The Analysis of Seattle City Crime

Yu’ang Hou, Ruobing Chen, Alec Gao

Summary of questions and results.

1. Question 1: What is the geographical distribution of crime incidents in Seattle city? Which area has the highest crime rate in Seattle city?
   a. The top three areas with the highest number of cases during 2016-2021 are Downtown commercial, Capitol Hill, and Northgate.
2. Question 2: What is the correlation between the income level and crime rate?
   a. With the higher income level, the area would have a higher crime rate.
3. Question 3: Which is the most common type of crime and trend of top ten most common types of crime in Seattle?
   a. The most common crime is larceny-theft and it decreased from 2016-2019 and rapidly increased from 2019-2021.
4. Question 4:Which is the most common type of crime and their trend of changes in the University district and what is the trend for the top ten most common crimes by timezone?
   a. The most common type of crime is still larceny-theft in the University district and larceny-theft usually occurs in the afternoon.
5. Question 5: How accurately can the ML model predict the type of crime that occurred given the time and location? Which information serves as the most important feature to the prediction accuracy?
   a. Given the time and location the ML model only had 30 percent accuracy, and the most important feature is time in this case.

Motivation:

Since the Covid-19 pandemic broke out, hundreds of incidents related to violence happened and were reported on TV and media. For example, hundreds of Asians encountered lethal attacks by these racist Asian haters. Numerous luxury stores were smashed and commodities that worthed millions of dollars were stolen. Many people have expressed their concern regarding the daily personal safety level given the current chaos in society. As students who study and live in Seattle, the well-being and safety level of where we live is crucially important. In this project, we would like to investigate how crimes vary in terms of locations, time, and types specifically in the Seattle city area. We are also interested in data particularly pertaining to the University district. The results from our research can provide our students recommendations and alerts on which area might need more attention when traveling and the statistics calculated from our dataset would be beneficial for the public to get closer attention to the safety issue.

Dataset

We used several datasets from public sources. The first one is SPD Crime data from 2008-present by the Seattle government. It has 17 columns that describe every single crime case in Seattle. The columns are respectively: report number, offense id, offense start date time, offense end date time, report date time, group A B, crime against category, offense parent group, offense, offense code, precinct, sector, MCPP, 100 block address, longitude, latitude. We will use several columns for our analysis. The dataset is public online and can be downloaded by clicking the top right corner.
https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5
We will also use data from ziptlas.com that has the average income by household in Seattle by zip codes. It is then easy for us to plot the average income in each area in Seattle and make the comparison with its crime rate.
http://zipatlas.com/us/wa/seattle/zip-code-comparison/median-household-income.htm

Method

1. Data cleaning and preprocessing
    a. Filter the first dataset only includes from 2016 to April, 2022.
    b. Filter the following columns from CSV that could be helpful for data manipulation: offense start date time, offense end date time, report date time, offense parent group, longitude, latitude,
    c. Add the additional column zip code as a category, which is calculated using longitude and latitude.
2. Subtasks for each question
    a. What is the geographical distribution of the crime rate? Which area has the highest crime rate in Seattle city?
        i. For this question, we will use latitude, longitude, and criminal cases. To calculate the crime, we would count the incidents for each district and plot them on maps. The color of the heat map is proportional to the crime cases in the area.
    b. What is the correlation between income level and crime rate?
        - For this question, we will need crime data from the first dataset and the income level and population data from the second data set. Due to a huge dataset and rate-limiting API, we narrow down the data to December 2020 (more details will be discussed in the limitation section). Using geolocator API to convert longitude and latitude into zip code in the first data set and we join the two datasets using Zip code. Then, we will calculate the crime rate (incident count/population*100000) for each county in the first data set. We will plot the scatter plot to showcase the relationship between crime rate and income level.

c. Which is the most common type of crime and trend of top ten most common types of crime in Seattle?
    - For this question, we used pandas, seanborn, and matplotlib as our fundamental tools. We first used pandas to find the counts of crimes using groupby and count, and then plotted them on a pie chart that shows the proportion of each type of crime. Then we used the size function to find the total counts of ten most common crimes by year and plot them on a line graph using seaborn and matplotlib.

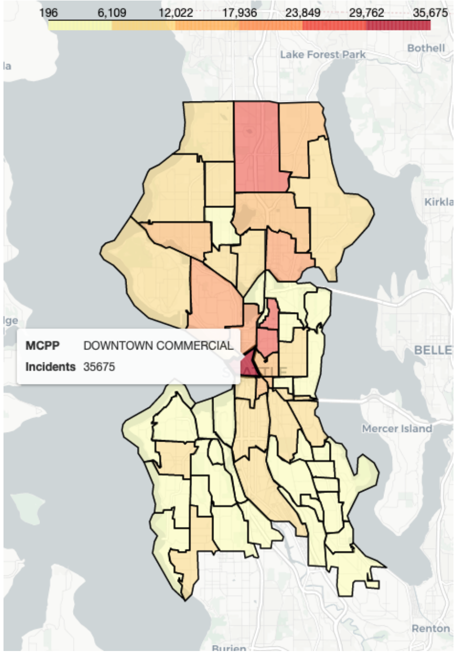
d. Which is the most common type of crime and their trend of changes in the University district and what is the trend for the top ten most common crimes by timezone?
    - For this question, we used pandas to filter the dataset that only includes the "University" MCPP and plotted a pie graph to show the proportion of each type of crime related to this area. We also selected the ten most common crimes and used the seaborn relplot to show the change of crimes by their types in these years. We loop through each case and add a new column that indicates what timezone that the case belongs to. We divide the whole day into 4 sections including early morning, morning, afternoon, and night. We used the size function to find the total counts of ten most common crimes by their timezone and plotted them on a line graph by seaborn and matplotlib.

e. How accurately can the ML model predict the type of crime that occured given the time and location? Which information serves as the most important feature to the prediction accuracy?
    - For this question, we aim to predict the crime type that might occur in a specific area within a particular time by using the ML model. To predict the crime type, we will have several related features of crime including the occurrence location, the occurrence time, the occurrence day of week. Then we will use the decision tree classifier to create the model to predict the type.

