CS03 - Vaping Behaviors in American Youth

Yohan Kim, Ming Qiu, Kevin Lam, Khiem Pham

Introduction

According to a report from the CDC, there has been a rise in tobacco/nicotine usage in American Youth, which is primarily attributed to the use of electronic cigarettes[1]. Because of how simple it is to use and hide, it can make it easier for youths to become dependent and harder for parents to intervene. For this case study, we will investigate e-cigarette usage by American youths from 2015 onwards to determine any trends between gender, vaping brands, and other tobacco use.

Load packages

```
library(OCSdata)
library(tidyverse)
library(broom)
library(tidymodels)
library(viridis)
library(scales)
library(srvyr)
```

Question

1. How has tobacco and e-cigarette/vaping use by American youths changed since 2015?
2. How does e-cigarette use compare between males and females?
3. What vaping brands and flavors appear to be used the most frequently?
4. Is there a relationship between e-cigarette/vaping use and other tobacco use?
5. Extended question: how have the current vaping/e-cigarette users who are under the age of 18 changed from 2015 to 2019 compared to adult users

The Data

The data comes from the National Youth Tobacco Survey, a survey that asks students from American high schools and middle schools about tobacco use every year[2]. Each year contains its own codebook for describing the data within the dataset, as the questions being asked slightly differ each year. For this case study, we'll be using the data from 2015-2019.

Data Import

```
# import the data
# OCSdata::load_simpler_import("ocs-bp-vaping-case-study", outpath = getwd())
```

Data Wrangling

To begin, we first read in all the CSV files and add names to them.

```
# check csv files in data/simpler_import/ and displays the path name
# goes through each csv file and reads it in
nyts_data <- list.files("data/simpler_import/",
                        pattern = "*.csv",
                        full.names = TRUE) |>
  map(~ read_csv(.))
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   psu = col_character(),
##   stratum = col_character()
## )
## ℹ Use `spec()` for the full column specifications.
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   psu = col_character(),
##   stratum = col_character(),
##   Q1 = col_character()
## )
## ℹ Use `spec()` for the full column specifications.
```

```
## Warning: 21 parsing failures.
##  row col expected actual                                file
## 2284  Q2 a double      * 'data/simpler_import//nyts2016.csv'
## 3406  Q3 a double      * 'data/simpler_import//nyts2016.csv'
## 4218  Q2 a double      * 'data/simpler_import//nyts2016.csv'
## 5725  Q2 a double      * 'data/simpler_import//nyts2016.csv'
## 5725  Q3 a double      * 'data/simpler_import//nyts2016.csv'
## .... ... ........ ...... ...................................
## See problems(...) for more details.
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   psu = col_character(),
##   stratum = col_character(),
##   Q1 = col_character(),
##   Q3 = col_character()
## )
## ℹ Use `spec()` for the full column specifications.
```

```
## Warning: 8 parsing failures.
##  row col expected actual                                file
## 1197  Q2 a double      * 'data/simpler_import//nyts2017.csv'
## 1892  Q2 a double      * 'data/simpler_import//nyts2017.csv'
## 8356  Q2 a double      * 'data/simpler_import//nyts2017.csv'
## 9041  Q2 a double      * 'data/simpler_import//nyts2017.csv'
## 9114  Q2 a double      * 'data/simpler_import//nyts2017.csv'
## .... ... ........ ...... ...................................
## See problems(...) for more details.
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   psu = col_character(),
##   stratum = col_character(),
##   Q1 = col_character(),
##   Q3 = col_character()
## )
## ℹ Use `spec()` for the full column specifications.
```

```
## Warning: 13 parsing failures.
##   row col expected actual                                file
##  3392  Q2 a double      * 'data/simpler_import//nyts2018.csv'
##  4488  Q2 a double      * 'data/simpler_import//nyts2018.csv'
##  6452  Q2 a double      * 'data/simpler_import//nyts2018.csv'
##  8594  Q2 a double      * 'data/simpler_import//nyts2018.csv'
## 10464  Q2 a double      * 'data/simpler_import//nyts2018.csv'
## ..... ... ........ ...... ...................................
## See problems(...) for more details.
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_character(),
##   psu = col_double(),
##   finwgt = col_double()
## )
## ℹ Use `spec()` for the full column specifications.
```

```r
# check csv files in data/simpler_import/ and displays file names
# get rid of csv extension from file names
nyts_data_names <- list.files("data/simpler_import/",
                              pattern = "*.csv") |>
  str_extract("nyts201[5-9]")

# adds names of files to tibbles in nyts_data
names(nyts_data) <- nyts_data_names
# glimpse(nyts_data)
```

We rename the variable names in the 2015 data based on what's specified in the data dictionary.

```r
nyts_data[["nyts2015"]] <- nyts_data[["nyts2015"]] |>
  rename(Age = Qn1,
         Sex = Qn2,
         Grade = Qn3)
```

We create a function, update_survey, that renames columns so they fit the data dictionary. We apply this function to the 2016-2018 data.

```r
# take in a dataset and rename columns
update_survey <- function(dataset) {
  dataset |>
    rename(Age = Q1,
           Sex = Q2,
           Grade = Q3,
           menthol = Q50A,
           clove_spice = Q50B,
           fruit = Q50C,
           chocolate = Q50D,
           alcoholic_drink = Q50E,
           candy_dessert_sweets = Q50F,
           other = Q50G)
}

# apply the update_survey function to all datasets, excluding 2015 and 2019
nyts_data <- nyts_data |>
  map_at(vars(-nyts2015, -nyts2019), update_survey)
```

We rename the variable names in the 2019 data so it matches the data dictionary.

```r
nyts_data[["nyts2019"]] <- nyts_data[["nyts2019"]] |>
  rename(brand_ecig = Q40,
         Age = Q1,
         Sex = Q2,
         Grade = Q3,
         menthol = Q62A,
         clove_spice = Q62B,
         fruit = Q62C,
         chocolate = Q62D,
         alcoholic_drink = Q62E,
         candy_dessert_sweets = Q62F,
         other = Q62G)
```

We check the data to see if all the names are consistent throughout the different years

```r
map(nyts_data, names)
```

```
## $nyts2015
##  [1] "psu"        "finwgt"     "stratum"    "Age"        "Sex"
##  [6] "Grade"      "ECIGT"      "ECIGAR"     "ESLT"       "EELCIGT"
## [11] "EROLLCIGTS" "EFLAVCIGTS" "EBIDIS"     "EFLAVCIGAR" "EHOOKAH"
## [16] "EPIPE"      "ESNUS"      "EDISSOLV"   "CCIGT"      "CCIGAR"
## [21] "CSLT"       "CELCIGT"    "CROLLCIGTS" "CFLAVCIGTS" "CBIDIS"
## [26] "CHOOKAH"    "CPIPE"      "CSNUS"      "CDISSOLV"
##
## $nyts2016
##  [1] "psu"                  "finwgt"               "stratum"
##  [4] "Age"                  "Sex"                  "Grade"
##  [7] "ECIGT"                "ECIGAR"               "ESLT"
## [10] "EELCIGT"              "EHOOKAH"              "EROLLCIGTS"
## [13] "EFLAVCIGAR"           "EPIPE"                "ESNUS"
## [16] "EDISSOLV"             "EBIDIS"               "CCIGT"
## [19] "CCIGAR"               "CSLT"                 "CELCIGT"
## [22] "CHOOKAH"              "CROLLCIGTS"           "CPIPE"
## [25] "CSNUS"                "CDISSOLV"             "CBIDIS"
## [28] "menthol"              "clove_spice"          "fruit"
## [31] "chocolate"            "alcoholic_drink"      "candy_dessert_sweets"
## [34] "other"
##
## $nyts2017
##  [1] "psu"                  "finwgt"               "stratum"
##  [4] "Age"                  "Sex"                  "Grade"
##  [7] "ECIGT"                "ECIGAR"               "ESLT"
## [10] "EELCIGT"              "EHOOKAH"              "EROLLCIGTS"
## [13] "EPIPE"                "ESNUS"                "EDISSOLV"
## [16] "EBIDIS"               "CCIGT"                "CCIGAR"
## [19] "CSLT"                 "CELCIGT"              "CHOOKAH"
## [22] "CROLLCIGTS"           "CPIPE"                "CSNUS"
## [25] "CDISSOLV"             "CBIDIS"               "menthol"
## [28] "clove_spice"          "fruit"                "chocolate"
## [31] "alcoholic_drink"      "candy_dessert_sweets" "other"
##
## $nyts2018
##  [1] "psu"                  "finwgt"               "stratum"
##  [4] "Age"                  "Sex"                  "Grade"
##  [7] "ECIGT"                "ECIGAR"               "ESLT"
## [10] "EELCIGT"              "EHOOKAH"              "EROLLCIGTS"
## [13] "EPIPE"                "ESNUS"                "EDISSOLV"
## [16] "EBIDIS"               "CCIGT"                "CCIGAR"
## [19] "CSLT"                 "CELCIGT"              "CHOOKAH"
## [22] "CROLLCIGTS"           "CPIPE"                "CSNUS"
## [25] "CDISSOLV"             "CBIDIS"               "menthol"
## [28] "clove_spice"          "fruit"                "chocolate"
## [31] "alcoholic_drink"      "candy_dessert_sweets" "other"
##
## $nyts2019
##  [1] "psu"                  "finwgt"               "stratum"
##  [4] "Age"                  "Sex"                  "Grade"
##  [7] "ECIGT"                "ECIGAR"               "ESLT"
## [10] "EELCIGT"              "EHOOKAH"              "EROLLCIGTS"
## [13] "EPIPE"                "ESNUS"                "EDISSOLV"
## [16] "EBIDIS"               "EHTP"                 "CCIGT"
## [19] "CCIGAR"               "CSLT"                 "CELCIGT"
## [22] "CHOOKAH"              "CROLLCIGTS"           "CPIPE"
## [25] "CSNUS"                "CDISSOLV"             "CBIDIS"
## [28] "CHTP"                 "brand_ecig"           "menthol"
## [31] "clove_spice"          "fruit"                "chocolate"
## [34] "alcoholic_drink"      "candy_dessert_sweets" "other"
```

