CS02 - Vaping Behaviors in American Youth

Xiaoqing Wang, Aleksandar Bumbalov, Zikang Chen

Introduction

In this case study, we aim to examine the prevalence of tobacco and e-cigarette use among American youths from 2015 to 2019. Tobacco is a significant public health issue in the United States, and it is the leading cause of preventable death and disease. Most tobacco product use begins during youth and young adulthood, making it critical to understand how its use has evolved over time. Recent studies suggest that e-cigarettes are the most common tobacco product among teens, and their rise in popularity makes this an issue worth studying. Consequently, we will examine e-cigarette use by gender and the most commonly used vaping brands and flavors among American youths in hopes of gaining some helpful insight into how we can begin to structure preventative measures.

To further extend the study, we want to explore how age influences tobacco and e-cigarette use among American youths and to examine the correlation between e-cigarette use and other tobacco products. Understanding the current trends of use is necessary for developing strategies to reduce the danger that these products cause. Therefore, this case study seeks to provide essential insights into the present state of tobacco and e-cigarette use and its implications for public health especially among American youths.

Load packages

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

Questions

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. What is the relationship between e-cigarette and tobacco/non-ecigarette use and age?

The Data

The data we are using come from National Youth Tobacco Survey (NYTS), in which students from high school and middle school answered their tobacco usage in the United States of America. For our report, we are using data from 2015-2019. In the survey, they gather information including demographic, tobacco use, e-cigarette use, flavor of e-cig, and e-cig brand. For the e-cig flavor data, the survey did not ask for flavor in 2015, so we only have flavor data from 2016-2019. Similarly, we also only have limited data for brand use.

Data Import

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

Data Wrangling

We start by import the data.

```
# read in CSVs
nyts_data <- list.files("data/simpler_import/",
                        pattern = "*.csv",
                        full.names = TRUE) |>
  map(~ read_csv(.))
```

```
## Rows: 17711 Columns: 29── Column specification ────────────────────────────────────────
──────────────
## Delimiter: ","
## chr  (2): psu, stratum
## dbl (27): finwgt, Qn1, Qn2, Qn3, ECIGT, ECIGAR, ESLT, EELCIGT, EROLLCIGTS, E...
## 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.
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 20675 Columns: 34── Column specification ────────────────────────────────────────
──────────────
## Delimiter: ","
## chr  (4): psu, stratum, Q1, Q2
## dbl (30): finwgt, Q3, ECIGT, ECIGAR, ESLT, EELCIGT, EHOOKAH, EROLLCIGTS, EFL...
## 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.
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 17872 Columns: 33── Column specification ────────────────────────────────────────
──────────────
## Delimiter: ","
## chr  (3): psu, stratum, Q1
## dbl (30): finwgt, Q2, Q3, ECIGT, ECIGAR, ESLT, EELCIGT, EHOOKAH, EROLLCIGTS,...
## 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.Rows: 2
0189 Columns: 33── Column specification ────────────────────────────────────────────────
──────
## Delimiter: ","
## chr  (5): psu, stratum, Q1, Q2, Q3
## dbl (28): finwgt, ECIGT, ECIGAR, ESLT, EELCIGT, EHOOKAH, EROLLCIGTS, EPIPE, ...
## 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.Rows: 1
9018 Columns: 36── Column specification ────────────────────────────────────────────────
──────
## Delimiter: ","
## chr (34): stratum, Q1, Q2, Q3, ECIGT, ECIGAR, ESLT, EELCIGT, EHOOKAH, EROLLC...
## dbl  (2): psu, finwgt
## 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.
```

```
# get names
nyts_data_names <- list.files("data/simpler_import/",
                              pattern = "*.csv") |>
  str_extract("nyts201[5-9]")

# apply names
names(nyts_data) <- nyts_data_names
```

We need to recode the names and columns from the survey using their code book.

```
# rename 2015 data variables
nyts_data[["nyts2015"]] <- nyts_data[["nyts2015"]] |>
  rename(Age = Qn1,
         Sex = Qn2,
         Grade = Qn3)

# function to rename variables in 2016-2018
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)
}

nyts_data <- nyts_data |>
  map_at(c("nyts2016", "nyts2017", "nyts2018"), update_survey)


# rename 2019 data
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) |>
  mutate_all(~ replace(., . %in% c(".N", ".S", ".Z", ".M", "M"), NA)) |>
  mutate_at(vars(starts_with("E", ignore.case = FALSE),
                 starts_with("C", ignore.case = FALSE),
                 menthol:other),
            list( ~ as.numeric(.)))
```

Next, we clean the values in each variable.

```r
# function to wrangle values
update_values <- function(dataset){
  dataset |>
    mutate_all(~ replace(., . %in% c("*", "**"), NA)) |>
    mutate(Age = as.numeric(Age) + 8,
           Grade = as.numeric(Grade) + 5) |>
    mutate(Age = as.factor(Age),
           Grade = as.factor(Grade),
           Sex = as.factor(Sex)) |>
    mutate(Sex = case_match(Sex,
                            "1" ~ "male",
                            "2" ~ "female")) |>
    mutate_all(~ replace(., . %in% c("*", "**"), NA)) |>
    mutate(Age = case_match(Age, "19" ~ ">18",
                            .default = Age),
           Grade = case_match(Grade,
                              "13" ~ "Ungraded/Other",
                              .default = Grade)) |>
    mutate_at(vars(starts_with("E", ignore.case = FALSE),
                   starts_with("C", ignore.case = FALSE)
    ), list( ~ case_match(., 1 ~ TRUE,
                             2  ~ FALSE,
                          .default = NA)))
}

nyts_data <- map(nyts_data, update_values)

# function to count how many males
count_sex <- function(dataset){dataset |>
    filter(Sex=='male') |>
    count(Sex) |>
    pull(n)}

# clean 2019 values
nyts_data[["nyts2019"]] <- nyts_data[["nyts2019"]]  |>
  mutate(psu = as.character(psu)) |>
  mutate(brand_ecig = case_match(brand_ecig,
                              "1" ~ "Other", # levels 1,8 combined to `Other`
                              "2" ~ "Blu",
                              "3" ~ "JUUL",
                              "4" ~ "Logic",
                              "5" ~ "MarkTen",
                              "6" ~ "NJOY",
                              "7" ~ "Vuse",
                              "8" ~ "Other"))
```

We clean up flavor data by creating a function.

```
# function to wrangle flavor data
update_flavors <- function(dataset){
  dataset |>
    mutate_at(vars(menthol:other),
              list(~ case_match(.,
                                1 ~ TRUE,
                                NA ~ FALSE))) }

nyts_data  <- nyts_data  |>
  map_at(vars(-nyts2015), update_flavors)
```

```
## Warning: Using `vars()` in .at was deprecated in purrr 1.0.0.
```

Here, we combine the data together matched by year.

```
# combine data
nyts_data <- nyts_data |>
  map_df(bind_rows, .id = "year") |>
  mutate(year = as.numeric(str_remove(year, "nyts")))
```

We clean up data for tobacco use and e-cigarette into new columns broken down by ever use and current use for future analysis.

```
# wrangle tobacco columns
nyts_data <- nyts_data %>% # use first letter from survey to recode
  mutate(tobacco_sum_ever = rowSums(select(., starts_with("E",
                                    ignore.case = FALSE)), na.rm = TRUE),
         tobacco_sum_current = rowSums(select(., starts_with("C",
                                    ignore.case = FALSE)), na.rm = TRUE))  |>
  mutate(tobacco_ever = case_when(tobacco_sum_ever > 0 ~ TRUE,
                                  tobacco_sum_ever == 0 ~ FALSE),
         tobacco_current = case_when(tobacco_sum_current > 0 ~ TRUE,
                                     tobacco_sum_current == 0 ~ FALSE))

# wrangle e-cigarette columns
nyts_data <- nyts_data %>%
  mutate(ecig_sum_ever = rowSums(select(., EELCIGT), na.rm = TRUE),
         ecig_sum_current = rowSums(select(., CELCIGT), na.rm = TRUE),
         non_ecig_sum_ever = rowSums(select(., starts_with("E",  ignore.case = FALSE),
                                            -EELCIGT), na.rm = TRUE),
         non_ecig_sum_current = rowSums(select(., starts_with("C", ignore.case = FALSE),
                                               -CELCIGT), na.rm = TRUE)) |>
  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))

# specify ever/current user
nyts_data <- nyts_data |>
             mutate(ecig_only_ever = case_when(ecig_ever == TRUE &
                                               non_ecig_ever == FALSE &
                                                ecig_current == FALSE &
                                             non_ecig_current == FALSE ~ TRUE,
                                                                TRUE ~ FALSE),
              ecig_only_current = case_when(ecig_current == TRUE &
                                               non_ecig_ever == FALSE &
                                             non_ecig_current == FALSE ~ TRUE,
                                                                TRUE ~ FALSE),
            non_ecig_only_ever = case_when(non_ecig_ever == TRUE &
                                                   ecig_ever == FALSE &
                                                ecig_current == FALSE &
                                             non_ecig_current == FALSE ~ TRUE,
                                                                TRUE ~ FALSE),
      non_ecig_only_current = case_when(non_ecig_current == TRUE &
                                                   ecig_ever == FALSE &
                                                ecig_current == FALSE ~ TRUE,
                                                                TRUE ~ FALSE),
                        no_use = case_when(non_ecig_ever == FALSE &
                                                   ecig_ever == FALSE &
                                                ecig_current == FALSE &
                                             non_ecig_current == FALSE ~ TRUE,
```

```
                                                        TRUE ~ FALSE)) %>%
                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 produc
ts"))

# sum up number of surveys in each year
nyts_data <- nyts_data |>
  add_count(year)
```

Here, we save the wrangled data so we don't need to run everything again.

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

Analysis

Exploratory Data Analysis

Before we can begin any analysis, we can first take a look at our data and understand what it contains. The following plot shows us the count of survey responses for each age 9 through 18 as well as those older than 18. There seems to be very few children aged 9 and 10 that took the survey so most of our findings will apply more to the ages of 11 through 18.

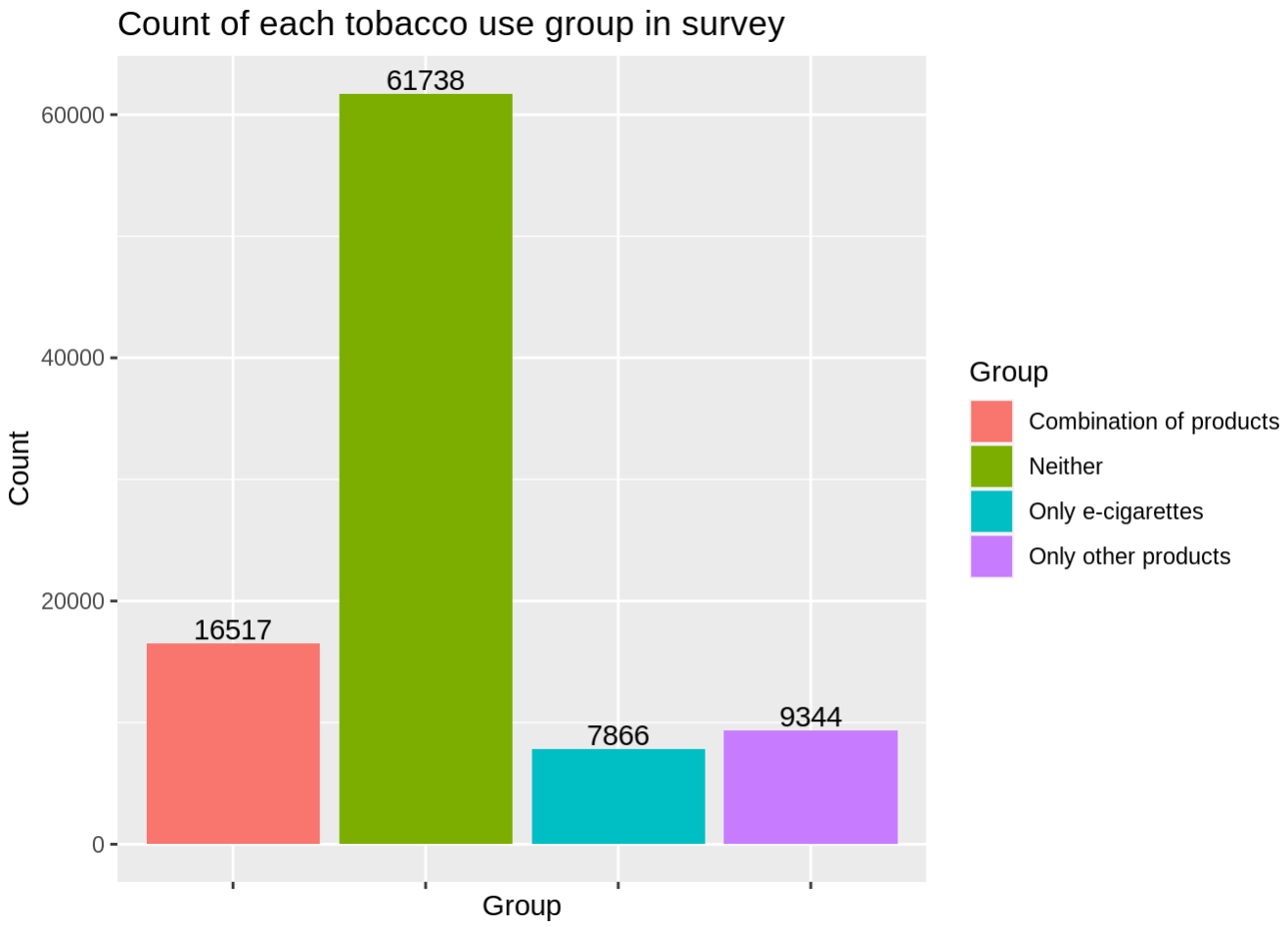
```
table(nyts_data$Age)
```

```
##
##   >18    10    11    12    13    14    15    16    17    18     9
##   727    50  5360 13499 14613 14036 13498 13205 12754  7108   198
```

The bar plot below shows us what category of tobacco use survey participants with the counts of each group. We can already see that most of the participants(61,738) don't use any form of tobacco, with the next largest group being users of a combination of products(16,517).

