data 607 project 2

Yina Qiao

2023-03-04

Introduction

The goal of this assignment is to practice in preparing different datasets for downstream analysis work

DataSet 1 NYC MTA Subway Ridership from 2013

load packages

```
# Load required libraries
library(tidyverse)
```

load data

```
url <- 'https://raw.githubusercontent.com/yinaS1234/data-607/main/project%202/Annual20Subway20Ridership.csv'

dfMTA <- read.csv(file = url)

glimpse(dfMTA)
```

```
## Rows: 434
## Columns: 10
## $ Station..alphabetical.by.borough. <chr> "The Bronx", "138 St-Grand Concourse…
## $ X2013                             <chr> "The Bronx", "957,984", "4,427,399",…
## $ X2014                             <chr> "The Bronx", "1,033,559", "4,536,888…
## $ X2015                             <chr> "The Bronx", "1,056,380", "4,424,754…
## $ X2016                             <chr> "The Bronx", "1,070,024", "4,381,900…
## $ X2017                             <chr> "The Bronx", "1,036,746", "4,255,015…
## $ X2018                             <chr> "The Bronx", "944,598", "3,972,763",…
## $ X2017.2018.Change                 <chr> "The Bronx", "-92,148", "-282,252", …
## $ X2017.2018.Change2                <chr> "The Bronx", "-8.9%", "-6.6%", "-2.4…
## $ X2018.Rank                        <chr> "The Bronx", "365", "121", "38", "16…
```

Tidy Data

```
# changing the column names
new_col_name <- c('Station', 2013, 2014, 2015, 2016, 2017, 2018, '2017 - 2018 Net Change',
                  '2017 - 2018 % Change', '2018 Rank')
colnames(dfMTA) <- new_col_name

# finding the rows where the boroughs are entered
borough <- c('The Bronx', 'Brooklyn', 'Manhattan', 'Queens')

rowvalues <- c()

for(i in 1:length(borough)){
  rowvalues[i] <- rownames(dfMTA[which(dfMTA$'2013' == borough[i]),])
}
rowvalues
```

```
## [1] "1"   "70"  "228" "350"
```

```
#now that we now where the boroughs dataset begins and ends, we can capture the
# data accordingly
dfBronx <- dfMTA[2:69,]
dfBronx['Borough'] <- borough[1]

dfBrooklyn <- dfMTA[71:227,]
dfBrooklyn['Borough'] <- borough[2]

dfManhattan <- dfMTA[229:349,]
dfManhattan['Borough'] <- borough[3]

dfQueens <- dfMTA[351:dim(dfMTA)[1],]
dfQueens['Borough'] <- borough[4]

# combined all sub datasets
dfMTA2 <- rbind(dfBronx, dfBrooklyn, dfManhattan, dfQueens)

# changed the columns from character to integer and removing commas
dfMTA2 <- dfMTA2 %>%
  mutate('2013' = as.integer(str_remove_all(dfMTA2$'2013', ',')),
         '2014' = as.integer(str_remove_all(dfMTA2$'2014', ',')),
         '2015' = as.integer(str_remove_all(dfMTA2$'2015', ',')),
         '2016' = as.integer(str_remove_all(dfMTA2$'2016', ',')),
         '2017' = as.integer(str_remove_all(dfMTA2$'2017', ',')),
         '2018' = as.integer(str_remove_all(dfMTA2$'2018', ',')),
         '2017 - 2018 Net Change' = as.integer(str_remove_all(dfMTA2$'2017 - 2018 Net Change', ',')),
         '2017 - 2018 % Change' = as.numeric(str_remove_all(dfMTA2$'2017 - 2018 % Change', '%')),
         '2018 Rank' = as.integer(dfMTA2$'2018 Rank')) %>%
  select(Borough, colnames(dfMTA2))
```

Data Analysis

Let's look at the data by boroughs.

```
# subset of the data we want to look at
colnames2 <- c('Borough', 2013, 2014, 2015, 2016, 2017, 2018)

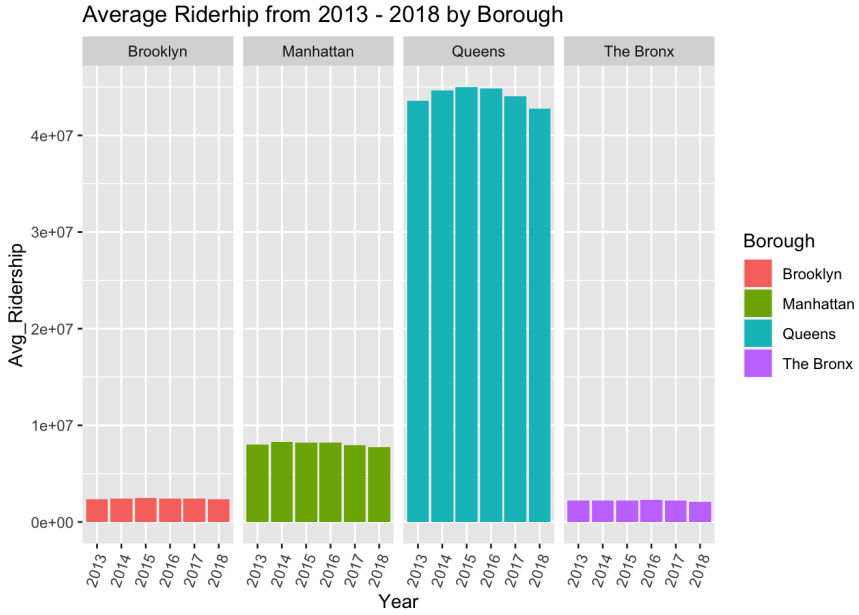
dfMTA3 <- dfMTA2 %>%
  select(colnames2)
```

```
## Note: Using an external vector in selections is ambiguous.
## ℹ Use `all_of(colnames2)` instead of `colnames2` to silence this message.
## ℹ See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
aggMTA <- dfMTA3 %>%
  pivot_longer(!Borough, names_to = 'Year', values_to = 'Ridership') %>%
  group_by(Borough, Year) %>%
  summarize(Avg_Ridership = mean(Ridership, na.rm = TRUE))
```

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

```
ggplot(data = aggMTA) +
  geom_bar(mapping = aes(x = Year, y = Avg_Ridership, fill = Borough), stat = 'identity') +
  facet_grid(~ Borough) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1)) +
  labs(title = 'Average Riderhip from 2013 - 2018 by Borough')
```



Conclusion for Dataset 1

There is only minor change for ridership by boroughs from 2013 - 2018. The Queens borough has the most riders among all other boroughs.

DataSet 2 MoviesOnStreamingPlatforms

I am curious to see which platform has better movies.

##load data

```
url <- 'https://raw.githubusercontent.com/yinaS1234/data-607/main/project%202/MoviesOnStreamingPlatforms.csv'

dfMovies <- read.csv(file = url)
glimpse(dfMovies)
```

```
## Rows: 16,744
## Columns: 17
## $ X                <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, …
## $ ID               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,…
## $ Title            <chr> "Inception", "The Matrix", "Avengers: Infinity War", "…
## $ Year             <int> 2010, 1999, 2018, 1985, 1966, 2018, 2002, 2012, 1981, …
## $ Age              <chr> "13+", "18+", "13+", "7+", "18+", "7+", "18+", "18+", …
## $ IMDb             <dbl> 8.8, 8.7, 8.5, 8.5, 8.8, 8.4, 8.5, 8.4, 8.4, 8.3, 8.3,…
## $ Rotten.Tomatoes  <chr> "87%", "87%", "84%", "96%", "97%", "97%", "95%", "87%"…
## $ Netflix          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ Hulu             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ Prime.Video      <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, …
## $ Disney.          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ Type             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ Directors        <chr> "Christopher Nolan", "Lana Wachowski,Lilly Wachowski",…
## $ Genres           <chr> "Action,Adventure,Sci-Fi,Thriller", "Action,Sci-Fi", "…
## $ Country          <chr> "United States,United Kingdom", "United States", "Unit…
## $ Language         <chr> "English,Japanese,French", "English", "English", "Engl…
## $ Runtime          <int> 148, 136, 149, 116, 161, 117, 150, 165, 115, 153, 114,…
```

Tidy Data and data manipulation

```
# Cleaning
dfMovies <- dfMovies[,-1] %>%
  rename(Rotten_Tomatoes = Rotten.Tomatoes, Prime_Video = Prime.Video, Disney = Disney.) %>%
  mutate(Rotten_Tomatoes = as.integer(str_remove(Rotten_Tomatoes, '%')))

# Transforming - we need to identify the platforms where the movies can be streamed.
# I created a subset for each platform and then combined them after
dfNetflix <- dfMovies %>%
  filter(Netflix == 1) %>%
  select(Title, IMDb, Rotten_Tomatoes)
dfNetflix['Platform'] <- 'Netflix'

dfHulu <- dfMovies %>%
  filter(Hulu == 1) %>%
  select(Title, IMDb, Rotten_Tomatoes)
dfHulu['Platform'] <- 'Hulu'

dfPrime_Video <- dfMovies %>%
  filter(Prime_Video == 1) %>%
  select(Title, IMDb, Rotten_Tomatoes)
dfPrime_Video['Platform'] <- 'Prime_Video'

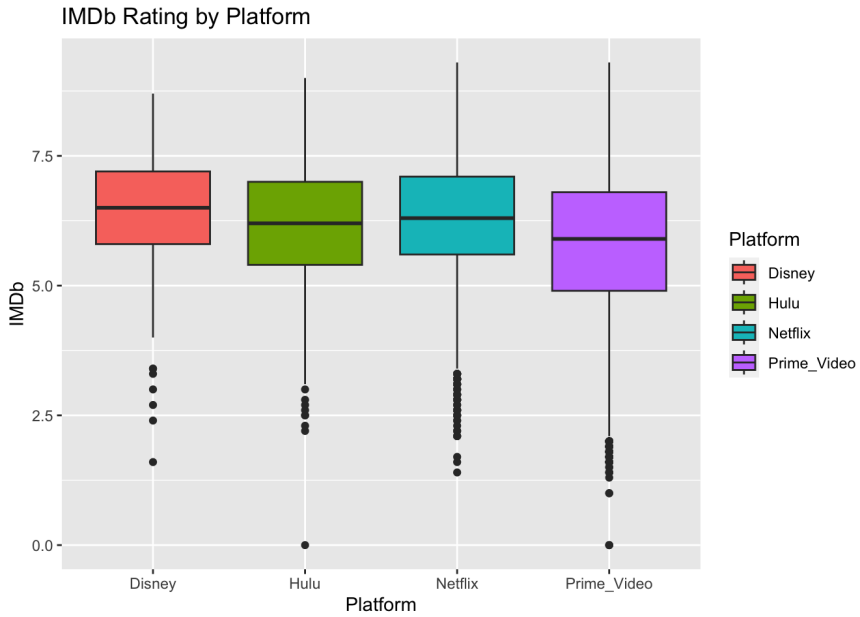
dfDisney <- dfMovies %>%
  filter(Disney == 1) %>%
  select(Title, IMDb, Rotten_Tomatoes)
dfDisney['Platform'] <- 'Disney'

dfMovies2 <- rbind(dfNetflix, dfHulu, dfPrime_Video, dfDisney)
```

Data Analysis

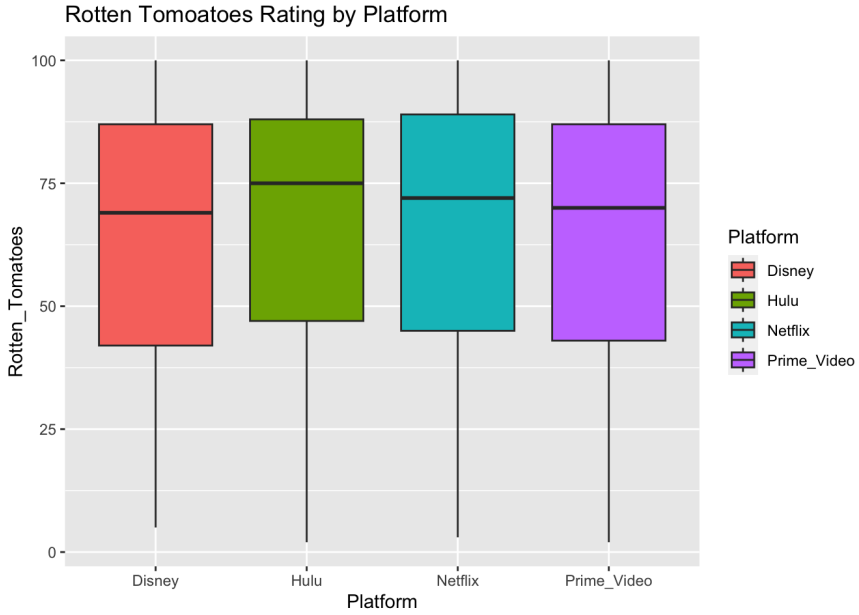
```
ggplot(data = dfMovies2, aes(x = Platform, y = IMDb, fill = Platform)) +
  geom_boxplot() +
  labs(title = 'IMDb Rating by Platform')
```

```
## Warning: Removed 576 rows containing non-finite values (`stat_boxplot()`).
```




```
ggplot(data = dfMovies2, aes(x = Platform, y = Rotten_Tomatoes, fill = Platform)) +
  geom_boxplot() +
  labs(title = 'Rotten Tomoatoes Rating by Platform')
```

```
## Warning: Removed 11895 rows containing non-finite values (`stat_boxplot()`).
```



Conclusion on Dataset 2

It seems like Rotton tomatoes rating has many NA and not enough data to support finding, where IMDb rating has less NA and therefore is a better indicator. Recommend to compare platforms based on IMDb rating, Disney has highest IMDb, while Prime-Video has lowest IMDb.

