DSNYPD

me

2022-06-11

#Introduction Going to look at a datasource on NYPD shooting. Why? Because I was told to do this as HW. :(the link i will be using for getting the data is https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Step1 - Indentify and import the Data

I will start by reading the data from a csv file

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD_shootings <- read_csv(url_in)
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tidying and Transforming Data

after looking at the data I would like to look into information about the perp and information about the victim and see if there is any relation

so i will be getting rid of columns that are not of interest to me.

```
NYPD_shootings <- NYPD_shootings %>%
  select(c(PERP_AGE_GROUP,PERP_SEX,PERP_RACE,VIC_AGE_GROUP,VIC_SEX,VIC_RACE))
```

I am also only interested in full information about the perp and vic and will filter out any rows that have NA

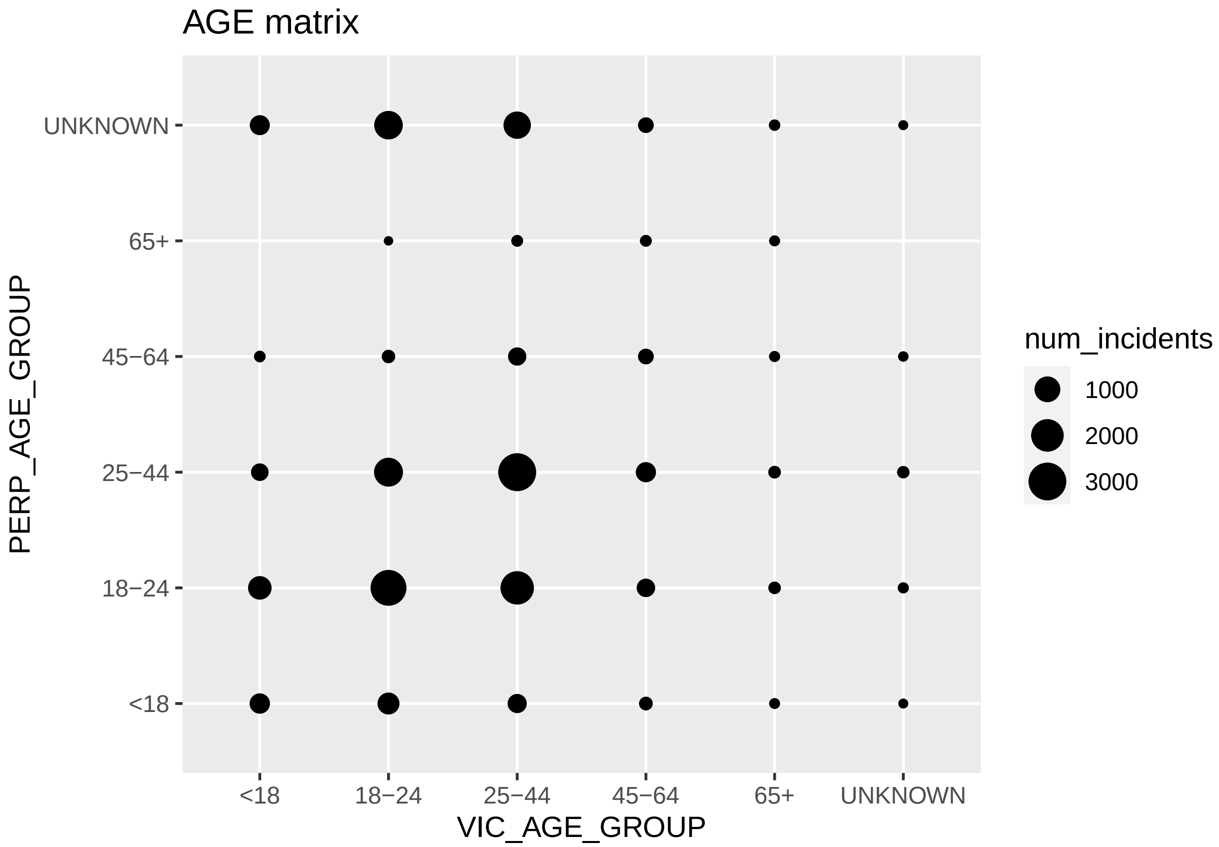
```
NYPD_shootings <- NYPD_shootings %>%
  drop_na()
```

