NOTProject1

Jared Brotamonte

1/26/2024

Introduction

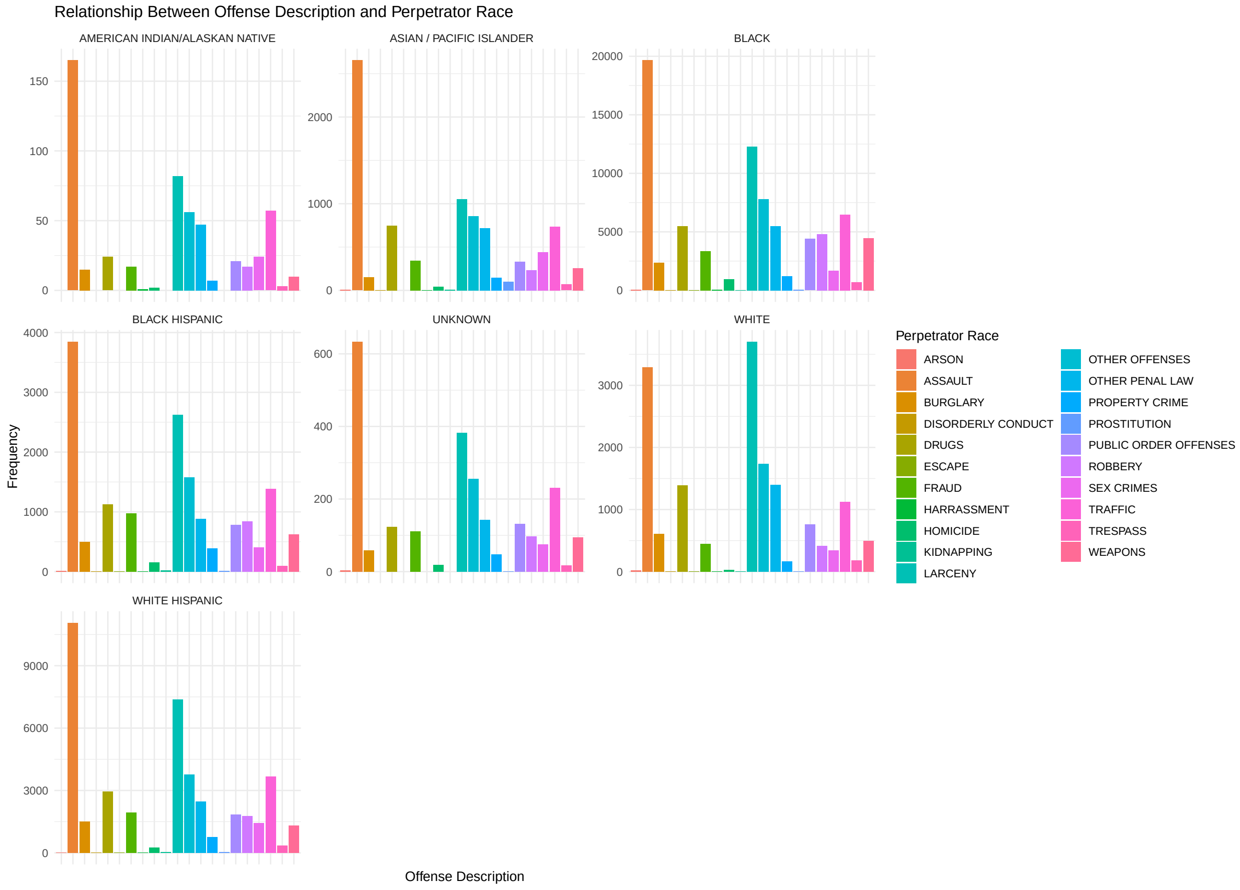
For this project, I decided to look into the topic of law enforcement and whether or not law enforecment truly targets minorities and people of color.
The difference in power when comparing law enforcement to the average citizen has always been hugely drastic. This imbalance of power has recently led to many conflicts arising between law enforcement and US citizens with the bases of these conflicts being that law enforcement has too much ability to abuse their power without enough consequences. In particular, I wanted to focus on claim that law enforcement abuses their power towards minorities or more specifically people of color