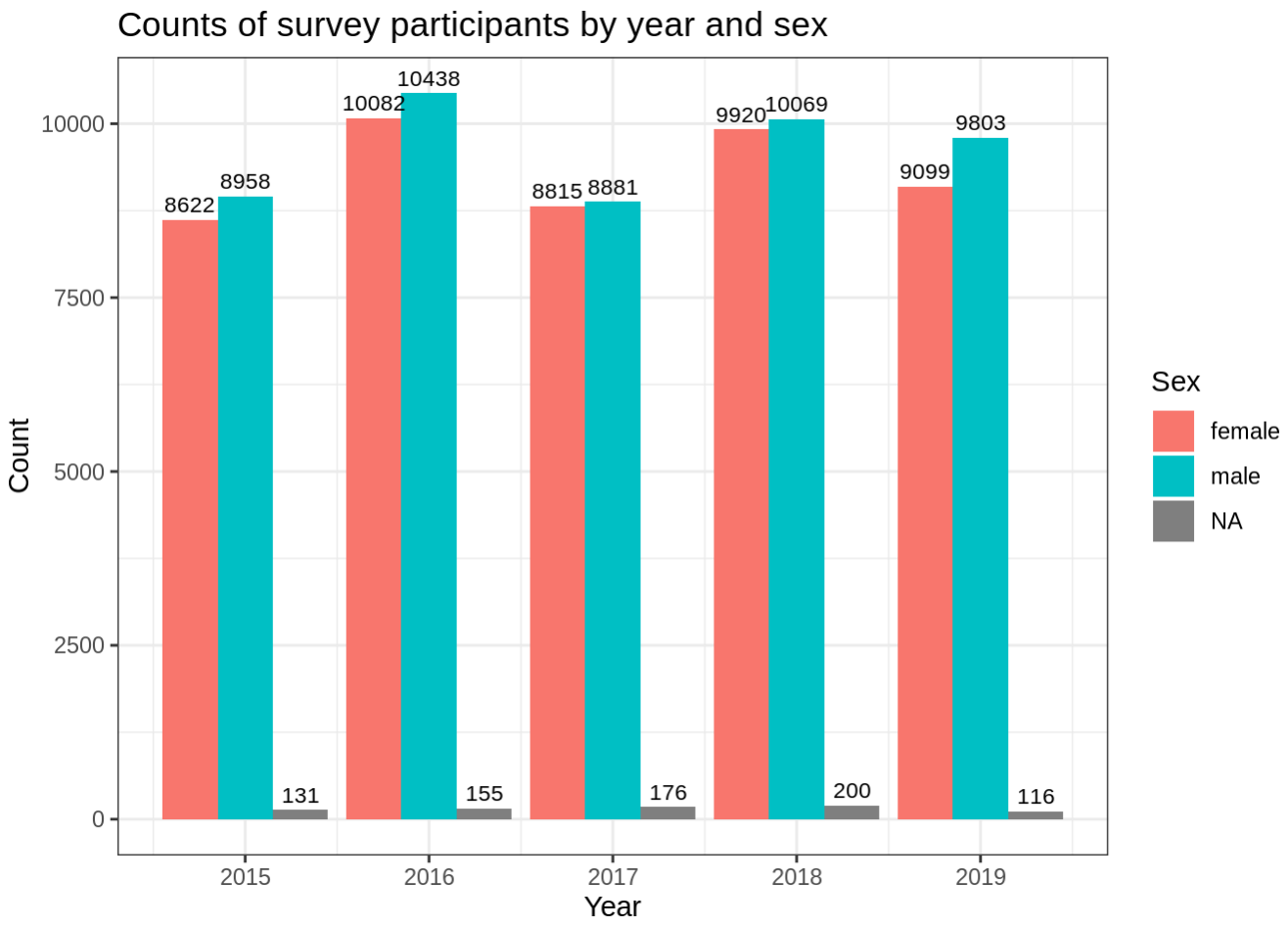
```
nyts_data |>
  group_by(Group) |>
  ggplot(aes(x=Group, fill=Group)) +
  geom_bar()+
  stat_count(aes(label=..count..), geom="text", position=position_dodge(0.9),vjust=-0.2) +
  labs(title = "Count of each tobacco use group in survey", x = "Group", y = "Count")  +
  theme(axis.text.x = element_blank())
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.

## Warning: Please use `after_stat(count)` instead.
```



Another important aspect of the data we can visualize is the roll that sex/gender plays in E-Cig use across the years in the survey.

```
nyts_data %>%
  group_by(year) %>%
  count(Sex) %>%
  ggplot(aes(x=year, y=n, fill=Sex)) +
  geom_col(position = "dodge") +
  geom_text(aes(label=n), position=position_dodge(width=0.9), vjust=-0.5, size = 3) +
  labs(title="Counts of survey participants by year and sex",
       x="Year",
       y="Count",
       fill="Sex") +
  theme_bw()
```

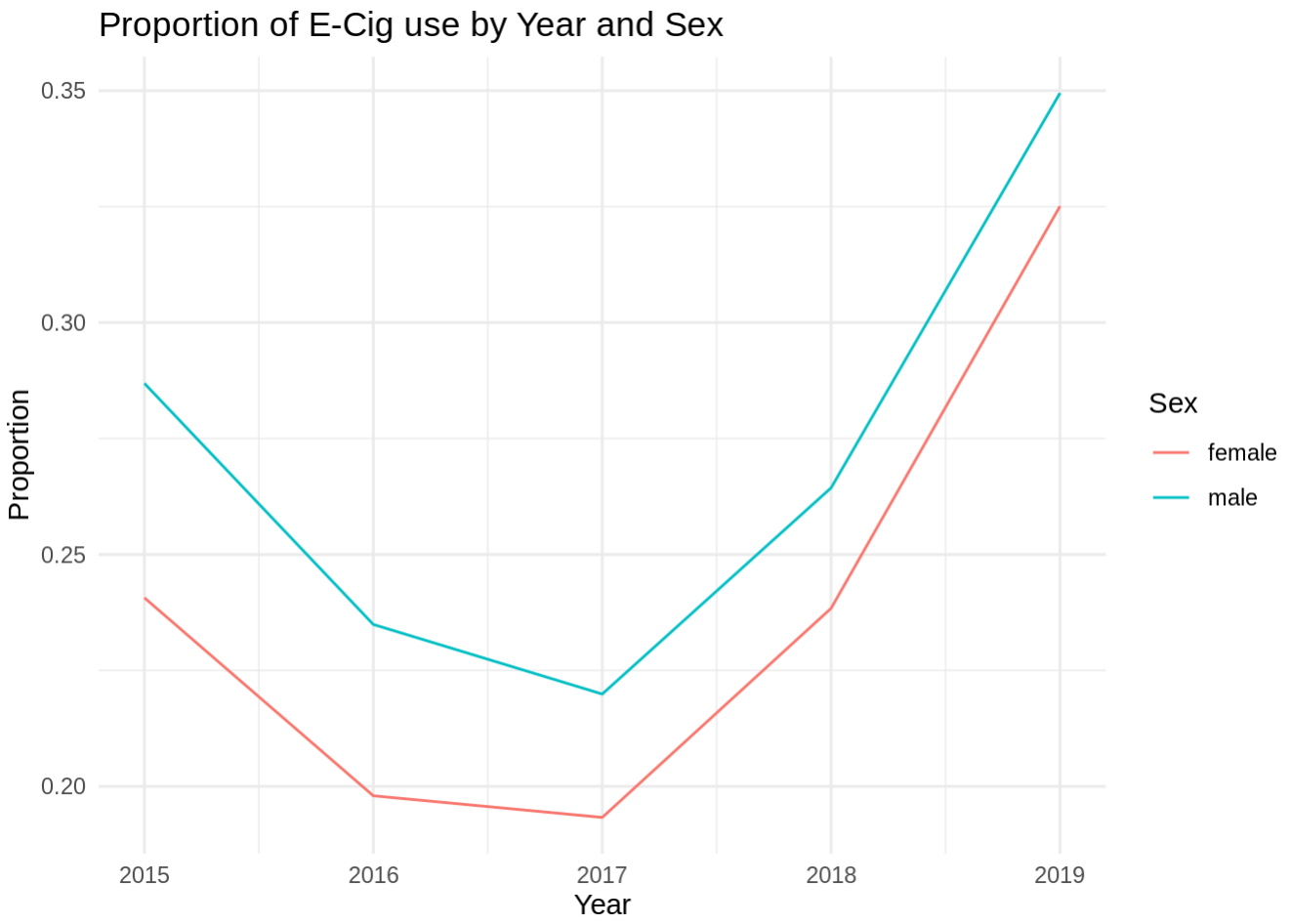


Here we can see that across our data, we have a pretty even response rate from both males and females for every year recorded. Knowing this, we should be able to accurately compare the E-Cig use across both genders.

We can now plot to see the proportion of E-Cig use for each gender by year.

```
nyts_data |>
  group_by(year, Sex) |>
  # count things
  summarize(mean_ever = mean(ecig_ever, na.rm=TRUE)) |>
  na.omit() |>
  ggplot(aes(x=year, y=mean_ever, group=Sex, color=Sex)) +
  geom_line() +
  labs(title="Proportion of E-Cig use by Year and Sex",
       x="Year", y="Proportion",
       color="Sex") +
  theme_minimal()
```

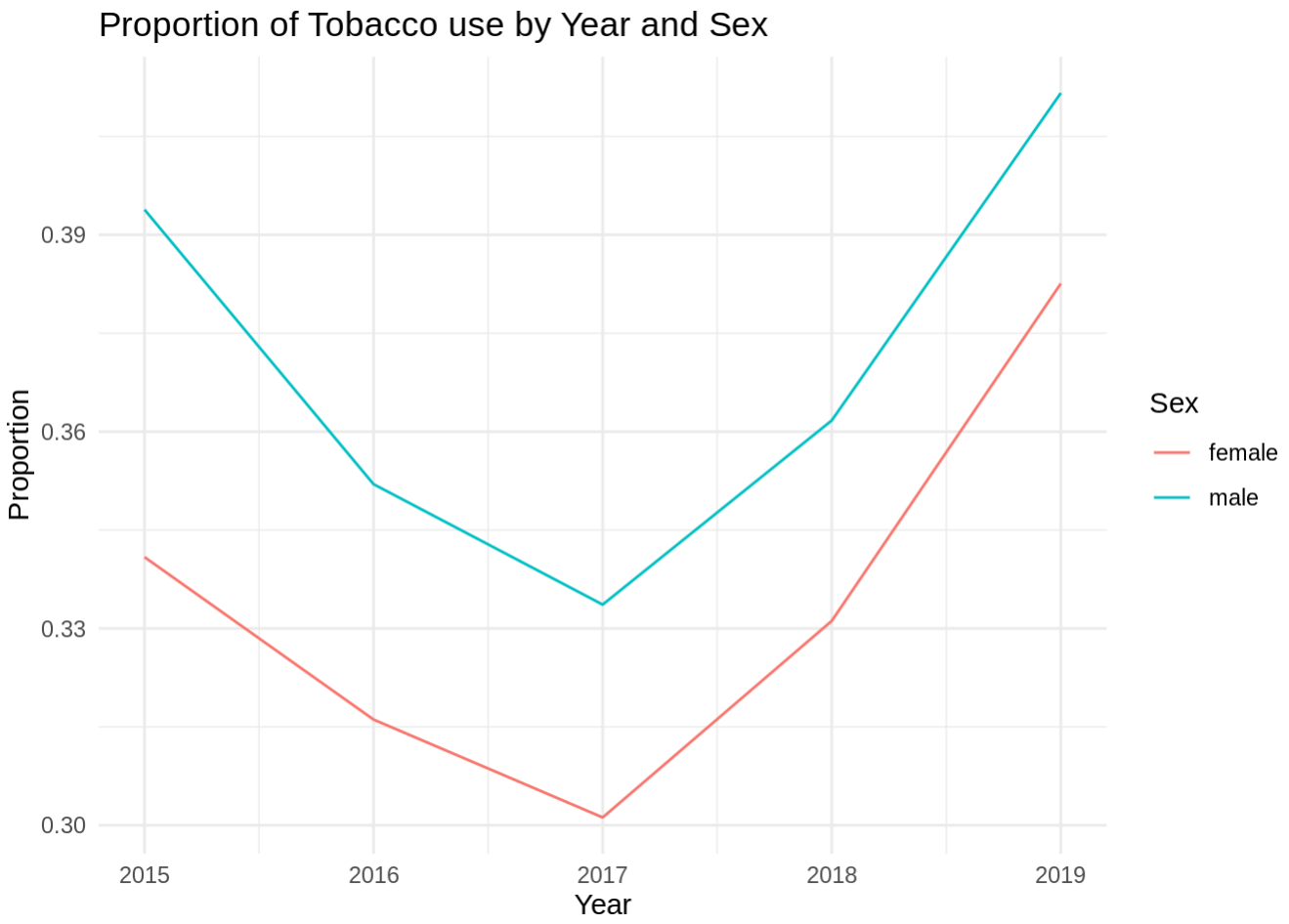
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups`
## argument.
```



We can see that from 2015 to 2019, the male proportion of E-cig use has been higher. We can also check to see if this is the case with tobacco use.