We create a function, update_values, that changes all Age and Sex values to more appropriate ones based on the data dictionary. We apply update_values to our data.

```
update_values <- function(dataset){
  dataset |>
    # adds 8 to each Age value
    mutate(Age = as.numeric(Age) + 8,
           # adds 5 to each Grade value
           Grade = as.numeric(Grade) + 5) |>
    # make Age, Grade, and Sex factors
    mutate(Age = as.factor(Age),
           Grade = as.factor(Grade),
           Sex = as.factor(Sex)) |>
    # change Sex values from 1 to male, 2 to female
    mutate(Sex = recode(Sex,
                        `1` = "male",
                        `2` = "female")) |>
    mutate_all(~ replace(., . %in% c("*", "**"), NA)) |>
    # change 19 to >18
    mutate(Age = recode(Age, `19` = ">18"),
           # change 13 to Ungraded/Other
           Grade = recode(Grade,
                          `13` = "Ungraded/Other")) |>
    # for all columns that start with E or C, if it's 1, change to TRUE, if it's 2, change to FALSE, if it has mi
ssing information, leave as NA
    mutate_at(vars(starts_with("E", ignore.case = FALSE),
                   starts_with("C", ignore.case = FALSE)
    ), list( ~ recode(., `1` = TRUE,
                      `2` = FALSE,
                      .default = NA,
                      .missing = NA)))
}

# apply the function to the data
nyts_data <- map(nyts_data, update_values)
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

We create a function, count_sex, to confirm that we get the same male count as the codebook in our data.

```
# filter Sex by males and count number of observations there are
count_sex <- function(dataset){dataset |>
    filter(Sex=='male') |>
    count(Sex) |>
    pull(n)}

# apply the count_sex function
map(nyts_data, count_sex)
```

```
## $nyts2015
## [1] 8958
##
## $nyts2016
## [1] 10438
##
## $nyts2017
## [1] 8881
##
## $nyts2018
## [1] 10069
##
## $nyts2019
## [1] 9803
```

We create a function, update_flavors, that changes the values of all columns between menthol and other to appropriate values.

```
update_flavors <- function(dataset){
  dataset |>
    # change values of all columns between menthol and other,
    mutate_at(vars(menthol:other),
              # if 1, chane to TRUE, everything else to FALSE
              list(~ recode(.,
                            `1` = TRUE,
                            .default = FALSE,
                            .missing = FALSE))) }
# apply to all years exclusing 2015
nyts_data  <- nyts_data  |>
  map_at(vars(-nyts2015), update_flavors)
```

We modify the 2019 data further so the brand codes match the actual brand.

```
nyts_data[["nyts2019"]] <- nyts_data[["nyts2019"]] |>
  mutate_all(~ replace(., . %in% c(".N", ".S", ".Z"), NA)) |>  # change all values with .N,.S,.Z to NA
  mutate(psu = as.character(psu)) |>                           # turn psu to characters
  mutate(brand_ecig = recode(brand_ecig,                       # change brand codes to actual brand
                             `1` = "Other",                    # levels 1,8 combined to `Other`
                             `2` = "Blu",
                             `3` = "JUUL",
                             `4` = "Logic",
                             `5` = "MarkTen",
                             `6` = "NJOY",
                             `7` = "Vuse",
                             `8` = "Other"))
```

We then combine the data and make a few modifications to the year column so it’s easy to operate on.

```
# combines everything into a single tibble
nyts_data <- nyts_data |>
  # creates an id column called year, removes the nyts from year column so it only shows year
  map_df(bind_rows, .id = "year") |>
  # make the year a numeric
  mutate(year = as.numeric(str_remove(year, "nyts"))) # clean-up year column

glimpse(nyts_data)
```

```
## Rows: 95,465
## Columns: 40
## $ year                 <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2…
## $ psu                  <chr> "015438", "015438", "015438", "015438", "015438",…
## $ finwgt               <dbl> 216.7268, 324.9620, 324.9620, 397.1552, 264.8745,…
## $ stratum              <chr> "BR3", "BR3", "BR3", "BR3", "BR3", "BR3", "BR3", …
## $ Age                  <fct> 18, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 1…
## $ Sex                  <fct> female, male, male, male, female, female, male, f…
## $ Grade                <fct> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
## $ ECIGT                <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE…
## $ ECIGAR               <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FAL…
## $ ESLT                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, T…
## $ EELCIGT              <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE…
## $ EROLLCIGTS           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, F…
## $ EFLAVCIGTS           <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, FA…
## $ EBIDIS               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ EFLAVCIGAR           <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ EHOOKAH              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ EPIPE                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ ESNUS                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, F…
## $ EDISSOLV             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CCIGT                <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, F…
## $ CCIGAR               <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, F…
## $ CSLT                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CELCIGT              <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, F…
## $ CROLLCIGTS           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CFLAVCIGTS           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CBIDIS               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CHOOKAH              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CPIPE                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CSNUS                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CDISSOLV             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ menthol              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ clove_spice          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ fruit                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ chocolate            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ alcoholic_drink      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ candy_dessert_sweets <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ other                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ EHTP                 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ CHTP                 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ brand_ecig           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
```

We create columns that contain the number of ways an individual has used tobacco ever and currently. Then, more columns are added containing whether the individual used tobacco at all or never.

```
nyts_data <- nyts_data %>%
  # Create new column by summing across "E" columns by adding up the TRUEs for each individual
  mutate(tobacco_sum_ever = rowSums(select(., starts_with("E",
                                    ignore.case = FALSE)), na.rm = TRUE),
         # Create new column by summing across "C" columns by adding up the TRUEs for each individual
         tobacco_sum_current = rowSums(select(., starts_with("C",
                                    ignore.case = FALSE)), na.rm = TRUE))  |>
  # Create new boolean columns of our sum columns by setting sums greater than 0 to TRUE and 0 to FALSE
  mutate(tobacco_ever = case_when(tobacco_sum_ever > 0 ~ TRUE,
                                  # If they never smoked, the value in tobacco_sum_ever is FALSE.
                                  tobacco_sum_ever == 0 ~ FALSE),
         # Repeat process for current smokers.
         tobacco_current = case_when(tobacco_sum_current > 0 ~ TRUE,
                                     tobacco_sum_current == 0 ~ FALSE))
```

Here we create new columns that contain the sum of the number of ways each individual has ever and currently use an e-cigarette and non e-cigarette. Then, we create boolean columns that contain whether or not they have ever or currently use an e-cigarette.

```
nyts_data <- nyts_data %>%
  # Creating new columns
  # if individual ever used e-cig add 1
  mutate(ecig_sum_ever = rowSums(select(., EELCIGT), na.rm = TRUE),
         # if individiual currently uses e-cig add 1
         ecig_sum_current = rowSums(select(., CELCIGT), na.rm = TRUE),
         # sums number of ways they ever use tobacco besides e-cig
         non_ecig_sum_ever = rowSums(select(., starts_with("E",  ignore.case = FALSE),
                                            -EELCIGT), na.rm = TRUE),
         # sums number of ways they currently use tobacco besides e-cig
         non_ecig_sum_current = rowSums(select(., starts_with("C", ignore.case = FALSE),
                                               -CELCIGT), na.rm = TRUE)) |>
  # Create new boolean columns of our sum columns by setting sums greater than 0 to TRUE and 0 to FALSE
  mutate(ecig_ever = case_when(ecig_sum_ever > 0 ~ TRUE,
                              ecig_sum_ever == 0 ~ FALSE),
         ecig_current = case_when(ecig_sum_current > 0 ~ TRUE,
                                  ecig_sum_current == 0 ~ FALSE),
         non_ecig_ever = case_when(non_ecig_sum_ever > 0 ~ TRUE,
                                   non_ecig_sum_ever == 0 ~ FALSE),
         non_ecig_current = case_when(non_ecig_sum_current > 0 ~ TRUE,
                                      non_ecig_sum_current == 0 ~ FALSE))
```

This gets the code into a usable format by groupings of how the individual uses tobacco.

