MTH 765P Mini-project : COVID-19 Impact on Digital Learning

KANCHAN KUMARI

1 Introduction

Over the past decades significant progress has been made globally on the way to more equitable education. However, these achievements have been at stake since the outbreak of COVID-19 and the associated lockdowns and school closures that had been implemented to limit the spread of the disease.

The COVID-19 Pandemic has disrupted learning for more than 56 million students in the United States. In the Spring of 2020, most states and local governments across the U.S. closed educational institutions to stop the spread of the virus. In response, schools and teachers have attempted to reach students remotely through distance learning tools and digital platforms. In this project we try to analyze how COVID-19 has affected digital learning and how experience of digital learning relates to factors such as district demographics, broadband access, state-level income, and state/national level policies and events.

2 Understanding the Data

The Data set for this project was acquired from Kaggle.

`https://www.kaggle.com/c/learnplatform-covid19-impact-on-digital-learning/data`

The data set contains engagement data for 233 school districts in USA in 2020.

We are provided with mainly 3 data sets: The "districts-info" file contains information about each school district and the "products-info" file contains information about the top 370 different tools used for digital learning. For each school district, there is an additional file that contains the "engagement-data" for each tool for everyday in 2020.

2.1 District Data

The district file districts-info.csv includes information about the characteristics of school districts. The file contains following columns-

- district-id - The unique identifier of the school district
- state - The state where the district resides in.
- locale - NCES locale classification that categorizes U.S. territory into four types of areas: City, Suburban, Town, and Rural.
- pct-black/hispanic - Percentage of students in the districts identified as Black or Hispanic based on 2018-19 NCES data.
- pct-free/reduced - Percentage of students in the districts eligible for free or reduced-price lunch based on 2018-19 NCES data.
- county-connections-ratio - ratio (residential fixed high-speed connections over 200 kbps in at least one direction/households)

  pp-total-raw - Per-pupil total expenditure (sum of local and federal expenditure) from Edunomics Lab's National Education Resource Database on Schools (NERDS) project. The expenditure data are school-by-school, and we use the median value to represent the expenditure of a given school district.

2.2 Products Data

The product file products-info.csv includes information about the characteristics of the top 372 products with most users in 2020. The file contains following columns-

- LP ID- The unique identifier of the product.
- URL- Web Link to the specific product.
- Product Name- Name of the specific product.
- Provider/Company Name- Name of the product provider.
- Sector(s)- Sector of education where the product is used.
- 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

2.3 Engagement Data

The engagement data are aggregated at school district level, and each file in the folder engagement-data represents data from one school district. The 4-digit file name represents district-id which can be used to link to district information in district-info.csv. The lp-id can be used to link to product information in product-info.csv. Each file in this folder has following columns:

- time - date in "YYYY-MM-DD" format
- lp-id - The unique identifier of the product.
- pct-access - Percentage of students in the district who have at least one page-load event of a given product and on a given day.
- engagement-index - Total page-load events per one thousand students of a given product and on a given day.

3 Objectives

Following are the key questions we will try to answer through the analysis of this data-

- What is the picture of digital connectivity and engagement in 2020?
- What is the effect of the COVID-19 pandemic on online and distance learning, and how might this also evolve in the future?
- How does student engagement with different types of education technology change over the course of the pandemic?
- How does student engagement with online learning technologies relate to different location-specific characteristics?
- How the student engagement relates to factors such as district demographics, broadband access, per pupil expenditure by state, ethnicity and socio-economic status ?

4 Analysis

For data analysis main python libraries used are- Pandas, Matplotlib and Seaborn.

⇒ The following steps were conducted to clean data:

- Remove districts with no state information.
- Remove districts with incomplete 2020 engagement data.
- Remove rows in engagement-data that are not in products-info. There are 8417 unique products in engagement data, but only 369 unique products in products-info, since we don't have any additional information about the majority of these products, we remove engagement data for these products.
- Clean values in 'Primary Essential Function' Eg.- make 'Sites, Resources References' and 'Sites, Resources Reference' into one.

⇒ The following feature engineering steps were conducted:

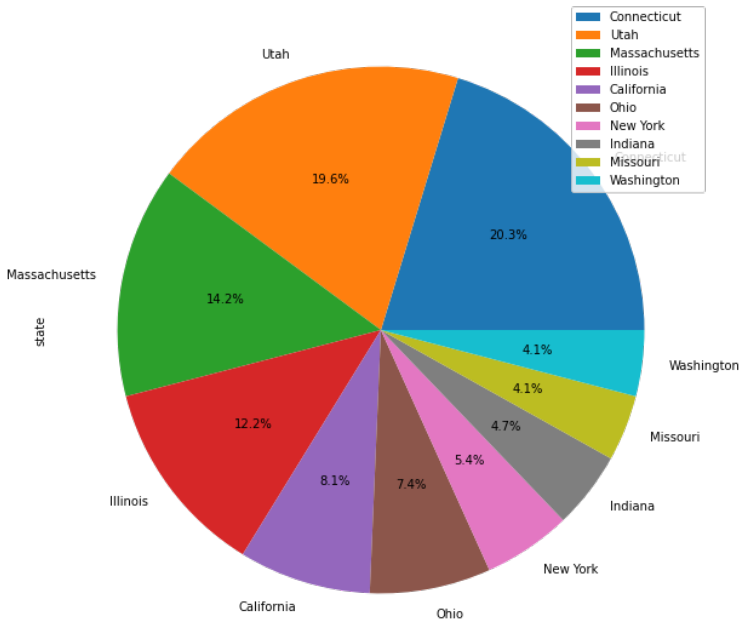
- Split columns 'Primary Essential Function' into main and sub functions
- Get dummy variables for sectors - Corporate Sector, Higher Education Sector, PreK-12 Sector and Create 3 new columns for each sector.
- Convert the 'time' column into datetime for engagement data
- Create new columns 'weekday' for engagement data.
- Filter only weekday data.
- Create new "Quarter "column using datetime function in python and visualize the change in percentage access and engagement from Quarter 1 (pre-covid) to Quarter 4 (after second wave).
- Virtual classroom products are best indicators for good educational experience . Filter these products and visualize how engagement and pct-access changed for them, from quarter 1 to quarter 4 in year 2020.

