Homework 1

Netasha

Homework 1

For this assignment, you may use R or Python, or a combination of both, to complete both case studies. You can use the code that we include in Labs 1 and 2 to answer these questions. You also may need to use other functions. I encourage you to make use of our textbook(s) and use the Internet to help you solve these problems. You can also work together with your classmates. If you do work together, you should provide the names of those classmates below.

Names of Collaborators (if any):

Case Study: New York Times Ad Impressions (Simulated)

There are 10 data sets in the `/data` subdirectory named `nyt1.csv`, `nyt2.csv`, ..., `nyt10.csv`. Each file represents one day's worth of simulated data on ad impressions and clicks on the New York Times homepage. Each row represents a single user. There are five columns:

- `Age` (user's age)
- `Gender` (user's gender, coded as `0` = female, `1` = male)
- `Impressions` (number of ads displayed during the user's visit)
- `Clicks` (number of clicks made by the user)
- `Signed_In` (whether or not the user was signed in as a member)

Packages I used

```
library(readr)
library(tidyverse)
```

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

```
library(psych)
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
library(cowplot)
```

```
Attaching package: 'cowplot'

The following object is masked from 'package:lubridate':

    stamp
```

```
library(gmodels)
```

Load **a single data file**. Then do the following.

```r
#this is me uploading day 7 of simulated data
nyt7 <- read_csv("data/nyt7.csv")
```

```
Rows: 452493 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Age, Gender, Impressions, Clicks, Signed_In

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.
```

1. Create a new variable, **age_group**, that categorizes users into the following age groups: $<$ 18, 18-24, 25-34, 35-44, 45-54, 55-64, and 65+.

```r
#this is me using the cut function to create 7 age group
nyt7$age_group<-cut(nyt7$Age,
             breaks = c(-1, 17,24,34,44,54,64,112),
            labels = c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))

# Print the result
print(table(nyt7$age_group))
```

```
   <18   18-24   25-34   35-44   45-54   55-64     65+
149474   39873   57018   69284   63543   44573   28728
```

2. Plot the distributions of impressions and "click-through rate" for all 6 age categories. (*Note:* Click-through rate is defined as the number of clicks divided by the number of impressions; it represents the proportion of ads that generated clicks.)

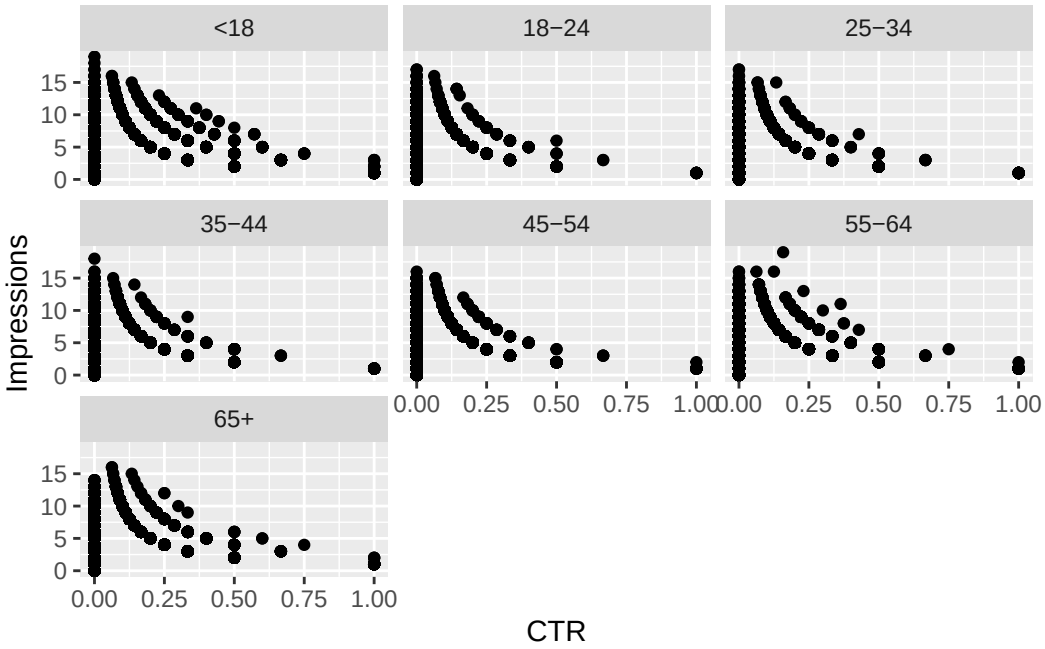
```r
summary(nyt7$Impressions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       3       5       5       6      19
```

```r
#this is me creating another variable for click through rate called CTR
nyt7$CTR <- (nyt7$Clicks/nyt7$Impressions)
nyt7$CTR[is.nan(nyt7$CTR)]<-0
#this is trying to determine the number of NAs using the summary function
summary(nyt7$CTR)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.0183  0.0000  1.0000
```

```
# this is me creating distributions of impressions and click through rate for all 6 age categ
ggplot(nyt7,
 aes(x = CTR, y = Impressions))+
   geom_point()+
   facet_wrap(~ age_group)
```



3. Create a new variable to categorize users based on their click behavior. (The name and categories for this variable are up to you. Explain what decision[s] you make and why.)

```
#this is me reviewing click information using the summary function
summary(nyt7$Clicks)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.09253 0.00000 4.00000
```

```
psych::describe(nyt7$Clicks)
```

```
   vars      n mean   sd median trimmed mad min max range skew kurtosis se
X1    1 452493 0.09 0.31      0       0   0   0   4     4  3.5    13.32  0
```

```r
#this is me requesting frequencies
table(nyt7$Clicks)
```

```
     0      1      2      3      4
413264  36755   2319    143     12
```

```r
nyt7$click_group<-cut(nyt7$Clicks,
             breaks = c(-1, 0,4),
            labels = c("no clicks", "clicks"))
# Print the result
print(table(nyt7$click_group))
```

```
no clicks    clicks
   413264     39229
```

The best approach to categorize participants based on their click behavior would be dichotomize observations into two groups given the summary statistics, which suggest that the max number of clicks is 4 and minimum is 0, and frequencies of the number of clicks. Leaving groups in 5 categories would be difficult to analyze as each there is not enough variation across the sample.

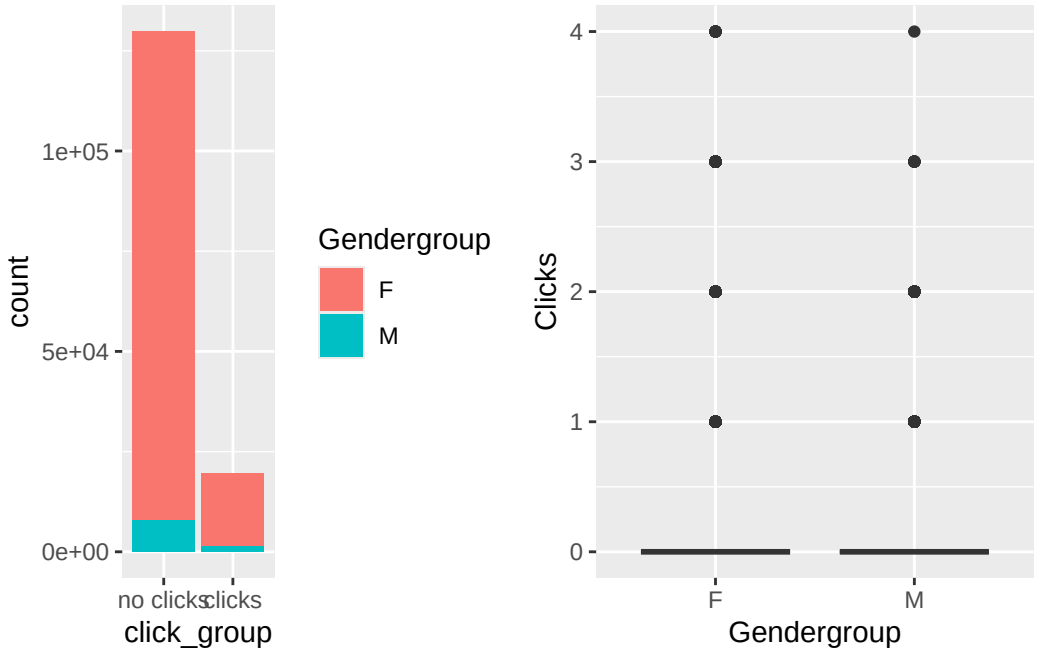
4. Explore the data and make visual and quantitative comparisons across user segments/demographics to answer the following:

    - How do <18 year old males differ from <18 year old females in terms of clicks and impressions?

```r
#this is me creating a new variable to dictomize age
nyt7$age_dichot<-cut(nyt7$Age,
             breaks = c(-1, 17,112),
            labels = c("<18", ">18"))

 # Print the result
print(table(nyt7$age_dichot))
```

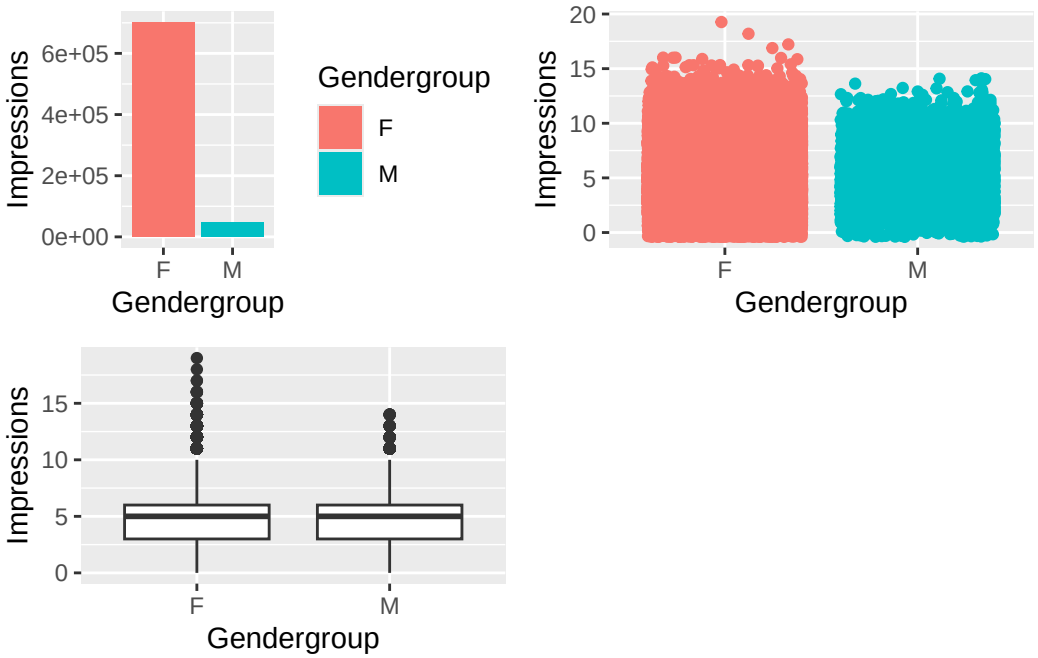
```
   <18    >18
149474 303019
```

```
#this is me creating a factor for gender and adding it to the dataset
nyt7$Gendergroup<-factor(nyt7$Gender)
levels(nyt7$Gendergroup) <- c("F", "M")

#this is me creating a bar graph and box ploots to visually see how males and females di
  Pc1 <-  nyt7 %>% filter(Age <18) %>%
ggplot +
geom_bar(aes(x = click_group, fill = Gendergroup))

   Pc2 <- nyt7 %>% filter(Age <18) %>%
   ggplot(
 aes(x = Gendergroup, y =  Clicks))+
  geom_boxplot()
  plot_grid(Pc1, Pc2, align = "h", rel_widths = c(1, 1))
```

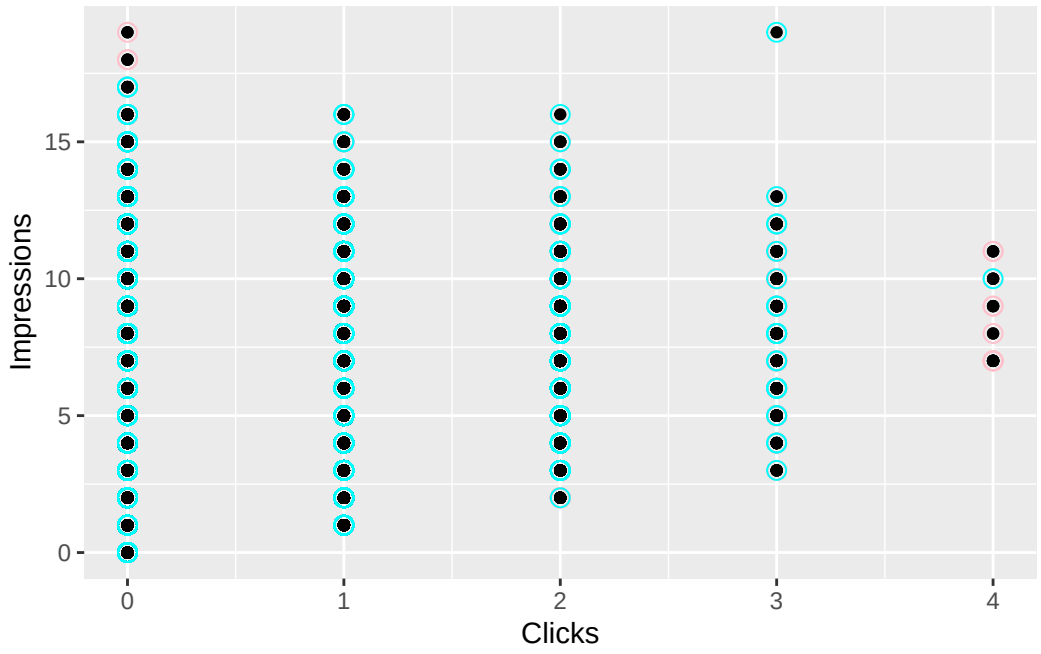



```
#this is me creating a bar graph and box plots to visually see how males and females dif

p1<- nyt7 %>% filter(Age <18) %>%
ggplot(aes(x = Gendergroup, y = Impressions, fill= Gendergroup)) +
geom_bar(stat = "identity")

p2<- nyt7 %>% filter(Age <18) %>%
ggplot(aes(x = Gendergroup, y = Impressions, color = Gendergroup))+
  geom_jitter() +
```

```
theme(legend.position = "none")

p3<- nyt7 %>% filter(Age <18) %>%
   ggplot(
 aes(x = Gendergroup, y =  Impressions))+
  geom_boxplot()

  plot_grid(p1, p2,p3, align = "h", rel_widths = c(1, 1))
```




```
#this is me ploting males and females under the age of 18 on clicks and impressions

ggplot(nyt7, aes(x = Clicks, y = Impressions)) +
 geom_point() +
 geom_point(
   data = nyt7 |> filter(age_dichot == "<18")
 ) +
 geom_point(
   data = nyt7 |> filter(Gender == "0"),
   shape = "circle open", size = 3, color = "pink")+
 geom_point(
   data = nyt7 |> filter(Gender == "1"),
   shape = "circle open", size = 3, color = "cyan")
```




```
 #this is me running quantitative comparisons
 #This is me running an anova
 under18<- subset(nyt7, age_dichot=='<18')

res_aov <- aov(Impressions ~ Gendergroup,
  data = under18
)
summary(res_aov)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
Gendergroup      1      5   4.944   0.989   0.32
Residuals   149472 747306   5.000
```

```
  #this is me looking at the tabulation of clicks for gender and individuals under 18
table_1 <- xtabs(~ as.factor(Clicks) + Gendergroup+ age_dichot, data=nyt7)
table_1
```

```
, , age_dichot = <18

                 Gendergroup
as.factor(Clicks)      F      M
                0 121816   7977
                1  16791   1204
                2   1445    108
                3    113      9
```

```
                4     10      1

, , age_dichot = >18

                 Gendergroup
as.factor(Clicks)      F      M
                0 142623 140848
                1   9852   8908
                2    419    347
                3      6     15
                4      1      0
```

```
#determining proportions and percentage  of clicks per each gendeer group
   round(prop.table(table_1,1), 2)
```

```
, , age_dichot = <18

                 Gendergroup
as.factor(Clicks)    F    M
                0 0.29 0.02
                1 0.46 0.03
                2 0.62 0.05
                3 0.79 0.06
                4 0.83 0.08

, , age_dichot = >18

                 Gendergroup
as.factor(Clicks)    F    M
                0 0.35 0.34
                1 0.27 0.24
                2 0.18 0.15
                3 0.04 0.10
                4 0.08 0.00
```

```
100*(round(prop.table(table_1,1), 2))
```

```
, , age_dichot = <18

                 Gendergroup
as.factor(Clicks)  F  M
                0 29  2
                1 46  3
```

```
                2 62  5
                3 79  6
                4 83  8

, , age_dichot = >18

                 Gendergroup
as.factor(Clicks)  F  M
                0 35 34
                1 27 24
                2 18 15
                3  4 10
                4  8  0
```

