

# Coursework Task Sheet 1

## Data Analytics Pipeline: Planning and Design

CMP5366: Data Management and Machine Learning Operations

---

Birmingham City University

Kacper Popis - 23161791 - [kacper.popis@mail.bcu.ac.uk](mailto:kacper.popis@mail.bcu.ac.uk)

# Contents Table

Contents Table .....	2
Step 1.1.....	3
Step 1.2.....	8
Step 2 .....	9
Data Ingestion .....	9
Data Pre-processing .....	9
Model Development .....	10
Model Deployment .....	10
Model Monitoring.....	10
Step 3 .....	11
References .....	12

## Step 1.1

One of the dataset collected (which I will reference as “**Dataset A**”) is a dataset that isn't used by the scientific report titled “Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method” [1]. The dataset that is used in the report isn't made public as it is data the researchers have collected themselves, but have outlined the time period that the data has come from (data collected from February 2006 to November 2016). As such, the only other option was to source / find a dataset that would contain data from the dates of February 2006 to November 2016. The dataset that will be used for this task will be a dataset found on [Kaggle](#) which covers data form the years 1980 up to 2020, but the data that will be used will be limited to years 2006 – 2016 (this leaves the dataset with 10481 rows out of the original 16598 – resulting in a 37% of rows being made redundant). The dataset itself only contains 11 columns:

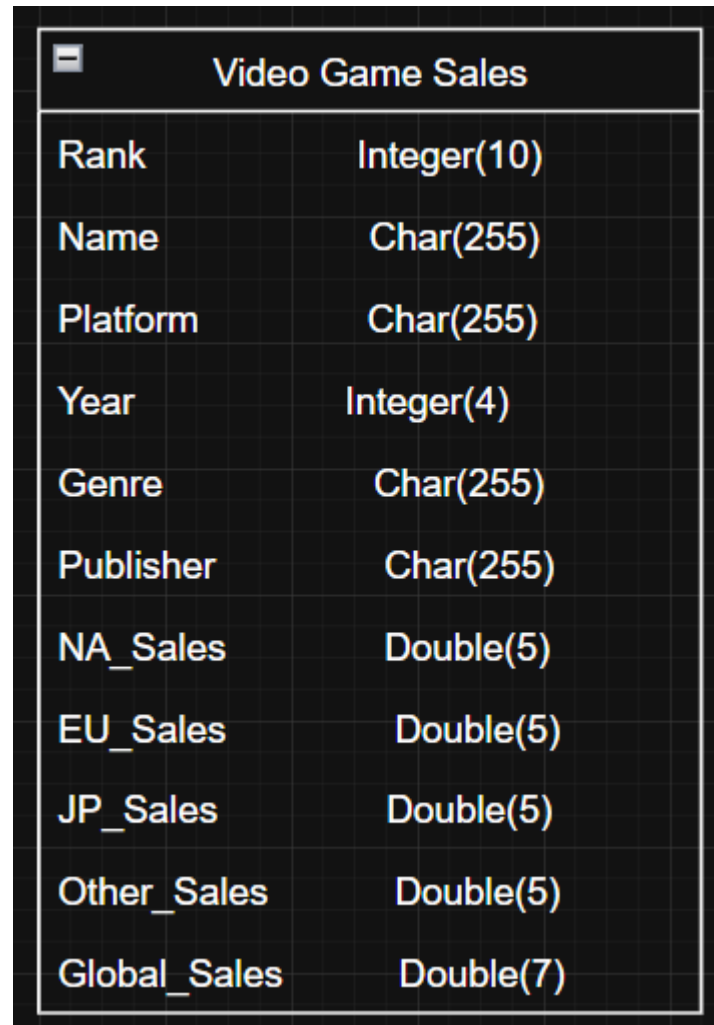
Column Name (Feature)	Description
Rank	Ranking of overall sales
Name	The games name
Platform	Platform of the games release (i.e. PC,PS4, etc.)
Year	Year of the game's release
Genre	Genre of the game
Publisher	Publisher of the game
NA_Sales	Sales in North America (in millions)
EU_Sales	Sales in Europe (in millions)
JP_Sales	Sales in Japan (in millions)
Other_Sales	Sales in the rest of the world (in millions)
Global_Sales	Total worldwide sales

*Table 1 - Description of features taken directly from the Kaggle page of **Dataset A***

It is noted on the page that the data collected for this dataset comes from a GitHub repository where a Python script is ran to collect data from a website that stores the data used in the dataset.

The purpose of this dataset is to show the difference in sales per world regions (the most consumer populated regions around the world) and to be used to compare which type of genre is the most popular in which region and to see overall which games have sold the most.