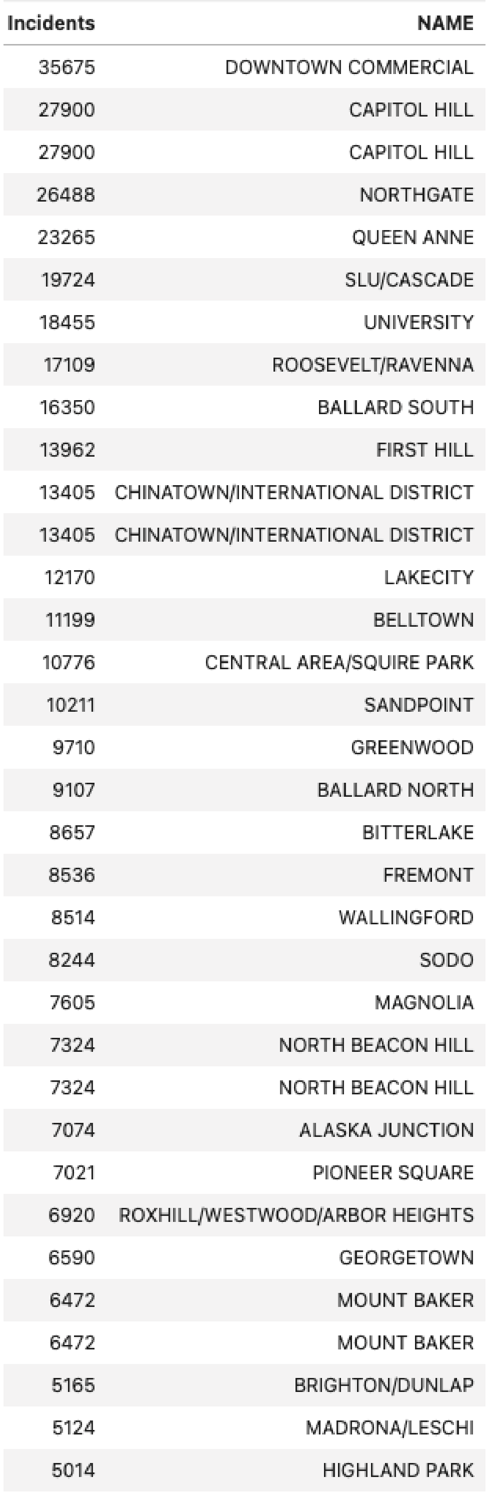
f. What type of crime do we need the most attention to in each district area?
    - For this question, we will first calculate the number of incidents based on different crime types and plot the crime distribution bar chart showing the most frequent crime type that occurs in Seattle City. For each district area, we will calculate the proportion of each type of crime happening in each area by using the coordinates columns, parent group column, and crime type column. Then we will have a time-series plot with the number of incidents based on crime types over time to see the overall trend.

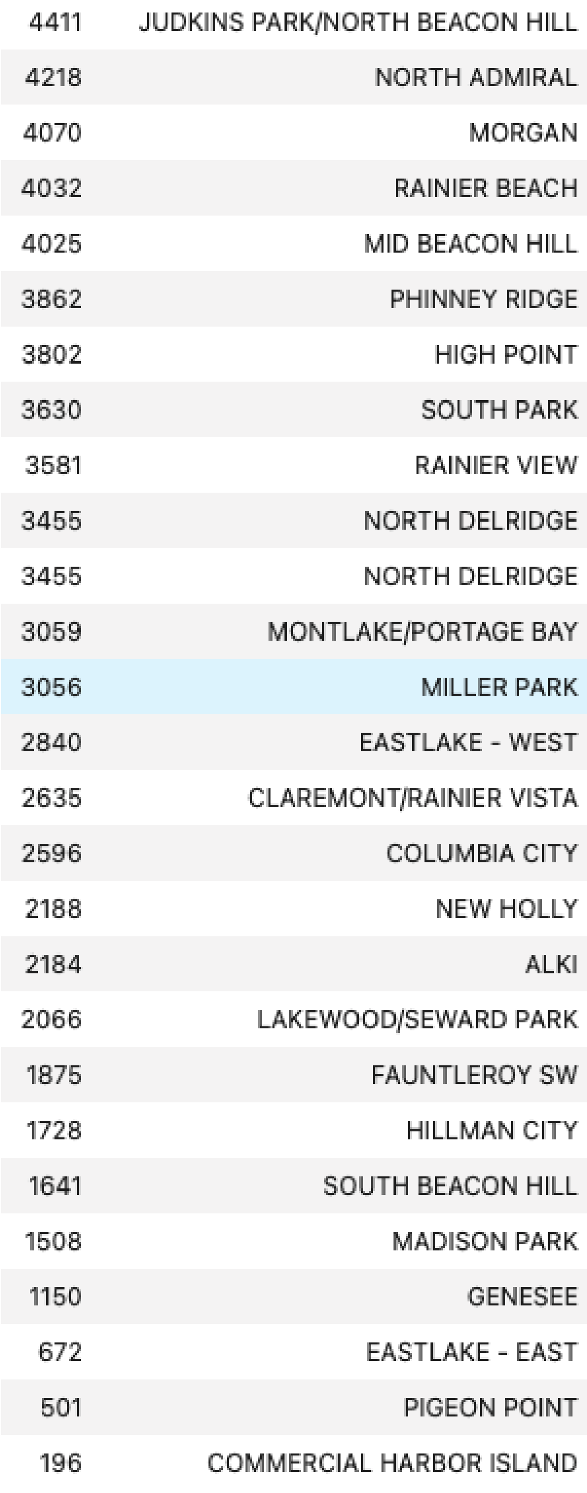
Results

1. **What is the geographical distribution of crime incidents in Seattle city? Which area has the highest crime cases in Seattle city?**
    a. From the interactive choropleth map we plotted with python below, we can see the top three areas with the highest number of cases during 2016-2021 are Downtown commercial, Capitol Hill, and Northgate. The answer is expected since the cities in dark red colors had the reputation of high crime cases compared to other areas. One thing that's interesting is that two of the areas right next to the Downtown Commercial area are only considered as a light orange level for criminal cases. Also, the southern Seattle areas have relatively less cases, which is somewhat unexpected given it has a history of lower security. The reason for this might be that we plot the number of cases instead of the crime rate.