visulaizing Data

I am going to attempt to visualize this data

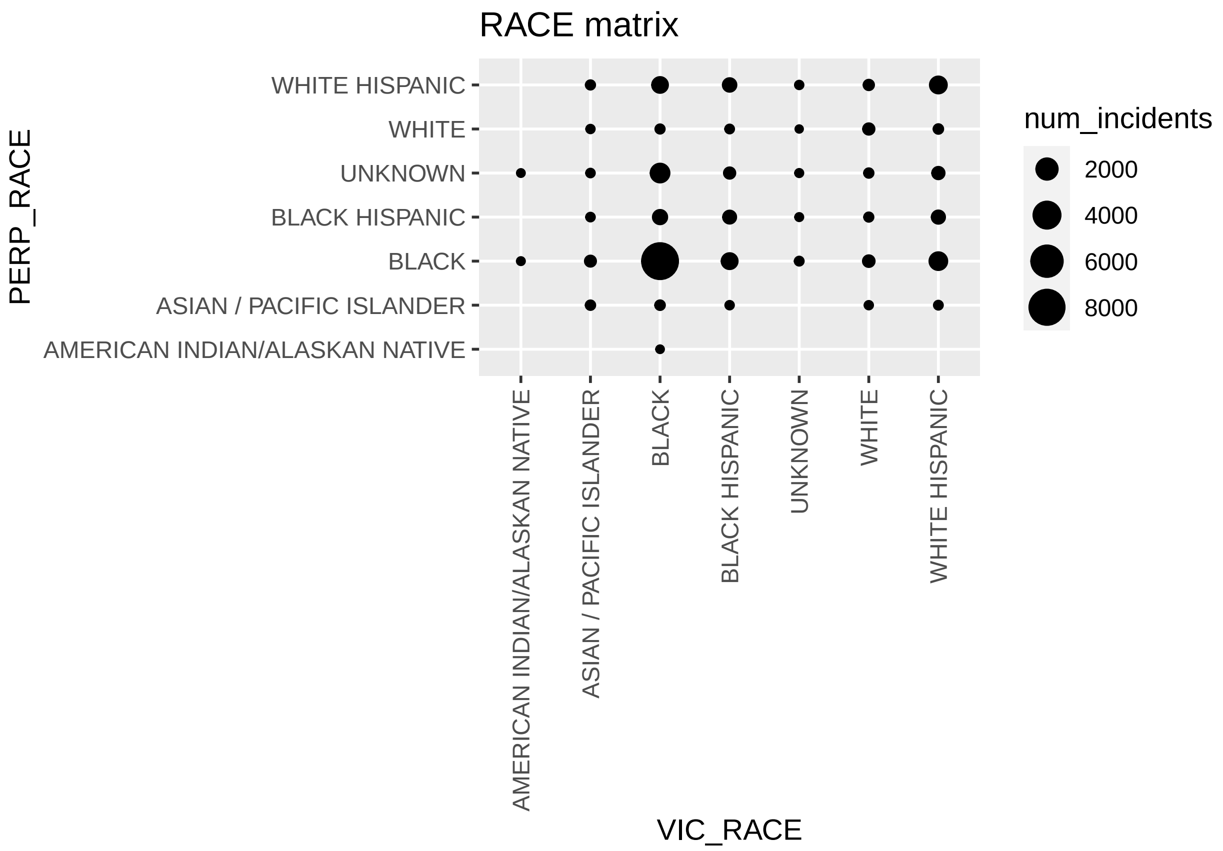
visualizing ages

```
AGE_NYPD <- NYPD_shootings %>%
  select(PERP_AGE_GROUP, VIC_AGE_GROUP) %>%
  group_by(PERP_AGE_GROUP, VIC_AGE_GROUP) %>%
  tally() %>%
  rename(num_incidents = n) %>%
  filter(PERP_AGE_GROUP != '1020', PERP_AGE_GROUP != 224, PERP_AGE_GROUP != 940) %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = PERP_AGE_GROUP)) +
  geom_point(aes(size = num_incidents)) +
  labs(title = "AGE matrix")
AGE_NYPD
```



visualizing races

```
RACE_NYPD <- NYPD_shootings %>%
  select(PERP_RACE, VIC_RACE) %>%
  group_by(PERP_RACE, VIC_RACE) %>%
  tally() %>%
  rename(num_incidents = n) %>%
  ggplot(aes(x = VIC_RACE, y = PERP_RACE)) +
  geom_point(aes(size = num_incidents)) +
  labs(title = "RACE matrix") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
RACE_NYPD
```



analysis

from the matrices above there seems to be a pattern that shows up: in age it apears that young and similiar age people are most likely to shoot each other.

in race it is a little harder to tell as there is no set order I will do test of independence and see which Race has the highest and lowest scores

Model

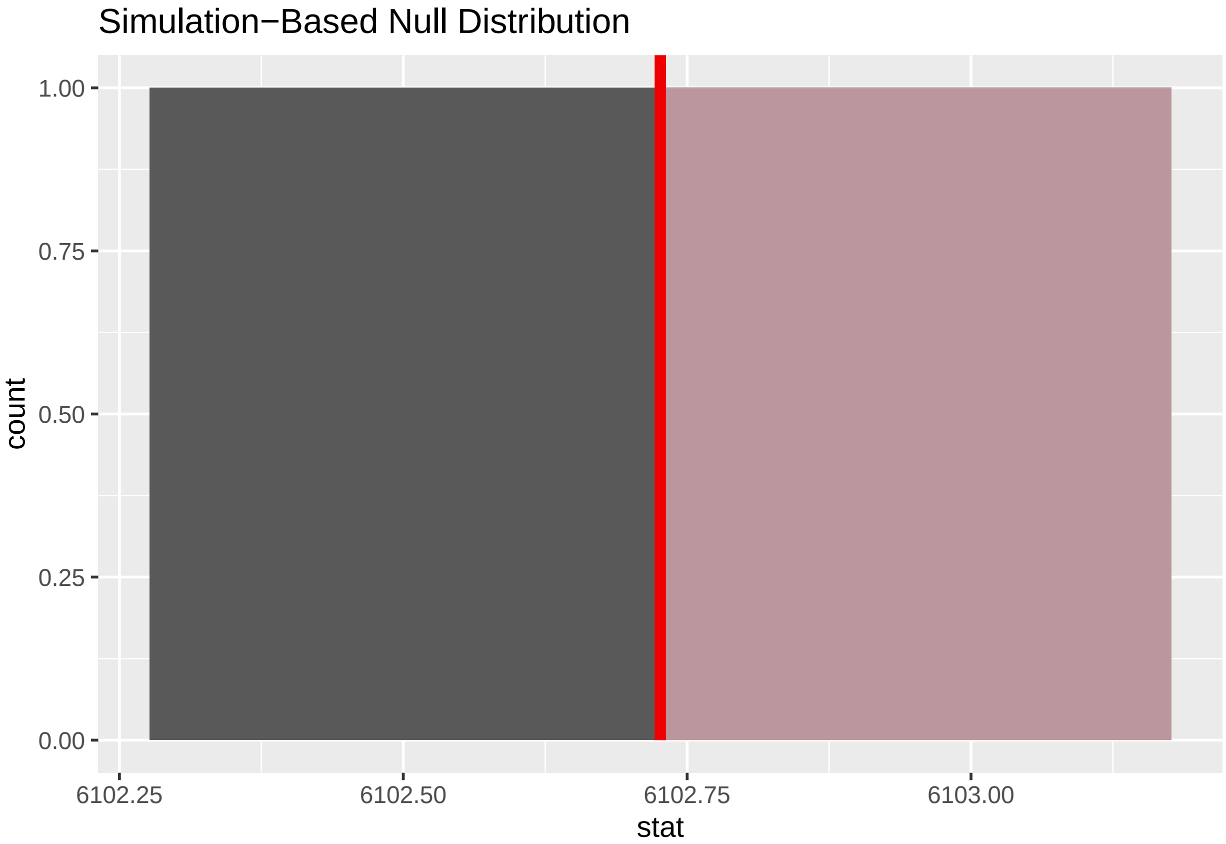
I am attempting to do a chisq independence test but it is not working and do not know why followed instructions on tidymodels

```
model_data <- NYPD_shootings %>%
  select(PERP_RACE, VIC_RACE)


observed_indep_statistic <- model_data %>%
  specify(VIC_RACE ~ PERP_RACE) %>%
  calculate(stat ="Chisq")
```

```
null_distribution_theoretical <- model_data %>%
  specify(VIC_RACE ~ PERP_RACE) %>%
  hypothesize(null = "independence") %>%
  calculate(stat = "Chisq")

null_distribution_theoretical %>%
  visualize(bins = 15, method = "simulation") +
  shade_p_value(observed_indep_statistic,direction = "greater")
```



bias

analysis of this data needs to be dealt with very carefully. I struggle to even bring my self to make any conclusion on the data in front of me other than being put in a disadvantaged position young people will go to great lengths in trying to break out of the system they were born into.

I have little to no information on how this information was collected by whom and how they came to reason which race each person was a part of.