Gabriela Masak

D596 – Task 1

9/1/2024

The Data Analytics Life Cycle

Data is everywhere. We collect data with our eyes and ears, with sensors and chips, and with algorithms and programs. This data can tell a story, lead to a decision, or make a prediction, but needs to be transformed first in order to be used. Data is transformed by the data analytics life cycle, which is defined by a series of seven steps, through various tools and techniques.

The first step in the data analytics life cycles involves meetings with business stakeholders to determine the scope and objectives of the analysis (Renze, 2022). Business stakeholders are anyone with a vested interest in the outcome of the analysis (Hoory, 2022). These meetings help a data analyst to understand where the business is and what they want to do with the data. The outcome of these meetings will align with the business’ goal and mission, which are imperative to not only the data used, but how it is transformed. Businesses have many uses for data analytics, such as predictive, prescriptive, diagnostic, or descriptive applications, which each define a separate and independent style of analysis (Johnson, 2022). These meetings help analysts narrow and pinpoint the appropriate data to use and how to transform it. Using the understanding of the business’ goals and mission is essential for determining what, how, and why data will be used and its transformation.

A data analyst can gain expertise in this phase of the data analytics life cycle by connecting with an organization that values collaborative, informed decision-making. Meeting with an organization’s stakeholders allows the analyst to understand diverse perspectives, gather comprehensive requirements, and ensure that the analysis aligns with the organization’s goals and needs. For example, the analyst could meet with a real estate company. The stakeholders for the real estate company, presumably realtors and developers, may ask the analyst to create a prediction for the growth of local commercial real estate in the upcoming year and how the expansion of public transportation into the area will affect pricing. The data analyst will extract several components of these details given during these meetings to drive subsequent steps of their data analysis, making sure to ask questions and stay involved with stakeholders to continually align the process with the needs and goals of the business.

The second step of the data analytics life cycle involves acquiring the raw data, if the data is not already procured. Data can be acquired directly by the business, by a secondary party paid by the business, or by a third party such as the government. Data is stored in local databases or can be accessed by various APIs or data streams (Renze, 2022). Data can exist in various forms and is divided into three categories: structured, semi-structured, or unstructured. Structured data is qualitative or quantitative data that can be organized into rows and columns. Semi structured data may contain structured data, but also includes additional metadata that is harder to organize in a similar fashion. Unstructured data includes images, videos, and other file formats that cannot be well organized or arranged (Marr, 2019). These types of data and how they will be used should be considered using the information collected in the first stage, in order to make sure that the analysis will align with the business’ goals and mission. If the business is focused on a quantitative analysis, structured, quantitative data should be collected. If the business wants to analyze digital media or documents, the analyst will focus on collecting unstructured data.

A data analyst can gain expertise in the data acquisition phase by using public or private sources to collect data, or using data furnished directly from an organization they are working with. For a data analyst working with a real estate company, they may use publicly accessible data from the United States Census Bureau or the local tax appraisal office. They may purchase data from a company such as Zillow or Batchleads. The real estate company may have its own private data for the analyst to use as well. Techniques such as web scraping or using APIs may be employed by the analyst at this step. Once the analyst feels like they have a sufficient amount of data, they will begin the next step, data cleaning. They may return to this step if subsequent steps show that the data is insufficient, inadequate, or that more data is necessary.

Data cleaning is the third step of the data analytics life cycle. At this stage, the analyst focuses on preparing and validating the data to ensure that data is accurate and formatted correctly for future analysis. When possible, the analyst will correct data that is incomplete, inaccurate, irrelevant, corrupt, or incorrectly formatted using a variety of techniques (Hillier, 2021). This includes deleting data that is duplicated or irrelevant. Remembering the company’s goals and mission is essential to refining the data and making sure that only relevant data is included in the dataset. Data that does not contribute to the analysis should be excluded and removed. The analyst will also correct structural errors, such as typos, inconsistent capitalization, or unnecessary white space. Punctuation, inconsistent labeling, and mislabeled categories are also scrutinized (Hillier, 2021). Fixing these errors allows for data to be standardized. Data should be standardized with the organization in mind, as the type of data storage, manipulation, and resulting analysis are largely dictated by the organization.

Once standardized, the analyst can focus on removing unwanted outliers and fixing contradictory data errors (Hillier, 2021). Outliers are figures that are dramatically different than the majority of the data set. They could indicate a broken gauge or sensor, or erroneous manual data entry, but could also represent a valid data point, which should be determined by the analyst with caution. The analyst will also address contradictory, type conversion, and syntax errors in order to make sure that the data is consistent, as well as figuring out whether to remove or impute missing data (Hillier, 2021). After these potential errors are addressed, the analyst can validate the dataset for subsequent analysis.

