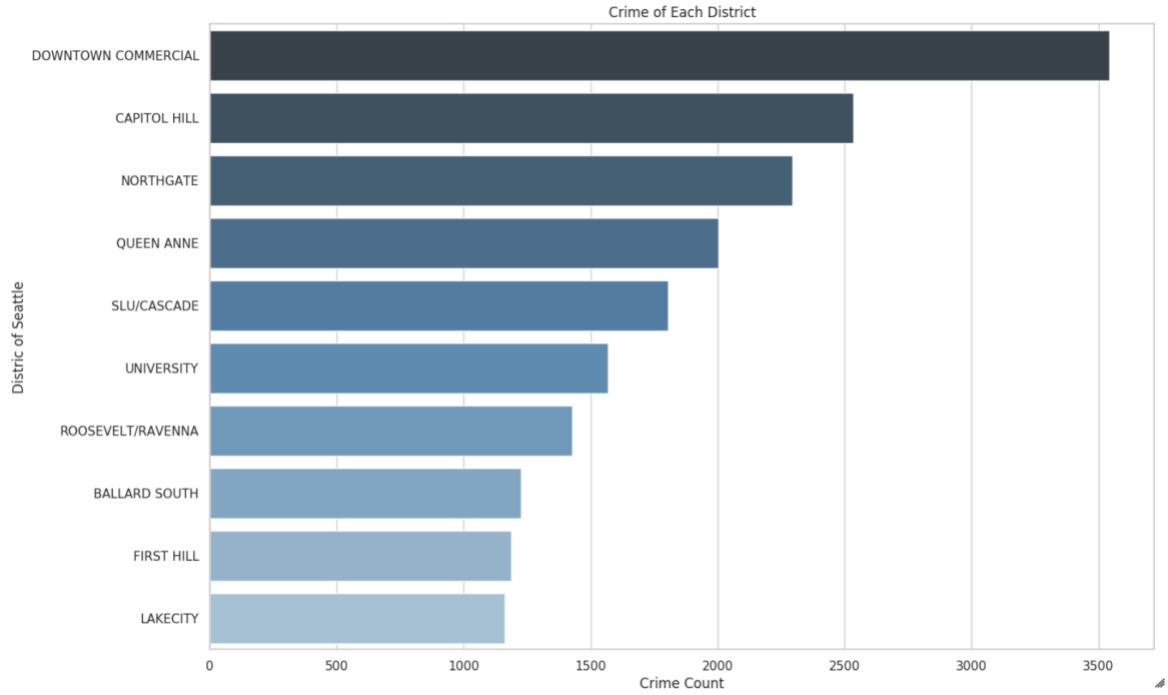


Combining two levels of categories together, the relationship is clearer. Through this sunburst, the proportions of each main category and corresponding parent groups can be easily found.



Section 3.2 Crimes vs. Districts

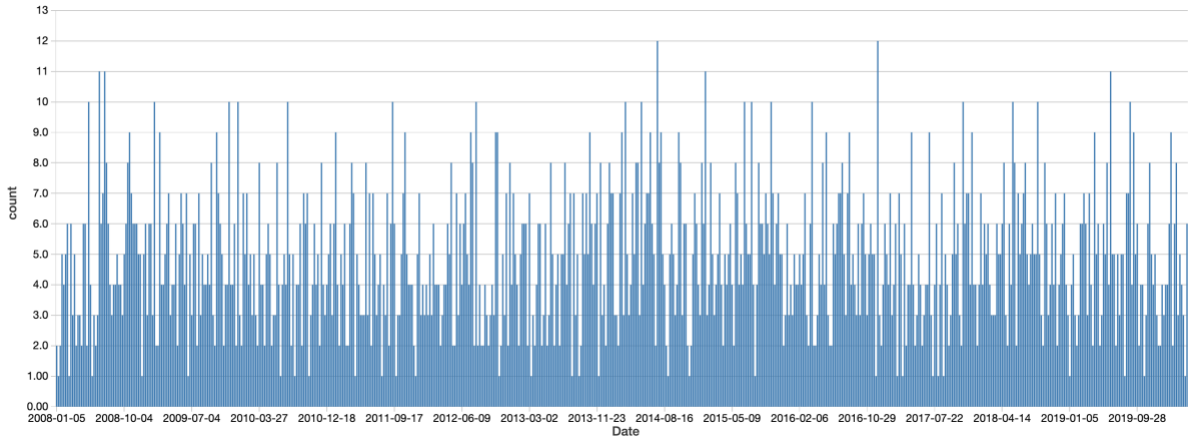
Then I want to detect the number of crimes in each district: among 59 districts, we show the top 10 districts with their crime information. In the chart below, Downtown Commercial is the area with the highest cases of offenses. This phenomenon is in line with our intuition. Because the population density in the city center is relatively large and the economy is relatively concentrated, the crime cases in the downtown area would be greater than in other regions.



