Descriptive analysis(Ayeda)

1. What is this dataset about?

The dataset is about various crimes which occur in Seattle under the jurisdiction of the Seattle Police Department, 2008-present.

2. Where did you get the dataset?

The dataset was downloaded from Seattle Open Data,
https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5

3. Who collected it?

The Seattle Police Department collected this data.

4. What can you say about the sampling? What is the sample and what is the population you want to analyze here? How exactly was the data collected? Do you think the sample is somehow biased or not?

This data was collected through cluster sampling because it is all the crime incidents in a designated area of the jurisdiction of the Seattle Police Department. In this dataset, there is not really a population because it is not describing people, it is describing crime incidents. The data was collected through the National Incident-Based Reporting System (NIBRS). The sample is biased because they intentionally chose all the crime incidents which occurred in a specific area (Seattle).

5. How trustworthy is the dataset?

The dataset is trustworthy because the Seattle Police Department provides it. The Seattle Police department is a government-funded organization. SPD was also very transparent about its reliable method of collecting the data through the NIBRS. The NIBRS is FBI-approved as well.

Data and Research Questions(Ethan)

1. Research Questions:
   Which precinct has the most criminal activity in the area?
   How many offense codes are considered non-violent?
   What crime against category is most common within the dataset?
   How many instances of Larceny-Theft from the offense parent group are there in the dataset?

2. Within the dataset, there are 17 different variables which include: Report Number, Offense ID, Offense ID Start DateTime, Offense End DateTime, Report DateTime, Group A B, Crime Against Category, Offense Parent Group, Offense, Offense Code, Precinct, Sector, Beat, MCPP, 100 Block Address, Longitude, and Latitude.

   The variables that are relevant to our project are:
   Precinct- Designated police precinct boundary where offense(s) occurred.
   Offense Code- Corresponding offense code for the previous category of offenses

Crime Against Category- Corresponding offense crime against category which entails what general type of crime was committed against society

Offense Parent group- Similar to the offense grouping but more generalized

3. For simplification, we narrowed down our data by removing and selecting relevant variables for our questions. The brief summary of our data includes:

```
> subset <- cc %>% select(Precinct, `Offense Code`, `Crime Against Category`, `Offense Parent Group`)
> summary(subset)
   Precinct         Offense Code       Crime Against Category Offense Parent Group
 Length:1012514     Length:1012514     Length:1012514         Length:1012514
 Class :character   Class :character   Class :character       Class :character
 Mode  :character   Mode  :character   Mode  :character       Mode  :character
>
```

All of our variables are considered categorical. The categories are listed below

```
> pre <- subset %>% pull(Precinct)
> oc <- subset %>% pull(`Offense Code`)
> cag <- subset %>% pull(`Crime Against Category`)
> opg <- subset %>% pull(`Offense Parent Group`)
> unique(pre)
[1] "W"       "N"       "SW"      "E"       "S"       "UNKNOWN" "OOJ"     NA        "<Null>"
> unique(oc)
 [1] "35A" "23G" "120" "290" "90D" "23C" "23F" "26E" "23D" "100" "250" "23H" "370" "210" "240"
[16] "11B" "280" "26A" "26B" "270" "26F" "26C" "520" "11D" "26G" "35B" "200" "64A" "90G" "90A"
[31] "23A" "11A" "11C" "40C" "23B" "90F" "23E" "90B" "720" "09A" "40A" "26D" "90H" "40B" "90E"
[46] "09C" "36A" "36B" "510" "39B" "39A" "09B" "39C" "90J" "13B" "13A" "13C" "220" "64B"
> unique(cag)
[1] "SOCIETY"     "PROPERTY"    "PERSON"      "NOT_A_CRIME"
> unique(opg)
 [1] "DRUG/NARCOTIC OFFENSES"                  "LARCENY-THEFT"
 [3] "ROBBERY"                                 "DESTRUCTION/DAMAGE/VANDALISM OF PROPERTY"
 [5] "DRIVING UNDER THE INFLUENCE"             "FRAUD OFFENSES"
 [7] "KIDNAPPING/ABDUCTION"                    "COUNTERFEITING/FORGERY"
 [9] "PORNOGRAPHY/OBSCENE MATERIAL"            "EXTORTION/BLACKMAIL"
[11] "MOTOR VEHICLE THEFT"                     "SEX OFFENSES"
[13] "STOLEN PROPERTY OFFENSES"                "EMBEZZLEMENT"
[15] "WEAPON LAW VIOLATIONS"                   "ARSON"
[17] "HUMAN TRAFFICKING"                       "LIQUOR LAW VIOLATIONS"
[19] "BAD CHECKS"                              "PROSTITUTION OFFENSES"
[21] "FAMILY OFFENSES, NONVIOLENT"             "CURFEW/LOITERING/VAGRANCY VIOLATIONS"
[23] "ANIMAL CRUELTY"                          "HOMICIDE OFFENSES"
[25] "PEEPING TOM"                             "DRUNKENNESS"
[27] "SEX OFFENSES, CONSENSUAL"                "BRIBERY"
[29] "GAMBLING OFFENSES"                       "TRESPASS OF REAL PROPERTY"
[31] "ASSAULT OFFENSES"                        "BURGLARY/BREAKING&ENTERING"
```

unique(pre) is under the Precinct variable with categories indicating cardinal directions and unknown, NA, and null data