Data Cleaning

In order to clean the data and make it more usable, I had to discern what variables where actually neccessary for analyzing the data. Many variables held similar information as other variables like how OFNS_DESC and PD_DESC both describe the type of offense that was commited, but one is just more specific than the other. There are also variables that are just hard to interpret, such as the variable LAW_CODE which is just a number which corresponds to the law broken, but because I was given just the law code number, it makes it hard to use this number and generate some type of analysis. Much of the data was cleaned out due to the reasons prior which in turn left me with 10 variables left over.
To clean the left over variables I had to deal with a couple different problems. Firstly I made sure to properly format the variable ARREST_DATE as dates. I then made sure to only keep valid data points, thus I got rid of data points containing null values as well as data points that did not contain valid information. Then lastly I had to make sure to turn the variables that needed to be factored into factors. This also included cleaning up factors. For the variable OFNS_DESC, since there where too many factors in for this variables, I had to go through and change the factors to be more generalized. As for the other factors, I had to make sure to check that those variables didn't contain any factors that shouldn't exist.

Data Visualization 1

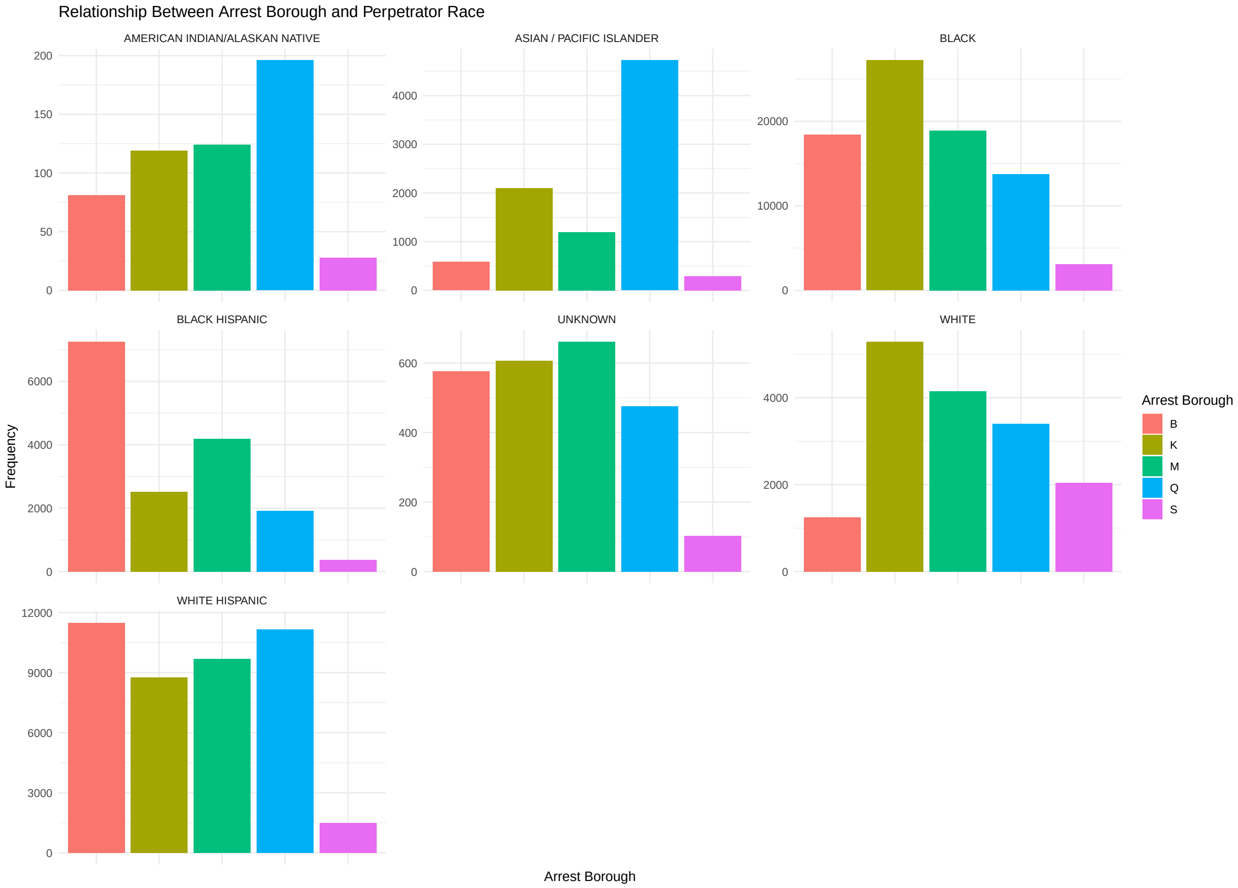
```
ggplot(NYPD_Arrest, aes(x = OFNS_DESC, fill = OFNS_DESC)) +
  geom_bar(position = position_dodge()) +
  facet_wrap(~PERP_RACE, scales = "free_y") +  # scales = "free_y" for better visualization
  labs(
    x = "Offense Description",
    y = "Frequency",
    title = "Relationship Between Offense Description and Perpetrator Race",
    fill = "Perpetrator Race"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank())  # Remove x-axis labels due to clutter
```