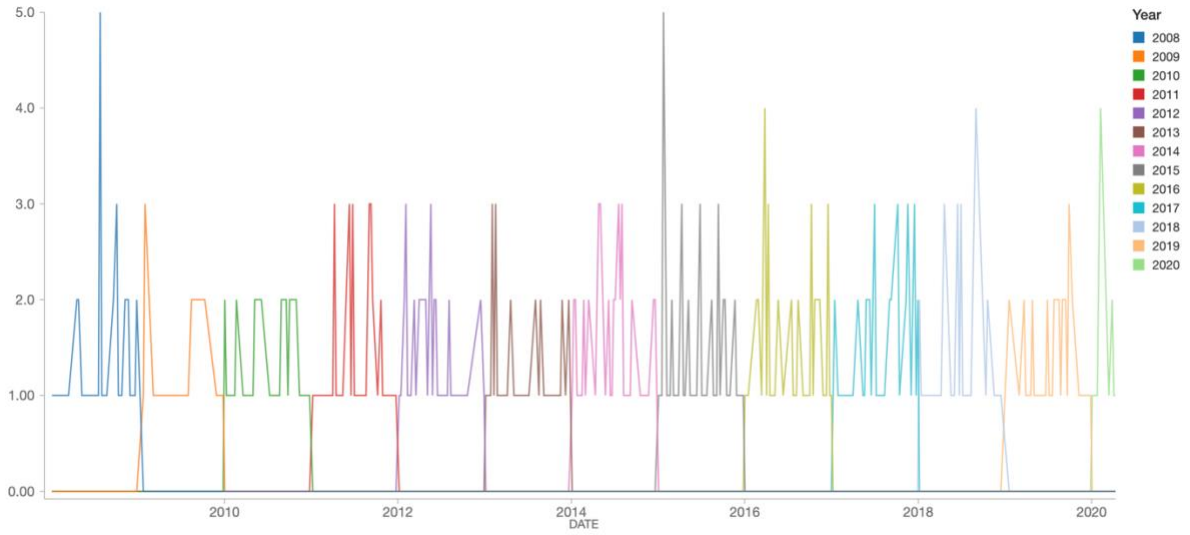
Section 3.3 Crimes in Downtown

Furthermore, knowing that downtown area has the largest cases of crimes, I want to dig more by seeking the patterns of crimes on each Sunday in the downtown area of Seattle. To do this, I split the StratTime of crimes into Date and Hour, respectively, and added a new column of DayofWeek. The maximal count is 12, the minimal count is 1, and the average count is 4.79. First, based on the exact location of the city center of Seattle (47.608013, -122.335167), I define the range of downtown within longitude (-122.285167, -122.385167) and latitude (47.558013,47.658013). On July 19, 2014 and Dec 10, 2016, the cases reached 12. The average cases on each Sunday is 4.79, but because I sampled the data into 5% subset, I assume the actual average number is around 95.

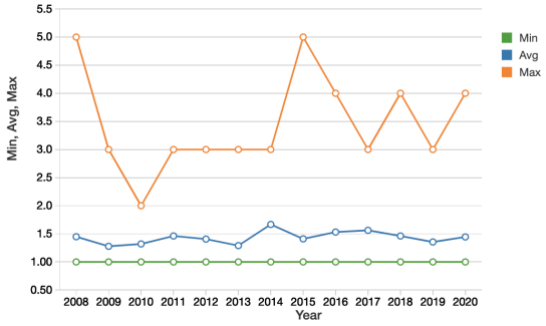
avg(count)	max(count)	min(count)
4.794348508634223	12	1

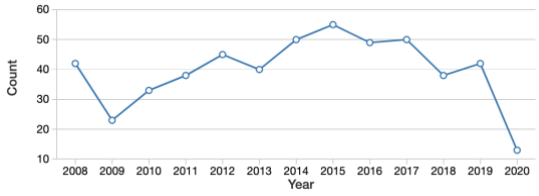


From another perspective, I regarded the 'Downtown Commercial' District as the downtown area of Seattle. For each Sunday of 2008 to 2020, I drew the graph below. The highest value goes to July 26, 2008 and Jan 24, 2015 with the number of cases of 5. From the graph below, we can compare the Sunday crimes for each year. Intuitively, years such as 2008, 2015, 2016, and 2018 have more cases on Sundays.

