PS2

Xiaowen Yue

2025-02-19

R Markdown

a. Consider the random variable X = age of the driver conditional on being stopped. Compute the sample probability mass function and the sample cumulative distribution function of X. Produce a graph. What is the maximum probability across all age groups, and for which age group?

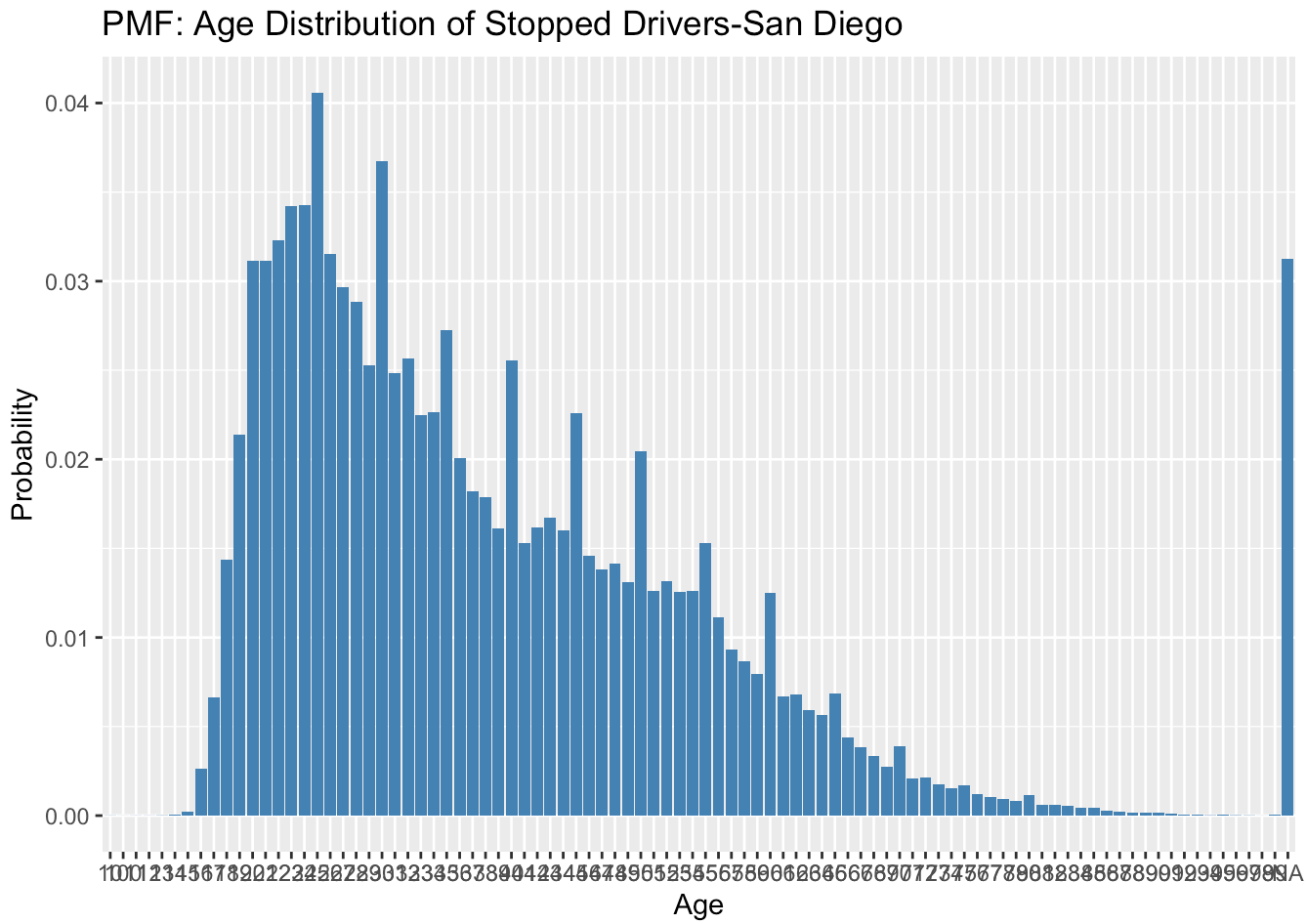
```
library(readxl)
library(tidyverse)
```

