TCGA_LUAD Demographic Analysis

Hsiao-Yu Peng

2023-12-14

```
# Load data
lung_phe <- read_tsv("~/Desktop/methylation/lung/TCGA-LUAD.GDC_phenotype.tsv")
```

```
## Rows: 877 Columns: 125
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (78): submitter_id.samples, additional_pharmaceutical_therapy, additiona...
## dbl (41): age_at_initial_pathologic_diagnosis, day_of_dcc_upload, day_of_for...
## lgl  (6): withdrawn, releasable.project, days_to_sample_procurement.samples,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(lung_phe)
```

```
## [1] 877 125
```

```
head(lung_phe)
```

```
## # A tibble: 6 x 125
##   submitter_id.samples additional_pharmaceutical_therapy additional_radiation_~1
##   <chr>                <chr>                             <chr>
## 1 TCGA-62-A46Y-01A     <NA>                              <NA>
## 2 TCGA-55-7903-01A     <NA>                              <NA>
## 3 TCGA-55-7903-11A     <NA>                              <NA>
## 4 TCGA-97-8179-01A     <NA>                              <NA>
## 5 TCGA-64-5774-01A     NO                                YES
## 6 TCGA-55-8092-01A     NO                                NO
## # i abbreviated name: 1: additional_radiation_therapy
## # i 122 more variables: additional_surgery_locoregional_procedure <chr>,
## #   additional_surgery_metastatic_procedure <chr>,
## #   age_at_initial_pathologic_diagnosis <dbl>,
## #   anatomic_neoplasm_subdivision_other <chr>, batch_number <chr>, bcr <chr>,
## #   bcr_followup_barcode <chr>, bcr_followup_uuid <chr>, submitter_id <chr>,
## #   day_of_dcc_upload <dbl>, day_of_form_completion <dbl>, ...
```

Explore Data Analsyis

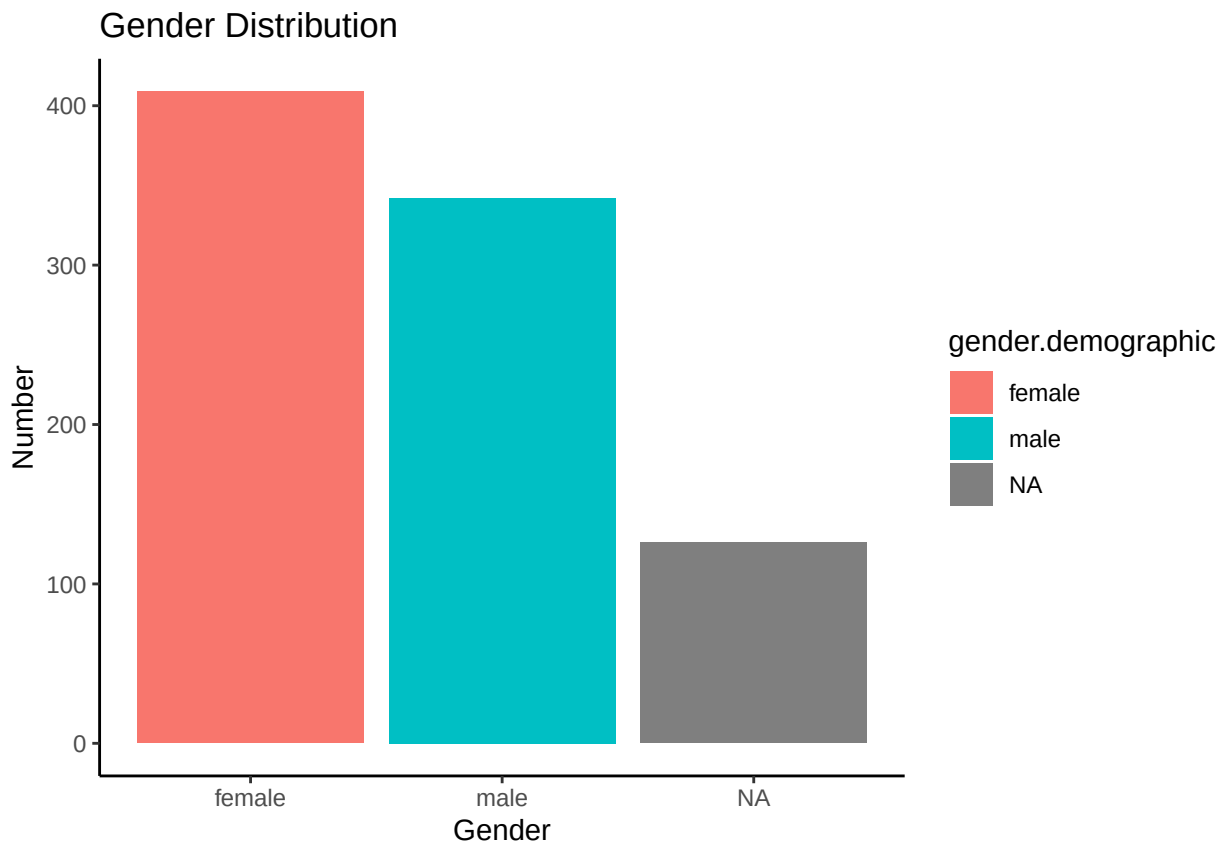
Gender Distribution

```
gender_dist <- lung_phe %>%
  group_by(gender.demographic) %>%
  summarize(number = n())

print(gender_dist)
```

```
## # A tibble: 3 x 2
##   gender.demographic number
##   <chr>                <int>
## 1 female                 409
## 2 male                   342
## 3 <NA>                   126
```

```
ggplot(gender_dist, aes(x = gender.demographic, y = number, fill = gender.demographic)) +
  geom_bar(stat = "identity") +
  labs(title = "Gender Distribution",
       x = "Gender",
       y = "Number") +
  theme_classic()
```