```
nyts_data <- nyts_data |>
             # Create column ecig_only_ever with value TRUE only if they ever used an e-cig, but is not a current
user.
             mutate(ecig_only_ever = case_when(ecig_ever == TRUE &
                                           non_ecig_ever == FALSE &
                                            ecig_current == FALSE &
                                        non_ecig_current == FALSE ~ TRUE,
                                                             TRUE ~ FALSE),
              # Create column ecig_only_current with TRUE only if they are a current e-cig user
              ecig_only_current = case_when(ecig_current == TRUE &
                                           non_ecig_ever == FALSE &
                                        non_ecig_current == FALSE ~ TRUE,
                                                            TRUE ~ FALSE),
            # Create column non_ecig_only_ever with TRUE if they have never used an e-cig
            non_ecig_only_ever = case_when(non_ecig_ever == TRUE &
                                               ecig_ever == FALSE &
                                            ecig_current == FALSE &
                                        non_ecig_current == FALSE ~ TRUE,
                                                            TRUE ~ FALSE),
      # Create column non_ecig_only_current with TRUE if they currently use tobacco but not with an e-cig
      non_ecig_only_current = case_when(non_ecig_current == TRUE &
                                               ecig_ever == FALSE &
                                            ecig_current == FALSE ~ TRUE,
                                                            TRUE ~ FALSE),
                        # Create column no_use with TRUE if they have never used tobacco
                        no_use = case_when(non_ecig_ever == FALSE &
                                               ecig_ever == FALSE &
                                            ecig_current == FALSE &
                                        non_ecig_current == FALSE ~ TRUE,
                                                            TRUE ~ FALSE)) %>%
                  # Create Group column to label how they use tobacco.
                 mutate(Group = case_when(ecig_only_ever == TRUE |
                                       ecig_only_current == TRUE ~ "Only e-cigarettes",
                                      non_ecig_only_ever == TRUE |
                                   non_ecig_only_current == TRUE ~ "Only other products",
                                                  no_use == TRUE ~ "Neither",
                                          ecig_only_ever == FALSE &
                                       ecig_only_current == FALSE &
                                      non_ecig_only_ever == FALSE &
                                   non_ecig_only_current == FALSE &
                                                  no_use == FALSE ~ "Combination of products"))
```

This adds a count column “n” that tells you the number of individuals surveyed for each year.

```
nyts_data <- nyts_data |>
  # Groups by the year variable and counts the number of rows in that group.
  add_count(year)

glimpse(nyts_data)
```

```
## Rows: 95,465
## Columns: 59
## $ year                  <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, …
## $ psu                   <chr> "015438", "015438", "015438", "015438", "015438"…
## $ finwgt                <dbl> 216.7268, 324.9620, 324.9620, 397.1552, 264.8745…
## $ stratum               <chr> "BR3", "BR3", "BR3", "BR3", "BR3", "BR3", "BR3",…
## $ Age                   <fct> 18, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, …
## $ Sex                   <fct> female, male, male, male, female, female, male, …
## $ Grade                 <fct> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, …
## $ ECIGT                 <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRU…
## $ ECIGAR                <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FA…
## $ ESLT                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, …
## $ EELCIGT               <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRU…
## $ EROLLCIGTS            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, …
## $ EFLAVCIGTS            <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, F…
## $ EBIDIS                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ EFLAVCIGAR            <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, F…
## $ EHOOKAH               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ EPIPE                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ ESNUS                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, …
## $ EDISSOLV              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CCIGT                 <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CCIGAR                <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ CSLT                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CELCIGT               <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, …
## $ CROLLCIGTS            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CFLAVCIGTS            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CBIDIS                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CHOOKAH               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CPIPE                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CSNUS                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ CDISSOLV              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ menthol               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ clove_spice           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ fruit                 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ chocolate             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ alcoholic_drink       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ candy_dessert_sweets  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ other                 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ EHTP                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ CHTP                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ brand_ecig            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
## $ tobacco_sum_ever      <dbl> 1, 4, 0, 3, 0, 2, 8, 4, 0, 0, 0, 1, 1, 0, 0, 4, …
## $ tobacco_sum_current   <dbl> 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ tobacco_ever          <lgl> TRUE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE…
## $ tobacco_current       <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, F…
## $ ecig_sum_ever         <dbl> 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, …
## $ ecig_sum_current      <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ non_ecig_sum_ever     <dbl> 1, 3, 0, 2, 0, 1, 7, 3, 0, 0, 0, 0, 1, 0, 0, 3, …
## $ non_ecig_sum_current  <dbl> 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ ecig_ever             <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRU…
## $ ecig_current          <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, …
## $ non_ecig_ever         <lgl> TRUE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE…
## $ non_ecig_current      <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ ecig_only_ever        <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ ecig_only_current     <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ non_ecig_only_ever    <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …
## $ non_ecig_only_current <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,…
## $ no_use                <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, FALSE, FALSE, F…
## $ Group                 <chr> "Only other products", "Combination of products"…
## $ n                     <int> 17711, 17711, 17711, 17711, 17711, 17711, 17711,…
```

Here we save the wrangled data into a file in the wrangled folder to be used for later.

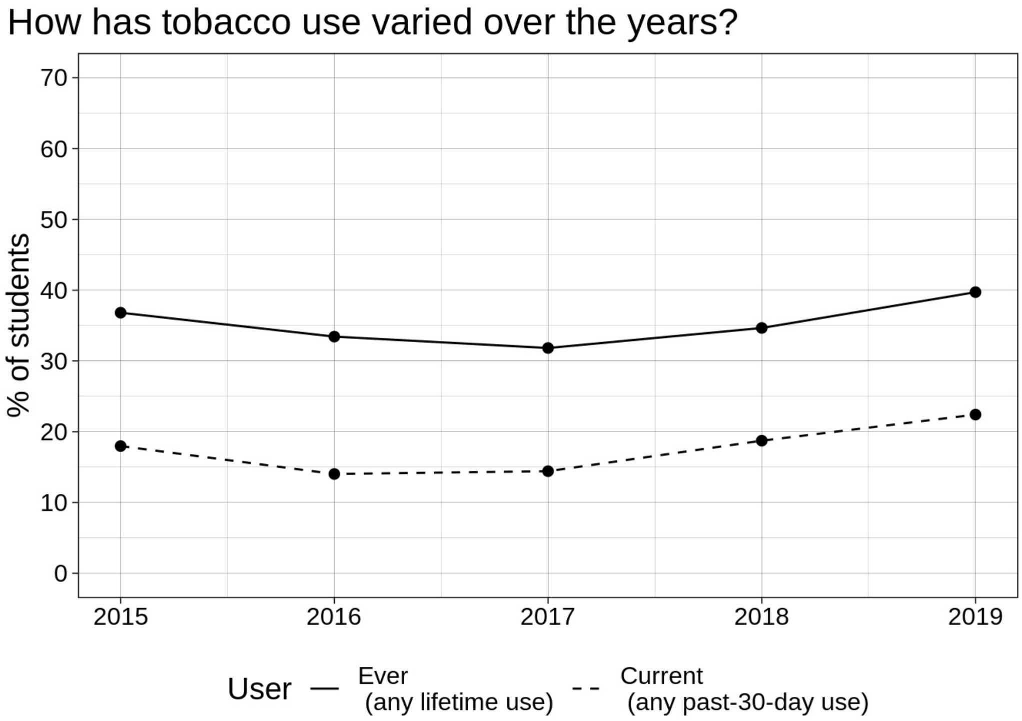
```
save(nyts_data, file="data/wrangled/wrangled_data_vaping.rda")
```

Analysis

Exploratory Data Analysis

Q1. How has tobacco and e-cigarette/vaping use by American youths changed since 2015?