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

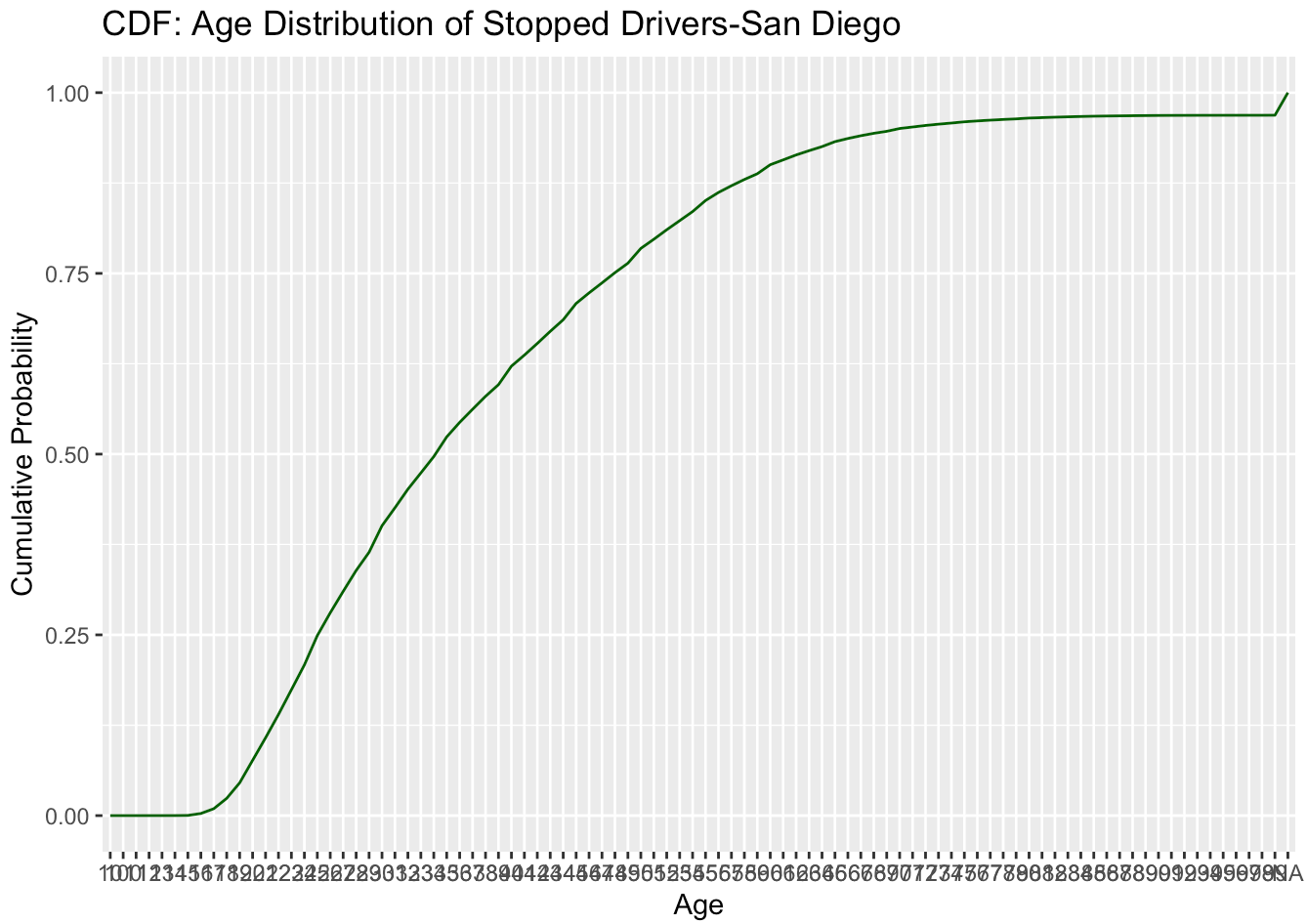
```
library(lubridate)
library(readr)
#for city Sandeigo
casand <- read_excel("~/Desktop/Research Data/ca_san_diego_2020_04_01.xlsx", col_types =
"text")
glimpse(casand)
```

```
## Rows: 383,027
## Columns: 21
## $ raw_row_number               <chr> "1", "2", "3", "4", "5", "6", "7", "8", "…
## $ date                         <chr> "2014-01-01", "2014-01-01", "2014-01-01",…
## $ time                         <chr> "01:25:00", "05:47:00", "07:46:00", "08:1…
## $ service_area                 <chr> "110", "320", "320", "610", "930", "820",…
## $ subject_age                  <chr> "24", "42", "29", "23", "35", "30", "19",…
## $ subject_race                 <chr> "white", "white", "asian/pacific islander…
## $ subject_sex                  <chr> "male", "male", "male", "male", "male", "…
## $ type                         <chr> "vehicular", "vehicular", "vehicular", "v…
## $ arrest_made                  <chr> "FALSE", "FALSE", "FALSE", "FALSE", "FALS…
## $ citation_issued              <chr> "TRUE", "FALSE", "FALSE", "TRUE", "TRUE",…
## $ warning_issued               <chr> "FALSE", "TRUE", "TRUE", "FALSE", "FALSE"…
## $ outcome                      <chr> "citation", "warning", "warning", "citati…
## $ contraband_found             <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA",…
## $ search_conducted             <chr> "FALSE", "FALSE", "FALSE", "FALSE", "FALS…
## $ search_person                <chr> "FALSE", "FALSE", "FALSE", "FALSE", "FALS…
## $ search_vehicle               <chr> "FALSE", "FALSE", "FALSE", "FALSE", "FALS…
## $ search_basis                 <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA",…
## $ reason_for_search            <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA",…
## $ reason_for_stop              <chr> "Moving Violation", "Moving Violation", "…
## $ raw_action_taken             <chr> "Citation", "Verbal Warning", "Verbal War…
## $ raw_subject_race_description <chr> "WHITE", "WHITE", "LAOTIAN", "WHITE", "HI…
```

```
casand_clean <- casand %>%
  filter(!is.na(subject_age))

pmf <- casand_clean %>%
  count(subject_age) %>%
  mutate(probability = n / sum(n))

cdf <- pmf %>%
  arrange(subject_age) %>%
  mutate(cumulative_probability = cumsum(probability))

ggplot(pmf, aes(x = subject_age, y = probability)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "PMF: Age Distribution of Stopped Drivers-San Diego",
       x = "Age",
       y = "Probability")
```




```
ggplot(cdf, aes(x = subject_age, y = cumulative_probability, group = 1)) +
  geom_line(color = "darkgreen") +
  labs(title = "CDF: Age Distribution of Stopped Drivers-San Diego",
       x = "Age",
       y = "Cumulative Probability")
```



