

metadata QUALITY

ASSIGNMENT THREE



November 7, 2020

jAMIE LU

W0441213

Table of Contents

[Introduction: 2](#_Toc55539232)

[TASK TWO: 3](#_Toc55539233)

[DATA LINEAGE 3](#_Toc55539234)

[DATA SET 1: Netflix Movies and TV Shows 4](#_Toc55539235)

[DATA SET 2: Airline Crash Data Since 1908 5](#_Toc55539236)

[DATA SET 3: 5](#_Toc55539237)

[TASK THREE: 5](#_Toc55539238)

[DATA QUALITY PLAN FOR 5](#_Toc55539239)

[References 6](#_Toc55539240)

# Introduction:

It is very important to maintain a high quality of data. This will ensure the usability of data and its relevance to all the data users. Having a high quality of data will create better decisions for a business or organization because they are able to create more educated ones based on the data presented. They will have more confidence in their decisions as high-quality data reduces risks and can produce consistent improvements in the results.[[1]](#footnote-1)

Data lineage is very helpful in ensuring high quality of data. This is because the user will be able to understand the source of the data and the transformations that it went through.

Data quality plan is helpful as well as it will help ensure that high data quality is established and maintained.

This document contains assignment three, task two and three. Task two showcases the knowledge learned regarding data lineage. It shows 3 data lineages created from 3 separate data sets. Task 3 showcases the knowledge learned regarding the creation of a data quality plan. It demonstrates the ability to create an effective data quality plan for a specific data set.

