Analysis of Traffic Stops in Montana

Chandana Pamidi, Tejas Ganesh Naik, Ujwala Munigela, Yogeshwar Pullagurla

2024-05-20

Introduction In this project, we will analyse on the patterns of traffic stops in the state of Montana across 9 years, from December 2008 to December 2017. This focused analysis aims to provide insights into law enforcement activity and potentially reveal any differences in how stops are conducted across the state.

We mainly focus to analyse on the following questions through our study:

1) Is there a statistically significant relationship between the age of subjects and the likelihood of receiving a warning during a stop?
2) How does the likelihood of receiving a warning vary across different age groups (e.g., youngsters, middle-aged, old)?
3) Is the mean age of the drivers who got arrested same as the mean age of driver got received warning ?
4) Is the time of the day a factor in determining the outcome of the traffic stop?
5) Are female drivers less at risk for violations compared to male drivers?

```r
# Load data from a CSV file
data <- readRDS("wb225bk3255_mt_statewide_2023_01_26.rds")

# Display the structure of the dataset
str(data)
```

```
## tibble [921,228 x 30] (S3: tbl_df/tbl/data.frame)
##  $ raw_row_number           : chr [1:921228] "1" "2" "3" "4" ...
##  $ date                     : Date[1:921228], format: "2009-01-01" "2009-01-02" ...
##  $ time                     : 'hms' num [1:921228] 02:10:53 11:34:19 11:36:42 10:33:11 ...
##   ..- attr(*, "units")= chr "secs"
##  $ location                 : chr [1:921228] "US 89 N MM10 (SB)" "HWY 93 SO AND ANNS LANE S/B" "POO
##  $ lat                      : num [1:921228] 47.6 46.8 46.7 46.7 46.7 ...
##  $ lng                      : num [1:921228] -112 -114 -114 -114 -114 ...
##  $ county_name              : chr [1:921228] "Cascade County" "Missoula County" "Missoula County" "
##  $ subject_age              : int [1:921228] 16 19 17 17 31 20 30 34 21 18 ...
##  $ subject_race             : Factor w/ 6 levels "asian/pacific islander",..: 4 4 4 NA NA NA 4 NA 4
##  $ subject_sex              : Factor w/ 2 levels "male","female": 2 1 1 2 1 1 1 2 1 2 ...
##  $ department_name          : chr [1:921228] "Montana Highway Patrol" "Montana Highway Patrol" "Mon
##  $ type                     : Factor w/ 2 levels "pedestrian","vehicular": 2 2 2 2 2 2 2 2 2 2 ...
##  $ violation                : chr [1:921228] "240 - INSURANCE|150 - HIT AND RUN|245 - OTHER NON-HAZA
##  $ arrest_made              : logi [1:921228] FALSE TRUE TRUE TRUE TRUE TRUE ...
##  $ citation_issued          : logi [1:921228] TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ warning_issued           : logi [1:921228] TRUE TRUE FALSE FALSE FALSE TRUE ...
##  $ outcome                  : Factor w/ 4 levels "warning","citation",..: 2 4 4 4 4 4 2 4 2 NA ...
##  $ frisk_performed          : logi [1:921228] FALSE FALSE FALSE NA NA NA ...
##  $ search_conducted         : logi [1:921228] FALSE FALSE FALSE TRUE TRUE TRUE ...
##  $ search_basis             : Factor w/ 5 levels "k9","plain view",..: NA NA NA NA NA NA NA NA NA N
```

```
##  $ reason_for_stop          : chr [1:921228] "--- - HIT AND RUN" "EXPIRED TAG ( - MONTHS OR LESS )"
##  $ vehicle_make             : chr [1:921228] "FORD" "GMC" "GMC" "HOND" ...
##  $ vehicle_model            : chr [1:921228] "EXPLORER" "TK" "YUKON" "CR-V" ...
##  $ vehicle_type             : chr [1:921228] "SPORT UTILITY" "TRUCK" "SPORT UTILITY" "SPORT UTILITY
##  $ vehicle_registration_state: Factor w/ 51 levels "AL","AK","AZ",..: 21 21 21 21 21 21 21 21 21 21
##  $ vehicle_year             : int [1:921228] 1994 1996 1999 2002 1992 1998 2006 2004 1992 1987 ...
##  $ raw_Race                 : chr [1:921228] "W" "W" "W" "W" ...
##  $ raw_Ethnicity            : chr [1:921228] "N" "N" "N" NA ...
##  $ raw_SearchType           : chr [1:921228] "NO SEARCH REQUESTED" "NO SEARCH REQUESTED" "NO SEARCH
##  $ raw_search_basis         : chr [1:921228] "" "" "" "" ...
```

```
data
```

```
## # A tibble: 921,228 x 30
##    raw_row_number date       time     location             lat   lng county_name
##    <chr>          <date>     <time>   <chr>              <dbl> <dbl> <chr>
##  1 1              2009-01-01 02:10:53 US 89 N MM10 (SB)   47.6 -112. Cascade Co~
##  2 2              2009-01-02 11:34:19 HWY 93 SO AND ANN~  46.8 -114. Missoula C~
##  3 3              2009-01-03 11:36:42 P007 HWY 93 MM 77~  46.7 -114. Missoula C~
##  4 4              2009-01-04 10:33:11 P007 HWY 93 MM 81~  46.7 -114. Missoula C~
##  5 5              2009-01-04 10:46:43 P007 HWY 93 MM 81~  46.7 -114. Missoula C~
##  6 6              2009-01-04 14:41:57 P007 HWY 93 MM 67~  46.5 -114. Ravalli Co~
##  7 7              2009-01-04 17:45:40 WESTBOUND TRUCK S~  45.9 -108. Yellowston~
##  8 8              2009-01-05 15:32:41 P007 HWY 93 MM 79~  46.7 -114. Missoula C~
##  9 9              2009-01-06 16:45:12 INTERSECTION OF H~  45.9 -108. Yellowston~
## 10 10             2009-01-06 16:45:17 INTERSECTION OF H~  45.9 -108. Yellowston~
## # i 921,218 more rows
## # i 23 more variables: subject_age <int>, subject_race <fct>,
## #   subject_sex <fct>, department_name <chr>, type <fct>, violation <chr>,
## #   arrest_made <lgl>, citation_issued <lgl>, warning_issued <lgl>,
## #   outcome <fct>, frisk_performed <lgl>, search_conducted <lgl>,
## #   search_basis <fct>, reason_for_stop <chr>, vehicle_make <chr>,
## #   vehicle_model <chr>, vehicle_type <chr>, ...
```

Then, we selected 17 columns to study and processed the corresponding filtered data for better and faster analysis.

```
data_filtered <- data %>%
  select(date, time, county_name, subject_sex, subject_age, citation_issued, warning_issued, arrest_made
```

Linear Regression Introduction: In this section we would like to study: 1.) Is there a statistically significant relationship between the age of subjects and the likelihood of receiving a warning during a stop? 2.) Are there any outliers in the data? 3.) Is there any influence or leverage of some instances? 5.) Does the data follow Equal variance condition? 6.) Does the data follow normal distribution? ### Linear Regression

Null Hypothesis (H0): There is no association(linear relationship) between subject age groups and arrest made during the incidents.

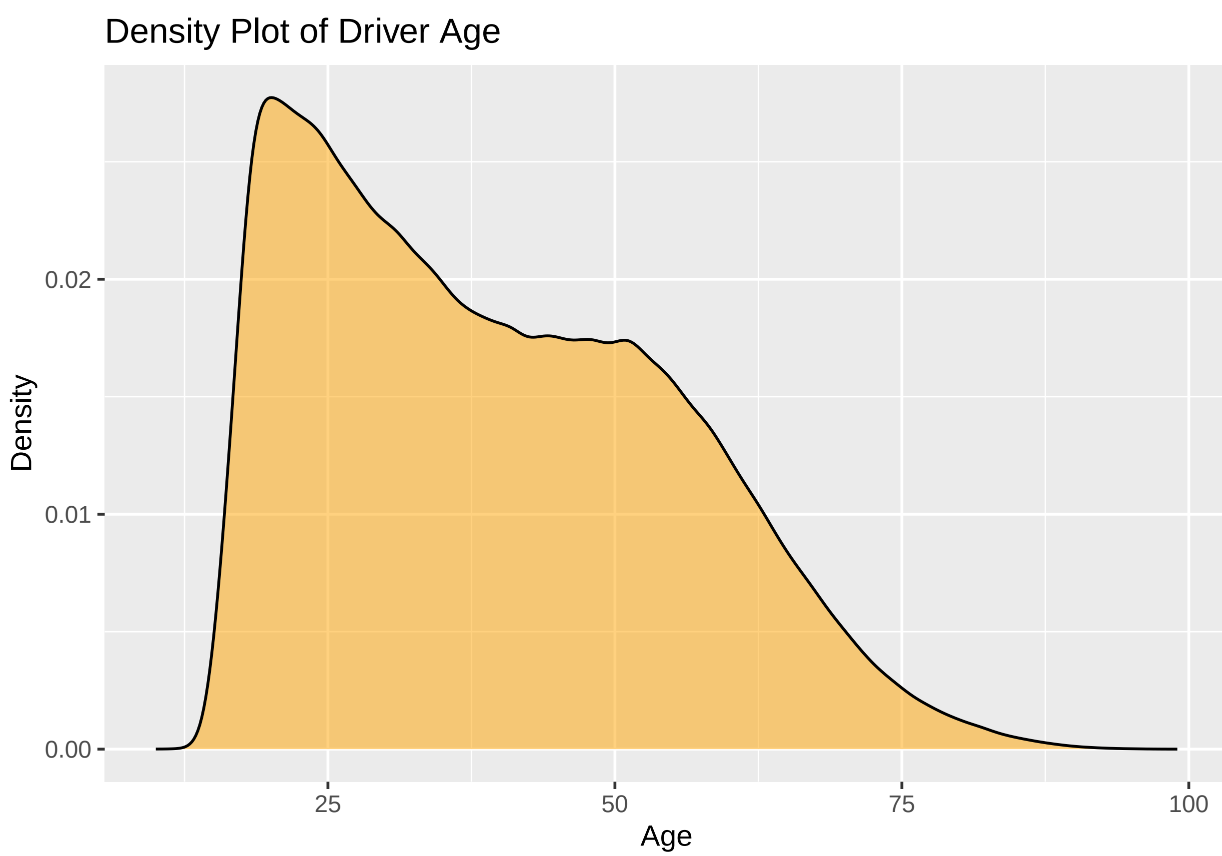
Alternative Hypothesis (H1): There is an association(linear relationship) between subject age groups and arrest made during the incidents.

Data Processing

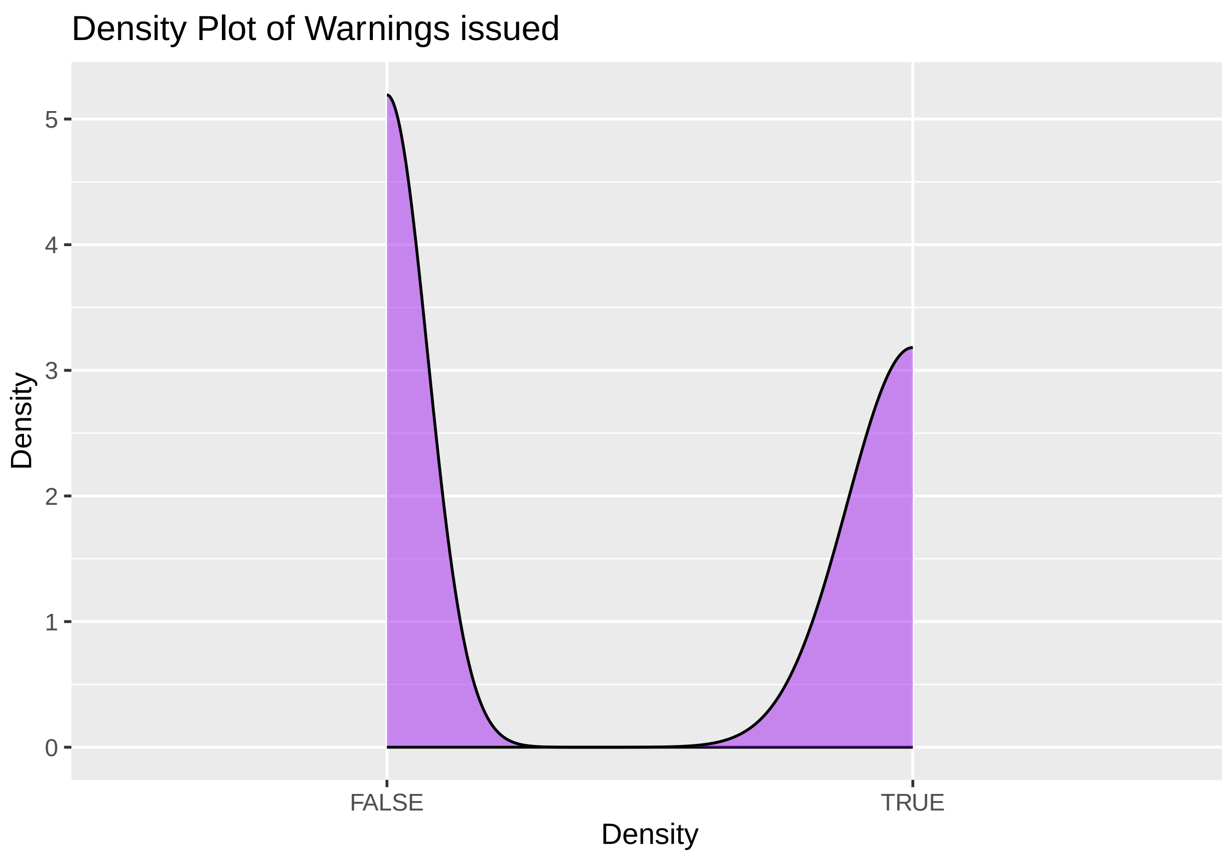
```r
data_filtered <- data_filtered %>%
  drop_na()
colSums(is.na(data_filtered))
```

```
##              date              time       county_name       subject_sex
##                 0                 0                 0                 0
##       subject_age  citation_issued    warning_issued       arrest_made
##                 0                 0                 0                 0
##           outcome  frisk_performed search_conducted  reason_for_stop
##                 0                 0                 0                 0
##      vehicle_make    vehicle_model      vehicle_type      vehicle_year
##                 0                 0                 0                 0
##         violation
##                 0
```

```r
ggplot(data_filtered, aes(x=subject_age)) +
  geom_density(fill = "orange", alpha=0.5) +
  labs(title = "Density Plot of Driver Age", x="Age", y="Density")
```




```r
ggplot(data_filtered, aes(x=warning_issued)) +
  geom_density(fill = "purple", alpha=0.5) +
  labs(title = "Density Plot of Warnings issued", x="Density", y="Density")
```

