Project Proposal

Due November 17 at 11:59pm

ALL group member names here

Load Packages

```r
library(tidyverse)
library(dplyr)
library(ggplot2)

rm(list = ls())
```

#add dataset 1 HERE

Dataset 2

Data source:

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Brief description:

This data set reflects incidents of crime in the City of Los Angeles dating from 2020 to 2023. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data.

This code book describes the data in more depth: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

Long description

The City of Los Angeles provides a crime dataset from 2020 to the 2023, covering incidents reported to the Los Angeles Police Department (LAPD). Here's a breakdown of key columns in the dataset:

1. DR_NO: Unique identifier assigned to each crime report, used to track individual cases.

2. Date Rptd: Date the incident was officially reported to the police.

3. DATE OCC: Date when the crime actually occurred.

4. TIME OCC: Time the incident occurred, which allows for time-of-day analysis.

5. AREA: Code representing the geographical area of Los Angeles where the incident took place.

6. AREA NAME: Name of the area corresponding to the AREA code, providing a more human-readable location.

7. Rpt Dist No: Reporting district number within the LAPD, which is a more specific geographical indicator within the area.

8. Part 1-2: Crime classification indicator distinguishing between Part 1 and Part 2 crimes, which helps in crime severity categorization.

9. Crm Cd: Crime code representing a specific type of crime.

10. Crm Cd Desc: Description of the crime type (e.g., robbery, assault), giving context to the "Crm Cd" column.

11. Mocodes: Modus operandi codes that describe the method or behavior pattern of the suspect.

12. Vict Age: Age of the victim, which can be used for demographic analysis.

13. Vict Sex: Gender of the victim, another demographic detail.

14. Vict Descent: Ethnic background or descent of the victim.

15. Premis Cd: Premise code indicating the type of location where the crime occurred.

16. Premis Desc: Description of the premise (e.g., street, residence), helping to understand crime locations.

17. Weapon Used Cd: Code indicating if a weapon was used in the crime, which can be used to assess weapon involvement trends.

18. Weapon Desc: Description of the weapon, if applicable, providing details on the weapon type.

19. Status: Code indicating the current status of the case (e.g., open, closed).

20. Status Desc: Description of the case status, complementing the “Status” code with a text explanation.

21. Crm Cd 1-4: Additional crime codes, capturing cases where multiple types of crimes were involved in a single incident. 22. LOCATION: General location description of the crime.

23. Cross Street: Cross street information for more precise location data. 24. LAT: Latitude coordinate of the crime location, useful for mapping. 25. LON: Longitude coordinate of the crime location, also useful for mapping.

This dataset allows for comprehensive analysis of crime trends in Los Angeles, with potential insights into crime types, locations, times, demographics of victims, weapon involvement, and case status. The dataset is valuable for identifying patterns, conducting demographic analysis, and mapping geographical hotspots of crime.

Research question 1:

Research Question: What is the relationship between the severity of reported crimes and their spatiotemporal distribution in L.A? More specifically how do the frequencies of Part 1 and Part 2 violent crimes vary across different geographical areas of Los Angeles and the time of the day between 2020 and 2023?

Outcome Variable: *Part 1-2* (binary variable indicating crime seriousness, with Part 1 crimes generally more serious than Part 2).

Predictor Variables: *LAT* and *LON* (location coordinates), *AREA* (area code), and *TIME OCC* (time of day). Maybe others too.

Inference Goal: This question seeks to determine if more serious crimes are more prevalent in certain areas and at specific times.

Research question 2:

Research Question: How do victim demographics (age, sex, descent) influence the likelihood of being involved in the most common crimes—vehicle theft, simple assault (battery), burglary from vehicle, theft of identity, and felony vandalism—in Los Angeles between 2020 and 2023, and how do these patterns vary across different geographical areas?

