Data Sci: Group 10 Draft

Matthew O'Donnell, Sean Sander, Daniel Forcade

4/9/2021

1. Introduction:

What is the source of the data?: The data is sourced from a database of about 12,000 civilian complaints against New York City police officers.

Where and when was it created?: The data was made public on July 26, 2020 and it spans from September 1985 to January 2020, detailing incidents from all around the city.

Is it a sample?: The data is not a sample, it is a thorough list of active-duty officers who have had at least one allegation against them that is substantiated by New York's Civilian Complaint Review Board (CCRB), meaning that a civilian provided a sworn statement to investigators.

Do you suspect any sampling bias?: Our data set is not originally a random sample, so biases may exist in the collection of the data. The data set is limited to complaints filed against officers that were still on the force as of June 2020, so it is not all-encompassing of the types and number of complaints filed against the NYPD between 1985 and 2020.

Experiment or Observational study?: The data is an observational study, so no causation can necessarily be inferred from our data.

How were measurements taken?: The data set includes the name, badge number, rank, and precinct (location) of each officer, and demographic information about both the complainant and the officer. Also contained is whether the CCRB found the officer's conduct violated NYPD rules. In total, there are 33,358 observations of 27 variables.

Do you suspect any bias in the measurements?: Whether bias exists in the measurements is an auxiliary part of our analysis. It is possible that the data may be incomplete, despite it theoretically being an all-inclusive database of information.

Why is this data of interest to you?: This data is very interesting to us for a number of reasons. For one, the NYPD has a well-documented history of alleged racism and discrimination towards minorities, and we wanted to see if those kinds of trends were present in our data as well. This data set is particularly interesting because it was just released; before mid-2020, complaints against officers were kept a secret, as were the outcomes of those complaints.

What kind of data cleaning was necessary?: In progress. **Will expand** this section when all transformations are complete. Data set needed significant cleaning and manipulation for what we set out to do. We have such a large number of transformations and setups that we are discussing (and likely will) include a separate section for data cleaning/manipulation for the final report. For right now, all of our r-code and exploratory work is included in data visualizations.

Our project is focused on exploring data relating to complaints made against New York City police officers, and the specific differences between Precinct 75 and the other precincts. New York has a long, well-documented history of police misconduct and discrimination towards minorities, much of which driven by the CompStat system, which introduced an over-reliance on data-driven methods that often came at the expense of people of color. Precinct 75 in particular is historically known for police corruption and preliminary analysis of our data showed that the "75" had the highest number of complaints, a majority of which were made by people of color against white officers.

This data set is particularly interesting because it was just released on July 26, 2020; before mid-2020, complaints against officers were kept a secret, as were the outcomes of those complaints. The data is sourced from a database of about 12,000 civilian complaints against New York City police officers, and it detailing incidents from all around the city spanning from September 1985 to January 2020. The data set includes the name, badge number, rank, and precinct (location) of each officer, and demographic information about both the complainant and the officer. Also contained is whether the CCRB found the officer's conduct violated NYPD rules. In total, there are 33,358 observations of 27 variables.

The data is not a sample; rather, it is a thorough list of active-duty officers who have had at least one allegation against them that is substantiated by New York's Civilian Complaint Review Board (CCRB), meaning that a civilian provided a sworn statement to investigators. Biases may exist in the collection of the data because the data is not a random sample. The data set is also limited to complaints filed against officers that were still on the force as of June 2020, so it is not all-encompassing of the types and number of complaints filed against the NYPD between 1985 and 2020. It is possible that the data may be incomplete, despite it supposedly being an all-inclusive database of information. In addition, the data is an observational study, so no causation can necessarily be inferred from our data.

2. Data Visualizations:

Notes: Split roughly into two sections, exploration/manipulation and presentation graphs. Both of these sections are works in progress; we wanted to set up a framework within the data where we gained familiarity with the data, and then organized it to produce the information we are looking for.

Data Exploration/Manipulation:

Our data-set was large, unwieldy, and filled with tricky and incomplete observations. We had several avenues of attack when cleaning and preparing the data, which we worked sequentially:

1. First, we converted all blank values (") into `NA` as the dataset had a substantial amount of missing values. However, we did not initially remove the `NA` values, as the initial set had almost 30 variables, and we wanted to be sure the missing variable was truly relevant to our analysis before removing the entire observation.
2. We then conducted research to define what each variable meant, since many were still coded (for example, `mos_ethnicity` is member of service ethnicity, which refers to the ethnicity of the responding officer).
3. Next, we identified and cleaned/prepared critically important variables, such as `board_disposition`, which refers to the complaint outcome (Substantiated, Unsubstantiated, or Exonerated). Another example in this phase was identifying that some officers had the same last or first name, which we solved by combining the First and Last officer name into a single identifier.

4. Then, we attacked the dataset in a manner in which to look for initial trends, as to provide a story or angle. We primarily used descriptive statistics in this phase, supported by visuals when an interesting trend developed. We narrowed our focus down to two specific aspects:

    - A few officers had a tremendous amount of complaints.

    - A single precinct (Precinct 75) had an overwhelming number of complaints when compared against any other single precinct.

5. Ultimately, we had a few ethical concerns with conducting and focusing on the officer component (as they had their full name exposed), as decided on exploring Precinct 75.

6. Once deciding on this category, we created several preliminary exploration graphs to take a closer look at the data. What we determined here, was that we needed to primarily focus on the last 10 years of data, as although the dataset technically covered 1986-2019, there was not a substantial or consistent amount of data entries prior to 2005.

```
## Read in data
allegations <- read.csv("allegations_202007271729.csv", na.strings = c("", "NA"))

## Check names
names(allegations)
```

```
##  [1] "unique_mos_id"            "first_name"
##  [3] "last_name"                "command_now"
##  [5] "shield_no"                "complaint_id"
##  [7] "month_received"           "year_received"
##  [9] "month_closed"             "year_closed"
## [11] "command_at_incident"      "rank_abbrev_incident"
## [13] "rank_abbrev_now"          "rank_now"
## [15] "rank_incident"            "mos_ethnicity"
## [17] "mos_gender"               "mos_age_incident"
## [19] "complainant_ethnicity"    "complainant_gender"
## [21] "complainant_age_incident" "fado_type"
## [23] "allegation"               "precinct"
## [25] "contact_reason"           "outcome_description"
## [27] "board_disposition"
```

```
###### New Column: Officer Full Name (combined first and last)
allegations$full_name <- paste(allegations$first_name, allegations$last_name)




## Board disposition transformation
## Substantiated variants merged into one category: 'Substantiated"

table(allegations$board_disposition)
```

```
##
##                              Exonerated
##                                    9609
```

```
##                   Substantiated (Charges)
##                                      3796
##     Substantiated (Command Discipline A)
##                                       964
##     Substantiated (Command Discipline B)
##                                       789
##       Substantiated (Command Discipline)
##                                       851
## Substantiated (Command Lvl Instructions)
##                                       454
##      Substantiated (Formalized Training)
##                                      1033
##             Substantiated (Instructions)
##                                       248
##         Substantiated (MOS Unidentified)
##                                         1
##       Substantiated (No Recommendations)
##                                       165
##                           Unsubstantiated
##                                     15448
```

```
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Charges)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Command Discipline A)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Command Discipline B)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Command Discipline)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Command Lvl Instructions)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Formalized Training)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (Instructions)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (MOS Unidentified)'] <- 'Substantiated'
allegations$board_disposition[allegations$board_disposition ==
                                'Substantiated (No Recommendations)'] <- 'Substantiated'


###
# Officer Name Frequency
###

officer_freq <- allegations %>%
  group_by(full_name) %>% tally

# Top down order
officer_freq <- officer_freq %>%
  arrange(desc(n))

head(officer_freq, 15)
```