# TASK TWO:

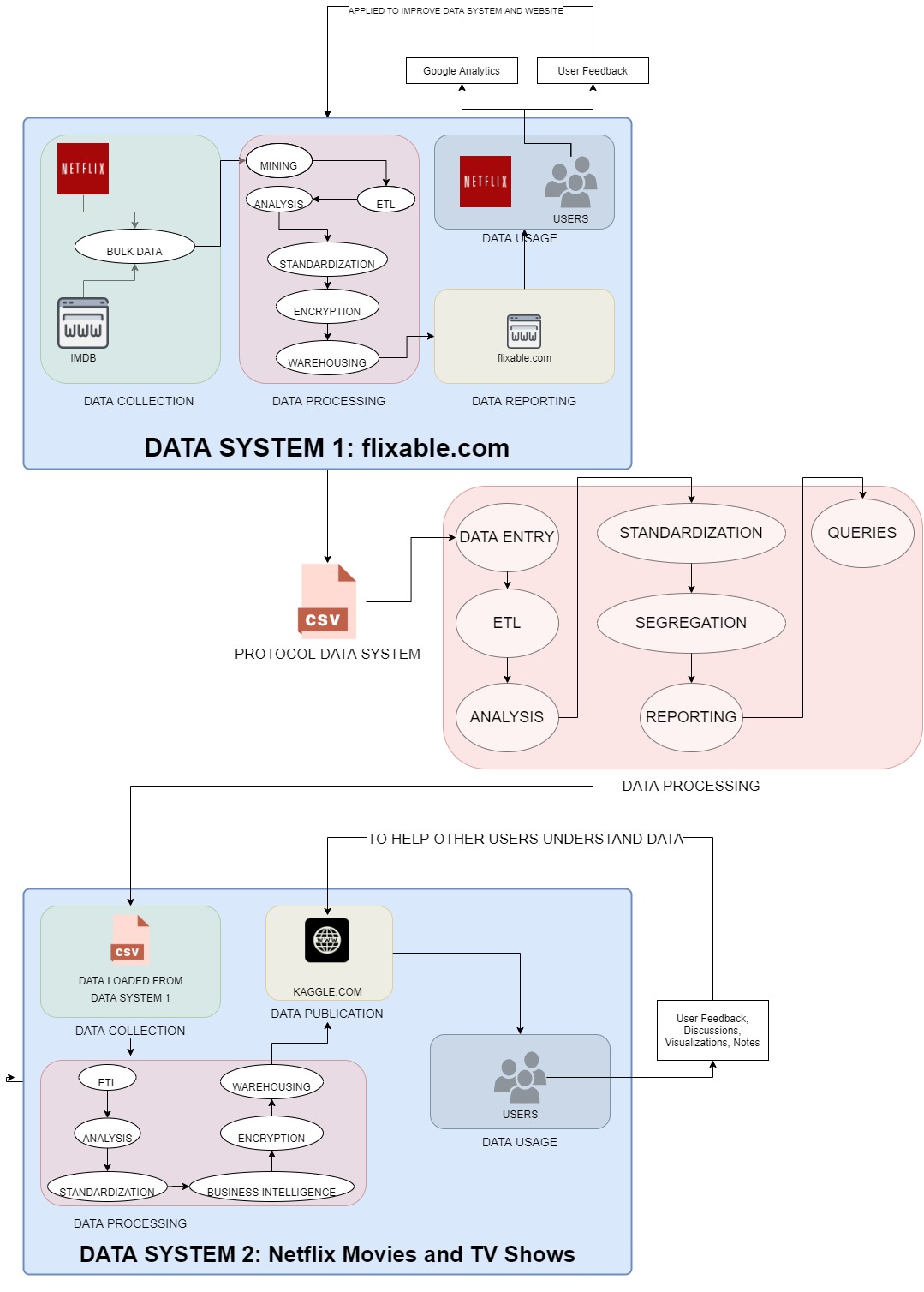
## DATA LINEAGE

Determining where data came from, how it was processed, and what processes went through throughout its lifecycle is essential. This is to help businesses or organizations create or improve their policies. It will help them understand data more and increase the quality of that data and its usability. This will also help make the business or organization flow better because it will make the flow of their data better. It will reduce the amount of errors or issues because it helps in finding where an issue is taking place or what is causing it because it is mapped out well.

#### DIAGRAM SHOWING USUAL LIFE CYCLE OF DATA

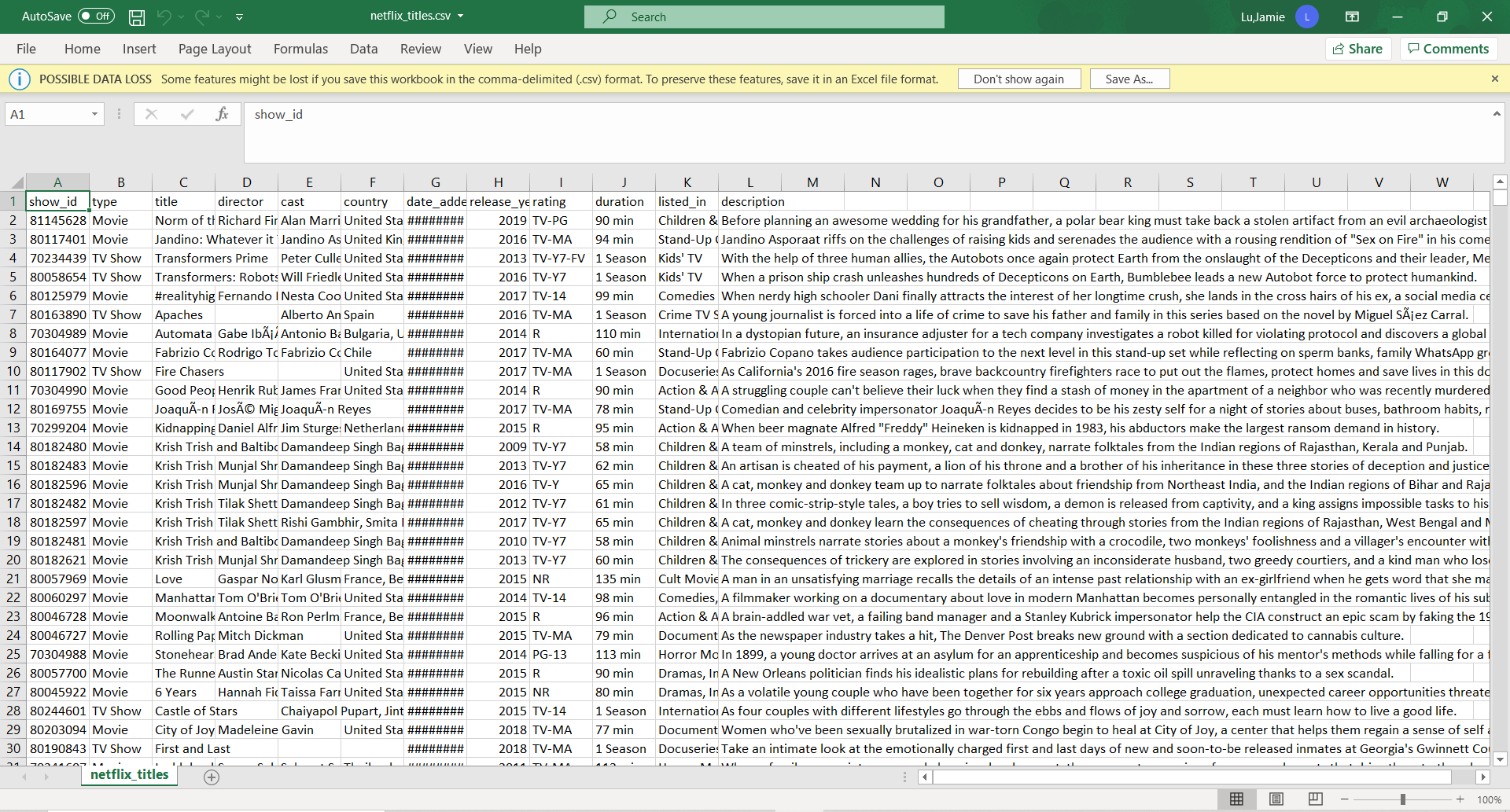
### DATA SET 1: Netflix Movies and TV Shows[[2]](#footnote-2)

