Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Impact of Demographics on Product Engagement

Programme	: B.Tech
Course Title	: Data Visualization
Course Code	: CSE3020
Slot	: D1
Faculty	: Prof. Joshan

Team Members:

Sakshi Saraogi - 20BCE1378

Saanjhi Saraogi - 20BCE1350

CONTENTS

ABSTRACT

The COVID-19 pandemic has brought about unprecedented changes to the global economy and society. The impact of the pandemic has been felt in almost all aspects of life, including education. The educational structure, which was largely based on physical classrooms and in-person interactions, had to adapt quickly to the new reality of remote learning in order to continue providing education to students. This shift to digital learning has highlighted the impact of demographic factors on educational systems, both before and after the pandemic.

This project aims to analyze the impact of demographic factors on the educational system before, through and after COVID-19. By examining the ways in which different demographic factors, such as socio-economic status, geographic location, and age, have affected the educational system, the project seeks to identify areas of potential improvement and innovation. Through this project, we will analyze how the pandemic has accelerated the shift towards digital learning and how it has affected different demographic groups. We will also examine how educational institutions and systems have adapted to the challenges posed by the pandemic and identify areas where improvements can be made to better support students.

By analyzing the impact of demographic factors on the educational system, this project can provide valuable insights into the inequalities and challenges faced by different demographic groups. This information can be used to inform policy and institutional changes that address these issues and create a more equitable educational system. Ultimately, this project aims to contribute to the development of a more inclusive and effective educational system that meets the needs of all students, regardless of their demographic background.

Keywords: *Demographics, statistics, COVID, predict, hypothesized.*

OBJECTIVES OF THE PROJECT

The objective is to explore:

1) the state of digital learning before, after and through COVID
2) How demography has an impact on digital learning

We have guided the analysis with questions that are related to the themes thatare described above (in bold font). Some questions that relate to our problem statement include:

- What is the picture of digital connectivity and engagement?
- 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 platforms relate to varied demography? (e.g., race/ethnicity, ESL, learning disability)?
- Does certain state interventions, practices, or policies (e.g., reopening, eviction) correlate with the increase or decrease of online engagement?

METHODS USED IN THE PROJECT

The following methods can be used in this project:

- Data Collection: The project can involve the collection of data from digital learning platforms, such as the number of logins, time spent on the platform, completion of assignments and assessments, and interaction with teachers and peers. The data can be collected through various means such as API access or data scraping.

- Statistical Analysis: The collected data can be analyzed using statistical methods such as regression analysis, correlation analysis, and descriptive statistics to identify patterns and trends in engagement across different demographic groups.

- Literature Review: The project can involve a review of existing literature on the impact of demographic factors on digital learning engagement. This can help provide a theoretical framework for the analysis and interpretation of the collected data.

- Visualization: Data visualization techniques such as graphs, charts, and heat maps can be used to present the analysis findings in a clear and concise manner.

The use of these methods can help provide a comprehensive understanding of the impact of demographic factors on digital learning engagement and inform the development of strategies and products to improve engagement and support for underrepresented students.

OUTCOME OF THE PROJECT

The outcome of this project is expected to provide valuable insights into the impact of demographic factors on digital learning engagement. By analyzing data from digital learning platforms, the project aims to identify patterns and trends in engagement across different demographic groups, such as age, gender, race, ethnicity, and socio-economic status.

The project's findings may highlight areas where underrepresented students require additional support to fully engage with digital learning platforms. This information can inform the development of targeted strategies to improve engagement and support for these students, ultimately contributing to reducing the educational gap between different socio-economic groups.

The project may also provide insights into the evolving needs and preferences of students and educators, as well as the development of new and innovative digital learning platforms that are tailored to meet these needs. This information can be used to inform the development of new digital learning products that better meet the needs of diverse student populations.

Additionally, the project aims to predict product engagement for the year 2023 based on trends and patterns identified in the analysis. This prediction can help guide the development and marketing of digital learning products, ensuring that they meet the needs and preferences of students and educators in the coming years.

Overall, the outcome of this project is expected to contribute to a better understanding of the impact of demographic factors on digital learning engagement, as well as informing the development of strategies and products to improve engagement and support for underrepresented students.

SCOPE OF THE PROJECT

The scope of this project is to analyze the impact of demographic data on digital learning, specifically product engagement. The project aims to investigate how demographic factors such as age, gender, race, ethnicity, and socio-economic status influence student engagement with digital learning platforms.

The project will involve collecting and analyzing data from digital learning platforms, such as the number of logins, time spent on the platform, completion of assignments and assessments, and interaction with teachers and peers. The data will be analyzed using statistical methods to identify patterns and trends in engagement across different demographic groups.

The project will also involve a review of existing literature on the impact of demographic factors on digital learning engagement, as well as an analysis of education policies and funding that may affect access to technology and digital learning platforms.

The goal of this project is to provide insights and recommendations that can inform strategies to improve engagement and support for underrepresented students in digital learning. The project will also aim to predict product engagement for the year 2023 based on trends and patterns identified in the analysis.

The project's scope will be limited to digital learning platforms and will not cover other forms of remote learning such as television or radio programs. The project will also focus on the United States education system, although insights from other countries may be used as a comparison.

1. INTRODUCTION

The COVID-19 pandemic has had a profound impact on education worldwide, disrupting learning for over 56 million students. In the Spring of 2020, most states and local governments closed educational institutions to stop the spread of the virus. This forced schools and teachers to attempt to reach students remotely through distance learning tools and digital platforms. The sudden shift to online learning posed significant challenges for educators, students, and parents, and highlighted the importance of access to technology, internet connectivity, and digital literacy skills.

The pandemic and subsequent lockdowns have affected students from all walks of life, but the impact has not been equal. The digital divide has exacerbated existing inequalities in access to education, with students from low-income families and rural areas being disproportionately affected. Many students lack access to devices such as computers, laptops, and tablets, as well as reliable internet connectivity. This has made it difficult for them to participate in remote learning activities and has widened the educational gap between different socio-economic groups.

To understand the impact of demography on digital learning, it is important to analyze student engagement with digital learning platforms. This includes metrics such as the number of logins, time spent on the platform, completion of assignments and assessments, and interaction with teachers and peers. By analyzing these metrics, we can identify patterns and trends in engagement across different demographic groups and use this information to develop strategies to improve engagement and support for underrepresented students.

Predicting product engagement for the year 2023 requires a thorough understanding of the factors that influence student engagement with digital learning platforms. This includes analyzing trends in technology adoption, changes in education policy and funding, and the impact of the pandemic on education. It also requires an understanding of the evolving needs and preferences of students and teachers, as well as the development of new and innovative digital learning platforms that are tailored to meet these needs.

In conclusion, the COVID-19 pandemic has disrupted education on an unprecedented scale, highlighting the importance of access to technology and digital literacy skills. The impact of the pandemic on education has been uneven, with students from low-income families and marginalized communities being disproportionately affected. Demographic data plays a significant role in student engagement with digital learning platforms, and analyzing this data can help identify patterns and trends that can inform strategies to improve engagement and support for underrepresented students. Predicting product engagement for the year 2023 requires a thorough understanding of the factors that influence engagement and the development of innovative solutions that meet the evolving needs of students and educators.

2. REVIEW OF LITERATURE

LITERATURE SURVEY

Paper 1: Islam, Md, Noor Asliza Abdul Rahim, Tan Chee Liang, and Hasina Momtaz. (2011)

For this paper, students were selected randomly from the University Malaysia Perlis, Malaysia to evaluate the effectiveness of the learning system in practice. After finding the results, it was confirmed that the age, program of study, and level of education have a significant effect on the effectiveness of E-learning. In this paper, they made some comparisons between the effectiveness of e- learning and traditional learning. First, it talks about the evolution and transformation of traditional learning methods to E-learning methods in the education sector followed by a discussion on all the demographic factors like gender, age, student's status, the program of study, level of education, race, marital status, employment, and the effectiveness of e-learning. One-way ANOVA and T-test were employed to identify the effect of demographic factors on the effectiveness of e-learning.

Paper 2: Muthuprasad, Thiyaharajan, S. Aiswarya, K. S. Aditya, and Girish K. Jha. (2021)

COVID-19 had a major impact worldwide, many countries were under lockdown and all educational institutes had been shut down for an indefinite period. To complete the syllabus in the stipulated time frame, most of them shifted to online mode using Blackboard, Microsoft Teams, Zoom, or other online platforms. According to a study by Adam et. al(2012), there is no significant difference between online learning and traditional learning, it is said to be as effective as a traditional class if structured properly. E-learning has made our educational journey easier. The concept of readiness for online learning in the Australian vocational education and training sector was described by Warner et al. which was further refined by several researches. They selected agricultural graduates as it has the most diverse subjects and conducted a literature survey. Discussions were held for students who were attending the online classes. Data was collected based on demographic features which were analysed and summarized into tables. The study depicted that majority of the respondents preferred online classes to catch up with the curriculum during the pandemic, they also mentioned that interaction is a major factor in the success of online classes.

Paper 3: Naresh, B., D. Bhanu Sree Reddy, and Uma Pricilda (2016)

In this study, they analysed the level of awareness of the student, degree of acceptance, and student's adoption of the e-learning environment. The study mainly focuses on identifying the demographic factors which influences other variables such as - the readiness of the student towards e-learning environment as well as knowledge and the level of comfort in the technology they use. E- learning has shown rapid growth as an alternative to traditional teaching. Many firms and educational institutes have come forward to train their employees or to promote education. Flexibility was the major factor that most people benefited from. But it has its own disadvantage too which is the lack of interaction and late feedback.

Not all e-learning techniques are suitable for developing countries, they should undergo upgradation in the assessment approach which'll make them up-to date with models of readiness of learners with new technology.

To analyse the awareness level of e-learning among students in higher education institutes and to also assess the degree of familiarity with e-learning technologies and tools, samples of a total of 130 respondents were collected - 84 male and 46 female respondents from the Vellore district of any college mainly focusing on students taking business administration or masters in business administration. Questionnaires were distributed amongst the students and used as samples for research and data analysis. SPSS v20.0 is used for data analysis and One-way ANOVA is used regarding determination of demographical variables. All the responses were analysed together and a conclusion was reached that the institute should consider students' awareness level and knowledge about e-learning technologies and demographic variables to have a higher influence through e-learning.

Paper 4: El Refae, Ghaleb Ghaleb A., Abdoulaye Kaba, and Shorouq Eletter(2021)

Students' grades and their GPAs were collected from the admission and registration unit at AL Ain University in the United Arab Emirates to analyse the impact of demographic characteristics on academic performance in face-to-face learning and distance learning, implemented during the pandemic.

It was observed that the number of students who were not able to cope up with the teaching was 11% less in front-to-front learning compared to distance learning. The framework consists of independent variables like gender, college, and status of the student and dependent variables like academic performance.

This framework determines the impact of demographic characteristics on academic performance which is its main objective. The table summarizes the differences between front-to-front learning and distance learning with respect to students' grades and also compares the semester grade point average against the point average in them. Lastly, it summarises the results of Spearman's rho correlation analysis applied to demographic characteristics and academic performance. Thus, as inferenced demographic characteristics have a significant impact on students' academic performance. Hence, educational institutes should offer distance learning program along with front-to-front learning programs to attract more students.

Paper 5: Associations between demographic factors and the academic trajectories of medical students in Japan

Similar to the previous study, this paper too examines the associations between demographic factors and academic performance trajectories using group-based trajectory modelling and categorizes students into GPA trajectories. They included 202 medical students admitted to Tokyo Medical and Dental University in Japan in 2013-14. In recent years in Japan, medical students are either repeating years or withdrawn from medical school.

Group-based trajectory modelling allows researchers to categorize each individual into a different group with a different outcome trajectory.

Demographic factors include the type of high school, geographical area, type of admission test, HS graduation year, and sex. Past academic performance was assessed using high school GPA scores that were reported from their high schools during the admission application process. Some of the findings provide information about high schools' geographical area being outside of the National Capital Region as a risk factor for a lower GPA trajectory, withdrawing or repeating the school year, and also students who entered university by the second exam were more likely to be in the highest GPA group.

It helped in identifying and monitoring students at risk of poor academic performance.

Paper 6: The influence of COVID-19 related psychological and demographic variables on the effectiveness of e-learning among health care students in the southern region of Saudi Arabia

E-learning is a modern and flexible mode of education and is being used as an alternative to conventional mode of education during the ongoing COVID-19 pandemic.

E-learning coupled with the ongoing pandemic, offered a great challenge to academic institutions and students to keep teaching and learning respectively, However, it proved to be relevant and effective.

With sudden introduction to e-learning, the swap from traditional to online teaching affected the social and psychological wellbeing of students. Pandemic as well as specific demographic characteristics have been found to have a major influence on the effectiveness of online teaching and the learning process.

Based on evidence in studying multiple variables, different aspects of students' perception which impacts the students learning were included for the analysis by using a Likert scale, which is as follows: P-value < 0.05 was considered significant.

To our surprise, more than half of the respondents (63.4%) reported that they had no previous experience in e-learning.

Paper 7: Research on Effects of Demographic Factors on the English LanguageLearning Among Tribal High School Students

In this paper the attempt was made to analyse the tribal student's interest towards learning English and how far their demographic situation affects them to learn a language.

The study examines whether demographic factors of students has any effect on the language learning process among the high school tribal students with special reference to a school in Udhagamandalam of Nilgiris district in Tamil Nadu.

According to the belief, it is true i.e., parent's education and family income also plays a similar role in a child’s education. When talking about learning a second language for the tribal student it is entirely different.

As learned from the questionnaire filled by the students. It was inferenced that only less than 40% of the students read English newspapers/ magazines in their day-to-day life and the rest 60% of the students never read or showed any interest towards reading the newspaper/ magazine.

Paper 8: Demographic factors matter to e-learning as determined in Nepal

COVID-19 a worldwide pandemic, had created a new environment for transformation in teachers from organizing traditional classroom learning to managing the behaviour of their students through electronic learning platforms, that also demanded skill of new technology.

It seemed Nepal had a weak e-learning setting that lacked some energetic presence of the middle- aged group. Newer generations were enjoying the virtual meetings and classes while older adults appeared to the e-learning setting with the intention of escaping from it. This study uses the Servage (2005) for the definition of e-learning which it has gotten from the US Report (2002) by the Commission on Technology and Adult Learning that refers to instructional content or learning experiences delivered or enabled by electronic technology. No statistical tools were used to analyse the data used in this descriptive study.

Structuring or physical arrangement of classroom environments helped to ease traffic flow, minimize distractions, and provided teachers with good access to students in order to respond to their questions and better control their behaviour. If e-learning system as such is implemented without consideration of people’s needs and their income level, social inequality will be prevalent.

Paper 9: The Demographic Factors and Decision on Selecting Open, Distanceand Online Learning

The decision to continue studying via open, distance and online learning is one of the many steps in a sequence of processes including an overview of needs, information retrieval, alternative evaluation, choices, and post-decision behaviours.

Many things come to the attention and consideration of students in deciding to continue their studies through open, distance and online learning. These factors can be grouped into internal factors and external factors. There are other factors that influence students to decide between them as well such as social environment, culture, psychology, marketing and systems control.

