Night percent, logistic regression

Amber Lee

3/7/2020

Set up

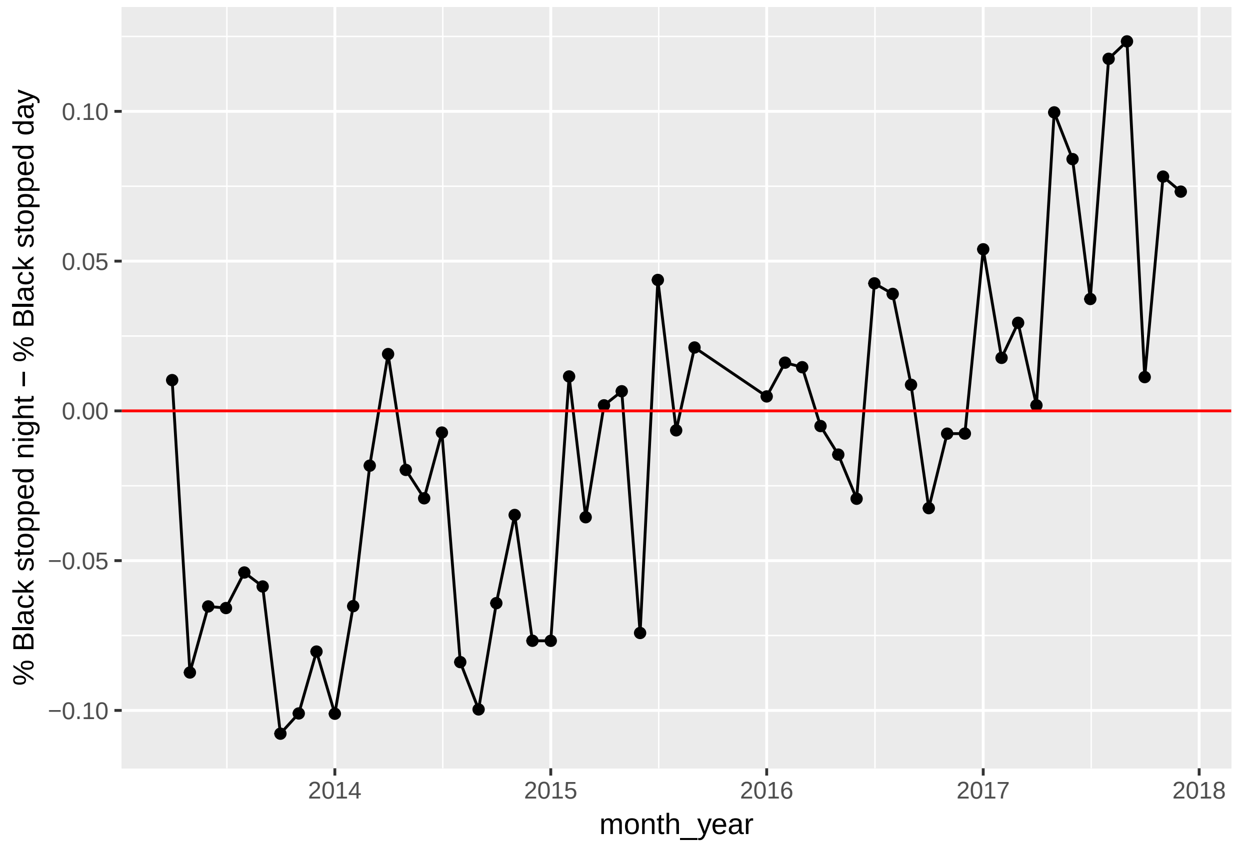
Night percent - day percent

```
# 1. fix the dates and lat/lng types. Check correct timezones for tz

CAoak <- CAoak %>%

  # optional: i have to filter out my NA's for date for POSIX to work
  filter(str_detect(date, "NA", negate = TRUE)) %>%

  mutate(nice_date = ymd(date),
         nice_year = year(nice_date),
         nice_month = month(nice_date),
         nice_day = day(nice_date),
         nice_time = hms(time),
         nice_day_of_year = yday(date),
         #for sunset sunrise:
         posix_date_time = as.POSIXct(paste(nice_date, time), tz = "America/Chicago", format = "%Y-%m-%
  mutate(lat_num = as.numeric(lat),
         lng_num = as.numeric(lng))
```

```
## Warning in .parse_hms(..., order = "HMS", quiet = quiet): Some strings
## failed to parse, or all strings are NAs
```

```
# 2. use sunrise/sunset function, again heeding the tz

oursunriseset <- function(latitude, longitude, date, direction = c("sunrise", "sunset")) {
  date.lat.long <- data.frame(date = date, lat = latitude, lon = longitude)
  if(direction == "sunrise"){
    getSunlightTimes(data = date.lat.long, keep=direction, tz = "America/Los_Angeles")$sunrise }else{
      getSunlightTimes(data = date.lat.long, keep=direction, tz = "America/Los_Angeles")$sunset } }

# 3. create variable for light (day/night)

CAoak <- CAoak %>%

  # use oursunriseset function to return posixct format sunrise and sunset times
  mutate(sunrise = oursunriseset(lat_num, lng_num, nice_date, direction = "sunrise"),
         sunset = oursunriseset(lat_num, lng_num, nice_date, direction = "sunset")) %>%

  # night and day!!
  mutate(light = ifelse(posix_date_time > sunrise & posix_date_time < sunset, "day", "night"))

# 4a. count the number of ALL DRIVERS and BLACK DRIVERS stopped during day and night.
```

```r
# 4b. calculate the percentage of black/all for day AND black/all for night

CAoakcheckpoint <- CAoak %>%

  # filter out the NA's for light variable
  filter(light == "day" | light == "night") %>%

  # group by month, year, and light
  group_by(nice_month, nice_year, light) %>%

  # count number of drivers stopped per month during night/day
  summarise(all_drivers_stopped = n(), black_drivers_stopped = sum(subject_race == "black")) %>%

  # find percent of black/all drivers stopped for day and night
  mutate(stops_black_percent = black_drivers_stopped/all_drivers_stopped) %>%

  #create arbitrary lubridate (first day of each month) for each year-month pair
  mutate(month_year = ymd(paste(nice_year, nice_month, "1", sep = "-")))

# 5. use filter to create two seperate day and night dataframes (to be joined later)

CAoak_day_stopcounts <- CAoakcheckpoint %>% filter(light == "day")
CAoak_night_stopcounts <- CAoakcheckpoint %>% filter(light == "night")

# 6. join and use mutate to calculate percents day/night and percent differences

# join by month_year
# do keep: all_drivers_stopped, black_drivers_stopped, and stops_black_percent for both day, night
# 6 variables in total

CAoak_join_stopcounts <- inner_join(CAoak_day_stopcounts, CAoak_night_stopcounts, by = c("month_year",

  # rename columns for clarity (day/night)
  rename(day_all_drivers_stopped = all_drivers_stopped.x,
         night_all_drivers_stopped = all_drivers_stopped.y,
         day_black_drivers_stopped = black_drivers_stopped.x,
         night_black_drivers_stopped = black_drivers_stopped.y,
         day_stops_black_percent = stops_black_percent.x,
         night_stops_black_percent = stops_black_percent.y) %>%

  # calculate the difference! OBSERVE that it is night percent difference
  mutate(racial_percent_diff = night_stops_black_percent - day_stops_black_percent)

CAoak_join_stopcounts %>%

  ggplot(mapping = aes(x = month_year, y = racial_percent_diff))+
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 0, color = "red") +
  labs(y = "% Black stopped night - % Black stopped day")
```