The visual above is a bunch of bar plots looking at the relationship between the offense committed and the race of the perpetrator. From the visual one can see that in general, the distribution of the plots lay similar except for the bar plot of the "White" race where Larceny is the highest committed offense compared to all the other races having Assault as the highest commited offense. From the visual, one can conclude that in general, there is not much of a relationship between the type of offense committed and the race of the perpetrator.

Data Visualization 2

```r
ggplot(NYPD_Arrest, aes(x = ARREST_BORO, fill = ARREST_BORO)) +
  geom_bar(position = position_dodge()) +
  facet_wrap(~PERP_RACE, scales = "free_y") +  # scales = "free_y" for better visualization
  labs(
    x = "Arrest Borough",
    y = "Frequency",
    title = "Relationship Between Arrest Borough and Perpetrator Race",
    fill = "Arrest Borough"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank())  # Remove x-axis labels due to clutter
```



Looking at the visual above, the race of the perpetrator does seem to have a relationship with the borough where the arrest occurred. However there is a lack of reasoning that can be made to explain this relationship. In general the (S)Staten Island borough conducts the least amount of arrests while the borough which conducts the most amount of arrests differ depending on race. For the races that are often classified as "Minorities", the Natives, Asians, Blacks, and Hispanics, the boroughs in which they are most often arrested in seems to be between (B)Bronx and (Q)Queens. In order to properly analyze this relationship its likely that more data would be needed. Data such as the race distribution in these boroughs and maybe even data on the general wealth of each borough could help explain why certain races are arrested more often in certain boroughs but for now this question cannot be answered.

Conclusions

If given more time I would love to look into making a cluster model to look into the specific coordinates of the arrests, and try to find a hot spots of crime and look to see if it differs depending on race. I would also love to try and gather more data to help explain why certain races get arrested more often in certain boroughs. It would also be great if I could create a multinomial logistic regression model in order to try and determine what variables are actually important in predicting what race the perpetrator was. Dealing with this data set helped me learn how to deal with data with variables that are mostly factors. Of course the biggest problem for this data set is that not only are most of my variables are factors, but also the fact that the variable that I'm trying to predict is also a factor which in turn restricts the types of models that can be used. There is also the challenge of there simply not being enough variables, and thus if I were to continue looking into this data I would also have to look into getting more data from elsewhere.

Link to GitHub Repository

```
devtools::install_github(‘JKBrotamonte/NOT-Project-1-STA486C’)
```