GameOn!

- Jesus Jimenez
- Shridhar Kamat
- Marina Mukhranskaia
- Shih-Hsien (Sherry) Ma
- Portia Pray
- Gagandeep Singh

Problem Overview

Gaming Sales (Console): Jesus and Gagan.

- Was there a increase in console gaming sales during covid?

Gaming and Mental Illness: Portia and Shridhar

- Does number of hours spent gaming affect mental health?
- Does gaming affect mental illness differently for male or female gamers?

Gaming Genre: Marina and Sherry

- Does votes affect the rating of games?
- What games/genres are people most commonly engaging with , and the trend of genre over time?

Data Challenges:

- Game, Sales, Genre Data
  - Majority of accurate data located behind paywalls or is published for media purposes thus, unable to extract.
  - Not enough time in overall project to merge sales data from various company in order to create a sample size large enough to properly represent all genres.
  - Each dataset we located had missing titles, genres, and regions thus, no 1 or 2 datasets were available that encompassed all the information we needed.

- Mental Data:
  - Scientific journal data requiring subscription or purchase.
  - Recent data was harder to find as psychological studies can last for many years before data is published. As a result the data itself can become outdated for the current period of time.

What data sources made the merging and sorting?

- Online Gaming Anxiety Data
  - Data collected as part of survey among gamers worldwide.
  - Questionnaire asked questions that psychologists generally ask people who are prone to anxiety, social phobia, and less to no life satisfaction.

- Video Games Sales for 2019 and 2020
  - Provides merged sales data from 2 different data sets.
  - Features columns with information such as, the Video Game Sales dataset and add new columns that highlight both critic and user scores and counts, develop name, and ESRB rating.

- IMDB Video Games
  - Purpose of the dataset is to gain insights into the trends of game genre popularity.
  - Provides game trends, plot, genre, and popularity.

CONSOLE GAMING SALES

What was our data process?

Step 1: Down the CSV files from our datasource websites and import each separate CSV file.

Step 2: Determine which columns we will keep for research, and create a dataframe for year 2019 and 2020.

Step 3: Group the data by genre and the sum of the sales for each year.

```
# !pip install matplotlib
# !pip install --upgrade numpy
# !conda install numpy
```

```
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as st
import numpy as np
from scipy.stats import linregress


file_path = "../InputData/vgchartz-6_23_2020.csv"
data2 = "../InputData/vgsales-12-4-2019.csv"
data3 = "../InputData/Video_Games.csv"


game_data_2020 = pd.read_csv(file_path)
game_data2019 = pd.read_csv(data2)
game_data3 = pd.read_csv(data3)
```

```
list(game_data_2020)
```

```
['Unnamed: 0',
 'img',
 'title',
 'console',
 'genre',
 'publisher',
 'developer',
 'vg_score',
 'critic_score',
 'user_score',
 'total_shipped',
 'total_sales',
 'na_sales',
 'jp_sales',
 'pal_sales',
 'other_sales',
 'release_date',
 'last_update']
```

```
clean_data_20 = game_data_2020.drop(['Unnamed: 0','img','last_update','release_date','console','publisher','developer'],axis=1)
clean_data_20
```

```
#groupping data for individual plotting with the sum of total sales

total_sales_20 = clean_data_20.groupby(["genre"])[["total_sales"]].sum()
na_sales_20 = clean_data_20.groupby(["genre"])[["na_sales"]].sum()
jp_sales_20 = clean_data_20.groupby(["genre"])[["jp_sales"]].sum()
pal_sales_20 = clean_data_20.groupby(["genre"])[["pal_sales"]].sum()
other_sales_20 = clean_data_20.groupby(["genre"])[["other_sales"]].sum()
total_sales_20
```

genre	total_sales
Action	1178.67
Action-Adventure	148.67
Adventure	341.21
Board Game	0.33
Education	0.09
Fighting	361.45
MMO	11.86
Misc	589.74
Music	52.56
Party	5.39
Platform	418.75
Puzzle	132.88
Racing	539.19
Role-Playing	483.08
Sandbox	1.89
Shooter	1047.95
Simulation	305.24
Sports	1221.48
Strategy	141.84
Visual Novel	5.78

Data process continued....

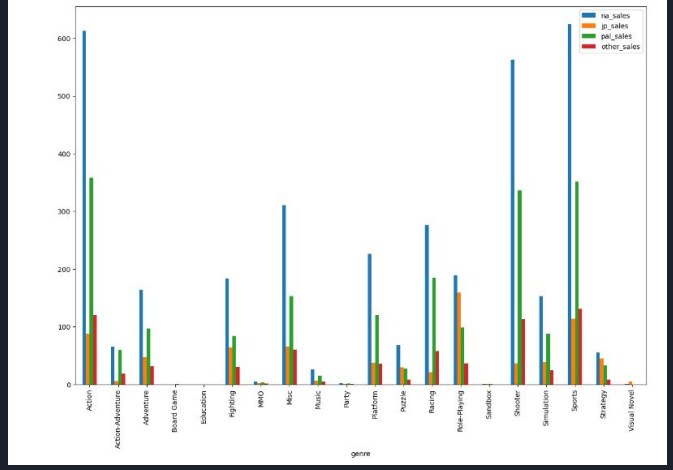
Out[7]:

genre	na_sales	jp_sales	pal_sales	other_sales	total_sales
Action	612.77	87.21	358.03	119.55	1178.67
Action-Adventure	65.11	5.46	59.63	18.40	148.67
Adventure	164.42	47.70	97.07	31.61	341.21
Board Game	0.06	0.04	0.22	0.02	0.33
Education	0.08	0.00	0.00	0.01	0.09
Fighting	183.36	63.29	84.40	30.33	361.45
MMO	4.88	2.36	3.42	1.19	11.86
Misc	310.42	64.93	153.19	60.48	589.74
Music	25.98	6.58	15.08	4.93	52.56
Party	2.35	0.95	1.61	0.47	5.39
Platform	226.30	36.86	119.54	35.74	418.75
Puzzle	67.57	29.76	27.10	7.97	132.88
Racing	276.71	20.68	184.14	57.41	539.19
Role-Playing	188.73	159.40	98.52	36.67	483.08
Sandbox	0.89	0.52	0.56	0.12	1.89
Shooter	562.55	36.04	336.09	113.09	1047.95
Simulation	153.01	39.30	87.80	24.48	305.24
Sports	624.50	113.55	351.50	131.31	1221.48
Strategy	54.98	44.73	33.37	8.70	141.84
Visual Novel	0.49	5.06	0.07	0.13	5.78