This dataset consists of only a singular CSV file containing nearly 17000 entries of data, as such the logical structure for this dataset would be a flat file. If this dataset was to be used then the data from the CSV file would have to be imported into the MariaDB engine to be used. This dataset is displayed in a logical diagram in *Figure 1* below.

A logical diagram of a database table titled "Video Game Sales". The diagram shows a table with 11 columns. The first four columns are categorical: Rank, Name, Platform, and Year. The remaining seven columns are numerical: Genre, Publisher, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, and Global\_Sales. Each column is followed by its data type and length in parentheses.

Video Game Sales	
Rank	Integer(10)
Name	Char(255)
Platform	Char(255)
Year	Integer(4)
Genre	Char(255)
Publisher	Char(255)
NA_Sales	Double(5)
EU_Sales	Double(5)
JP_Sales	Double(5)
Other_Sales	Double(5)
Global_Sales	Double(7)

*Figure 1 – Logical Diagram of **Dataset A***

If this dataset was to be used to develop a machine learning model, then the most likely approach for this dataset would be to use a regression model to predict values as majority of the data in this dataset is numerical, and the non-numerical data could then also be converted into numerical using encoding – using a classification approach wouldn't be suitable as there isn't features for classification. Below in *Figure 2* are outlined the features that would be used to train the model (shown in red) and the target variables (shown in green).

Video Game Sales	
Rank	Integer(10)
Name	Char(255)
Platform	Char(255)
Year	Integer(4)
Genre	Char(255)
Publisher	Char(255)
NA_Sales	Double(5)
EU_Sales	Double(5)
JP_Sales	Double(5)
Other_Sales	Double(5)
Global_Sales	Double(7)

*Figure 2 – Predictor and Target Variables*

Another dataset that has been collected (which I will reference as “**Dataset B**”) is the dataset used and is backing up the claims made in the scientific report titled “Global trends in dietary micronutrient supplies and estimated prevalence of inadequate intakes” [3]. The report makes the dataset used for it publicly available, which would aid in checking the accuracy of claims made in the report. The data from the dataset covers various micronutrient supplied to countries all over the world. The dataset contains 5 different CSV files (each containing different information) but all of the CSV files contain data linking each other. Below in *Table 2.1*, are some of the columns of the main file (S4) in the dataset with a brief description.

Column Name (Feature)	Description
Zone	Region the country is in
Country	Country name
ISO3	3 letter country code
Year	Year the data is from
Population	Population of that country
Fortification	Increase of micronutrient contents <b>[4]</b>
Tagname	Shortened name of the micronutrient
Micronutrient	Name of micronutrient
Units	Units of measurement for micronutrient
Estimated Intake	Estimated intake of micronutrients
Requirements	Required intake of micronutrients

*Table 2 – Description of features from **Dataset B***

The purpose of this data is to show which countries are lacking what specific micronutrients and how these challenges have been tackled over time (if they have been at all), with this trends can be monitored and can potentially predict the future outcomes.

The data in the dataset comes in a total of 5 CSV files ranging in file size, however, as previously mentioned the CSV files are linked with each other as some files contain keys / codes from another file resulting in a link from table to table. As such, the format for this set of data would be a relational database where all 5 files are treated as individual tables and are linked to each other (where appropriate). This is displayed visually in *Figure 3* below.

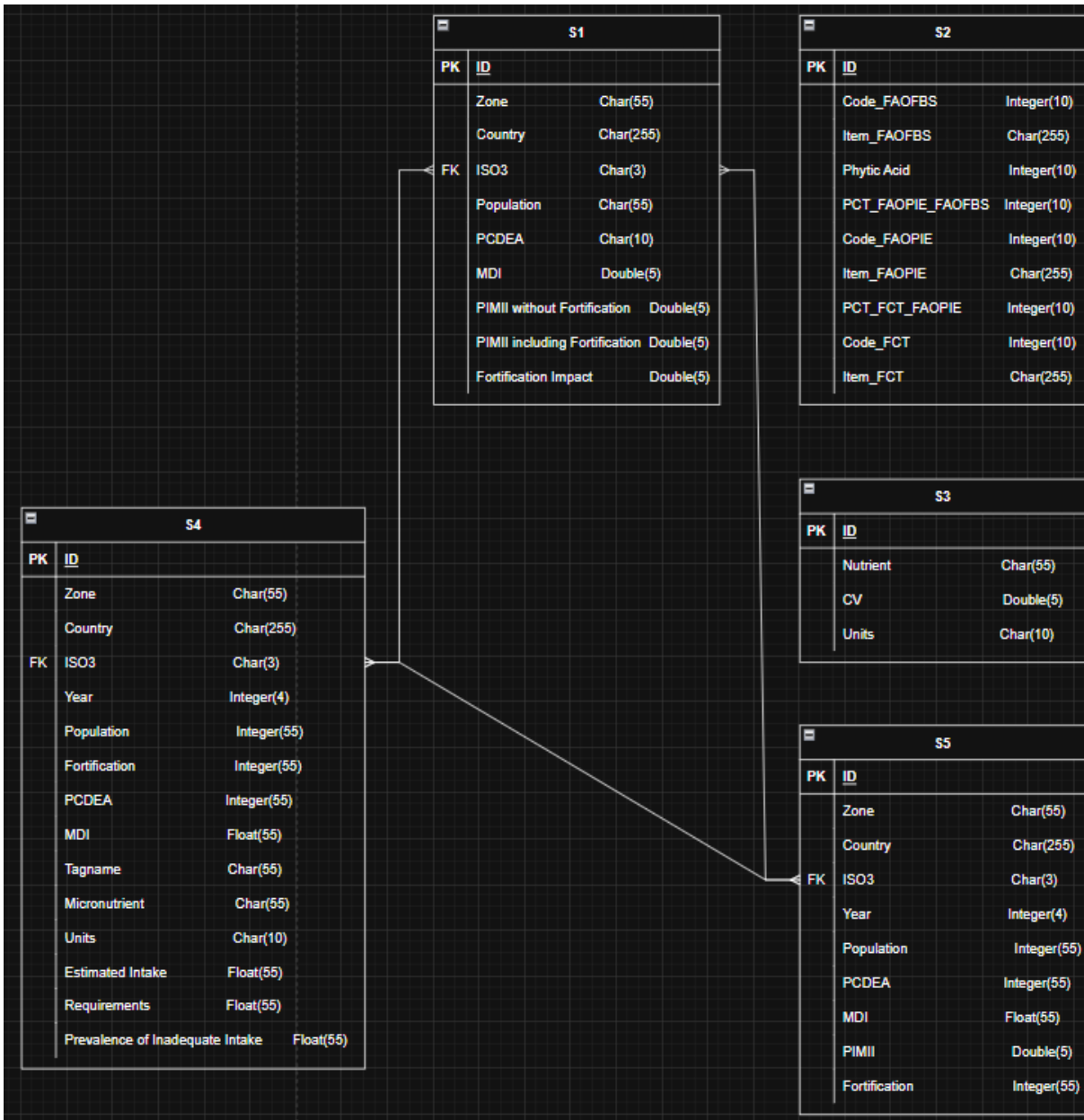


Figure 3 – ERD for **Dataset B**

If this dataset gets chosen for this task, then the main table that will be in charge of providing the training data would be “S4” as it is the table with the most features. For the training, data like “Year”, “Population”, “Micronutrient”, “Units” and “Estimated Intake” would be used to give the model data to predict with, meanwhile the target data would be “Requirements” and evaluating the disparity between the two using “Prevalence of Inadequate Intake”

## Step 1.2

For the task of creating an AI model, I believe that **Dataset B** would be the best choice as the dataset itself contains a lot of data (over 8 million rows per table) compared to **Dataset A**, also there are more tables than just 1 to work with, allowing for a more complex data structure and a more in-depth analysis of data for the model, which could increase the accuracy of prediction for the model. Another reason why this dataset is being chosen is because the scientific report using this dataset doesn't utilize an AI model on it but rather attempts to predict the future trends using visualization of graphs and the incline and growth of numbers present, this way the AI model can be compared to what the researchers have predicted. Whereas for **Dataset A** the data is most likely not the one used in the scientific report mentioned.



## Step 2

### Data Ingestion

The first step of the data pipeline will be data ingestion, where the data that will be used on the AI model will be imported from a source (in this case, the data of **Dataset B** is made available on the website figShare [3]) into a database (which will be hosted and managed using MariaDB). On the website there are 5 CSV files, and 5 TIF images, but the model will only use the CSV files.

One of the possible ways of accomplishing this would be to create a python script that utilize the site's API to download the 5 files – this method will always gather the most recent data (if there has been any changes made). Another way would be to download the data directly onto the machine that will be developing the model and directly access those files (in this case this is the most optimal method as the other method will increase the time to get to desired outcome, and use up more processing power)

Once the files have been retrieved, the data from these CSV files will then have to be imported into the actual SQL database for the model to be able to have data to be trained – additionally, this way the raw data can be queried (if there is a need) directly before any changes have been as this step is before the “Pre-processing” stage. The data from of these files do not follow a specific order, they are mostly unordered except some that follow an alphabetical order based on the country name (such as file S4 and file S5).

### Data Pre-processing

This stage will focus on eliminating any noise / invalid data inside of the data stored, as any anomalies inside of the dataset can cause the accuracy of the AI model to be drastically changed and make the prediction of values different from the “expected” results.