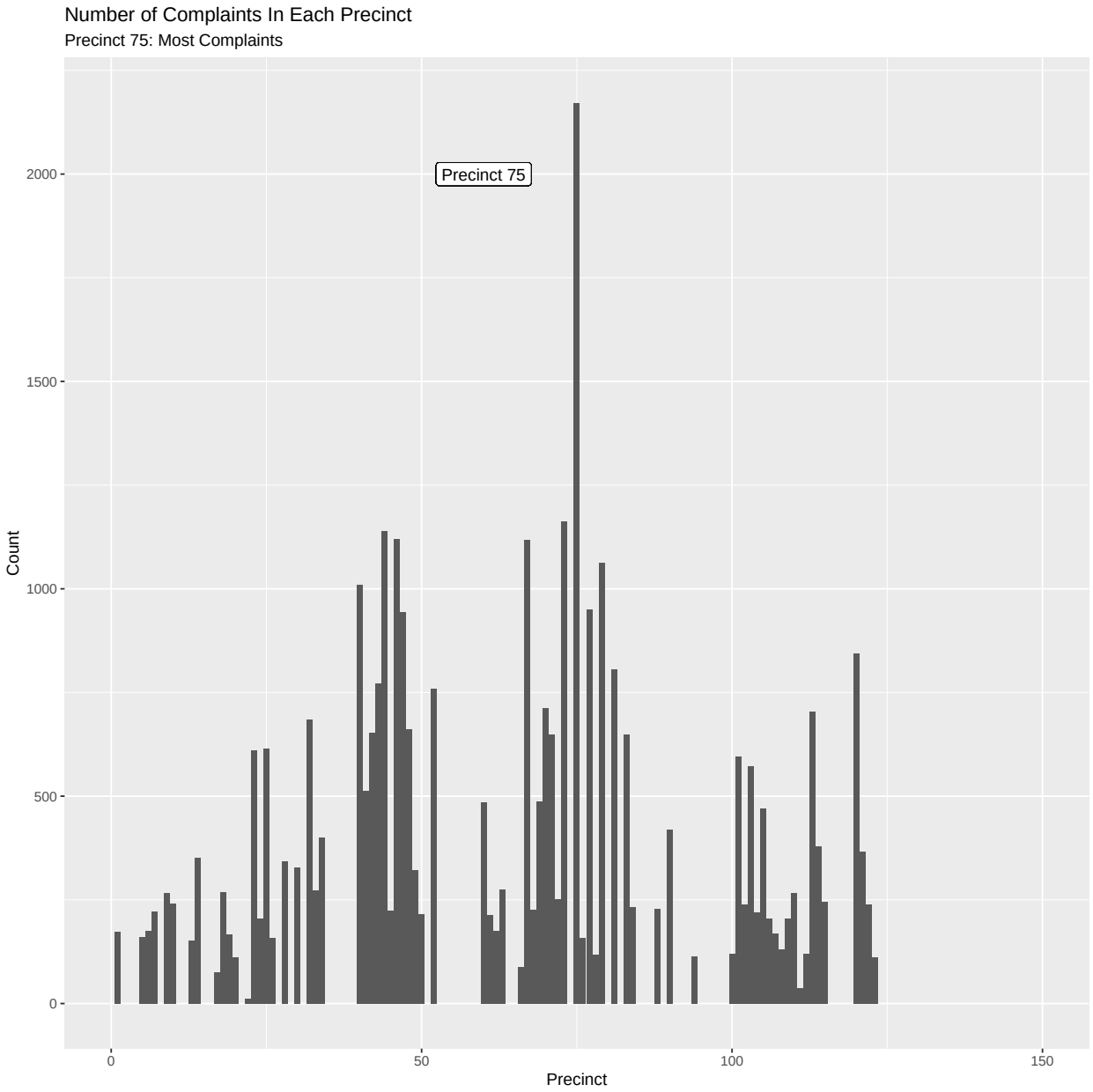
```
## # A tibble: 15 x 2
##    full_name               n
##    <chr>               <int>
##  1 Daniel Sbarra          75
##  2 Mathew Reich           75
##  3 Gary Messina           73
##  4 Joseph Tallarine       73
##  5 Christophe Mccormack   72
##  6 William Taylor         65
##  7 David Cheesewright     63
##  8 Mike Civil             56
##  9 Paul Mcmahon           56
## 10 Michael Raso           50
## 11 Matthew Lewis          47
## 12 Robert Delaney         47
## 13 Trevor Baronette       47
## 14 David Grieco           46
## 15 Michael Miller         46
```

```
###
# Officer ID Frequency
###

# id_freq <- allegations %>%
#  group_by(unique_mos_id) %>% tally

# Top down order
# id_freq <- id_freq %>%
#  arrange(desc(n))

# head(id_freq, 15)



### - Notes: Names and IDs to not match: 3996 to 3958.
### Some officers likely have overlapping names; id is better identifier


###
# Precinct frequency
###

precinct_freq <- allegations %>%
  group_by(precinct) %>% tally

# Top down order
precinct_freq <- precinct_freq %>%
  arrange(desc(n))


head(precinct_freq, 15)
```

```
## # A tibble: 15 x 2
##    precinct     n
```

```
##       <int> <int>
##  1       75  2172
##  2       73  1163
##  3       44  1139
##  4       46  1120
##  5       67  1119
##  6       79  1062
##  7       40  1009
##  8       77   950
##  9       47   944
## 10      120   844
## 11       81   806
## 12       43   772
## 13       52   759
## 14       70   713
## 15      113   704
```

```
## geom_point of complaints in each precinct
## Quick visualiation of precinct complaints
## Precinct 75 identified as leading complaint area
ggplot(data = precinct_freq,
       mapping = aes(x = precinct, y = n))+
  geom_col()+
  geom_label(x = 60, y =  2000, label = "Precinct 75")+
  labs(title = "Number of Complaints In Each Precinct", subtitle = "Precinct 75: Most Complaints",
       y = "Count", x = "Precinct")+
  xlim(0,150)
```




```
################## Precinct 75 ################

## Precinct 75 Notes:

# * Historically corrupt precinct in NYPD
# * Covers an area in Brooklyn, NY
# * By far the most numerous complain precinct in our data-set
#
#
# NYPD Page: https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts/75th-precinct.page
#
# Documentary Wikipedia Info: https://en.wikipedia.org/wiki/The_Seven_Five
#
# Recent article on misconduct complaints: https://theintercept.com/2020/08/23/nypd-75th-precinct-polic
```