In [8]:
```
#outliers plot
```

In [9]:
```
data3.plot.bar(figsize = (15,10))
```

Out[9]: <Axes: xlabel='genre'>

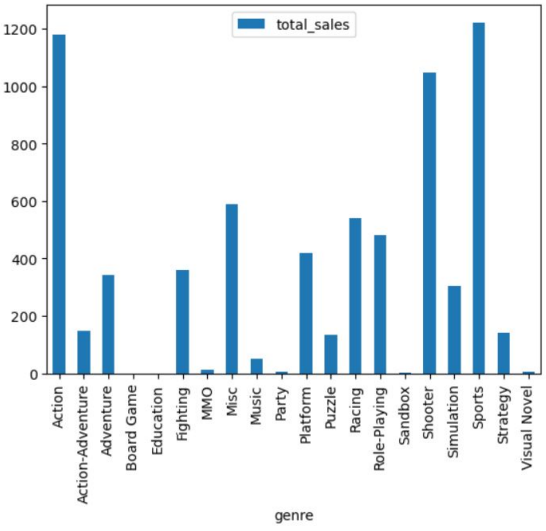


Step 4: Merge all sales by region.

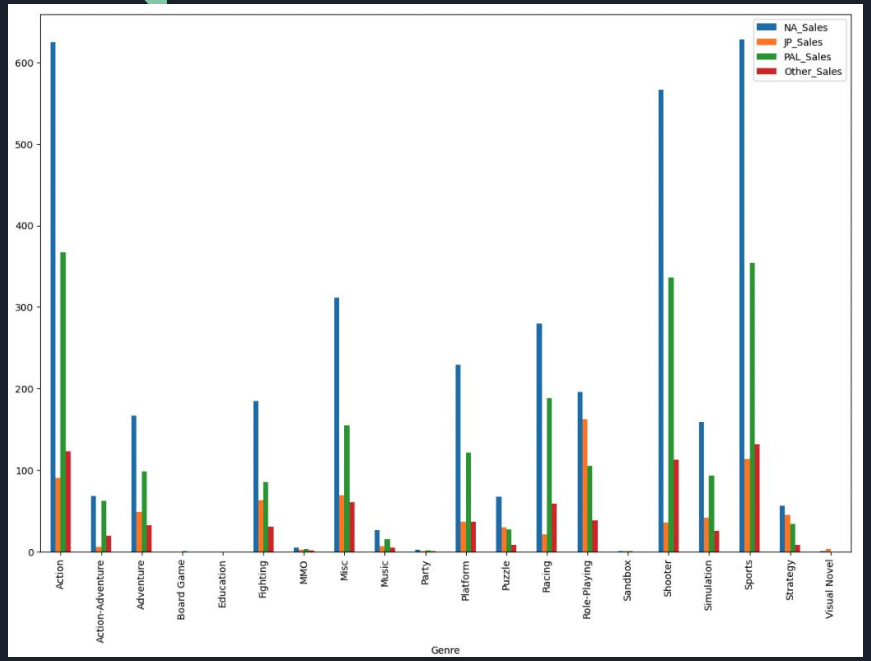
Step 5: Merge both 2019 and 2020 datasets to compare years side by side.

