ML_project

samuel Mweni

2023-11-10

```
library(ggplot2)
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.3.2

library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(caret)

## Loading required package: lattice

wes_cleaned_stop_data <- read.csv("wes_cleaned_stop_data.csv")
```

Data cleaning

```
# Sum of NA values in each column
na_count <- colSums(is.na(wes_cleaned_stop_data))
print(na_count)

##        stop_outcome        ticket_count       warning_count
stop_district
##                   0                   0                   0
624
##   stop_duration_mins     person_searched   property_searched
traffic_involved
##                   0                   0                   0
0
##              gender           ethnicity                 age
```

```
primary_stop_reason
##                  0                   0                2036
0
##         day_of_week       time_category
##                  27                   0
```

handling Negative Duration

```
# Investigating negative values in 'stop_duration_mins'
summary(wes_cleaned_stop_data$stop_duration_mins)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -543.00    5.00   10.00   16.17   16.00  443.00

# Removing rows with negative 'stop_duration_mins'
wes_cleaned_stop_data <-
wes_cleaned_stop_data[wes_cleaned_stop_data$stop_duration_mins >= 0, ]
```

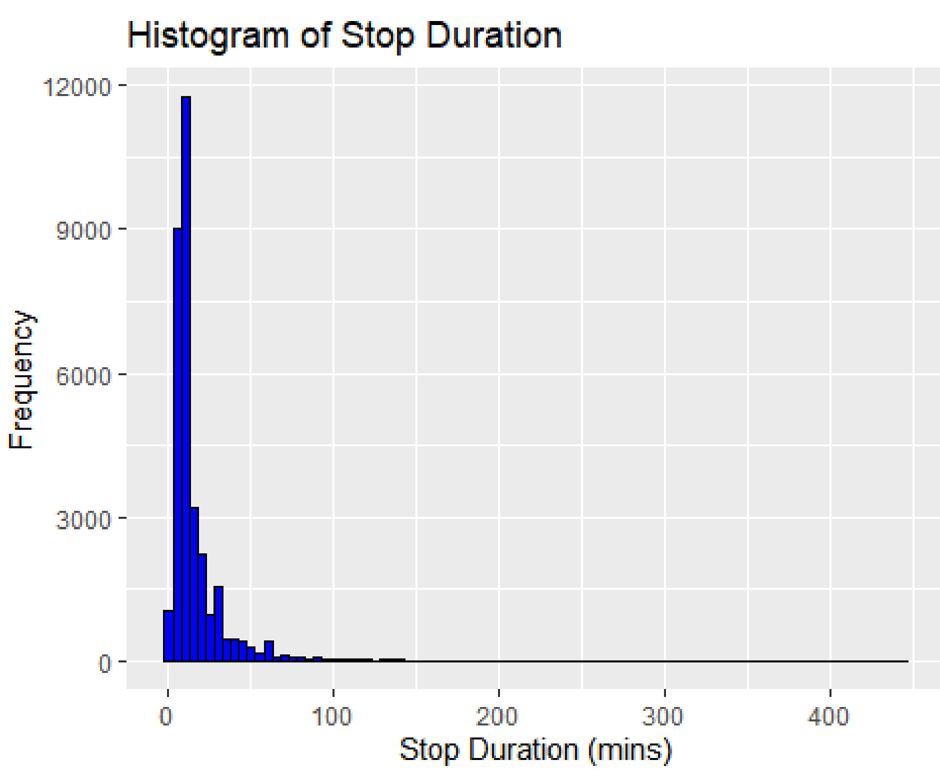
Data Type Conversion for Categorical Variables

```
library(caret)
dummies <- dummyVars(" ~ .", data = wes_cleaned_stop_data)
wes_cleaned_stop_data_transformed <- predict(dummies, newdata =
wes_cleaned_stop_data)

glimpse(wes_cleaned_stop_data)

## Rows: 32,981
## Columns: 14
## $ stop_outcome        <chr> "Arrest", "No Action", "No Action", "Arrest",
"Arr…
## $ ticket_count        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ warning_count       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ stop_district       <chr> "2D", "6D", "4D", "3D", "6D", "5D", "3D",
"3D", "6…
## $ stop_duration_mins  <int> 60, 10, 15, 13, 20, 10, 7, 5, 5, 20, 2, 31,
40, 5,…
## $ person_searched     <int> 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0,
0, 0,…
## $ property_searched   <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ traffic_involved    <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0,…
## $ gender              <chr> "Male", "Male", "Male", "Male", "Male",
"Male", "M…
## $ ethnicity           <chr> "Black", "Other", "Black", "Black", "Black",
"Blac…
## $ age                 <int> NA, 36, 46, 55, 46, 29, 38, 24, 23, 26, 26,
31, 26…
## $ primary_stop_reason <chr> "call for service", "demeanor during a field
```

```
conta…
## $ day_of_week          <chr> "Sun", "Sun", "Sun", "Sun", "Sun", "Sun",
"Sun", "…
## $ time_category        <chr> "Night", "Night", "Night", "Morning", "Night",
"Ni…
```

##2. Exploratory Data Analysis (EDA)

```
# Summary statistics for 'stop_duration_mins'
summary(wes_cleaned_stop_data$stop_duration_mins)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   10.00   16.19   16.00  443.00

# Histogram to see the distribution of 'stop_duration_mins'
library(ggplot2)
ggplot(wes_cleaned_stop_data, aes(x = stop_duration_mins)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Histogram of Stop Duration", x = "Stop Duration (mins)", y =
"Frequency")
```

