Final Project Report

Clean Data

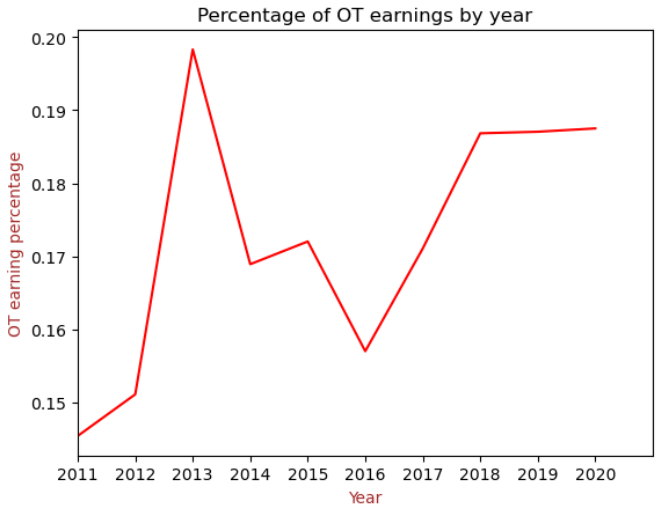
There hasn't been much necessary pre-processing of the data on our part so far, to make year-wise analysis easier we did separate the employee earnings dataset and special events dataset into multiple datasets containing only records from each of the years we were looking at. The datasets had no missing or null values which needed to be cleaned. For the analysis we have done until now we did not require non-numerical data to be converted to numerical quantities.

Explore Data

Our project is largely centered around data exploration. We are trying to uncover any trends in the data that illustrate police misuse of the special events overtime system.

Overtime Earning as a Percentage of Total Earnings

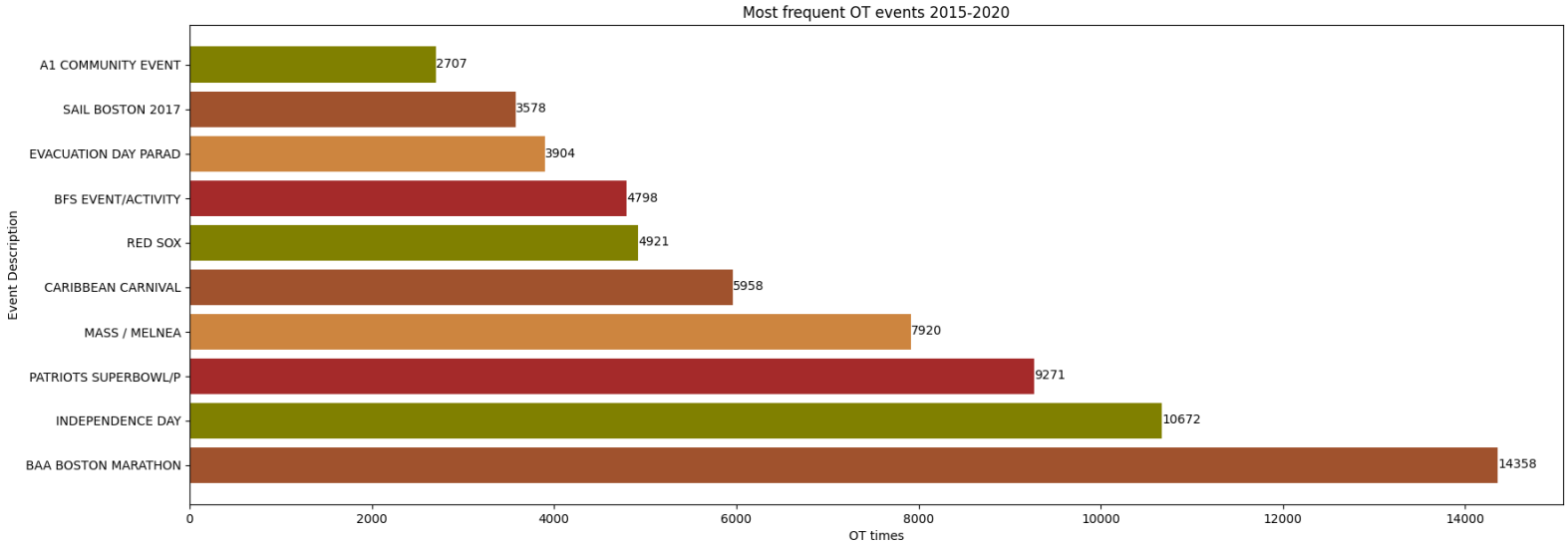
We began by looking at the employee earnings data to see what percentage of total annual earnings was overtime earnings, to see if there were any outlier years we should look at more closely.



