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 …	…	…

Each of these rows represents a crime with a description of the offense, a time, date(s), a categorization, a location, as well as a city and state of where the crime occurred. Below is a categorization of the aforementioned table based on the options presented in A through D.

CASENO	Category	Explanation
21014296	Categorical/Qualitative: Ordinal	Since this is a crime that can be categorized based off of a value (qualitatively, theft under $950, a misdemeanor), it falls in the categorization of ordinal.
21014391	Categorical/Qualitative: Ordinal	The reasoning for this is the same as case No. 21014296.
21090494	Categorical/Qualitative: Ordinal	The reasoning for this is the same as cases No. 21014296 and 21014391.
21090204	Categorical/Qualitative: Ordinal	Similar to the misdemeanors, this is a crime that can be categorized base off of a value (qualitatively, theft over $950, a felony), it falls in the categorization of ordinal.

CASENO	Category	Explanation
21090179	Categorical/Qualitative: Nominal	This crime can be categorized, but not based off of an exact value, because this case does not include a value of the property that was stolen, so this falls in the categorization of nominal.

Crimes can be categorized based off of the severity, e.g. in theft, based upon the value of the property that was stolen. In the context of the auto burglary, we don't know the value of the property that was stolen, so this can't necessarily be categorized on a fixed value like the severity of the thefts.

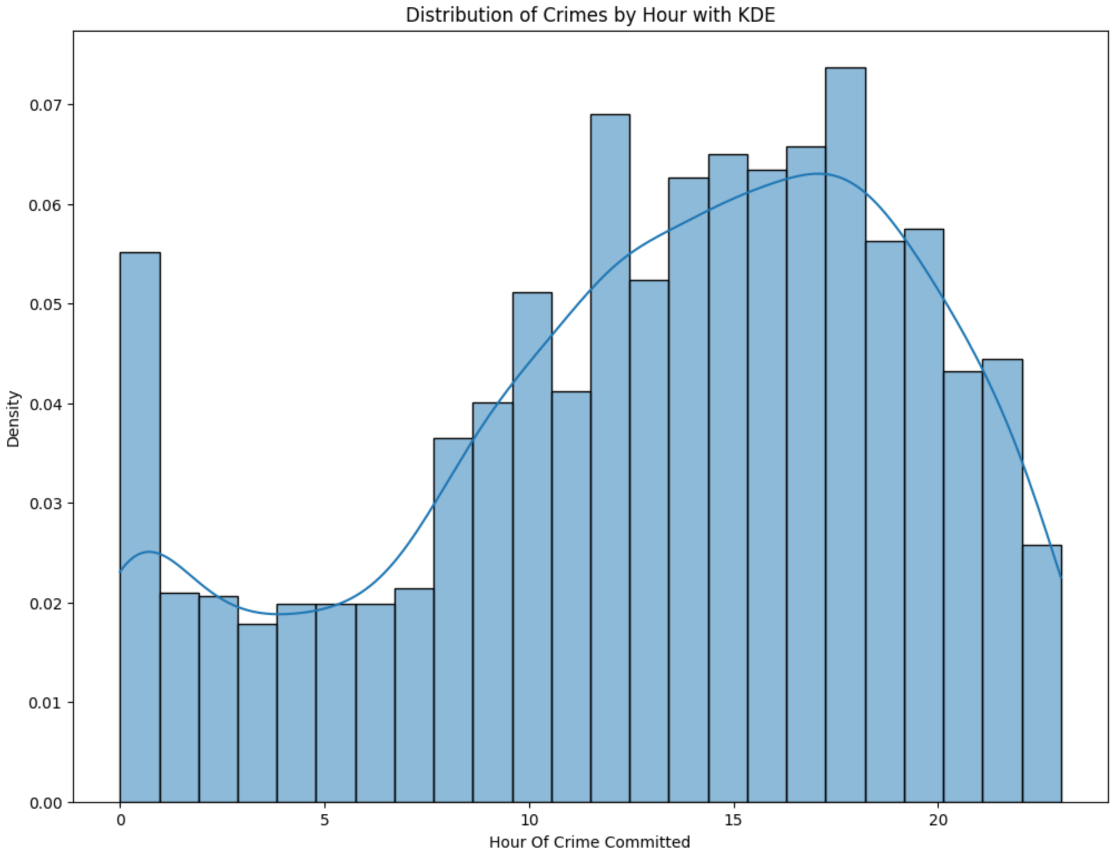
0.1 Question 2c

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 [19]: calls_hour = calls["Hour"].value_counts()
         sns.histplot(calls['Hour'], kde=True, bins=24, stat="density")
         plt.ylabel("Density")
         plt.xlabel("Hour Of Crime Committed")
         plt.title("Distribution of Crimes by Hour with KDE")
         # 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.

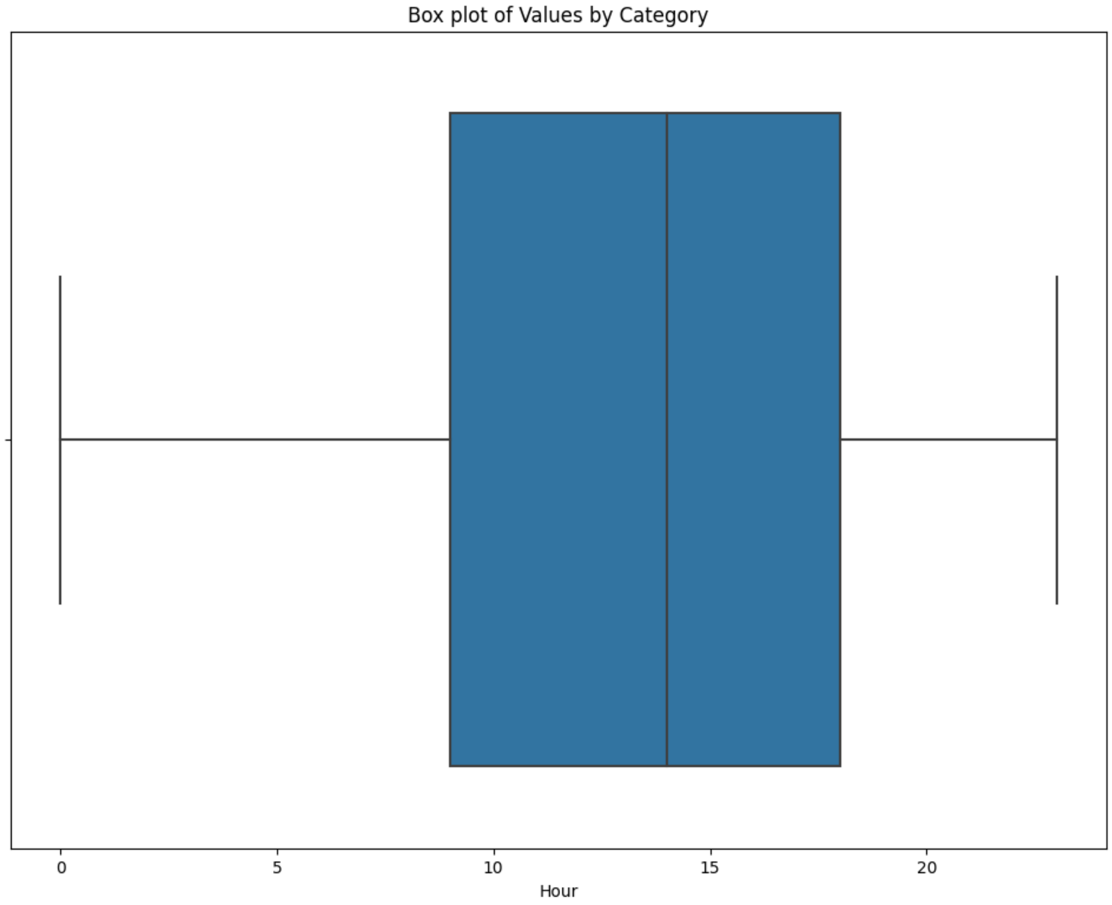
ii). To better understand the time of day a report occurs we could **stratify the analysis by DayType (i.e. by weekday vs weekends).**

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)

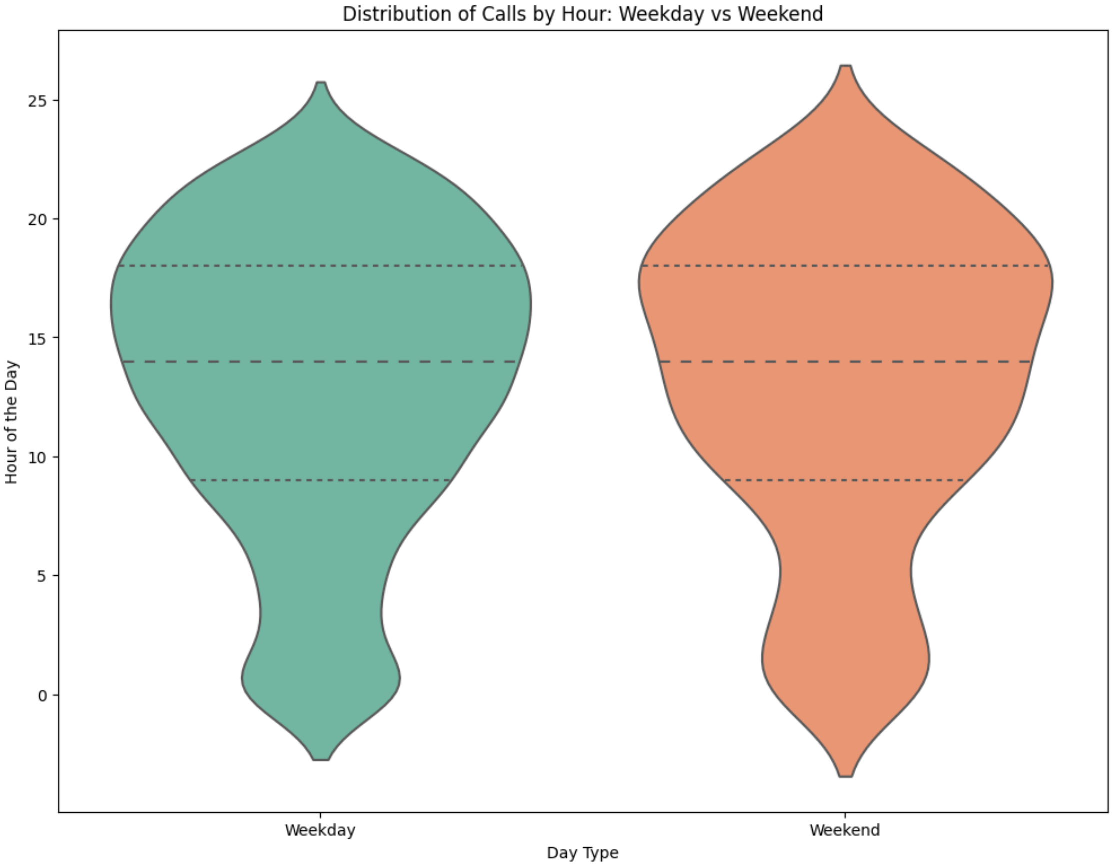
Note: For aesthetic purposes only the violin plot continues past the end of the whiskers (i.e. past 0 and 24 hours); however it is not possible to get data points outside of the whiskers for this distribution.

```
In [22]: sns.boxplot(x=calls['Hour'])
         plt.title('Box plot of Values by Category')
         plt.show()

         # Your code for boxplot above this line
```



```
In [23]: sns.violinplot(x='DayType', y='Hour', data=calls, split=True, inner="quart", palette='Set2')
         plt.xlabel("Day Type")
         plt.ylabel("Hour of the Day")
         plt.title("Distribution of Calls by Hour: Weekday vs Weekend")
         plt.show()
         # Your code for side-by-side violin plots above this line
```


Distribution of Calls by Hour: Weekday vs Weekend
Hour of the Day
25
20
15
10
5
0
Weekday
Weekend
Day Type

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:

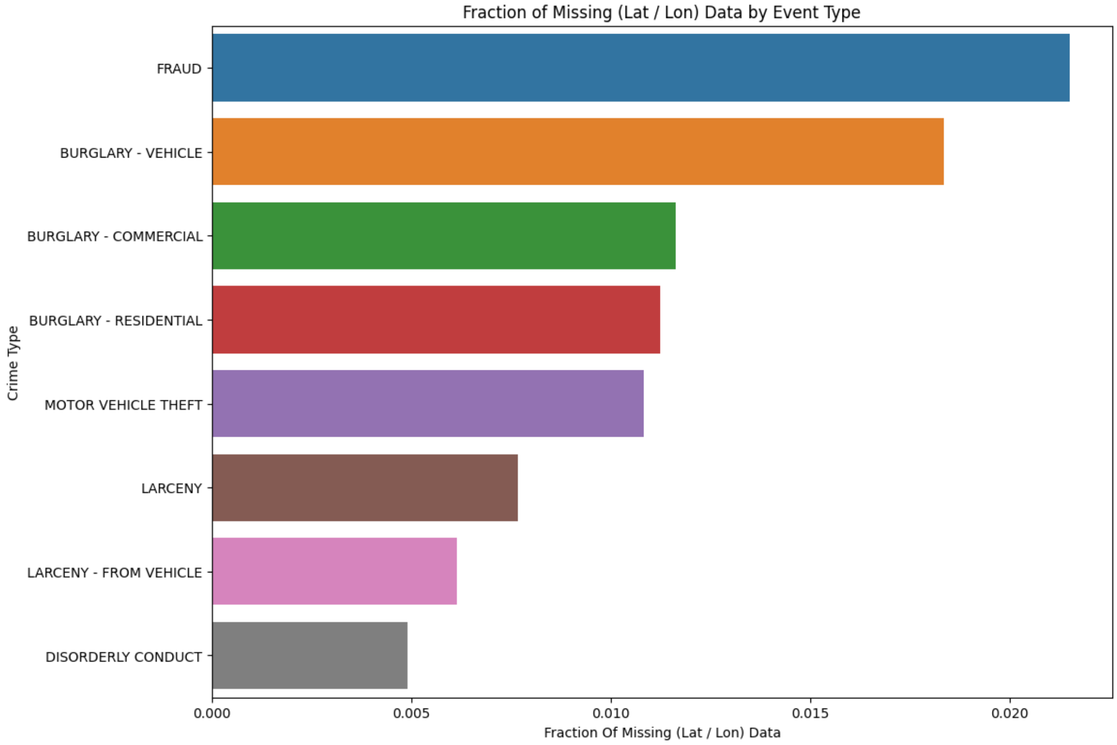
- Are there more calls in the day or at night?
- What are the most and least popular times?
- Do call patterns and/or IQR vary by weekend vs weekday?

- There tends to be more crime that is committed at night, this is kind of expected, of course we need to take into account the time of the year for if for instance, 6 PM is in the night or in the day still.
- From part 2c, we see that the most common time for a crime to be committed is around 6 PM. The least popular time for crimes to be committed is around 3 AM.
- There are slight difference between the weekend and the weekdays, but there isn't enough of a difference to really make a definitive statement here.

```
In [33]: missing_by_crime = missing_lat_lon.groupby("CVLEGEND").size()
         total_crime = calls.groupby("CVLEGEND").size()
         missing_by_crime = missing_by_crime / total_crime
         missing_by_crime = missing_by_crime[missing_by_crime.notnull()]
         missing_by_crime = missing_by_crime.sort_values(ascending=False)
         # Your code above this line
         missing_by_crime
```

```
Out[33]: 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
         dtype: float64
```

```
In [37]: sns.barplot(x=missing_by_crime.values, y=missing_by_crime.index, orient="h")
         plt.xlabel("Fraction Of Missing (Lat / Lon) Data")
         plt.ylabel("Crime Type")
         plt.title("Fraction of Missing (Lat / Lon) Data by Event Type")
         plt.show()
         # Your code to create the barplot above this line
```



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

The most common crime type where there are invalid GPS coordinates is Fraud. The other types of crime where there is invalid GPS coordiantes tend to be crimes where it would be hard to pinpoint exact locations of the crime. For instance, one may not be able to get specific GPS coordinates of a disorderly conduct crime. Let's examine which of these options would make the most sense:

- Option 1: This doesn't make a lot of sense, because we could have skewed crime statistics if we just completely removed the entries with missing / invalid data.
- Option 2: This is maybe a good idea, but it could still potentially skew the data from representing what is actually happening in regards to crimes committed.
- Option 3: This makes the most sense to me, represent the crimes where an exact location could not be found as a subset of the crimes where exact valid locations could be found. In this case, we keep the data for all crimes, we are just putting them in different 'buckets'.

So from the above, I believe that Option 3 is most likely the best option out of all three.

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.

- This is probably the easiest way for someone to pinpoint the location of where a crime was being committed. For instance, when police officers arrive on the seen, they probably just look at street signs and report the crime being committed at the corner of “x and y”. It would be kind of awkward to say that this crime was committed 32.6 ft south from some street, so they just go with the intersection as the location of the crime.
- It’s possible that the crimes that would be committed on campus would not be handled by the city of Berkely, instead, they would be handled by the University Police. This heat map does not imply that the campus is the safest place in the city, but rather just a location where the city police may be out of their jurisdiction.