In choosing University, some students also try to determine the best one based on the above, and this paper analyses all such influential factors of UT students in Jakarta Regional Office.

The table shows the Educational Background of the Jakarta regional students in 2019 and based on the data, most Jakarta students have graduated from high school, and almost 60% of the students from high school has the highest percentage of marks, and some students also graduated with master degree (around 0.4%).

Paper 10: Impact of demographic trends on the achievement of the millennium development goal of universal primary education

Education has an impact on both human and economic development, its own development can be impeded by the influence of a range of demographic, economic, social, cultural and political factors

The second part of the paper discusses the demographic pressures and the achievement of universal primary education (UPE) Given that the aim here is to deal with the issue of achieving the Millennium Development Goals, the third part examines how the type of population data used to calculate the indicators measuring progress towards those goals can affect perceptions in regard to achieving UPE

While demographic statistics may serve as the underpinnings of educational plans, enabling planners to reckon the ways in which population growth could hinder achievement of the goal of UPE, it should be stressed that population education is a key aspect of population policies and programmes whose aim is to bring population growth under control.

In the monitoring of the Millennium Development Goals (MDGs), population data served as the basis for calculating the net enrolment ratio (MDG 2, target 3: UPE)

Table VII.3 showed the discrepancies between United Nations Population Division (UNPD) population data estimates and those produced by the countries participating in the World Education Indicators (WEI) programme.

The discrepancies vary from -6.5 to 16 percentage points, the negative values indicating that the national population data has a negative impact on net enrolment ratio (NER) because they are overestimated in comparison to those of UNPD.

3. MATERIALS AND METHODS

3.1 ALGORITHM / PSEUDOCODE

1. First use list.files() to display all the input data files.

2. Load all the libraries into the environment and put the location of all the input files into a variable.

3. Create a function to split all the values and remove all the extra characters from it.

4. Load the district information and starting from the “pct” column, using the function, reformat the variable using the function and remove “\\[” and also replace “,” with “-”.

5. Plot 3 different graphs for pct_black_hispanic, pct_free_reduced, and pp_total_raw and then combine them together to have an overview of District demographics for each area type.

6. Then load the product information and separate values between “-” and split the value.

7. Check the summary statistics of districts and products.

8. Load DataExplorer and lubridate to plot for missing for both datasets.

9. Cleans pp_total_raw column

10. Separate pp_total_raw into pp_total_raw-high and pp_total_raw- low and convert it as integer values.

11. Check the median and average for both pp_total_raw-high and pp_total_raw-low for imputing and impute the NA values with median

12. Merge both pp_total_raw-high and pp_total_raw-low.

13. Plot a bar graph for both district and product. Analyse district data distribution by plotting a bar graph based on State and a pie chart based on locale.

14. Then Analyse product data distribution by plotting bar graphs based on Sectors and Providers.

15. Visualization is done for each sector for 327 products and providers for each product.

16. Plot a pie chart for the functionalities of all those online products.

17. Make new columns minorities, free reduced lunch, and expenditures, and the values for each range are modified and inserted in those columns

18. List all the files for product engagement, read the engagement file, and add district id from the district dataset.

19. Adding district_id from district dataset and lp_id, product_name, provifer_company_name, sector_s, main_fun, sub_fun from product dataset into engagement dataset

20. Find the top 5 products based on their engagement index for each of the primary categories and plot a graph based on it.

21. Replace NA values present in the engagement dataset with the mean of the neighbouring values.

22. Scales engagement_index and pct_access and add it as columns in engagement.

23. Select the district_id of all whose state is NA and store it and add a new column in engagement.

24. Monthly engagement is then (Engagement and student access change)

25. Omit all the NAs from engagement and then check the correlation coefficient between pct_access and engagement_index and plot monthly engagement access.

26. Plot a line graph on the weekly engagement index.

27. Separating the timeline into before and mid-pandemic and collect the 3 highest engagement and plot bar graphs highlighting the first three highest engagement index

28. Check how the engagement index and pct access differ during after and mid-pandemic using t-test.

29. Then plot a line graph for engagement and access change across Districts grouped based on month and locale.

30. Analyse engagement and access change based on demographics like minority level, and free/reduced lunch by categorizing them into less than 40 % and more than 40%.

31. Read the device internet availability dataset in an availability variable and set the column names

32. Correcting the values by replacing Connecticut to Connecticut and Massachusetts to Massachusetts in the availability dataset

33. Categorize each state with its count along with the mean of its engagement index and pct access

34. Removing the states with the lowest observations.

35. Find the mean of engagement index and pct access.

36. Using the Correlation test, check the correlation between Device_always, Device_usually, Device_sometimes, Device_rarely, Device_never, Device_DNR, Internet_always, Internet_usually, Internet_sometimes, Internet_rarely, Internet_never, Internet_DNR with respect to engagement and pct access and store it as a data frame.

37. Analyse the engagement and access based on expenditure by pupils and recode expenditure to certain distinct values.

38. Join the engagement dataset with the products dataset ignoring the NA values and take the mean of the scaled engagement index and scaled pct access. Analyse the engagement and access change across each product.

39. Group engagement dataset based on lp_id,provider_company_name, and product_name, and store its count sorted in a descending manner.

40. Store 20 highest observations separately on another dataset.

41. Plot a bar graph for the Top 20 Google Products used. Plot a bar graph for the Top 20 Non-Google Products used. Plot a bar graph for the Top 20 Digital Learning products used.

42. Store the product information in a dataset whose lp_id=29322 i.e., Khan Academy and then its user engagement is analysed using a graph. Similarly, information is stored in a dataset for lp_id = 90153, i.e., Netflix and analyse thoroughly using graphs.

43. To analyse the mean daily page load events of the top 10 tools per student, filter all the digital learning tools from the product dataset group by their product name and its time of access, and plot a graph to observe.

44. Plot a graph to observe the mean daily page load events for the top 10 tools by locale, per student for each digital learning platform

45. Again, plot the same for minority (Black/Hispanic) students and for students who are eligible for free or reduced-price lunch.

46. Plot a Map for State data distribution for the district dataset.

47. Remove all the redundant values from lp_id and product_name and store the mean of pct_access(nm_pa) and engagement_index(nm_ei) in descending order in a dep_product dataset.

48. Group dep_product based on the district_id and store its count and mean of nm_pa and combine the district_id together into nwea_data.

49. Remove all the states with NA as a value and group by state, its count, and the mean of nm_ei is stored in sd_ep in descending order

50. Plot a state map to analyse the percentage of access in each region.

51. Plot state maps to analyse the engagement index of each region

52. Filter some states like Arizona, California, Connecticut, the District of Columbia, and Florida and analyse the mean daily page-load events for the top 10 digital learning platform tools for each state.

53. The same is done for the rest of the remaining states.

54. Load pct_access, engagement_index, minority, free reduced lunch, and expenditure into a dataset and pre-processed and then plot a corr- plot.

55. List the dates that we consider as early and mid-pandemic.

56. Include a column that includes a date with a break of 1 week and store the mean of engagement and pct_access on that particular date for each region

57. Remove all the data except the list of dates that were considered as early and mid-pandemic and check if all the regions have the exact number of observations as the number of dates for both early and mid- pandemic and remove the one with a lesser number

58. Extract and categorize the state data and store them in early-state and mid-state along with the mean of engagement index and pct access. Then add the rank of engagement and access for both early and mid- values as a column.

59. Check the change in the engagement and pct_access from early to min pandemic and is stored it as a column. Observe the states that showed improvement.

60. To check if any state has issued any policies which played a role in the huge engagement, load the state response and extract only those state's responses whose engagement index improved a lot.

61. Analyse any other factor that affected other states for distance learning. Read childhood trends, covid database, and covid health.

62. Remove all the missing values from these three datasets.

63. Group the state data by state and include the mean of engagement_index and pct_access and then combine all these datasets by state and store it in a new dataset state_combined

64. Replace all the missing values from state_combined with the mean of neighbouring data.

65. Find the Pearson correlation for each column present in the state_combine with respect to the engagement_index and pct_access and store it in two different datasets.

66. Access all the files from engagement data and store time, pct_access and engagement_index.

67. Converting the columns of the district dataset from character to factor and then adding the district dataset to engagement dataset.

68. Calculate the mean engagement index for each district and add it to a new dataset and imputing the missing values with NA.

69. Plot a scatter plot for the mean engagement index.

70. Find the linear mixed effect of the new engagement dataset with respect to to locale, minorities, pct free reduced, pp total raw, and then find the summary for each of them.

3.2 DATASET DESCRIPTION

We have used a set of daily edtech engagement data from over 200 schooldistricts. The other sets of files that we have used include:

- The **engagement_ data** folder is based on a Chrome Extension that collects page load events of over 10K education technology products, including websites, apps, web apps, software programs, extensions, eBooks, hardware, and services used in educational institutions. The engagement data have been aggregated at the school district level, and each file represents data from one school district.

- The **products_info.csv** file includes information about the characteristics of the top 372 products.

- The **districts_info.csv** file includes information about the characteristics of school districts, including data from NCES and FCC.

The definitions of each column in the three data sets are detailed below:

- **Engagement Data**

The engagement data is aggregated at school district level, and each file in the folder 'engagement_data' represents data from a 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'.

➢ *Name* - Description

➢ *time* - date in "YYYY-MM-DD"

➢ *lp_id* - The unique identifier of the product

➢ *pct_access* - Percentage of students in the district 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 ofa given product and on a given day.

- **District Information Data**

The district file `districts_info.csv` includes information about the characteristics of school districts, including data from:

➢ *district_id* - The unique identifier of the school district

- *state* - The state where the district resides in

- *locale* - NCES locale classification that categorizes 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 NCES data

- *pct_free/reduced* - Percentage of students in the districts eligible for free or reduced-price lunch based on NCES data

- *county_connections_ratio* - `ratio` (residential fixed high-speed connections over 200 kbps in at least one direction/households) based onthe county level data from FCC From 477

- *pp_total_raw* - Per-pupil total expenditure (sum of local and federal expenditure). The expenditure data are school-by-school, and we use the median value to represent the expenditure of a given school district.

- **Product Information Data**

  The product file 'products_info.csv' includes information about the characteristics of the top 372 products with most users in 2020.

- *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 labelled as one of these threecategories:

- *LC* = Learning & Curriculum,

- *CM* = Classroom Management

- *SDO* = School & District Operations.

  Each of these categories have multiple sub-categories with which the products have been labeled.

3.3 ARCHITECTURE AND EXPLAINATION

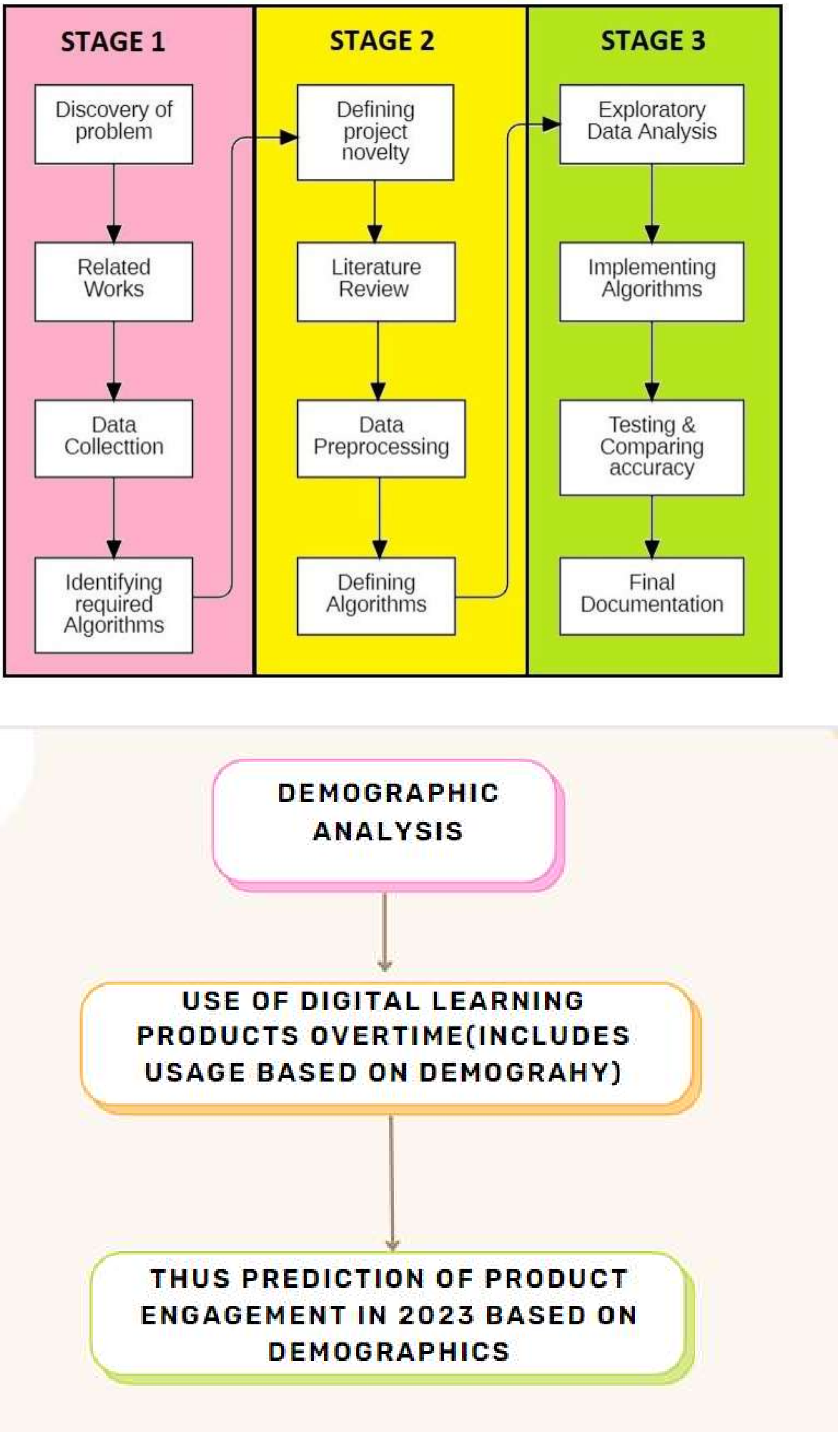

STAGE 1
Discovery of problem
Related Works
Data Collecttion
Identifying required Algorithms
STAGE 2
Defining project novelty
Literature Review
Data Preprocessing
Defining Algorithms
STAGE 3
Exploratory Data Analysis
Implementing Algorithms
Testing & Comparing accuracy
Final Documentation
DEMOGRAPHIC ANALYSIS
USE OF DIGITAL LEARNING PRODUCTS OVERTIME(INCLUDES USAGE BASED ON DEMOGRAHY)
THUS PREDICTION OF PRODUCT ENGAGEMENT IN 2023 BASED ON DEMOGRAPHICS

Figure: (A) Generic Architecture Diagram, (B) Architecture Diagram for the Project

First in order to understand the various attributes of our dataset, we'll be performing exploratory data analysis. Before proceeding with modelling and relation finding, we'll clean our dataset of NA values if there are any, tidy the data, convert the needed numeric/character values to factors, filter out the useless attributes, create groups for "age" columns, etc. After data pre- processing, we'll work around with demography and product engagement data.