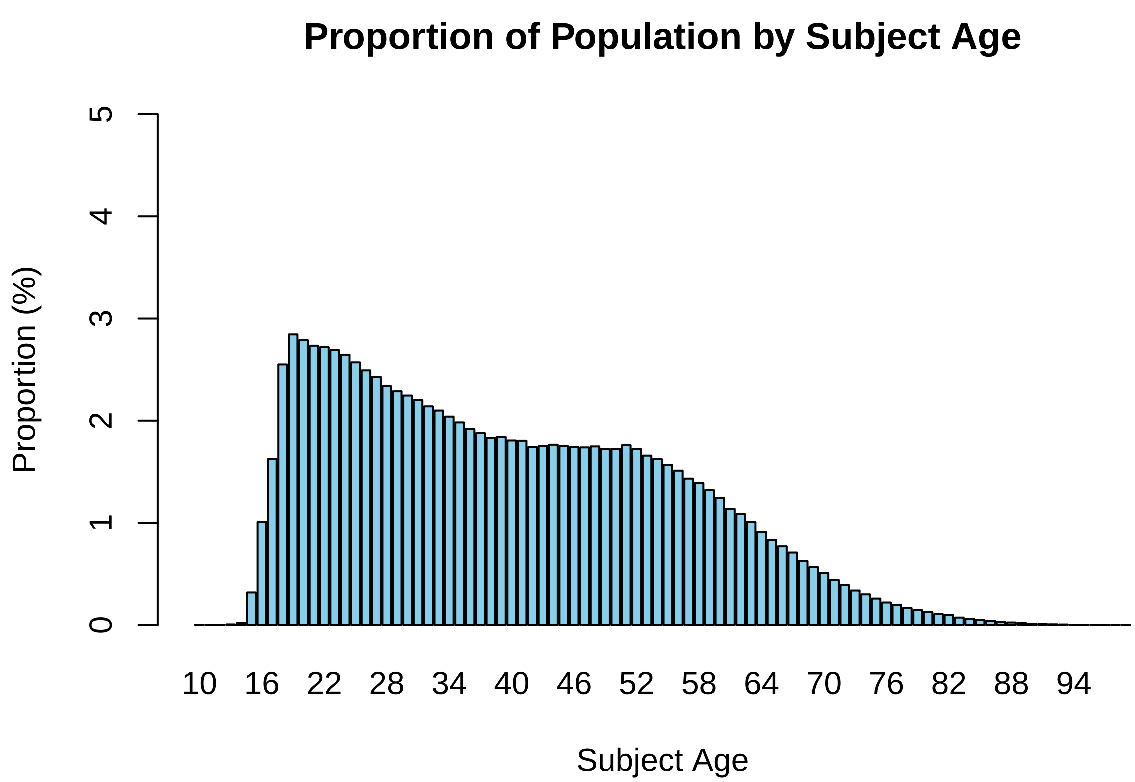



```r
# Calculate the total number of stops
total_warnings <- nrow(data_filtered)

# Calculate the proportion of warnings issued by subject age
proportion_by_age <- prop.table(table(data_filtered$subject_age)) * 100

# Print the proportion by age
print(proportion_by_age)
```

```
##
##           10           11           12           13           14           15
## 0.0007222555 0.0004815036 0.0010833832 0.0048150364 0.0191397697 0.3179127780
##           16           17           18           19           20           21
## 1.0077871176 1.6230283931 2.5501636510 2.8444827507 2.7880264490 2.7339776655
##           22           23           24           25           26           27
## 2.7188103008 2.6895589547 2.6453809958 2.5705071798 2.4921424625 2.4288247339
##           28           29           30           31           32           33
## 2.3367371628 2.2875034157 2.2460941027 2.2003512569 2.1399225501 2.0988743649
##           34           35           36           37           38           39
## 2.0394086654 1.9825912359 1.9187920037 1.8775030666 1.8311583413 1.8401865345
##           40           41           42           43           44           45
## 1.8057590243 1.8037126338 1.7407560329 1.7503861057 1.7649515908 1.7497842262
##           46           47           48           49           50           51
## 1.7399134016 1.7382281388 1.7477378357 1.7230607742 1.7239034056 1.7594142990
##           52           53           54           55           56           57
## 1.7217366392 1.6576966551 1.6236302726 1.5676554745 1.5105972932 1.4325937036
```

```
##            58            59            60            61            62            63
## 1.3888972483 1.3202829797 1.2419182623 1.1361078376 1.0841054445 1.0082686213
##            64            65            66            67            68            69
## 0.9106437583 0.8334828001 0.7698039438 0.7090141093 0.6253528519 0.5661279042
##            70            71            72            73            74            75
## 0.5101531061 0.4400943266 0.3886938130 0.3370525477 0.2993748879 0.2583267026
##            76            77            78            79            80            81
## 0.2196860356 0.1960923572 0.1644334929 0.1446918437 0.1253113222 0.1046066657
##            82            83            84            85            86            87
## 0.0961803520 0.0717440423 0.0598268272 0.0476688603 0.0404463057 0.0297328497
##            88            89            90            91            92            93
## 0.0235936783 0.0166118756 0.0115560873 0.0081855619 0.0054169159 0.0038520291
##            94            95            96            97            98            99
## 0.0019260146 0.0016852627 0.0008426314 0.0004815036 0.0001203759 0.0001203759
```

```r
# Create a bar plot
barplot(proportion_by_age,
        main = "Proportion of Population by Subject Age",
        ylab = "Proportion (%)",
        xlab = "Subject Age",
        col = "skyblue",
        ylim = c(0, 5))  # Adjust the y-axis limits if needed
```




```r
# Create a contingency table
linear_reg_table <- table(data_filtered$subject_age, data_filtered$warning_issued)
linear_reg_table
```

```
##
##      FALSE  TRUE
##   10     1     5
##   11     2     2
##   12     1     8
##   13    13    27
##   14    37   122
##   15   571  2070
##   16  1932  6440
##   17  3171 10312
##   18  5965 15220
##   19  6744 16886
##   20  6566 16595
##   21  6530 16182
##   22  6409 16177
##   23  5992 16351
##   24  5819 16157
##   25  5629 15725
##   26  5352 15351
##   27  5303 14874
##   28  5055 14357
##   29  5065 13938
##   30  4978 13681
##   31  4875 13404
##   32  4699 13078
##   33  4667 12769
##   34  4484 12458
##   35  4356 12114
##   36  4201 11739
##   37  4186 11411
##   38  4086 11126
##   39  4057 11230
##   40  4052 10949
##   41  4127 10857
##   42  3867 10594
##   43  4051 10490
##   44  3979 10683
##   45  3949 10587
##   46  4074 10380
##   47  4014 10426
##   48  3978 10541
##   49  3899 10415
##   50  3905 10416
##   51  3949 10667
##   52  3846 10457
##   53  3709 10062
##   54  3512  9976
##   55  3409  9614
##   56  3320  9229
##   57  3052  8849
##   58  3115  8423
##   59  2810  8158
##   60  2664  7653
##   61  2416  7022
```

```
##   62  2406  6600
##   63  2170  6206
##   64  1899  5666
##   65  1727  5197
##   66  1568  4827
##   67  1481  4409
##   68  1301  3894
##   69  1182  3521
##   70  1010  3228
##   71   917  2739
##   72   768  2461
##   73   640  2160
##   74   587  1900
##   75   494  1652
##   76   419  1406
##   77   356  1273
##   78   273  1093
##   79   232   970
##   80   222   819
##   81   177   692
##   82   145   654
##   83   112   484
##   84    80   417
##   85    73   323
##   86    57   279
##   87    32   215
##   88    30   166
##   89    22   116
##   90    16    80
##   91    15    53
##   92     6    39
##   93     4    28
##   94     2    14
##   95     5     9
##   96     1     6
##   97     0     4
##   98     0     1
##   99     0     1
```

```
# Convert the contingency table into a data frame
linear_reg_table_df <- as.data.frame.table(linear_reg_table)

# Rename the columns for clarity
names(linear_reg_table_df) <- c("Subject_Age", "Warning_Issued", "Frequency")

linear_reg_table_df
```

```
##     Subject_Age Warning_Issued Frequency
## 1            10          FALSE         1
## 2            11          FALSE         2
## 3            12          FALSE         1
## 4            13          FALSE        13
## 5            14          FALSE        37
## 6            15          FALSE       571
```

```
## 7            16          FALSE      1932
## 8            17          FALSE      3171
## 9            18          FALSE      5965
## 10           19          FALSE      6744
## 11           20          FALSE      6566
## 12           21          FALSE      6530
## 13           22          FALSE      6409
## 14           23          FALSE      5992
## 15           24          FALSE      5819
## 16           25          FALSE      5629
## 17           26          FALSE      5352
## 18           27          FALSE      5303
## 19           28          FALSE      5055
## 20           29          FALSE      5065
## 21           30          FALSE      4978
## 22           31          FALSE      4875
## 23           32          FALSE      4699
## 24           33          FALSE      4667
## 25           34          FALSE      4484
## 26           35          FALSE      4356
## 27           36          FALSE      4201
## 28           37          FALSE      4186
## 29           38          FALSE      4086
## 30           39          FALSE      4057
## 31           40          FALSE      4052
## 32           41          FALSE      4127
## 33           42          FALSE      3867
## 34           43          FALSE      4051
## 35           44          FALSE      3979
## 36           45          FALSE      3949
## 37           46          FALSE      4074
## 38           47          FALSE      4014
## 39           48          FALSE      3978
## 40           49          FALSE      3899
## 41           50          FALSE      3905
## 42           51          FALSE      3949
## 43           52          FALSE      3846
## 44           53          FALSE      3709
## 45           54          FALSE      3512
## 46           55          FALSE      3409
## 47           56          FALSE      3320
## 48           57          FALSE      3052
## 49           58          FALSE      3115
## 50           59          FALSE      2810
## 51           60          FALSE      2664
## 52           61          FALSE      2416
## 53           62          FALSE      2406
## 54           63          FALSE      2170
## 55           64          FALSE      1899
## 56           65          FALSE      1727
## 57           66          FALSE      1568
## 58           67          FALSE      1481
## 59           68          FALSE      1301
## 60           69          FALSE      1182
```

```
## 61            70          FALSE     1010
## 62            71          FALSE      917
## 63            72          FALSE      768
## 64            73          FALSE      640
## 65            74          FALSE      587
## 66            75          FALSE      494
## 67            76          FALSE      419
## 68            77          FALSE      356
## 69            78          FALSE      273
## 70            79          FALSE      232
## 71            80          FALSE      222
## 72            81          FALSE      177
## 73            82          FALSE      145
## 74            83          FALSE      112
## 75            84          FALSE       80
## 76            85          FALSE       73
## 77            86          FALSE       57
## 78            87          FALSE       32
## 79            88          FALSE       30
## 80            89          FALSE       22
## 81            90          FALSE       16
## 82            91          FALSE       15
## 83            92          FALSE        6
## 84            93          FALSE        4
## 85            94          FALSE        2
## 86            95          FALSE        5
## 87            96          FALSE        1
## 88            97          FALSE        0
## 89            98          FALSE        0
## 90            99          FALSE        0
## 91            10           TRUE        5
## 92            11           TRUE        2
## 93            12           TRUE        8
## 94            13           TRUE       27
## 95            14           TRUE      122
## 96            15           TRUE     2070
## 97            16           TRUE     6440
## 98            17           TRUE    10312
## 99            18           TRUE    15220
## 100           19           TRUE    16886
## 101           20           TRUE    16595
## 102           21           TRUE    16182
## 103           22           TRUE    16177
## 104           23           TRUE    16351
## 105           24           TRUE    16157
## 106           25           TRUE    15725
## 107           26           TRUE    15351
## 108           27           TRUE    14874
## 109           28           TRUE    14357
## 110           29           TRUE    13938
## 111           30           TRUE    13681
## 112           31           TRUE    13404
## 113           32           TRUE    13078
## 114           33           TRUE    12769
```

```
## 115              34            TRUE     12458
## 116              35            TRUE     12114
## 117              36            TRUE     11739
## 118              37            TRUE     11411
## 119              38            TRUE     11126
## 120              39            TRUE     11230
## 121              40            TRUE     10949
## 122              41            TRUE     10857
## 123              42            TRUE     10594
## 124              43            TRUE     10490
## 125              44            TRUE     10683
## 126              45            TRUE     10587
## 127              46            TRUE     10380
## 128              47            TRUE     10426
## 129              48            TRUE     10541
## 130              49            TRUE     10415
## 131              50            TRUE     10416
## 132              51            TRUE     10667
## 133              52            TRUE     10457
## 134              53            TRUE     10062
## 135              54            TRUE      9976
## 136              55            TRUE      9614
## 137              56            TRUE      9229
## 138              57            TRUE      8849
## 139              58            TRUE      8423
## 140              59            TRUE      8158
## 141              60            TRUE      7653
## 142              61            TRUE      7022
## 143              62            TRUE      6600
## 144              63            TRUE      6206
## 145              64            TRUE      5666
## 146              65            TRUE      5197
## 147              66            TRUE      4827
## 148              67            TRUE      4409
## 149              68            TRUE      3894
## 150              69            TRUE      3521
## 151              70            TRUE      3228
## 152              71            TRUE      2739
## 153              72            TRUE      2461
## 154              73            TRUE      2160
## 155              74            TRUE      1900
## 156              75            TRUE      1652
## 157              76            TRUE      1406
## 158              77            TRUE      1273
## 159              78            TRUE      1093
## 160              79            TRUE       970
## 161              80            TRUE       819
## 162              81            TRUE       692
## 163              82            TRUE       654
## 164              83            TRUE       484
## 165              84            TRUE       417
## 166              85            TRUE       323
## 167              86            TRUE       279
## 168              87            TRUE       215
```

```
## 169          88            TRUE       166
## 170          89            TRUE       116
## 171          90            TRUE        80
## 172          91            TRUE        53
## 173          92            TRUE        39
## 174          93            TRUE        28
## 175          94            TRUE        14
## 176          95            TRUE         9
## 177          96            TRUE         6
## 178          97            TRUE         4
## 179          98            TRUE         1
## 180          99            TRUE         1
```

```r
# Filter rows where Warning_Issued is TRUE
filtered_linear_reg_table_df <- linear_reg_table_df %>% filter(Warning_Issued == TRUE)

# Print the filtered data frame
print(filtered_linear_reg_table_df)
```

```
##    Subject_Age Warning_Issued Frequency
## 1           10           TRUE         5
## 2           11           TRUE         2
## 3           12           TRUE         8
## 4           13           TRUE        27
## 5           14           TRUE       122
## 6           15           TRUE      2070
## 7           16           TRUE      6440
## 8           17           TRUE     10312
## 9           18           TRUE     15220
## 10          19           TRUE     16886
## 11          20           TRUE     16595
## 12          21           TRUE     16182
## 13          22           TRUE     16177
## 14          23           TRUE     16351
## 15          24           TRUE     16157
## 16          25           TRUE     15725
## 17          26           TRUE     15351
## 18          27           TRUE     14874
## 19          28           TRUE     14357
## 20          29           TRUE     13938
## 21          30           TRUE     13681
## 22          31           TRUE     13404
## 23          32           TRUE     13078
## 24          33           TRUE     12769
## 25          34           TRUE     12458
## 26          35           TRUE     12114
## 27          36           TRUE     11739
## 28          37           TRUE     11411
## 29          38           TRUE     11126
## 30          39           TRUE     11230
## 31          40           TRUE     10949
## 32          41           TRUE     10857
## 33          42           TRUE     10594
## 34          43           TRUE     10490
```

```
## 35             44            TRUE     10683
## 36             45            TRUE     10587
## 37             46            TRUE     10380
## 38             47            TRUE     10426
## 39             48            TRUE     10541
## 40             49            TRUE     10415
## 41             50            TRUE     10416
## 42             51            TRUE     10667
## 43             52            TRUE     10457
## 44             53            TRUE     10062
## 45             54            TRUE      9976
## 46             55            TRUE      9614
## 47             56            TRUE      9229
## 48             57            TRUE      8849
## 49             58            TRUE      8423
## 50             59            TRUE      8158
## 51             60            TRUE      7653
## 52             61            TRUE      7022
## 53             62            TRUE      6600
## 54             63            TRUE      6206
## 55             64            TRUE      5666
## 56             65            TRUE      5197
## 57             66            TRUE      4827
## 58             67            TRUE      4409
## 59             68            TRUE      3894
## 60             69            TRUE      3521
## 61             70            TRUE      3228
## 62             71            TRUE      2739
## 63             72            TRUE      2461
## 64             73            TRUE      2160
## 65             74            TRUE      1900
## 66             75            TRUE      1652
## 67             76            TRUE      1406
## 68             77            TRUE      1273
## 69             78            TRUE      1093
## 70             79            TRUE       970
## 71             80            TRUE       819
## 72             81            TRUE       692
## 73             82            TRUE       654
## 74             83            TRUE       484
## 75             84            TRUE       417
## 76             85            TRUE       323
## 77             86            TRUE       279
## 78             87            TRUE       215
## 79             88            TRUE       166
## 80             89            TRUE       116
## 81             90            TRUE        80
## 82             91            TRUE        53
## 83             92            TRUE        39
## 84             93            TRUE        28
## 85             94            TRUE        14
## 86             95            TRUE         9
## 87             96            TRUE         6
## 88             97            TRUE         4
```

