Consultancy Project with LearnPlatform (LP): Development of analytics models to assess distance learning and effects from the COVID-19 pandemic

Aaron M. Carr

University of San Diego

Master of Science, Applied Data Science

ADS-501: Foundations of Data Science

Section 02

02/21/22

Table of Contents

List of Tables

List of Figures

CRISP-DM Phases 1 and 2

Business Understanding

Background

Organization

CarR^2, LLC is a data science and analytics firm that specializes in assisting not-for-profit and public organizations to use their data for strengthening and expanding the efficacy and reach of social services, through providing a comprehensive approach to analytical and predictive modeling. CarR^2 uses a full-stack data science approach, which includes working with the customer to assess true business and research problems; performing necessary data manipulation (e.g., wrangling and transformation); creating robust descriptive and predictive analytical models; and working with customers to fully deploy a final, stable analytics solution product for ongoing business use.

The CarR^2 Data Science Team (DST) is proud and excited to partner with LearnPlatform (LP) on this innovative, critical, and timely project. LP launched in 2014 and has since won several awards for their work, which focuses on enabling education organizations to easily access and assess the best digital learning products (DLPs) to serve the needs of their student population (LearnPlatform, n.d., "About Us" webpage). There is a vast array of DLPs on the market today, with more deploying all the time. This, combined with each education district having criteria they need to meet in order to effectively serve their specific student population, LP serves an important need by providing one resource with an integrated standard assessment process that enables education organizations to leverage their finite money to choose the best DLP for them.

LP is headed by Karl Rectanus, CEO and the organization's Co-founder (LearnPlatform, n.d., "Our Team" webpage). Several divisions report directly to him, including People & Culture, Sales & Customer Success, Product, Marketing, Regional Sales for key locations, and Finance & Operations (see Appendix A for the full organizational chart). In consultation with Mr. Rectanus, along with divisional leaders and associated departmental directors, the DST has identified the need to focus the bulk of collaboration between the following groups and individuals within each organization:

DST project leader: Aaron Carr, CarR^2 CEO and lead data scientist

- For this project, Mr. Carr will be directly responsible for providing the data science expertise, including liaising with key team members within LP to ensure a thorough understanding of the exact business problem and goals. Mr. Carr will also oversee the research, creation, and development of analytic models for final deployment to LP.

LP lead project contacts: Maggie Smith, Director of Research; John Watson, Director of Engineering

- Together, Ms. Smith and Mr. Watson will ultimately be responsible for regularly assessing progress of the project from LP's standpoint, including ascertaining whether agreed-upon parameters and milestones are satisfactorily being met. Day-to-day, Mr. Carr and staff will work directly with team members in both Departments within the Product Division (see below for breakdown), who will also help the DST liaise with other LP divisional staff.

LP Product Division contacts: Kristie Lindell, Senior VP; Ms. Smith and Team; Mr. Watson and Team

- With the Product Division being fundamental to the successful deployment of analytic models, Mr. Carr will regularly meet directly with Ms. Lindell to provide broadly focused updates on model development and the meeting of milestones.
- A large of amount of coordination and work will be done with the Programming Team, led by LeVon James, within the Engineering Department. They will work with Mr. Carr and his team, using both Agile Data Science and Agile Software methods to avoid the waterfall trap that creates timing bottlenecks (based on siloed linear task management) between the Data Science and Programming Teams (Jurney,

2017). The goal of these Teams working together will be to turn the descriptive and predictive data analytic models researched and created by the DST into a fully deployable software application product. During all phases it will also be necessary to work closely with the heads of Network Systems Administration, Database Administration, and IT, in order to ensure all project models can be implemented within LP's data infrastructure.

- Mr. Carr and his staff will also work closely with the Research Team, led by Nancy Free, within the Research Department. Team staff includes two generalized research associates, one analyst, and one statistician. Coordination between the DST and the LP Research Team will facilitate transfer of critical inside knowledge and expertise related to the nature of existing data, including where is it stored, what individual features mean in context of the business problem, and what additional features could be derived to add value to the developed models.

LP Marketing Division: Gail Strong, Director of Market Research; Wendy Lite, Director of Outreach

- Both the Market Research and the Marketing Outreach Teams will provide invaluable input from the perspective of LP's customers. Understanding what the customer really needs and wants will ensure that time and resources are used only for value-added components, as opposed to working from misinformed assumptions that may threaten the viability of the entire project. As a note, the Marketing Research Team will coordinate closely with the Customer Success Department within the Sales Division as well, but generally speaking, Customer Success focuses more on implementation and customer understanding of existing tools, whereas Marketing is more focused on future directions. There is definitely some overlap, but it was determined that it would save time to assign one Department to liaise with the DST on a day-to-day basis. There is, however, a member of the Customer Success Department on the Steering Committee (detailed below).

LP Finance & Operations Division: Tom Walks, Director of Accounting

- It will be necessary to work with the Accounting & Finance Team to ensure that the analytics solution under development is being done is a cost-effective way, such that the eventual cost transferred to the LP's customers is within fair market value (FMV).

LP project Steering Committee

- The project Steering Committee will be co-chaired by Ms. Smith and Mr. Watson. Other members will include at least one director from each main division, including: HR from People & Culture; Customer Success from Sales & Customer Success; Market Research from Marketing; Marketing Outreach from Marketing; and Accounting from Finance & Operations.

Problem Area

The COVID-19 pandemic has had serious and long-lasting impacts on all aspects of life in the United States. The virus' potential to cause wide-spread hospitalization and death, along with how quickly it moved across the country starting in early 2020, resulted in federal, state, and local governments creating policies mandating restrictions on how and where people could congregate. Though these measures were done to limit the health effects and contagion rate, they had other social ramifications, which included closure of service industry businesses like restaurants and an unprecedented closure of public and private schools. According to a survey performed by the United States Census Bureau (2020), 93% of people who self-reported as having school-aged children in their house also reported that those children relied on "distance learning" in some capacity. For many schools and districts, this meant putting a great deal of effort into pivoting to online or digital resources as a way of continuing to reach students and meet their learning needs. As much as these efforts were critical in attempts to minimize the negative impacts from such a substantial shift in learning environments, there was also a potential to widen the digital gaps between children of differing socio-economic backgrounds. Referring again to the Census Bureau's survey, whether or not the distance learning was online or paper-based was dependent on household

income. Those households with income under $100K were far less likely to use digital resources than those whose income was above. Moreover, in terms of general education access, the schools closing created a potential for long-term whole-sale reductions in achievement of learning outcomes.

The main business problem areas are focused on identifying: The nature of online engagement in relation to several key factors, including district demographics, access to reliable internet, and government policy decisions; how engagement has been impacted by the pandemic, and whether some groups have been hit harder than others; and whether certain measures can improve online engagement for all students, but also more specifically for groups identified as facing serious and significant gaps in outcomes. As LP's primary business focus is on facilitating the use of digital and online learning tools for education organizations, they both have a key stake in providing government agencies with information to improve digital learning engagement and outcomes, but are also uniquely poised to quickly develop additional resources and capabilities to add to their "edtech effectiveness system" (EES) in order to guide those policy makers on making informed, evidence-based decisions about how better to serve the needs of all students (LearnPlatform, n.d., "Districts" webpage). See the Business Objectives section for more details.

Given the nature of LP's business model, as well as their strong placement in the market, there is a general acknowledgment and acceptance of digital techniques to strengthen business processes and increase service capabilities; this includes seeing the value of investing in data mining, machine learning, and predictive analytics. They are in fact currently working towards the creation of a new Data Science Department within LP, but given that they are only in the planning stages they still require input and expertise from an outside organization. All of this together means the company understands the potential value of this project, but currently just lacks the full knowledgebase to execute it on their own. As a result, the project itself is still in the middle of the planning process, though they did have a solid formulation of the business problem by the time CarR^2 was retained.

Reporting

Regular project presentations and reports

Steering Committee: Project update reports will be presented to the Steering Committee on a monthly basis. As this group includes non-technical stakeholders, the reports will cover general techniques and progress, but will avoid overly in-depth coverage of methodologies and jargon.

Project Teams: Regular meetings will occur based on need determined by the DST and all LP project teams. They may range from daily to weekly, and will include in-depth review of technical issues, statistical analyses methods and discoveries, research outcomes, and project and timeline progress. As an example of code that would be reviewed during a DST and Research Team meeting, see Appendix C (Jupyter Notebook code).

Final project presentation and report

The final project result will be presented at a meeting of LP senior-level management, including the CEO and all Division Senior Vice Presidents. The expected details will include high-level reiteration of the business problems and goals, as well as a non-technical description of the descriptive and predictive analytics models in the context of how they provide a solution to the problems. The focus will be on conveying broad concepts and value-add to the company and their customers, without going too deep into methodology (e.g., specific computer engineering or statistical methods used).

Current Solution

As noted above, though LP's main business expertise is in the digital product arena, they do not yet have a dedicated team of data scientists. Their Research Team in the Product Division does have some responsibilities related to analytics, but due to the relatively small size of the Research Department, they are teaming up with the

CarR^2 DST to provide additional assistance in researching and developing the predictive analytics tool to solve the project business and technical problems.

Business Objectives and Success Criteria

Business Objectives

From a business standpoint, the aim is to provide more targeted services that elucidate gaps in online engagement caused by the pandemic, along with being able to predict whether certain district or state-level policy and practice changes can mitigate or decrease those gaps. The new services will be incorporated into LP's existing EES, which is already specifically "used by districts and states to continuously improve the safety, equity, and effectiveness of their educational technology" (Kaggle, n.d., "Overview" section). As the pandemic and its impacts may continue into the foreseeable future, it is seen as critical by LP—whose central mission is to provide critical digital learning resources to educators to best serve students—to contribute to the activities that increase understanding of digital learning capacity and shortfalls.

In consultation with LP's Product Division Research Team and Marketing Division Market Research Team, the CarR^2 DST has outlined several business questions that will serve as the basis for expanded services in the LP portfolio.

Firstly, in order to know where to best utilize existing efforts to maximize gains, the following business questions related to descriptive analytics need to be addressed:

1. What are the measured characteristics of online and digital learning in 2020?
2. How does access to distance learning differ across the United States, focusing on the district and state levels?
3. What differences can be seen in terms of socio-economic status? What differences can be identified based on race/ethnicity and income level? What general differences can be seen across localities or between districts?
4. What measurable impact has the COVID-19 pandemic had on distance learning and specifically digital learning? Have the affects been different for students of differing socio-economic status?

Secondly, predictive analytics can used to examine the following business problems:

5. Which student populations are at risk of having low levels of online engagement?
6. Do governmental/public policy changes or state interventions change engagement in online learning? What about specifically for those groups already at a digital disadvantage?

Expected benefits from this project will include:

1. First implementation of robust data science and predictive analytics functionalities to the existing EES.
2. Utilization of specific predictive models focused on issues fundamental to increasing digital learning access and engagement.
3. A reduction in online engagement gaps among geographically and socio-economically diverse student populations.
4. Strengthening of existing relationships with customers, as well as an increase of the overall customer base.

Business Success Criteria

Outlined in the Business Objectives section are two main areas of focus for the new services being developed for this project. Below specific success criteria are addressed, as well as what LP groups will be responsible for their continued assessment.

Main objective 1: Descriptive analytics measures

Descriptive analytics methods will entail summarizing the data and creating visualizations appropriate to both the data's structure and the specific business objectives. The key success criteria will be based on whether the displays of the data create insights into its underlying nature, as well as point to key relationships between features that can be acted upon within the predictive analytics sphere. Success will mainly be evaluated by the Research Team in the Product Division, but they will also need to rely on expertise from the two teams in the Marketing Division, who will be able to speak directly to the needs of the customers. Post-deployment, all three teams will meet several times a year (e.g., quarterly) to review summaries and plots related to online engagement measures.

Main objective 2: Predictive analytics models

In terms of the predictive model development for deployment to the EES, the main criterion for success is ensuring that the models developed by the DST meet pre-determined thresholds for accuracy, precision, and long-term performance reliability. The value-add of the final models delivered to the education organizations, DLP designers, and state officials must result in actionable information in terms of changing and guiding policy to sustain and grow digital learning capabilities, as well as close gaps between those who currently have access and those whose access is less reliable. The models must be deployable in digestible formats via the existing LP inventory dashboard, meaning that the corresponding information summaries and visualizations must be readable and easily interpretable. Evaluation of success will be done within the Product Division Research Team, in consultation with the Programming and the Marketing Division Outreach Teams. Post-deployment, teams will meet regularly (e.g., monthly) to review reports related to model health and stability.

Inventory of Resources

Hardware

Project research and development

Mr. Carr and his staff have a network of PCs running Windows 10 that they use to access CarR^2 internal servers, LP's databases (where applicable), and the internet, as well as perform data science programming in Python, R, and MySQL.

As a digital service provider, LP maintains a full-stack hardware structure, including multiple web servers for front-end functionality and database servers for back-end processes and storage that rely on the Apache Spark framework. Individual LP employees also use a networked PC that can access and query company databases and warehouses using SQL, based on permission hierarchies specific to their department and role.

An assessment of conflicts between project access needs and network system maintenance schedules has been determined and an access schedule has been implemented to avoid unnecessary delays.

Project deployment

LP will host the analytics solution product on their servers in line with their general procedures. At this time no need for any special handling guidelines has been identified, as their servers are already capable of running a variety of software, including Python and R. Any relevant dashboards will be available via LP's generally accepted standardized methods for delivery of information to customers.

Data and Knowledge

Data sources

As noted in the Hardware section, LP internally maintains several databases for storage of data relevant to its aim to deliver digital services. Due to general security precautions, CarR^2 staff will not be given permission to directly download data from the mainframe. Instead, they will be able to review relevant features from pre-determined database tables and then they must request datasets from the Research Team, which will be provided in .csv format. Currently three datasets of interest available from the LP warehouse have been identified:

- Engagement Table, which contains data related to student access of DLPs, aggregated at the school district level.
- District Information Table, which includes information pertaining to individual school districts, such as selected measures related to race and socio-economic attributes.
- Product Information Table, which includes data related to the top 372 products pulled from LP's database of over 10,000 products.

Additional electronic resources and data will be available via the internet as needed, including the following examples: Excel files from the COVID-19 US State Policy Database; reference materials and datasets from the Annie E. Casey Foundation 2020 Kids Count Data Book; aggregated summary statistics from the Kaiser Family Foundation (KFF) State COVID-19 Data and Policy Action site; etc.

Other data and knowledge resources include expertise from LP staff and documentation related to standard business processes and policies.

Personnel

Though most of the following information is noted above in the organization contact section, it is worth reiterating that the following people will be critical resources for the continued progress and success of this project:

- Ms. Smith and Mr. Watson have been identified as project co-sponsors on the LP side.
- The DST will also need to work closely with Ms. Beth Jones' Network Administration Team and Mr. Nick Bubbles' Database Administration Team to ensure seamless integration of the new services into the company's data infrastructure. Ms. Jones and Mr. Bubbles will also assist with facilitation of data requests during research and development. Ms. Joan Tone's Information Technology (IT) Team will be integral if software or hardware utilized by any of the project teams requires updates, maintenance, or repair.
- The Research Team also contains key staff that will facilitate providing the DST with background information related to LP and its business model, along with specific details about features available to build the analytics models. The DST will also work especially closely with the statistician to ensure all statistical methodologies fit the analytics goals and are then implemented appropriately.

Requirements, Assumptions, Constraints, and RESOLVEDD Strategy

Requirements

Project output conceptualization

As noted in the Business Objectives and Success Criteria section—under main objective 2—it is a requirement that value be clearly added for each of the three main LP client target groups, with the following specific details for each:

Education organizations: The new information must allow districts to more efficiently comply with local and state policies surrounding digital learning and the effects of COVID-19, as well empower individual educators to tailor their own teaching methodologies to the students in their classrooms (LearnPlatform, n.d., "Education Services Agencies" webpage).

State education agencies: The new information must enable state public officials and decision-makers to assess where additional funding and resources should be allocated, or where new legislation is required to best serve students in the state by minimizing gaps in digital learning access and outcomes, while also strengthening digital learning for all K-12 students to mitigate the effects of the pandemic (LearnPlatform, n.d., "State Education Agencies" webpage).

EdTech product providers: The new information must enable DLP developers to provide evidence to education organizations and state agencies regarding use of their service/product as an effective digital learning tool (LearnPlatform, n.d., "EdTech Providers" webpage). It should also enable exploration of additional services as add-ons to providers' existing tools that will further the aim of addressing the pandemic and its effects on digital learning.

Project timeline

It has been deemed necessary to complete the project, excluding deployment, within 15 weeks. Table 1 is a broad outline of the project life cycle, developed using the framework from Wrike (n.d.); note that several items could have been listed in both the Execution and Controlling & Monitoring Phases (e.g., daily meetings with Research and Programming Teams) since they heavily overlap, but to avoid redundancies they are listed only in one section. For details related to a justification of the timeline, along with more specific milestones based on the CRISP-DM process, see the Project Plan / Order of Tasks section.

Table 1

LearnPlatform COVID-19 Project Life Cycle

	Weeks														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Initiation & Planning Phases															
Define Project															
ID Key Stakeholders															
Develop Resource Plan															
Create Project Plan															
Define Goals & Evaluation Measures															
Communicate Roles to Teams															
Anticipate Risks & Develop Contingencies															
Execution Phase															
Exploratory Data Analysis															
Data Preparation															
Develop Initial Models															
Ongoing Iteration to Model Improvement															
Coordinate Agile Data Science Plan															
Controlling & Monitoring Phase															
Daily Meetings w/:															
Research & Programming Teams															
Weekly Meetings w/:															
LP Sponsors (Ms. Smith & Mr. Watson)															
Marketing Team															
Network and Database Admin Teams															
Ethical Frameworks Forums (EFFs)															
Update Meetings w/ Senior-level Mgrs.:															
Director of Accounting															
Director of Market Research															
Sr. VP of Product Division															
Steering Committee															
LearnPlatform Legal Counsel Review															
Update Presentation to CEO & VPs															
Monitor:															
Project Timeline															
Risk Measures															
Overall Project Progress															
Project Closure Phase															
Inventory of Deliverables															
Training															
Review Initial Customer Response															
Evaluate Model Post-Deployment															
Organize Project Files															

Model evaluation

Performance: An evaluation measures framework (EMF) to test model performance will be created concurrently with the development of the models themselves. The EMF will not only be developed towards its use both during the Execution and Controlling & Monitoring Phases of the project life cycle, but also as an ongoing tool for assessing the health and sustainability of the models post-deployment. As such, the framework will be robust, but also flexible, to maximize its own performance in order to quickly pinpoint areas of risk that require timely addressing and adaptation of the predictive models.

Ethics: It is imperative that a guiding ethics framework be mapped to this project, which will allow for frank assessment of how model development decisions can introduce bias. Knowing that decisions will have an impact on different people differently is not enough; every effort should be made to minimize the models' negative impacts on one group over another. This specific area is discussed more the in the Constraints, RESOLVEDD Strategy section.

Other requirements

As the models developed as part of the project could help shape state and local policy, and therefore will potentially have an impact on a great deal of people, it is critical that they all go through a multi-stage, comprehensive review process that includes the entire DST, the LP Research Team (especially the statistician), Programming Team, Marketing Department, and senior-level LP managers. Additionally, the project must be reviewed by LP's independent legal counsel firm, Taylor, Anthony, Remington, Dierdre, Ingle, & Simpson, (TARDIS, LLC) in order to provide an appraisal of whether the methods or deployment of the new services adhere to state and local law, or if they pose a future potential legal risk.

As with all services provided over the internet, security will be an ongoing concern. Overall, this project must be compatible with LP's existing databases and internet security protocols. To ensure this is the case, security will be a standing agenda item during regular meetings between the Data Science, Programming, Network Admin, and Database Admin Teams. In terms of the data extract, transform, load (ETL) process, restrictions on internal data will not change, and any features that are deemed sensitive will continue to be transformed based on existing methods, e.g., binning the continuous data for the pct_black/hispanic feature to maintain maximum anonymization. Some of the data used for analytics is external, but assessment methods will be established to ensure that incorporation of such data into the analytics base table (ABT) will not allow any level of re-identification of personally identifiable information (PII).

As noted in the Background, Reporting section, regular meetings will be held to review project progress and depending on the attendees, group-specific levels of detail.

Assumptions

Assumptions are critical to every project. They are the parameters by which a finite, measurable product results. Without assumptions, all outcomes would have the same weight and meaning, and would therefore be devoid of specific use. The downside of needing assumptions is that choosing the right ones to guide the process is critical, and whatever assumptions are made automatically rule out a set of complimentary assumptions that by nature must be assumed not to be. Consequently, these choices introduce bias. The goal of any data science project is to ensure that any present bias leads the models to predict success without sacrificing the needs of one group over those of another.

Assumptions specific to this project

- There is a sufficient amount of data to produce the expected descriptive and predictive models.
- The relationship between descriptive features (e.g., pct_black/hispanic) and target features (e.g., pct_access) are measurable, repeatable, and can be used to generalize a predictive model for deployment into the wild, while also maintaining a high-level of continually assessable predictive performance.
- Any biases included in the data can be evaluated and attempts can be made to mitigate their effects.
- The data contains minimum noise, or if not, the noise can be corrected well enough to produce high-performing descriptive and predictive models.
- Converting time-based descriptive feature values into binary values (i.e., "before" and "after") will enable representative data instances for model building.

- An ethical framework (e.g., adapted RESOLVEDD, with a focus on meaningfulness, accountability, responsibility, transparency [MART]) can be implemented so as to reduce the impact of bias, both in the process of developing the models, as well as in the models' performance/outcomes (Vakkuri & Kemell, 2019).
- There is a sufficient amount of measurable value-add for new services to existing ones to justify the cost; value-add can include direct benefits (e.g., increase in client satisfaction, increase in client base) or indirect benefits (larger emphasis on promoting digital learning by state policymakers).

Constraints

After general research performed as part of a literature review, it is believed that the new services developed under this project are novel, and as such there are some naturally occurring restraints, which include: Finding data that sufficiently allow for predictive models to support the project's goals; making sure whatever data is available does not have too high a percentage of missing instances; determining whether use of the data within the planned context is permissible; and being able to establish an efficient data pipeline and ETL process between LP's internal databases and the DST. An additional key constraint related to being the first to perform a particular analysis is whether or not project teams have the expertise to build predictive models using the available data. As noted in the Background, Problem Area section, all data science techniques and methods will be coming from the CarR^2 DST, and as such they will be constrained by how well they are able to formulate an accurate sense of business domain knowledge and full comprehension of the true problem.

As it is anticipated the predictive analytics models will use time-based measures to delineate differences in outcomes before and after specific events, the project will be constrained based on whether clear segments can be achieved based on available time data, and that sufficient data exists in each partition to allow for production of adequately predictive models.

As noted in the Requirements section, this project is constrained by a relatively short life cycle (15 weeks), so everything must move quickly and efficiently. Any major issues or obstacles may endanger the overall success of the project if they cannot be addressed and resolved in a timely manner. Also, due to somewhat untested nature of the service, the production cost must be kept to a minimum. Fortunately, the product is a web service and will not require any additional costs for capital equipment. The majority of project costs will go to carve-out time for LP staff—though some of it will naturally overlap with general job responsibilities—as well as the consultancy fee for CarR^2, which has been negotiated as a $10,000 flat-fee. The overall project cost must not exceed $15,000.