TCGA_LUAD data set has more female individuals than male. Because we focus on studying female nonsmoker with LUAD, we select female for further investigation.

Smoking History Distribution

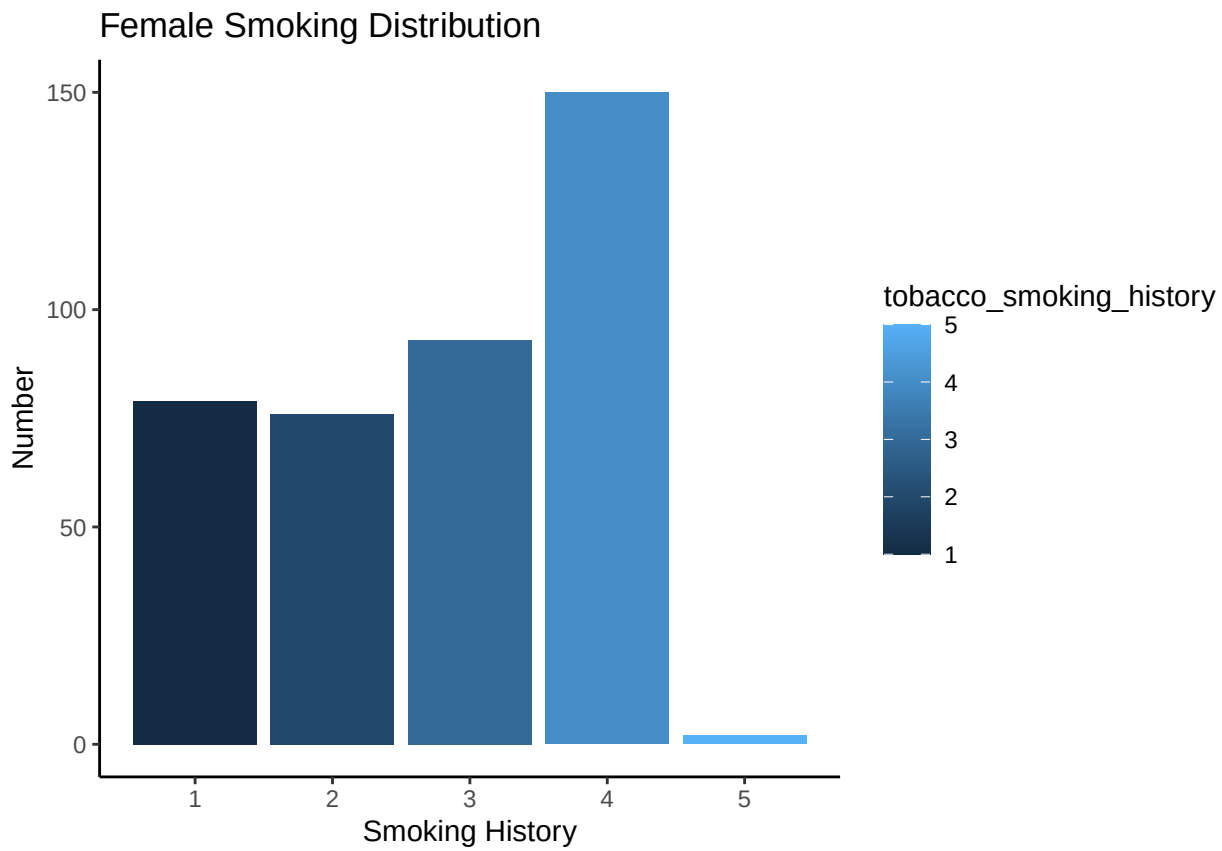
```
# Select female and gropy by smoking history
smoking_dist <- lung_phe %>%
  filter(gender.demographic == "female") %>%
  group_by(tobacco_smoking_history) %>%
  summarize(number = n())

print(smoking_dist)
```

```
## # A tibble: 6 x 2
##   tobacco_smoking_history number
##                     <dbl>  <int>
## 1                       1     79
## 2                       2     76
## 3                       3     93
## 4                       4    150
## 5                       5      2
## 6                      NA      9
```

```
# Create bar plot
ggplot(smoking_dist, aes(x = tobacco_smoking_history, y= number, fill = tobacco_smoking_history )) +
  geom_bar(stat = "identity") +
  labs(title = "Female Smoking Distribution",
       x = "Smoking History",
       y = "Number") +
  theme_classic()
```

```
## Warning: Removed 1 rows containing missing values (`position_stack()`).
```



Category "1" of Smoking History means non-smokers, others are smokers. We continue to study race distribution of female nonsmokers and smokers

