Steam

The Gaming World

**Team Members:**

Zheng Qi

Oscar Hui

Chad Molenaar

Joe Quinn

Contents

[ETL Project Proposal 2](#_Toc66125220)

[Project Overview 2](#_Toc66125221)

[Extract - Proposed data sources: 3](#_Toc66125222)

[Transform - Proposed clean-up and analysis: 4](#_Toc66125223)

[Load - Data storage: 4](#_Toc66125224)

[Potential limitations: 5](#_Toc66125225)

[Steam Data Generation Project 6](#_Toc66125226)

[Executive Summary (ETL Process) 6](#_Toc66125227)

[Workflow Process – Initial ETL Overview 6](#_Toc66125228)

[Dependencies 7](#_Toc66125229)

[Part One Processes: - Genre’s and Feedback 8](#_Toc66125230)

[Extraction Process 8](#_Toc66125231)

[Transformation Process 8](#_Toc66125232)

[Loading Process 9](#_Toc66125233)

[Summary Output Diagrams (Part One) 10](#_Toc66125234)

[Part Two Processes: - Steam Game Price Index 12](#_Toc66125235)

[Extraction Process 12](#_Toc66125236)

[Transformation Process 12](#_Toc66125237)

[Loading Process 13](#_Toc66125238)

[Summary Output Diagrams (Part Two) 14](#_Toc66125239)

[ETL: - Process Summary of Parts One & Two 15](#_Toc66125240)

[Part One: - Genre’s and Feedback ETL Diagramatic Process 15](#_Toc66125241)

[Part One: - Steam Game Price Index ETL Diagramatic Process 16](#_Toc66125242)

[Conclusion 16](#_Toc66125243)

# ETL Project Proposal

## Project Overview

* What is your question of interest?

How can both gamers and game publishers access data throughSQL/MongoDBto better understand and/or for analyses for information such as:

* + Steam game sales?
  + Most successful publishers?
  + Most popular genres?
  + Size of online player base?
  + Regional and GDP per capita players.
* What is the expected outcome of your ETL project? (i.e., how will this data ETL project add value?) Think about how your ETL project will enrich the data and how this data might create value.

As a group, we are looking to allow both to have the information available to them that will help them analyse and extract data. Through the extraction, transformation and data storage processes used, anyone will be able to save resources such as time, money, and trusted information. An important value-add for both gamers and game developers.

This data will essentially be a summary of multiple different data sets to give an accurate picture of the game sales of Steam.

Game publishers can use this data to gain a better understanding of what the most popular genres are so they can choose to develop a game targeting a wider audience or aim at a more niche fanbase. The size of the online player base can help both gamers and publishers.

Gamers can use this data to look at the popularity of different publishers so they can get a better idea of how good the game will be before making a purchase. In addition to this, gamers can use this to help them decide on which game to buy. This is especially useful for users who want to be a part of an online community and play in larger online games.

## Extract - Proposed data sources:

* Where is the data located?
  + Data sources are for Part One are steamspy and steamstore.
  + For Part Two
* What are the data set formats?
  + Part One: - Json.
  + Part Two: - .csv file
* How will you get this data?
  + Combination of API and downloaded data from steamspy and steamstore, Kaggle, isthereanydeal and foreignexchangerates. The links of which are all located in the readme file.
* What are your assumptions about the data?
  + Potentially messy and may require pretty print. Utilising JSON can have its pitfalls as it comes out as a dictionary.
  + CSV files tend to hold a lot of duplicates, misspellings, and null values as examples**.**
  + We are additionally holding the assumption that it will hold the Regional and GDP data we are looking for; however, this will be identified with further investigation of other potential databases/sets.
* What does it cover?
  + We are expecting to locate common values or headers (such as geographical data or demographical data) that will enable us to answer the above questions through the cleansing and transformation process.
* What is it missing?
  + Accurate data on owners’ numbers or I.D.’s for games is an expectation of missing data.