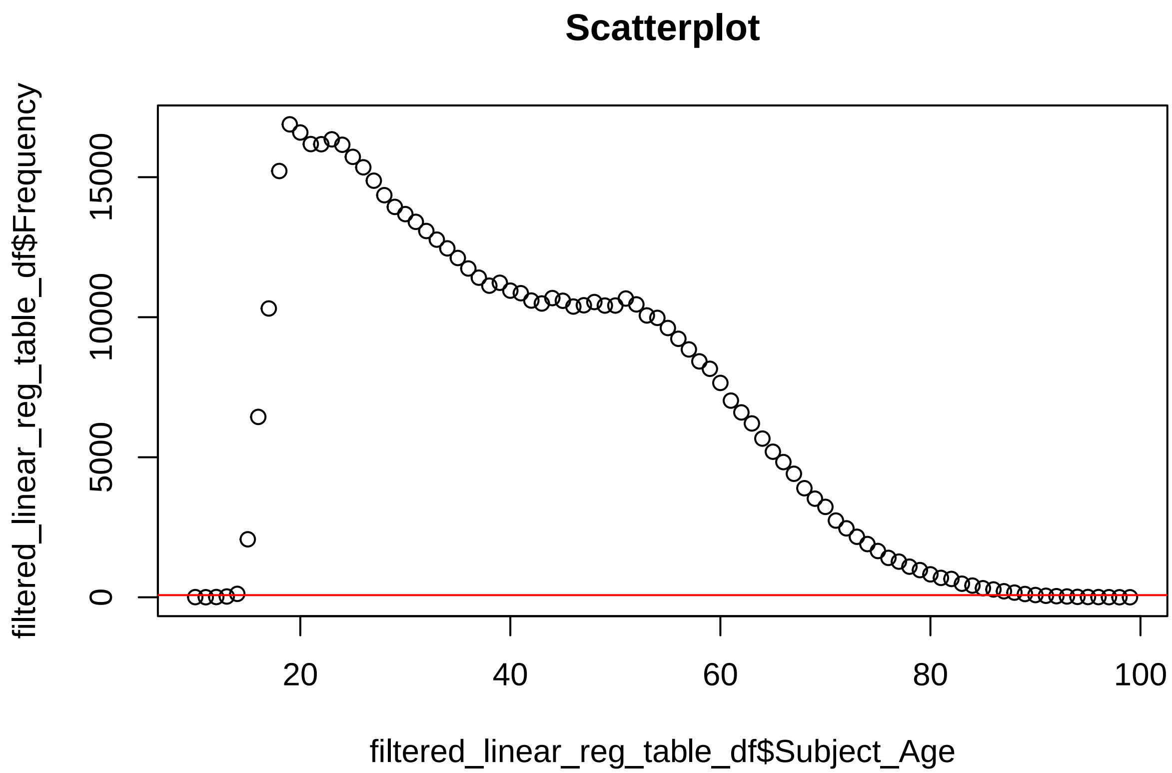
```
## 89          98          TRUE          1
## 90          99          TRUE          1
```

```r
# Ensure 'Subject_Age' and 'Frequency' are numeric
filtered_linear_reg_table_df$Subject_Age <- as.numeric(as.character(filtered_linear_reg_table_df$Subjec
filtered_linear_reg_table_df$Frequency <- as.numeric(as.character(filtered_linear_reg_table_df$Frequenc
```

Analysis

```r
# Create a scatterplot with jittering
ggplot(filtered_linear_reg_table_df, aes(x = Subject_Age, y = Frequency)) +
  geom_point(position = position_jitter(width = 0.1, height = 0.1)) +
  labs(title = "Scatterplot of Subject Age vs. Warnings Issued",
       x = "Subject Age", y = "Warnings Issued")
```


```r
plot(filtered_linear_reg_table_df$Subject_Age , filtered_linear_reg_table_df$Frequency, main="Scatterpl
abline(lm(Subject_Age ~ Frequency, data = filtered_linear_reg_table_df),col="red")
```




```r
# Fit the linear regression model
model <- lm(filtered_linear_reg_table_df$Subject_Age ~ filtered_linear_reg_table_df$Frequency)
```

```r
# Summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = filtered_linear_reg_table_df$Subject_Age ~ filtered_linear_reg_table_df$Frequency)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.862  -2.582   4.371   8.838  23.126
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            75.8774446  3.1247253  24.283  < 2e-16
## filtered_linear_reg_table_df$Frequency -0.0031548  0.0003535  -8.924 5.88e-14
##
## (Intercept)                            ***
## filtered_linear_reg_table_df$Frequency ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.04 on 88 degrees of freedom
```

```
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4691
## F-statistic: 79.64 on 1 and 88 DF,  p-value: 5.876e-14
```

With this table we can construct the least square regression line: Subject_Age=75.8774446-0.0031548×Frequency

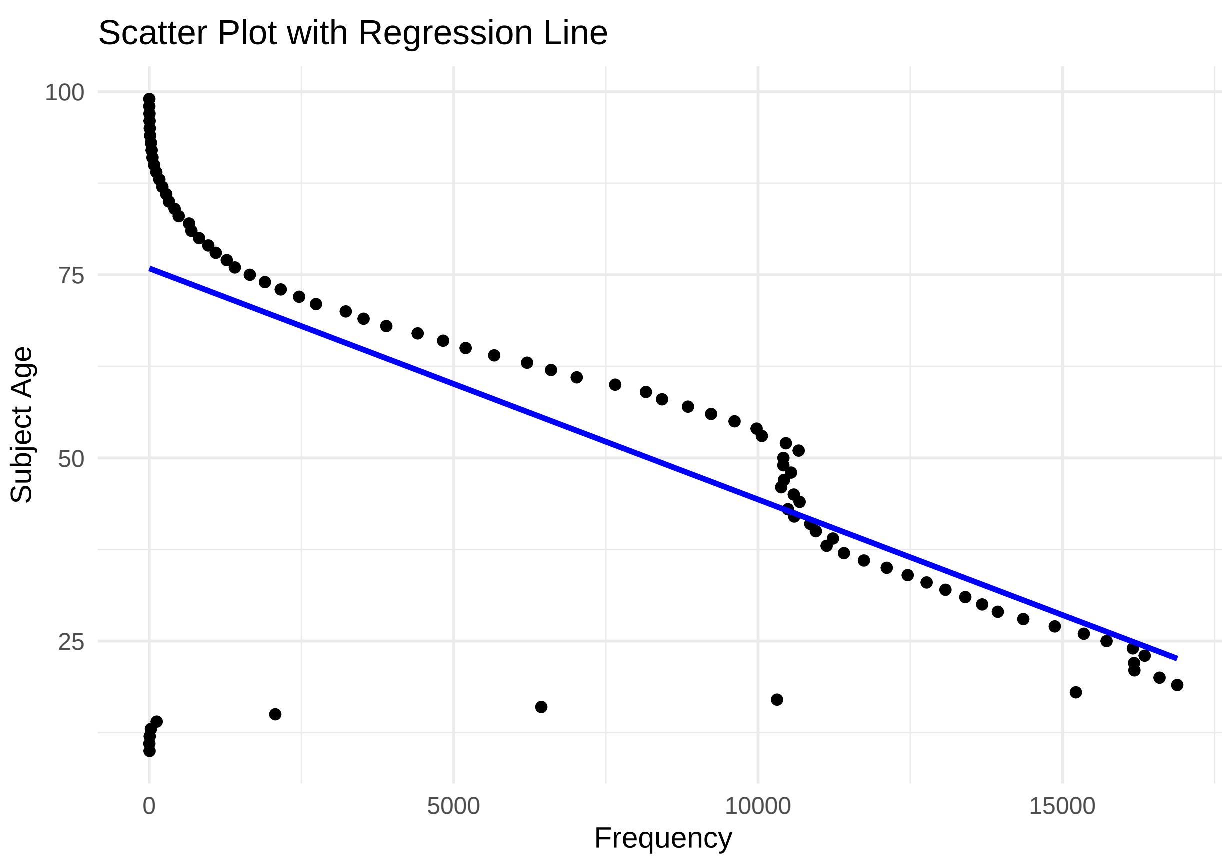
Where Frequency is the number of warnings issued corresponding to age.

Prediction and prediction errors

A scatterplot with the least squares line laid on top.

```
# Assuming filtered_contingency_df is your dataframe and it has columns 'Subject_Age' and 'Frequency'
# Create the scatter plot with regression line
ggplot(data = filtered_linear_reg_table_df, aes(x = Frequency, y = Subject_Age)) +
  geom_point() + # Scatter plot
  geom_smooth(method = "lm", se = FALSE, color = "blue") + # Regression line
  labs(title = "Scatter Plot with Regression Line",
       x = "Frequency",
       y = "Subject Age") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This line can be used to predict y at any value of x. When predictions are made for values of x that are beyond the range of the observed data, it is referred to as *extrapolation* and is not usually recommended.

However, predictions made within the range of the data are more reliable. They're also used to compute the residuals.

Model Diagnostics

```
correlation_coefficient <- cor(filtered_linear_reg_table_df$Subject_Age , filtered_linear_reg_table_df$I
correlation_coefficient
```

```
## [1] -0.6892581
```

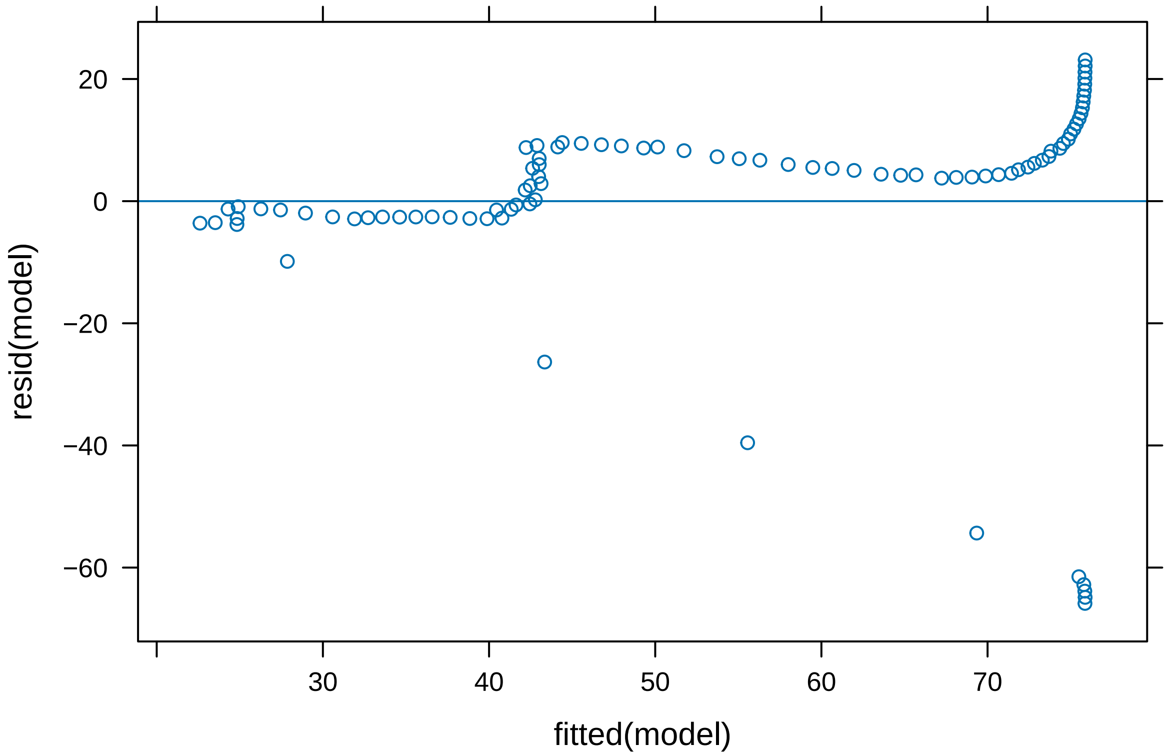
```
sum(residuals(model)^2)
```

