Distribution of HIV/ADIS diagnoses in NYC

Sai Pavani Cheruku

December 2021

PreReq : Loading the libraries required

```
library(ggplot2)
library(tidyverse)
library(sf)
```

1 : Introduction - HIV and AIDs Analysis Project.

The term HIV stands for human immunodeficiency virus. It is an infection that weakens the immune system and make the body more vulnerable to other diseases and infections. Many people consider 'HIV' and 'AIDS' synonymous because these two diseases are associated together. However, HIV and AIDS are not the same thing. HIV disease has 3 stages. The third and the most severe stage is when the patient develop Acquired Immunodeficiency Syndrome (AIDS). It appears when the patient does not receive treatment for the disease and their immune system get severely compromised.

In this project, we'll be studying the infection rate of HIV and AIDS diagnosis in different genders and ethnicity in New York city between 2010 and 2013. We will also study some of the city's efforts to prevent and reduce the spread of this disease.

```
data = read.csv('HIV_AIDS_Diagnoses_by_Neighborhood__Sex__and_Race_Ethnicity.csv')
```

2 : Cleaning the data

The first step in our is to clean our data in order to get better, more accurate results and analysis.

- First, we are changing all columns to their appropriate data types. In this process few '*' values present in the file will be coerced into NA's during this step.

```
data$TOTAL.NUMBER.OF.HIV.DIAGNOSES =
  strtoi(data$TOTAL.NUMBER.OF.HIV.DIAGNOSES)
data$HIV.DIAGNOSES.PER.100.000.POPULATION =
  as.double(data$HIV.DIAGNOSES.PER.100.000.POPULATION)
```

```
## Warning: NAs introduced by coercion
```

```
data$TOTAL.NUMBER.OF.CONCURRENT.HIV.AIDS.DIAGNOSE =
  strtoi(data$TOTAL.NUMBER.OF.CONCURRENT.HIV.AIDS.DIAGNOSE)
data$PROPORTION.OF.CONCURRENT.HIV.AIDS.DIAGNOSES.AMONG.ALL.HIV.DIAGNOSES =
  strtoi(data$PROPORTION.OF.CONCURRENT.HIV.AIDS.DIAGNOSES.AMONG.ALL.HIV.DIAGNOSES)
data$TOTAL.NUMBER.OF.AIDS.DIAGNOSES =
  strtoi(data$TOTAL.NUMBER.OF.AIDS.DIAGNOSES)
data$AIDS.DIAGNOSES.PER.100.000.POPULATION =
  as.double(data$AIDS.DIAGNOSES.PER.100.000.POPULATION)
```

```
## Warning: NAs introduced by coercion
```

```
data$TOTAL.NUMBER.OF.CONCURRENT.HIV.AIDS.DIAGNOSES =
  as.double(data$TOTAL.NUMBER.OF.CONCURRENT.HIV.AIDS.DIAGNOSES)
```

```
## Warning: NAs introduced by coercion
```

- Next, we are filtering the NA's from Total number of HIV diagnoses,race ethnicity, SEX and Total number of AIDS diagnoses.

```
#Fiter NA's from Total number of HIV diagnoses:
data = data %>%
  filter(!is.na(TOTAL.NUMBER.OF.HIV.DIAGNOSES))

#Filter NA from race ethnicity:
data = data %>%
  filter(!is.na(RACE.ETHNICITY))

#Filter NA from SEX:
data = data %>%
  filter(!is.na(SEX))

#For AIDS diagnosis filter NA's from Total number of AIDS diagnoses:
data = data %>%
  filter(!is.na(TOTAL.NUMBER.OF.AIDS.DIAGNOSES))
```

Analysis

After cleaning our data, It is ready for us to begin our analysis.

3 : Comparing the average diagnosis of HIV and AIDS in NYC neighberhoods

Looking at the different neighborhoods provided in the data set, we were curious about which neighborhood has the highest average of HIV patients and AIDS patients.

```
N = data %>%
  group_by(Neighborhood..U.H.F.)%>%
  summarise(Average = mean(TOTAL.NUMBER.OF.HIV.DIAGNOSES, na.rm= TRUE))
head(arrange(N, desc(Average)), 3)
```

```
## # A tibble: 3 x 2
##   Neighborhood..U.H.F.                Average
##   <chr>                                 <dbl>
## 1 All                                  3100
## 2 Unknown                                54.6
## 3 Bedford Stuyvesant - Crown Heights     36.4
```

This Bedford Stuyvesant - Crown Heights neighborhood has the largest average of HIV patients in NYC. 22 people were diagnosed with HIV between 2010 and 2013.

```
Aids = data %>%
  group_by(Neighborhood..U.H.F.)%>%
  summarise(AvG = mean(TOTAL.NUMBER.OF.AIDS.DIAGNOSES, na.rm= TRUE))
head(arrange(Aids, desc(AvG)), 3)
```

```
## # A tibble: 3 x 2
##   Neighborhood..U.H.F.                  AvG
##   <chr>                               <dbl>
## 1 All                                2181.
## 2 Unknown                              38.5
## 3 Bedford Stuyvesant - Crown Heights   30.4
```