Now out of the relations obtained, we'll try to brainstorm and come up with predictive models that will work best with our dataset. After that, we'll do detailed research on these models and look for any scope of improvement. Then the finalized model will be built and trained on the dataset taken. The trained models will be tested using the testing dataset and various metrics willalso be evaluated (like confusion matrix, accuracy, RMSE, etc.). Based on the metrics all the different models will be compared, to find out the best- performing one. In the end, results will be visualized and presented.

Hence, we will explore how digital learning was before the covid period, how much impact it had during this period and finally how it is right now. Then we will survey the impact of demographics of the available districts on digital learning, and combine it with the data obtained for digital learning engagement. Finally, we will predict the product engagement for digitallearning for the year 2023 based on demographics.
The modules in the architecture are as follows:

- Discovery – The defined problem statement for our project is defined as "Impact of Demography on Digital Learning". We will collect our data by filtering through existing datasets about demography's impact on learningand how digital learning has been before, after and through COVID.

- Data Preparation – The data will be cleaned following ELT procedure i.e., Extract, Load, Transform. Our data will be cleaned of all NA values and the missing records will be imputed.

- Model Planning – We have picked the statistical model for the analysis,depiction as well as future prediction based on our dataset.

- Model Building – We will execute our model at this stage and train it withour dataset.

- Communication of Results – At this step we need to execute 4 major substeps such as: (i) Key Findings; (ii) Quantify the Business Value; (iii) Summarize Results; (iv) Convey the Findings to Stakeholders.

- Operationalize – A pilot project is executed to confirm the results obtained before in real life conditions. If successful, the model will beexecuted in production environment.

Thus, to sm up –

- District Distribution and Demography - we will evaluate the effect of demography of a region on digital learning, based on the available district. Thus, we will determine districts' demographics by area type.

- Top Digital Learning Products By engagement - Digital learning products are categorized according to their functions and sub-functions and their engagement is evaluated over the years before, through and after COVID.

- Engagement in Digital Products for 2023 based on different Demographics

  We will combine engagement data of the year, grouped based on demography and predict product engagement for the year ahead on this basis.

3.4 HARDWARE & SOFTWARE USED

HARDWARE REQUIREMENTS:

- Operating system- Windows 7,8,10
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)
- RAM-4GB

SOFTWARE REQUIREMENTS:

- Jupyter Notebooks - Jupyter Notebook is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

- R studio - RStudio is an integrated development environment (IDE) for R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging, and workspace management. RStudio is available in open-source and commercial editions and runs on the desktop (Windows, Mac, and Linux).

- R - R is a Programming Language that is mostly used for machine learning, data analysis, and statistical computing. It is an interpreted language and is platform independent that means it can be used on platforms like Windows, Linux, and macOS.

- Chrome - Google Chrome is a cross-platform web browser developed by Google. It was first released in 2008 for Microsoft Windows, built with free software components from Apple WebKit and Mozilla Firefox. Versions were later released for Linux, macOS, iOS, and also for Android, where it is the default browser.

4. PROPOSED WORK

4.1 NOVELTY

The proposed work is novel in the following ways:

1. Analyzing the impact of demographic factors on digital learning engagement: While previous studies have examined the impact of various factors on digital learning engagement, this project focuses specifically on the impact of demographic factors such as age, gender, race, ethnicity, and socio-economic status. This approach provides a more targeted and nuanced understanding of the factors that influence student engagement with digital learning.

2. Predicting product engagement for the year 2023: The use of statistical analysis to predict product engagement for the year 2023 is a novel approach that can provide valuable insights into the future needs and preferences of students and educators. This information can inform the development and marketing of digital learning products to better meet the evolving needs of diverse student populations.

3. Integrating multiple methods of analysis: The project will integrate multiple methods of analysis, including statistical analysis, literature review and education policy analysis. This approach provides a comprehensive understanding of the impact of demographic factors on digital learning engagement and informs the development of strategies and products to improve engagement and support for underrepresented students.

Overall, the proposed work is novel in its approach to analyzing the impact of demographic factors on digital learning engagement and its use of multiple methods of analysis to provide a compr

ehensive understanding of this complex issue. The project's focus on predicting product engagement for the year 2023 and the use of ethnographic methods to gain a deeper understanding of underrepresented students are also novel contributions to the field of digital learning research.

4.2 PROJECT CONTRIBUTIONS

The project can make significant contributions in the following areas:

- Understanding the impact of demographic factors on digital learning engagement: By analyzing data from digital learning platforms, the project can provide insights into how demographic factors such as age, gender, race, ethnicity, and socio-economic status impact student engagement with digital learning. This understanding can inform the development of targeted strategies to improve engagement and support for underrepresented students.

- Informing the development of new digital learning products: The project can provide insights into the evolving needs and preferences of students and educators. This information can be used to inform the development of new and innovative digital learning products that better meet the needs of diverse student populations.

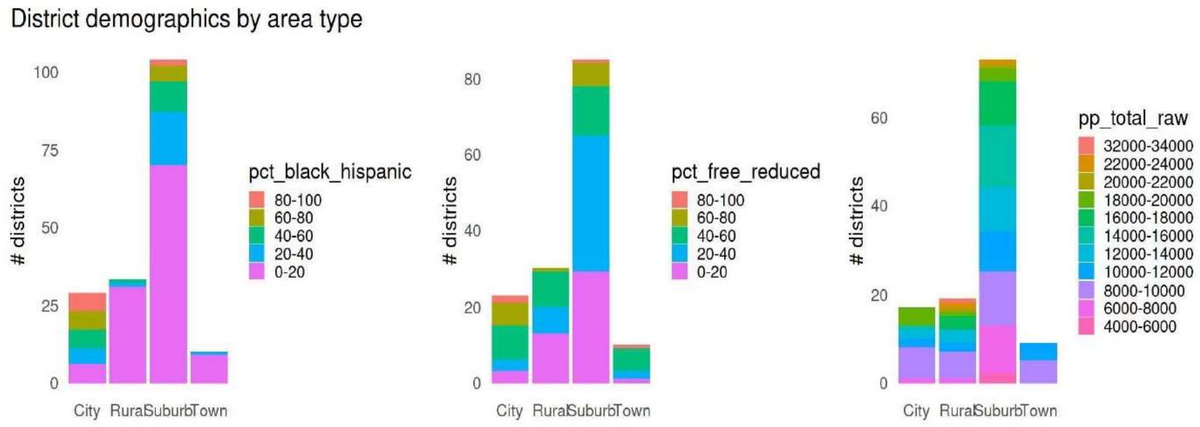
- Predicting product engagement for the year 2023: The project can use machine learning algorithms to predict product engagement for the year 2023 based on trends and patterns identified in the analysis. This prediction can help guide the development and marketing of digital learning products, ensuring that they meet the needs and preferences of students and educators in the coming years.

- Reducing the educational gap: The project's findings may highlight areas where underrepresented students require additional support to fully engage with digital learning platforms. This information can inform the development of targeted strategies to improve engagement and support for these students, ultimately contributing to reducing the educational gap between different socio-economic groups.

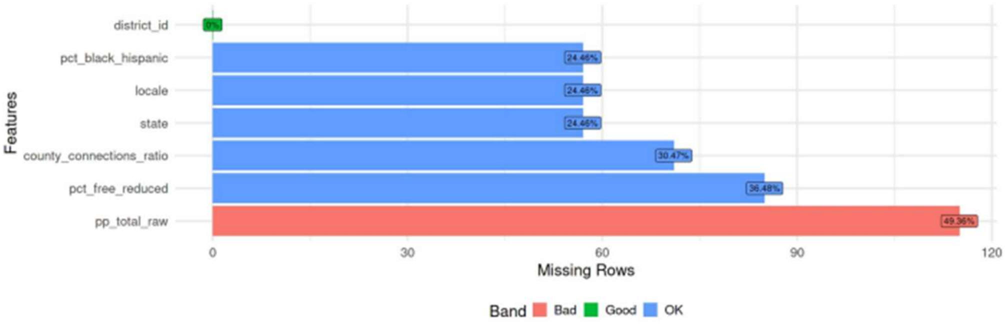
Overall, the project's contributions can lead to a better understanding of the impact of demographic factors on digital learning engagement and inform the development of strategies and products to improve engagement and support for underrepresented students. The project can contribute to reducing the educational gap and ensuring that digital learning products meet the evolving needs and preferences of students and educators.

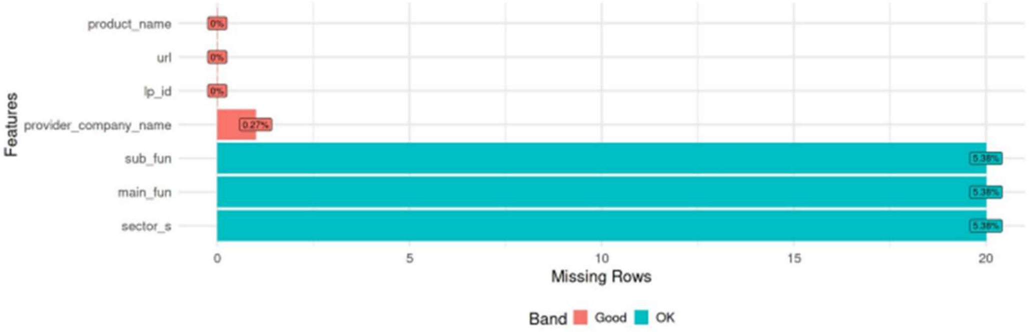
5. RESULTS AND DISCUSSION

5.1 RESULTS – FIGURES, COMPARISON TABLES AND EXPLAINATION

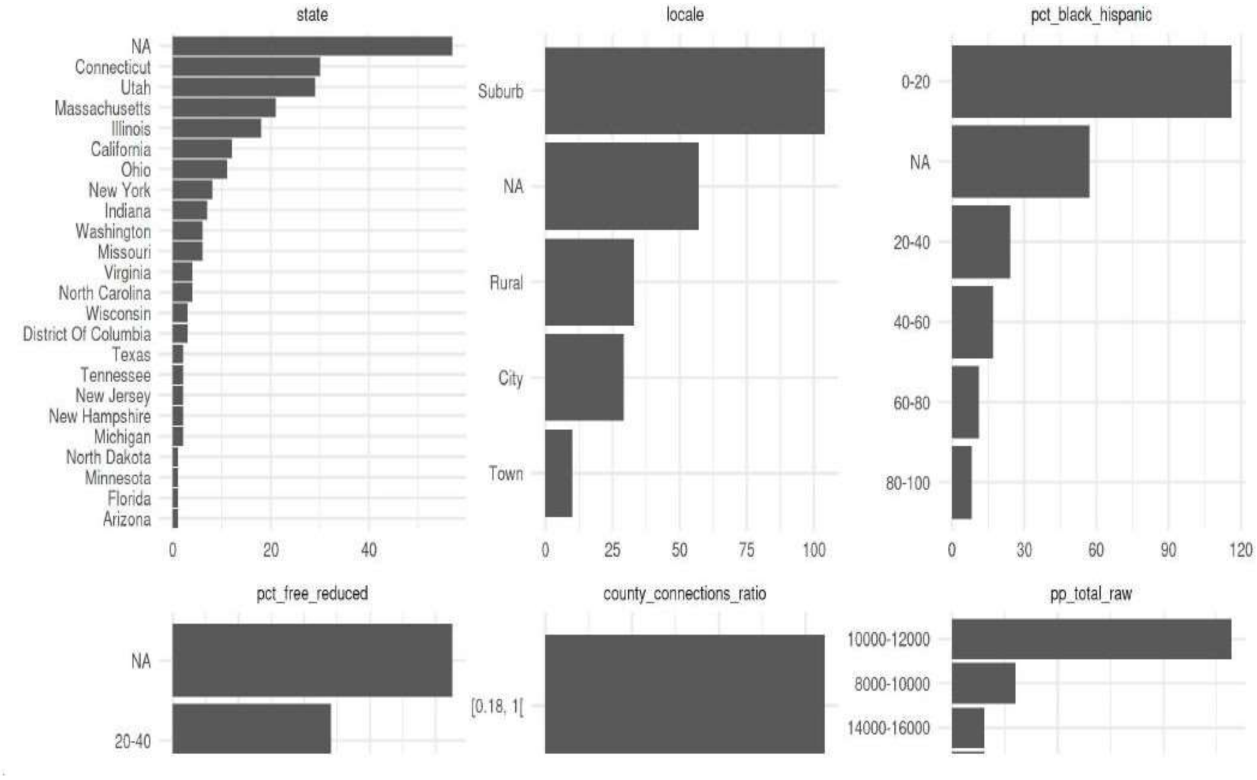


INFERENCE: Demographic data collected shows that the different area types differ substantially in their proportion of blacks & Hispanics, and in their proportion of pupils with access to digital learning. No very obvious patterns emerge for expenditure per pupil from this bar chart.