```
nyts_data |>
  group_by(year) |>
  summarize("Ever \n (any lifetime use)" = (mean(tobacco_ever, na.rm = TRUE) * 100),
            "Current \n (any past-30-day use)" = (mean(tobacco_current, na.rm = TRUE) * 100)) |>
  pivot_longer(cols = -year, names_to = "User", values_to = "Percentage of students") |>
  ggplot(aes(x = year, y = `Percentage of students`)) +
  geom_line(aes(linetype = User)) +
  geom_point(show.legend = FALSE, size = 2) +
  # this allows us to choose what type of line we want for each line
  scale_linetype_manual(values = c(1, 2),
                        breaks = c("Ever \n (any lifetime use)",
                                   "Current \n (any past-30-day use)")) +
  # this allows us to specify how the y-axis should appear
  scale_y_continuous(breaks = seq(0, 70, by = 10),
                     labels = seq(0, 70, by = 10),
                     limits = c(0, 70)) +
  # this adjusts the background style of the plot
  theme_linedraw() +
  labs(title = "How has tobacco use varied over the years?",
       y = "% of students") +
  # this moves the legend to the bottom of the plot and removes the x axis title
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```

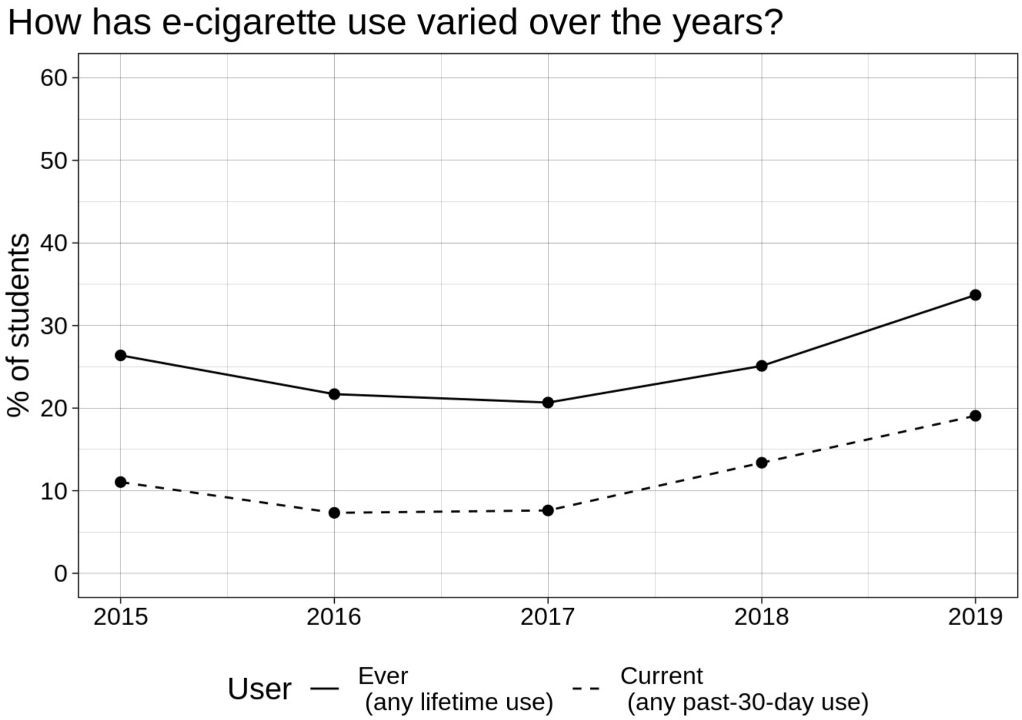


The above graph shows tobacco use from 2015 to 2019 for lifetime and any past-30 day use. Examining the lifetime use, we see that from 2015 to 2017, there’s a decline. However, from 2017 onwards, the use of tobacco increases steadily and by 2019, the use of tobacco has surpassed the percentage of 2015.

Similarly, students who have used tobacco in the last 30 days from 2015 to 2016 shows a decreasing trend, but 2016 to 2017 seems to be about the same. From 2017 to 2019, we see an increase in percentage, where 2018 and 2019 surpasses the 2015 percentage.

Despite the initial decline in earlier years, tobacco use overall among students has increased over the years for lifetime and past-30 day users.

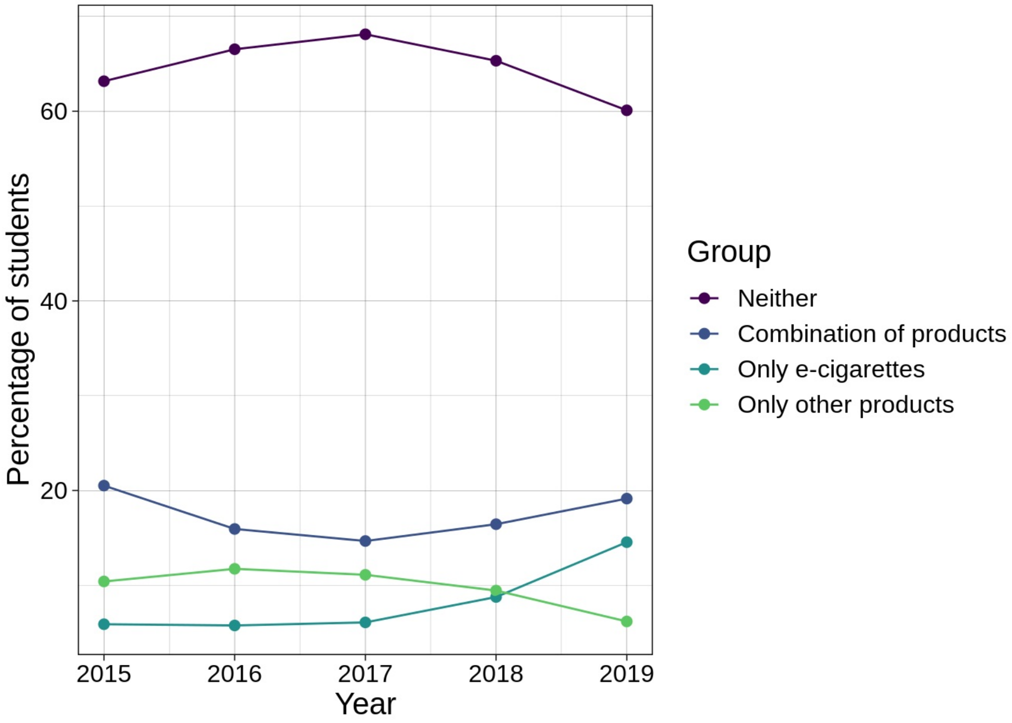
```
nyts_data |>
  group_by(year) |>
  summarize("Ever \n (any lifetime use)" = (mean(ecig_ever, na.rm = TRUE) * 100),
            "Current \n (any past-30-day use)" = (mean(ecig_current, na.rm = TRUE) * 100)) |>
  pivot_longer(cols = -year, names_to = "User", values_to = "Percentage of students") |>
  ggplot(aes(x = year, y = `Percentage of students`)) +
  geom_line(aes(linetype = User)) +
  geom_point(show.legend = FALSE, size = 2) +
  # this allows us to choose what type of line we want for each line
  scale_linetype_manual(values = c(1, 2),
                        breaks = c("Ever \n (any lifetime use)",
                                   "Current \n (any past-30-day use)")) +
  # this allows us to specify how the y-axis should appear
  scale_y_continuous(breaks = seq(0, 60, by = 10),
                     labels = seq(0, 60, by = 10),
                     limits = c(0, 60)) +
  # this adjusts the background style of the plot
  theme_linedraw() +
  labs(title = "How has e-cigarette use varied over the years?",
       y = "% of students") +
  # this moves the legend to the bottom of the plot and removes the x axis title
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```



Above graph shows the same graph but this time, this is use of e-cigarette among students in their lifetime/any past 30-day use. We see the same trend as the previous graph showed, except that the general percentages of both past 30-day use/lifetime of e-cigarettes have lower percentages than the tobacco.

```
v_colors =  viridis(5)[1:4]  #specify color palatte
nyts_data |>
  group_by(Group, year, n) |>
  summarize(group_count = n()) |>
  mutate("Percentage of students" = group_count / n * 100) |>
  ggplot(aes(x = year, y = `Percentage of students`, color = Group)) +
  geom_point(size = 2) +
  geom_line() +
  scale_color_manual(breaks = c("Neither", "Combination of products",
                                "Only e-cigarettes", "Only other products"),
                     values = v_colors) +
  theme_linedraw() +
  labs(x = "Year") +
  theme(text = element_text(size = 15),
        plot.title.position = "plot")
```