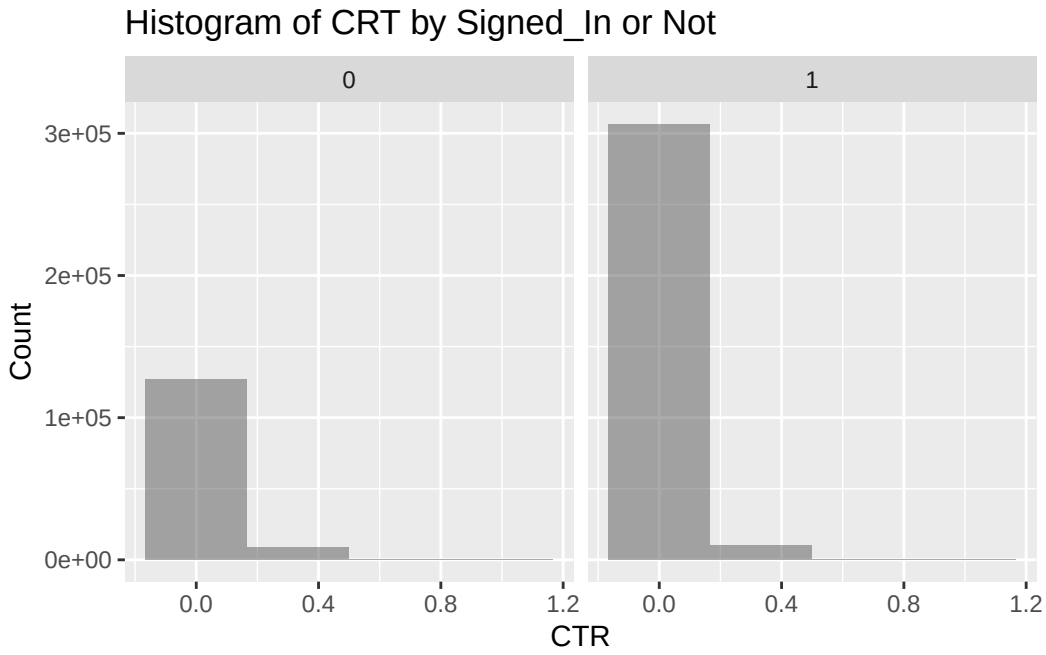
```
  stu_data = data.frame(nyt7$Clicks, nyt7$Age,nyt7$Gender)
  stu_data<- stu_data %>%
  filter(nyt7.Age <18)
   print(chisq.test(stu_data$nyt7.Clicks, stu_data$nyt7.Gender))
```

```
Warning in chisq.test(stu_data$nyt7.Clicks, stu_data$nyt7.Gender): Chi-squared
approximation may be incorrect


    Pearson's Chi-squared test

data:  stu_data$nyt7.Clicks and stu_data$nyt7.Gender
X-squared = 9.9257, df = 4, p-value = 0.0417
```

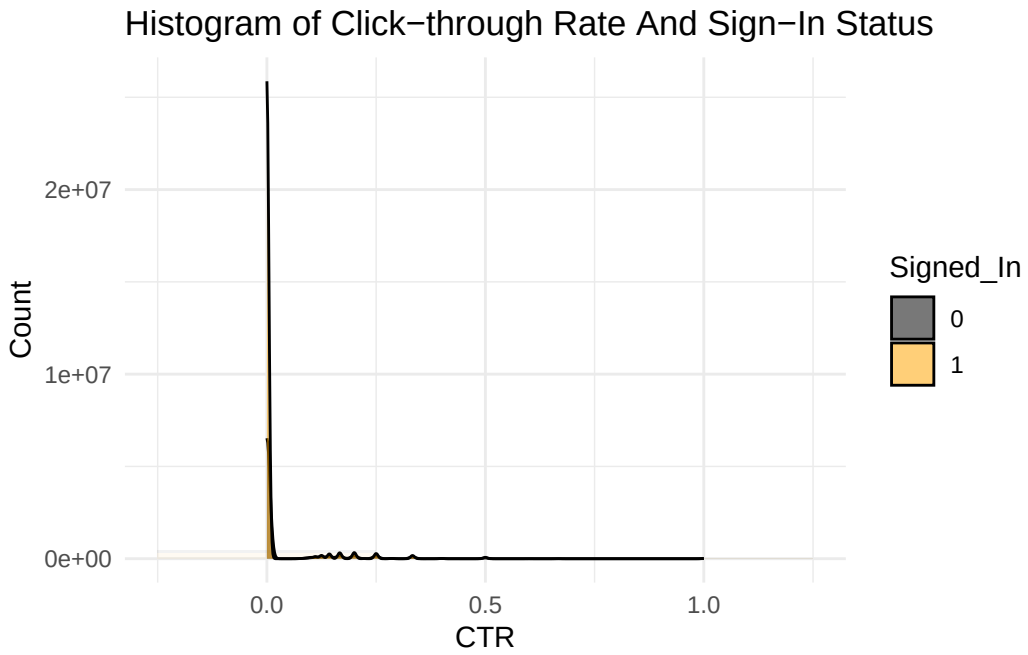
- Through a visual inspection of bar graphs and boxplots the relation between gender and the number of impression for adolescents and children (age <18) does not appear to differ. An analysis of varience further confirms that that no difference we detected. A visual inspection of bar graph and box plots of the relation between clicks and gender for adolescents and children may be unique such at the bar graph suggests that more females than males engage in clicks. A chi-square revealed that clicks significantly differ across gender
- How does the distribution of click-through rate for users who are signed in differ from the distribution for those who are **not** signed in?

```
#this is me creating two histograms displaying click-through rate for signed-in and sign
p4<- nyt7 %>%
  ggplot(aes(x = CTR)) +
  geom_histogram(bins = 4, alpha = 0.5) +
  labs(title = "Histogram of CRT by Signed_In or Not",
```

```
   x = "CTR",
   y = "Count")+
facet_wrap(~ Signed_In)
#this is me printing the plot
p4
```




```
#this is me making Signed_in a factor
nyt7$Signed_In<-factor(nyt7$Signed_In)

#This is me printing the two distributions on top of each other
nyt7 %>%
  ggplot(aes(x = CTR, fill = Signed_In)) +
  geom_histogram(bins = 3, alpha = 0.05) +
  labs(title = "Histogram of Click-through Rate And Sign-In Status",
   x = "CTR",
   y = "Count",
   fill = "Signed_In") +
  theme_minimal() +
  geom_density(aes(y = after_stat(count)), position = "identity", alpha = 0.5) +
  scale_fill_manual(values=c("black", "orange"))
```



- To better understand the relation between click-through rate and signed-in status distributions were created. The distributions suggests that a click-through rate of zero was visibly more frequent in the signed-in condition incomparison to the not signed in conditions.

