Crime Trends: Victim Characteristics and Crime Feature Analysis

Deepthi Gangiredla

In this section, we will examine the characteristics of crime victims and analyze how these traits align with the cities that have the highest crime rates. By exploring this connection, we hope to better understand the factors contributing to crime in these areas and identify potential patterns.

```
# load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(ranger)

#load data set
crime_data <- read.csv("Crime_Data_from_2020_to_Present.csv")
head(crime_data)
```

```
##       DR_NO              Date.Rptd               DATE.OCC TIME.OCC AREA
## 1 190326475 03/01/2020 12:00:00 AM 03/01/2020 12:00:00 AM     2130    7
## 2 200106753 02/09/2020 12:00:00 AM 02/08/2020 12:00:00 AM     1800    1
## 3 200320258 11/11/2020 12:00:00 AM 11/04/2020 12:00:00 AM     1700    3
## 4 200907217 05/10/2023 12:00:00 AM 03/10/2020 12:00:00 AM     2037    9
## 5 220614831 08/18/2022 12:00:00 AM 08/17/2020 12:00:00 AM     1200    6
## 6 231808869 04/04/2023 12:00:00 AM 12/01/2020 12:00:00 AM     2300   18
##   AREA.NAME Rpt.Dist.No Part.1.2 Crm.Cd
## 1  Wilshire         784        1    510
## 2   Central         182        1    330
## 3 Southwest         356        1    480
## 4  Van Nuys         964        1    343
## 5 Hollywood         666        2    354
## 6 Southeast        1826        2    354
##                            Crm.Cd.Desc             Mocodes Vict.Age
## 1                       VEHICLE - STOLEN                           0
```

```
## 2                    BURGLARY FROM VEHICLE      1822 1402 0344       47
## 3                              BIKE - STOLEN           0344 1251       19
## 4 SHOPLIFTING-GRAND THEFT ($950.01 & OVER)           0325 1501       19
## 5                         THEFT OF IDENTITY 1822 1501 0930 2004       28
## 6                         THEFT OF IDENTITY 1822 0100 0930 0929       41
##   Vict.Sex Vict.Descent Premis.Cd                                Premis.Desc
## 1        M            O       101                                     STREET
## 2        M            O       128            BUS STOP/LAYOVER (ALSO QUERY 124)
## 3        X            X       502 MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)
## 4        M            O       405                                CLOTHING STORE
## 5        M            H       102                                      SIDEWALK
## 6        M            H       501                        SINGLE FAMILY DWELLING
##   Weapon.Used.Cd Weapon.Desc Status  Status.Desc Crm.Cd.1 Crm.Cd.2 Crm.Cd.3
## 1             NA                AA Adult Arrest      510      998       NA
## 2             NA                IC  Invest Cont      330      998       NA
## 3             NA                IC  Invest Cont      480       NA       NA
## 4             NA                IC  Invest Cont      343       NA       NA
## 5             NA                IC  Invest Cont      354       NA       NA
## 6             NA                IC  Invest Cont      354       NA       NA
##   Crm.Cd.4                              LOCATION Cross.Street     LAT
## 1       NA  1900 S  LONGWOOD                  AV              34.0375
## 2       NA  1000 S  FLOWER                   ST              34.0444
## 3       NA  1400 W  37TH                     ST              34.0210
## 4       NA 14000    RIVERSIDE                 DR              34.1576
## 5       NA                         1900    TRANSIENT              34.0944
## 6       NA  9900    COMPTON                   AV              33.9467
##         LON
## 1 -118.3506
## 2 -118.2628
## 3 -118.3002
## 4 -118.4387
## 5 -118.3277
## 6 -118.2463
```

Clean data and extract names of the top 10 cities with the highest crime in the dataset