- Bedford Stuyvesant - Crown Heights is also the neighborhood who has the largest average of patients diagnosed with AIDS. From this neighborhood, the average people diagnosed with AIDS is 17.
- The average people diagnosed with HIV is 22 and the average people diagnosed with AIDS is 17. This can be explained by the fact that not everyone has HIV can get AIDS but whoever has aids definitely has HIV.
- There are a number of reasons that makes the average patients of both HIV and AIDS high in this area -
  - It can because the patients were afraid from the stigma of having HIV and AIDS so he/she decided to stay quiet about it, instead of getting the treatment they need and warn the people who they had unprotected contact with.
  - This situation explains the wide spread of HIV in this area and the high average of patients who had AIDS too.
  - Another reason could be that the patients in this area did not have access to the proper medication to treat HIV which lead their immune system getting badly damaged and get AIDS.

4 : Visual representation of the distributions of male and female who are diagnosed with HIV and with AIDS

```
BySexHIV = data %>%
  group_by(SEX) %>%
  summarise(Total = sum(TOTAL.NUMBER.OF.HIV.DIAGNOSES))
```

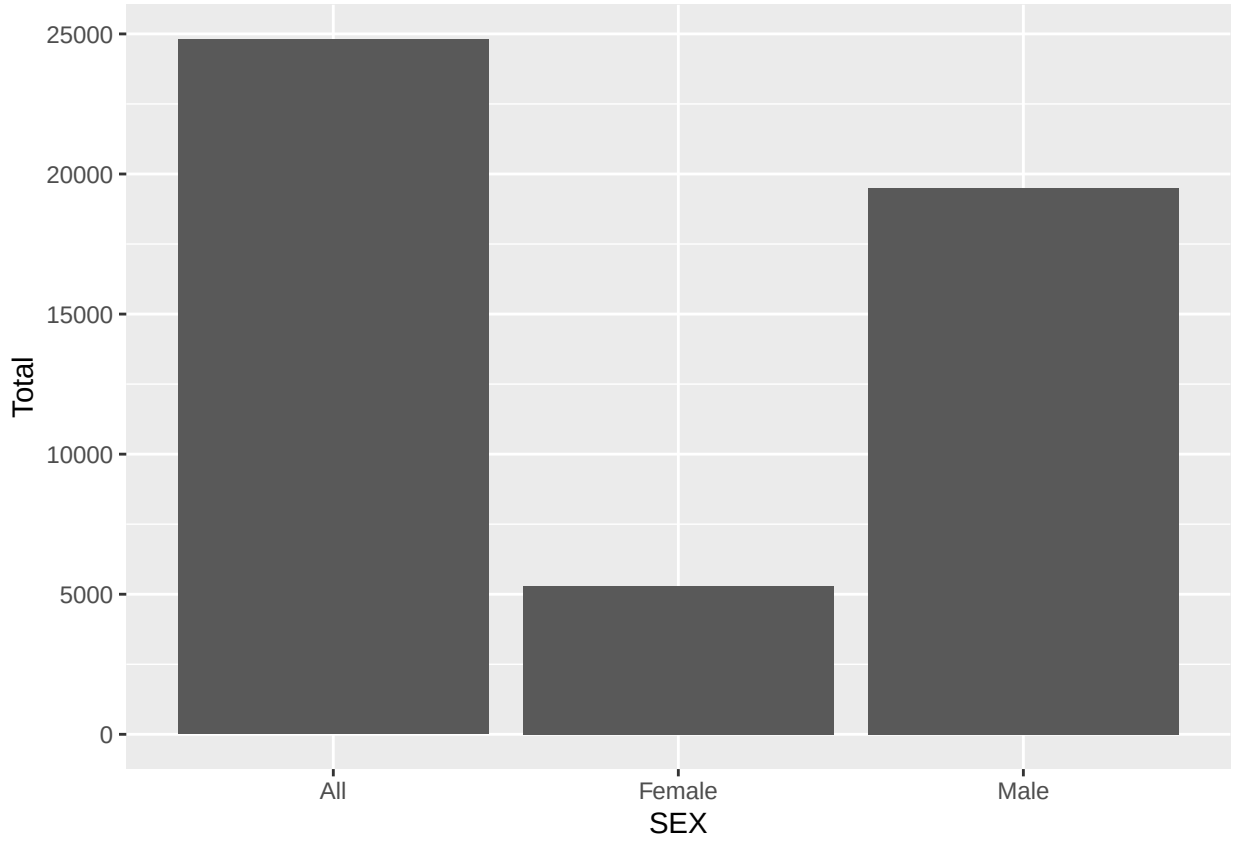
In this step, we created a data set that is grouped by the gender and gives the total number of HIV diagnoses for each gender.

```
BySexAIDS = data %>%
  group_by(SEX) %>%
  summarise(Total = sum(TOTAL.NUMBER.OF.AIDS.DIAGNOSES))
```

In this step, we created a dataset that is grouped by the gender and gives the total number of AIDS diagnoses for male and female.

