**Task 1: The Data Analytics Life Cycle**

Joseph Emmanuel Cayetano

Western Governors University

D596: The Data Analytics Journey Task 1

September 4, 2024

1. **Describe all the Data Analytics Life Cycle Phases with Own Reflection for Each**

The Data Analytics Life Cycle (DALC) is a data framework that data analysts and scientists usually follow when working with data to ensure that data-driven initiatives are successful and achieve business goals. The DALC has seven phases, which are: business understanding, data acquisition, data cleaning, data exploration, predictive modeling, data mining, and reporting and visualization. The first phase, which is business understanding, is about defining the business questions that need to be addressed and understanding the scope of the analysis. I have worked as a data analyst at various healthcare organizations for years, and I think this is the most important step in the DALC primarily because it sets the foundation for the entire data analytics process. I had some experience working with key stakeholders in the healthcare industry, and I made sure that I understood their expectations and needs so there would not be misunderstanding later in the data analysis project. The second phase, which is data acquisition, is about collecting data from various sources such as databases, surveys, and web scraping. I typically view electronic health records that contain patient medical histories and diagnoses, and I have exported such data source to an analytics software like R before. The process was easy and straight-forward. The third phase is data cleaning, and it involves transforming the collected raw data into a clean and structured format that can be used for analysis. Cleaning data is a process that should not be skipped because it is important that the data follow consistent formats. Like for example, I regularly make sure that I standardize patient’s date of birth, unit measurements, etc. so I won’t run into trouble later in the analyzation part. I also delete data that I think won’t be necessary for analysis. The fourth phase is data exploration, and it involves understanding the data and its relationships. This phase relies on descriptive statistics and data visualization. Data exploration is my favorite phase in the DALC because I enjoy identifying patterns and trends in the data. For example, working with medical data, I like seeing trends in patient admissions and the outcomes of treatments. I also like using histograms and bar charts to show distribution of certain variable in the medical dataset. The fifth phase is predictive modeling, and it involves creating models to predict outcomes that an organization is interested in. While I do not have much experience building models that can predict future outcomes, I have seen some models in medical organizations that can predict the likelihood of patient readmissions, and I thought it was interesting that predictive modeling is possible in health datasets. The sixth phase is data mining, and it involves computers discovering patterns and useful information from large amounts of data. I always enjoy looking for patterns and relationship in medical datasets because I get to learn new information about patients and healthcare operations. Also, learning such information is crucial because it can help improve the way hospitals treat its patients. The last phase is reporting and visualization, and it involves using visualizations like graphs, dashboards, and charts to present the finding from the analysis to the stakeholders. I have created many visualizations from both the patient and operation data in healthcare for years, and I have presented them to hospital administrators in hopes to improve operation efficiencies in the hospital. This phase is very important because it turns data into actionable insights.

**A1. Provide an Example for Each DALC Phase Where You Might Gain Expertise**

For the business understanding phase, I will try to be more involved in communicating with stakeholders (healthcare providers and administrators) by conducting meetings and interviews with them to better understand their business goals and ensure these goals align with data analytics projects. For the data acquisition phase, I will continue importing data from various sources such as spreadsheets and CSV files. I will also learn to web scrape, as I believe this skill is crucial whenever I want to extract large volumes of data from websites. For the data cleaning phase, I will continue to do data validation and standardization to make it easier to do data analysis. I will learn to deal with outliers in the data because they can have a significant negative impact on the data results. It is important that the data is accurate as possible. For the data exploration phase, I will learn to use central measures more often so I could understand the distribution of the data more and spot potential outliers. For the predictive modeling phase, I shall familiarize myself with the common algorithms like logistic regression, decision trees, etc., as these machine learning algorithms are useful to predict future outcomes and trends in the healthcare industry. For the data mining phase, I will learn to differentiate and use both supervised and unsupervised models because they help understand patterns and make predictions from large data sets. Lastly, for the reporting and visualization phase, I will continue to create graphs and charts that showcase trends, as they help the stakeholders make informed decisions. Also, I will learn to use proper scales in graphs to avoid misinterpreting the story to the stakeholders.

**A2. How the Goal and Mission of the Organization Help Analyst Identify the Business Requirement**

One of the goals of our healthcare organization is to reduce hospital readmission rates by at least 15% over the next year. This helps the analyst identify the business requirements by shifting its focus on understanding the factors that contribute to readmissions. The analyst would need to identify patient demographics and cases, and determine which groups are at most risk. Also, the analyst would need to analyze the effectiveness of clinical processes. With this goal in mind, it helps the data analyst to prioritize business requirements that contribute to lowering hospital admission rates.

1. **Select One Data Analytics Tool and Describe its Use in One DALC Phase**

One data analytics tool that I always use when working with data is Jupyter Notebook. Jupyter Notebooks can be used “for all sorts of data science tasks including data cleaning and transformation, numerical simulation, exploratory data analysis, data visualization, statistical modeling, machine learning, deep learning, and much more” (Science, 2020, para. 4). Basically, Jupyter Notebook can be used in any phase of the DALC, but the phase I will specifically focus on is data cleaning. It is common to see missing values, duplicates, and unstandardized data in medical datasets. For example, it is common to see different date formats, unit of measurements, text cases, etc. With the help of Jupyter Notebook, it allows me to clean and standardized the data to make it easier to do data analysis.

**B1. Three Risks of the Selected Tool for Data Analytics**

While Jupyter Notebook comes with many benefits, it also comes with some risks. One of the risks is it can get unresponsive when working with large datasets. Jupyter Notebook only runs in a single kernel, which means its performance is limited, and it doesn’t help when it is exploring large amounts of data. It is common for the Jupyter Notebook to crash when working with large datasets. Another risk is Jupyter Notebook shows both data and code, and having this combination pose a risk of exposing sensitive information like patient records in hospital data, if not anonymized properly. The third risk is the reproducibility issues of Jupyter Notebook. Lack of proper documentation for the data and code dependencies can make it challenging to reproduce the notebooks.

**B2. Describe an Organizational or Technical Problem Using the Selected Tool**

One technical problem for using Jupyter Notebook is that it is difficult to use it for collaboration with other data analysts. Jupyter Notebook is not designed for multi-user collaboration and does not have real-time editing like Google Docs do. This means data analysts must take turns making changes, which is very inefficient.

**C. Describe the Decision-Making Process of Selecting Data Analytics Tool from Part B**

Jupyter Notebook is the best data analytics tool that our organization could use because it is easy to explore and analyze data using Jupyter Notebook. Also, Jupyter Notebook is capable of creating informative graphs and charts, which is useful in medical research. Additionally, it is much easier to clean the data in Jupyter Notebook than in any other analytical software. It is very common to see missing values, duplicates, unstandardized data in healthcare datasets, so having a tool that can easily help clean the data truly helps data analysts working in healthcare. Lastly, most data analysts in our organization are proficient with Jupyter Notebook, so it is reasonable to use it as our main data analytical tool.

**C1. Justify the Organizational or Technical Need for the Selected Tool**

My organization decided to use Jupyter Notebook for analyzing data because our organization had limited budget, and they wanted to use an analytical tool that provides not only the most significant cost savings, but also the most accessible. This is the reason why they chose Jupyter Notebook as their main data analytical tool. It also helps that Jupyter Notebook does not have additional licensing fees.

**C2. Summarize the Results of Using the Selected Tool in the Life Cycle Phase You Selected in B**

In the data cleaning phase, we cleaned the patient data in Jupyter Notebook by formatting the admission dates correctly and ensuring that all dates in the dataset are consistent with each other. We also imputed missing data and eliminated duplicates to have an accurate data analysis. These methods made our patient data clean and consistent.

**C3. Evaluate the 3 Potential Ethical Problems of Using the Selected Data Analytics Tool in Part B1**

There are ethical concerns when using Jupyter Notebook. First, since Jupyter Notebook combines both data and code, there is a possibility that sensitive information can be seen by unauthorized individuals. This could lead to violations of confidentiality and privacy. Second, bias can exist when working with data in Jupyter Notebook. When data or algorithms for machine learning are not carefully examined for fairness, it can lead to biased results. We must always strive to produce the most accurate results as possible. Third, there may be a temptation that the data will be used beyond consent. For example, the analyst may use exploratory data analysis on patient data in ways that might not align with patients’ expectations. This is definitely unethical and violates the use of patient data.

**References**

Science, O.-O. D. (2020, July 15). Why you should be using Jupyter Notebooks. Medium. https://odsc.medium.com/why-you-should-be-using-jupyter-notebooks-ea2e568c59f2