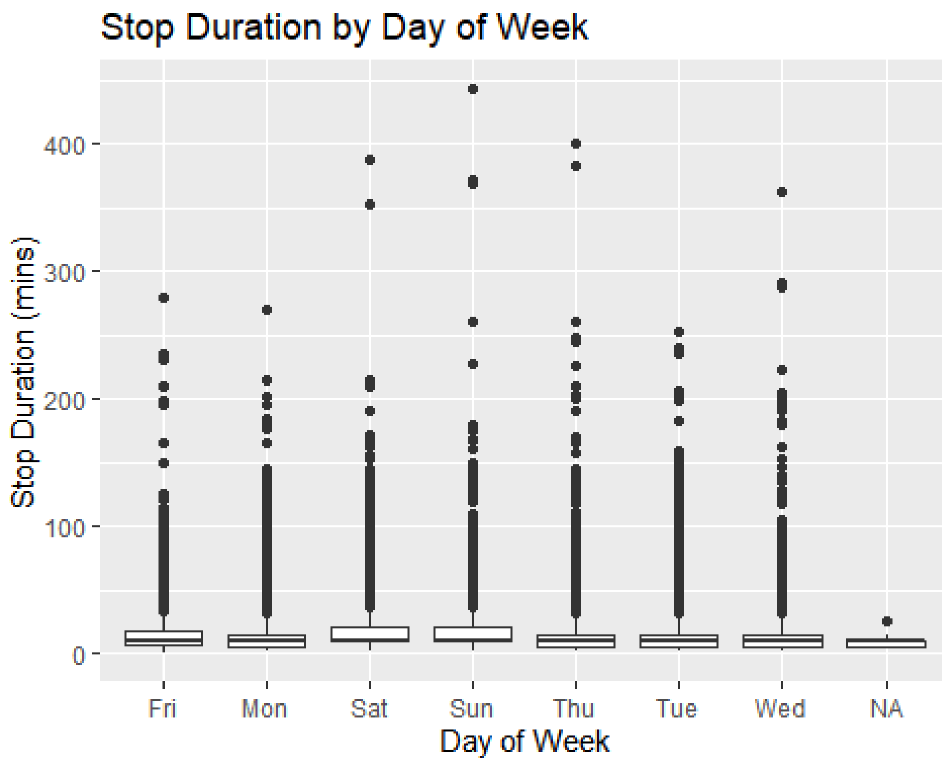


Relationships Between Predictors and Stop Duration

a) Stop Duration by Day of Week

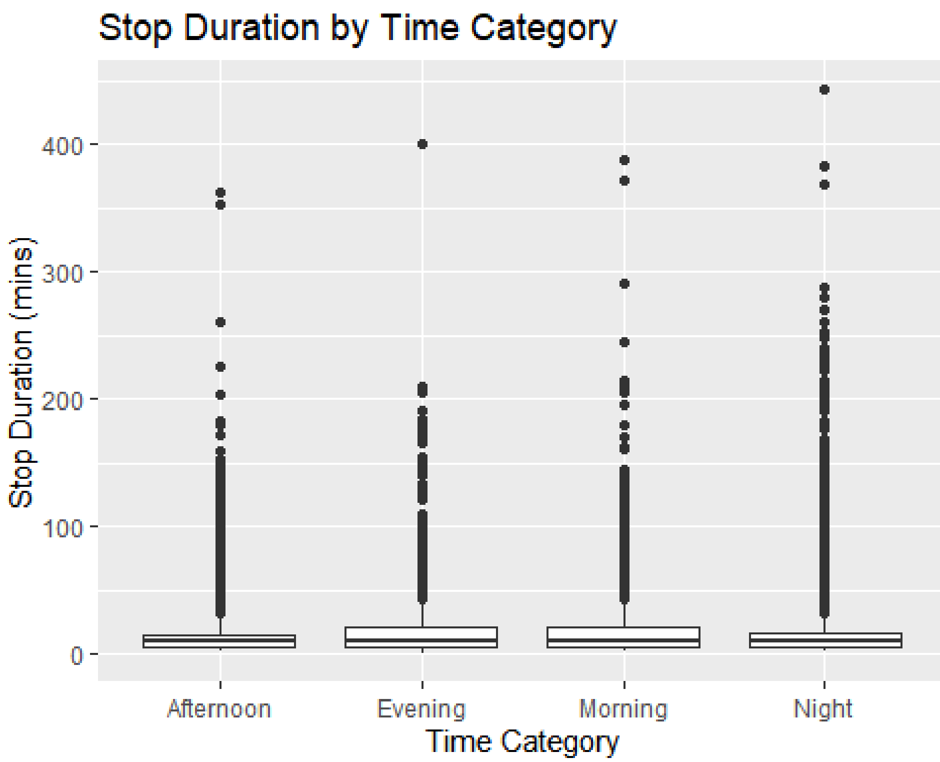
```
ggplot(wes_cleaned_stop_data, aes(x = day_of_week, y = stop_duration_mins)) +
  geom_boxplot() +
```

```
  labs(title = "Stop Duration by Day of Week", x = "Day of Week", y = "Stop Duration (mins)")
```



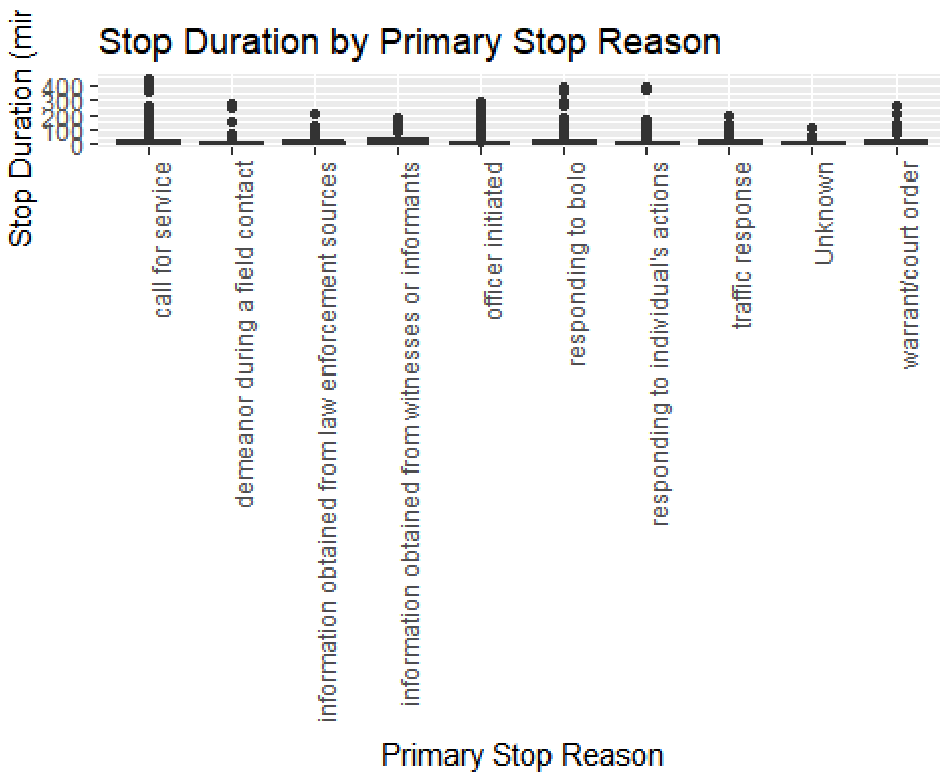
b) Stop Duration by Time Category

```
ggplot(wes_cleaned_stop_data, aes(x = time_category, y = stop_duration_mins))
+
  geom_boxplot() +
  labs(title = "Stop Duration by Time Category", x = "Time Category", y =
"Stop Duration (mins)")
```



c) Stop Duration by Primary Stop Reason

```
ggplot(wes_cleaned_stop_data, aes(x = primary_stop_reason, y =
stop_duration_mins)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Stop Duration by Primary Stop Reason", x = "Primary Stop
Reason", y = "Stop Duration (mins)")
```