5 Findings

The following insights were derived from data:

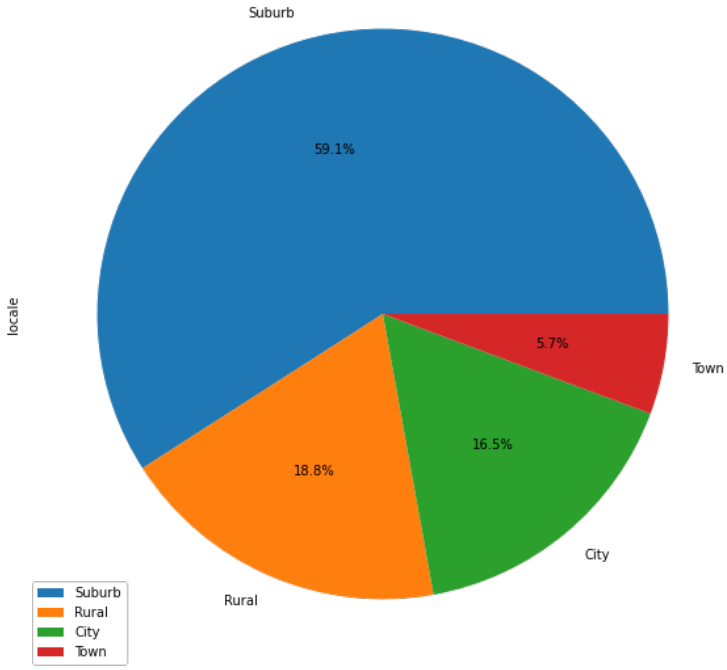
1. State wise distribution of given data:

   We are provided with data from 233 school districts around the USA. For every school district we are given the state it belongs to, and we can see from the graph below that we are given most school districts from Connecticut followed by Utah and Massachusetts.



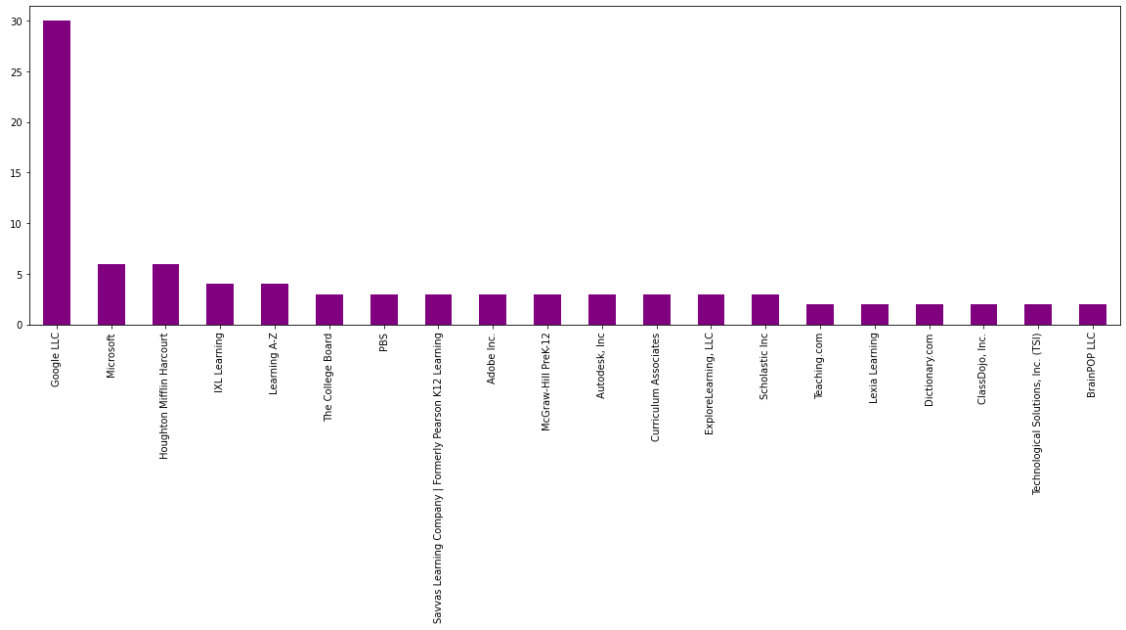
2. Distribution of locale - (suburban, rural, city, town):

   According to NCES, US territory is divided in 4 locales and we are given for every district which locale it belongs to. It is clear from the graph below that for our data most districts belong to suburban areas.