(i) HIV diagnoses by Gender

```
BySexHIV %>%
  ggplot(aes(x = SEX, y = Total)) +
  geom_bar(stat="identity")
```

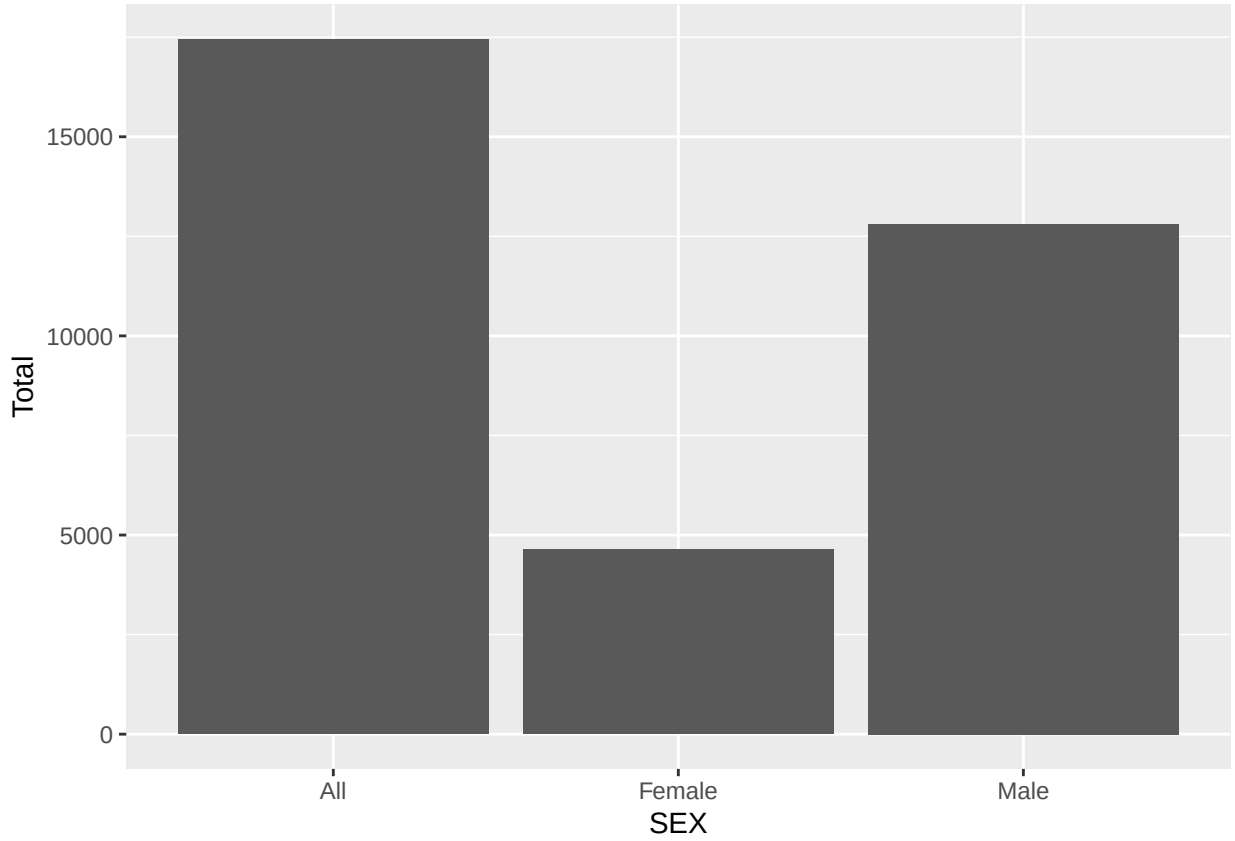


- This graph represents the total number of people who have HIV under Male and Female categories.
- There are about 5500 females who have HIV and about 22500 males who have HIV.
- The difference between the number of male who have HIV and Female is significant.
  - This could be explained by an assumption that females tend to be more careful when it comes to taking the measure to be protected from this disease.
  - Another hypothesis could be that males are more aware than females about this disease and got themselves tested.