In [6]:
```
total_sales_20.plot.bar()
```

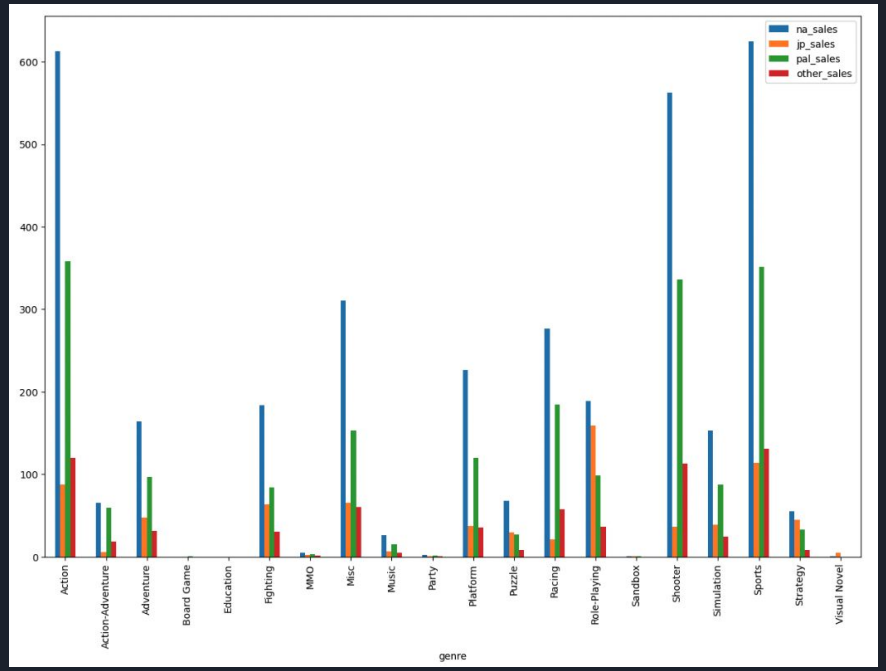
Out[6]: <Axes: xlabel='genre'>



2019 vs 2020 Sales per Genre



2019 Sales



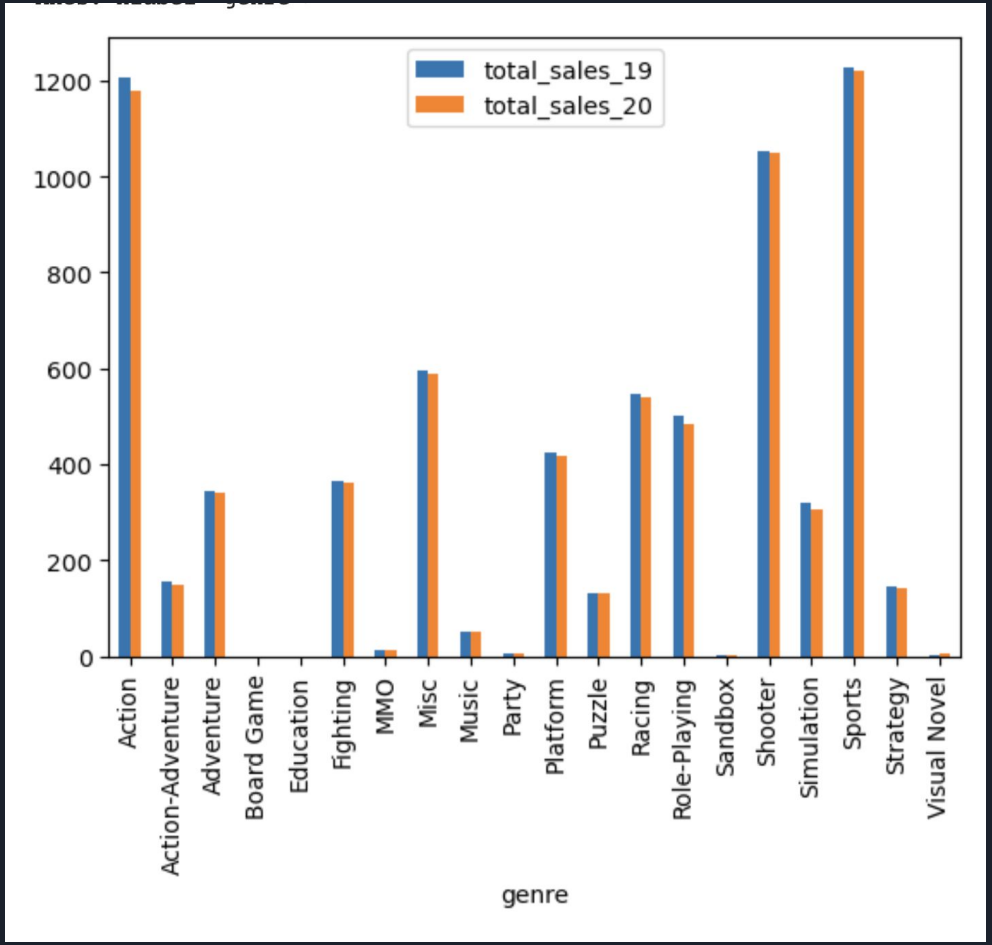
2020 Sales

Dominant Genres over 2 year period.

- Here we took the 2 years of sales data to compare sales amount grouped by genre.
- Top 5 performing genres:
  - Sports
  - Action
  - Shooter
  - Misc
  - Racing

Conclusion:

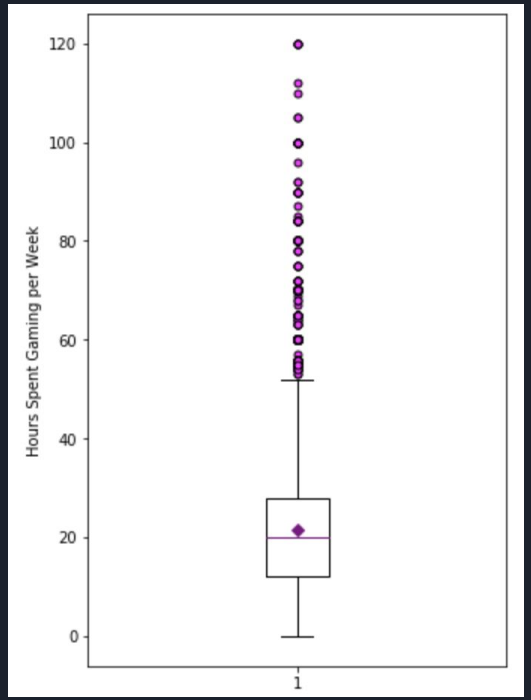
- Sales did not increase during Covid.