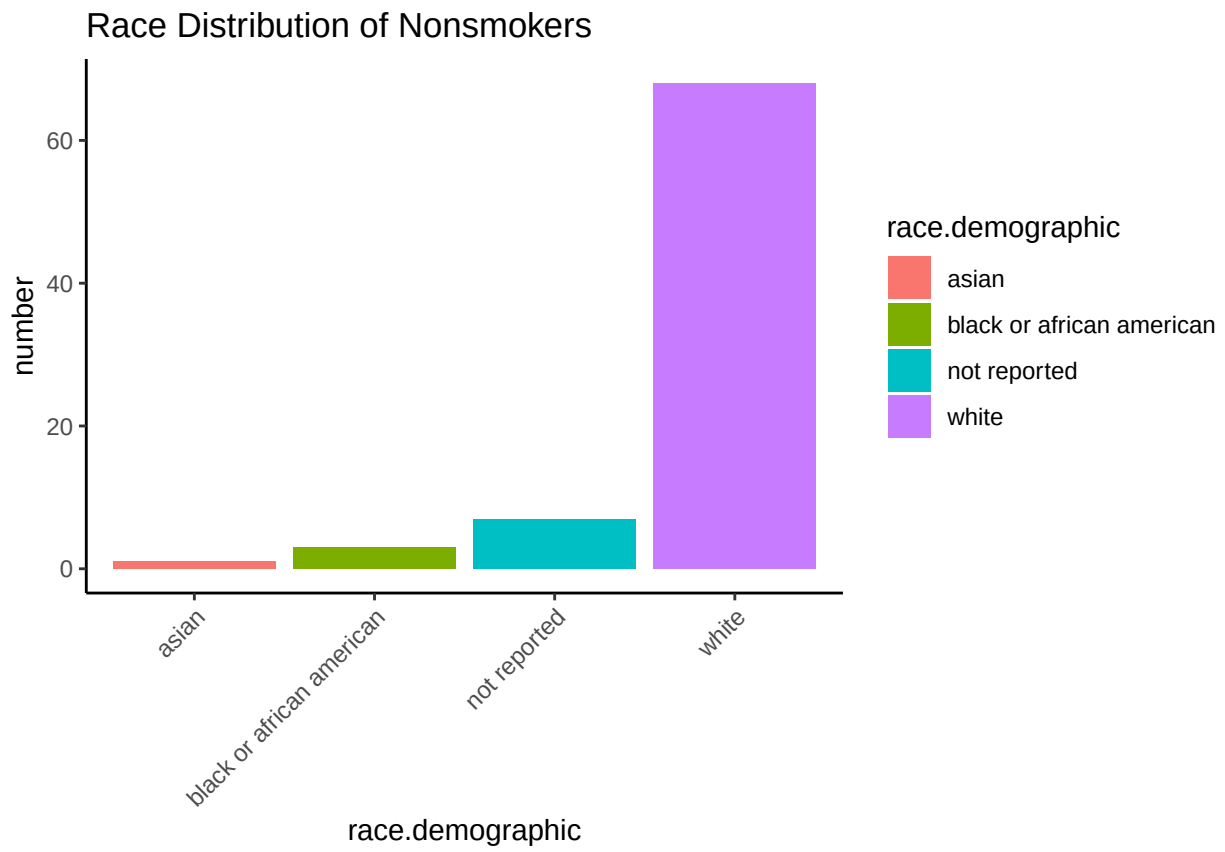
Race Distribution

```
# Select female nonsmokers
race_dist <- lung_phe %>%
  filter(gender.demographic == "female",
         tobacco_smoking_history == 1) %>%
  group_by(race.demographic) %>%
  summarize(number = n())

print(race_dist)
```

```
## # A tibble: 4 x 2
##   race.demographic          number
##   <chr>                      <int>
## 1 asian                          1
## 2 black or african american      3
## 3 not reported                   7
## 4 white                         68
```

```
ggplot(race_dist, aes(x = race.demographic, y = number, fill = race.demographic)) +
  geom_bar(stat = "identity") +
  labs(title = "Race Distribution of Nonsmokers") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

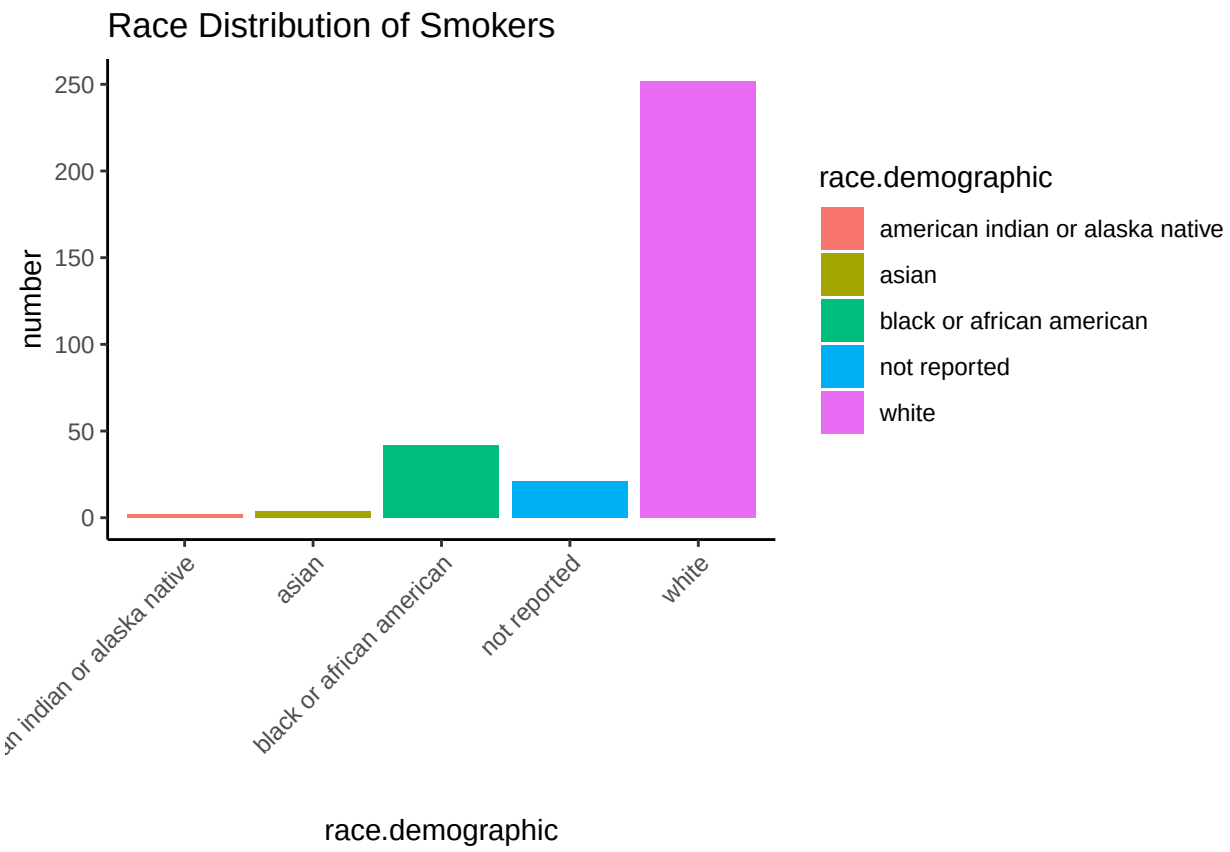



```
# Select female smokers
race_dist <- lung_phe %>%
  filter(gender.demographic == "female",
         tobacco_smoking_history != 1) %>%
  group_by(race.demographic) %>%
  summarize(number = n())


print(race_dist)
```

```
## # A tibble: 5 x 2
##   race.demographic                 number
##   <chr>                             <int>
## 1 american indian or alaska native      2
## 2 asian                                 4
## 3 black or african american            42
## 4 not reported                         21
## 5 white                               252
```

```r
ggplot(race_dist, aes(x = race.demographic, y = number, fill = race.demographic)) +
  geom_bar(stat = "identity") +
  labs(title = "Race Distribution of Smokers") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We only study race regarding Asian, Black or African American and White. Next step we explore age distribution.

Non-smoker Age Distribution

```r
# Select female nonsmokers
nonsmokers_age <- lung_phe %>%
  filter(tobacco_smoking_history ==1,
         gender.demographic == "female",
         race.demographic %in% c("asian", "black or african american", "white", "not reported")) %>%
  select(submitter_id.samples, age_at_index.demographic, gender.demographic,
         race.demographic, tobacco_smoking_history)

# Find non-smokers' age range
nonsmokers_age %>%
  group_by(race.demographic) %>%
  summarize(age_range = paste(min(age_at_index.demographic, na.rm = T),
                              max(age_at_index.demographic, na.rm = T),
```

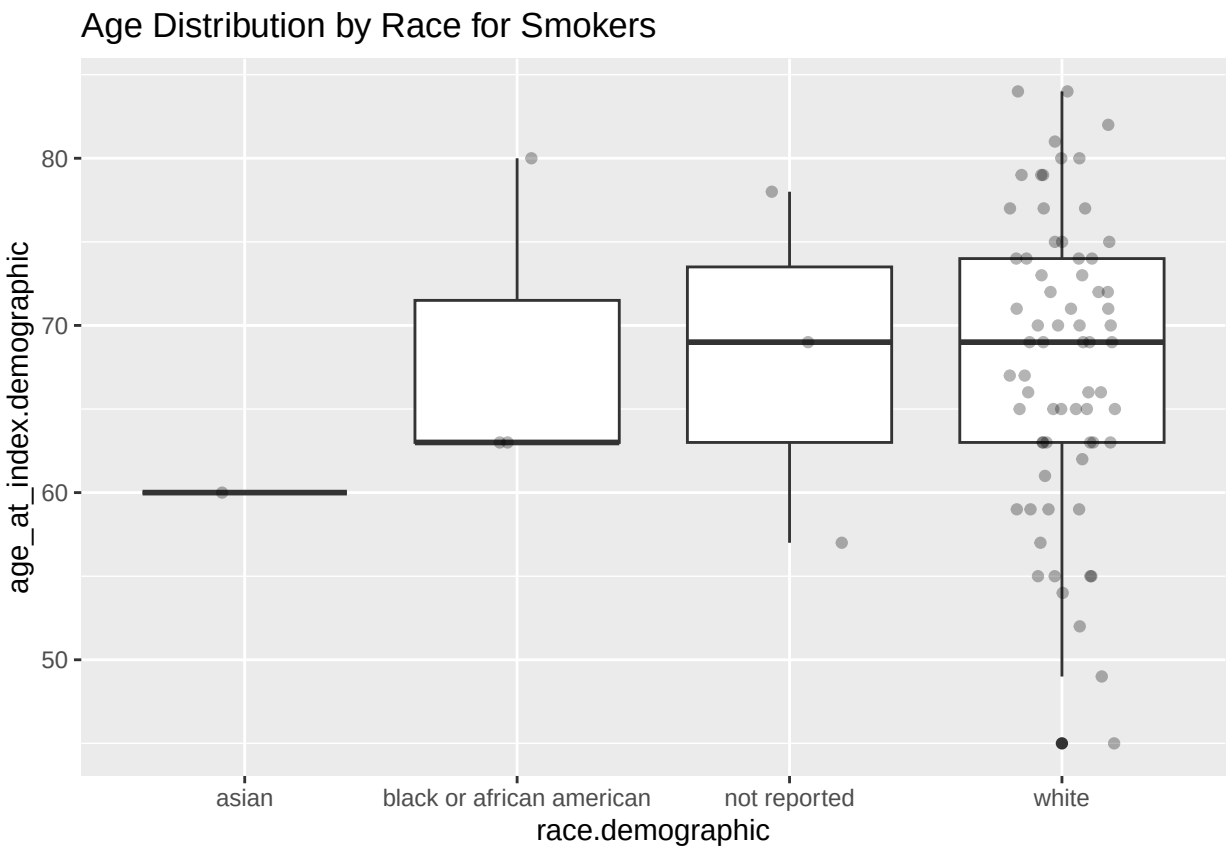
```
                              sep = " - "),
            count = n())
```

```
## # A tibble: 4 x 3
##   race.demographic          age_range count
##   <chr>                     <chr>     <int>
## 1 asian                     60 - 60       1
## 2 black or african american 63 - 80       3
## 3 not reported              57 - 78       7
## 4 white                     45 - 84      68
```

```
ggplot(nonsmokers_age, aes(x = race.demographic, y = age_at_index.demographic)) +
  geom_boxplot() +
  geom_jitter(position = position_jitter(width = 0.2, height = 0), alpha = 0.3) +
  labs(title = "Age Distribution by Race for Smokers")
```

```
## Warning: Removed 4 rows containing non-finite values (‘stat_boxplot()‘).
```

```
## Warning: Removed 4 rows containing missing values (‘geom_point()‘).
```



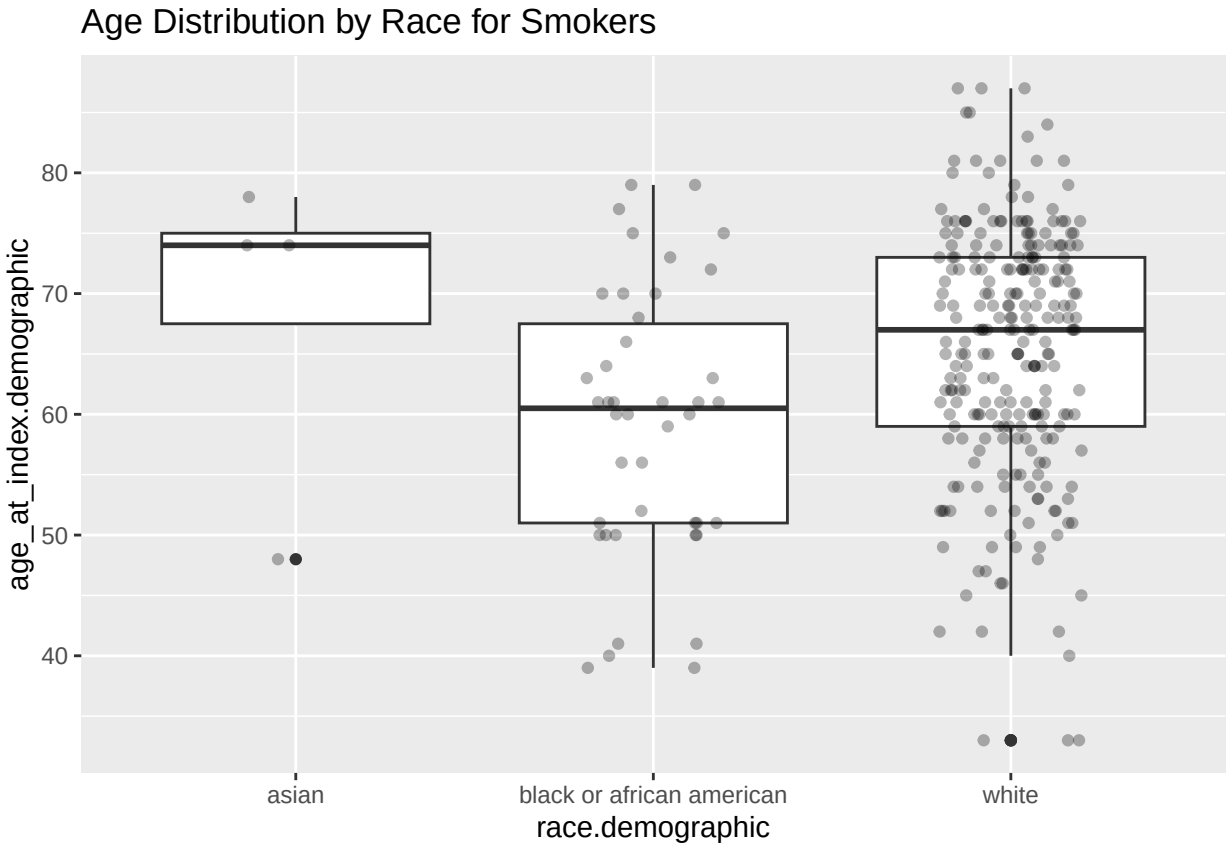
Smokers Age Distribution

```r
smokers_age <- lung_phe %>%
  filter(tobacco_smoking_history !=1,
         gender.demographic == "female",
         race.demographic %in% c("asian", "black or african american", "white")) %>%
  select(submitter_id.samples, age_at_index.demographic, gender.demographic,
         race.demographic, tobacco_smoking_history)

# Find smokers' age range
smokers_age %>%
  group_by(race.demographic) %>%
  summarize(age_range = paste(min(age_at_index.demographic, na.rm = T),
                              max(age_at_index.demographic, na.rm = T),
                              sep = " - "),
            count = n())
```

```
## # A tibble: 3 x 3
##   race.demographic          age_range count
##   <chr>                     <chr>     <int>
## 1 asian                     48 - 78       4
## 2 black or african american 39 - 79      42
## 3 white                     33 - 87     252
```

```r
ggplot(smokers_age, aes(x = race.demographic, y = age_at_index.demographic)) +
  geom_boxplot() +
  geom_jitter(position = position_jitter(width = 0.2, height = 0), alpha = 0.3) +
  labs(title = "Age Distribution by Race for Smokers")
```


Age Distribution by Race for Smokers
age_at_index.demographic
80
70
60
50
40
asian
black or african american
white
race.demographic