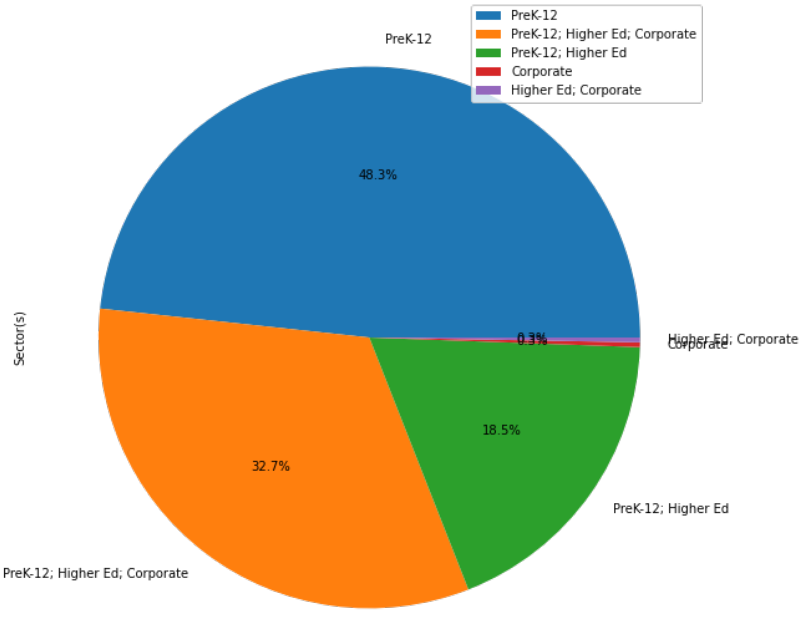


3. Companies that are providing products for online education:

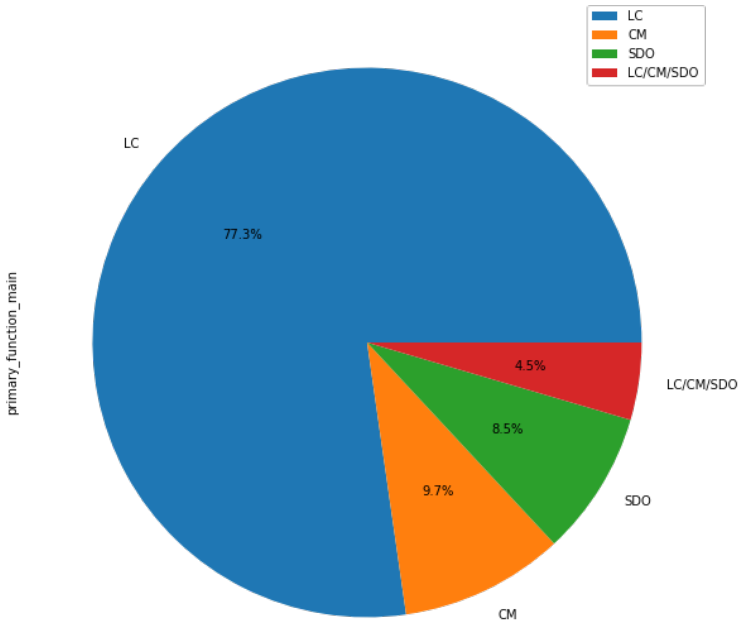
   It is clear from the graph below that Google provided the most learning products(33) followed by Microsoft (6), Houghton (6).



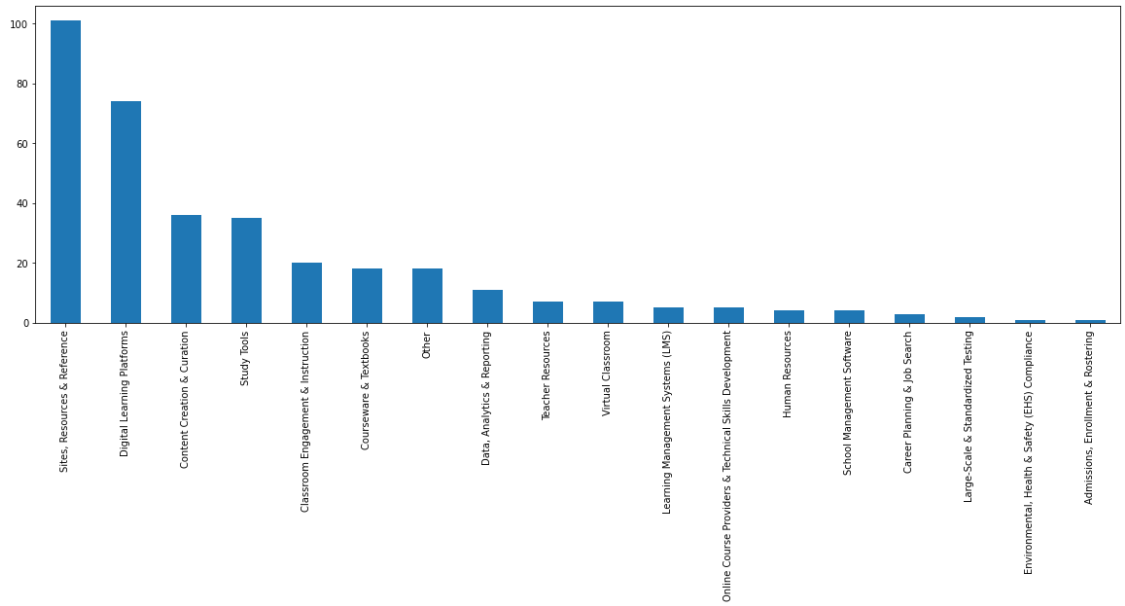
4. Sectors in which companies are active: The graph below shows that most digital learning providers are active in sector PreK-12 (48.3%+32.7%+18.5%=99.5)and very few(0.5%) in the Corporate sector.