Dataset is from Kaggle.com which was originally sourced from Flixable.com, a third-party Netflix (and other video streaming applications) search engine.



#### DATA LINEAGE FOR DATA SET – NETFLIX MOVIES AND TV SHOWS

(Placeholder only, please see separate **jpg file** named **TASK2\_DATASET1\_Netflix.jpg**)



#### SCREENSHOT OF DATA SET - NETFLIX MOVIES AND TV SHOWS

This data set contains information on the TV shows and movies on Netflix in 2019. The data was collected from Flixable, a third-party Netflix (and other video streaming applications) search engine. This data set shows which countries the TV show and movies are available in. This data set was created to help answer data users’ questions such as, what content is available in the different countries, or when combined with other datasets (such as IMBD rating data sets) show findings on what TV shows and movies are rated high or low.

### DATA SET 2: Airline Crash Data Since 1908[[3]](#footnote-3)

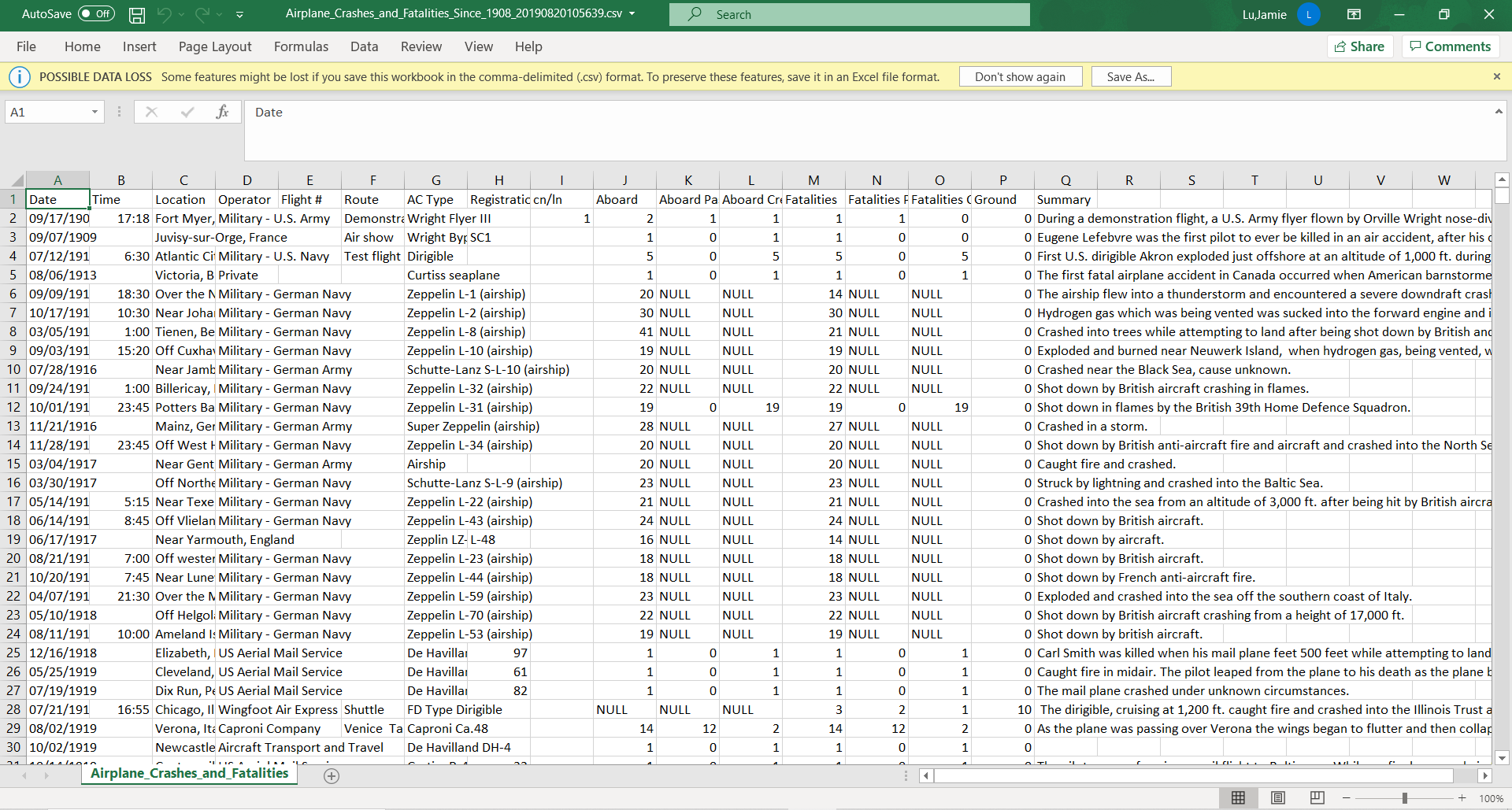
Data was sourced from website, planecrashinfo.com which has multiple sources from magazines, internet news agencies, articles, etc.