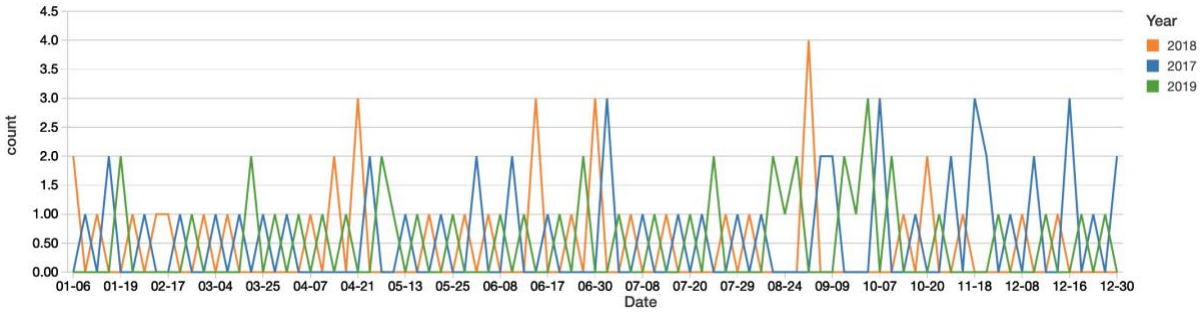


Then I looked at the average, max, and min number of Sunday crimes for each year. As we observed, the average crime number was in 2008 and 2015, which were arong 5, and the second largest numebr was around 4 in 2016, 2018 and 2020 so far. The total count graph shows the similat result: crimes in 2015 was the most.