- The ALL gender is not defined in the data set. It can mean the total number of LGBT+ but since this info is not provided, we will be ignoring this gender in our analysis

(ii) AIDS diagnoses by Gender

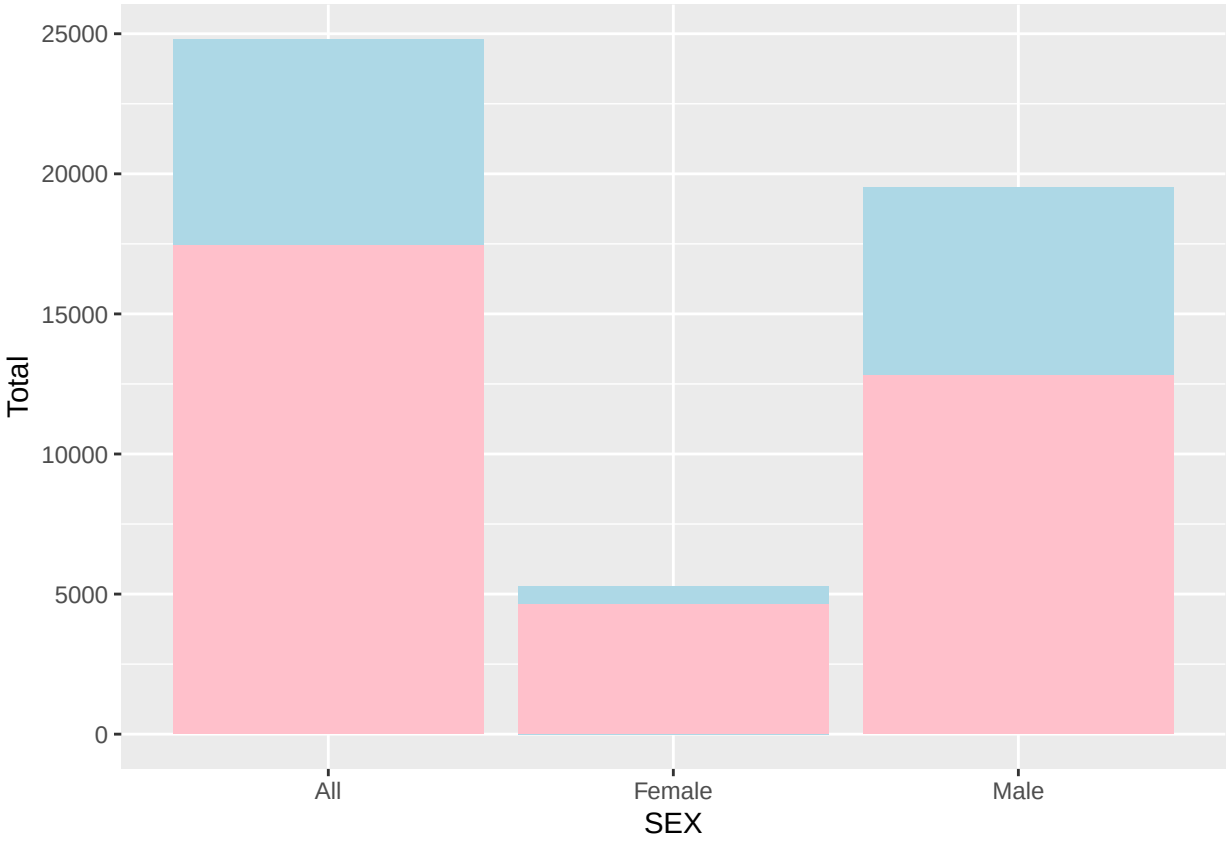
```
BySexAIDS %>%
  ggplot(aes(x = SEX, y = Total)) +
  geom_bar(stat="identity")
```



- This graph represents the total number of people who have AIDS for Male and Female.
- In line with our previous analysis result, the total number of males who have AIDS is higher than the total number of female who has AIDS.

(iii) Overlayed bar graphs of AIDs and HIV diagnoses by Gender

```
BySexHIV$TotalAIDS = BySexAIDS$Total

ggplot(data=BySexHIV, aes(x=SEX)) +
  geom_bar(aes(y=Total), stat="identity", fill='lightblue') +
  geom_bar(aes(y=TotalAIDS), stat="identity", fill='pink')
```



- This graph represents the total number of people who have HIV, AIDS plotted by Gender.
- Overall, the total diagnosis of AIDS disease for each gender is low comparing to the total diagnosis of HIV disease. This shows that if an HIV patient receive the treatment, the help and medication he/she needs, he will likely not develop symptom for AIDS.
- According to the graph, the number of females who have AIDS is very low comparing to the number of female who have HIV. This is a good sign because it means that they are treating their disease correctly.
- However, the number of males who have AIDS is relatively high comparing it with the number of males who have HIV. This means that our earlier hypothesis that males are more aware of the disease and its symptoms seems wrong. Males seem to be not giving this disease the attention it needs.