- Are certain age groups more likely to be signed in than others? Which ones?

```
#this is me creating Signed-in into a factor
nyt7$FS_I<- as.factor(nyt7$Signed_In)

#this is me creating a table of both variables
stu_data = table(nyt7$age_group, nyt7$FS_I)
stu_data
```

```
            0      1
  <18   135670  13804
  18-24      0  39873
  25-34      0  57018
  35-44      0  69284
  45-54      0  63543
  55-64      0  44573
  65+        0  28728
```

```
print(chisq.test(stu_data))
```

```
    Pearson's Chi-squared test

data:  stu_data
X-squared = 392811, df = 6, p-value < 2.2e-16
```

- A cursory glance of the data using a table suggest that children and adolescents were the only group that both signed_in or no signed in as a member. All other age_groups were signed in as members. A pearson chi-square futher suggest that signed_in status is dependent on age_group status $x^2(6) = 392811, p < .001$

5. Calculate summary statistics for the click-through rate. These should include (1) quartiles, (2) mean, (3) median, (4) minimum and maximum, and (5) variance. Choose two user segments to compare these statistics across (for example, compare the mean, median, and quartiles for users 25-34 to those for users 65+).

```
nyt7 %>%
  group_by(age_group) %>%
  summarise(Mean=mean(CTR), SD = sd(CTR), Variance = (sd(CTR)^2), InterQuartile = IQR(CTR), M
```

```
# A tibble: 7 x 9
  age_group    Mean     SD Variance InterQuartile Median Minimum Maximum na.rm
  <fct>       <dbl>  <dbl>    <dbl>         <dbl>  <dbl>   <dbl>   <dbl> <lgl>
1 <18       0.0283  0.0840  0.00706             0      0       0       1 TRUE
2 18-24     0.0107  0.0517  0.00268             0      0       0       1 TRUE
3 25-34     0.0100  0.0514  0.00264             0      0       0       1 TRUE
4 35-44     0.0102  0.0522  0.00273             0      0       0       1 TRUE
5 45-54     0.00979 0.0497  0.00247             0      0       0       1 TRUE
6 55-64     0.0194  0.0693  0.00481             0      0       0       1 TRUE
7 65+       0.0300  0.0871  0.00758             0      0       0       1 TRUE
```

```
#define quantiles of interest
q = c(.25, .5, .75)

#calculate quantiles by grouping variable
nyt7 %>%
  group_by(age_group) %>%
  summarize(quant25 = quantile(CTR, probs = q[1]),
            quant50 = quantile(CTR, probs = q[2]),
            quant75 = quantile(CTR, probs = q[3]))
```

```
# A tibble: 7 x 4
  age_group quant25 quant50 quant75
  <fct>       <dbl>   <dbl>   <dbl>
1 <18             0       0       0
2 18-24           0       0       0
3 25-34           0       0       0
4 35-44           0       0       0
5 45-54           0       0       0
6 55-64           0       0       0
7 65+             0       0       0
```

```
GenderT<- table(nyt7$Gendergroup)
GenderT
```

```
     F      M
293076 159417
```

```
AgeT<- table(nyt7$age_group)
AgeT
```

```
   <18  18-24  25-34  35-44  45-54  55-64    65+
149474  39873  57018  69284  63543  44573  28728
```

```
AgeT2 <- summary(nyt7$Age)
AgeT2
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00   31.00   29.52   48.00  112.00
```

```
FSI_T<- table(nyt7$FS_I)
FSI_T
```

```
     0      1
135670 316823
```

6. Summarize your findings in a brief (1-2 paragraph) report intended for a New York Times (NYT) advertising team.

- Visual and statistical analysis were used to explore the data related to ad impressions and clicks during wave seven. Of primary concern was click-through rate across different demographic variables (e.g., age, gender, and member status), which is defined as the proportion of ads that are generated given the number clicks. There were a total 452493 observations during this epoch with 293076 identifying as females and 159417 identifying as males. On average participants were 29.50 years of age with the minimum of 0-years and 112-years old. Notably, the majority of observations were under the age of 18 (n = 149474). On average participants clicked .09 (SD =0.31) times and 5 ad impressions were generated. On average, .018 (sd = 0) ads were generated per click. Visual aides, such as bargraphs and box plots revealed that age and gender played a role in clicks, impressions, and click through-rate. For instance, a bar graph showed that more females than males did not click and those that did click were mostly female in participants under the age of 18. A pearson chi-square revealed that the relationship was significant $X^2(4) = 9.9257, p-value = 0.0417$.Commensurate, more female children and adolescents than males were exposed to ads. However, an analysis of variance suggested that the differences were not significantly different. A further evaluation of the data revealed, that membership status was related to age group, $X^2(6) = 392811, p - value < .001$, such that fewer children and adolescents were members of the sight. In addition to children and adolescents, older adults (65+) also had on average higher click-through rate.

Additional Questions for 231 Students

*Now read in at least **three to four more** of these data files and extend your analyses.*

```
nyt8 <- read_csv("~/Library/CloudStorage/GoogleDrive-nkpizano@ucsb.edu/My Drive/1. Fall 2024/
```

```
Rows: 463196 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Age, Gender, Impressions, Clicks, Signed_In

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.
```

```
nyt9 <- read_csv("~/Library/CloudStorage/GoogleDrive-nkpizano@ucsb.edu/My Drive/1. Fall 2024/
```

```
Rows: 459472 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Age, Gender, Impressions, Clicks, Signed_In
```

```
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.
```

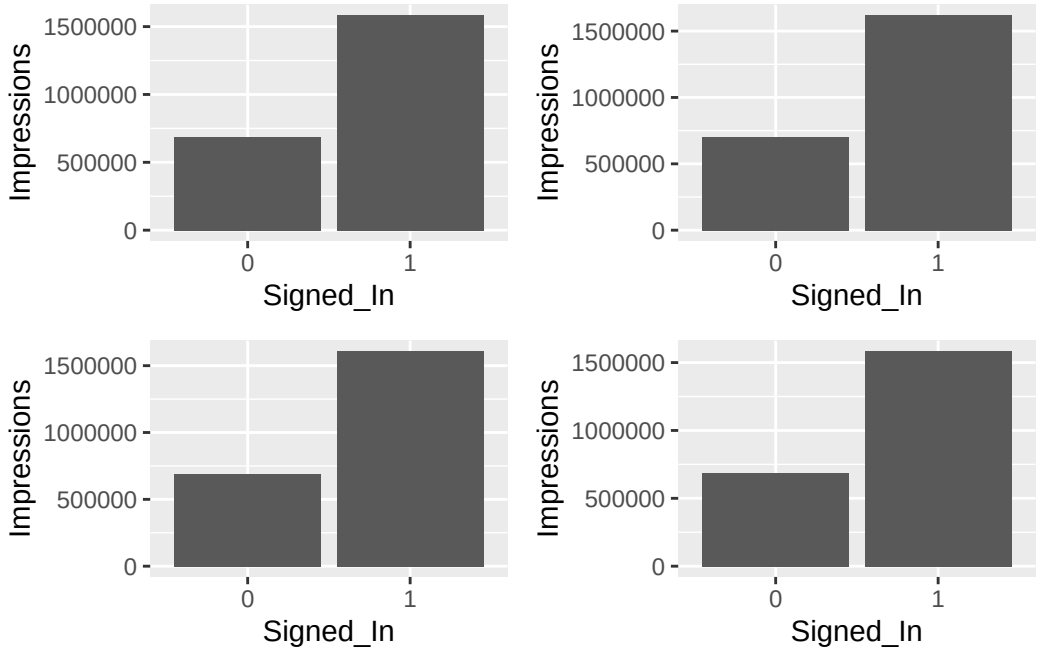
```
nyt10 <- read_csv("~/Library/CloudStorage/GoogleDrive-nkpizano@ucsb.edu/My Drive/1. Fall 202
```

```
Rows: 452766 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Age, Gender, Impressions, Clicks, Signed_In

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.
```

