University Libraries

UNIVERSITY OF COLORADO **BOULDER**

Data Cleanup Report

March 17, 2023

Team 1: Jozie Wille, Lexi Dingle, Charlie Rudy,
Micaiah Lowe, Quincy Marcotte

Table of Contents

Table of Contents

Data Cleaning

Introduction and Strategy

To ensure that our analysis is relevant and usable, we outlined our data cleanup steps below. We started by identifying the necessary and helpful features and removing irrelevant ones from the dataset. This involved checking for duplicates, correcting errors, and deciding on the best approach to handle missing values. We also used the Pandas report to identify any inconsistencies or outliers that required further cleaning. Through the steps detailed below, we ensured that the final dataset was clean and accurate, enabling us to provide reliable pricing recommendations to the Universityof Colorado Boulder Bookstore (CU Bookstore).

Our data cleaning process involved four steps. Step 1 was removing duplicate and unnecessary data. Step 2 involved handling outliers and missing data. In Step 3, we made structural adjustments to the data. Finally, in Step 4, we performed data validation. Throughout the project, we have only used books marked as "required" and excluded any optional books, based on the book_status column.

In the following sections, we will describe our steps to clean the CU Bookstore data.

Dropped Columns

To streamline our dataset and make it more efficient for analysis, we began our data cleaning process by dropping duplicative columns, had a significant number of null or zero values, or were not relevant for our analysis.

We first identified the columns that were duplicates of another column and removed them to reduce redundancy in our dataset. We then assessed the percentage of null or zero values in the remaining columns and eliminated those that did not meet our criteria. Lastly, we identified columns that were not necessary for our analysis, which we removed as well.

Table 1 provides a list of the columns we dropped and the reasoning behind each removal. By removing these columns, we were able to focus our analysis on the most relevant data points and improve the accuracy of our results.

Feature Name	Reason	Details
undergrad_course	No Necessary Information	Because all values are 'True,' this column does not provide any useful information
author	No Necessary Information	Unnecessary data for analysis and 34% missing values
title	No Necessary Information	Unnecessary data for analysis and 34% missing values
publisher	No Necessary Information	Unnecessary data for analysis and 41% missing values
term_hours	Redundant Column	This column is highly correlated with full_time so it may not provide any additional information.
hours_online_remote	Redundant Column	This column is highly correlated with hours_on_campus so it may not provide any additional information.
deg_seek_flg	Redundant Column	This column is highly correlated with primary_college so it may not provide any additional information.
section_code	Redundant Column	This column provides the same information that we could get from course_title.
section	No Necessary Information	Do not need the section number.
isbn	Missing Values	34% of the data for this column was missing.
ce	No Necessary Information	This column is for Continuing Education students and they will be excluded from the pricing bundle.
dept	No Necessary Information	We do not need to know the department where the course is taught.
class_level	No Necessary Information	The pricing bundle will apply for all undergraduate students.
term_cd	Redundant Column	This column is highly correlated with term_desc and so we can get term information from term_desc.

Table 1: All dropped columns from our dataset

Dropped Rows

After we dropped all the columns we will not be using, we moved on to remove unnecessary rows in our data. We first checked for duplicates in our rows to ensure that no rows were repeated and there were none. Once we confirmed that each row only appeared once, we removed the unnecessary rows for our analysis. These rows contained observations for textbooks that are not

required for a class. We have included a table below (See Table 2) describing the rows we removed and our reasoning.

Rows	Reason
All non-required textbooks	We are only looking at required textbooks in order to give a more accurate prediction of price.
When course_hours was zero	Because these observations showed no course hours being taken, they are not beneficial for analysis.

Table 2: All dropped rows from our dataset

Dealing with Missing Values

To address the missing values in our data set, we first looked at the amount of missing data, which was 22.1% of all cells. Our solution is to impute the means into all the numeric columns. We will do this by replacing the missing values with the average of the non-missing values in the columns. We are aware that imputing the means could affect the distribution of the original data, and it may lead to bias if the missing data is not missing at random. Removing rows with missing data was another option our group considered, but we chose to impute the means instead because it would be less disruptive and preserve the accuracy of our analysis better than dropping these rows. For the non-numeric columns, we will be dropping the rows with missing values.