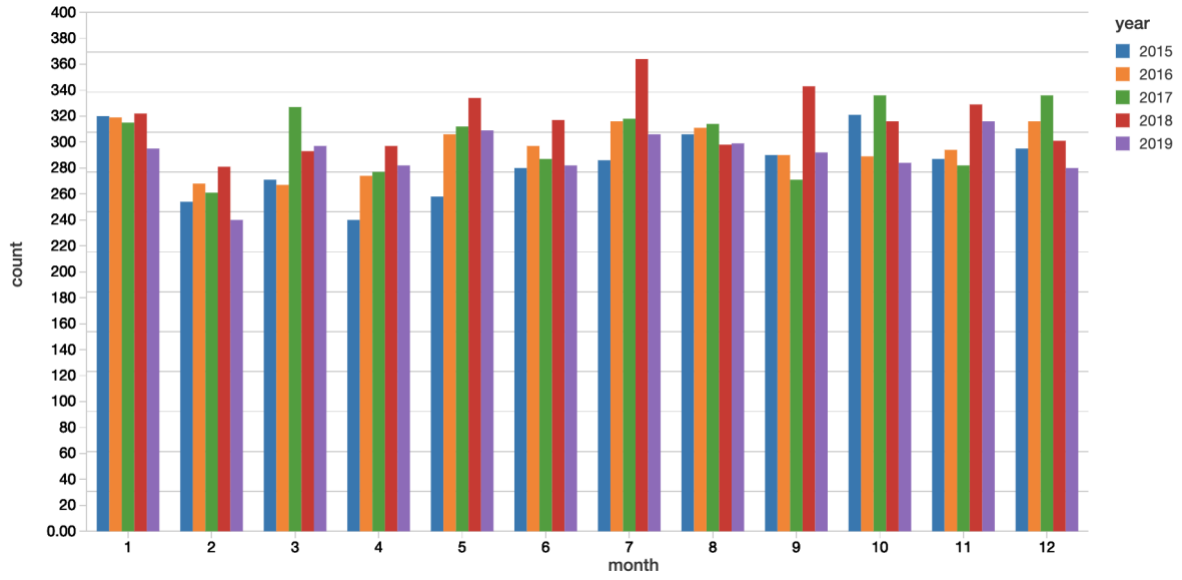


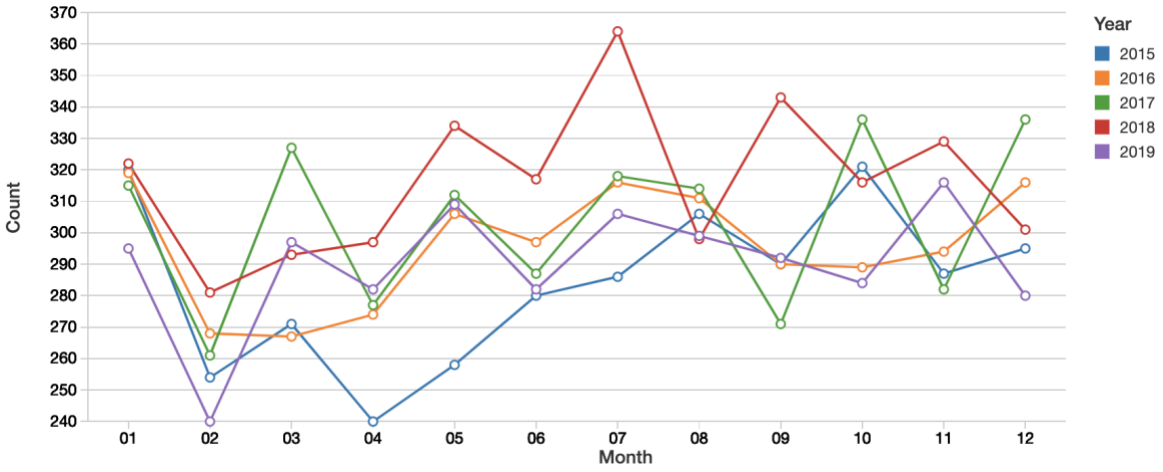
More specifically, I narrowed down into the nearest three years: 2017 to 2019 to see recent patterns of each month. Within these years, the crime rate in the first half of the year was relatively stable, but there may be greater fluctuations in the second half, especially in the August to October. In 2017, the November and December have relatively high crime cases as well.



Section 3.4 Monthly Crime

In addition, I showed the number of crimes in each month from 2015 to 2019. Combining the following figures, the overall crime rate in July was relatively high, especially in 2018. In addition, for 2018, July, May and September had higher crime rates. For 2017, the crime rate fluctuated greatly, which was lower in January, September and November, but suddenly increased in February, October and December. In the first half of 2015, the overall crime rate was lower than in other years. In the whole, the fluctuation of the crime rate in the second half of the year is smaller than that in the first half. In January and August, the crime rate in the same period in different years was not much different, but in other years it was quite different.

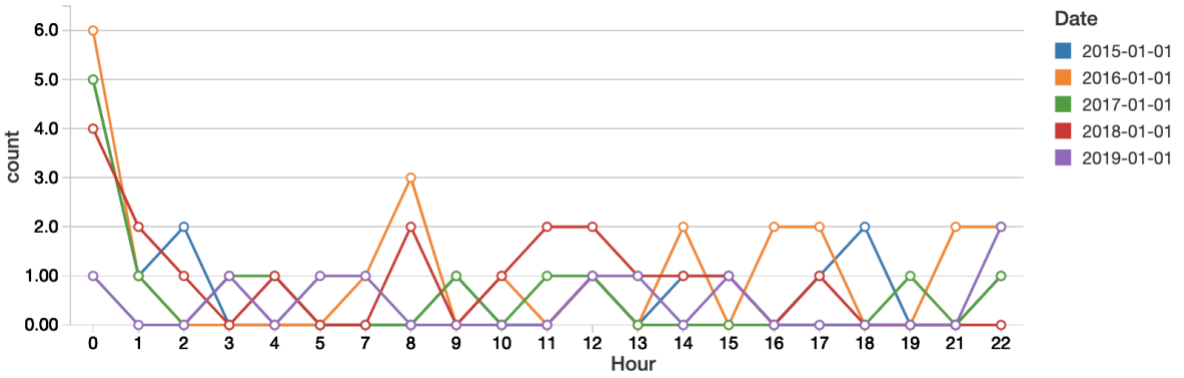




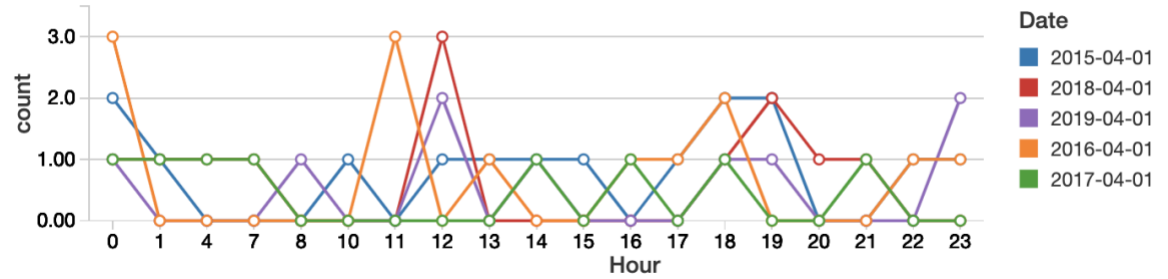
Section 3.5 Hourly Crime

In this section, I looked at the hourly crime in certain days January 1st, April 1st, July 1st, and then October 1st between 2015 to 2019.