DataSet3 School Diversity

I am curious to see the difference racial average per school.

##load data

```
url <- 'https://raw.githubusercontent.com/yinaS1234/data-607/main/project%202/School_Diversity.csv'

dfSchool <- read.csv(file = url)
str(dfSchool)
```

```
## 'data.frame':    27944 obs. of  16 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ LEAID       : int  100002 100005 100005 100006 100006 100007 100007 100008 100011 100012 ...
##  $ LEA_NAME    : chr  "alabama youth services" "albertville city" "albertville city" "marshall county" ...
##  $ ST          : chr  "AL" "AL" "AL" "AL" ...
##  $ d_Locale_Txt: chr  NA "town-distant" "town-distant" "rural-distant" ...
##  $ SCHOOL_YEAR : chr  "1994-1995" "1994-1995" "2016-2017" "1994-1995" ...
##  $ AIAN        : num  0 0 0.294 0.104 0.492 ...
##  $ Asian       : num  0.589 0.321 0.551 0.134 0.299 ...
##  $ Black       : num  71.709 1.283 3.194 0.373 1.073 ...
##  $ Hispanic    : num  0.196 4.522 46.741 0.909 21.294 ...
##  $ White       : num  27.5 93.9 46.8 98.5 75.8 ...
##  $ Multi       : num  NA NA 2.44 NA 1.04 ...
##  $ Total       : int  509 3118 5447 6707 5687 7671 13938 10440 1973 2389 ...
##  $ diverse     : chr  "Diverse" "Extremely undiverse" "Diverse" "Extremely undiverse" ...
##  $ variance    : num  NA NA 0.0116 NA NA ...
##  $ int_group   : chr  NA NA "Highly integrated" NA ...
```

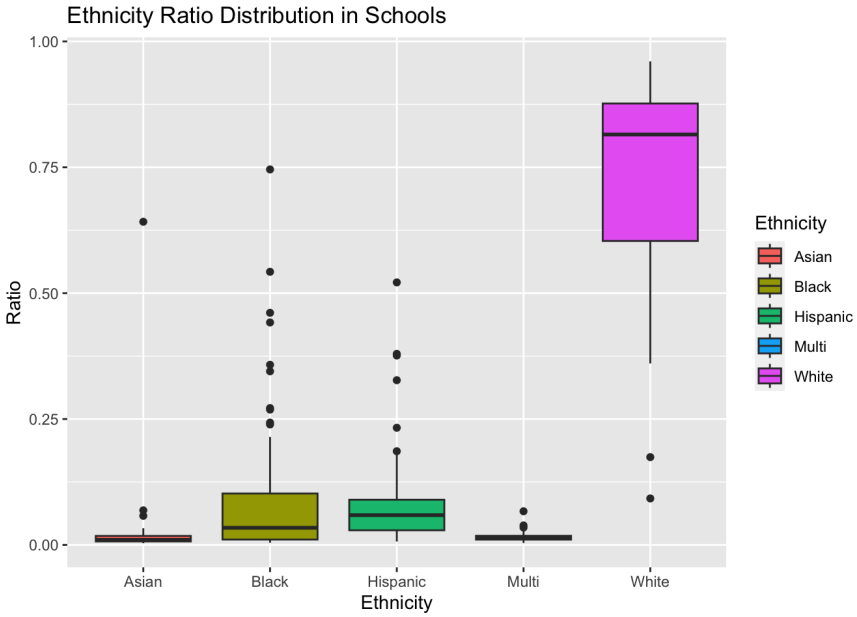
Data Cleaning and Manipulation

```
dfSchool2 <- dfSchool %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  filter(Total > 100) %>%
  mutate(Asian = Asian / 100,
         Black = Black / 100,
         Hispanic = Hispanic / 100,
         White = White / 100,
         Multi = Multi / 100,
         ) %>%
  group_by(ST) %>%
  summarize(Asian = mean(Asian),
            Black = mean(Black),
            Hispanic = mean(Hispanic),
            White = mean(White),
            Multi = mean(Multi)) %>%
  pivot_longer(!ST, names_to = 'Ethnicity', values_to = 'Ratio')
```

```
## Data Analysis

The above data shows the ratio average across school.  Let's plot it for better analysis.


```r
ggplot(data = dfSchool2, aes(x = Ethnicity, y = Ratio, fill = Ethnicity)) +
  geom_boxplot() +
  labs(title = 'Ethnicity Ratio Distribution in Schools')
```



Conclusion Dataset 3

From our analysis, we see that the overall the spread of White student ratio is the highest by a wide margin, where the other race ratios are more similars.
```