Diagram

Description automatically generated

#### DATA LINEAGE FOR DATA SET – AIRLINE CRASH DATA SINCE 1908

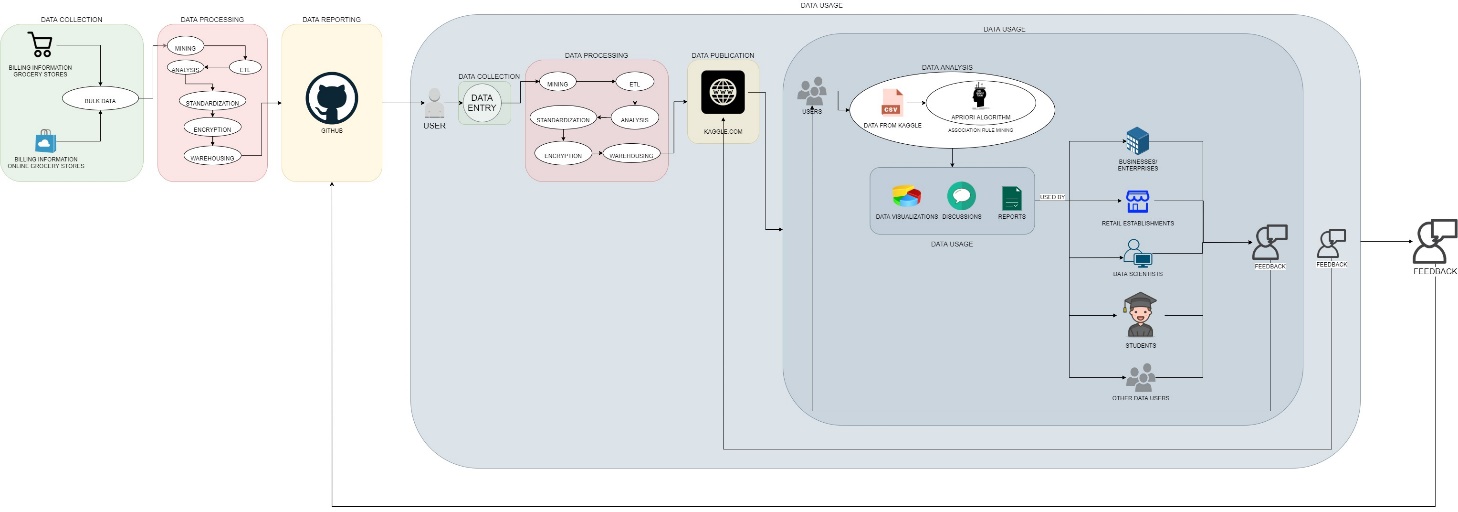
(Placeholder only, please see separate **jpg file** named **TASK2\_DATASET2\_Airplane.jpg**)