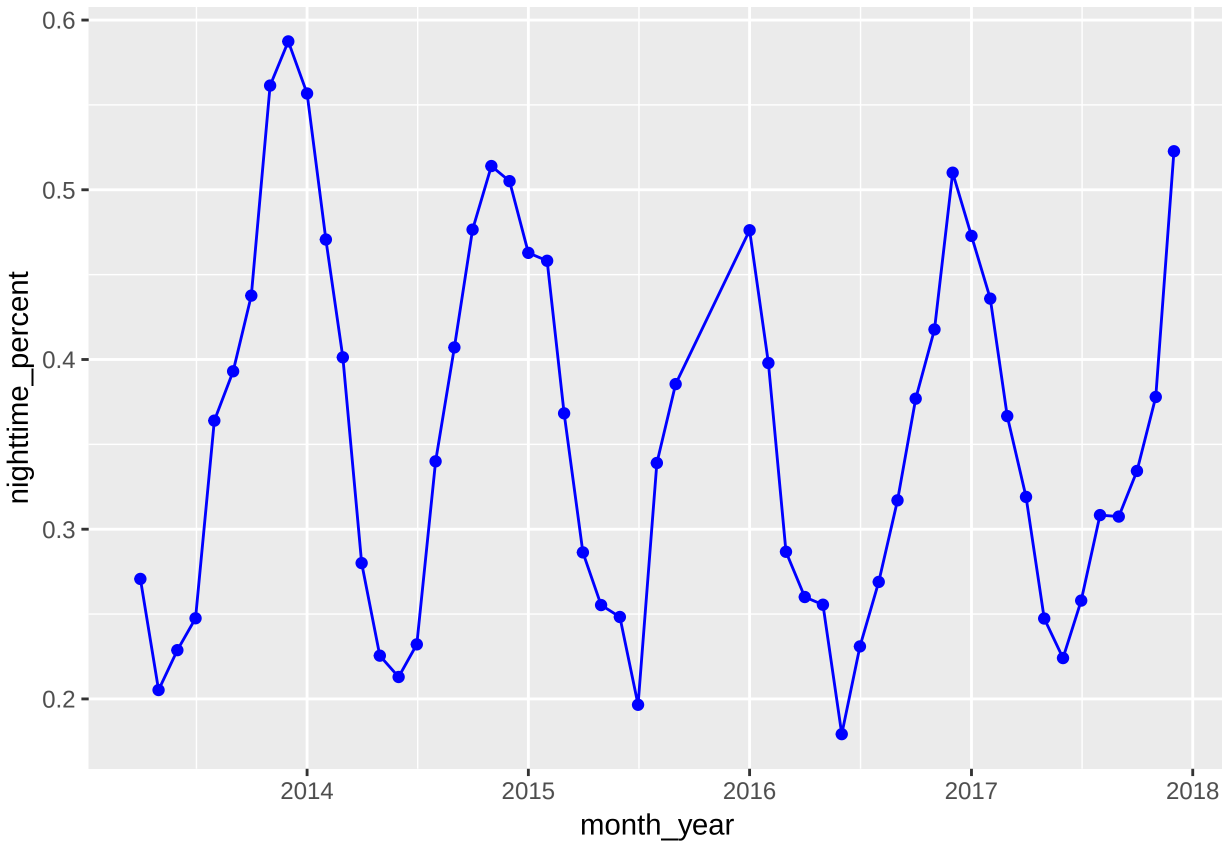



```r
#ggsave("CAoak_daynightpercent_tidyversemethod.png")
```

Questions: Have the day traffic stops and night traffic stop relative proportions stayed the same?

To answer this question, I build off of the already-cleaned CAoak_join_stopcounts

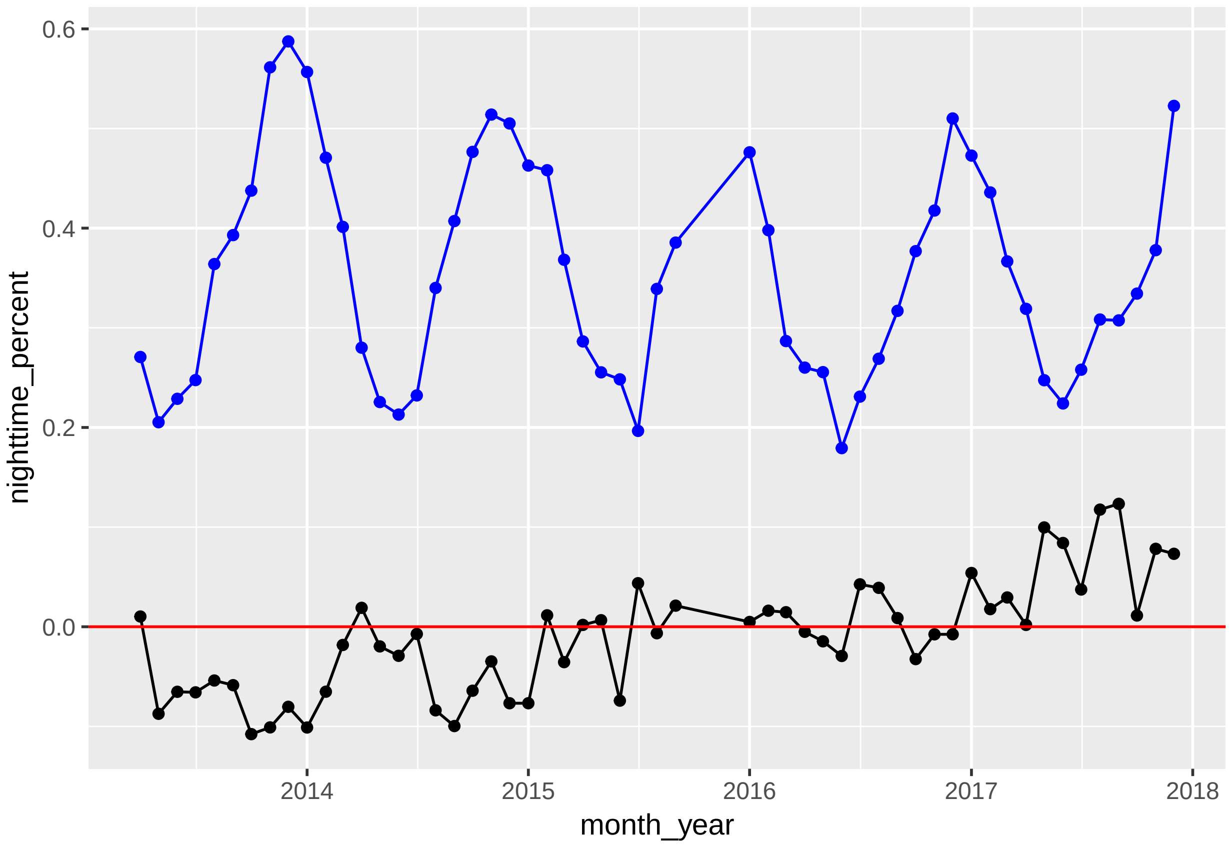
```r
CAoak_join_stopcounts <- CAoak_join_stopcounts %>%

  # Find all stop counts
  mutate(total_stop_count = day_all_drivers_stopped + night_all_drivers_stopped,

         # Find percentage of night-time stops
         nighttime_percent = night_all_drivers_stopped/total_stop_count)

CAoak_join_stopcounts %>%

  ggplot(mapping = aes(x = month_year, y = nighttime_percent)) +
  geom_point(color = "blue") +
  geom_line(color = "blue")
```




```r
ggsave("CAoak_nightpercent.png")
```

```
## Saving 6.5 x 4.5 in image
```

