Saleha Bakht
Springboard - Data Science
Capstone Two

NY Daily Inmates in Custody

September 22, 2018

OVERVIEW

The dataset analyzed was provided by the City of New York on their Daily Inmates in Custody. While the City has not posed any questions with the release of its data, our exploration will look into exploring the patterns presented in the data to see if there are any unexpected findings. The goal is also to use machine learning processes to predict missing values and verify the findings in the dataframe.

GOALS

1. Does race or gender have overly large influence on custody level?
2. Is there a pattern in infractions amongst certain groups of inmates?
3. Attempt a predictive model to see if the custody level of the inmate can be predicted.

CLEANING THE DATASET

To start us off, we replaced the original column names 'BRADH' and 'SRG_FLG' with 'MENTAL_OBSERVATION' and 'GANG_AFFILIATION' respectively. The original names we did not find intuitive. The new names are longer but easier to remember for the explorer.

Next we dropped the columns 'DISCHARGED_DT' and 'SEALED'. Since all of the values in the former were null values and all the values in the latter were the same (in that none of the inmate's files were sealed) the columns would have no influence on the outcomes of our exploration. They were removed from the dataframe as considerable features.

For the preliminary purposes of EDA all rows that contained a single null value in the columns of 'GENDER', 'RACE', and 'AGE' were dropped. The rows that had null values in the 'TOP_CHARGE' column were imputed with '-1' so that the TOP_CHARGE feature could be used in a future predictive model and the null values would be treated as a separate category. The total percentage of rows dropped is 0.2% leaving us with more than 99% of the original data.

The last thing we did to clean the data was to break up the time stamps in 'ADMITTED_DT' into columns for year, month, day, and hour. The 'ADMITTED_DT' column was then dropped.

EXPLORATORY DATA ANALYSIS

A preliminary exploration into gender distinctions first, where we found that:

- There were 14.45 times as many male inmates as there were women.
- Low numbers of gang involvement.
- Proportionally higher numbers of women being under mental observation than men.
- More men had unrecorded values in their top charge but the proportion disparity is getting smaller and smaller by the year.

We found almost 100 records belong to inmate 16-17 years of age. Most of them are given maximum security custody levels. A higher relative proportion of them among their own set have gang affiliations than the adults. More than half of their top ten most common charges involve robbery or burglary.

In relevance to race, there was nothing noteworthy when comparing the genders in incarceration. Race had no bearing on the existence of infractions either. There are proportionally less of each race under mental observation except for the White and Islander populations. Those races have more people under mental observation than not.

Using the NY website http://ypdcrime.com/ the charges were matched with their numbers.

The top ten charges of the people aged 18 and over are as follows:

1. 125.25 is 'Murder in the second degree'.
2. 160.15 is 'Robbery in the first degree'.
3. 110-125.25, if treated as a range, is 'Attempt to commit a crime' to 'Murder in the first degree'. No singular offense was found for this charge.
4. 120.05 is 'Assault in the second degree'.
5. 265.03 is 'Criminal possession of a weapon in the second degree'.
6. 140.25 is 'Burglary in the second degree'.
7. 220.39 is 'Criminal sale of a controlled substance in the third degree'.
8. 220.16 is 'Criminal possession of a controlled substance in the third degree'.
9. 160.10 is 'Robbery in the second degree'.
10. 140.20 is 'Burglary in the third degree'.

We can see that four of ten of these most frequent charges are of or relating to theft or the intention to commit theft. Six of the charges require face-to-face interaction with another person, presumably a victim.

The top ten charges of the people under the age of 18 are as follows:

1. 160.15 is 'Robbery in the first degree'.
2. 160.10 is 'Robbery in the second degree'.

3. 110-125.25, if treated as a range, is 'Attempt to commit a crime' to 'Murder in the first degree'. No singular offense was found for this charge.
4. 265.03 is 'Criminal possession of a weapon in the second degree'.
5. 125.25 is 'Murder in the second degree'.
6. 105.15 is 'Conspiracy in the second degree'.
7. 140.25 is 'Burglary in the second degree'.
8. 120.00 is 'Assault in the third degree'.
9. 110-160.15, if treated as a range, is 'Attempt to commit a crime' to 'Robbery in the first degree'. No singular offense was found for this charge.
10. 110-160.10, if treated as a range, is 'Attempt to commit a crime' to 'Robbery in the second degree'. No singular offense was found for this charge.