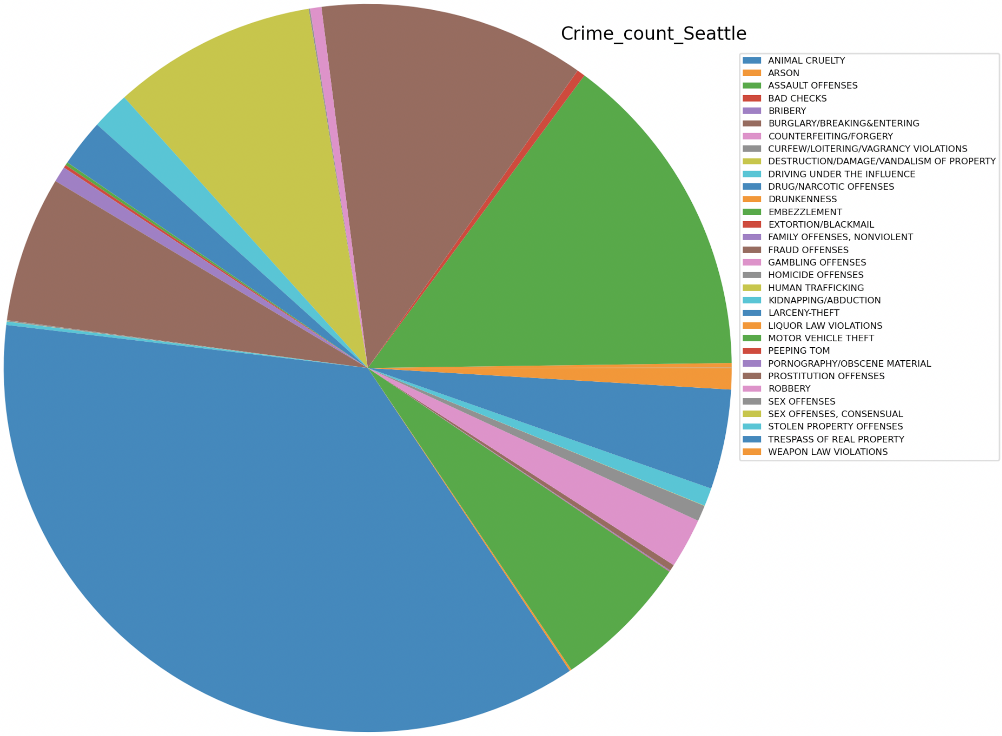
MCPP DOWNTOWN COMMERCIAL
Incidents 35675

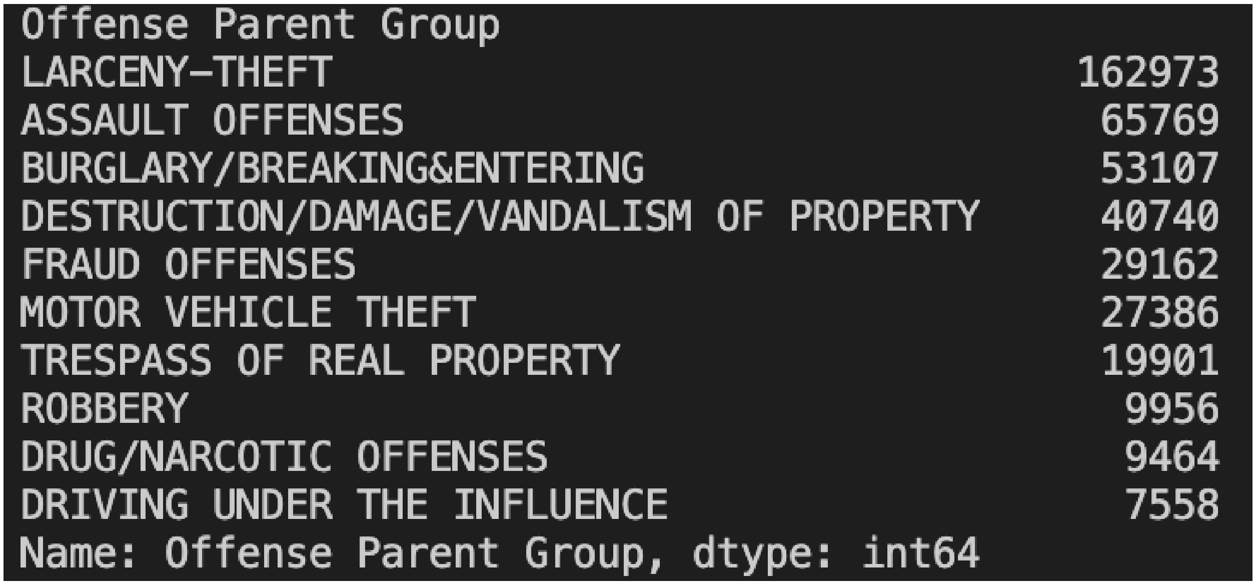




Incidents	NAME
35675	DOWNTOWN COMMERCIAL
27900	CAPITOL HILL
27900	CAPITOL HILL
26488	NORTHGATE
23265	QUEEN ANNE
19724	SLU/CASCADE
18455	UNIVERSITY
17109	ROOSEVELT/RAVENNA
16350	BALLARD SOUTH
13962	FIRST HILL
13405	CHINATOWN/INTERNATIONAL DISTRICT
13405	CHINATOWN/INTERNATIONAL DISTRICT
12170	LAKECITY
11199	BELLTOWN
10776	CENTRAL AREA/SQUIRE PARK
10211	SANDPOINT
9710	GREENWOOD
9107	BALLARD NORTH
8657	BITTERLAKE
8536	FREMONT
8514	WALLINGFORD
8244	SODO
7605	MAGNOLIA
7324	NORTH BEACON HILL
7324	NORTH BEACON HILL
7074	ALASKA JUNCTION
7021	PIONEER SQUARE
6920	ROXHILL/WESTWOOD/ARBOR HEIGHTS
6590	GEORGETOWN
6472	MOUNT BAKER
6472	MOUNT BAKER
5165	BRIGHTON/DUNLAP
5124	MADRONA/LESCHI
5014	HIGHLAND PARK
4411	JUDKINS PARK/NORTH BEACON HILL
4218	NORTH ADMIRAL
4070	MORGAN
4032	RAINIER BEACH
4025	MID BEACON HILL
3862	PHINNEY RIDGE
3802	HIGH POINT
3630	SOUTH PARK
3581	RAINIER VIEW
3455	NORTH DELRIDGE
3455	NORTH DELRIDGE
3059	MONTLAKE/PORTAGE BAY
3056	MILLER PARK
2840	EASTLAKE - WEST
2635	CLAREMONT/RAINIER VISTA
2596	COLUMBIA CITY
2188	NEW HOLLY
2184	ALKI
2066	LAKEWOOD/SEWARD PARK
1875	FAUNTLEROY SW
1728	HILLMAN CITY
1641	SOUTH BEACON HILL
1508	MADISON PARK
1150	GENESEE
672	EASTLAKE - EAST
501	PIGEON POINT
196	COMMERCIAL HARBOR ISLAND