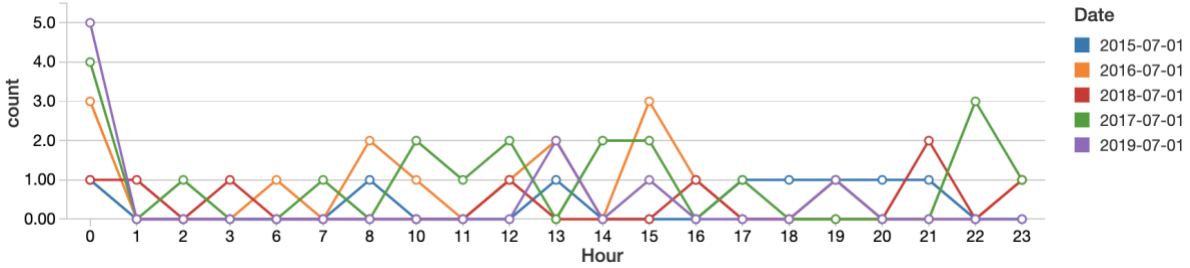
January 1st



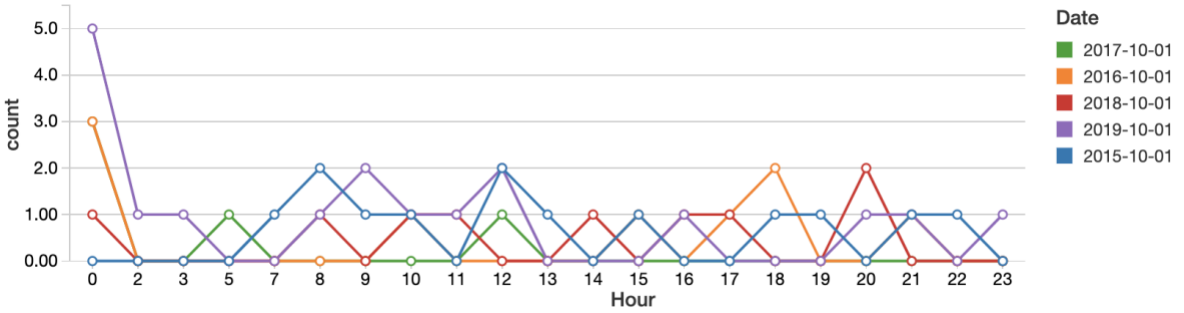
April 1st



July 1st



October 1st



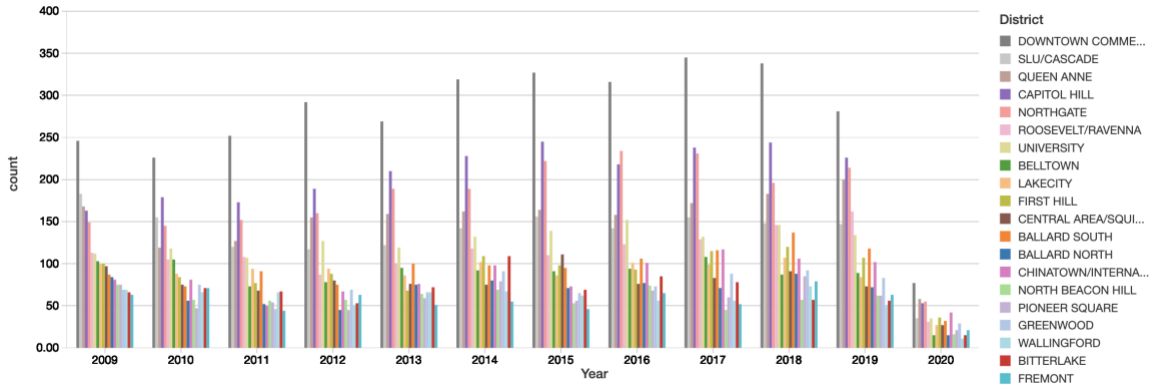
These four pictures show that midnight is when Seattle has the highest crime rate without exception, which means residents should be advised not to go out at midnight if possible. In

addition, local theft cases are known to be the most, and property crimes including burglary and theft also frequently occur at midnight. Therefore, residents are advised to protect their property, such as locking doors and windows, and paying attention to anti-theft.