## Transform - Proposed clean-up and analysis:

* What are the transformations you will apply to the data? (e.g. filtering, aggregation, derived columns)
  + Filtering, aggregation.
* What steps will you take to clean the data and ensure its validity? (e.g. messy data, duplicated data, incorrectly formatted data)
  + Drop the duplicated data. Convert the data to the right format.

* How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc)

* + Exploratory data analysis.
* How will the data be integrated? (e.g. joins, merges)

* + Mergers and joins.
* How will you apply these transformations (e.g. jupyter notebook, python script)
  + Pandas, Jupyter notebook, Json, all are listed under dependencies.
* IMPORTANT → Why did you apply these transformations? How did this enrich your data?
  + The data requested through API is quite messy and non-structured. To investigate the relationship between different factors, the data needs to be converted from non-structured data to structured data.

## Load - Data storage:

* What type of database (relational, document) will you store the data?
  + Originally we had planned on using an SQL database, however given the parameters and data found, we decided to use Mongo.
* Why did you choose this database over another database?
  + Structured data for easier comparison was the original thought for using SQL, however due to the use of Json, we thought that Mongo would be a more appropriate system to normalize the data we are storing with the query mechanism.
* What are your expected tables / documents and relationships between tables / documents in your database?
  + The database tables and documents should be reflective of the proposal brief. These will include the process of transforming the extracted data to ensure relationships such as income and geographical locations are included in both parts. In addition, these tables will also represent the required genre’s and popularity.

### Potential limitations:

* What are the potential limitations of your above proposed steps?
  + As the publishing time of different games is different, the aggregation results may be limited to a particular time.
* How can you control these potential issues?

* + We will not be considering the recent published game and specify the selection to a certain period.

# Steam Data Generation Project

## Executive Summary (ETL Process)

This is ETL project is focused on the gaming sales platform Steam. As per our title page, it really has become one of the major platforms for “the gaming world”. We have found the process of extracting, transforming, and loading data to allow others to search and visualise relevant data (accessibility) has many crossovers, as it is not as simple as just extracting, transforming and loading as individual processes (as shown in the workflow process). Through each step, there has been some form of general transformation and additional extraction.

## Workflow Process – Initial ETL Overview

As we can see below, the overall simplicity of an ETL process, looks straight forward. But as we continued the process after the initial adding of dependencies, some sort of transformation was needed immediately to obtain the related/relevant data. This inherently has allowed the group to take two differing approaches to maximise the raw data and ensure better transformation process’.

### Dependencies

|  |  |
| --- | --- |
| **Part One -** | **Part Two -** |
| **Extraction** | **Extraction** |
| import csv  import datetime as dt  import json  import os  import statistics  import time  # *third-party imports*  Import numpy as np  import pandas as pd  import requests | import pandas as pd  import requests |
| **Transform & Load** | **Transform & Load** |
| import csv  import json  import pandas as pd  import pymongo  import matplotlib.pyplot as plt  import scipy.stats as st | import numpy as np  import pandas as pd  import requests  import time  from matplotlib import pyplot as plt  from pprint import pprint  from iso3166 import countries  import pymongo |

**Table 1:** Dependencies for both Part One and Part Two shows two very differing approaches the group has taken.

The two differing approaches were taken to ensure that the Rubic (initial) was meet. Primarily utilising several different tools/programs and dependencies would give the group a greater understanding and effective working knowledge of the ETL process. Given the diffing dependencies, we can make a relative distinction between the two process’ that would enable the extraction and transformation processes.

## Part One Processes: - Genre’s and Feedback

### Extraction Process

# Customisations - ensure tables show all columns.

# Create the link for steamspy api.

# Create an empty list to store the first 10 pages of games.

# The games are stored in steamspy API by the order of popularity.

# Compile the data together to a new dataframe for better analysis.

# Export the dataframe to Data folder. This allows us to revert and analyse at any time.

# Create a function to continuously retrieve data from: -

<http://store.steampowered.com/api/appdetails/>.

By doing this, we allow all data to be updated through each iteration and ensure all data is collected.