One additional constraint is related to the pandemic and its effects on LP staff. Most staff are currently working either full-time remotely or in a hybrid model. Though overall this issue is not insurmountable, not having the team in one location could potentially cause some issues in terms of ease of meetings, presentations, and discussions. Also problematic is whether or not individual project staff need to be out for an extended time in order to attend to their own health. This will need to be dealt with as a risk, along with corresponding contingencies, as it is critical employees take the time they need to address health concerns, but some overlap of responsibilities will need to be accounted for from the start in case someone must be out for an extensive length of time (see Risks and Contingencies section).

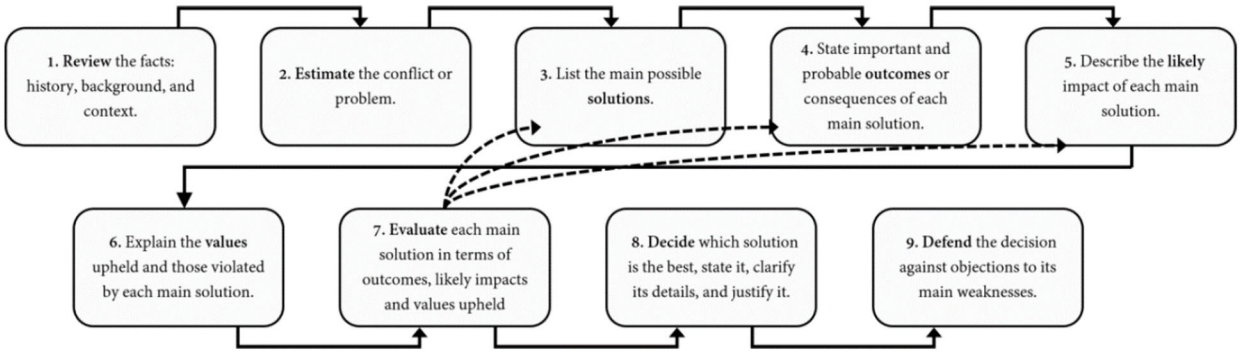
RESOLVEDD Strategy

A key aspect of any data science project lead by CarR^2 is the implementation of an ethics framework, mapped to the particulars of the project. Mr. Carr and his staff believe that consideration of ethics and bias are of central importance in order to prevent the deployment of models and systems that end up benefitting some groups at the expense of others. As noted in the Assumptions section, it is unavoidable that decisions of which features, data, and inductive bias methods to utilize will naturally exclude others, but every effort must be made to ensure that those choices do not "bake-in" systemic prejudices and methods of oppression.

As standard operating procedure, the CarR^2 DST uses the RESOLVEDD Strategy, as adapted from Vakkuri and Kemell. Figure 1 is a diagram showing the nine steps of the process.

Figure 1

The Nine Steps of the RESOLVEDD Strategy



Note. From "Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy," by V. Vakkuri and K.-K. Kemell, 2019, in S. Hyrynsalmi, M. Suoranta, A. Nguyen-Duc, P. Tyrväinen, & P. Abrahamsson (Eds.), *Software Business* (Vol. 370, pp. 260–275). Copyright 2019 by Springer International Publishing.

Mr. Carr and his staff have been found that using RESOLVEDD helps maintain the importance of ethically aligned design (EAD) considerations at every stage of the project (Vakkuri & Kemell). Moreover, emphasizing its importance as a tool to confront and combat biases makes it not feel "tacked-on". Vakkuri and Kemell also express the importance of acknowledging the key part that MART plays in the process of implementing a data science project using EAD:

- **M**eaningfulness deals with contextualizing the need for introducing EAD into the work process and making its incorporation necessary to the eventual output; it is not just a superfluous aspect that has to be crammed in at the end.
- **A**ccountability is related to assigning leaders that ensure methods and product outputs are meeting agreed-up metrics for EAD, but it also has to do with emphasizing to individual staff that their decisions and output must adhere to ethical standards.
- **R**esponsibility is more amorphous but has to do with instilling a pride-of-ownership in the final product and a strong sense of social justice that CarR^2 strives for.
- **T**ransparency is key to entire RESOLVEDD process, and its emphasis serves to communicate trust in the development process, model design, and predictive output. Transparency is about clearly outlining and making public the assumptions used to build models, with the goal of gaining a diversity of perspectives and attitudes to ultimately create data science models that do good for all.

In the current project, Ethical Framework Forums (EFFs) will be implemented starting in week six as part of the Controlling & Monitoring Phase of the project life cycle (See Table 1). In the first EFF session, the RESOLVEDD strategy and anticipated steps toward its implementation will be presented and discussed. In following weeks, specific issues from this project can be brought forward by any staff member, and each item will be mapped to the RESOLVEDD strategy, which will also be generally reviewed to ascertain progress on previous issues and overall progress. CarR^2 has historically found that the EFF is key, whatever the project, for allowing open dialogue of concerns where every effort is made to ensure that everyone has a voice, and that voice is heard. Additionally, readings will be assigned on regular basis from journal articles pertaining to ethics in design, general sciences, and

data science to prompt broader discussions. Though the goal of the EFFs is to produce actionable items, it is not meant to be a space where technical details are discussed.

An example of an initial concern is related to lack of data on specific demographic makeup from some states. This creates a potential for selection bias, such that data needed to train the predictive models is missing, thus skewing the results. One possible solution is to perform an in-depth search for additional resources online. However, if the underlying issue is that the states are simply not collecting that data, a more involved alternative would be to reach out to public representatives within states where data is lacking and express the need for and importance of implementing stronger data collection methods. While this may result in hit-or-miss outcomes, a concerted campaign may gain some results, especially if evidence is provided for the usefulness of digital learning services in a country dealing with a pandemic. The values upheld by these solutions, if successful in obtaining more data, would be those around working to make data representative of the populations this project is working to help and shed light on inequities they are experiencing. It is estimated by the DST and LP Research Team that more than likely the states that are behind in data collection are those that also have large gaps in digital learning access and outcomes. Getting more data would further elucidate the breadth of the problem. Both solutions have drawbacks and opportunities for failure, but both may be critical in order to gain a clearer picture of the current state of digital learning

Risks and Contingencies

Identify Risks / Develop Contingency Plans

Anticipated risks and subsequent contingency plans are outlined below, broken down by several different categories:

1. ***Product Value***
    a. Risk: Customers do not see the value-add of the new services, or the information is not intuitive to act upon.
    b. Contingencies: This could occur if the descriptive displays and predictive models do not address actual customer needs and wants. To prevent this, regularly collaboration between the Data Science, Programming, Research, and Marketing Outreach Teams will be important to assess whether current project progress is meeting the actual customer requirements. Action items would be evaluating customer needs vs. model functionalities within team expertise, as well as outreaching directly to customers for feedback.
2. ***Financial***
    a. Risk: Customers do not have funds to continue, and consequently the return on investment (ROI) is little or none.
    b. Contingencies: This could be an issue if funding is cut, either at the educational organization, district, or state level. The DST must work closely with the LP Accounting & Finance Team to ensure that established cost thresholds are being met to minimize loss and maximize potential revenue, even if the revenue available is less than anticipated. While LP has no direct say in public budget efforts, creating marketing materials that emphasize the reward vs. cost to the customer (for example, in a cost-effective way the new services serve an integral societal need that cannot simply be ignored) may help convince them the services are a high priority for use of limited funds.
3. ***Model Health***
    a. Risk: Prediction models are not very robust.
    b. Contingencies: This may occur in several situations:
        i. The DST is unable to develop a sufficiently predictive model, given the available data. To avoid this, the DST will work closely with the statistician to maximally leverage statistical

methods to ensure models meet strict predictive metrics. See Data Integrity for a discussion specifically concerning data quality.

ii. A model works in the lab but does not meet expectations once deployed to the wild. Firstly, making sure to use test and training datasets will help prevent model under- or overfitting. Secondly, model development is an iterative process, so if it does not work as anticipated, the Data Science, Research, and Programming Teams can reassess why it did not work with the new data and revise as necessary. There may even be a need to add an additional step of model development that includes a validation dataset, which allows refining of a model pre-deployment.

iii. Even if the model works at first, concept drift creeps in relatively quickly. This can occur again because the data did not sufficiently enable a model that would adapt to additional data input, or could also occur because relationships between features shift over time. To mitigate concept drift, regular and continual evaluation of the models will need to occur post-deployment.

4. Data Integrity

a. Risk: The data was insufficient to develop effective models, or there were clear errors/noise in the data.

b. Contingencies: This could occur for a variety of reasons, including data entry error, a large number of outliers that are skewing statistical measures, or just lacking either enough data or specific features to build useful models. The DST will work with the Research Team in the Product division, including with the researchers and statistician, to identify issues with the data and resolve accordingly.

5. Process Issues

a. Risk: Problems with project process can occur at multiple time points or across multiple domains.

b. Contingencies: A prime example would be the software development waterfall problem, where one team cannot continue with their project efforts because they are waiting on another team to complete their tasks (i.e., development becomes linear instead of iterative and segmentable). There are three main methods that will be deployed to reduce risk and/or provide contingency if an issue arises:

i. Create an overall broad plan for how to add issues that may arise throughout the project. This should be done during the planning process and should group potential obstacles and challenges into conceptually similar sections so that key stakeholders and resolution leaders can be identified.

ii. Use of Agile Data Science methods, which rely on breaking tasks into smaller segments, along with iteration with each team, so that tasks can be managed and reexamined while waiting for more input from other teams.

iii. Regular meetings with key teams, along with project leaders (i.e., Mr. Carr, Ms. Smith, and Mr. Watson), to discuss project progress, to include specific forums for addressing identified potential issues. Regular may mean weekly, daily, or ad hoc, dependent on project stage and what teams are involved.

iv. Create some overlap or redundancies of team staff responsibilities in case anyone has to be out for an extended period. As the overall project must fit within the strict 15-week deadline, there is little margin for delay.

Terminology

Business Terminology

Digital learning gap – Measurable and significant disparity between groups differentiated by key features, such as locale, district, state, race, ethnicity, access to reliable internet, etc.

Digital learning products (DLPs) – Any product that can be used in a digital environment—including locally on a computer, on a smart phone, or in a browser on the internet—that focuses on achieving a learning outcome for a specific student population.

Digital platform – Any device or service that enables access to digital resources.

Distance learning – Any occurrence of an educational experience outside the traditional learning environment in which educators are in the same location as the students.

District – One of the main customer bases for LearnPlatform (LP). They are educational entities made from dividing states into segments that serve geographically localized student populations.

Education and teaching technology (EdTech) – The broad term for technology related to providing education opportunities through a digital platform. It can include apps, software, web services, extensions, eBooks, and more.

Education organization – Includes public run educational entities, as well as foundations and private companies. Focus is on K-12 schools and includes a differentiation between public institutions and Charter schools.

EdTech effectiveness system (EES) – As stated by LP, a digital resource to provide education agencies with the ability to "ensure the safety, equity and value of their edtech investments" ("Districts" webpage). This is the core product of LearnPlatform and is meant to customize and maximize the customer's access to appropriate and applicable EdTech.

EdTech product providers – The companies and organizations that develop and deploy DLPs for use by education organizations and the public.

Educator – A provider of education services, including teachers, principals, and education organization administrators.

Inventory dashboard – A central location for displaying visualizations based on EdTech data. Loaded via a user installed browser extension. Per the platform, there is no cost for a district to download and install it.

Online engagement – The measurable access of digital and internet resources and products by students. For this specific project, this value, which is recorded in the engagement_index raw feature described in the Initial Data Collection report section, is based on page load events.

State education agency – A governmental entity responsible for administration and management of state-level education policy and guidelines.

Students – Individuals who are targeted to benefit and gain knowledge from the learning process.

Data Mining and Data Science Terminology (Kelleher et al., 2020)

Aggregate – Perform a summarization of a feature's values to obtain a single representative value. Common methods include summation, counting, and averaging.

Algorithm – A set of steps used by the machine learning (ML) model to process data to solve a problem. The steps can either be strictly pre-defined or based on guiding parameters, depending on the model and desired outcomes.

Analytics base table (ABT) – The "working" table once data has been accumulated from multiple sources, wrangled, transformed, and condensed in a single unified structure.

Analytics solution – A solution to the business problems that uses analytics methods, which is based on use of the scientific method to achieve clear and repeatable results.

Concept drift– The idea that over time even the most high-performing predictive models will lose predictive efficacy.

CRISP-DM (CRoss-Industry Standard Process for Data Mining) – A comprehensive, generalized, and broadly applicable procedural framework for implementing data mining and ML models to solve complex business and research problems.

Data manipulation – An important step in working towards creating the ABT; it is the process of joining different tables, creating derived features, and filtering out instances or features. This is often considered an extract, transform, and load (ETL) process.

Dataset – A collection of data instances (records) that contain key informational elements (e.g., attributes) about a sample of individuals for which each element contains a datapoint for a measured value of that element. An example could be a tabular Excel document for product users, with columns representing attributes (like name or user id) and rows representing instances (for each user).

Derived feature – A feature that is created using the values from other existing or raw features (e.g., aggregate values).

Descriptive feature – Also known as an independent variable, this is a feature that is used by itself or in combination with other descriptive features to predict the value of the target feature.

Domain concepts – High level conceptual categories used to delineate and assign individual features within the analytics solution.

Ethically aligned design (EAD) – "Refers to the involvement of decision-making in practice and ethical consideration in the practice and design [of] AI and autonomous systems and technologies" (Vakkuri & Kemell).

Evaluation measures framework (EMF) – An adaptive set of standards used to assess a model's performance from multiple perspectives (e.g., relative to business and/or data mining objectives) and time points (e.g., pre- and/or post-deployment).

Extract, transform, and load (ETL) – The process that involves accessing data, transforming it, and then transferring it to a new location (Microsoft, 2021).

False negative predictions – The number of actual positive instances predicted to be negative by an ML algorithm.

False positive predictions – The number of actual negative instances predicted to be positive by an ML algorithm.

Feature – Also known as a variable, attribute, or column (in the case of a tabular dataset). It contains values within a range particular to the thing that is represents. For example, an "age" feature should ideally contain a range greater than zero and less than a reasonable threshold (e.g., 110 years of age). Anything outside of that range would ostensibly be an outlier, and may even be the result of recording error.

Generalize – To take a developed ML model and apply it to data that was not part of either the train or test datasets. The model's effectiveness and long-term stability will need to be evaluated regularly to determine whether concept drift has occurred.

Hold-out test dataset – See ***test dataset***.

Inductive bias – The data scientist must set up rules or guard rails for the ML algorithm in order for it to differentiate between models. See restriction bias and preference bias.

Insight – A particular new idea achieved through assessment of data, which can be used to guide decision-making.

Instance – Also known as a row of data, which corresponds to an observation of particular values.

Machine learning (ML) – As stated by Kelleher, ML is at its most basic an automated process with the goal of extracting patterns from data that are impossible (or nearly so) to achieve using conventional computer algorithm development.

Model testing – Using a subset of a data sample, withheld from the model training process, to assess the accuracy and proficiency of a predictive model.

Model training – Using a subset of a data sample to develop the prediction model, which can then be tested using the model test dataset.

Noise – Data values included in a dataset that have issues that introduce potential to skew analysis results and thus end up with a poor performing predictive model. Issues could include errors or substantial outliers.

Overfitting – Production of a predictive analytics model that too closely resembles the underlying data, in which case it is more likely to perform poorly given new data inputs. See also ***underfitting***.

Precision – A performance measure of classification that divides the number of true positive predictions by the sum of the true positive and false positive predictions. It relays confidence in how good the model is at predicting the nature of the positive class accurately.

Prediction – Using existing data to determine a value that has not yet been observed.

Prediction model – A conceptual framework for ML that represents a set of rules that describe the relationship between data inputs and prediction outputs.

Prediction subject – This is the elemental level of the data, at which each instance is referring to one prediction subject. In the case of LP's engagement data table, the prediction subject would be districts.

Preference bias – A type of inductive bias. It steers the ML algorithm to value some models over others.

Proxy feature – A feature that is used as a stand-in for another attribute that is either unmeasurable or otherwise unattainable—the latter could be due to prohibitive cost or lack of permission to access the desired data.

Raw feature – A variable that is used in the modeling process that comes directly from a data source and has not been derived.

Recall – Also known as sensitivity or true positive rate (TPR), it is a performance measure of classification that divides the number of true positive predictions by the sum of the true positive and false negative predictions. It relays confidence in how good the model is at identifying all of the instances with the positive target level.

RESOLVEDD strategy – A nine-step process outline used to lead consideration of professional ethics problems and solutions (see Figure 1). It is used generally as a project's ethics evaluation framework by the CarR^2 DST, with emphasis on promoting meaningfulness, accountability, responsibility, and transparency (MART) throughout the project process and ML methods.

Restriction bias – A type of inductive bias. It sets limitations that dictates which models the ML algorithm will use during the learning phase.

Sample – A dataset of instances that are pulled from the population of interest.

Sampling method – The process by which samples are pulled from a larger dataset of instances. The exact method used determines the final sample makeup.

Sensitivity – See ***recall***.

Supervised machine learning – An ML method that involves using a dataset with delineations (or labels) between descriptive features and a target feature to develop a predictive model. Examples include regression and classification.

Target feature – The dependent feature/variable that you are attempting to predict.

Test dataset – Also known as a hold-out test dataset. It is used to test the trained model for performance. Using two different datasets (one for training and one for testing) prevents overfitting. See ***train dataset***.

Train dataset – The dataset used to train the ML model. Using two different datasets (one for training and one for testing) prevents overfitting. See ***test dataset***.

True negative predictions – The number of actual negative instances predicted to be negative by an ML algorithm.

True positive predictions – The number of actual positive instances predicted to be positive by an ML algorithm.

True positive rate (TPR) – See ***recall***.

Underfitting – Production of a predictive analytics model that does not closely resemble all of the underlying data and therefore performs poorly. See also ***overfitting***.

Unsupervised machine learning – An ML method that involves examining data where the relationship between features is unknown or too difficult to ascertain. An example method includes k-means clustering.

Validation dataset – Sometimes this additional step can be placed between the data training and testing phases. This is generally done to fine-tune models using additional data elements.

Validation scheme – A set of measures used to evaluate the ongoing performance of a predictive model post-deployment.

Data Mining Goals and Success Criteria

Main objective 1: Descriptive analytics measures

Goal: The main goal for objective 1 is to use CRISP-DM data description and summarization methods to generate displays of data that provide insight into the data relative to the business objectives stated in the Business Objectives and Success Criteria section (Chapman et al., 2000, p. 66). The expected outputs related to this goal are outlined in Table 2.

Table 2

Descriptive Analytics Data Mining Outputs

Tabular Summaries		
Type	**Purpose**	**Example of Variables Affected**
Measures of central tendency and variability Table	Provide statistical measures that provide location (e.g., mean, median, mode, max, min) and dispersion (e.g., standard deviation, range, IQR) of features with numerical values.	pct_access; engagement_index; pct_black/hispanic*; pct_free/reduced*; county_connections_ratio*; pp_total_raw*
Contingency Table	Method to compare two discrete features, whereby expected frequency values can be calculated and used to test for feature independence.	pct_access / engagement_index
Correlation Table	A cross-wise table displaying the correlation coefficient value for each pair of features (row v. column), which indicates the relative likelihood for a potential linear relationship to exist between them.	pct_access / engagement_index
Association Tables	Compare frequencies of two categorical features based on count of specific categories for each.	state / locale; district_id / pct_free/reduced; district_id / Product
Visualizations		
Type	**Purpose**	**Example of Variables Affected**
Scatterplots	Graph comparison of two features with numerical values, either discrete or continuous, used to determine whether a relationship is discernable.	pct_access / engagement_index
Density plots	Graph showing the probability density of a feature's values.	pct_access; engagement_index
Boxplots	Graph showing the measures of centrality and dispersion for a numerical feature. Items displayed include the 25th percentile, median, 75th percentile, whiskers connecting the IQR to the lowest non-outlier value, and outliers.	pct_access; engagement_index
K-Nearest Neighbor (KNN) Classification Plots	Provide a scatterplot of two variables with the additional characteristic of visual clustering by a third variable. Clusters are generated by a KNN algorithm and are based on distance of points to measures of mean.	pct_access / engagement_index; district
Histograms	Provide frequency/density plots for either discrete or continuous data.	pct_access; engagement_index

* For features with binned numerical values, mode and mean can be calculated

Success Criteria – Success will be measured based subjectively on the insights gained from reviewing the data, particularly how they provide context to the predictive models. Such subjectivity will generally be based on the perspective of the Research Team, including the statistician.

Main objective 2: Predictive analytics models

There are two main data mining goals that correspond to business objective 2:

Goal 1: Use classification methods, such as k-nearest neighbor, to predict what districts face shortfalls in terms of being able to provide services to their students (Chapman et al., pp. 69-70). Features examined would include the level of engagement (engagement_index); the ratio of County-level broad band high-speed connections (county_connections_ratio); and the median value of per pupil spending by schools within districts (pp_total_raw).

Success Criteria 1: Classification models will be assessed using performance evaluation measures such as precision and recall, or by using F_1, which is their combined harmonic mean; see the Data Mining and Data Science Terminology section.

Goal 2**:** Use prediction methods (e.g., regression), in conjunction with out-of-time sampling, to develop models looking at changes over time to predict the level of student percent access (pct_accesss) based on what district, state, and locale they live in; the percentage of people who are Black and Hispanic and who reside in the district; and the percentage of students eligible for meal cost reduction, all relative to state policy changes enacted at measured intervals (Chapman et al., pp. 70-71).

Success Criteria 2: An example of success would be creating linear regression models that compare time periods and are able to achieve coefficient of determination levels between 0.8 and 0.9. Additionally, once deployed to the wild using real world data, they would maintain statistically similar levels of precision and accuracy as seen in development.

Project Plan / Order of Tasks

Project timeline

In consultation with LP Team staff, as well as management level VPs, a relatively short timeline has been allocated for this project. The provided justification points are as follows:

1. According to the NCES (n.d.), there are currently approximately 49.4 million pre-kindergarten to high school aged children in the US. As such, effects of the pandemic on nation-wide learning are potentially adding up to an overall tremendous loss.
2. After a thorough assessment of project needs, and relying heavily on Agile Data Science methods, the DST estimate that a productionable set of models can be ready within 15 weeks.
3. Even once the customers have access to the information, the process to review the information (let alone enact changes to organizational and public policy) can often require a great length of time, so the earlier the product goes to market the better.

A series of Gantt charts have been developed to detail each required step of the project (see Appendix B, Tables 10-14). Gantt charts are generally helpfully for providing a clear visualization of a project lifecycle, especially how each task temporally relates to all of the others. The charts will be referred to throughout the project, and will even be updated should modifications to the timeline be necessary.

Data Understanding

Initial Data Collection Report

Data Requirements Planning

To begin with, the DST has been provided with three data sources from LP's internal databases, which are detailed in the Selection Criteria section.