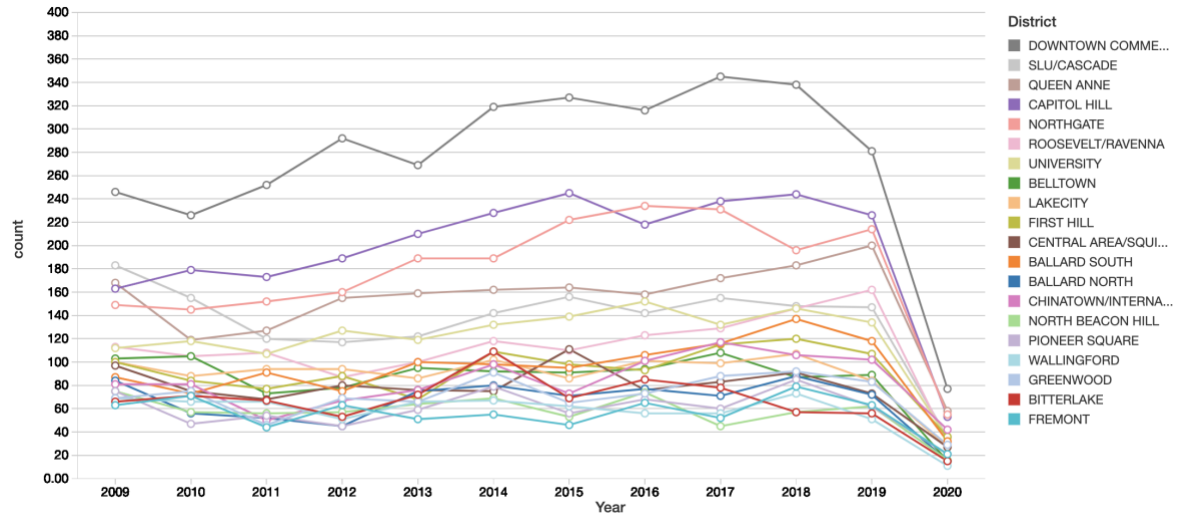
In addition, the theft rate may be different in different seasons of the year, which may be related to the local climate and temperature. Studies have shown that high temperatures can easily cause higher crime rates (Jagannathan, 2019). For example, temperatures in July are higher, especially between 2 and 4 pm, and we can speculate that the higher crime rate is due to the negative impact of high temperatures on people's emotions. Comparing the first two figures, the average crime rate in April is less than that in January. Combined with previous statistics, the crime rate in April tends to be lower than in other months. After that, I will conduct further research on this discovery.

Section 3.6 District vs. Time vs. Category

In this section I compared the crime in different districts over years, with respect to time and categories. I define the areas with the highest annual crime cases as the most dangerous areas. From the following two graphs, we can see Downtown Commercial was always the area with the most crime cases. As we discussed, this area with diverse population and business tends to provide opportunities for crimes. The second and third highest crime rate occurs in Northgate and Capitol Hill. In most years except for 2016, the crime cases in Capitol hill was larger than that of Northgate. At the same time, from 2008 to 2017, the overall crime rate showed an upward trend. The number of crimes in downtown increased year by year. Since 2018, this number has declined. It is worth noting that the data for 2020 includes only four months, and it is currently impossible to speculate whether the number of crimes in 2020 will continue to decline.