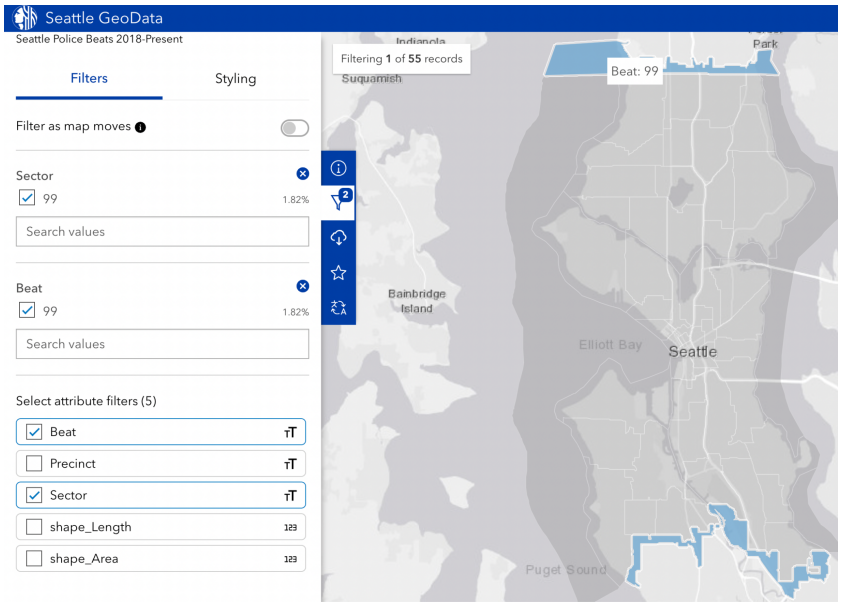
unique(oc) is under the Offense Code variable, which has over 46 different combinations of numbers and letters

unique(cag) is under the Crime Against Category variable with 4 categories, notably not a crime also listed as a category

unique(opg) is under the Offense Parent Group variable with 31 unique categorizations of crime

4. Only Precinct contains 4 NA values and various different forms of unknown data. Upon observing the unique categories for Precinct, further research indicates that there are 5 distinct precincts that SPD uses which is North, East, South, West, and Southwest. The value OOJ is used as a notation for all areas outside of the 5 major precincts which is represented as sector/beat 99 and located on the North and South edge of the greater Seattle area. Otherwise, OOJ is also used along with the unknown value which would suggest that the area of crime may or may not be within the SPDs jurisdiction. The values unknown, NA, and null represent an area outside of legal jurisdiction, not applicable to the type of crime, and no data on the location respectively.

Precinct	Sector	Beat
OOJ	OOJ	OOJ
OOJ	OOJ	OOJ
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	UNKNOWN	UNKNOWN
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	99	99
OOJ	UNKNOWN	UNKNOWN




```
> is.na(pre) %>% sum()
[1] 4
> is.na(oc) %>% sum()
[1] 0
> is.na(cag) %>% sum()
[1] 0
> is.na(opg) %>% sum()
[1] 0
```

Answer your questions(Amanda/Harrison)

3 Answer your questions (95pt)

1. (4 × 20pt) Next, answer your questions based on data. Why?

Which precinct has the most criminal activity in the area?

- N is the precinct that has the most criminal activity in the area, with a total of 324652. However, with a total of 6683 unknown values, the result might change not for the Precinct with the most criminal activity, but for those second, third and fourth will be greatly affected.

```
> SPD %>%
+   pull(Precinct) %>%
+   table()
.
 <Null>       E       N     OOJ       S      SW UNKNOWN       W
      2  158643  324652       1  145465  101261    6683  273942
```

How many offense codes are considered non-violent?

There are 103 non-violent offense codes.

There are 103 offense codes that are considered non-violent.

Some examples of non-violent offenses include: trespassing, theft, property damage, and public intoxication.

What crime against category is most common within the dataset?

Theft is the most common crime against category.

Theft is the most common crime against category because it includes all types of theft, such as shoplifting, pickpocketing, and car theft.

How many instances of Larcney-Theft from the offense parent group are there in the dataset?

- 377954 of the cases is Larceny-Theft.

```
FAMILY OFFENSES, NONVIOLENT
                      10278
          GAMBLING OFFENSES
                         27
          HUMAN TRAFFICKING
                         53
              LARCENY-THEFT
                     377954
       MOTOR VEHICLE THEFT
                      63197
PORNOGRAPHY/OBSCENE MATERIAL
                        368
                    ROBBERY
                      23566
  SEX OFFENSES, CONSENSUAL
                        191
 TRESPASS OF REAL PROPERTY
```

• If you need additional data (e.g. population figures), explain where do you get those (e.g. from census), and how reliable you find those.

• If you cannot answer a question, then explain why: is the question too unclear, or is your data

not suited for answering it? Be specific!

2. **(15pt) Discuss the limitations of your answers, data, analysis. Normally the answers do not answer everything. Why? What kind of data would you need to get better answers? Would you need better analysis methods?**

The limitations are there are thousands of unknown data in the SPD dataset, which leads to an inaccuracy of the results. For instance, if I am analyzing which Precinct has the second and third most criminal activity, those unknown values will have a huge impact on determining the results. For the offense code, I think it is hard to understand for the first time reading it. It is better and more efficient to use categorical data instead of numerical values. So that we can pull up the table of the distribution of offense code.