5. Primary Function (main) of the products provided by different companies:
   LC = Learning Curriculum, CM = Classroom Management, and SDO = School District Operations

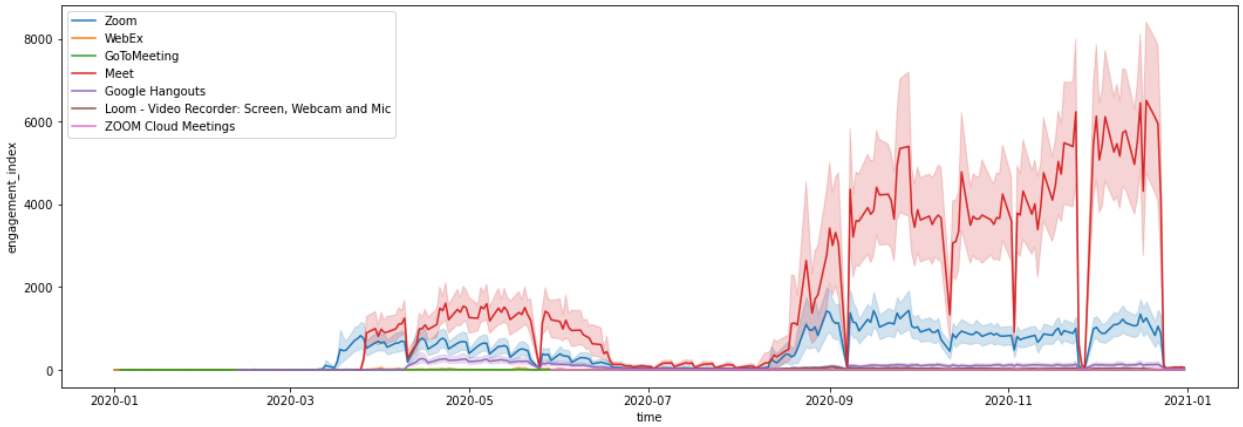


6. Primary Function (Sub) of the products:
   For every product their sub function is also defined which is sub category of main function, distribution of which is shown in graph below:



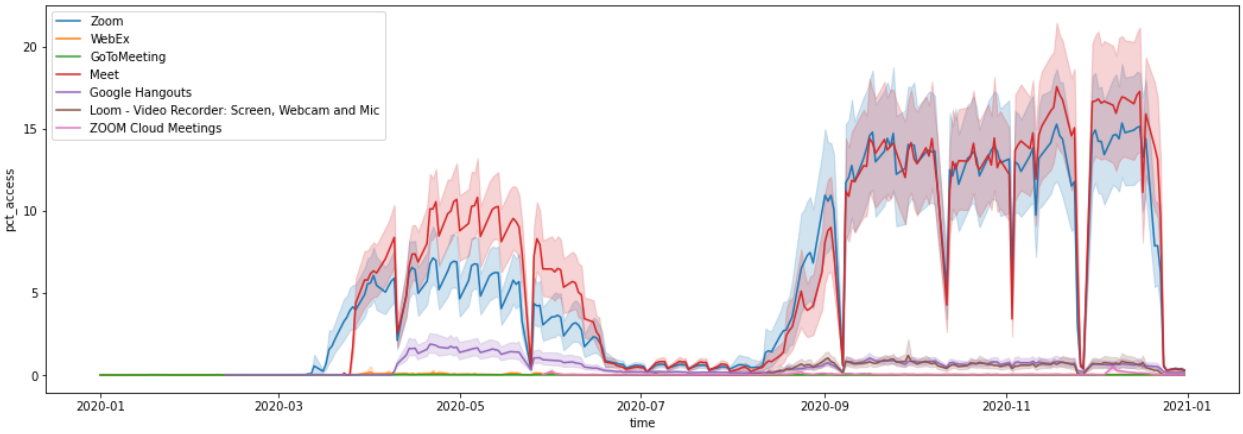
7. Engagement of Virtual classroom Provider products over given period of time :

   Engagement is defined as - Total page-load events per one thousand students of a given product and on a given day



