### 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 ... | ... | ... |

i) Each row represents each case that had been called into to Berkeley Police Department.

| Column Name | Category | Explanation/Reasoning |
|---|---|---|
| **CASENO** | C | This is because the data can be catgorized through crime type, but not ranked. |
| **OFFENSE** | C | This is because, crime is random, and cannot be ordinal, but nominal |
| **EVENTDT** | B | This is Discrete as it counts to the day to the hour |
| **EVENTTM** | A | This is because time is continous, and doesn't stop. Furthermore it is consistent in increments. |
| **CVLEGEND** | D | As this shows what kind of crime they committed, we can rank them when categorized for the type of crime. |

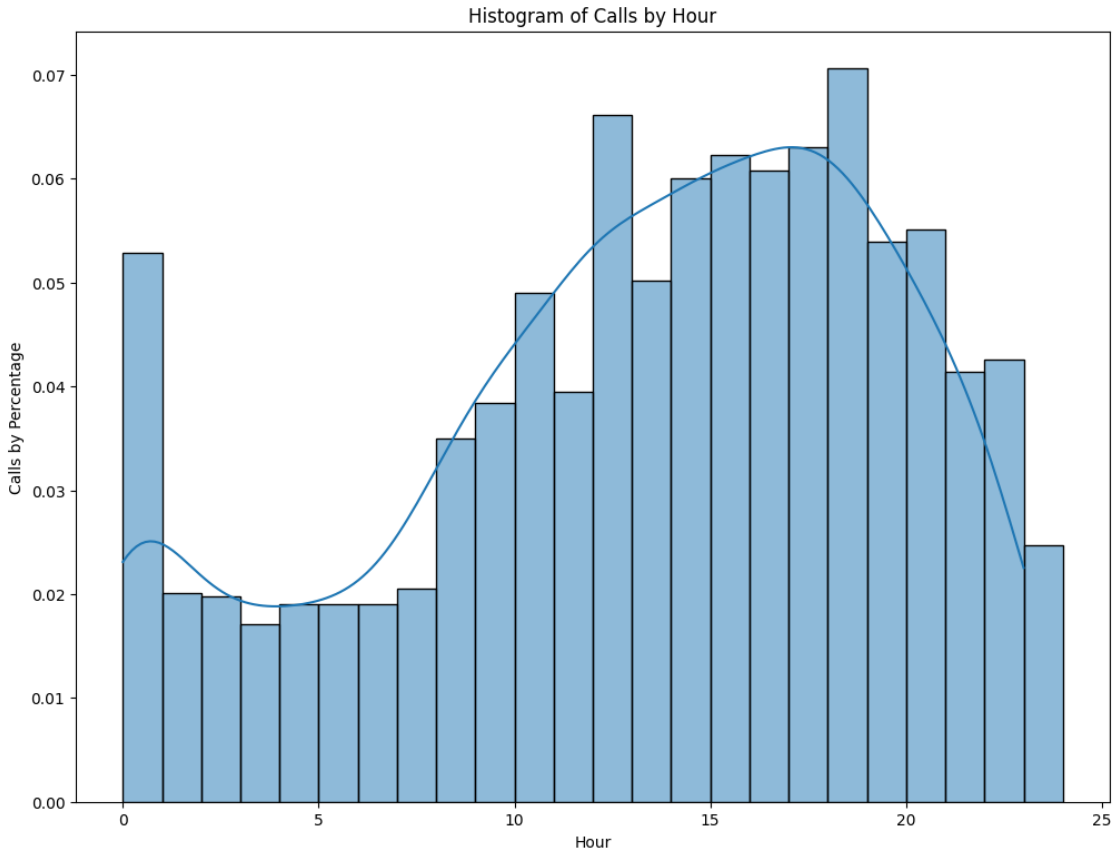
| Column Name | Category | Explanation/Reasoning |
|---|---|---|
| **CVDOW** | B | As this is countable, to see the level of crime committed we can set this as quantitative discrete |
| **InDbDate** | B | This has days of the week, and then to the hour. |
| **Block_Location** | C | The location of the crime occured is random |
| **BLKADDA** | C | Each crime occured is random |
| **City** | C | No order, but within the City of Berkeley, thus not ranked |
| **State** | C | No order, but is within City of Berkely, which is within California. Thus it is nominal |

## 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]: ax_3d = sns.histplot(data = calls, x = "Hour", stat = 'density', binwidth = 0.999999,
                              kde=True)
         ...
         # Your code above this line
         plt.xlabel("Hour")
         plt.ylabel("Calls by Percentage")
         plt.title('Histogram of Calls by Hour')
         # 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.