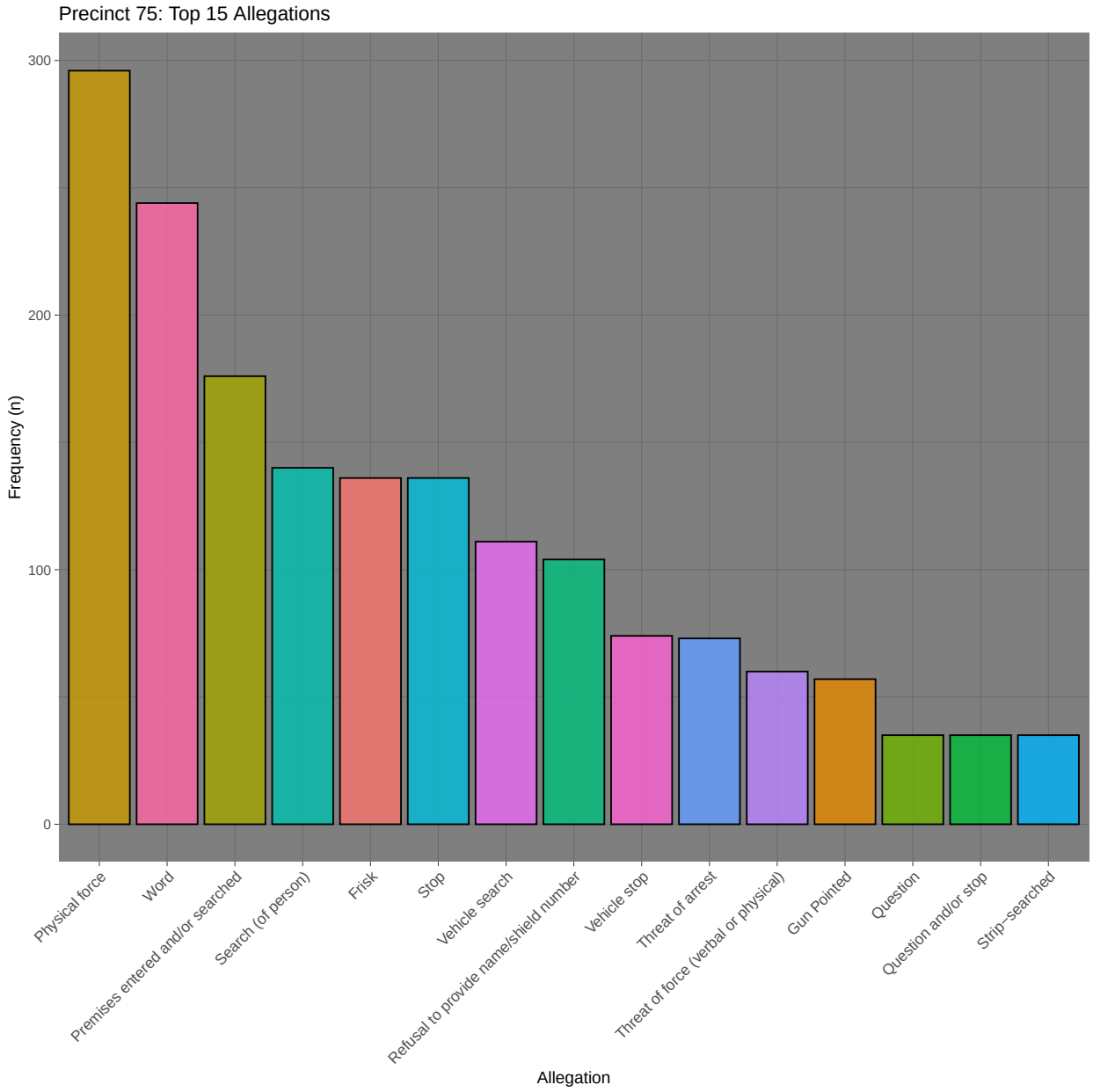
```
## High score champion: Precinct 75

## Allegation data for Precinct 75
precinct_75 <- allegations %>%
  filter(allegations$precinct == 75)


## Allegations Data minus Precinct 75
allegations_no_75 <- allegations %>%
  filter(allegations$precinct != 75)




## Precinct 75 allegation freq
precinct_75_allegation <- precinct_75 %>%
  group_by(allegation) %>% tally %>%
  arrange(desc(n))

head(precinct_75_allegation, 15)
```

```
## # A tibble: 15 x 2
##    allegation                                n
##    <chr>                                 <int>
##  1 Physical force                          296
##  2 Word                                    244
##  3 Premises entered and/or searched        176
##  4 Search (of person)                      140
##  5 Frisk                                   136
##  6 Stop                                    136
##  7 Vehicle search                          111
##  8 Refusal to provide name/shield number   104
##  9 Vehicle stop                             74
## 10 Threat of arrest                         73
## 11 Threat of force (verbal or physical)     60
## 12 Gun Pointed                              57
## 13 Question                                 35
## 14 Question and/or stop                     35
## 15 Strip-searched                           35
```

```
##top 15
top_15 <- precinct_75_allegation %>%
  filter(n >= 35)
```