Dealing with Outliers

We took several steps to detect and handle outliers in our dataset. Our first step was to examine the minimum and maximum values, 1st and 3rd quartile values, box plots, and scatter plots to identify true outliers accurately and determine their significance. We examined all categories that had cost and price for outliers to determine how to set prices accurately and which cost is most accurate for students. We also investigated outliers for full-time status, class level, online hours, and on/off-campus hours to ensure our data is consistent and accurate.

For some of our variables, we used bar graphs to visualize outliers. In Figure 1, we observed outliers equal to zero, representing non-full-time students, and removed them from the dataset since they represented a small percentage.

Figure 1: Bar Graph for Full Time

In Figure 2, there were a small fraction of data points that fell into the ‘Other values’ category, and we removed them from the dataset as they could muck up our data results.

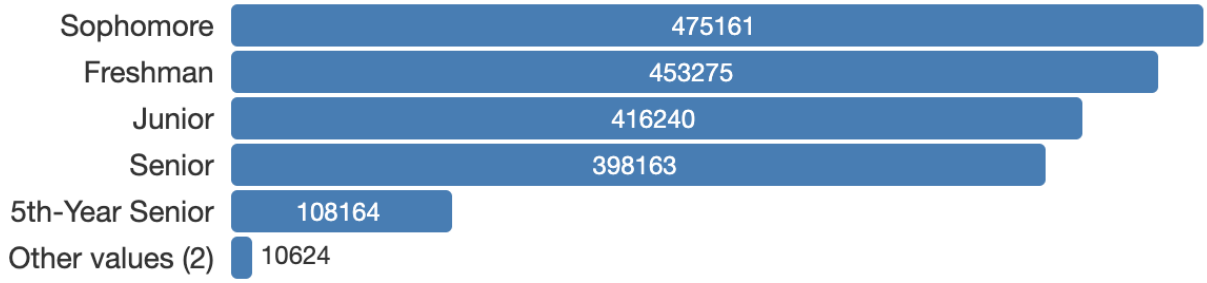

Figure 2: Bar Graph for Class Level

For the new print price and cost, seen in Figures 3 and 4, the distributions of price and cost were extremely similar, and there was an outlier present with a skewed distribution for the cost (Figure 3). We removed any data points for costs above $250 in this dataset.