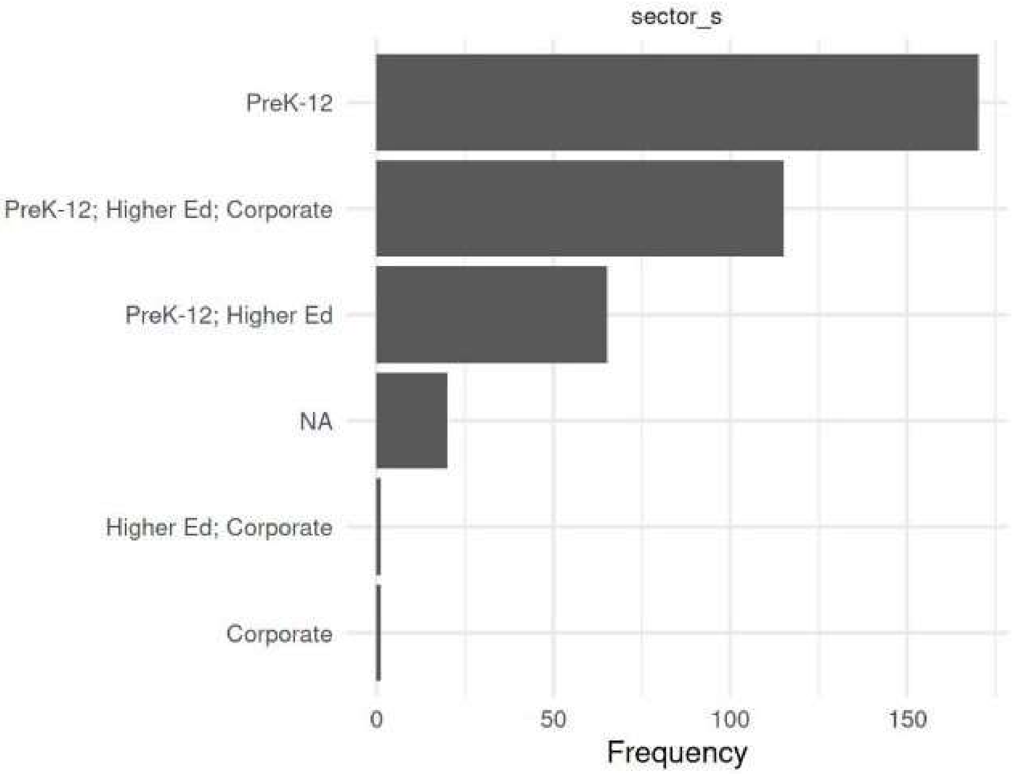




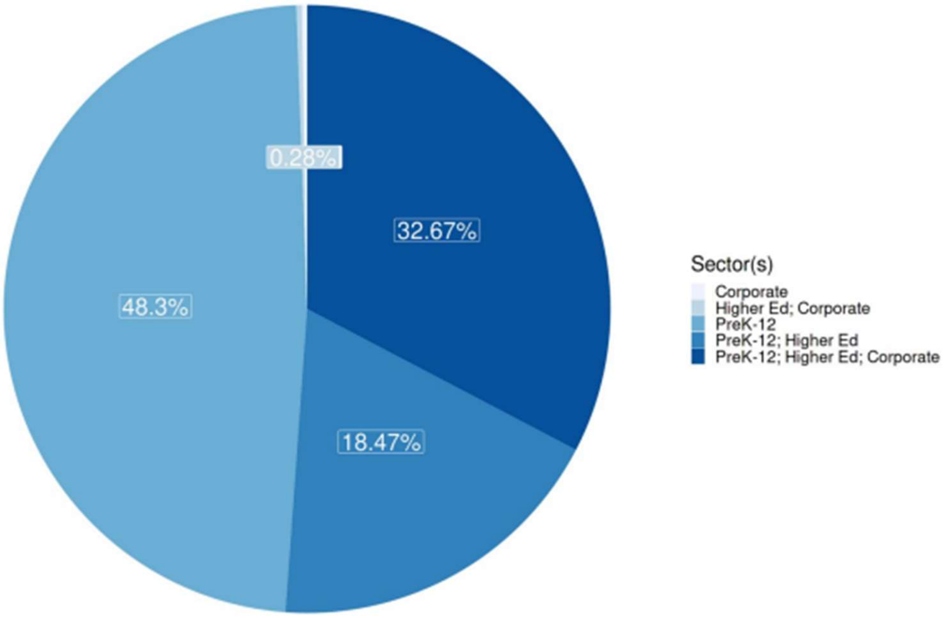
INFERENCE: An account of the missing values is taken in for consideration. The amount of missing data in both data sets is acceptable except pp_total_raw column in districts dataframe.



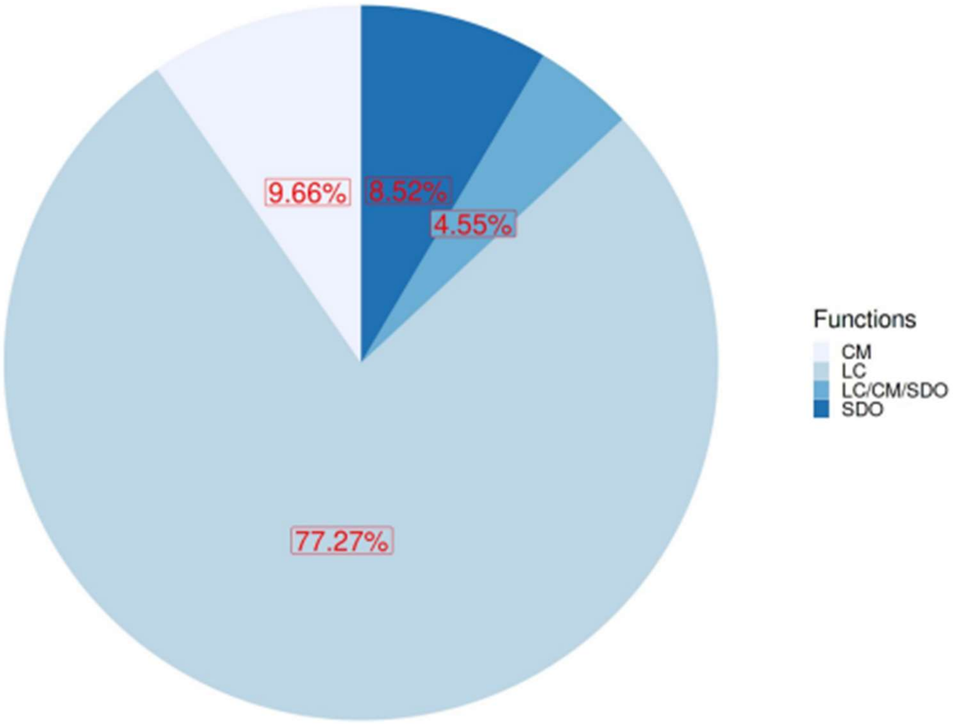
INFERENCE: Demographic data distribution based on available districts.



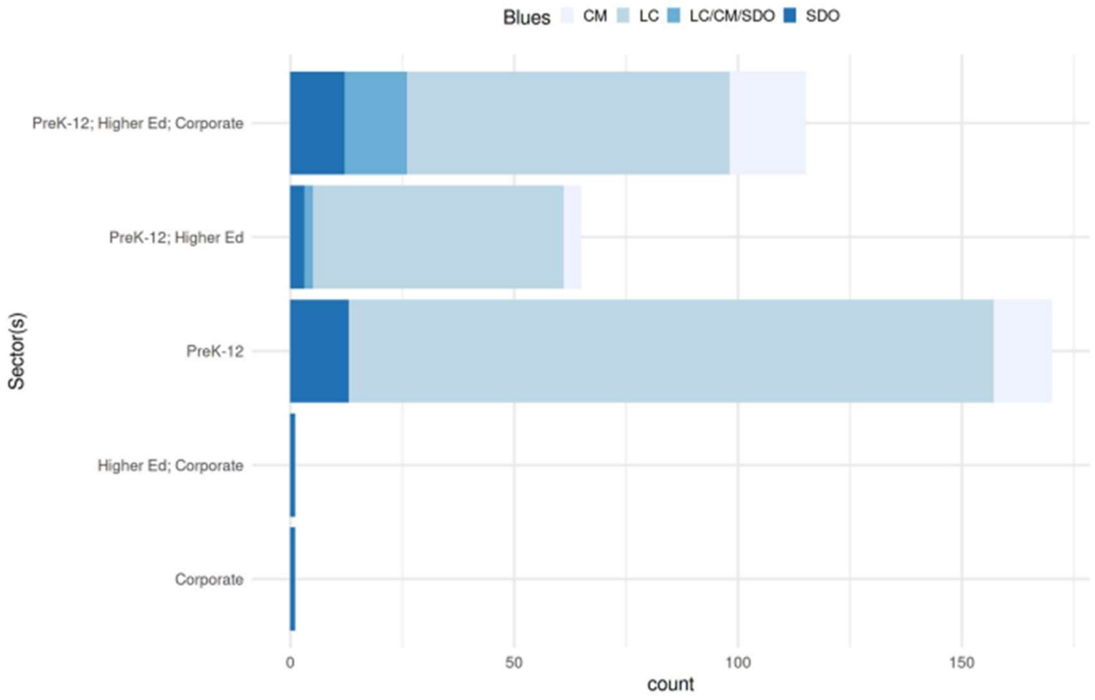
INFERENCE: Products Data Distribution based on Sectors



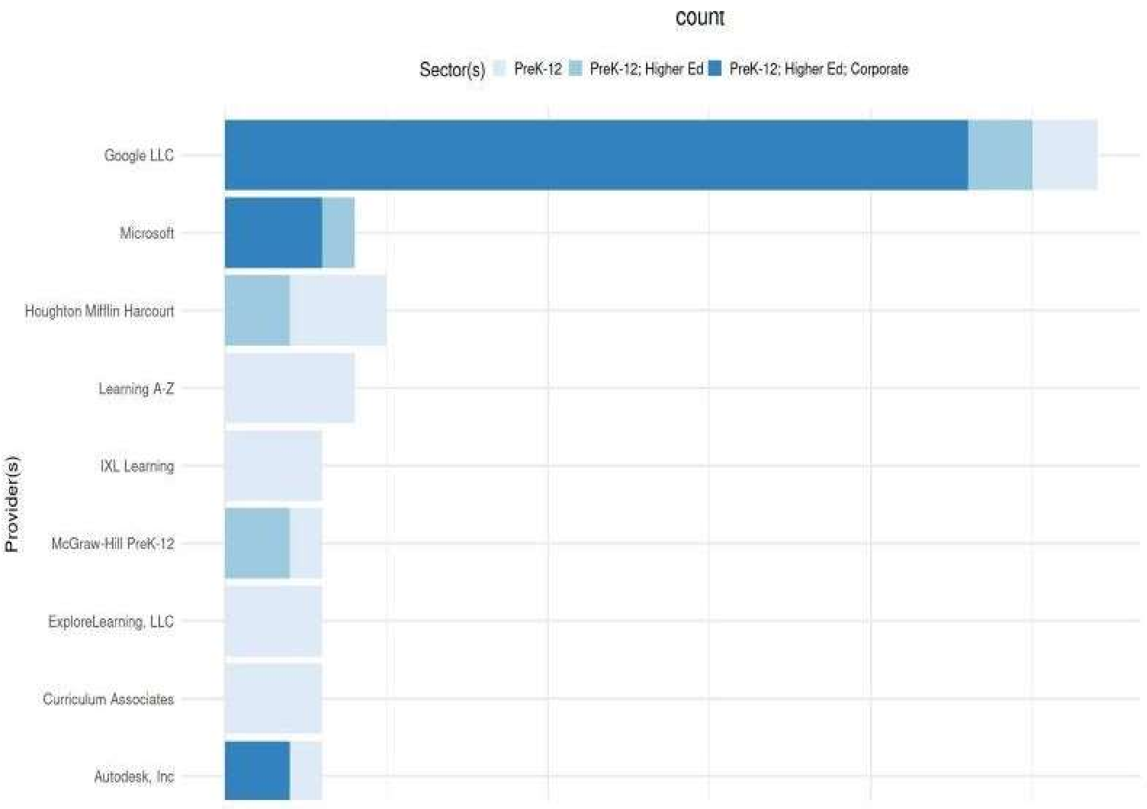
INFERENCE: Visualization of Sectors. The most prevalent sector(s) of those products is the Prek-12(48.3%), and in contrast, those products with Corporate sector(s) are not so popular in students.



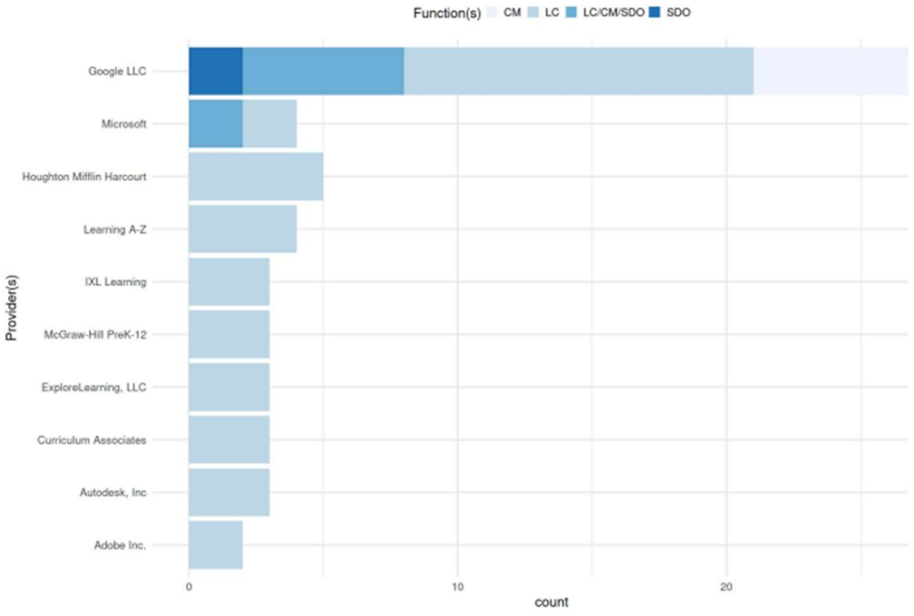
INFERENCE: Visualization of Functions. Besides, the functions of those products mainly consist of 4 parts: CM, LC, LC/CM/SDO and SDO. Them LC is the most frequent function of those products.



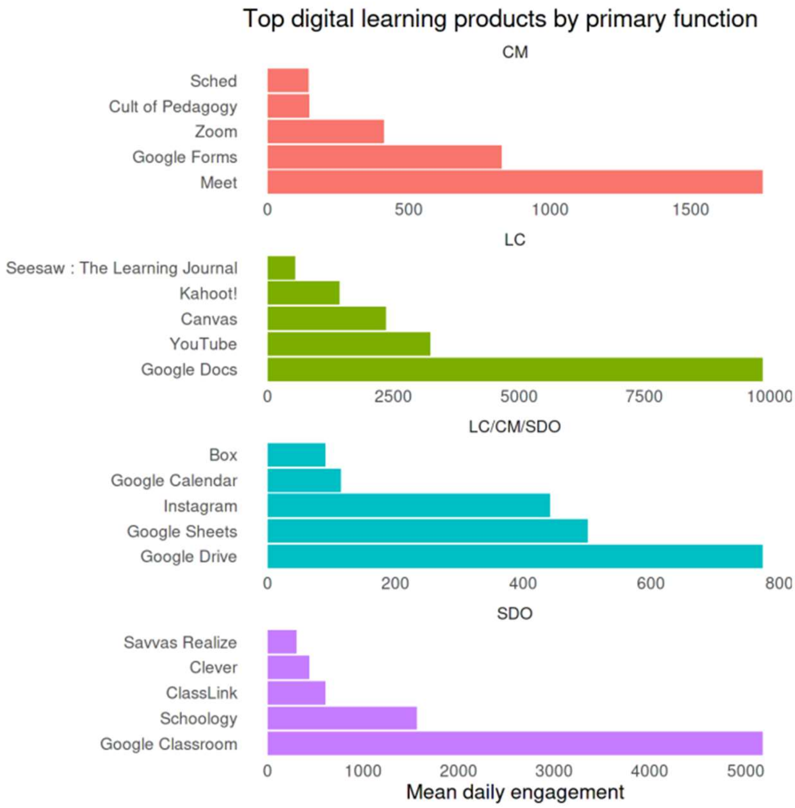
INFERENCE: Visualization of Sectors based on functions



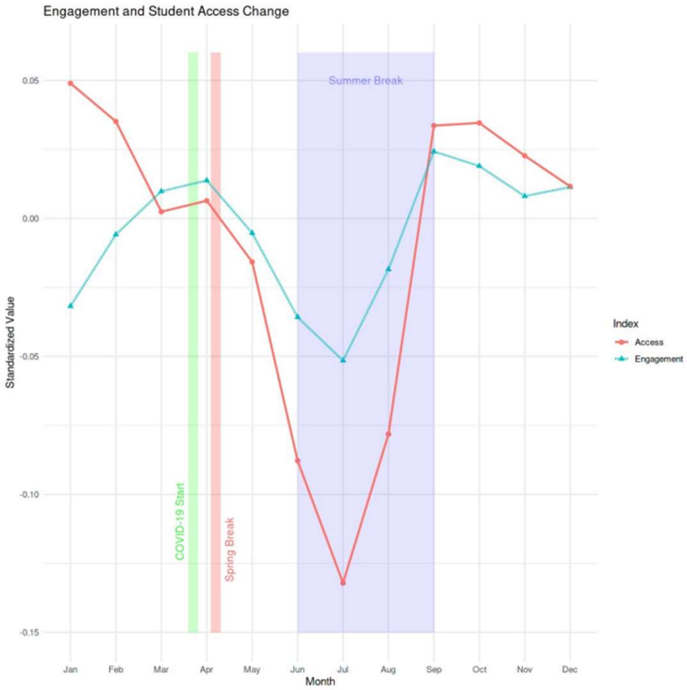
INFERENCE: Visualization of Providers based on Sectors. There are 327 top popular online products in those districts. Most of them are provided by Google LLC, and the next one is Microsoft.