#### SCREENSHOT OF DATA SET – AIRLINE CRASH DATA SINCE 1908

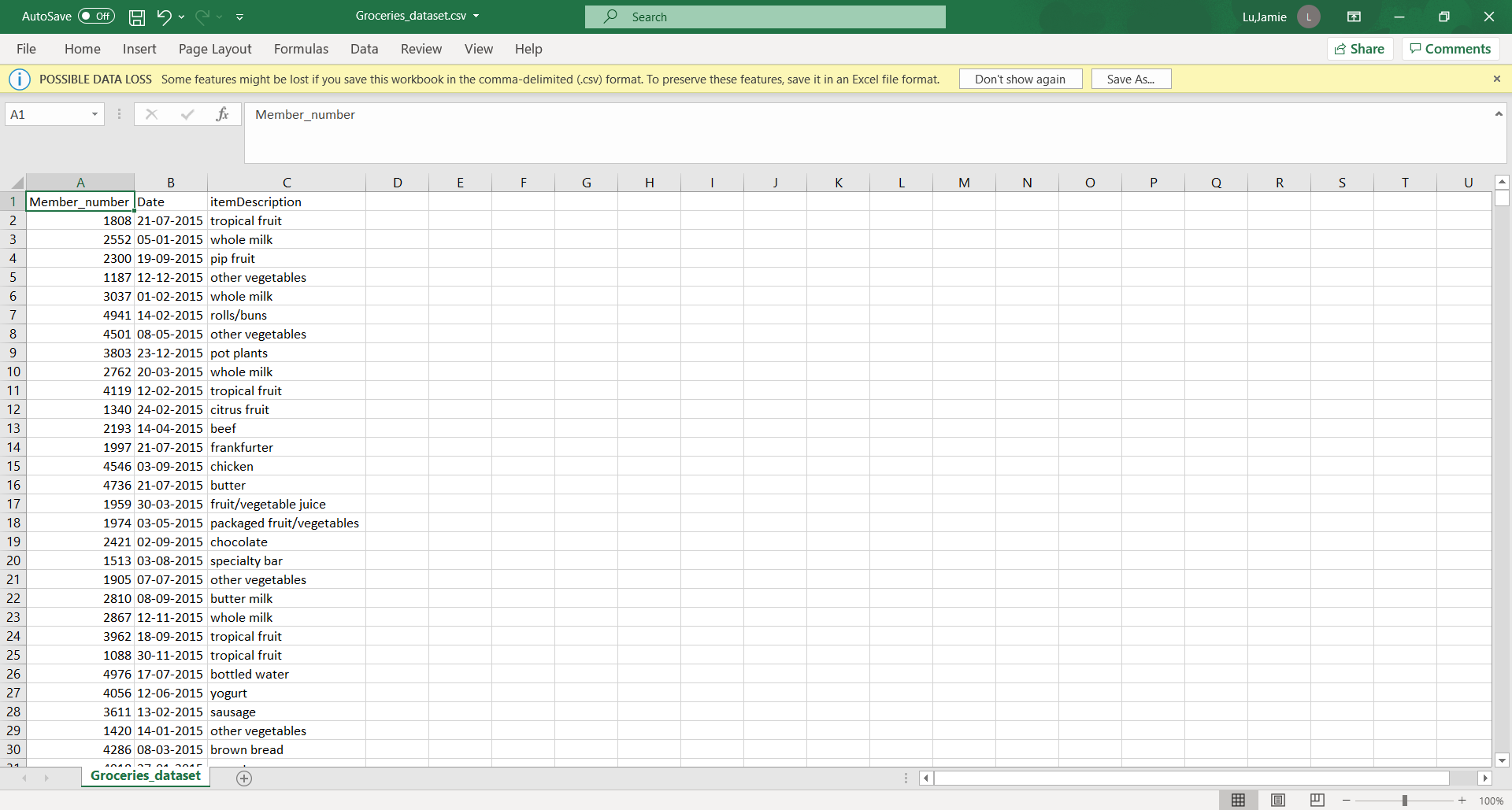
This data set contains information extracted from planecrashinfo.com and loaded into the csv in Kaggle to show the aviation accidents throughout the world from 1908-2019. This data set was made to determine the causes of plane crashes, which ones are the most common, and get insights like which are the worst operators and aircrafts.

### DATA SET 3: Groceries[[4]](#footnote-4)



#### DATA LINEAGE FOR GROCERIES DATA SET

(Placeholder only, please see separate **jpg file** named **TASK2\_DATASET3\_Groceries.jpg**)



#### SCREENSHOT OF DATA SET - GROCERIES DATA SET

This data set shows the purchase orders from multiple people from various grocery stores. This is suggested to go through Market Basket Analysis like the Apriori Algorithm. This can be used by retail stores to determine associations between products. For example, how often is milk bought when customers buy cereal.