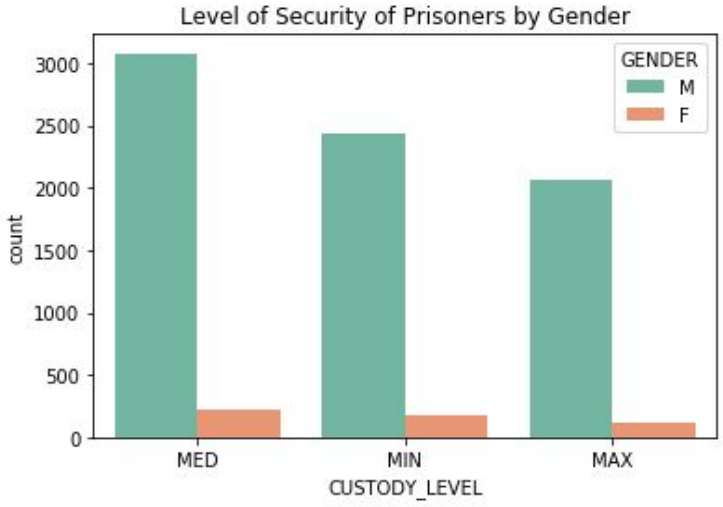
Five of these top ten most frequent charges involve theft or the intent to commit theft. Four of the top ten required face-to-face interaction with, again presumably, a victim.

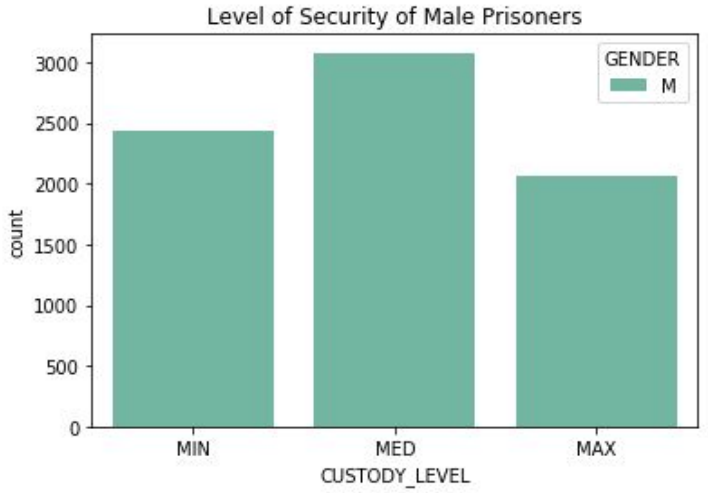
Two of the most frequent charges for 16-17 year olds were of the murder variety. Same with the 18+ year olds. However, we can see in the graphs that a majority of the adults are under a medium CUSTODY_LEVEL whereas a majority of the 16-17 year olds are under a maximum CUSTODY_LEVEL.

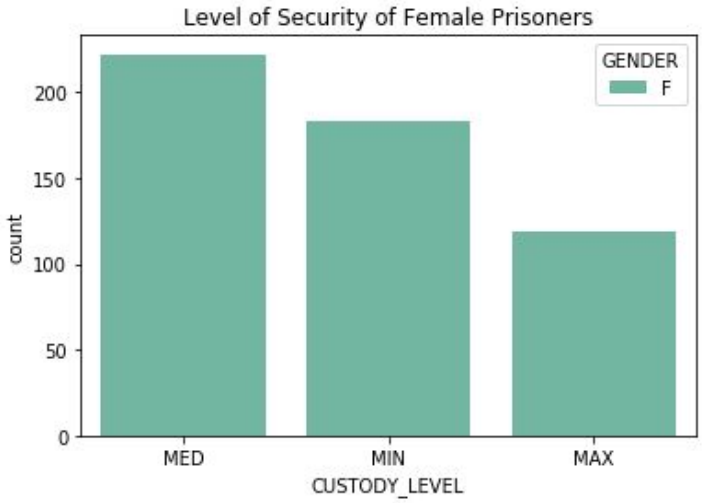
ANSWERING THE GOAL QUESTIONS

Does race or gender have overly large influence on custody level?