```
## [1] 31885.15
```

1.) Is there a statistically significant relationship between the age of subjects and the likelihood of receiving a warning during a stop?\ With correlation coefficient = -0.6879639 and from the above plots we can say that subject_age and number of Warnings issued are negatively correlated and have relationship is moderate because correlation coefficient is not much closer to -1.

To check Equal Variance

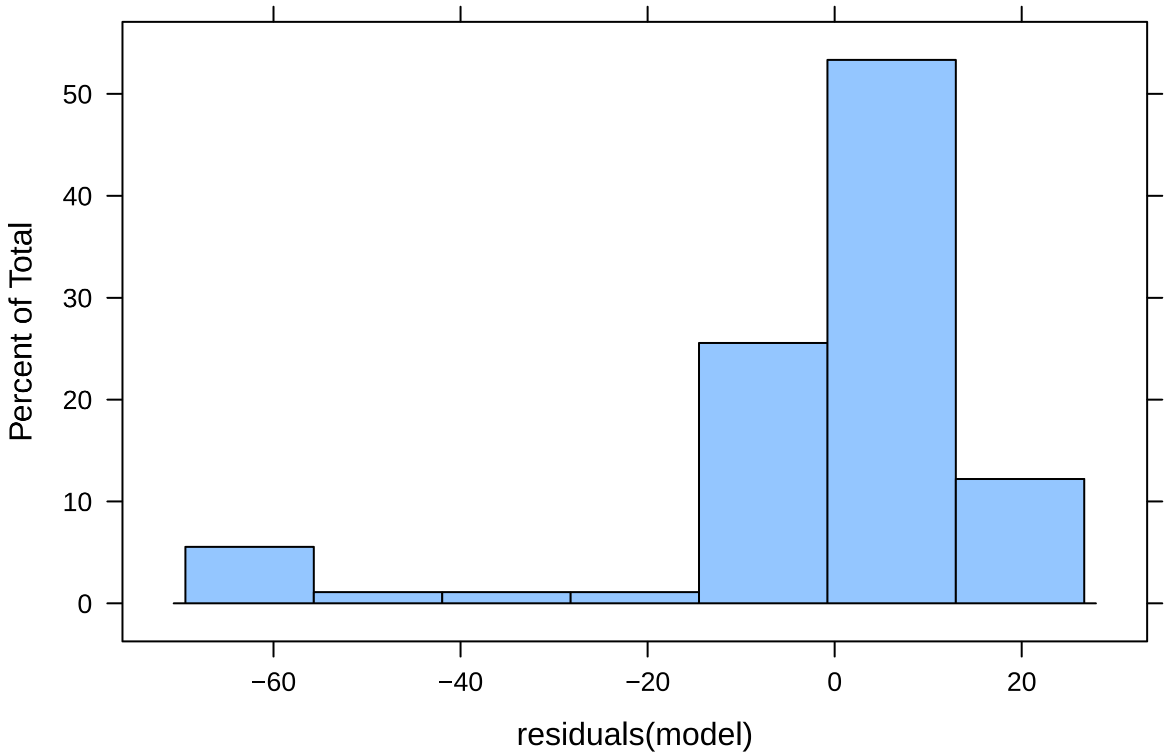
```
xyplot(resid(model) ~ fitted(model), data=filtered_linear_reg_table_df, type=c("p", "r"))
```



5.) Does the data follow Equal variance condition? From the plot Equal Variance is not met.

To check Normal Errors

```
histogram(~residuals(model), width=50)
```




```
qqmath(~resid(model))
ladd(panel.qqmathline(resid(model)))
```

2.) Are there any outliers in the data? 3.) Is there any influence or leverage of some instances? 4.) Does the data follow normal distribution? From the plot we can say that the model is normally distributed with few outliers but there is no high influence or high leverage.

Linear Regression Assumptions

Random Sampling : The data is collected randomly and this conditions is assumed to be met. Independence : This condition is also assumed to be met. From linear_reg_table, we can also see that expected cell frequencies is also met.

```
linear_reg_table
```

```
##
##      FALSE  TRUE
##   10     1     5
##   11     2     2
##   12     1     8
##   13    13    27
##   14    37   122
##   15   571  2070
##   16  1932  6440
```

```
##    17  3171 10312
##    18  5965 15220
##    19  6744 16886
##    20  6566 16595
##    21  6530 16182
##    22  6409 16177
##    23  5992 16351
##    24  5819 16157
##    25  5629 15725
##    26  5352 15351
##    27  5303 14874
##    28  5055 14357
##    29  5065 13938
##    30  4978 13681
##    31  4875 13404
##    32  4699 13078
##    33  4667 12769
##    34  4484 12458
##    35  4356 12114
##    36  4201 11739
##    37  4186 11411
##    38  4086 11126
##    39  4057 11230
##    40  4052 10949
##    41  4127 10857
##    42  3867 10594
##    43  4051 10490
##    44  3979 10683
##    45  3949 10587
##    46  4074 10380
##    47  4014 10426
##    48  3978 10541
##    49  3899 10415
##    50  3905 10416
##    51  3949 10667
##    52  3846 10457
##    53  3709 10062
##    54  3512  9976
##    55  3409  9614
##    56  3320  9229
##    57  3052  8849
##    58  3115  8423
##    59  2810  8158
##    60  2664  7653
##    61  2416  7022
##    62  2406  6600
##    63  2170  6206
##    64  1899  5666
##    65  1727  5197
##    66  1568  4827
##    67  1481  4409
##    68  1301  3894
##    69  1182  3521
##    70  1010  3228
```

```
##    71    917   2739
##    72    768   2461
##    73    640   2160
##    74    587   1900
##    75    494   1652
##    76    419   1406
##    77    356   1273
##    78    273   1093
##    79    232    970
##    80    222    819
##    81    177    692
##    82    145    654
##    83    112    484
##    84     80    417
##    85     73    323
##    86     57    279
##    87     32    215
##    88     30    166
##    89     22    116
##    90     16     80
##    91     15     53
##    92      6     39
##    93      4     28
##    94      2     14
##    95      5      9
##    96      1      6
##    97      0      4
##    98      0      1
##    99      0      1
```

Chi-Square Introduction: In this section we would like to study: 1.) How does the likelihood of receiving a warning vary across different age groups (e.g., youngsters, middle-aged, old)? 2.) Are there specific age ranges that are more likely to receive warnings compared to others? 3.) How does the rate of warnings issued to younger subjects compare to the rate of warnings issued to older subjects?

Chi-Square Test

Null Hypothesis (H0): There is no association between subject age groups and arrest made during the incidents.

Alternative Hypothesis (H1): There is an association between subject age groups and arrest made during the incidents.

Categorize Age groups

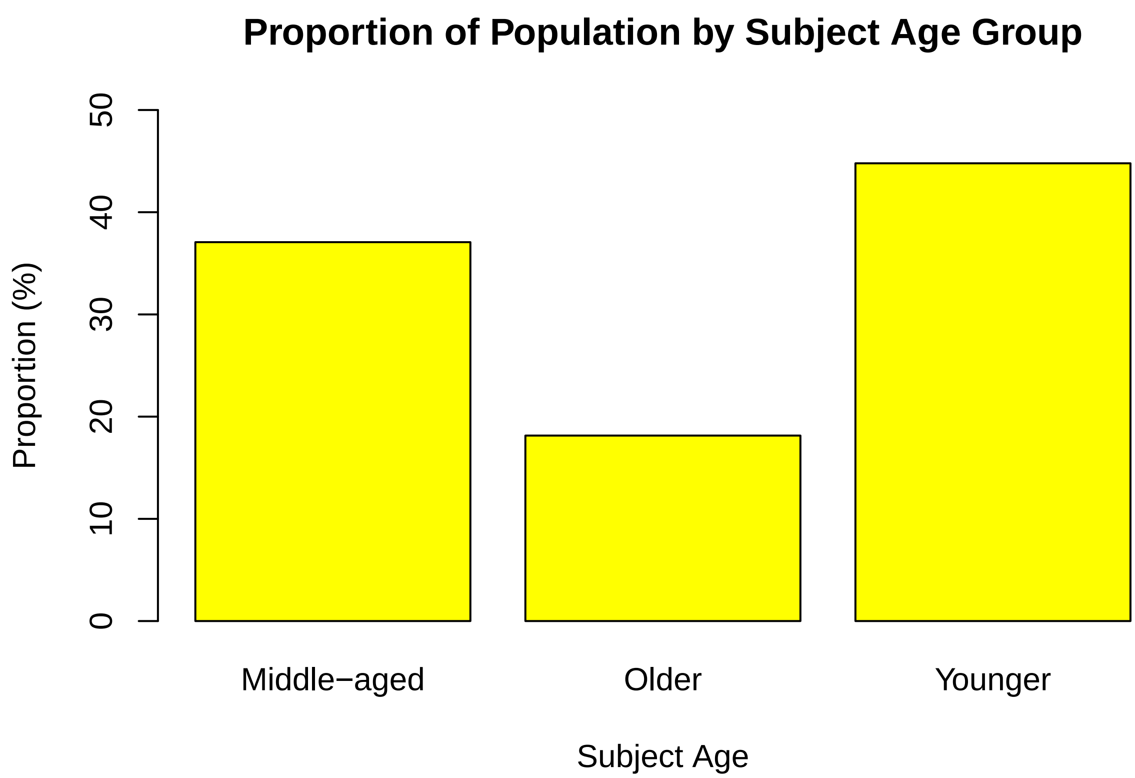
```
chi_sq_data <- data_filtered %>%
  mutate(subject_age = case_when(
    subject_age < 35 ~ "Younger",
    subject_age >= 35 & subject_age <= 55 ~ "Middle-aged",
    subject_age > 55 ~ "Older"
  ))
```

```r
# Calculate the total number of stops
total_warnings <- nrow(chi_sq_data)

# Calculate the proportion of warnings issued by subject age
proportion_by_age_group <- prop.table(table(chi_sq_data$subject_age)) * 100

# Print the proportion by age
print(proportion_by_age_group)
```

```
##
## Middle-aged       Older     Younger
##    37.06856    18.14571    44.78574
```

With approximately 37.06% falling within the middle-aged category, this segment represents a significant portion of the population. In contrast, the older age group, comprising about 18.14%, constitutes a smaller proportion. Conversely, the younger age group, with a proportion of approximately 44.78%, emerges as the largest segment, indicating a substantial presence within the population. Collectively, these proportions depict the age structure of the population, crucial for understanding demographic trends and informing various societal and policy considerations.

```r
# Create a bar plot
barplot(proportion_by_age_group,
        main = "Proportion of Population by Subject Age Group",
        ylab = "Proportion (%)",
        xlab = "Subject Age",
        col = "yellow",
        ylim = c(0, 50))  # Adjust the y-axis limits if needed
```




```r
# Create a contingency table
chi_sq_table <- table(chi_sq_data$subject_age, chi_sq_data$warning_issued)
chi_sq_table
```

```
##
##               FALSE   TRUE
##   Middle-aged  83206 224734
##   Older        37806 112936
##   Younger      99860 272189
```

1.) How does the likelihood of receiving a warning vary across different age groups (e.g., youngsters, middle-aged, old)? Out of 410264 young drivers, 73.16% received warning. Out of 341298 middle-aged drivers, 72.97% received warning. Out of 165853 older drivers, 74.92% received warning.

2.) Are there specific age ranges that are more likely to receive warnings compared to others? From this data we can say that middle-aged people more likely to receive warnings compared others.

Analysis

```r
# Perform the chi-square test of independence
chi_sq_test <- chisq.test(chi_sq_table)