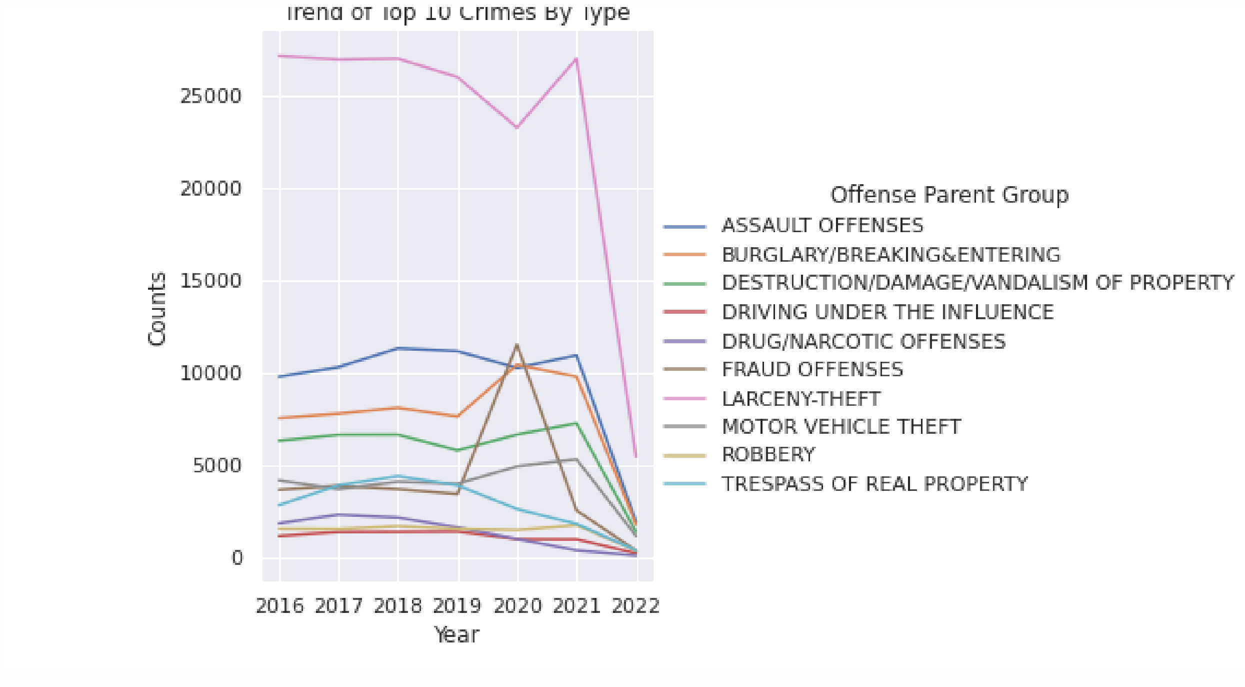
2. **Which is the proportion and trend of each type of crime in Seattle?**
   a. For the Seattle city area, the proportion of each type of crime and ten most common crimes are:



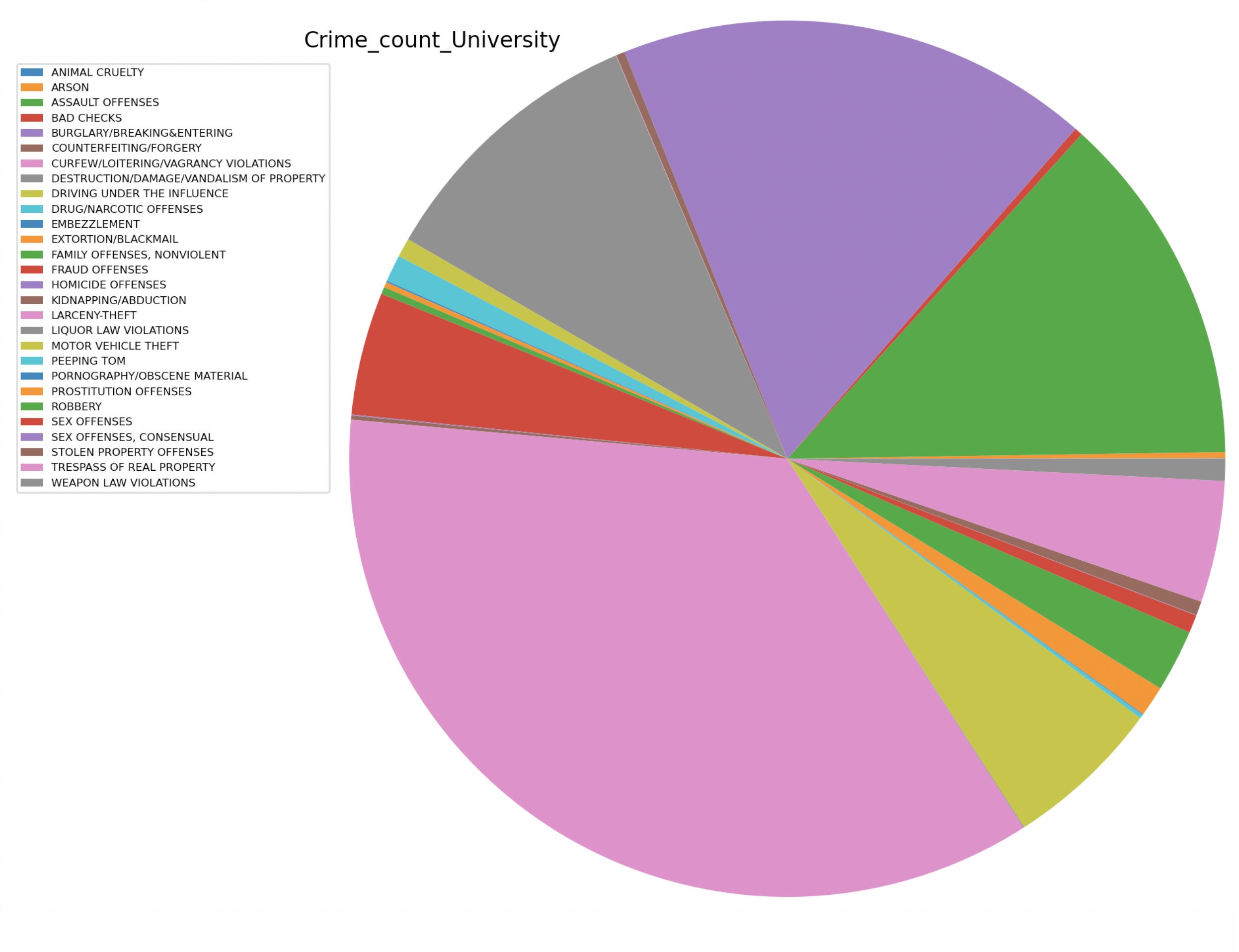


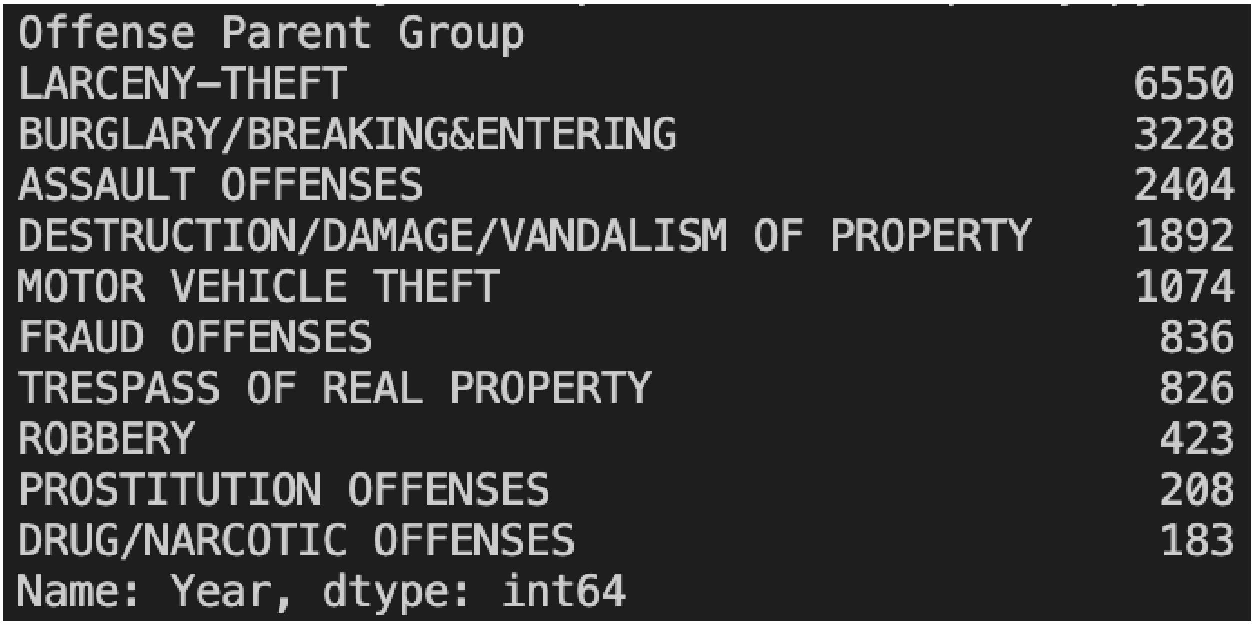
```
Offense Parent Group
LARCENY-THEFT                                   162973
ASSAULT OFFENSES                                 65769
BURGLARY/BREAKING&ENTERING                       53107
DESTRUCTION/DAMAGE/VANDALISM OF PROPERTY         40740
FRAUD OFFENSES                                   29162
MOTOR VEHICLE THEFT                              27386
TRESPASS OF REAL PROPERTY                        19901
ROBBERY                                           9956
DRUG/NARCOTIC OFFENSES                            9464
DRIVING UNDER THE INFLUENCE                       7558
Name: Offense Parent Group, dtype: int64
```

From the result we can conclude that larceny-theft took the major part of all crimes in Seattle from 2016 to April 2021. Also, Assault offense, Burglary, Damage of property, Fraud offenses, and Motor vehicle theft took major components of the all crimes. Looking at the trend, noticeably larceny-theft followed a decreasing trend from 2016-2020 and rapidly increased in 2021 Other types including Burglary, Damage to property, Assault offenses, Fraud offenses also followed an increasing trend from 2019-2021. However, due to the incomplete data from 2022, we cannot see the trend for 2022. The sudden increase from 2019-2021 was primarily due to the burst of Covid 19 pandemic. Some radical activists were strongly opposed to government quarantine policy and lockdown of the city, thus trying to draw more public attention by crimes. The increase of crimes also might be due to the reduction of police force so that people could make money by theft or other crimes.



3. **Which is the proportion of each type of crime in the University district and what time of the day has the most common crimes?**
    a. The overall proportion for each type of crime for the university district is shown in the following pie chart. And noticeably prostitution offense replaced the position of driving under the influence. The high number of prositution offense in the University district, from our analysis, was unexpected. The potential reason might be the weak public attention on the prositution offense in the region.