# TASK THREE:

## DATA QUALITY

There are several factors that affect data quality. Some of them are accuracy, completeness, validity, relevancy, timeliness, and consistency (Lotame, 2019). To ensure that all these factors are considered, a data quality plan must be made for the business or organization.

### TRIP ADVISOR HOTEL REVIEWS – ABOUT THE DATA SET[[5]](#footnote-5)

This data set was created to be used for review rating prediction, topic modelling, and exploring what the key aspects are that define a good or bad hotel. The data used is from Kaggle.com and was collected from the travel company site, Tripadvisor.com.

### DATA QUALITY PLAN FOR TRIP ADVISOR HOTEL REVIEWS

**DEFINITION**

Business Goals

The **Trip Advisor Hotel Reviews** contains data from the hotel reviews of different hotels by the guests that have stayed there, or people who are curious about the hotel and decide to post questions.

The main goals of the data set are to maintain a high data quality and usability, to achieve this, several factors must be considered:

* Data must be accurate, such as the hotel information, even the spelling
* There must be a standard format or syntax for the values in the data set
* Data must be timely. Data must be recorded as close to when the person stayed at the hotel and must be continuously updated.
* Ensure that the non-nullable fields are filled and that all the fields are completed
* Determine and set what data needs to be collected. Ensure that all the data collected is relevant to the data set
* Ensure that the data is consistent. For example, if the data is being stored in multiple places, all of them should be consistent with each other.

**ASSESSMENT AND ANALYSIS**

Existing Data State

The values in the fields in the data set are complete. However, the data collected seems to be accurate but lacks fields, it only collects the actual reviews and the ratings by TripAdivsor.com’s users. There should be more fields added to the data set. It should collect more information, such as which hotel the user is talking about, the kind of room they stayed in, the dates they stayed at the hotel, where this hotel is located.

With regards to the standardization of the data, it seems like there is a standard format for the data in the data set. Trip advisor has a rating scale of 1-5 for reviews as well as a 200-20,000 character-standard for the actual review. There is no way to determine the timeliness of the data since there are no dates collected in the data set on when each review was made. In conclusion, the existing data state is lacking, it is incomplete and very difficult to create visualizations and collect findings. It is very difficult and tedious to search for the data as the data lineage is incomplete. It also looks like the dataset has never been updated.

**IMPROVEMENT**

Improvement Plan

The gaps in the data set must be filled to achieve the business goals that are set.

*STEPS FOR IMPROVEMENT:*

1. Determine the business goals of the data set.
2. Establish a data quality team composed of data stewards.
3. Create a metadata strategy.
4. Create a data quality plan based off the goals set.
5. Decide what data needs to be collected.
6. Add the necessary fields that were missing from the existing data set to make it more complete.
7. Process the data.
   1. ETL. Extract the data from the existing data set, transfer and load them into the improved data set.
   2. Set the standards for formatting and syntax.
   3. Encrypt the data to secure data and ensure confidentiality
   4. Data Mapping
   5. Store the data in a data warehouse
8. Analyze the data.
9. Create an ERD and Data Dictionary.
10. Create a data lineage.
11. Test the data set. The best way to test the data set is to use it.
    1. Create visualizations, reports based on the data set
    2. Ensure that the data can be mapped out and that all users of the data can locate the information they need.

**IMPLEMENTATION AND CONTROL**

Implementation and control of the Plan

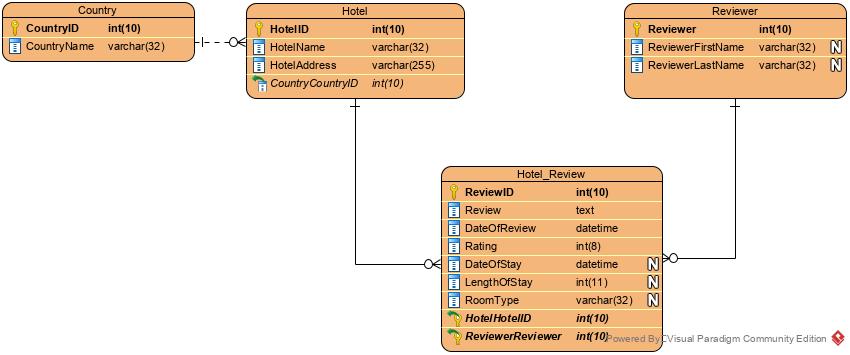
To start the implementation of the data quality plan, the data quality team must be established. They will be composed of the data stewards and data custodians. They are responsible following the steps in the improvement plan. They are also responsible for continuously checking in periodic intervals that the data is consistent with the business goals. They are also responsible for constantly updating the data set to improve it.