# Print the result
print(chi_sq_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chi_sq_table
## X-squared = 217.27, df = 2, p-value < 2.2e-16
```

Given the p-value is significantly less than 0.05, we reject the null hypothesis. This means: There is strong evidence to suggest that there is a significant association between the age groups (subject_age) and whether a warning was issued (warning_issued).

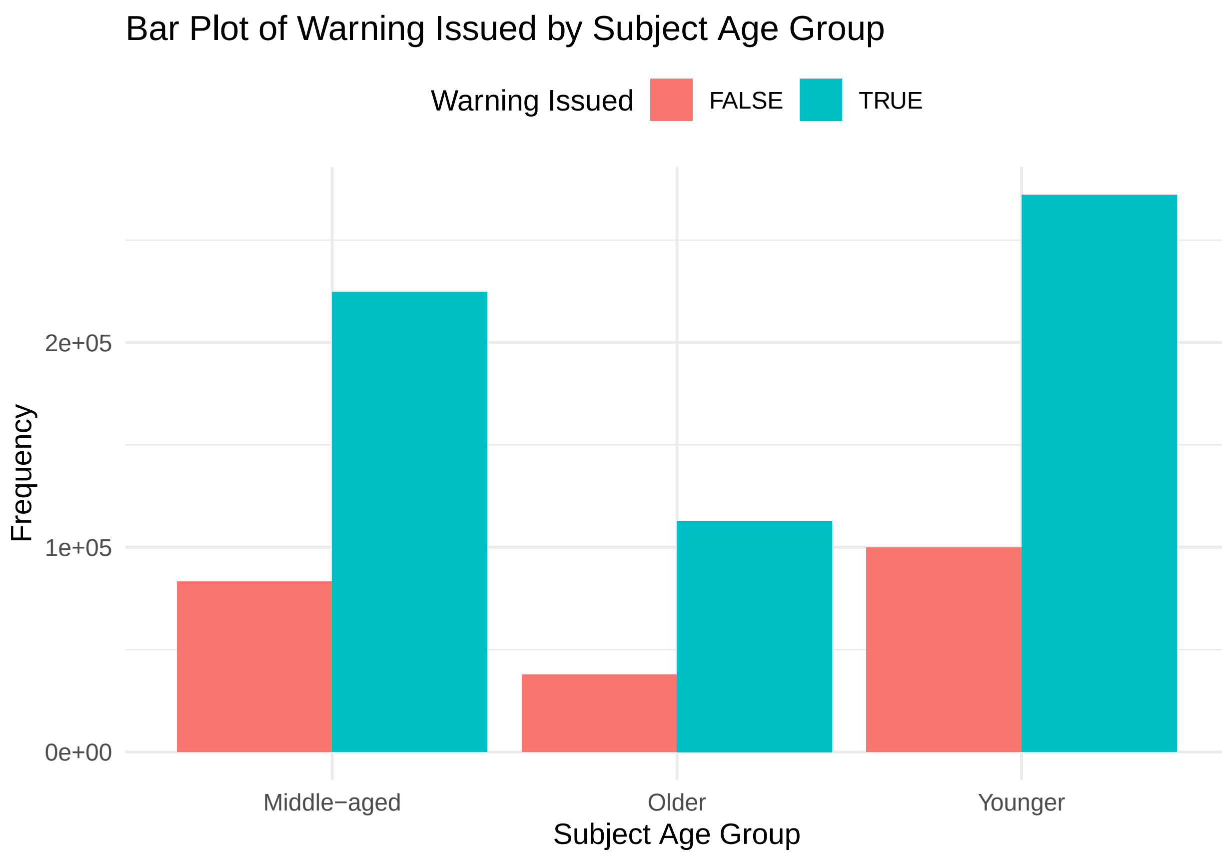
```r
# Convert the contingency table into a data frame
chi_sq_table_df <- as.data.frame.table(chi_sq_table)

# Rename the columns for clarity
names(chi_sq_table_df) <- c("Subject_Age", "Warning_Issued", "Frequency")

chi_sq_table_df
```

```
##   Subject_Age Warning_Issued Frequency
## 1 Middle-aged          FALSE     83206
## 2       Older          FALSE     37806
## 3     Younger          FALSE     99860
## 4 Middle-aged           TRUE    224734
## 5       Older           TRUE    112936
## 6     Younger           TRUE    272189
```

```r
# Ensure 'Frequency' is numeric
chi_sq_table_df$Frequency <- as.numeric(as.character(chi_sq_table_df$Frequency))
```

The below bar graph depicts the clear relationship between the warnings_issued and the subject_age of the driver

```r
# Create a bar plot
ggplot(chi_sq_table_df, aes(x = Subject_Age, y = Frequency, fill = Warning_Issued)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Bar Plot of Warning Issued by Subject Age Group",
       x = "Subject Age Group",
       y = "Frequency",
       fill = "Warning Issued") +
  theme_minimal() +
  theme(legend.position = "top")
```



3.) How does the rate of warnings issued to younger subjects compare to the rate of warnings issued to older subjects? We can see that younger people have received warnings more the 2x the warnings received by older people.

Chi - Square Assumptions:

Random Sampling : The data is collected randomly and this conditions is assumed to be met. Independence : This condition is also assumed to be met. Counted Data Condition: this condition is met as we have frequencies of individual categories. From chi_sq_table, we can also see that expected cell frequencies is also met.

```
chi_sq_table
```

```
##
##               FALSE   TRUE
##   Middle-aged 83206 224734
##   Older       37806 112936
##   Younger     99860 272189
```

Is the mean age of the drivers who got arrested same as the mean age of driver got received warning ?

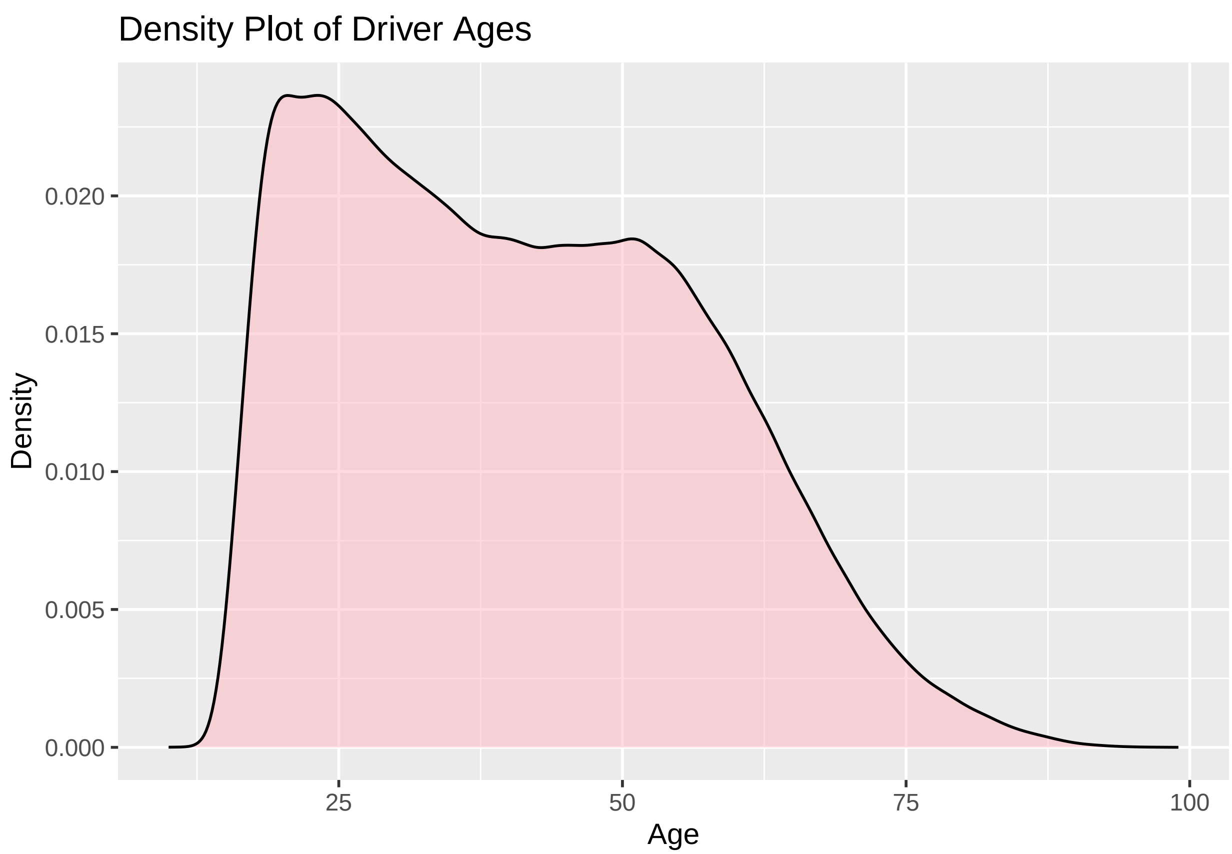
Considering that an arrest is more severe than a warning, it is likely possible that a driver would have received a warning before getting arrested. If younger population is more likely to be arrested, the law can enforce programs in schools to educate students on violations and address the specific behavior.

Since only arrests and warnings are studied, we have removed the traffic stops against citations

```
arrests_warning_filtered <- data_filtered %>%
  filter(outcome %in% c("arrest", "warning"))
```

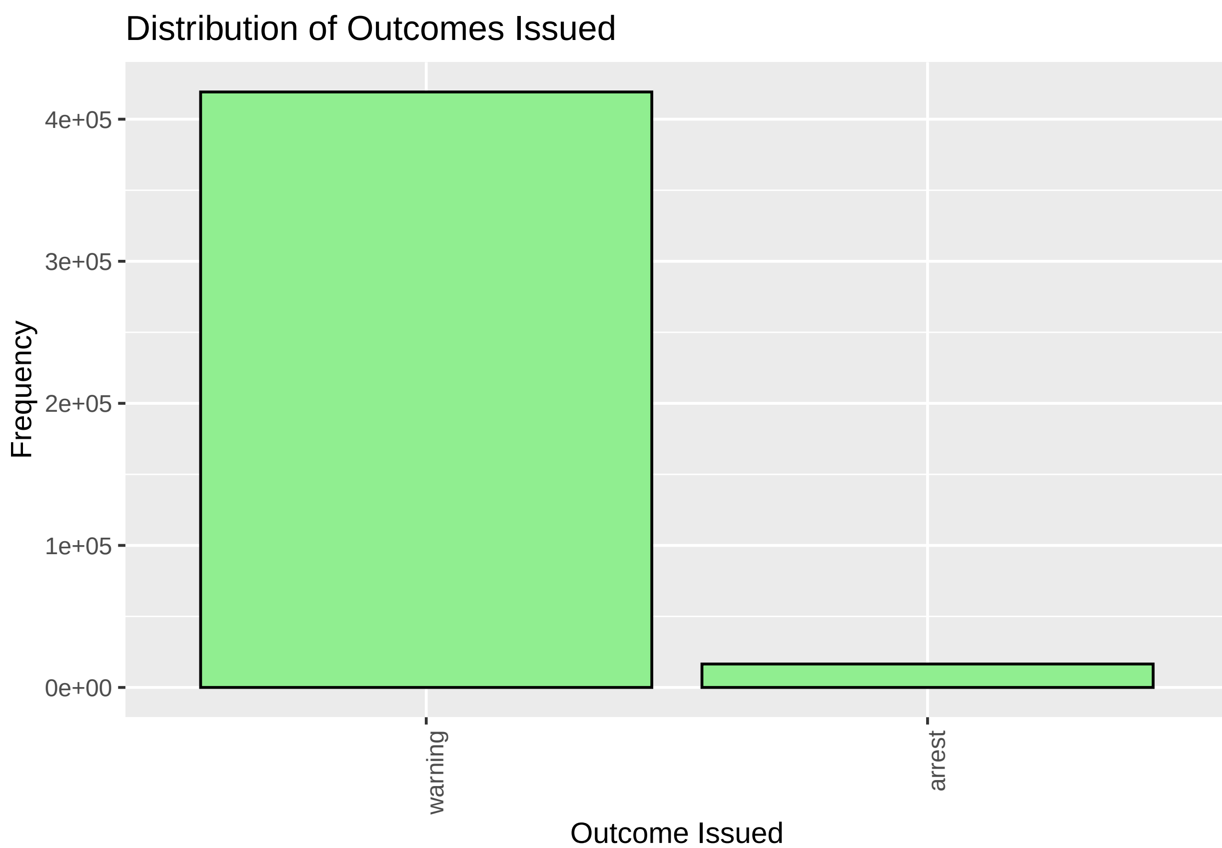
Analyzing the columns "subject_age" and "outcome", we observe that age is numerical, continuous data where outcome is categorical data with only 2 values.

```
ggplot(arrests_warning_filtered, aes(x = subject_age)) +
  geom_density(fill = "lightpink", alpha = 0.5) +
  labs(title = "Density Plot of Driver Ages", x = "Age", y = "Density")
```



The graph is right-skewed with two peaks, and the age of majority of the drivers are in the range of 20-50 years.

```
ggplot(arrests_warning_filtered, aes(x = outcome)) +
  geom_bar(fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Outcomes Issued", x = "Outcome Issued", y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



To run the test, a few assumptions are made:

1) Data is sampled randomly
2) Data is independent of one another
3) Large sample size

Hypothesis

Problem Statement : The average age of drivers involved in traffic stops that result in arrests does not differ from the average age of drivers involved in stops that result in warnings.

Null Hypothesis : True difference in means between group warning and group arrest is equal to 0

Alternate Hypothesis : True difference in means between group warning and group arrest is not equal to 0.

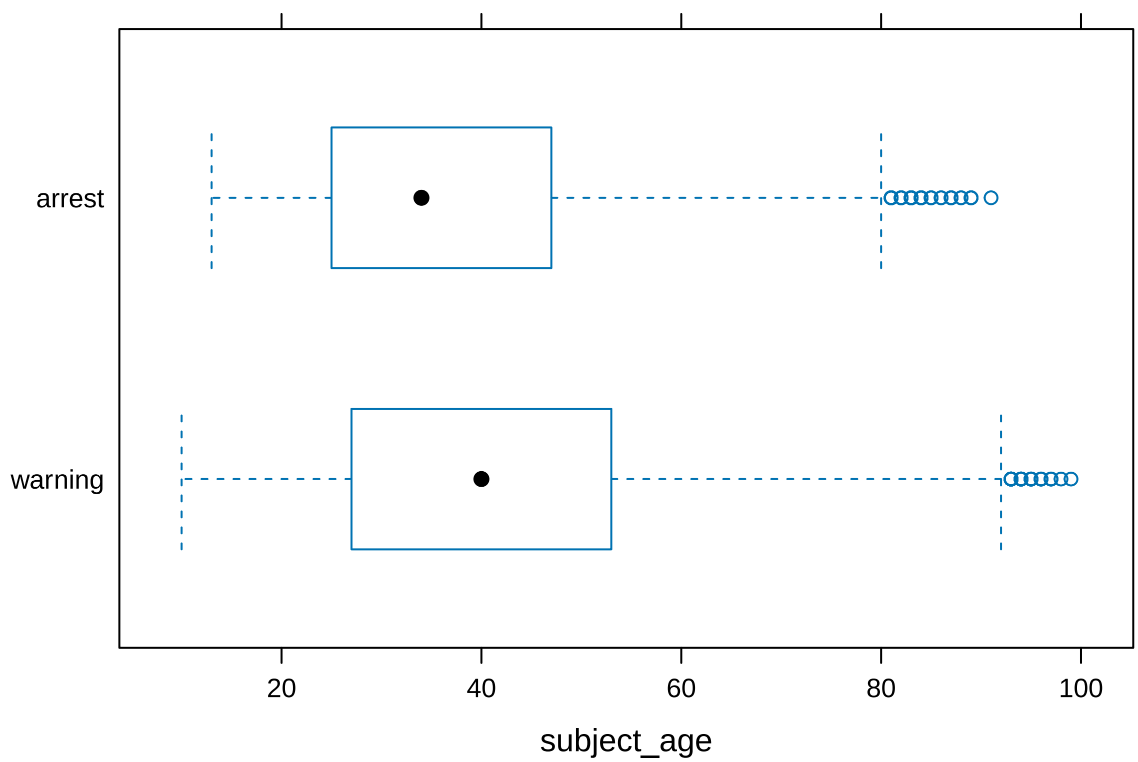
2 tests are run to study, t-test and anova

```
arrests_warning_test <- t.test(subject_age ~ outcome, data = arrests_warning_filtered)
print(arrests_warning_test)
```

Method 1 : T - test

```
##
##  Welch Two Sample t-test
##
## data:  subject_age by outcome
## t = 36.469, df = 18160, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group warning and group arrest is not equal
## 95 percent confidence interval:
##  3.980863 4.433083
## sample estimates:
## mean in group warning  mean in group arrest
##              41.06908              36.86211
```

```
bwplot(outcome ~ subject_age, data=arrests_warning_filtered)
```



Interpretation: Since 0 does not lie in the confidence interval, the difference in mean can never be 0, therefore rejecting the null hypothesis. The average age of drivers involved in traffic stops that result in arrests differs significantly from the average age of drivers involved in stops that result in warnings.

```
anova_arrests_warning_anova <- aov(subject_age ~ outcome, data = arrests_warning_filtered)
summary(anova_arrests_warning_anova)
```

Method 2 : Anova

```
##                  Df    Sum Sq Mean Sq F value Pr(>F)
## outcome           1    281180  281180    1092 <2e-16 ***
## Residuals    435683 112181886     257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anova_arrests_warning_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = subject_age ~ outcome, data = arrests_warning_filtered)
##
## $outcome
##                      diff      lwr       upr p adj
## arrest-warning -4.206973 -4.45649 -3.957455     0
```

Interpretation: With very high f-value and very less p-value, the anova results are rejecting the null hypothesis. The difference in the mean is nearly 4.2 years and we are 95% confident that the difference in the age lies between 3.95 to 4.45 years.