Outcome Variable Name: `Crm.Cd.Desc` (Crime Code Description)
Type: Categorical Variable (Nominal)
Description:
The outcome variable for this research question is the **type of crime committed**, specifically focusing on the top five most common crimes in Los Angeles between 2020 and 2023. These crimes are:

1. **Vehicle - Stolen**
2. **Battery - Simple Assault**
3. **Burglary From Vehicle**
4. **Theft of Identity**
5. **Vandalism - Felony ($400 & Over, All Church Vandalisms)**

This variable represents the specific crime associated with each incident report in the dataset. It is a nominal categorical variable because the categories (crime types) are names without an inherent order or ranking.

Predictor Variables: - Vict.Age - Vict.Sex - Vict.Descent - AREA.NAME

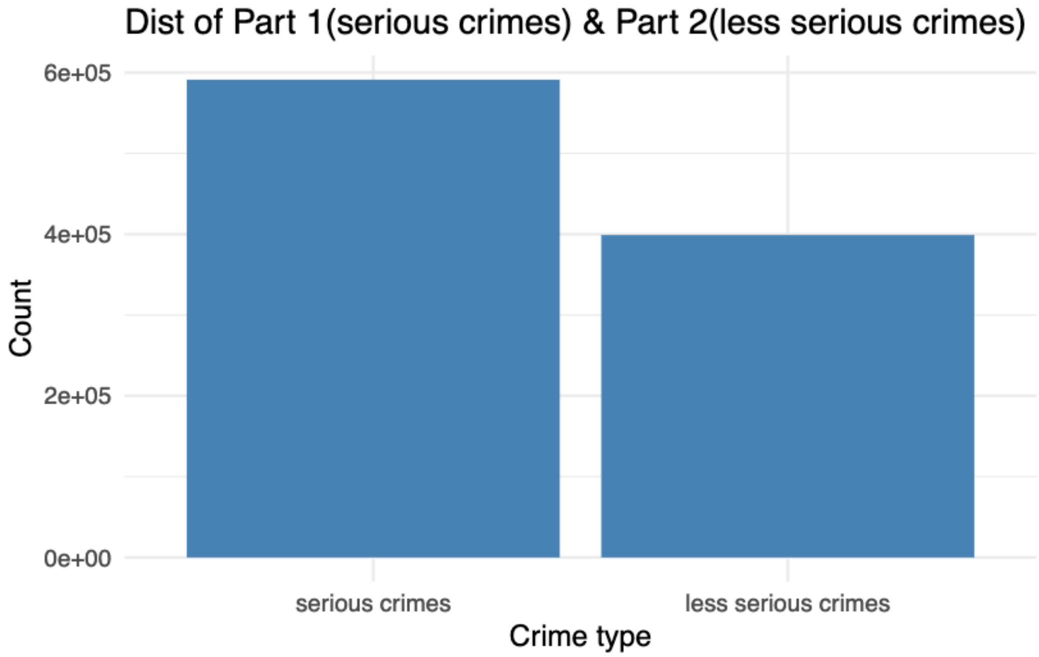
Inference Goal: This question aims to analyze the relationship between victim demographics and the likelihood of being involved in the five most common crimes in Los Angeles. It also seeks to investigate how these patterns differ across various geographical areas within the city. By focusing on these specific crimes and demographic factors, we can identify potential vulnerabilities and trends in victimization across different population groups and locations.

Load the data and provide a `glimpse()`:

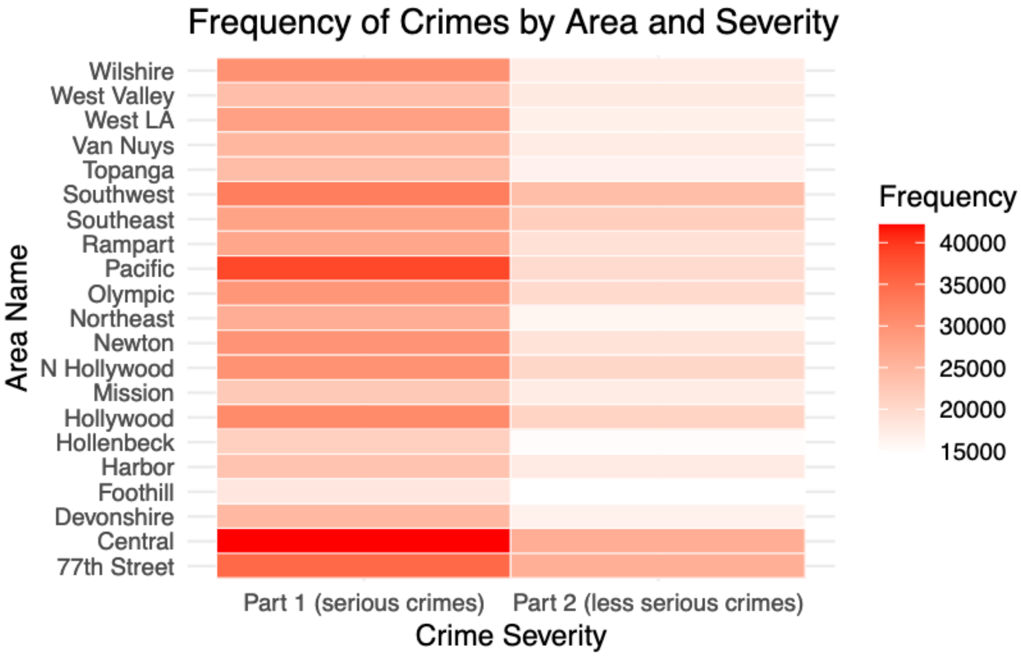
```r
data <- read.csv("~/r\ stuff/Stats_final/Stats_Final_Project/Crime_Data_from_2020_to_Present


glimpse(data)
```

```
Rows: 990,293
Columns: 28
$ DR_NO          <int> 190326475, 200106753, 200320258, 200907217, 220614831, ~
$ Date.Rptd      <chr> "03/01/2020 12:00:00 AM", "02/09/2020 12:00:00 AM", "11~
$ DATE.OCC       <chr> "03/01/2020 12:00:00 AM", "02/08/2020 12:00:00 AM", "11~
$ TIME.OCC       <int> 2130, 1800, 1700, 2037, 1200, 2300, 900, 1110, 1400, 12~
$ AREA           <int> 7, 1, 3, 9, 6, 18, 1, 3, 13, 19, 18, 19, 2, 10, 3, 18, ~
$ AREA.NAME      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
$ Rpt.Dist.No    <int> 784, 182, 356, 964, 666, 1826, 182, 303, 1375, 1974, 18~
$ Part.1.2       <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2~
$ Crm.Cd         <int> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
$ Crm.Cd.Desc    <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
$ Mocodes        <chr> "", "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
$ Vict.Age       <int> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 0, 56,~
$ Vict.Sex       <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
$ Vict.Descent   <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
$ Premis.Cd      <int> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
$ Premis.Desc    <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
$ Weapon.Used.Cd <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
$ Weapon.Desc    <chr> "", "", "", "", "", "", "", "", "", "STRONG-ARM (HANDS,~
$ Status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
$ Status.Desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
$ Crm.Cd.1       <int> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
$ Crm.Cd.2       <int> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
$ Crm.Cd.3       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Crm.Cd.4       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ LOCATION       <chr> "1900 S  LONGWOOD                     AV", "1000 S  FLO~
$ Cross.Street   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "VA~
$ LAT            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
$ LON            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
```