To determine the answer to this question, let us look at some graphs. Overall we found that there were 14.45 times as many men as there were women incarcerated.

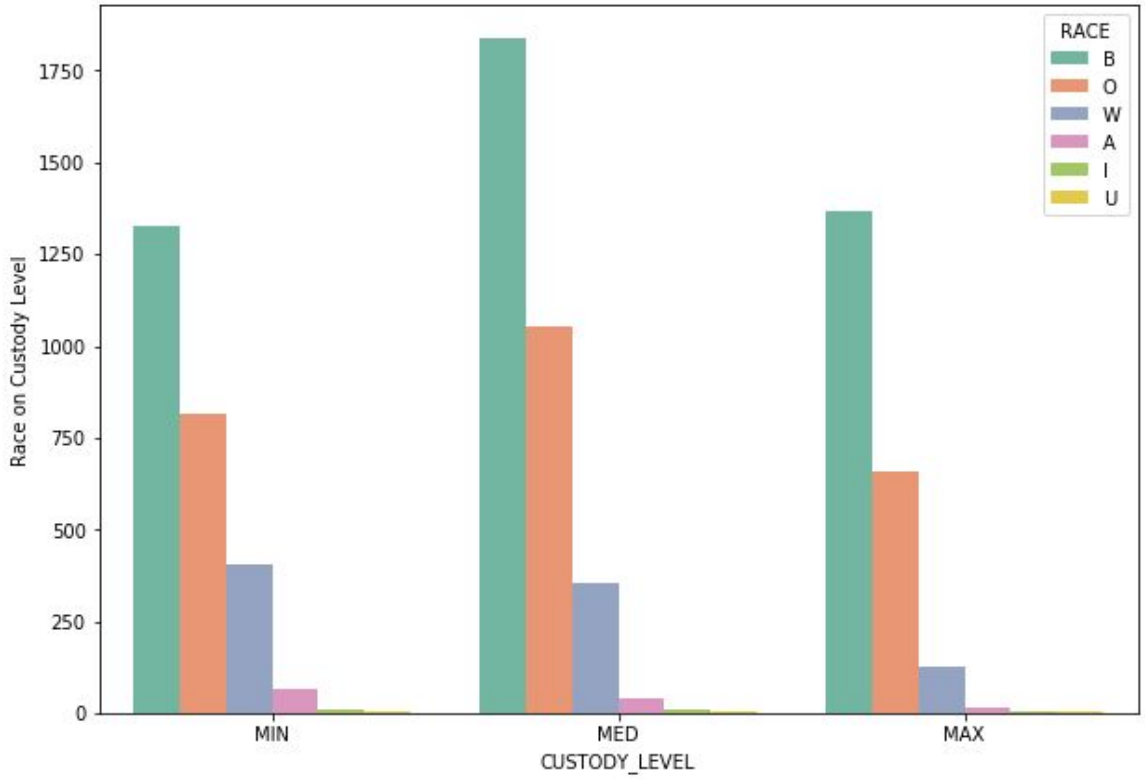






These graphs shows us two things. One, the difference in the number of male and female inmates. Two, The highest proportion of men and women are at medium security custody level and that the least proportion is in maximum security. This holds true for both genders.

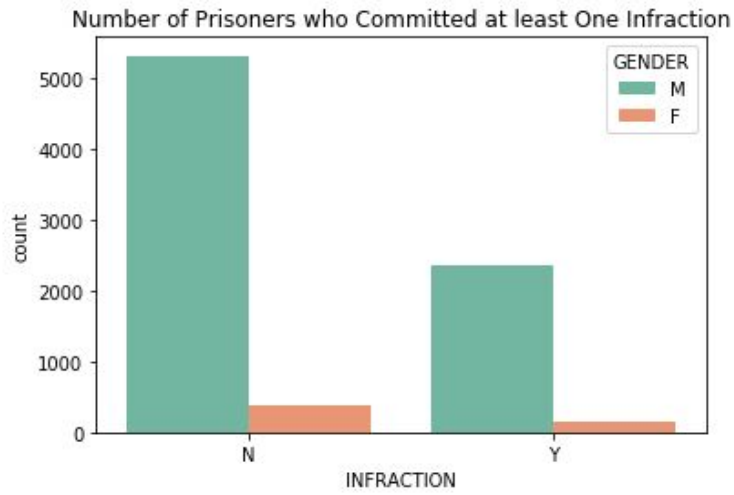
```
Text(0,0.5,'Race on Custody Level')
```

