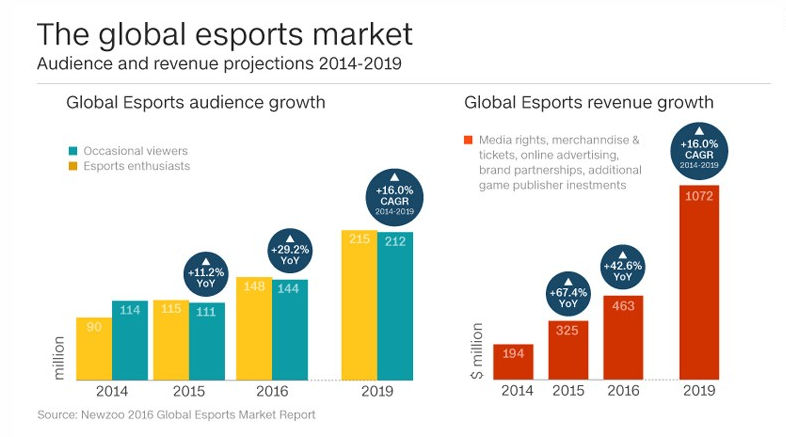
**ETL Project Report: eSports Revenues and Earnings**

Due: October 13, 2020

***What is eSports?***

eSports have been gaining exponential traction in the entertainment industry over the last few years. With approximately 380 million viewers worldwide, the genre easily rivals traditional sports entertainment.¹ Yet, while there is such enthusiasm around eSports, many are still in the dark on what this emerging subcategory of sports entertainment is. eSports refers to a category of virtual, competitive gameplay that functions much like traditional championship sports.¹ Players from all over the world are able to engage with one-another in competitive league-style gaming tournaments. Among some of the more popular game titles are League of Legends, Fortnite, Madden NFL, and Call of Duty.¹ The participation in eSports is growing and with it comes an increase in revenues from the sales, marketing, and promotion of the games and tournaments (Figure 1). These revenues lend to earnings won in-tournament by players making the sport a potentially lucrative endeavor for those that are able to put in the time and effort to practice and advance in championships.

*Figure 1. eSports Market Audience and Revenue Growth Projections (CNN.com, 2018)*



***Project Purpose and Scope***

The purpose of this project was to apply Extract, Transform, and Load (ETL) procedures to eSport and video games sales datasets to prepare them for an analysis of the revenues and earnings and of select, competitive games. From two separate data sources, information on eSport tournament earnings and global game sales revenues were gathered, cleaned, and loaded into a relational database as preparation for analysis. The analysis and report of findings of the data collected is not within the scope of this project.

***Extract Process***

Two data sources were identified for this project. The first dataset was sourced from Kaggle.com. This dataset contained the following: Name, Release Date, Genre, Total Earning, Online Earnings, Totally Player, and Total Tournaments for 492 unique games that are played competitively in eSport tournaments. This data was originally sourced from EsportEarnings.com, a site compiling the history, highlights, and status of eSport tournaments. This data was able to be downloaded from Kaggle.com as a CSV file and stored to a local drive.

Our second dataset was sourced from Data.world. This dataset contained the following: Rank, Name, Platform, Release Year, Genre, Publisher, North American Sales, European Union Sales, Japanese Sales, Other Sales, and Global Sales for 16,598 rows of games that are sold and played worldwide on different platforms. This data was able to be downloaded from Data.world as a CSV file and stored to a local drive.

These datasets were imported into Jupyter Notebooks using Pandas to read the CSV files and convert them to dataframes.

***Transformation Process***

The datasets used were of high quality and already very clean. Some additional cleaning, aggregating, and joining took place using Pandas and pgAdmin. The joining of the data will be described in the Loading PRocess section of this report. To prepare the data for analysis, the following steps were completed:

* An index was defined on both data sets to be used as a primary key when loaded into pgAdmin. This index was created using the .reset\_index() function. Using this function on the dataframes created a new column called “index” that was renamed to “id” and set as the index using the .set\_index() function.
* Extraneous columns were removed from the Data.world dataset so that only the index, Name, and Global Sales columns remained. The decision to keep the Kaggle.com dataset was made to retain the eSpport tournament information should it be of use for analysis.
* The Global Sales column in the Data.world dataset was reformatted to convert values from millions (i.e. 85.52 million) by multiplying the column by 10^6. This resulted in a numerical format that matched that of the Kaggle.com earnings format.
* The column names in both datasets were changed to match the formatting of pgAdmin tables.
* The values in the Data.world dataset were aggregated by the sum of all sales across gaming platforms and grouped by the game name. While troubleshooting the transformation and loading process, it was recognized that the Data.world dataset delineated sales by the platform on which the game was sold. This resulted in additional rows for each game title. By aggregating by the sum of the global sales, we were able to display the data with a single line for each game.

After the transformation process, the resulting dataframes were loaded into pgAdmin, a PostgreSQL program.

***Loading Process***

A relational database was chosen as the loading and consolidation of the datasets was anticipated to result in three tables, the consolidated table being fairly small. A connection from Jupyter Notebooks to pgAdmin was made using SQLAlchemy’s create\_engine function. The dataframes were then imported to pgAdmin using the .to\_sql() function.

In pgAdmin, two initial tables were created. One table was made to hold the eSport tournament earnings data (“eport”), and another was made to hold the game global sales data (“game”). Each table was designated an ID column to be used as a primary key. Subsequent columns were made to hold the data from the dataframes created in Pandas. The dataframes from Jupyter Notebooks were loaded to these tables as described above.

These tables were then merged in pgAdmin to combine the earnings and global sales data for games that represented in both datasets. This was accomplished using an inner join between the tables based on the game names. The data in the table was then ordered by the game names from the “esport” table. The resulting table was named “merged\_esport”. As a final step toward preparing for data analysis, Pandas was used read the “merged\_esport” table from pgAdmin using the connection made to the database. This table was transformed into a dataframe and the “id” column was then removed.

***Conclusion***

Through the Extract, Transform, and Load process described above, a final dataframe was created that contains eSport tournament and global sales data for select eSport games. This data can be used to analyze the benefits to players, sales to earnings ratios, and the overall volume of tournament cash flows. This information can also be expanded upon using similar processes to integrate additional data sources to this dataset.