7. Visualize impressions and click-through rate for signed-in versus signed-out users over time (ggplot).

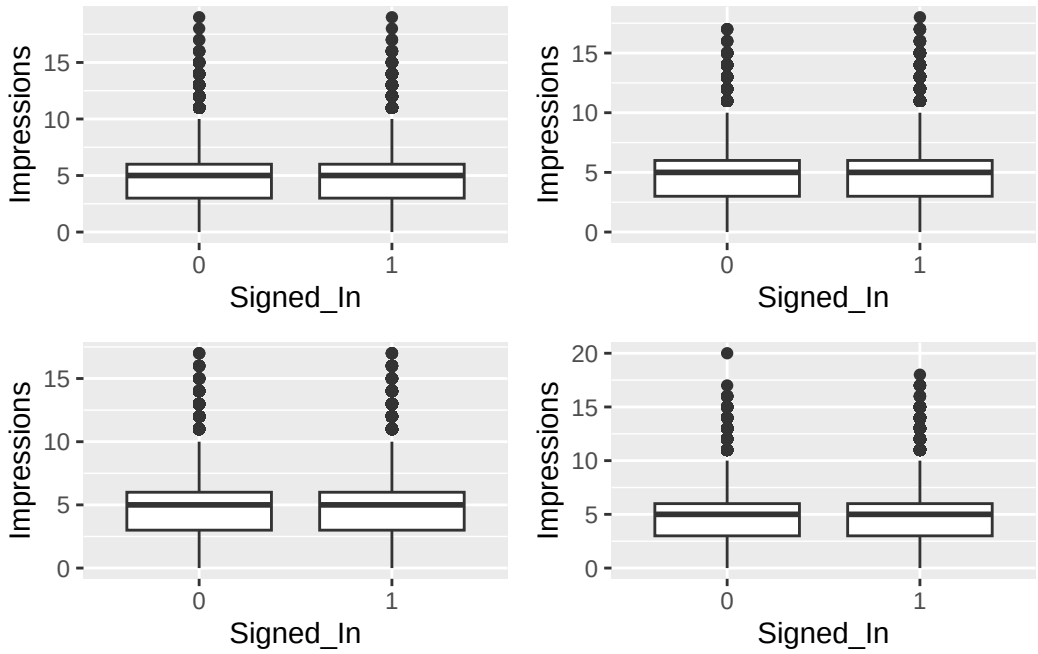
```
#this is me creating a factor of sign in membership status across waves
nyt10$Signed_In<-factor(nyt10$Signed_In)
nyt9$Signed_In<-factor(nyt9$Signed_In)
nyt8$Signed_In<-factor(nyt8$Signed_In)
nyt7$Signed_In<-factor(nyt7$Signed_In)

#this is me creating click through rate for waves 8,9,10
nyt10$CTR <- (nyt10$Clicks/nyt10$Impressions)
nyt10$CTR[is.nan(nyt10$CTR)]<-0
nyt9$CTR <- (nyt9$Clicks/nyt9$Impressions)
nyt9$CTR[is.nan(nyt9$CTR)]<-0
nyt8$CTR <- (nyt8$Clicks/nyt8$Impressions)
nyt8$CTR[is.nan(nyt8$CTR)]<-0

#Impressions across time for signed in vs. not


Imp7B <- nyt7 %>%
       ggplot(
     aes(x = Signed_In, y =  Impressions))+
       geom_bar(stat = "identity")
Imp8B <- nyt8 %>%
       ggplot(
     aes(x = Signed_In, y =  Impressions))+
       geom_bar(stat = "identity")
   Imp9B <- nyt9 %>%
```

```
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
     geom_bar(stat = "identity")
   Imp10B <- nyt10 %>%
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
      geom_bar(stat = "identity")

 plot_grid(Imp7B, Imp8B, Imp9B, Imp10B, align = "h", rel_widths = c(1, 1))
```




```
Imp7 <- nyt7 %>%
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
     geom_boxplot()
Imp8 <- nyt8 %>%
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
     geom_boxplot()
  Imp9 <- nyt9 %>%
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
     geom_boxplot()
   Imp10 <- nyt10 %>%
```

```
      ggplot(
    aes(x = Signed_In, y =  Impressions))+
     geom_boxplot()

plot_grid(Imp7, Imp8, Imp9, Imp10, align = "h", rel_widths = c(1, 1))
```



8. Calculate summary statistics to compare signed-in versus signed-out users over time.

```
Signed7<- table(nyt7$Signed_In)
Signed7<- round(prop.table(Signed7), 2)
Signed7<-100*(round(prop.table(Signed7), 2))
Signed7
```

```
 0  1
30 70
```

```
Signed8<- table(nyt8$Signed_In)
Signed8<-round(prop.table(Signed8), 2)
Signed8<-100*(round(prop.table(Signed8), 2))
Signed8
```

```
 0  1
30 70
```

```
Signed9<- table(nyt9$Signed_In)
Signed9<- round(prop.table(Signed9), 2)
Signed9<- 100*(round(prop.table(Signed9), 2))
Signed9
```

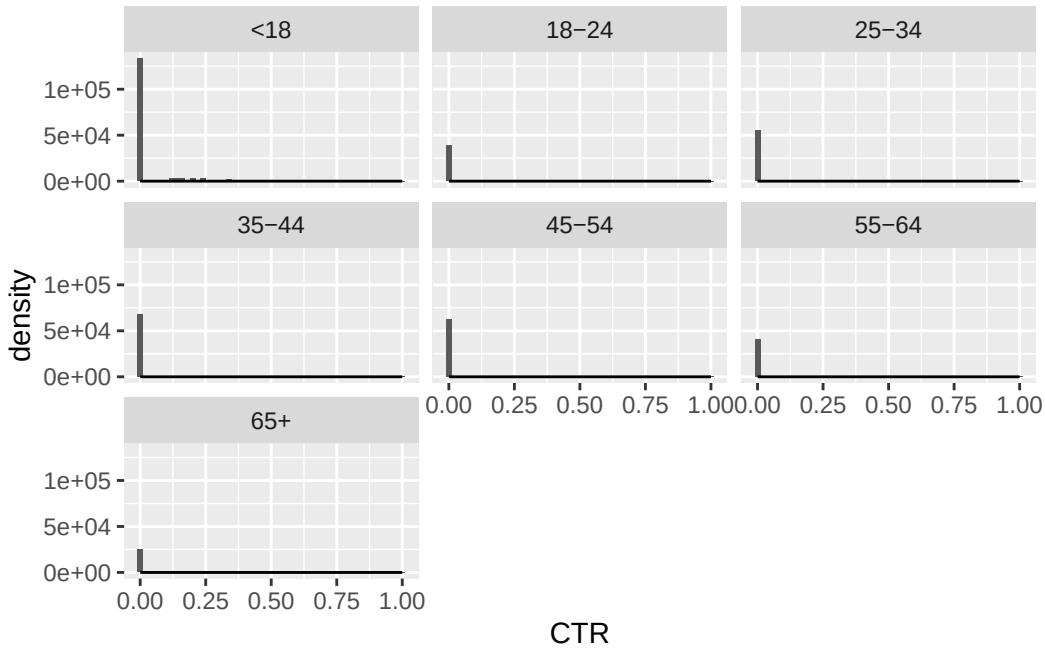
```
 0  1
30 70
```

```
Signed10<- table(nyt10$Signed_In)
Signed10<-round(prop.table(Signed10), 2)
Signed10<-100*(round(prop.table(Signed10), 2))
Signed10
```