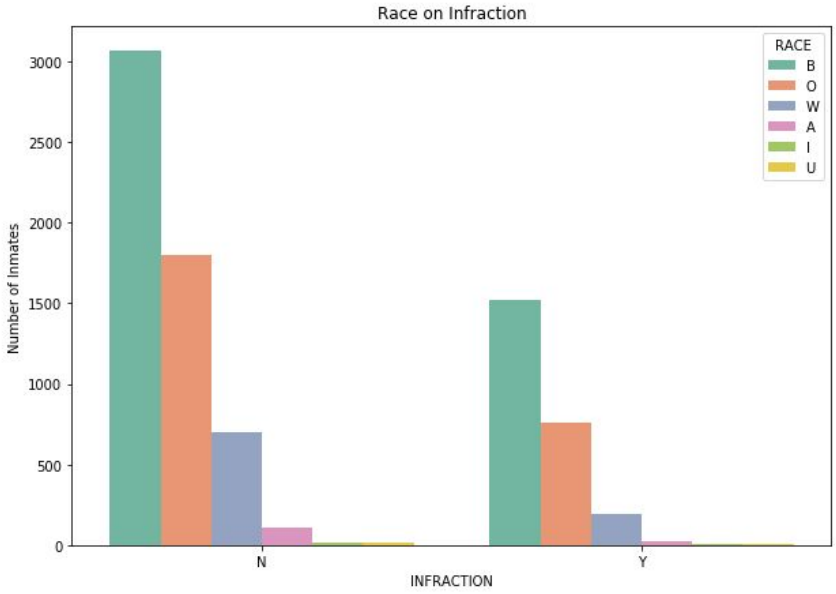


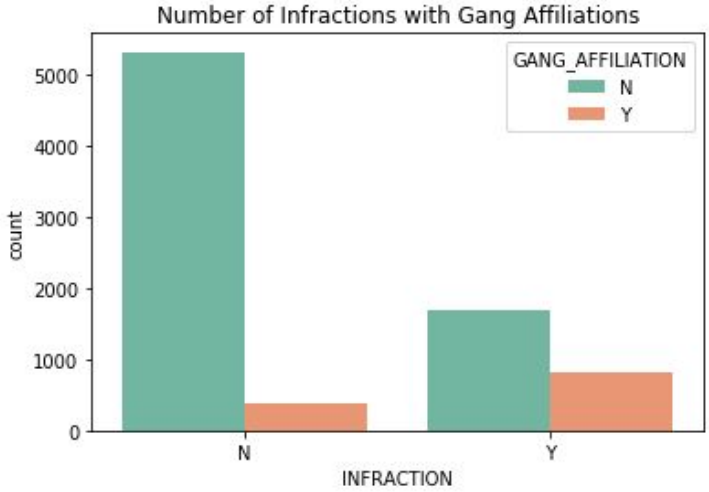
Most of the races follow the same trend as identified in the previous three graphs; a majority of the inmates are under a medium custody level with the least numbers of inmates being under maximum. This holds true for all of the races except for White and Asian (these races were determined using other legends in other datasets also by NYC Open Data). Race does not appear to have an 'overly large influence on custody level' but there is a difference in races and custody level in the dataset.

Is there a pattern in infractions amongst certain groups of inmates?

This question will also be resolved through graphs.


Number of Prisoners who Committed at least One Infraction
GENDER
M
F
count
5000
4000
3000
2000
1000
0
N
Y
INFRACTION


Race on Infraction
RACE
B
O
W
A
I
U
Number of Inmates
3000
2500
2000
1500
1000
500
0
N
Y
INFRACTION


Number of Infractions with Gang Affiliations
GANG_AFFILIATION
N
Y
count
5000
4000
3000
2000
1000
0
N
Y
INFRACTION

There are no patterns between race and gender and the number of infractions written. Of the people who do commit infractions, a majority of them are not in gangs. Although more gang affiliated inmates commit infractions than don't.