In the initial planning stages, it is believed by project staff that additional data will be needed to address all of the stated business problems. This may include the websites referenced in the Inventory of Resources, Data and Knowledge section. Additionally, some existing measures are based on the National Center for Education Statistics website (NCES) and it may prove useful for further data.

Selection Criteria

Engagement data tables

This is a series of individual .csv files, with the prediction subject level being school districts. This table is key to both the descriptive and prediction analytics process, as it records event details for students directly accessing DLPs. Mr. Carr and his staff were presented with a folder that contained data on 233 different districts. The dataset has the following features that will be incorporated into the ABT:

- **time**: The date of an individual engagement event performed by a student. This will be critical for examining time dependent relationships.

- **lp_id**: The identification number of the product being accessed, which can be used a foreign key when joining this table with the production information data table.
- **pct_access**: The percentage of students in a district that accessed a product at least once in a given day.
- **engagement_index**: Total number of page load events per 1,000 students in a district relative to the product in each instance.

At this stage of the project, it is anticipated that all of the provided features will be needed to either build the final model, either directly (through use as a raw feature) or indirectly (through use to derive new features).

All of the individual district data tables will need to be joined, and a newly derived feature will need to be created designating which district the data was from—this new feature will be a foreign key for the district information data table.

District information data table

This is one .csv file that contains several variables that provide values for certain socio-economic measures. Some of the features have been anonymized either through use of a specific anonymization tool, or by being generalized into a range. The dataset has the following features that will be incorporated into the ABT:

- **district_id**: The district identification number to be used as the table's primary key.
- **state**: The state in which the district resides. This will be useful for comparing correlations across US states.
- **locale**: Categorical classification based on NCES; the four types are "City", "Suburban", "Town", and "Rural".
- **pct_black/hispanic**: Data from NCES on percentage of students identified as Black or Hispanic.
- **pct_free/reduced**: Data from NCES on percentage of students eligible for meal cost reduction.
- **county_connections_ratio**: Extrapolated from December 2018 data on the FCC website. It represents the ratio of county-level high-speed connections. This feature will be key for assessing relative internet connection capacities.
- **pp_total_raw**: Raw data was from the National Education Resource Database on Schools project, but was transformed by LP as the median value of per-pupil spending by schools within a district.

At this stage of the project, it is anticipated that all of the provided features will be needed to either build the final model, either directly (through use as a raw feature) or indirectly (through use to derive new features).

Product information data table

This is one .csv file that contains several variables related to products contained in LP's expansive database. In was decided that to efficiently utilize existing resources, analyses would focus on only the top 372 products, as determined based relative to the number of users in 2020. The dataset has the following features that will be incorporated into the ABT:

- **LP ID**: Product identification number to be used as the table's primary key.
- **Product Name**: Name of the product.
- **Sector**: Education sector in which the product is used.
- **Primary Essential Function**: Two-tiered classification, with subcategories under the following main ones: "LC" (learning and curriculum); "CM" (classroom management); "SDO" (school district operations).

The following features were removed, as they are either redundant or it is believed they are not relevant to the predictive modeling process: URL; Provider/Company Name.

COVID-19 US state policy database data table

This dataset was downloaded from openICPSR in Excel format. It includes data related to state-level policies enacted to slow or prevent the spread of COVID-19 infections (Raifman, 2021). The dataset has the features in Table 3 that will be incorporated into the ABT.

Table 3

COVID-19 US State Policy Selected Feature List

Variable Name	Variable Label	Description	Unit	Values/Value Range	Value Labels
STATE	State	US State names	text		
CLSCHOOL	Date closed K-12 public schools	The date a state closed K-12 public schools statewide. Only included directives/orders. Did not include guidance or recommendations. Order must apply to entire state	date	0: policy not implemented	
END_BSNS	Began to reopen businesses	The date a state began to reopen businesses that were previously closed due to COVID-19 statewide. Order must apply to entire state.	date	0: policy not implemented	
FM_ALL	Mandate face mask use by all individuals in public spaces	The date a state mandated face mask use in public spaces by all individuals statewide. The order does not have to apply to all public spaces, but must apply statewide. Only included directives/orders. Did not include guidance or recommendations. Order must apply to entire state	date	0: policy not implemented	
QR_END	Date all mandated quarantines ended	The date a state ended all mandated quarantines for individuals arriving from out of state. If any statewide quarantines for out of state individuals is still in effect or if the state never had a quarantine	date	0: policy not implemented	

		in effect the column will bear a 0.			
EMSTART	Overall eviction moratorium start	The earliest date a state prohibited some aspect of the court process of eviction, ejectment, forcible entry and detainer, FED. This may include suspending notices of eviction to tenants, suspending filing of eviction claims, suspending hearings on eviction, entering judgments or issuing writs of eviction, or suspending enforcement of new order of eviction. Although different states give this legal process different names, all states allow for landlords to use the court system and law enforcement personnel to remove tenants.	date	0: policy not implemented	
EMEND	Overall eviction moratorium expiration	The latest date a state lifted an order that prohibited all aspects of the court process of eviction, ejectment, forcible entry and detainer, FED	date	0: policy not implemented	
EMSTART2	Second overall eviction moratorium start	The earliest date a state prohibited some aspect of the court process of eviction, ejectment, forcible entry and detainer, FED for the second point in time during the COVID-19 pandemic	date	0: policy not implemented	

EMEND2	Second overall eviction moratorium end	The latest date a state lifted an order that prohibited all aspects of the court process of eviction, ejectment, forcible entry and detainer, FED for the second point in time during the COVID-19 pandemic	date	0: policy not implemented	
EMSTART3	Third overall eviction moratorium start	The earliest date a state prohibited some aspect of the court process of eviction, ejectment, forcible entry and detainer, FED for the third point in time during the COVID-19 pandemic	date	0: policy not implemented	
EMEND3	Third overall eviction moratorium end	The latest date a state lifted an order that prohibited all aspects of the court process of eviction, ejectment, forcible entry and detainer, FED for the third point in time during the COVID-19 pandemic	date	0: policy not implemented	
SMSTART	Utilities shutoff moratorium start	The date a state prohibited utility companies from disconnecting tenants from utilities	date	0: policy not implemented	
SMEND	Utilities shutoff moratorium expiration	The date a state lifted an order that prohibited utility companies from disconnecting tenants from utilities	date	0: policy not implemented	
SMSTART2	Second utilities shutoff moratorium start	The date a state prohibited utility companies from disconnecting tenants from utilities for the second point in time during the COVID-19 pandemic	date	0: policy not implemented	

WV_WTPRD	Waived one week waiting period for unemployment insurance	The date a state waived the one week waiting period for unemployment insurance benefits. Did not include guidance or recommendations. Order must apply to entire state.	date	0: policy not implemented; 1: Could not locate the date the policy started but did locate evidence that the policy was in place.	
REI_WTPRD	Reinstated one week waiting period for unemployment insurance	The date a state reinstated the one week waiting period for unemployment insurance benefits. Did not include guidance or recommendations. Order must apply to entire state.	date	0: policy not implemented; 1: Could not locate the date the policy started but did locate evidence that the policy was in place.	
WV_WKSR	Waive work search requirement for unemployment insurance	If a state waived the work search requirement for unemployment insurance benefits at some point during the pandemic. Did not include guidance or recommendations. Order must apply to entire state.	flag	1, 0	1="YES"; 0="NO"
REI_WKSR	Reinstated work search requirement for unemployment insurance	If a state reinstated the work search requirement for unemployment benefits at some point during the pandemic. Did not include guidance or recommendations. Order must apply to the entire state.	flag	0, 1, 2	0 = never waived work search requirement; 1 = work search requirement remains waived; 2 = reinstated work search requirement
UIQUAR	Expand eligibility of unemployment insurance to anyone who is quarantined and/or taking care of someone who is quarantined	If a state expanded eligibility of unemployment insurance to anyone who is quarantined and/or taking care of someone who is quarantined during the pandemic. Did not include guidance or recommendations.	flag	1, 0	1="YES"; 0="NO"

		Order must apply to entire state.			
UIHIRISK	Expand eligibility to high-risk individuals in preventative quarantine	If a state explicitly defines individuals at high risk of COVID-19 who are undergoing preventive quarantine during the pandemic to be eligible for unemployment insurance. Did not include guidance or recommendations. Order must apply to entire state	flag	1, 0	1="YES"; 0="NO"
UICLDCR	Expand eligibility of unemployment insurance to those who have lost childcare/school closures	If a state expanded eligibility of unemployment insurance to those who have lost childcare during the pandemic in response to daycare/school closures. Did not include guidance or recommendations. Order must apply to entire state.	flag	1, 0	1="YES"; 0="NO"
UIEXTND	Extend the amount of time an individual can be on unemployment insurance	If a state extended the amount of time an individual can be on unemployment insurance. Did not include guidance or recommendations. Order must apply to entire state.	flag	1, 0	1="YES"; 0="NO"
EBSTART	Extended Benefits program activated	The date of issue of Department of Labor Trigger Notice that first showed the state's Extended Benefits (EB) program to have activated. If a state never activated EB, it will be marked as 0.	date	0: policy not implemented	

EBEND	Extended Benefits program deactivated	The date of issue of Department of Labor Trigger Notice that first showed the state's Extended Benefits (EB) program to have deactivated. If a state never activated EB or activated EB but has yet to deactivate EB, it will be marked as 0.	date	0: policy not implemented;	

Note. Adapted from Raifman (2021).

The remainder of features were removed, as they are either redundant or it is believed they are not relevant to the predictive modeling process. A full list has not been included due to the number of excluded features, which is almost 200.

Insertion of data

Additional features that provide data on governmental policies surrounding efforts to address the COVID-19 pandemic (e.g., reopening, stimulus, etc.) were obtained from a source external to LP. Some of the data will need to be processed as there are a lot of dates specific to when policies or mandates were enacted.

Data Description Report

Volumetric Analysis of Data

The three LP datasets (.csv files) were sent to the DST from the Research Team. Additionally, Mr. Carr and his staff downloaded the fourth dataset (Excel file) via the openICPSR website (Raifman). The four files were then imported into Python 3.7 using a function defined to accept a file input and generate specific initial descriptive details. These details include:

- Number of data frame rows (N):
    - Engagement (all files combined) = 22,324,190;
    - District Information = 233;
    - Product Information = 372;
    - COVID-19 US State Policy DB = 53.
- Count of null values for each feature.
- List of Python datatype for each feature.
- Descriptive statistics (using the Python .describe method), which by default are displayed for all numerical features.

See Appendix C.1 for additional information and a display of the first 10 rows of each dataset.

Besides the COVID-19 US State Policy DB that initially had 222 features (which were trimmed for analysis purposes), no other table shows a high level of complexity. The features that will require the most amount of processing and consideration are the date/time features.

Attribute Types and Values

The majority of features that will be using modeling are categorical in nature. The main exceptions are pct_access and engagement_index from the engagement data table. There are some features that were originally continuous in nature, but were generalized to a range in order to anonymize the data. This transformation necessitates their use as categorical features, which along with the others will need to be discretized in order to perform prediction modeling. The features that link the tables (see Keys section) are being used consistently across each table, so there is no anticipated conflict in terms of value match.

Tables 4-7 are the Data Quality Reports (DQRs) for the COVID-19 State Policy DB, District Info, Engagement, and Product Info data tables, respectively. Analysis of the details will be outlined in the Data Exploration Report section. For every feature, the DQRs include a column for the feature name (Feature), data type (Type); total number of records (N); number of records with null values (Miss.); number of records without null values (n); percent of records with a null value (% Miss.); cardinality, or number of unique values (Card.); and mode. For features with numerical values, there are five additional columns: minimum value (Min.); average value (Mean); median value (Median); maximum value (Max.); and standard deviation (Stan. Dev.).

Table 4

Data Quality Report (DQR)

COVID-19 State Policy DB Table

Feature	Type	N	Miss.	n	% Miss.	Card.	Mode	Min.	Mean	Median	Max.	Stan. Dev.
REI_WKSR	continuous	53	2	51	3.8	4	1	0.00	1.06	1.00	2.00	0.54
UICLDCR	continuous	53	2	51	3.8	3	0	0.00	0.37	0.00	1.00	0.48
UIEXTND	continuous	53	2	51	3.8	3	0	0.00	0.06	0.00	1.00	0.24
UIHIRISK	continuous	53	2	51	3.8	3	0	0.00	0.22	0.00	1.00	0.41
UIQUAR	continuous	53	2	51	3.8	3	1	0.00	0.82	1.00	1.00	0.38
WV_WKSR	continuous	53	2	51	3.8	3	1	0.00	0.88	1.00	1.00	0.32
CLSCHOOL	date/time	53	2	51	3.8	12	2020-03-16	na	na	na	na	na
EBEND	date/time	53	2	51	3.8	22	0	na	na	na	na	na
EBSTART	date/time	53	2	51	3.8	12	2020-05-03	na	na	na	na	na
EMEND	date/time	53	2	51	3.8	34	0	na	na	na	na	na
EMEND2	date/time	53	2	51	3.8	5	0	na	na	na	na	na
EMEND3	date/time	53	1	52	1.9	4	0	na	na	na	na	na
EMSTART	date/time	53	2	51	3.8	20	2020-03-16	na	na	na	na	na
EMSTART2	date/time	53	2	51	3.8	8	0	na	na	na	na	na
EMSTART3	date/time	53	2	51	3.8	3	0	na	na	na	na	na
END_BSNS	date/time	53	2	51	3.8	21	2020-05-01	na	na	na	na	na
FM_ALL	date/time	53	2	51	3.8	31	0	na	na	na	na	na
QR_END	date/time	53	2	51	3.8	17	0	na	na	na	na	na
REI_WTPRD	date/time	53	1	52	1.9	7	0	na	na	na	na	na
SMEND	date/time	53	2	51	3.8	29	0	na	na	na	na	na
SMSTART	date/time	53	2	51	3.8	27	0	na	na	na	na	na
SMSTART2	date/time	53	2	51	3.8	3	0	na	na	na	na	na
STATE	key / string	53	2	51	3.8	52	0	na	na	na	na	na
WV_WTPRD	date/time	53	2	51	3.8	22	2020-03-17	na	na	na	na	na

Table 5

Data Quality Report (DQR)

District Info Table

Feature	Type	N	Miss.	n	% Miss.	Card.	Mode	Min.	Mean	Median	Max.	Stan. Dev.
county_connections_ratio	binned	233	71	162	30.5	3	[0.18, 1[	na	na	na	na	na
district_id	key / cotinuous	233	0	233	0.0	233	8,815	1,000.00	5,219.78	4,937.00	9,927.00	2,590.18
locale	string	233	57	176	24.5	5	Suburb	na	na	na	na	na
pct_black/hispan	binned	233	57	176	24.5	6	[0, 0.2[	na	[0.14, 0.34[	na	na	na
pct_free/reduced	binned	233	85	148	36.5	6	0	na	[0.2, 0.37[	na	na	na
pp_total_raw	binned	233	115	118	49.4	12	0	na	[7693.18, 9034.09[	na	na	na
state	string	233	57	176	24.5	24	0	na	na	na	na	na

Note . * Mean for binned features calculated based on method outlined in Table 9

Table 6

Data Quality Report (DQR)

Engagement Table

Feature	Type	N	Miss.	n	% Miss.	Card.	Mode	Min.	Mean	Median	Max.	Stan. Dev.
engagement_index	continuous	500,000	120,205	379,795	24.0	38,315	0	0.01	169.60	1.90	125,969.17	1,693.94
pct_access	continuous	500,000	307	499,693	0.1	3,906	0	0.00	0.50	0.02	89.19	3.18
district_id	continuous	500,000	0	500,000	0.0	233	2,956	1,000.00	5,236.48	4,929.00	9,927.00	2,643.99
lp_id	continuous	500,000	0	500,000	0.0	6,294	95,731	10,003.00	54,689.93	54,827.00	99,984.00	26,475.10
time	date	500,000	0	500,000	0.0	366	2020-11-19	2020-01-01	na	na	2020-12-31	na

Note . * Sample chosen at random from *N* = 22,324,190

Table 7

Data Quality Report (DQR)

Product Info Table

Feature	Type	N	Miss.	n	% Miss.	Card.	Mode	Min.	Mean	Median	Max.	Stan. Dev.
LP ID	key / continous	372	0	372	0.0	372	13,117	10,533.00	54,565.80	53,942.50	99,916.00	26,212.25
Primary Essential Function	string	372	20	352	5.4	36	LC - Digital Learning Platforms	na	na	na	na	na
Product Name	string	372	0	372	0.0	372	SplashLearn	na	na	na	na	na
Sector(s)	string	372	20	352	5.4	6	PreK-12	na	na	na	na	na

Keys

Each of the four tables has at least one feature in common with another table. The table relationships are:

- The engagement data table includes: A product identification number (lp_id) that is a foreign key for the product information data table LP ID feature and a district identification number (derived feature) that is a foreign key for the district information data table district_id feature.
- The district information data table includes: A state feature (state) that is a foreign key for the COVID-19 US state policy database data table STATE feature.

Review Assumptions/Goals

No new assumptions have been determined at this stage of the project.

Data Exploration Report

Data Exploration

The DQRs (Tables 4-7) include the descriptive statistical information for the four main data tables. Each table was examined separately in order to correctly derive the measures of centrality (e.g., mean, mode) and dispersion (standard deviation). To minimize processing time and computing resources, a random sample (n = 500,000) was taken from the engagement data table due to its large size (N = 22,324,190).

During review of the DQRs and several generated visualizations, some items of interest and further exploration were noted:

COVID-19 State Policy DB Table

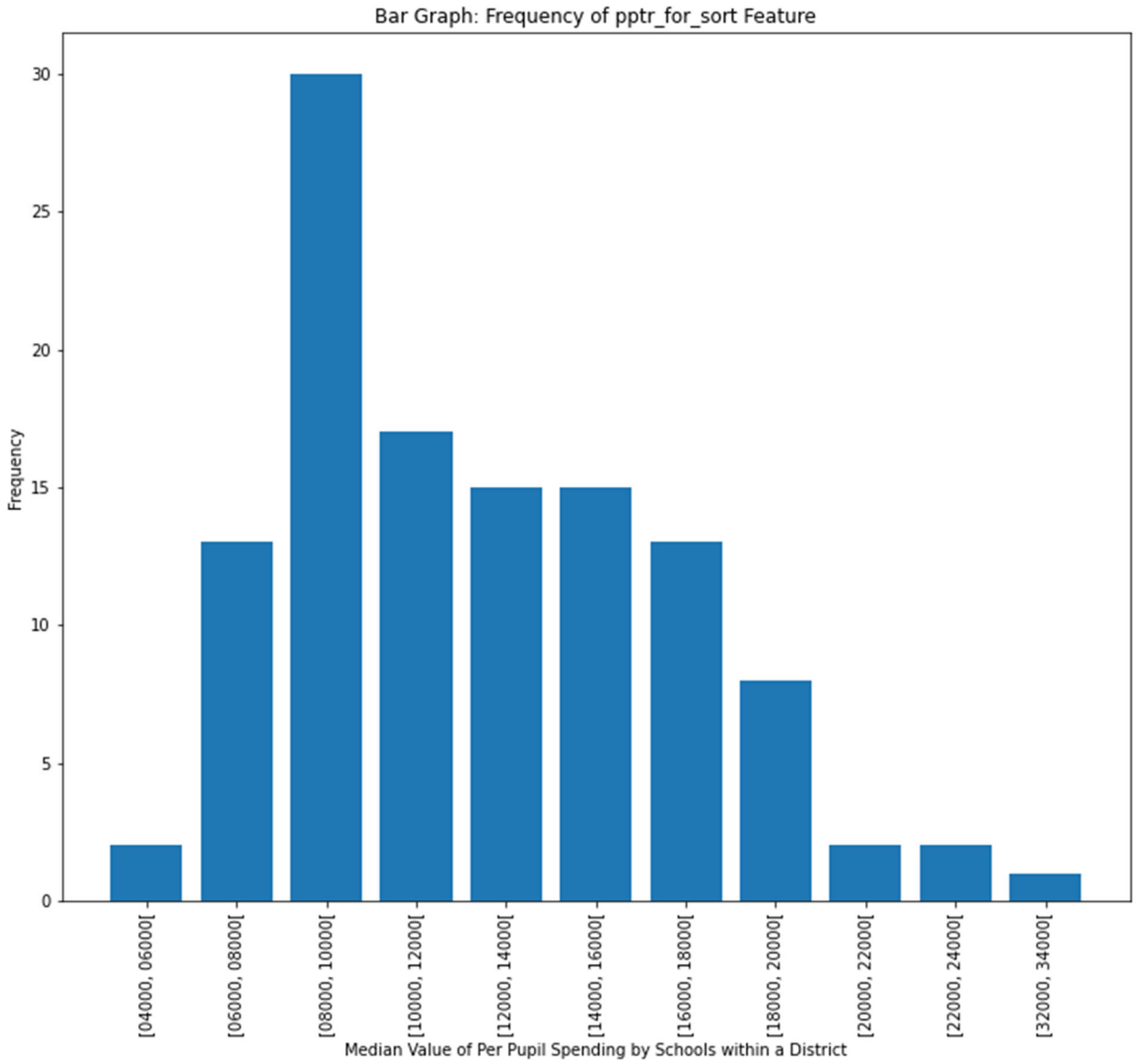
The majority of features in this table are categorical in nature. Additionally, the six that are continuous are actually flag features, with values assigned in accordance with Table 3. How these should be treated is outlined in the Data Quality Plan (DQP; Table 8). The highest count of null values for any of the features is low (n = 2), which is fortunate because the overall size of the dataset is small (N = 53). As a result, the percentage of missing values for each feature is either 1.9% or 3.8%.

District Info Table

This table contains mostly categorical features, but unlike the COVID-19 State Policy DB data table, four of the categorical features are binned transformations of continuous values. As noted in the Initial Data Collection Report section, this was done by LP's Research Team in order to anonymize the data as much as possible. The DST was still able to calculate a mean based on the method outlined in Stack Exchange (2016), and can be seen in Appendix B, Table 9.

For plotting purposes, the pp_total_raw feature was transformed from a number to a string for values under 10,000 in order to make sorting consistent (the derived feature is pptr_for_sort). A bar graph was then generated (Figure 2) to show the frequency of each category. Since pp_total_raw is a binned continuous feature, the result actually fairly represents a distribution of the values, which can be seen to be unimodal, but also highly skewed to the right (positive)—the mean is [7693.2, 9034.1[.

Figure 2



A potential issue for most of the features is the high percentage of missing values—the only exception is district_id (0.0%). Some are as high as 49.4%. They will need to be addressed, possibly using one or more of the following methods: 1) Discuss with LP to see if additional data is available either internally or externally; 2) Fill in the missing values using transformation methods, such central tendency (e.g., mean) substitution; 3) Where steps #2-3 do not derive an appropriate stand-in value, remove the feature if possible, or remove only the instances with nulls. Additional bar graphs of categorical features are in Appendix C.2.