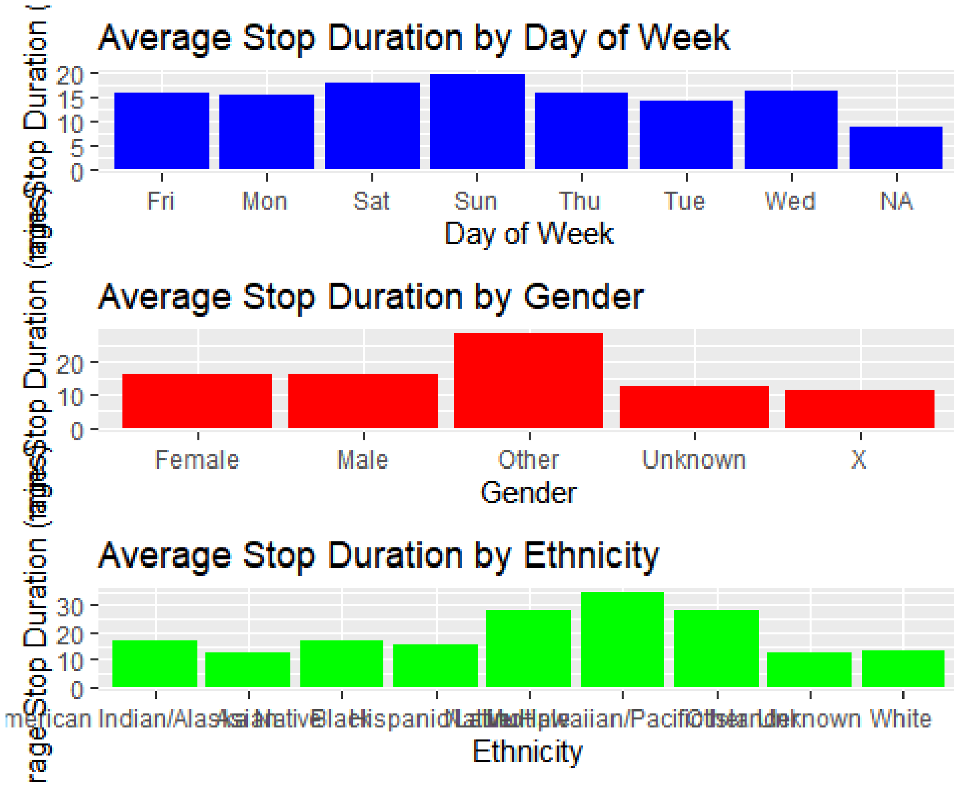


Correlation Analysis for Numerical Variables

```
# Selecting only numerical columns for correlation analysis
numerical_data <- wes_cleaned_stop_data %>%
  select_if(is.numeric)

# Calculating correlation
correlation_matrix <- cor(numerical_data, use = "complete.obs")

## Warning in cor(numerical_data, use = "complete.obs"): the standard
deviation is
## zero

# Displaying the correlation matrix
print(correlation_matrix)

##                    ticket_count warning_count stop_duration_mins
## ticket_count                  1            NA                 NA
## warning_count                NA             1                 NA
## stop_duration_mins           NA            NA         1.00000000
## person_searched              NA            NA         0.17056107
## property_searched            NA            NA         0.09608616
## traffic_involved             NA            NA        -0.32211716
## age                          NA            NA        -0.10829688
##                    person_searched property_searched traffic_involved
## ticket_count                    NA                NA               NA
## warning_count                   NA                NA               NA
## stop_duration_mins       0.1705611        0.09608616       -0.3221172
```

```
## person_searched          1.0000000       0.29094195       -0.3494251
## property_searched        0.2909419       1.00000000       -0.1660923
## traffic_involved        -0.3494251      -0.16609231        1.0000000
## age                     -0.1468781      -0.05139838        0.2480415
##                             age
## ticket_count                 NA
## warning_count                NA
## stop_duration_mins -0.10829688
## person_searched    -0.14687815
## property_searched  -0.05139838
## traffic_involved    0.24804146
## age                 1.00000000
```

```r
# Plot for 'day_of_week'
plot1 <- ggplot(wes_cleaned_stop_data, aes(x = day_of_week, y =
stop_duration_mins)) +
  geom_bar(stat = "summary", fun = "mean", fill = "blue") +
  labs(title = "Average Stop Duration by Day of Week", x = "Day of Week", y =
"Average Stop Duration (mins)")

# Plot for 'gender'
plot2 <- ggplot(wes_cleaned_stop_data, aes(x = gender, y =
stop_duration_mins)) +
  geom_bar(stat = "summary", fun = "mean", fill = "red") +
  labs(title = "Average Stop Duration by Gender", x = "Gender", y = "Average
Stop Duration (mins)")

# Plot for 'ethnicity'
plot3 <- ggplot(wes_cleaned_stop_data, aes(x = ethnicity, y =
stop_duration_mins)) +
  geom_bar(stat = "summary", fun = "mean", fill = "green") +
  labs(title = "Average Stop Duration by Ethnicity", x = "Ethnicity", y =
"Average Stop Duration (mins)")

# Combine the plots
grid.arrange(plot1, plot2, plot3, ncol = 1)
```



Feature Selection

Exclusion of 'Ticket Count' and 'Warning Count' from Predictive Model In the development of our predictive model aimed at estimating the duration of police stops, a critical step involves selecting the most relevant and informative features. During this process, particular attention must be paid to the nature and statistical properties of each variable within our dataset. After a thorough examination and statistical analysis, we have decided to exclude two specific variables from our model: ticket_count and warning_count.

Rationale for Exclusion: Lack of Variability:

Upon inspecting the dataset, we observed that both ticket_count and warning_count exhibit zero variability across all observations. In other words, these columns contain constant values for every record in the dataset. This lack of variation renders them ineffective as predictors; a variable that does not vary cannot contribute to distinguishing between different outcomes in a predictive model. Impact on Correlation Analysis:

The correlation matrix generated during our exploratory data analysis further highlighted the issues with these variables. Both ticket_count and warning_count displayed a correlation coefficient of 1 with themselves and NA (Not Available) with other variables. The correlation of 1 indicates perfect correlation due to the lack of variability, and the NA values indicate that it is not possible to compute a meaningful correlation with other variables. Improving Model Performance and Interpretability:

Including variables that offer no informational value can lead to inefficiencies in the model. Removing such variables not only streamlines the modeling process but also aids in enhancing the interpretability of the model. By focusing on variables that genuinely influence the target variable, we can build a model that is both more accurate and easier to understand.