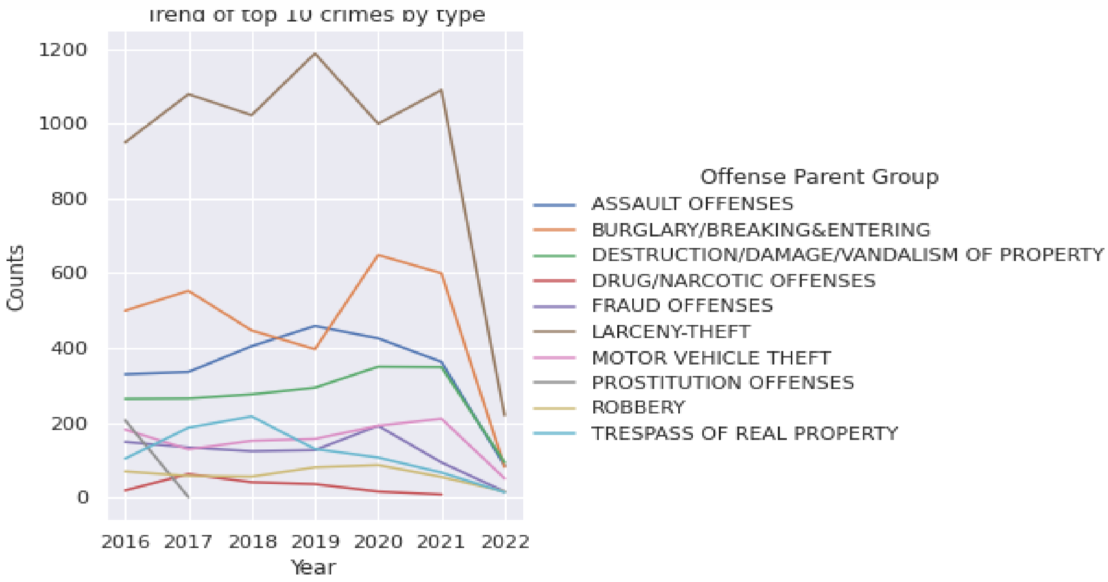




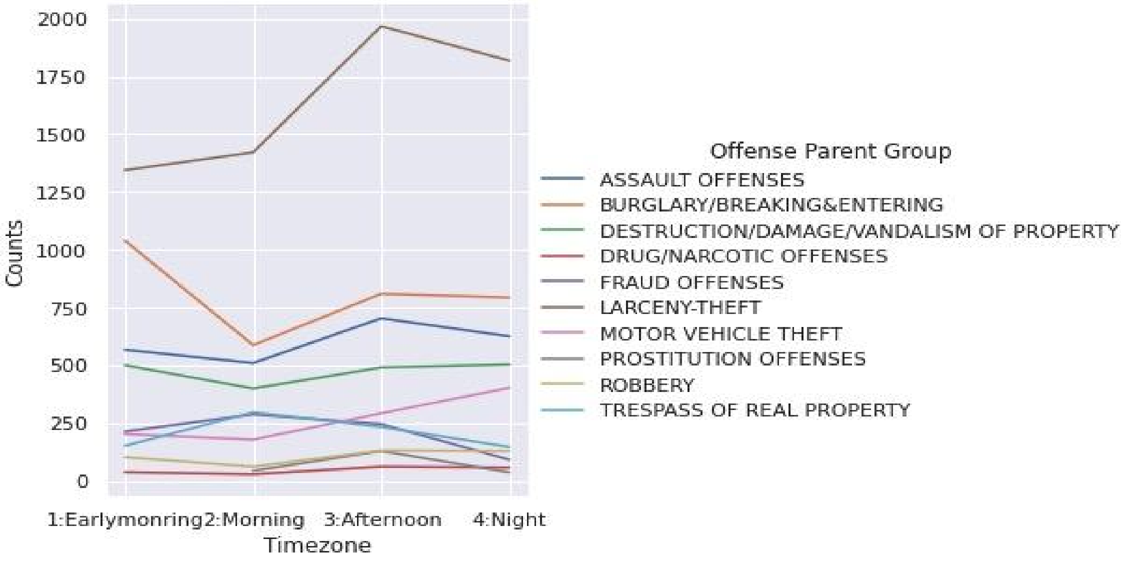
```
Offense Parent Group
LARCENY-THEFT                                   6550
BURGLARY/BREAKING&ENTERING                      3228
ASSAULT OFFENSES                                2404
DESTRUCTION/DAMAGE/VANDALISM OF PROPERTY        1892
MOTOR VEHICLE THEFT                             1074
FRAUD OFFENSES                                   836
TRESPASS OF REAL PROPERTY                        826
ROBBERY                                          423
PROSTITUTION OFFENSES                            208
DRUG/NARCOTIC OFFENSES                           183
Name: Year, dtype: int64
```

Regarding the trend of most common crimes in the University district, we found the uniform increase of every type of crime besides Drug offense and Trespass of real property. And during

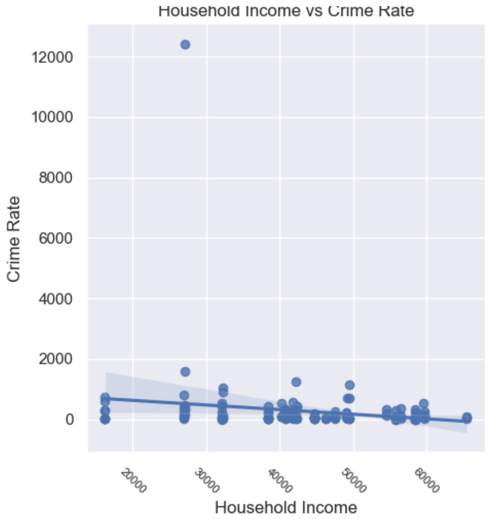
2020-2021, Larceny-theft, Burglary, Damage of property, Motor vehicle theft, and Fraud offense reached their highest level. This was also caused by radical activists and those who tried to make money for their own interests. The data for 2022 is still updating. Thus, we cannot extend our conclusion regarding 2022.

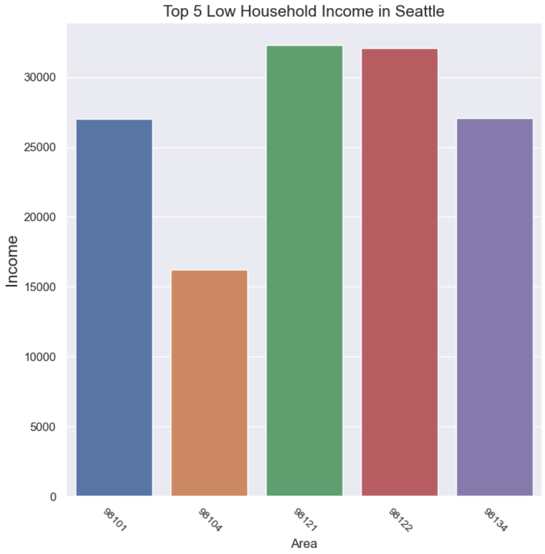


Regarding the change of most common crime by their timezone, Larceny-theft commonly happened in the afternoon and night. Other kinds of crimes have a low rate of happening in the morning. Noticeably, burglaries usually happened in the early morning. This was primarily because in the early morning people are sleeping and unaware of the situation around them. Also, Motor vehicle theft usually occurred at night. People tend not to use motor vehicles at night and at this time motor vehicles are at risk of getting stolen.

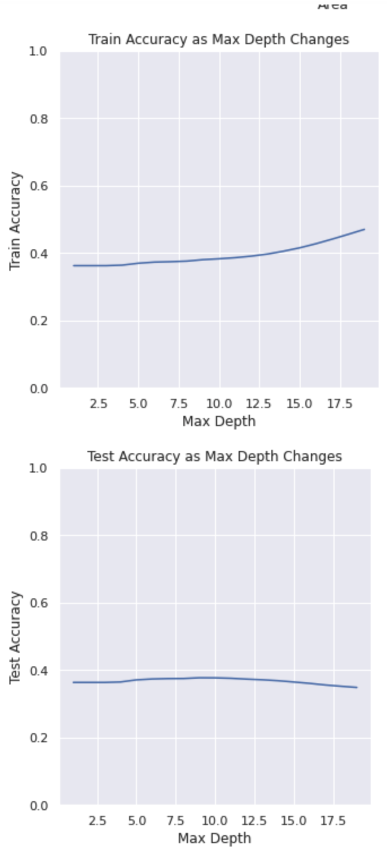


4. **What is the correlation between income level and crime rate?**
    a. In many cities, poverty is associated with a high crime rate. When calculating the crime rate, the population in each area is also put into account. In terms of crime rate per 100000, the first scatterplot below has shown the negative relationship between income and crime rate with outliers. The correlation value is -0.166 which is far from 1 or -1. It indicates that income level and the crime rate are negatively correlated at a low level. To dig more into this, according to the crime map in the first question, the area with the high crime rate is Downtown commercial, Capitol Hill, and Northgate where the household income is relatively low in the area of Downtown (98121) and Capitol Hill (98122) in the bar chart. It is interesting to discover that both results show that with the higher household income, the area in Seattle would have a higher crime rate with an increased risk of violence.





