COVID19 IMPACT ON DIGITAL LEARNING

Bea Barros

2023-09-08

## OVERVIEW

The global education landscape underwent profound changes due to the COVID-19 pandemic. This analysis seeks to offer a comprehensive understanding of how the pandemic has specifically shaped digital learning experiences for students in the United States.

## DATA WRANGLING AND CLEANING

**I. Installation of Libraries**

The following libraries are loaded for any necessary data manipulations and visualizations.

library(tidyverse)  
library(dplyr)  
library(readr)  
library(zoo)  
library(ggplot2)  
library(viridis)

**II. Manipulation of District Engagement Files**

The code below lists and combines the “district\_info.csv” and district engagement files (1000.csv, 1039.csv,1044.csv,1052.csv,1131.csv) in order to conduct a more meaningful and comprehensive Engagement Analysis of students.

# Identify and list all district csv files in folder using lapply  
selected\_files <- c("1000.csv", "1131.csv", "1052.csv", "1039.csv", "1044.csv")  
data\_list <- lapply(selected\_files, read.csv)  
  
#Loop through the file paths  
for (selected\_files in selected\_files) {  
 district\_id <- sub("\\.csv$", "", selected\_files) # Extract district ID from file name  
 data <- read.csv(selected\_files) # Read CSV file into a data frame  
  
 # Add a new "district" column with the district ID  
 data$district <- district\_id  
   
 # Assign data frame to a variable  
 assign(paste("district\_", district\_id, "\_data", sep = ""), data)  
   
 # Store data frame in the list  
 data\_list[[district\_id]] <- data  
   
}  
  
combined\_dataframe <- rbind(  
 district\_1000\_data,  
 district\_1039\_data,  
 district\_1044\_data,  
 district\_1052\_data,  
 district\_1131\_data  
)  
#show unique districts  
unique(combined\_dataframe$district)

## [1] "1000" "1039" "1044" "1052" "1131"

**III. Manipulation of “districts\_info.csv”**

**A.**Within the ‘District\_Info’ file, certain columns contain categorical values enclosed within brackets. These values require cleaning to ensure they are well-structured and formatted.

# Read the 'districts\_info.csv' file into a data frame  
districts\_info <- read\_csv(file = "districts\_info.csv")

## Rows: 233 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): state, locale, pct\_black/hispanic, pct\_free/reduced, county\_connect...  
## dbl (1): district\_id  
##   
## 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.

# Display a summary of the 'districts\_info' data frame  
summary(districts\_info)

## district\_id state locale pct\_black/hispanic  
## Min. :1000 Length:233 Length:233 Length:233   
## 1st Qu.:2991 Class :character Class :character Class :character   
## Median :4937 Mode :character Mode :character Mode :character   
## Mean :5220   
## 3rd Qu.:7660   
## Max. :9927   
## pct\_free/reduced county\_connections\_ratio pp\_total\_raw   
## Length:233 Length:233 Length:233   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

# Show the first 6 entries of the 'districts\_info' data frame  
head(districts\_info)

## # A tibble: 6 × 7  
## district\_id state locale `pct\_black/hispanic` `pct\_free/reduced`  
## <dbl> <chr> <chr> <chr> <chr>   
## 1 8815 Illinois Suburb [0, 0.2[ [0, 0.2[   
## 2 2685 NaN NaN NaN NaN   
## 3 4921 UTAH Suburb [0, 0.2[ [0.2, 0.4[   
## 4 3188 NaN NaN NaN NaN   
## 5 2238 NaN NaN NaN NaN   
## 6 5987 Wisconsin Suburb [0, 0.2[ [0, 0.2[   
## # ℹ 2 more variables: county\_connections\_ratio <chr>, pp\_total\_raw <chr>

**B.** The following columns have values in range. (ex.[0, 0.2[):

\*1. pct\_black/hispanic  
\*2. pct\_free/reduced  
\*3. pp\_total\_raw  
\*4. county\_connections\_ratio

The left and right brackets must be dropped to make the values well-structured.

# Columns to modify  
columns\_to\_modify <- c(  
 'pct\_black/hispanic',  
 'pct\_free/reduced',  
 'pp\_total\_raw',  
 'county\_connections\_ratio'  
)  
  
for (column\_name in columns\_to\_modify) {  
 districts\_info[[column\_name]] <- gsub("\\[|\\]", "", districts\_info[[column\_name]])  
   
}  
 # Print the updated dataframe  
 head(districts\_info)

## # A tibble: 6 × 7  
## district\_id state locale `pct\_black/hispanic` `pct\_free/reduced`  
## <dbl> <chr> <chr> <chr> <chr>   
## 1 8815 Illinois Suburb 0, 0.2 0, 0.2   
## 2 2685 NaN NaN NaN NaN   
## 3 4921 UTAH Suburb 0, 0.2 0.2, 0.4   
## 4 3188 NaN NaN NaN NaN   
## 5 2238 NaN NaN NaN NaN   
## 6 5987 Wisconsin Suburb 0, 0.2 0, 0.2   
## # ℹ 2 more variables: county\_connections\_ratio <chr>, pp\_total\_raw <chr>

**C.** There are five districts being provided to provide analysis on the engagement scores. The code below selects only the mentioned district IDs - 1000, 1039, 1044, 1052, 1131.

# Values to select  
selected\_district\_ids <- c(1000, 1039, 1044, 1052, 1131)  
  
# Select rows with specific district\_id values  
selected\_rows <- districts\_info[districts\_info$district\_id %in% selected\_district\_ids, ]  
  
# Display selected rows  
print(selected\_rows)

## # A tibble: 5 × 7  
## district\_id state locale `pct\_black/hispanic` `pct\_free/reduced`  
## <dbl> <chr> <chr> <chr> <chr>   
## 1 1044 Missouri Suburb 0, 0.2 0, 0.2   
## 2 1131 <NA> NaN NaN NaN   
## 3 1000 Connecticut Suburb 0.6, 0.8 0.2, 0.4   
## 4 1052 Illinois Suburb 0.2, 0.4 0.2, 0.4   
## 5 1039 NaN NaN NaN NaN   
## # ℹ 2 more variables: county\_connections\_ratio <chr>, pp\_total\_raw <chr>

**III. Consolidating all data files into one single file**

The following code combines multiple dataframes, including engagement files, product information, and district information, into a single consolidated file. This consolidation enhances efficiency and enables more insightful analysis of various variables.

#read products\_info file  
products\_info <- read.csv("products\_info.csv")  
  
# Merge the different dataframe with the combined\_dataframe based on district\_id  
merged\_engagement\_district <- merge(combined\_dataframe, selected\_rows, by.x = "district", by.y = "district\_id",all.x = TRUE)  
  
# Merge the data based on the common "lp\_id" column  
consolidated\_engagement\_data <- merge(merged\_engagement\_district, products\_info, by.x = "lp\_id", by.y = "LP.ID", all.x = TRUE)  
  
head(consolidated\_engagement\_data)

## lp\_id district time pct\_access engagement\_index state locale  
## 1 10003 1052 19/11/2020 0.00 NA Illinois Suburb  
## 2 10006 1044 12/11/2020 0.00 0.23 Missouri Suburb  
## 3 10006 1044 11/11/2020 0.00 0.86 Missouri Suburb  
## 4 10032 1044 2/03/2020 0.01 0.10 Missouri Suburb  
## 5 10032 1044 25/03/2020 0.01 0.10 Missouri Suburb  
## 6 10032 1044 27/04/2020 0.01 0.10 Missouri Suburb  
## pct\_black/hispanic pct\_free/reduced county\_connections\_ratio pp\_total\_raw  
## 1 0.2, 0.4 0.2, 0.4 0.18, 1 16000, 18000  
## 2 0, 0.2 0, 0.2 <NA> 10000, 12000  
## 3 0, 0.2 0, 0.2 <NA> 10000, 12000  
## 4 0, 0.2 0, 0.2 <NA> 10000, 12000  
## 5 0, 0.2 0, 0.2 <NA> 10000, 12000  
## 6 0, 0.2 0, 0.2 <NA> 10000, 12000  
## URL Product.Name Provider.Company.Name Sector.s. Primary.Essential.Function  
## 1 <NA> <NA> <NA> <NA> <NA>  
## 2 <NA> <NA> <NA> <NA> <NA>  
## 3 <NA> <NA> <NA> <NA> <NA>  
## 4 <NA> <NA> <NA> <NA> <NA>  
## 5 <NA> <NA> <NA> <NA> <NA>  
## 6 <NA> <NA> <NA> <NA> <NA>

**IV. Cleaning of Data Values**

The following actions were undertaken to clean the data:

1. Remove some missing Values  
2. Remove irrelevant columns  
3. Change datatypes  
4. Remove duplicates  
5. Change mispelled values

**A. Manipulating the Time Column**

The code below converts the ‘time’ column values from character to date format while also removing incorrect values. This conversion is essential for enabling time series analysis of student engagement index over the course of time.

#check for unique values in time column  
unique(consolidated\_engagement\_data$time)

## [1] "19/11/2020" "12/11/2020" "11/11/2020" "2/03/2020" "25/03/2020"  
## [6] "27/04/2020" "3/03/2020" "11/03/2020" "28/02/2020" "5/03/2020"   
## [11] "10/03/2020" "9/03/2020" "12/03/2020" "13/03/2020" "13/12/2020"  
## [16] "22/03/2020" "26/03/2020" "23/03/2020" "9/04/2020" "10/04/2020"  
## [21] "21/12/2020" "16/09/2020" "30/04/2020" "23/09/2020" "28/09/2020"  
## [26] "15/05/2020" "20/05/2020" "6/10/2020" "8/10/2020" "1/06/2020"   
## [31] "2/06/2020" "11/09/2020" "24/09/2020" "29/09/2020" "6/03/2020"   
## [36] "29/07/2020" "17/11/2020" "11/08/2020" "23/11/2020" "31/03/2020"  
## [41] "2/12/2020" "26/08/2020" "31/08/2020" "14/04/2020" "7/10/2020"   
## [46] "17/04/2020" "18/04/2020" "9/10/2020" "21/04/2020" "6/06/2020"   
## [51] "10/09/2020" "12/09/2020" "26/04/2020" "14/09/2020" "22/09/2020"  
## [56] "25/09/2020" "14/05/2020" "4/06/2020" "19/10/2020" "21/10/2020"  
## [61] "3/12/2020" "4/03/2020" "4/05/2020" "2/10/2020" "10/12/2020"  
## [66] "15/07/2020" "30/09/2020" "27/10/2020" "14/01/2020" "7/03/2020"   
## [71] "8/03/2020" "29/01/2020" "17/03/2020" "19/03/2020" "27/03/2020"  
## [76] "28/03/2020" "30/11/2020" "30/03/2020" "7/02/2020" "1/04/2020"   
## [81] "25/08/2020" "15/03/2020" "24/03/2020" "28/08/2020" "8/04/2020"   
## [86] "12/04/2020" "13/04/2020" "6/04/2020" "7/04/2020" "15/04/2020"  
## [91] "16/04/2020" "20/04/2020" "23/04/2020" "24/04/2020" "22/04/2020"  
## [96] "20/12/2020" "28/04/2020" "1/05/2020" "31/12/2020" "3/05/2020"   
## [101] "19/09/2020" "5/05/2020" "8/05/2020" "9/05/2020" "12/05/2020"  
## [106] "1/01/2020" "2/01/2020" "17/05/2020" "6/01/2020" "7/01/2020"   
## [111] "8/01/2020" "21/05/2020" "9/01/2020" "23/05/2020" "11/01/2020"  
## [116] "13/01/2020" "15/01/2020" "16/01/2020" "17/01/2020" "19/01/2020"  
## [121] "21/01/2020" "22/01/2020" "8/06/2020" "26/01/2020" "30/01/2020"  
## [126] "31/01/2020" "2/02/2020" "4/02/2020" "20/10/2020" "21/06/2020"  
## [131] "11/02/2020" "13/02/2020" "15/02/2020" "18/02/2020" "20/02/2020"  
## [136] "22/10/2020" "21/02/2020" "24/02/2020" "27/02/2020" "29/02/2020"  
## [141] "27/12/2020" "29/12/2020" "16/08/2020" "6/05/2020" "7/05/2020"   
## [146] "29/03/2020" "1/12/2020" "2/04/2020" "24/08/2020" "3/04/2020"   
## [151] "4/04/2020" "5/12/2020" "7/12/2020" "4/01/2020" "8/12/2020"   
## [156] "9/12/2020" "11/04/2020" "2/09/2020" "3/09/2020" "31/05/2020"  
## [161] "4/09/2020" "19/04/2020" "11/10/2020" "9/06/2020" "25/04/2020"  
## [166] "22/12/2020" "13/09/2020" "1/02/2020" "15/09/2020" "3/02/2020"   
## [171] "29/04/2020" "17/09/2020" "28/11/2020" "10/05/2020" "11/05/2020"  
## [176] "17/02/2020" "13/05/2020" "26/09/2020" "16/05/2020" "18/05/2020"  
## [181] "19/05/2020" "1/10/2020" "22/05/2020" "3/10/2020" "17/07/2020"  
## [186] "12/01/2020" "25/05/2020" "26/05/2020" "20/01/2020" "23/01/2020"  
## [191] "28/01/2020" "15/10/2020" "16/10/2020" "6/02/2020" "9/02/2020"   
## [196] "25/02/2020" "10/02/2020" "24/06/2020" "14/02/2020" "22/08/2020"  
## [201] "27/08/2020" "23/02/2020" "25/06/2020" "26/02/2020" "18/07/2020"  
## [206] "21/07/2020" "29/05/2020" "5/09/2020" "16/12/2020" "17/12/2020"  
## [211] "9/09/2020" "16/11/2020" "31/07/2020" "1/08/2020" "1/09/2020"   
## [216] "20/11/2020" "20/03/2020" "21/11/2020" "18/09/2020" "27/11/2020"  
## [221] "1/07/2020" "2/11/2020" "17/10/2020" "14/12/2020" "30/12/2020"  
## [226] "13/10/2020" "24/12/2020" "26/12/2020" "27/09/2020" "21/08/2020"  
## [231] "23/10/2020" "26/10/2020" "18/11/2020" "12/12/2020" "14/10/2020"  
## [236] "11/12/2020" "8/11/2020" "10/11/2020" "13/11/2020" "7/06/2020"   
## [241] "6/11/2020" "29/10/2020" "4/12/2020" "29/11/2020" "18/12/2020"  
## [246] "8/09/2020" "12/10/2020" "16/03/2020" "12/02/2020" "21/03/2020"  
## [251] "30/08/2020" "7/09/2020" "23/07/2020" "30/05/2020" "23/08/2020"  
## [256] "25/01/2020" "6/09/2020" "20/09/2020" "21/09/2020" "3/11/2020"   
## [261] "8/08/2020" "24/11/2020" "14/03/2020" "7/08/2020" "18/08/2020"  
## [266] "19/08/2020" "15/11/2020" "10/08/2020" "24/10/2020" "10/06/2020"  
## [271] "23/06/2020" "17/08/2020" "18/03/2020" "5/11/2020" "5/06/2020"   
## [276] "6/08/2020" "14/08/2020" "4/11/2020" "20/08/2020" "12/08/2020"  
## [281] "11/06/2020" "15/12/2020" "9/11/2020" "24/07/2020" "7/07/2020"   
## [286] "10/01/2020" "24/01/2020" "27/01/2020" "3/08/2020" "5/02/2020"   
## [291] "19/02/2020" "15/08/2020" "5/04/2020" "13/06/2020" "16/06/2020"  
## [296] "28/05/2020" "29/08/2020" "18/06/2020" "22/06/2020" "10/10/2020"  
## [301] "25/10/2020" "22/07/2020" "30/07/2020" "13/08/2020" "23/12/2020"  
## [306] "5/10/2020" "28/10/2020" "14/11/2020" "4/08/2020" "5/08/2020"   
## [311] "25/11/2020" "30/10/2020" "14/07/2020" "29/06/2020" "3/01/2020"   
## [316] "22/02/2020" "22/11/2020" "18/10/2020" "6/12/2020" "27/05/2020"  
## [321] "17/06/2020" "30/06/2020" "3/07/2020" "5/07/2020" "10/07/2020"  
## [326] "19/07/2020" "27/07/2020" "2/05/2020" "20/07/2020" "12/06/2020"  
## [331] "4/10/2020" "1/11/2020" "7/11/2020" "26/11/2020" "9/08/2020"   
## [336] "19/12/2020" "8/07/2020" "13/07/2020" "26/07/2020" "2/08/2020"   
## [341] "9/07/2020" "6/07/2020" "11/07/2020" "12/07/2020" "15/06/2020"  
## [346] "16/07/2020" "18/01/2020" "28/07/2020" "5/01/2020" "1/03/2020"   
## [351] "24/05/2020" "25/07/2020" "14/06/2020" "27/06/2020" "16/02/2020"  
## [356] "3/06/2020" "31/10/2020" "8/02/2020" "19/06/2020" "28/06/2020"  
## [361] "20/06/2020" "28/12/2020" "2/07/2020" "26/06/2020" "4/07/2020"   
## [366] "25/12/2020" "1/01/2022" "1/01/2044" "1/01/2050" "31/12/1020"  
## [371] "2/01/2050" "1/01/2033"

#change character to date  
consolidated\_engagement\_data$time <- as.Date(consolidated\_engagement\_data$time, format = "%d/%m/%Y")  
  
#drop years not in 2020  
dates\_to\_drop <- c("1/01/2022", "1/01/2044", "1/01/2050", "31/12/1020", "2/01/2050", "1/01/2033")  
  
consolidated\_engagement\_data <- consolidated\_engagement\_data %>%  
 filter(!time %in% as.Date(dates\_to\_drop, format = "%d/%m/%Y"))  
  
#check updated dataframe  
unique(consolidated\_engagement\_data$time)

## [1] "2020-11-19" "2020-11-12" "2020-11-11" "2020-03-02" "2020-03-25"  
## [6] "2020-04-27" "2020-03-03" "2020-03-11" "2020-02-28" "2020-03-05"  
## [11] "2020-03-10" "2020-03-09" "2020-03-12" "2020-03-13" "2020-12-13"  
## [16] "2020-03-22" "2020-03-26" "2020-03-23" "2020-04-09" "2020-04-10"  
## [21] "2020-12-21" "2020-09-16" "2020-04-30" "2020-09-23" "2020-09-28"  
## [26] "2020-05-15" "2020-05-20" "2020-10-06" "2020-10-08" "2020-06-01"  
## [31] "2020-06-02" "2020-09-11" "2020-09-24" "2020-09-29" "2020-03-06"  
## [36] "2020-07-29" "2020-11-17" "2020-08-11" "2020-11-23" "2020-03-31"  
## [41] "2020-12-02" "2020-08-26" "2020-08-31" "2020-04-14" "2020-10-07"  
## [46] "2020-04-17" "2020-04-18" "2020-10-09" "2020-04-21" "2020-06-06"  
## [51] "2020-09-10" "2020-09-12" "2020-04-26" "2020-09-14" "2020-09-22"  
## [56] "2020-09-25" "2020-05-14" "2020-06-04" "2020-10-19" "2020-10-21"  
## [61] "2020-12-03" "2020-03-04" "2020-05-04" "2020-10-02" "2020-12-10"  
## [66] "2020-07-15" "2020-09-30" "2020-10-27" "2020-01-14" "2020-03-07"  
## [71] "2020-03-08" "2020-01-29" "2020-03-17" "2020-03-19" "2020-03-27"  
## [76] "2020-03-28" "2020-11-30" "2020-03-30" "2020-02-07" "2020-04-01"  
## [81] "2020-08-25" "2020-03-15" "2020-03-24" "2020-08-28" "2020-04-08"  
## [86] "2020-04-12" "2020-04-13" "2020-04-06" "2020-04-07" "2020-04-15"  
## [91] "2020-04-16" "2020-04-20" "2020-04-23" "2020-04-24" "2020-04-22"  
## [96] "2020-12-20" "2020-04-28" "2020-05-01" "2020-12-31" "2020-05-03"  
## [101] "2020-09-19" "2020-05-05" "2020-05-08" "2020-05-09" "2020-05-12"  
## [106] "2020-01-01" "2020-01-02" "2020-05-17" "2020-01-06" "2020-01-07"  
## [111] "2020-01-08" "2020-05-21" "2020-01-09" "2020-05-23" "2020-01-11"  
## [116] "2020-01-13" "2020-01-15" "2020-01-16" "2020-01-17" "2020-01-19"  
## [121] "2020-01-21" "2020-01-22" "2020-06-08" "2020-01-26" "2020-01-30"  
## [126] "2020-01-31" "2020-02-02" "2020-02-04" "2020-10-20" "2020-06-21"  
## [131] "2020-02-11" "2020-02-13" "2020-02-15" "2020-02-18" "2020-02-20"  
## [136] "2020-10-22" "2020-02-21" "2020-02-24" "2020-02-27" "2020-02-29"  
## [141] "2020-12-27" "2020-12-29" "2020-08-16" "2020-05-06" "2020-05-07"  
## [146] "2020-03-29" "2020-12-01" "2020-04-02" "2020-08-24" "2020-04-03"  
## [151] "2020-04-04" "2020-12-05" "2020-12-07" "2020-01-04" "2020-12-08"  
## [156] "2020-12-09" "2020-04-11" "2020-09-02" "2020-09-03" "2020-05-31"  
## [161] "2020-09-04" "2020-04-19" "2020-10-11" "2020-06-09" "2020-04-25"  
## [166] "2020-12-22" "2020-09-13" "2020-02-01" "2020-09-15" "2020-02-03"  
## [171] "2020-04-29" "2020-09-17" "2020-11-28" "2020-05-10" "2020-05-11"  
## [176] "2020-02-17" "2020-05-13" "2020-09-26" "2020-05-16" "2020-05-18"  
## [181] "2020-05-19" "2020-10-01" "2020-05-22" "2020-10-03" "2020-07-17"  
## [186] "2020-01-12" "2020-05-25" "2020-05-26" "2020-01-20" "2020-01-23"  
## [191] "2020-01-28" "2020-10-15" "2020-10-16" "2020-02-06" "2020-02-09"  
## [196] "2020-02-25" "2020-02-10" "2020-06-24" "2020-02-14" "2020-08-22"  
## [201] "2020-08-27" "2020-02-23" "2020-06-25" "2020-02-26" "2020-07-18"  
## [206] "2020-07-21" "2020-05-29" "2020-09-05" "2020-12-16" "2020-12-17"  
## [211] "2020-09-09" "2020-11-16" "2020-07-31" "2020-08-01" "2020-09-01"  
## [216] "2020-11-20" "2020-03-20" "2020-11-21" "2020-09-18" "2020-11-27"  
## [221] "2020-07-01" "2020-11-02" "2020-10-17" "2020-12-14" "2020-12-30"  
## [226] "2020-10-13" "2020-12-24" "2020-12-26" "2020-09-27" "2020-08-21"  
## [231] "2020-10-23" "2020-10-26" "2020-11-18" "2020-12-12" "2020-10-14"  
## [236] "2020-12-11" "2020-11-08" "2020-11-10" "2020-11-13" "2020-06-07"  
## [241] "2020-11-06" "2020-10-29" "2020-12-04" "2020-11-29" "2020-12-18"  
## [246] "2020-09-08" "2020-10-12" "2020-03-16" "2020-02-12" "2020-03-21"  
## [251] "2020-08-30" "2020-09-07" "2020-07-23" "2020-05-30" "2020-08-23"  
## [256] "2020-01-25" "2020-09-06" "2020-09-20" "2020-09-21" "2020-11-03"  
## [261] "2020-08-08" "2020-11-24" "2020-03-14" "2020-08-07" "2020-08-18"  
## [266] "2020-08-19" "2020-11-15" "2020-08-10" "2020-10-24" "2020-06-10"  
## [271] "2020-06-23" "2020-08-17" "2020-03-18" "2020-11-05" "2020-06-05"  
## [276] "2020-08-06" "2020-08-14" "2020-11-04" "2020-08-20" "2020-08-12"  
## [281] "2020-06-11" "2020-12-15" "2020-11-09" "2020-07-24" "2020-07-07"  
## [286] "2020-01-10" "2020-01-24" "2020-01-27" "2020-08-03" "2020-02-05"  
## [291] "2020-02-19" "2020-08-15" "2020-04-05" "2020-06-13" "2020-06-16"  
## [296] "2020-05-28" "2020-08-29" "2020-06-18" "2020-06-22" "2020-10-10"  
## [301] "2020-10-25" "2020-07-22" "2020-07-30" "2020-08-13" "2020-12-23"  
## [306] "2020-10-05" "2020-10-28" "2020-11-14" "2020-08-04" "2020-08-05"  
## [311] "2020-11-25" "2020-10-30" "2020-07-14" "2020-06-29" "2020-01-03"  
## [316] "2020-02-22" "2020-11-22" "2020-10-18" "2020-12-06" "2020-05-27"  
## [321] "2020-06-17" "2020-06-30" "2020-07-03" "2020-07-05" "2020-07-10"  
## [326] "2020-07-19" "2020-07-27" "2020-05-02" "2020-07-20" "2020-06-12"  
## [331] "2020-10-04" "2020-11-01" "2020-11-07" "2020-11-26" "2020-08-09"  
## [336] "2020-12-19" "2020-07-08" "2020-07-13" "2020-07-26" "2020-08-02"  
## [341] "2020-07-09" "2020-07-06" "2020-07-11" "2020-07-12" "2020-06-15"  
## [346] "2020-07-16" "2020-01-18" "2020-07-28" "2020-01-05" "2020-03-01"  
## [351] "2020-05-24" "2020-07-25" "2020-06-14" "2020-06-27" "2020-02-16"  
## [356] "2020-06-03" "2020-10-31" "2020-02-08" "2020-06-19" "2020-06-28"  
## [361] "2020-06-20" "2020-12-28" "2020-07-02" "2020-06-26" "2020-07-04"  
## [366] "2020-12-25"

**B. Manipulating the “engagement\_index” Column**

The chunks of code below presents the process of changing all NAs from the “engagement\_index” column to “0”. Preserving the NAs will help measure if there are 0 or lack of engagement from students.

#Check for NAs in engagement\_index column  
sum(is.na(consolidated\_engagement\_data$engagement\_index))

## [1] 135792

#Change NAs to 0.00  
consolidated\_engagement\_data$engagement\_index <- ifelse(is.na(consolidated\_engagement\_data$engagement\_index), 0, consolidated\_engagement\_data$engagement\_index)

**C. Manipulating the “state” Column**

There are missing values in State column. The NA and NaN values have been transformed instead to State 1131 and State 1039 to preserve the dataset.

#show unique values in state column  
unique(consolidated\_engagement\_data$state)

## [1] "Illinois" "Missouri" "Connecticut" NA "NaN"

consolidated\_engagement\_data$state <- ifelse(consolidated\_engagement\_data$state == "NaN", "State 1131", consolidated\_engagement\_data$state)  
consolidated\_engagement\_data$state <- ifelse(is.na(consolidated\_engagement\_data$state), "State 1039", consolidated\_engagement\_data$state)  
  
#Check unique values again  
unique(consolidated\_engagement\_data$state)

## [1] "Illinois" "Missouri" "Connecticut" "State 1039" "State 1131"

**D. More Manipulations to Columns**

The ‘URL’ and ‘locale’ columns were excluded from the analysis due to their limited relevance.

#Check unique column headers  
unique(names(consolidated\_engagement\_data))

## [1] "lp\_id" "district"   
## [3] "time" "pct\_access"   
## [5] "engagement\_index" "state"   
## [7] "locale" "pct\_black/hispanic"   
## [9] "pct\_free/reduced" "county\_connections\_ratio"   
## [11] "pp\_total\_raw" "URL"   
## [13] "Product.Name" "Provider.Company.Name"   
## [15] "Sector.s." "Primary.Essential.Function"

#Remove URL and Locale  
consolidated\_engagement\_data <- consolidated\_engagement\_data[, !(names(consolidated\_engagement\_data) %in% c("URL", "locale", ""))]  
  
#Check unique column headers  
unique(names(consolidated\_engagement\_data))

## [1] "lp\_id" "district"   
## [3] "time" "pct\_access"   
## [5] "engagement\_index" "state"   
## [7] "pct\_black/hispanic" "pct\_free/reduced"   
## [9] "county\_connections\_ratio" "pp\_total\_raw"   
## [11] "Product.Name" "Provider.Company.Name"   
## [13] "Sector.s." "Primary.Essential.Function"

The code below changes the header names to make it more presentable.

#Rename Column Headers)  
consolidated\_engagement\_data <- consolidated\_engagement\_data %>%  
 rename("LP ID" = `lp\_id`,  
 "Timestamp" = `time`,  
 "Engagement Index" = `engagement\_index`,  
 "% of Black/Hispanic" = `pct\_black/hispanic`,  
 "Product Name" = `Product.Name`,  
 "County Connection" = `county\_connections\_ratio`,  
 "Sector" = `Sector.s.`,  
 "District" = `district`,  
 "State" = `state`,  
 "% Free or Reduced" = `pct\_free/reduced`,  
 "District Expenditure" = `pp\_total\_raw`,  
 "Company Provider" = `Provider.Company.Name`,  
 "Primary Product Function" = `Primary.Essential.Function`,  
 "% of Page Access" = `pct\_access`)  
  
#Check updated header  
unique(names(consolidated\_engagement\_data))

## [1] "LP ID" "District"   
## [3] "Timestamp" "% of Page Access"   
## [5] "Engagement Index" "State"   
## [7] "% of Black/Hispanic" "% Free or Reduced"   
## [9] "County Connection" "District Expenditure"   
## [11] "Product Name" "Company Provider"   
## [13] "Sector" "Primary Product Function"

There are also some incorrect input on the “Sector” column which require cleaning.

# Show unique values in the "Sector" column  
unique(consolidated\_engagement\_data$Sector)

## [1] NA "PreK-12"   
## [3] "PreK-12; Higher Ed; Corporate" "PreK-12; Higher Ed"   
## [5] "PreK-122" "Corporate"   
## [7] "PreK-112" ""   
## [9] "PreK-12; Higher; Corporate" "PPreK-12"   
## [11] "Higher Ed; Corporate"

# Replace specific values in the "Sector(s)" column  
consolidated\_engagement\_data$Sector <- gsub("PreK-112", "PreK-12", consolidated\_engagement\_data$Sector)  
  
consolidated\_engagement\_data$Sector <- gsub("PreK-122", "PreK-12", consolidated\_engagement\_data$Sector)  
  
consolidated\_engagement\_data$Sector <- gsub("PreK-12; Higher; Corporate", "PreK-12; Higher Ed; Corporate", consolidated\_engagement\_data$Sector)  
  
consolidated\_engagement\_data$Sector <- gsub("PPreK-12", "PreK-12", consolidated\_engagement\_data$Sector)  
  
# Show the modified "Sector(s)" column  
unique(consolidated\_engagement\_data$Sector)

## [1] NA "PreK-12"   
## [3] "PreK-12; Higher Ed; Corporate" "PreK-12; Higher Ed"   
## [5] "Corporate" ""   
## [7] "Higher Ed; Corporate"

**F. Drop NAs in other columns**

Rows with missing values in the ‘Sector.s.’, ‘Provider.Company.Name’, and ‘Primary.Essential.Function’ columns were removed from the analysis due to their lack of relevance.

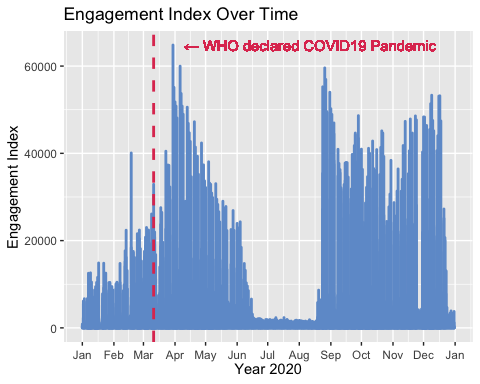
# Define a logical condition to check for NAs in specific columns  
na\_condition <- complete.cases(consolidated\_engagement\_data[, c("Sector", "Company Provider", "Primary Product Function")])  
  
# Use the condition to subset your DataFrame  
consolidated\_engagement\_data <- consolidated\_engagement\_data[na\_condition, ]  
head(consolidated\_engagement\_data)

## LP ID District Timestamp % of Page Access Engagement Index State  
## 1476 10533 1131 2020-09-28 0.04 8.17 State 1039  
## 1477 10533 1052 2020-07-21 0.04 1.09 Illinois  
## 1478 10533 1000 2020-04-09 0.14 6.09 Connecticut  
## 1479 10533 1039 2020-11-26 0.08 4.51 State 1131  
## 1480 10533 1052 2020-04-08 1.23 175.27 Illinois  
## 1481 10533 1000 2020-09-09 0.46 12.77 Connecticut  
## % of Black/Hispanic % Free or Reduced County Connection  
## 1476 NaN NaN NaN  
## 1477 0.2, 0.4 0.2, 0.4 0.18, 1  
## 1478 0.6, 0.8 0.2, 0.4 0.18, 1  
## 1479 NaN NaN NaN  
## 1480 0.2, 0.4 0.2, 0.4 0.18, 1  
## 1481 0.6, 0.8 0.2, 0.4 0.18, 1  
## District Expenditure Product Name Company Provider Sector  
## 1476 NaN Scholastic Scholastic Inc PreK-12  
## 1477 16000, 18000 Scholastic Scholastic Inc PreK-12  
## 1478 <NA> Scholastic Scholastic Inc PreK-12  
## 1479 NaN Scholastic Scholastic Inc PreK-12  
## 1480 16000, 18000 Scholastic Scholastic Inc PreK-12  
## 1481 <NA> Scholastic Scholastic Inc PreK-12  
## Primary Product Function  
## 1476 LC - Sites, Resources & References - Learning Materials & Supplies  
## 1477 LC - Sites, Resources & References - Learning Materials & Supplies  
## 1478 LC - Sites, Resources & References - Learning Materials & Supplies  
## 1479 LC - Sites, Resources & References - Learning Materials & Supplies  
## 1480 LC - Sites, Resources & References - Learning Materials & Supplies  
## 1481 LC - Sites, Resources & References - Learning Materials & Supplies

**DATA VISUALIZATION**

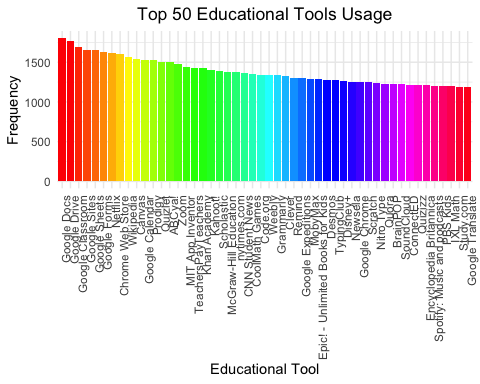
**I.Engagement Index of Students over time**

A time series graph is a powerful tool for tracking student Engagement Index over time, enabling us to uncover patterns and evaluate the pandemic’s impact on student engagement. The graph below illustrates the student engagement trend for the year 2020



**II. Top 50 Educational Tools Usage**

This bar chart below showcases the top 50 educational tools that are most frequently utilized by students. These tools play a pivotal role in enhancing student engagement within the realm of digital learning.

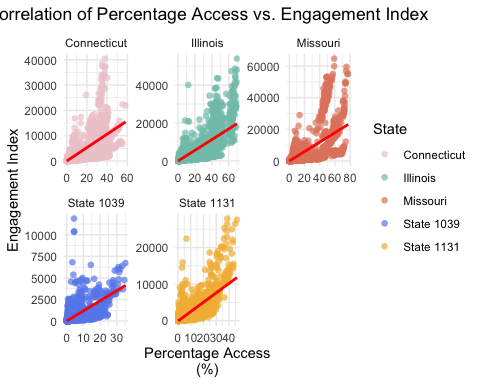


**III. Correlation of PCT Access & Engagement Score**

The scatterplot is used to explored correlations between the engagement index and percent access. These metrics are pimportant in assessing the effectiveness of educational products in the context of digital learning.

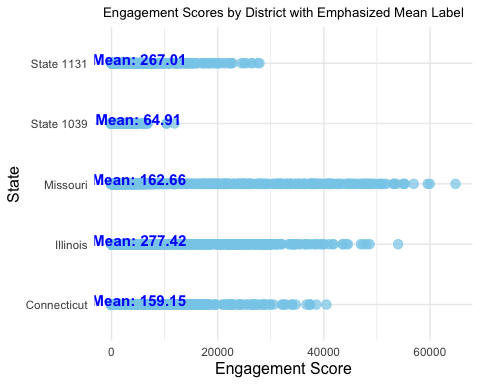
# Pick Colors  
custom\_colors <- c("#EDCBD2", "#80C4B7", "#E3856B", "#678CEC", "#F3B941")  
  
# Create separate scatter plots per district  
ggplot(consolidated\_engagement\_data, aes(x = `% of Page Access`, y = `Engagement Index`, color = State)) +  
 geom\_point(shape = 16, size = 2, alpha = 0.7) +  
 geom\_smooth(method = "lm", se = TRUE, color = "red") +  
 labs(x = "Percentage Access\n(%)",  
 y = "Engagement Index",  
 title = "Correlation of Percentage Access vs. Engagement Index") +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(hjust = 0.5)  
 ) +  
 facet\_wrap(~State, nrow = 2, scales = "free") +   
 scale\_color\_manual(values = custom\_colors)

## `geom\_smooth()` using formula = 'y ~ x'



**IV. Distribution of Engagement per District**

The Dot Plot is a good way to show and compare distribution of Engagement index on each district, and identify the mean per district category.



**V. Product Usage Trend of Top 5 Products in 2020**

The multiple bar graph illustrates the usage trends of the top 5 educational tools used by students throughout the year 2020. This visualization offers valuable insights into the evolving patterns of product utilization over time, shedding light on the most frequently employed tools by students during this period.

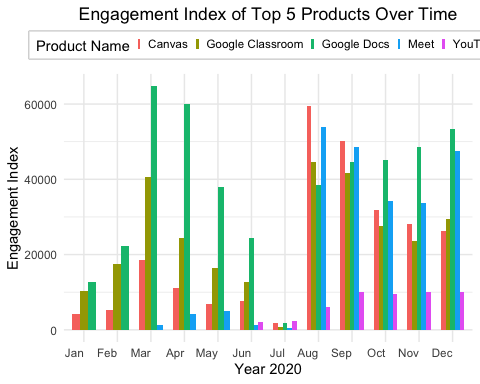
# Filter for the top 5 products based on timestamp  
top\_products <- consolidated\_engagement\_data %>%  
 group\_by(`Product Name`) %>%  
 summarise(Engagement\_Index = mean(`Engagement Index`)) %>%  
 arrange(desc(Engagement\_Index)) %>%  
 top\_n(5)

## Selecting by Engagement\_Index

# Left join with consolidated\_engagement\_data to get additional data  
top\_products <- top\_products %>%  
 left\_join(consolidated\_engagement\_data, by = "Product Name")  
  
# Add a new column "Month" with the name of each month  
top\_products <- top\_products %>%  
 mutate(Month = month(Timestamp, label = TRUE))  
  
# Create a grouped bar plot  
bar\_plot <- ggplot(top\_products, aes(x = Month, y = `Engagement Index`, fill = `Product Name`)) +  
 geom\_bar(stat = "identity", position = "dodge", width = 0.7) +  
 labs(title = "Engagement Index of Top 5 Products Over Time",  
 x = "Year 2020",  
 y = "Engagement Index",  
 fill = "Product Name") +  
 theme\_minimal() +  
 theme(  
 legend.position = "top",   
 legend.box.background = element\_rect(color = "black", size = 0.1),   
 legend.direction = "horizontal",  
 legend.key.size = unit(0.1, "cm"),   
 legend.key.width = unit(0.1, "cm"),  
 plot.title = element\_text(hjust = 0.5)  
 )

## Warning: The `size` argument of `element\_rect()` is deprecated as of ggplot2 3.4.0.  
## ℹ Please use the `linewidth` argument instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

# Rotate x-axis labels for better readability  
bar\_plot <- bar\_plot + theme(axis.text.x = element\_text(angle = 0, hjust = 1))  
  
# Display the grouped bar plot  
print(bar\_plot)



## FINDINGS

**I. Observations & Insights**

**A. Time-Series Analysis on Engagement Index**

Analyzing the time series plot reveals significant insights in engagement levels of students. The following findings from the Time-Series Plot have been observed:

1. Page-load events were doubled after COVID-19 was declared a pandemic.  
2. Engagement significantly declined between June to August 2020 (summer)   
3. By the end of 2020, engagement index is still higher compared to pre covid.  
4. Usage of educational tools during US summer/holiday are the lowest.

**B. Top 50 Education Product Usage**

Following the declaration of COVID-19 as a pandemic, the usage of certain products experiences a rapid surge. The following findings from the Bar Chart have been observed:

1. Several educational tools and digital platforms also managed to grow during Pandemic.  
2. Google is the most successful company that benefited during pandemic   
3. Google is much more effective in terms of engaging students during Pandemic  
4. Youtube and Google Meet had peaked rapidly during Pandemic.

**C. Correlation of PCT Access & Engagement Score**

The scatter plot revealed a positive correlation coefficient between PCT Access and Engagement Index. The following findings from the Scatter Plot have been observed:

1. There is an upward-sloping pattern  
2. The data points are closely clustered around the line  
3. Access to the online platform is important for student engagement  
4. Higher access rates are associated with higher engagement levels

**D. Distribution of Engagement per District**

The following findings from the Dot Plot Chart have been observed:

1. District 1033 has the lowest mean engagement indexes from the rest of the districts.  
2. Illinois has the highest mean engagement index   
3. From the mean scores, Illinois, Missouri, Connecticut may have effectively adapted to digital learning during the pandemic.

**E. Product Usage Trend of Top 5 Products in 2020**

The following findings from the Multiple Bar Chart have been observed:

1. Before the pandemic, the Engagement Index of all top 5 products were near the 20,000 mark.  
2. Google Meet and Youtube had lower page loads before pandemic compared to Google Doc, Google Classroom and Canvas.  
3. After the start of pandemic on March, Google Meet started to grow its page loads per days as most students started to use video conferencing platforms for attending class.  
4. It is also interesting to note that Youtube started to increase its page loads during summer or holiday break as students are using it more for entertainment purposes. But it steadily grew as well after the holiday until the end of Decemeber 2020.

**III. Recommendations**

The following recommendations below can be used to provide meaningful solutions to different stakeholders:

**ENGAGEMENT INDEX OF STUDENTS OVER TIME**

Pandemic had a positive effect on student engagement, and increased the reliance of students on remote learning. With learning methods shifting towards digital formats at an unprecedented pace, it becomes crucial, particularly for educational institutions, to embrace these evolving trends. By doing so, they can enhance their teaching approaches and better support students in their pursuit of quality education.

**RECOMMENDATIONS REGARDING EDUCATIONAL TOOLS**

For companies operating in the digital learning sector, this analysis underscores the significance of designing and offering products that resonate with students and contribute to their learning success. Digital platforms trying to increase growth and penetration to student market can gain valuable market intelligence and competitive analysis from these top Educational Providers and Products to be able craft strategic strategies, and enhance products that cater to the evolving needs of students. This understanding is instrumental in achieving a competitive edge and fostering growth in the education technology market.

**DISTRIBUTION OF ENGAGEMENT PER DISTRICT**

Having the lowest mean of Engagement among the given districts, District 1033 might require additional support or intervention to improve digital learning outcomes.

**PERCENTAGE OF ACCESS AND ENGAGEMENT**

For educational institutions and product developers, understanding the positive correlation of PCT Access & Engagement Score is valuable. These companies and institution must put efforts to increase product adoption and access to produce greater student engagement and, consequently, improved learning outcomes.

They should also give focus on designing and promoting educational tools that not only attract users but also encourage their active participation, ultimately contributing to enhanced educational experiences.

**IV. Conclusion**

In conclusion, embracing these insights and recommendations can lead to more effective teaching approaches, better support for students, and ultimately, higher-quality education.