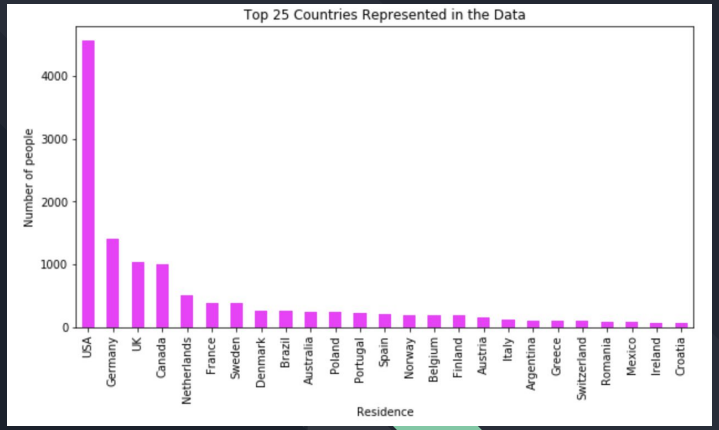


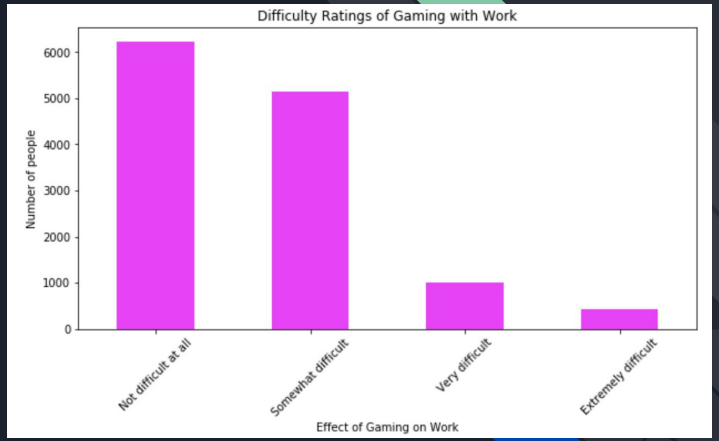
GAMING AND MENTAL HEALTH

Our Data/Data Process

- 55 Columns to 12
- Removed all players who gamed impossible amounts
  - (>120 per week)
- 108 countries
  - Mostly United States
- > 10 different games played by respondents
  - ~85% League of Legends players
- Most did not have trouble with balancing work and gaming

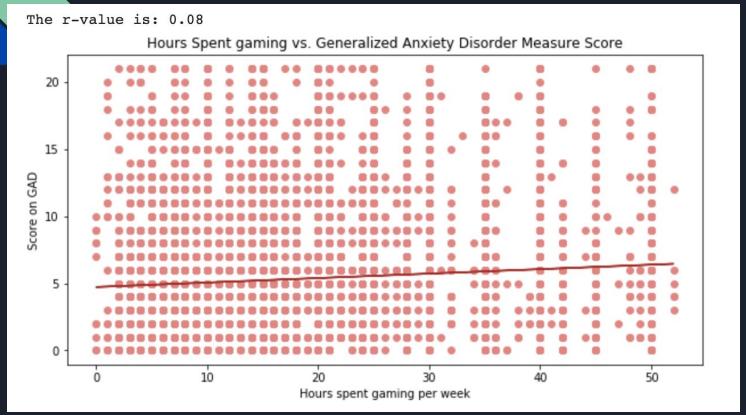




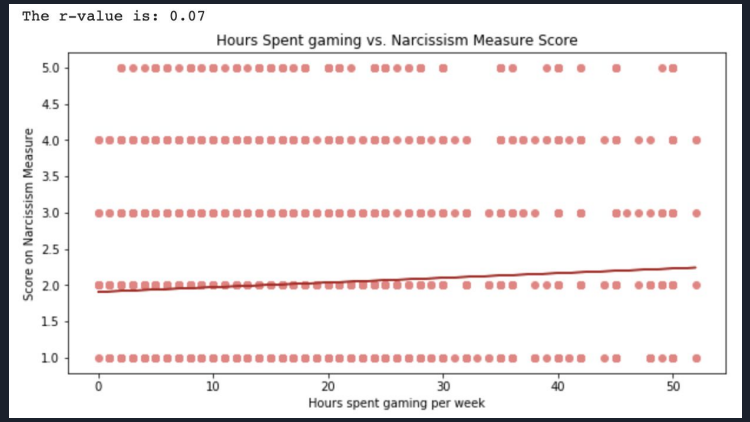


Null Hypothesis was Supported

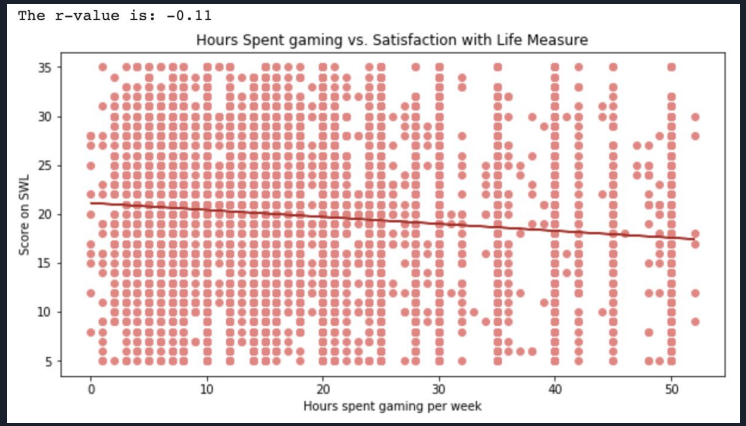
The r-value is: 0.08



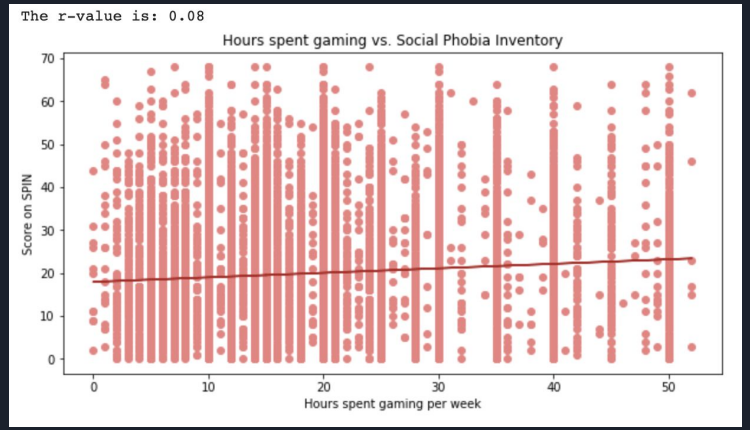
The r-value is: 0.07



The r-value is: -0.11



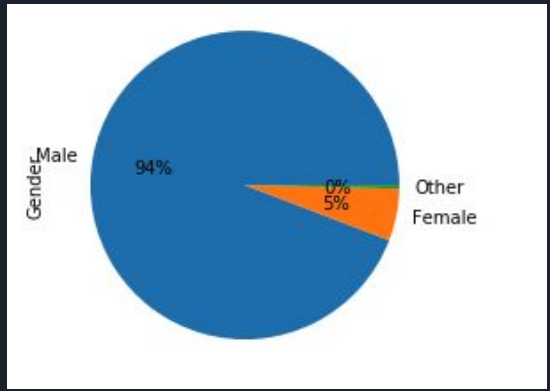
The r-value is: 0.08



Null Hypothesis:

Gaming affect mental illness differently for male or female gamers

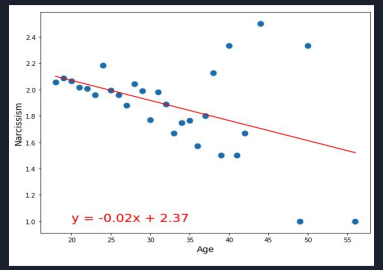
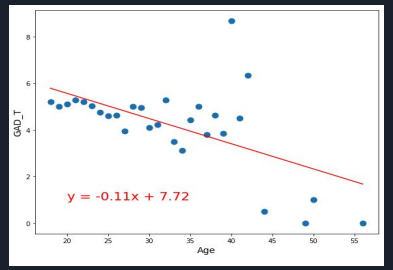
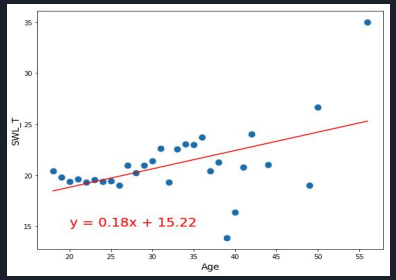
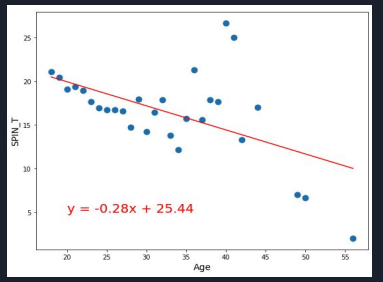
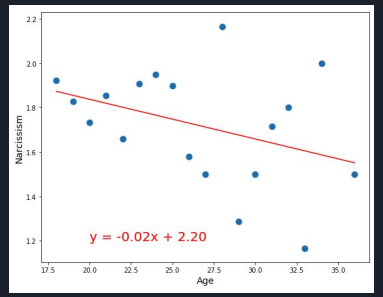
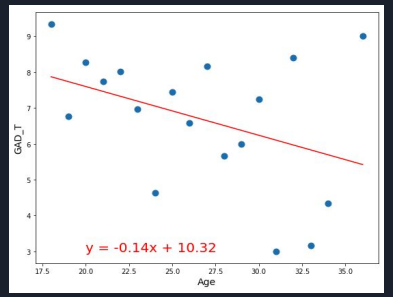
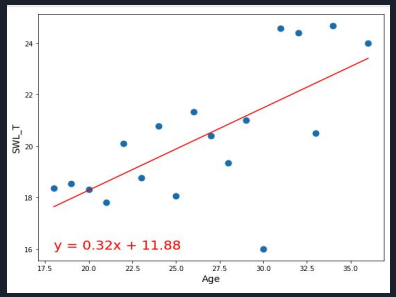
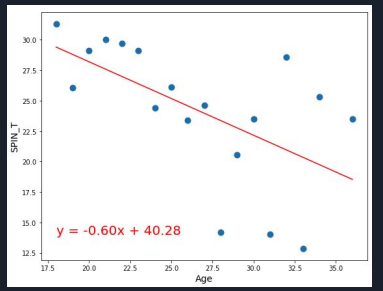
- Data Cleansing
  - Removed unwanted columns and only focused into the Gender, Age and the four mental disorders viz. Narcissism, General Anxiety Disorder (GAD), Satisfaction With Life (SWL) and Social Phobia Inventory (SPIN)
  - Removed the Gender “Other” as it does not conclude if Male or Female so focused only on the Male and Female Genders
- Data Preparation
  - Created separate DataFrames for Male and Female Genders
  - Grouped by Age and took the average for Narcissism, GAD, SWL and SPIN
  - Prepared the separate charts for Male and Female gamers




```
mental_illness_in_male_df = cleaned_health_df.loc[(cleaned_health_df['Gender'] == "Male"), ['Age', 'Narcissism', 'GAD_T', 'SWL_T', 'SPIN_T
mental_illness_in_male_df = mental_illness_in_male_df.groupby("Age").mean()
mental_illness_in_male_df
```

```
mental_illness_in_female_df = cleaned_health_df.loc[(cleaned_health_df['Gender'] == "Female"), ['Age', 'Narcissism', 'GAD_T', 'SWL_T', 'SI
mental_illness_in_female_df = mental_illness_in_female_df.groupby('Age').mean()
mental_illness_in_female_df
```

Observation: H_A was supported

	Narcissism	General Anxiety Disorder (GAD)	Satisfaction With Life (SWL)	Social Phobia Inventory (SPIN)
Male				
Female				

GAMING GENRE

What was our data process?

```
# Dependencies and Setup
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

```
# import IMDB video games csv file
data_url = 'https://raw.githubusercontent.com/shriparna/GameOn/main/InputData/imdb-videogames.csv'
df = pd.read_csv(data_url)
df.head()
```

```
# drop the columns
df.drop(['Unnamed: 0', 'url', 'certificate', 'plot'], axis=1, inplace=True)
df.info()
```

Step 1: Down the CSV files from our datasource websites and import each separate CSV file.

Step 2: Determine which columns we will keep for research, and drop the null values

```
# exclude all the missing values
df = df[df['year'].notna() & df['rating'].notna() & df['votes'].notna()]
df.isna().sum()
```

```
name         0
year         0
rating       0
votes        0
Action       0
Adventure    0
Comedy       0
Crime        0
Family       0
Fantasy      0
Mystery      0
Sci-Fi       0
Thriller     0
dtype: int64
```

Data process continued....