```
nyts_data |>
  group_by(year, Sex) |>
  # count things
  summarize(mean_ever = mean(tobacco_ever, na.rm=TRUE)) |>
  na.omit() |>
  ggplot(aes(x=year, y=mean_ever, group=Sex, color=Sex)) +
  geom_line() +
  labs(title="Proportion of Tobacco use by Year and Sex",
       x="Year", y="Proportion",
       color="Sex") +
  theme_minimal()
```

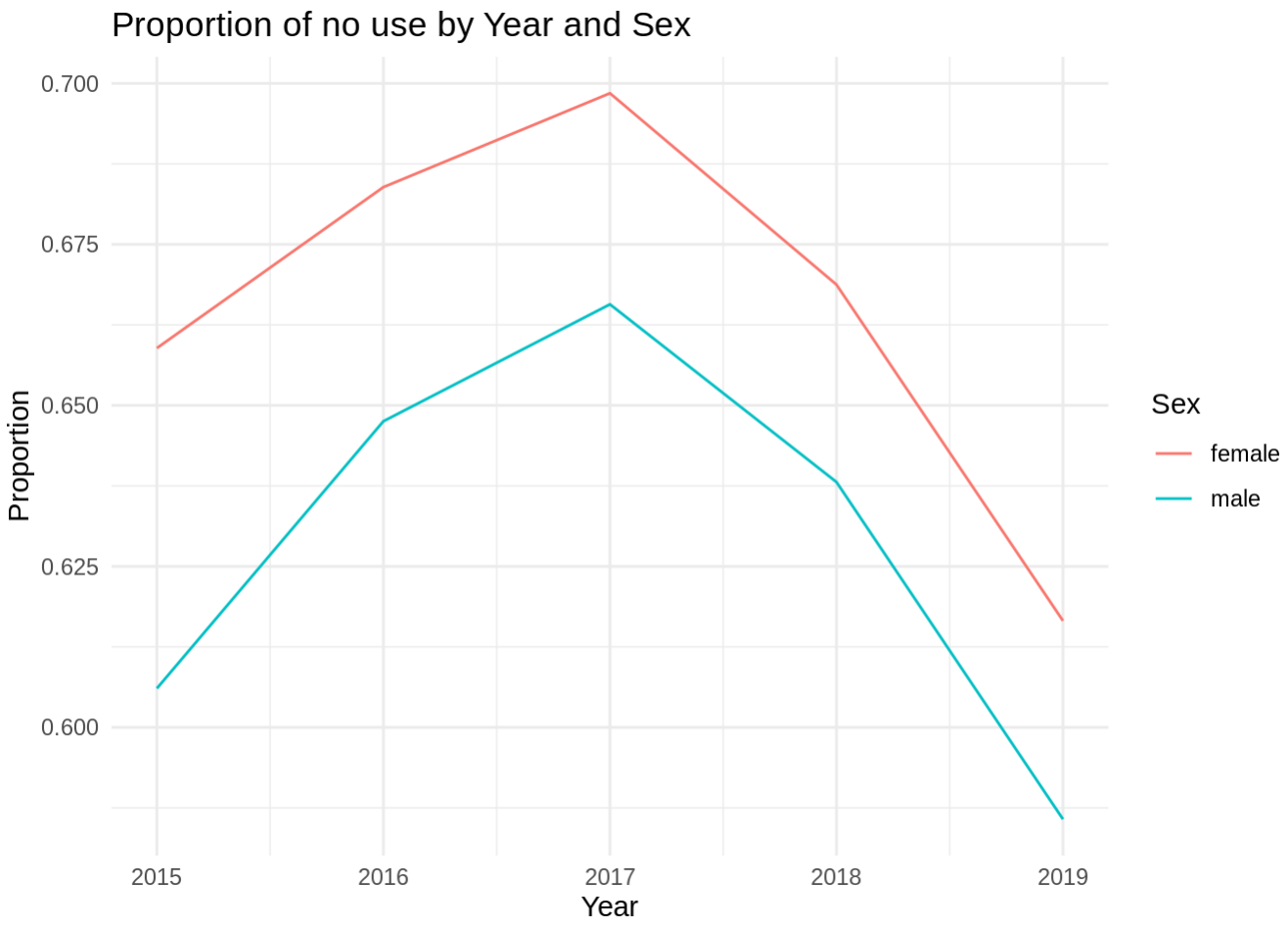
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups`
## argument.
```



Again, we see a similar trend of male tobacco use being high across all years from 2015 to 2019. Interestingly, in both plots we can see a dip in the proportion of use for both tobacco and e-cig around 2017. Looking below at the plot of non-users, we see that proportions increased during that 2017 dip in tobacco and e-cig users. This is consistent with what we would think happens.

```
nyts_data |>
  group_by(year, Sex) |>
  # count things
  summarize(mean_ever = mean(no_use, na.rm=TRUE)) |>
  na.omit() |>
  ggplot(aes(x=year, y=mean_ever, group=Sex, color=Sex)) +
  geom_line() +
  labs(title="Proportion of no use by Year and Sex",
       x="Year", y="Proportion",
       color="Sex") +
  theme_minimal()
```

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

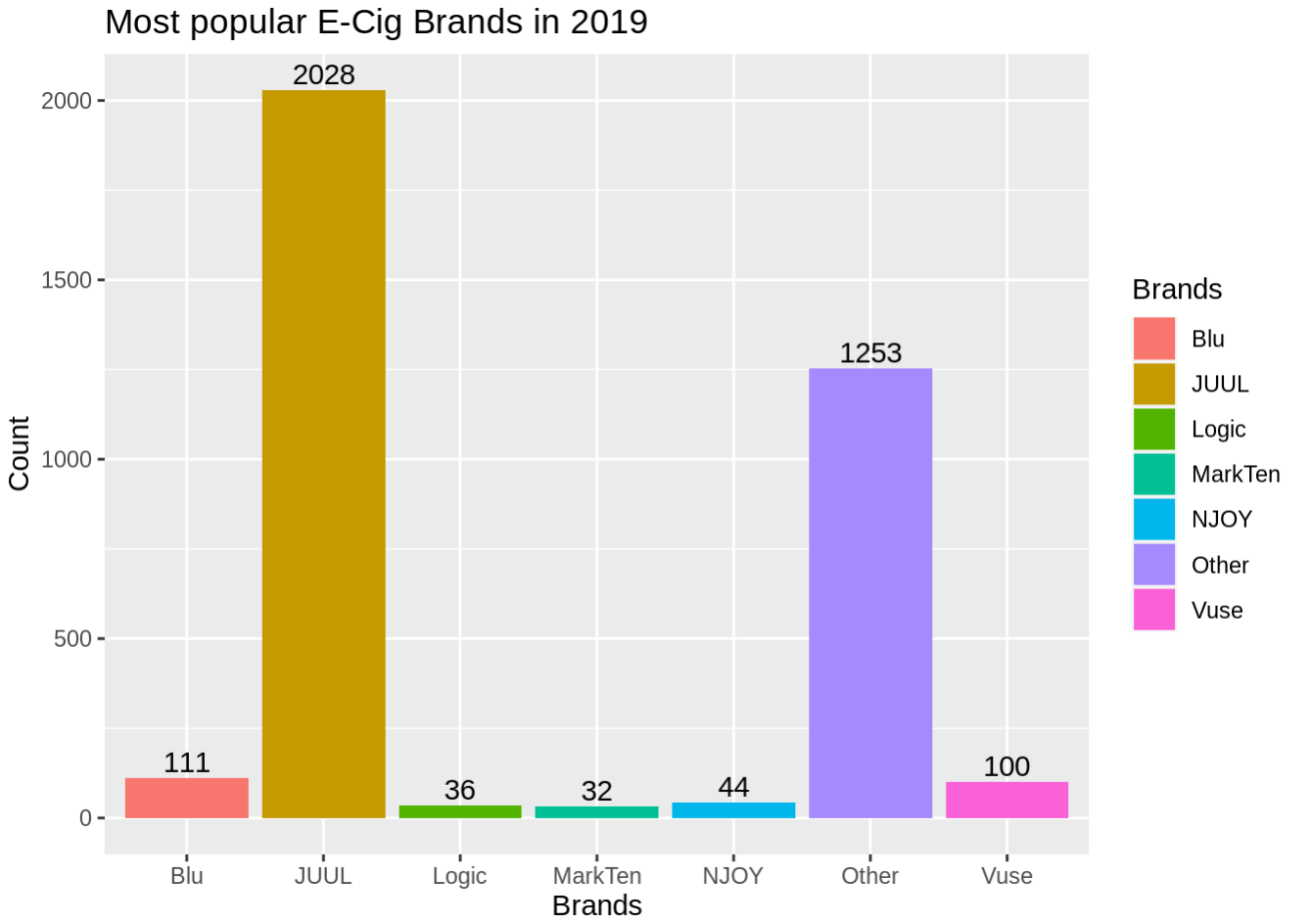


Next we can visualize the distribution of the most popular brands of e_cig. We should first point out that there is no brand information prior to 2019, so four our plot, we only see the distribution of brands for that year.

```
nyts_data |>
  group_by(year) |>
  count(brand_ecig)
```

```
## # A tibble: 12 × 3
## # Groups:   year [5]
##     year brand_ecig     n
##    <dbl> <chr>      <int>
##  1  2015 <NA>       17711
##  2  2016 <NA>       20675
##  3  2017 <NA>       17872
##  4  2018 <NA>       20189
##  5  2019 Blu          111
##  6  2019 JUUL        2028
##  7  2019 Logic         36
##  8  2019 MarkTen       32
##  9  2019 NJOY          44
## 10  2019 Other       1253
## 11  2019 Vuse         100
## 12  2019 <NA>       15414
```