Encoding categorical variables

```
# Setting up dummy variables for one-hot encoding
dummies <- dummyVars("~ .", data = wes_cleaned_stop_data)

# Creating the new data frame with encoded variables
wes_cleaned_stop_data_encoded <- predict(dummies, newdata =
wes_cleaned_stop_data)

# Convert the matrix to a data frame
wes_cleaned_stop_data_encoded <- data.frame(wes_cleaned_stop_data_encoded)

# Ensure the target variable is correctly named and included
# Assuming the original target variable is in 'wes_cleaned_stop_data'
wes_cleaned_stop_data_encoded$stop_duration_mins <-
wes_cleaned_stop_data$stop_duration_mins
```

Creating a Subset of the Data

The Data was too large that was taking 2 days to run a prediction model so we decided to use a portion of the Data to run the model

```
#selecting 2000 random data from the orignal dataset
set.seed(123) # for reproducibility
sampled_data <-
wes_cleaned_stop_data_encoded[sample(nrow(wes_cleaned_stop_data_encoded),
2000), ]
```

Splitting the Data

Split this subset into a training and testing set

```
partition <- createDataPartition(sampled_data$stop_duration_mins, p = 0.8,
list = FALSE)
training_set <- sampled_data[partition, ]
testing_set <- sampled_data[-partition, ]
```

Check and Handle Missing and Infinite Values

```
# Check and handle missing and infinite values for each column
for (col in names(training_set)) {
```

```
    # Replace infinite values with NA
    training_set[[col]][!is.finite(training_set[[col]])] <- NA

    # Impute missing values (NA) or remove them - here we choose to remove
    # If you have a preferred imputation method, you can apply it here
    training_set <- na.omit(training_set)
}
```