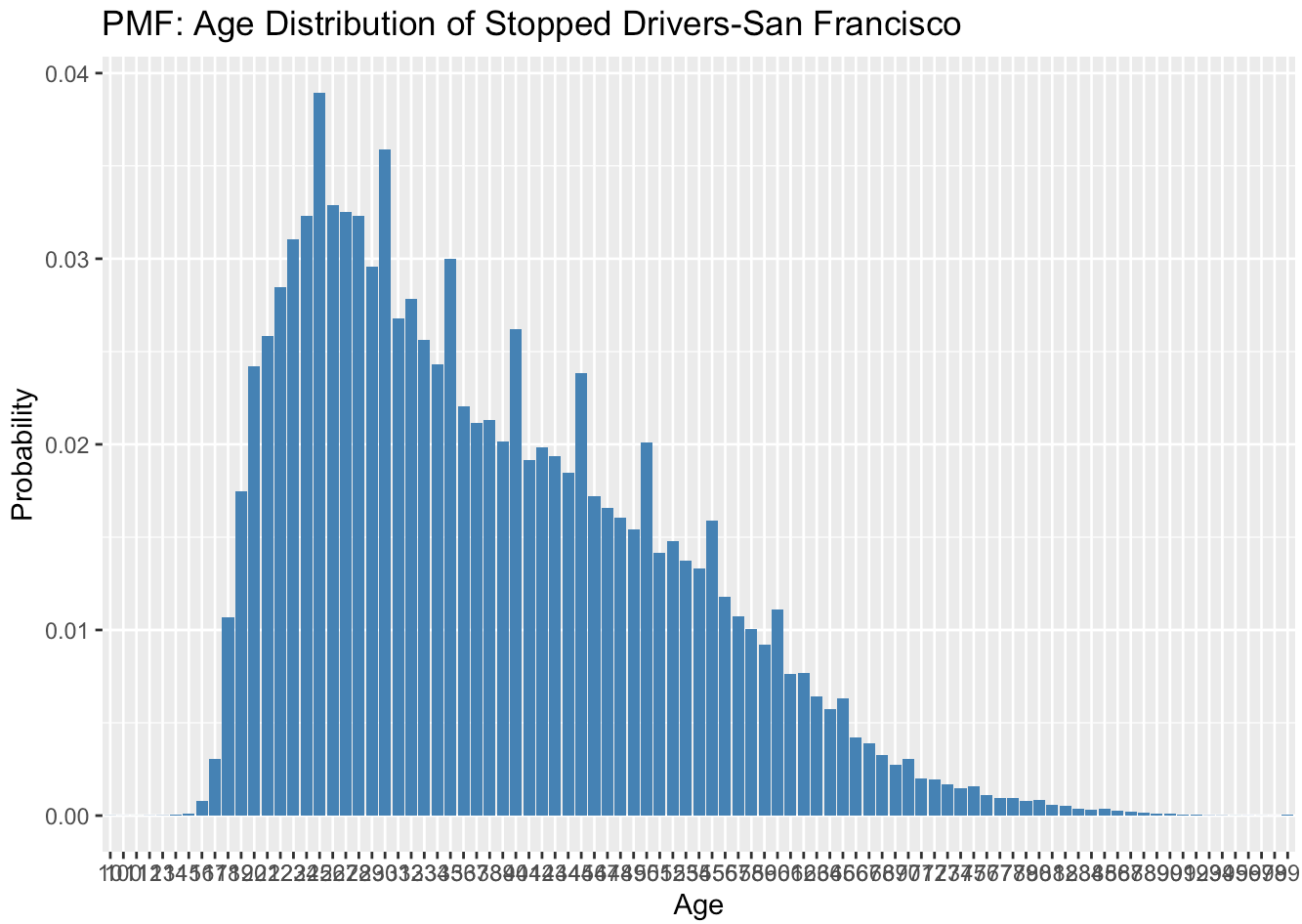

```
max_age <- pmf %>% filter(probability == max(probability))
print(max_age)
```

```
## # A tibble: 1 × 3
##   subject_age     n probability
##   <chr>       <int>       <dbl>
## 1 25          15545      0.0406
```