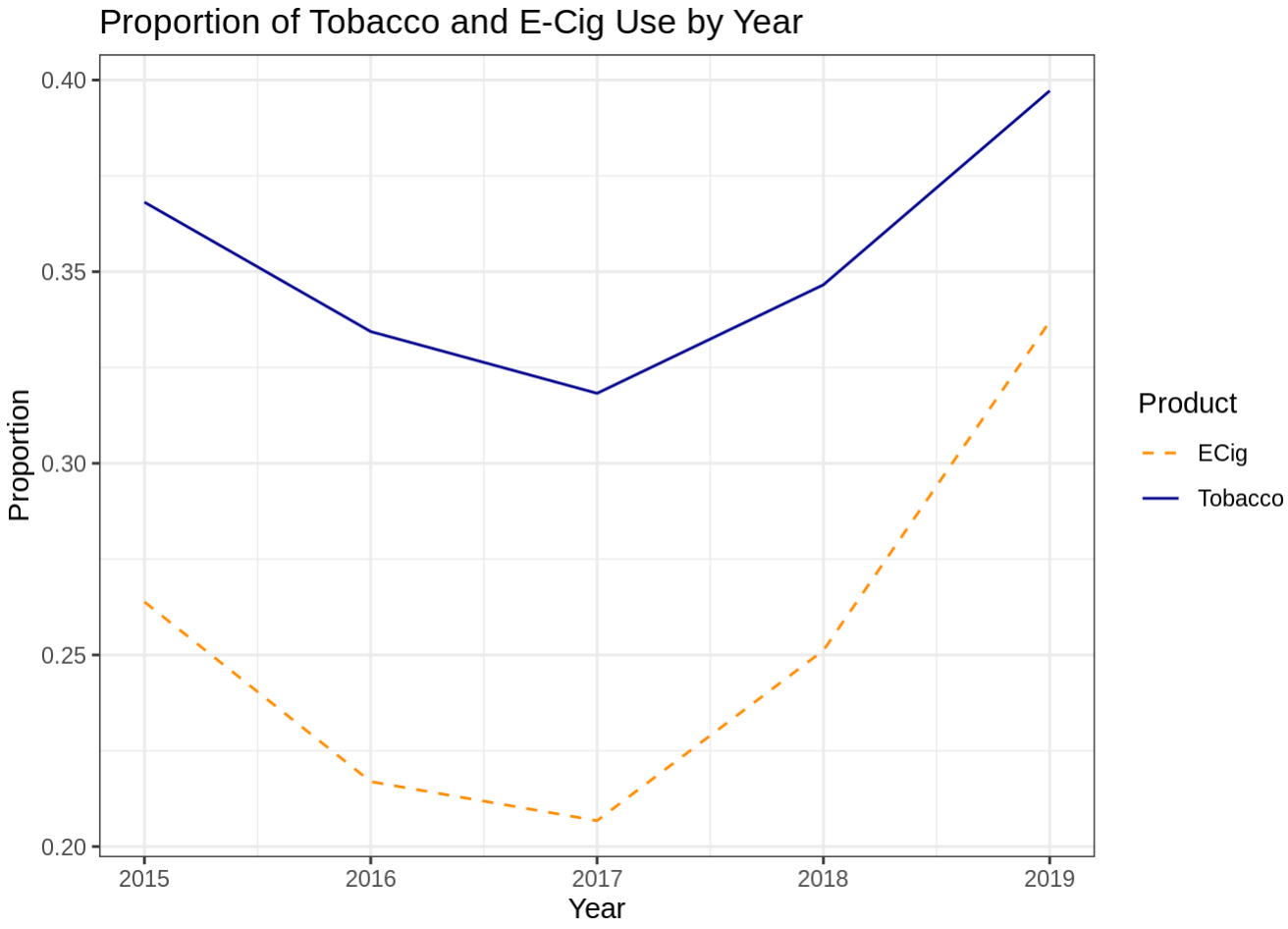
```
nyts_data |>
  filter(year == 2019 & !is.na(brand_ecig)) |>
  ggplot(aes(x=brand_ecig, fill = brand_ecig)) +
  geom_bar() +
  geom_text(stat='count', aes(label=..count..), vjust=-.3) +
  labs(title="Most popular E-Cig Brands in 2019", fill="Brands", x="Brands", y="Count")
```



Juul looks to be the most popular brand by a fairly large margin.

Another trend we can look at is tobacco and e-cig use across time. In the following plot, we can see that the proportion of e-cig use has generally been lower than that of tobacco use, and the two products have followed a similar pattern of increased and decreased use across between 2015 and 2019. However, we can also see that the difference in their individual proportion of use seems to have grown much closer around the year 2018 and especially 2019. This indicates a more rapid increase in use for e-cigs which is consistent with their increased popularity in recent years.

```
nyts_data |>
  group_by(year) |>
  summarize(ECig = mean(ecig_ever, na.rm = TRUE),
            Tobacco = mean(tobacco_ever, na.rm = TRUE)) |>
  pivot_longer(-year, names_to = "variable", values_to = "values") |>
  ggplot(aes(x = year, y = values, group = variable, linetype = variable, color = variabl
e)) +
  geom_line() +
  scale_linetype_manual(values = c("dashed", "solid"), name = "Product") +
  scale_color_manual(values = c("darkorange", "darkblue"), name = "Product") +
  labs(x = "Year", y = "Proportion", title = "Proportion of Tobacco and E-Cig Use by Yea
r") +
  theme_bw()
```



Data Analysis

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

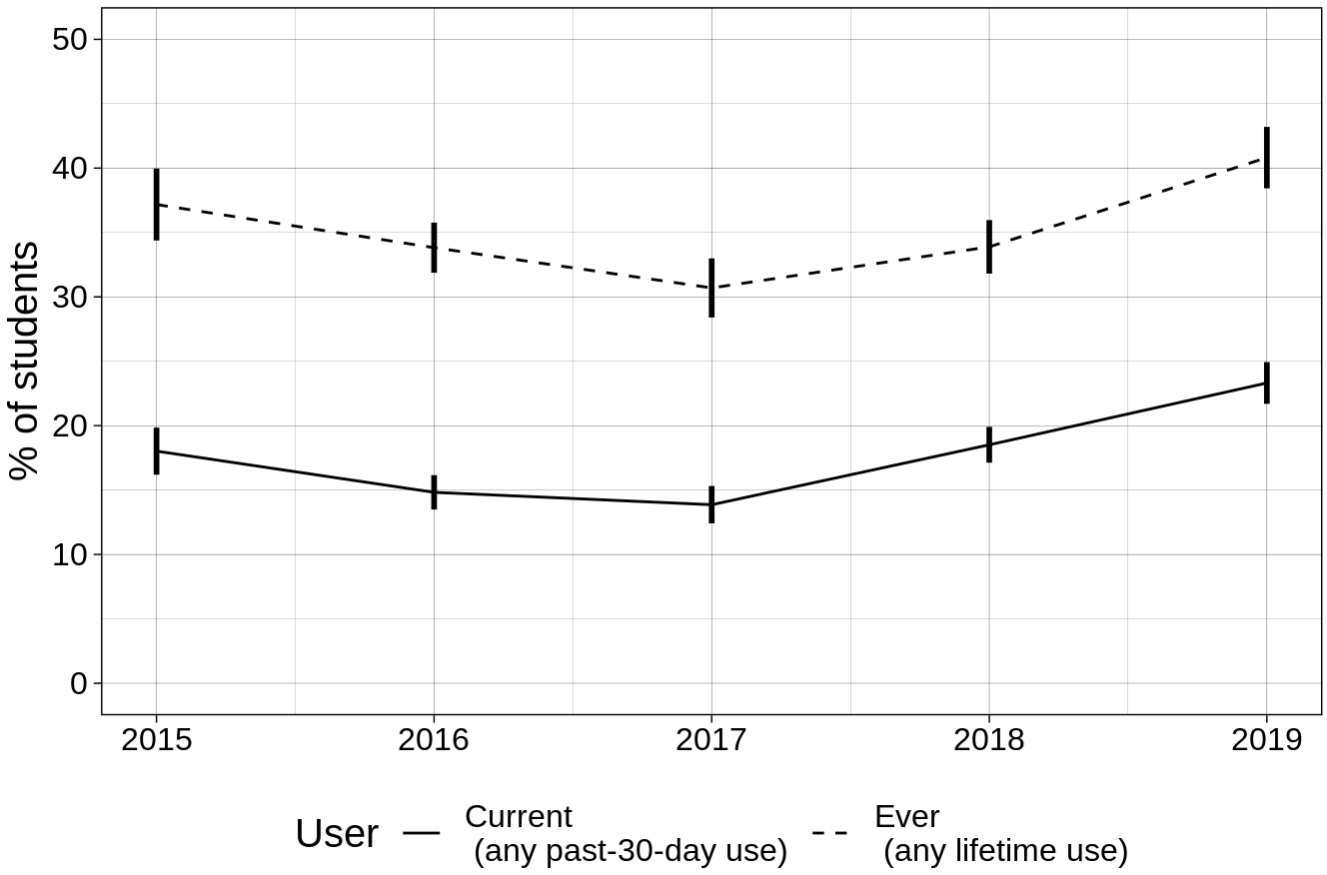
Tobacco Use

To better understand questions and perform analysis using our survey data, we need to adjust our data according to the survey design. Here, we generate a function to get 95% confidence interval and calculate the weighted mean.

```
# function to calculate averages of user proportions based on the survey design.
surveyMeanA <- function(currYear) {
  options(survey.lonely.psu = "adjust")
  currYear |>
    # specifies strata, cluster IDs and survey weights of the survey design
    as_survey_design(strata = stratum,
                     ids = psu,
                     weight  = finwgt,
                     nest = TRUE) |>
    summarize(tobacco_ever = survey_mean(tobacco_ever,
                                         vartype = "ci",
                                         na.rm = TRUE),
              # get confidence interval with upper and lower end
              tobacco_current = survey_mean(tobacco_current,
                                            vartype = "ci",
                                            na.rm = TRUE))  |>
    mutate_all("*", 100) |>
    pivot_longer(everything(),
                 names_to = "Type", # rename the variables
                 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 \n (any lifetime use)",
                            str_detect(Type, "current") ~ "Current \n (any past-30-day us
e)",
                            TRUE ~ "Mean"))}
```

Now, we could use that weighted mean to graph tobacco use. We have our 95% confidence interval here, which means that we are 95% sure that the true value fall between the lower bound and upper bound of our estimation.

```
# apply our function from above for each year and graph tobacco use
nyts_data |>
  group_by(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)) +
  # shows the confidence interval of each data point
  geom_linerange(aes(ymin = Lower,
                     ymax = Upper),
                     linewidth = 1,
                show.legend = FALSE) +
  # this allows us to choose what type of line we want for each line
  scale_linetype_manual(values = c(1, 2)) +
  # this allows us to specify how the y-axis should appear
  scale_y_continuous(breaks = seq(0, 50, by = 10),
                     labels = seq(0, 50, by = 10),
                     limits = c(0, 50)) +
    theme_linedraw() +
    labs(title = "Tobacco product users more prevalent after 2017",
         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 plot show that there is a larger proportion of students who have ever used tobacco than students who currently use tobacco from 2015 to 2019. This makes sense because current users of tobacco are a subset of ever users. The proportion of ever users of tobacco decreased from 2015 to 2017 and increased from 2017 to 2019. Similarly, the proportion of current users of tobacco decreased from 2015 to 2017 and increased from 2017 to 2019.

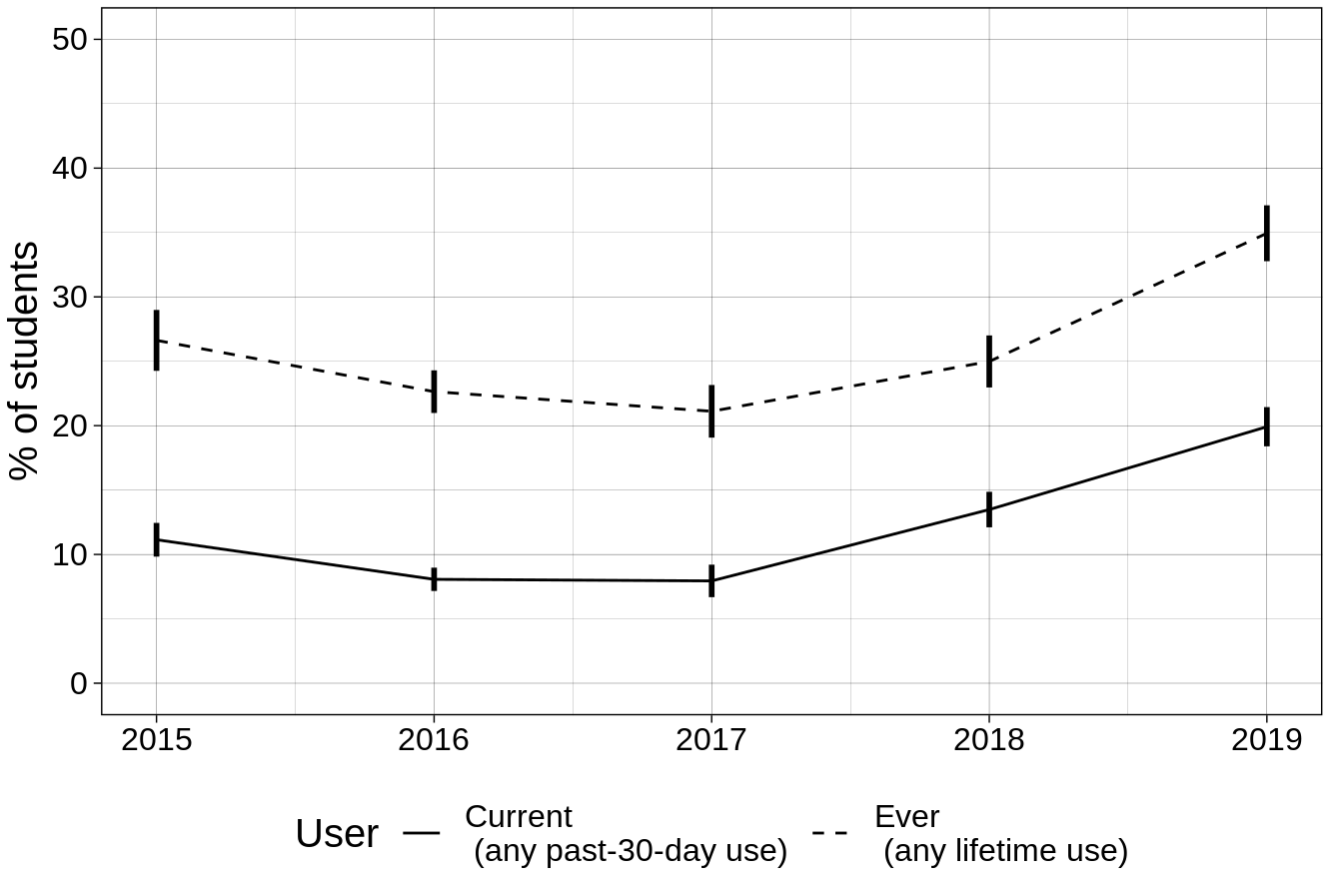
E-cig Use

We use the same function again to adjust the E-cigarette data according to the survey design.