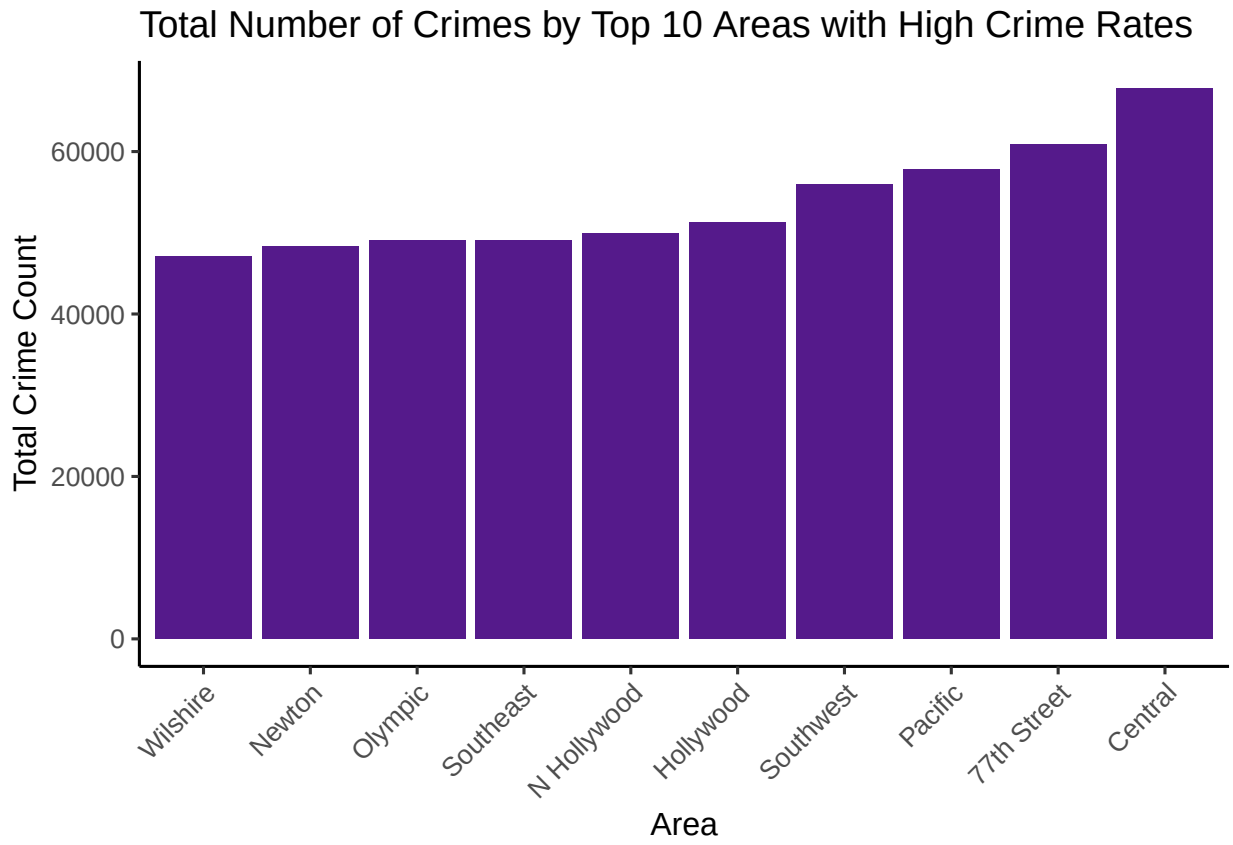
```
# create dictionary for all ethnicity and their codes in the crime data
ethnicity_dict <- c(
  "A" = "Other Asian",
  "B" = "Black",
  "C" = "Chinese",
  "D" = "Cambodian",
  "F" = "Filipino",
  "G" = "Guamanian",
  "H" = "Hispanic/Latin/Mexican",
  "I" = "American Indian/Alaskan Native",
  "J" = "Japanese",
  "K" = "Korean",
  "L" = "Laotian",
  "O" = "Other",
  "P" = "Pacific Islander",
  "S" = "Samoan",
  "U" = "Hawaiian",
  "V" = "Vietnamese",
```

```
  "W" = "White",
  "X" = "Unknown",
  "Z" = "Asian Indian"
)

# Add ethnicity column that maps to ethnicity code
crime_data <- crime_data %>%
  mutate(Ethnicity = ethnicity_dict[Vict.Descent])


#Get count of how many crimes are documented for each area
crime_top<-crime_data %>%
    count(AREA.NAME, sort = TRUE, name= "crime_count") %>%
    arrange(desc(crime_count))


# get table of top 10 areas of crime and number of crimes total per area
top_10_area <- head(crime_top,10)
top_10_area
```

```
##      AREA.NAME crime_count
## 1      Central       67774
## 2  77th Street       60865
## 3      Pacific       57810
## 4    Southwest       55978
## 5    Hollywood       51324
## 6  N Hollywood       49978
## 7    Southeast       49119
## 8      Olympic       49023
## 9       Newton       48268
## 10    Wilshire       47090
```

It appears the Central, 77th Street, Pacific, Southwest, Hollywood, N Hollywood, Southeast, Olympic, Newton, Wilshire area in LoS Angeles have the highest rate of crime.

```
#plot top 10 areas and count

ggplot(top_10_area, aes(x = reorder(AREA.NAME, crime_count), y = crime_count)) +
   geom_bar(stat = "identity", fill = "purple4") +
   labs(x = "Area", y = "Total Crime Count", title = "Total Number of Crimes by Top 10 Areas with High C
   theme_classic(base_size = 12) +
   theme(axis.text.x = element_text(angle = 45, hjust = 1),
         legend.position = "none")
```



In this plot we can visually see the distribution of total crimes across the top 10 area.