Feedback from data users should also be collected to grab an outside perspective from other people using the data. This will help in continuously improving the data set and its usability.

Errors are inevitable but the goal is to reduce them and to be able to troubleshoot them quickly and timely. This will be another responsibility of the data quality team. They will continuously check feedback from data users on any errors or issues they are encountering with the data set.

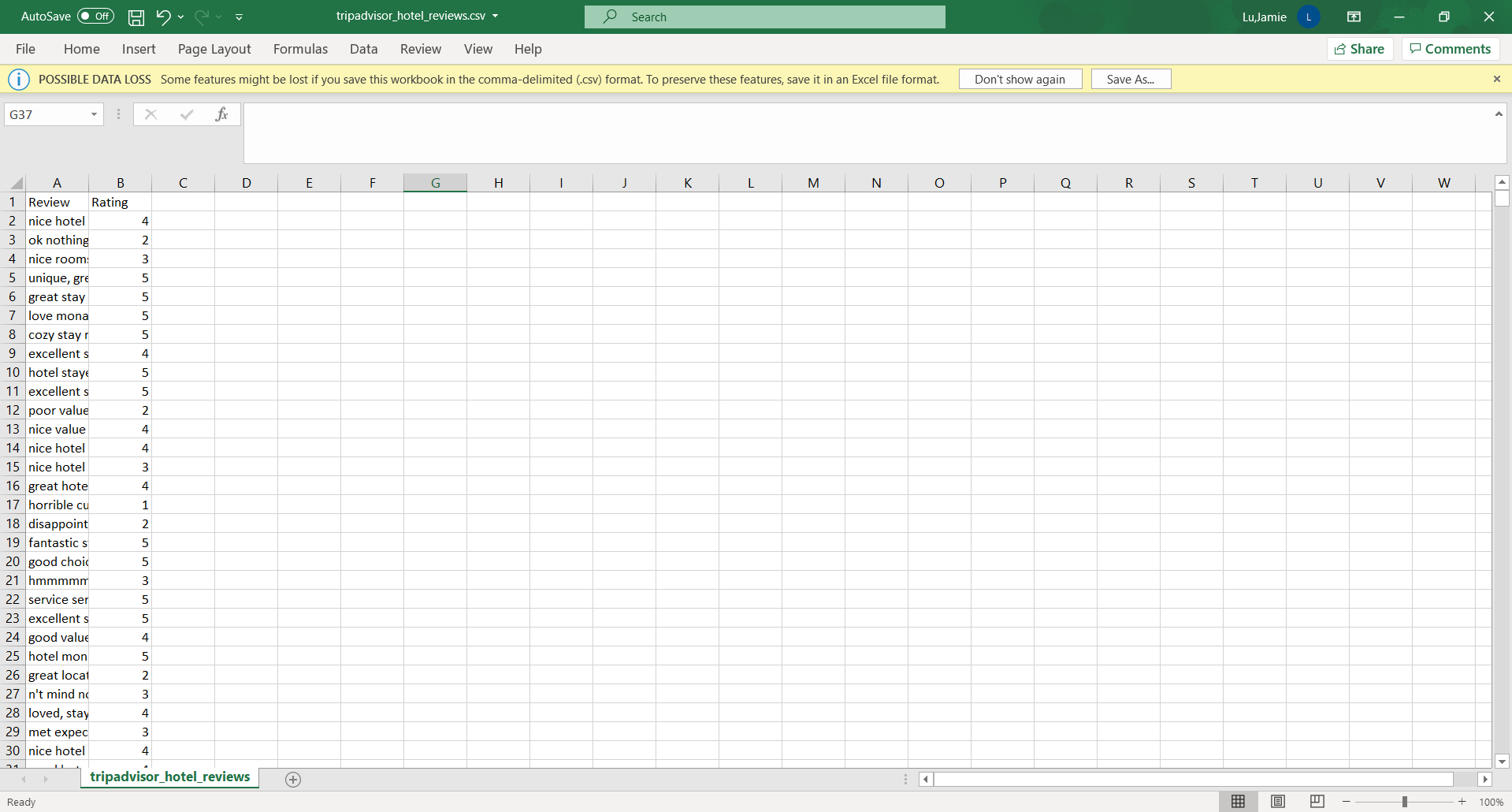
**ERD**

I have created a skeleton on what the ERD can possibly look like. This is to be reviewed and updated by the data quality team.