5. **How accurately can the ML model predict the type of crime that occurred given the time and location? Which information serves as the most important feature to the prediction accuracy?**
    1. We used the type of crime as a label and all other variables as features in this model. By splitting the data into 20 percent train and 80 percent test, we got an accuracy of about 30 percent. Which is not ideal. The most important factor here is time of the day, this is reasonable since crimes tend to happen in the evening. One possible reason for this might be that the data we have is not large enough to train a dataset with 30 plus labels. In the future, we might try to get a larger dataset or eliminate some labels that only have a small percentage of incidents.
        a. On the other hand, the best max in depth will be around 10, since the test accuracy will decrease with a max depth higher than 10.



Impact and Limitations

Impact and benefit

By analyzing results, we could draw implications including the trends of crime geographically, the occurrence of different crime types, the likelihood of crime incidents at different times of the day, and the main aspect that contributes to predicting the crime types.

The primary purpose of crime analysis is to assist the operation of the police officer. Seattle police can analyze the timezone of different types of crime in each area (MCPP). It helps police to evaluate the distribution of the police force, develop more efficient strategies to prevent crime, and improve public safety in the neighborhood. Since our analysis focuses more on the crime that occurred between 2017 to 2020, it may not be able to provide a recent crime summary.

Seattle residents will be the ones that benefit from having more information on the geographical distribution of the crime incidents that happened in Seattle. By referring to the color which indicates the severity of crimes on the map, people will raise their attention when they travel to the related areas. Also, Seattle residents can learn about the proportion of each type of crime and raise their attention on the types of crime that occur particularly frequently in their neighborhood. The analysis also draws more information for students who live near the University district. They can know how many and what kind of crimes happened every day and during what timezone. Regarding the rapid increase of larceny-theft in the morning, students and University district residents may put special attention to their house property or private belongings in the morning. The information regarding this might help prevent loss from happening.

Limitation

There are several limitations of our data that may result in biased results:

Missing Data: The data for 2022 is still updating. Therefore we are unable to draw full conclusions for the trend and timezone for crimes in 2022 since we only have data until April. However, this would not negatively affect our graph for the proportion of each type of crime since the possibility that one kind of crime occurs is independent and will affect the possibility of another kind of crime from happening.

API limitation: To discover the relationship between income and crime rate, we plan to join two datasets using zip codes. By using geolocator API, the longitude and latitude in the first dataset can be converted into zip codes. However, due to the limited rate of API (1 data point per sec), we are unable to process all the 40k data from 2017 to 2020 in an efficient time. Therefore, we narrowed the range down to December 2020 which contains 2k data information. The relatively small sample might lead to a weak correlation between income level and crime rate but not an extremely biased relationship. The conclusion can still be used to estimate the crime rate based on the relative income level.

Challenge Goals

We will have 3 challenge goals for our project including **multiple datasets** and **machine learning**, and a **new library**. First of all, we will calculate the crime rate in each zip code in Seattle city and combine the results with the educational attainment level and income level tables. This would allow us to make the regression analysis between crime rate and the other indexes in that area. We will also use machine learning to predict the type of crime in a particular time range and location. First, we will add columns that convert each incident starting time into the occurrence time and the occurrence day of the week. In this case, we will use the decision tree and classifier as our model. And then we will use the type of crime as labels and other factors including location and time information as features to train our model. In that way, we can predict if a particular type of crime usually happens at which time zone and location. At last, we will learn a new python library, plotly as our plot tool. By inputting the zip code or longitude and latitude we can see the density of crimes that happened in an area.

Work Plan Evaluation

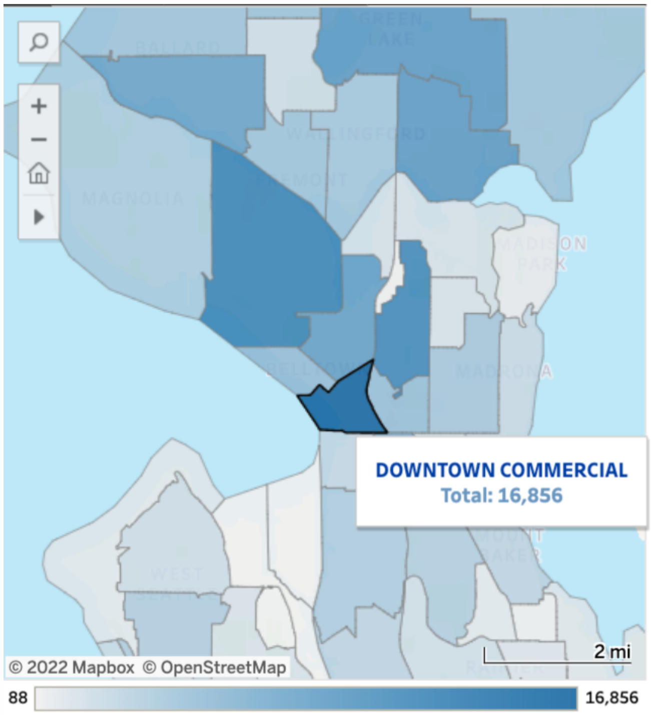
1. (Alec) Set up Github repository to share access and work - less than an hour
2. (Ruobing) Download the original CSV files and preprocessing dataset including clean up, filter and join datasets. - 1 hour
3. (All) Exploratory Analysis - estimated 3 hour for each question (expected to have initial writeup answers by 2/26)
    a. Develop code, complete the data manipulation and visualization to answer the question
    b. Responsibility:
        i. Yu'ang Hou: Finish question 3 and 5
        ii. Ruobing Chen: Finish question 2
        iii. Alec Gao: Finish question 1
4. (All) Machine Learning Model and Verification - Few hours (finished by 3/5)
    a. Since it is a challenging part for our team, all the team members would like to work together to dive deep into the method and spend more time on research question 4
5. (All) Project Writeup - 4 hours (finished by 3/10)

We followed most of our proposed work plan but we did spend more time than expected on cleaning data due to the large dataset and limited rate of API.