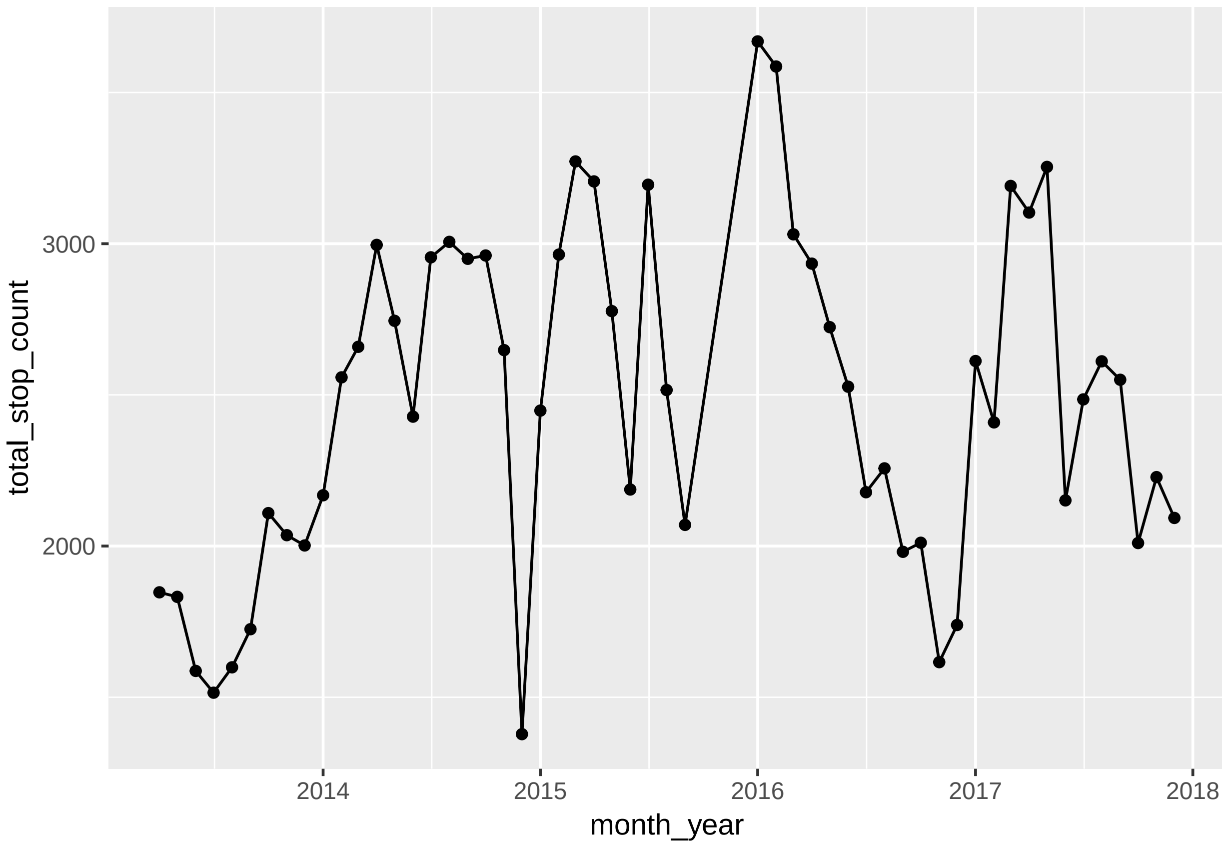
```r
CAoak_join_stopcounts %>%

  # Find all stop counts
  mutate(total_stop_count = day_all_drivers_stopped + night_all_drivers_stopped,

         # Find percentage of night-time stops
         nighttime_percent = night_all_drivers_stopped/total_stop_count) %>%

  ggplot(mapping = aes(x = month_year, y = nighttime_percent)) +
  geom_point(color = "blue") +
  geom_line(color = "blue") +

  # Overlay the racial_percent_diff from earlier chunk
  geom_point(mapping = aes(x = month_year, y = racial_percent_diff)) +
  geom_line(mapping = aes(x = month_year, y = racial_percent_diff)) +
  geom_hline(yintercept = 0, color = "red")
```




```r
ggsave("CAoak_overlaydaynightpercent_tidyversemethod.png")
```

```
## Saving 6.5 x 4.5 in image
```

```r
# number of stops per month

CAoak_join_stopcounts %>%

  ggplot(mapping = aes(x = month_year, y = total_stop_count)) +
  geom_point() +
  geom_line()
```

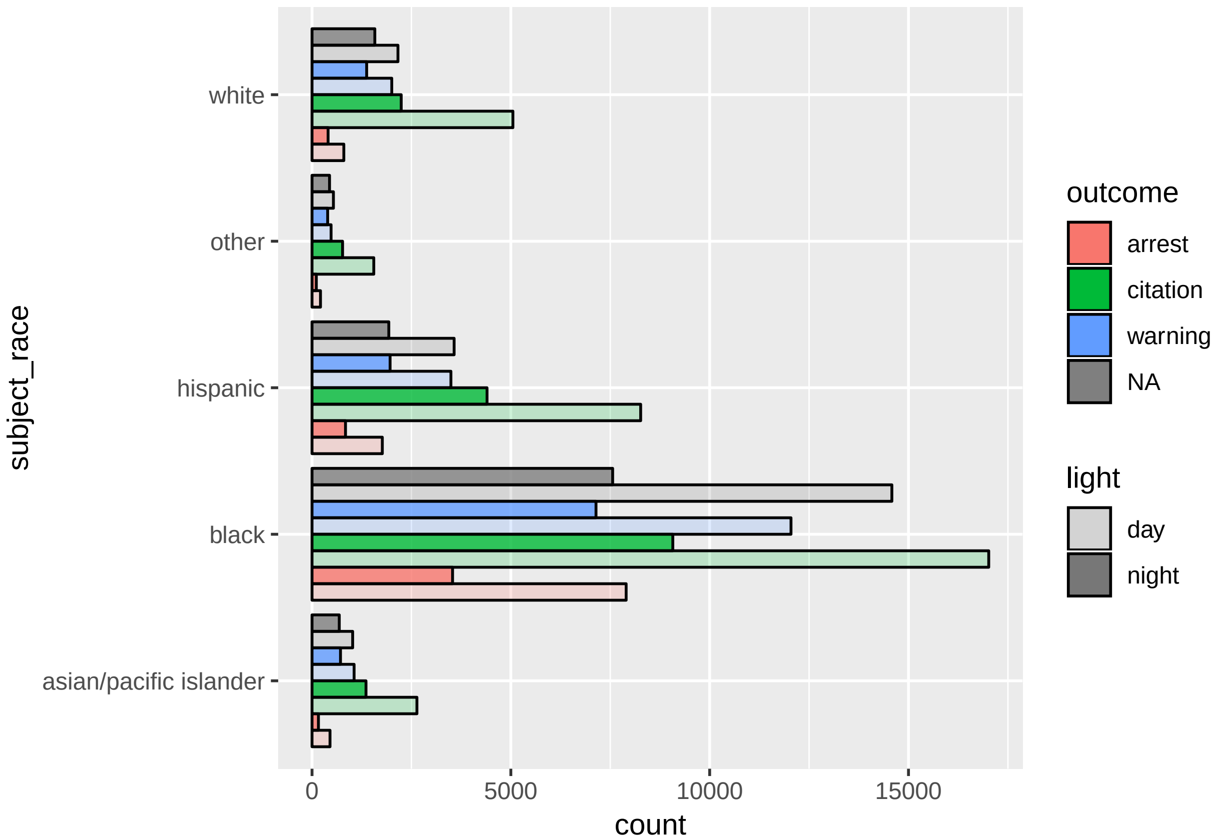



```
# Count the total number of stops: 41k. 41k out of 133k is about 30%, so using this data to model searc
CAoak %>%

  filter(search_conducted == "TRUE") %>%
  summarise(n())
```