#### DIAGRAM OF SKELETON OF ERD

*(If unable to see clearly, please see separate file “****HotelReviews.jpg”****)*



#### SCREENSHOT OF TRIP ADVISOR HOTEL REVIEWS DATA SET

# References

Bansal, B. (2018, April 17). *TripAdvisor Hotel Review Dataset*. Retrieved November 07, 2020, from Zenodo: https://zenodo.org/record/1219899#.X6dLPWhKhPY

Bansal, S. (2019, December 04). *Netflix Movies and TV Shows*, 3. Retrieved November 05, 2020, from Kaggle: https://www.kaggle.com/shivamb/netflix-shows

Cem. (2019, August 20). *Airplane Crash Data Since 1908*, 1. Retrieved November 05, 2020, from Kaggle: https://www.kaggle.com/cgurkan/airplane-crash-data-since-1908

Dedhia, H. (2020, September 17). *Groceries Dataset*. Retrieved November 06, 2020, from Kaggle: https://www.kaggle.com/heeraldedhia/groceries-dataset

Global Data Store LLC. (2017, February 11). *Data Management - Data Quality*. Retrieved November 07, 2020, from Youtube: https://www.youtube.com/watch?v=kDOelMaTOuM&feature=youtu.be

Jean. (2019, April 30). *HOW TO TRACK AND VISUALIZE DATA LINEAGE*. Retrieved November 05, 2020, from Linkurious: https://linkurio.us/blog/how-to-track-and-visualize-data-lineage/

Khomov, A. (. (2020, February 02). *Netflix Content Analysis*. Retrieved November 04, 2020, from NYC Data Science Academy: https://nycdatascience.com/blog/student-works/netflix-content-analysis/

Larxel. (2020, September 28). *Trip Advisor Hotel Reviews*. Retrieved November 07, 2020, from Kaggle: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews

Lotame. (2019, April 30). *Why is Data Quality Important?* Retrieved November 06, 2020, from Lotame: https://www.lotame.com/why-is-data-quality-important/

Mercado, M. (2018, January 25). *This New Netflix Search Engine Lets You Get SUPER Specific About What Kind Of Movie You Want*. Retrieved November 04, 2020, from Bustle: https://www.bustle.com/p/flixable-a-netflix-search-engine-is-a-much-more-intuitive-way-to-find-a-movie-for-your-next-night-in-8021370

Moran, D. (2020, 26 October). *Data Lineage: Key Elements to Understand*. Retrieved November 04, 2020, from Canto: https://www.canto.com/blog/data-lineage/

Multiple Authors compiled - Science Direct. (n.d.). *Data Quality Requirement*. Retrieved November 07, 2020, from Science Direct: https://www.sciencedirect.com/topics/computer-science/data-quality-requirement

Parish, B. (2018, February 26). Understanding Data Lineage. Retrieved November 04, 2020, from https://www.youtube.com/watch?v=lUxgWb6WpF0&t=629s

PeerXPTeam. (2017, October 17). *The 6 Stages of Data Processing Cycle*. Retrieved November 04, 2020, from Medium: https://medium.com/@peerxp/the-6-stages-of-data-processing-cycle-3c2927c466ff

Ringlead. (n.d.). *How To Set Up An Effective Data Quality Plan*. Retrieved November 11, 2020, from Ringlead: https://www.ringlead.com/blog/how-to-set-up-an-effective-data-quality-plan/

Romadhoni, F. (n.d.). *What is Data Lifecycle Management? and What phases would it pass through?* Retrieved November 04, 2020, from Medium: https://medium.com/jagoanhosting/what-is-data-lifecycle-management-and-what-phases-would-it-pass-through-94dbd207ff54

Steenbeek, I. (n.d.). *The Basics of Data Lineage*. Retrieved November 05, 2020, from EW Solutions: https://www.ewsolutions.com/the-basics-of-data-lineage/

1. Referenced from (Lotame, 2019) [↑](#footnote-ref-1)
2. Referenced from (Bansal S. , 2019) [↑](#footnote-ref-2)
3. Referenced from (Cem, 2019) which states it was originally sourced from planecrashinfo.com [↑](#footnote-ref-3)
4. Referenced from (Dedhia, 2020) [↑](#footnote-ref-4)
5. Referenced from (Larxel, 2020) which states it was sourced from (Bansal B. , 2018) [↑](#footnote-ref-5)