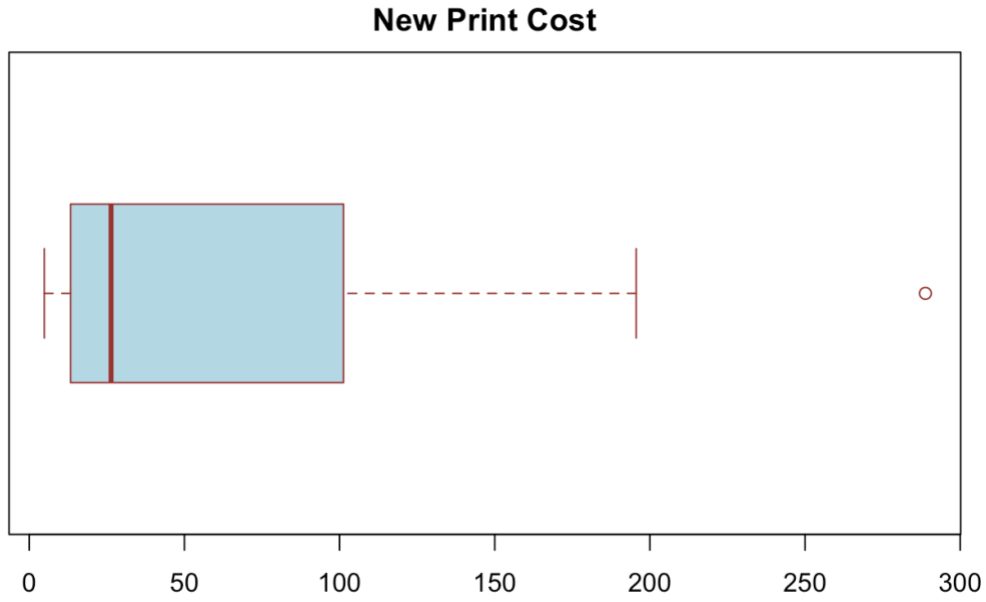

Figure 3: New Print Cost

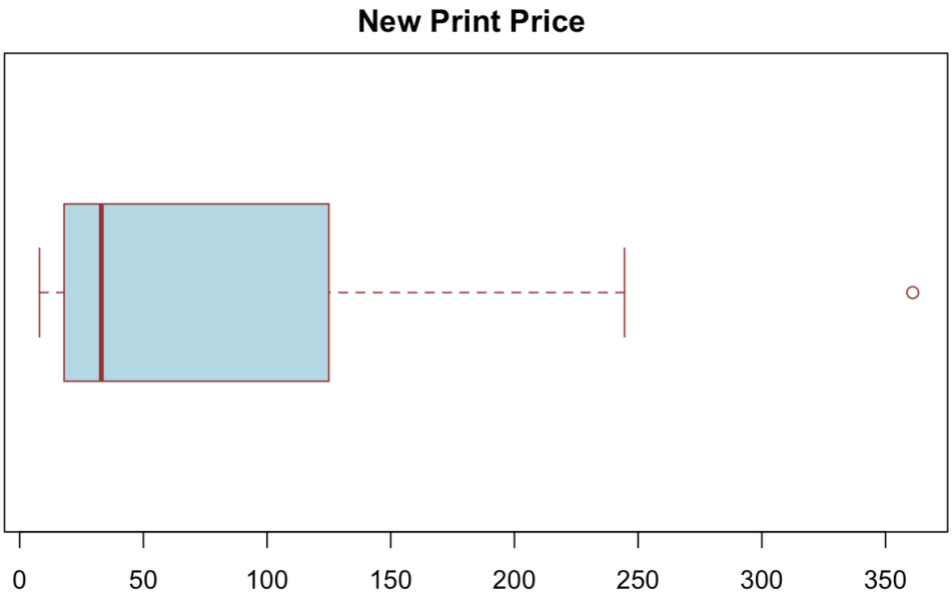

Figure 4: New Print Price

We analyzed all the price and cost columns in the dataset since they were numeric and crucial to the analysis. The results showed that there were no outliers to deal with in the Day One Access Price and Cost, Digital Perpetual Access Price and Cost, and Digital 180 Access Price and Cost columns, as seen in Figures 5-10. The distributions appeared relatively normal in these cases.