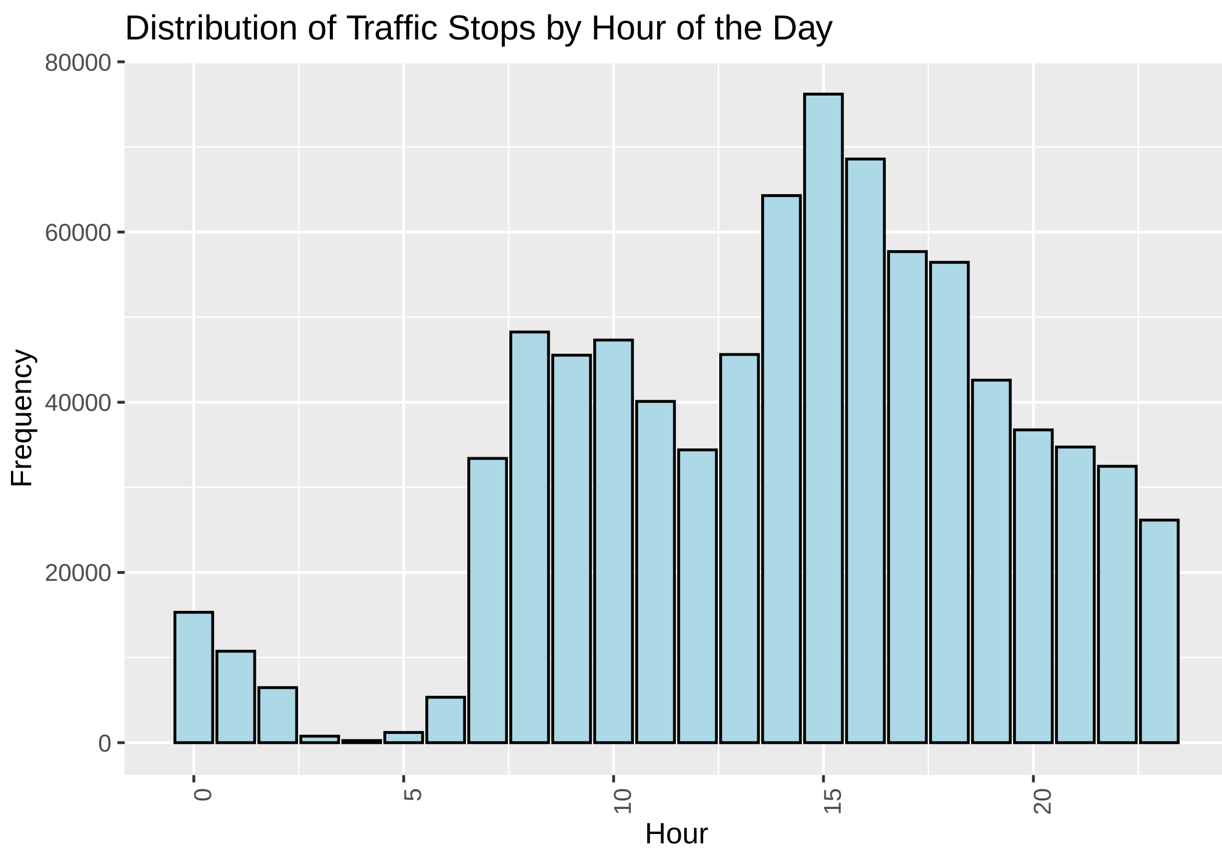
Is the time of the day a factor in determining the outcome of the traffic stop?

To ease the analysis, we have grouped the time such that all the traffic stops that have occurred in an hour will be group to it's corresponding hour. For example, if the traffic stop is issued at "02:45:89", the value under the column "hour" will be 2. Also we have considered only the one violated against every event that has occurred, assuming that the first violation entered has the highest severity.

```
data_filtered$hour <- hour(data_filtered$time)
remove_alpha_rows <- function(data) {
  alpha_rows <- grep("^[a-zA-Z]", data$violation)
  if (length(alpha_rows) > 0) {
    data <- data[-alpha_rows, , drop = FALSE]
  }
  return(data)
}

retrieve_values_until_pipe <- function(data) {
  split_values <- strsplit(data$violation, "|", fixed = TRUE)
  data$violation <- sapply(split_values, "[[", 1)
  return(data)
}

data_filtered <- remove_alpha_rows(data_filtered)
#mt_data_filtered$violation_code <- substring(mt_data_filtered$violation, 1, 3)
data_filtered <- retrieve_values_until_pipe(data_filtered)
```

Distribution of traffic stops by hour of the day

```r
ggplot(data_filtered, aes(x = hour)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "Distribution of Traffic Stops by Hour of the Day", x = "Hour", y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



To further study the time of the day, we have to group the time into 4 categories, namely, Night, Morning, Afternoon and Evening.

```r
data_filtered$hour_category <- cut(data_filtered$hour, breaks = c(-Inf, 6, 12, 17, 20, Inf), labels = c
```

```r
summary(data_filtered)
```

```
##       date                 time           county_name         subject_sex
##  Min.   :2009-01-01   Length:830611      Length:830611      male  :558387
##  1st Qu.:2011-11-10   Class1:hms         Class :character   female:272224
##  Median :2013-11-09   Class2:difftime    Mode  :character
##  Mean   :2013-11-14   Mode  :numeric
##  3rd Qu.:2015-10-15
##  Max.   :2017-12-31
##   subject_age    citation_issued warning_issued  arrest_made
##  Min.   :10.00   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.:26.00   FALSE:433249    FALSE:220829    FALSE:814107
##  Median :37.00   TRUE :397362    TRUE :609782    TRUE :16504
##  Mean   :39.37
##  3rd Qu.:51.00
```

```
##  Max.   :99.00
##      outcome       frisk_performed search_conducted reason_for_stop
##  warning :419117   Mode :logical   Mode :logical    Length:830611
##  citation:394990   FALSE:830598    FALSE:827218     Class :character
##  summons :      0   TRUE :13        TRUE :3393       Mode  :character
##  arrest  : 16504
##
##
##  vehicle_make       vehicle_model      vehicle_type        vehicle_year
##  Length:830611      Length:830611      Length:830611      Min.   :1915
##  Class :character   Class :character   Class :character   1st Qu.:2000
##  Mode  :character   Mode  :character   Mode  :character   Median :2005
##                                                             Mean   :2004
##                                                             3rd Qu.:2009
##                                                             Max.   :2019
##   violation              hour         hour_category
##  Length:830611      Min.   : 0.00   Night     :133447
##  Class :character   1st Qu.:10.00   Morning   :248992
##  Mode  :character   Median :15.00   Afternoon:312382
##                     Mean   :14.14   Evening   :135790
##                     3rd Qu.:18.00
##                     Max.   :23.00
```

Since there are no records of summons issued, lets remove the label summons from outcome column.

```r
# Drop unused levels from the factor
data_filtered$outcome <- droplevels(data_filtered$outcome)
print(unique(data_filtered$outcome))
```

```
## [1] citation warning  arrest
## Levels: warning citation arrest
```

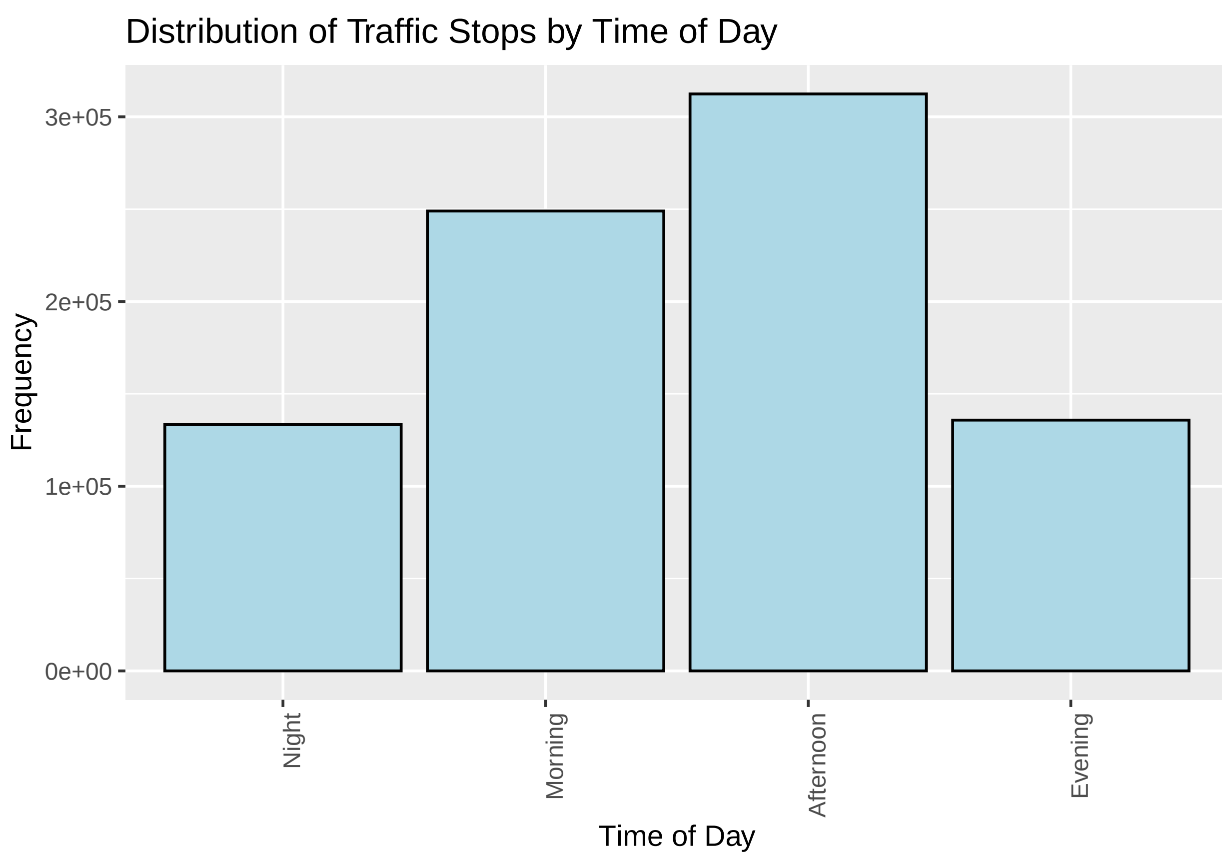
```r
summary(data_filtered)
```

```
##       date                 time           county_name        subject_sex
##  Min.   :2009-01-01   Length:830611      Length:830611      male  :558387
##  1st Qu.:2011-11-10   Class1:hms         Class :character   female:272224
##  Median :2013-11-09   Class2:difftime    Mode  :character
##  Mean   :2013-11-14   Mode  :numeric
##  3rd Qu.:2015-10-15
##  Max.   :2017-12-31
##   subject_age    citation_issued warning_issued  arrest_made
##  Min.   :10.00   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.:26.00   FALSE:433249    FALSE:220829    FALSE:814107
##  Median :37.00   TRUE :397362    TRUE :609782    TRUE :16504
##  Mean   :39.37
##  3rd Qu.:51.00
##  Max.   :99.00
##      outcome       frisk_performed search_conducted reason_for_stop
##  warning :419117   Mode :logical   Mode :logical    Length:830611
##  citation:394990   FALSE:830598    FALSE:827218     Class :character
##  arrest  : 16504   TRUE :13        TRUE :3393       Mode  :character
##
```

```
##
##
##   vehicle_make       vehicle_model      vehicle_type        vehicle_year
##   Length:830611      Length:830611      Length:830611      Min.   :1915
##   Class :character   Class :character   Class :character   1st Qu.:2000
##   Mode  :character   Mode  :character   Mode  :character   Median :2005
##                                                            Mean   :2004
##                                                            3rd Qu.:2009
##                                                            Max.   :2019
##    violation              hour         hour_category
##   Length:830611      Min.   : 0.00   Night    :133447
##   Class :character   1st Qu.:10.00   Morning  :248992
##   Mode  :character   Median :15.00   Afternoon:312382
##                      Mean   :14.14   Evening  :135790
##                      3rd Qu.:18.00
##                      Max.   :23.00
```

Considering that the time of the day can be a factor in determining the outcome of the traffic stop, we have to study the columns "hour_category" and "outcome", where the former is categorical and the latter is also categorical with 3 values.

```r
ggplot(data_filtered, aes(x = hour_category)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "Distribution of Traffic Stops by Time of Day", x = "Time of Day", y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



To run the test, a few assumptions are made:

1) Data is sampled randomly
2) Data is independent of one another
3) Large sample size

Hypothesis

Problem Statement : The time of the day is not a factor in determining the outcome of the traffic stop.