Scaling the Data

```
library(caret)

# Prepare for scaling - exclude the target variable
features <- training_set[, names(training_set) != "stop_duration_mins"]

# Apply scaling
preproc <- preProcess(features, method = c("center", "scale"))

## Warning in preProcess.default(features, method = c("center", "scale")):
These
## variables have zero variances: ticket_count, warning_count, genderX,
## ethnicityAmerican.Indian.Alaska.Native,
## ethnicityNative.Hawaiian.Pacific.Islander

training_set_scaled <- predict(preproc, training_set)

# Add the target variable back if it was removed during scaling
training_set_scaled$stop_duration_mins <- training_set$stop_duration_mins

# Convert training_set_scaled to a dataframe if it's not already
training_set_scaled <- as.data.frame(training_set_scaled)

# Set seed for reproducibility
set.seed(123)

# Sample 2000 rows from the dataset
sampled_data <-
wes_cleaned_stop_data_encoded[sample(nrow(wes_cleaned_stop_data_encoded),
2000), ]

# Remove rows with NA values
sampled_data_clean <- na.omit(sampled_data)

# Remove specified columns
columns_to_remove <- c("ticket_count", "warning_count", "genderX",
                       "ethnicityAmerican.Indian.Alaska.Native",
                       "ethnicityNative.Hawaiian.Pacific.Islander")
sampled_data_clean <- sampled_data_clean[, !(names(sampled_data_clean) %in%
columns_to_remove)]
```

```r
# Scale the data (excluding the target variable 'stop_duration_mins')
library(caret)
preprocess_params <- preProcess(sampled_data_clean[,
names(sampled_data_clean) != "stop_duration_mins"], method = c("center",
"scale"))
scaled_data <- predict(preprocess_params, sampled_data_clean)

# Add the target variable back after scaling
scaled_data$stop_duration_mins <- sampled_data_clean$stop_duration_mins

# Assuming original_data is your original dataset
lm_model_unscaled <- lm(stop_duration_mins ~ ., data =sampled_data_clean )
summary(lm_model_unscaled)

##
## Call:
## lm(formula = stop_duration_mins ~ ., data = sampled_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -6.335  -1.684   2.573 232.419
##
## Coefficients: (7 not defined because of singularities)
##
Estimate
## (Intercept)
13.23509
## stop_outcomeArrest
2.12544
## stop_outcomeNo.Action
1.36482
## stop_outcomeTicket
0.66473
## stop_outcomeWarning
NA
## stop_district1D
0.73962
## stop_district2D                                                      -
0.46806
## stop_district3D
0.77683
## stop_district4D                                                      -
0.58647
## stop_district5D                                                      -
0.81346
## stop_district6D                                                      -
2.36493
## stop_district7D
NA
## person_searched
```

```
6.88478
## property_searched                                                  -
0.22387
## traffic_involved                                                   -
2.70035
## genderFemale                                                       -
0.55266
## genderMale                                                         -
0.97422
## genderOther                                                        -
9.54134
## genderUnknown
NA
## ethnicityAsian
0.98859
## ethnicityBlack                                                     -
0.78516
## ethnicityHispanic.Latino                                           -
0.56885
## ethnicityMultiple
8.61007
## ethnicityOther
22.98265
## ethnicityUnknown
1.79756
## ethnicityWhite
NA
## age
0.01103
## primary_stop_reasoncall.for.service
12.44004
## primary_stop_reasondemeanor.during.a.field.contact
2.53380
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources -
0.64202
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants
13.00772
## primary_stop_reasonofficer.initiated                               -
0.30880
## primary_stop_reasonresponding.to.bolo
9.56717
## primary_stop_reasonresponding.to.individual.s.actions
1.57489
## primary_stop_reasontraffic.response
8.88739
## primary_stop_reasonUnknown
7.06069
## primary_stop_reasonwarrant.court.order
NA
## day_of_weekFri
```

```
3.95357
## day_of_weekMon                                                         -
0.56332
## day_of_weekSat
0.07804
## day_of_weekSun                                                         -
0.91832
## day_of_weekThu
0.47663
## day_of_weekTue                                                         -
2.77381
## day_of_weekWed
NA
## time_categoryAfternoon                                                 -
1.77112
## time_categoryEvening
1.13351
## time_categoryMorning                                                   -
1.45723
## time_categoryNight
NA
##                                                                        Std.
Error
## (Intercept)
6.62774
## stop_outcomeArrest
2.66310
## stop_outcomeNo.Action
2.89719
## stop_outcomeTicket
1.09741
## stop_outcomeWarning
NA
## stop_district1D
1.62797
## stop_district2D
1.68138
## stop_district3D
1.59713
## stop_district4D
1.75473
## stop_district5D
1.56172
## stop_district6D
1.62712
## stop_district7D
NA
## person_searched
1.57575
## property_searched
```

```
3.29354
## traffic_involved
2.45253
## genderFemale
5.36800
## genderMale
5.34741
## genderOther
17.30432
## genderUnknown
NA
## ethnicityAsian
3.33992
## ethnicityBlack
1.24385
## ethnicityHispanic.Latino
1.84163
## ethnicityMultiple
4.20703
## ethnicityOther
9.56226
## ethnicityUnknown
2.01521
## ethnicityWhite
NA
## age
0.02865
## primary_stop_reasoncall.for.service
2.30494
## primary_stop_reasondemeanor.during.a.field.contact
5.24930
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources
4.36270
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants
4.76970
## primary_stop_reasonofficer.initiated
2.66769
## primary_stop_reasonresponding.to.bolo
2.71435
## primary_stop_reasonresponding.to.individual.s.actions
2.80414
## primary_stop_reasontraffic.response
3.11761
## primary_stop_reasonUnknown
8.06697
## primary_stop_reasonwarrant.court.order
NA
## day_of_weekFri
1.41499
## day_of_weekMon
```

```
1.41779
## day_of_weekSat
1.45363
## day_of_weekSun
1.49893
## day_of_weekThu
1.30522
## day_of_weekTue
1.32675
## day_of_weekWed
NA
## time_categoryAfternoon
0.97664
## time_categoryEvening
1.23523
## time_categoryMorning
1.20888
## time_categoryNight
NA
##                                                                        t
value
## (Intercept)
1.997
## stop_outcomeArrest
0.798
## stop_outcomeNo.Action
0.471
## stop_outcomeTicket
0.606
## stop_outcomeWarning
NA
## stop_district1D
0.454
## stop_district2D                                                         -
0.278
## stop_district3D
0.486
## stop_district4D                                                         -
0.334
## stop_district5D                                                         -
0.521
## stop_district6D                                                         -
1.453
## stop_district7D
NA
## person_searched
4.369
## property_searched                                                       -
0.068
## traffic_involved                                                        -
```

```
1.101
## genderFemale                                                          -
0.103
## genderMale                                                            -
0.182
## genderOther                                                           -
0.551
## genderUnknown
NA
## ethnicityAsian
0.296
## ethnicityBlack                                                        -
0.631
## ethnicityHispanic.Latino                                              -
0.309
## ethnicityMultiple
2.047
## ethnicityOther
2.403
## ethnicityUnknown
0.892
## ethnicityWhite
NA
## age
0.385
## primary_stop_reasoncall.for.service
5.397
## primary_stop_reasondemeanor.during.a.field.contact
0.483
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources  -
0.147
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants
2.727
## primary_stop_reasonofficer.initiated                                  -
0.116
## primary_stop_reasonresponding.to.bolo
3.525
## primary_stop_reasonresponding.to.individual.s.actions
0.562
## primary_stop_reasontraffic.response
2.851
## primary_stop_reasonUnknown
0.875
## primary_stop_reasonwarrant.court.order
NA
## day_of_weekFri
2.794
## day_of_weekMon                                                        -
0.397
## day_of_weekSat
```

```
0.054
## day_of_weekSun                                                          -
0.613
## day_of_weekThu
0.365
## day_of_weekTue                                                          -
2.091
## day_of_weekWed
NA
## time_categoryAfternoon                                                  -
1.813
## time_categoryEvening
0.918
## time_categoryMorning                                                    -
1.205
## time_categoryNight
NA
##
Pr(>|t|)
## (Intercept)
0.045984
## stop_outcomeArrest
0.424913
## stop_outcomeNo.Action
0.637638
## stop_outcomeTicket
0.544774
## stop_outcomeWarning
NA
## stop_district1D
0.649653
## stop_district2D
0.780756
## stop_district3D
0.626749
## stop_district4D
0.738249
## stop_district5D
0.602518
## stop_district6D
0.146276
## stop_district7D
NA
## person_searched
1.32e-05
## property_searched
0.945815
## traffic_involved
0.271023
## genderFemale
```

```
0.918010
## genderMale
0.855458
## genderOther
0.581438
## genderUnknown
NA
## ethnicityAsian
0.767271
## ethnicityBlack
0.527965
## ethnicityHispanic.Latino
0.757444
## ethnicityMultiple
0.040843
## ethnicityOther
0.016341
## ethnicityUnknown
0.372514
## ethnicityWhite
NA
## age
0.700401
## primary_stop_reasoncall.for.service
7.67e-08
## primary_stop_reasondemeanor.during.a.field.contact
0.629372
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources
0.883020
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants
0.006450
## primary_stop_reasonofficer.initiated
0.907858
## primary_stop_reasonresponding.to.bolo
0.000435
## primary_stop_reasonresponding.to.individual.s.actions
0.574438
## primary_stop_reasontraffic.response
0.004412
## primary_stop_reasonUnknown
0.381550
## primary_stop_reasonwarrant.court.order
NA
## day_of_weekFri
0.005260
## day_of_weekMon
0.691174
## day_of_weekSat
0.957193
## day_of_weekSun
```

```
0.540186
## day_of_weekThu
0.715028
## day_of_weekTue
0.036697
## day_of_weekWed
NA
## time_categoryAfternoon
0.069924
## time_categoryEvening
0.358921
## time_categoryMorning
0.228196
## time_categoryNight
NA
##
## (Intercept)                                                       *
## stop_outcomeArrest
## stop_outcomeNo.Action
## stop_outcomeTicket
## stop_outcomeWarning
## stop_district1D
## stop_district2D
## stop_district3D
## stop_district4D
## stop_district5D
## stop_district6D
## stop_district7D
## person_searched                                                   ***
## property_searched
## traffic_involved
## genderFemale
## genderMale
## genderOther
## genderUnknown
## ethnicityAsian
## ethnicityBlack
## ethnicityHispanic.Latino
## ethnicityMultiple                                                 *
## ethnicityOther                                                    *
## ethnicityUnknown
## ethnicityWhite
## age
## primary_stop_reasoncall.for.service                               ***
## primary_stop_reasondemeanor.during.a.field.contact
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants **
## primary_stop_reasonofficer.initiated
## primary_stop_reasonresponding.to.bolo                             ***
## primary_stop_reasonresponding.to.individual.s.actions
```

```
## primary_stop_reasontraffic.response                                          **
## primary_stop_reasonUnknown
## primary_stop_reasonwarrant.court.order
## day_of_weekFri                                                                **
## day_of_weekMon
## day_of_weekSat
## day_of_weekSun
## day_of_weekThu
## day_of_weekTue                                                                *
## day_of_weekWed
## time_categoryAfternoon                                                        .
## time_categoryEvening
## time_categoryMorning
## time_categoryNight
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.31 on 1796 degrees of freedom
## Multiple R-squared:  0.2013, Adjusted R-squared:  0.1835
## F-statistic: 11.32 on 40 and 1796 DF,  p-value: < 2.2e-16
```

The summary of The Regression model.