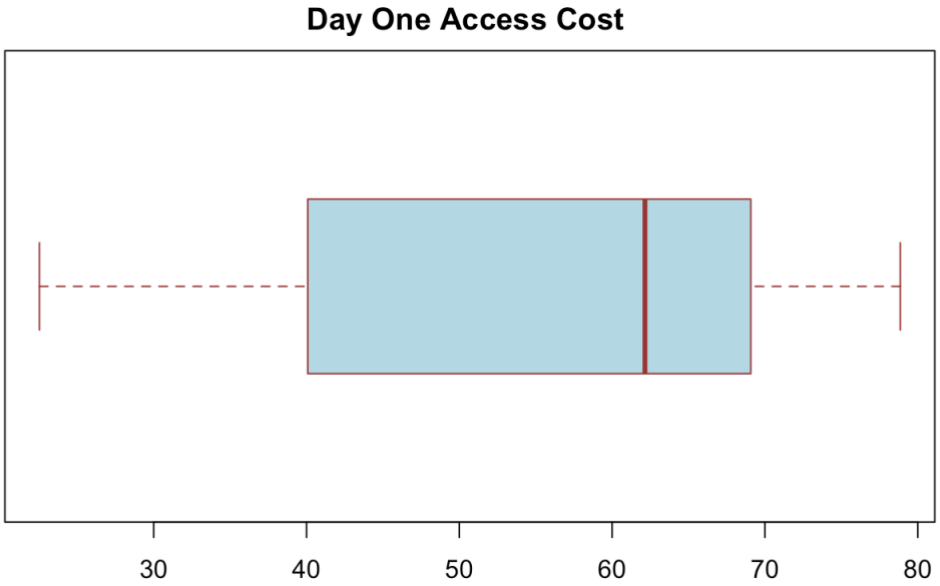

Figure 5: Day One Access Cost

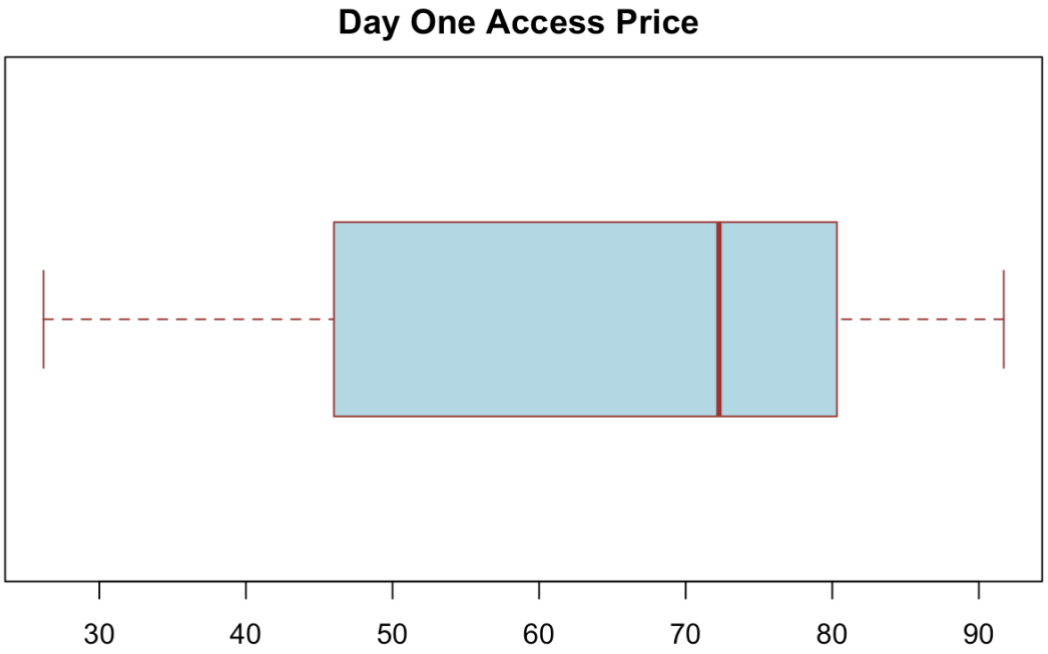

Figure 6: Day One Access Price

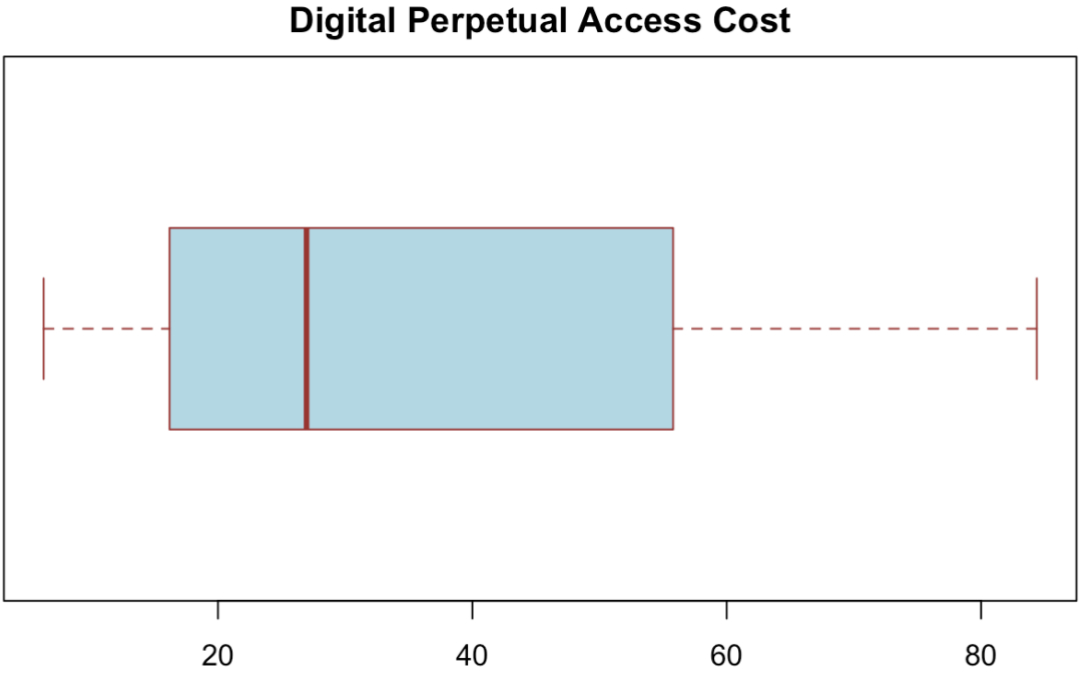

Figure 7: Digital Perpetual Access Cost

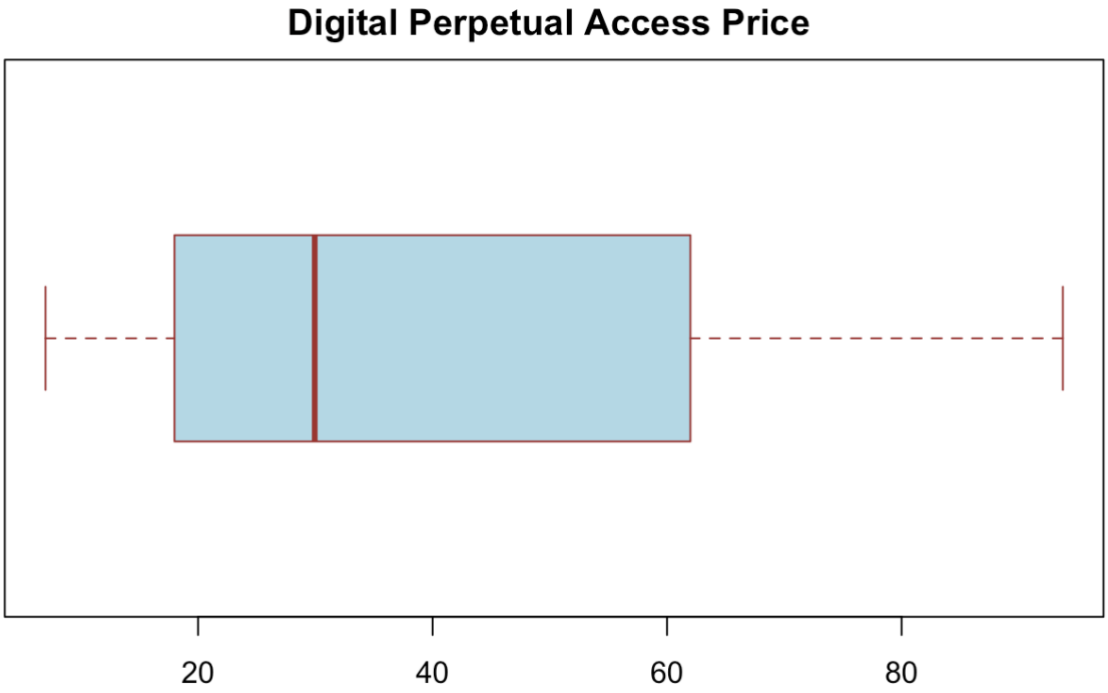

Figure 8: Digital Perpetual Access Price

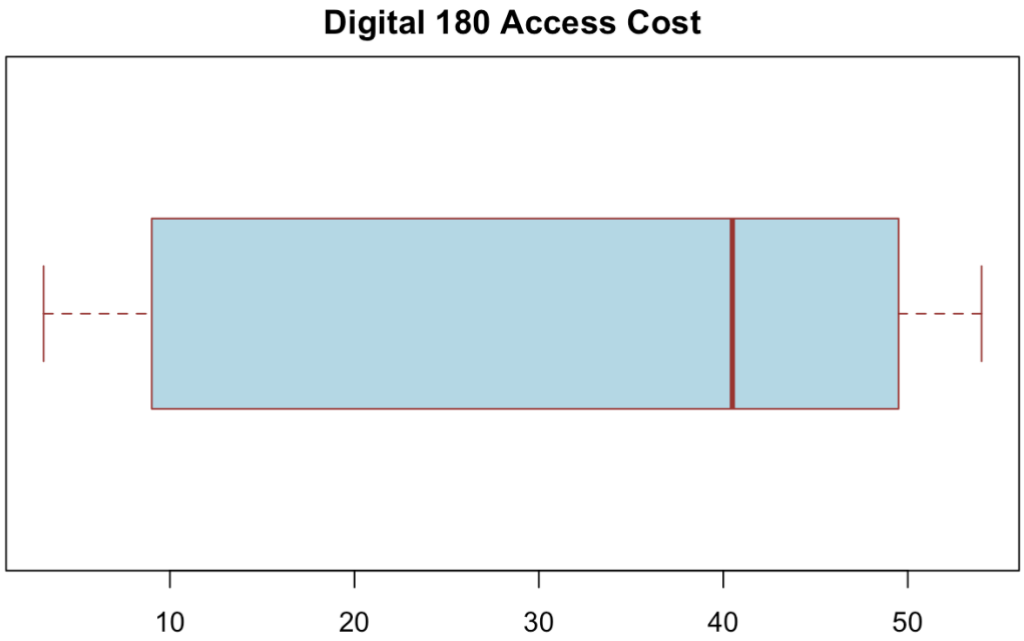

Figure 9 : Digital 180 Access Cost

Figure 10 : Digital 180 Access Price

Variable Manipulation

The bookstore data we received contained a large number of variables listed under the ‘object’ type. This datatype ‘object’ is basically an umbrella term for columns containing a mix of values with different data types or that cannot be easily classified into any other data type. The object is

difficult to work with in analysis so we decided to change the variable types for relevant columns to more manageable datatypes. For all binary categorical variables (Y/N, 1 or 2) we converted the datatype to ‘boolean’. This operation affected our columns containing data regarding students who are degree seeking or not, students residency status, and weather or not the student is listed as an undergraduate. These variables being converted to boolean expressions will make more advanced analysis techniques of more use to our group. Additionally, several variables were converted to categorical variables which are not useful in regression analysis but can be more easily managed. Columns for students’ primary major, section code, and course title, were among the affected variables. Going forward, when we have our methods for testing further thought out, the team may decide to re-code some of the columns listed in the section above as we find fitting for our analysis.

Feature Engineering