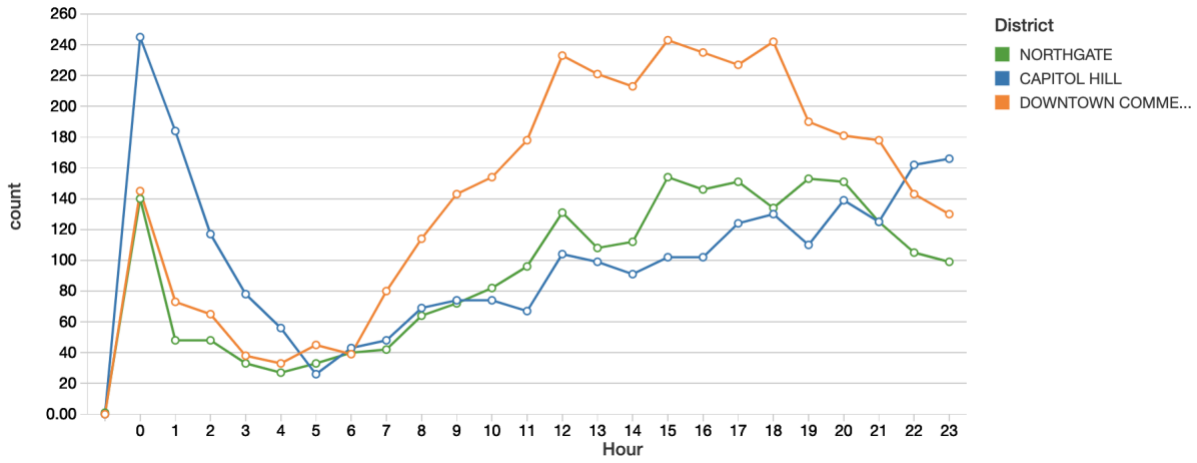
More intuitively, the crime rate in the city center is much higher than in other areas, ranging from 220 to 320. Since I randomly sampled 5% of the data, we can assume that the actual number of crimes in the city center was around 4000 to 6000 cases per year.



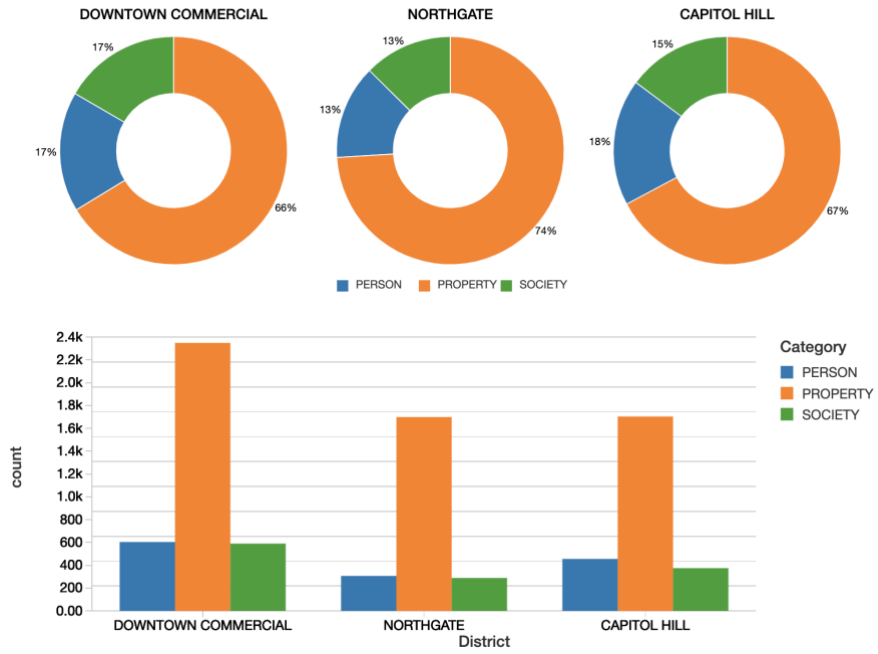
Looking further, I extracted the relationship between crime cases and time in the three most dangerous areas. As we mentioned earlier, the crime rate in these areas as a whole is relatively high at midnight. In addition, from the perspective of the Downtown Commercial, the crime rate continued to increase from 6 a.m. until it reached its peak around 12 p.m. and continued until about 6 p.m. After 6 p.m., the crime rate has declined, possibly because people in the city center come home from work and the population density has dropped rapidly.

For Capitol Hill and Northgate, the number of crimes has gradually increased from 5am without significant decline. Capitol Hill has a very high crime rate at midnight. Capitol Hill is the seat of US government agencies, including the US Capitol, the Senate, the House of Representatives, and the Neoclassical High Court. At the same time, vendors of agricultural

products, meat and cheese are open every day. The combination of politics and commerce determines the special nature of this area, so it is also a place where crime rates are relatively concentrated. Northgate is a block in the north of Seattle. It is named after and surrounds the first Northgate Mall in the United States, so Northgate has a higher crime rate as a commercial center.



In addition, I explored the crime categories in these three regions. Not surprisingly, property crimes accounted for the largest share, with Northgate's property crime rate reaching 74%. In Northgate and Downtown Commercial, the number of person and society crime cases are almost the same.