# create a new list to iterate through each row of app\_list, confined by start and stop.

# Retrieve app data for a row, handled by supplied parser, and append to list.

# Write the data to a json file due to the unstructured nature.

# Examine the data file.

### Transformation Process

# Customisations - ensure tables show all required columns. Once this is achieved we can then move forward with the loading of the data to action further transformation of the data retrieved.

# Load the app\_list data extracted from steamspy API.

# Load the app\_data data extracted from steamspy API.

Data Merging

# Rename the app id column for merging. As one was different, without this process it would be difficult to merge.

# Merge the two dataframes based on the unique app id.

Data selection

# Select the games identified from the app list. This will allow us to further narrow down our data extracted to allow us to continue the transformation process.

Data filtration

# Drop the useless columns.

Data Cleaning

# Drop the duplicated games.

Data Splitting

# Drop the index columns. This allows us to aggregate more effectively in the next step.

Data aggregating

# We then aggregate the data of price/release data/recommendations together with previous dataframe.

# Drop the old unsplit columns. As we are moving forward, we can see a clearer picture and then we can rename the columns for easier reading.

Data filtering and operating

# Filtering the null value. For any null values, this will inherently alter the output of the data a skew the results we receive,

# Convert the price to readable value.

# Reset the index to allow for a more readable and useable database to be loaded.

### Loading Process

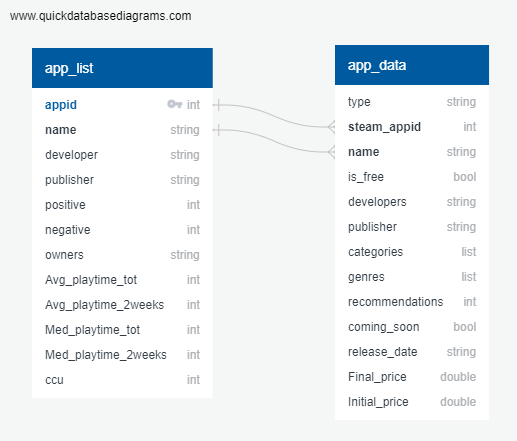
# Initialize PyMongo to work with MongoDBs.

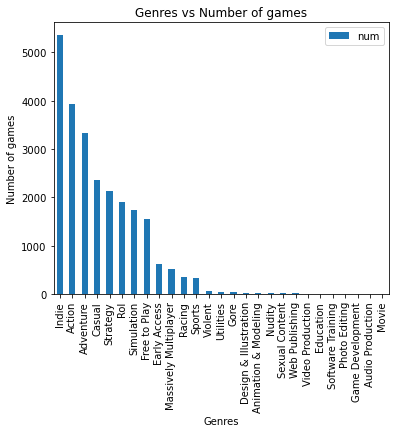
# Define the 'steam\_game\_database' database in Mongo.

# Drop the game collection if exists as we do not want any duplicates.

# Load the data frame to game collection in database.

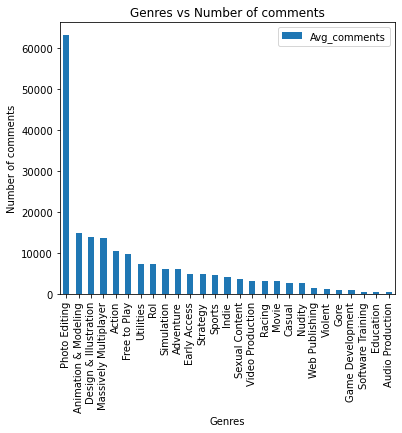
### Summary Output Diagrams (Part One)





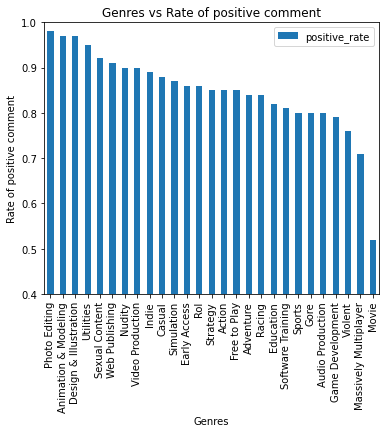
**Summary ERD of Genres Vs Number of games.**

The summary provides both gamers and developers a great tool to analyse. For a developer as an example – they may wish to make a new style of game in a popular (or unpopular) genre. For a gamer – they may wish to analyse which are the most popular and make decisions from there. It is a great example of adding value to both.