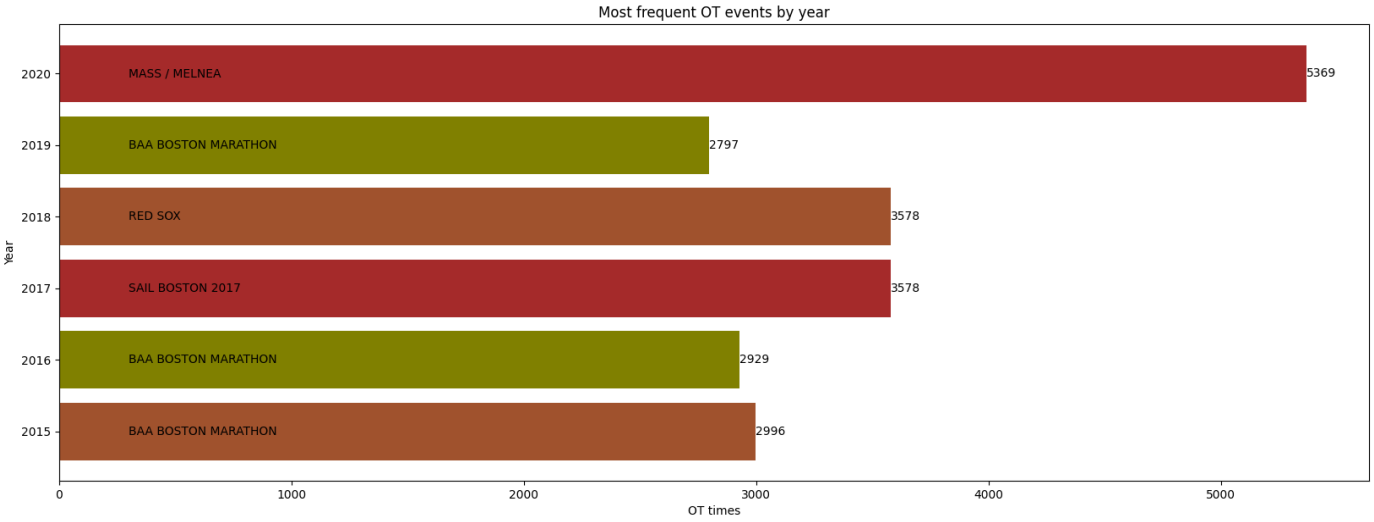
Every year the percentage of total earnings from overtime earnings was between 15 and 20 percent, this variance is not very large and didn't indicate any year as being an outlier.

Most Frequent Overtime Events

To understand the data better we wanted to see which overtime events were occurring most often in the dataset over the entire timeline of records in the dataset.