8. Pct-access for Virtual classroom Provider products over given period of time:

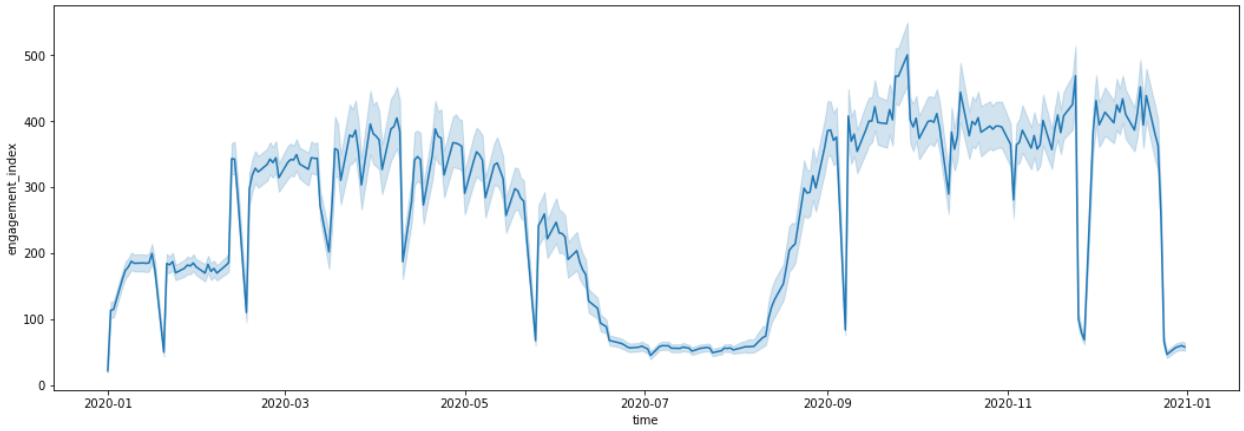
   Pct-access is defined as- Percentage of students in the district who have at least one page-load event of a given product and on a given day.



   We can observe from the graphs above that meet, zoom and hangouts are the most prominent products here and show the most engagement and pct-access.

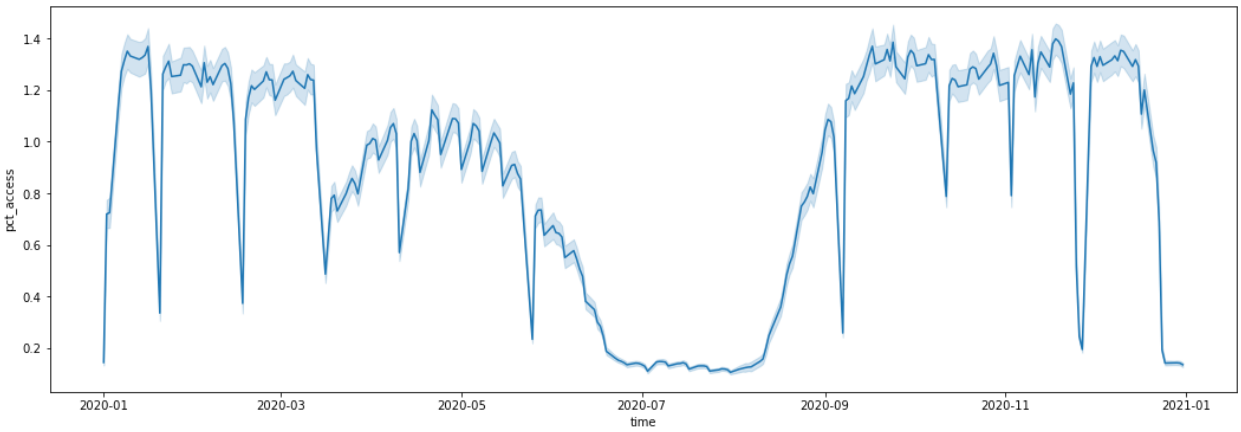
9. Engagement over the given time period:

   In the graph below there is a clear jump in engagement index around march 2020 (Start of pandemic). There is a dip around July 2020 due to summer holidays and then it increases again.There are random dips in the graph at a few places , which are probably because of national holidays/bank holidays. It is also clear form the graph that engagement is higher in second half of 2020.