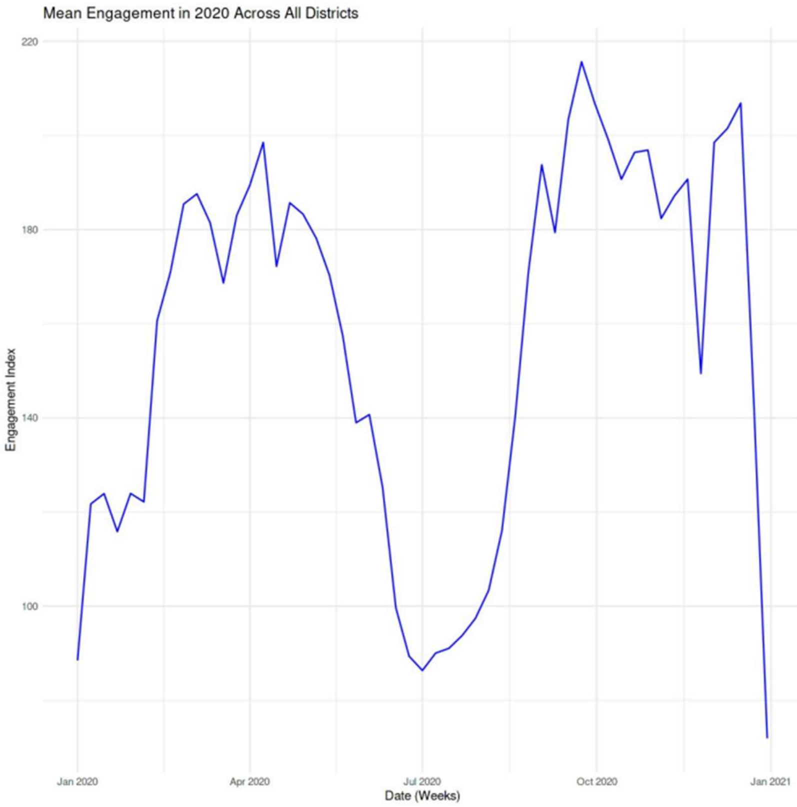
INFERENCE: Visualization of Providers based on Functions



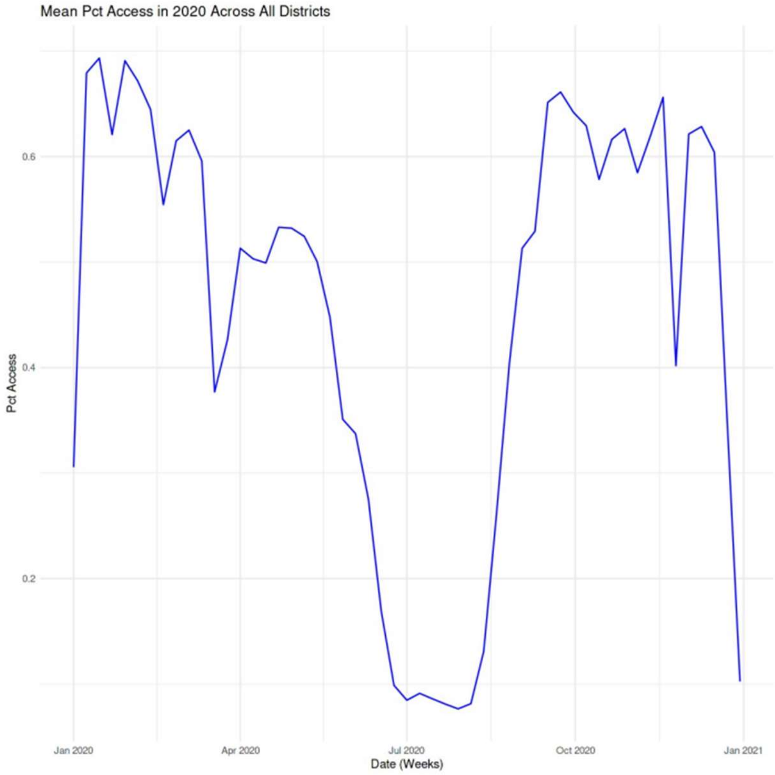
INFERENCE: Top Digital learning products by primary function based on mean daily engagement



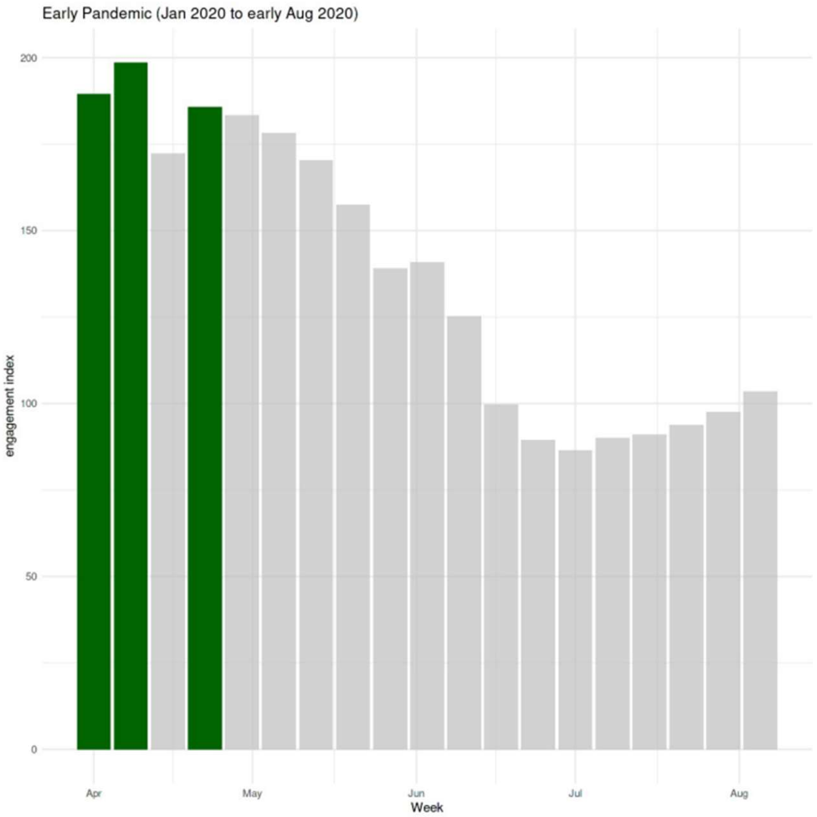
INFERENCE: Engagement and Student Access Change



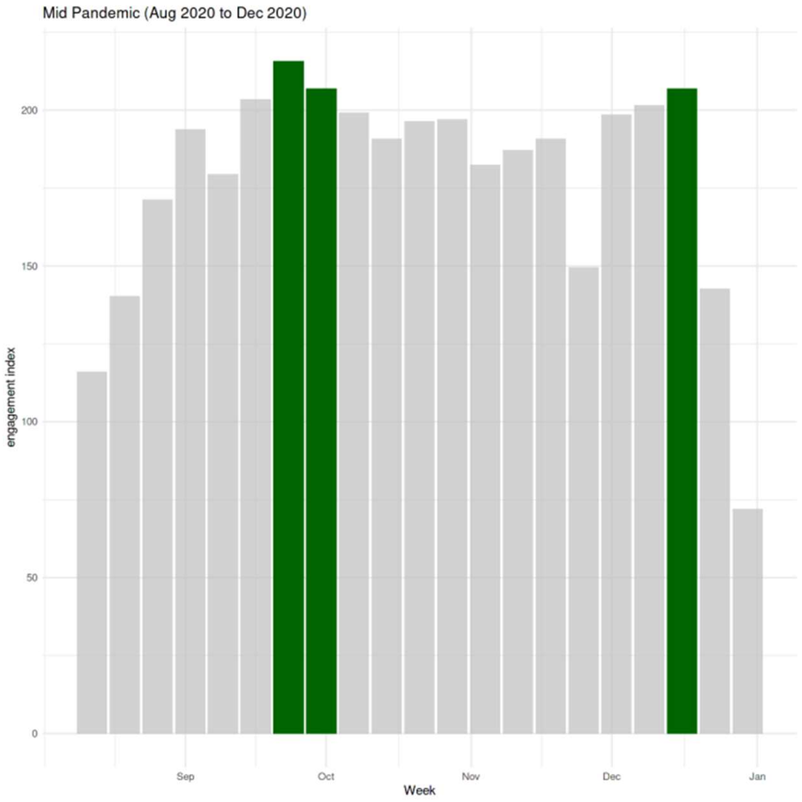
INFERENCE: Mean Engagement Across All Districts



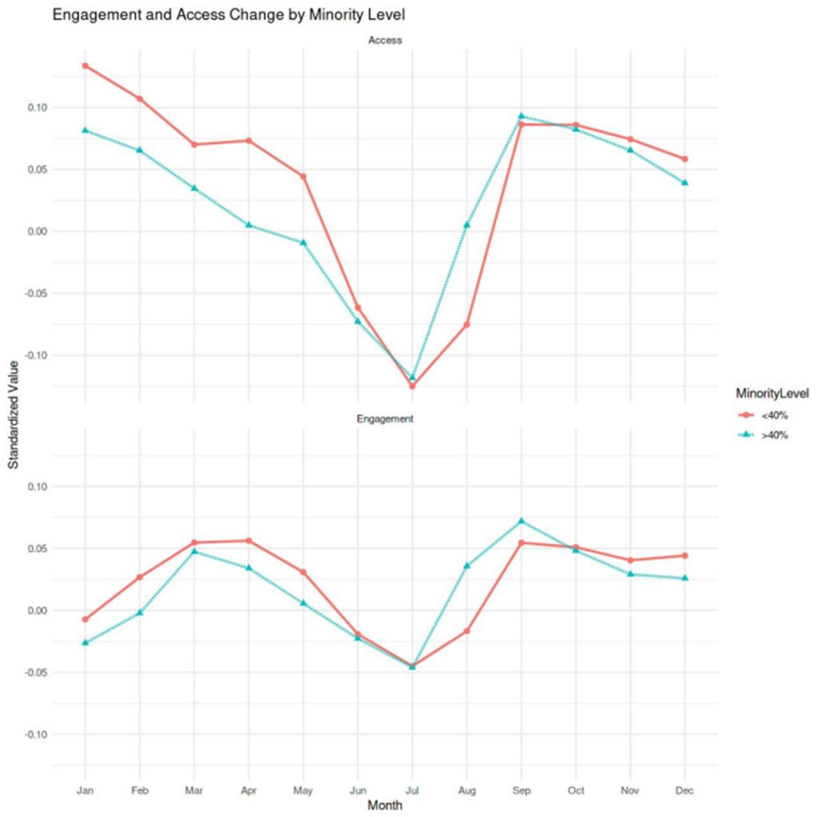
INFERENCE: Mean Pct Access Across All Districts



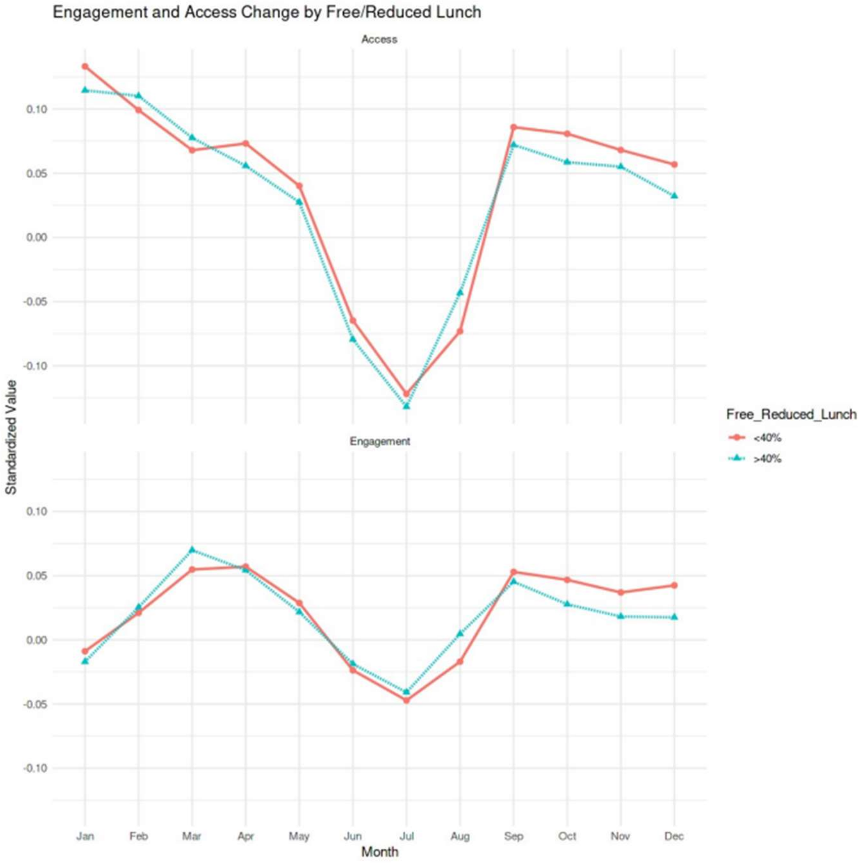
INFERENCE: Engagement change during early pandemic



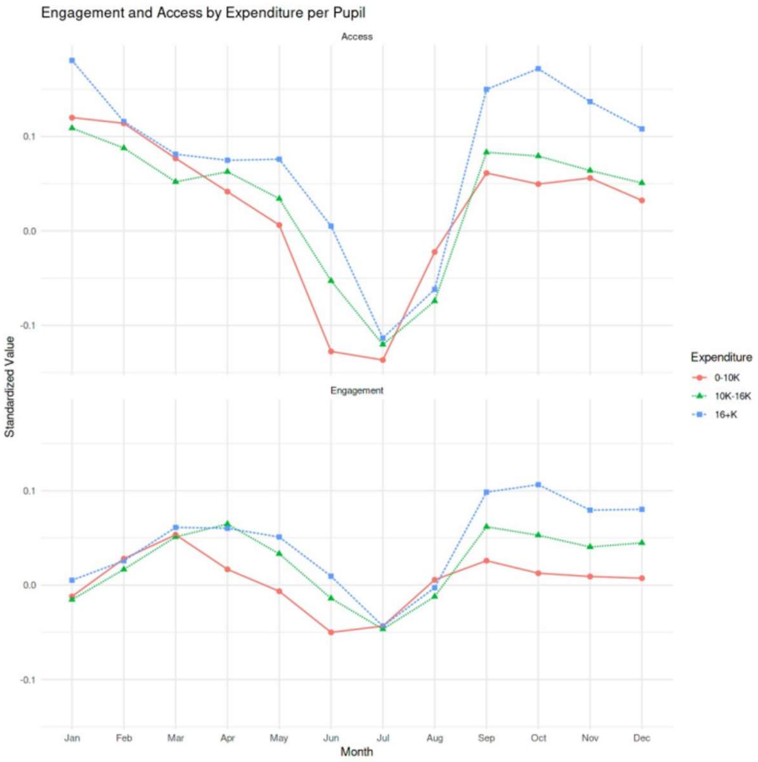
INFERENCE: Engagement change during mid pandemic



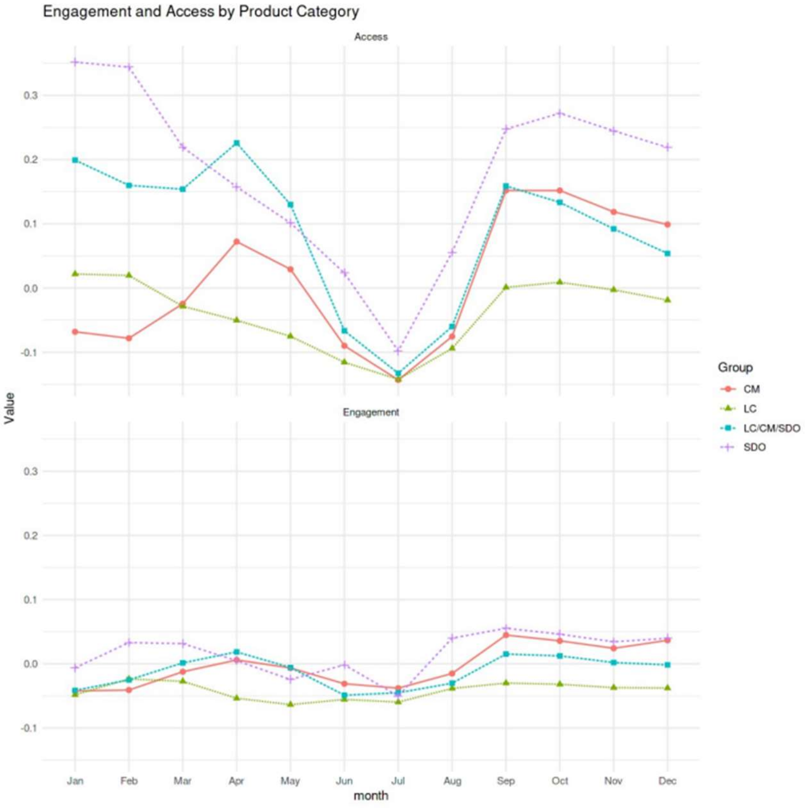
INFERENCE: Engagement and Access Change by Minority Level



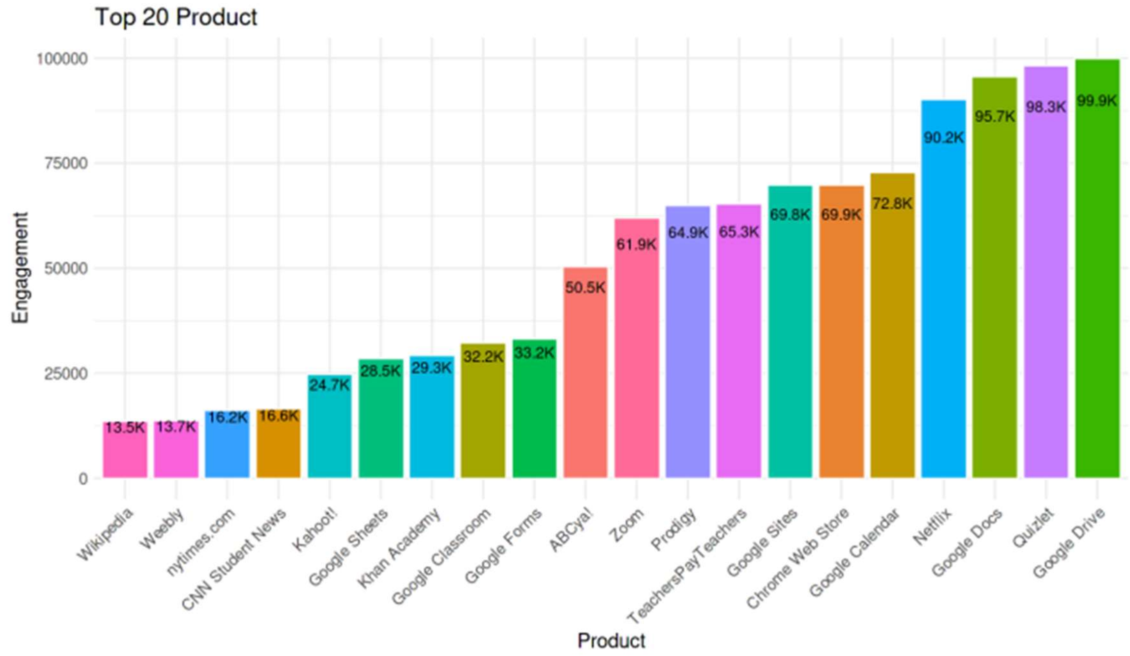
INFERENCE: Engagement and Access Change by Free/Reduced Lunch



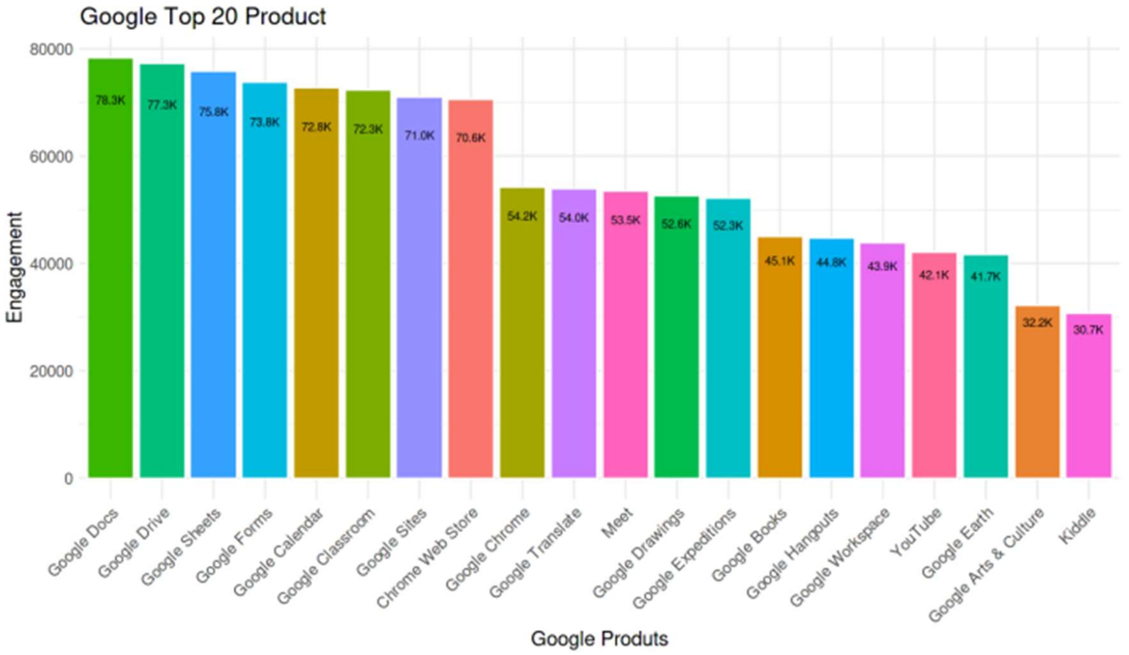
INFERENCE: Engagement and Access by Expenditure per pupil



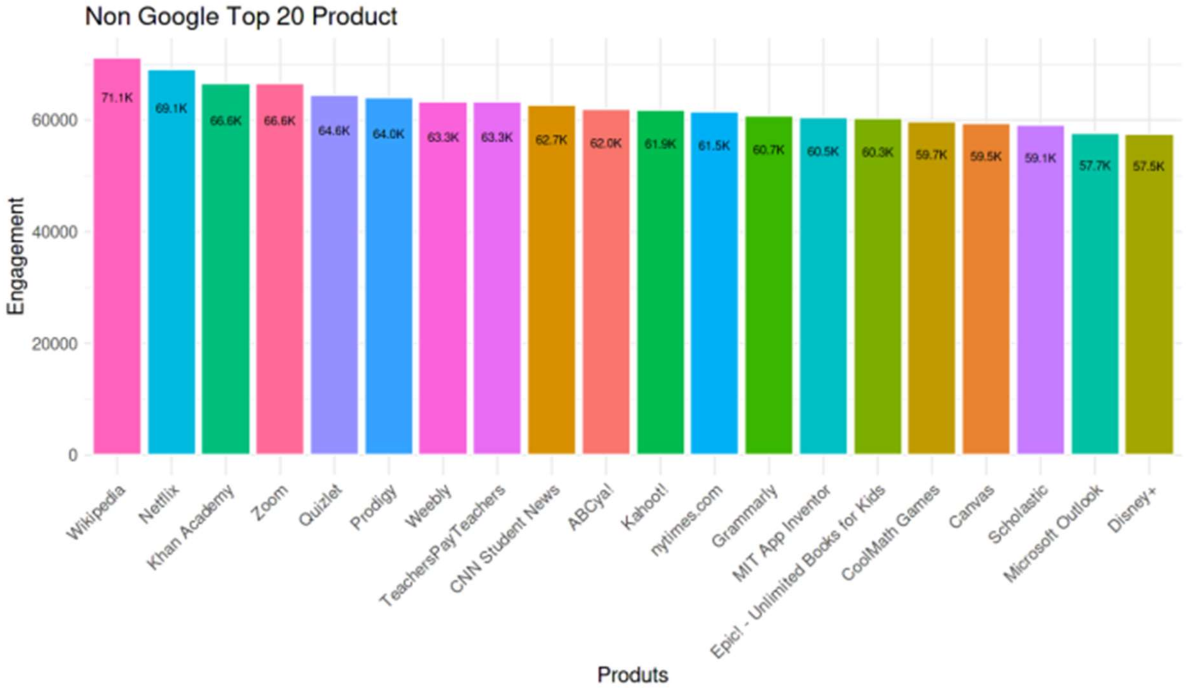
INFERENCE: Engagement and Access by Product Category



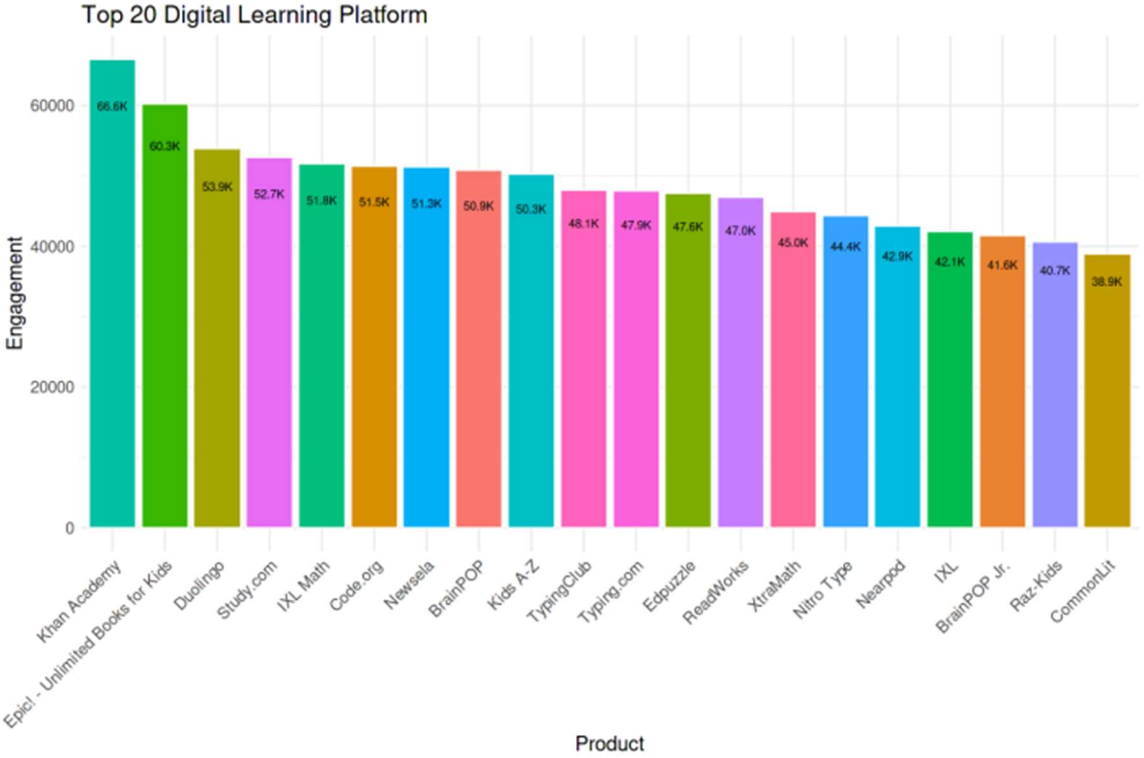
INFERENCE: Top 20 Products based on Engagement



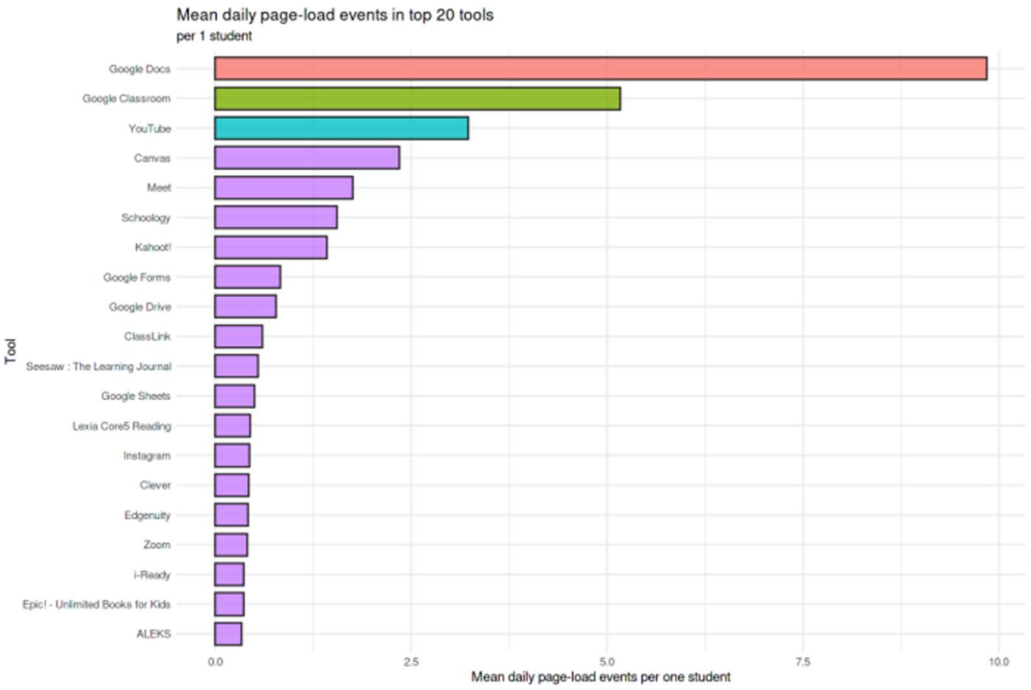
INFERENCE: Google Top 20 Products based on Engagement



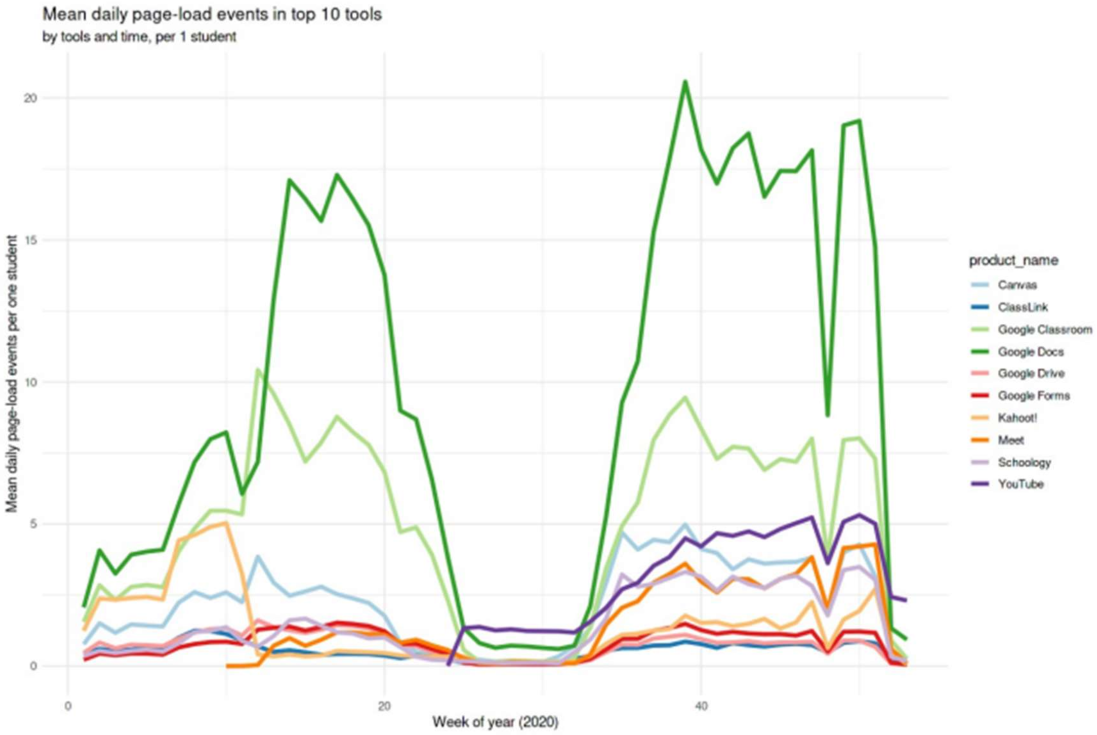
INFERENCE: Non Google Top 20 Products based on Engagement



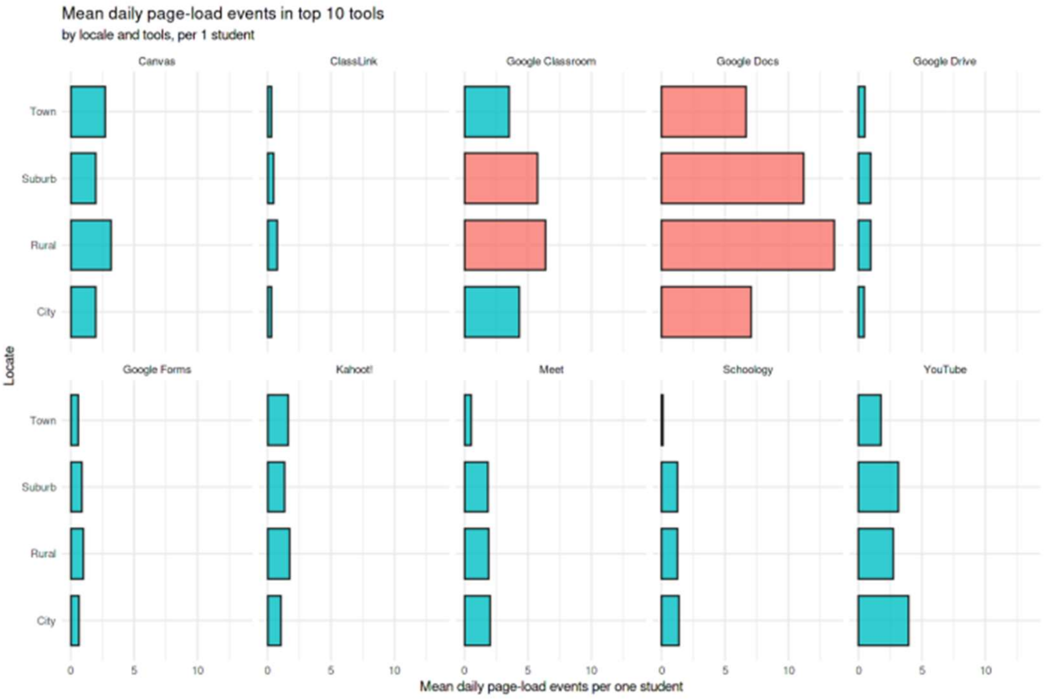
INFERENCE: Top 20 Digital Learning Platform based on Product Engagement



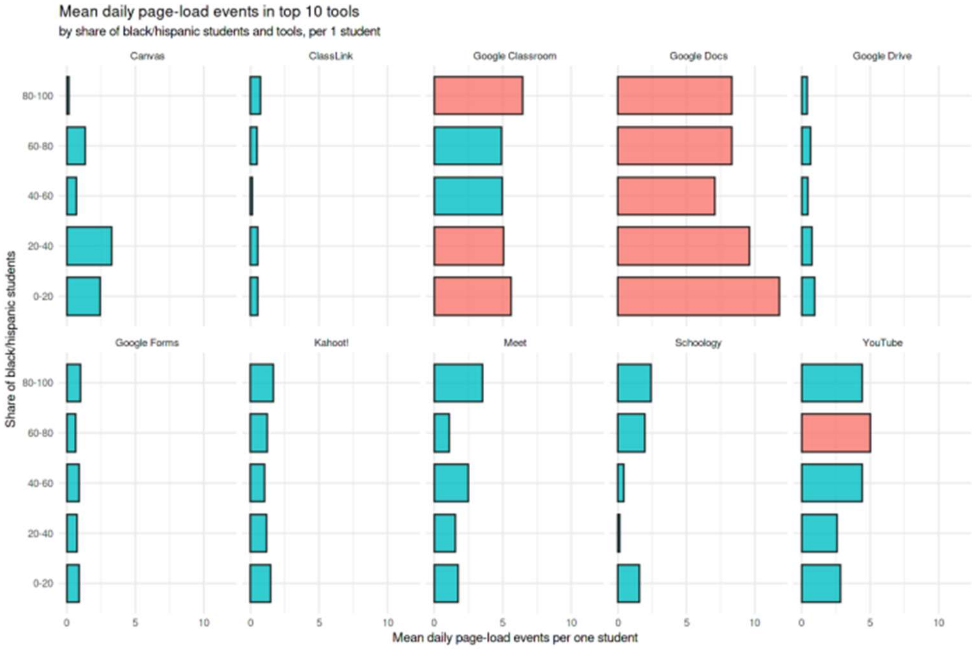
INFERENCE: Mean daily page-load events in top 20 Tools, per student



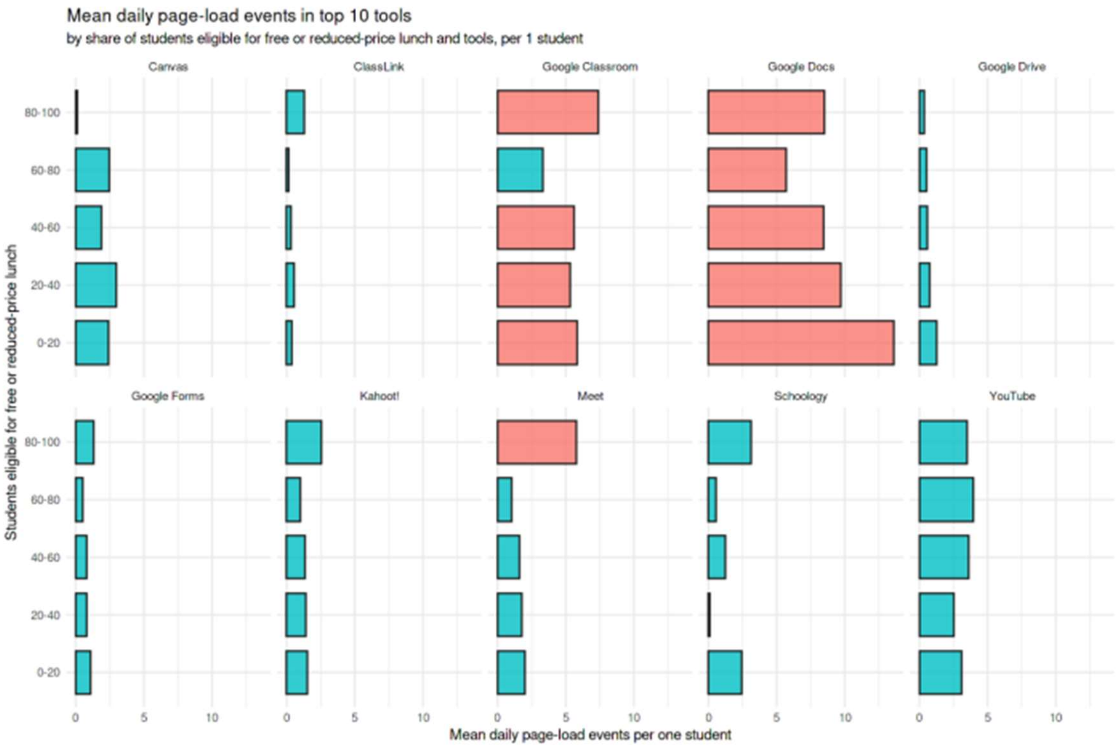
INFERENCE: Mean daily page-load events in top 10 Tools by tools and time per student



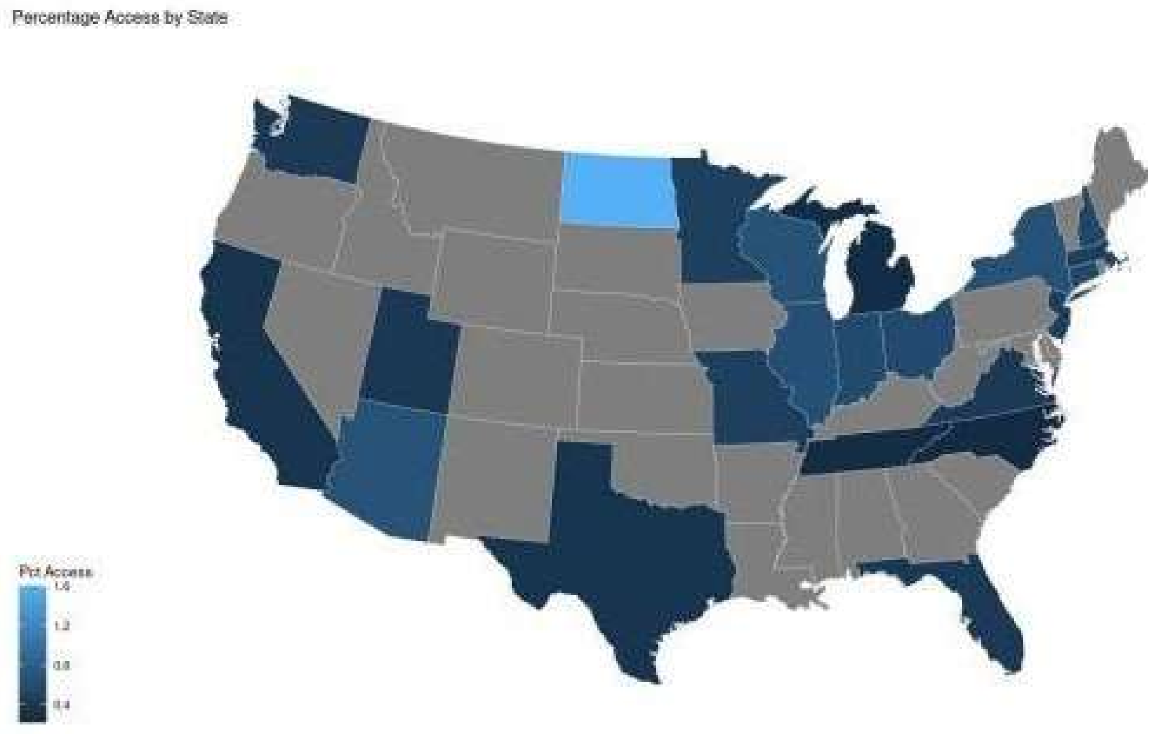
INFERENCE: Mean daily page-load events in top 10 Tools by locale and tools, per student



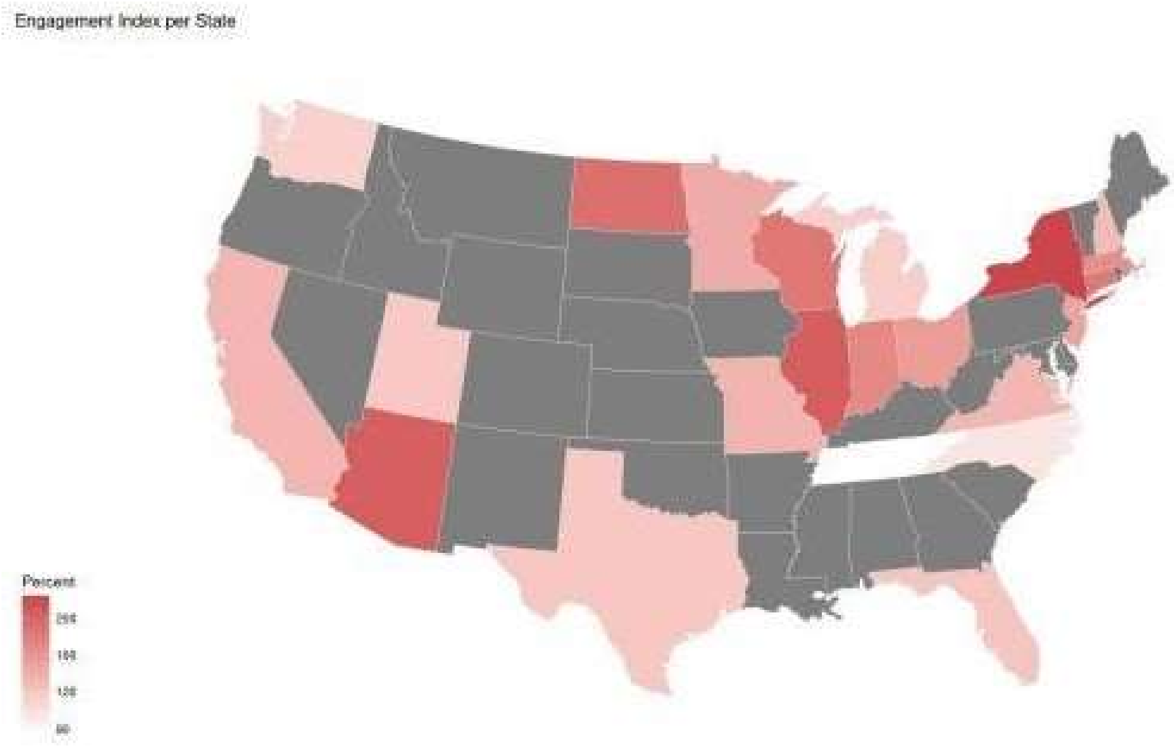
INFERENCE: Mean daily page-load events in top 10 Tools by share of black/Hispanic students and tools, per student



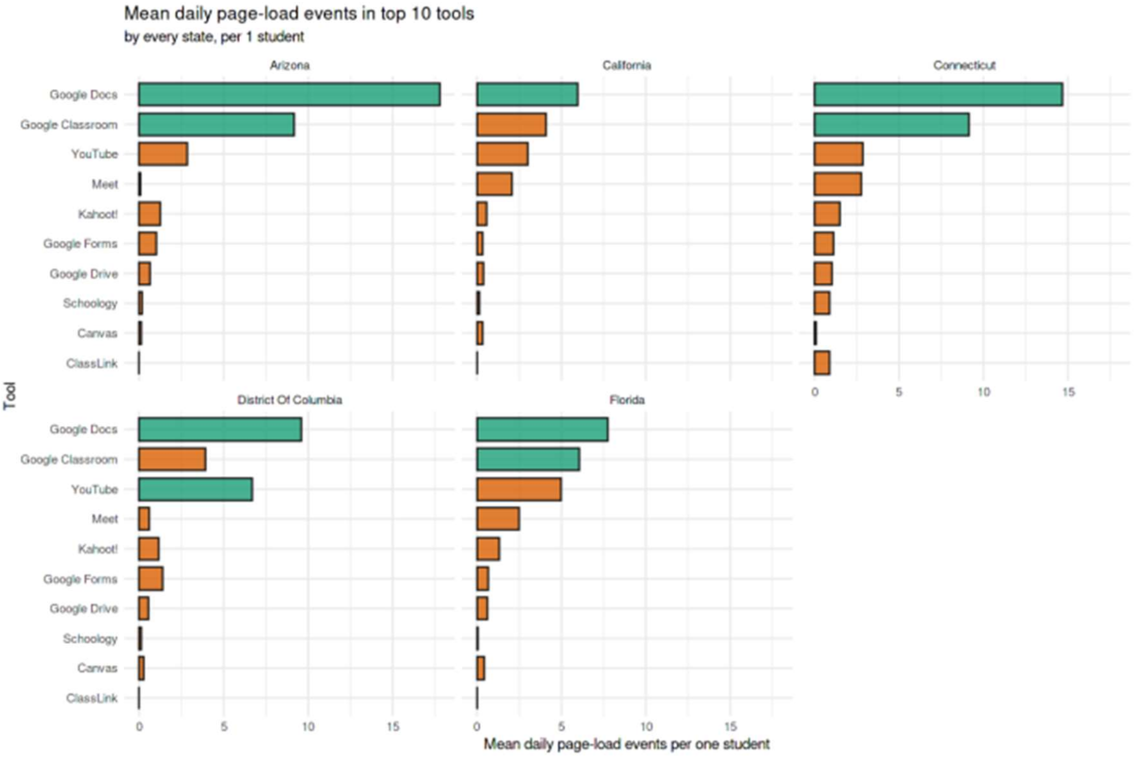
INFERENCE: Mean daily page-load events in top 10 Tools by share of students eligible for free or reduced- price lunch and tools, per student



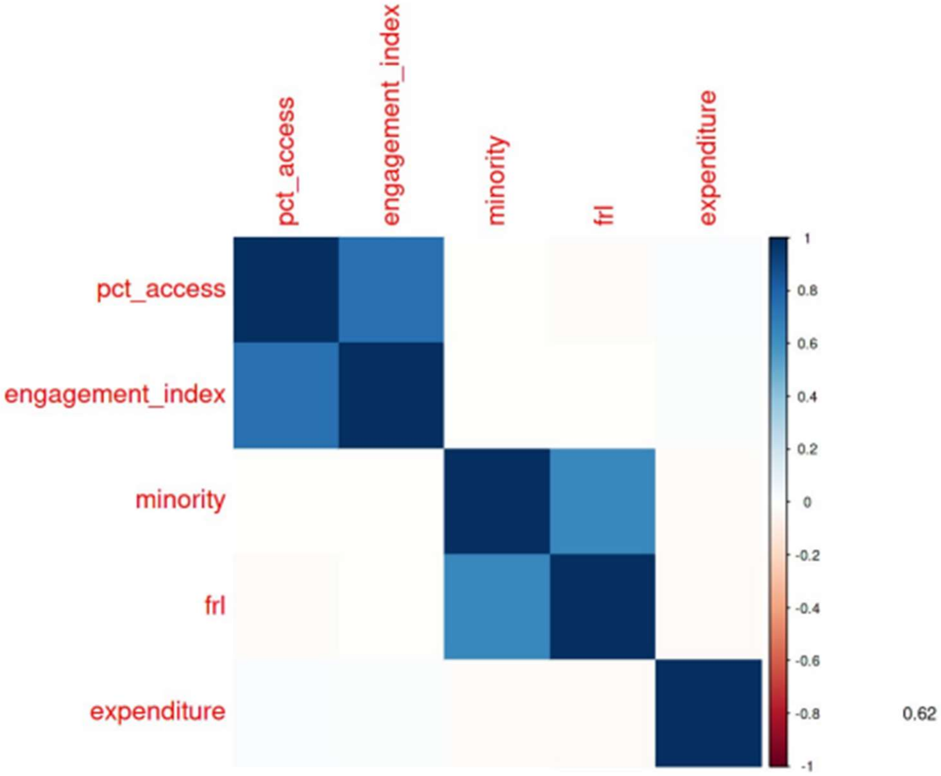
INFERENCE: Percentage of Pct Access by State



INFERENCE: Engagement in Digital Learning by State

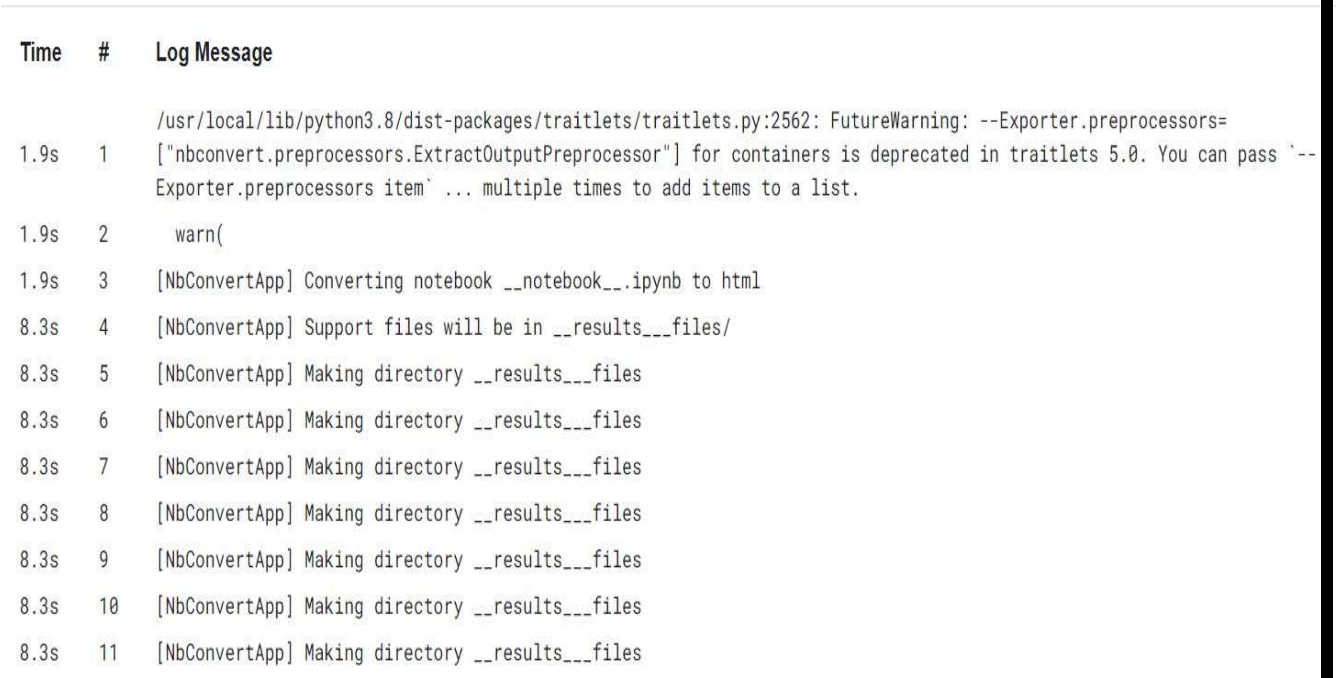


INFERENCE: Mean daily page-load events in top 10 Tools by every state, per student



INFERENCE: Correlation Analysis

5.2 PERFORMANCE ANALYSIS



```
Time   #    Log Message

            /usr/local/lib/python3.8/dist-packages/traitlets/traitlets.py:2562: FutureWarning: --Exporter.preprocessors=
1.9s   1    ["nbconvert.preprocessors.ExtractOutputPreprocessor"] for containers is deprecated in traitlets 5.0. You can pass `--
            Exporter.preprocessors item` ... multiple times to add items to a list.
1.9s   2      warn(
1.9s   3    [NbConvertApp] Converting notebook __notebook__.ipynb to html
8.3s   4    [NbConvertApp] Support files will be in __results___files/
8.3s   5    [NbConvertApp] Making directory __results___files
8.3s   6    [NbConvertApp] Making directory __results___files
8.3s   7    [NbConvertApp] Making directory __results___files
8.3s   8    [NbConvertApp] Making directory __results___files
8.3s   9    [NbConvertApp] Making directory __results___files
8.3s   10   [NbConvertApp] Making directory __results___files
8.3s   11   [NbConvertApp] Making directory __results___files
```

From the above snip we can conclude that the entire program execution is completed in a matter of 8.3 secs and hence, the code completes execution in record time.

However, the code uses large imported libraries to process the data acquired for computing product engagement, and impact of demographics on digital learning.Thus, the program has high space complexity despite having low time complexity.

6. CONCLUSION & FUTURE SCOPE

In conclusion, the project's analysis of the impact of demographic factors on digital learning engagement, as well as its prediction of product engagement for the year 2023, can provide valuable insights into the current state and future of digital learning. The project's findings can inform the development of targeted strategies and products to improve engagement and support for underrepresented students, ultimately contributing to reducing the educational gap between different socio-economic groups.

By understanding the impact of demographic factors on digital learning engagement, the project can help introduce new digital learning platforms focused on the demographics that have shown weaker interest and growth in adapting to digital learning mode. These platforms can be tailored to meet the specific needs and preferences of these groups, ensuring that they have access to high-quality education and opportunities for personal and professional growth.

Moreover, the project's findings can help existing digital learning platforms improve themselves in order to better accommodate neglected demographics. By taking into account the specific needs and preferences of underrepresented students, these platforms can provide better support and engagement, leading to improved results in the future. Overall, the project's contributions can help drive the growth and innovation of digital learning, ensuring that it continues to meet the evolving needs and preferences of students and educators. By addressing the impact of demographic factors on digital learning engagement, the project can help create a more equitable and inclusive education system that benefits all students, regardless of their socio-economic background.

In conclusion, the project's findings have the potential to make a significant impact on the field of digital learning research, and contribute to the development of more effective and inclusive education systems for the future.

(GITHUB LINK: https://github.com/Sakshi-saraogi/IMPACT-OF-DEMOGRAPHICS-ON-PRODUCT-ENGAGEMENT)

7. REFERENCES

- https://www.indiatoday.in/education-today/featurephilia/story/covid-19-impact-how-has-the-pandemic-affected-teaching-profession-1845333-2021-08-26
- https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

- https://openlab.bmcc.cuny.edu/inquirer/2022/09/19/the-effects-of-covid-19-on-students-learning-reshaping-online-practices/

- https://www.jstor.org/stable/1502821

- Islam, Md, Noor Asliza Abdul Rahim, Tan Chee Liang, and Hasina Momtaz. "Effect of demographic factors on e-learning effectiveness in a higher learning Institution in Malaysia." *International Education Studies* 4, no. 1 (2011): 112-121.

- Muthuprasad, Thiyaharajan, S. Aiswarya, K. S. Aditya, and Girish K. Jha. "Students' perception and preference for online education in India duringCOVID-19 pandemic." *Social Sciences & Humanities Open* 3, no. 1 (2021):100101.

- Naresh, B., D. Bhanu Sree Reddy, and Uma Pricilda. "A Study on the Relationship Between Demographic Factor and e-Learning Readiness among Students in Higher Education." *Global Management Review* 10, no. 4 (2016).

- El Refae, Ghaleb A., Abdoulaye Kaba, and Shorouq Eletter. "The impact of demographic characteristics on academic performance: face-to-face learning versus distance learning implemented to prevent the spread of COVID-19." *The International Review of Research in Open and DistributedLearning* 22, no. 1 (2021): 91-110.

- Nawa, N., Numasawa, M., Nakagawa, M., Sunaga, M., Fujiwara, T., Tanaka, Y. and Kinoshita, A., 2020. Associations between demographic factors and the academic trajectories of medical students in Japan. *PloS one*, *15*(5), p.e0233371.

- Alavudeen, Sirajudeen Shaik, Vigneshwaran Easwaran, Javid Iqbal Mir, Sultan M. Shahrani, Anas Ali Aseeri, Noohu Abdullah Khan, Ahmed Mohammed Almodeer, and Abdulaziz Abdullah Asiri. "The influence of COVID-19 related psychological and demographic variables on the effectiveness of e-learning among health care students in the southern region of Saudi Arabia." *Saudi Pharmaceutical Journal* 29, no. 7 (2021): 775-780.

- Radhika, S., and Ms S. Nivedha. "Research on effects of demographic factors on the English language learning among Tribal High School students." *International Journal of Applied Engineering Research* 15, no. 2 (2020): 111-113.

- Riady, Yasir. "The Demographic Factors and Decision on Selecting Open, Distance and Online Learning: Case study in Jakarta Regional office in 2017." *Management* 1: 496.

- Nicole, B., and B. Said. "Impact of demographic trends on the achievement of the Millennium Development Goal of universal primary education." In *Seminar on Population Aspects for the Achievement of the Millennium Development Goals*. 2004.