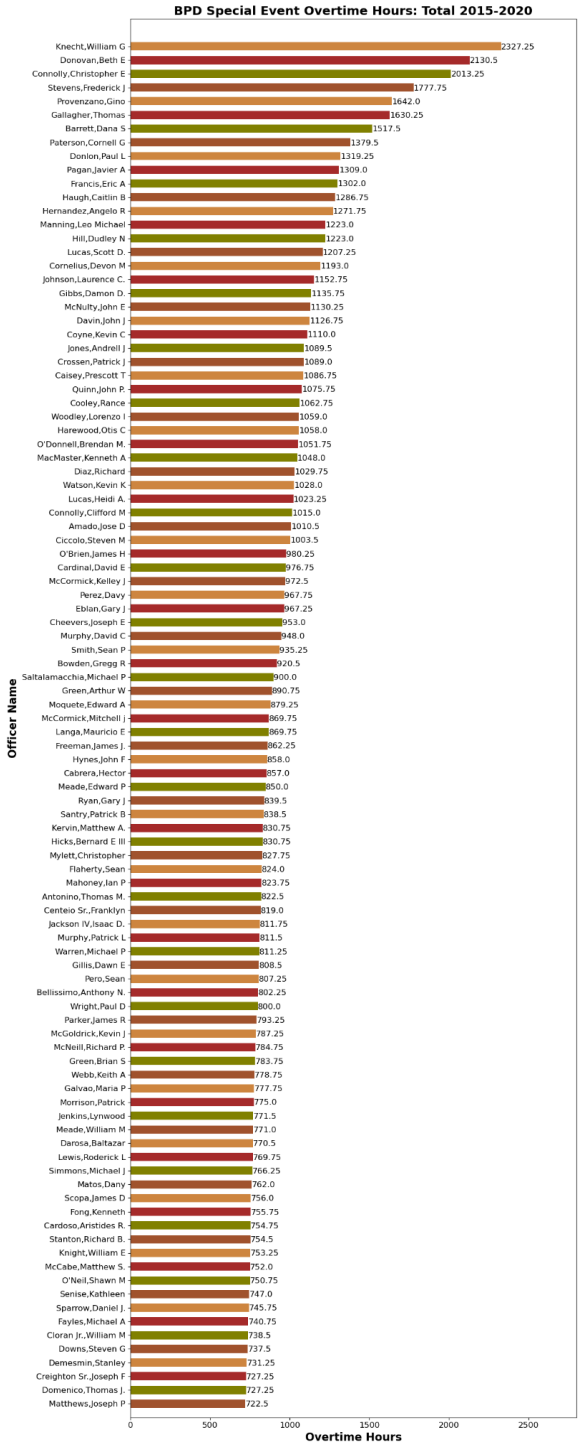
We then also looked at which event was the most frequently occurring each year.



Mass/Melnea was an event that raised a lot of questions. Why was it so frequently occuring in 2020? Was it because of the reduction of in-person events the police could not make overtime in?

Officers with Most Logged Overtime Hours

We then looked at which officers were logging the most overtime. So we could get an indication of whether there were specific officers we should look into more.