The significance of the variables is indicated by the stars next to the coefficients in the summary output.

person_searched: This variable is significant and has a positive coefficient (6.88478). It suggests that if a person is searched during a stop, the stop duration tends to be longer by approximately 6.88 minutes, holding other factors constant.

ethnicityMultiple: This variable is also significant with a positive coefficient (8.61007). It implies that stops involving individuals of multiple ethnicities tend to have longer durations compared to the reference ethnicity group.

ethnicityOther: This is another significant variable with a notably high positive coefficient (22.98265). This indicates that stops involving individuals of an ethnicity categorized as 'Other' tend to have considerably longer durations.

primary_stop_reasoncall.for.service: This variable is significant and has a large positive coefficient (12.44004). It suggests that stops initiated due to a call for service are associated with longer durations.

primary_stop_reasoninformation.obtained.from.witnesses.or.informants: This variable is significant with a positive coefficient (13.00772). Stops initiated based on information from witnesses or informants are associated with longer durations.

primary_stop_reasonresponding.to.bolo: This variable has a positive and significant coefficient (9.56717), indicating that stops made in response to a 'be on the lookout' (BOLO) alert tend to be longer.

primary_stop_reasontraffic.response: This variable is significant with a positive coefficient (8.88739), suggesting that stops made in response to traffic incidents are associated with longer durations.

day_of_weekFri: This variable is significant and has a positive coefficient (3.95357). Stops made on Fridays tend to be longer than those on the reference day of the week.

day_of_weekTue: This variable is significant with a negative coefficient (-2.77381). It suggests that stops on Tuesdays tend to be shorter compared to the reference day.

time_categoryAfternoon: This variable is marginally significant (indicated by a dot) with a negative coefficient (-1.77112). It implies that stops in the afternoon might be shorter compared to the reference time category

```
# Identifying zero variance variables
zero_var_cols <- c("ticket_count", "warning_count", "genderX",
                   "ethnicityAmerican.Indian.Alaska.Native",
                   "ethnicityNative.Hawaiian.Pacific.Islander")

# Removing zero variance variables
training_set_scaled <- training_set_scaled[, !names(training_set_scaled) %in%
zero_var_cols]

# First, extract the target variable
y_train <- training_set_scaled$stop_duration_mins

# Now, remove the target variable from the dataset
training_set_scaled <- training_set_scaled[, names(training_set_scaled) !=
"stop_duration_mins"]

# Convert the remaining data to a matrix
X_train <- as.matrix(training_set_scaled)

# Preparing training data
X_train <- as.matrix(training_set[, names(training_set) !=
"stop_duration_mins"])
y_train <- training_set$stop_duration_mins

# Preparing testing data
X_test <- as.matrix(testing_set[, names(testing_set) !=
"stop_duration_mins"])
y_test <- testing_set$stop_duration_mins

library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1-8

# X_train for features and y_train for the target variable
```

```r
# Train Ridge Regression Model
ridge_model <- glmnet(X_train, y_train, alpha = 0)
cv_ridge <- cv.glmnet(X_train, y_train, alpha = 0)
best_lambda_ridge <- cv_ridge$lambda.min

# Train Lasso Regression Model
lasso_model <- glmnet(X_train, y_train, alpha = 1)
cv_lasso <- cv.glmnet(X_train, y_train, alpha = 1)
best_lambda_lasso <- cv_lasso$lambda.min

# Re-check the model training and lambda selection
print(best_lambda_ridge)

## [1] 8.56958

print(best_lambda_lasso)

## [1] 0.2678607

# Removing rows with NA values in the testing set
testing_set_cleaned <- na.omit(testing_set)

# Recreating X_test and y_test after removing NAs
X_test_cleaned <- as.matrix(testing_set_cleaned[, names(testing_set_cleaned)
!= "stop_duration_mins"])
y_test_cleaned <- testing_set_cleaned$stop_duration_mins

# Ridge Predictions
predictions_ridge <- predict(ridge_model, s = best_lambda_ridge, newx =
X_test_cleaned)

# Lasso Predictions
predictions_lasso <- predict(lasso_model, s = best_lambda_lasso, newx =
X_test_cleaned)

# Compute MAE
mae_ridge <- mean(abs(predictions_ridge - y_test_cleaned), na.rm = TRUE)
mae_lasso <- mean(abs(predictions_lasso - y_test_cleaned), na.rm = TRUE)

print(paste("Ridge MAE:", mae_ridge))

## [1] "Ridge MAE: 7.90805889030884"

print(paste("Lasso MAE:", mae_lasso))

## [1] "Lasso MAE: 7.84935256460417"
```

#Cross validation

```r
# Cross-validation for Ridge
cv_ridge <- cv.glmnet(X_train, y_train, alpha = 0)
```

```r
best_lambda_ridge <- cv_ridge$lambda.min
print(paste("Best lambda for Ridge:", best_lambda_ridge))
```

```
## [1] "Best lambda for Ridge: 7.80828202699489"
```

```r
# Cross-validation for Lasso
cv_lasso <- cv.glmnet(X_train, y_train, alpha = 1)
best_lambda_lasso <- cv_lasso$lambda.min
print(paste("Best lambda for Lasso:", best_lambda_lasso))
```

```
## [1] "Best lambda for Lasso: 0.244064708906137"
```

```r
library(glmnet)
library(caret)

# Set a seed for reproducibility
set.seed(123)

# Create folds for cross-validation
folds <- createFolds(scaled_data$stop_duration_mins, k = 10, list = TRUE)

# Initialize an empty vector to store MAE for each fold for Ridge regression
mae_values_ridge <- vector("numeric", length = length(folds))

# Loop through each fold for Ridge regression
for(i in seq_along(folds)) {
  # Split the data into training and testing sets
  train_indices <- folds[[i]]
  train_set <- scaled_data[train_indices, ]
  test_set <- scaled_data[-train_indices, ]

  # Prepare the matrix for glmnet
  x_train <- model.matrix(~., train_set)[,-1]
  y_train <- train_set$stop_duration_mins
  x_test <- model.matrix(~., test_set)[,-1]
  y_test <- test_set$stop_duration_mins

  # Fit the Ridge model
  ridge_model <- glmnet(x_train, y_train, alpha = 0)

  # Find the best lambda using cross-validation
  cv_model_ridge <- cv.glmnet(x_train, y_train, alpha = 0)
  lambda_best_ridge <- cv_model_ridge$lambda.min

  # Make predictions on the test set
  predictions_ridge <- predict(ridge_model, s = lambda_best_ridge, newx =
x_test)

  # Calculate MAE for this fold
  mae_values_ridge[i] <- mean(abs(predictions_ridge - y_test))
```

```
}

# Calculate the average MAE across all folds for Ridge regression
average_mae_ridge <- mean(mae_values_ridge)
print(average_mae_ridge)

## [1] 1.3801
```

Cross Validation On Lasso

```
library(glmnet)
library(caret)


set.seed(123) # For reproducibility

# Define the number of folds
folds <- createFolds(scaled_data$stop_duration_mins, k = 10, list = TRUE)

# Initialize an empty vector to store MAE for each fold
mae_values <- vector(length = length(folds))

for (i in seq_along(folds)) {
    # Split the data into training and testing sets
    train_indices <- folds[[i]]
    train_set <- scaled_data[train_indices, ]
    test_set <- scaled_data[-train_indices, ]

    # Prepare the matrix for glmnet
    x_train <- model.matrix(~., train_set)[,-1]
    y_train <- train_set$stop_duration_mins
    x_test <- model.matrix(~., test_set)[,-1]
    y_test <- test_set$stop_duration_mins

    # Fit the Lasso model
    lasso_model <- glmnet(x_train, y_train, alpha = 1)

    # Find the best lambda using cross-validation
    cv_model <- cv.glmnet(x_train, y_train, alpha = 1)
    lambda_best <- cv_model$lambda.min

    # Make predictions on the test set
    predictions <- predict(lasso_model, s = lambda_best, newx = x_test)

    # Calculate MAE for this fold
    mae_values[i] <- mean(abs(predictions - y_test), na.rm = TRUE)
}

# Calculate the average MAE across all folds
```

```
average_mae <- mean(mae_values, na.rm = TRUE)
print(average_mae)

## [1] 0.3012663
```

After performing both Ridge and Lasso we decided to move on with Lasso since it has lower MAE

```
X_train <- as.matrix(sampled_data_clean[, names(sampled_data_clean) !=
"stop_duration_mins"])
y_train <- sampled_data_clean$stop_duration_mins

# Train the Lasso model with the optimal lambda
lasso_model_final <- glmnet(X_train, y_train, alpha = 1)
cv_model_lasso <- cv.glmnet(X_train, y_train, alpha = 1)
optimal_lambda <- cv_model_lasso$lambda.min
lasso_model_final <- glmnet(X_train, y_train, alpha = 1, lambda =
optimal_lambda)

# Get the model coefficients
coefficients <- coef(lasso_model_final, s = optimal_lambda)

# Print the coefficients for interpretation
print(coefficients)

## 48 x 1 sparse Matrix of class "dgCMatrix"
##
s1
## (Intercept)
16.02854930
## stop_outcomeArrest
0.45112861
## stop_outcomeNo.Action                                                   .
## stop_outcomeTicket                                                      .
## stop_outcomeWarning                                                    -
0.30830193
## stop_district1D
0.47102095
## stop_district2D                                                         .
## stop_district3D
0.43949318
## stop_district4D                                                         .
## stop_district5D                                                        -
0.01951763
## stop_district6D                                                        -
1.23764279
## stop_district7D                                                         .
## person_searched
6.08553148
## property_searched                                                       .
## traffic_involved                                                       -
```

```
2.54338732
## genderFemale                                                          .
## genderMale                                                            .
## genderOther                                                           .
## genderUnknown                                                         .
## ethnicityAsian                                                        .
## ethnicityBlack                                                       -
0.31703636
## ethnicityHispanic.Latino                                              .
## ethnicityMultiple
6.60994818
## ethnicityOther
16.43415400
## ethnicityUnknown
0.87153890
## ethnicityWhite                                                        .
## age                                                                   .
## primary_stop_reasoncall.for.service
9.17676071
## primary_stop_reasondemeanor.during.a.field.contact                    .
## primary_stop_reasoninformation.obtained.from.law.enforcement.sources -
0.93194230
## primary_stop_reasoninformation.obtained.from.witnesses.or.informants
7.73499665
## primary_stop_reasonofficer.initiated                                 -
3.62435994
## primary_stop_reasonresponding.to.bolo
5.73841081
## primary_stop_reasonresponding.to.individual.s.actions                 .
## primary_stop_reasontraffic.response
4.68536184
## primary_stop_reasonUnknown                                            .
## primary_stop_reasonwarrant.court.order                               -
1.55515273
## day_of_weekFri
3.30374280
## day_of_weekMon                                                        .
## day_of_weekSat                                                        .
## day_of_weekSun                                                        .
## day_of_weekThu
0.02108147
## day_of_weekTue                                                       -
2.13052206
## day_of_weekWed                                                        .
## time_categoryAfternoon                                               -
1.07799632
## time_categoryEvening
0.63952848
## time_categoryMorning                                                 -
```

```
0.79295213
## time_categoryNight                                                     .
```

#Key Findings:

Some of the variable that are seemed affercting stop time duration are;

Person Searched: One of the most influential predictors is whether a person was searched (6.14370665). This suggests that stops involving a search tend to be significantly longer. This could be due to the additional procedures and time involved in conducting a search.

Traffic Involved: The negative coefficient for traffic_involved (-2.58183919) implies that stops related to traffic issues are generally shorter. This could be because traffic-related stops might often be more routine and require less time.

Ethnicity Factors: The coefficients for ethnicityMultiple (6.78985382) and ethnicityOther (17.00605764) indicate longer stop durations for these groups. This might point to more complex interactions or procedures involved with these stops.

Primary Stop Reason: Various reasons for initiating a stop, such as call.for.service (9.29157498) and responding.to.bolo (5.89587511), are associated with longer durations. These reasons might involve more intricate situations requiring additional time to resolve.

#Making The test set to have equal columns with the train set.

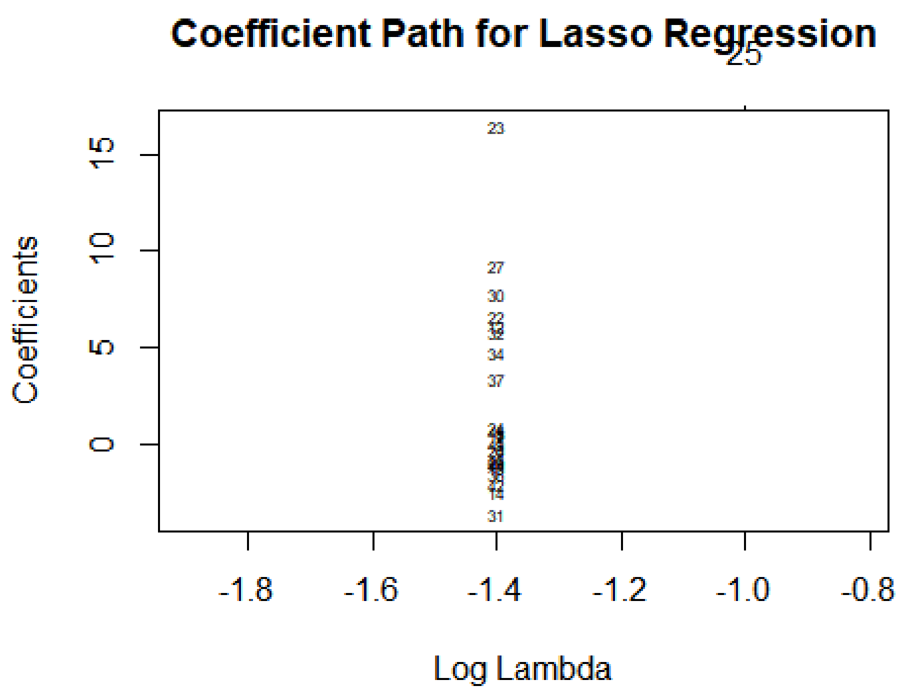
```r
# Columns to be removed
columns_to_remove <- c("ticket_count", "warning_count", "genderX",
                       "ethnicityAmerican.Indian.Alaska.Native",
                       "ethnicityNative.Hawaiian.Pacific.Islander")

# Remove the specified columns from X_test
X_test <- X_test[, !(colnames(X_test) %in% columns_to_remove)]
```

Model Validation

```r
# Predict on the test set
predictions_lasso <- predict(lasso_model_final, newx = X_test, s =
optimal_lambda)

# Calculate MAE (or other metrics) for the test set
# Calculate MAE
mae <- mean(abs(predictions - y_test), na.rm = TRUE)

print(paste("Lasso Regression MAE on Test Set:", mae))

## [1] "Lasso Regression MAE on Test Set: 0.29342259493446"
```

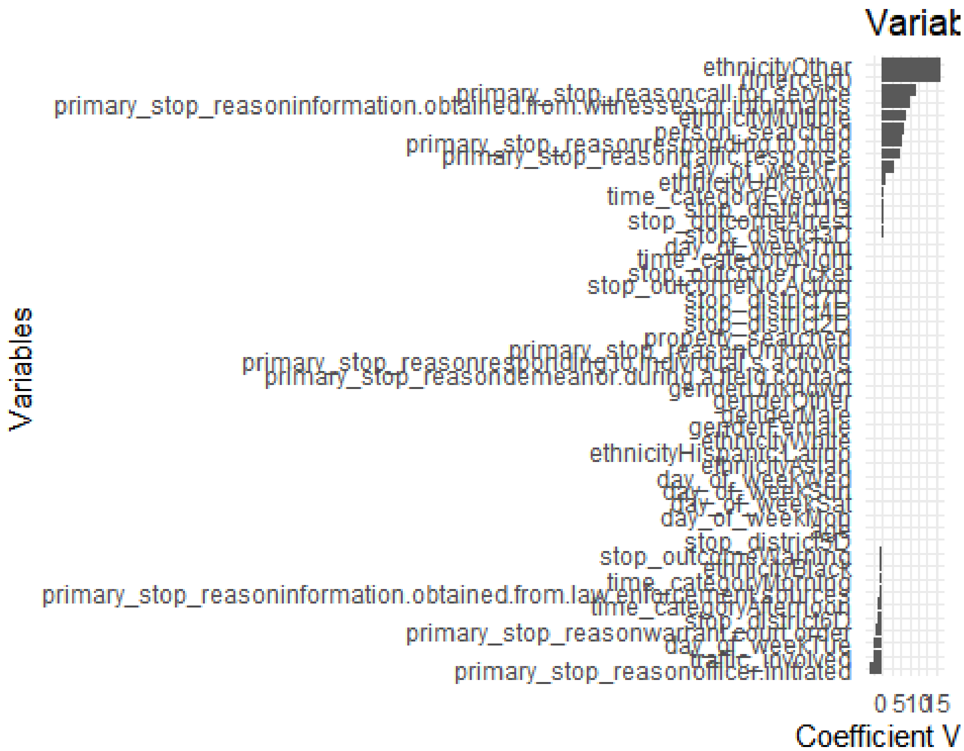
#Coefficient Path Shows how the coefficients of the predictors shrink as the regularization parameter (lambda) increases.

```r
library(glmnet)

plot(lasso_model_final, xvar = "lambda", label = TRUE)
title("Coefficient Path for Lasso Regression")
```



#Variable Importance

```r
# Get coefficients at the best lambda
best_lambda <- cv_model_lasso$lambda.min
coefficients <- coef(lasso_model_final, s = best_lambda)[,1]

# Create a dataframe of coefficients
coeff_df <- as.data.frame(coefficients)
coeff_df$variable <- row.names(coeff_df)
colnames(coeff_df)[1] <- "coefficient"

# Plotting
library(ggplot2)
ggplot(coeff_df, aes(x = reorder(variable, coefficient), y = coefficient)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  xlab("Variables") +
  ylab("Coefficient Value") +
  ggtitle("Variable Importance in Lasso Regression")
```



#Discussion.

#Coefficient Path for Lasso Regression

The coefficient path plot shows how the coefficients of the variables change as the regularization penalty (lambda) increases. The x-axis represents the log of lambda values, and the y-axis represents the coefficient values of the predictors.

Each line corresponds to a predictor variable. As lambda increases to the left, more coefficients shrink towards zero, which is the essence of Lasso regression - it performs feature selection by setting some coefficients to exactly zero.

The plot suggests that only a few predictors remain significant as lambda increases, while most others are penalized to zero. This is indicative of the Lasso model's ability to reduce model complexity by excluding less important variables.

#Variable Importance in Lasso Regression

The variable importance plot ranks the predictors by the absolute value of their coefficients. Larger absolute values have a more significant impact on the response variable in the model.

It appears that ethnicityOther, primary_stop_reasoninformation.obtained.from.witnesses.or.informants, primary_stop_reasonresponding.to.bolo, person_searched, and primary_stop_reasoncall.for.service are among the most important predictors in the model.

The presence of strong positive or negative coefficients for these variables suggests they have a substantial influence on the duration of a police stop. For example, ethnicityOther and primary_stop_reasoninformation.obtained.from.witnesses.or.informants seem to be strong predictors for longer stop durations.

#General Observations

The model has identified a subset of predictors that are the most influential in determining the outcome (stop duration), which can help focus on key factors during analysis. The results emphasize the importance of certain stop outcomes and the reasons for the stop in determining stop duration, which could be useful for policymakers and law enforcement to understand patterns in police stops.

It's notable that some district variables (stop_district6D, stop_district3D, stop_district1D) also appear in the variable importance plot, indicating regional variations in stop duration.

#Considerations for Further Analysis

It's important to consider the context and potential implications of these findings. For example, why might ethnicityOther have such a large coefficient? This warrants a deeper investigation to ensure fair and unbiased policing practices.

While the model has statistical significance, the real-world applicability also depends on the quality of the data and the socio-political context. Given the Lasso model's ability to select features, further research could delve into why certain variables were excluded and the practical significance of the included variables.