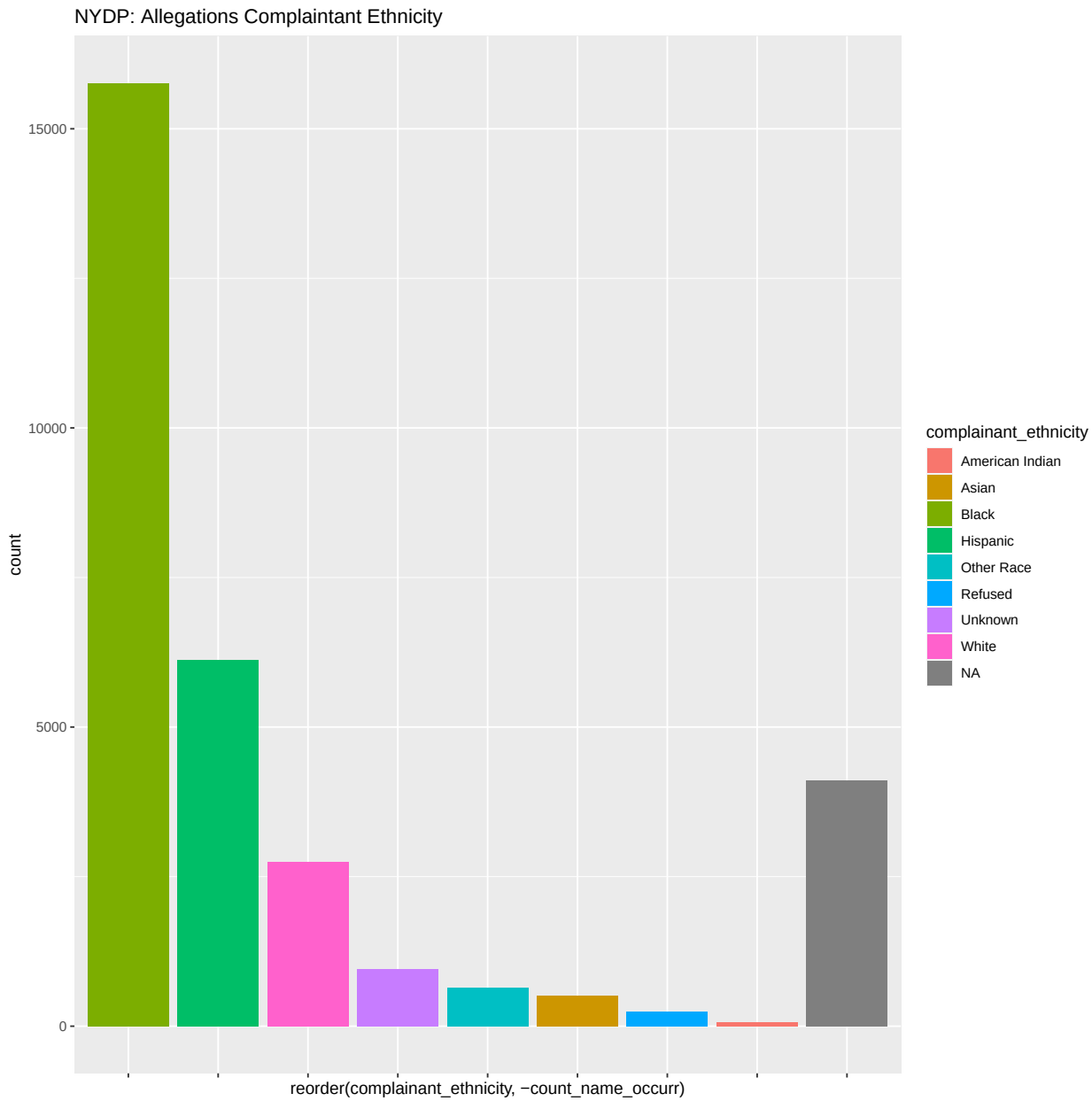
```
## Precinct 75: Top 15 Allegations
ggplot(data = top_15,
       mapping = aes(x = reorder(allegation, -n), y = n, fill = allegation))+
  geom_bar(stat = 'identity', alpha = .8, color = 'black')+
  theme_dark()+
  scale_fill_discrete()+
```

```
  guides(fill = FALSE, alpha = FALSE, size = FALSE, color = FALSE)+
  labs(title = "Precinct 75: Top 15 Allegations", x = "Allegation", y = "Frequency (n)")+
  theme(axis.text.x=element_text(angle = 45, hjust = 1, size = 10))
```



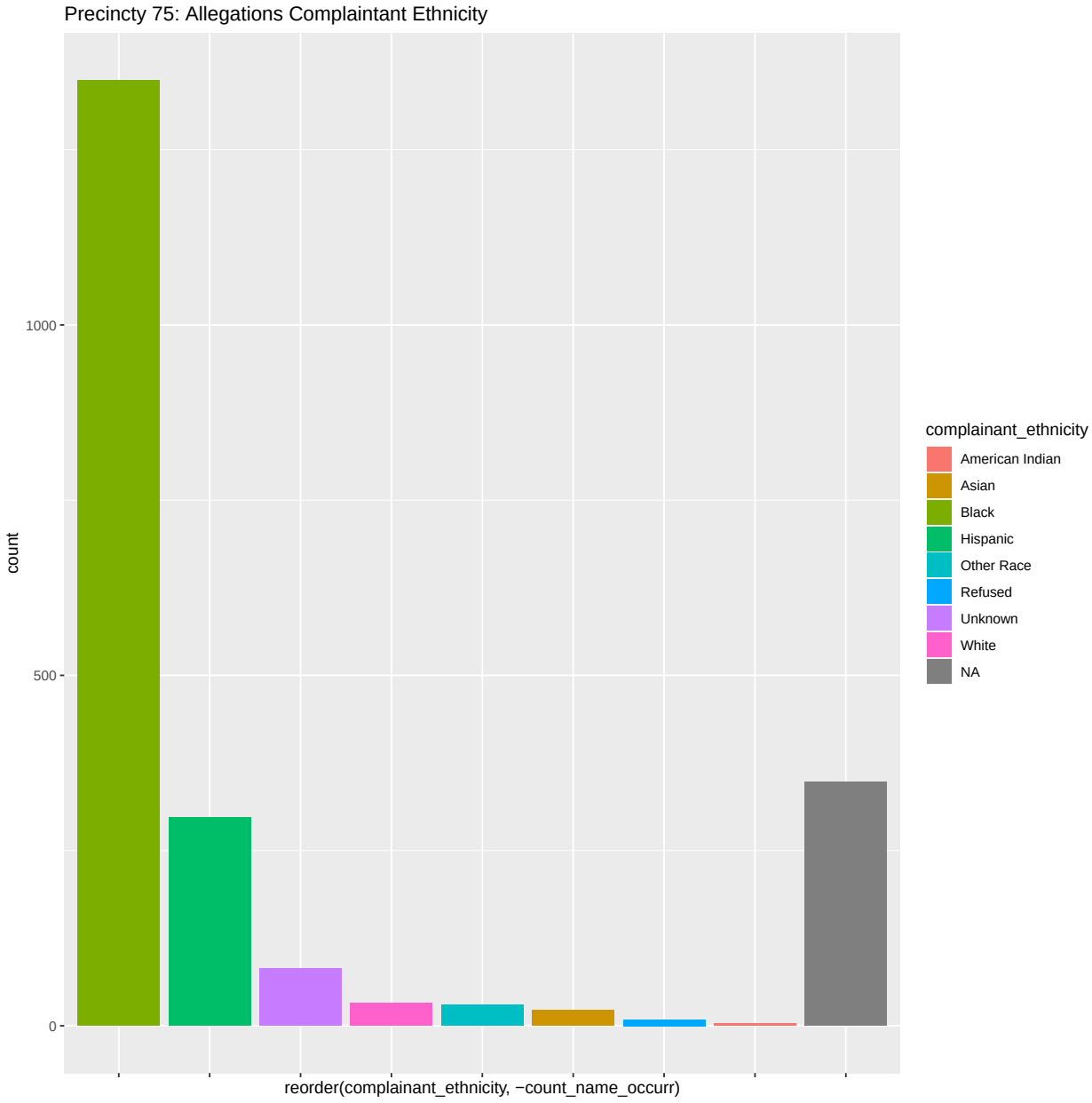

```
## Allegations w/o 75 Order Count Mutate
allegations_no_75 <- allegations_no_75 %>%
  group_by(complainant_ethnicity) %>%
  mutate(count_name_occurr = n())

## Precinct 75 Order Count Mutate
precinct_75 <- precinct_75 %>%
  group_by(complainant_ethnicity) %>%
  mutate(count_name_occurr = n())
```

```
## Allegations no 75 Complaint Ethnicity
ggplot()+
  geom_bar(data = allegations_no_75,
           mapping = aes(x = reorder(complainant_ethnicity, -count_name_occurr), y = ..count..,
                         fill = complainant_ethnicity))+
  geom_bar()+
  labs(title = "NYDP: Allegations Complaintant Ethnicity")+
  theme(axis.text.x=element_blank())
```




```
## Precinct 75 Complaint Ethnicity
ggplot()+
  geom_bar(data = precinct_75,
           mapping = aes(x = reorder(complainant_ethnicity, -count_name_occurr), y = ..count..,
```

```
                        fill = complainant_ethnicity))+
  geom_bar()+
  labs(title = "Precincty 75: Allegations Complaintant Ethnicity")+
  theme(axis.text.x=element_blank())
```




```
############ Last 10 years ##############

# Preliminary graphs and analysis revealed lack of data for earlier years
# Solution: Focus on bulk of data: last 10 years


## last 10 years
recent_allegations <- allegations %>%
```

```
  filter(allegations$year_received >= 2009)


## Allegations for Precinct 75 ##
precinct_75_last10 <- recent_allegations %>%
  filter(recent_allegations$precinct == 75)



##### Allegations Data minus Precinct 75  #####
## 77 total precincts included
allegations_no_75_last10 <- recent_allegations %>%
  filter(recent_allegations$precinct != 75)


## Number of precincts in last 10 years
number_of_precincts <- allegations_no_75_last10 %>%
  group_by(precinct) %>% tally

## length = 77
length(number_of_precincts$precinct)
```

```
## [1] 77
```

