# Data staging steps, encountered issues with data quality and how we handled them

## Updates to the Dimensional Model

We realized there was an issue with our dimensional model after trying to load our data into the data mart (as discussed later): we set up the *Branch Attendance* Outrigger Dimension to contain a branch's attendance for the year a particular review was written. We could not pull any date information because the outrigger dimension was only linked to the *Branch* Dimension through a foreign key reference to the outrigger dimension by the *Branch* Dimension. To resolve this problem, we replaced the outrigger dimension's single *Number of Attendees* attribute with all the yearly park attendances from 2006 to 2021.

## Extraction

During the extraction process, we had some issues loading the datasets properly. We later identified that the issue was that we had to load each dataset with its proper encoding in mind. For example, we initially loaded all three of our datasets with *latin-1* encoding, but none of them were actually encoded in *latin-1*. The Disneyland Reviews and Universal Studios Reviews datasets were downloaded from Kaggle, and we compiled the Park Attendance dataset ourselves (so we already knew it was encoded in UTF-8). We identified the encoding of the Disneyland Reviews dataset based on the information provided on Kaggle, and we identified the encoding of the Universal Studios Reviews dataset using an online encoding detector.

The Park Attendance data came from PDFs that contained attendance figures for the world's most popular parks for a given year. We manually created a dataset containing the desired parks' yearly attendance for this analysis. No data was recorded from 2006 to 2009 for Universal Studios Singapore because the park opened in 2010.

## Transformation

### Data cleaning

Both datasets contain written reviews by customers. We used Neural Language Processing (NLP) to clean them and ensure consistency by:

- stripping whitespaces and extra blank spaces,
- normalizing reviews (remove capitalization and punctuation),
- removing Unicode and special characters,
- removing stop words,
- stemming each word in the reviews (replace all words with their root word).

For both datasets, we checked for and dropped data with undefined (null or missing) *Month* or *Year* values from their dates; we found no way to determine these missing values because this information is not publicly available. We decided to not consider data without Month or Year values since they are necessary to calculate our measures.

Ratings in the Universal Studios Reviews dataset were converted from floats to integers (to be consistent with the Disneyland Reviews dataset and because all ratings are supposed to be integers between 1 and 5).

We also removed all entire rows from both datasets that were duplicated (we kept only one of the duplicated rows and dropped the other). Rows with the same Review_Text values (but different date values – *Month* or *Year* – or different *Branch* values) were dropped.

After cleaning, the Disneyland Reviews dataset contained rows with the following *Review_Text* values: “activex vt error.” We dropped these rows because these are not actual reviews but noisy data in the dataset.

## Feature engineering

To display a consistent date format across the datasets, we separated each dataset’s respective columns pertaining to the date (*Year_Month* for the Disneyland Reviews and *written_date* for the Universal Studios Reviews) into separate *Month* and *Year* attributes.

Note that in the final table we loaded into the data mart, there are still null values; null values only appear for the *2006 Attendance (millions)*, *2007 Attendance (millions)*, *2008 Attendance (millions)* and *2009 Attendance (millions)* attributes of our *Branch Attendance* Outrigger Dimension. As stated earlier, no attendance data was recorded from 2006 to 2009 for Universal Studios Singapore because the park did not open until 2010. These null values indicate rows in the table that represent Universal Studios Singapore reviews. Keeping these null values should not have a negative impact on the analysis we conduct on our data later.

We also created a *Quarter* column in both datasets based on the values from the *Month* attribute, and generated measures for all rows in each dataset (according to our dimensional model) without issues.

## Data integration

We dropped all the distinct columns from the Disneyland and Universal Studios Reviews datasets because there was no way to generate such data and ensure that the datasets could be easily integrated. We could not scale up each dataset to conform to one another because the information required to do so is not publicly available. We dropped the *Review_ID* and *Reviewer_Location* columns from the Disneyland Reviews dataset since they are not found in the Universal Studios Reviews dataset. We removed the *reviewer* and *title* columns from the Universal Studios Reviews dataset. We also did not keep the *day* information from this dataset's *written_date* column because the Disneyland Reviews dataset only provides the month and year (of the review).

Before the Disneyland Reviews and Universal Studios Reviews were integrated together, all their column names were renamed to conform to the same format; both datasets had the same set of columns upon integration.

We then integrated the Disneyland Reviews and Universal Studios Reviews datasets together using an outer join on all columns in each dataset before integrating the Park Attendance Dataset with the former two datasets (using an outer join on the *Branch* column).

## Load

We encountered issues converting the data for our data mart to a .csv file. The issue was that we were not correctly extracting the source datasets before integrating them (the encoding issue mentioned in the Extraction process that we fixed).

After researching how to use pgAdmin 4, we created the fact table, surrogate keys and dimensions using PostgreSQL queries with no issues. During this step, we realized our problem with the outrigger dimension in the dimensional model (as discussed earlier) and made changes accordingly (we repeated our ETL process as described above after making the necessary changes to our dimensional model).