0.0.1 Question 1a

As with any good EDA, you try to understand the variables included.

i). What is the granularity of the data (i.e. what does each row represent)?

ii). As we discussed in class, classifications of variable conceptual types can sometimes be subjective depending on what we are doing with the dataset. Categorize each of the variables in this dataset as either

A). Quantitative: Continuous

B). Quantitative: Discrete

C). Categorical/Qualitative: Nominal

D). Categorical/Qualitative: Ordinal

Give your answer as a table in the following form:

Column Name	Category	Explanation/Reasoning
CASENO	category letter here	reasoning here
cont'd ...	...	...

Column Name	Category	Explanation/Reasoning
CASENO	Categorical/Qualitative: Nominal	This is like an ID. Since they do not have any numerical value and the order does not mean anything, it is nominal.
OFFENSE	Categorical/Qualitative: Nominal	This is describing an attribute and has no order.
EVENTDT	Quantitative: Discrete	The date is numerical, but a date cannot be negative or fractional so it is not continuous.
EVENTTM	Quantitative: Continuous	The time is numerical and continuous since it can have many decimal places
CVLEGEND	Categorical/Qualitative: Nominal	This describes the crime and has no order.
CVDOW	Categorical/Qualitative: Nominal	This seems to encode the day of the week, so it has no numerical value and no order.
InDbDate	Quantitative: Discrete	This is the date again. It is numerical, but a date cannot be negative or fractional so it is not continuous.
Block_Location	Categorical/Qualitative: Nominal or continuous	Even though the coordinates are numerical, we cannot graph with this value of the location yet.
BLKADDR	Categorical/Qualitative: Nominal	This is not a numerical value and has no order.

Column Name	Category	Explanation/Reasoning
City	Categorical/Qualitative Nominal	This is not a numerical value and has no order.
State	Categorical/Qualitative Nominal	This is not a numerical value and has no order.