Null Hypothesis : The time of the day is not a factor in determining the outcome of the traffic stop.

Alternate Hypothesis : The time of the day is a factor in determining the outcome of the traffic stop.

Chi-square test

```
time_outcome_table <- table(data_filtered$hour_category, data_filtered$outcome)
time_outcome_table
```

```
##
##               warning citation arrest
##   Night         78298    52139   3010
##   Morning      121580   123207   4205
##   Afternoon    148155   157967   6260
##   Evening       71084    61677   3029
```

```
chisq.test(time_outcome_table)
```

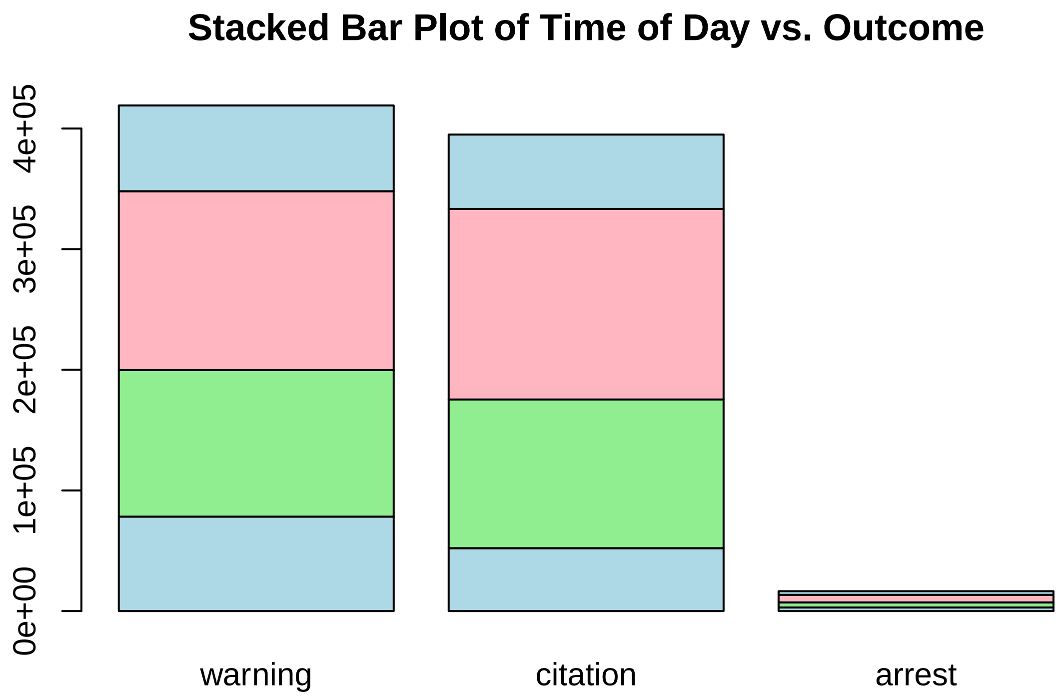
```
##
##  Pearson's Chi-squared test
##
## data:  time_outcome_table
## X-squared = 5722.6, df = 6, p-value < 2.2e-16
```

Interpretation: With a p-value even less than 0.0001, we reject the null hypothesis. The time of the day is a factor in determining the outcome of the traffic stop.

Visualization

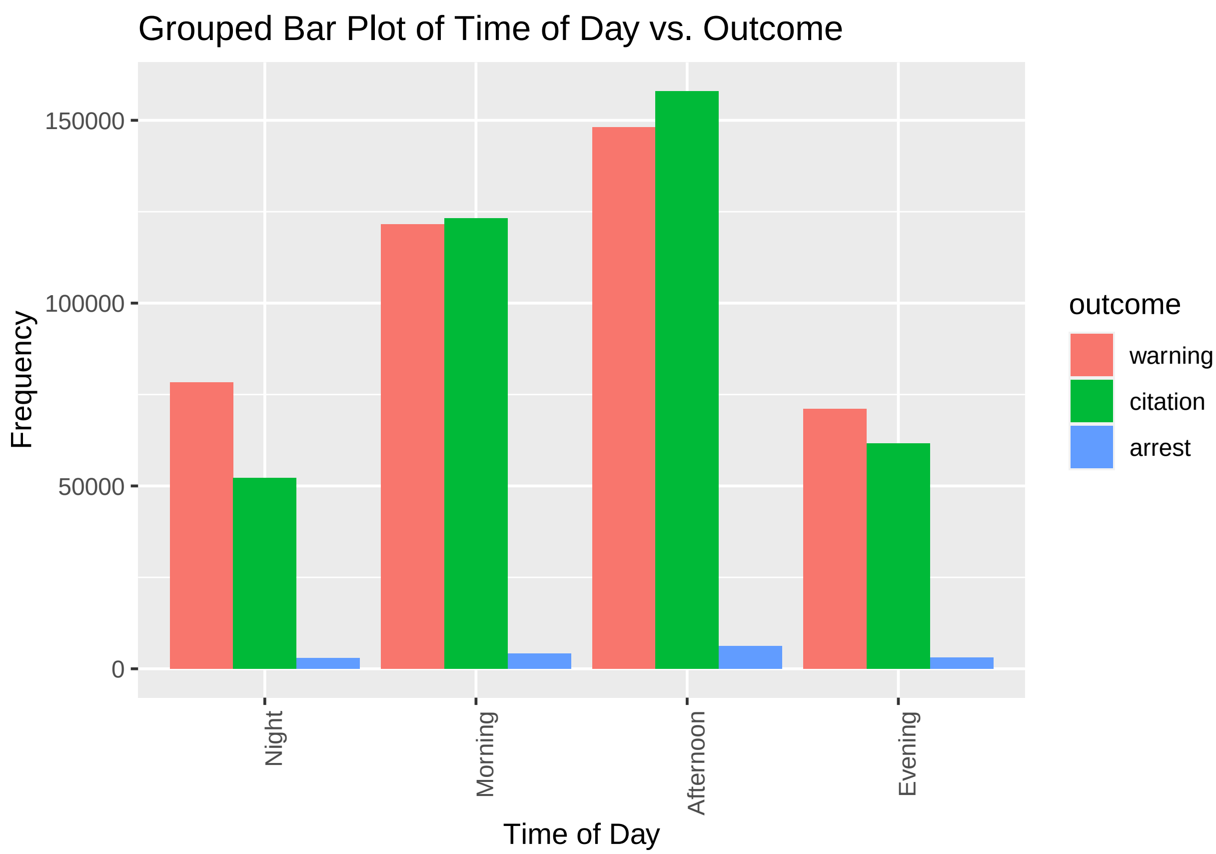
Using a stacked bar plot, we can visualize the relationship between the time of the day and the outcome of the traffic stop.

```
barplot(time_outcome_table, main = "Stacked Bar Plot of Time of Day vs. Outcome", col = c("lightblue",
```



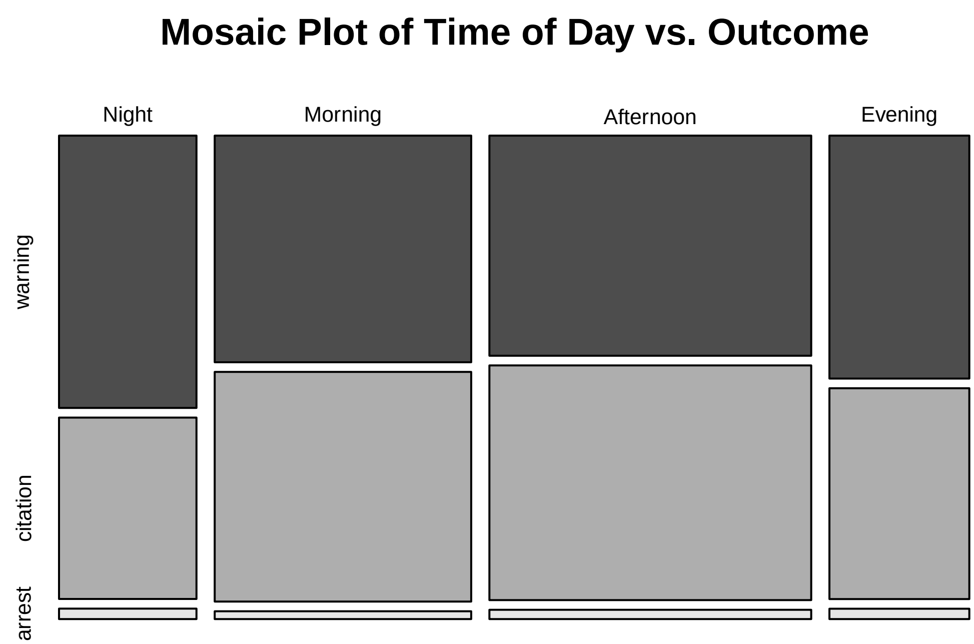
Using a grouped bar plot, we can visualize the relationship between the time of the day and the outcome of the traffic stop.

```
ggplot(data_filtered, aes(x = hour_category, fill = outcome)) +
  geom_bar(position = "dodge") +
  labs(title = "Grouped Bar Plot of Time of Day vs. Outcome", x = "Time of Day", y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



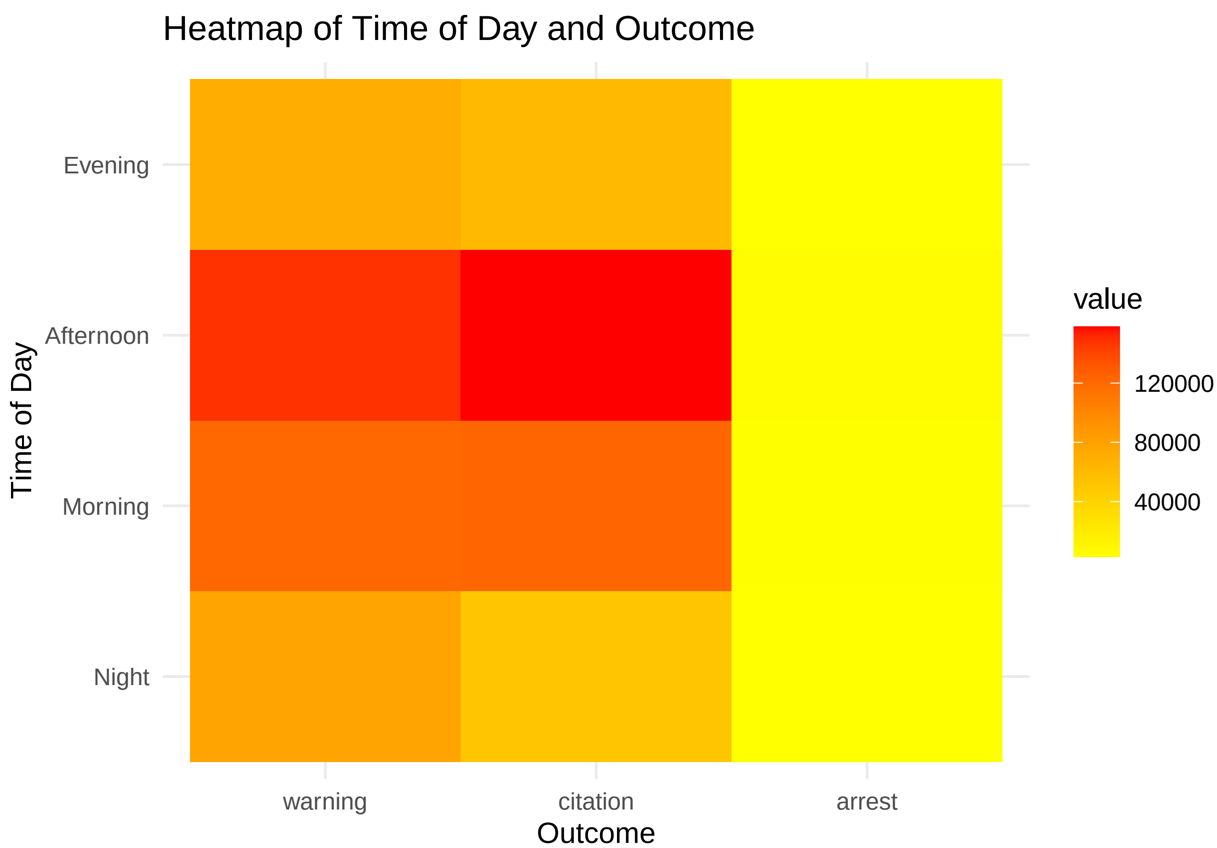
Using a mosaic plot, we can visualize the relationship between the time of the day and the outcome of the traffic stop.

```
mosaicplot(time_outcome_table, main = "Mosaic Plot of Time of Day vs. Outcome", color = TRUE)
```



Using a heat map, we can visualize the relationship between the time of the day and the outcome of the traffic stop.