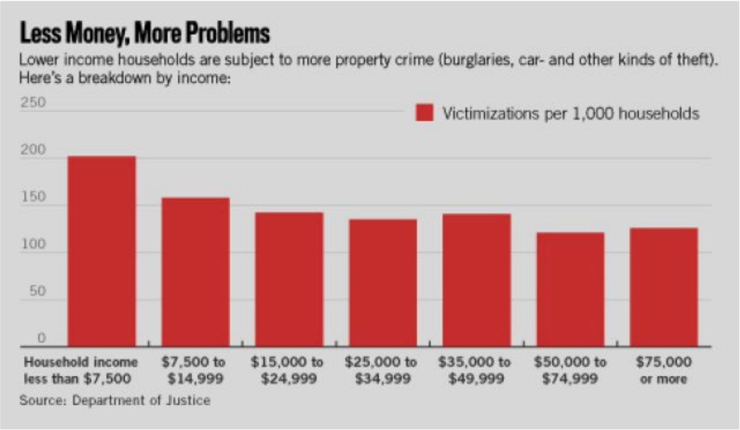
Testing Describe how you tested your code.

1. Test large crime data cleaning
    a. Use assert_equals to test the data after cleaning is valid (len(data) > 0) and the record year is after 2016.
2. Test the geographical distribution of crime incidents in Seattle city
    a. We compare our data visualization and the incident count in each area with the Seattle Government dashboard (we share the same dataset but we did not directly

use their visualization). The geographical distribution of crime incidents could mostly match with the official dashboard as below.



3. Test the relationship between income level and crime rate
   a. Test the negative relationship by comparing with other authorized sources. The resource also indicates that low income may lead to higher crime rate.


Less Money, More Problems

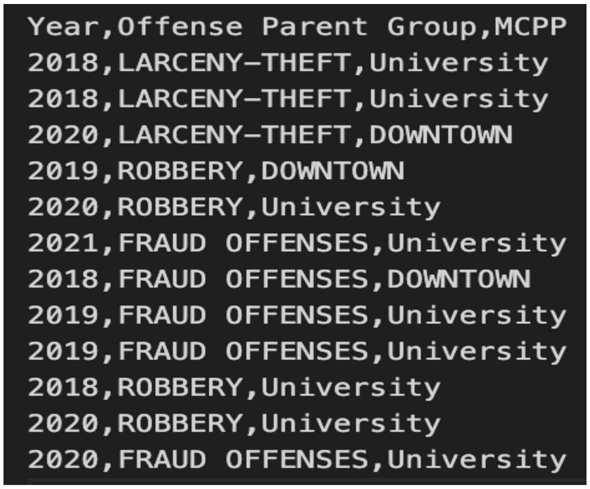
Lower income households are subject to more property crime (burglaries, car- and other kinds of theft). Here's a breakdown by income:

Source: Department of Justice

      i. https://www.usnews.com/opinion/articles/2010/10/19/property-crime-rates-by-income-level

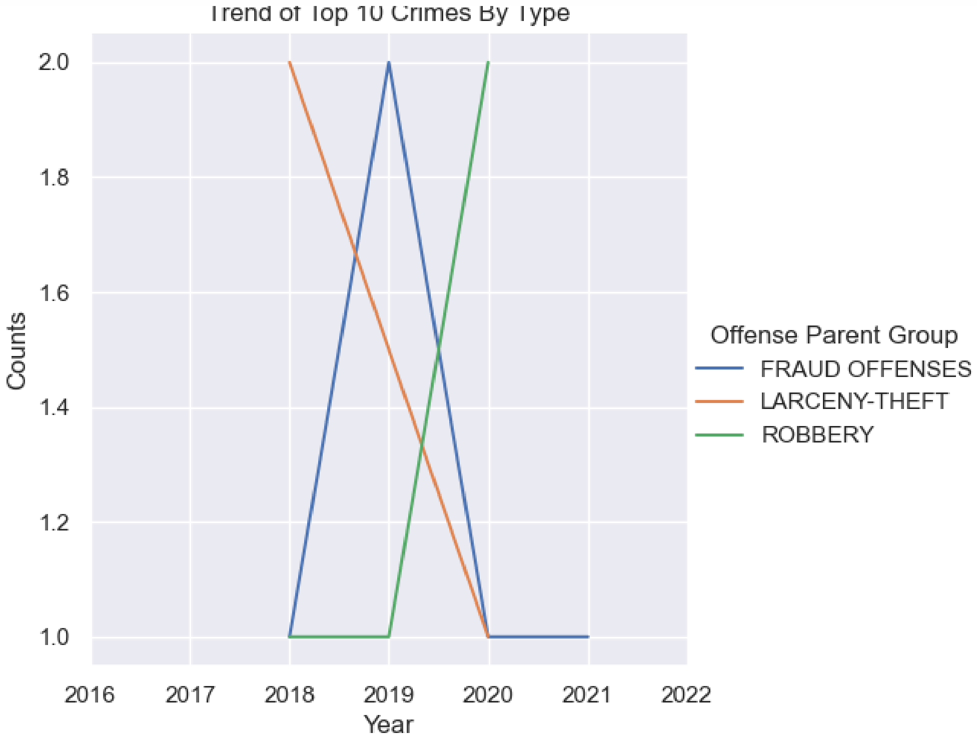
4. Test the Trend graph for each type of crime by years.
   a. We manually created a small csv dataset that has columns of Year, Offense Parent Group, and MCPP. And because our test dataset has only 12 cases we can easily

see if the graph created by codes is correct. The test data is shown in the following graph:



```
Year,Offense Parent Group,MCPP
2018,LARCENY-THEFT,University
2018,LARCENY-THEFT,University
2020,LARCENY-THEFT,DOWNTOWN
2019,ROBBERY,DOWNTOWN
2020,ROBBERY,University
2021,FRAUD OFFENSES,University
2018,FRAUD OFFENSES,DOWNTOWN
2019,FRAUD OFFENSES,University
2019,FRAUD OFFENSES,University
2018,ROBBERY,University
2020,ROBBERY,University
2020,FRAUD OFFENSES,University
```

From the graph we can see that Larceny-theft has 2 in 2018, 1 in 2020. Robbery has 1 in 2018, 1 in 2019, 2 in 2020. Fraud offenses have 1 in 2018, 2 in 2019, 1 in 2020, 1 in 2021. And we used the exact same code used to create trends for most common crime in Seattle to create graphs for our test dataset. The result is shown in the following graph.



The graph produced from the test dataset exactly matches our predetermined calculation which proves that the original codes to graph the trend for most common crime are correct and reliable. The function to graph the trend for most common crime is the fundamental function in question 3 and 4. For question 4 that focuses on University district area, we only used a mask to filter out the data that pertains to University district and create a new csv dataset for extensional use. Therefore, once we prove that the function to graph the trend for most common crime is correct, then we can also conclude that the further functions that graph for a particular area are also correct.

Collaboration State

We asked our TA for suggestions about data cleaning. Otherwise, we did not work with others from outside.