```
## `summarise()` has grouped output by 'Group', 'year'. You can override using the `.groups` argument.
```



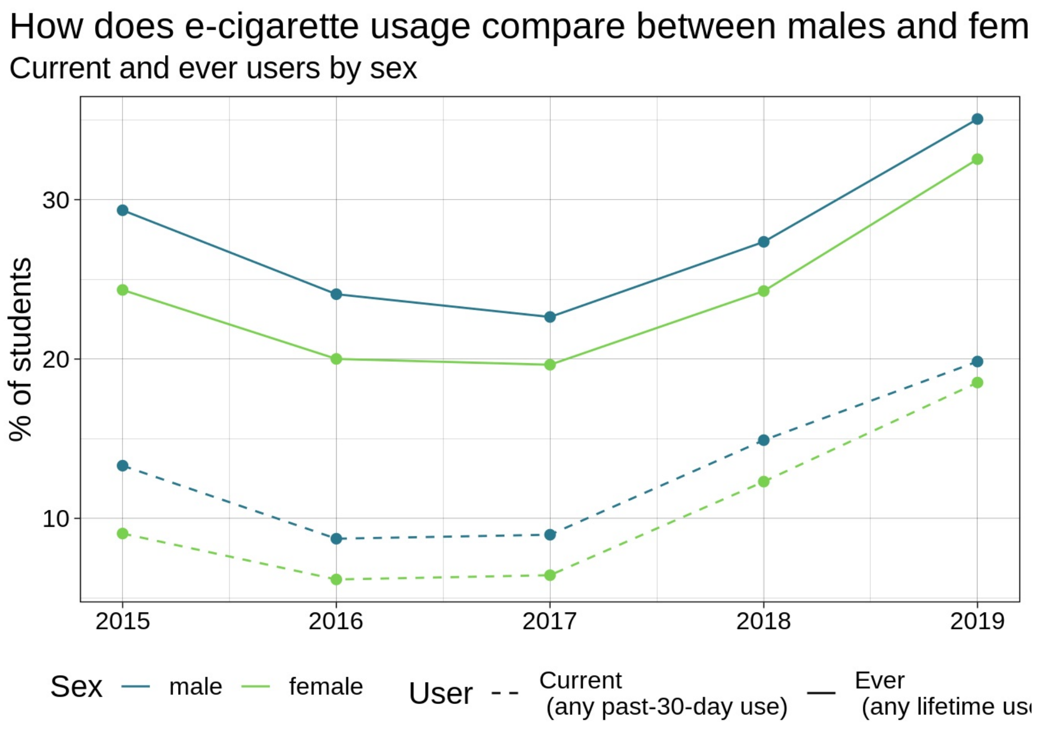
The above graph shows a percent of students (lifetime) in each year who have smoked before (only e-cigarette, tobacco, or both) or not. As we can see from the graph above, *Neither* group shows an opposite trend as the previous graph has shown, since *Neither* group trend should be opposite concept of use of tobacco group. Also, we can see that the *Only e-cigarette* group shows the most increased trend over the years. From this graph, we can argue that the *Combination of products (both tobacco and e-cigarette)* group is affected by the *Only e-cigarette* group since *Only other products* group do show decreasing trend but overall, we see an increasing trend from *Combination of products* group.

Given all plots above, we can argue that the use of e-cigarette/tobacco use by American youths changed since 2015 have increased over the years. Specifically, from 2017 to 2019, the use of e-cigarette/tobacco has gradually increased.

Q2. How does e-cigarette use compare between males and females?

```
v_colors =  viridis(6)[c(3, 5)]
nyts_data |>
  filter(!is.na(Sex)) |>
  group_by(year, Sex) |>
  summarize("Ever \n (any lifetime use)" = (mean(EELCIGT, na.rm = TRUE) * 100),
            "Current \n (any past-30-day use)" = (mean(CELCIGT, na.rm = TRUE) * 100)) |>
  pivot_longer(cols = "Ever \n (any lifetime use)":"Current \n (any past-30-day use)",
               names_to = "User",
               values_to = "Percentage of students") |>
  ggplot(aes(x = year, y = `Percentage of students`, color = Sex)) +
  geom_line(aes(linetype = User)) +
  geom_point(show.legend = FALSE, size = 2) +
  scale_linetype_manual(values = c(2, 1)) +
  scale_color_manual(values = v_colors) +
  theme_linedraw() +
  labs(title = "How does e-cigarette usage compare between males and females?",
       subtitle = "Current and ever users by sex",
       y = "% of students") +
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

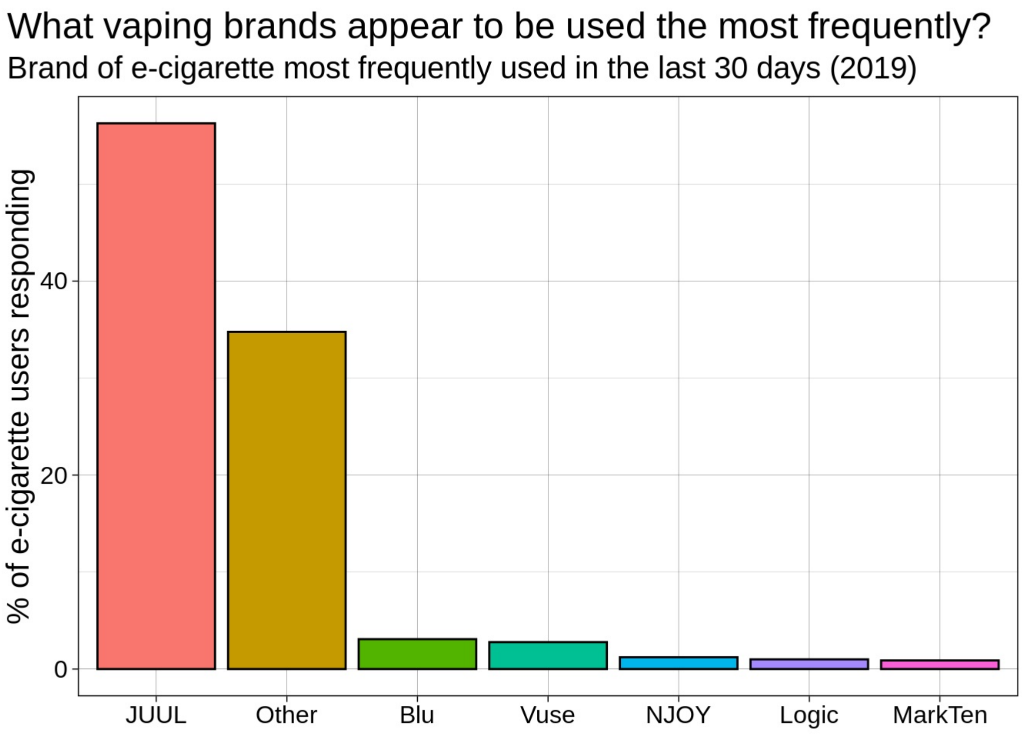


The graph above shows the percent of students who used e-cigarette in the past 30 days and lifetime divided by sex. We can say that generally, we see a similar trend that from 2015 to 2017, the use of e-cigarettes has declined, but from 2017 to 2019, the use of e-cigarettes has increased more than what it has declined before. In addition, males tend to use e-cigarette (whether past 30-day use or lifetime use) more than female, by around 5 percent more.

Using the graph above, we can conclude that the use of e-cigarettes between male and females show the same trend from both sex. However, males overall tend to use more e-cigarettes than females.

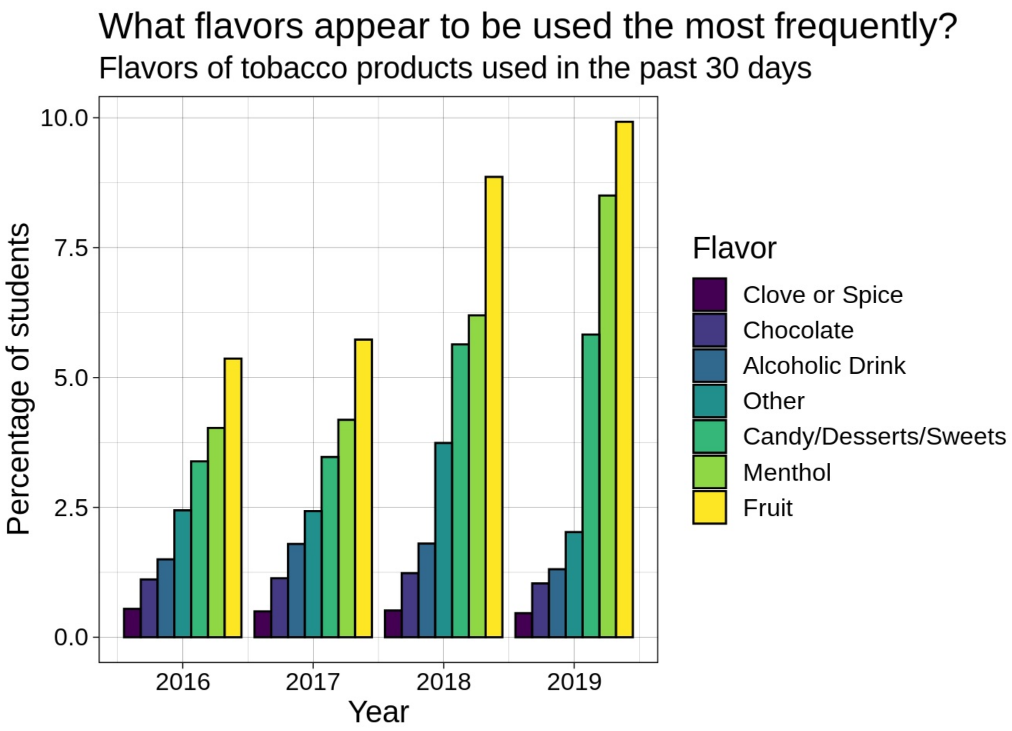
Q3. What vaping brands and flavors appear to be used the most frequently?

```
nyts_data |>
  filter(year == 2019) |>
  group_by(brand_ecig) |>
  filter(!is.na(brand_ecig)) |>
  summarize(n = n()) |>
  mutate(total = sum(n),
         Percent = n * 100 / total) |>
  mutate(brand_ecig = fct_reorder(brand_ecig, desc(Percent))) |>
  ggplot(aes(x = brand_ecig, y = Percent, fill = brand_ecig)) +
  geom_bar(stat = "identity", color = "black") +
  theme_linedraw() +
  labs(title = "What vaping brands appear to be used the most frequently?",
       subtitle = "Brand of e-cigarette most frequently used in the last 30 days (2019)",
       y = "% of e-cigarette users responding") +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```



The graph above shows the percentage of users using a brand of e-cigarette. We can see that *JUUL* is the major brand that people who smoke e-cigarettes use, with a frequency of around 50%. The second highest, *Other*, has a frequency of around 35%. The rest of the brands are used significantly less, with their frequency being under 5%.

