**Data Curation and Data Type Analysis**

Laurence T. Burden

Purdue Global University

IN498: Bachelor’s Capstone in Analytics

Robert Kayl

November 13, 2023

**Part 1**

Data curators are responsible for selecting, organizing, and presenting data to the rest of the data team (Pratt, 2022). The data is selected based on its relevancy to the problem set being analyzed. The data is then organized by cleaning and transforming it into a usable state. Finally, the data is presented to the rest of the data team. This usually involves building a data catalog or an API for accessing the data in other analysis tools.

Data curators need to have a strong sense of organization. This will help them find the best way to compile multiple data sources into usable data pools. They should have a degree that gives academic knowledge in curation techniques.

Curators may fall into a few places in an organization. They may sit between management and a data science team. This allows them to have more control over how data is stored and transformed before analysis. The control can enforce better standards for a data analytics group and result in multiple benefits for them.

Data curators may be given a range of responsibilities. They may be tasked with designing and implementing tools that ingest and analyze data. They may be given the role of writing standards documents. They may also be asked to act as a bridge between an operations team and an analytics team.

**Part 2**

There are many important elements of data curation. Three important elements are data availability, data quality, and data governance. These three aspects all help the team create the best analysis possible.

Data availability is the process of ensuring data is readily accessible when it is needed. This quality means that data analysts will have access to the information when and where they need it. If the data is not available, then analysis cannot be completed.

Data quality is the accuracy and completeness of the data. Any analysis of inaccurate or incomplete data will result in incorrect insights being found. A data curator must ensure the data conforms to a high standard of data quality to prevent this.

Data governance is a high-level process within an organization to levy and enforce standards on data across all departments. A strong data governance program will ensure that less work needs to be done on data before analysis is performed. Data curators can assist in overseeing the data governance program within their sphere of influence.

**Part 3**

The data provided for the app store contains multiple CSV files. Each file has 14 attributes for each observation. The first attribute is the Date. This would be stored in a date data type. The Package\_Name attribute can be stored as a string. The Country attribute is a two-digit country code that can be stored as a string.

The Store\_Listing\_Visitors represents the number of people to visit the store listing and can be represented as an integer. The next attribute is for Installers. This represents the number of people who installed the app after visiting the store listing. It can be represented as an integer. The Visitor-to-Install\_convsersion\_rate represents the percentage of listing visitors that end up installing the app. This can be represented with a floating-point number.

The last eight attributes represent the number of installs retained for 1, 7, 15, and 30 days, along with the retention rate for the same 1, 7, 15, and 30-day marks. The number of installs retained can each be represented by an integer since a fractional install cannot exist. The retention rate will need to be represented by a floating-point number.

I would preserve the original data by storing it in a separate folder on a shared drive. This allows us to retrieve the original data in the case of unexpected corruption on our working set. The working set will be stored in an access-controlled folder on the shared drive, as well. The access will be limited to only those on the data team and any IT admins that require higher-level access. This can be accomplished through using Microsoft’s Active Directory®.

The data will be usable by any analysis system able to ingest CSV files. No system should aim to overwrite the original data files. This will allow for reuse of the information by other systems.

The data contains missing data. This can be filled in by imputing the missing values. Any missing dates should be inferred from their most likely value based on the surrounding dates and the country they are next to. Missing numerical data can be calculated using the mean of all available data.

I would format any dates based on the ISO date format which is “YYYY-MM-DD” (Javascript Date Formats, n.d.). The values will be separated into individual cells using MS Excel®’s Text to Columns feature. All data can then be combined into one file by copying and pasting every value into one file.

**References**

Pratt, M. K. (2022, January 31). *What is data curation? - definition from Searchbusinessanalytics*. Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-curation

Javascript date formats. (n.d.). https://www.w3schools.com/js/js\_date\_formats.asp