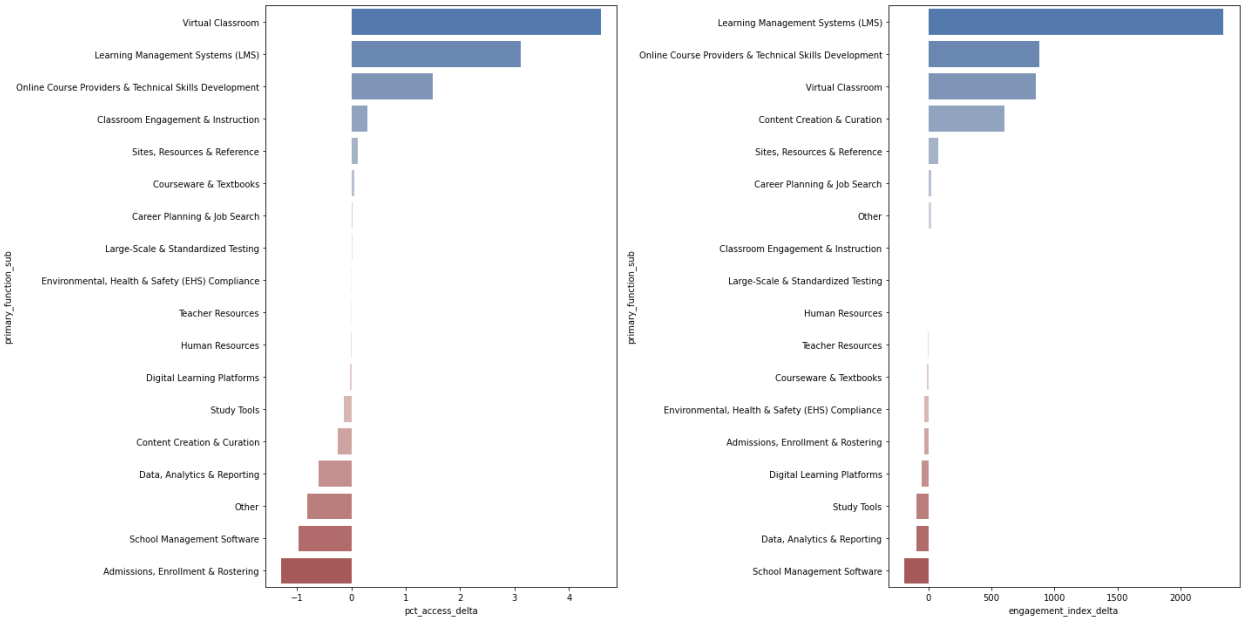


10. Pct-access over the given time period:

    From the graph below, we can see that march 2020 was the start of Covid-19 pandemic which reduced pct-access a bit. July 2020-august 2020 was summer holidays and that is why pct-access is close to zero around that time and after that it has increased. There are random dips in the graph at a few places , which are probably due to national holidays/bank holidays.

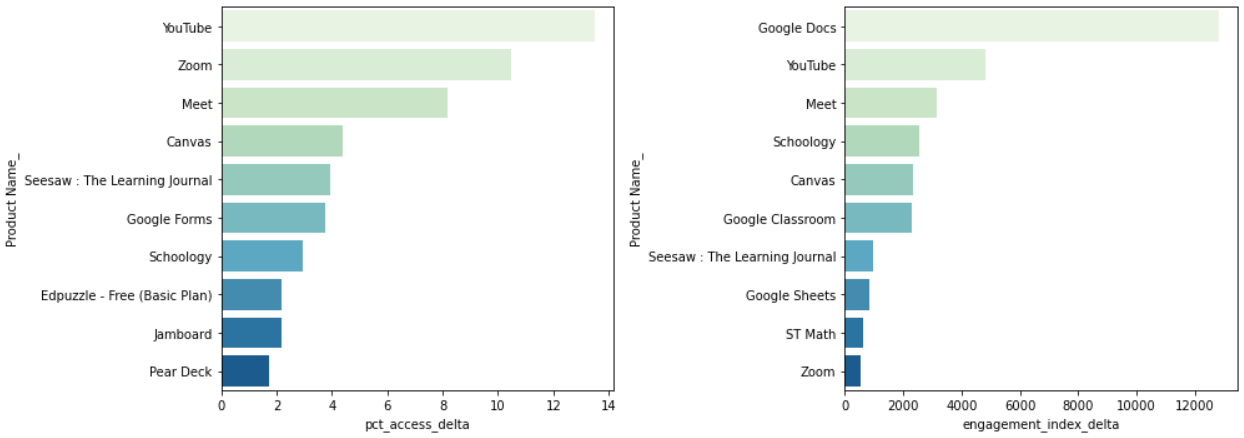


11. Change in pct-access and engagement index from first quarter of 2020 (before Covid) to last quarter of 2020(after second wave of Covid) of primary function sub of products:



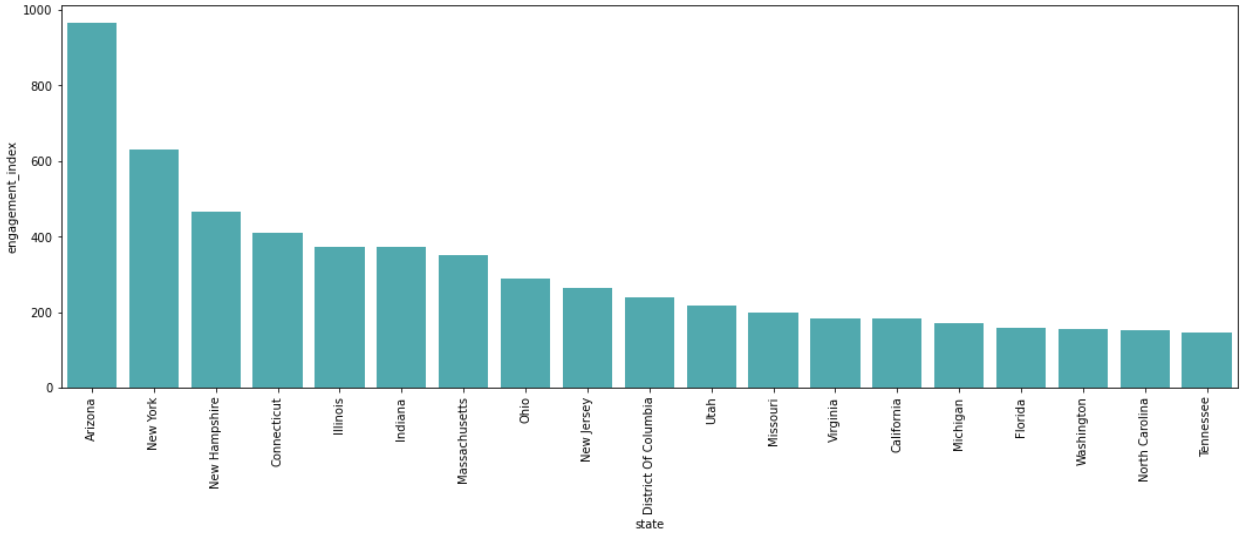
12. Change in pct-access and engagement index from first quarter of 2020 (before Covid) to last quarter of 2020(after second wave of Covid) of products:

    Here Youtube is in top in both graphs, but that could be because it was used for entertainment purposes as well.