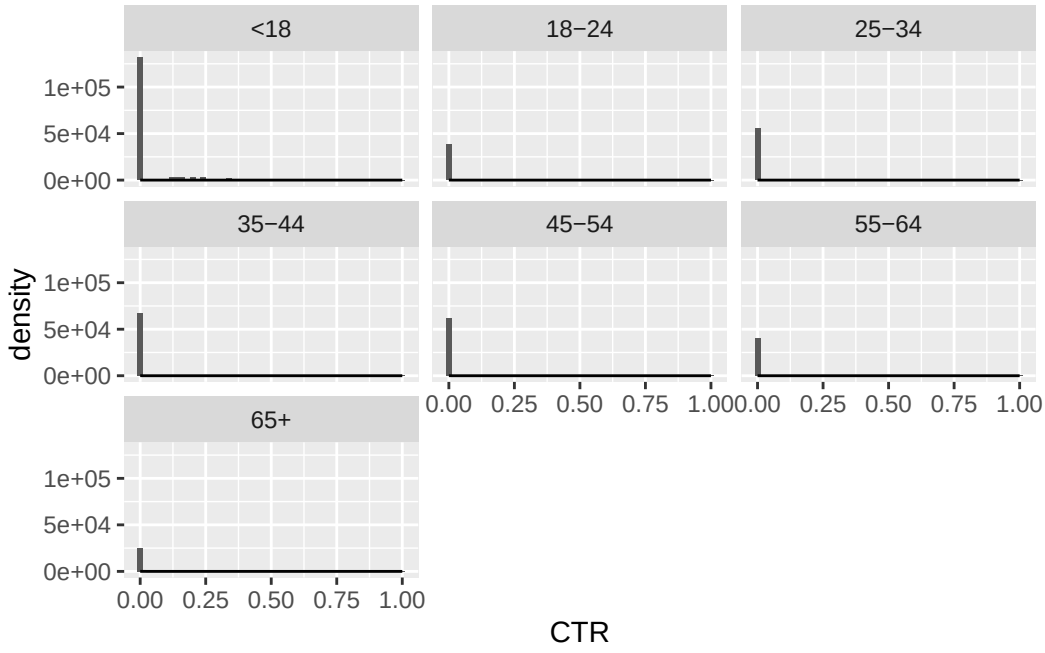
```
 0  1
30 70
```

9. Visualize click-through rate for the six different age groups over time.

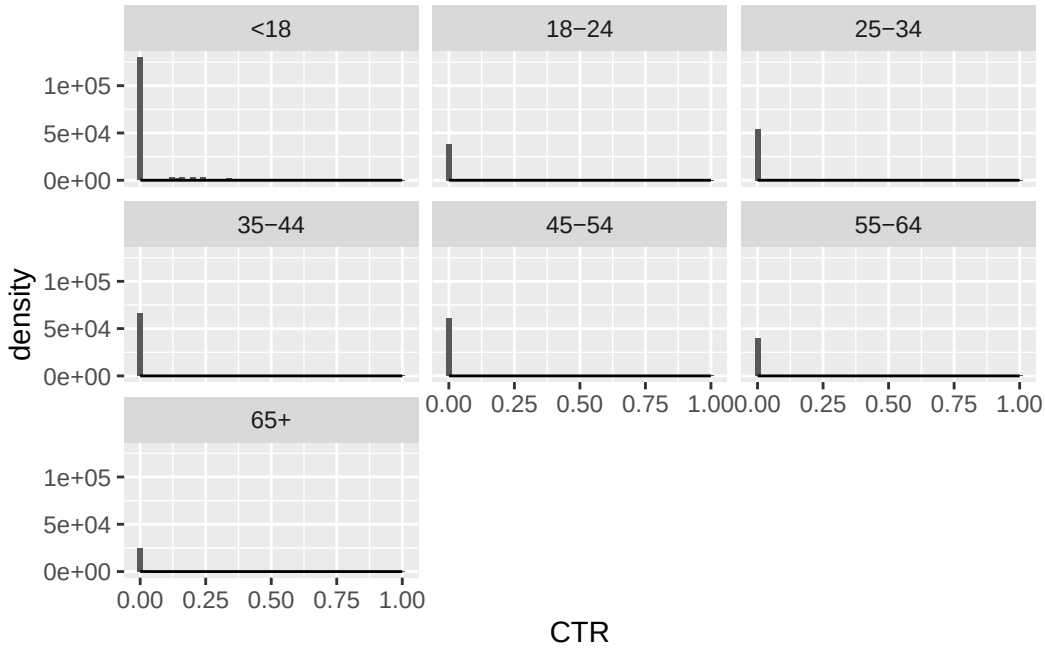
```
nyt8$age_group<-cut(nyt8$Age,
             breaks = c(-1, 17,24,34,44,54,64,112),
            labels = c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))
nyt9$age_group<-cut(nyt9$Age,
             breaks = c(-1, 17,24,34,44,54,64,112),
            labels = c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))
nyt10$age_group<-cut(nyt10$Age,
             breaks = c(-1, 17,24,34,44,54,64,112),
            labels = c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))
ggplot(nyt8,
   aes(x=CTR)) + geom_histogram(binwidth=.02)+
  geom_density()+
   facet_wrap(~ age_group)
```




```
ggplot(nyt9,
   aes(x=CTR)) + geom_histogram(binwidth=.02)+
  geom_density()+
   facet_wrap(~ age_group)
```



```
ggplot(nyt10,
   aes(x=CTR)) + geom_histogram(binwidth=.02)+
  geom_density()+
   facet_wrap(~ age_group)
```



Case Study: Social Media Engagement (Simulated)

The data file `Time-Wasters on Social Media.csv` contains a considerable amount of simulated data intended to mimic real-world social media usage scenarios. It comes from this source on Kaggle: https://www.kaggle.com/datasets/zeesolver/dark-web

Read through and familiarize yourself with the variables in the dataset. Then answer the following.

10. Produce a summary of the user data (the information about users: age, gender, location, debt, whether they own property, their profession). If you were asked to describe the "average user," what would you say?

```
Social_Media <- read_csv("~/Library/CloudStorage/GoogleDrive-nkpizano@ucsb.edu/My Drive/1. Fa
```

```
Rows: 1000 Columns: 31
-- Column specification --------------------------------------------------------
```

```
Delimiter: ","
chr  (12): Gender, Location, Profession, Demographics, Platform, Video Categ...
dbl  (16): UserID, Age, Income, Total Time Spent, Number of Sessions, Video ...
lgl   (2): Debt, Owns Property
time  (1): Watch Time

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.
```

```
Social_Media %>%
  summarise(Mean=mean(Age), SD = sd(Age), Minimum = min(Age), Maximum = max(Age),  na.rm = TR
```

```
# A tibble: 1 x 5
   Mean    SD Minimum Maximum na.rm
  <dbl> <dbl>   <dbl>   <dbl> <lgl>
1  41.0  13.5      18      64 TRUE
```

```
table(Social_Media$Gender)
```

```
Female   Male  Other
   322    514    164
```

```
table(Social_Media$Location)
```

```
       Barzil       Germany         India     Indonesia         Japan
           78            59           228            77            75
       Mexico      Pakistan   Philippines United States       Vietnam
           73            76            78           174            82
```

```
table(Social_Media$Debt)
```

```
FALSE  TRUE
  401   599
```

```
table(Social_Media$Profession)
```

```
      Artist      Cashier       driver     Engineer Labor/Worker
          47           56          113           65          186
     Manager     Students      Teacher Waiting staff
          54          246           39          194
```

```
table(Social_Media$`Owns Property`)
```

```
FALSE  TRUE
  458   542
```

- On aveage social media users in this dataset were 40.986 (SD = 13) years old with the youngest participants 18 years old and the more senior 64 years of age. Most social media users in this data set were from India (n = 228) and the US =174. Interestingly, social media users were in debt (n = 599),owned a home (n = 542), and students (n = 246)

11. What video categories are more popular with younger users (up to or below age 20)?

```
 Social_Media$age_dichot<-cut(Social_Media$Age,
                 breaks = c(-1, 19,64),
                labels = c("<20", ">20"))
smav_data = table(Social_Media$`Video Category`, Social_Media$age_dichot)
smav_data
```

```
               <20 >20
  ASMR           4  75
  Comedy         1  34
  Entertainment  8  94
  Gaming         3 116
  Jokes/Memes    8 171
  Life Hacks     5 157
  Pranks         7 103
  Trends         4  96
  Vlogs          9 105
```

- In participants that are younger than 20 three categories appear to be popular entertainment, jokes, and vlogs as the most popular.

What categories are more popular with older users (age 50 or above)?

