Joshlynn Hilburn

Analyzing Nintendo Switch game sales
Data Cleaning Log

• Raw data scraped directly from VGChartz using ParseHub is organized into a Microsoft Excel document organized by genre with one sheet per genre, due to the way the source website is structured

> Genres include Action, Action-Adventure, Adventure, Board Game, Education, Fighting, Misc, MMO, Music, Party, Platform, Puzzle, Racing, Role-Playing, Sandbox, Shooter, Simulation, Sports, Strategy, and Visual Novel

• Sheets will eventually be combined into one large table in a .CSV

• Removed all blank ("junk") rows that only contain the line "Read the review" that appear occasionally throughout each sheet due to web scraping, using the sort function to find these rows

• Deleted boxart column from all sheets because it says "Boxart Missing" for all observations; boxart will not be included in the cleaned dataset

• Deleted console column because all games in this dataset are already for the Nintendo Switch and this field is the same throughout every observation; is redundant information

• Added header row for column names of all sheets corresponding to how the data was structured on VGChartz which the scraper did not collect

> Column names include position, name, publisher, developer, VGChartz score, critic score, user score, total shipped, total sales, NA sales, PAL sales, Japan sales, release date

> Units for all shipped and sales variables are in millions; units for score are 1 through 10

• Added column for genre to each sheet after developer column with information filled in for matching genres

• Exported each sheet in the Excel workbook to a .CSV for combination in RStudio (20 total .CSV)

• Used R functions to combine all individual .CSV files into a single data frame and check the data types of each column; code used is shown below:

```
library("dplyr")
library("plyr")
library("readr")
SwitchGames <- list.files(path="C:/Users/Jaime Luong/Downloads/games", pattern=".csv", full.names = TRUE) %> % lapply(read_csv) %> % bind_rows
str(SwitchGames)
```

• This resulted in a data frame called SwitchGames where data types were incorrect; corrections made are shown below along with the code to clean data and export as a cleaner .CSV

> position (num), name, publisher, developer, genre (chr) remained the same data type

> vgchartz_score, critic_score, user_score, total_shipped, total_sales, na_sales, pal_sales, japan_sales converted from chr to num

> release_date converted from chr to date with month and year parsed from release_date

```
library(lubridate)
SwitchGames <- SwitchGames %> % mutate(release_date = as_date(dmy(release_date, tz = NULL, format = NULL))) %> % mutate(month = month(release_date, label = TRUE)) %> % mutate(year = year(release_date)) %> % mutate(vgchartz_score = as.numeric(gsub("N/A","", vgchartz_score))) %> % mutate(critic_score = as.numeric(gsub("N/A","",critic_score))) %> % mutate(user_score = as.numeric(gsub("N/A","",user_score))) %> % mutate(total_shipped =
```

```
as.numeric(gsub("m","",total_shipped))) %> % mutate(total_sales = as.numeric(gsub("m","",total_sales))) %> % mutate(na_sales = as.numeric(gsub("m","",na_sales))) %> % mutate(pal_sales = as.numeric(gsub("m","",pal_sales))) %> % mutate(japan_sales = as.numeric(gsub("m","",japan_sales)))
str(SwitchGames)
write.csv(SwitchGames,"C:/Users/Jaime Luong/Downloads/games/SwitchGames.CSV", row.names = FALSE)
```

• Opened up new SwitchGames.CSV generated by R output in Excel to change the position column to one where each value was unique in order to create a primary key

> As the existing order of games is arbitrary, the numbers for the position column functioning as a primary key are numbered sequentially (1 through 1979)

• Fix typos in game, publisher, and developer names using spell check, find and replace when common errors were found, and manually once when characters did not appear

> Specific changes are described in the table below:

Old name	Corrected name
All instances of <U+0092> ; 7 times	Changed to an apostrophe (')
All instances of <e9> ; 12 replacements	Changed to e with an accent (é)
All instances of <U+0096> ; 5 times	Changed to a dash (-)
All instances of <f6> ; 1 replacement	Changed to o with an umlaut (ö)
All instances of <fc> ; 1 replacement	Changed to a u with an umlaut (ü)

• Created a pivot table using all the data to check for validity in data constraints for relevant variables

> All positions were 1 through 1979, one for each game, such that position functioned as a primary key

> All games were only 1 of 20 total genres

> vgchartz_score, critic_score, and user_score were all between 1 and 10

> total_shipped, total_sales, na_sales, pal_sales, and japan_sales were all reasonable values

> One game, Assault Android Cactus, is stated to have released in September 23rd, 2015 which is impossible because the Nintendo Switch came out in 2015; this was corrected directly in the sheet after looking up the correct release date (March 8th, 2019)

> month and year values were all within range (after correction above)

• Exported final cleaned dataset as SwitchGames.CSV