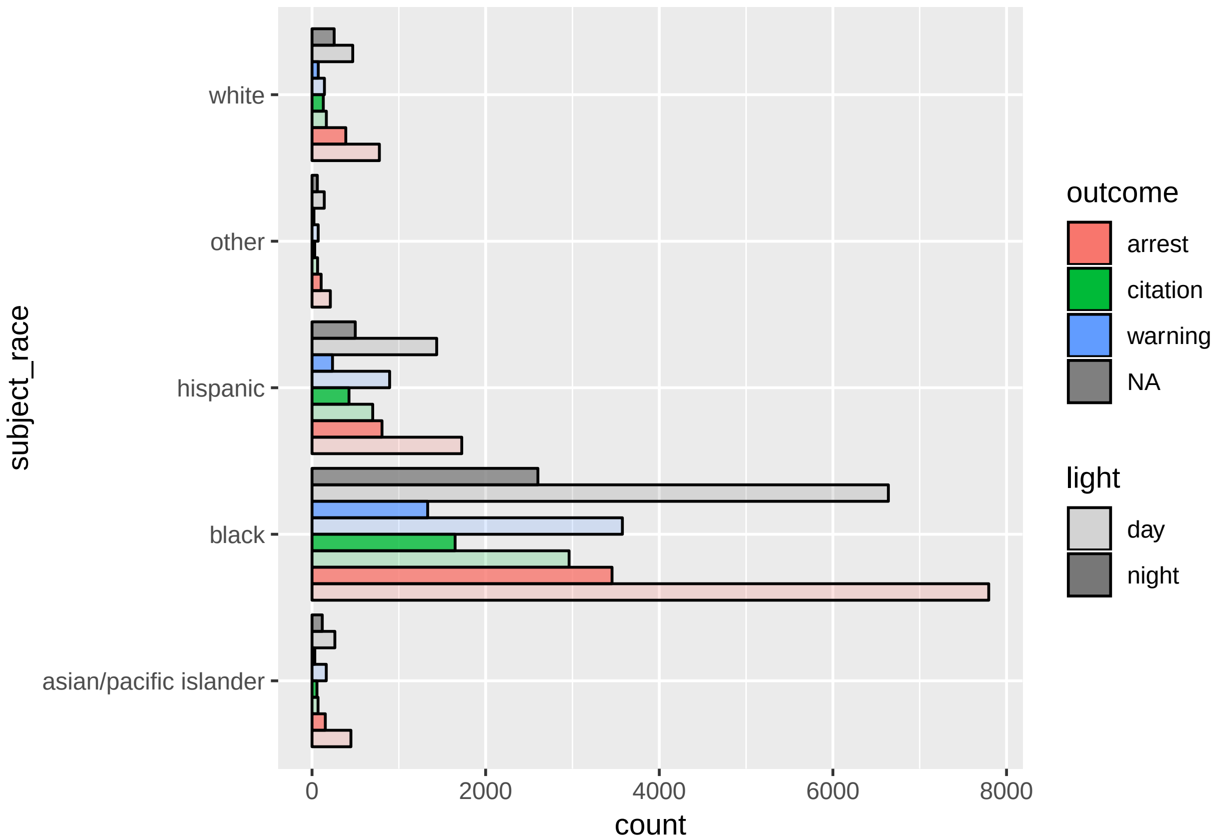
```
##   n()
## 1   0
```

Interlude: light, subject race, outcome

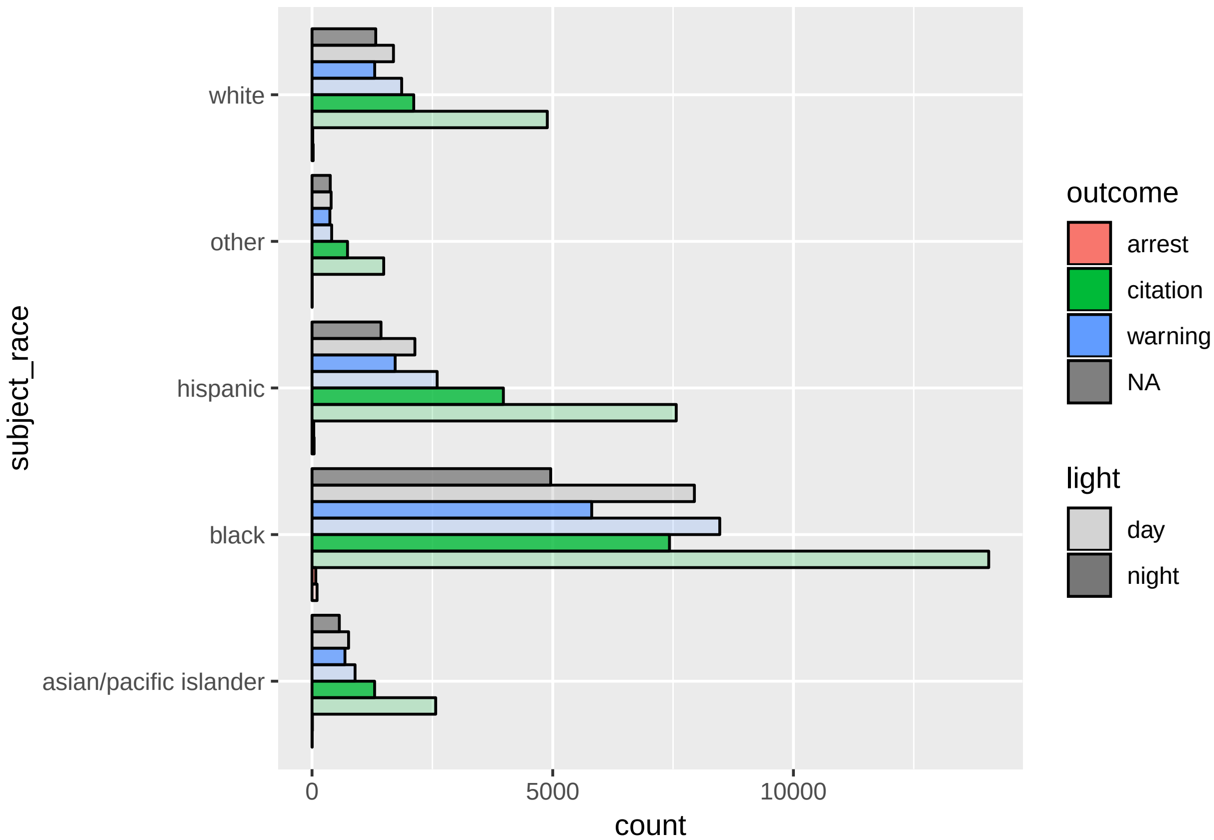
```
CAoak %>%

  #filter out NA's for readability
  filter(!is.na(light)) %>%
  ggplot(aes(x = subject_race, fill = outcome, alpha = light)) +
  geom_bar(position="dodge", colour="black") + coord_flip() + scale_alpha_manual(values=c(.2, .8))
```