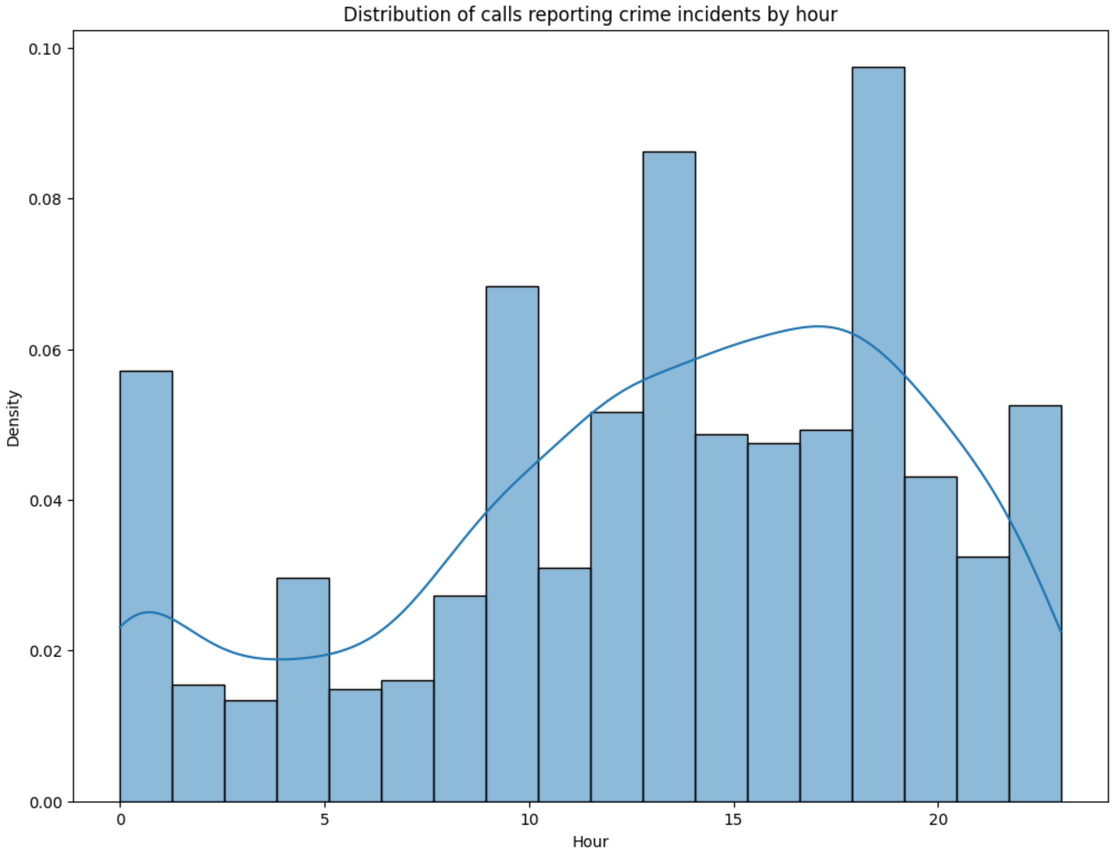
0.1 Question 2c

Use seaborn to create a **density** histogram showing the distribution of calls by hour.
Include the Kernal Density Estimate (KDE) graph on your histogram.

Be sure that your axes are labeled and that your plot is titled.

```
In [20]: sns.histplot(data = calls, x = "Hour", stat = "density", kde = True)
         plt.title("Distribution of calls reporting crime incidents by hour")
         # Your code above this line

         # Leave this for grading purposes
         ax_3d = plt.gca()
```



0.1.1 Question 2e

i). Use seaborn to construct a box plot showing the distribution of calls by hour.

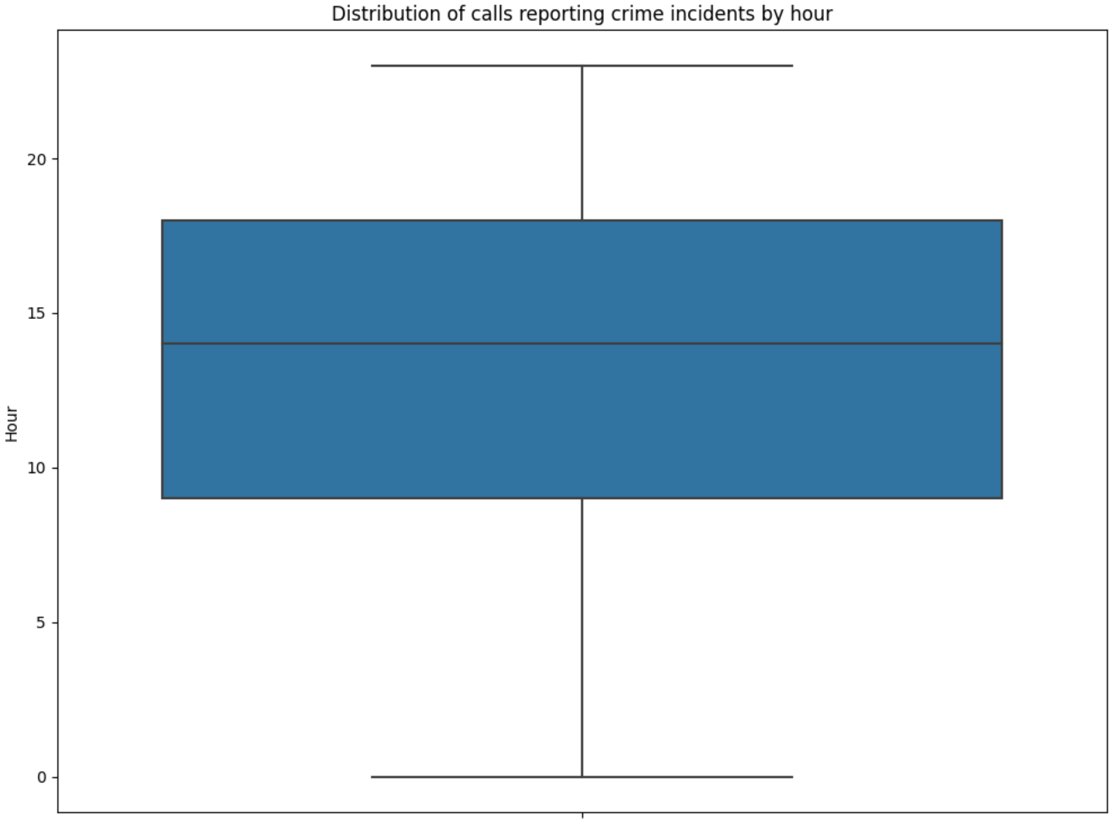
ii). To better understand the time of day a report occurs we could **stratify the analysis by DayType (i.e. by weekday vs weekends).**