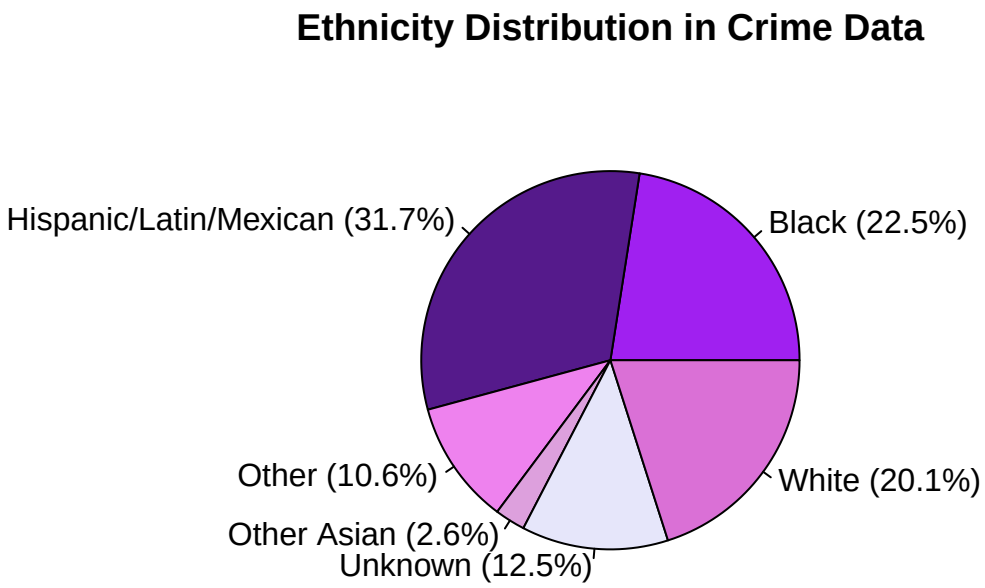
```
# Filter crime table to only include data from top 10 areas with highest crime count
crime_top_10 <-crime_data[crime_data$AREA.NAME %in% c(top_10_area$AREA.NAME), ]
```

VICTIM ETHNICITY

In high crime areas, how is the distribution of crime victims by ethinicities? Is one ethnicity more frequently targeted than the other? To explore these questions, we can use the code below:

```
# Summarize the data by counting the occurrences of each ethnicity
ethnicity_table <- table(crime_top_10$Ethnicity)

# Convert the table into a data frame
ethnicity_df <- data.frame(
  Ethnicity = names(ethnicity_table),
  Count = as.vector(ethnicity_table)
)

# Calculate percentages for each ethnicity group
ethnicity_df$Percentage <- round((ethnicity_df$Count / sum(ethnicity_df$Count)) * 100, 1)

# Group ethnicity with less than 2% into the "Other" category
ethnicity_df$Ethnicity <- ifelse(ethnicity_df$Percentage < 2, "Other", ethnicity_df$Ethnicity)

# Recalculate percentages
ethnicity_df <- ethnicity_df %>%
```

```r
  group_by(Ethnicity) %>%
  summarize(Count = sum(Count)) %>%
  mutate(Percentage = round((Count / sum(Count)) * 100, 1))


# Create a pie chart that shows distribution of victim ethnicity across top 10 high crime rate cites
labels <- paste(ethnicity_df$Ethnicity, " (", ethnicity_df$Percentage, "%)", sep = "")
pie(ethnicity_df$Count, labels = labels,
    col = c("purple", "purple4", "violet", "plum","lavender", "orchid"),
    main = "Ethnicity Distribution in Crime Data")
```



This pie chart shows that in Los Angeles, the majority of crime victims are Hispanic/Latino/Mexican. The next two largest groups are Black and White victims.