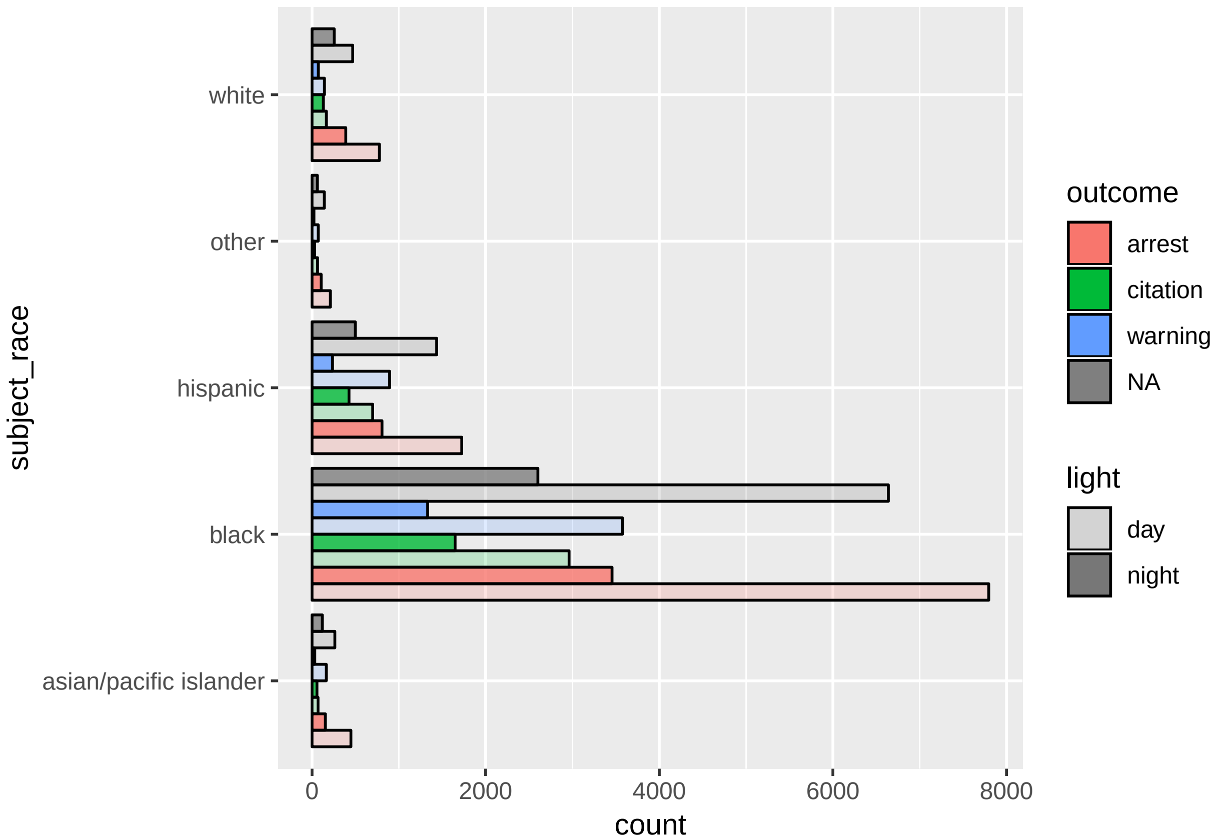

```
CAoak %>%

  #filter out NA's for readability
  filter(!is.na(light)) %>%
  filter(search_conducted == "1") %>%
  ggplot(aes(x = subject_race, fill = outcome, alpha = light)) +
  geom_bar(position="dodge", colour="black") + coord_flip() + scale_alpha_manual(values=c(.2, .8))
```




```
CAoak %>%

  #filter out NA's for readability
  filter(!is.na(light)) %>%
  filter(search_conducted == "0") %>%
  ggplot(aes(x = subject_race, fill = outcome, alpha = light)) +
  geom_bar(position="dodge", colour="black") + coord_flip() + scale_alpha_manual(values=c(.2, .8))
```




```
CAoak %>%

  #filter out NA's for readability
  filter(!is.na(light)) %>%
  filter(search_conducted == "1") %>%
  ggplot(aes(x = subject_race, fill = outcome, alpha = light)) +
  geom_bar(position="dodge", colour="black") + coord_flip() + scale_alpha_manual(values=c(.2, .8))
```



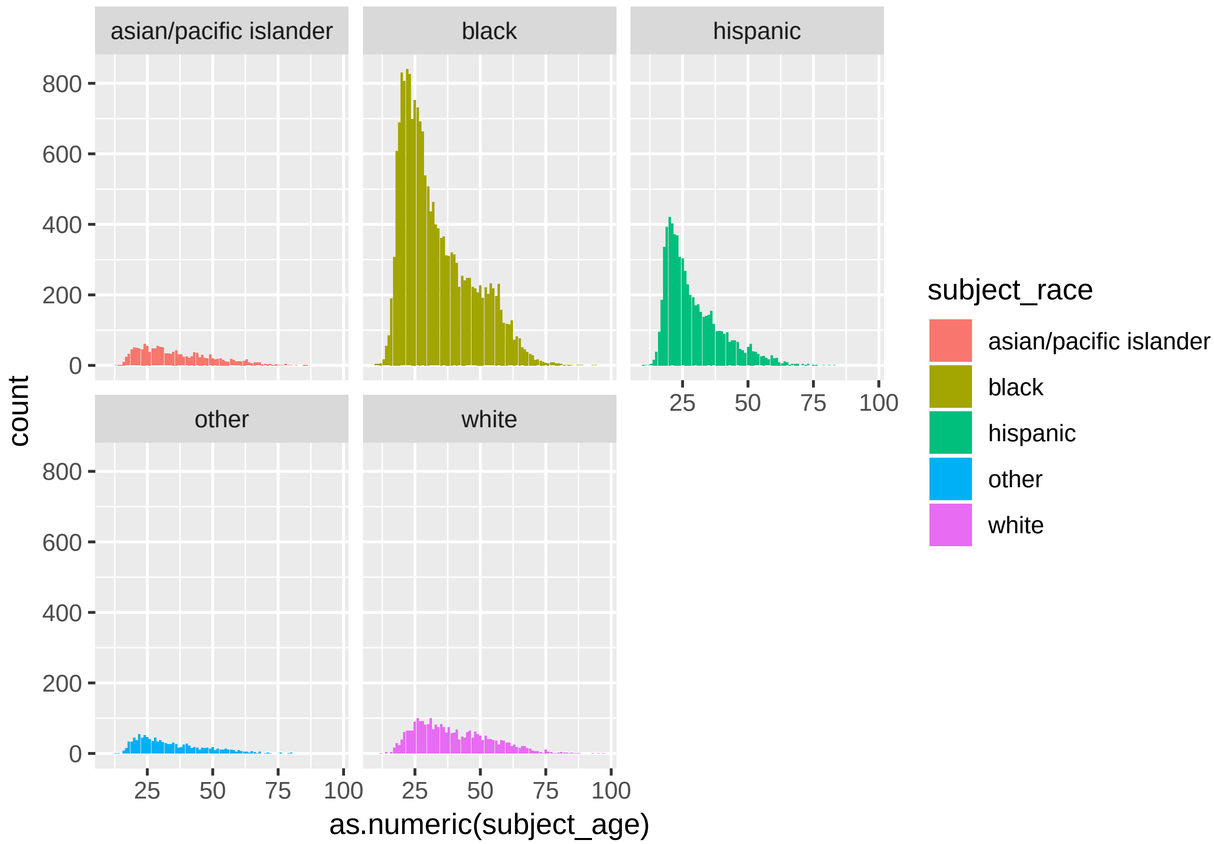

```r
ggsave("outcome, nightday, race progress.png")
```

```
## Saving 6.5 x 4.5 in image
```

perhaps can look at searches, then outcome :) searches precede citation, arrest. read more about the process of getting into the criminal justice system