Use seaborn to create side-by-side violin plots comparing the distribution of calls by hour on the weekend vs weekday (hint: see the violin plot documentation on how to stratify by a column in the dataframe https://seaborn.pydata.org/generated/seaborn.violinplot.html)

Note: For aesthetic purposes only the violin plot continues past the end of the whiskers (i.e. past 0 and 24 hours); however it is not possible to get data points outside of the whiskers for this distribution.

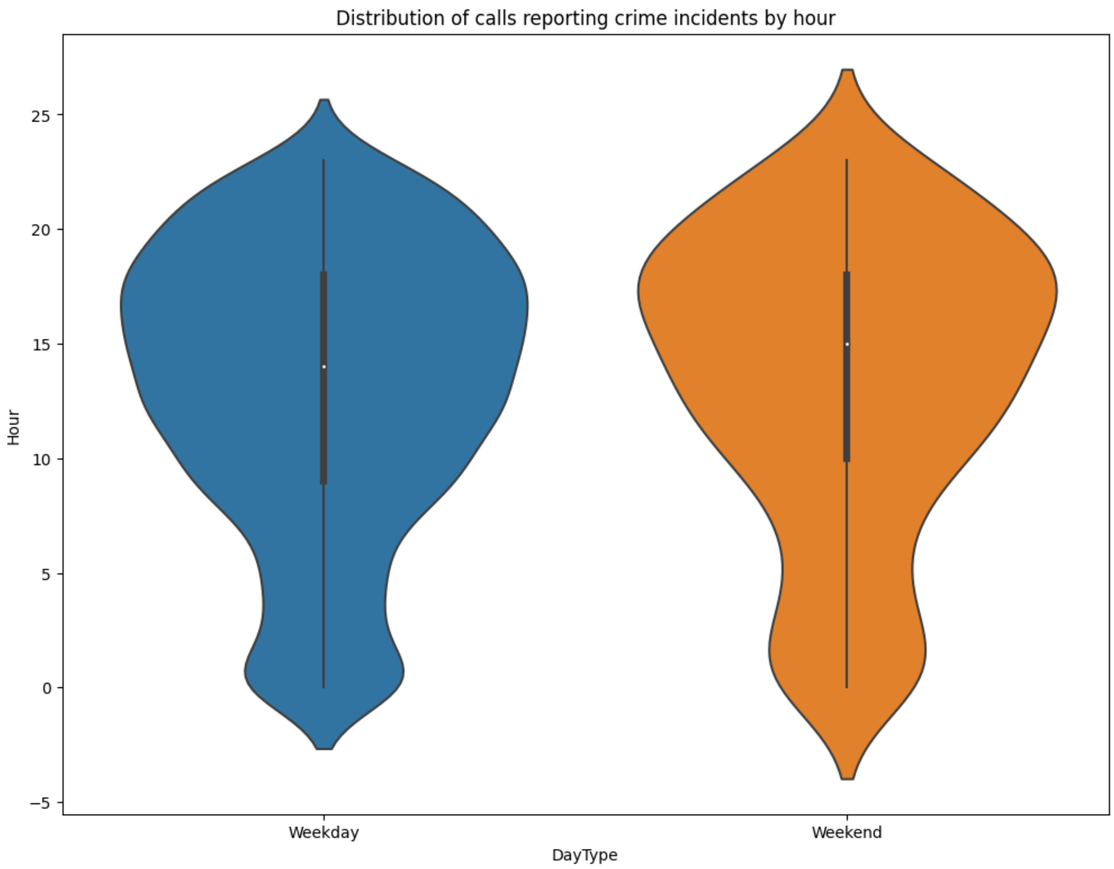
```
In [23]: sns.boxplot(y = "Hour", data = calls )
         plt.title("Distribution of calls reporting crime incidents by hour")
         # Your code for boxplot above this line
```

```
Out[23]: Text(0.5, 1.0, 'Distribution of calls reporting crime incidents by hour')
```



```
In [24]: sns.violinplot(data=calls, x = 'DayType', y = 'Hour')
         plt.title("Distribution of calls reporting crime incidents by hour")
         # Your code for side-by-side violin plots above this line
```

```
Out[24]: Text(0.5, 1.0, 'Distribution of calls reporting crime incidents by hour')
```


Distribution of calls reporting crime incidents by hour
Hour
25
20
15
10
5
0
−5
Weekday
Weekend
DayType

0.2 Question 2f

Based on your histogram, boxplot, and violin plots above, what observations can you make about the patterns of calls? Answer each of the following questions:

- Are there more calls in the day or at night?
- What are the most and least popular times?
- Do call patterns and/or IQR vary by weekend vs weekday?

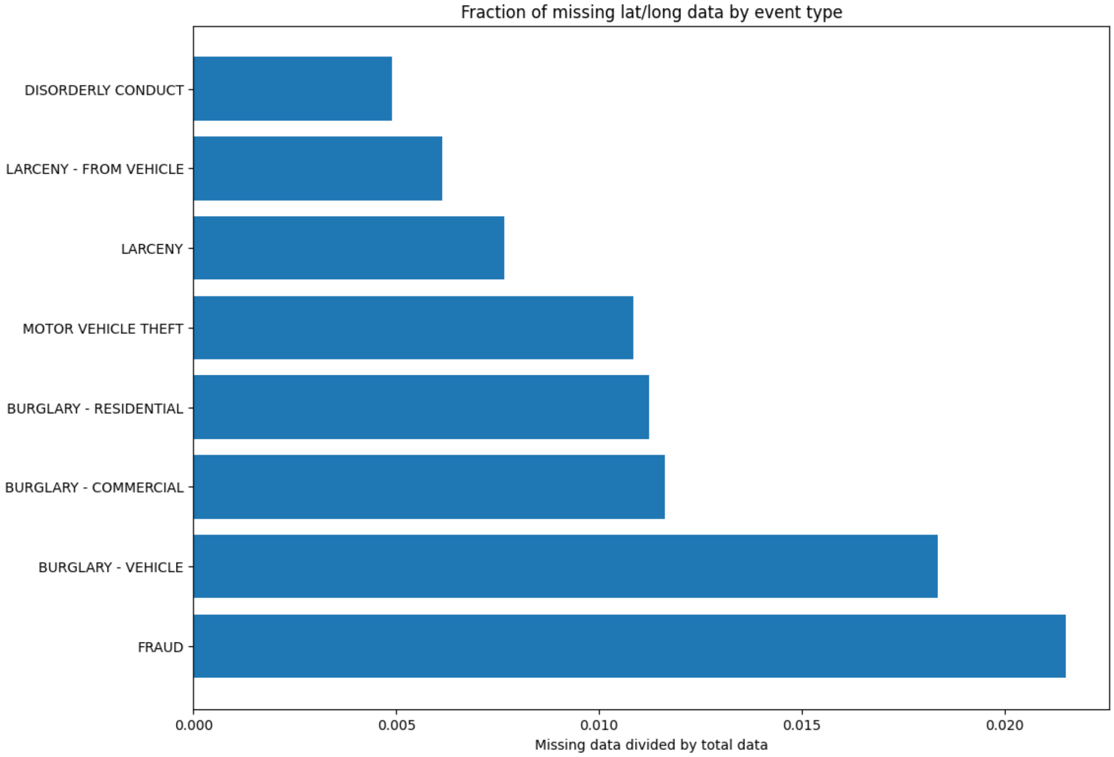
- There are more calls at night / evening.
- The most popular time is around 17:30 and the least popular time is around 4:00
- Patterns and IQR don't really vary by weekend vs weekday. They stay fairly similar

```
In [35]: missing = missing_lat_lon["CVLEGEND"].value_counts()
         total = calls["CVLEGEND"].value_counts()
         missing_by_crime = missing / total
         missing_by_crime = missing_by_crime.dropna()
         missing_by_crime = missing_by_crime.sort_values(ascending = False)
         # Your code above this line
         missing_by_crime
```

```
Out[35]: CVLEGEND
         FRAUD                     0.021505
         BURGLARY - VEHICLE        0.018349
         BURGLARY - COMMERCIAL     0.011628
         BURGLARY - RESIDENTIAL    0.011236
         MOTOR VEHICLE THEFT       0.010830
         LARCENY                   0.007673
         LARCENY - FROM VEHICLE    0.006135
         DISORDERLY CONDUCT        0.004902
         Name: count, dtype: float64
```

```
In [36]: plt.barh(missing_by_crime.index,missing_by_crime)
         plt.title("Fraction of missing lat/long data by event type")
         plt.xlabel("Missing data divided by total data")
         # Your code to create the barplot above this line
```

```
Out[36]: Text(0.5, 0, 'Missing data divided by total data')
```