Another step in this process will be standardizing different types of data to allow for the model to utilise as much data as possible. As the model used on this dataset will most likely be a regression model, this model will only work with numerical values, as such any non-numerical data will have to be changed into a numerical form. This can be achieved by utilizing encoding techniques such as label encoding to categorize the data appropriately and correctly – although it may make reading the data slightly more complex for humans. To combat this, the data will be visualize using different graphs like bar charts, box plot graphs and scatter graphs.

## Model Development

The model in itself will most likely be a regression model and will utilize different regression techniques such as linear regression, decision tree regression, random forest regression and lasso regression (to name a few) and each of these regression techniques will use the same instance of data (the data used to train the AI model is randomly picked without a value assigned, but for testing there will be a set instance of data used to accurately compare the different techniques) to compare how accurately each of the regression methods predict the same value to find the most usable method (the method with the lowest Mean Square Error, Mean Absolute Error and R2 Squared).

## Model Deployment

Once the most accurate regression technique has been found for the model, the model can then be used and tested by users by providing their own input and receiving a prediction from the model based on unseen data.

The model will ask the user for inputs on specific data that it will process using all of the previously mentioned regression methods and once that data has been processed, an output is expected. The input that will be needed from the user is the same data that is used to train the model – for example, the user will need to input the “Year”, “Population”, “Micronutrient”, “Units” and “Estimated Intake” (if these are the features that will be used to train the model) and the output should be the “Required Intake” and then the estimated and required should be compared to show the disparity between the two calculations.

## Model Monitoring

Once the AI model has been deployed and is being utilized by users, then more data can potentially be added to the database to help develop the accuracy of the models (as it may alter the most accurate regression method) as the data stored only has data from 1961 up to 2011 – which as of current time is missing 14 years worth of data.

Every time new data is added all the model should undergo pre-processing to ensure the new data is not in a different format than the already stored data, and the regression techniques should be reassessed to check for changes in accuracy.

## Step 3

As previously mentioned, the data of **Dataset B** comes directly from the accompanying report [3] that has utilized the same data – but this time the data will be used on an AI model. The data itself will most likely be stored directly on the device for this task, as the data will most likely not be updated as the report has already published its findings and won't be updated therefore retrieving the data from a web service will be redundant and will only use up more processing power only to give the same output as storing the data locally.

Also previously mentioned, the data will be converted from a CSV file format and imported directly into SQL tables (each file will have its own table and all the links between the files – which are now tables – will have to be reinstated). Additionally, I believe that the most appropriate approach to extracting and loading the data into this database and model would be to use the ETL (Extract, Transform, Load) [5] approach rather than the ELT (Extract, Load, Transform) [5] approach as the data that will have to be transformed as a part of the “Pre-processing” stage therefore it would be most optimal to transform the raw data then load it directly into database which will save repeating steps like applying pre-processing and then either create new tables for the transformed data or overriding the current tables.

The best way to describe the logical storage structure for this database would be a star schema as there will be one main table currently titled “S4” with supporting tables [6] that link to the main table with foreign keys (like ISO3 or micronutrient ID's). This way, once the AI model has been deployed and new data is being added to the database, it will eliminate the amount of duplicate data being added – which will only slow down the model or potentially even reduce the accuracy of it.

## References

- [1]** - [ieeexplore.ieee.org](https://ieeexplore.ieee.org). (n.d.). Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method | IEEE Conference Publication | IEEE Xplore. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/9755343>. (Accessed: 14 May 2025)
- [2]** - Smith, G. (2016). Video Game Sales. [online] [www.kaggle.com](http://www.kaggle.com). Available at: <https://www.kaggle.com/datasets/gregorut/videogamesales>. (Accessed: 14 May 2025)
- [3]** - Beal, T., Massiot, E., Arsenault, J.E., Smith, M.R. and Hijmans, R.J. (2017). Global trends in dietary micronutrient supplies and estimated prevalence of inadequate intakes. figshare. [online] Available at: <https://doi.org/10.1371/u002Fjournal.pone.0175554>. (Accessed: 14 May 2025)
- [4]** - World Health Organization (2022). Food fortification. [online] [www.who.int](http://www.who.int). Available at: [https://www.who.int/health-topics/food-fortification#tab=tab\\_1](https://www.who.int/health-topics/food-fortification#tab=tab_1). (Accessed: 14 May 2025)
- [5]** - Bartley, K. (2020). ETL vs ELT: Key Differences, Side-by-Side Comparisons, & Use Cases. [online] Rivery. Available at: <https://rivery.io/blog/etl-vs-elt/>. (Accessed: 14 May 2025)
- [6]** - Databricks (n.d.). What is a star schema? [online] Databricks. Available at: <https://www.databricks.com/glossary/star-schema>. (Accessed: 14 May 2025)