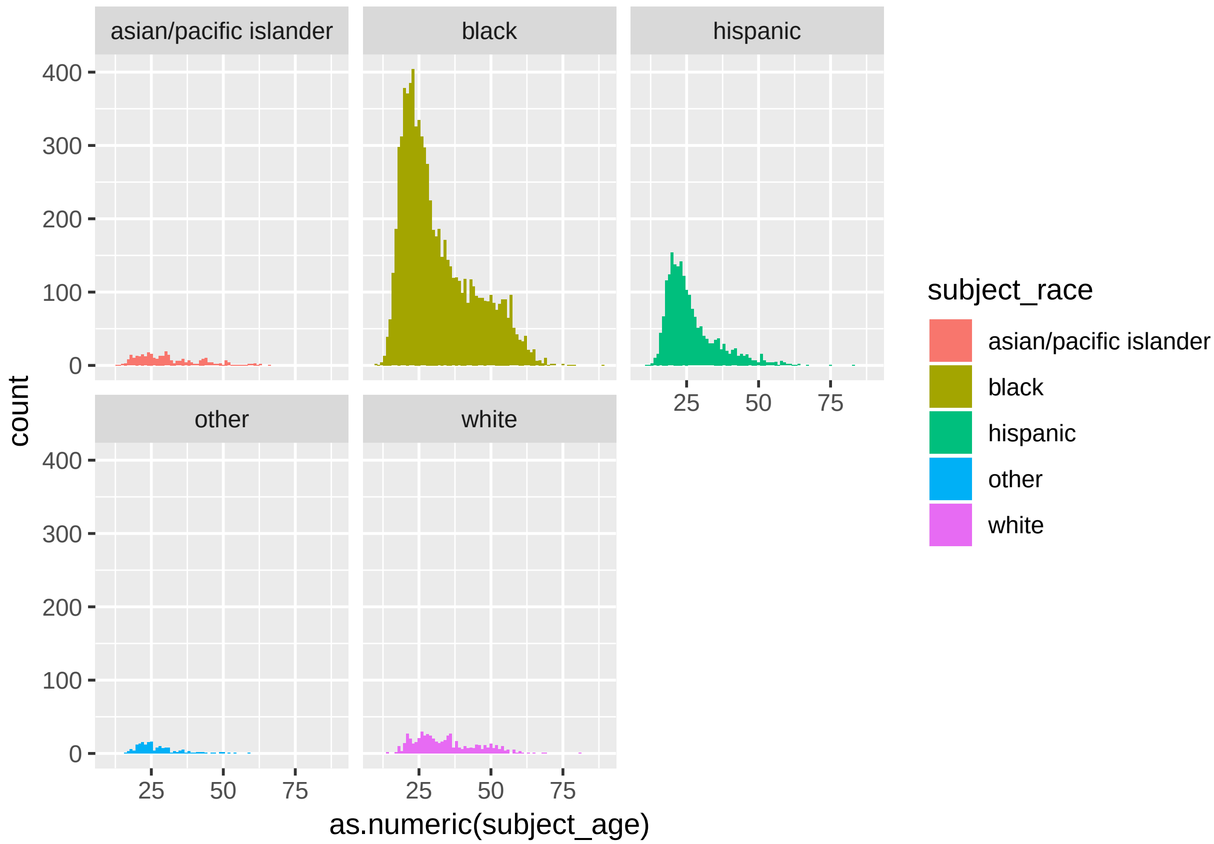
```r
ggplot(data = CAoak) +
   geom_bar(mapping = aes(x = as.numeric(subject_age), fill = subject_race)) +
  facet_wrap(~ subject_race)
```

```
## Warning: Removed 102722 rows containing non-finite values (stat_count).
```



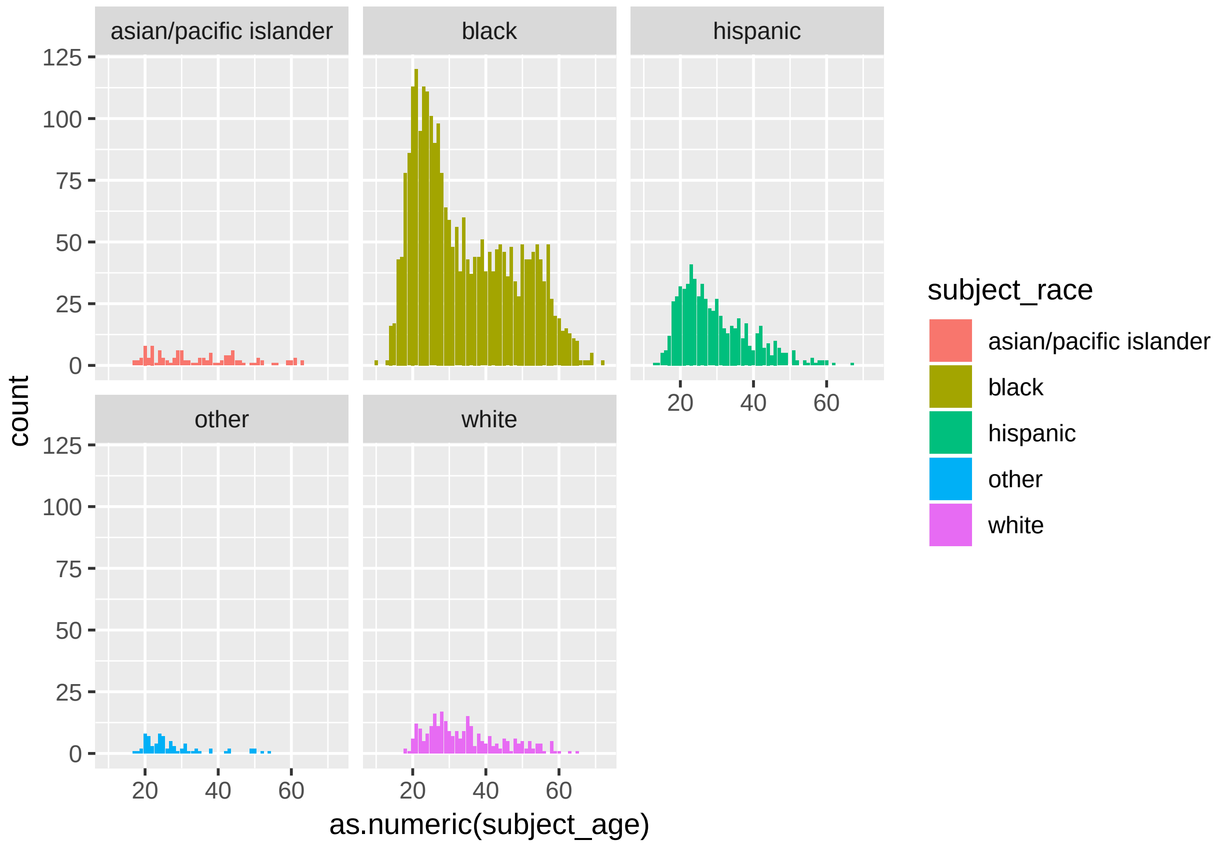

```r
CAoak %>%
  filter(search_conducted == "1") %>%
  ggplot() +
  geom_bar(mapping = aes(x = as.numeric(subject_age), fill = subject_race)) +
  facet_wrap(~ subject_race)
```

```
## Warning: Removed 30280 rows containing non-finite values (stat_count).
```




```r
CAoak %>%
  filter(search_conducted == "1", arrest_made == "1") %>%
  ggplot() +
  geom_bar(mapping = aes(x = as.numeric(subject_age), fill = subject_race)) +
  facet_wrap(~ subject_race)
```

```
## Warning: Removed 12105 rows containing non-finite values (stat_count).
```




```
# Question: do these distributions reflect census data?
```

Logistic Regression

```
logreg_oak1 <- CAoak %>%

  #only 30k out of 133k of my data records subject age
  filter(subject_age != "NA") %>%

  #use case_when to recode character variables to binary levels
  mutate(
         # assigned day = 1
         light_binary = case_when(light == "day" ~ 1,
                                  light == "night" ~ 0),
         subject_age = as.numeric(subject_age)) %>%
  select(subject_age, search_conducted, search_conducted, light, light_binary, subject_race, arrest_made

all_output1 <- glm(formula = search_conducted ~ subject_age*subject_race + factor(light_binary), data =

summary(all_output1)
```

```
##
## Call:
## glm(formula = search_conducted ~ subject_age * subject_race +
##     factor(light_binary), family = binomial, data = logreg_oak1)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2120  -0.9952  -0.7553   1.2556   2.6208
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.0861590  0.1821551  -0.473 0.636215
## subject_age                      -0.0359989  0.0052561  -6.849 7.44e-12
## subject_raceblack                 0.3525380  0.1859483   1.896 0.057974
## subject_racehispanic              0.2281283  0.1990810   1.146 0.251834
## subject_raceother                 0.3318093  0.3217224   1.031 0.302375
## subject_racewhite                 0.0611644  0.2340115   0.261 0.793805
## factor(light_binary)1            -0.0101902  0.0257883  -0.395 0.692734
## subject_age:subject_raceblack     0.0184709  0.0053817   3.432 0.000599
## subject_age:subject_racehispanic  0.0015488  0.0059629   0.260 0.795058
## subject_age:subject_raceother    -0.0256453  0.0103505  -2.478 0.013224
## subject_age:subject_racewhite    -0.0004428  0.0065765  -0.067 0.946314
##
## (Intercept)
## subject_age                      ***
## subject_raceblack                .
## subject_racehispanic
## subject_raceother
## subject_racewhite
## factor(light_binary)1
## subject_age:subject_raceblack    ***
## subject_age:subject_racehispanic
## subject_age:subject_raceother    *
## subject_age:subject_racewhite
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39877  on 30666  degrees of freedom
## Residual deviance: 38144  on 30656  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 38166
##
## Number of Fisher Scoring iterations: 5
```

concerns: 1) this is looking at all stops vs. all stops + searches. may want to look at all stops vs. searches THAT DIDN'T RESULT IN AN ARREST 2) may want to bin ages into rough age groups

```r
# count the number of searches conducted that did and didn't result in arrests
CAoak %>%
  select(search_conducted, arrest_made) %>%
  group_by(search_conducted, arrest_made) %>%
  summarise(n())
```

```
## # A tibble: 4 x 3
## # Groups:   search_conducted [2]
##   search_conducted arrest_made `n()`
##              <dbl>       <dbl> <int>
## 1                0           0 91929
```

```
## 2                0           1   320
## 3                1           0 25286
## 4                1           1 15870
```

```r
# out of 41,156 searches conducted, 15870 resulted in arrests made. that is 40%

# conduct logistic regression looking at search conducted but arrest not made

logreg_oak2 <- logreg_oak1 %>%
  filter(arrest_made == "0")

all_output2 <- glm(formula = search_conducted ~ subject_age*subject_race + factor(light_binary), data =

summary(all_output2)
```

```
##
## Call:
## glm(formula = search_conducted ~ subject_age * subject_race +
##     factor(light_binary), family = binomial, data = logreg_oak2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1398  -0.8669  -0.6472   1.3069   2.9139
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.1645689  0.2272255  -0.724 0.468910
## subject_age                      -0.0516224  0.0070314  -7.342 2.11e-13
## subject_raceblack                 0.1766233  0.2311830   0.764 0.444869
## subject_racehispanic              0.2599537  0.2476049   1.050 0.293777
## subject_raceother                -0.1105944  0.4055405  -0.273 0.785077
## subject_racewhite                -0.6250167  0.2971249  -2.104 0.035418
## factor(light_binary)1             0.1618909  0.0305709   5.296 1.19e-07
## subject_age:subject_raceblack     0.0253226  0.0071697   3.532 0.000413
## subject_age:subject_racehispanic -0.0005015  0.0079283  -0.063 0.949565
## subject_age:subject_raceother    -0.0145676  0.0135071  -1.079 0.280805
## subject_age:subject_racewhite     0.0147832  0.0087541   1.689 0.091275
##
## (Intercept)
## subject_age                      ***
## subject_raceblack
## subject_racehispanic
## subject_raceother
## subject_racewhite                *
## factor(light_binary)1            ***
## subject_age:subject_raceblack    ***
## subject_age:subject_racehispanic
## subject_age:subject_raceother
## subject_age:subject_racewhite    .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31024  on 26833  degrees of freedom
```

```
## Residual deviance: 29233  on 26823  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 29255
##
## Number of Fisher Scoring iterations: 5
```

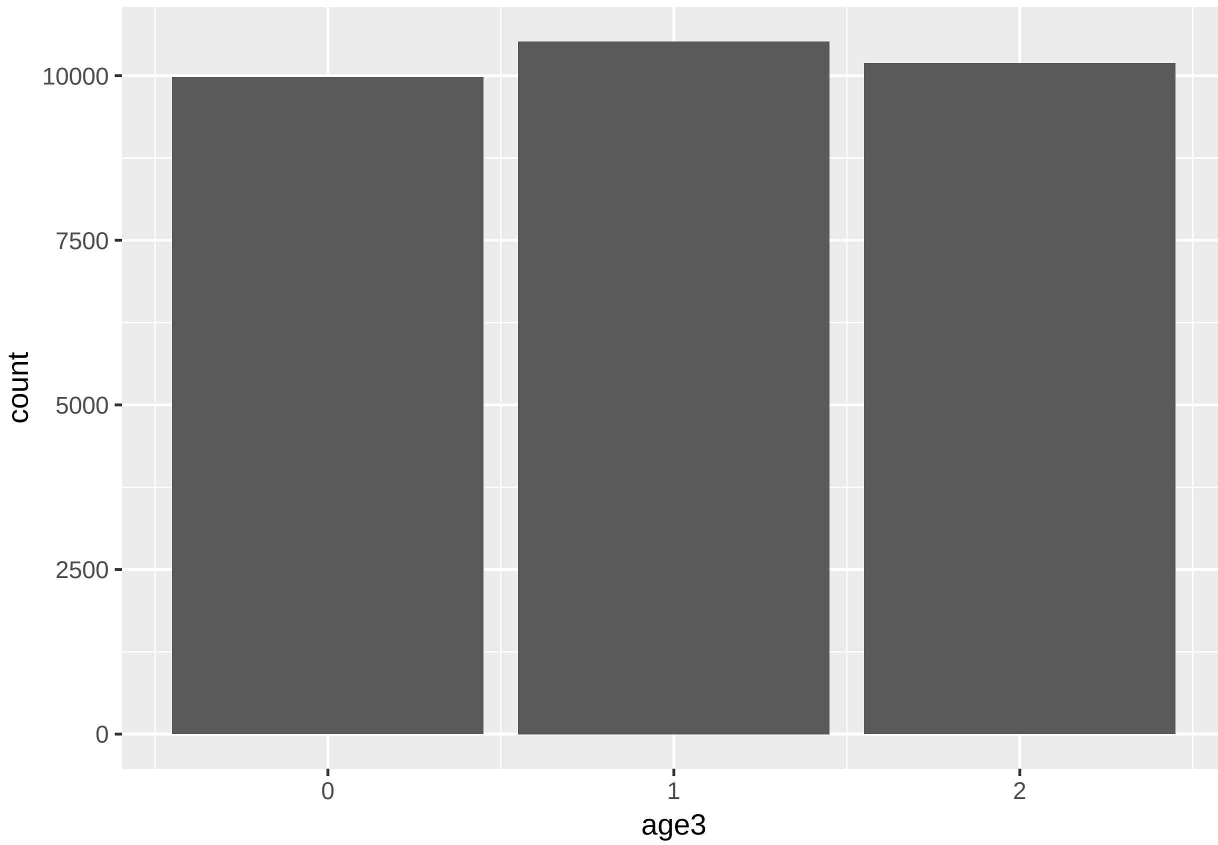
- note the coefficients that become statistically significant when looking only at discretionary searches:

** all searches > **Coefficients: Estimate Std. Error z value Pr(>|z|)**
subject_age -0.0359989 0.0052561 -6.849 7.44e-12 *subject_raceblack 0.3525380 0.1859483 1.896 0.057974 .*
subject_racewhite 0.0611644 0.2340115 0.261 0.793805
factor(light_binary)1 -0.0101902 0.0257883 -0.395 0.692734
subject_age:subject_raceblack 0.0184709 0.0053817 3.432 0.000599 ** subject_age:subject_raceother -0.0256453 0.0103505 -2.478 0.013224 *
subject_age:subject_racewhite -0.0004428 0.0065765 -0.067 0.946314

** discretionary searches only > **Coefficients: Estimate Std. Error z value Pr(>|z|)**
subject_age -0.0516224 0.0070314 -7.342 2.11e-13 *subject_raceblack 0.1766233 0.2311830 0.764 0.444869*
subject_racewhite -0.6250167 0.2971249 -2.104 0.035418
factor(light_binary)1 0.1618909 0.0305709 5.296 1.19e-07 ***subject_age:subject_raceblack 0.0253226 0.0071697 3.532 0.000413*** subject_age:subject_raceother -0.0145676 0.0135071 -1.079 0.280805
subject_age:subject_racewhite 0.0147832 0.0087541 1.689 0.091275 .

- the magnitude of subject_age coefficient increases for discretionary searches (-.03 to -.05, more significant)
- magnitude of subject_racewhite coefficient goes from .06 insignificant to -.625 statistically significant when limiting to discretionary searches
- factor(light_binary) becomes positive .16 and statistically significant when limiting to discretionary searches. day = 1 and night = 0, so how to interpret the +.16 coefficient?
- subject_age::subject_raceblack goes from .018 to .025 (1.4x increase) when limiting to discretionary searches

```r
logreg_oak3 <- logreg_oak1 %>%
  mutate(age3 = case_when(subject_age <= 24 ~ 0,
                          subject_age > 24 & subject_age <= 36 ~ 1,
                          subject_age > 36 ~ 2))

# about uniformly distributed... is that good? 1/3 of each stops are in each age cut off
ggplot(data = logreg_oak3) +
  geom_bar(mapping = aes(x = age3))
```




```r
all_output3 <- glm(formula = search_conducted ~ age3*subject_race + factor(light_binary), data = logreg_

summary(all_output3)
```

```
##
## Call:
## glm(formula = search_conducted ~ age3 * subject_race + factor(light_binary),
##     family = binomial, data = logreg_oak3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1497  -0.9579  -0.7131   1.3128   2.2722
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.80491    0.10316  -7.802 6.07e-15 ***
## age3                       -0.47111    0.07925  -5.944 2.77e-09 ***
## subject_raceblack           0.73926    0.10447   7.076 1.48e-12 ***
## subject_racehispanic        0.27421    0.10802   2.538  0.01113 *
## subject_raceother          -0.16543    0.15500  -1.067  0.28584
## subject_racewhite           0.10151    0.13755   0.738  0.46054
## factor(light_binary)1      -0.01028    0.02575  -0.399  0.68977
## age3:subject_raceblack      0.22388    0.08138   2.751  0.00594 **
## age3:subject_racehispanic   0.05175    0.08708   0.594  0.55236
## age3:subject_raceother     -0.28997    0.13578  -2.136  0.03272 *
## age3:subject_racewhite     -0.05470    0.10254  -0.533  0.59369
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39877  on 30666  degrees of freedom
## Residual deviance: 38288  on 30656  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 38310
##
## Number of Fisher Scoring iterations: 4
```

```
logreg_oak4 <- logreg_oak2 %>%
    mutate(age3 = case_when(subject_age <= 24 ~ 0,
                            subject_age > 24 & subject_age <= 36 ~ 1,
                            subject_age > 36 ~ 2))

all_output4 <- glm(formula = search_conducted ~ age3*subject_race + factor(light_binary), data = logreg_

summary(all_output4)
```

```
##
## Call:
## glm(formula = search_conducted ~ age3 * subject_race + factor(light_binary),
##     family = binomial, data = logreg_oak4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0438  -0.8429  -0.6655   1.3171   2.5421
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.196372   0.117360 -10.194  < 2e-16 ***
## age3                       -0.652232   0.096661  -6.748 1.50e-11 ***
## subject_raceblack           0.712537   0.118355   6.020 1.74e-09 ***
## subject_racehispanic        0.284208   0.122217   2.325  0.02005 *
## subject_raceother          -0.392612   0.184243  -2.131  0.03309 *
## subject_racewhite          -0.243093   0.164488  -1.478  0.13944
## factor(light_binary)1       0.161272   0.030532   5.282 1.28e-07 ***
## age3:subject_raceblack      0.284051   0.099013   2.869  0.00412 **
## age3:subject_racehispanic   0.008423   0.106426   0.079  0.93691
## age3:subject_raceother     -0.148729   0.169752  -0.876  0.38094
## age3:subject_racewhite      0.089788   0.128096   0.701  0.48334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31024  on 26833  degrees of freedom
## Residual deviance: 29373  on 26823  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 29395
##
## Number of Fisher Scoring iterations: 5
```

```r
#Each query returns an R dataframe
DBI::dbGetQuery(con, "SHOW TABLES")
```

```
##    Tables_in_traffic
## 1          AZgilbert
## 2             AZmesa
## 3        AZstatewide
## 4       CAlosangeles
## 5          CAoakland
## 6    CAsanbernardino
## 7         CAsandiego
## 8     CAsanfrancisco
## 9           COaurora
## 10          COdenver
## 11       COstatewide
## 12        CThartford
## 13       CTstatewide
## 14           FLsaint
## 15       FLstatewide
## 16           FLtampa
## 17       GAstatewide
## 18       IAstatewide
## 19      IDidahofalls
## 20         ILchicago
## 21       ILstatewide
## 22       INfortwayne
## 23         KSwichita
## 24      KYlouisville
## 25       KYowensboro
## 26      LAneworleans
## 27       MAstatewide
## 28       MDbaltimore
## 29       MDstatewide
## 30       MIstatewide
## 31       MNsaintpaul
## 32       MSstatewide
## 33       MTstatewide
## 34       NCcharlotte
## 35          NCdurham
## 36         NCraleigh
## 37      NDgrandforks
## 38       NDstatewide
## 39       NEstatewide
## 40       NHstatewide
## 41          NJcamden
## 42       NJstatewide
## 43          NYalbany
## 44       NYstatewide
## 45      OHcincinnati
## 46        OHcolumbus
## 47       OHstatewide
## 48    OKoklahomacity
## 49       TNnashville
## 50           TNstate
```

```
## 51          TXaustin
## 52         TXgarland
## 53      TXsanantonio
## 54         WAseattle
## 55          WAtacoma
```