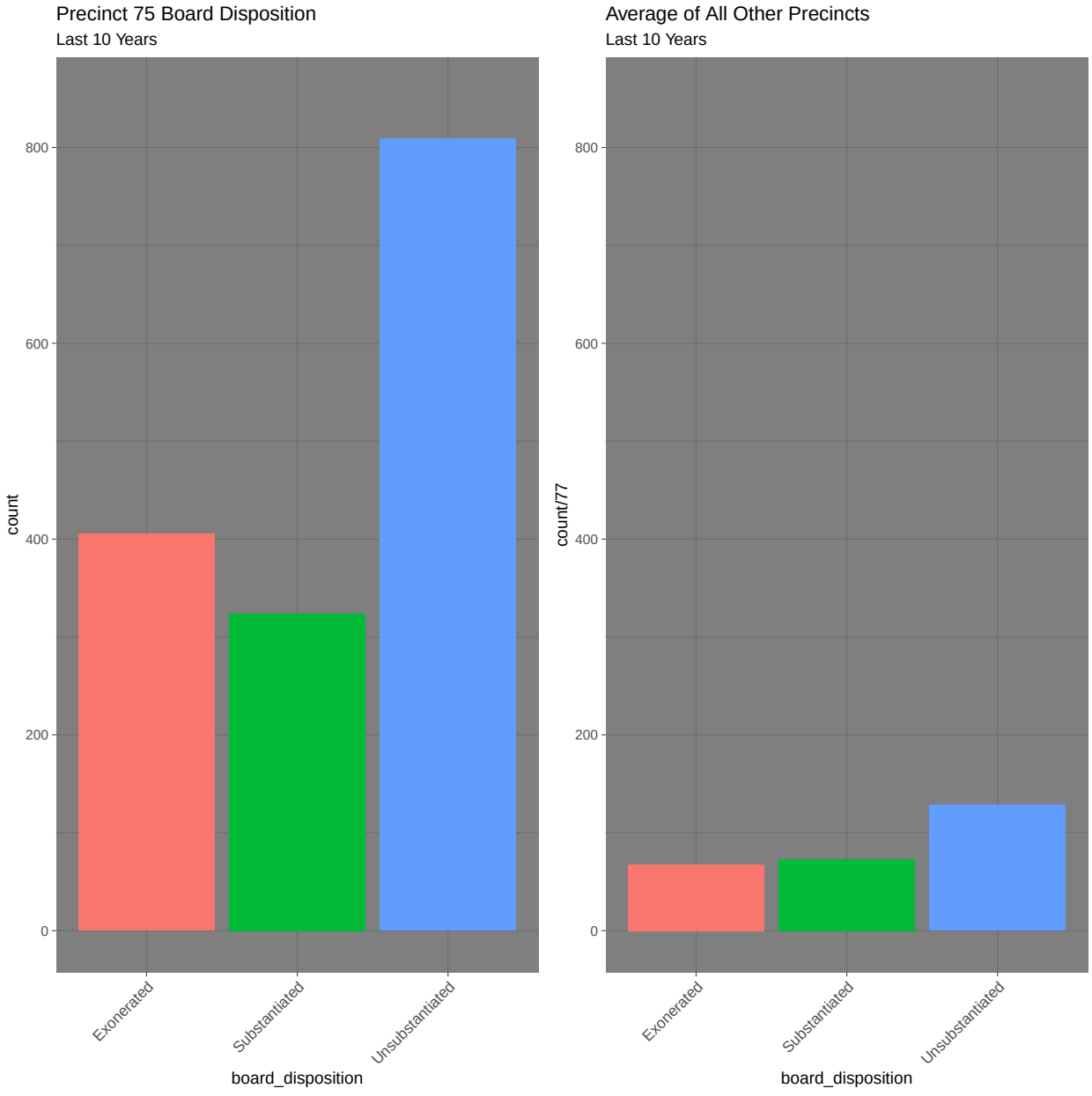
Presentation Graphs:

Note: Not final version. Colors/formatting will be updated.