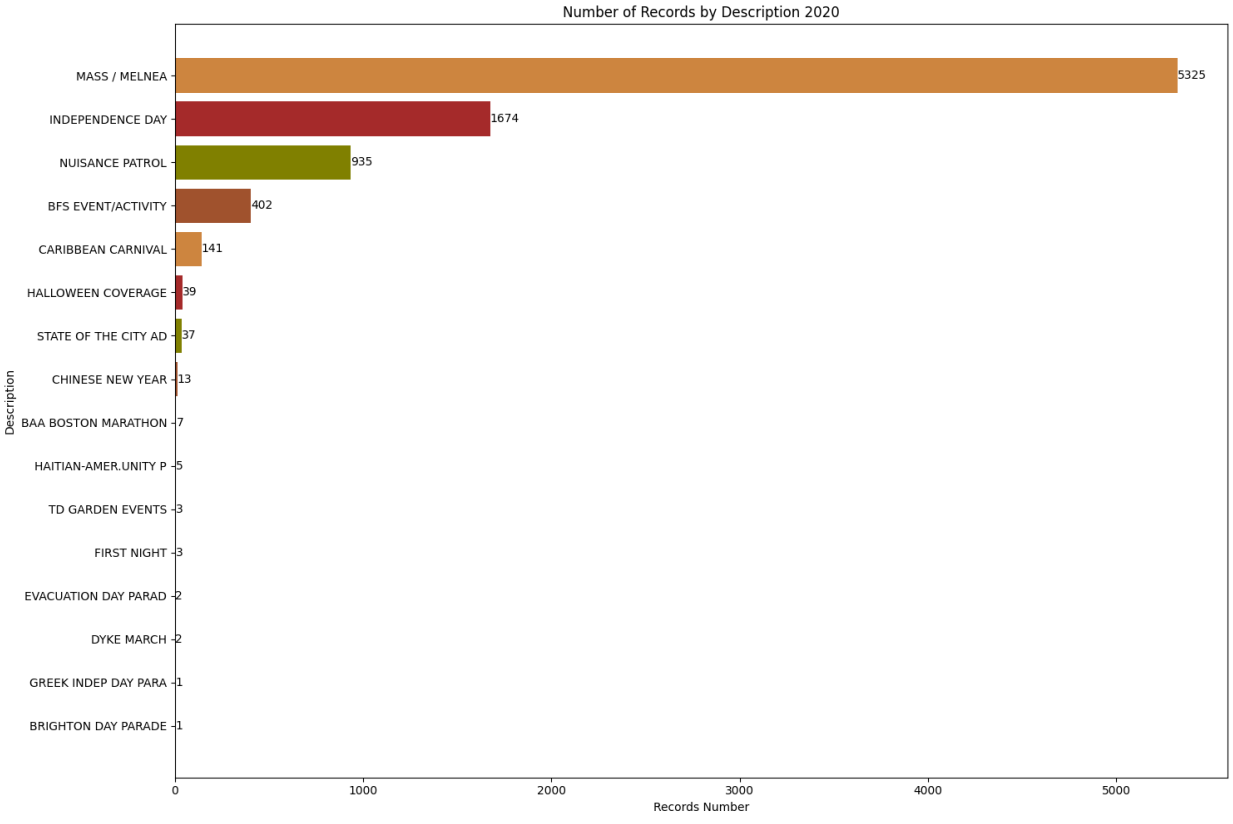


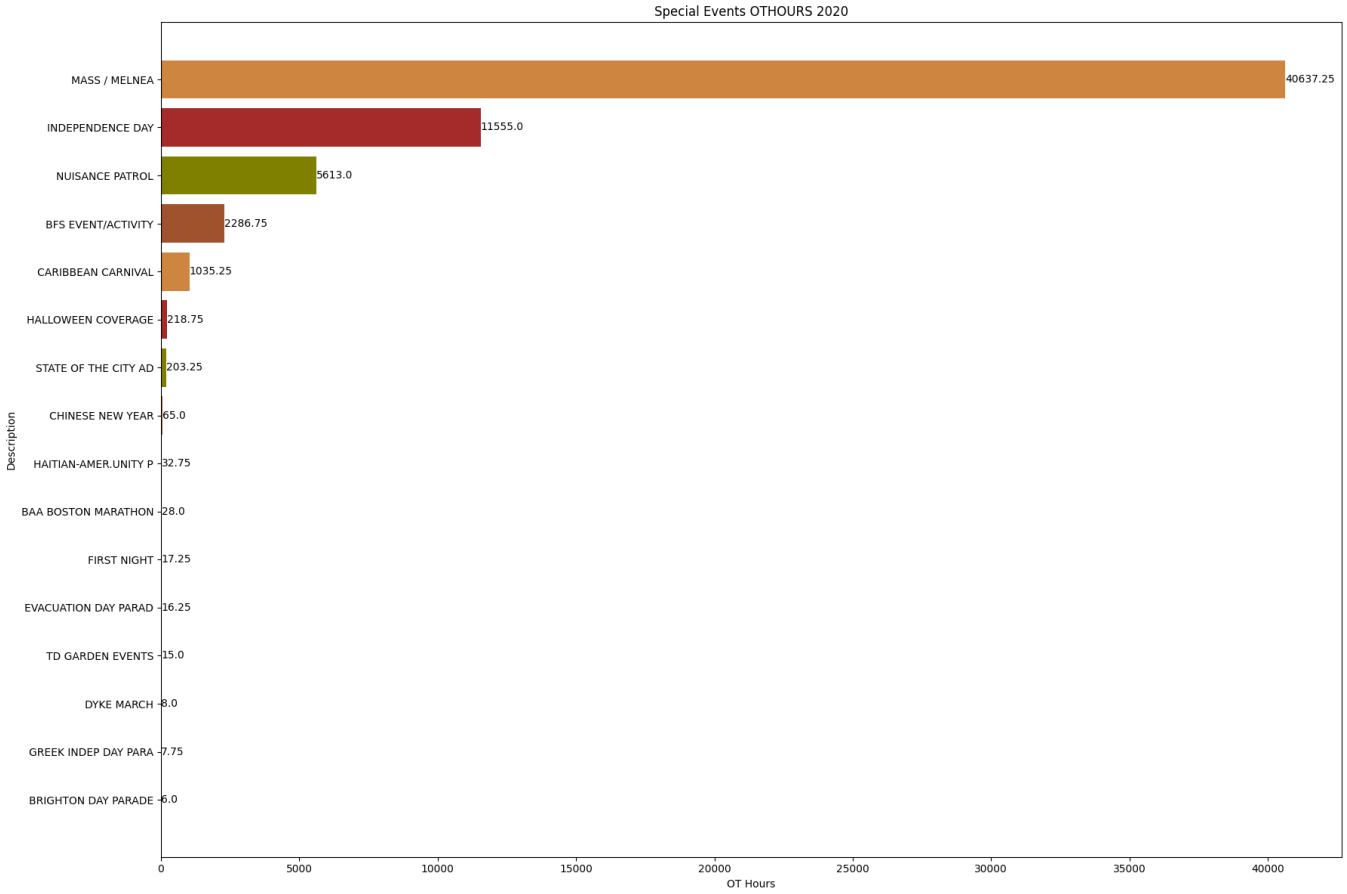
Certain officers had very inflated overtime hours logged, especially the top seven. We took note of these names as we explored the data further.