```
nyts_data |>
  filter(year != 2015) |>
  group_by(year) |>
  summarize(Menthol = (mean(menthol) * 100),
            `Clove or Spice` = (mean(clove_spice) * 100),
            Fruit = (mean(fruit) * 100),
            Chocolate = (mean(chocolate) * 100),
            `Alcoholic Drink` = (mean(alcoholic_drink) * 100),
            `Candy/Desserts/Sweets` = (mean(candy_dessert_sweets) * 100),
            Other = (mean(other) * 100)) |>
  pivot_longer(cols = -year,
               names_to = "Flavor",
               values_to = "Percentage of students") |>
  rename(Year = year) |>
  ggplot(aes(y = `Percentage of students`,
             x = Year,
             fill = reorder(Flavor, `Percentage of students`))) +
  geom_bar(stat = "identity",
           position = "dodge",
           color = "black") +
  scale_fill_viridis(discrete = TRUE) +
  theme_linedraw() +
  guides(fill = guide_legend("Flavor")) +
  labs(title = "What flavors appear to be used the most frequently?",
       subtitle = "Flavors of tobacco products used in the past 30 days") + theme(text = element_text(size = 15))
```

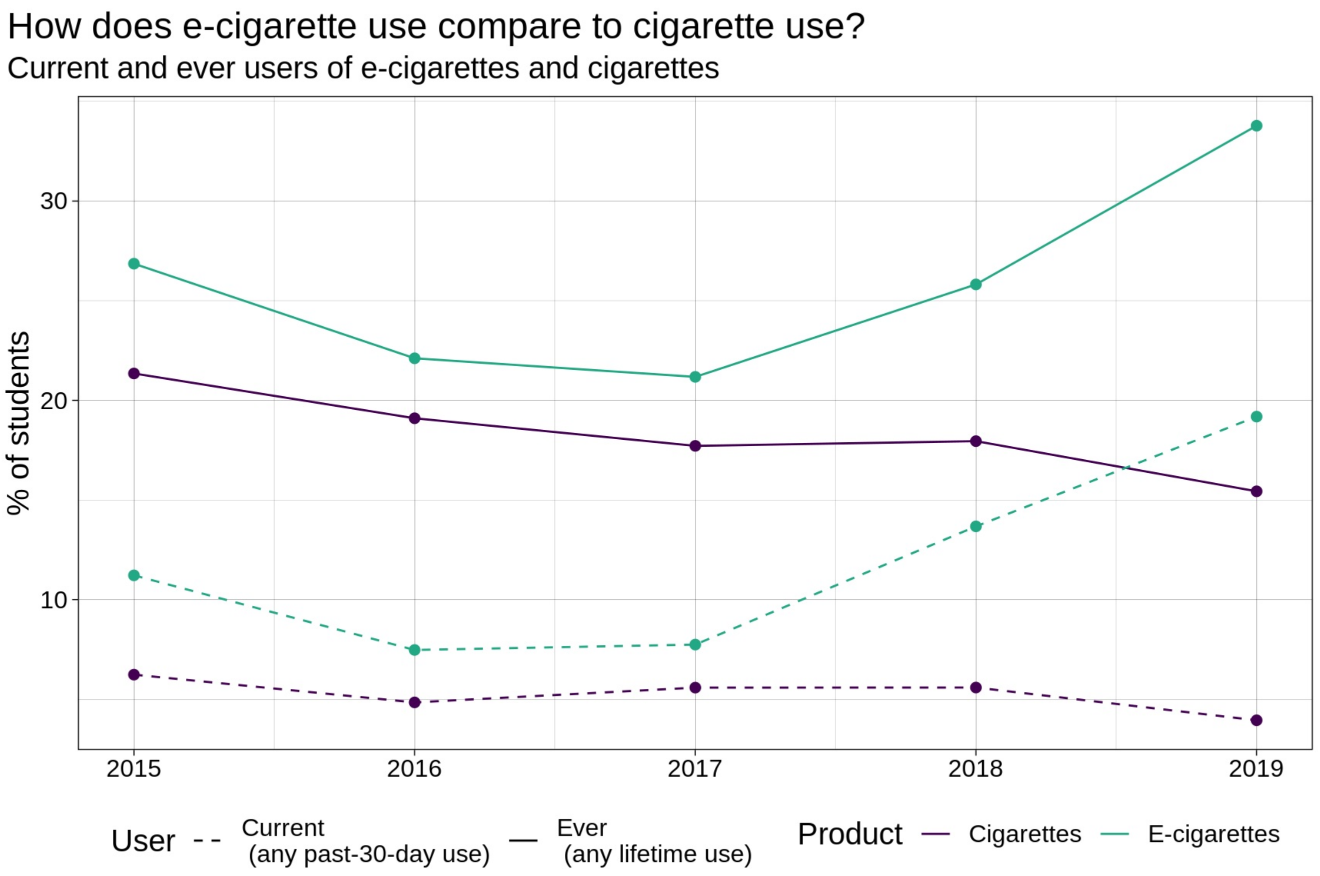


The above graph shows which flavors (Clove or Spice, Chocolate, Alcoholic Drink, Other, Candy/Desserts/Sweets, Menthol, and Fruit) students are using in their tobacco over the years. *Fruit*, having the highest percentage of students, increases over the years, with a dramatic increase from 2017-2018. This being the highest might be attributed to how universal fruity flavors are to different ages compared to flavors like *Alcoholic Drink* and *Clove or Spice*. *Menthol*, having the next highest percentage, also increases over the years and has a similar increase from 2017 to 2018. *Candy/Desserts/Sweets*, having the next highest percentage, follows a similar trend, increasing over the year and a dramatic increase from 2017 to 2018. *Other*, having the next highest percentage, also increases significantly in 2017 - 2018, but decreases from 2018-2019. *Alcoholic Drink*, having the next highest percentage, barely changes and increases from 2016-2018, but decreases from 2018-2019. *Chocolate* also follows this trend. The least percentage, *Clove and Spice*, seems to stay about the same throughout the years.

From those graphs above, we can say that the vaping brand that’s used the most is *JUUL*, and for flavored tobaccos, *Fruit* is used the most.

Q4. Is there a relationship between e-cigarette/Vaping use and other tobacco use?

```
v_colors =  viridis(6)[c(1, 4)]
nyts_data |>
  group_by(year) |>
  summarize(
    "Cigarettes, Ever \n (any lifetime use)" = (mean(ECIGT, na.rm = TRUE) * 100),
    "E-cigarettes, Ever \n (any lifetime use)" = (mean(EELCIGT, na.rm = TRUE) * 100),
    "Cigarettes, Current \n (any past-30-day use)" = (mean(CCIGT, na.rm = TRUE) * 100),
    "E-cigarettes, Current \n (any past-30-day use)" = (mean(CELCIGT, na.rm = TRUE) * 100)
  ) |>
  pivot_longer(cols = -year,
               names_to = "Category",
               values_to = "Percentage of students") |>
  separate(Category, into = c("Product", "User"), sep = ", ") |>
  ggplot(aes(
    x = year,
    y = `Percentage of students`,
    color = Product,
    linetype = User
  )) +
  geom_line() +
  geom_point(show.legend = FALSE, size = 2) +
  scale_linetype_manual(values = c(2, 1)) +
  scale_color_manual(values = v_colors) +
  theme_linedraw() +
  labs(title = "How does e-cigarette use compare to cigarette use?",
       subtitle = "Current and ever users of e-cigarettes and cigarettes",
       y = "% of students") +
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```



How does e-cigarette use compare to cigarette use?