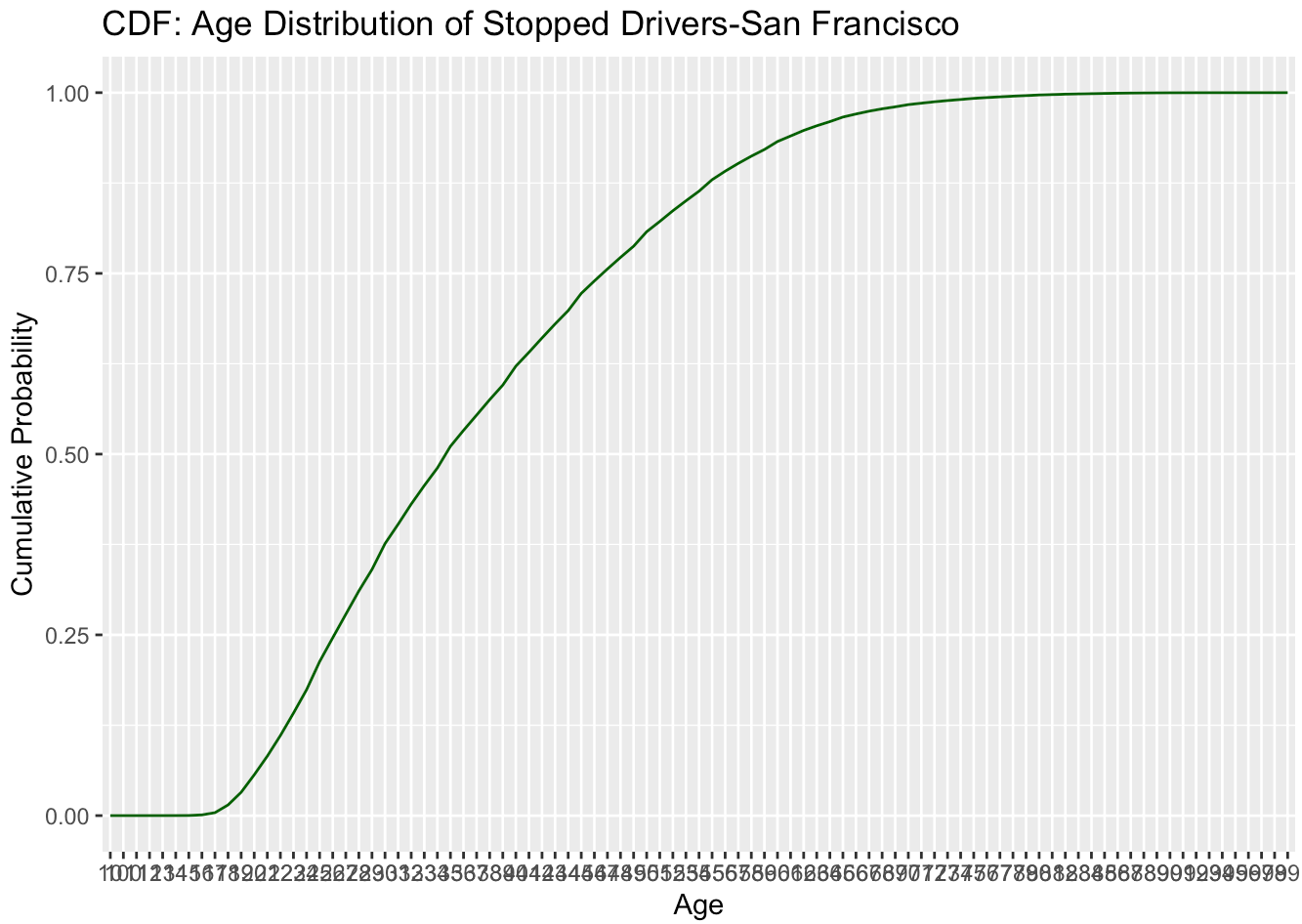
```
#For city san francisco
casanf <- read_excel("~/Desktop/Research Data/ca_san_francisco_2020_04_01.xlsx", col_typ
es = "text")
glimpse(casanf)
```

```
## Rows: 905,070
## Columns: 22
## $ raw_row_number                    <chr> "869921", "869922", "869923", "86992…
## $ date                              <chr> "2014-08-01", "2014-08-01", "2014-08…
## $ time                              <chr> "00:01:00", "00:01:00", "00:15:00", …
## $ location                          <chr> "MASONIC AV & FELL ST", "GEARY&10TH …
## $ lat                               <chr> "37.7730037", "37.7808985", "37.7869…
## $ lng                               <chr> "-122.4458727", "-122.4685858", "-12…
## $ district                          <chr> "NA", "NA", "NA", "NA", "NA", "NA", …
## $ subject_age                       <chr> "NA", "NA", "NA", "NA", "NA", "NA", …
## $ subject_race                      <chr> "asian/pacific islander", "black", "…
## $ subject_sex                       <chr> "female", "male", "male", "male", "m…
## $ type                              <chr> "vehicular", "vehicular", "vehicular…
## $ arrest_made                       <chr> "FALSE", "FALSE", "FALSE", "FALSE", …
## $ citation_issued                   <chr> "FALSE", "TRUE", "TRUE", "FALSE", "T…
## $ warning_issued                    <chr> "TRUE", "FALSE", "FALSE", "TRUE", "F…
## $ outcome                           <chr> "warning", "citation", "citation", "…
## $ contraband_found                  <chr> "NA", "NA", "NA", "NA", "NA", "NA", …
## $ search_conducted                  <chr> "FALSE", "FALSE", "FALSE", "FALSE", …
## $ search_vehicle                    <chr> "FALSE", "FALSE", "FALSE", "FALSE", …
## $ search_basis                      <chr> "NA", "NA", "NA", "NA", "NA", "NA", …
## $ reason_for_stop                   <chr> "Mechanical or Non-Moving Violation …
## $ raw_search_vehicle_description    <chr> "No Search", "No Search", "No Search…
## $ raw_result_of_contact_description <chr> "Warning", "Citation", "Citation", "…
```

```
casanf_clean <- casanf %>%
  filter(!is.na(subject_age), subject_age != "NA")

pmf <- casanf_clean %>%
  count(subject_age) %>%
  mutate(probability = n / sum(n))


cdf <- pmf %>%
  arrange(subject_age) %>%
  mutate(cumulative_probability = cumsum(probability))


ggplot(pmf, aes(x = subject_age, y = probability)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "PMF: Age Distribution of Stopped Drivers-San Francisco",
       x = "Age",
       y = "Probability")
```




```
ggplot(cdf, aes(x = subject_age, y = cumulative_probability, group = 1)) +
  geom_line(color = "darkgreen") +
  labs(title = "CDF: Age Distribution of Stopped Drivers-San Francisco",
       x = "Age",
       y = "Cumulative Probability")
```




```
max_age <- pmf %>%
  filter(!is.na(subject_age)) %>% # Ensure NAs are removed
  filter(probability == max(probability))

print(max_age)
```

```
## # A tibble: 1 × 3
##   subject_age     n probability
##   <chr>       <int>       <dbl>
## 1 25          32958      0.0389
```