5 : Visual representation of the distributions of population ethnicities and their HIV, AIDS diagnoses.

```
ByEthnicity_HIV = data %>%
  group_by(RACE.ETHNICITY) %>%
  summarise(Total = sum(TOTAL.NUMBER.OF.HIV.DIAGNOSES))
```

In this step, we created a data set that is grouped by the ethnicity and gives the total number of HIV diagnoses for different ethnicity.

```
ByEthnicity_AIDS = data %>%
  group_by(RACE.ETHNICITY) %>%
  summarise(Total = sum(TOTAL.NUMBER.OF.AIDS.DIAGNOSES))
```

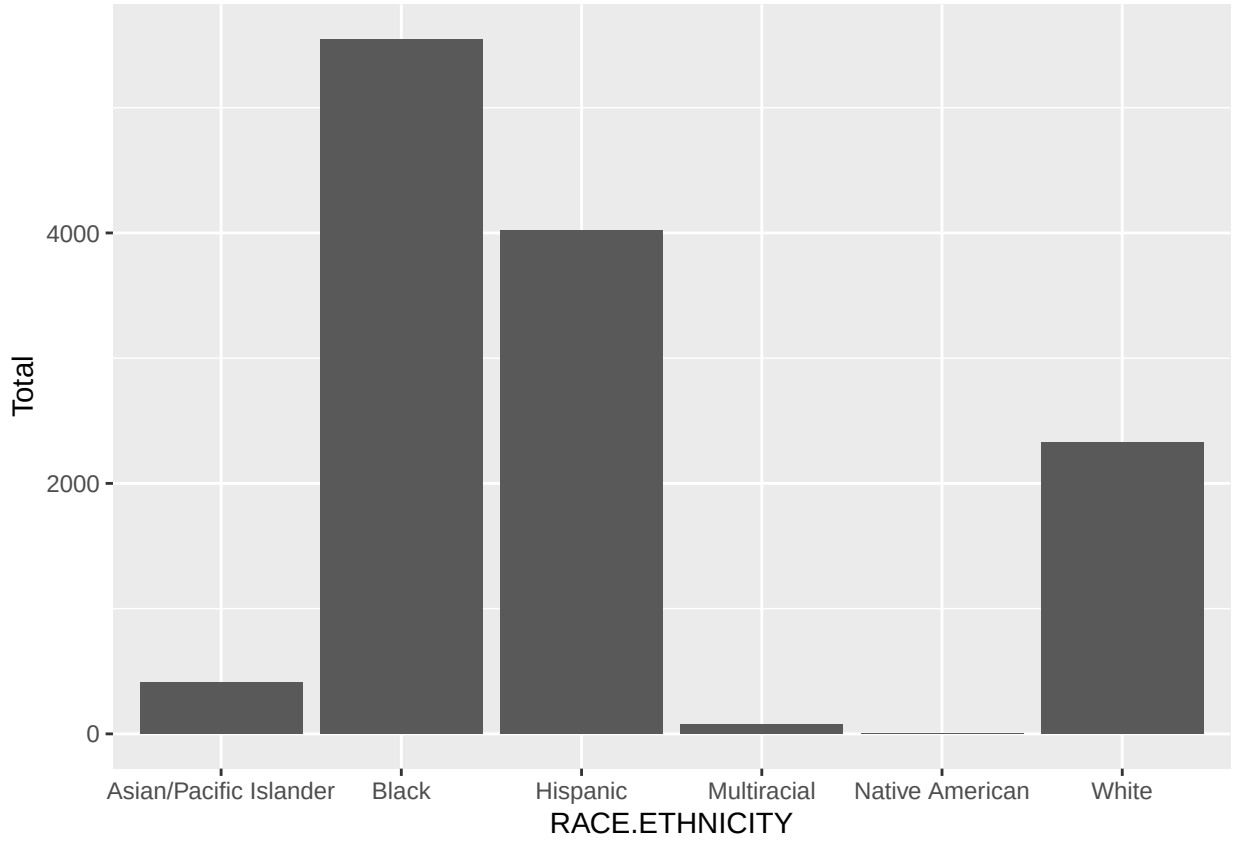
In this step, we created a data set that is grouped by the ethnicity and gives the total number of AIDS diagnoses for different ethnicity.

```
ByEthnicity_HIV_filter = ByEthnicity_HIV[c(2,3,4,5,6,8),c("RACE.ETHNICITY","Total")]

ByEthnicity_AIDS_filter= ByEthnicity_AIDS[c(2,3,4,5,6,8),c("RACE.ETHNICITY","Total")]
```

The Unknown ethnicity is actually a null value, so we deleted it to have a better graph for our analysis in both the ByEthnicity_HIV and ByEthnicity_AIDS data sets. We also deleted the ALL ethnicity because its meaning was not specified.