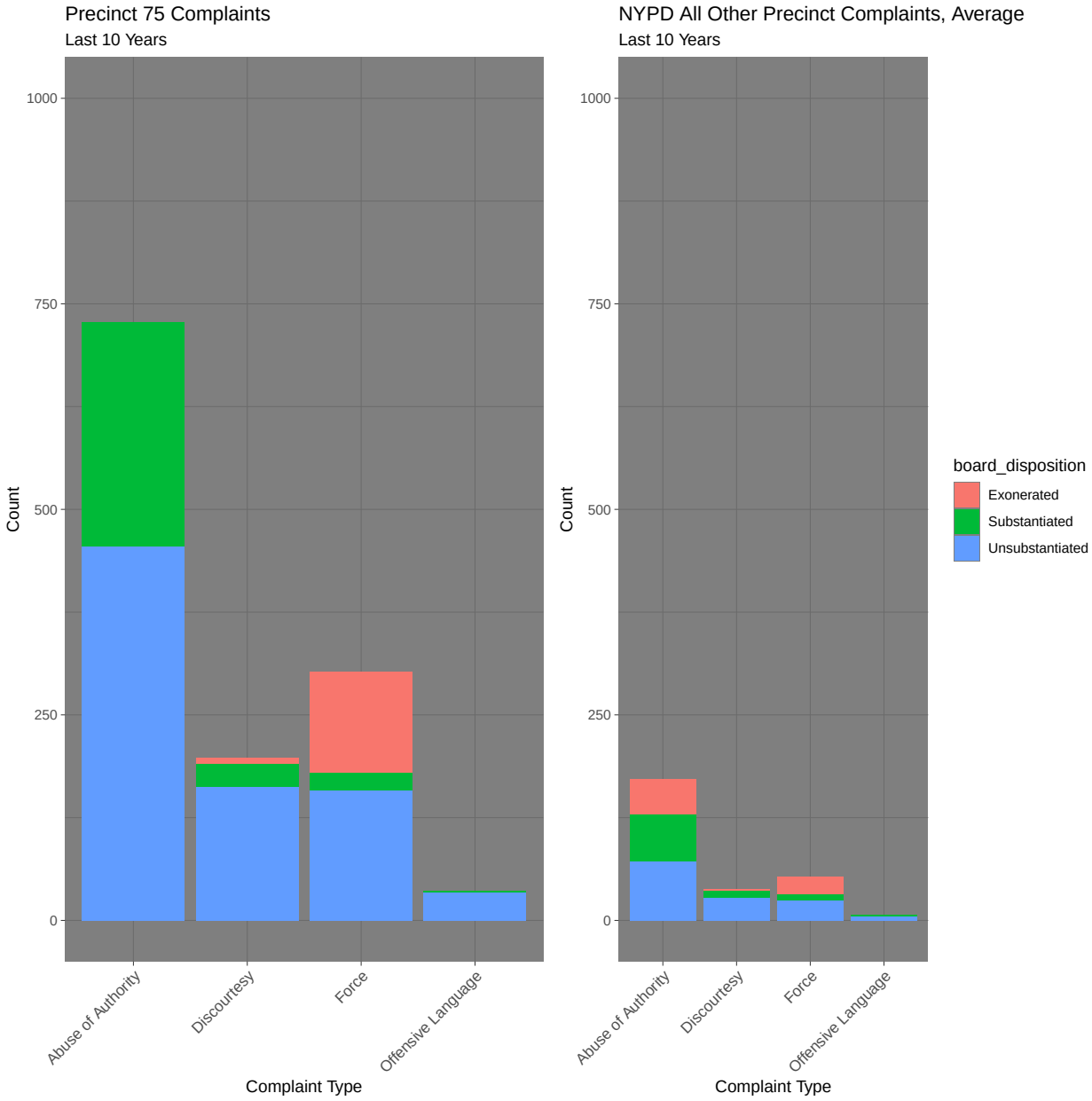
This graph (board disposition comparison) is a visualization of complaint outcome of Precinct 75 versus the average of the rest of the NYPD precincts.

```
## Board disposition 75 vs Average NYPD Precinct (..count../77)
## Last 10 years
## Board disposition == outcome

#75
g1 <- ggplot(data = precinct_75_last10,
       mapping = aes(x = board_disposition, y = ..count.., fill = board_disposition))+
  geom_bar()+
  labs(title = "Precinct 75 Board Disposition", subtitle = "Last 10 Years")+
  theme_dark()+
  theme(axis.text.x=element_text(angle = 45, hjust = 1, size = 10))+
  guides(fill = FALSE)+

  ylim(0,850)

#NYPD
g2 <- ggplot(data = allegations_no_75_last10,
       mapping = aes(x = board_disposition, y = ..count../77, fill = board_disposition))+
  geom_bar()+
  labs(title = "Average of All Other Precincts", subtitle = "Last 10 Years")+
```

```
  theme_dark()+
  theme(axis.text.x=element_text(angle = 45, hjust = 1, size = 10))+
  guides(fill = FALSE)+
  ylim(0,850)

ggarrange(g1,g2)
```