Based on the above research, I make the following suggestions for the most dangerous areas in Seattle. First of all, always pay attention to property safety, especially in these areas. Secondly, if it is not necessary, you can try to minimize living or staying in these areas. Third, avoid going out at midnight. At the same time, as crime rates tend to rise, people should be more vigilant about crime in the afternoon.

Section 4 Unsupervised Learning

This part is to be continued…

Conclusion

In this section of data exploration, I studied on the dataset of crime in Seattle from 2008 to 2020 in six aspects to seek crime patterns in Seattle and give recommendations for residents on how to avoid crimes and stay safe. This process is conducted by Apache Spark, using multiple tools such as Spark SQL, Dataframe, OLAP, visualization, etc. after essential data sampling and processing. From the data above, there are some main points:

- The crime rate in the whole Seattle increased until 2017 and then tended to decline.
- Among three main categories of crimes, the property crime accounted for 73.7%.
- Among 30 parent groups of crimes, the larceny theft had the most cases.
- Downtown commercial, Capitol Hill, and Northgate were the most dangerous areas.
- Crime rates on Sundays were high in 2008 and 2015.
- April had relatively low crime rate and July had relatively high crime rate.
- Midnight was the most dangerous time.
- Afternoons in summer tended to have higher crime cases.

Reference

Dataset: SPD Crime Data: 2008-Present

https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5

Jagannathan.,M (2019) High temperatures can lead to more violent crime, study finds https://nypost.com/2019/06/18/high-temperatures-can-lead-to-more-violent-crime-study-finds/

Appendix

Explanation of A/B Offense Group

Group "A" Offenses

NIBRS OFFENSES	NIBRS CODES	NIBRS OFFENSES	NIBRS CODES
Arson	200	**Human Trafficking**	
		-Commercial Sex Acts	64A
Assault Offenses		-Involuntary Servitude	64B
-Aggravated Assault	13A		
-Simple Assault	13B	**Kidnapping/Abduction**	100
-Intimidation	13C		
		Larceny/Theft Offenses	
Bribery	510	-Pocket Picking	23A
		-Purse Snatching	23B
Burglary/B&E	220	-Shoplifting	23C
		-Theft from Building	23D
Counterfeiting/Forgery	250	-Theft from Coin-Operated Machine or Device	23E
		-Theft from Motor Vehicle	23F
Destruction/Damage/Vandalism of Property	290	-Theft of Motor Vehicle Parts or Accessories	23G
		-All Other Larceny	23H
Drug/Narcotic Offenses			
-Drug/Narcotic Violations	35A	**Motor Vehicle Theft**	240
-Drug/Narcotic Equip. Violations	35B		
		Pornography/Obscene Material	370
Embezzlement	270		
		Prostitution Offenses	
Extortion/Blackmail	210	-Prostitution	40A
		-Assisting or Promoting Prostitution	40B
Fraud Offenses		-Purchasing Prostitution	40C
-False Pretenses/Swindle/ Confidence Games	26A		
-Credit Card/Automatic Teller Machine Fraud	26B	**Robbery**	120
-Impersonation	26C		
-Welfare Fraud	26D	**Sex Offenses (Forcible)**	
-Wire Fraud	26E	-Forcible Rape	11A
		-Forcible Sodomy	11B
		-Sexual Assault with An Object	11C
Gambling Offenses		-Forcible Fondling	11D
-Betting/Wagering	39A	**Sex Offenses (Non-Forcible)**	
-Operating/Promoting/ Assisting Gambling	39B	-Incest	36A
-Gambling Equip. Violations	39C	-Statutory Rape	36B
-Sports Tampering	39D		
		Stolen Property Offenses	280
Homicide Offenses			
-Murder/Non-Negligent Manslaughter	09A	**Weapon Law Violations**	520
-Negligent Manslaughter	09B		
-Justifiable Homicide	09C		

Group "B" Offenses

Group B's MUST have an arrest to be NIBRS Reportable

NIBRS OFFENSES	NIBRS CODES	NIBRS OFFENSES	NIBRS CODES
Bad Checks	90A	**Family Offenses, Non-Violent**	90F
Curfew/Loitering/Vagrancy Violations	90B	**Liquor Law Violations**	90G
Disorderly Conduct	90C	**Peeping Tom**	90H
Driving Under the Influence	90D	**Trespassing**	90J
Drunkenness	90E	**All Other Offenses**	90Z

Source: Association of State Uniform Crime Reporting Programs (ASUCRP). Accessed on June 6, 2014.