13. Engagement Index (mean) for Different States:

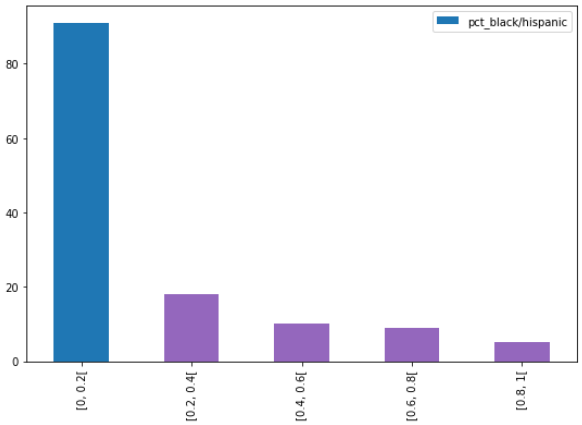
    Arizona and Newyork have the highest engagement index among all the states.

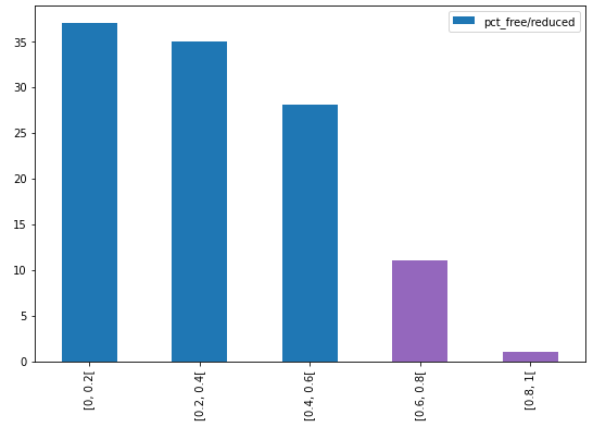


14. No. of districts vs pct-black/hispanic and pct-free/reduced:

    We can see that most of the given district have black/hispanic percentage between range 0-20%. Very few district have the black/hispanic percentage between 80-100%.

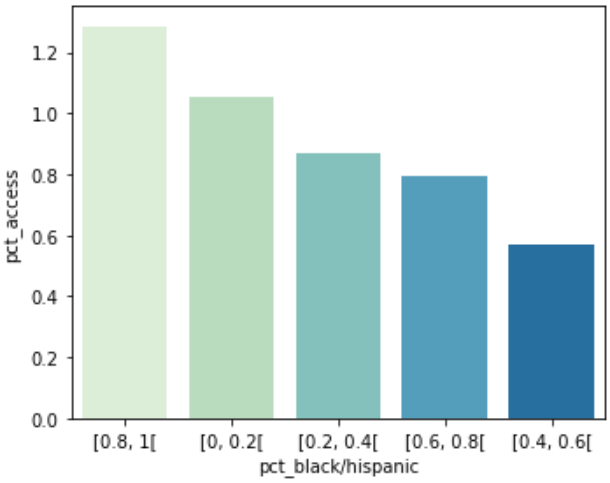
    For Percentage of students eligible for free or reduced-price lunch, we can see that very few districts have percentage between 80-100%.

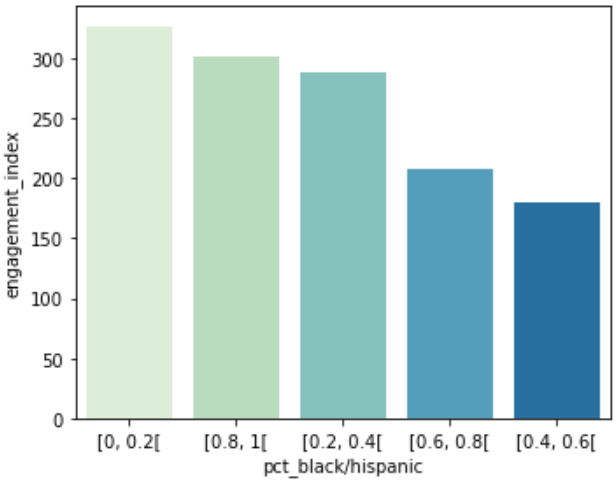




15. Pct-access and Engagement vs Percentage of students in the districts identified as Black or Hispanic-

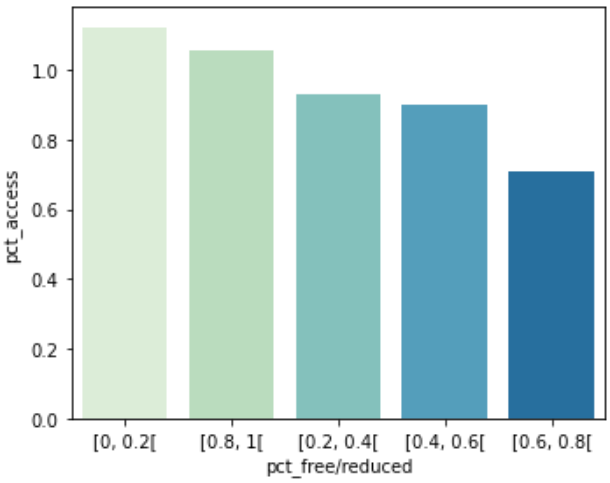
    We are getting districts with the most black/hispanic % in top here because we are given very few districts in that bin. Rest, We can observe that districts with least black/hispanic student percentage have the maximum engagement index and pct-access, and districts with more black/hispanic student % have the least engagement and pct-access, and are the most vulnerable.

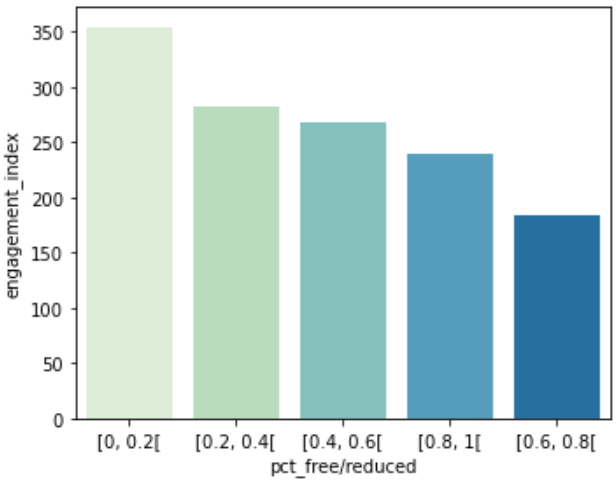




16. Pct-access and Engagement vs Percentage of students in the districts eligible for free or reduced-price lunch:

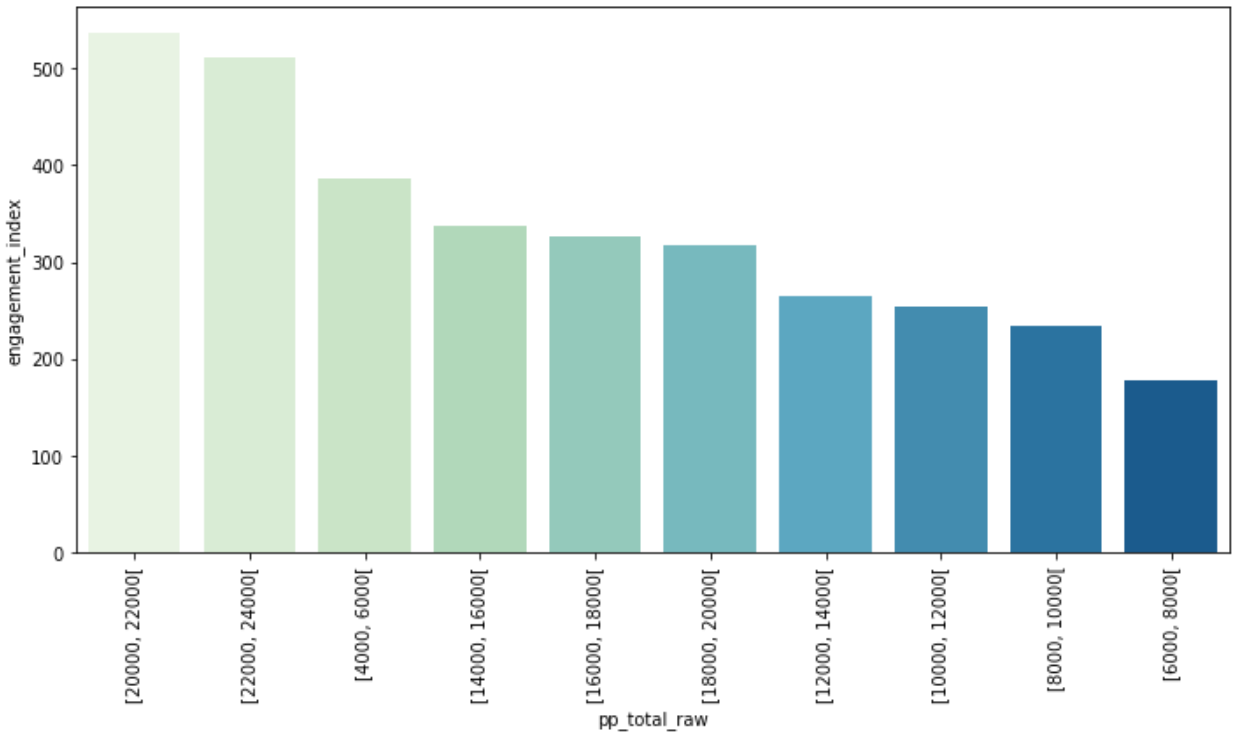
    There is an inverse correlation between the percentage of students who are eligible for free lunch and the engagement index. The less the pct-free/reduced the more the engagement and pct-access. We see the that 80-100 % bin is also in the top but that is because we are given very few districts in that bin.





17. Engagement index vs Per-pupil total expenditure :

    We can observe that there is a direct relationship between engagement and per-pupil total expenditure. The more the expenditure the more the engagement.



6 Conclusion

From above analysis we have found some interesting insights regarding the given data. We can see that digital learning engagement is affected by black/Hispanic percentage, socio-economic status, per-pupil expenditure by state and state itself. It is also clear that engagement is higher in second half of 2020.

Further analysis can be conducted using publicly available data on COVID-19. How much a school district was affected by Covid-19 will also effect the state of digital learning. Also the data available for county-connection-ratio was only for one class[0.18,1], if we can get more data for different class, further analysis can be done. We only analyzed engagement for virtual classroom products, engagement for other products can also be analysed to get a better understanding of state of digital learning.