A. The seven Data Analytics life cycle phases are: Business Understanding, Data Acquisition, Data Cleaning, Data Exploration, Predictive Modeling, Data Mining, and Reporting and Visualization.

1. Business understanding is a crucial phase in the data analytics life cycle because it provides context and gives meaning to every other phase. This phase is where we identify the significant questions that need to be answered, formally define the needs of all involved stakeholders, and ensure that we have ample resources to complete the analysis on time. Keeping the organization's goal and mission in mind during this phase is crucial. Knowing the organization's goal and mission helps analysts refine processes to align with these desires. Currently, I do not have much experience in the business world, as I come from an educational background. To gain expertise in this phase, I will engage with case studies and find real-world data analytics problems analysts have solved to understand why data analysis was necessary and what techniques analysts utilized to solve those problems based on their organizations' goals and missions.

2. Data Acquisition is the phase where the data itself is collected. The analyst may find this data within the company in a database, in which case SQL would be a valuable tool to access the data. Alternatively, analysts may need to acquire data from outside sources, in which case, web scraping tools or mass surveys may be necessary. Currently, I have introductory experience in SQL and Python. To gain expertise in this phase, I will sharpen my skills in SQL, work with web scraping tools in Python, and continue to gain experience working with large amounts of data.

3. Data Cleaning is when the analyst turns raw data collected in the Data Acquisition phase into clear, relevant data. We use tools like SQL, Python, R, and Excel to perform actions like deleting duplicate responses and pruning irrelevant data sets and outliers. I have some basic experience cleaning data in Excel for publication during my college years. To gain expertise in this phase, I will continue working with large datasets to accurately identify extraneous data points and create processes to eliminate them.

4. Data Exploration is the phase where we begin visualizing our data and identifying basic patterns and trends. We use tools such as Tableau and more straightforward calculations, like R-squared values, to identify patterns and ensure the patterns are accurate. I have a fair amount of experience identifying relationships and patterns from my physics studies in college. To gain expertise in this phase, I will gain experience in Tableau and continue strengthening my knowledge of graphing and statistical analysis.

5. Predictive modeling is the phase where we expand on the patterns we discovered in the Data Exploration phase. We build on those patterns using tools

like R and Python to create predictive models for the future based on the data we already have. I have some experience constructing basic models during physics experiments in college. To gain experience in this phase, I will continue strengthening my skills in Python and apply them specifically to different modeling techniques to familiarize myself with the optimal circumstances of each method.

6. Data mining is when we allow computers to look deeper into the data and identify patterns we did not easily visualize or discover. Many different data mining algorithms in R and Python can provide a much deeper and more detailed analysis of the data than humans can alone. Currently, I don't have much experience with large-scale data mining techniques. To gain expertise in this phase, I will work with several of these different data mining algorithms and explore various fields of machine learning.

7. Reporting and Visualization is the phase where we take all of the conclusions we've drawn from other life cycle phases and put them into terms that involved stakeholders will understand. This phase is when we communicate our findings to all the parties involved to provide them with the insight required to achieve our desired goal. Currently, I have some experience presenting large sets of data from my physics research in college. To gain expertise in this phase, I will continue creating graphs, presentations, and interactive dashboards in tools like Tableau to work on turning extensive data sets into visuals that teams can easily understand without a statistical background.

B. Currently, I work as a teacher in the 9th biggest school district in the United States. We have over 200,000 students in the district, and there has been a massive push in recent years to make data-driven decisions when planning our curriculums for our students. Due to this push, analysts must store and analyze large amounts of student data. Since the COVID-19 pandemic, we've seen substantial learning losses in reading and mathematics scores on all standardized tests compared to grade-level scores before 2020. We recognized these trends and have been very interested in identifying different trends in student performance to see how we. To analyze this much data, tools like Python would be necessary to properly go through the Data Mining phase of the cycle to get a clean look at the data of hundreds of thousands of students.

Some of the risks involved in using Python to look more deeply at some of the patterns in student data include concerns about the privacy of student data, how algorithms handle the racial and socio-economic differences throughout the school district, and how accurate the algorithms used to go through this large of a data set are. When dealing with student data, privacy is a significant concern. Many laws like FERPA (Family Educational Rights and Privacy Act) are in place to ensure that only schools and students' families can view student performance metrics. Analysts would require an extensive, accessible database of student data to have a tool like Python mine through student performance data to look for patterns. While historically, student data has not been the subject of major attacks and data breaches, as we've

seen at several online retailers and financial institutions, it still leaves the district vulnerable to an attack and risks being liable for FERPA violations.

Each school in our district has vastly different racial and socio-economic demographics. I've worked at three different high schools in the district, and their differences are apparent. Analysts must be careful only to apply algorithms trained to account for data like socio-economic status when analyzing student performance to avoid any biases in their conclusions.

Student performance has always been an important data point for schools, and analyzing something like this must be precise. When these data points are driving instruction, inaccurate data can make the lives of students, teachers, and parents more difficult in the long run. It is crucial that data is accurate and that proper models are applied.

C. A tool like Python is a significant need for an organization like a large school district because there is simply too much student data to analyze without the assistance of a tool like Python. Despite these risks, we have been presented with large amounts of student data to analyze and use to make prescriptive choices in our instruction to improve student understanding. Student performance has slowly improved since the 2020 COVID pandemic. The district's strong push for data-driven instruction began during the 2021-2022 school year. On Florida's standardized tests, achieving a Level 3 or higher score is considered passing. According to the Florida School Report Card for Orange County Public Schools, after the first year, 52.7% of our district's students scored a level 3 or higher on the state's standardized Mathematics tests, and 52.5% scored a level 3 or higher on the state's standardized English Language Arts tests. During the 2023-2024 school year, 3 years after student performance data became widely available to teachers and the emphasis on data-driven instruction began, 58.8% of students scored a level 3 or higher on the state's Mathematics tests, and 56.2% of students scored a level 3 or higher on the state's standardized English Language Arts texts. While the additional data trends provided to teachers via data mining programs like Python may not have been the sole cause of the student performance increase, a correlation exists between the amount of data provided to teachers and students' performance on state exams.

Using Python does provide some ethical risks. First of all, student privacy is essential both legally and ethically. Analysts must take all steps to protect student data from escaping the eyes of the schools and families. We also must keep in mind that we are keeping data on children who are not yet legally allowed to consent to analysis, so analysts must keep parents in the loop and receive their permission to run these analyses to attempt to improve student performance. Analysts must also keep data points in mind, like socio-economic status, to prevent bias while closely guarding these data points to protect students from discrimination.

Works Cited

1. “2023-24 ORANGE SCHOOL DISTRICT REPORT CARD.” *Florida Department of Education*, Florida Department of Education, 24 July 2024, edudata.fldoe.org/ReportCards/Schools.html?school=0000&district=48.