Exploratory Plots: ***Research question 1 plots:***

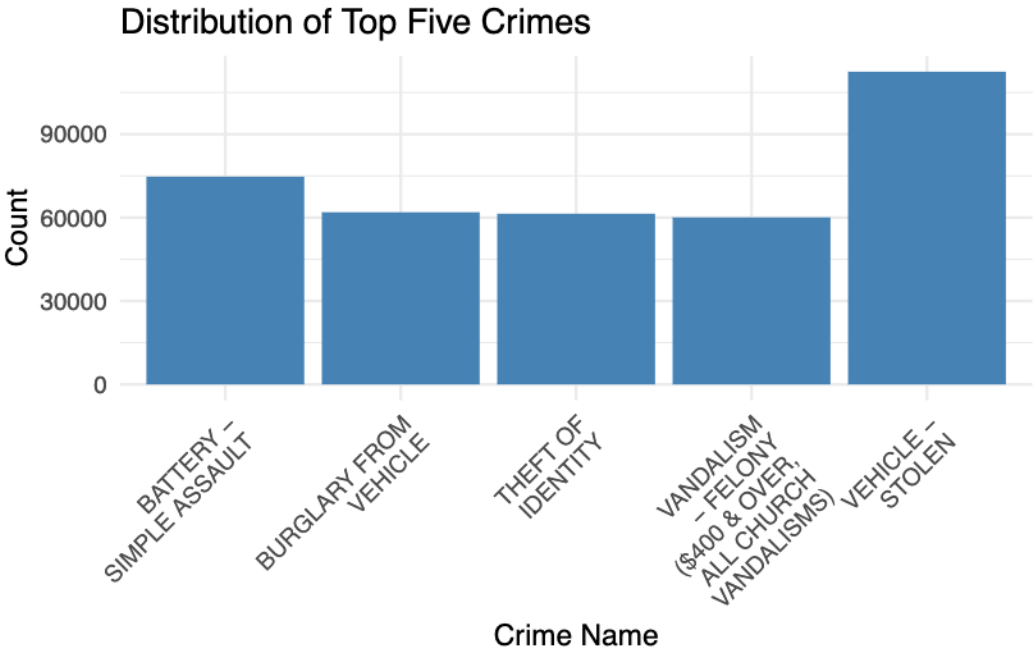


Distribution of Part 1 (Serious Crimes) & Part 2 (Less Serious Crimes) This bar chart categorizes crimes into Part 1 (serious crimes) and Part 2 (less serious crimes) and shows the frequency of each category. The two bars illustrate the relative proportions of serious versus less serious crimes, providing insight into the overall severity distribution within the dataset. EDA Insights: Part 1 crimes have a higher frequency than Part 2 crimes, indicating that serious crimes make up a larger portion of the reported incidents. The more severe crimes have about 591,254 observations and less severe crimes are 399,039. This distribution helps understand the nature of crime severity in Los Angeles, which may influence policing or resource allocation.



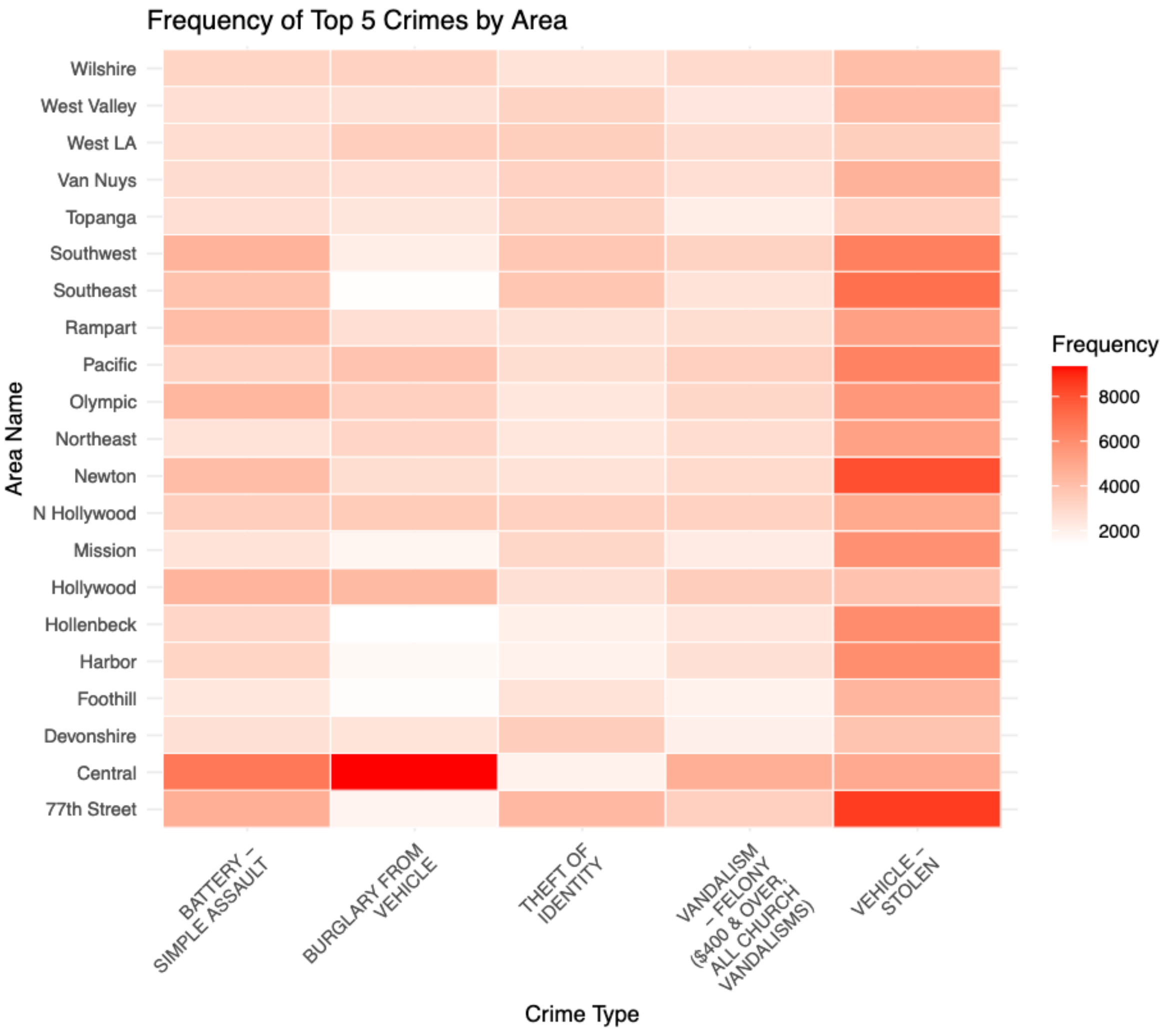
Frequency of Crime Severity(more severe and less severe) Description: This heatmap shows the frequency of the top five crimes across various geographical areas. Each cell's color intensity represents the count of a particular crime type within an area, with darker shades indicating higher frequencies. The y-axis lists the areas, while the x-axis lists the severity of crime(binary: less severe and more severe), allowing for a spatial view of crime distribution. EDA Insights: Areas with darker cells have higher crime counts, highlighting regions with potentially higher crime rates for each type. Certain areas like "Pacific", "Central","77 Street", etc. have a consistently high frequency for more severe crimes, indicating general high-crime zones. Areas with high frequencies for severity of the crime could benefit from targeted crime prevention programs or policing strategies.

Research question 2 plots:

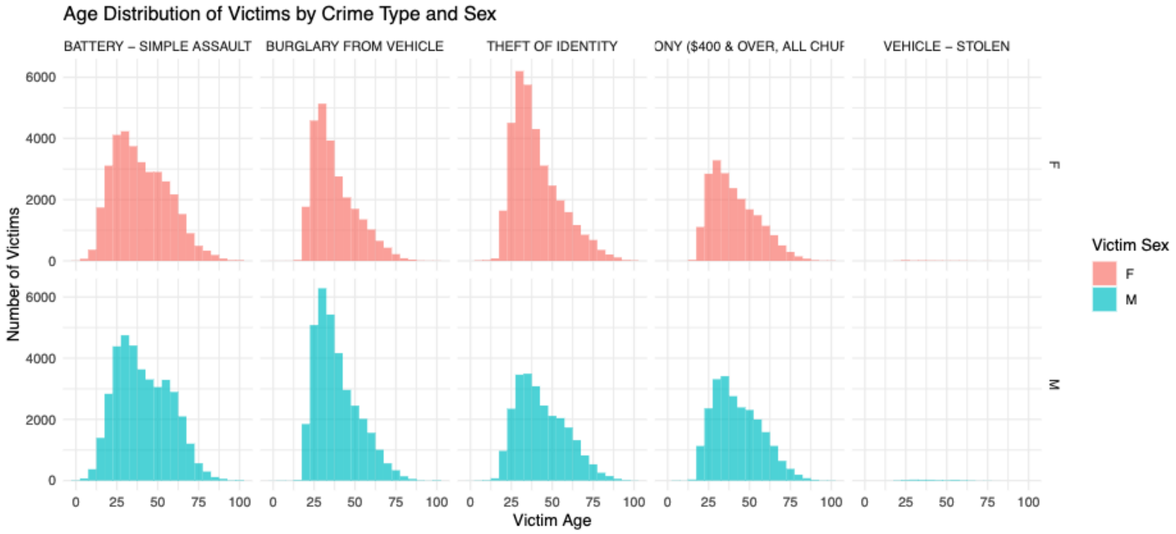


Distribution of Top Five Crimes Description: This bar chart shows the frequency of the top five most common crimes in Los Angeles. Each bar represents a specific crime type, with the height indicating the total count of occurrences. This visualization provides an overview of the relative prevalence of each crime, helping to identify which types are most frequent in the dataset. EDA Insights: The chart clearly shows that "Vehicle - Stolen" has the highest frequency among the top five crimes, followed closely by "Battery - Simple Assault." This distribution allows for quick comparison across crime types, highlighting which crimes are more common. Observing high counts for specific crimes could indicate priority areas for law enforcement or community awareness programs.

```
`summarise()` has grouped output by 'AREA.NAME'. You can override using the
`.groups` argument.
```



Frequency of Top 5 Crimes by Area Description: This heatmap shows the frequency of the top five crimes across various geographical areas. Each cell's color intensity represents the count of a particular crime type within an area, with darker shades indicating higher frequencies. The y-axis lists the areas, while the x-axis lists the crime types, allowing for a spatial view of crime distribution. EDA Insights: Areas with darker cells have higher crime counts, highlighting regions with potentially higher crime rates for each type. Certain areas like "Pacific", "Central","77 Street", etc. have a consistently high frequency for crime types like "Vandalism" and "Central" area has highest " Burglary from Vehicle" crime type cases. Areas with high frequencies for specific crime types could benefit from targeted crime prevention programs or policing strategies.



Age Distribution of Victims by Crime Type and Sex Description: This faceted plot shows the age distribution of victims across different crime types, separated by victim sex (female and male). Each facet represents a unique combination of crime type and sex, with the x-axis showing victim age and the y-axis showing the count. This plot reveals age-based patterns in victimization for each crime type, broken down by gender. EDA Insights: Certain age groups may have higher victimization rates for specific crimes, suggesting patterns of vulnerability related to age. Comparing male and female distributions within each crime type can reveal gender-based differences in victimization, potentially indicating targeted or biased victimization. There is a noticeably higher count of female victims across multiple crime types, particularly in simple assault and identity theft. Male victim counts are generally lower than females for most crimes; however, certain crimes, such as vehicle theft and burglary from vehicles, show a more balanced distribution between genders. This could suggest that these crime types are less influenced by the victim's gender. Across most crimes, young adults appear to have the highest victimization rates, particularly visible in crimes like simple assault and identity theft. This age group might be more exposed to environments or activities associated with these crimes. Different crimes show unique distributions by age and gender, which can guide more focused safety or awareness campaigns for specific demographic groups.