**Summary ERD of Genres Vs Number of comments.**

The content that garners the most comments, or the least and anywhere in between can be a useful tool, especially if they are positive. This not only will allow a developer to read the commentary for their analysis but allow them to sift through positive and negative ones of other developers, especially if it is the genre you are working with.



**Summary ERD of Genres Vs positive comments.**

As a developer, you can clearly see that the ERD shows the varying, yet surprising values attached to each genre. A particularly important added benefit for those targeting specific positive feedback for their business or for analysing competitors. In particular – if you joined a genre such as sexual content and nudity, it would show a clear trend for analysts.

## Part Two Processes: - Steam Game Price Index

### Extraction Process

# Read in the annual income by country from csv file.

# Then find the latest income year and filter the data. This was done primarily for two reasons. Firstly, to obtain the most current data, and secondly to test how up to date the gaming/developer community/industries are.

# Drop duplicate countries. This ensures no duplicates that will eventually affect the output.

### Transformation Process

# Transform the data from source to standardise USD. This allows a common currency to be used to represent a more equal standardisation across all regions.

# API call to convert all currencies to USD.

# Running first day/ last day of the year, then get the average (mean) exchange rate. This links in with the most recent data call made in the extraction process to ensure not just the accuracy- but an overall average is better than a specific point in time.

# Save the currencies name to columns for merging.

Data Merging

# Merge the currencies together with the view to merge the countries again in the aggregation section.

Data selection

# Get the region list from isthereanydeal API.

# As the European countries separate into 2 regions, we pulled the data and called EU1/ EU2.

We then had to add a column to identify which EU region.

# Append both EU1/EU2 together. This now will give a much cleaner dataset to continue to transform.

Data filtration

# Look up which EU region countries located in Annual income dataframe.

# API call to isthereanydeal and pull the top number of games. This can then be used to merge or analyse from the finalised Mongo upload.

# Convert json to dataframe for easier visualisation

Data Cleaning

# To assist with the cleaning process, we then used API call to look up game price by region using game's name.

# Add back the currency/ region/game to the dataframe as it is under different dictionary of json.

# Copy the dataframe to avoid another API call.

# Calculate the average game price by region using groupby. This again, will link in with the required game prices per region to be analysed if required by utilising both parts one and two in the database.

# Rename the country code from api to standard format as it has different naming.

Data aggregating

# Copy dataframe to avoid API call.

# Merging the dataframe with the average (mean) game price to the annual income dataframe.

# Rename dataframe columns for easier merging.

# Merging the average game price just for European countries and we saved as euro\_df. This now gives us a new dataframe to work from.

Data filtering and operating

# Drop Null values for non-European countries dataframe.

# Joining European & non-European countries result.

# Calculation of game price index by dividing the average game price with annual income.

# Export the annual income to bar graph.

# Export the game price index to bar graph.

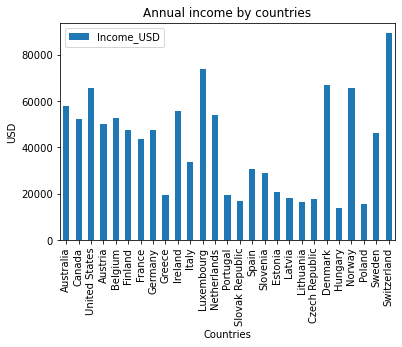
### Loading Process

# Initialize PyMongo to work with MongoDBs.

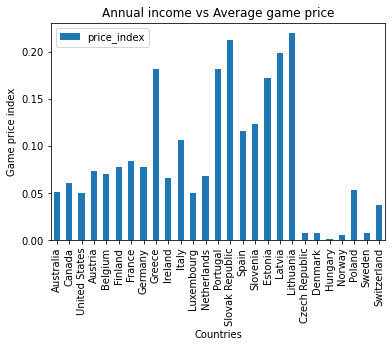
# Export result to mongo.

# Export result to csv.

### Summary Output Diagrams (Part Two)



**Summary Annual income by countries.**

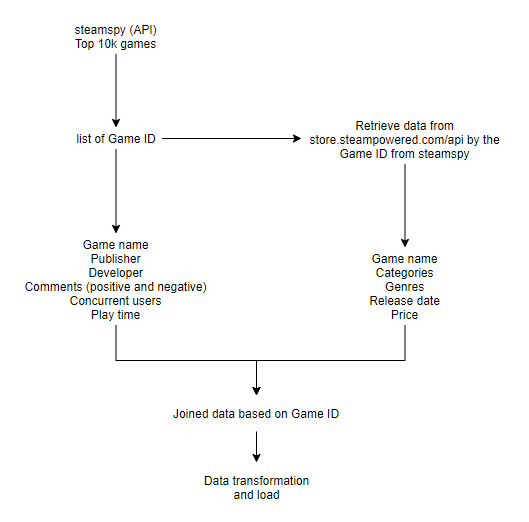


**Summary Annual income by Average game price.**

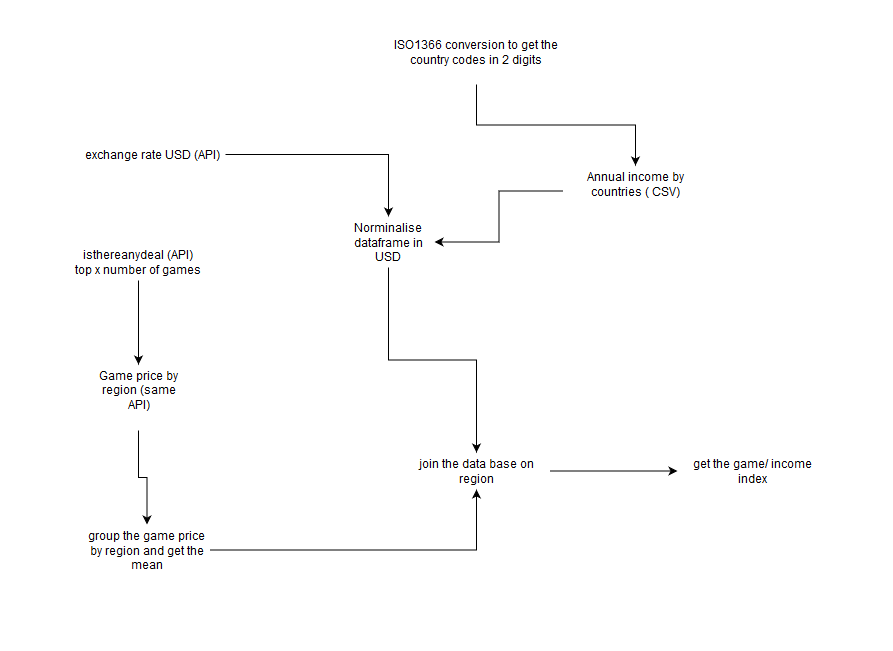
## ETL: - Process Summary of Parts One & Two

As already noted in the dependencies section, there are many differing ways to extract, transform and load data. This greatly depends upon the subject matter, complexities and the individual performing these tasks. Below are two perfect examples of each individual group member went about the project based upon the initial proposal.

### Part One: - Genre’s and Feedback ETL Diagramatic Process



### Part One: - Steam Game Price Index ETL Diagramatic Process



# Conclusion

As clearly shown, our group has made a step-by-step ETL utilising many standard and third-party dependencies. This is also shown throughout both Parts One and Two processes listing each individual step used to explain how and why the project produced the outcome it has. Once it is loaded into the MongoDB, any given gamer or developer can run their own analysis on Steam game sales, most successful publishers, most popular genres, size of online player base and regional and GDP per capita players.