```
# function to calculate averages of user proportions based on the survey design.
surveyMeanB <- function(currYear) {
  options(survey.lonely.psu = "adjust")
  currYear |>
    # specifies strata, cluster IDs and survey weights of the survey design
    as_survey_design(strata = stratum,
                     ids = psu,
                     weight  = finwgt,
                     nest = TRUE) |>
    summarize(ecig_ever = survey_mean(ecig_ever,
                                      vartype = "ci",
                                      na.rm = TRUE),
              # get confidence interval with upper and lower end
              ecig_current = survey_mean(ecig_current,
                                         vartype = "ci",
                                         na.rm = TRUE))  |>
    mutate_all("*", 100) |>
    pivot_longer(everything(),
                 names_to = "Type", # rename the variables
                 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 \n (any lifetime use)",
                            str_detect(Type, "current") ~ "Current \n (any past-30-day us
e)",
                            TRUE ~ "Mean"))}
```

Now, we could plot E-cigarette use data to see the trend.

```
nyts_data |>
  group_by(year) |>
  group_modify(~ surveyMeanB(.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)) +
  # shows the confidence interval of each data point
  geom_linerange(aes(ymin = Lower,
                     ymax = Upper),
                     linewidth = 1,
                show.legend = FALSE) +
  # this allows us to choose what type of line we want for each line
  scale_linetype_manual(values = c(1, 2)) +
  # this allows us to specify how the y-axis should appear
  scale_y_continuous(breaks = seq(0, 50, by = 10),
                     labels = seq(0, 50, by = 10),
                     limits = c(0, 50)) +
    theme_linedraw() +
    labs(title = "E-cigarette product users more prevalent after 2017",
         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")
```



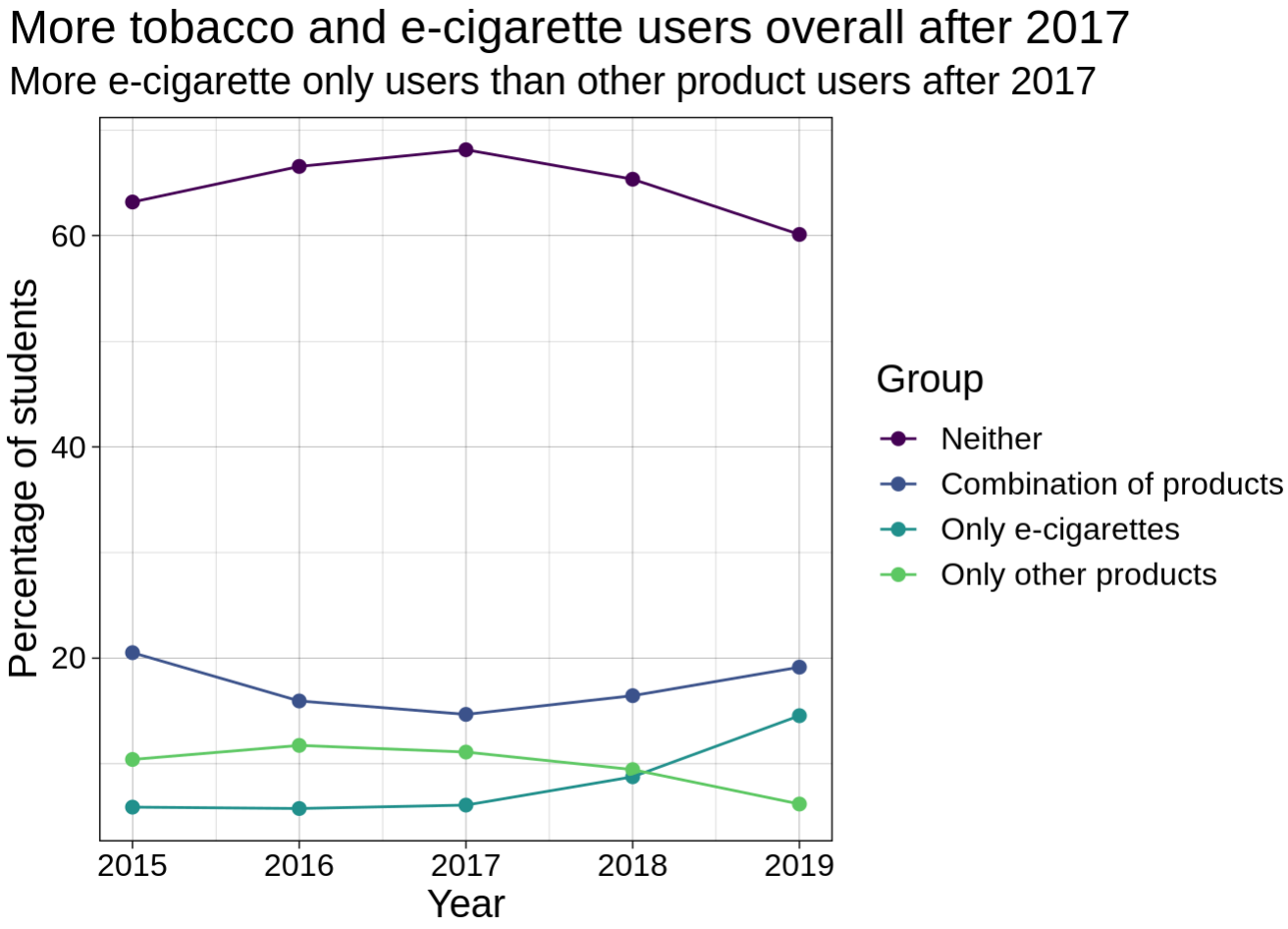
From the above graph, we again see the same pattern of higher ever use line compared with current use line because current users are included in ever users. Moreover, we could see that for both ever user and current user, the proportion has increased significantly from 2017 to 2019. We also noticed that from 2017 to 2019, the increase in the proportion of e-cigarette users is more significant than the increase in that of tobacco users. This might suggest that e-cigarette might have become more popular than tobacco products among American youths.

Product Usage

After examining the data separately, we would want to combine all the product usage data together into one graph to see the trend.

```
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",
       title = "More tobacco and e-cigarette users overall after 2017",
       subtitle = "More e-cigarette only users than other product users after 2017") +
  theme(text = element_text(size = 15),
        plot.title.position = "plot")
```

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



We could see that, overall, there are more e-cigarette and tobacco users after 2017. The proportion of students who use neither of the tobacco products decreased after 2017, though it is still the highest, which is expected as most people probably don't use any products. There is a higher proportion of students who use a combination of products than those who use a single category of product. In the combination of products curve, there is a drop from 2015 and 2017 and an increase from 2017 and 2019. We see that same increase from 2017 and 2019 in the e-cigarette curve, suggesting that it may contribute to the overall use increase during that period. On the other hand, the proportion of users of other products increased slightly from 2015 to 2016, but decreased from 2016 to 2019. This indicate that e-cigarette might have become more popular than any other tobacco products among youths during this period.

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

E-cig usage by Sex

To answer this question, we could plot e-cig use broken down by sex. We first use the function we had previously created to get the weighted data according to the survey design.

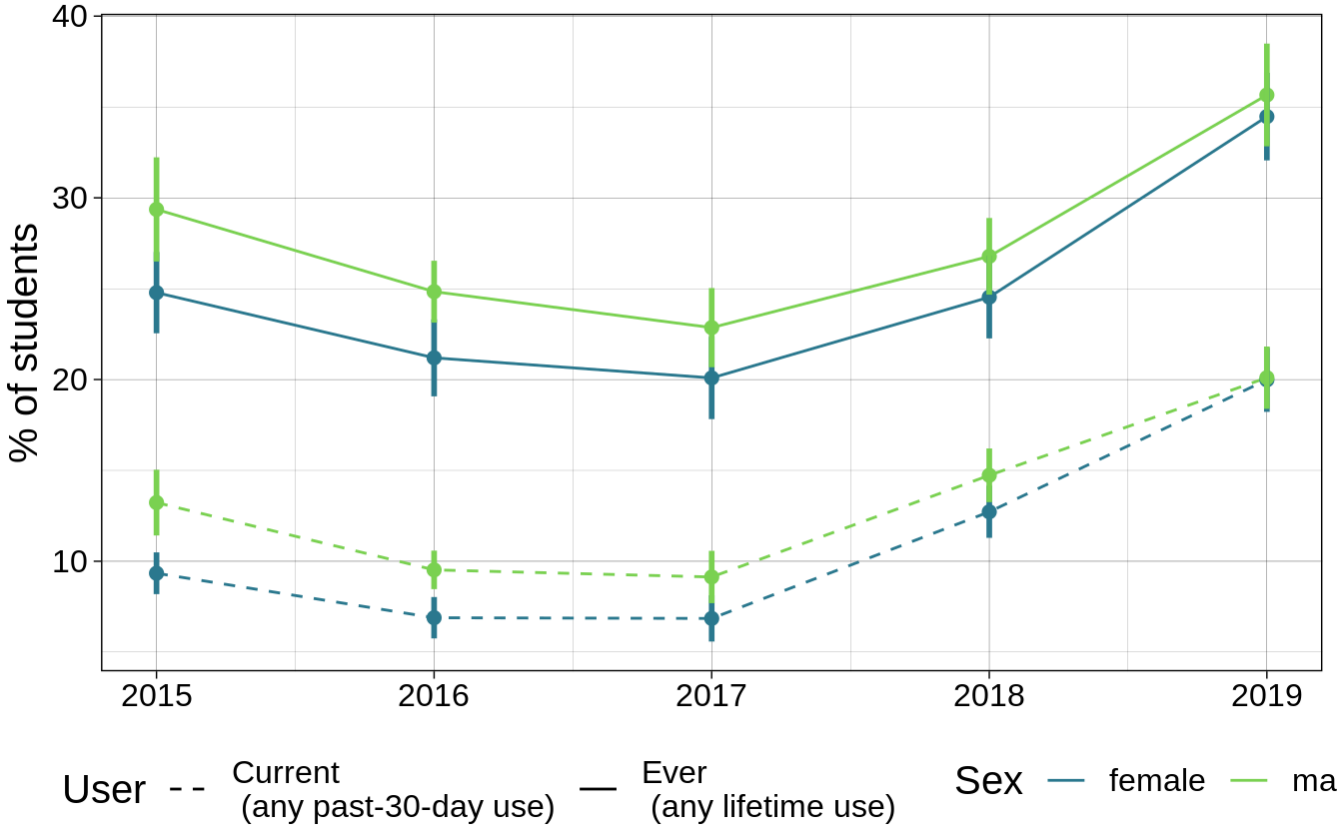
```
# function to calculate averages of user proportions based on the survey design.
surveyMeanC <- function(currYear) {
  options(survey.lonely.psu = "adjust")
  currYear |>
    # specifies strata, cluster IDs and survey weights of the survey design
    as_survey_design(strata = stratum,
                     ids = psu,
                     weight  = finwgt,
                     nest = TRUE) |>
    summarize(ecig_ever = survey_mean(EELCIGT,
                                      vartype = "ci",
                                      na.rm = TRUE),
              # get confidence interval with upper and lower end
              ecig_current = survey_mean(CELCIGT,
                                         vartype = "ci",
                                         na.rm = TRUE))  |>
    mutate_all("*", 100) |>
    pivot_longer(everything(),
                 names_to = "Type", # rename the variables
                 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 \n (any lifetime use)",
                            str_detect(Type, "current") ~ "Current \n (any past-30-day us
e)",
                            TRUE ~ "Mean"))}
```

Then, we make the plot broken down by sex.