```
# convert year column to integer
df['year'] = df['year'].astype('int')
df.info()
```

```
# convert votes column from string to numeric
df['votes'] = df['votes'].str.replace(',', '')
df['votes'] = df['votes'].astype('int')
df.info()
```

Step 3: Convert the data type

Step 4: Drop the duplicates in the dataframe

```
# some video game name and year are duplicates
# drop those duplicates
df.drop_duplicates(subset='name', inplace=True, ignore_index=True)
df.shape
```

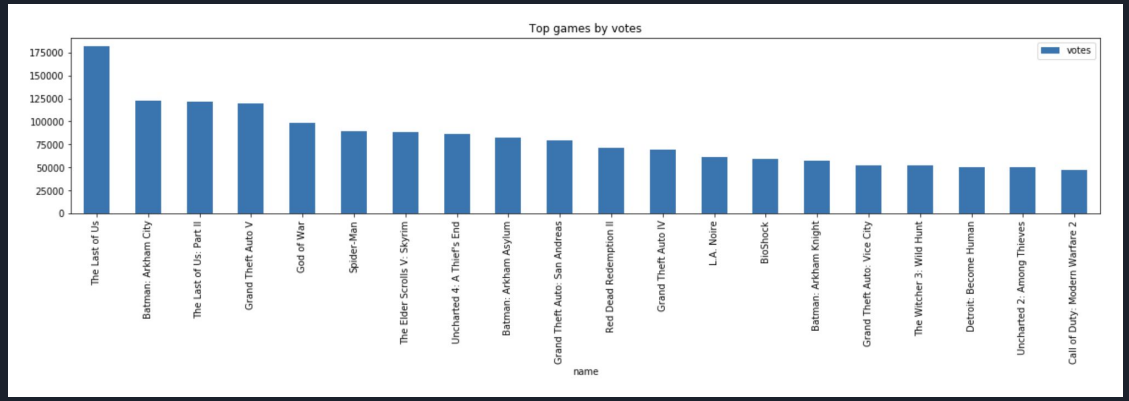
```
(10680, 13)
```

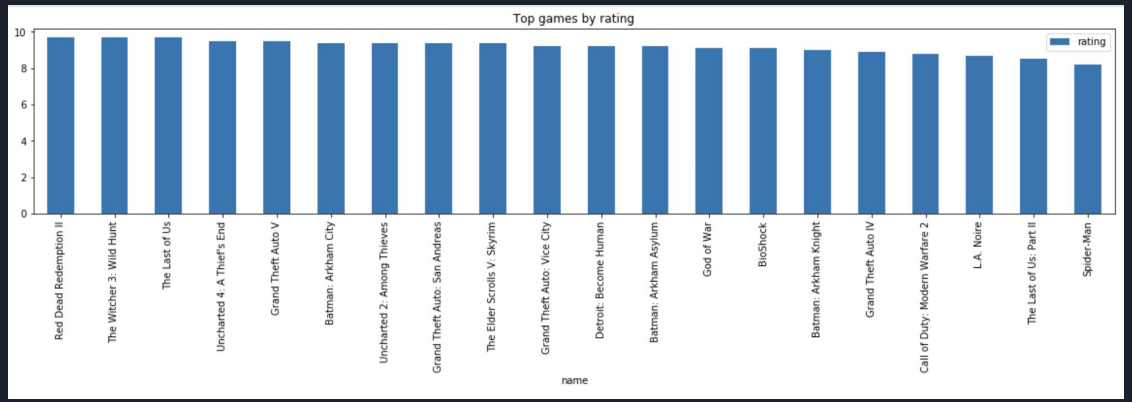
```
len(df['name'].unique())
```