(i) HIV diagnoses by Ethnicity

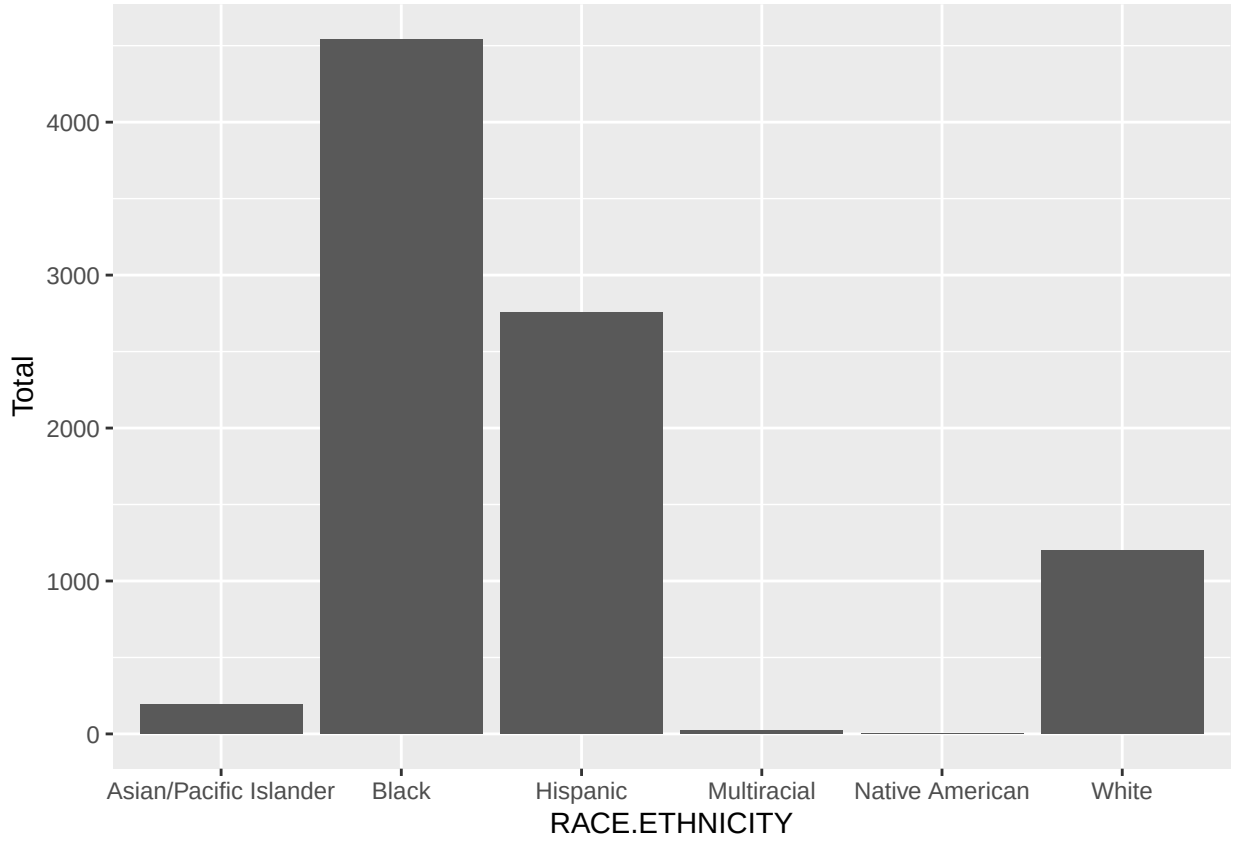
```
ByEthnicity_HIV_filter %>%
  ggplot(aes(x = RACE.ETHNICITY, y = Total)) +
  geom_bar(stat="identity")
```



- This graph represent the total patients who got diagnosed with HIV by ethnicity.
- The Black community has the highest total number of HIV patients.
- Native Americans have the lowest total of people diagnosed with HIV (There is exactly four Native American who have HIV according to the data set). This very low number of people could be due to limited access to testing in this community.

(ii) AIDS diagnoses by Ethnicity

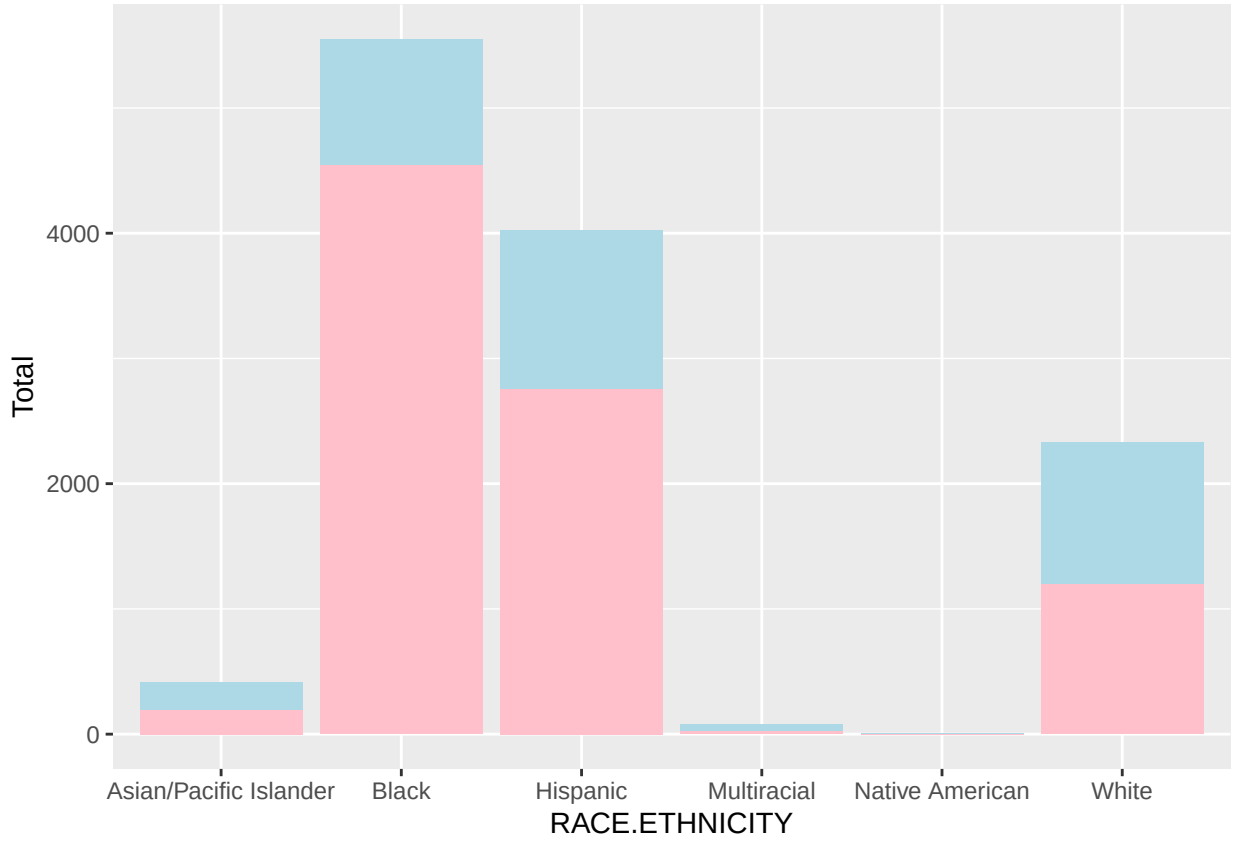
```
ByEthnicity_AIDS_filter %>%
  ggplot(aes(x = RACE.ETHNICITY, y = Total)) +
  geom_bar(stat="identity")
```



- This graph represent the total patients who got tested for AIDS by ethnicity.
- The Black community has also the majority of total patients who tested positive for HIV and the Native American community has the lowest total of AIDS patients.

(iii) Overlayed bar graphs of AIDs and HIV diagnoses by Ethnicity

```
ByEthnicity_HIV$TotalAIDS = ByEthnicity_AIDS$Total

ByEthnicity_HIV_filt = ByEthnicity_HIV[c(2,3,4,5,6,8),c("RACE.ETHNICITY","Total","TotalAIDS")]

ggplot(data=ByEthnicity_HIV_filt, aes(x=RACE.ETHNICITY)) +
  geom_bar(aes(y=Total), stat="identity", fill='lightblue') +
  geom_bar(aes(y=TotalAIDS), stat="identity", fill='pink')
```



- Using this graph, we can compare the number of HIV patients and AIDS patients for each ethnicity.
- Overall, the total number of AIDS patients is low when compared to the total number of HIV patients, which means that most patients are getting their disease treated.
- The black community is the community who has the less number of AIDS patients comparing to the number of HIV patients it has and the white community has the highest total number of AIDS comparing it to the total number of its HIV patients.