b. We now want to investigate the relative probability of being stopped con- ditional on different ethnicity. In particular, we want to compute the prob- ability of being stopped conditional on being black, Hispanic, and Asian relative to the probability of being stopped conditional on being white. De- rive the formula for a generic pair of ethnicity (E1, E2) first, making use of the Bayes Theorem.

```
#For city San Diego
casand_clean <- casand_clean %>%
  filter(!is.na(subject_race))


race_counts <- casand_clean %>%
  count(subject_race) %>%
  mutate(proportion_stopped = n / sum(n))
print(race_counts)
```

```
## # A tibble: 6 × 3
##   subject_race                n proportion_stopped
##   <chr>                   <int>              <dbl>
## 1 NA                       1234            0.00322
## 2 asian/pacific islander  32541            0.0850
## 3 black                   42705            0.111
## 4 hispanic               117083            0.306
## 5 other                   27238            0.0711
## 6 white                  162226            0.424
```

```
p_black <- race_counts %>% filter(subject_race == "black") %>% pull(proportion_stopped)
p_white <- race_counts %>% filter(subject_race == "white") %>% pull(proportion_stopped)
p_hispanic <- race_counts %>% filter(subject_race == "hispanic") %>% pull(proportion_sto
pped)
p_asian <- race_counts %>% filter(subject_race == "asian") %>% pull(proportion_stopped)

#white：45.1%
#Black：6.31%
#hispanic：28.76%
#Asian：15.63%

#black relative to white = p(B|stopped)*p(w)/p(white|stopped)*p(b)=(0.111*0.451)/(0.424*
0.0631)=1.87
#asian relative to white = p(A|stopped)*p(w)/p(white|stopped)*p(A)=(0.085*0.451)/(0.424*
0.1563)= 0.58
#hispanic relative to white = p(H|stopped)*p(w)/p(white|stopped)*p(H)=(0.306*0.451)/(0.4
24*0.2876)=1.13



#For City San Francisco
casanf_clean <- casanf_clean %>%
  filter(!is.na(subject_race))


race_counts <- casanf_clean %>%
  count(subject_race) %>%
  mutate(proportion_stopped = n / sum(n))
print(race_counts)
```

```
## # A tibble: 5 × 3
##   subject_race               n proportion_stopped
##   <chr>                  <int>              <dbl>
## 1 asian/pacific islander 146746              0.173
## 2 black                  142759              0.169
## 3 hispanic               108302              0.128
## 4 other                   97647              0.115
## 5 white                  350728              0.414
```

```
p_black <- race_counts %>% filter(subject_race == "black") %>% pull(proportion_stopped)
p_white <- race_counts %>% filter(subject_race == "white") %>% pull(proportion_stopped)
p_hispanic <- race_counts %>% filter(subject_race == "hispanic") %>% pull(proportion_sto
pped)
p_asian <- race_counts %>% filter(subject_race == "asian") %>% pull(proportion_stopped)

#As of the 2020 census, the racial makeup and population of San Francisco included: 361,
382 Whites (41.3%), 296,505 Asians (33.9%), 46,725 African Americans (5.3%), 86,233 Mult
iracial Americans (9.9%), 6,475 Native Americans and Alaska Natives (0.7%), 3,476 Native
Hawaiians and other Pacific Islanders (0.4%) and 73,169 persons of other races (8.4%). T
here were 136,761 Hispanic or Latino residents of any race (15.6%).

#black relative to white = p(B|stopped)*p(w)/p(white|stopped)*p(b)=(0.168*0.413)/(0.411*
0.053)=3.19
#asian relative to white = p(A|stopped)*p(w)/p(white|stopped)*p(A)=(0.174*0.413)/(0.411*
0.339)=0.516
#hispanic relative to white = p(H|stopped)*p(w)/p(white|stopped)*p(H)=(0.128*0.413)/(0.4
11*0.156)=0.825
```

c. How does the results change if, in addition to conditioning on a particular race we also condition on the gender of the driver?

```
# The probability of being stopped conditioning on black female:p(stopped|Black & femal
e)
# Count by race and gender
race_gender_counts <- casanf_clean %>%
  count(subject_race, subject_sex) %>%
  mutate(proportion_stopped = n / sum(n))

print(race_gender_counts)
```

```
## # A tibble: 10 × 4
##    subject_race           subject_sex      n proportion_stopped
##    <chr>                  <chr>        <int>              <dbl>
##  1 asian/pacific islander female       44579             0.0527
##  2 asian/pacific islander male        102167             0.121
##  3 black                  female       42363             0.0501
##  4 black                  male        100396             0.119
##  5 hispanic               female       24905             0.0294
##  6 hispanic               male         83397             0.0986
##  7 other                  female       20209             0.0239
##  8 other                  male         77438             0.0915
##  9 white                  female      116901             0.138
## 10 white                  male        233827             0.276
```

```
# Extracting specific race-gender proportions
p_black_female <- race_gender_counts %>% filter(subject_race == "black", subject_sex == "female") %>% pull(proportion_stopped)
p_black_male <- race_gender_counts %>% filter(subject_race == "black", subject_sex == "male") %>% pull(proportion_stopped)

p_white_female <- race_gender_counts %>% filter(subject_race == "white", subject_sex == "female") %>% pull(proportion_stopped)
p_white_male <- race_gender_counts %>% filter(subject_race == "white", subject_sex == "male") %>% pull(proportion_stopped)

p_hispanic_female <- race_gender_counts %>% filter(subject_race == "hispanic", subject_sex == "female") %>% pull(proportion_stopped)
p_hispanic_male <- race_gender_counts %>% filter(subject_race == "hispanic", subject_sex == "male") %>% pull(proportion_stopped)

p_asian_female <- race_gender_counts %>% filter(subject_race == "asian/pacific islander", subject_sex == "female") %>% pull(proportion_stopped)
p_asian_male <- race_gender_counts %>% filter(subject_race == "asian/pacific islander", subject_sex == "male") %>% pull(proportion_stopped)

# Displaying results
p_black_female
```

```
## [1] 0.0500637
```

```
p_black_male
```

```
## [1] 0.1186459
```

```
p_white_female
```

```
## [1] 0.1381511
```

```
p_white_male
```

```
## [1] 0.2763318
```

```
p_hispanic_female
```

```
## [1] 0.0294322
```

```
p_hispanic_male
```

```
## [1] 0.09855681
```

```
p_asian_female
```

```
## [1] 0.05268252
```

```
p_asian_male
```

```
## [1] 0.1207388
```

```
#According to the 2010 U.S. Census, the population of San Diego city, California, was 1,307,402. The gender distribution was approximately 50.2% male and 49.8% female.
```

d. Compute the probability of conducting a search conditional on being stopped for each ethnicity.

```
search_counts <- casanf_clean %>%
  group_by(subject_race) %>%
  summarize(
    total_stops = n(),
    searches_conducted = sum(search_conducted == TRUE)
  ) %>%
  mutate(proportion_searched = searches_conducted / total_stops)

print(search_counts)
```

```
## # A tibble: 5 × 4
##   subject_race           total_stops searches_conducted proportion_searched
##   <chr>                        <int>              <int>               <dbl>
## 1 asian/pacific islander      146746               2692              0.0183
## 2 black                       142759              22213              0.156
## 3 hispanic                    108302              10960              0.101
## 4 other                        97647               3569              0.0366
## 5 white                       350728              11207              0.0320
```

```
p_black_searched <- search_counts %>% filter(subject_race == "black") %>% pull(proportio
n_searched)
p_white_searched <- search_counts %>% filter(subject_race == "white") %>% pull(proportio
n_searched)
p_hispanic_searched <- search_counts %>% filter(subject_race == "hispanic") %>% pull(pro
portion_searched)
p_asian_searched <- search_counts %>% filter(subject_race == "asian/pacific islander") %
>% pull(proportion_searched)

p_black_searched
```

```
## [1] 0.1555979
```

```
p_white_searched
```

```
## [1] 0.03195354
```

```
p_hispanic_searched
```

```
## [1] 0.1011985
```

```
p_asian_searched
```

```
## [1] 0.01834462
```

```
#Compute the probability of contraband found conditional on search for each ethnicity.
#p(contraband | search conducted and white)

contraband_rate_by_race <- casanf_clean %>%
  filter(search_conducted == TRUE) %>% # Only consider rows where a search was conducted
  group_by(subject_race) %>%
  summarize(
    total_searches = n(),
    contraband_found = sum(contraband_found == TRUE),
    p_contraband_given_search = contraband_found / total_searches
  )

print(contraband_rate_by_race)
```

```
## # A tibble: 5 × 4
##   subject_race           total_searches contraband_found p_contraband_given_se…¹
##   <chr>                           <int>            <int>                    <dbl>
## 1 asian/pacific islander           2692              942                    0.350
## 2 black                           22213             2027                    0.0913
## 3 hispanic                        10960             1073                    0.0979
## 4 other                            3569              708                    0.198
## 5 white                           11207             2658                    0.237
## # ℹ abbreviated name: ¹p_contraband_given_search
```

```
p_contraband_white <- contraband_rate_by_race %>% filter(subject_race == "white") %>% pu
ll(p_contraband_given_search)
p_contraband_asian <- contraband_rate_by_race %>% filter(subject_race == "asian/pacific
islander") %>% pull(p_contraband_given_search)
p_contraband_black <- contraband_rate_by_race %>% filter(subject_race == "black") %>% pu
ll(p_contraband_given_search)
p_contraband_hispanic <- contraband_rate_by_race %>% filter(subject_race == "hispanic")
%>% pull(p_contraband_given_search)

p_contraband_white
```

```
## [1] 0.2371732
```

```
p_contraband_asian
```

```
## [1] 0.3499257
```

```
p_contraband_black
```

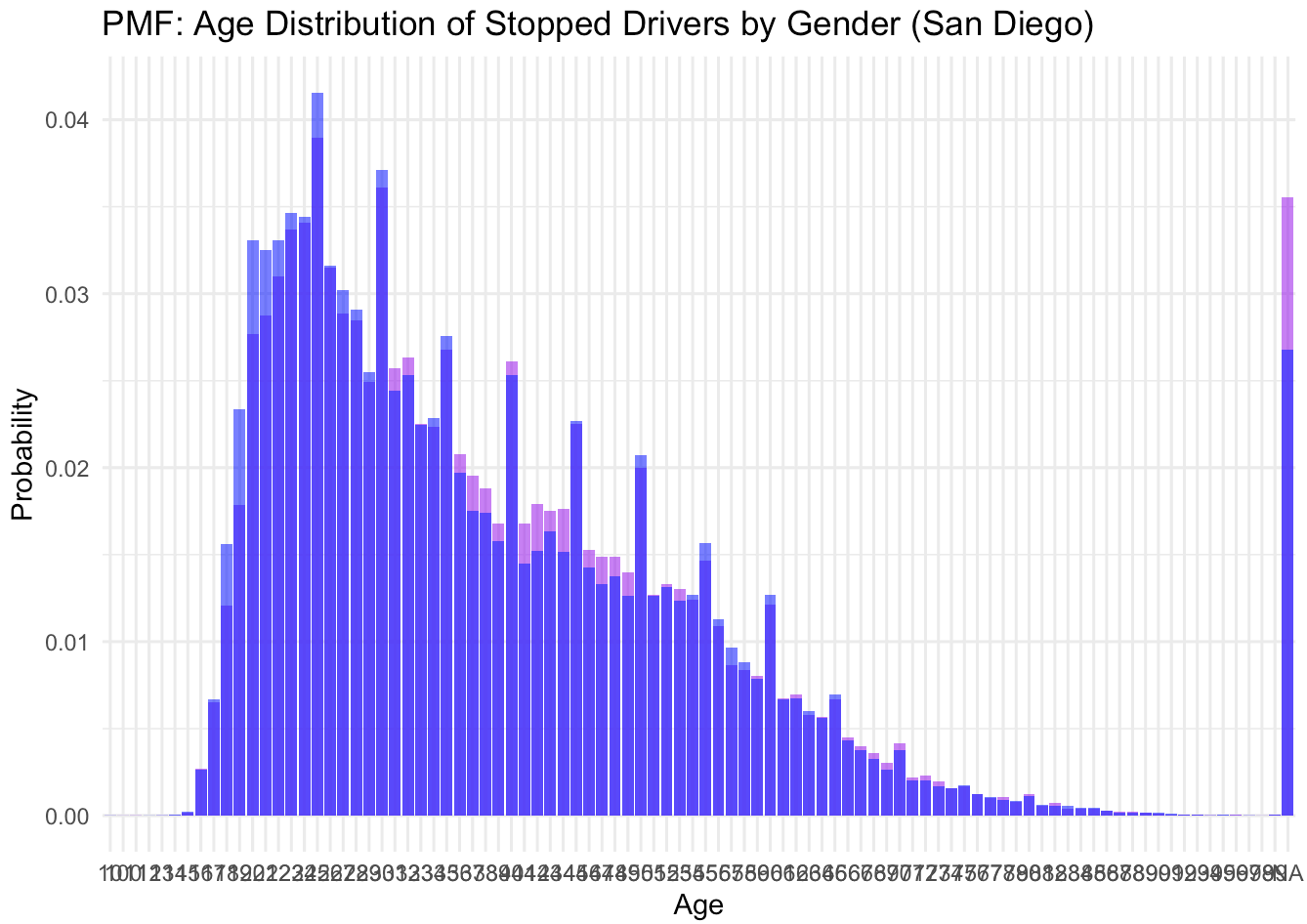
```
## [1] 0.09125287
```

```
p_contraband_hispanic
```

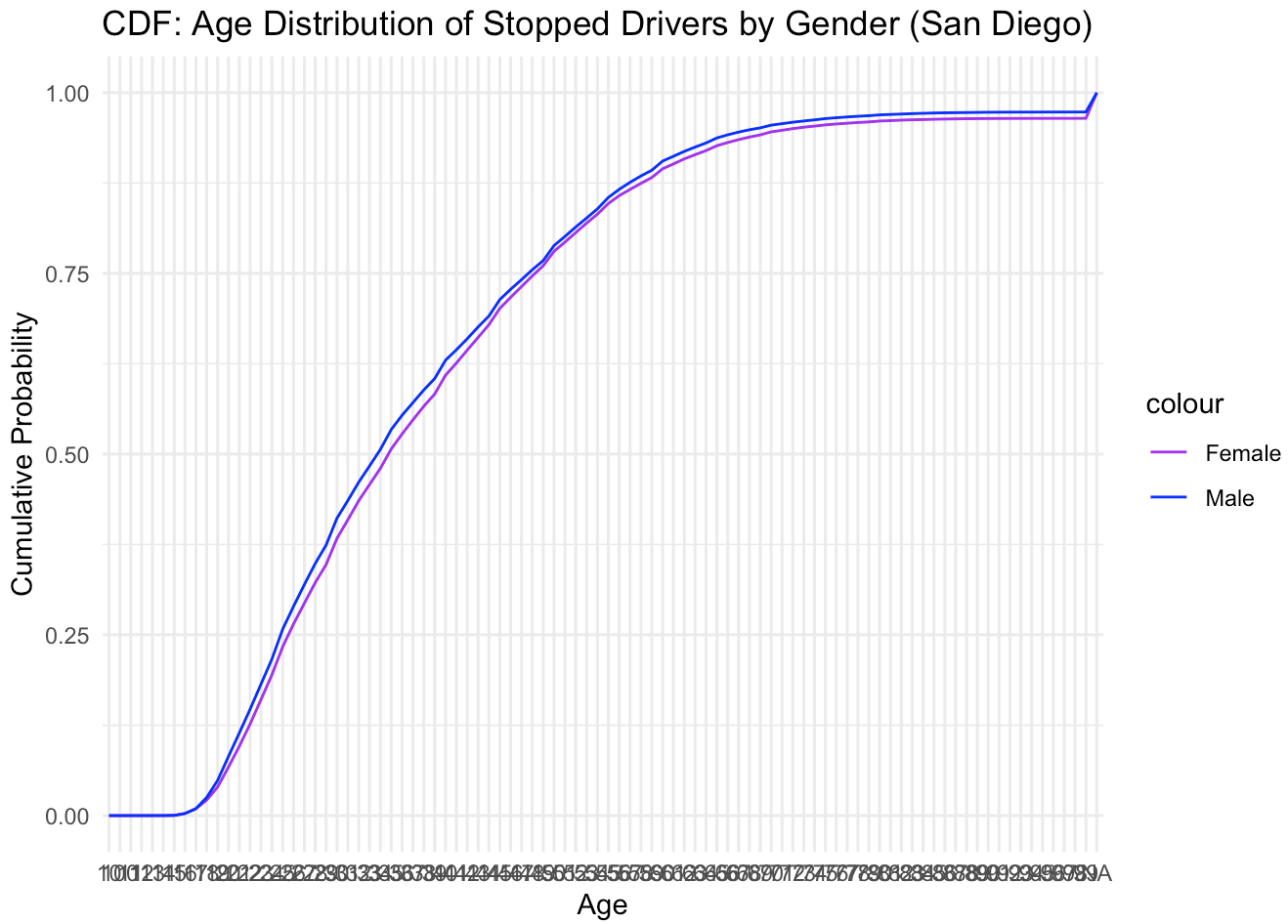
```
## [1] 0.09790146
```

#question 4

```
# Load data
casand <- read_excel("~/Desktop/Research Data/ca_san_diego_2020_04_01.xlsx")

# Clean data
casand_clean <- casand %>%
  filter(!is.na(subject_age), !is.na(subject_sex))

# PMF for females
pmf_female <- casand_clean %>%
  filter(subject_sex == "female") %>%
  count(subject_age) %>%
  mutate(probability = n / sum(n))

# PMF for males
pmf_male <- casand_clean %>%
  filter(subject_sex == "male") %>%
  count(subject_age) %>%
  mutate(probability = n / sum(n))

# CDF for females
cdf_female <- pmf_female %>%
  arrange(subject_age) %>%
  mutate(cumulative_probability = cumsum(probability))

# CDF for males
cdf_male <- pmf_male %>%
  arrange(subject_age) %>%
  mutate(cumulative_probability = cumsum(probability))

# Plot PMF
ggplot() +
  geom_bar(data = pmf_female, aes(x = subject_age, y = probability), stat = "identity",
fill = "purple", alpha = 0.6) +
  geom_bar(data = pmf_male, aes(x = subject_age, y = probability), stat = "identity", fi
ll = "blue", alpha = 0.6) +
  labs(title = "PMF: Age Distribution of Stopped Drivers by Gender (San Diego)",
       x = "Age",
       y = "Probability") +
  theme_minimal()
```

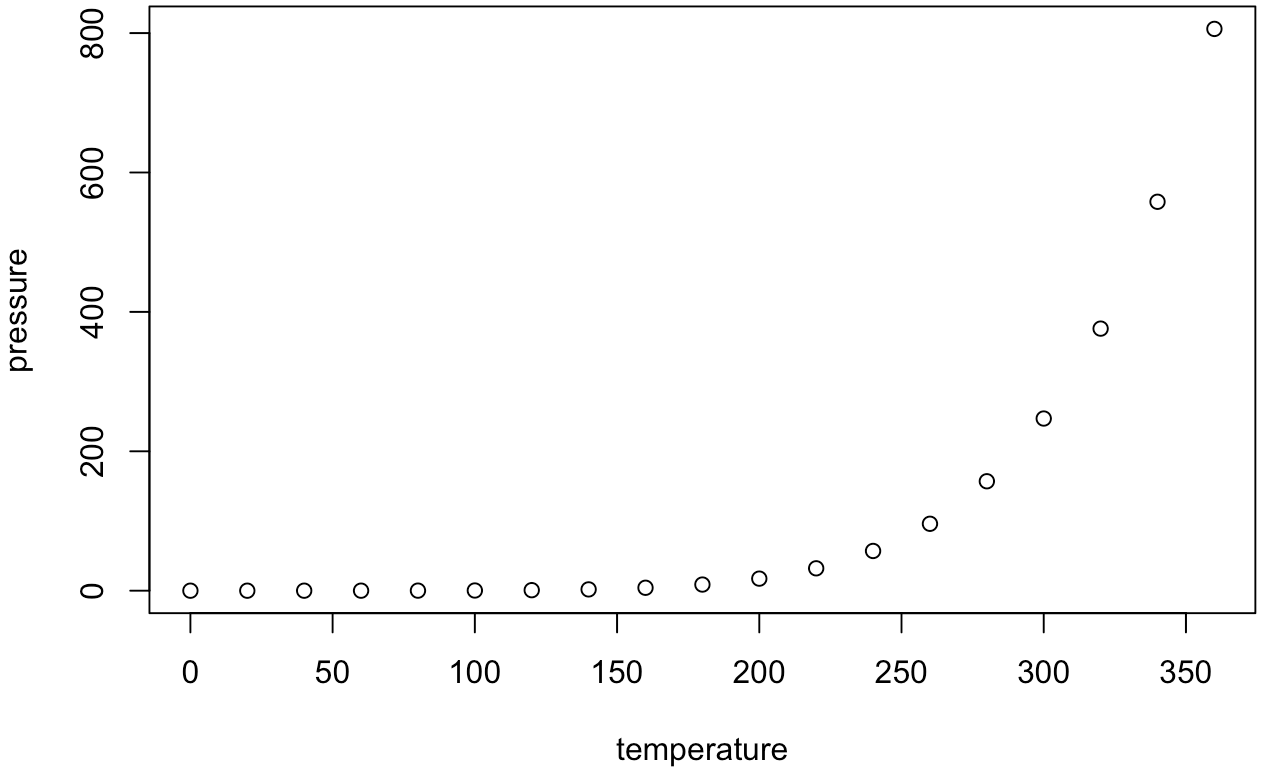



```
# Plot CDF
ggplot() +
  geom_line(data = cdf_female, aes(x = subject_age, y = cumulative_probability, group =
1, color = "Female")) +
  geom_line(data = cdf_male, aes(x = subject_age, y = cumulative_probability, group = 1,
color = "Male")) +
  labs(title = "CDF: Age Distribution of Stopped Drivers by Gender (San Diego)",
       x = "Age",
       y = "Cumulative Probability") +
  scale_color_manual(values = c("Female" = "purple", "Male" = "blue")) +
  theme_minimal()
```



Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.