Current and ever users of e-cigarettes and cigarettes

The above graph shows the comparison of e-cigarette to cigarette use by showing the point with line graph of percentage of students over the years. We can see that from 2015 to 2017, both cigarette and e-cigarette have declined slowly (except cigarette at any past 30-days use). However, from 2017 to 2019, we see that e-cigarette use has suddenly increased (both lifetime and past 30-day use), but cigarette use has declined. In fact, the usage of e-cigarettes for past-30-day use surpassed cigarette lifetime use in 2019.

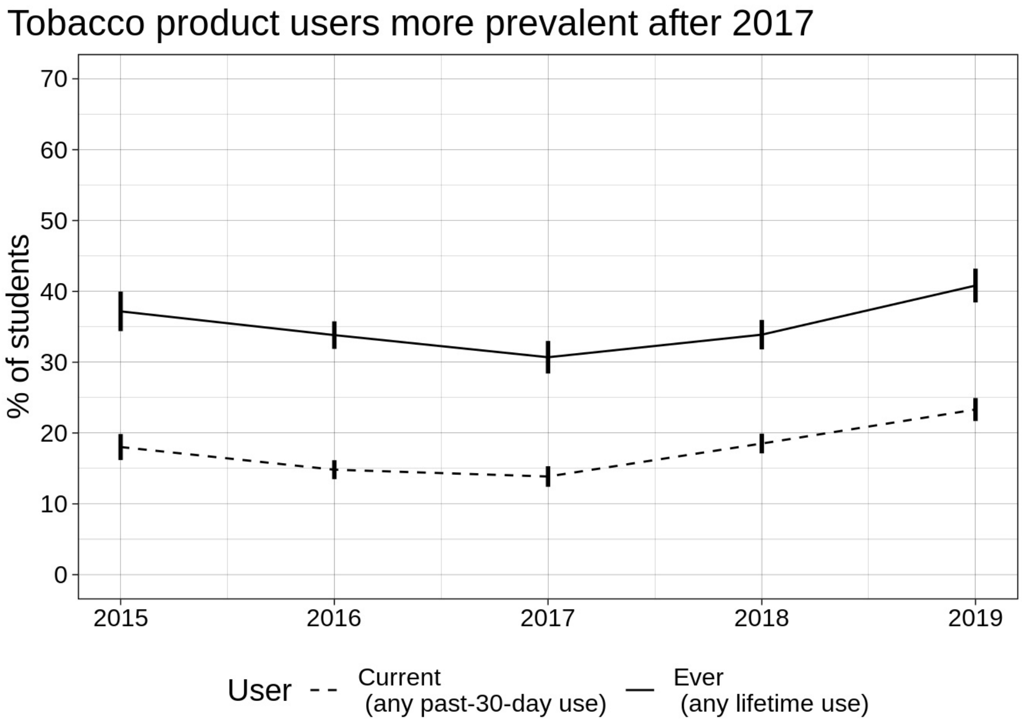
From the graph above, we can say that from 2015 to 2017, cigarettes and e-cigarettes had a direct relationship (they both decreased), and from 2017-2018, they had a inverse relationship (while e-cigarette use increased, cigarette use decreased). We hypothesize that starting from 2017 onwards, there were cultural changes that led to e-cigarettes having a more positive reputation than cigarettes, causing them to be used more than cigarettes.

Data Analysis

```
surveyMeanA <- function(currYear) {
  options(survey.lonely.psu = "adjust")
  currYear |>
    # create survey object, taking into account strata, weight, and psu
    as_survey_design(strata = stratum,
                     ids = psu,
                     weight  = finwgt,
                     nest = TRUE) |>
    # calculate confidence interval for tobacco_ever and tobacco_current
    summarize(tobacco_ever = survey_mean(tobacco_ever,
                                         vartype = "ci",
                                         na.rm = TRUE),
              tobacco_current = survey_mean(tobacco_current,
                                            vartype = "ci",
                                            na.rm = TRUE))  |>
    # convert into percentage
    mutate_all("*", 100) |>
    pivot_longer(everything(),
                 names_to = "Type",
                 values_to = "Percentage of students") |>
    mutate(Estimate = case_when(str_detect(Type, "_low") ~ "Lower",
                                str_detect(Type, "_upp") ~ "Upper",
                                TRUE ~ "Mean"),
           User = case_when(str_detect(Type, "ever") ~ "Ever",
                            str_detect(Type, "current") ~ "Current",
                            TRUE ~ "Mean"))}
```

The above code creates a function **survyMeanA**, where it creates a survey object (from srvyr) that contains strantum (categorical variable that indicates subsets of the data that include respondents from differnet PSU), psu (Primary Sampling unit), finwgt (Survey Weight), and with nest=True to force cluster ids to be nested within the strata. We use that to make sure every rows are on different weight when calculating mean of all rows, to ensure no swaying occurs in the dataset. We then use pivot_longer() and mutate() function to estimate the lower and upper bounds of estimates (for both current and ever users).

```
nyts_data |>
  group_by(year) |>
  # apply surveyMeanA function to each year
  group_modify(~ surveyMeanA(.x)) |>
  dplyr::select(-Type) |>
  pivot_wider(names_from = Estimate,
             values_from = `Percentage of students`) |>
  ggplot(aes(x = year, y = Mean)) +
  geom_line(aes(linetype = User)) +
  geom_linerange(aes(ymin = Lower,
                     ymax = Upper),
                     size = 1,
              show.legend = FALSE) +
  scale_linetype_manual(values = c(2, 1),
                        labels = c("Current \n (any past-30-day use)",
                                   "Ever \n (any lifetime use)")) +
  scale_y_continuous(breaks = seq(0, 70, by = 10),
                     labels = seq(0, 70, by = 10),
                     limits = c(0, 70)) +
    theme_linedraw() +
    labs(title = "Tobacco product users more prevalent after 2017",
         y = "% of students") +
    theme(legend.position = "bottom",
          axis.title.x = element_blank(),
          text = element_text(size = 15),
          plot.title.position = "plot")
```



The above graph shows the tobacco use when we group lifetime use and use of tobacco in last 30 days by using survey means method. This graph is different from the original graph that we have created earlier, because in each year, there is a line created which represents the estimates of lower and higher bounds of percent of students (95% confidence), based on survey means method calculation. We first updated estimates (middle of line) based on our survey weights and added the 95% confidence intervals. This indicates that we're 95% confident that the true value for the percent of students who have ever used tobacco for a certain year is between these lines. As we can see, tobacco products seem to be used more after 2017. Another important thing to note is the overlap between confidence intervals. For the users who have ever used tobacco, there's some overlap between consecutive years and overlap between 2015 and 2019, indicating the there isn't a drastic change for tobacco users. For the last 30-day tobacco users, there's overlap between 2015-2017, but from 2017-2019, there's no overlap so there's a much drastic change for 30-day tobacco users.

```
dat2015 <- nyts_data %>%
  filter(year == 2015, !is.na(Sex)) # Remove unnecessary rows like unknown sex
currEcigSex <- logistic_reg() |> # Using logistic regression
  set_engine("glm") |>
  fit(as.factor(ecig_current) ~ Sex, data = dat2015, family = "binomial") # Use ecig_current as logit(p), Sex as
X
(currEcigSexTidy <- tidy(currEcigSex))
```

```
## # A tibble: 2 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -1.89     0.0313    -60.4  0
## 2 Sexfemale     -0.425    0.0490     -8.66 4.73e-18
```

We are first only looking at 2015.

The tibble above shows the equation of logistic regression. We can write this as

log (odds of current e-cig uses) = (beta)0 + (beta)1 * Sex = -1.89 - 0.425 * (Sex == Female)

-0.425 is (beta)1 and is log(OR) (log of odds ratio). This means that the log odds of being a current e-cigarette user is 0.425 lower for females compared to males.

If we want it out of log scale, we could apply e to both sides. This results in e^(-0.425) = 0.65 = OR. This tells us that the odds of being a current e-cigarette user for female is 0.65 times lower than the odds for males. Since logistic regression can be used to estimate further happenings, we can also say that the chances of being a current e-cigarette user for female will likely to be 35% lower than the chances for males.

```
dat2015_survey_design <- dat2015 %>%
                          as_survey_design(strata = stratum,
                                            ids = psu,
                                            weight  = finwgt,
                                            nest = TRUE)
currEcigSex_svy <- survey::svyglm(ecig_current ~ Sex,
                                  family = quasibinomial(link = 'logit'),
                                  design = dat2015_survey_design)
tidy(currEcigSex_svy)
```

```
## # A tibble: 2 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -1.90     0.0788    -24.1  3.95e-34
## 2 Sexfemale     -0.383    0.0700     -5.48 7.52e- 7
```

The code above shows similar concept as the code above, where we create a logistic regression tibble for tobacco use for sex. However, for this logistic regression, we take survey weight into account.