MACHINE LEARNING

Attempt a predictive model to see of the custody level of the inmate can be predicted.

This goal question will be answered in the machine learning section of this report since that is what was required to meet the goal. For this we dropped the AGE feature from the dataframe. This is because if they are included in this dataset, the inmate was tried as an adult and not in a juvenile court. Therefore the 16-17 year olds would have been given the same level of consideration as the adults.

We also dropped the MONTH, DAY, and HOUR features from the dataframe as we do not believe that any pattern will emerge distinctly on a month-to-month or smaller basis that has a bearing on the CUSTODY_LEVEL. We believed that should a pattern occur, it would likely be found and visible in the YEAR feature. Any changes in law or the policies of the presiding judge would be more evident by the YEAR feature.

We also dropped INMATE_ID as that would not help us determine any distinctions in CUSTODY_LABEL.

One round of RandomForestClassifier() was used to attempt to predict the CUSTODY_LEVEL value of the inmates. This functions had to be transformed with a one-hot encoder since all of our features were categorical and labelled with more than just 1's and 0's.

RACE_W	GENDER_F	GENDER_M	...	INFRACTION_N	INFRACTION_Y	YEAR_1991
0	0	1	...	1	0	0
0	0	1	...	1	0	0
0	0	1	...	1	0	0
0	0	1	...	0	1	0
0	0	1	...	1	0	0

A classification report and a confusion matrix were formed to help evaluate the performance of the classifier.

```
print(classification_report(y_train[:1642], y_test))
```

	precision	recall	f1-score	support
0	0.27	0.26	0.26	449
1	0.39	0.40	0.39	667
2	0.29	0.29	0.29	506
avg / total	0.32	0.33	0.32	1622

```
confusion_matrix(trained[:1642], predictions)
```

```
array([[121, 199, 149],
       [187, 266, 214],
       [134, 225, 147]], dtype=int64)
```

The precision of 0.32, recall of 0.33, and the f1-score of 0.32 show that the classifier did not perform well. Higher numbers would have been better predicting the custody level of the inmate based on the categorical variables provided in the dataframel. Also note that the confusion matrix is not diagonal and, therefore, we can see that the classifier did not perform its absolute best in its predictions.

An alternative to this classifier would be to use a random search for parameter tuning instead of the grid search that was used. It is not expected that this would give highly different results; its run time would be lower and scikit-learn details that "performance is slightly worse for the randomized search"[1].

While the outcome of the classifier appears to put the same proportions of inmates in respective custody levels, it does not appear as if the classifier made excellent use of the dataframe's categorical variables to arrive at its conclusions.

A variable importance chart was made for the variables involved in the regressor.

	importance	variable
312	TOP_CHARGE_125.25	0.055100
313	INFRACTION_N	0.085351
314	GANG_AFFILIATION_N	0.121676
315	GANG_AFFILIATION_Y	0.186137
316	INFRACTION_Y	0.212501

As we can see the features that had the most influence on the custody level of the inmate was whether or not they were gang affiliated and had committed any infractions. The top charge value 125.25 is 'Murder in the first degree'. Even that was taken into less of consideration by the model than gang affiliation and infractive behaviour.

[1] http://scikit-learn.org/stable/auto_examples/model_selection/plot_randomized_search.html

This was the extent of the exploration into the NY Daily Inmates in Custody dataset as found on Kaggle and NYC Open Data.

CONCLUSION

The following are some suggestions based off of the exploration.

1- Future datasets should have also have the discharge dates in them so as to create a more robust analysis. An analysis in the length of time spent incarcerated against, race, gender, age, custody level, infraction, top charge, mental observation, and gang affiliation would be more enlightening into patterns in NY daily inmates in custody.

2- The number of missing top charges in the dataset is going down by the year. Continuing that, the missing top charges should not be a problem by the end of 2019.

3- There are such a disproportionate number of men towards women in the dataset. 14.45 times as many men than there are women incarcerated. NY should look into creating and implementing social programs that aim to reduce the amount of criminal behaviour found 14x more often in men than women.