Engagement Table

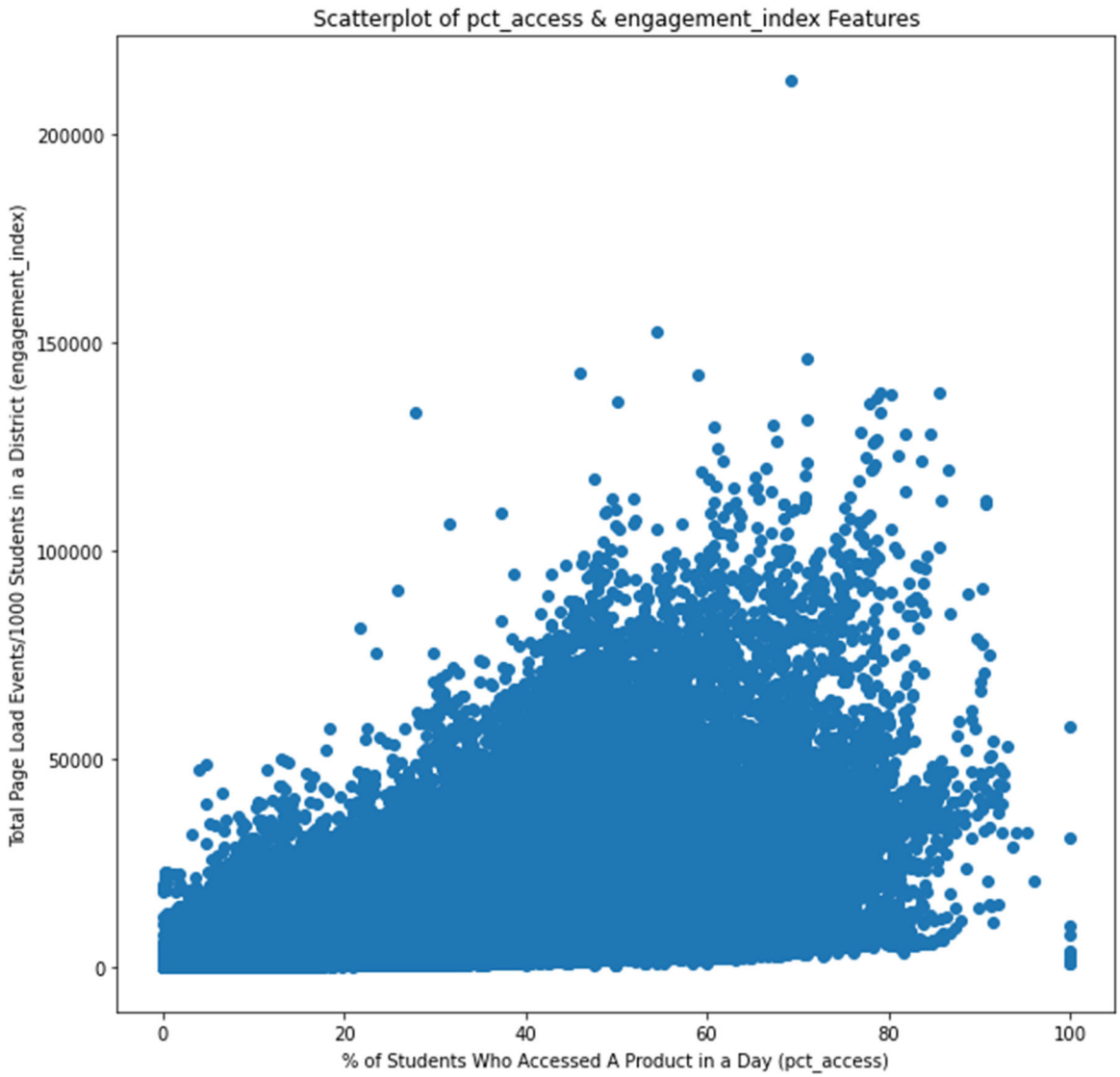
This table mostly contains features with numerical values, though only two are strictly usable for statistical analyses. The other two are district_id and lp_id, both of which are identification numbers and therefore should be treated as categorical (since a specific identification number is more like a category than a continuous value). As previously noted, it was necessary to pull a random sample as the Python pandas package did not have the

capability to run the full description process using the defined method. Additional steps will need to be taken to specifically examine the descriptive statistics of all 22,324,190 instances.

When looking at the two truly continuous features (engagement_index and pct_access), it was interesting to note that they both had mean values (168.29 and 0.50, respectively) that were significantly larger than the median values (1.92 and 0.02). This indicates that there is a very large positive skew for both features, with some instances containing very large outlier values. Further evidence for this is provided by comparing the mean to the minimum and maximum values of each: For engagement_index, the mean is 168.29, the minimum is 0.01, and the maximum is 98,710.37; for pct_access the mean is 0.50, the minimum is 0.00, and the maximum is 100.00. It can clearly be seen that the means are closer to the minimum values than the maximum ones.

The correlation coefficient was calculated for engagement_index and pct_access and it indicated a medium to strong linear relationship between the two features (r = .7515). However, when a scatterplot was created (Figure 3) the relationship between them appears less clear, most likely highly influenced by outliers for both features (see Appendix C.2 for boxplots of both).

Figure 3



One potential issue is the high number of missing or null values for the engagement_index (24.2%). The methods discussed above will be used to address them.

Product Info Table

All of the features in this table are categorical. The percentage of missing values is relatively low for each feature. Of note, the Primary Essential feature has a cardinality of 36 (out of 372 instances) and the mode is "LC – Digital Learning Platforms".

Form Suppositions for Future Analysis

After reviewing the DQRs and performing the initial review and analysis during data exploration, the project group formulated a hypothesis that there is a measurable relationship between percentage access (pct_access), per pupil spending (pp_total_raw), and percent of people who are Black and/or Hispanic living in a district (pct_black/hispanic). Additionally, there is a strong negative correlation between the percent of Black and Hispanic people in a district and per pupil spending, such that as pct_black/hispanic values increase pp_total_raw values decrease.

Through the course of this project, the stated hypothesis will be converted into a data mining goal through the process of determining a linear regression model, whereby pct_access is the target feature of the predictive model, and the other features are the descriptive (independent) features. The eventual model will most likely be a first-order model, either with or without interactions, but the possibility of second-order full quadradic models given the number of categorical features has not been ruled out.

Data Quality Report

Review Keys and Attributes, Data Quality in Flat Files, and Noise and Inconsistencies Between Sources

Data stored in flat files are in both .csv and Excel formats. The former has been checked for consistency of delimiters and it has been confirmed they are structurally sound (i.e., they do not have other delimiters that might cause errors).

There is potentially a great deal of noise for some of the key features. For instance, engagement_index and pct_access both have a significant number of outliers that are influencing the statistical measures. In order to outline the potential pitfalls, as well as the anticipated methods for handling them, the Data Science and Research Teams have developed the DQP, as seen in Table 8.

Table 8

Data Quality Plan (DQP)

Data Table	Feature	Data Quality Issue	Potential Handling Strategies
COVID-19 State Policy DB Table	CLSCHOOL	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EBEND	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EBSTART	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EMEND	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once

			transformed, this feature will be removed.
COVID-19 State Policy DB Table	EMEND2	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EMEND3	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EMSTART	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.

COVID-19 State Policy DB Table	EMSTART2	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	EMSTART3	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	END_BSNS	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	FM_ALL	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.

COVID-19 State Policy DB Table	QR_END	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	REI_WKSR	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the Data Quality Report (DQR) from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.
COVID-19 State Policy DB Table	REI_WTPRD	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	SMEND	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	SMSTART	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the

			Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	SMSTART2	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
COVID-19 State Policy DB Table	UICLDCR	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the DQR from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.
COVID-19 State Policy DB Table	UIEXTND	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the DQR from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.
COVID-19 State Policy DB Table	UIHIRISK	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the DQR from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.
COVID-19 State Policy DB Table	UIQUAR	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the DQR from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.

COVID-19 State Policy DB Table	WV_WKSR	This feature is designated as continuous even though it is a flag field, for which each value is non-ordinal.	Re-designate type on the DQR from "continuous" to "flag"; split into multiple (c) dummy columns, where the value of each is 0 or 1 based on the value of the original feature.
COVID-19 State Policy DB Table	WV_WTPRD	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be transformed to a categorical feature with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once transformed, this feature will be removed.
District Info Table	county_connections_ratio	1) This feature has a high rate of missing/null values (30.5%); 2) the categories are not clearly defined.	1) Work with LP to identify if additional data is available from internal or external sources to fill in the gaps; 2) work with LP to determine the meaning and significance of the two categories.
District Info Table	district_id	This feature has been miscategorized as continuous due to its nature as a numerical value. However, as ID numbers represent a thing and not a measurement, they should be treated as categorical.	Change type from "key / continuous" to "key / ID".
District Info Table	locale	This feature has high rate of missing/null values (24.5%).	Work with LP to identify if additional data is available from internal or external sources to fill in the gaps.
District Info Table	pct_black/hispanic	This feature has high rate of missing/null values (24.5%).	Work with LP to identify if additional data is available from internal or external sources to fill in the gaps.
District Info Table	pct_free/reduced	This feature has high rate of missing/null values (36.5%).	Work with LP to identify if additional data is available from internal or external sources to fill in the gaps.
District Info Table	pp_total_raw	This feature has high rate of missing/null values (49.4%).	Work with LP to identify if additional data is available from internal or external sources to fill in the gaps.

District Info Table	state	This feature has high rate of missing/null values (24.5%).	Work with LP to identify if additional data is available from internal or external sources to fill in the gaps.
Engagement Table	district_id	This feature has been miscategorized as continuous due to its nature as a numerical value. However, as ID numbers represent a thing and not a measurement, they should be treated as categorical.	Change type from "continuous" to "ID".
Engagement Table	lp_id	This feature has been miscategorized as continuous due to its nature as a numerical value. However, as ID numbers represent a thing and not a measurement, they should be treated as categorical.	Change type from "continuous" to "ID".
Engagement Table	time	This feature is date/time type, which can be particularly difficult to manage and use in a predictive model.	As this is a descriptive feature, and may play prominently in one of the final predictive models, the date will be used to transform the COVID-19 State Policy DB Table date features into categorical features with two values ("before"/"after") which relatively correspond to the time feature from the Engagement Table. Once used for transformed, this feature will be removed.
Product Info Table	LP ID	This feature has been miscategorized as continuous due to its nature as a numerical value. However, as ID numbers represent a thing and not a measurement, they should be treated as categorical.	Change type from "key / continuous" to "key / ID".

References

Census Bureau (2020, August 26). *Schooling during the COVID-19 pandemic*. U.S. Department of Commerce. https://www.census.gov/library/stories/2020/08/schooling-during-the-covid-19-pandemic.html

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Cross-industry standard process for data mining* (1st ed.; CRISP-DM 1.0). SPSS.

GeeksforGeeks. (2022, January 18). *Python | List frequency of elements*. https://www.geeksforgeeks.org/python-list-frequency-of-elements/

Jurney, R. (2017). *Agile data science 2.0: Building full-stack data analytics applications with Spark*. O'Reilly Media.

Kaggle. LearnPlatform COVID-19 Impact on Digital Learning. Analytics Competition. https://www.kaggle.com/c/learnplatform-covid19-impact-on-digital-learning/overview

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (2nd ed.). The MIT Press.

LearnPlatform. (n.d.). https://learnplatform.com/

Microsoft. (2021, November 30). *What is Apache Spark?* https://docs.microsoft.com/en-us/dotnet/spark/what-is-spark

NCES. (n.d.). *Back-to-school statistics*. US Department of Education and the Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/fastfacts/display.asp?id=372

Nyakundi, H. (2021, June 15). *What is full stack? How to become a full stack developer*. FreeCodeCamp. https://www.freecodecamp.org/news/what-is-a-fullstack-developer/

Phibbs, C., & Carr, A. (2021). *ADS-500B-02-FA21 final data science programming project: Bank marketing dataset* [Jupyter Notebook].

Raifman, J., Nocka, K., Jones, D., Bor, J., Lipson, S., Jay, J., Cole, M., Krawczyk, N., Benfer, E. A., Chan, P., & Galea, S. (2021). *COVID-19 US State Policy Database* [Unpublished raw data]. Retrieved January 24, 2022, from https://www.openicpsr.org/openicpsr/project/119446/version/V75/view?path=/openicpsr/119446/fcr:versions/V75&type=project

Stack Exchange. (2016). *How to calculate average of something based on a range of another thing?* https://math.stackexchange.com/questions/1983462/how-to-calculate-average-of-something-based-on-a-range-of-another-thing

Stack Overflow. (2015). *Extracting text from elements in pandas column, writing to new column*. https://stackoverflow.com/questions/33408403/extracting-text-from-elements-in-pandas-column-writing-to-new-column

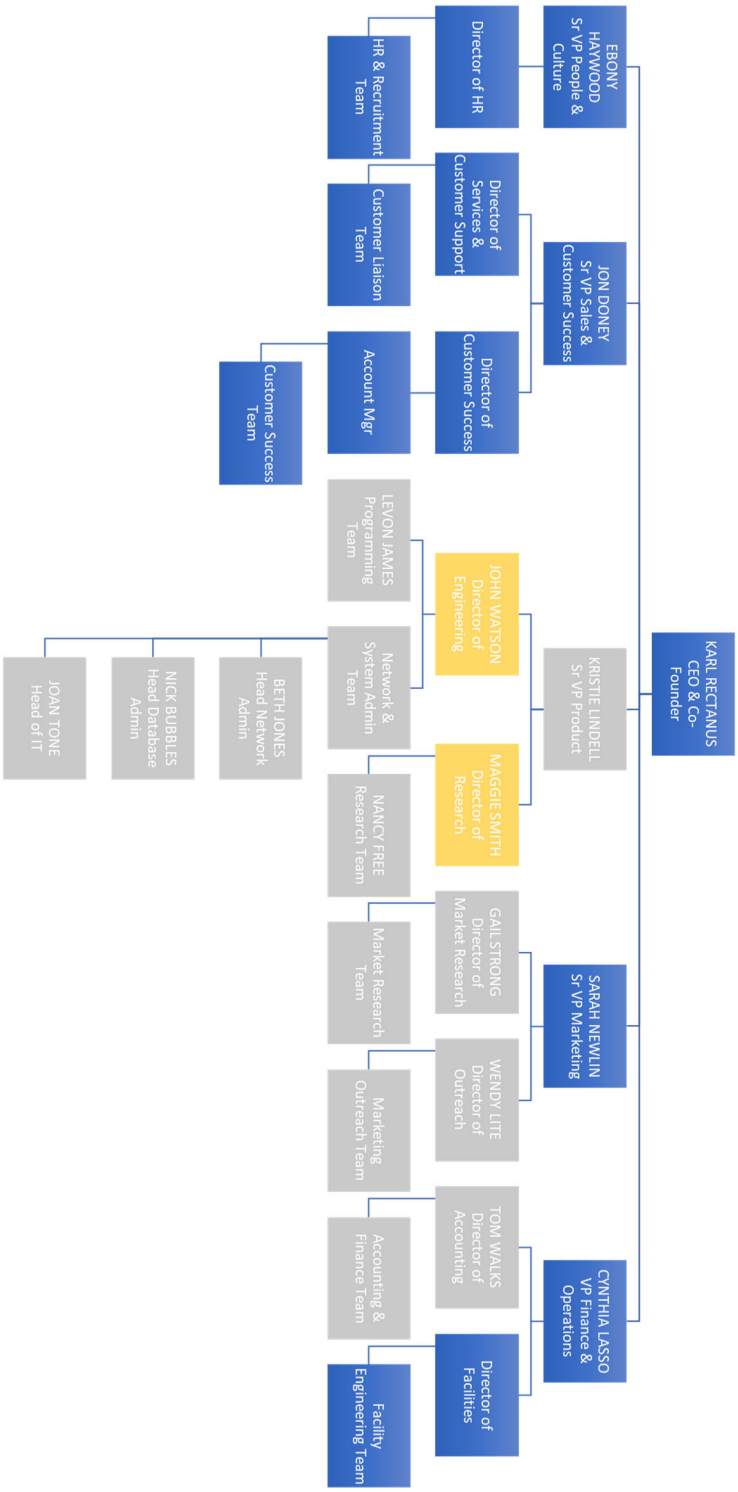
Vakkuri, V., & Kemell, K.-K. (2019). Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy. In S. Hyrynsalmi, M. Suoranta, A. Nguyen-Duc, P. Tyrväinen, & P. Abrahamsson (Eds.), *Software Business* (Vol. 370, pp. 260–275). Springer International Publishing. https://doi.org/10.1007/978-3-030-33742-1_21

Wrike. (n.d.). *Project management guide: Project lifecycle*. https://www.wrike.com/project-management-guide/project-lifecycle/

Appendix A

LearnPlatform's organizational chart

Figure 4



Appendix B

Selected Tables

Table 9

Examples of Average Calculation Method for Binned Variables (Stack Exchange, 2019)

pct_black/hispanic

Pandas Index	Count	Lower Bin Bound (LBB)	Upper Bin Bound (UBB)	Bin Range	Lower Bin Weight (Count * LBB)	Lower Bin Weight (Count * UBB)
0*	116	0.0	0.2	0.2	0.00	23.20
1*	24	0.2	0.4	0.2	4.80	9.60
2	17	0.4	0.6	0.2	6.80	10.20
3	11	0.6	0.8	0.2	6.60	8.80
4	8	0.8	1.0	0.2	6.40	8.00
Sums	176				24.60	59.80
				Mean Bin Bounds	**0.14**	**0.34**
				Mean Bin	**[0.14, 0.34[**	

pct_free/reduced

Pandas Index	Count	Lower Bin Bound (LBB)	Upper Bin Bound (UBB)	Bin Range	Lower Bin Weight (Count * LBB)	Lower Bin Weight (Count * UBB)
0	46	0.0	0.2	0.2	0.00	9.20
1*	48	0.2	0.4	0.2	9.60	19.20
2	37	0.4	0.6	0.2	14.80	22.20
3	13	0.6	0.8	0.2	7.80	10.40
4	4	0.8	1.0	0.2	3.20	4.00
Sums	148				35.40	65.00
				Mean Bin Bounds	**0.20**	**0.37**
				Mean Bin	**[0.2, 0.37[**	

pp_total_raw

Pandas Index	Count	Lower Bin Bound (LBB)	Upper Bin Bound (UBB)	Bin Range	Lower Bin Weight (Count * LBB)	Lower Bin Weight (Count * UBB)
8	2	4,000	6,000	2,000	8,000.0	12,000.0
9*	13	6,000	8,000	2,000	78,000.0	104,000.0
10*	30	8,000	10,000	2,000	240,000.0	300,000.0
0	17	10,000	12,000	2,000	170,000.0	204,000.0
1	15	12,000	14,000	2,000	180,000.0	210,000.0
2	15	14,000	16,000	2,000	210,000.0	240,000.0
3	13	16,000	18,000	2,000	208,000.0	234,000.0
4	8	18,000	20,000	2,000	144,000.0	160,000.0
5	2	20,000	22,000	2,000	40,000.0	44,000.0
6	2	22,000	24,000	2,000	44,000.0	48,000.0
7	1	32,000	34,000	2,000	32,000.0	34,000.0
Sums	118				1,354,000.0	1,590,000.0
				Mean Bin Bounds	**7,693.2**	**9,034.1**
				Mean Bin	**[7693.2, 9034.1[**	

* Bins that overlap with the mean bin range

Table 10

Gantt Chart: Project Phase 1 (Weeks 1-3)

Task	Duration (In Days)	Week 1					Week 2					Week 3				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Phase 1: Create a dashboard with student digital learning access data, district demographical data, product data, and COVID-19 state policy data	15															
Business Understanding																
Identify organizational structure, resources, and key LP project stakeholders and contacts	5															
Data Understanding																
Setup initial execute, transform, load (ETL) pipeline:	5															
Review available data tables from LP internal databases, factoring in access permissions	2															
Identify external data sources	2															
Develop plan with Research Team to establish standard procedure for Data Science Team (DST) to access data, i.e., outline: process steps, initial transformation/anonymization, data refresh/update schedule, retrieval location	3															
Create queries to link the 4 data tables to a dashboard	2															
Identify features of interest	2															
Describe features and data from each of the available sources	5															
Create visualizations for each of the individual data tables. These include: Tabular summaries of central tendency and dispersion measures; histograms and boxplots for single key numeric features; scatterplot matrices for comparing two numerical features; bar graphs for categorical feature frequency; etc.	3															
Generate initial Data Quality Reports (DQRs)	3															
Create feature relationship mapping between tables	2															
Test dashboard	3															
Deploy dashboard into production and train business users, including specifically the Research, Marketing Research, Marketing Outreach, and Programming Teams	3															

Table 11

Gantt Chart: Project Phase 2, Business Understanding (Weeks 4-6)

Task	Duration (In Days)	Week 4 16 17 18 19 20	Week 5 21 22 23 24 25	Week 6 26 27 28 29 30
Phase 2: Combine individual data tables to create a first full dataset (analytics base table [ABT]) that can be used to build analytics outputs that can be added to dashboard to predict: 1) Which student populations are at risk for low online engagement; and 2) What types of public policy actions can positively impact online engagement, both overall and for the groups identified in #1.	60			
Business Understanding				
Conduct interviews with Domain Experts to understand nature of the data, such as what specifically what features like engagement really measure. Catalogue business terminology	5			
Define general business problem areas	5			
Specify business objectives for: 1) Descriptive analytics; 2) Predictive analytics	3			
Set evaluation criteria to track business success for each type of analytics objective	3			
Establish project requirements, assumptions, and constraints	3			
Map RESOLVEDD and MART strategies to project goals	3			
Inventory exhaustive list of risks, along with contingency plans, for the five main categories (product value, financial, model health, data integrity, and process issues)	3			
Specify data mining objectives for: 1) Descriptive analytics; 2) Predictive analytics	3			
Determine data mining problem types	3			
Set evaluation criteria to track data mining success for each type of analytics objective	2			
Perform literature review	5			

Table 12

Gantt Chart: Project Phase 2, Data Understanding/Preparation (Weeks 7-10)

Task	Duration (In Days)	Week 7					Week 8					Week 9					Week 10				
		31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Phase 2: Continued	60																				
Data Understanding																					
Evaluate the signals and variables that may predict online engagement. Explore distribution and signals, and perform general exploratory data analysis (EDA):	4																				
Review visualizations, statistical descriptive summaries, and DQRs in-depth	2																				
Review table relationships (e.g., keys) to confirm intimal assumptions, as well as identify issues with matching	2																				
Identify correlations between numerical features	2																				
Reassess features of interest for inclusion in predictive models	1																				
Form hypotheses and create initial general model formula	3																				
Generate Data Quality Plan (DQP)	4																				
Data Preparation																					
Collect additional data, if needed	3																				
Create derived features, including: Discretizing categorical values; aggregating measures of centrality and dispersion for use in the ABT; convert COVID-19 state policy dates to binary values	4																				
Address noise in the data: Impute missing values where appropriate; address outliers; eliminate features and instances not usable for modeling	6																				
Create training and testing datasets from the ABT	3																				