To streamline the dataset and facilitate a more comprehensive analysis of book prices, we performed feature engineering by creating new columns that capture the minimum costs and prices for both print and digital options.

We first identified the minimum cost and price for digital book options, including day one digital access, digital perpetual access, and digital 180-day access. The resulting minimum digital cost and price values were stored in two new columns: one capturing the minimal digital cost and the other the minimal digital price. Additionally, we created columns to store the source of the minimum digital cost and price values, indicating which of the original columns the minimum value was taken from. For instance, values in a source cost column labeled "digital_180_day_access_cost" indicate that the minimum cost value for that observation was taken from the 180-day access cost feature.

Since we needed to consider print book options when digital options were unavailable, we filled any null values in the minimal digital cost and price columns with the corresponding values from the new print cost and price columns. We also updated the minimum cost and price source columns to reflect the changes.

Next, we renamed the columns for digital cost and price as 'min_cost' and 'min_price', to store the minimum cost and price values for each book, regardless of whether it was a digital or print option. We also renamed the corresponding columns for the minimum cost and price source. After creating these new columns, we dropped the temporary columns that were no longer needed.

It is important to note that 98,677 cost and 36,839 price observations for print were cheaper than their digital counterparts, with an average difference of $11.23 for the former and $9.11 for the latter. However, we opted to capture the minimum values for digital over print when print was cheaper because the CU Bookstore explicitly stated their preference for digital options over print.