```
v_colors =  viridis(6)[c(3, 5)]

nyts_data |>
  filter(!is.na(Sex)) |>
  group_by(year, Sex) |>
  group_modify(~ surveyMeanC(.x)) |>
  dplyr::select(-Type) |>
  pivot_wider(names_from = Estimate,
              values_from = `Percentage of students`) |>
  ggplot(aes(x = year, y = Mean, color = Sex)) +
  geom_line(aes(linetype = User)) +
  geom_point(show.legend = FALSE, size = 2) +
  # shows the confidence interval of each data point
  geom_linerange(aes(ymin = Lower,
                     ymax = Upper),
                     linewidth = 1,
                show.legend = FALSE) +
  scale_linetype_manual(values = c(2, 1)) +
  scale_color_manual(values = v_colors) +
  theme_linedraw() +
  labs(title = "E-cigarette usage between males and females",
       subtitle = "More male users than female users",
       y = "% of students") +
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        text = element_text(size = 15),
        plot.title.position = "plot")
```

E-cigarette usage between males and females

More male users than female users

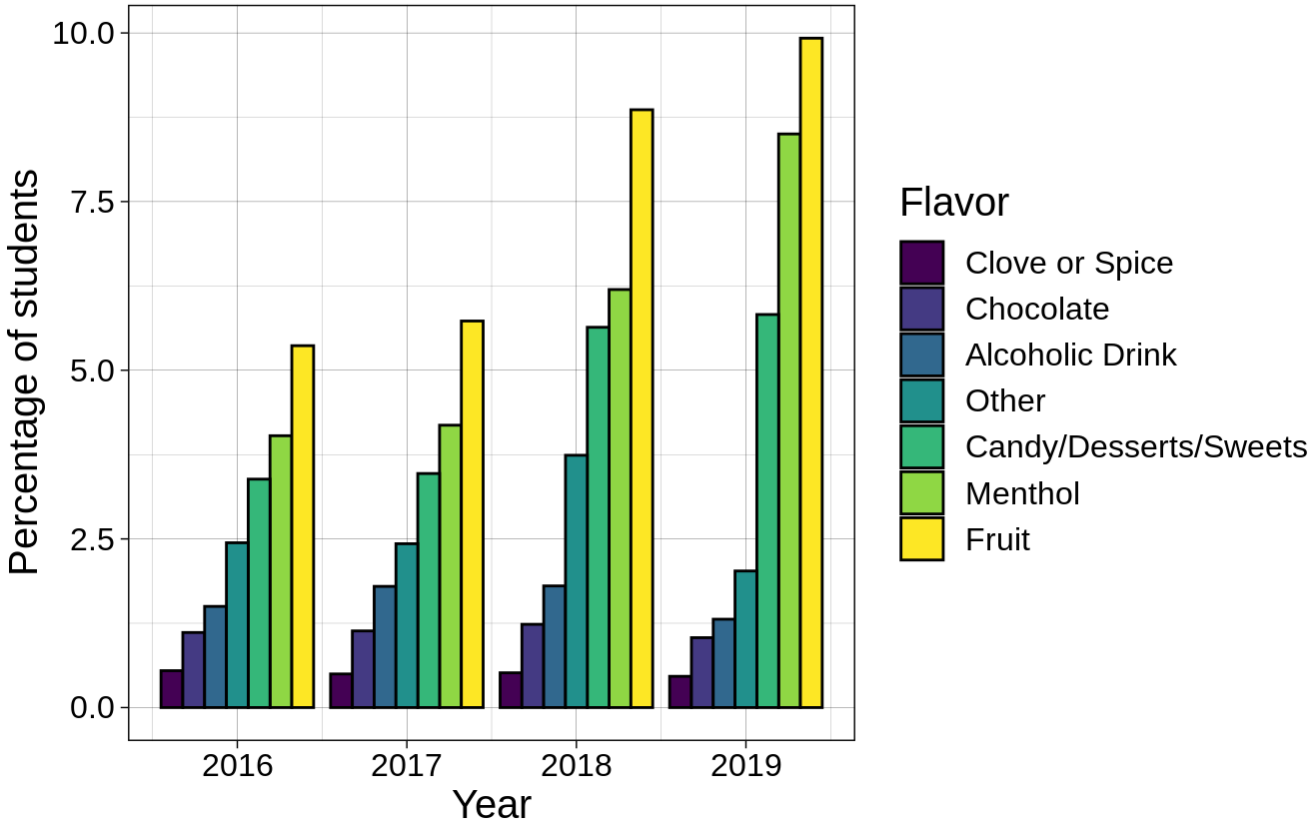


The plot shows male and female users follow a similar trend: they both decreased before 2017 and increased after 2017. We could see that on average, male tends to have a higher rate of being a current user and ever user for e-cigarette by about 5%, although the proportion of current female users have caught up with current male users in 2019.

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

To answer this question, we get the flavor data and calculate the percentage of students who are using these flavors.

```
nyts_data |>
  filter(year != 2015) |>
  group_by(year) |> # get the flavor percentage
  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 = "Fruit and Menthol are the most pupular vaping flavors",
       subtitle = "Flavors of tobacco products used in the past 30 days") +
  theme(text = element_text(size = 15),
        plot.title.position = 'plot')
```



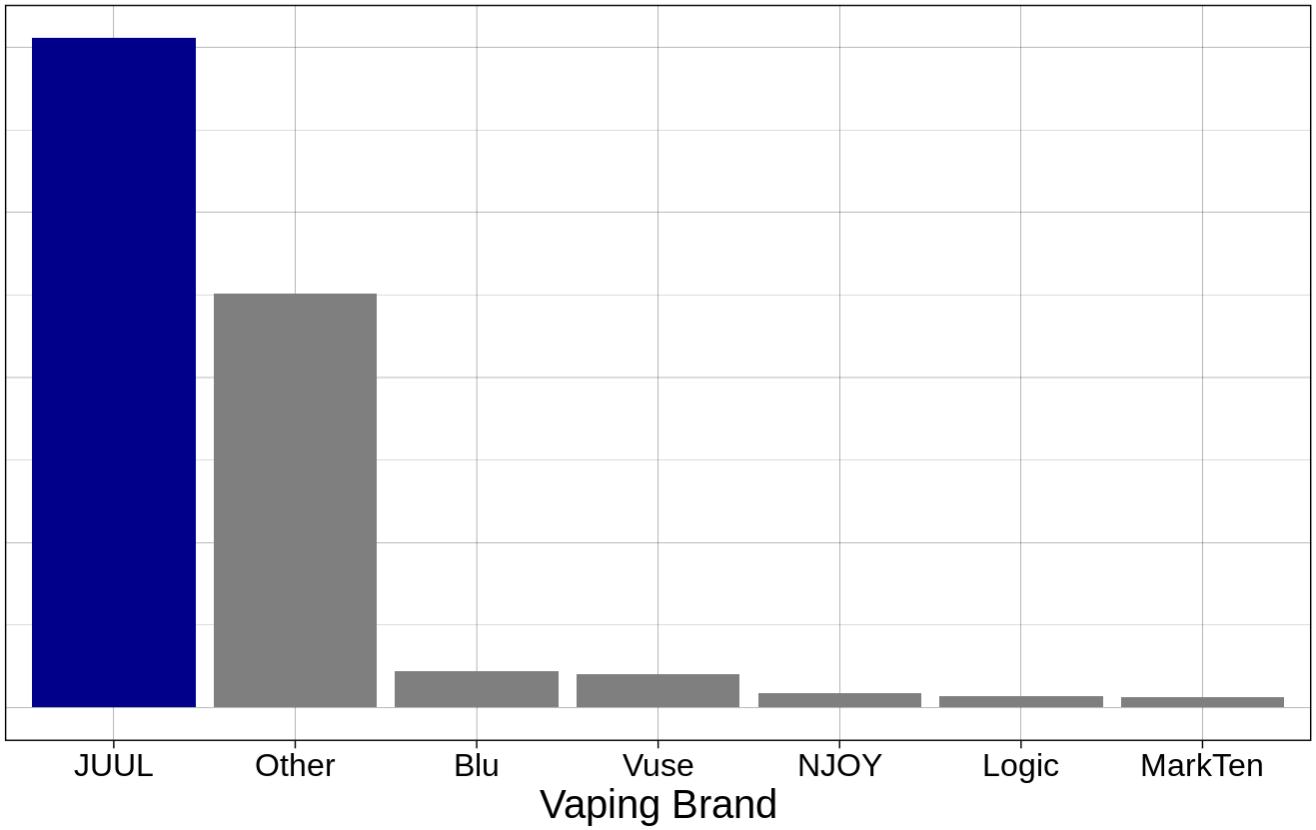
From the above graph, we could see that fruit, menthol, and candy/desserts/sweets are being used the most, and the ranking of popular flavors did not change from 2016 to 2019. With this graph, we also see the dramatic increase of e-cig use in 2018 and 2019, when the use of fruit flavor well surpassed other flavors.

Then, we want to plot and see what vaping brand is being used the most.

```
nyts_data |>
  filter(!brand_ecig %in% c(NA)) |>
  ggplot(aes(x = fct_infreq(brand_ecig), fill = brand_ecig)) +
  geom_bar() +
  scale_fill_manual(values = c("JUUL" = "darkblue")) +
  theme_linedraw() +
  theme(text = element_text(size = 15),
        legend.position = "none",
        plot.title.position = 'plot',
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
   labs(title = "JUUL is the most popular vaping brand amoung America Youth",
       subtitle = "Vaping brand data from 2019",
       x = "Vaping Brand",
       y = NULL)
```

JUUL is the most popular vaping brand amoung America Youth

Vaping brand data from 2019



From the data collected in 2019, we could see that JUUL is the most popular vaping brand in which the number reported is a lot higher than the other brands.

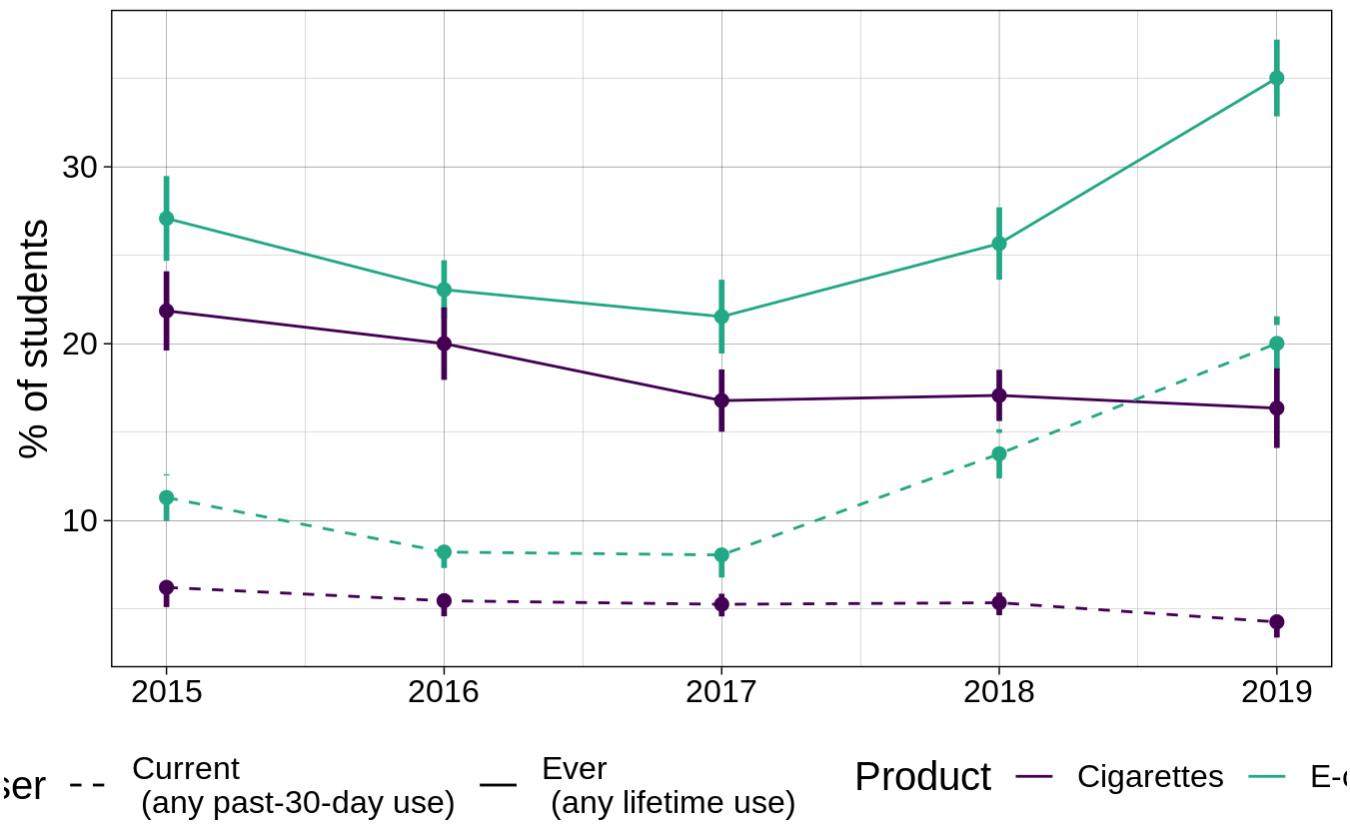
Q4: Is there a relationship between e-cigarette/vaping use and other tobacco use?

We want to see if there's a relationship between e-cig use and other tobacco use. Ultimately, we will fit a model trying to predict e-cig use. We first use apply the function to calculate the weighted data.