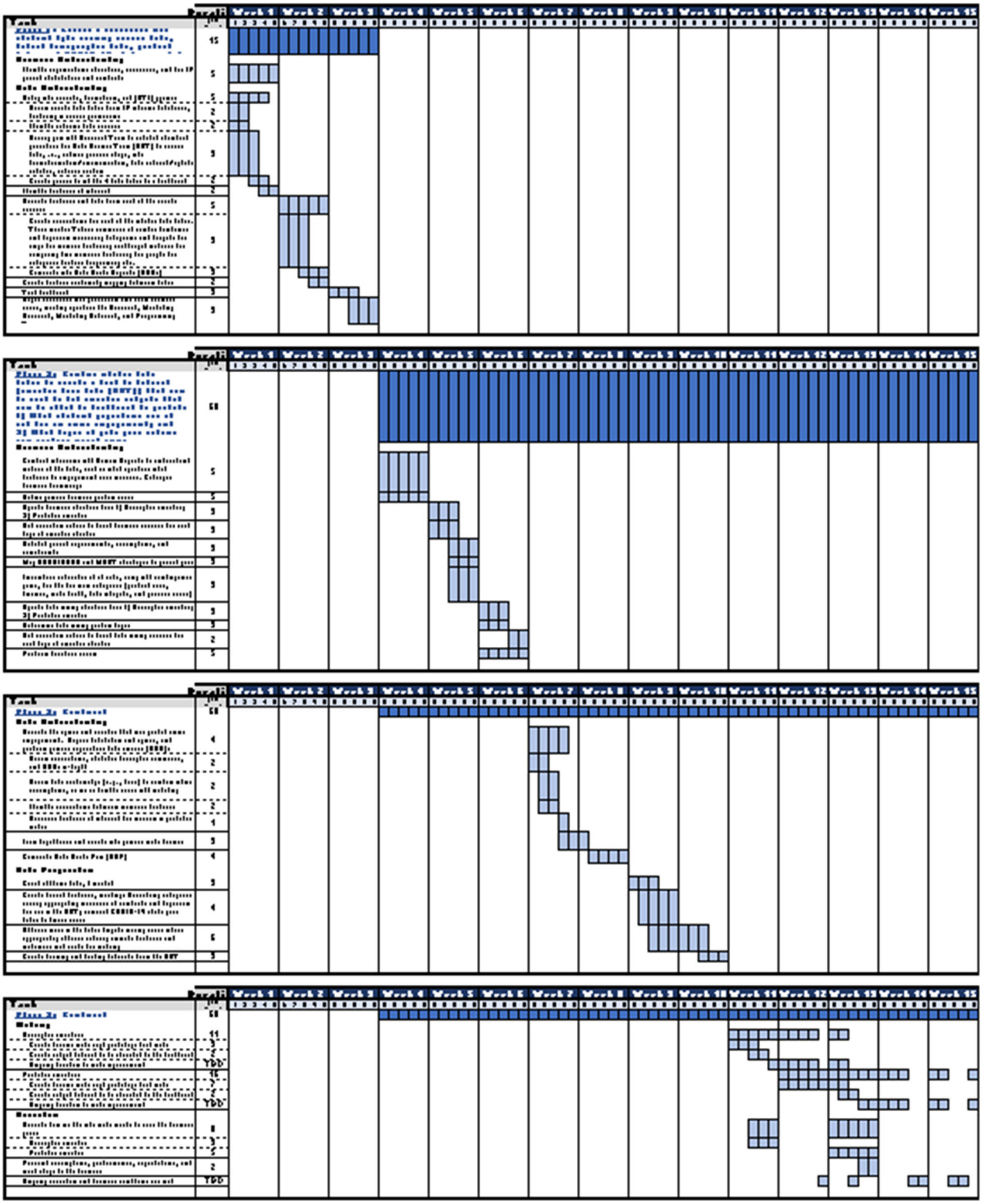
Table 13

Gantt Chart: Project Phase 2, Modeling & Evaluation (Weeks 11-15)

Task	Duration (In Days)	Week 11					Week 12					Week 13					Week 14					Week 15				
		51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Phase 2: Continued	60																									
Modeling																										
Descriptive analytics:	11																									
Create baseline model rapid prototype first model	3																									
Create output dataset to be absorbed by the dashboard	2																									
Ongoing iteration to model improvement	TBD																									
Predictive analytics:	16																									
Create baseline model rapid prototype first model	7																									
Create output dataset to be absorbed by the dashboard	2																									
Ongoing iteration to model improvement	TBD																									
Evaluation																										
Evaluate how well the initial model works to solve the business goals:	8																									
Descriptive analytics	3																									
Predictive analytics	5																									
Present assumptions, performance, expectations, and next steps to the business	2																									
Ongoing evaluation until business conditions are met	TBD																									

Table 14

Gantt Chart: Project Phases 1-2 (Weeks 1-15)



ADS501_Module_6_v1

February 19, 2022

Appendix C

1. Initial Review of Datasets

Introduction Dataset from kaggle.com for a competition hosted by LearnPlatform

[1]:
```
# Import_master_modules
import numpy as np
import pandas as pd
import statistics
import matplotlib as mpl
import matplotlib.pyplot as plt
import os
import os.path as op
import datetime as dt
import copy
from textwrap import wrap
```

[2]:
```
# Set start time to determine time length of script
start_time = dt.datetime.today()
print(f'Start Time = {start_time}')
```

```
Start Time = 2022-02-19 13:52:22.129547
```

Define functions to convert external data to pandas df

[3]:
```
# Define function to accept input of csv file and output df sample
def ld_csv(file=None, head=0, sep=','):
    df = pd.read_csv(file, sep=sep, header=head)
    print(f'\nFirst 10 rows:\n{df.head(10)}') # display first 10 rows of df
    df_len = len(df)
    print(f'\nNumber of df rows = {df_len}') # display df length
    return df
```

[4]:
```
# Define function to accept input of Excel file and output df sample
def ld_xl(file=None, head=0):
    df = pd.read_excel(file, header=head)
    print(f'\nFirst 10 rows:\n{df.head(10)}') # display first 10 rows of df
    df_len = len(df)
```

```
    print(f'\nNumber of df rows = {df_len}') # display df length
    return df
```

[5]:
```
# Define function to accept input of multiple csv files and output df sample
def comb_csv(files=[], head=0, sep=','):
    df = pd.DataFrame()
    for f in files:
        df_sub01 = pd.read_csv(f, sep=sep, header=head)
        df_sub01['district_id'] = f[:4]
        df = pd.concat([df, df_sub01])
    print(f'\nFirst 10 rows:\n{df.head(10)}') # display first 10 rows of df
    df_len = len(df)
    print(f'\nNumber of df rows = {df_len}') # display df length
    return df
```

[6]:
```
# Check current folder path
%pwd
```

[6]:
```
'C:\\Users\\acarr\\Sync\\Programming\\Python 3\\ADS Markdown Files\\ADS-501
Markdown\\Module 1'
```

[7]:
```
# Change directory to main kaggle datta files location
%cd C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501␣
 ↪Markdown\learnplatform-covid19-impact-on-digital-learning
```

```
C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501
Markdown\learnplatform-covid19-impact-on-digital-learning
```

[8]:
```
# Derive features for sorting of values for graph presentation
district_df01 = ld_csv('districts_info.csv')
district_df01['pptr_lb'] = district_df01['pp_total_raw'].str.
 ↪extract(r"\[([A-Za-z0.0-9.9 _]+)\,")
district_df01['pptr_ub'] = district_df01['pp_total_raw'].str.extract(r"\,␣
 ↪([A-Za-z0.0-9.9 _]+)\[")
district_df01.loc[(district_df01['pptr_lb'].astype(float) < 10000), 'pptr_lb']␣
 ↪= '0' + district_df01['pptr_lb']
district_df01.loc[(district_df01['pptr_ub'].astype(float) < 10000), 'pptr_ub']␣
 ↪= '0' + district_df01['pptr_ub']
district_df01.loc[:, 'pptr_for_sort'] = '[' + district_df01['pptr_lb'] + ', ' +␣
 ↪district_df01['pptr_ub'] + '['
```

```
First 10 rows:
   district_id          state  locale pct_black/hispanic pct_free/reduced  \
0         8815       Illinois  Suburb           [0, 0.2[         [0, 0.2[
1         2685            NaN     NaN                NaN              NaN
2         4921           Utah  Suburb           [0, 0.2[       [0.2, 0.4[
3         3188            NaN     NaN                NaN              NaN
```

```
4         2238             NaN     NaN                NaN              NaN
5         5987       Wisconsin  Suburb          [0, 0.2[         [0, 0.2[
6         3710            Utah  Suburb          [0, 0.2[       [0.4, 0.6[
7         7177  North Carolina  Suburb        [0.2, 0.4[       [0.2, 0.4[
8         9812            Utah  Suburb          [0, 0.2[       [0.2, 0.4[
9         6584  North Carolina   Rural        [0.4, 0.6[       [0.6, 0.8[

  county_connections_ratio    pp_total_raw
0                [0.18, 1[  [14000, 16000[
1                      NaN             NaN
2                [0.18, 1[    [6000, 8000[
3                      NaN             NaN
4                      NaN             NaN
5                [0.18, 1[  [10000, 12000[
6                [0.18, 1[    [6000, 8000[
7                [0.18, 1[   [8000, 10000[
8                [0.18, 1[    [6000, 8000[
9                [0.18, 1[   [8000, 10000[

Number of df rows = 233
```

[9]:
```
product_df01 = ld_csv('products_info.csv') # load df from csv file

pd_drop_lst01 = ['URL', 'Provider/Company Name'] # features to remove
product_df02 = product_df01.drop(pd_drop_lst01, axis=1)
```

```
First 10 rows:
   LP ID                            URL    Product Name  \
0  13117     https://www.splashmath.com     SplashLearn
1  66933           https://abcmouse.com    ABCmouse.com
2  50479          https://www.abcya.com          ABCya!
3  92993          http://www.aleks.com/           ALEKS
4  73104   https://www.achieve3000.com/     Achieve3000
5  37600  http://www.activelylearn.com/  Actively Learn
6  18663     http://www.adaptedmind.com     AdaptedMind
7  65131        http://www.amplify.com/         Amplify
8  26491        http://www.answers.com/         Answers
9  56441         http://www.audible.com         Audible

     Provider/Company Name                      Sector(s)  \
0            StudyPad Inc.                        PreK-12
1   Age of Learning, Inc                          PreK-12
2          ABCya.com, LLC                        PreK-12
3     McGraw-Hill PreK-12           PreK-12; Higher Ed
4             Achieve3000                        PreK-12
5          Actively Learn                        PreK-12
6                GloWorld                        PreK-12
```

```
7  Amplify Education, Inc.                          PreK-12
8                  Answers                PreK-12; Higher Ed
9        Amazon.com, Inc.   PreK-12; Higher Ed; Corporate

                          Primary Essential Function
0                   LC - Digital Learning Platforms
1                   LC - Digital Learning Platforms
2  LC - Sites, Resources & Reference - Games & Si…
3                   LC - Digital Learning Platforms
4                   LC - Digital Learning Platforms
5                   LC - Digital Learning Platforms
6                   LC - Digital Learning Platforms
7                       LC - Courseware & Textbooks
8                            LC - Study Tools - Q&A
9  LC - Sites, Resources & Reference - Streaming …

Number of df rows = 372
```

Add additional dataset from an external source

[10]:
```
# Change directory to district data files location
%cd C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501␣
↪Markdown\COVID-19-US-State-Policy-Database-master
```

```
C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501
Markdown\COVID-19-US-State-Policy-Database-master
```

[11]:
```
st_policy_df01 = ld_xl('COVID-19 US state policy database 3_29_2021_der.xlsx')␣
↪# load df from Excel

# List of selected features of interest
st_cols_lst = ['STATE',
               'CLSCHOOL',
               'END_BSNS',
               'FM_ALL',
               'QR_END',
               'EMSTART',
               'EMEND',
               'EMSTART2',
               'EMEND2',
               'EMSTART3',
               'EMEND3',
               'SMSTART',
               'SMEND',
               'SMSTART2',
               'WV_WTPRD',
               'REI_WTPRD',
               'WV_WKSR',
```

```
                'REI_WKSR',
                'UIQUAR',
                'UIHIRISK',
                'UICLDCR',
                'UIEXTND',
                'EBSTART',
                'EBEND']

# Create new df w/ only featuers of interest
st_policy_df02 = st_policy_df01[st_cols_lst]
```

```
First 10 rows:
                  STATE POSTCODE  FIPS    STEMERG           STEMERGEND  \
0               Alabama       AL   1.0 2020-03-13                    0
1                Alaska       AK   2.0 2020-03-11  2021-02-14 00:00:00
2               Arizona       AZ   4.0 2020-03-11                    0
3              Arkansas       AR   5.0 2020-03-11                    0
4            California       CA   6.0 2020-03-04                    0
5              Colorado       CO   8.0 2020-03-11                    0
6           Connecticut       CT   9.0 2020-03-10                    0
7              Delaware       DE  10.0 2020-03-13                    0
8  District of Columbia       DC  11.0 2020-03-11                    0
9               Florida       FL  12.0 2020-03-09                    0

              CLSCHOOL               CLDAYCR             OPNCLDCR  \
0  2020-03-20 00:00:00  2020-03-20 00:00:00  2020-05-23 00:00:00
1  2020-03-16 00:00:00                    0                    0
2  2020-03-16 00:00:00                    0                    0
3  2020-03-17 00:00:00                    0                    0
4  2020-03-23 00:00:00                    0                    0
5  2020-03-23 00:00:00                    0                    0
6  2020-03-17 00:00:00                    0                    0
7  2020-03-16 00:00:00  2020-04-06 00:00:00  2020-06-15 00:00:00
8  2020-03-16 00:00:00                    0                    0
9  2020-03-17 00:00:00                    0                    0

              CLNURSHM             STAYHOME  … MINWAGEMAR2019  \
0  2020-03-19 00:00:00  2020-04-04 00:00:00  …           7.25
1                    0  2020-03-28 00:00:00  …           9.89
2                    0  2020-03-31 00:00:00  …          11.00
3  2020-03-13 00:00:00                    0  …           9.25
4                    0  2020-03-19 00:00:00  …          11.00
5  2020-03-12 00:00:00  2020-03-26 00:00:00  …          11.10
6  2020-03-09 00:00:00                    0  …          10.10
7                    0  2020-03-24 00:00:00  …           8.75
8                    0  2020-04-01 00:00:00  …          13.25
9  2020-03-15 00:00:00  2020-04-03 00:00:00  …           8.46
```

```
  MINWAGEJUL2019 MINWAGEOCT2019 MINWAGEJAN2020 MINWAGEJUL2020  MINWAGESEP2020  \
0           7.25           7.25           7.25           7.25            7.25
1           9.89           9.89          10.19          10.19           10.19
2          11.00          11.00          12.00          12.00           12.00
3           9.25           9.25          10.00          10.00           10.00
4          11.00          11.00          12.00          12.00           12.00
5          11.10          11.10          12.00          12.00           12.00
6          10.10          11.00          11.00          11.00           12.00
7           8.75           9.25           9.25           9.25            9.25
8          14.00          14.00          14.00          15.00           15.00
9           8.46           8.46           8.56           8.56            8.56

  MINWAGEOCT2020 TIPMINWAGE2020  MINWAGE2021  SMALLBUSMINWAGE
0           7.25           2.13         7.25              0.0
1          10.19          10.19        10.34              0.0
2          12.00           9.00        12.15              0.0
3          10.00           2.63        11.00              1.0
4          12.00          12.00        13.00              1.0
5          12.00           8.98        12.32              1.0
6          12.00           6.38        12.00              0.0
7           9.25           2.23         9.25              0.0
8          15.00           5.00        15.00              0.0
9           8.56           5.54         8.65              0.0

[10 rows x 222 columns]

Number of df rows = 53
```

[12]:
```
# Change directory to district data files location
%cd C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501␣
 ↪Markdown\learnplatform-covid19-impact-on-digital-learning\engagement_data
```

```
C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501
Markdown\learnplatform-covid19-impact-on-digital-learning\engagement_data
```

[13]:
```
# List files in current directory
dir_list = os.listdir()
print(dir_list)
```

```
['1000.csv', '1039.csv', '1044.csv', '1052.csv', '1131.csv', '1142.csv',
'1179.csv', '1204.csv', '1270.csv', '1324.csv', '1444.csv', '1450.csv',
'1470.csv', '1536.csv', '1549.csv', '1558.csv', '1570.csv', '1584.csv',
'1624.csv', '1705.csv', '1712.csv', '1742.csv', '1772.csv', '1791.csv',
'1857.csv', '1877.csv', '1904.csv', '1965.csv', '2017.csv', '2060.csv',
'2074.csv', '2106.csv', '2130.csv', '2165.csv', '2167.csv', '2172.csv',
'2201.csv', '2209.csv', '2238.csv', '2257.csv', '2285.csv', '2321.csv',
'2339.csv', '2393.csv', '2439.csv', '2441.csv', '2517.csv', '2549.csv',
'2567.csv', '2598.csv', '2601.csv', '2685.csv', '2729.csv', '2779.csv',
```

```
'2870.csv', '2872.csv', '2940.csv', '2956.csv', '2991.csv', '3080.csv',
'3160.csv', '3188.csv', '3222.csv', '3228.csv', '3248.csv', '3266.csv',
'3301.csv', '3314.csv', '3322.csv', '3371.csv', '3390.csv', '3393.csv',
'3412.csv', '3471.csv', '3550.csv', '3558.csv', '3580.csv', '3640.csv',
'3668.csv', '3670.csv', '3692.csv', '3710.csv', '3732.csv', '3772.csv',
'3864.csv', '3936.csv', '3959.csv', '3986.csv', '4029.csv', '4031.csv',
'4051.csv', '4083.csv', '4165.csv', '4183.csv', '4203.csv', '4314.csv',
'4348.csv', '4373.csv', '4408.csv', '4516.csv', '4520.csv', '4550.csv',
'4569.csv', '4591.csv', '4602.csv', '4629.csv', '4666.csv', '4668.csv',
'4683.csv', '4744.csv', '4749.csv', '4775.csv', '4808.csv', '4921.csv',
'4929.csv', '4936.csv', '4937.csv', '4949.csv', '5006.csv', '5042.csv',
'5057.csv', '5150.csv', '5231.csv', '5257.csv', '5380.csv', '5404.csv',
'5422.csv', '5479.csv', '5510.csv', '5524.csv', '5527.csv', '5600.csv',
'5604.csv', '5627.csv', '5802.csv', '5882.csv', '5890.csv', '5903.csv',
'5934.csv', '5970.csv', '5987.csv', '6046.csv', '6049.csv', '6055.csv',
'6066.csv', '6104.csv', '6131.csv', '6144.csv', '6165.csv', '6194.csv',
'6250.csv', '6345.csv', '6418.csv', '6512.csv', '6577.csv', '6584.csv',
'6640.csv', '6665.csv', '6721.csv', '6762.csv', '6774.csv', '6919.csv',
'6998.csv', '7086.csv', '7164.csv', '7177.csv', '7305.csv', '7308.csv',
'7342.csv', '7352.csv', '7387.csv', '7457.csv', '7541.csv', '7614.csv',
'7660.csv', '7675.csv', '7723.csv', '7741.csv', '7752.csv', '7767.csv',
'7785.csv', '7798.csv', '7829.csv', '7858.csv', '7964.csv', '7970.csv',
'7975.csv', '7980.csv', '8017.csv', '8076.csv', '8103.csv', '8127.csv',
'8160.csv', '8184.csv', '8256.csv', '8328.csv', '8425.csv', '8433.csv',
'8515.csv', '8520.csv', '8539.csv', '8556.csv', '8685.csv', '8702.csv',
'8723.csv', '8748.csv', '8784.csv', '8796.csv', '8815.csv', '8845.csv',
'8884.csv', '8902.csv', '8937.csv', '9007.csv', '9043.csv', '9120.csv',
'9140.csv', '9230.csv', '9303.csv', '9357.csv', '9463.csv', '9478.csv',
'9515.csv', '9536.csv', '9537.csv', '9553.csv', '9589.csv', '9729.csv',
'9778.csv', '9812.csv', '9839.csv', '9899.csv', '9927.csv']
```

[14]:
```
engagement_comb_df01 = comb_csv(dir_list) # Combine multiple csv files & load df
```

```
First 10 rows:
         time    lp_id  pct_access  engagement_index district_id
0  2020-01-01  93690.0        0.00               NaN        1000
1  2020-01-01  17941.0        0.03              0.90        1000
2  2020-01-01  65358.0        0.03              1.20        1000
3  2020-01-01  98265.0        0.57             37.79        1000
4  2020-01-01  59257.0        0.00               NaN        1000
5  2020-01-01  90153.0        0.06              3.90        1000
6  2020-01-01  41587.0        0.00               NaN        1000
7  2020-01-01  29322.0        0.06              5.10        1000
8  2020-01-01  37479.0        0.00               NaN        1000
9  2020-01-01  51340.0        0.09              1.20        1000

Number of df rows = 22324190
```

[15]:
```
# Minor transformation of values to get consistent Python data types
engagement_comb_df02 = engagement_comb_df01[engagement_comb_df01.lp_id.
 ↪notnull()]
engagement_comb_df02['lp_id'] = engagement_comb_df02['lp_id'].astype(int)
engagement_comb_df02['district_id'] = engagement_comb_df02['district_id'].
 ↪astype(int)
```

```
C:\Users\acarr\anaconda3\envs\ds_py37-1\lib\site-
packages\ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  This is separate from the ipykernel package so we can avoid doing imports
until
C:\Users\acarr\anaconda3\envs\ds_py37-1\lib\site-
packages\ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  after removing the cwd from sys.path.
```

2. Exploratory Data Analysis (EDA)

[16]:
```
# Change directory back to Jupyter Notebook location
%cd C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501␣
 ↪Markdown\Module 1
```

```
C:\Users\acarr\Sync\Programming\Python 3\ADS Markdown Files\ADS-501
Markdown\Module 1
```