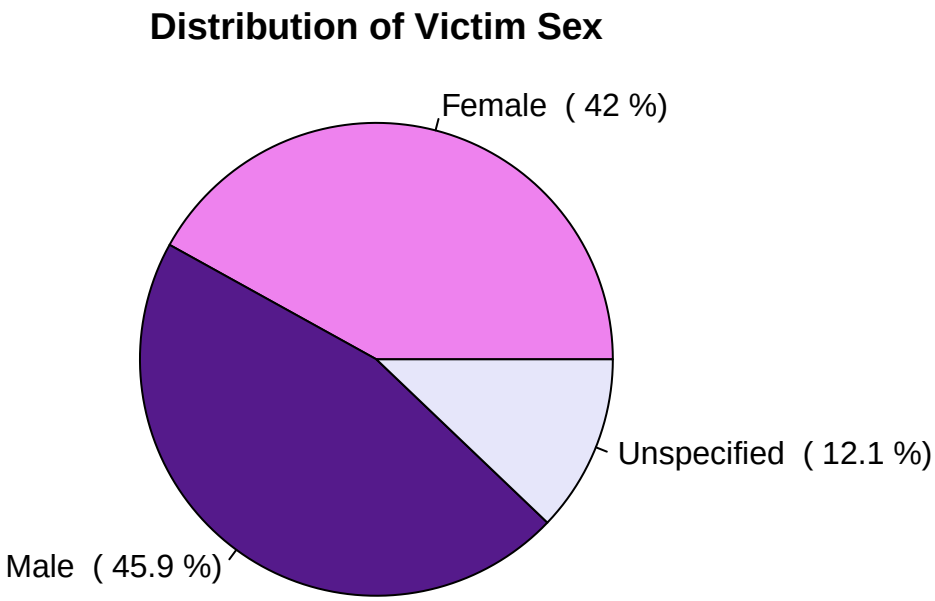
VICTIM SEX In high-crime areas, how is the distribution of crime victims by sex? Is one sex more frequently targeted than the other? To explore these questions, we can use the code below

```r
#table(crime_data[crime_data$AREA.NAME == "Central",])
sex_table <-table(crime_top_10 [, "Vict.Sex"])
unique(crime_top_10$Vict.Sex)
```

```
## [1] "M" "X" "F" ""  "H" "-"
```

```r
collapsed_table_sex <- c(
  Female = sex_table["F"],  # F = Female
  Male = sex_table["M"],    # M = Male
```

```
  Other = sum(sex_table[c("H", "X")] ) )
names(collapsed_table_sex) <- c("Female", "Male","Unspecified")


# create percentage of sex ratios
total_count <- sum(collapsed_table_sex)
percentages <- round((collapsed_table_sex / total_count) * 100, 1)  # Round to 1 decimal place

# Create the pie chart for the sex distribution
pie(collapsed_table_sex,
    main = "Distribution of Victim Sex",
    col = c( "violet","purple4","lavender"),
    labels = paste(names(collapsed_table_sex), " (",percentages,"%)"),
  radius = 1)
```



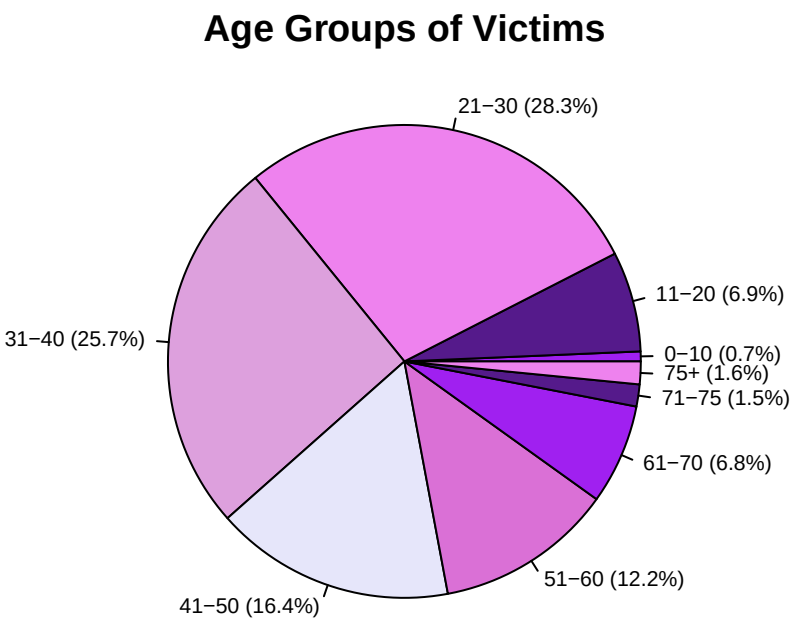
Based on the pie chart above, it appears that the majority of crimes are committed against men, though the difference is relatively small. This indicates that, overall, the sex ratio of crime victims is fairly balanced.