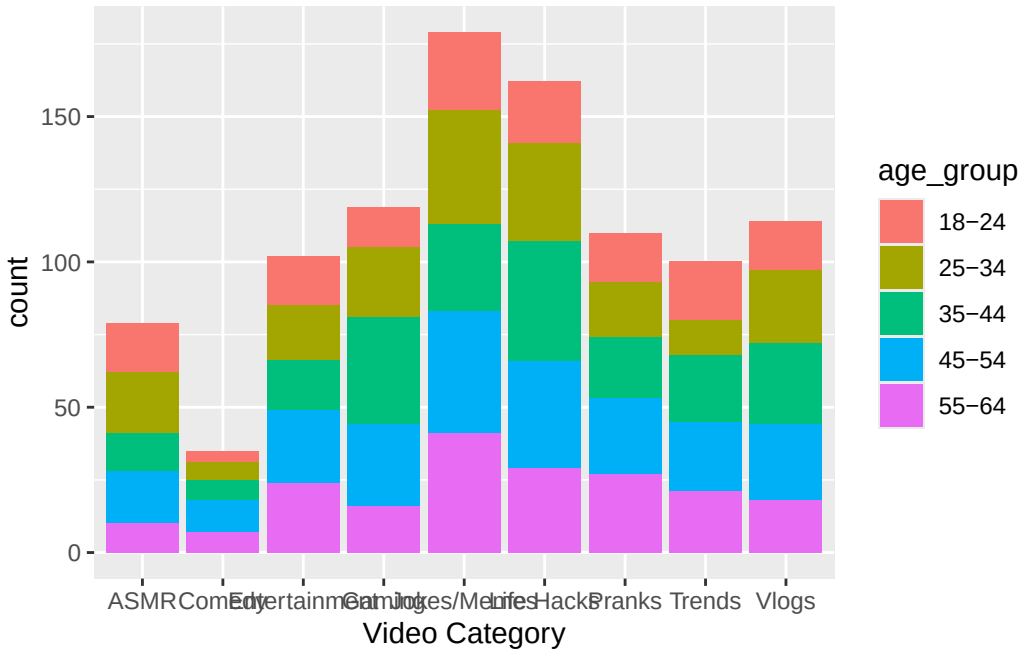
```
Social_Media$age_dichot_senior<-cut(Social_Media$Age,
                 breaks = c(-1, 49,64),
                labels = c("<50", ">50"))
smavsenior_data = table(Social_Media$`Video Category`, Social_Media$age_dichot_senior)
smavsenior_data
```

```
               <50 >50
  ASMR          63  16
  Comedy        23  12
  Entertainment 64  38
  Gaming        88  31
  Jokes/Memes  114  65
  Life Hacks   114  48
  Pranks        66  44
  Trends        67  33
  Vlogs         82  32
```

- In participants that are older than 50 it appears that jokes are the most popular video categories.

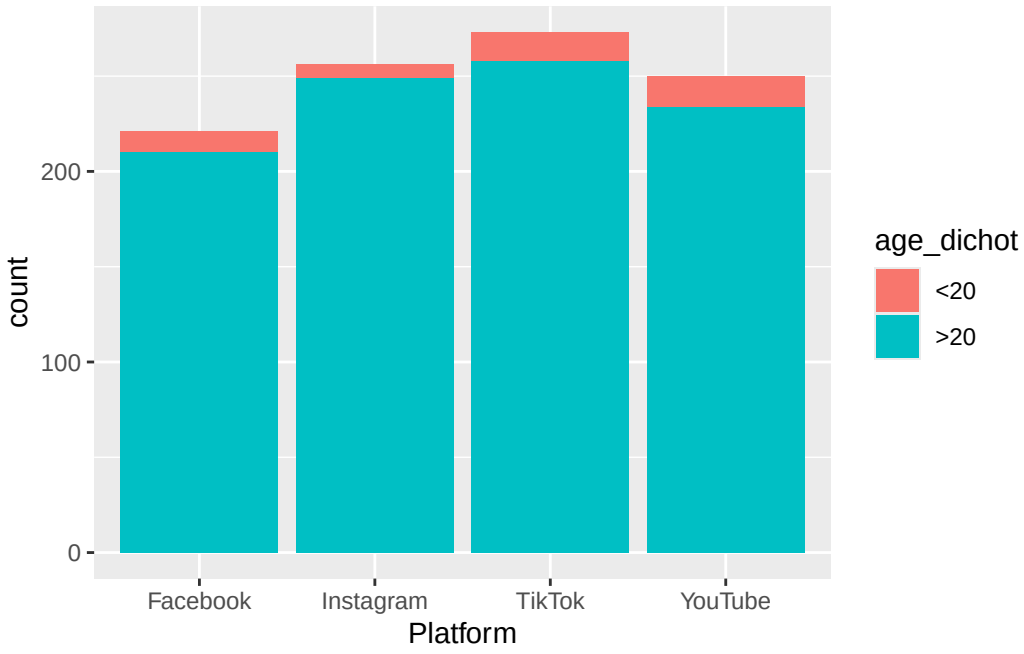
Create a plot or table of the distribution of video categories preferred by younger vs. older users.

```
Social_Media$age_group<-cut(Social_Media$Age,
             breaks = c(-1, 17,24,34,44,54,64),
            labels = c("<18", "18-24", "25-34", "35-44", "45-54", "55-64"))

 p1sm<- Social_Media %>%
    ggplot(aes(x = `Video Category`, fill= age_group)) +
    geom_bar()
 p1sm
```



12. What platforms are more popular with younger users (up to or below age 20)?

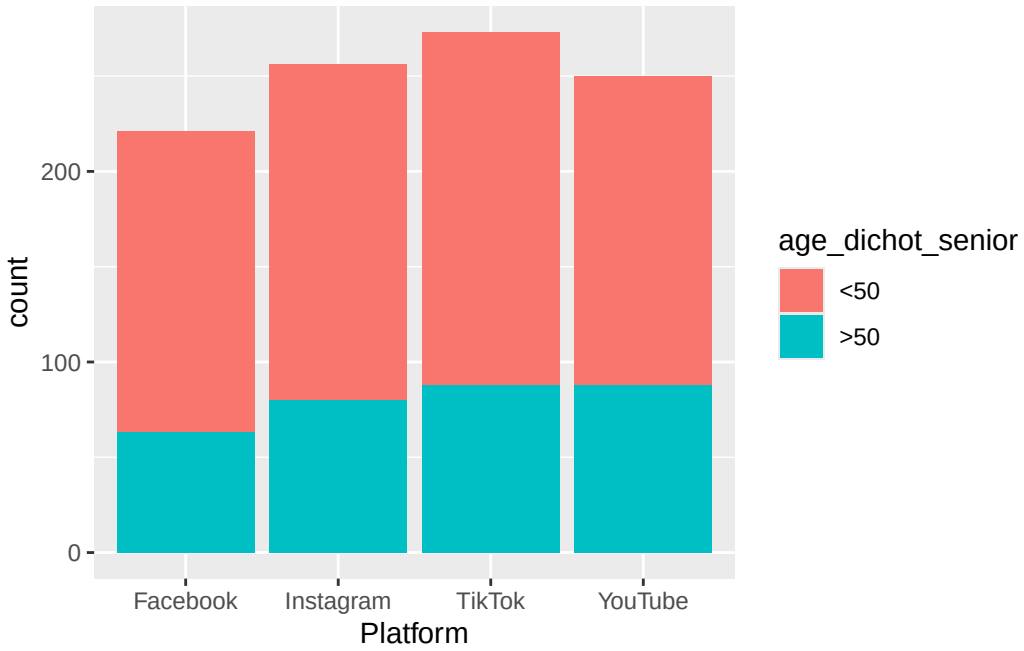
```
p2sm<- Social_Media %>%
   ggplot(aes(x = Platform, fill= age_dichot)) +
   geom_bar()
p2sm
```



- It appears that tiktok is the most popular platform for people under 20

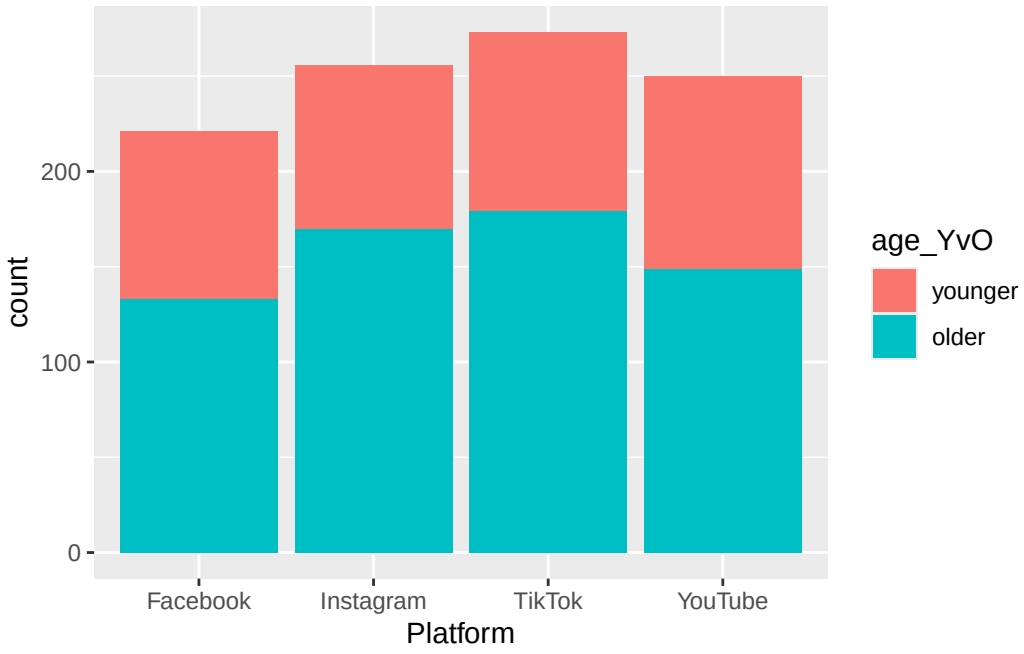
What platforms are more popular with older users (age 50 or above)?

```
p3sm<- Social_Media %>%
    ggplot(aes(x = Platform, fill= age_dichot_senior)) +
    geom_bar()
 p3sm
```



Create a plot or table of the distribution of platforms preferred by younger vs. older users.

```
Social_Media$age_YvO<-cut(Social_Media$Age,
             breaks = c(-1,35,64),
            labels = c("younger","older"))
p4sm<- Social_Media %>%
    ggplot(aes(x = Platform, fill= age_YvO)) +
    geom_bar()
 p4sm
```

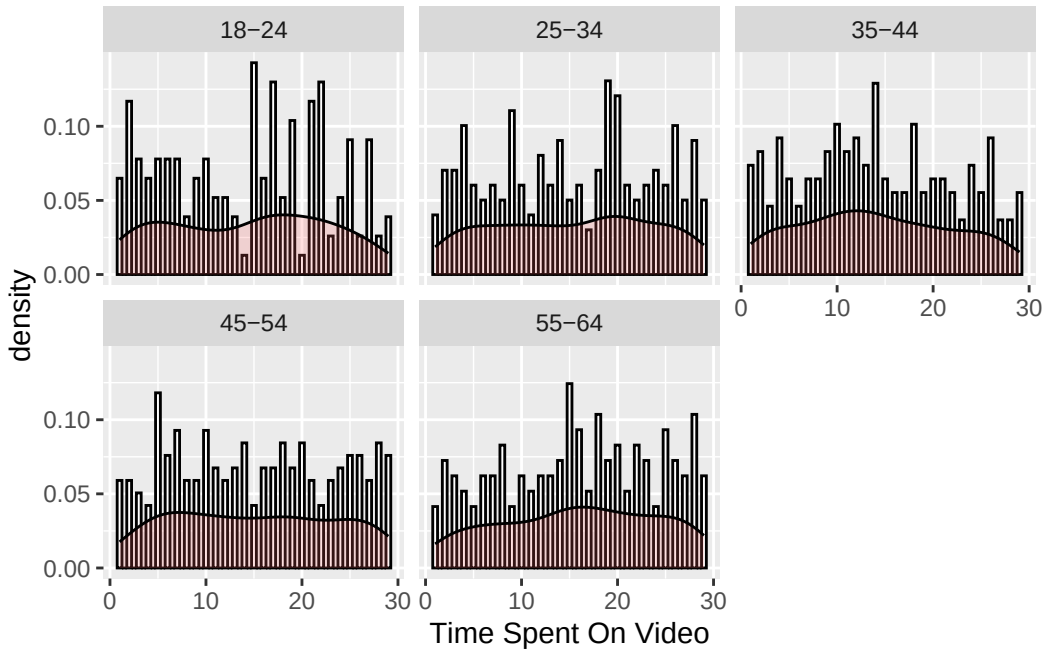


Additional Questions for 231 Students

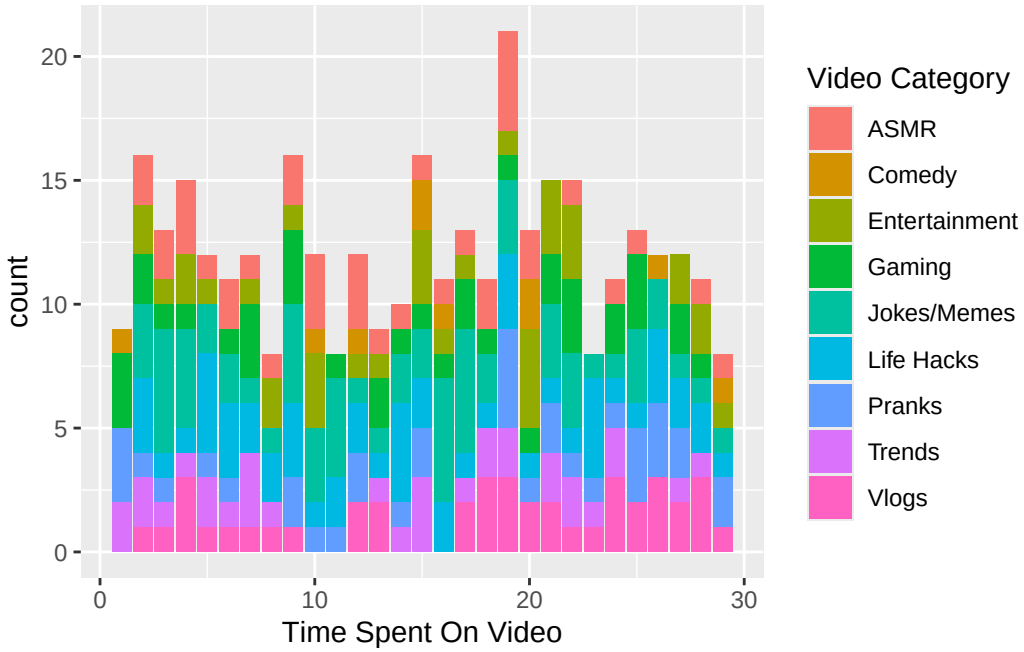
13. Explore the data. What are some patterns that you notice? Create one to two visualizations.

```
ggplot(Social_Media,
   aes(x=`Time Spent On Video`)) +
  geom_histogram(aes(y=..density..),
                   binwidth=.5,
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666")+
   facet_wrap(~ age_group)
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```




```
p6sm<- Social_Media %>%
  filter(Age <35) %>%
  ggplot(aes(x = `Time Spent On Video`, fill= `Video Category`)) +
  geom_bar()
 p6sm
```



14. Summarize your findings in a brief (1-2 paragraphs) report.

- In the previous visualizations, age, video categories, and time spent were evaluated.A histogram revealed that time spent on video appeared to differ, such that younger individuals (under 35) spend more time watching social media videos in comparison to 35 through 54 year olds. However, there is an increase in time spent watching videos for individuals between 55-64. Additionally, pranks,hack and ASMR are categories that were especially popular when younger individuals would watch an average amount of video.