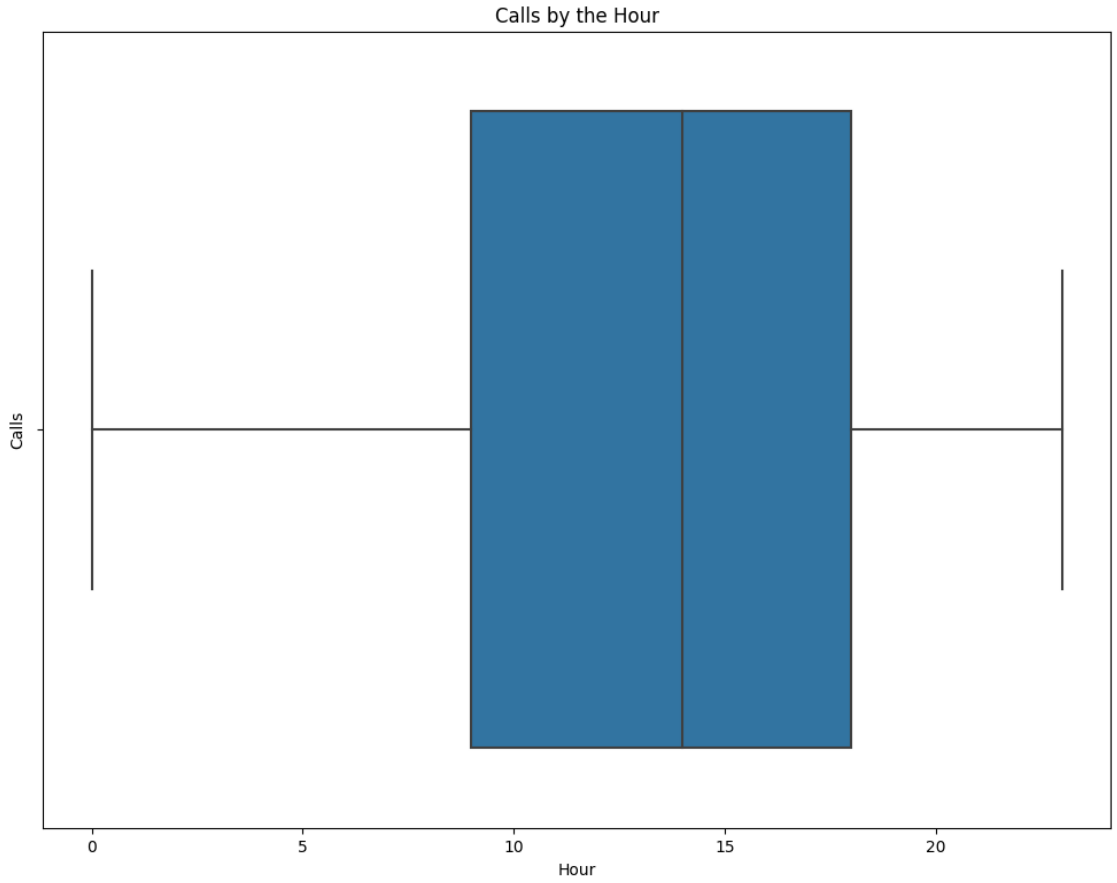
ii). To better understand the time of day a report occurs we could **stratify the analysis by DayType (i.e. by weekday vs weekends).**

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.

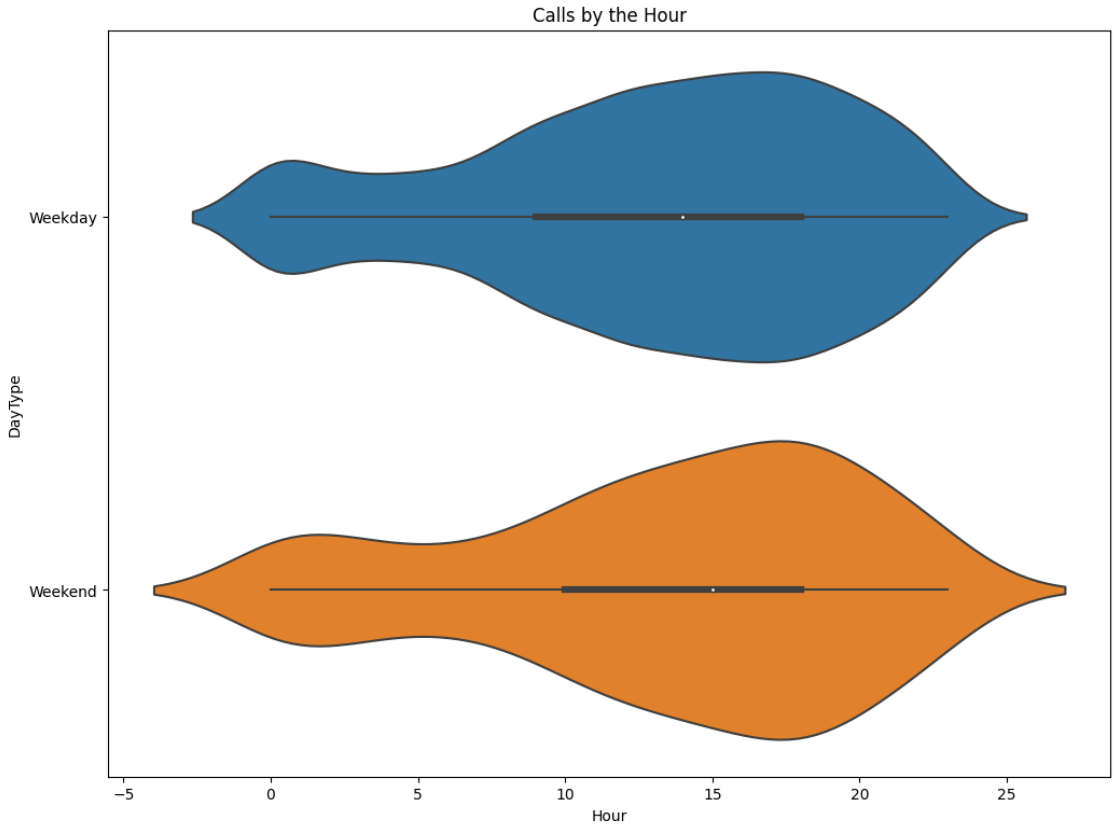
```
In [23]: ax_3d = sns.boxplot(data = calls, x = "Hour")
         plt.xlabel("Hour")
         plt.ylabel("Calls")
         plt.title("Calls by the Hour")
         # Your code for boxplot above this line
```

```
Out[23]: Text(0.5, 1.0, 'Calls by the Hour')
```




```
In [24]: ax_3d = sns.violinplot(data = calls, x = "Hour", y = "DayType")
         plt.xlabel("Hour")
         plt.ylabel("DayType")
         plt.title("Calls by the Hour")
         # Your code for boxplot above this line# Your code for side-by-side violin plots above this li
```

```
Out[24]: Text(0.5, 1.0, 'Calls by the Hour')
```