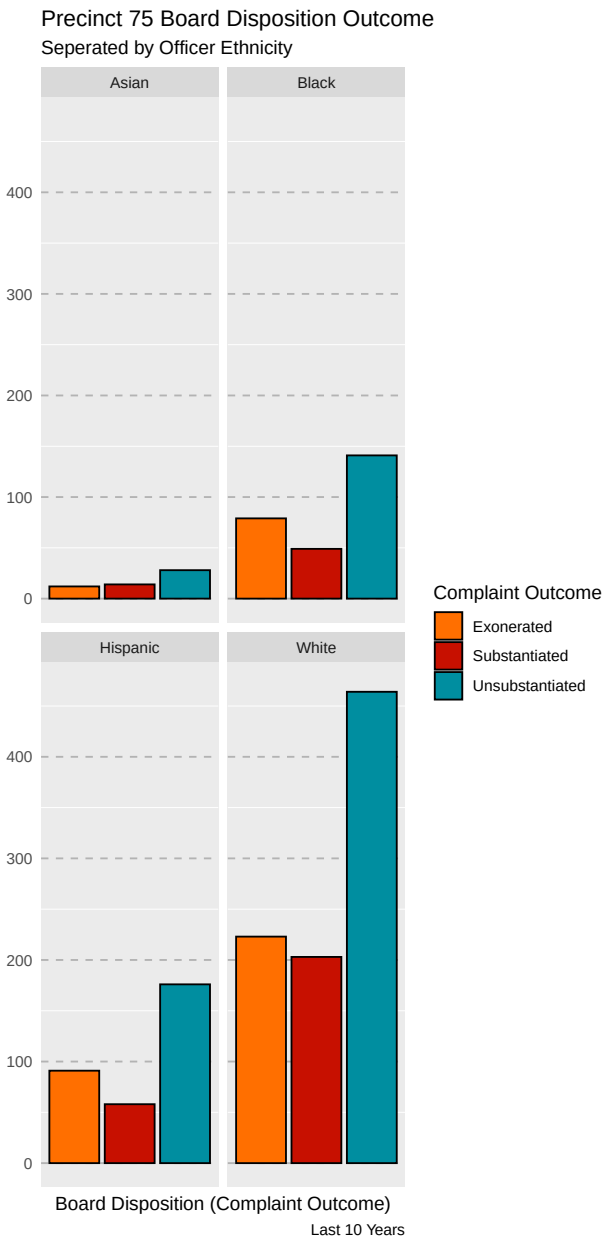


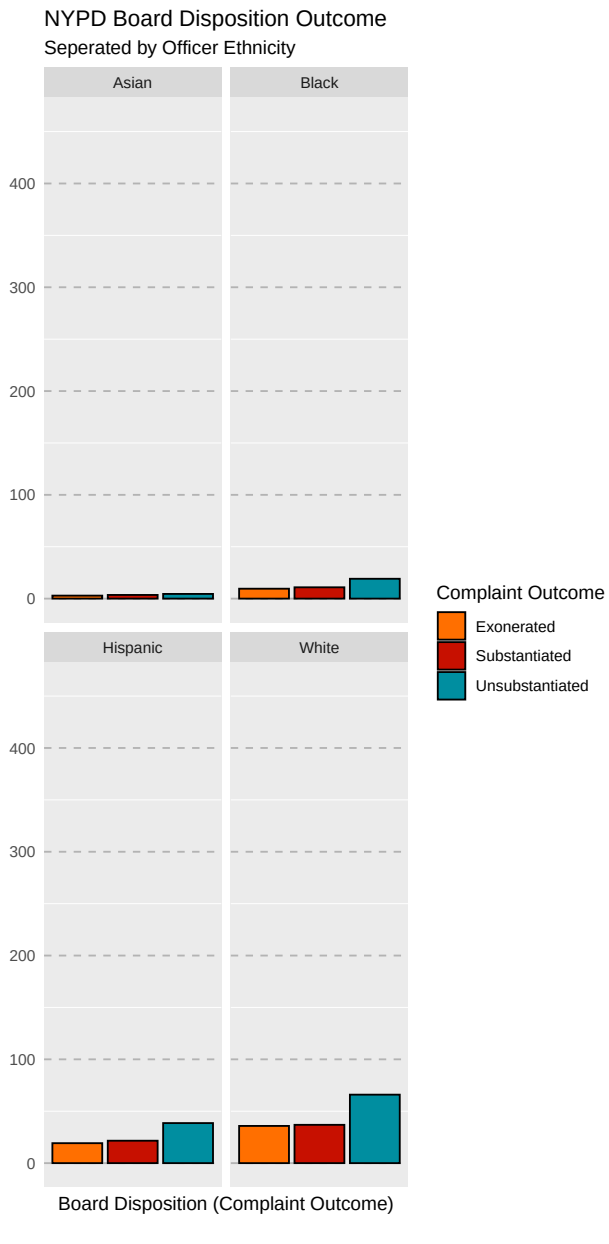
This plot (Complaints) shows the distribution of complaint types for Precinct 75 compared to all other precincts, on average, as well as the outcome of the complaint case. Again, the total complaint disparity is present here. There is little to no difference between Precinct 75 and other precincts in terms of complaint type. The most common complaint across all other precincts was Abuse of Authority on behalf of the officer, followed by Force, Discourtesy, and Offensive Language. This distribution keeps its shape for the 75. The main difference is that out of all Abuse of Authority cases in Precinct 75, none were Exonerated, only Substantiated or Unsubstantiated. This differs from Abuse of Authority complaints across all other precincts, where there is about a 1/3 split for what the outcome of the case will be.



This plot (Board Disposition Outcome by Officer Ethnicity) shows the breakdown of complaint outcomes for Precinct 75 vs all other precincts, on average, faceted by the ethnicity of the officer. Overall, the majority of officers with complaints are white, with fewer black or hispanic and an even smaller number of asian officers. This ethnicity distribution is the same across all precincts including 75, and is in line with what we would expect in a large city like New York, that likely has a large number of minority police officers. Looking at total complaints and not the proportion of total officers in the NYPD who have received a complaint, it is hard to see whether white officers are more likely to receive a complaint than black, asian, or hispanic officers. It is, however, evident that white police officers tend to be exonerated more across the board, especially in Precinct 75. In that precinct, there was not one incident of a white officer having a complaint marked as unsubstantiated.

```
##### Board Disposition by Officer Ethnicity | Last 10 Years


## Filter extremely low value
allegations_no_75_last10_OE <- allegations_no_75_last10 %>%
  filter(mos_ethnicity != "American Indian")


#75
g1 <- ggplot(data = precinct_75_last10,
       mapping = aes(x = board_disposition, fill = board_disposition))+
  geom_bar(color = "black")+
  facet_wrap(~mos_ethnicity)+
  theme_cleveland()+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  labs(title = "Precinct 75 Board Disposition Outcome", subtitle = "Seperated by Officer Ethnicity",
       caption = "Last 10 Years", x = "Board Disposition (Complaint Outcome)")+
  scale_fill_futurama(name = "Complaint Outcome")+
  ylim(0,470)

#NYPD
g2<- ggplot(data = allegations_no_75_last10_OE,
       mapping = aes(x = board_disposition, y = ..count../77, fill = board_disposition))+
  geom_bar(color = "black")+
  facet_wrap(~mos_ethnicity)+
  theme_cleveland()+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  labs(title = "NYPD Board Disposition Outcome", subtitle = "Seperated by Officer Ethnicity",
       caption = "", x = "Board Disposition (Complaint Outcome)")+
  scale_fill_futurama(name = "Complaint Outcome")+
  ylim(0,460)


ggarrange(g1,g2)
```