```r
# function to calculate averages of user proportions based on the survey design.
surveyMeanD <- function(currYear) {
  options(survey.lonely.psu = "adjust")
  currYear |>
    # specifies strata, cluster IDs and survey weights of the survey design
    as_survey_design(
      strata = stratum,
      ids = psu,
      weight  = finwgt,
      nest = TRUE
    ) |>
    # get confidence interval with upper and lower end
    summarize(
      ecig_ever = survey_mean(EELCIGT,
                              vartype = "ci",
                              na.rm = TRUE),
      ecig_current = survey_mean(CELCIGT,
                                 vartype = "ci",
                                 na.rm = TRUE),
      cigt_ever = survey_mean(ECIGT,
                              vartype = "ci",
                              na.rm = TRUE),
      cigt_current = survey_mean(CCIGT,
                                 vartype = "ci",
                                 na.rm = TRUE)
    )  |>
    mutate_all("*", 100) |>
    pivot_longer(everything(),
                 names_to = "Type", # rename the variables
                 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 \n (any lifetime use)",
        str_detect(Type, "current") ~ "Current \n (any past-30-day use)",
        TRUE ~ "Mean"
      ),
      Product = case_when(
        str_detect(Type, "ecig") ~ "E-cigarettes",
        str_detect(Type, "cigt") ~ "Cigarettes",
        TRUE ~ "Mean"
      )
    )
}
```

We want to first make a plot to see the overall trend of e-cig use and tobacco use.

```
v_colors =  viridis(6)[c(1, 4)]

nyts_data |>
  group_by(year) |>
  group_modify(~ surveyMeanD(.x)) |>
  dplyr::select(-Type) |>
  pivot_wider(
              names_from = Estimate,
              values_from = `Percentage of students`) |>
  ggplot(aes(
    x = year,
    y = Mean,
    color = Product,
    linetype = User
  )) +
  geom_line() +
  geom_point(show.legend = FALSE, size = 2) +
  # shows the confidence interval of each data point
  geom_linerange(aes(ymin = Lower,
                     ymax = Upper),
                     linewidth = 1,
              show.legend = FALSE) +
  scale_linetype_manual(values = c(2, 1)) +
  scale_color_manual(values = v_colors) +
  theme_linedraw() +
  labs(title = "E-cigarette use more prevalent than cigarette use after 2017",
       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")
```



From the above graph, we can see both e-cigarette and cigarette use decreased before 2017. However, e-cigarette use increased significantly after 2017, while cigarette use largely follows a decreasing trend from 2015 to 2019.

Next, we create a survey-weighted logistic regression model for each year's data to quantify the relationship between e-cigarette use and other tobacco products use.

```
# save data for faster knitting
save(currEcigToba15, currEcigToba16, currEcigToba17, currEcigToba18, currEcigToba19, file
="data/wrangled/q4models.rda")
```

```
# load data for faster knitting
load("data/wrangled/q4models.rda")
```

We decided to use `ecig_current`, which represents the students who currently use e-cigarettes, as the outcome variable and `non_ecig_current`, which represents the students who currently use tobacco products other than e-cigarettes, as the main predictor. We chose the current users because an ever user may not be a current user, and it is more urgent to address the issue of the current use of tobacco products. We also included sex and grade as predictors, as we see e-cigarette use is different between males and females in Question 2, and we see e-cigarette use differs across users of different ages in Question 5, our extension of analysis.

```
# create survey object with specific survey design
dat2015_survey_design <- nyts_data |>
  filter(year == 2015) |>
  as_survey_design(
    strata = stratum,
    ids = psu,
    weight  = finwgt,
    nest = TRUE
  )

# create e-cig model for 2015
currEcigToba15 <-
  survey::svyglm(
    ecig_current ~ non_ecig_current + Sex + Grade,
    family = quasibinomial(link = 'logit'),
    design = dat2015_survey_design
  )
```

```
# select to include only the slope of `non_ecig_currentTRUE` for comparison later
currEcigToba15Tidy <- tidy(currEcigToba15) |>
  select(term, estimate) |>
  filter(term == "non_ecig_currentTRUE") |>
  mutate(year = 2015)

tidy(currEcigToba15)
```

```
## # A tibble: 10 × 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)          -2.74       0.122   -22.6    4.29e-30
##  2 non_ecig_currentTRUE  2.81       0.0911   30.9    2.89e-37
##  3 Sexmale               0.165      0.0916    1.80   7.66e- 2
##  4 Grade11              -0.00779    0.113    -0.0691 9.45e- 1
##  5 Grade12              -0.0699     0.105    -0.666  5.08e- 1
##  6 Grade6               -1.30       0.161    -8.07   5.18e-11
##  7 Grade7               -0.804      0.143    -5.62   5.87e- 7
##  8 Grade8               -0.255      0.139    -1.83   7.22e- 2
##  9 Grade9               -0.118      0.110    -1.08   2.87e- 1
## 10 GradeUngraded/Other   0.285      0.664     0.429  6.69e- 1
```

We do the same thing for 2016.

```
dat2016_survey_design <- nyts_data |>
  filter(year == 2016) |>
  as_survey_design(
    strata = stratum,
    ids = psu,
    weight  = finwgt,
    nest = TRUE
  )

currEcigToba16 <-
  survey::svyglm(
    ecig_current ~ non_ecig_current  + Sex + Grade,
    family = quasibinomial(link = 'logit'),
    design = dat2016_survey_design
  )
```

```
currEcigToba16Tidy <- tidy(currEcigToba16) |>
  select(term, estimate) |>
  filter(term == "non_ecig_currentTRUE") |>
  mutate(year = 2016)

tidy(currEcigToba16)
```

```
## # A tibble: 10 × 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)             -2.96     0.154     -19.2  1.37e-26
##  2 non_ecig_currentTRUE     2.74     0.0821     33.4  4.08e-39
##  3 Sexmale                  0.199    0.0900      2.21 3.13e- 2
##  4 Grade11                 -0.161    0.132      -1.22 2.28e- 1
##  5 Grade12                 -0.181    0.122      -1.48 1.43e- 1
##  6 Grade6                  -1.14     0.194      -5.89 2.20e- 7
##  7 Grade7                  -0.958    0.205      -4.67 1.85e- 5
##  8 Grade8                  -0.363    0.158      -2.30 2.53e- 2
##  9 Grade9                  -0.243    0.117      -2.08 4.21e- 2
## 10 GradeUngraded/Other      1.34     0.458       2.93 4.82e- 3
```

2017 model.

```
dat2017_survey_design <- nyts_data |>
  filter(year == 2017) |>
  as_survey_design(
    strata = stratum,
    ids = psu,
    weight  = finwgt,
    nest = TRUE
  )

currEcigToba17 <-
  survey::svyglm(
    ecig_current ~ non_ecig_current + Sex + Grade,
    family = quasibinomial(link = 'logit'),
    design = dat2017_survey_design
  )
```

```
currEcigToba17Tidy <- tidy(currEcigToba17) |>
  select(term, estimate) |>
  filter(term == "non_ecig_currentTRUE") |>
  mutate(year = 2017)

tidy(currEcigToba17)
```

```
## # A tibble: 10 × 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)           -2.80      0.161    -17.4   1.73e-24
##  2 non_ecig_currentTRUE   2.63      0.0976    26.9   4.40e-34
##  3 Sexmale                0.209     0.0816     2.57  1.29e- 2
##  4 Grade11               -0.246     0.148     -1.66  1.03e- 1
##  5 Grade12               -0.0277    0.131     -0.211 8.33e- 1
##  6 Grade6                -1.25      0.174     -7.17  1.64e- 9
##  7 Grade7                -1.42      0.199     -7.12  2.04e- 9
##  8 Grade8                -0.704     0.196     -3.59  6.85e- 4
##  9 Grade9                -0.231     0.134     -1.73  8.93e- 2
## 10 GradeUngraded/Other    0.895     0.645      1.39  1.71e- 1
```

2018 model.

```
dat2018_survey_design <- nyts_data |>
  filter(year == 2018) |>
  as_survey_design(
    strata = stratum,
    ids = psu,
    weight  = finwgt,
    nest = TRUE
  )

currEcigToba18 <-
  survey::svyglm(
    ecig_current ~ non_ecig_current + Sex + Grade,
    family = quasibinomial(link = 'logit'),
    design = dat2018_survey_design
  )
```

```
currEcigToba18Tidy <- tidy(currEcigToba18) |>
  select(term, estimate) |>
  filter(term == "non_ecig_currentTRUE") |>
  mutate(year = 2018)

tidy(currEcigToba18)
```

```
## # A tibble: 10 × 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)            -2.01      0.0997   -20.2   7.01e-31
##  2 non_ecig_currentTRUE    2.31      0.0793    29.1   8.18e-41
##  3 Sexmale                 0.0769    0.0483     1.59  1.16e- 1
##  4 Grade11                 0.242     0.0938     2.58  1.21e- 2
##  5 Grade12                 0.126     0.105      1.21  2.31e- 1
##  6 Grade6                 -1.98      0.164    -12.1   7.61e-19
##  7 Grade7                 -1.24      0.163     -7.63  8.60e-11
##  8 Grade8                 -0.908     0.129     -7.03  1.11e- 9
##  9 Grade9                 -0.177     0.146     -1.21  2.31e- 1
## 10 GradeUngraded/Other    -0.221     0.552     -0.400 6.90e- 1
```

2019 model.

```
options(survey.adjust.domain.lonely=TRUE)
options(survey.lonely.psu="adjust")

dat2019_survey_design <- nyts_data |>
  filter(year == 2019) |>
  as_survey_design(
    strata = stratum,
    ids = psu,
    weight  = finwgt,
    nest = TRUE
  )

currEcigToba19 <-
  survey::svyglm(
    ecig_current ~ non_ecig_current + Sex + Grade,
    family = quasibinomial(link = 'logit'),
    design = dat2019_survey_design
  )
```

```
currEcigToba19Tidy <- tidy(currEcigToba19) |>
  select(term, estimate) |>
  filter(term == "non_ecig_currentTRUE") |>
  mutate(year = 2019)

tidy(currEcigToba19)
```

```
## # A tibble: 10 × 5
##    term                  estimate std.error statistic  p.value
##    <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)           -1.42      0.0842   -16.8   1.51e-25
##  2 non_ecig_currentTRUE   2.56      0.0825    31.0   4.21e-41
##  3 Sexmale               -0.172     0.0599    -2.88  5.35e- 3
##  4 Grade11                0.0651    0.0886     0.735 4.65e- 1
##  5 Grade12                0.337     0.0837     4.02  1.51e- 4
##  6 Grade6                -1.55      0.151    -10.3   2.21e-15
##  7 Grade7                -1.11      0.126     -8.88  7.11e-13
##  8 Grade8                -0.545     0.126     -4.34  5.04e- 5
##  9 Grade9                -0.134     0.0910    -1.47  1.46e- 1
## 10 GradeUngraded/Other   -1.03      0.578     -1.78  7.94e- 2
```

After running the models, we combine all model outputs from all years for comparison.

```
comparing_year <- currEcigToba15Tidy |>
  bind_rows(currEcigToba16Tidy, currEcigToba17Tidy, currEcigToba18Tidy, currEcigToba19Tid
y)

comparing_year
```

```
## # A tibble: 5 × 3
##   term                 estimate  year
##   <chr>                   <dbl> <dbl>
## 1 non_ecig_currentTRUE     2.81  2015
## 2 non_ecig_currentTRUE     2.74  2016
## 3 non_ecig_currentTRUE     2.63  2017
## 4 non_ecig_currentTRUE     2.31  2018
## 5 non_ecig_currentTRUE     2.56  2019
```

The slope for `non_ecig_currentTRUE` in all models are positive, which suggests a positive relationship between current use of other tobacco products and the current use of e-cigarettes. We then take the average of all slopes to get a bigger picture.

```
mean_log_odds = mean(comparing_year$estimate)
(mean_log_odds)
```

```
## [1] 2.609281
```

```
# exponentiates mean_log_odds to obtain the original odds
mean_odds = exp(mean_log_odds)
mean_odds
```

```
## [1] 13.58928
```

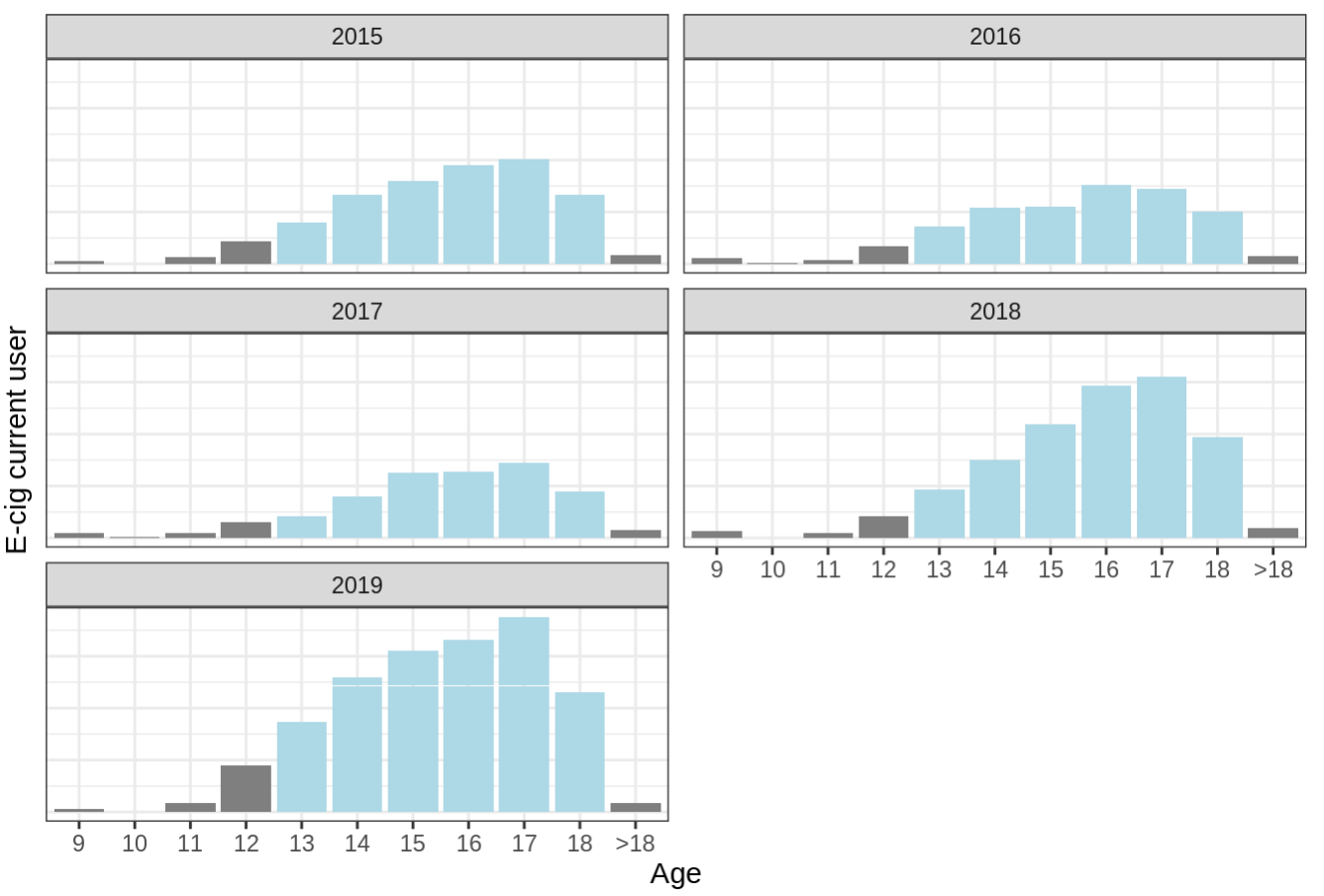
The average log odds of currently using e-cigarettes is 2.609 higher for someone who also currently uses other tobacco products than someone who does not currently uses other tobacco products, taking survey weights into account.

In other words, the average odds of currently using e-cigarettes for someone who also currently uses other tobacco products is 13.589 times the odds for someone who does not currently uses other tobacco products, taking survey weights into account.

Q5: What is the relationship between e-cigarette and tobacco/non-ecigarette use and age?

To answer this question, we plot e-cig current use and break down by age and year.

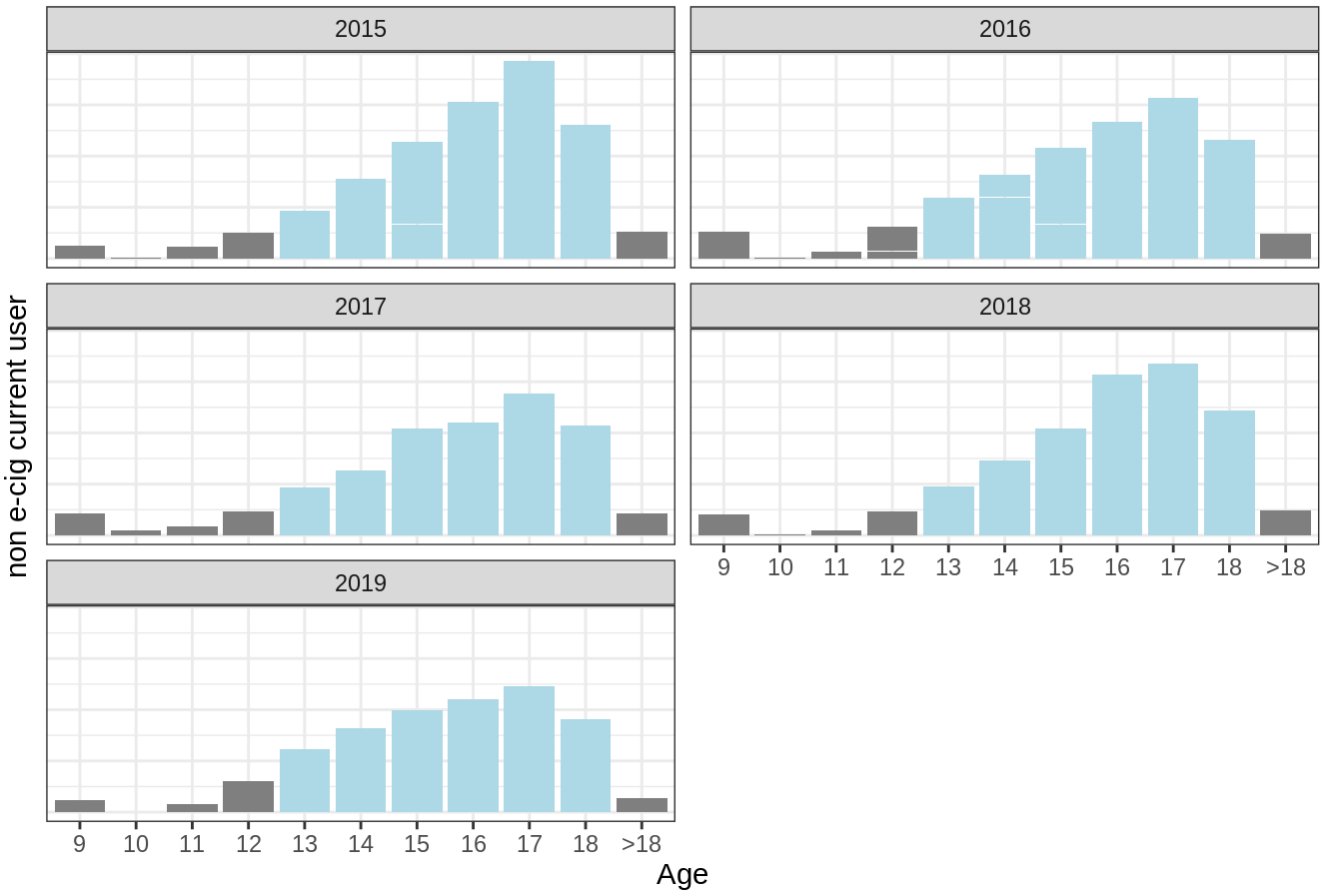
```
nyts_data <- nyts_data |> # relevel variable
  mutate(Age = fct_relevel(Age, "9","10","11","12","13",
                           "14","15","16","17","18",">18"))

nyts_data |>
  filter(!Age %in% c(NA)) |>
  ggplot(aes(x = Age, y = ecig_sum_current, fill = Age)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=c("13" = "lightblue",
                             "14" = "lightblue",
                             "15" = "lightblue",
                             "16" = "lightblue",
                             "17" = "lightblue",
                             "18" = "lightblue")) +
  facet_wrap(~ year, nrow = 3) +
  theme_bw() +
  labs(title = "Current e-cig usage is generally higher in the 13-18 year age group across
all years",
       y = "E-cig current user") +
  theme(legend.position = "none",
        plot.title.position = 'plot',
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```



We can see from the graph above, a trend of increased e-cig use climbing rapidly in tennagers between the ages of 13 to 18. We can also see a sharp increase in e-cig use in 2018 and 2019 compared to previous years with the same trend of increasing use up until the age of 18. We first see a major jump in e-cig usage around the age of 13. This could indicate that teenagers at that age start to have a much greater exposure to potential pressures that may cause them to try e-cigs. As we know about the addictive qualities of nicotine, it is very easy to see how the usage would only continue to grow from that age forward. Another thing to note is that while usage grows until around the age of 17, it sees a decline from ages 18 and onward. This could potentially indicate that as teenagers begin to reach adulthood, they start to better realize the health hazards that e-cigs can cause, and begin to decrease their usage as a result.

```
nyts_data |>
  filter(!Age %in% c(NA)) |>
  ggplot(aes(x = Age, y = non_ecig_sum_current, fill = Age)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=c("13" = "lightblue",
                             "14" = "lightblue",
                             "15" = "lightblue",
                             "16" = "lightblue",
                             "17" = "lightblue",
                             "18" = "lightblue")) +
  facet_wrap(~ year, nrow = 3) +
  theme_bw() +
  labs(title = "Non e-cigarette usage(tobacco and other) in teenagers by year",
       y = "non e-cig current user") +
  theme(legend.position = "none",
        plot.title.position = 'plot',
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```

Non e-cigarette usage(tobacco and other) in teenagers by year



Now, when looking at non e-cig usage we see a similar increasing trend from the ages of 13-18, however, the total usage seems to have decreased from 2015 to 2019. This is presumably due to the increased popularity of e-cigs as an alternative to tobacco and other products. Again, we notice the most drastic increase in usage begin at the age of 13, then continue to grow until the age of 17 where it then drops of by 18 and beyond. Just like e-cig usage, non-ecig usage shows the same relationship of growth by age beginning at 13 when presumably teenagers see an increased exposure and pressure to try non-ecig/tobacco products.

Results

Q1: Our analyses have shown that both ever and current tobacco users have decreased from 2015 to 2017 and increased from 2017 to 2019. Both ever and current e-cig users have decreased slightly from 2015 to 2017, while ecig current and ever usage increased significantly from 2017 to 2019 in American youths. Overall, users of e-cigarettes and tobacco products increased after 2017. Besides, users of e-cigarettes only have surpassed users of other tobacco products only in 2019, which might suggest that e-cigarettes have become more popular than any other tobacco products among youths during this period.

Q2: Moreover, our analysis shows that male students tend to have a higher usage of e-cigarette by about 5% compared to female students, though female users increased at a faster rate than male users from 2017 to 2019. Overall, both male and female users follow a similar trend: they decreased from 2015 to 2017 and increased from 2017 to 2019.

Q3: Fruit is the most popular vaping flavor, and menthol is the second popular flavor. With the data we have about vaping brand, JUUL appears to be used most frequently.

Q4: We found that while e-cigarette use increased rapidly after 2017, cigarette use largely follows a decreasing trend from 2015 to 2019. This indicates e-cigarettes became more popular among youths than cigarettes after 2017. After running a logistic regression model that takes survey weights into account, we found that the odds of being a current e-cigarette user for a current user of other tobacco products is 1258.9% higher than the odds for a user who does not currently use other tobacco products.

Q5: e-cig use has increased since 2017, especially for students aged between 13-18. We could see a dramatic increase in vaping among these age groups. There are a lot more young adults in late middle school and high school who are using e-cig in recent years. On the other hand, non e-cig use has decreased slightly for after 2017 but the decrease is not as sharp as the increase in vaping.

Conclusion

Based on the analysis and visualization of the National Youth Tobacco Survey data, it is clear that the use of e-cigarettes has become a concerning issue among young American students in middle school and high school. Our findings suggest that male students tend to use e-cigarettes more frequently than female students. Despite this difference, the sharp increase in e-cigarette use after 2017 in both groups, is a trend that requires immediate attention from public health authorities and policymakers. The popularity of fruity, menthol, and candy/dessert/sweet flavors, as well as the use of JUUL products, among young students is also a matter of concern. The availability of these attractive flavors and products could be a major factor contributing to the increase in e-cigarette use among young students. This use is particularly high among students aged 13-18. Prevention needs to occur as early as possible given the growth of use only increases from 13 onward. Additionally, the positive relationship between current e-cigarette use and current other tobacco product use suggests that e-cigarette use could be a gateway to other forms of tobacco use among young students. This finding highlights the need for comprehensive tobacco prevention and cessation programs that address all forms of tobacco use, including e-cigarettes. In conclusion, the results of this analysis emphasize the urgent need for public health authorities and policymakers to take action to prevent and reduce e-cigarette use among young students in the United States. This could include increasing regulations and restrictions on the marketing, sale, and use of e-cigarettes and related products, as well as implementing comprehensive tobacco prevention and cessation programs that address all forms of tobacco use.