Calls by the Hour
Weekday
Weekend
DayType
−5
0
5
10
15
20
25
Hour

## 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:

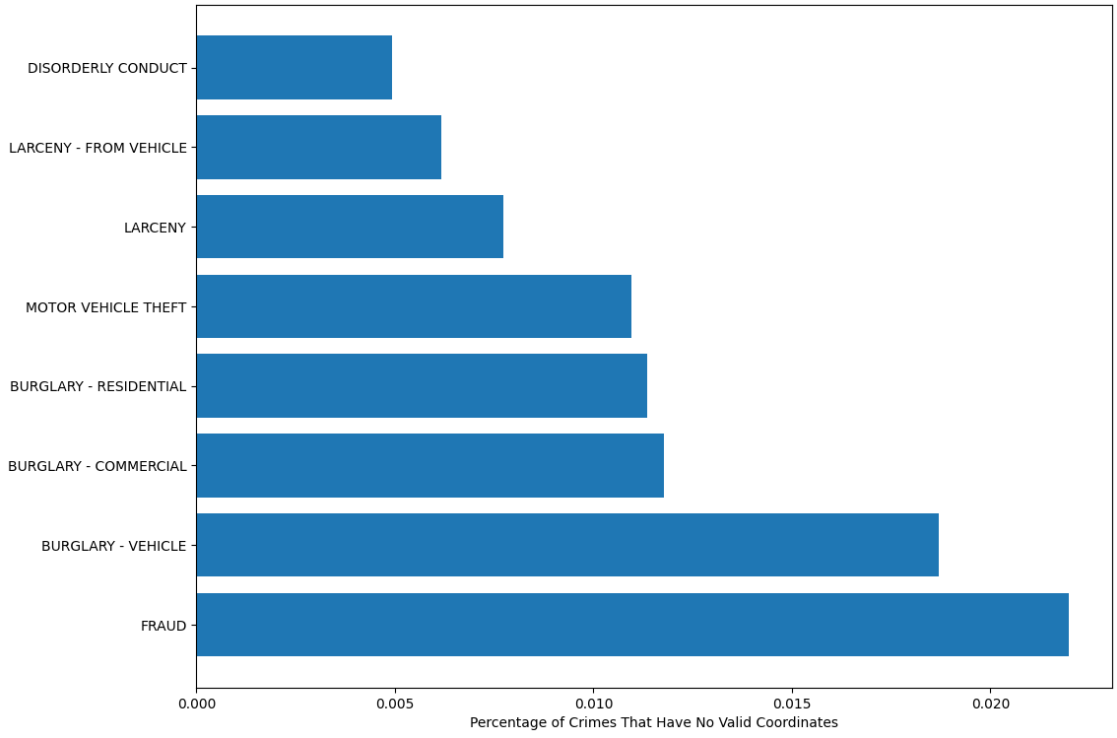
- Are there more calls in the day or at night?
- What are the most and least popular times?
- Do call patterns and/or IQR vary by weekend vs weekday?

There are more calls in the afternoon around 3 to 8. The most times in both weekdays and weekends are around 3:00pm to 8:00 pm. The patterns seem the same, except during the weekend, there is a slight decrease in calls around 7am to 9am, creating a narrower violin graph.

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

In [36]:
```
plt.barh(missing_by_crime.index, missing_by_crime)
plt.xlabel("Percentage of Crimes That Have No Valid Coordinates")
# Your code to create the barplot above this line
```

Out[36]:
```
Text(0.5, 0, 'Percentage of Crimes That Have No Valid Coordinates')
```



### 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

There are no patterns among the entries in the missing lat,long data. I believe that we should "Drop rows with missing data" (option1). This is because that Nan values aren't useful in this case, we will not go with option2. Finally imputing data will substitue missing data with a different value such as median/mean, which may cause errors as they may or may not be miscalculated.

### 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.

Some drawbacks could be that most calls occur in the streets and intersections. This may pose a problem as the crime could've been occuring inside a building, or something non-street located, which is not represented in the dataset. This will not be a pinpoint precision of the crime's location which might make it challenging where the incident occured. Some factors that could explain why there were no calls around the "Observatory Hill" could be that there is another police force that covers campus and the Observatory Hill, or the cell tower doesn't cover the observatory hill. Furthermore, it could be that calling from the Observatory Hill could send a signal to another police force. Another case could be that calls from the Observatory Hill is confidential or restricted, through some contract thus isn't in the dataset.