Finally, we removed any rows with all missing values for the specified price-related, either digital or print, columns to clean up the dataset further.

Data Partitioning

To facilitate our analysis, we partitioned the dataset into separate sets for each academic year. We began by filtering our dataset to include only books marked as "required." Next, we created four separate dataframes, one for each academic year: 2019-2020, 2020-2021, 2021-2022, and 2022-2023. We also created a fifth dataset incorporating all academic years for required purchases.

The new datasets will allow us to analyze the total cost of required book purchases, the number of books purchased, and the average price students paid for books per academic year. Additionally, we can perform further partitioning by semester to see if certain semesters have higher total book costs or prices.

This partitioning will enable us to gain insights into book purchase trends, identify any patterns or outliers by the year, and make data-driven recommendations to the CU Bookstore.

Conclusion

In this report, we outlined our four-step process for cleaning the the data provided to us by the CU Bookstore. In that four step process, we removed duplicate values, irrelevant columns, and rows that were not required, and addressed missing values by imputing the means of the column in which the missing value was found. We also handled outliers by using the Pandas report to identify any inconsistencies that required further cleaning. Now that we have a cleaned and formatted dataset, modeling and predicting future outcomes or possibilities should be much more manageable. Going forward, the group may decide to change the format of various columns depending on the analysis in question. All new changes will be detailed in later reports.