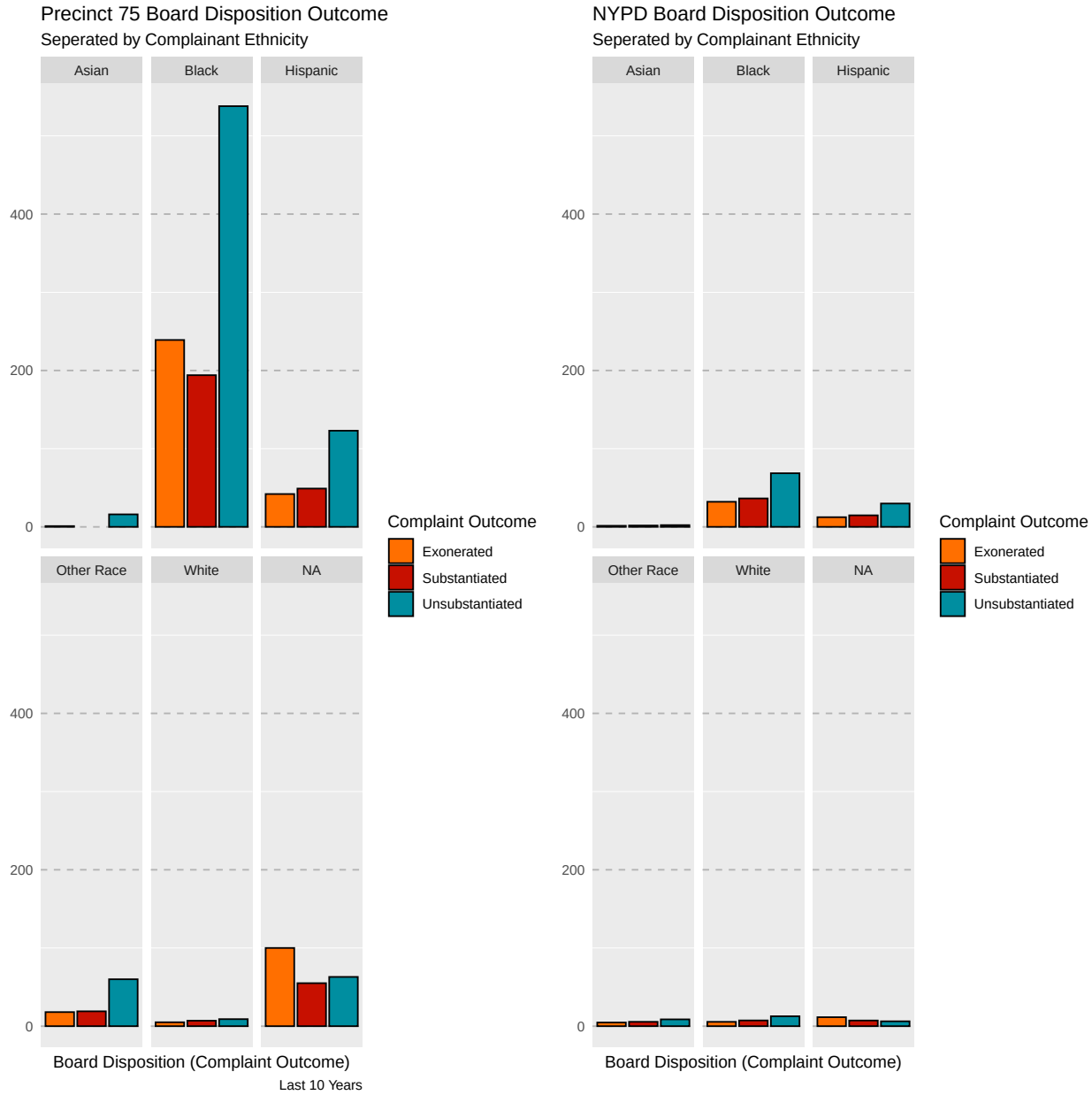

Precinct 75 Board Disposition Outcome
Seperated by Officer Ethnicity
Asian
Black
Hispanic
White
400
300
200
100
0
Complaint Outcome
Exonerated
Substantiated
Unsubstantiated
Board Disposition (Complaint Outcome)
Last 10 Years


NYPD Board Disposition Outcome
Seperated by Officer Ethnicity
Asian
Black
Hispanic
White
400
300
200
100
0
Complaint Outcome
Exonerated
Substantiated
Unsubstantiated
Board Disposition (Complaint Outcome)

This plot (Outcome by Complainant Ethnicity) clearly shows that the majority of complainants to the NYPD are black. There are some hispanic complainants in the data set, as well as other minorities, who have been grouped into the "Other Races" category. Additionally, there are very few white and asian complaints, and this is especially true in precinct 75. This is extremely telling of the kinds of people that police have incidents with and tend to target for various reasons. Overall, black complainants tend to have their complaints marked as unsubstantiated, and arguably more so in the 75th precinct. Officers in Precinct 75 also have a much higher chance of being exonerated than other precincts, on average.

```
######## Outcome by Complainant Ethnicity
## Considered filter(), decided was better to include in combined group

## Low Value Clean
#75
precinct_75_last10_CE <- precinct_75_last10
precinct_75_last10_CE$complainant_ethnicity[precinct_75_last10_CE$complainant_ethnicity ==
                                              'American Indian'] <- 'Other Race'
precinct_75_last10_CE$complainant_ethnicity[precinct_75_last10_CE$complainant_ethnicity ==
                                              'Refused'] <- 'Other Race'
precinct_75_last10_CE$complainant_ethnicity[precinct_75_last10_CE$complainant_ethnicity ==
                                              'Unknown'] <- 'Other Race'
precinct_75_last10_CE$complainant_ethnicity[precinct_75_last10_CE$complainant_ethnicity ==
                                              ''] <- 'Other Race'
## Rest of NYPD
allegations_no_75_last10_CE <- allegations_no_75_last10
allegations_no_75_last10_CE$complainant_ethnicity[allegations_no_75_last10_CE$complainant_ethnicity ==
                                                    'American Indian'] <- 'Other Race'
allegations_no_75_last10_CE$complainant_ethnicity[allegations_no_75_last10_CE$complainant_ethnicity ==
                                                    'Refused'] <- 'Other Race'
allegations_no_75_last10_CE$complainant_ethnicity[allegations_no_75_last10_CE$complainant_ethnicity ==
                                                    'Unknown'] <- 'Other Race'
allegations_no_75_last10_CE$complainant_ethnicity[allegations_no_75_last10_CE$complainant_ethnicity ==
                                                    ''] <- 'Other Race'

#75
g1 <- ggplot(data = precinct_75_last10_CE,
             mapping = aes(x = board_disposition, fill = board_disposition))+
  geom_bar(color = "black")+
  facet_wrap(~complainant_ethnicity)+
  theme_cleveland()+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  labs(title = "Precinct 75 Board Disposition Outcome", subtitle = "Seperated by Complainant Ethnicity"
       caption = "Last 10 Years", x = "Board Disposition (Complaint Outcome)")+
  scale_fill_futurama(name = "Complaint Outcome")+
  ylim(0,540)

#NYPD
g2<- ggplot(data = allegations_no_75_last10_CE,
            mapping = aes(x = board_disposition, y = ..count../77, fill = board_disposition))+
  geom_bar(color = "black")+
  facet_wrap(~complainant_ethnicity)+
  theme_cleveland()+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
```

```
  labs(title = "NYPD Board Disposition Outcome", subtitle = "Seperated by Complainant Ethnicity",
       caption = "", x = "Board Disposition (Complaint Outcome)")+
  scale_fill_futurama(name = "Complaint Outcome")+
  ylim(0,540)


ggarrange(g1,g2)
```



3. Machine Learning: To Do Later:

Ideas:

- Compare precincts against normal distribution
- Compare precincts against each other
- Compare officer race vs complaint outcome
- Compare complainant race vs complaint outcome

4. Conclusions:

This was/is a challenging data set, and we had several avenues of attack. First, we created a brainstorm and some rough visualizations to see what we were working with. We wanted to become familiar with the data, and created several probes through dplyr to check for potential interesting developments. We found several, but the prize-horse development was the finding of the incredible `Precinct` outlier, `Precinct 75`. This prompted outside research into the area, where we found that 75 has a lengthy and troubled history in the city of New York.

Our goal is to explore if the reputation of Precinct 75 as described by publications and journals of racism, excessive complaints, and unwarranted promotions matches our data. We have set our data up with several transformations and exploratory visualizations and tables in order to serve as a launch-pad for more in-depth statistical analysis, including machine learning. Currently, our presentation graphs are basic bar charts, but we intend to push our analysis further in the final version. We believe that we have successfully created a solid framework and foundation to begin with deeper statistical tests and analysis regarding Precinct 75 in comparison with the rest of the NYPD.

Additionally, we have some light, surface analysis on the NYPD as a whole, but our focus and story is and will be on Precinct 75.

5. Limitations/Recommendations:

There are a few limitations to our analysis; for one, our data set is incomplete by only containing complaints made against officers who were still on the force as of June 2020. There are likely many data points that were removed from the data set as the officer in question retired. It is also possible that the data set might not be entirely complete.

Notes:

- NYPD had a large amount of time to manipulate data before public release
- Second organization accessed data-set before public release (increased potential for manipulation)
- Research into precinct locations would allow borough or location partition for data analysis