Fraction of missing lat/long data by event type
DISORDERLY CONDUCT
LARCENY - FROM VEHICLE
LARCENY
MOTOR VEHICLE THEFT
BURGLARY - RESIDENTIAL
BURGLARY - COMMERCIAL
BURGLARY - VEHICLE
FRAUD
0.000
0.005
0.010
0.015
0.020
Missing data divided by total data

0.2.1 Question 3d

Based on the plots above, are there any patterns among entries that are missing latitude/longitude data?

Based on the plots above, give your recommendation as to how we should handle the missing data, and justify your answer:

Option 1). Drop rows with missing data

Option 2). Set missing data to NaN

Option 3). Impute data

Based on the plots above, there seems to be no pattern among entries that are missing lat/long data. Because the data is missing at random, we should drop the rows with missing data because the percentages are so small, meaning there is still enough observations to result in a reliable analysis. Dropping these rows will not take away a lot of information.

0.3 Question 3e

Based on the above map, what could be some **drawbacks** of using the location fields in this dataset to draw conclusions about crime in Berkeley? Here are some sub-questions to consider:

- Zoom into the map. Why are all the calls located on the street and often at intersections?
- UC Berkeley campus is on the area of the map titled “Observatory Hill”, which appears to have no calls. What are some factors about our data that could explain this? Is it really the case that their campus is the safest place to be in the area? The dataset information linked at the top of this notebook may also give more context.

Based on the above map, one potential drawback of using the location fields in this dataset to draw conclusions about crime in Berkeley is that the database is storing events based on the block level location so calls are only located on streets and intersections and might not be the most precise location. Another potential drawback is that the dataset does not consider if other police (not BPD) are present at the location. For example, UC Berkeley appears to have no calls, but it is not reliable to say that the campus is the safest area. Since UC Berkeley has its own university cops on campus, crimes would be reported to them instead of the BPD. Therefore, the BPD does not have access to this data.