```
10680
```

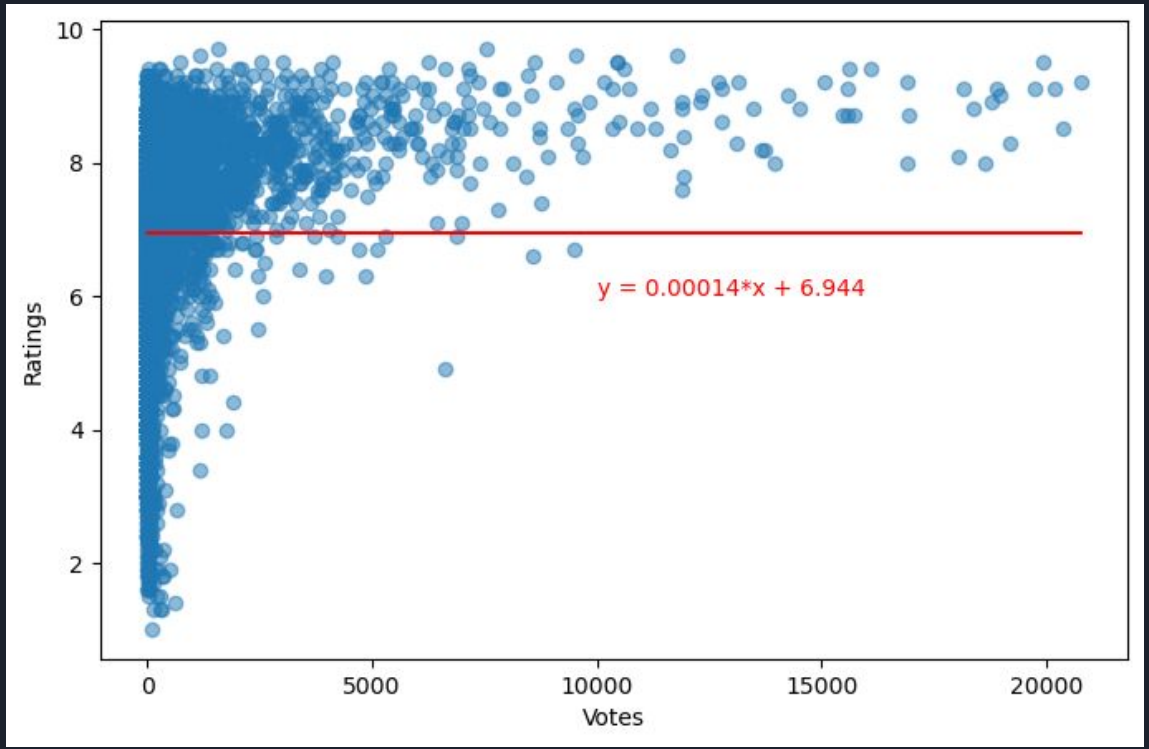
Top 20 games. Ratings vs Votes

Top 20 games. Ratings vs Votes

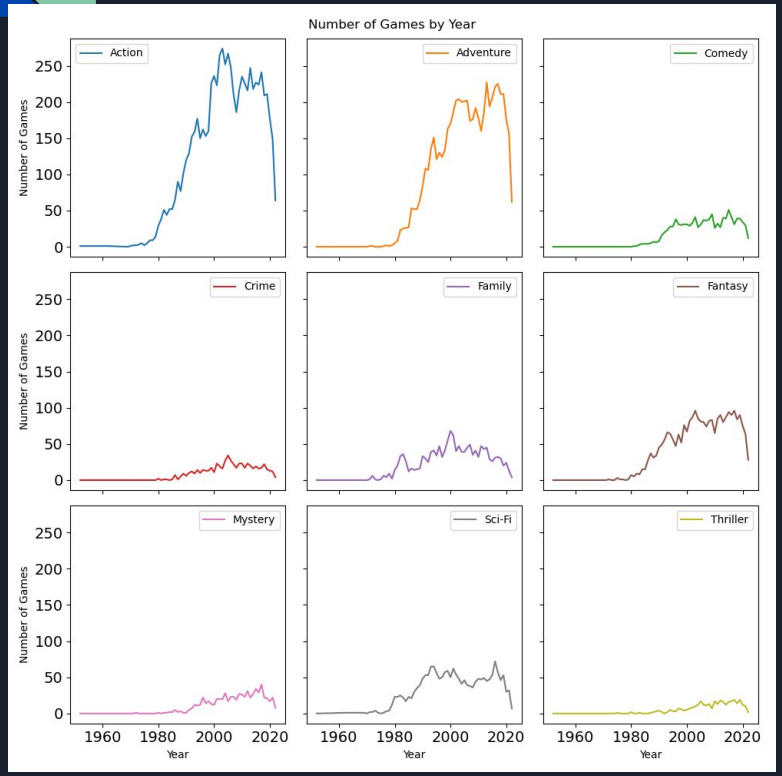


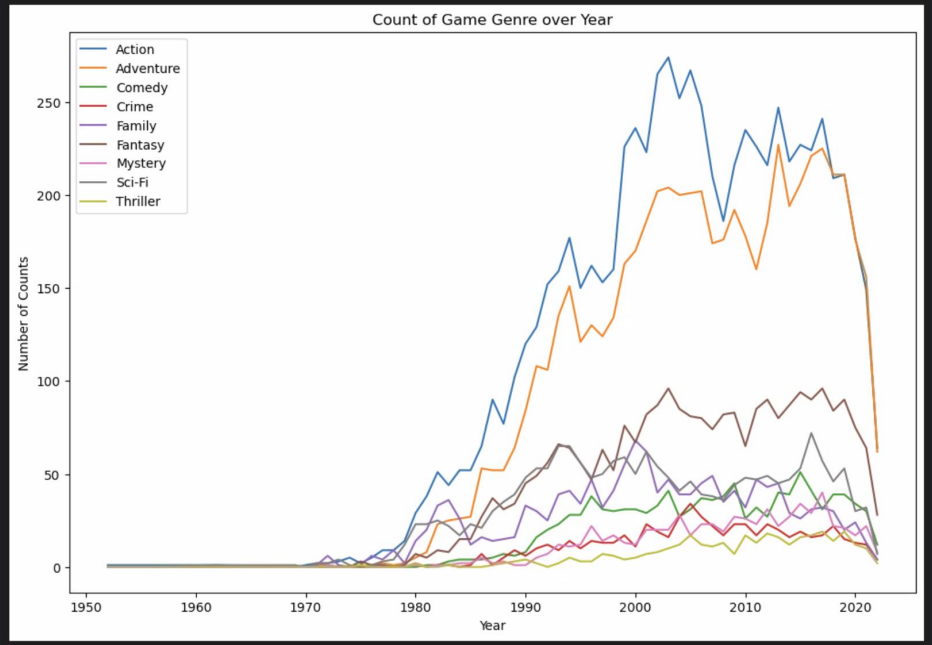


Linear Regression of Rating & Votes

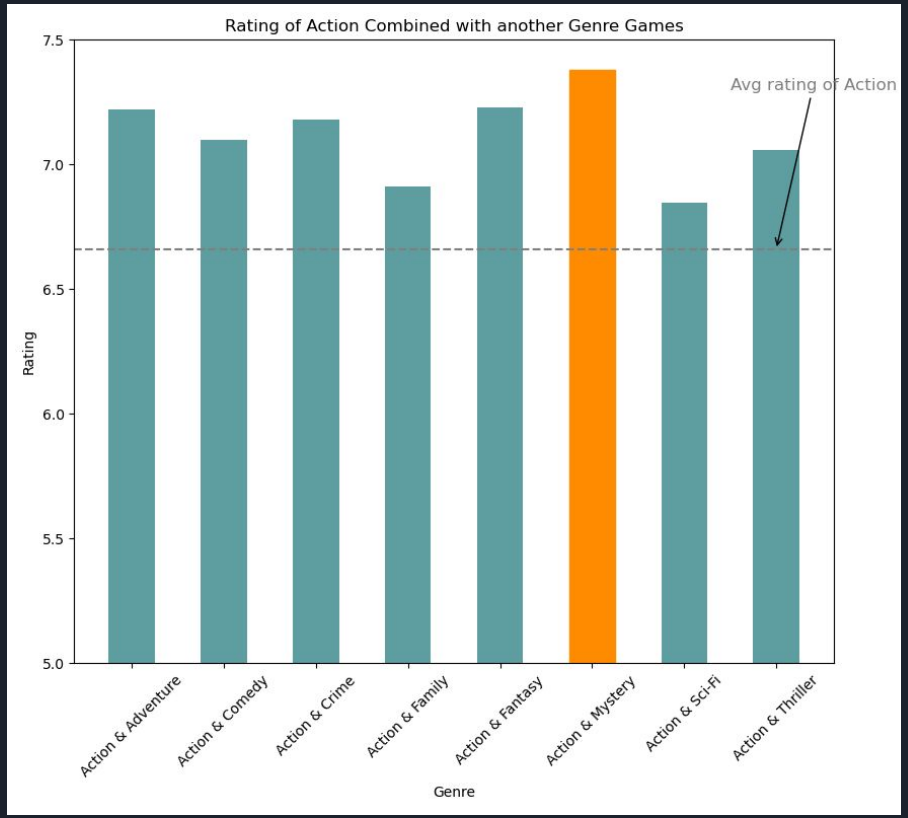


Game Genre over Time



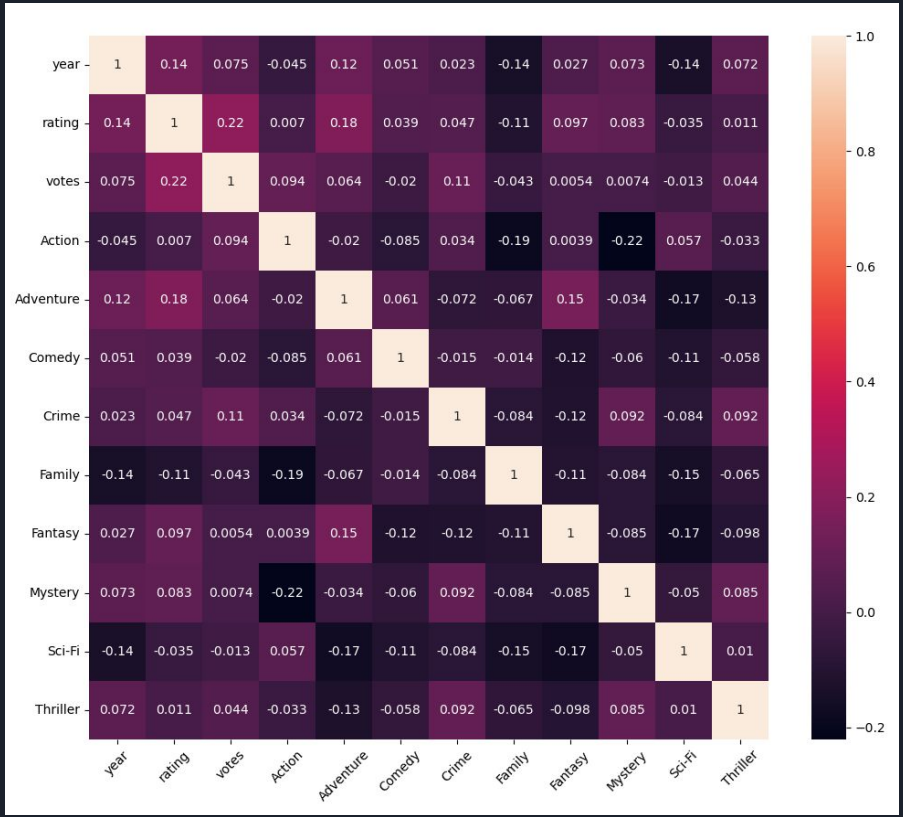


Rating of Action & another Genre



- Action: 6.66 / 10

Correlation Heatmap



Conclusion

- Gaming Sales:
  - During COVID, sales did not increase for console based games.
- Gaming and Mental Illness:
  - Hours spent gaming did not have an effect on mental health
  - Gaming affects male and female gamers the same way
- Gaming Genre:
  - Vote did not affect the rating of games
  - Action is the genre people most commonly engaging with, but a combination of Action and another genre will reach a higher rating

Future Plans:

Ways to expand and enhance our project?

- **Locating API's for data.**
  - **Having an API to pull gaming information from would give us a better insight.**
- **Utilizing accurate and present data.**
  - **Having current data from sales companies within the gaming industry.**
- **Having expanded data over larger years and gaming platforms**
  - **More generalizable data to draw larger conclusions.**
- **Bigger case studies.**
  - **mental health had a lot of limitations, a more varied study would provide a better insight.**

Questions & Answers

Data Sources with links.

- Online Gaming Anxiety Data
  - https://www.kaggle.com/datasets/divyansh22/online-gaming-anxiety-data?resource=download
- Video Games Sales Dataset
  - https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv
- Video Game Sales
  - https://www.kaggle.com/datasets/gregorut/videogamesales
- IMDB Video Games
  - https://www.kaggle.com/datasets/muhammadadiltalay/imdb-video-games?resource=download
- Video Game Sales Dataset
  - https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat