Assesssment 1 - COVID-19 impact on digital learning

Priyam and 47750731

2024-04-12

Data Description

Over 56 million students in the US had disruptions in their education because of the COVID-19 pandemic. To stop the virus from spreading, the majority of states and local governments in the United States closed their educational institutions in the spring of 2020. Schools and teachers have responded by attempting to connect with pupils virtually using digital platforms and tools for distance learning. Even now, worries about the widening digital divide and long-term learning loss among the most disadvantaged students in America are becoming more and more pressing.

Data upload

- Now we'll load the entire dataset—distributions, products, and engagements of the districts.

Districts Dataset

- District-specific data, including information from Edunomics Lab, the FCC (Dec. 2018), and NCES (2018–19), is included in the district file districts_info.csv.
- We eliminated the school districts' personally identifiable information from this data collection. In order to alter many data fields and lower the possibility of re-identification, we additionally employed the open-source program ARX (Prasser et al. 2020).
- Certain data points are released with a range that the actual value falls under in order to aid with data generalization. Furthermore, a large number of missing data points are designated as "NaN," suggesting that the data was suppressed to increase the dataset's anonymity.

```
##   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
##
##
##
```

- The above shows the dimension of the ‘Districts’ dataset.
- We can see the districts file has the column variable district_id as the numeric, and the rest of the variables are characters.
- Let’s do data cleaning and pre-processing of Data.districts info Summary of the data in terms of data type and summary statistics of the numeric variable Now, we will take each column and do the cleaning and pre processing of the variable.

District’s State

Let’s take the state column first.

```
##    Length     Class      Mode
##       233 character character
```

- The state column has 233 rows which has a character type.

```
## # A tibble: 33 × 2
##    state                  count
##    <chr>                  <int>
##  1 "Arizona"                  1
##  2 "California"              12
##  3 "ConnectiCUT"              1
##  4 "Connecticut"             29
##  5 "District Of Columbia"     3
##  6 "Florida"                  1
##  7 "Illinois"                18
##  8 "Indiana"                  7
##  9 "Massachusetts"           21
## 10 "Michigan"                 2
## 11 "Minnesota"                1
## 12 "Missouri"                 6
## 13 "NY City"                  1
## 14 "New Hampshire"            2
## 15 "New Jersey"               2
## 16 "New Y0rk"                 1
## 17 "New York"                 6
## 18 "North Carolina"           4
## 19 "North Dakota"             1
## 20 "Ohi0"                     1
## 21 "Ohio"                    10
## 22 "Tennessee"                2
## 23 "Texas"                    2
## 24 "UTAH"                     1
## 25 "Utaah"                    1
## 26 "Utah"                    26
## 27 "Virginia"                 4
## 28 "Washington"              6
## 29 "Wisconsin"               3
## 30 "don\x92t know"           1
## 31 "uTtah"                   1
## 32 "whereabouts"             1
## 33  <NA>                    55
```

- As we can see, we don’t know, and whereabouts seem vague.
- We will check on all the districts that are part of the engagement, as they will be crucial.
- As we can also see, there are 55 NA, and without data cleaning, there are 32 states, but that’s incorrect as there are discrepancies like Utah, New York, and Ohio. It will be correct in this.
- Before that, let’s see the districts 1000, 1039, 1044, 1052, and 1131. Since these are engagement files, let’s see these districts.

```
## # 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 "don\x92t know" <NA>   <NA>                 <NA>
## 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  <NA>           <NA>   <NA>                 <NA>
## # ℹ 2 more variables: county_connections_ratio <chr>, pp_total_raw <chr>
```

- As we can see, 1131 doesn’t know the state, which means we don’t know the state, but in merging the engagement districts, we cannot keep it as this as we want this row. Similarly, we need 1039 rows as well. Therefore, we will keep 1039 rows by keeping their state name as the district ID.

- Correcting the spelling errors by first entering the title in title format and then using the replace function.

```
## # A tibble: 27 × 3
##    state                count percentage
##    <chr>                <int>      <dbl>
##  1 1039                     1      0.429
##  2 1131                     1      0.429
##  3 Arizona                  1      0.429
##  4 California              12      5.15
##  5 Connecticut             30     12.9
##  6 District Of Columbia     3      1.29
##  7 Florida                  1      0.429
##  8 Illinois                18      7.73
##  9 Indiana                  7      3.00
## 10 Massachusetts           21      9.01
## 11 Michigan                 2      0.858
## 12 Minnesota                1      0.429
## 13 Missouri                 6      2.58
## 14 New Hampshire            2      0.858
## 15 New Jersey               2      0.858
## 16 New York                 8      3.43
## 17 North Carolina           4      1.72
## 18 North Dakota             1      0.429
## 19 Ohio                    11      4.72
## 20 Tennessee                2      0.858
## 21 Texas                    2      0.858
## 22 Utah                    29     12.4
## 23 Virginia                 4      1.72
## 24 Washington               6      2.58
## 25 Whereabouts              1      0.429
## 26 Wisconsin                3      1.29
## 27 <NA>                    54     23.2
```

```
## # A tibble: 27 × 3
##    state         count percentage
##    <chr>         <int>      <dbl>
##  1 <NA>             54      23.2
##  2 Connecticut      30      12.9
##  3 Utah             29      12.4
##  4 Massachusetts    21       9.01
##  5 Illinois         18       7.73
##  6 California       12       5.15
##  7 Ohio             11       4.72
##  8 New York          8       3.43
##  9 Indiana           7       3.00
## 10 Missouri          6       2.58
## # ℹ 17 more rows
```

- Removing the NA of the district file because they are 23.17%, which is a huge number, but without knowing the state name of the district, we cannot analyze the further district demographics of it. The government will not be able to identify further over the districts that have na as the district.

District's Locale

- Let's take the second column of the NCES locale classification that categorizes U.S. territory into four types of areas: city, suburban, town, and rural.

```
## # A tibble: 8 × 3
##   locale count percentage
##   <chr>  <int>      <dbl>
## 1 C1ty       1      0.559
## 2 Cit        1      0.559
## 3 City      28     15.6
## 4 Rural     33     18.4
## 5 Sub        2      1.12
## 6 Suburb   101     56.4
## 7 Town      10      5.59
## 8 <NA>       3      1.68
```

- There are 3 NA’s, of which 1.6% we will replace. As there’s no point in deleting rows, we will replace them with the most common locale.
- We can see there are some similarities between cities; there are a few of these; we will replace them. Set them into the required categories.

```
## # A tibble: 5 × 3
##   locale count percentge
##   <chr>  <int>     <dbl>
## 1 City      30     16.8
## 2 Rural     33     18.4
## 3 Suburb   103     57.5
## 4 Town      10      5.59
## 5 <NA>       3      1.68
```

- Now we will replace NA’s and see the results.

```
## [1]  23  24 128

## # A tibble: 4 × 2
##   locale count
##   <chr>  <int>
## 1 City      30
## 2 Rural     33
## 3 Suburb   106
## 4 Town      10
```

District’s Minority Range

- Let’s take the next column. Pct black/hispanic Percentage of students in the districts identified as black or Hispanic based on 2018–19 NCES data As it seems like a long variable name, we will cut it to the minority range.

```
## # A tibble: 6 × 3
##   minority_range count percentge
##   <chr>          <int>     <dbl>
## 1 [0, 0.2[         116     64.8
## 2 [0.2, 0.4[        24     13.4
## 3 [0.4, 0.6[        17      9.50
## 4 [0.6, 0.8[        11      6.15
## 5 [0.8, 1[           8      4.47
## 6 <NA>               3      1.68
```

- “[0, 0.2]” is determined as 0 to 20%. We will replace this with the middle number of 10% for easy understanding. We will convert this to a numeric data type. - We will replace the na values with the median of the entire column.

```
## # A tibble: 5 × 3
##   minority_range count percentge
##            <dbl> <int>     <dbl>
## 1            0.1   119     66.5
## 2            0.3    24     13.4
## 3            0.5    17      9.50
## 4            0.7    11      6.15
## 5            0.9     8      4.47
```

District’s pct_free/reduced

*Now, we will take the next column, pct_free/reduced. Percentage of students in the districts eligible for free or reduced-price lunch based on 2018-19 NCES data As it seems like a long variable name, we will cut it to the free range.

```
## # A tibble: 6 × 3
##   free_range count percentge
##   <chr>      <int>     <dbl>
## 1 [0, 0.2[      46     25.7 
## 2 [0.2, 0.4[    48     26.8 
## 3 [0.4, 0.6[    37     20.7 
## 4 [0.6, 0.8[    13      7.26
## 5 [0.8, 1[       4      2.23
## 6 <NA>          31     17.3 
```

- “[0, 0.2[” determines as 0 to 20 % we will replace this as the middle number 10% for easy understanding. we will convert this to numeric data type.
- We will replace the na values with the median of the entire column.

```
## # A tibble: 5 × 3
##   free_range count percentge
##        <dbl> <int>     <dbl>
## 1        0.1    46     25.7 
## 2        0.3    79     44.1 
## 3        0.5    37     20.7 
## 4        0.7    13      7.26
## 5        0.9     4      2.23
```

District’s Expense per pupil

- pp_total_raw Per-pupil total expenditure (sum of local and federal expenditure) from Edunomics Lab’s National Education Resource Database on Schools (NERD$) project. The expenditure data are school-by-school, and we use the median value to represent the expenditure of a given school district
- As it seems the long variable name we will cut it to the free range

```
## # A tibble: 12 × 3
##    Expenseperpupil count percentge
##    <chr>           <int>     <dbl>
##  1 [10000, 12000[     17     9.50 
##  2 [12000, 14000[     15     8.38 
##  3 [14000, 16000[     15     8.38 
##  4 [16000, 18000[     13     7.26 
##  5 [18000, 20000[      8     4.47 
##  6 [20000, 22000[      2     1.12 
##  7 [22000, 24000[      2     1.12 
##  8 [32000, 34000[      1     0.559
##  9 [4000, 6000[        2     1.12
## 10 [6000, 8000[       13     7.26
## 11 [8000, 10000[      30    16.8
## 12 <NA>               61    34.1
```

- “[10000, 12000[”, determines as 10000 to 12000 we will replace this as the mean value number 1 for easy understanding. we will convert this to numeric data type.
- we will replace the na values with the median of the entire column.

```
## # A tibble: 11 × 3
##    Expenseperpupil count percentge
##              <dbl> <int>     <dbl>
##  1            5000     2     0.858
##  2            7000    13     5.58
##  3            9000    30    12.9
##  4           11000    78    33.5
##  5           13000    15     6.44
##  6           15000    15     6.44
##  7           17000    13     5.58
##  8           19000     8     3.43
##  9           21000     2     0.858
## 10           23000     2     0.858
## 11           33000     1     0.429
```

District’s Connection Ratio

let’s take the next column county_connections_ratio ratio (residential fixed high-speed connections over 200 kbps in at least one direction/households) based on the county level data from FCC From 477 (December 2018 version).

```
## # A tibble: 3 × 3
##   Ratio     count percentge
##   <chr>     <int>     <dbl>
## 1 [0.18, 1[   161    69.1
## 2 [1, 2[        1     0.429
## 3 <NA>         17     7.30
```

- “[0.18, 1[”, determines as 18% to 100% we will replace this as the 18% number easy understanding. we will convert this to numeric data type
- We will replace the na values with the median of the entire column.

#Product Dataset *The product file products_info.csv includes information about the characteristics of the top 372 products with most users in 2020. The categories listed in this file are part of LearnPlatform’s product taxonomy. Data were labeled by our team. Some products may not have labels due to being duplicate, lack of accurate url or other reasons

*Lets take each column individually and do the data cleaning and data preprocessing wherever required.

*The first column is the Sector. Sector(s) of education where the product is used ## Product's Sector

```
## # A tibble: 13 × 3
##    Sector                        count percentge
##    <chr>                         <int>     <dbl>
##  1 Corporate                         1     0.269
##  2 Higher Ed; Corporate              1     0.269
##  3 PPreK-12                          1     0.269
##  4 PreK-112                          2     0.538
##  5 PreK-12                         165    44.4
##  6 PreK-122                          1     0.269
##  7 PreK-12; Higher Ed               65    17.5
##  8 PreK-12; Higher Ed; Corporate   114    30.6
##  9 PreK-12; Higher; Corporate        1     0.269
## 10 not sure                          1     0.269
## 11 pre kindergarten to year 12       1     0.269
## 12 pre kindergarten to yr 12         1     0.269
## 13 <NA>                             18     4.84
```

- There are 18 na which is 4.8% of the column. Lets also try to minimise the categories by combining the similar categories name.
- Not sure is replaced by NA, Further NA is replaced by most common category.

```
## # A tibble: 3 × 3
##   Sector    count percentage
##   <chr>     <int>      <dbl>
## 1 Corporate   117       31.5
## 2 Higher Ed    65       17.5
## 3 Prek-12     190       51.1
```

Product's Primary Function

- lets take the next column
- Primary Essential Function The basic function of the product. There are two layers of labels here. Products are first labeled as one of these three categories: LC = Learning & Curriculum, CM = Classroom Management, and SDO = School & District Operations. Each of these categories have multiple sub-categories with which the products were labeled.

```
## [1] "LC - Digital Learning Platforms"
## [2] "LC - Digital Learning Platforms"
## [3] "LC - Sites, Resources & Reference - Games & Simulations"
## [4] "LC - Digital Learning Platforms"
## [5] "LC - Digital Learning Platforms"
## [6] "LC - Digital Learning Platforms"
```

- we can see they have combined the function we will separate into two columns, Primary and secondary for category creation as part of data processing.

```
## Warning: Expected 2 pieces. Additional pieces discarded in 112 rows [3, 9,
10, 12, 20,
## 21, 26, 27, 37, 44, 45, 46, 52, 53, 54, 56, 57, 59, 64, 66, ...].
```

```
## # A tibble: 6 × 3
##   Primary      count percentage
##   <chr>        <int>      <dbl>
## 1 "CL "            1      0.269
## 2 "CM "           34      9.14
## 3 "LC "          271     72.8
## 4 "LC/CM/SDO "    16      4.30
## 5 "SDO "          30      8.06
## 6  <NA>           20      5.38
```

N

```
## # A tibble: 3 × 3
##   Primary count percentage
##   <chr>   <int>      <dbl>
## 1 "CM "      35       9.41
## 2 "LC "     307      82.5
## 3 "SDO "     30       8.06
```

- Now we have the 3 category data of CM, LC and SDO as the primary function
- Now let's look into the secondary function colum that we have created

Product's Secondary Function

```
## # A tibble: 23 × 3
##    Secondary                                                 count
percentage
##    <chr>                                                     <int>
<dbl>
##  1 " Digital Learning Platforms"                                74
19.9
##  2 " Sites, Resources & Reference "                             50
13.4
##  3 " Sites, Resources & Reference"                              47
12.6
##  4 " Content Creation & Curation"                               36
9.68
##  5 " Study Tools"                                               25
6.72
##  6 " Classroom Engagement & Instruction "                       20
5.38
##  7  <NA>                                                        20
5.38
##  8 " Courseware & Textbooks"                                    18
4.84
##  9 " Other"                                                     18
4.84
```

```
## 10 " Study Tools "                                               10
2.69
## 11 " Data, Analytics & Reporting "                                8
2.15
## 12 " Teacher Resources "                                          7
1.88
## 13 " Virtual Classroom "                                          7
1.88
## 14 " Learning Management Systems (LMS)"                          5
1.34
## 15 " Online Course Providers & Technical Skills Development"     5
1.34
## 16 " Human Resources"                                            4
1.08
## 17 " School Management Software "                                4
1.08
## 18 " Sites, Resources & References "                             4
1.08
## 19 " Career Planning & Job Search"                               3
0.806
## 20 " Data, Analytics & Reporting"                                3
0.806
## 21 " Large"                                                      2
0.538
## 22 " Admissions, Enrollment & Rostering"                         1
0.269
## 23 " Environmental, Health & Safety (EHS) Compliance"            1
0.269
```

- There are 23 rows with 20 NA since there are too many secondary function we will replace them as unknown category as top secondary function is just the 19%.
- There was the one text error in which extra space was used because of this it converted into two different categories. they are same.

```
## # A tibble: 22 × 3
##    Secondary                                                 count
percentage
##    <chr>                                                     <int>
<dbl>
##  1 " Sites, Resources & Reference"                              97
26.1
##  2 " Digital Learning Platforms"                                74
19.9
##  3 " Content Creation & Curation"                               36
9.68
##  4 " Study Tools"                                               25
6.72
##  5 " Classroom Engagement & Instruction "                       20
5.38
##  6 "Unknown"                                                    20
```

```
5.38
##  7 " Courseware & Textbooks"                                    18
4.84
##  8 " Other"                                                     18
4.84
##  9 " Study Tools "                                              10
2.69
## 10 " Data, Analytics & Reporting "                               8
2.15
## 11 " Teacher Resources "                                         7
1.88
## 12 " Virtual Classroom "                                         7
1.88
## 13 " Learning Management Systems (LMS)"                          5
1.34
## 14 " Online Course Providers & Technical Skills Development"     5
1.34
## 15 " Human Resources"                                            4
1.08
## 16 " School Management Software "                                4
1.08
## 17 " Sites, Resources & References "                             4
1.08
## 18 " Career Planning & Job Search"                               3
0.806
## 19 " Data, Analytics & Reporting"                                3
0.806
## 20 " Large"                                                      2
0.538
## 21 " Admissions, Enrollment & Rostering"                         1
0.269
## 22 " Environmental, Health & Safety (EHS) Compliance"            1
0.269
```

- Now, Let’s take the next column Provider/Company Name Name of the product provider

```
## # A tibble: 292 × 2
##    `Provider/Company Name`    count
##    <chr>                     <int>
##  1 Google LLC                   29
##  2 Houghton Mifflin Harcourt     6
##  3 Microsoft                     6
##  4 IXL Learning                  4
##  5 Learning A-Z                  4
##  6 Adobe Inc.                    3
##  7 Autodesk, Inc                 3
##  8 Curriculum Associates         3
##  9 ExploreLearning, LLC          3
## 10 McGraw-Hill PreK-12           3
## # ℹ 282 more rows
```

- There are 292 products in which top 5 provider or companies govt will be interested in dealing with this.

- Lets look product id which is like a key between enagagemmnet district files and the product files. To give consistent names we will change the name of the column to match it with district engagement file.

Product’s ID

#Engagement Data

- Let’s do the merging of the files Let’s do the engagemnt of the files

- To do this we will provide the extra column to our engagement district file to merge with dirstrict file.

```
## [1] "Extra Column is added"
```

*We will use the inner join to take the common intersection of the districts dataset and the engagement files of district with the common key of district_id

- We will use the inner join to take the common intersection of the dataset and the engagement files of district by the common key of lp_id

Engage Data Date

- let’s look into the date format of the files and make this consistent throught out the dataset. *we will look into each data files date column

##Merging the Data

*Now, we will combine all the engagement files and merge all the rows to the one engageDataset

*Now let’s do the data processing and add Year and month from the time column for better analysis

- We are dealing with 2020 year only therefore we need dataset of only 2020

```
## # A tibble: 6 × 2
##    Year  count
##   <dbl>  <int>
## 1  1020      1
## 2  2020 266990
## 3  2022      1
## 4  2033      1
## 5  2044      1
## 6  2050      1
```

- Now, as we only need 2020 we filter out the year 2020

```
## # A tibble: 1 × 2
##    Year  count
##   <dbl>  <int>
## 1  2020 266990
```

Engagement Pct_Access

- Let’s move to the next column pct_access #Percentage of students in the district have at least one page-load event of a given product and on a given day

```
##    Length     Class      Mode 
##    266990 character character
```

- As we can see they are denoted as character which is incorrect we need to change the datatype of this, replace na with the median values

```
## [1] "character"

## Warning: NAs introduced by coercion

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0000  0.0000  0.0500  0.8978  0.2300 77.6200       2

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.0500  0.8978  0.2300 77.6200
```

Engaagement Index

- Let’s take the engagement index as the next column to clean; similarly to the pct_access, we will convert it into a numeric and then put the na’s to the median.
- engagement_index Total page-load events per one thousand students of a given product and on a given day

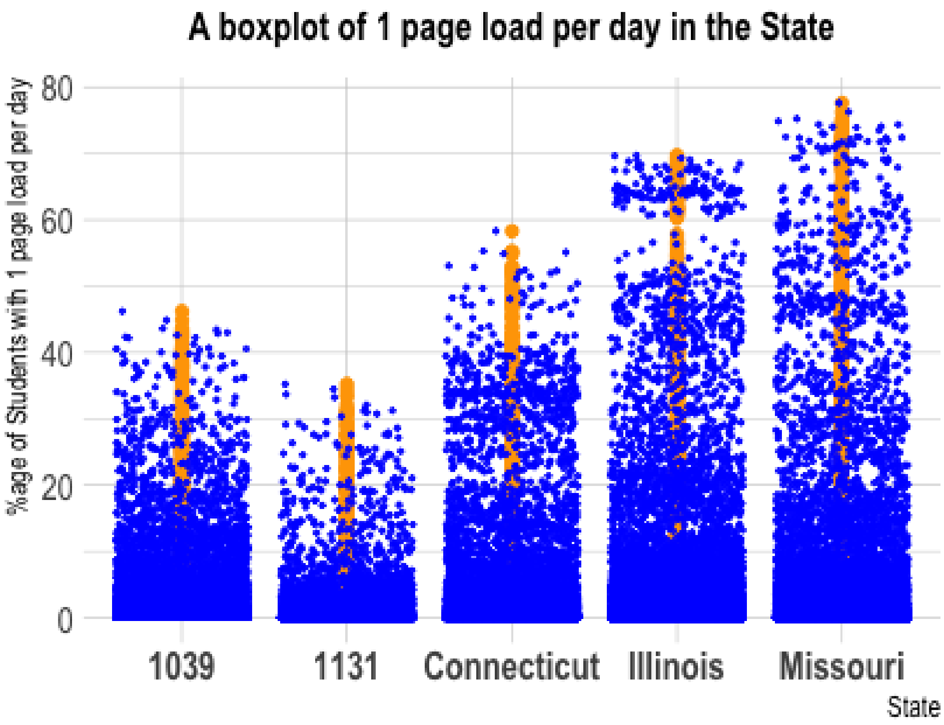
```
## [1] "character"

##    Length     Class      Mode 
##    266990 character character

## Warning: NAs introduced by coercion

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##     0.05     1.57     5.27   189.59    20.11 64818.66
```

**We can see there is a huge difference between the 3rd quartile and the maximum in the Pct_Access, so now we can consider them as the outliers. But since this is the page load, there is a chance that some pages can get unrealistic higher page loads in comparison to very less used pages. Therefore, despitethe fact that the column has uneven variability in the engagement and the page load access, we will keep this uneven variability in the dataset. Let’s check up on the NA’s in the Engage dataset.*

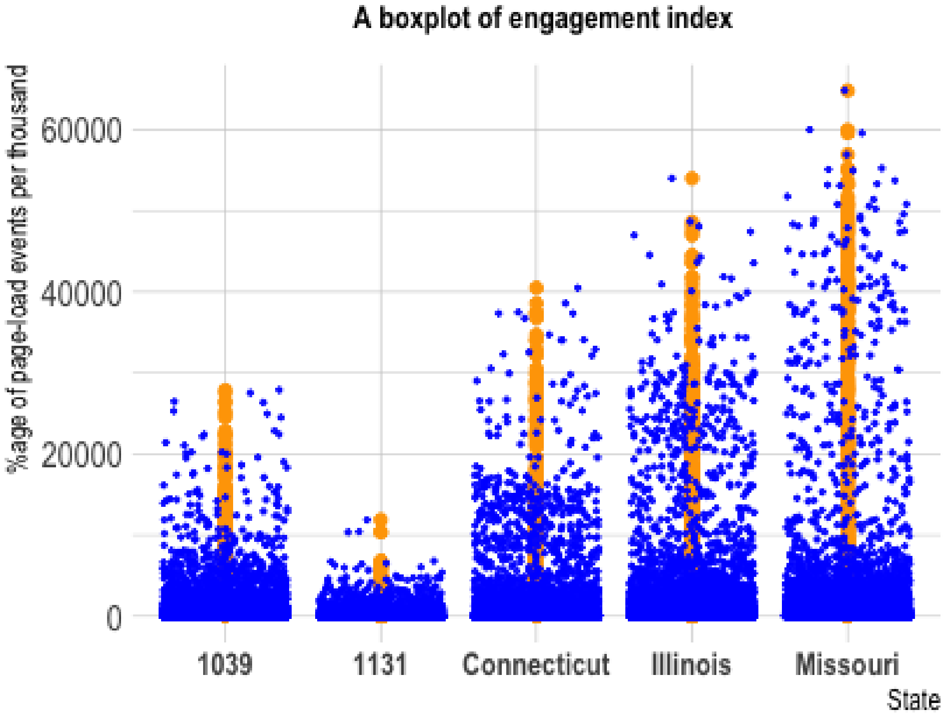
```
##                 time                lp_id           pct_access 
##                    0                    0                    0 
##     engagement_index          district_id                state 
##                    0                    0                    0 
##               locale       minority_range           free_range 
##                    0                    0                    0 
##                Ratio       Expenseperpupil                  URL
##                     0                      0                      0
##          Product Name Provider/Company Name                 Sector
##                     0                      0                      0
##               Primary              Secondary                   Year
##                     0                      0                      0
##                   Day                  Month
##                     0                      0
```

Engage Dataset is now cleaned and let's look for the anlaysis through visualisation # Visualisation

Now, we will plot some graphs and do quick analysis of them .



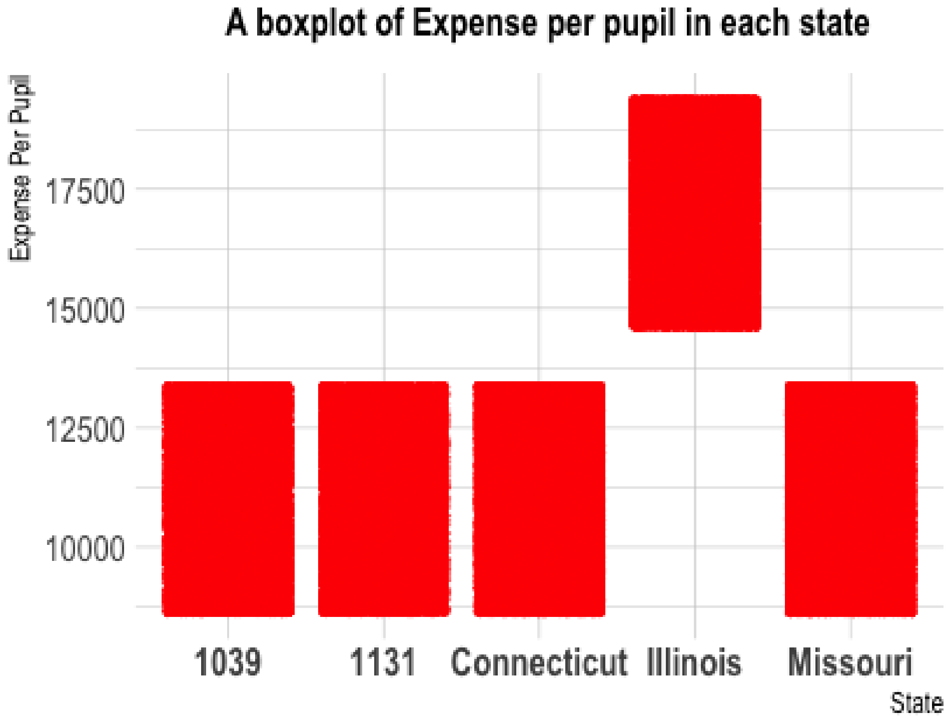
As per the above graph, we can easily say Missouri State has the highest percentage of students with 1 page load per day in the state. Therefore, the Missouri government has performed well in this state in terms of getting the page load in their district. Now, they should share their strategy with 1039, 1131 districts regarding the upliftment of page loads. 1 page load shows the good connectivity available in this state of Missouri and Illinois.



Similarly, as per the above graph, we can easily say Missouri State has the highest percentage of students with 1 page load per day in the state. Therefore, the Missouri government has performed well in this state in terms of getting the page load in their district. Now, they should share their strategy with 1039, 1131 districts regarding the upliftment of page loads. The engagement index also suggests the products used in Missouri were great, and other states can follow the same products for better results.

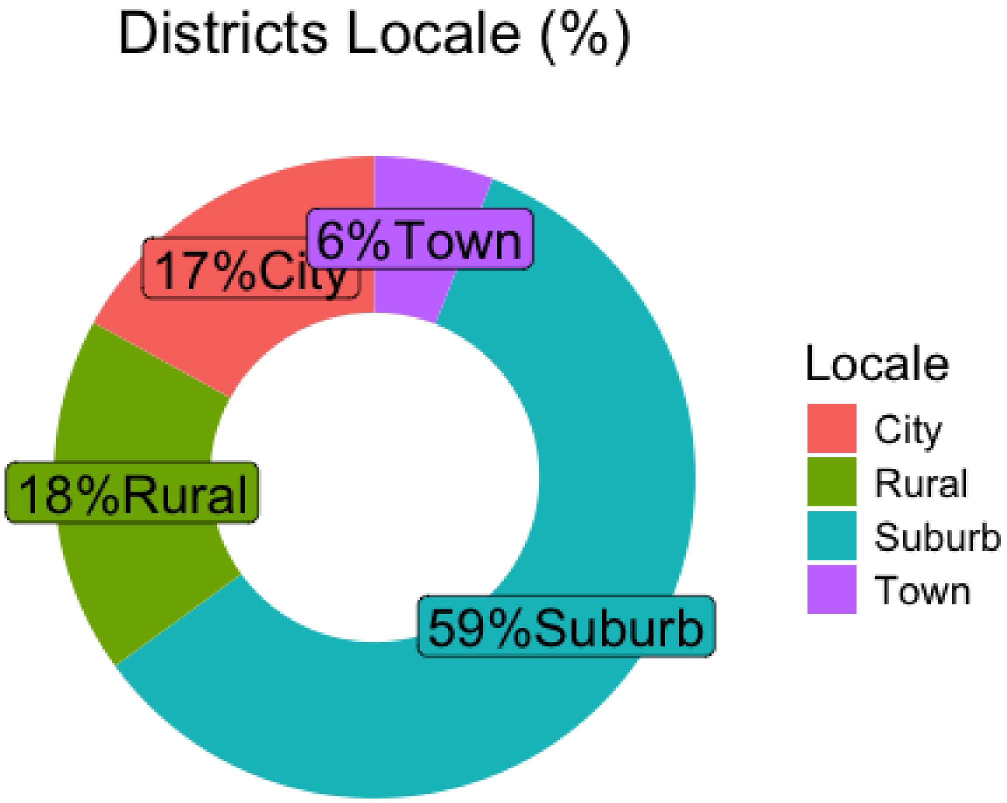
```
## # A tibble: 365 × 2
##    `Product Name`    count
##    <chr>             <int>
##  1 ABCya!              366
##  2 Big Ideas Math      366
##  3 BrainPOP            366
##  4 CNN Student News    366
##  5 Canvas              366
##  6 Chrome Web Store    366
##  7 Clever              366
##  8 CoolMath Games      366
##  9 Desmos              366
## 10 Disney+             366
## # ℹ 355 more rows
```

Therefore Missouri used ABCya!, Big Ideas Math product a lot therefore other states should learn from their products and try to implemnet these



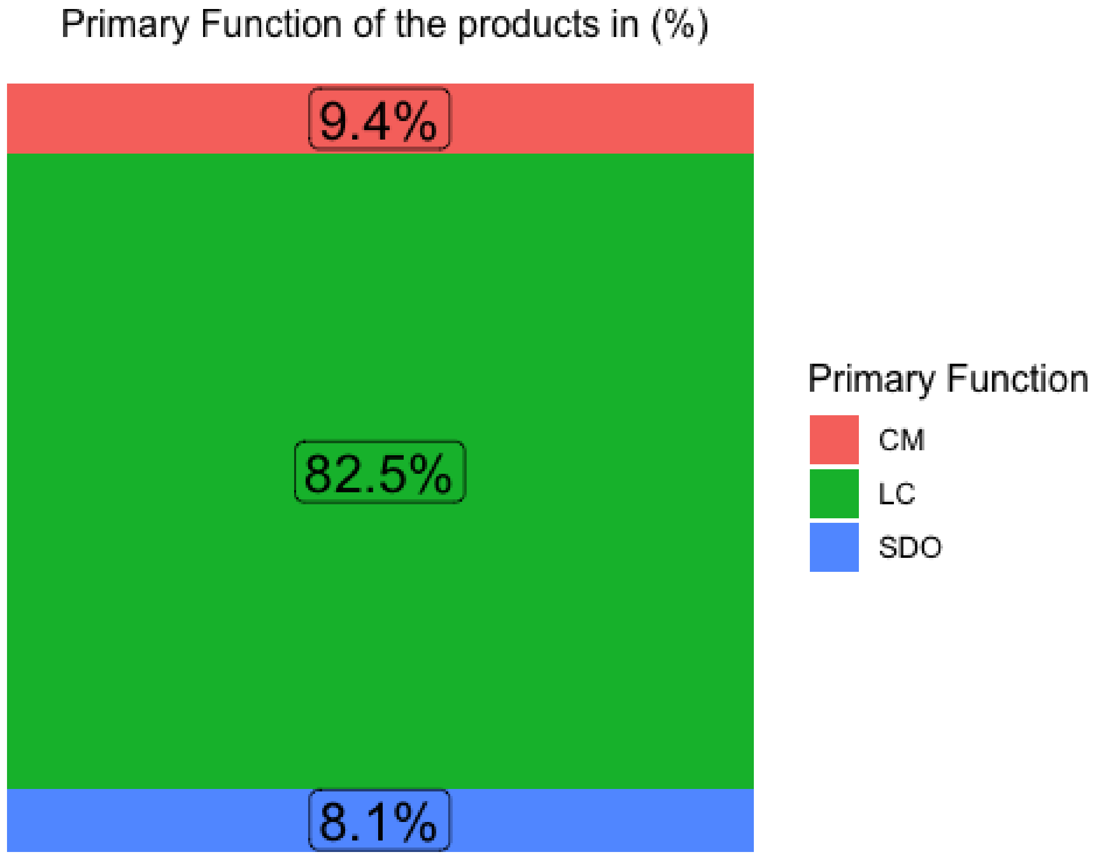
Per-pupil total expenditure (sum of local and federal expenditure) from Edunomics Lab's National Education Resource Database on Schools (NERD$) project. The expenditure data are school-by-school, and we use the median value to represent the expenditure of a given school district. As we can see, the Illinois government . in the respective district has done a lot of the expenditure. Further, they have gotten the results, as we can see that Illinois State has gotten a good amount of engagement as well. Therefore, if other states have the budget with them, they can consult Illinois

for a better strategy on how to spend the money to get better benefits.



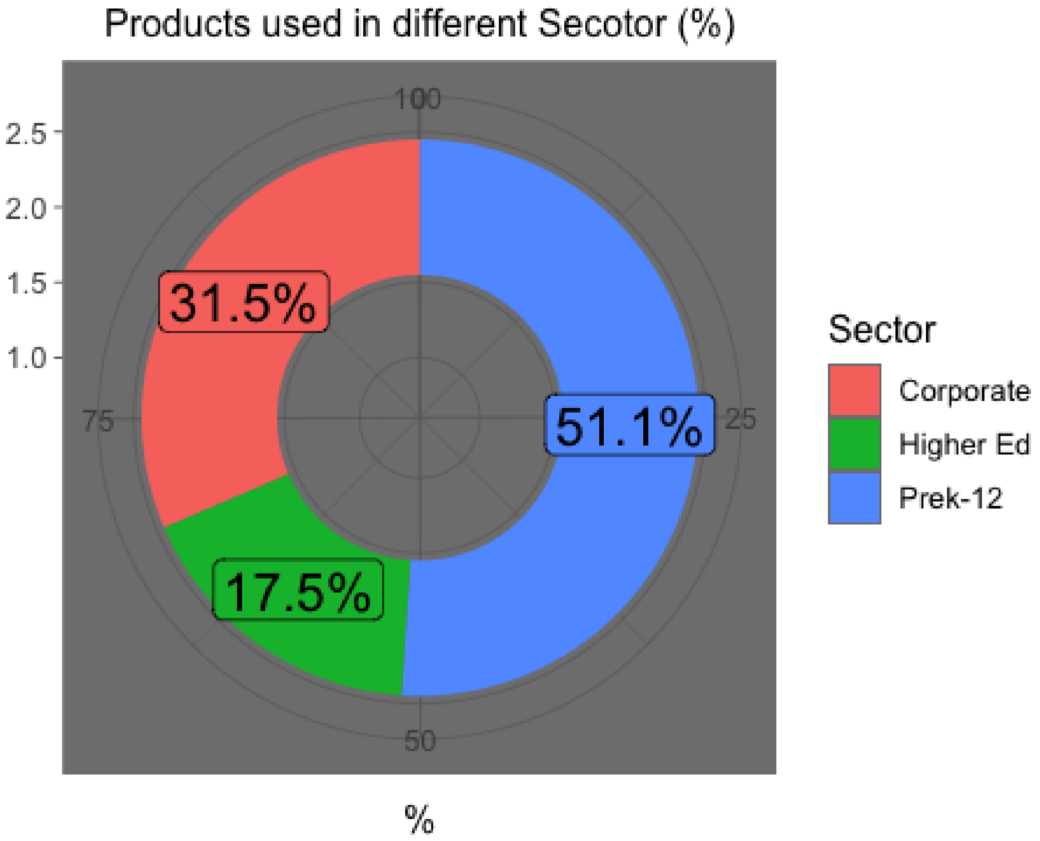
From the above graph, we can see the demographics of the entire USA: 59% of the districts are suburbs. Therefore, government planning for future digital empowerment should include suburbs strategists as well. When they are planning to take some bigger change or subsidy, they should open the gates for suburbs first.
18% rural is the second highest. Therefore, each small or big plan towards Didgita

should include representatives of suburbs and locales.

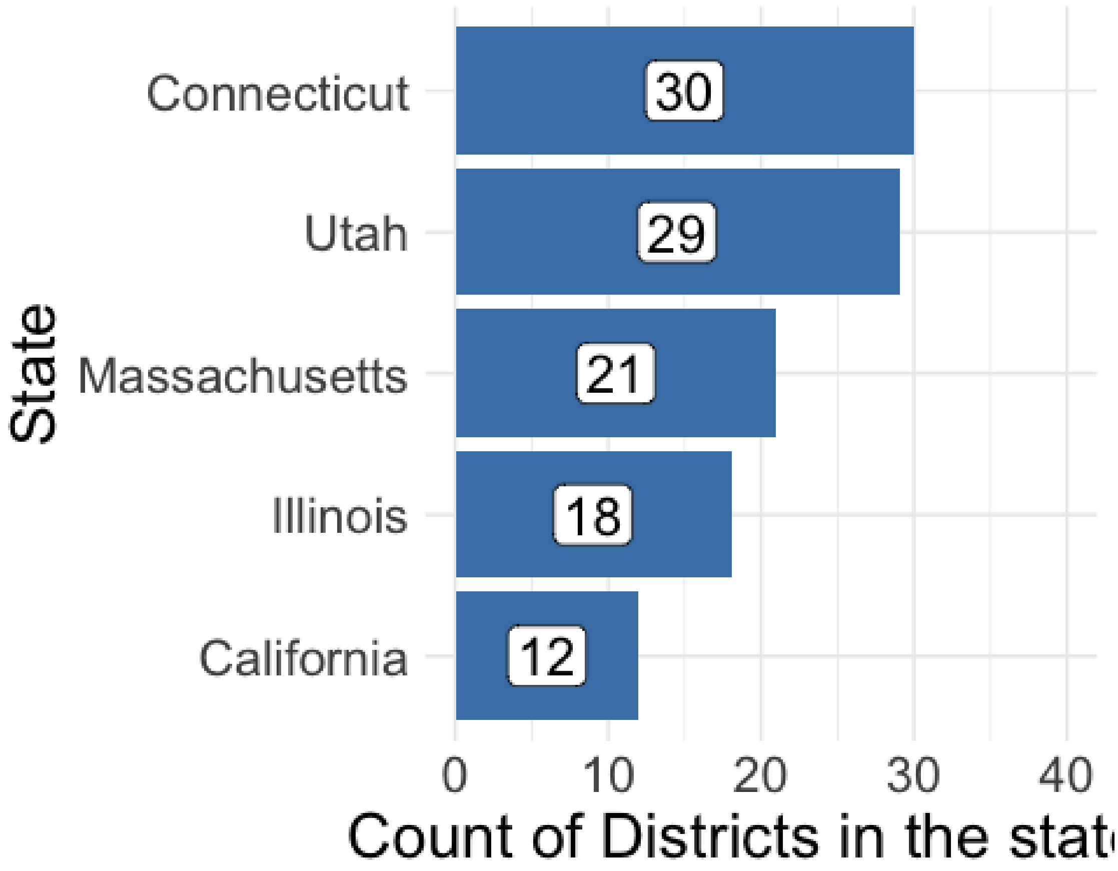


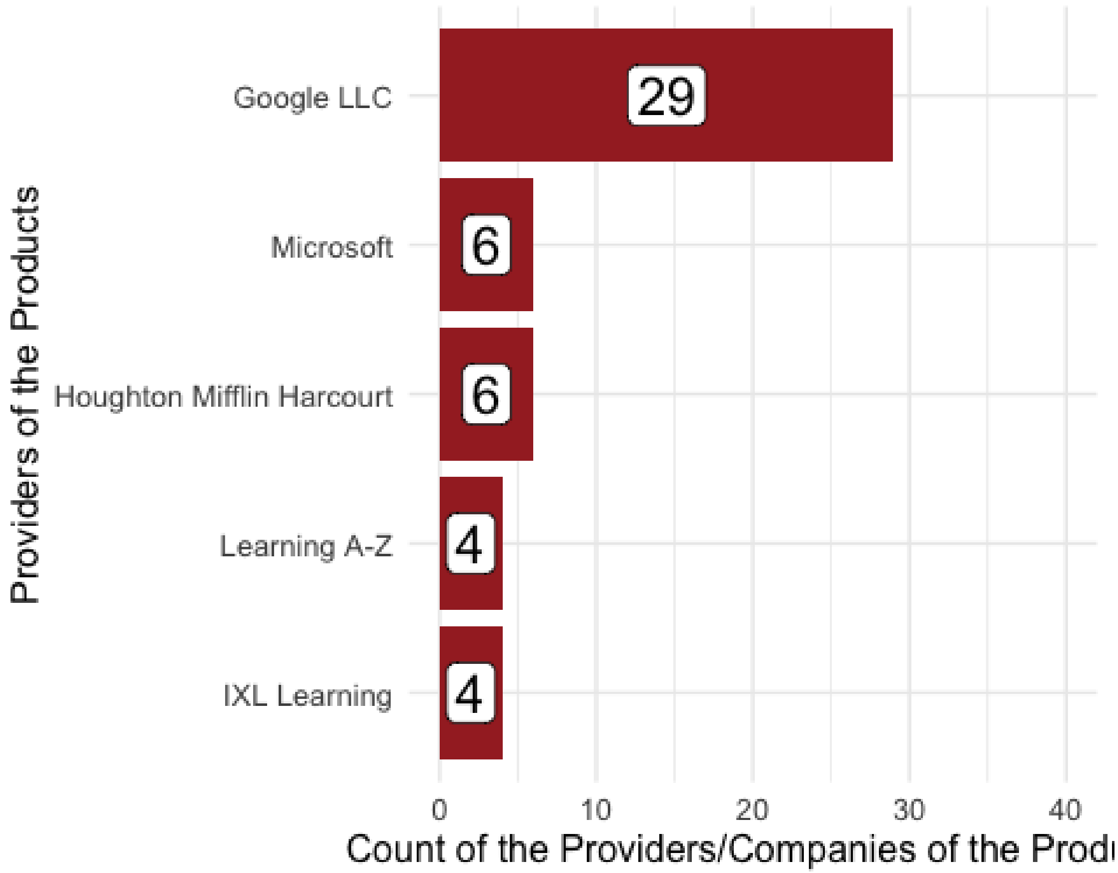
The basic function of the product. There are two layers of labels here. Products are first labeled as one of these three categories: LC = learning and curriculum, CM = classroom management, and SDO = school and district operations. As we can see, the majority of products were focused on LC, which is learning and curriculum. LC Products should be given more importance while designing the future strategy of

digital engagement, as in the past this function was very popular in the market.