For a data analyst working with a real estate company, expertise in data cleaning can be gleaned by using various tools and techniques to clean the data. If data is derived from different sources, it must be standardized. The analyst would check to make sure that dates or addresses are formatted similarly, for example. Typos and grammatical errors would be addressed. They would also check for missing data, such as if the price or neighborhood were not included for a data entry. The analyst could research to see if they can figure out what the missing values are and update the data or determine whether to exclude these data points if it is not available. The data would also be checked for relevancy, as the real estate company may only want to focus on a localized area, rendering data from other areas as unnecessary. Once the data is standardized, the analyst can take a more comprehensive look at the data and find outliers such as pricing that was incorrectly entered and missing a decimal point, fixing the data, or determining whether an outlier such as a unique block sale that could skew the analysis should be removed. Once the dataset is cleaned and standardized, the analyst will proceed with data exploration.

The fourth step in the data analytics life cycle is data exploration. This step involves informally identifying patterns, trends, and relationships within the dataset where the data analyst will develop a greater understanding of the data’s quality and key characteristics by employing various statistical techniques and data visualization tools. Data exploration allows the data analyst to identify issues that may lead them to revert to a prior step or use the trends and patterns discovered to guide future analysis, ultimately leading to more reliable and accurate analytical outcomes. It includes several steps: variable identification, univariate analysis, bi-variate analysis, missing value treatment, outlier treatment, variable transformation, and variable creation (Ray, 2016). Outlier and missing value treatment are also stages in the data cleaning step, but may become more apparent as data is manipulated and explored. Variable exploration is defined by identifying the predictor and target variables, then selecting the data type and category of each variable to determine whether they are univariate or bivariate. Univariate variable are characterized by continuous, which represent quantitative data, or categorical variables, which represent qualitative data. Bivariate analysis explored the relationship between two variables (Ray, 2016). Scatter plots, two-way tables, stacked column charts and chi-square tests may be used by a data analyst to visually or statistically represent the relationships within the data for exploration. Results of this analysis may lead the analyst to transform the variables or create derived or dummy variables. Ultimately, the results of the exploration will guide the analyst in determining which mathematical and computational models to use in subsequent steps.

Within the real estate example, the data analyst can gain expertise in the data exploration phase by creating a scatter plot to determine whether a correlation exists and how strong that correlation is between various variables. They may find a strong correlation between real estate prices and local income or proximity to public water features. The analyst would use the company’s goals and mission to narrow in on key points and variables for future analysis, which would have them selecting sale prices as the dependent variable for future predictive analysis, with several independent variables such as proximity to public transportation and lot size. The exploration may lead them to determine that there are several outliers that were missed during the data cleaning step and create a strategy to correct or remove them. Once the exploration is complete and the analyst is confident in their understanding of the patterns, trends, and relationships within the data, they will begin the data modeling step.

Data modeling represents the fifth step of the data analytics life cycle. The data analyst will use mathematical or computational models to make predictions or derive insights from the data. Tools such as Python and R feature several libraries and functions to carry out the automated computations of these models. These models offer more insight than the data exploration step and can extend the correlations of that step, predict future values or probabilities of a certain event, or cross-validate predictive models to ensure accuracy (Renze, 2022).

A data analyst can gain experience in the data modeling by familiarizing themselves with the different packages and libraries for R and Python. When using R, ggplot2, tidyverse, and caret are essential for data analysts, or Numpy, pandas, and TensorFlow for Python. Learning how to correctly write code to import data, call functions, and create visuals and models can be practiced and applied through dummy data provided online. Many online courses and projects simulate real-life scenarios and their solutions and models can be used on similar projects and data. For the real estate example, a data analyst may research which models were used for similar data collections and recreate the analysis, using the previous work to guide and check their work, before applying their skills to their real data.

Data mining is the sixth step in the data analytics life cycle. Machine learning, also called artificial intelligence, is applied to data via supervised and unsupervised models. With supervised learning the data is broken into a group of data used to trained the classification or regression algorithm and the group of data that the subsequent algorithm will use to produce results. Unsupervised learning does not need data to be grouped or separated beforehand. Clustering, association, and dimensionality reduction can be derived with machine learning algorithms without needing to use data for training (Delua, 2024). The type of machine learning method applied depends on the data and goals of the organization. The results are analyzed and sent to the shareholders.

Expertise can be gained in the data mining step by exploring the different machine learning techniques and researching how they can be implemented with Python code. In the real estate company example, the data analyst might use supervised machine learning regression methods to predict future house prices. The analyst could explore how similar data has been used in past studies. By experimenting with different algorithms, such as linear regression, decision trees, and random forests, the analyst can determine which method provides the most accurate predictions, while honing their expertise and gaining experience too.

The final step of the data analytics life cycle is reporting and visualization. The data analyst will keep the stakeholders in mind while crafting a summary with actionable insights from the analytical findings. Their background and expectations should be considered. The report should not include overly technical jargon and should focus on conveying the findings using captivating graphs and interactive dashboards, which can be creating using tools such as Tableau (Renze, 2022). Ultimately, the analyst should appropriately convey their findings and what they mean for the organization in a manner that is relevant to the organization’s goals and mission.

Expertise can be gained in the reporting and visualization step from practice, research, and exploring Tableau and its features. An analyst can research how different data is reported and visualized, noting how different methods convey different information more effectively. They can consider how they could communicate the same data to different stakeholders and how the degree of statistical education, type of organization, or visual impairments should be considered. For the analyst in the real estate example, the analyst would consider how to communicate how different variables affect pricing and which types of visuals can convey that. Instead of focusing on the statistical methods used, the analyst should focus on how the real estate company can use the findings to drive business decisions.

As a data analyst, I will use these steps iteratively while I use data to tell stories. While mostly inexperienced, I have completed a few educational projects applying data analytics steps, tools, and techniques. I have personally used a Microsoft tool called Excel to clean data and complete basic data analysis. Features such as the filter function allow for duplicate data to be found and removed. Dates can be standardized easily with various functions and formatting tools. Excel also provides features that allow for whitespace and Unicode characters to be removed or replaced. A spell checker can locate and suggest improvements for typos. Unnecessary columns and rows can be easily selected and deleted. While several features and tools are relevant and useful for a data analyst, using Excel poses several major risks. If the organization keeps all data in a singular, shared Excel file, data can irreversibly become compromised or deleted with a single user error, and this could affect the entire organization and any future analysis. Strategies for backing up older versions or limiting user capabilities in modifying data would allow an organization to avoid some technical issues posed under data accuracy and quality. As Excel is within the Microsoft family of tools, an organization may risk finding itself overly dependent on Microsoft. This may hamper innovation and impact flexibility if a new tool, technique, or system is developed outside of the vendor that is necessary for evaluating and analyzing data. For example, if a company wanted to integrate unstructured data such as photos of houses into its pricing analysis, Excel would not allow the inclusion or analysis of photos, much less an opportunity for the images to be cleaned. Lastly, there is a major risk associated with privacy when using Excel. Personal information such as full names, addresses, phone numbers, related financial transactions are not fully secure in an Excel spreadsheet. There are some features that hinder access to certain sheets and data, but Excel lacks the ability to encrypt sensitive information. Lack of encryption could lead to privacy breaches and set the business up for legal consequences.

While Excel has its limitations, it can be appropriate for more basic applications of data analytics. For a real estate company attempting to make basic predictions about pricing, Excel may fit its needs for a basic regression analysis. Information about real estate sales is publicly available and can usually be exported to an Excel file or converted from a .csv file. The data can be cleaned for errors, outliers, and missing values. A pivot table could be created or graphs and various functions employed to model and analyze the data. The data analyst could find a high correlation between pricing and school ratings, then apply basic statistical functions and features within Excel to calculate a prediction for how the most recent school rating will affect the local real estate prices. This information would be packaged and communicated to real estate agents to help drive sales or purchases. If the data shows that the prices may decrease, agents are suggested to focus on marketing to sellers. If the data shows that the prices may increase, agents are encouraged to focus on marketing to buyers.

There may be some ethical concerns involved, which should be acknowledged. Firstly, historical data can reflect discriminatory practices and predictive models can perpetuate this bias. In real estate, there are areas with noticeable racial or income disparities. A predictive model may continue to undervalue minority neighborhoods if this historical bias is not addressed. Secondly, there may be ethical issues with using subjective or inaccurate data. Using subjective data such as school ratings may introduce misleading predictions, as school ratings can be based on opinion, with different criteria weighed variably or inaccurately, depending on the individual or institution creating the metric. Thirdly, there is an ethical concern that the data could be used unethically to manipulate or exploit consumers into accepting deals they are not comfortable with due to artificially inflated or deflated prices.

In conclusion, data has its own story as it is analyzed and transformed to tell a story. Data analysts use several tools and techniques throughout the seven-step data analytics life cycle. Understanding the risks and ethical concerns are just as important as understanding and demonstrating expertise of each step.

Sources Cited

1. Delua, J. (2024, May 13). *Supervised vs. unsupervised learning: What’s the difference?* IBM. <https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning>
2. Hillier, W. (2021, November 12). What Is Data Cleaning and Why Does It Matter? Careerfoundry.com; Career Foundry. <https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/>
3. Hoory, L., & Bottorff, C. (2022, August 7). *What is a stakeholder analysis? Everything you need to know*. Forbes; Forbes. <https://www.forbes.com/advisor/business/what-is-stakeholder-analysis/>
4. Johnson, J. (2022, June 28). *Business leaders and data analytics - Data analytics for business professionals*[Video].LinkedIn. https://www.linkedin.com/learning/data-analytics-for-business-professionals-14936642/business-leaders-and-data-analytics?resume=false&u=2045532
5. Marr, B. (2019, October 18). *What’s The Difference Between Structured, Semi-Structured And Unstructured Data?* Forbes.com; Forbes. <https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/>
6. Ray, S. (2016, January 10). *A Complete Tutorial which teaches Data Exploration in detail*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
7. Renze, M. (2022, November 30). Data science: the big picture[Video]. Pluralsight. https://app.pluralsight.com/ilx/video-courses/data-science-big-picture