```
time_melted <- melt(time_outcome_table)

# Plotting the heatmap
ggplot(time_melted, aes(x=Var2, y=Var1, fill=value)) +
  geom_tile() +
  scale_fill_gradient(low="yellow", high="red") +
  labs(title="Heatmap of Time of Day and Outcome", x="Outcome", y="Time of Day") +
  theme_minimal()
```



For this hypothesis, there is a large skew in the data, with most of the traffic stops occurring in the morning and afternoon. This could be due to various factors such as rush hour traffic, school zones, and work schedules.

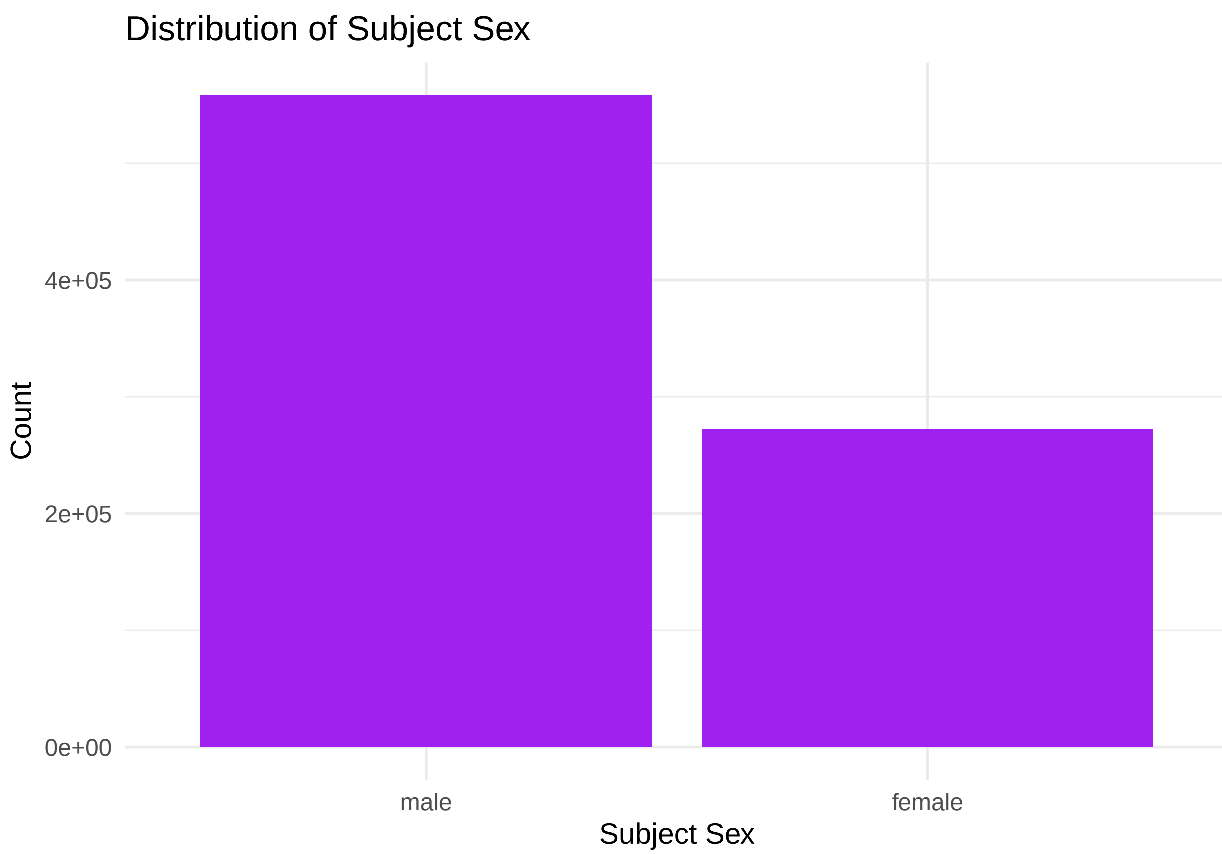
The outcome of the traffic stops is also skewed, with most stops resulting in warnings and citations.

The chi-square test results indicate that the time of the day is a factor in determining the outcome of the traffic stop. This could be due to various factors such as law enforcement practices, traffic patterns, and driver behavior at different times of the day.

But given the skew in the data, further analysis is needed to determine the specific factors that influence the outcome of traffic stops at different times of the day. As none of the plots are showing a clear relationship between the time of the day and the outcome of the traffic stop, further analysis is needed to understand the underlying patterns and trends in the data.

Are female drivers less at risk for violations compared to male drivers?

```
library(ggplot2)

ggplot(data_filtered, aes(x = subject_sex)) +
  geom_bar(fill = "purple") +
  labs(title = "Distribution of Subject Sex", x = "Subject Sex", y = "Count") +
  theme_minimal()
```



To run the test, we have made a few assumptions. They are:

1) We have randomly sampled data.
2) Data is independent of one another
3) We have Large sample size of data

Hypothesis

Null Hypothesis : Female drivers are more at risk for violations compared to male drivers.

Alternate Hypothesis : Female drivers are less at risk for violations compared to male drivers.

```
risk_table <- table(data_filtered$subject_sex, data_filtered$outcome)
columns_with_zeros <- apply(risk_table, 2, function(col) all(col == 0))
risk_table <- risk_table[, !columns_with_zeros]
print(risk_table)
```

```
##
##          warning citation arrest
##   male    277242   269677  11468
##   female  141875   125313   5036
```

```r
chi_sq_test <- chisq.test(risk_table)
print(chi_sq_test)
```

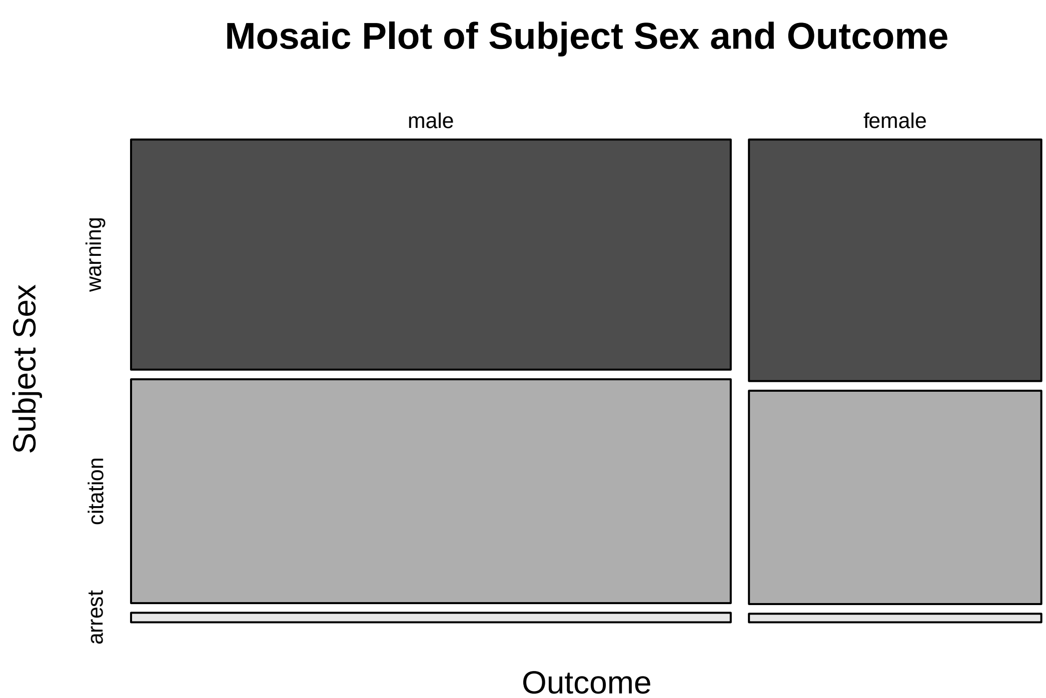
```
##
##  Pearson's Chi-squared test
##
## data:  risk_table
## X-squared = 455.93, df = 2, p-value < 2.2e-16
```

Interpretation: With a very small p-value as shown in the results above, we can reject the null hypothesis, and conclude that female drivers are less at risk for violations compared to male drivers.

Visualizations

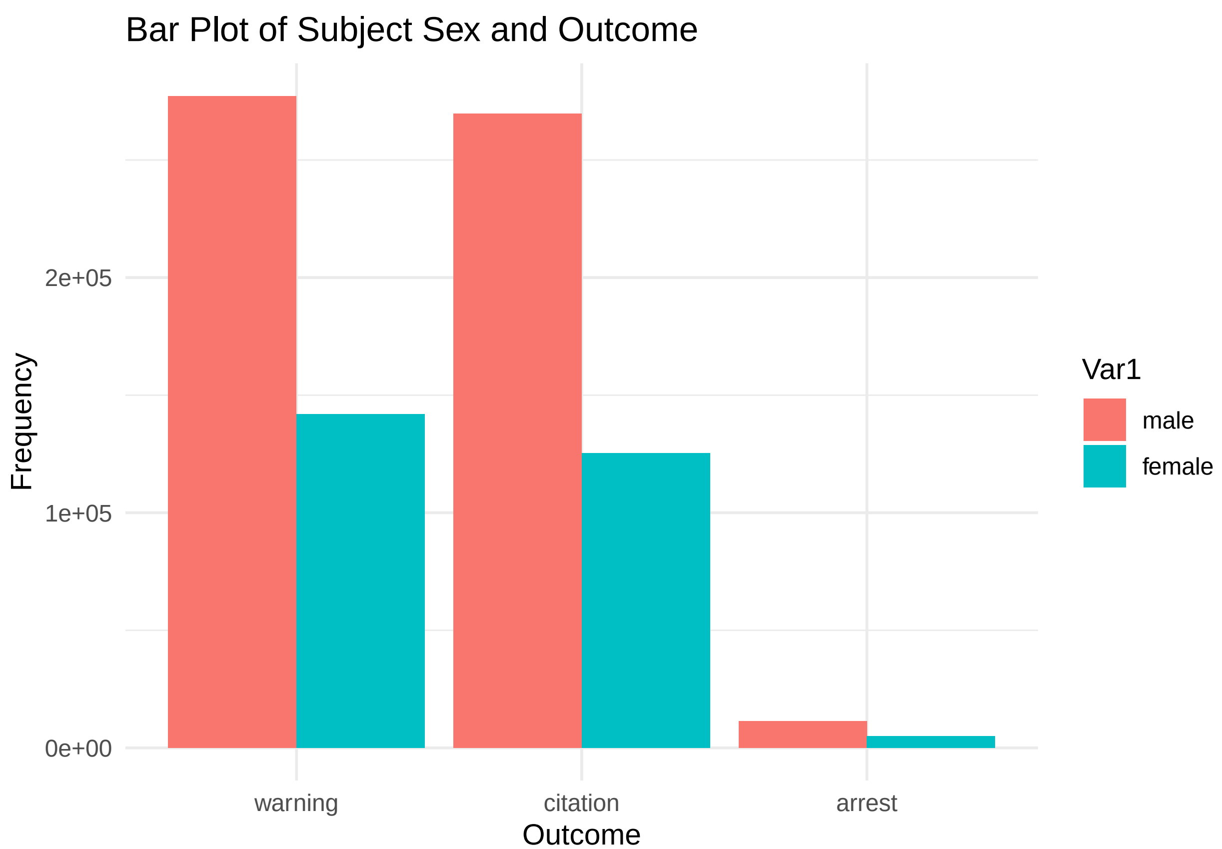
A mosaic plot is useful for visualizing the relationship between Subject Sex and Outcome.

```r
mosaicplot(risk_table, main="Mosaic Plot of Subject Sex and Outcome",
           xlab="Outcome", ylab="Subject Sex", color=TRUE)
```



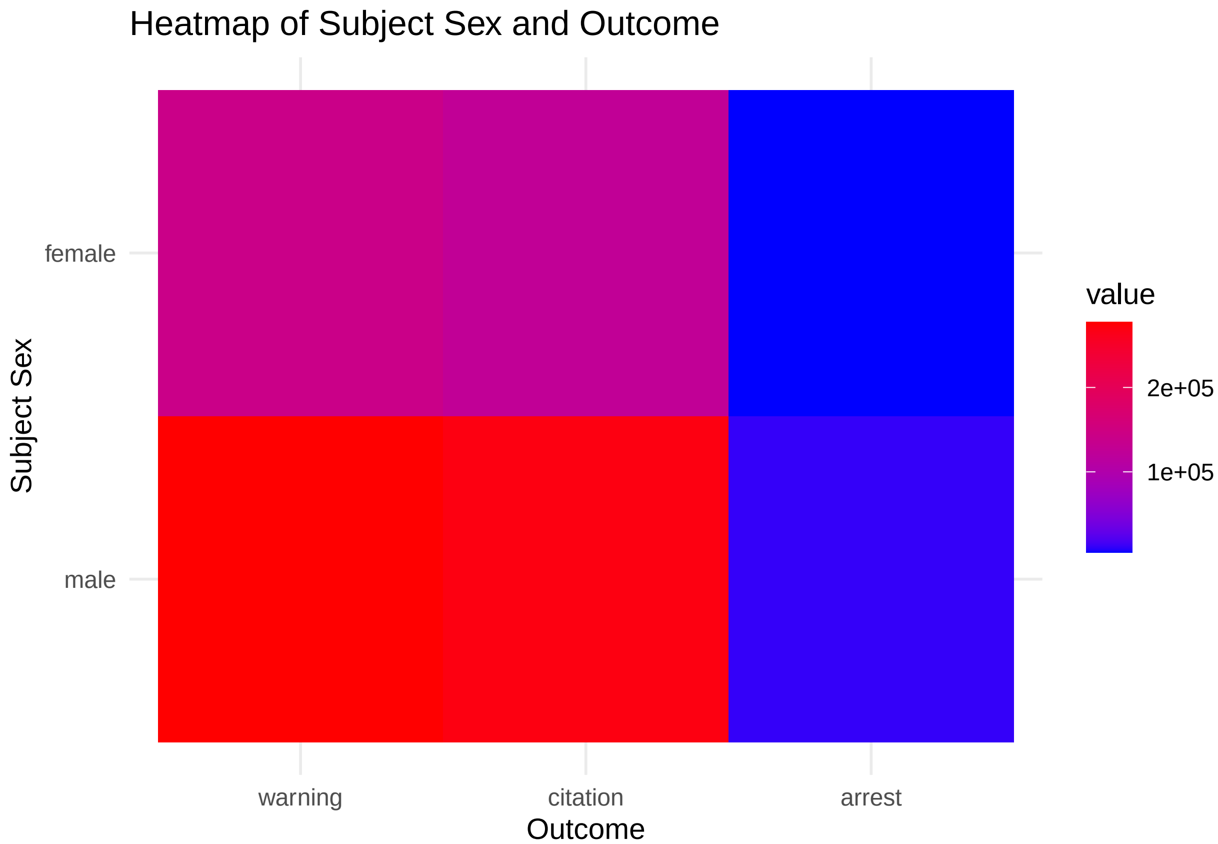
A bar plot can also be helpful to visualize the frequencies of these categories.

```r
# Converting the table to a data frame
risk_df <- as.data.frame(risk_table)

# Plotting using ggplot2
library(ggplot2)
ggplot(risk_df, aes(x=Var2, y=Freq, fill=Var1)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Bar Plot of Subject Sex and Outcome", x="Outcome", y="Frequency") +
  theme_minimal()
```



Heatmaps are also great visualiztion tool that can reveal patterns and relationships between these variables that may not be immediately apparent in the data.

```r
# Melting the table
risk_melted <- melt(risk_table)

# Plotting the heatmap
ggplot(risk_melted, aes(x=Var2, y=Var1, fill=value)) +
  geom_tile() +
  scale_fill_gradient(low="blue", high="red") +
  labs(title="Heatmap of Subject Sex and Outcome", x="Outcome", y="Subject Sex") +
  theme_minimal()
```



For this hypothesis, the data shows a significant skew, with most outcomes being either warnings or citations. The visualizations also clearly indicate that males have a higher number of violations and are at greater risk compared to females, corroborating the chi-square test results.