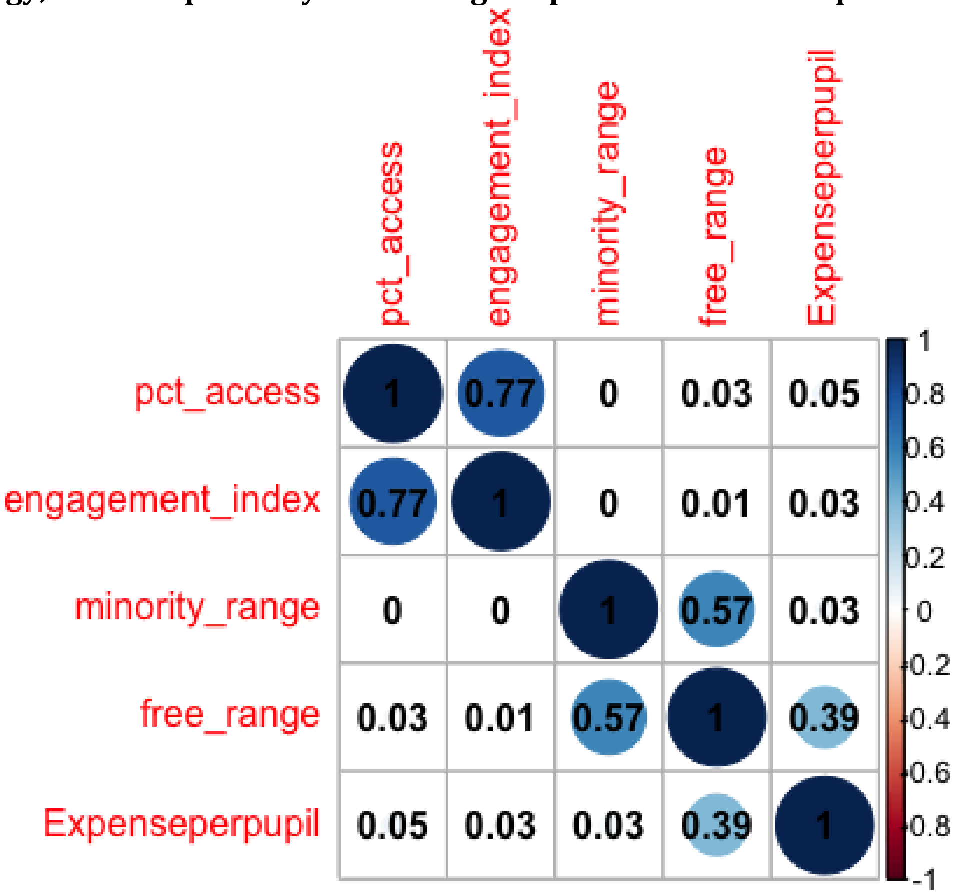
The chart's 'PreK-12' sector is very visible, indicating that the majority of items are designed with the kindergarten through twelfth grade (K–12) educational sector in mind. Therefore, we can say mostly students were using these products, as schools were a good part of the dataset. Therefore, those products were a part of them in a greater sense.


State
Connecticut
Utah
Massachusetts
Illinois
California
30
29
21
18
12
0
10
20
30
40
Count of Districts in the stat



This shows the top companies involved in the digital engagement. Google LLC has launched various product. Government should consult Google for future upcoming

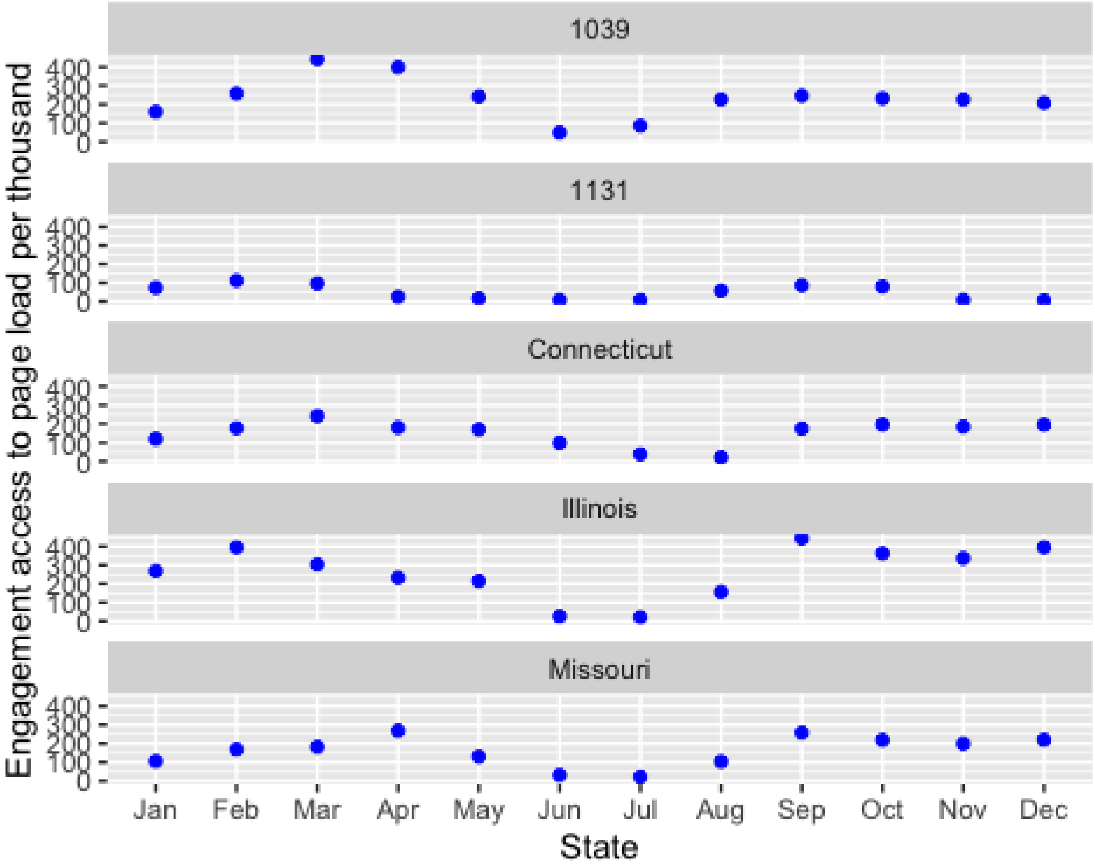
strategy, as in the past they are solving the problem with their products



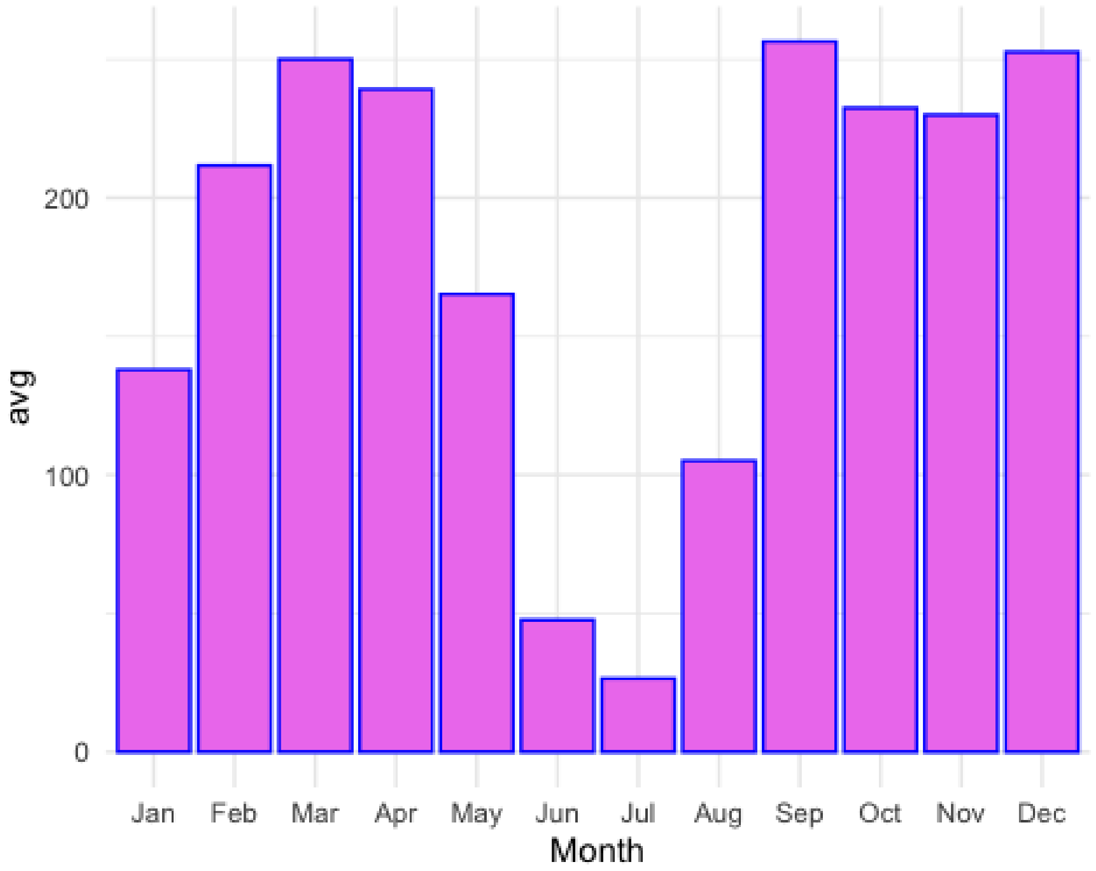
	pct_access	engagement_index	minority_range	free_range	Expenseperpupil
pct_access	1	0.77	0	0.03	0.05
engagement_index	0.77	1	0	0.01	0.03
minority_range	0	0	1	0.57	0.03
free_range	0.03	0.01	0.57	1	0.39
Expenseperpupil	0.05	0.03	0.03	0.39	1

The above is the correlation plot between all the numeric variables. We can see the strong relationship between engagement index and Pct_access, which is 0.77. It indicates that if there's a higher one-page load on any of the products, then there's a high chance of engagement as well. Therefore, if a district gets a good product, good connectivity, and a good strategy, they can excel in digital engagement. There's also the chance of a correlation between free range and minority range, which means areas with a higher number of minorities were given a higher percentage of free or

reduced services as well.



The above graph clearly indicates the trend is similar in all the districts, which means that in the month of June or July, the average decreases. But among all the states, we can say Connecticut has the lowest index throughout 2020



```
## $x
## [1] "Month"
##
## $y
## [1] "Average Engagement Index"
##
## $title
## [1] "Average Engagement by Month"
##
## attr(,"class")
## [1] "labels"
```

Now, as per the above graph, we can say the maximum engagement is in March and September. It can be because of the school students having exams during these months.Similarly, for the months of June and July, the engagement is very low. It's because of the holidays. Since the maximum product category is learning curriculum, we can conclude this is a summer holiday. Students don't have exams; therefore, they don't use products during their vacation.

```
##
##                 11000 17000
##   1039          27629     0
##   1131          22093     0
##   Connecticut 63354     0
##   Illinois        0 55637
##   Missouri    98277     0
```

By this we can say Illinois state has done the highest expenditure in it’s district as compared to other’s state

```
## 
##         1039  1131 Connecticut Illinois Missouri
##   1000     0     0       63354        0        0
##   1039 27629     0           0        0        0
##   1044     0     0           0        0    98277
##   1052     0     0           0    55637        0
##   1131     0 22093           0        0        0
```

Every District ID has gone the one state.Therefore, There is no discrepancies

```
## 
##                 0.1   0.3   0.7
##   1039        27629     0     0
##   1131        22093     0     0
##   Connecticut     0     0 63354
##   Illinois        0 55637     0
##   Missouri    98277     0     0
```

70% of Connecticut has the highest minority population. From the correlation graph, we can say they were offered a free lunch, but the school engagement index was very low in this area. Therefore, Connecticut State should improve their strategy and look for better engagement, either by products or from Missouri

```
## 
##                 0.1   0.3
##   1039            0 27629
##   1131            0 22093
##   Connecticut     0 63354
##   Illinois        0 55637
##   Missouri    98277     0
```

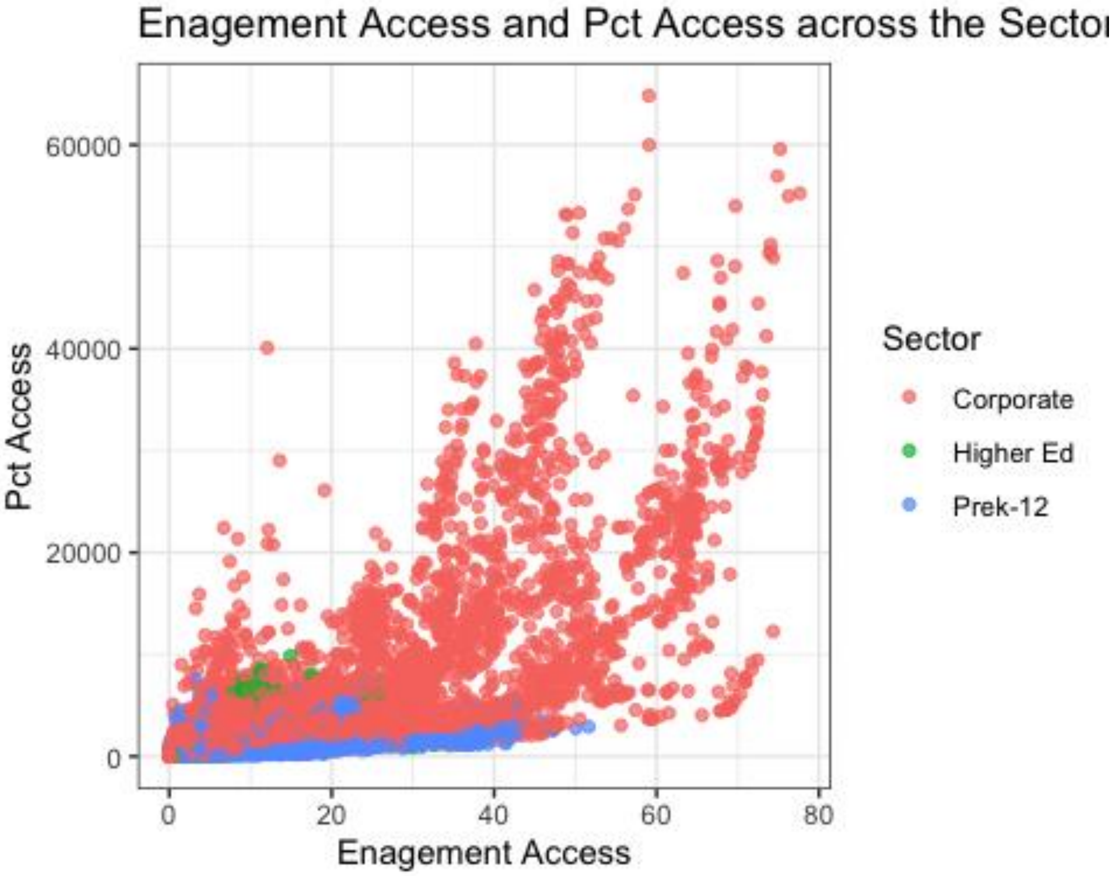
Missouri was given less of the free lunch in comparison to the other districts

```
## 
##         0.18
##   1000 63354
##   1039 27629
##   1044 98277
##   1052 55637
##   1131 22093
```

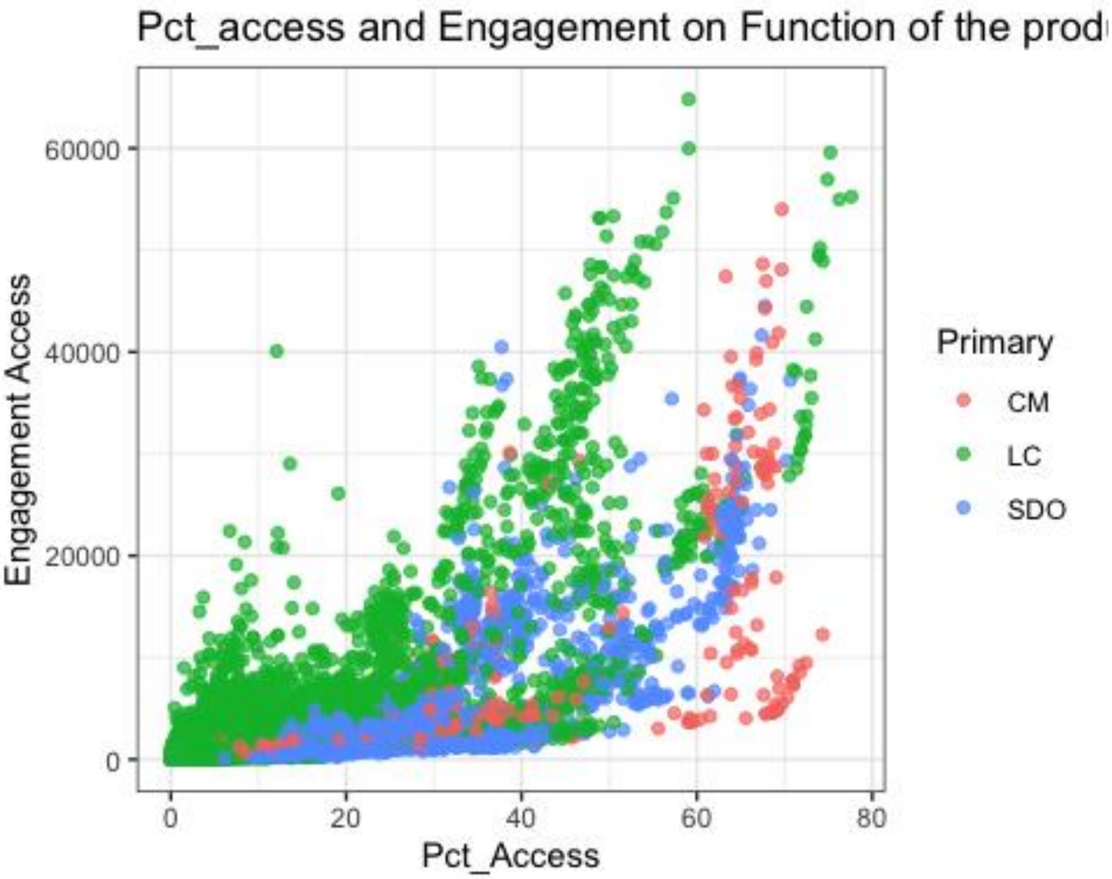
All the districts had the same connectivity ratio (residential fixed high-speed connections over 200 kbps in at least one direction/household

```
##
##        Suburb
##   1000  63354
##   1039  27629
##   1044  98277
##   1052  55637
##   1131  22093
```

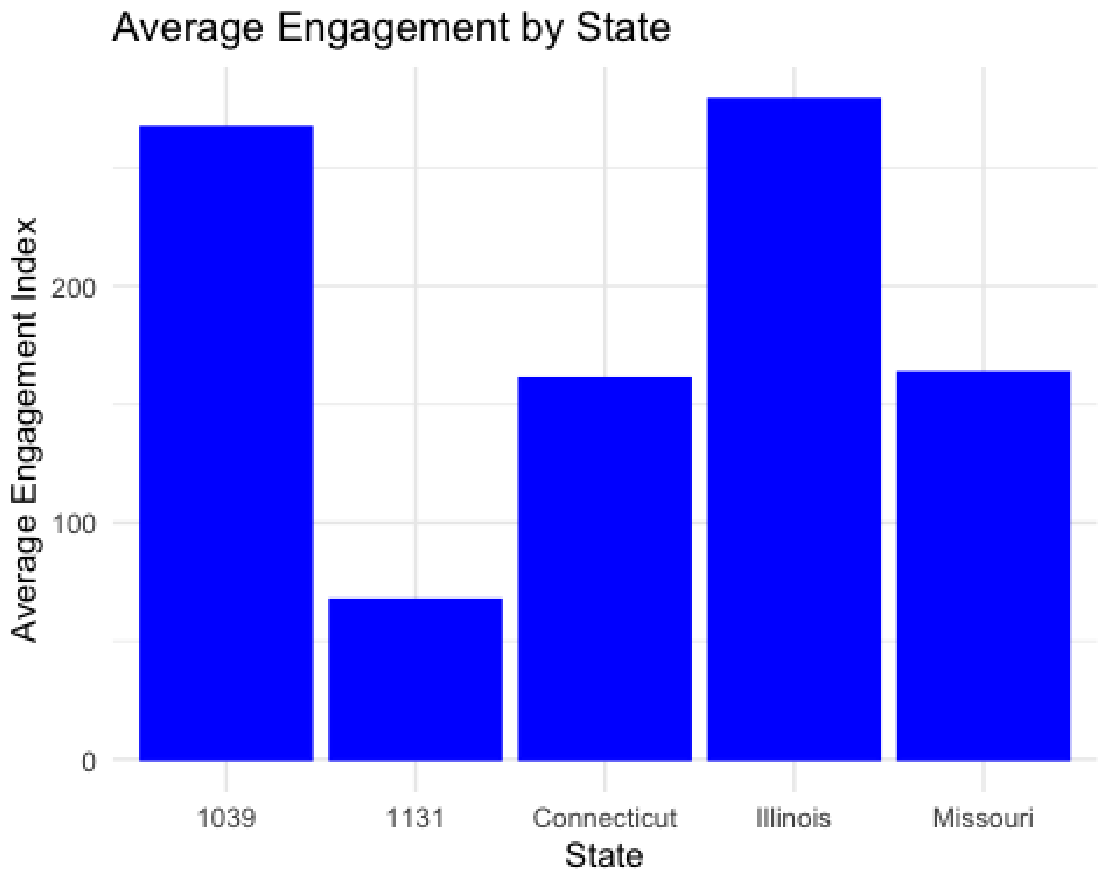
ALL the enaged data had the Suburban region . So, no other classification or analysis can be done in this.



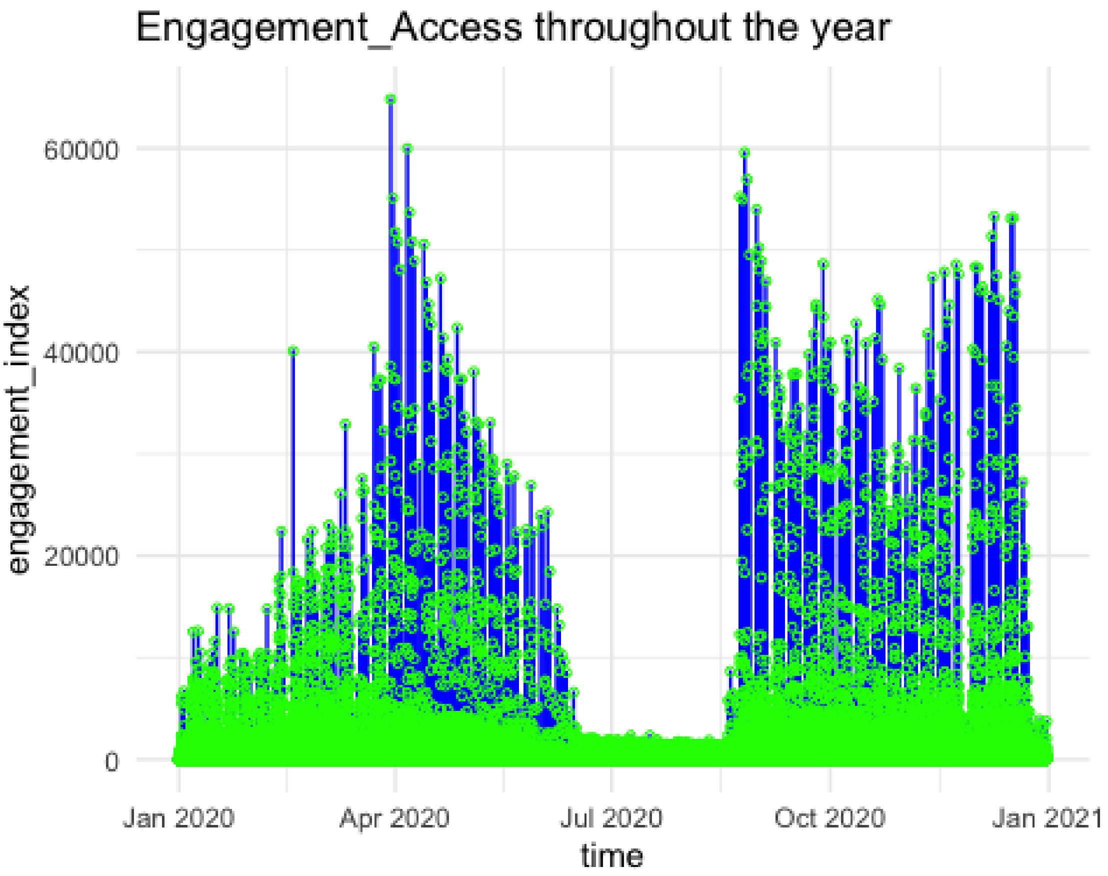
We can conclude this higher the Pct_Access, The engagement index will be higher. Towards the higher use of the page loads or engagement index, we can see more of the engagement was from the corporate, but the data had more of the percentage of the data from prek–12. Therefore, we can say that despite the lower number of corporate employees, we can assume corporate people have the products on which they have spent a lot of time in comparison to prek-12.



From the above graph we can say that Learning And Curricullum as grows rapidly as soon as it takes the page load of 40% people in the district it goes viral. Therefore, Government should invest on LC type product, as they are going viral



Therefore we can say that 1039 and Illinois state has the maximum engagement, whereas 1131 has the lowest state



** Above Graph shows the engagement index throughout the year. It’s the continous chart the explanation of month is already provided in the above **

** Recommendation are given below each of the graph as per the analysis **