Define function to output descriptive statistics for a df

[17]:
```
# Define function to accept input of csv file and output data descriptions
file_count = 1
na_flag = 'na'
def df_stats(df, file_count):
    df_sub01 = df.copy()
    df_sub02 = df_sub01.head(10)
    print(f'\nFirst 10 rows:\n{df_sub02}') # display first 10 rows of df
    df_sub01_cols_lst = df_sub01.columns.values.tolist() # review column names
    '''Initialize empty dict & lists for stats df output to Excel'''
    new_df_dict = {}
    n_lst = []
    n_miss_lst = []
```

```
    n_not_miss_lst = []
    perc_miss_lst = []
    card_lst = []
    mode_lst = []
    min_lst = []
    q1_lst = []
    mean_lst = []
    med_lst = []
    q3_lst = []
    max_lst = []
    stdd_lst = []
    for c in df_sub01_cols_lst:
        '''Iterate thru list of col names & generate stats'''
        df_sub01_c_len = len(df_sub01[c])
        df_sub01_c_null_len = len(df[df_sub01[c].isnull()])
        n_lst.append(df_sub01_c_len)
        n_miss_lst.append(df_sub01_c_null_len)
        n_not_miss_lst.append(df_sub01_c_len - df_sub01_c_null_len)
        perc_miss_lst.append(round((df_sub01_c_null_len / df_sub01_c_len) *␣
↪100, 1))
        card_lst.append(len(df_sub01[c].unique()))
        try:
            mode_lst.append(statistics.mode(df_sub01[c]))
        except:
            mode_lst.append(na_flag)
        try:
            min_lst.append(df_sub01[c].min())
        except:
            min_lst.append(na_flag)
        try:
            mean_lst.append(np.mean(df_sub01[c]))
        except:
            mean_lst.append(na_flag)
        try:
            med_lst.append(df_sub01[c].median())
        except:
            med_lst.append(na_flag)
        try:
            max_lst.append(df_sub01[c].max())
        except:
            max_lst.append(na_flag)
        try:
            stdd_lst.append(np.std(df_sub01[c]))
        except:
            stdd_lst.append(na_flag)
    new_df_dict['cols'] = df_sub01_cols_lst
    new_df_dict['n'] = n_lst
```

```
    new_df_dict['n_miss'] = n_miss_lst
    new_df_dict['n_not_miss'] = n_not_miss_lst
    new_df_dict['perc_miss'] = perc_miss_lst
    new_df_dict['card'] = card_lst
    new_df_dict['mode'] = mode_lst
    new_df_dict['min'] = min_lst
    new_df_dict['mean'] = mean_lst
    new_df_dict['median'] = med_lst
    new_df_dict['max'] = max_lst
    new_df_dict['stand_dev'] = stdd_lst
    #print(f'\nNew DF Dictionary:\n{new_df_dict}')
    df_sub01_len = len(df_sub01)
    df_sub03 = df_sub01[df_sub01.isnull().any(axis=1)]
    df_sub03_len = len(df_sub03)
    print(f'\nNumber of df rows = {df_sub01_len}') # display df length
    df_sub04 = df_sub01.isnull().sum()
    print(f'\nNull count per variable:\n{df_sub04}') # review dataset for cols
↪w/ null value
    data_type = df_sub01.dtypes
    print(f'\nData type per variable:\n{data_type}')
    stats_df_sub01 = df_sub01.describe()
    print(f'\nDescriptive stats for numerical variables:\n{stats_df_sub01}')
    #print(f'\nMissing:\n{df_sub03_len}')
    #print(f'\n% Missing:\n{df_sub03_len/df_sub01_len}')
    stats_df_sub01.to_excel(f'stats-{file_count}.xlsx')
    df_sub02.to_excel(f'head-{file_count}.xlsx')
    df_sub04.to_excel(f'nulls-{file_count}.xlsx')
    df_sub05 = pd.DataFrame.from_dict(new_df_dict)
    df_sub05.to_excel(f'dqr-{file_count}.xlsx')
    file_count += 1
    return file_count, df_sub05
```

Define function to output boxplots for numerical features

[18]:
```
# Create function to generate comparison boxplots (Phibbs & Carr, 2021)
def box_comp(df, var=[(None, 1.5)]):
    '''Create function to id outliers & generate compartive boxplots;
    var input uses column string & outlier threshold as x,y tuple'''
    df_sub01 = df.dropna()
    df_sub01['outlier'] = 0
    for i, j in var:
        q3, q1 = np.percentile(df_sub01[i], [75, 25]) # calculate quartiles 1 &
↪3
        iqr = q3 - q1 # calculate interquartile range
        print('\nIQR: {}-{} = {}'.format(round(q1, 4),round(q3, 4),round(iqr,
↪4))) # display IQR
```

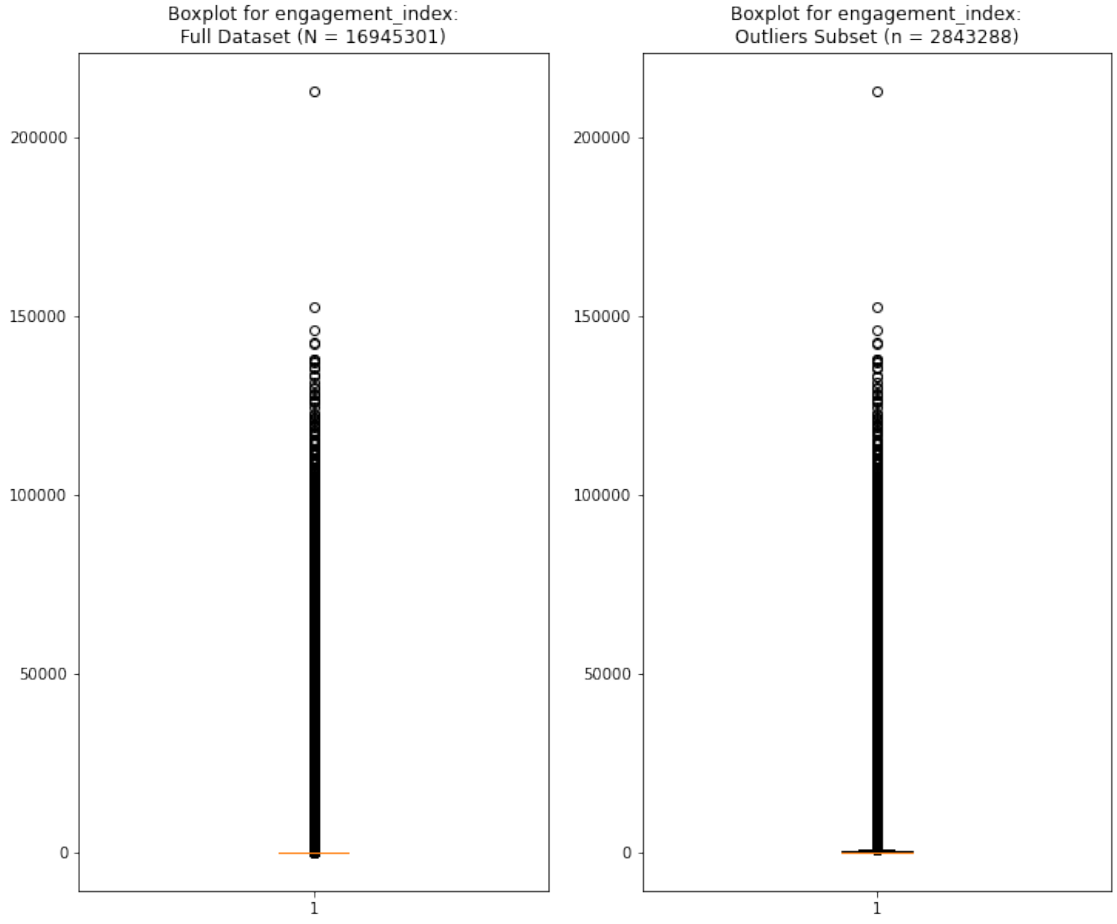
```
        iqr_out = iqr * j # calculate outlier threshold
        otlr_low = q1 - iqr_out # calculate lower outlier limit
        otlr_high = q3 + iqr_out # calculate upper outlier limit
        df_sub01_sub1 = df_sub01.loc[(df_sub01[i] < otlr_low) | (df_sub01[i] >␣
↪otlr_high)] # use .loc method to search for records that are outliers;␣
↪assign to new dataframe
        df_sub01_sub2 = df_sub01.loc[(df_sub01[i] >= otlr_low) & (df_sub01[i]␣
↪<= otlr_high)] # use .loc method to search for records that are outliers;␣
↪assign to new dataframe
        df_sub01.loc[(df_sub01[i] < otlr_low) | (df_sub01[i] > otlr_high),␣
↪'outlier'] = 1
        len01 = len(df_sub01)
        len02 = len(df_sub01_sub1)
        len03 = len(df_sub01_sub2)

        fig1, axs1 = plt.subplots(1, 2, sharey=False, figsize=(12 , 10)) # set␣
↪figure fram
        axs1[0].boxplot(df_sub01[i].dropna()) # subplot 1 for full dataset
        axs1[0].set_title('\n'.join(wrap(f'Boxplot for {i}: Full Dataset (N =␣
↪{len01})', 30)))
        axs1[1].boxplot(df_sub01_sub1[i].dropna()) # subplot 2 for outlier␣
↪dataset
        axs1[1].set_title('\n'.join(wrap(f'Boxplot for {i}: Outliers Subset (n␣
↪= {len02})', 30)))
        plt.show()

        fig2 = plt.figure(figsize=(12 , 10)) # set figure fram
        plt.boxplot(df_sub01_sub2[i].dropna()) # subplot 2 for outlier dataset
        plt.title('\n'.join(wrap(f'Boxplot for {i}: w/o Outliers Subset (n =␣
↪{len03})', 30)))
        plt.show()

        df_sub01.describe()
        print(df_sub01[i].describe()) # descriptive stats for varibale
        print('\n', df_sub01_sub1[i].describe()) # display descriptive stats of␣
↪data subset
        print('\n', df_sub01_sub2[i].describe()) # display descriptive stats of␣
↪data subset

        print('\nmean {} = {}'.format(i, round(df_sub01[i].mean(), 4))) #␣
↪average age for full dataset
        print('median {} = {}'.format(i, round(df_sub01[i].median(), 4))) #␣
↪median age for full dataset

        print('\noutliers mean {} = {}'.format(i, round(df_sub01_sub1[i].
↪mean(), 4))) # average age for outliers
```

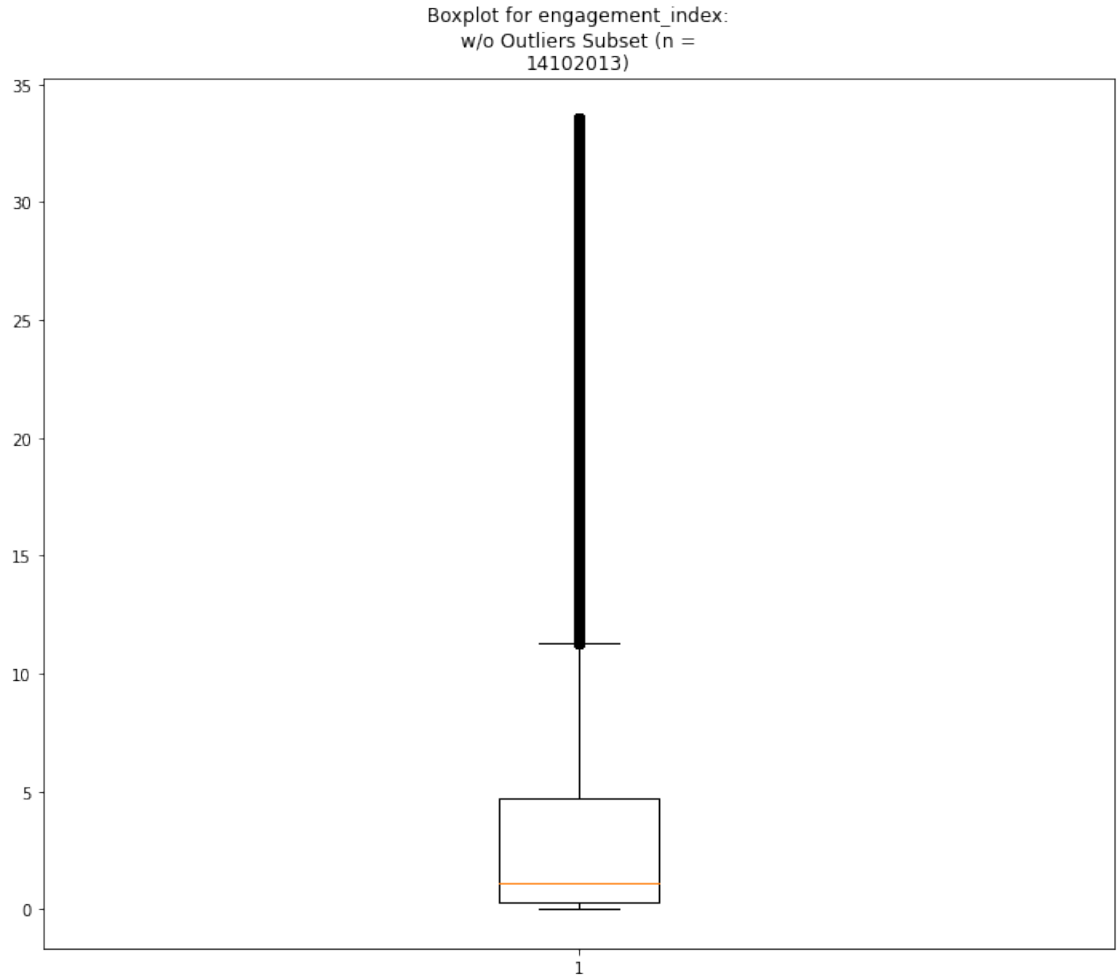
```
        print('outliers median {} = {}'.format(i, round(df_sub01_sub1[i].
↪median(), 4))) # median age for outliers
        print('Sub1 Column count = {}'.format(len(df_sub01_sub1.columns))) #␣
↪alternative way to print only number of columns
        print('Sub1 Row count = {}'.format(len(df_sub01_sub1))) # alternative␣
↪way to print only number of rows

        print('\nw/o outliers mean {} = {}'.format(i, round(df_sub01_sub2[i].
↪mean(), 4))) # average age for outliers
        print('w/o outliers median {} = {}'.format(i, round(df_sub01_sub2[i].
↪median(), 4))) # median age for outliers
        print('Sub2 Column count = {}'.format(len(df_sub01_sub2.columns))) #␣
↪alternative way to print only number of columns
        print('Sub2 Row count = {}'.format(len(df_sub01_sub2))) # alternative␣
↪way to print only number of rows

    df_sub01_sub3 = df_sub01.loc[(df_sub01['outlier'] == 0), :]
    len04 = len(df_sub01_sub3)
    for k, l in var:
        fig6 = plt.figure(figsize=(12 , 10)) # set figure fram
        plt.boxplot(df_sub01_sub3[k].dropna()) # subplot 1 for full dataset
        plt.title('\n'.join(wrap(f'Boxplot for {k}: Total Subset (n =␣
↪{len04})', 30)))
        plt.show()

    return df_sub01_sub3
```

[19]:
```
# Run function to review boxplots for selected numerical features
engagement_comb_df03 = box_comp(engagement_comb_df02, [('engagement_index', 1.
↪5), ('pct_access', 1.5)]) # display revised df to omit records with outliers
```

```
C:\Users\acarr\anaconda3\envs\ds_py37-1\lib\site-
packages\ipykernel_launcher.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy



IQR: 0.37-13.65 = 13.28

C:\Users\acarr\anaconda3\envs\ds_py37-1\lib\site-
packages\pandas\core\indexing.py:1817: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

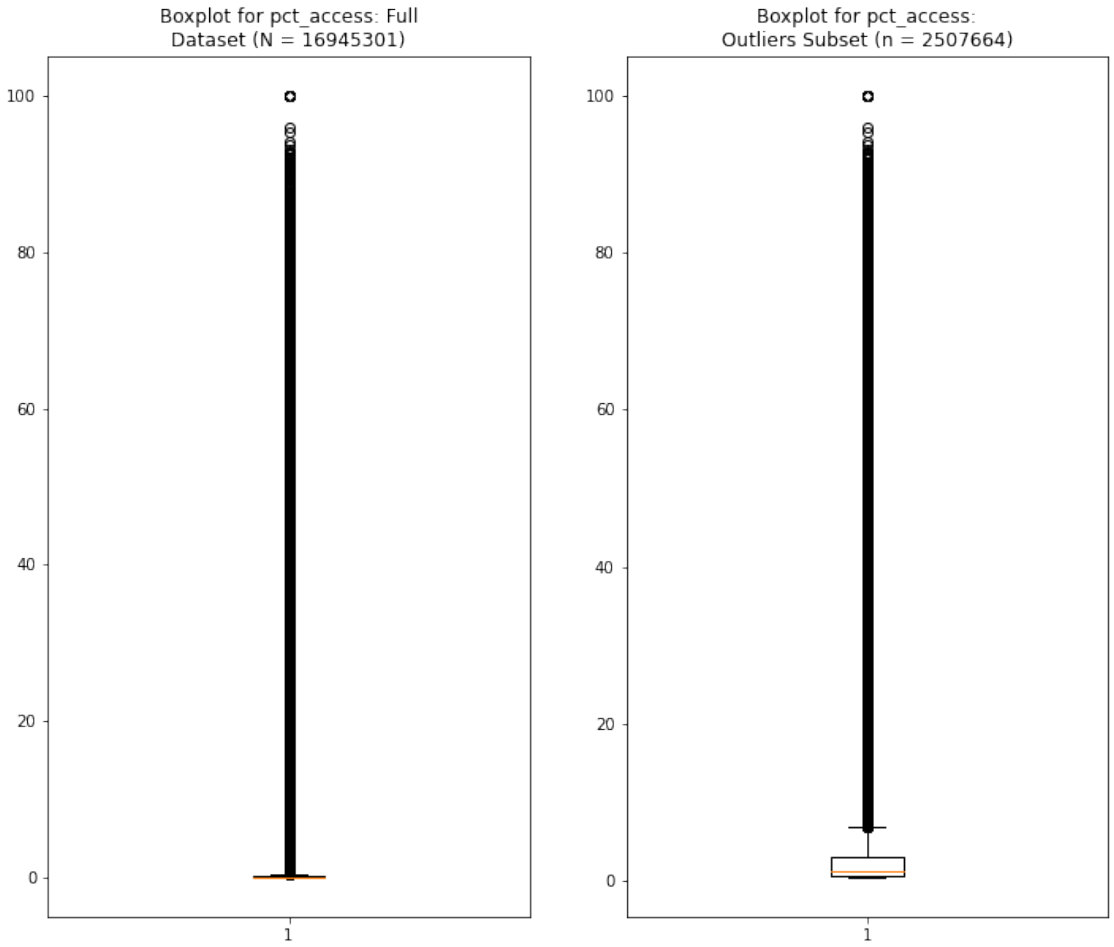
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_column(loc, value, pi)
```

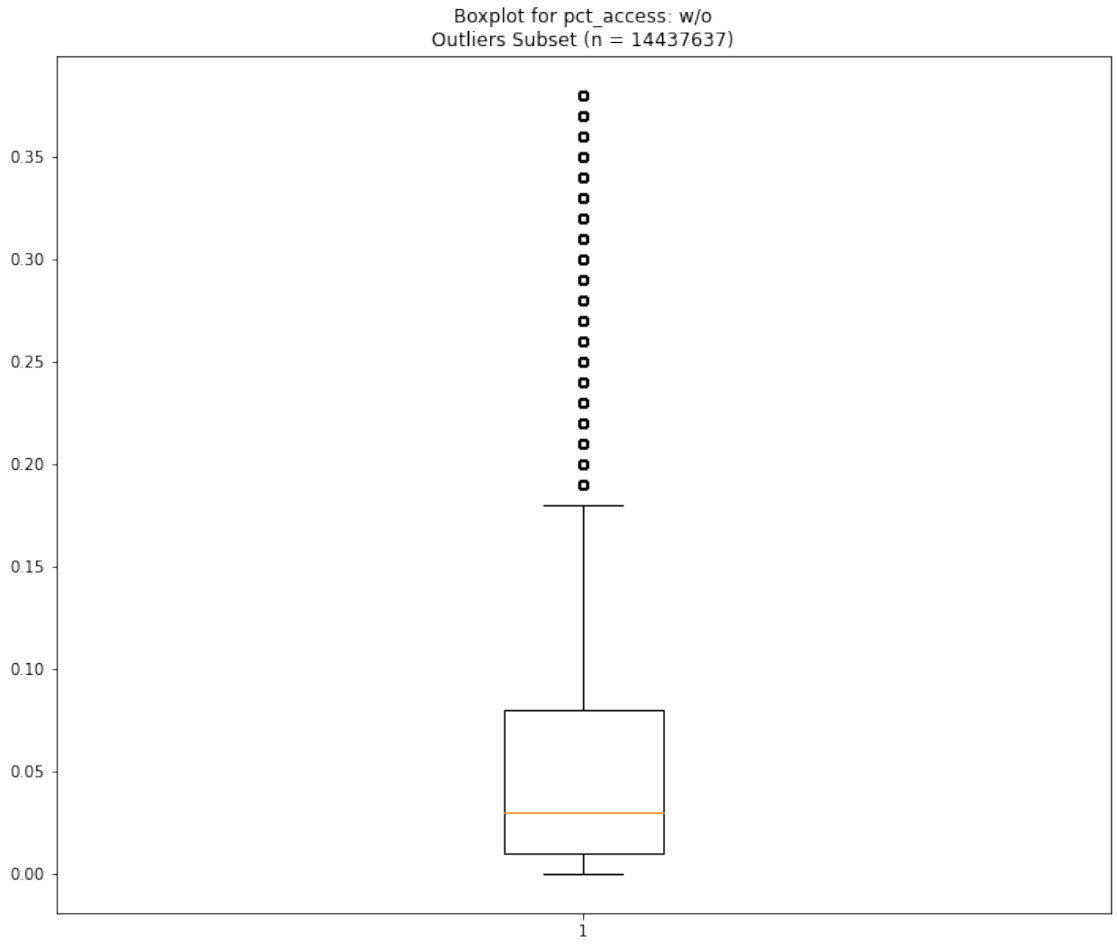





```
count    1.694530e+07
mean     1.676106e+02
std      1.682247e+03
min      1.000000e-02
25%      3.700000e-01
50%      1.920000e+00
75%      1.365000e+01
max      2.130455e+05
Name: engagement_index, dtype: float64

 count    2.843288e+06
mean     9.782071e+02
std      4.009498e+03
min      3.358000e+01
25%      5.992000e+01
50%      1.255150e+02
75%      3.836300e+02
```