VICTIM AGE

In high-crime areas, how is the distribution of crime victims by age? Is one age group more frequently targeted than the other? To explore these questions, we can use the code below

```
# define age groups
age_groups <- cut(crime_top_10$Vict.Age,
                  breaks = c(0, 10, 20, 30, 40, 50, 60, 70, 75, Inf),
                  right = TRUE,
```

```
                  labels = c("0-10", "11-20", "21-30", "31-40", "41-50", "51-60", "61-70", "71-75", "75-

#frequency table of the age groups
age_group_table <- table(age_groups)

# convert the table into a data frame for ggplot
age_group_df <- data.frame(
  AgeGroup = names(age_group_table),
  Count = as.vector(age_group_table)
)

#add percentages for plot
age_group_df$Percentage <- round((age_group_df$Count / sum(age_group_df$Count)) * 100, 1)

# Create a pie chart that shows distribution of victim age groups across top 10 high crime rate cites
labels <- paste(age_group_df$AgeGroup, " (", age_group_df$Percentage, "%)", sep = "")
pie(age_group_df$Count, labels = labels,
    col = c("purple", "purple4", "violet", "plum","lavender", "orchid"),
    main = "Age Groups of Victims",cex=0.7,radius = 1)
```



Based on the pie chart above, it appears that the majority of crimes are committed against individuals in the 21-30 age group, with the second largest group of victims falling within the 31-40 age range. This suggests that younger adults are more frequently targeted, followed by those in early middle age.

CRIME FEATURE ANALYSIS

Next, we aim to explore which features of crimes in high-crime areas are most strongly correlated with or

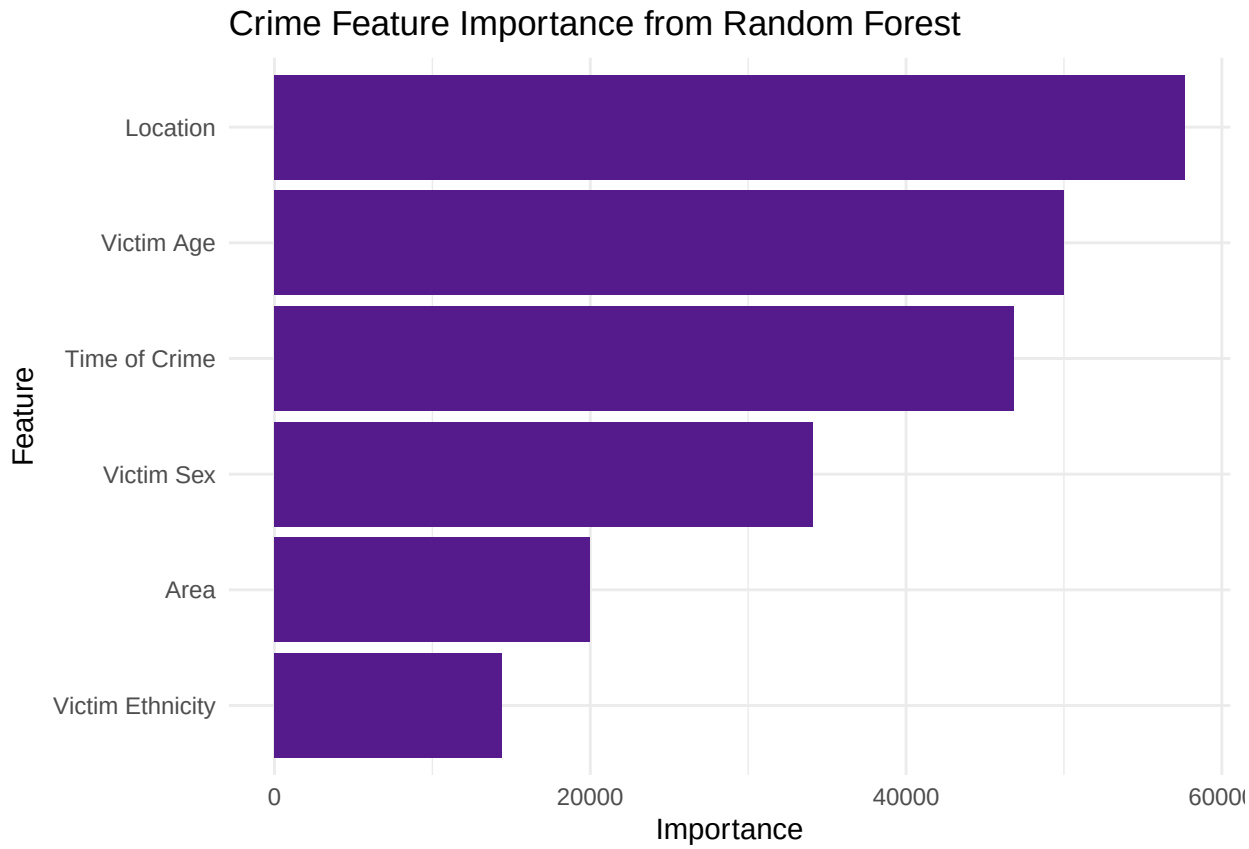
associated with different crime types. Do specific characteristics of the crime or the victim influence the likelihood of a particular crime occurring?"

We address this question by first imputing all missing values, and then using a random forest model to conduct our feature analysis.

```
# imputation  of missing data
# Use the ethnicity group with the highest number of victims to replace missing values
crime_top_10$Ethnicity[is.na(crime_top_10$Ethnicity)] <- names(sort(table(crime_top_10$Ethnicity), decr
# Use the median of all victim ages to replace missing vlaues
crime_top_10$Age[is.na(crime_top_10$Age)] <- median(crime_top_10$Age)

# convert crime type as factor
crime_top_10$Crm.Cd.Desc <- as.factor(crime_top_10$Crm.Cd.Desc)

# Fit a random forest model
rf_model <- ranger(Crm.Cd.Desc ~ AREA + LOCATION+Vict.Age + Vict.Sex+ TIME.OCC+ Ethnicity,
                   data = crime_top_10,
                   importance = 'impurity',
                   num.trees = 500)
```

```
## Growing trees.. Progress: 5%. Estimated remaining time: 10 minutes, 27 seconds.
## Growing trees.. Progress: 12%. Estimated remaining time: 8 minutes, 7 seconds.
## Growing trees.. Progress: 19%. Estimated remaining time: 6 minutes, 53 seconds.
## Growing trees.. Progress: 26%. Estimated remaining time: 6 minutes, 12 seconds.
## Growing trees.. Progress: 32%. Estimated remaining time: 5 minutes, 40 seconds.
## Growing trees.. Progress: 39%. Estimated remaining time: 5 minutes, 1 seconds.
## Growing trees.. Progress: 46%. Estimated remaining time: 4 minutes, 22 seconds.
## Growing trees.. Progress: 53%. Estimated remaining time: 3 minutes, 47 seconds.
## Growing trees.. Progress: 60%. Estimated remaining time: 3 minutes, 12 seconds.
## Growing trees.. Progress: 66%. Estimated remaining time: 2 minutes, 41 seconds.
## Growing trees.. Progress: 73%. Estimated remaining time: 2 minutes, 9 seconds.
## Growing trees.. Progress: 79%. Estimated remaining time: 1 minute, 41 seconds.
## Growing trees.. Progress: 86%. Estimated remaining time: 1 minute, 9 seconds.
## Growing trees.. Progress: 92%. Estimated remaining time: 36 seconds.
## Growing trees.. Progress: 99%. Estimated remaining time: 3 seconds.
```

```
# save feature importance
importance_values <- importance(rf_model)

# put values into a data frame
importance_df <- data.frame(
  Feature = names(importance_values),
  Importance = importance_values)

#format features labels
importance_df$Feature_Name <- c("Area", "Location", "Victim Age", "Victim Sex","Time of Crime" ," Victi

# Plot the all features and their importance to the crime type
ggplot(importance_df, aes(x = reorder(Feature_Name, Importance), y = Importance)) +
  geom_bar(stat = "identity", fill = "purple4") +
  coord_flip() +
  labs(title = "Crime Feature Importance from Random Forest", x = "Feature", y = "Importance") +
  theme_minimal()
```



This plot reveals that the top three features most strongly associated with a crime are the specific location, the victim's age, and the time of day the crime occurs. These factors appear to play an important in determining the likelihood and nature of different crime types.