Categorical Clustering

After this initial exploration we decided that it would be useful to split the dataset categorically, first we decided to split the data based on how the events were charged in Column K of the special events dataset. This, however, created some confusion because it may be that every event in the special events overtime dataset isn't actually being paid through the special events overtime budget. We then clustered data together by their event descriptions filtering out all events not charged as special events in column K. After the description based categorical clustering we realized there is another issue. These events that are charged as special events only cover 2020. Before that they were logged differently but we don't know which ones to include or not include. But we were still able to raise interesting questions using the data focused solely on 2020.

We raised the question earlier, under the most frequent overtime events exploration, why did Mass/Melnea events spike in 2020. Here we were able to visualize that spike more.





Mass/Melnea was logged 3 times more records than any other event in 2020 and logged 4 times more hours than any other event in 2020. The Mass/Melnea event refers to policing of the area known as the “Methadone Mile” in Boston. This area has a high concentration of drug activity but this is not a phenomenon that suddenly emerged in 2020, it has been happening in years past, which begs the question why did it need an extremely high amount of extra policing in 2020?

Strategic Questions

---

1. How are events that are charged to the special events overtime budget labeled/categorized? We thought they may only be events charged in Column K as SPECIAL EVENTS, but those events only cover special events in 2020, how were those events logged before 2020?

2. Why was there such a massive spike in MASS/MELNEA special events in 2020?

Early Insights

Limitations

---

1. The interpretation of this data is steeped in political context. The exploration has been a difficult process because there are no objective metrics of misuse or any obvious discrepancies in the data. Rather, we are looking at the data from many angles looking at the relationship between different fields, looking at how the data is distributed, and from these altering perspectives trying to understand the qualitative political meaning of our extractions. Exploring this dataset from the data science angle is looking for a needle, painted the color of hay, in a haystack. It is a slow and adventurous process.

2. The special events data has records categorized in Column K (Charged) very liberally, this is creating confusion about which events are actually being paid for with the overtime events budget. We thought maybe only events tagged as SPECIAL EVENTS in Column K are coming from the special events budget but events were only tagged with SPECIAL EVENTS in 2020. We have no clue how events being paid for from the special events overtime budget were labeled/categorized in years before 2020. If everything in this dataset is charged to the special events overtime budget then there are a lot categories that begin looking like they don't belong (e.g.

INTERNAL INVESTIGATIONS UNIT, JOINT TERRORISM TASK FORCE, EXPLOSIVE ORDNANCE UNIT...).