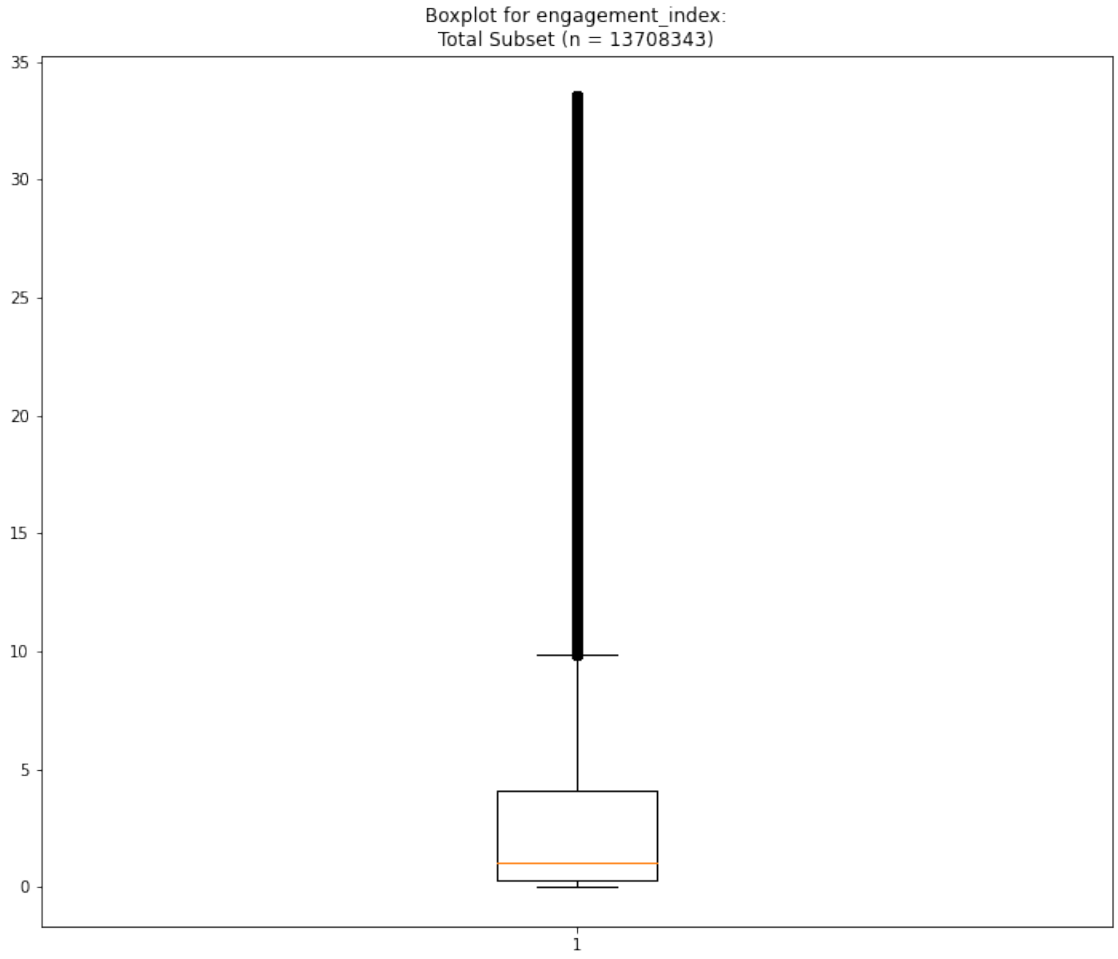
```
max      2.130455e+05
Name: engagement_index, dtype: float64

 count    1.410201e+07
mean     4.175885e+00
std      6.728099e+00
min      1.000000e-02
25%      2.700000e-01
50%      1.110000e+00
75%      4.680000e+00
max      3.357000e+01
Name: engagement_index, dtype: float64

mean engagement_index = 167.6106
median engagement_index = 1.92

outliers mean engagement_index = 978.2071
outliers median engagement_index = 125.515
Sub1 Column count = 6
Sub1 Row count = 2843288

w/o outliers mean engagement_index = 4.1759
w/o outliers median engagement_index = 1.11
Sub2 Column count = 6
Sub2 Row count = 14102013

IQR: 0.01-0.16 = 0.15

C:\Users\acarr\anaconda3\envs\ds_py37-1\lib\site-
packages\pandas\core\indexing.py:1817: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_column(loc, value, pi)
```

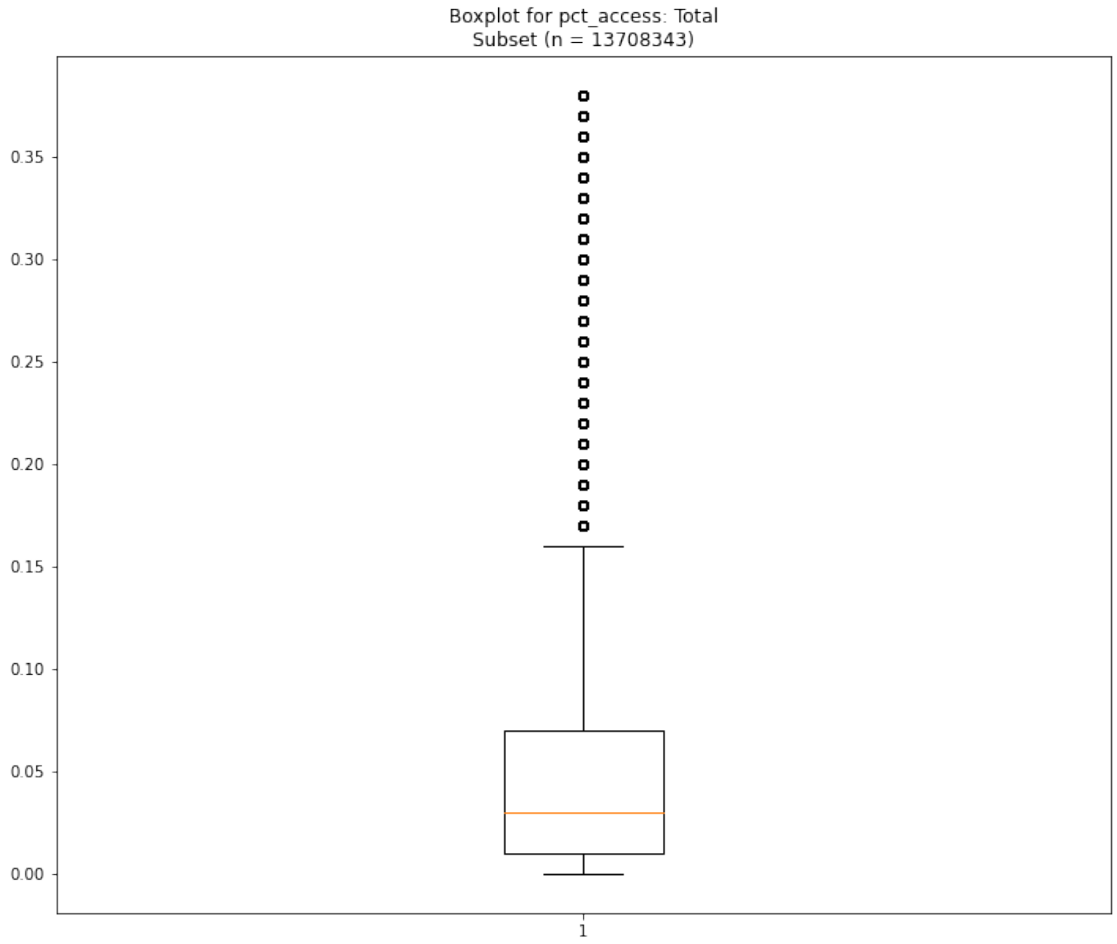

Boxplot for pct_access: Full Dataset (N = 16945301)
Boxplot for pct_access: Outliers Subset (n = 2507664)
100
80
60
40
20
0
1




```
count    1.694530e+07
mean     6.638985e-01
std      3.634980e+00
min      0.000000e+00
25%      1.000000e-02
50%      4.000000e-02
75%      1.600000e-01
max      1.000000e+02
Name: pct_access, dtype: float64

 count    2.507664e+06
mean     4.129040e+00
std      8.669304e+00
min      3.900000e-01
25%      6.200000e-01
50%      1.150000e+00
75%      3.100000e+00
max      1.000000e+02
```

```
Name: pct_access, dtype: float64

 count    1.443764e+07
mean     6.204028e-02
std      7.974911e-02
min      0.000000e+00
25%      1.000000e-02
50%      3.000000e-02
75%      8.000000e-02
max      3.800000e-01
Name: pct_access, dtype: float64

mean pct_access = 0.6639
median pct_access = 0.04

outliers mean pct_access = 4.129
outliers median pct_access = 1.15
Sub1 Column count = 6
Sub1 Row count = 2507664

w/o outliers mean pct_access = 0.062
w/o outliers median pct_access = 0.03
Sub2 Column count = 6
Sub2 Row count = 14437637
```


Boxplot for engagement_index:
Total Subset (n = 13708343)
35
30
25
20
15
10
5
0
1



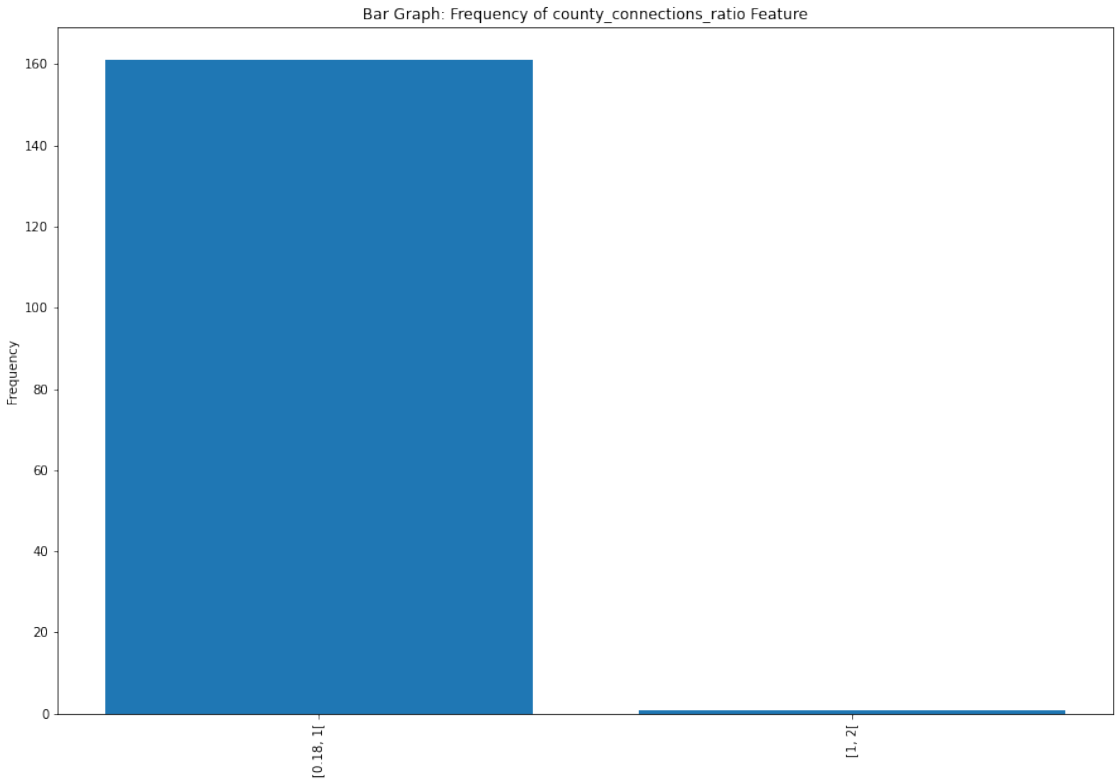
Define function to output bar charts for categorical features

[20]:
```
# Define function to output a bar diagram based on frequency
def bar_freq(df, var=None, var_des=None, sort='Frequency', figw=15, figh=10):
    '''Inputs are dataframe; col; and "Frequency" or "Element" for sorting'''
    df_sub01 = df.copy()
    cat_var_01_lst = df_sub01[var].unique()
    print(cat_var_01_lst)

    # Mini-dataframe for generating bar graph (GeeksforGeeks, 2022)
    df01_sub01_var_freq = pd.Series(df_sub01[var]).value_counts().sort_index().
↪reset_index().reset_index(drop=True)
    df01_sub01_var_freq.columns = ['Element', 'Frequency']
    if sort == 'Frequency':
        df01_sub01_var_freq.sort_values(by=['Frequency','Element'])
    else:
        df01_sub01_var_freq.sort_values(by=['Element', 'Frequency'])
```

```
    print(df01_sub01_var_freq)
    fig = plt.figure(figsize=(figw, figh))
    plt.bar(df01_sub01_var_freq['Element'], df01_sub01_var_freq['Frequency'])
    plt.ylabel('Frequency')
    plt.xlabel(f'{var_des}')
    plt.xticks(rotation='vertical')
    plt.title(f'Bar Graph: Frequency of {var} Feature')
    plt.show()
```

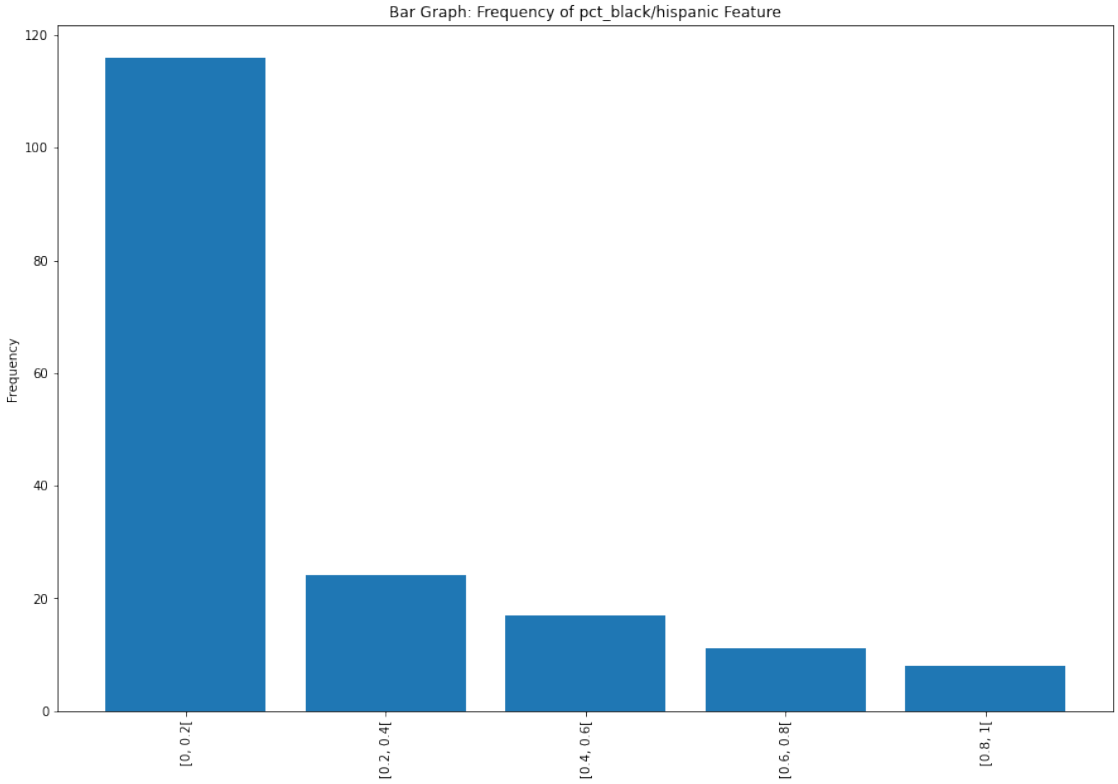
[21]:
```
# Run function to generate bar charts
bar_freq(district_df01, 'county_connections_ratio', var_des = '')
bar_freq(district_df01, 'pct_black/hispanic', var_des = '')
bar_freq(district_df01, 'pct_free/reduced', var_des = '')
bar_freq(engagement_comb_df02, 'lp_id', sort = 'Element', var_des = '')
bar_freq(st_policy_df02, 'REI_WKSR', sort = 'Element', var_des = '')
```

```
['[0.18, 1[' nan '[1, 2[']
     Element  Frequency
0  [0.18, 1[        161
1     [1, 2[          1
```

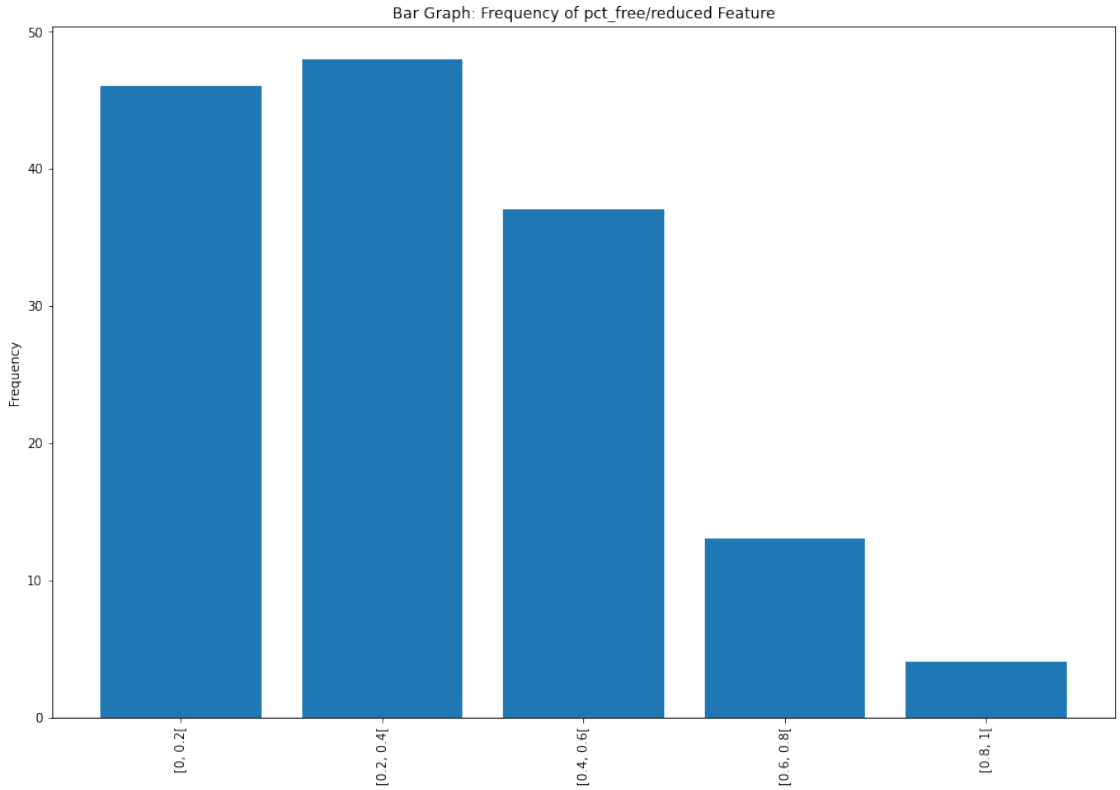



```
['[0, 0.2[' nan '[0.2, 0.4[' '[0.4, 0.6[' '[0.8, 1[' '[0.6, 0.8[']
```

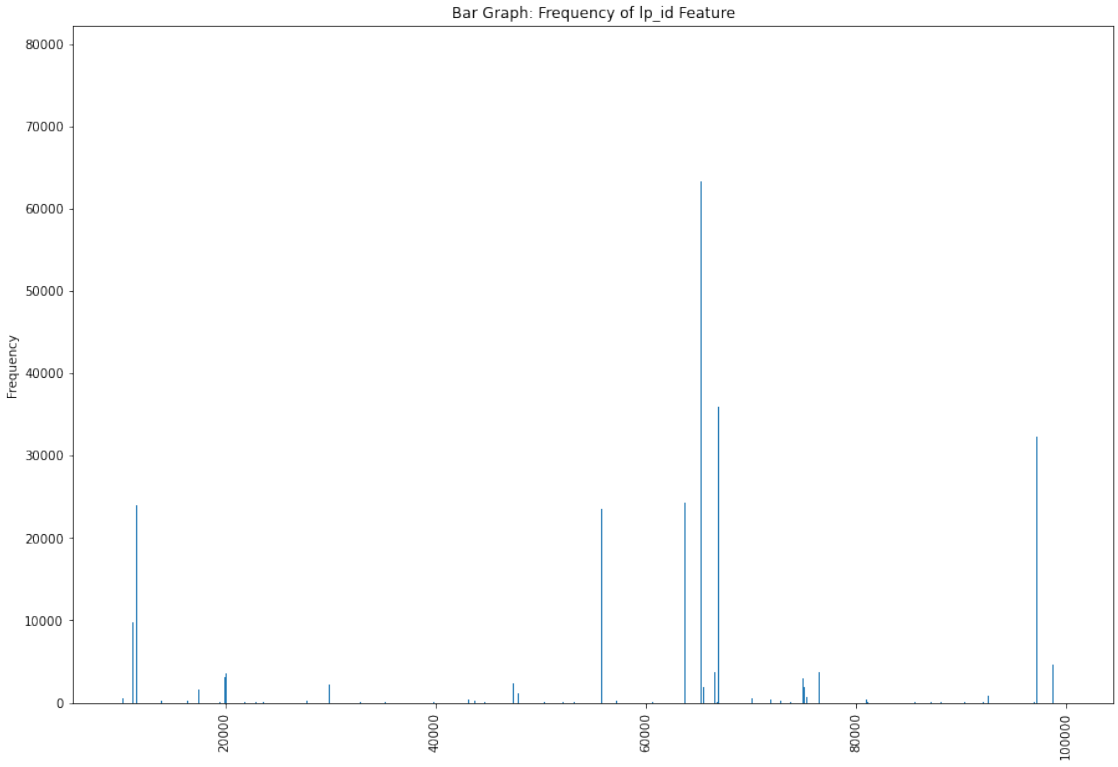
```
      Element  Frequency
0    [0, 0.2[        116
1  [0.2, 0.4[         24
2  [0.4, 0.6[         17
3  [0.6, 0.8[         11
4    [0.8, 1[          8
```



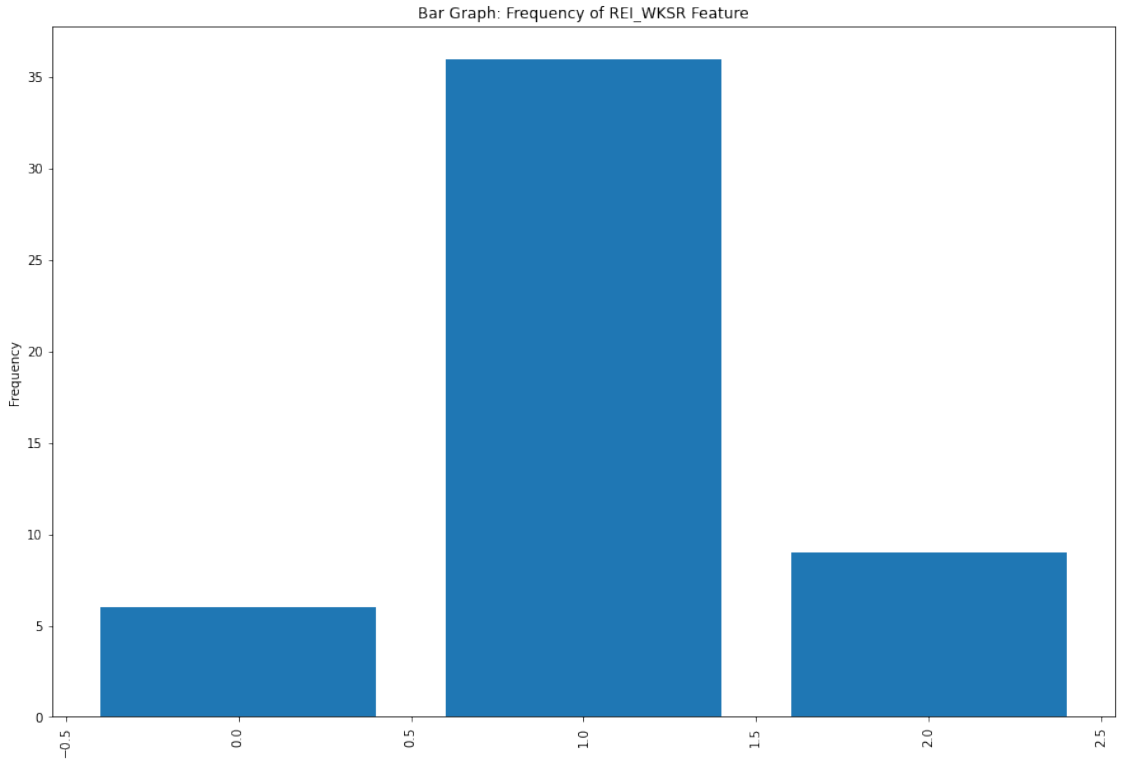

```
['[0, 0.2[' nan '[0.2, 0.4[' '[0.4, 0.6[' '[0.6, 0.8[' '[0.8, 1[']
      Element  Frequency
0    [0, 0.2[         46
1  [0.2, 0.4[         48
2  [0.4, 0.6[         37
3  [0.6, 0.8[         13
4    [0.8, 1[          4
```




```
[93690 17941 65358 … 63113 38604 43864]
      Element  Frequency
0       10003         27
1       10006        102
2       10024          2
3       10032         68
4       10035          7
…         …          …
8641    99953        938
8642    99968         29
8643    99972         40
8644    99984      14330
8645    99991          2

[8646 rows x 2 columns]
```

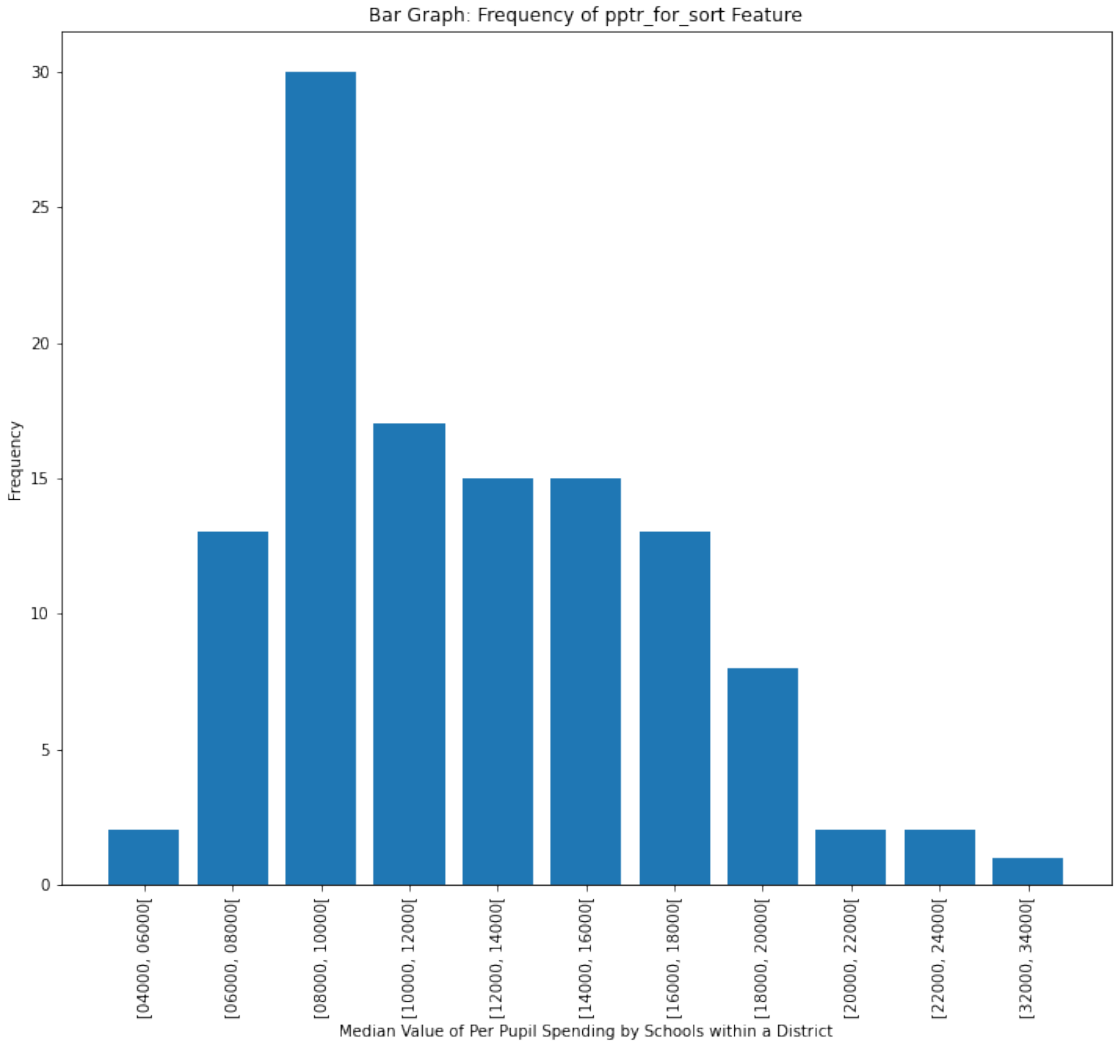



```
[ 1.  2.  0. nan]
   Element  Frequency
0      0.0          6
1      1.0         36
2      2.0          9
```



[22]:
```
bar_freq(district_df01, 'pptr_for_sort', var_des='Median Value of Per Pupil␣
↪Spending by Schools within a District', figw=12, figh=10)
```

```
['[14000, 16000[' nan '[06000, 08000[' '[10000, 12000[' '[08000, 10000['
 '[12000, 14000[' '[16000, 18000[' '[20000, 22000[' '[18000, 20000['
 '[22000, 24000[' '[04000, 06000[' '[32000, 34000[']
          Element  Frequency
0   [04000, 06000[          2
1   [06000, 08000[         13
2   [08000, 10000[         30
3   [10000, 12000[         17
4   [12000, 14000[         15
5   [14000, 16000[         15
6   [16000, 18000[         13
7   [18000, 20000[          8
8   [20000, 22000[          2
9   [22000, 24000[          2
10  [32000, 34000[          1
```



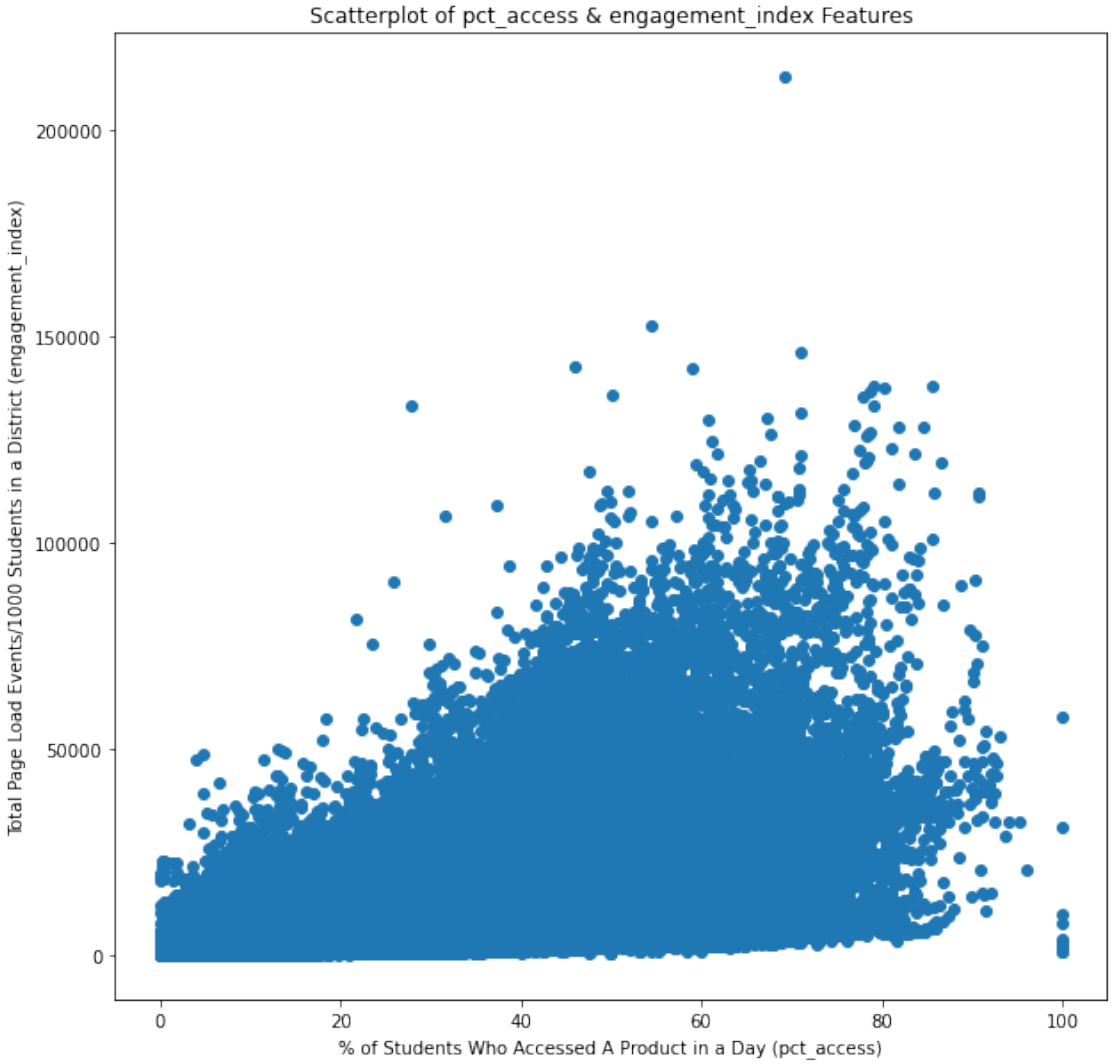
[23]:
```
engagement_comb_df02_corr_lst = ['engagement_index', 'pct_access']
engagement_comb_df02.loc[:, engagement_comb_df02_corr_lst].corr() # generate␣
↪correlation matrix
```

[23]:

	engagement_index	pct_access
engagement_index	1.000000	0.751541
pct_access	0.751541	1.000000

[24]:
```
# Display scatter plot for 2 truly continuous vars
feat_x = '% of Students Who Accessed A Product in a Day'
feat_y = 'Total Page Load Events/1000 Students in a District'
var_x = 'pct_access'
var_y = 'engagement_index'
fig10 = plt.figure(figsize = (10 , 10))
plt.scatter(engagement_comb_df02[var_x], engagement_comb_df02[var_y])
```

```
plt.xlabel(f'{feat_x} ({var_x})')
plt.ylabel(f'{feat_y} ({var_y})')
plt.title(f'Scatterplot of {var_x} & {var_y} Features')
plt.show()
```



[25]:
```
#Run function to display descriptive statsitics for individual df's
file_count, district_df01_stats = df_stats(district_df01, file_count)

file_count, st_policy_df02_stats = df_stats(st_policy_df02, file_count)

file_count, product_df02_stats = df_stats(product_df02, file_count)

engagement_comb_df02s = engagement_comb_df02.sample(n = 500000)
```

```
file_count, engagement_comb_df02s_stats = df_stats(engagement_comb_df02s,␣
 ↪file_count)
```

```
First 10 rows:
   district_id           state  locale pct_black/hispanic pct_free/reduced  \
0         8815        Illinois  Suburb           [0, 0.2[         [0, 0.2[
1         2685             NaN     NaN                NaN              NaN
2         4921            Utah  Suburb           [0, 0.2[       [0.2, 0.4[
3         3188             NaN     NaN                NaN              NaN
4         2238             NaN     NaN                NaN              NaN
5         5987       Wisconsin  Suburb           [0, 0.2[         [0, 0.2[
6         3710            Utah  Suburb           [0, 0.2[       [0.4, 0.6[
7         7177  North Carolina  Suburb         [0.2, 0.4[       [0.2, 0.4[
8         9812            Utah  Suburb           [0, 0.2[       [0.2, 0.4[
9         6584  North Carolina   Rural         [0.4, 0.6[       [0.6, 0.8[

  county_connections_ratio    pp_total_raw pptr_lb pptr_ub   pptr_for_sort
0                [0.18, 1[  [14000, 16000[   14000   16000  [14000, 16000[
1                      NaN             NaN     NaN     NaN             NaN
2                [0.18, 1[    [6000, 8000[   06000   08000  [06000, 08000[
3                      NaN             NaN     NaN     NaN             NaN
4                      NaN             NaN     NaN     NaN             NaN
5                [0.18, 1[  [10000, 12000[   10000   12000  [10000, 12000[
6                [0.18, 1[    [6000, 8000[   06000   08000  [06000, 08000[
7                [0.18, 1[   [8000, 10000[   08000   10000  [08000, 10000[
8                [0.18, 1[    [6000, 8000[   06000   08000  [06000, 08000[
9                [0.18, 1[   [8000, 10000[   08000   10000  [08000, 10000[

Number of df rows = 233

Null count per variable:
district_id                  0
state                       57
locale                      57
pct_black/hispanic          57
pct_free/reduced            85
county_connections_ratio    71
pp_total_raw               115
pptr_lb                    115
pptr_ub                    115
pptr_for_sort              115
dtype: int64

Data type per variable:
district_id                 int64
state                      object
locale                     object
```

```
pct_black/hispanic          object
pct_free/reduced            object
county_connections_ratio    object
pp_total_raw                object
pptr_lb                     object
pptr_ub                     object
pptr_for_sort               object
dtype: object

Descriptive stats for numerical variables:
       district_id
count   233.000000
mean   5219.776824
std    2595.751581
min    1000.000000
25%    2991.000000
50%    4937.000000
75%    7660.000000
max    9927.000000

First 10 rows:
                  STATE             CLSCHOOL             END_BSNS  \
0               Alabama  2020-03-20 00:00:00  2020-04-30 00:00:00
1                Alaska  2020-03-16 00:00:00  2020-04-24 00:00:00
2               Arizona  2020-03-16 00:00:00  2020-05-08 00:00:00
3              Arkansas  2020-03-17 00:00:00  2020-05-04 00:00:00
4            California  2020-03-23 00:00:00  2020-05-08 00:00:00
5              Colorado  2020-03-23 00:00:00  2020-05-01 00:00:00
6           Connecticut  2020-03-17 00:00:00  2020-05-20 00:00:00
7              Delaware  2020-03-16 00:00:00  2020-05-08 00:00:00
8  District of Columbia  2020-03-16 00:00:00  2020-05-29 00:00:00
9               Florida  2020-03-17 00:00:00  2020-05-18 00:00:00

                 FM_ALL               QR_END             EMSTART  \
0  2020-07-16 00:00:00                    0  2020-04-03 00:00:00
1  2020-04-24 00:00:00  2021-02-14 00:00:00  2020-03-23 00:00:00
2                    0  2020-05-12 00:00:00  2020-03-24 00:00:00
3  2020-07-20 00:00:00  2020-06-15 00:00:00                    0
4  2020-06-18 00:00:00                    0  2020-03-27 00:00:00
5  2020-07-16 00:00:00                    0  2020-04-30 00:00:00
6  2020-04-20 00:00:00  2021-03-19 00:00:00  2020-03-16 00:00:00
7  2020-04-28 00:00:00  2020-06-01 00:00:00  2020-03-17 00:00:00
8  2020-04-17 00:00:00                    0  2020-03-15 00:00:00
9                    0  2020-08-05 00:00:00  2020-04-02 00:00:00

                 EMEND             EMSTART2               EMEND2 EMSTART3  \
0  2020-06-01 00:00:00                    0                    0        0
1  2020-07-01 00:00:00                    0                    0        0
```

```
2  2020-10-31 00:00:00                   0                   0        0
3                    0                   0                   0        0
4  2020-09-02 00:00:00                   0                   0        0
5  2020-06-14 00:00:00  2020-10-21 00:00:00  2021-01-01 00:00:00        0
6  2021-02-10 00:00:00                   0                   0        0
7  2020-07-01 00:00:00                   0                   0        0
8  2021-01-31 00:00:00                   0                   0        0
9  2020-10-01 00:00:00                   0                   0        0

   …             WV_WTPRD REI_WTPRD WV_WKSR REI_WKSR UIQUAR UIHIRISK  \
0  …  2020-03-16 00:00:00         0     1.0      1.0    1.0      0.0
1  …  2020-03-18 00:00:00         0     1.0      1.0    1.0      0.0
2  …  2020-03-20 00:00:00         0     1.0      1.0    1.0      1.0
3  …  2020-03-17 00:00:00         0     1.0      1.0    1.0      0.0
4  …  2020-03-12 00:00:00         0     1.0      1.0    1.0      1.0
5  …  2020-03-20 00:00:00         0     1.0      2.0    0.0      0.0
6  …                    0         0     1.0      1.0    0.0      0.0
7  …                    1         0     1.0      1.0    0.0      0.0
8  …  2020-03-17 00:00:00         0     1.0      1.0    1.0      0.0
9  …  2020-03-31 00:00:00         0     1.0      1.0    1.0      0.0

   UICLDCR  UIEXTND              EBSTART                EBEND
0      0.0      0.0  2020-05-31 00:00:00  2020-09-26 00:00:00
1      1.0      0.0  2020-05-03 00:00:00                    0
2      1.0      0.0  2020-06-14 00:00:00  2020-12-12 00:00:00
3      0.0      0.0  2020-05-31 00:00:00  2020-10-17 00:00:00
4      0.0      0.0  2020-05-10 00:00:00                    0
5      0.0      0.0  2020-05-31 00:00:00  2020-11-28 00:00:00
6      1.0      0.0  2020-04-26 00:00:00                    0
7      1.0      0.0  2020-05-24 00:00:00  2021-01-09 00:00:00
8      0.0      0.0  2020-05-24 00:00:00                    0
9      0.0      0.0  2020-06-07 00:00:00  2020-11-07 00:00:00

[10 rows x 24 columns]

Number of df rows = 53

Null count per variable:
STATE        2
CLSCHOOL     2
END_BSNS     2
FM_ALL       2
QR_END       2
EMSTART      2
EMEND        2
EMSTART2     2
EMEND2       2
EMSTART3     2
```

```
EMEND3       1
SMSTART      2
SMEND        2
SMSTART2     2
WV_WTPRD     2
REI_WTPRD    1
WV_WKSR      2
REI_WKSR     2
UIQUAR       2
UIHIRISK     2
UICLDCR      2
UIEXTND      2
EBSTART      2
EBEND        2
dtype: int64

Data type per variable:
STATE         object
CLSCHOOL      object
END_BSNS      object
FM_ALL        object
QR_END        object
EMSTART       object
EMEND         object
EMSTART2      object
EMEND2        object
EMSTART3      object
EMEND3        object
SMSTART       object
SMEND         object
SMSTART2      object
WV_WTPRD      object
REI_WTPRD     object
WV_WKSR      float64
REI_WKSR     float64
UIQUAR       float64
UIHIRISK     float64
UICLDCR      float64
UIEXTND      float64
EBSTART       object
EBEND         object
dtype: object

Descriptive stats for numerical variables:
         WV_WKSR   REI_WKSR     UIQUAR   UIHIRISK    UICLDCR    UIEXTND
count  51.000000  51.000000  51.000000  51.000000  51.000000  51.000000
mean    0.882353   1.058824   0.823529   0.215686   0.372549   0.058824
std     0.325396   0.544491   0.385013   0.415390   0.488294   0.237635
```

```
min     0.000000   0.000000   0.000000   0.000000   0.000000   0.000000
25%     1.000000   1.000000   1.000000   0.000000   0.000000   0.000000
50%     1.000000   1.000000   1.000000   0.000000   0.000000   0.000000
75%     1.000000   1.000000   1.000000   0.000000   1.000000   0.000000
max     1.000000   2.000000   1.000000   1.000000   1.000000   1.000000

First 10 rows:
   LP ID    Product Name                      Sector(s)  \
0  13117     SplashLearn                        PreK-12
1  66933    ABCmouse.com                        PreK-12
2  50479          ABCya!                        PreK-12
3  92993           ALEKS             PreK-12; Higher Ed
4  73104     Achieve3000                        PreK-12
5  37600  Actively Learn                        PreK-12
6  18663     AdaptedMind                        PreK-12
7  65131         Amplify                        PreK-12
8  26491         Answers             PreK-12; Higher Ed
9  56441         Audible  PreK-12; Higher Ed; Corporate

                         Primary Essential Function
0                   LC - Digital Learning Platforms
1                   LC - Digital Learning Platforms
2  LC - Sites, Resources & Reference - Games & Si…
3                   LC - Digital Learning Platforms
4                   LC - Digital Learning Platforms
5                   LC - Digital Learning Platforms
6                   LC - Digital Learning Platforms
7                       LC - Courseware & Textbooks
8                            LC - Study Tools - Q&A
9  LC - Sites, Resources & Reference - Streaming …

Number of df rows = 372

Null count per variable:
LP ID                          0
Product Name                   0
Sector(s)                     20
Primary Essential Function    20
dtype: int64

Data type per variable:
LP ID                          int64
Product Name                  object
Sector(s)                     object
Primary Essential Function    object
dtype: object

Descriptive stats for numerical variables:
```

```
             LP ID
count    372.000000
mean   54565.795699
std    26247.551437
min    10533.000000
25%    30451.000000
50%    53942.500000
75%    77497.000000
max    99916.000000

First 10 rows:
              time  lp_id  pct_access  engagement_index  district_id
1853    2020-01-06  72932        0.06              0.78         5510
53370   2020-05-07  40508        0.07              3.49         2549
260885  2020-12-11  57513        0.01              1.67         7177
6197    2020-02-14  32340        0.09              8.71         1131
25896   2020-11-21  61292        7.73           2246.36         2017
40847   2020-02-27  40278        0.00              0.07         2956
234892  2020-11-20  25267        0.10              1.65         6919
59811   2020-03-25  24460        0.01              0.11         4051
16781   2020-03-17  69827        2.37            174.57         6998
138389  2020-08-24  45716        0.04              1.24         9537

Number of df rows = 500000

Null count per variable:
time                    0
lp_id                   0
pct_access            312
engagement_index    120339
district_id              0
dtype: int64

Data type per variable:
time                 object
lp_id                 int32
pct_access          float64
engagement_index    float64
district_id           int32
dtype: object

Descriptive stats for numerical variables:
               lp_id     pct_access  engagement_index    district_id
count  500000.000000  499688.000000     379661.000000  500000.000000
mean    54736.898062       0.513166        167.030046    5236.748712
std     26457.051819       3.216250       1615.139747    2644.572542
min     10003.000000       0.000000          0.010000    1000.000000
25%     31027.000000       0.000000          0.370000    2956.000000
```

50%	55007.000000	0.020000	1.930000	4929.000000
75%	77656.000000	0.100000	13.760000	7675.000000
max	99984.000000	91.070000	93508.740000	9927.000000

3. Initial Step Towards Builing ABT - Initial Combination of Datasets

[26]:
```
#Join District Info Table w/ State Policy Table to create ABT sub1
lp_abt_df01 = district_df01.merge(st_policy_df02, how='left', left_on='state',␣
 ↪right_on='STATE')

#Join Engagment Table w/ ABT sub 1 to create ABT sub 2
lp_abt_df01 = engagement_comb_df02.merge(lp_abt_df01, how='left',␣
 ↪left_on='district_id', right_on='district_id')

#Join ABT sub 2 w/ Product Info Table to create ABT sub 3 (final)
lp_abt_df01 = lp_abt_df01.merge(product_df02, how='left', left_on='lp_id',␣
 ↪right_on='LP ID')
```

[27]:
```
end_time = dt.datetime.today()
time_elapse = end_time - start_time
print(f'Start Time = {start_time}')
print(f'End Time = {end_time}')
print(f'Script Time = {time_elapse}')
```

```
Start Time = 2022-02-19 13:52:22.129547
End Time = 2022-02-19 14:07:17.490483
Script Time = 0:14:55.360936
```