6 : NYC Maps - Visually representation of the safer sex product availability across NYC

IN order to prevent the spread of HIV disease and reduce the chance of patients with HIV to develop AIDS as well, New York city has worked on several solutions. One of them is to provide free safer sex products to all the neighborhoods in the city.

```
solution_map = read.csv('NYC_Condom_Availability_Program_-_HIV_condom_distribution_locations.csv')
```

The solution_map data set is a list of over 325 venues, across the five boroughs for New York city that actively distribute free safer sex products.

```
loc = "DataSet/ZIP_CODE_040114"
NYCmap = st_read(dsn = loc, layer = "ZIP_CODE_040114")
```

```
## Reading layer 'ZIP_CODE_040114' from data source
```

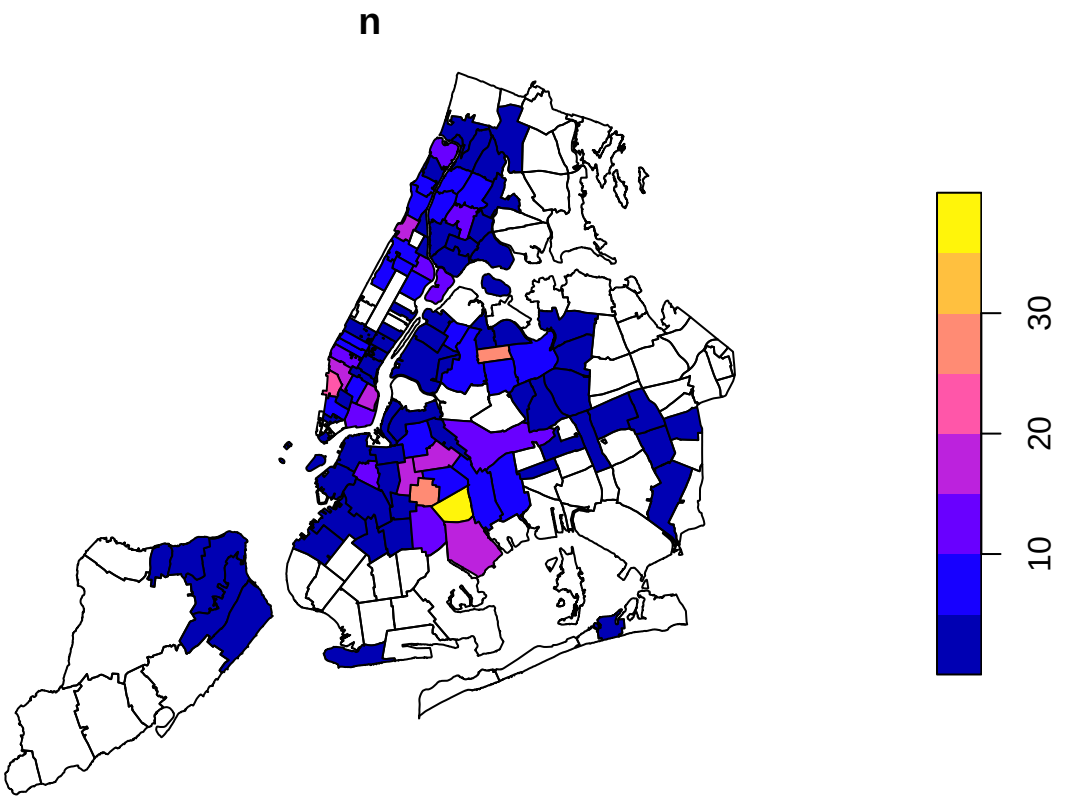
```
##   'C:\Users\pavan\Desktop\Baruch\Baruch_College_Masters_in_Business_Analytics\Semester_1\OPR9750_Soft
##   using driver 'ESRI Shapefile'
## Simple feature collection with 263 features and 12 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 913129 ymin: 120020.9 xmax: 1067494 ymax: 272710.9
## Projected CRS: NAD83 / New York Long Island (ftUS)
```

```
plot(st_geometry(NYCmap))
```

```
serviceCount = solution_map %>%
    group_by(Zipcode) %>%
    count()

serviceCount = serviceCount %>%
    rename(ZIPCODE = Zipcode)

NYCmap$ZIPCODE =
  strtoi(NYCmap$ZIPCODE)

NYCZipServiceCount = left_join(x = NYCmap, y=serviceCount)
```

```
## Joining, by = "ZIPCODE"
```

```
plot(NYCZipServiceCount["n"])
```



- This Map represents the venues that distribute free safer sex products.
- Venues are located in all 5 boroughs. However, they are not located in all neighborhoods.
- In fact, in Staten Island, Bronx and Queens the venues are only located in one part of the borough. For example, on Staten Island, we see that the venues are located only the north of the island. There isn't any the other venues in the other areas which makes hard to people to have access to these products.
- The distribution of free safer sex products that can prevent the transmission of the HIV is not sufficient in those borough.
- In Manhattan and Brooklyn, the venues' locations are better and most people who live in those borough can get the help that is offered.
- So the hypothesis we made at the beginning of our project about the people in Bedford Stuyvesant - Crown Heights is not correct. This neighborhood has the opportunity to get help and for free.

Conclusion:

To conclude, between 2010 and 2013, max proportion of people who were tested positive for HIV are male. The black community has the highest people diagnosed with HIV and Bedford Stuyvesant - Crown Heights is the neighborhood that has the most HIV and AIDS patients.