The equation is the following:

-0.383 = (beta)1 = log(OR) - We can put exponential function to both side to make simpler equation: 0.68 = e^(-0.383) = e^(Beta1) = OR

We can interpret this equation as the log odds of being a current e-cigarette user is 0.383 lower for females than for males.

If we want it out of log scale, we apply e to both sides. This results in e^(-.383) = 0.68 = OR. This tells us that the odds of being a current e-cigarette user for females is 0.68 times and 32% lower than the odds for males.

Results

Here, we see a difference between logistic regression equation with simple dataset compared to dataset that was normalized with survey means method. For the original dataset, we see 0.425 lower chances of female smoking e-cigarette compared to males. On the other hand, our normalized with survey means methods show that we see 0.383 lower chances of female smoking e-cigarette compared to males. From here, we see a lower chance of females smoking e-cigarette from the dataset used with Survey means method. This tells us that we do have some rows that weigh more than it should be, and/or rows that weigh less than it should be. By using the survey means method, we are able to make our prediction more accurate.

Extended Analysis

Additional Question:

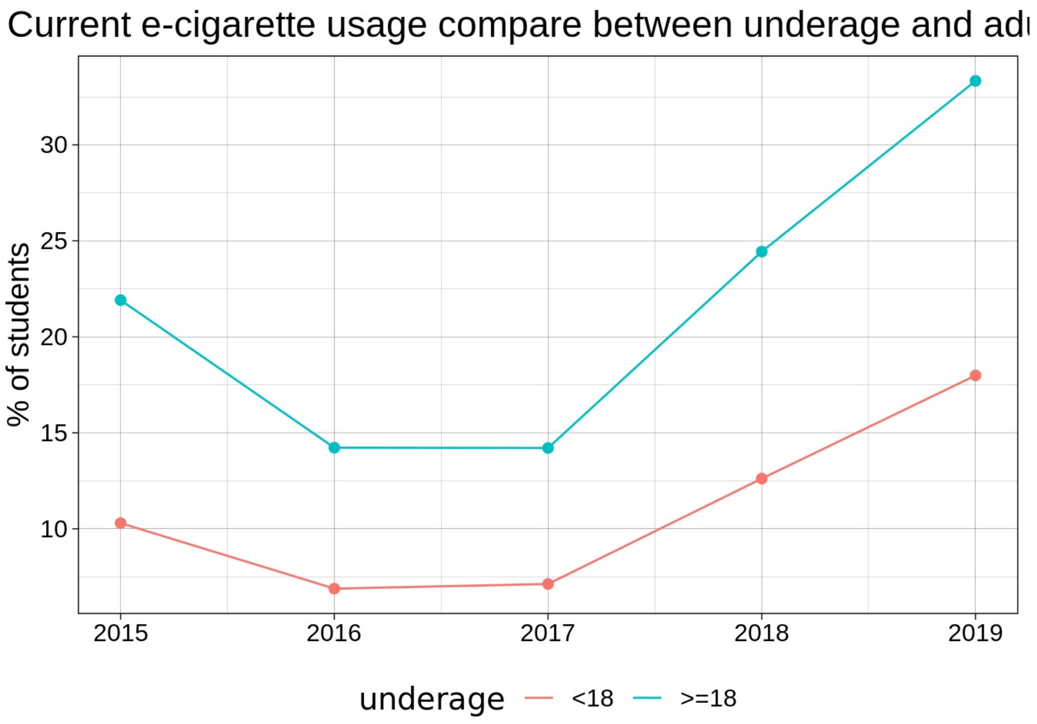
According to the Centers for Disease Control and Prevention (CDC), starting in 2014, vaping/e-cigarette became the most popular tobacco product among teenagers. Therefore, we are interested in how have the current vaping/e-cigarette users who are under the age of 18 changed from 2015 to 2019 compared to adult users.

```
df_underage <- nyts_data %>%
  add_column(underage = ifelse(.$Age == '>18', '>=18', # set underage as '<18' for Age is less than 18, set under
age as '>=18' for age is more than or equal to 18
                               ifelse(.$Age == 18, '>=18', '<18')))
```

The code above creates a new column that labels underage and adult data points.

```
v_colors =  viridis(6)[c(3, 5)]
df_underage |>
  filter(!is.na(underage)) |> # remove rows that don't have underage value
  group_by(year, underage) |>
  summarize("Current \n (any past-30-day use)" = (mean(CELCIGT, na.rm = TRUE) * 100)) |>
  pivot_longer(cols = "Current \n (any past-30-day use)",
               names_to = "User",
               values_to = "Percentage of students") |>
  ggplot(aes(x = year, y = `Percentage of students`, color = underage)) +
  geom_line() +
  geom_point(show.legend = FALSE, size = 2) +
  scale_linetype_manual(values = c(2, 1)) +
  theme_linedraw() +
  labs(title = "Current e-cigarette usage compare between underage and adults",
       y = "% of students") +
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot",
        legend.title = element_text('Age'))
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```



Above visualization shows the percentage of current e-cigarette usage between 2015 and 2019 by age. Both age groups have decreased in e-cigarette usage from 2015 to 2016. From 2016-2017, usage is pretty stagnant. From 2017 to 2019, there’s a drastic increase in usage. The percentage of underage users remains less than the percentage of adult users over the years.

```
# model fitting
currEcigAge <- logistic_reg() |> # Using logistic regression
  set_engine("glm") |>
  fit(as.factor(ecig_current) ~ underage, data = df_underage, family = "binomial") # Use ecig_current as logit(p)
, Underage as X
(currEcigAgeTidy <- tidy(currEcigAge))
```

```
## # A tibble: 2 × 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    -2.11     0.0109    -193.  0
## 2 underage>=18    0.792    0.0297      26.6 1.82e-156
```

Equation: log(p/(1-p)) = -2.1064647 + 0.7920089*(underage == ‘>=18’), for ‘<18’ = 0 and ‘>=18’=1

Based on above equation, we can calculate that the probability of being a current e-cigarette user for people whose age is less than 18 is 0.1084701, and the probability of being a current e-cigarette user for people whose age is larger than or equal to 18 is 0.2117422. By comparing the probabilities, we can say that the adult people have a larger likelihood of being a current e-cigarette user than underage people.

```
# Survey-weighted model
df_underage_survey_design <- df_underage %>%
                          as_survey_design(strata = stratum,
                                            ids = psu,
                                            weight  = finwgt,
                                            nest = TRUE)
currEcigAge_svy <- survey::svyglm(ecig_current ~ underage,
                                  family = quasibinomial(link = 'logit'),
                                  design = df_underage_survey_design)
tidy(currEcigAge_svy)
```

```
## # A tibble: 2 × 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    -2.07     0.0361     -57.4 1.34e-199
## 2 underage>=18    0.822    0.0457      18.0 6.50e- 54
```

Equation: log(p/(1-p)) = -2.072122 + 0.822048*(underage == '>=18'), for '<18' = 0 and '>=18'=1

Based on above equation, we can calculate that, by taking survey weights into consideration, the probability of being a current e-cigarette user for people whose age is less than 18 is 0.1118361, and the probability of being a current e-cigarette user for people whose age is larger than or equal to 18 is 0.2226873. Both probabilities with survey weights into taken to account has a small increase compare to the probabilities without taken it into account. We can say that the adult people still more likely to be a current e-cigarette user than underage people when we take survey weights into consideration.

Discussion of Extended Analysis

Since Vaping became more and more popular in the youth population over the years, we came into analyzing the change of current e-cigarette user between adults and underage people. Based on our analysis, we can see that while the trend of the percentage of underage e-cigarette users and adult users is similar, adult users have larger percentage than underage users from 2015 and 2019. By fitting the logistic model to the data with and without survey weights involved, we see that adult people tend to have a greater chance to be a current e-cigarette user than people who are under 18.

Conclusion

In this case study, we used NYTS (National Youth Tobacco Survey) dataset to compare and contrast use of tobacco/e-cigarette among age group, gender, and which flavors are most popular for certain groups of people. From wrangling and analyzing dataset, we found out that

1. Over the years, the percentage of tobacco/e-cigarette usage have drastically increased over the years, specifically from 2017 to 2019.
2. Generally, males tend to smoke e-cigarette more than females.
3. E-cigarette company **JUUL** is used the most, and flavor **Fruit** was the most popular flavor for tobacco.
4. From 2015 to 2017, there was a direct relationship between tobacco and e-cigarettes (both decreased). From 2017 to 2019, there was an inverse relationship (e-cigarette usage increased, tobacco usage decreased).

To answer our additional question, we provided more analysis on the current e-cigarette usage between underage people (<18) and adults (>=18). We found that over the years, adults seem to have larger percentage of current e-cigarette users than underage people.

References

(1) https://www.cdc.gov/mmwr/volumes/68/wr/mm6806e1.htm (https://www.cdc.gov/mmwr/volumes/68/wr/mm6806e1.htm)

(2) https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm (https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm)