COVID-19 Impact on the Video Game Industry

Permissions

Place an `X` in the appropriate bracket below to specify if you would like your group's project to be made available to the public. (Note that student names will be included (but PIDs will be scraped from any groups who include their PIDs).

- [X] YES - make available
- [ ] NO - keep private

Overview

This project focuses on investigation between COVID-19 and the video game community, specifically on how in- game player counts, discount frequencies, and number of games releases react to global quarantine mandates. Results showed there is a significance increase in player counts after the U.S. quarantine for one month followed by a slight drop. On the other hand, both discount frequencies and number of games releases have no effects to the quarantine mandates which may imply there are no developers trying to generate sales in the start of the quarantine.

Names

- Brandon Vazquez
- Ernesto Escusa
- Chung En Pan
- Manuel Rodriguez Nunez
- Eric Estabaya

Research Question

Big Question: ***How has the COVID-19 infection rate influenced the video game community?***

1. How has the rise in COVID-19 infection rates affect user playtime among popular online video-games?
2. From the start of COVID-19 infections, what relationship can be seen develop with video game discounts and quarantine length?
3. Ever since COVID-19 has quarantined everyone what has it done to major game release dates?

Background & Prior Work

Amidst the COVID-19 pandemic, it is essential that we pay attention to how people are spending their time in this ongoing crisis. With this in mind people have turned to different types of entertainment since the availability of going outside has become limited in the sense that quarantine has restricted the ability to have fun in the sun. The ability to talk to our friends and have each other's presence has now been pushed to the internet and through our screens we have one another and instead of playing sports a majority of us have turned to virtual gaming to connect to one another in these trying times of the COVID-19 quarantine.

Our research question emerged from our group's interest in gaming and how the Corona Virus has affected our ability to enjoy our habits due to lockdown orders of COVID-19. As people who are currently experiencing this lockdown and crisis we use different videogames to fill in the time and to cope with the stress coming from lockdown, instead of the usual freedom we have outside. Our group is interested in how the lockdown has affected the gaming community and how major corporations are responding due to the free time and ability to play more video games within our homes. In most cases we believe that this increase in staying at home will result in more time being used to play games and create more purchases toward gaming entertainment henceforth the idea leads to our hypothesis that COVID-19 will positively influence the gaming community.

The prior works we found relating to our project demonstrate that the increase in COVID-19 cases that resulted in more persons being quarantined at home occurred during 2019-2021, which makes sense due to the quarantines starting around that time. The first source shows us that at the middle of the pandemic an increase of freetime had occurred and during this freetime people of all age groups began testing out new gaming service (Arkenberg, 2020). In this source there were increases in COVID-19

cases that then created an increase of freetime at home that then is converted into people trying out new digital activities to pass the time. With everything being closed and unavailable outside people began turning to gaming to fill the hours stuck at home. An increase in mobile games and video game popularity began to skyrocket in response to the new playerbase stepping into the market. As we can see this source is highly related to our research question as we are looking for the kind of impacts COVID-19 had on the gaming community. On the other hand, in response to this massive increase in playersize many companies tried to push more game releases during quarantine so that their games can have a better chance of being played. It also created more sale revenue for online gaming because many people have spare income due to public entertainment being closed, as well as not being able to go out as often. With reference 2 we can see the boom in the gaming business model as it stepped through 2020 with major increases in sales and the market exploding due to the newcomers(Hall, 2020). We were able to see that the market grew 48% compared to last year in the total of 77 billion dollars for gaming revenue globally. We also saw an increase in game downloads from internet providers that saw an increase in data usage across all parts of the world. With both these sources in mind, we are expecting to have a positive correlation between COVID-19 cases and gaming revenue. In reference 3 we were able to see that during normal peak hours of gaming an increase of 12 percent had occured.This was due to the overall increase of time spent at home, coming from online learning and recess students spent their time playing games during breaks (). It also shows that since most of the time students were at home anyways they would spend their time playing games during school as well because they are not physically in class so they could do whatever they wanted. Other than an increase and social media consumption the overall server increase to most major gaming platforms had increased overall in all time periods.

References (include links):

- 1)Arkenberg Chris. Will gaming keep growing when the lockdowns end? (n.d.). Retrieved March 15, 2021, from https://www2.deloitte.com/us/en/insights/industry/technology/video-game-industry-trends.html
- 2)Hall Stefan. COVID-19 is taking gaming and esports to the next level. Retrieved March 15, 2021, from https://www.weforum.org/agenda/2020/05/covid-19-taking-gaming-and-esports-next-level/
- 3)Shanley, P. (2020, March 18). Gaming usage up 75 PERCENT Amid coronavirus Outbreak, Verizon reports. Retrieved March 15, 2021, from https://www.hollywoodreporter.com/news/gaming-usage-up-75-percent-coronavirus-outbreak-verizon-reports-1285140

Hypothesis

Overall Hypothesis:

We believe COVID-19 affects gaming communities in a positive way.

1. We believe that as COVID-19 infection rates rise the playtime for popular online video-games shall increase as well so they should have a positive correlation to one another as seeing the time quarantined will be used as time playing games.

2. Looking at the start of the COVID-19 infection we can predict that video game discounts begin to become more frequent the longer the quarantine goes on for the more sales they try to make by slashing prices since people are more likely to buy a game if it is cheaper and since they have more free time.

3. We predict that major game release dates are pushed out as quick as possible so to create more sales since quarantine is giving everyone more free time if gaming companies can release games during this time they can make more money.

Dataset(s)

- **1)** Dataset Name: COVID-19-data (ouroworldindata.org) 16.1MB
- Link to the dataset: https://github.com/owid/covid-19-data/tree/master/public/data
- Number of observations: 62827 rows x 59 columns

This dataset is the global COVID dataset, containing 192 countries information about the total of cases, new cases each day, total death, new death each day, total case per million, new case per million, new death/ total death per million, etc.

- **2)** Dataset Name: Steam Games Dataset : Player count history, Price history and data about games 6.53GB
- Link to the dataset: https://data.mendeley.com/datasets/ycy3sy3vj2/1
- files Name: PlayercountHistoryPart1, PlayercountHistoryPart2, PriceHistory, Developers, Genres, Information, Packages, Supportedlanguages, Tags
- Number of observations:
  - PlayercountHistoryPart1: containing 1000 (game id) csvs, each of them 280224 rows x 2 column containing player count in each 5 minute interval

    - PlayercountHistoryPart2: containing 1000 (game id) csvs, each of them 23352 rows x 2 column containing player count in each 1 hour interval
    - PriceHistory: Containing 1513 (game id) csvs, each of them 493 rows x 4 column containing date, initial price, discounted price, and discounted rate
    - Developers: 2000 rows x 2 columns (game_id, developers)
    - Genres: 2000 rows x 2 columns (game_id, generes)
    - Information: 2000 rows x 5 columns(game_id, type, name, releasedate, freetoplay)
    - Packages: 1797 rows x 3 columns(game_id, package_id)
    - Supportedlanguages: 1993 rows x 2 columns (game_id, languages)
    - Tags: 2000 rows x 21 columns(game_id, tag1- 20)

This dataset provides mainly about the price and discount history of games by each game.

- **3)** Dataset Name: Lifetime concurrent users on Steam
- Link to the dataset: https://steamdb.info/app/753/graphs/
- Variables: Date, users (count), in-game (count)
- Number of observations: 1211 rows x 2 columns

This dataset provides a comprehensive player count for the games.

- **4)** Dataset Name: release
- Link to the dataset: https://steamspy.com/year/2021
- Variables: Game, Date
- Number of observations: 38380 rows × 2 columns

This dataset provides numbers of games released in each date.

Setup

In [ ]:
```
import os
import glob
import datetime as dt

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from matplotlib.dates import DateFormatter
import matplotlib.dates as mdates
import geopandas as gpd
import folium
import branca.colormap as cm
from folium.plugins import TimeSliderChoropleth

import seaborn as sns
sns.set()
sns.set_context('talk')
```

Data Cleaning

Step 1:

For easier data wrangling and visualization later, 'date' variable are transformed to datetime object, and the average playercount are groupby by the day. After transformation, the tables are saved as csv file and later merge with the information csv for later use. This process not only transforms the regular string object to datetime object, but also reduced the observation by a lot (original has more than 280000 observations per file, now to around 1000 observations per file for 2000 files).

In [ ]:
```
def wrangle_datetime(folder):

    empty_file= np.array([])
    filepath = '~/Github/covid_game_analysis/data/steaminfo/'+ folder

    for filename in os.listdir(filepath):
        filename_path = filepath + '/' + filename
        if filename.endswith('.csv'):
            if os.stat(filename_path).st_size !=0:
                with open(os.path.join(filepath, filename),'r+') as f:
                    print(filename)
                    temp= pd.read_csv(f,dtype={"Time": str, "PlayerCount": f
                    temp['Time']= pd.to_datetime(temp['Time'], errors='coerc
                    temp = temp.dropna(axis=0)
                    result= temp.groupby([temp['Time'].dt.date]).mean()
                    result['Playercount']= result['Playercount'].apply(int)
                    result.to_csv(path_or_buf= filename_path)

        else:
            empty_file= np.append(empty_file,filename)
    print('empty_file', empty_file)
#wrangle_datetime('PlayerCountHistoryPart1')
#wrangle_datetime('PlayerCountHistoryPart2')
```

Step 2:

There are several files needed to be merged based on game id, so we extracted the game_id from information csv and went through the PlayerCountHistory1/PlayerCountHistory2 to extract the time and player count based on each game_id and then merge them to the corresponding game_id for around 2000 game_ids. Eventually what we obtained is a single table with game_ids, names, released date, time, playercount, etc.

In [ ]:
```
def mergeinfo():
    temp_info = pd.read_csv('data/steaminfo/Information.csv', encoding='unic
    df = pd.DataFrame()
    empty_file= np.array([])
    for id in temp_info.get('appid'):

        file_path = 'data/steaminfo/PlayerCountHistory/' + str(id) + '.csv'
        if os.stat(file_path).st_size !=0:
            with open(file_path, 'r') as f:
                f_csv = pd.read_csv(f,dtype={"Time": str, "Playercount": flo
                f_csv['appid'] = id
                temp = pd.DataFrame()
                temp = temp.append(temp_info[temp_info.get('appid') == id])
                temp = temp.merge(f_csv, how = 'inner', on = 'appid')
                df = df.append(temp)

        else:
            empty_file= np.append(empty_file,id)
    print('empty_file:',empty_file)
    df['Time']= pd.to_datetime(df['Time'], errors='coerce', format='%Y-%m-%d
    return df
df = mergeinfo()
```

```
empty_file: [397100.]
```

Here we see that game_id 397100 is empty, therefore we will omit this observation (game).

Step 3:

Since the 'date' variable type in file 'PriceHistory' is an object, so for easier wrangling purposes and merging purposes, we shall turn the 'date' variable type to datetime object. Then we will merge the price information into the merged dataframe based on game id and the date. Adding the initial price, discounted price, and discount correspond to each date for each game to the merged dataframe.

Note: the data from 'PriceHistory' are recorded since 2019-04-07.

In [ ]:
```
def merge_price(dataframe):
    temp = pd.DataFrame()
    empty_file = np.array([])
    for game_id in dataframe.get('appid').unique():
        file_path = 'data/steaminfo/PriceHistory/' + str(game_id) + '.csv'
        if os.path.exists(file_path):
            if os.stat(file_path).st_size !=0:
                with open(file_path,'r') as f:
                    f_csv = pd.read_csv(f,dtype={"Date": str, "Initialprice"
                    f_csv = f_csv.rename(columns= {'Date':'Time'})
                    f_csv['Time']= pd.to_datetime(f_csv['Time'], errors='coe
                    f_csv['appid'] = game_id
                    temp = temp.append(f_csv)
            else:
                empty_file= np.append(empty_file,str(game_id))
        else:
            empty_file = np.append(empty_file,str(game_id))
    dataframe = dataframe.merge(temp, how = 'outer', on = ['appid','Time'])
    print('Number of empty_file:',len(empty_file))
    return dataframe
```

In [ ]:
```
new_df = merge_price(df)
```

```
Number of empty_file: 485
```

All of the functions above will notify me whether the files corresponding to the appid existed or has no content in the file. It seems that there are a lot of empty (485) files from the price history that are empty.

Note: Missing price history of games are due to some games are free to play.

In [ ]:
```
month_cases = {'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6, '
def transform_datetime(dt_entry):
    if dt_entry is np.nan or dt_entry is 'NaN':
        return np.nan

    parts = dt_entry.split('-')

    try:
        month = month_cases[parts[1]]
        day = int(parts[0])
        year = int(parts[2])

        if year > 50:
            year = year + 1900
        else:
            year = year + 2000
    except:
        return np.nan

    dt_form = dt.datetime(year, month, day)
    return dt_form
new_df['releasedate'] = new_df['releasedate'].apply(transform_datetime)
```

```
<>:3: SyntaxWarning: "is" with a literal. Did you mean "=="?
<>:3: SyntaxWarning: "is" with a literal. Did you mean "=="?
<ipython-input-6-70cf06173533>:3: SyntaxWarning: "is" with a literal. Did yo
u mean "=="?
  if dt_entry is np.nan or dt_entry is 'NaN':
```

Step 4:

With some adjustment of the variables, eventually what we have is a huge dataframe containing the appid,type of content, name, whether it is free_to_play, player count, initial price, discounted price, and discount percentage corresponding to the date:

Steam is known for its wide array of gaming library, however, Steam application content has way more application than we think. Thus we filter out the unnecessary part.

In [ ]:
```
game_df = new_df[new_df['type'] == 'game']
game_df.head(5)
```

Out[ ]:

	appid	type	name	releasedate	freetoplay	Time	Playercount	Initialpric
0	578080	game	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	0.0	2017-12-14	1248227.0	Na
1	578080	game	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	0.0	2017-12-15	1427167.0	Na
2	578080	game	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	0.0	2017-12-16	1540028.0	Na
3	578080	game	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	0.0	2017-12-17	1451095.0	Na
4	578080	game	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	0.0	2017-12-18	1231938.0	Na

Step 5:

We've noticed that there are a lot of missing information on the release date and the player count of the games, therefore we pulled the player counts and the release date from other two datasets.

In [ ]:
```
release = pd.read_csv('data/release.csv', names = ['games','date'])
release ['date'] = pd.to_datetime(release['date'], format='%Y-%d-%m')
release.head(5)
```

Out[ ]:

	games	date
0	Destiny 2	2019-10-01
1	Black Squad	2019-06-26
2	Sekiro: Shadows Die Twice - GOTY Edition	2019-03-21
3	Football Manager 2020	2019-11-19
4	NBA 2K20	2019-09-05

In [ ]:
```
chart = pd.read_csv('data/steaminfo/chart.csv', names = ['date','user','inga
chart['date'] = pd.to_datetime(chart['date'], errors='coerce', format='%Y-%m
chart.head(5)
```

Out[ ]:

	date	user	ingame
0	2017-11-13	14054902	3948383
1	2017-11-14	13983638	5007635
2	2017-11-15	14048277	5063418
3	2017-11-16	13762149	5042709
4	2017-11-17	15128594	5738598

Step 6:

Lastly, the COVID- 19 Dataset

In [ ]:
```
covid = pd.read_csv('data/owid-covid-data.csv')
covid['date'] = pd.to_datetime(covid['date'], errors='coerce', format='%Y-%m
covid.head(5)
```

Out[ ]:

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	t
0	AFG	Asia	Afghanistan	2020-02-24	1.0	1.0	NaN	
1	AFG	Asia	Afghanistan	2020-02-25	1.0	0.0	NaN	
2	AFG	Asia	Afghanistan	2020-02-26	1.0	0.0	NaN	
3	AFG	Asia	Afghanistan	2020-02-27	1.0	0.0	NaN	
4	AFG	Asia	Afghanistan	2020-02-28	1.0	0.0	NaN	

5 rows × 59 columns

Data Analysis & Results

EDA

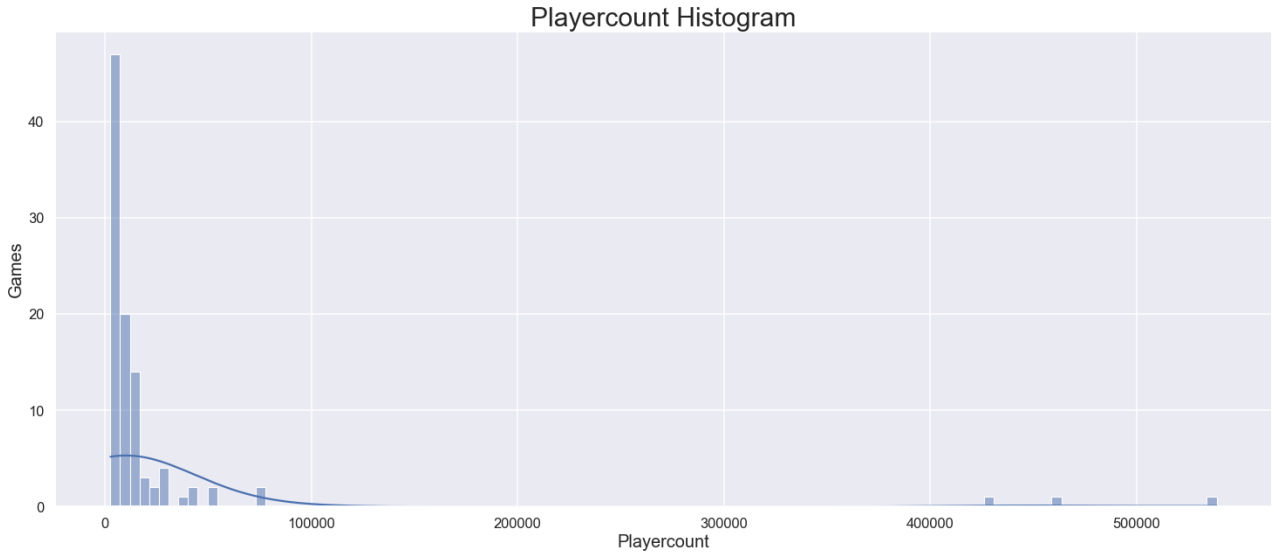
Since most of the player typically only focus playing on AAA title, games that have higher development and marketing budgets, we will look at the top 100 games.

Average player count across top 100 games:

In [ ]:
```
top_df = game_df.groupby('name').mean().sort_values('Playercount', ascending
top_df = top_df.head(100)
```

In [ ]:
```
fig, axes = plt.subplots(figsize = (25, 10))
temp = top_df.get('Playercount')
fig = sns.histplot(temp, kde = True)
axes.set_title("Playercount Histogram", fontsize = 30)
axes.set_ylabel("Games", fontsize = 20)
axes.set_xlabel("Playercount", fontsize = 20)
display(temp.describe())
```

```
count       100.000000
mean      26119.818060
std       81156.919603
min        2691.229188
25%        3995.698099
50%        8003.358684
75%       14975.329137
max      538642.069887
Name: Playercount, dtype: float64
```



From the histogram above, we can see that most of the data revolve below 100000 player count. But there are also few outliers. Let's take a look at what they are.

In [ ]:
```
z_scores = np.abs(stats.zscore(top_df.get(['Playercount'])))
outlier = (z_scores >= 3).all(axis =1)
outliers = top_df[outlier]
outliers
```

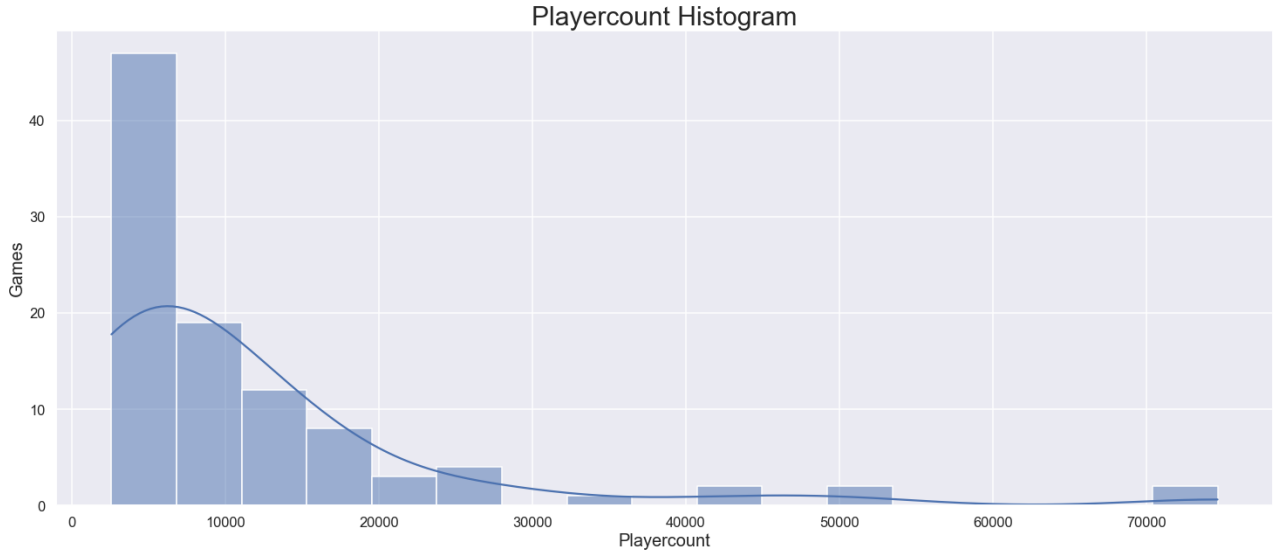
Out[ ]:

name	appid	freetoplay	Playercount	Initialprice	Finalprice	Discount
PLAYERUNKNOWN'S BATTLEGROUNDS	578080.0	0.0	538642.069887	29.99	27.555923	8.11359
Dota 2	570.0	1.0	462112.040082	NaN	NaN	NaN
Counter-Strike: Global Offensive	730.0	1.0	429399.503597	NaN	NaN	NaN

In [ ]:
```
top_df = game_df.groupby('name').mean().sort_values('Playercount', ascending
top_df = top_df.head(100)
```

In [ ]:
```
fig, axes = plt.subplots(figsize = (25, 10))
temp = top_df.get('Playercount')
fig = sns.histplot(temp, kde = True)

axes.set_title("Playercount Histogram", fontsize = 30)
axes.set_ylabel("Games", fontsize = 20)
axes.set_xlabel("Playercount", fontsize = 20)

display(temp.describe())
```

```
count      100.000000
mean     11896.909725
std      13514.983063
min       2575.463515
25%       3848.751799
50%       7380.550360
75%      14121.581706
max      74642.961973
Name: Playercount, dtype: float64
```

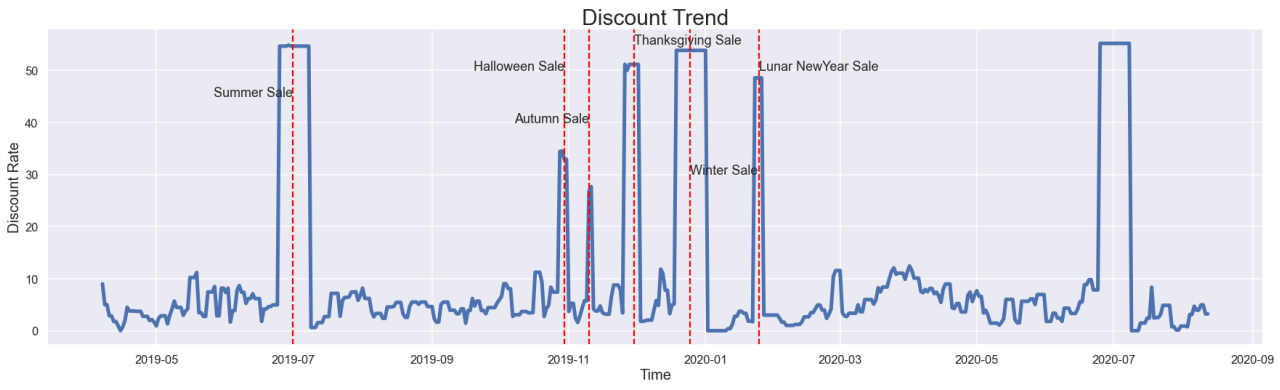


From the graph above, we can see there the plot and the mean of the data and the skewness has shrunk down by a lot after removing the outliers.

Discount history among 100 games:

In [ ]:
```
#plotting the lineplot
fig, axes = plt.subplots(figsize =( 30, 8))
top_name = game_df.groupby('name').max().sort_values('Playercount', ascendin
discount = game_df[game_df['name'].isin(top_name)]
discount = discount.get(['Time','Discount']).groupby('Time').mean()
sns.lineplot(data = discount, x = discount.index, y = discount.get('Discount
axes.set_title('Discount Trend', fontsize = 30)
axes.set_ylabel('Discount Rate',fontsize = 20)
axes.set_xlabel('Time',fontsize = 20)

#plotting the unusual findings
axes.axvline("2019-07-01", linestyle = '--', color = '#FF0000')
axes.text(x = "2019-07-01",y= 45, s ='Summer Sale', ha = 'right')
axes.axvline("2019-10-30", linestyle = '--', color = '#FF0000')
axes.text(x = "2019-10-30",y= 50, s ='Halloween Sale', ha = 'right')
axes.axvline("2019-11-10", linestyle = '--', color = '#FF0000')
axes.text(x = "2019-11-10",y= 40, s ='Autumn Sale', ha = 'right')
axes.axvline("2019-11-30", linestyle = '--', color = '#FF0000')
axes.text(x = "2019-11-30",y= 55, s ='Thanksgiving Sale', ha = 'left')
axes.axvline("2019-12-25", linestyle = '--', color = '#FF0000')
axes.text(x = "2019-12-25",y= 30, s ='Winter Sale', ha = 'left')
axes.axvline("2020-01-25", linestyle = '--', color = '#FF0000')
axes.text(x = "2020-01-25",y= 50, s ='Lunar NewYear Sale', ha = 'left')
```

Out[ ]:
```
Text(2020-01-25, 50, 'Lunar NewYear Sale')
```

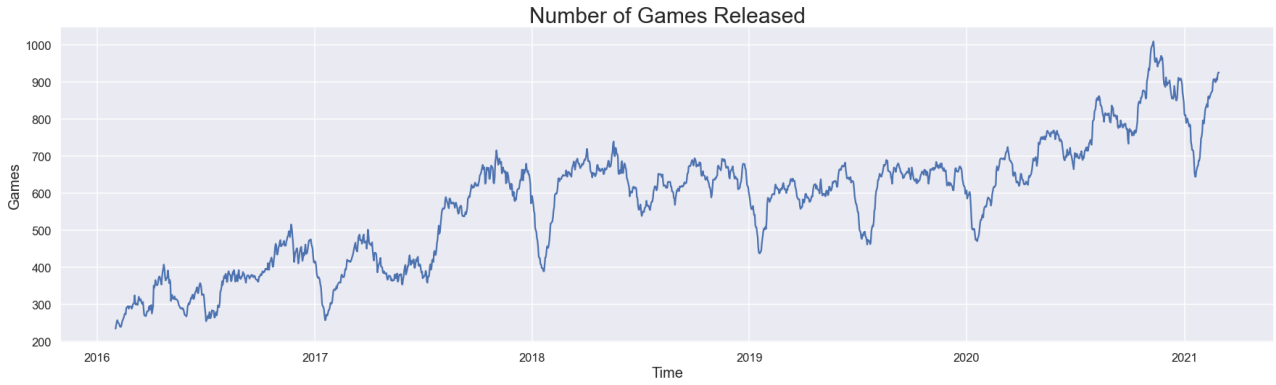


From the line plot above, there seems to be multiple unusual spikes during 2019. If we investigated further, we can see the unusual spikes are due to the discount sale on games.

Release Date for games frequency:

In [ ]:
```
fig, axes = plt.subplots(figsize =( 30, 8))
release_graph = release.groupby('date').count()
sns.lineplot(data = release_graph,x = release_graph.index ,y = release_graph

axes.set_title("Number of Games Released", fontsize = 30)
axes.set_ylabel("Games", fontsize = 20)
axes.set_xlabel("Time", fontsize = 20)
```

Out[ ]:
```
Text(0.5, 0, 'Time')
```



from the graphs above, we can see that there is a gradual increase overall, the only thing that is significant is that there is alway a slight drop between Q4 and Q1. One of the reason might be that some games entertainment might release the game align with Christmas, so more player can purchase their games.

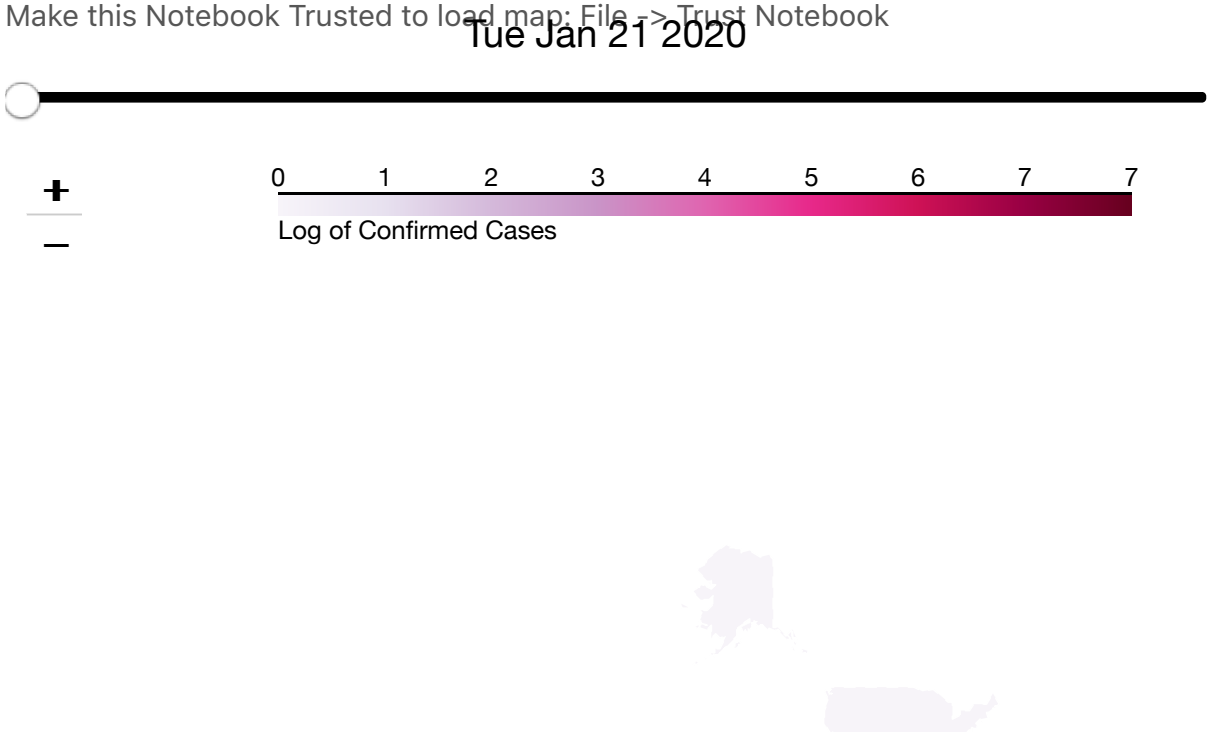
COVID- 19 infection rate

In [ ]:
```
#importing the shape file to map the countries, containing some small countr
#and the geometry of the countries.
world = gpd.read_file('Longitude_Graticules_and_World_Countries_Boundaries-s
world = world.rename(columns = {'CNTRY_NAME':'location'})

#merging the shape file based on the location
geo = covid.merge(world, on = 'location')
geo = geo.get(['OBJECTID','location','date','total_cases','geometry'])

#converting the Date to unix time in nanoseconds
geo['date'] = geo['date'].astype(int) / 10**9
geo['date'] = geo['date'].astype(str)

#we are going to plot the log of number of cases since there are coutries wi
#such as US and China, and Italy
geo['log_Confirmed'] = np.log10(geo['total_cases'])
geo.dropna(inplace = True)

#defining the color map
max_colour = max(geo['log_Confirmed'])
min_colour = min(geo['log_Confirmed'])
```

```
cmap = cm.linear.PuRd_09.scale(min_colour, max_colour)
geo['colour'] = geo['log_Confirmed'].apply(cmap)

#constructing the style dictionary for the timesliderchoropleth module
country_list = geo['location'].unique().tolist()
country_idx = range(len(country_list))
style_dict = {}
for i in country_idx:
    country = country_list[i]
    result = geo[geo['location'] == country]
    inner_dict = {}
    for _, r in result.iterrows():
        inner_dict[r['date']] = {'color': r['colour'], 'opacity': 1}
    style_dict[str(i)] = inner_dict

#creating a Geodataframe to plot the map
gdf = gpd.GeoDataFrame(geo.get(['location','geometry']))
gdf = gdf.drop_duplicates().reset_index()

#initialize the folium map
slider_map = folium.Map(min_zoom=2, max_bounds=True,tiles='cartodbpositron')

#create map and add slider bar and captions
_ = TimeSliderChoropleth(
    data=gdf.to_json(),
    styledict=style_dict,

).add_to(slider_map)

_ = cmap.add_to(slider_map)
cmap.caption = "Log of Confirmed Cases"
#save the file in ouput folder
slider_map.save(outfile='output/TimeSliderChoropleth.html')

print("1) Toggle the slider on top to see the trend overtime.")
print("2) Upper left corner to zoom in & out.")
print("3) Drag to locate.")
slider_map
```

```
1) Toggle the slider on top to see the trend overtime.
2) Upper left corner to zoom in & out.
3) Drag to locate.
```

Out[ ]: Make this Notebook Trusted to load map: File -> Trust Notebook



Using the slider bar, we can see that China has the most cases as of 2020/1/21, and the virus rapidly spread to different countries. Eventually we can see U.S. and India have two of the most confirmed cases among the world as of 2021/03/07.

Notes: informations are missing for the countries that are appeared to be white.

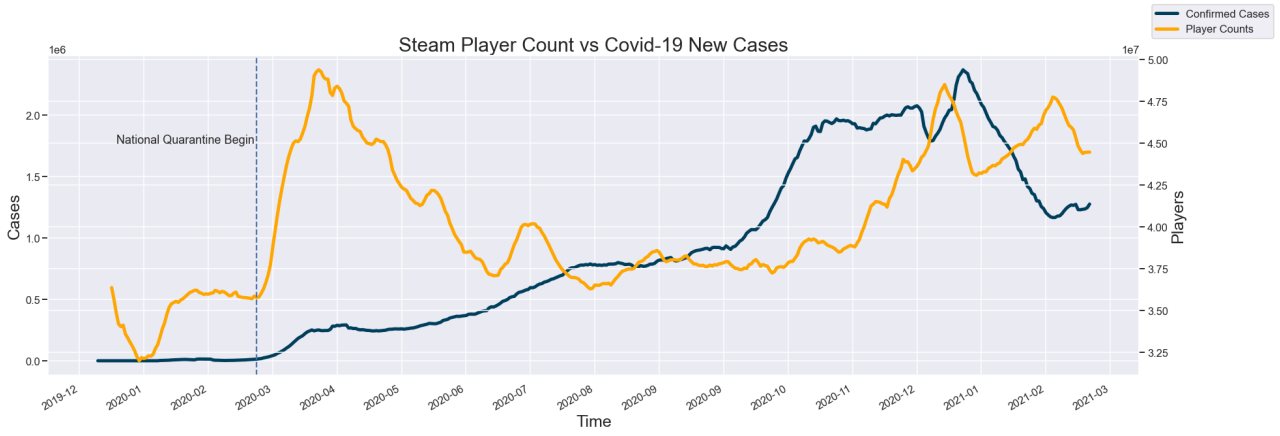
Result

We will now narrow down to different question to answer the big question: **How has the COVID-19 infection rate influenced the video game community.**

How has the rise in COVID-19 infection rates affect user playtime among popular online video-games?

In [ ]:
```
chart = chart.set_index('date')
```

In [ ]:
```
#extracting cases
covid_count = covid.get(['iso_code','continent','location','date','new_cases
covid_count = covid_count.groupby('date').sum()

#merging the playercount based on date
temp = covid_count.join(chart,how= 'left')

#plotting
fig, axes = plt.subplots(figsize = (30,10))
axes.set_title('Steam Player Count vs Covid-19 New Cases', fontsize = 30)
axes.set_xlabel("Time", fontsize =25)
axes.plot(temp.index, temp.get('new_cases_smoothed'),color = '#003f5c', line
axes.set_ylabel("Cases", fontsize = 25)

ax2=axes.twinx()
ax2.plot(temp.index, temp.get('ingame').rolling(7).sum(),color = '#ffa600',
ax2.set_ylabel("Players", fontsize = 25)
ax2.axvline("2020-03-10", linestyle = '--')
ax2.text(x = "2020-03-09",y= 45000000, s ='National Quarantine Begin', ha =

date_form = mdates.DateFormatter('%Y-%m')
axes.xaxis.set_major_formatter(date_form)
axes.xaxis.set_major_locator(mdates.WeekdayLocator(interval= 4 ))
fig.autofmt_xdate()
fig.legend()
plt.show()
```



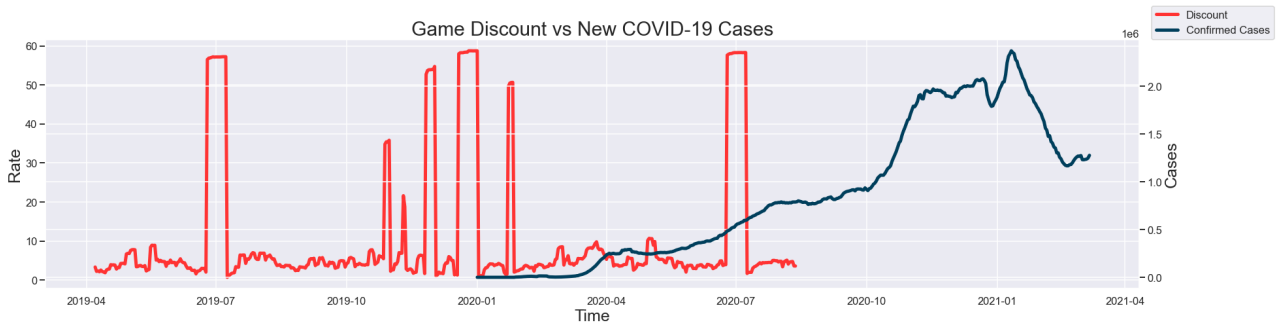
From the graphs above, we can see that since the national quarantine begin, the player count on Steam dramatically increased for a month. It might be due to the amount of citizens across the globe are staying at home for work, school, etc. Without leisure activities outside the house and decreased travel time between destinations, people can only turn to computer and play games. However, after one month, we see a gradual drop which might be due to the law of diminishing return then followed by increases player count again. Thus, we **failed to reject** the hypothesis that COVID-19 infection rates raise the playtime for popular online video-games.

From the start of COVID infections, what relationship can be seen develop with video game discounts and quarantine length?

In [ ]:
```
# Create dataframe to examine discounts over time
covid_discount = game_df.copy()
covid_discount = covid_discount.groupby('Time').mean().get(['Discount'])
```

In [ ]:
```
# Declare figure plane
fig, axes = plt.subplots(figsize = (30,7))

# Label figure and plot discount frequency over time
axes.set_title('Game Discount vs New COVID-19 Cases', fontsize= 30)
axes.plot(covid_discount.index,covid_discount.get('Discount'),color = '#ff33
axes.set_ylabel('Rate', fontsize = 25)

# Plot COVID-19 cases
ax2 = axes.twinx()
ax2.plot(temp.index, temp.get('new_cases_smoothed'),color = '#003f5c', linew
ax2.set_ylabel("Cases", fontsize = 25)
axes.set_xlabel('Time', fontsize = 25)
fig.legend()
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x7ff568ca9550>
```



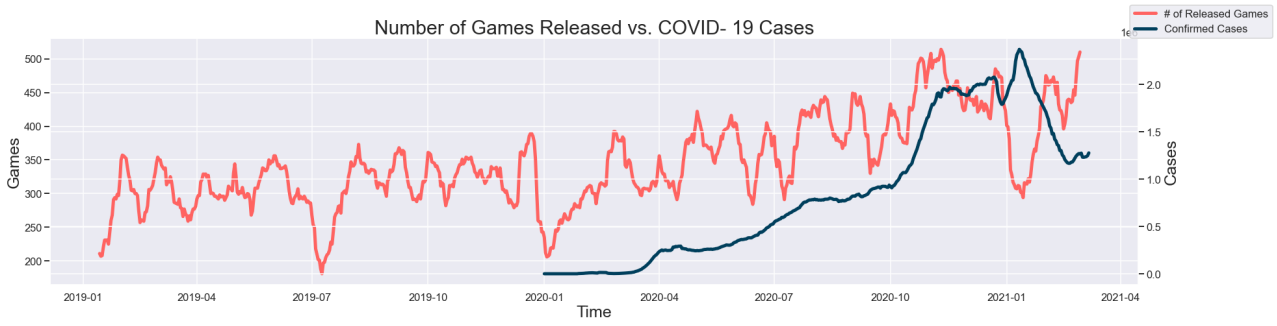
Based on this graph on discounts over time, COVID-19 did not really affect the frequency of the discounts that Steam had to offer nor the rate of the discount. We have similar trends from 2019-04 to 2019-07 compared against 2020-04 to 2020-07. The zone of these dates encompass normal, yearly, holiday discounts that are offered. Therefore we **reject** the hypothesis that video game discounts begin to become more frequent the longer the quarantine goes on for the more sales they try to make by slashing prices since people are more likely to buy a game if it is cheaper and since they have more free time.

What impact has COVID had on major game release dates since the start of quarantine?

In [ ]:
```
release_graph.reset_index(inplace = True)
release_graph['year'] = release_graph.get('date').dt.year
release_graph = release_graph[release_graph.get('year') >= 2019]
release_graph.set_index('date', inplace = True)
```

In [ ]:
```
# Declare figure plane
fig, axes = plt.subplots(figsize = (30,7))

# Plot number of games released on a year/month basis
axes.plot(release_graph.index, release_graph.get('games').rolling(14).sum(),
axes.set_ylabel('Games', fontsize = 25)

# Plot COVID-19 cases
ax2 = axes.twinx()
ax2.plot(temp.index, temp.get('new_cases_smoothed'),color = '#003f5c', linew
ax2.set_ylabel("Cases", fontsize = 25)

axes.set_xlabel('Time', fontsize = 25)
axes.set_title("Number of Games Released vs. COVID- 19 Cases", fontsize = 30
fig.legend()
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x7ff568ddb5e0>
```



From this, we noticed that the number of games released did not follow the trend in COVID- 19 cases. Though there are a slight overall increases, but we believe this is just due to the nature of the communities. We can ensure this by looking at the timeline before virus. There is already a minimal increase thus we **reject** the hypothesis that that major game release dates are pushed out as quick as possible so to create more sales.

Ethics & Privacy

When considering data privacy and security our main concern was to not leak any private information that could hurt any individual or gaming community. Given that the datasets we will be utilizing consist of largely mass-gathered public data, which does not take into account personal information, we do not believe this project risks the privacy of any individual or large gaming industries. Data security is also not of great concern as the data collected is publicly available and lacks personal information. For our data analyzes we were using anonymized information in order to witness the changes, if any, regarding video game sales and covid. In terms of identifying potentially affected parties, our aim is to identify how the ongoing pandemic has influenced the gaming community, release dates and price fluctuations as a whole, so we do not foresee our analysis having an impact on any particular stakeholder. Through conducting our analysis, we expect to find a positive correlation between the increase in COVID-19 infection rates and increase in engagement with online video games, discounts, and game release dates. However, we are aware that our analysis merely focuses on the relationship between the pandemic and video game engagement, and does not take into account other covariates which may account for the results found. Other factors that may contribute to flawed results is the collection bias found in our data, which only focuses on the most popular video games, without taking into account the rest of the market, or any possible situations that could affect the relationship we are looking at. In an effort to prevent dishonest representations in our analysis, we plan on modeling a representation which is reproducible and strictly details the results found within our datasets, while acknowledging its biases and limitations, since there are many factors in a real environment that can change certain outcomes.

Conclusion & Discussion

For our research project we analyzed the correlation between COVID-19 infection rate and the video game community. When going over what we could base the project off of, we hypothesized that playtime and infection rates would increase simultaneously, while game discounts throughout quarantine become more frequent from the start of the virus. We also hypothesized that game release sales would increase due to companies wanting to make as many sales as possible. Looking at the results of our hypothesis and data, we could conclude that rising COVID-19 infection rates do in fact have a positive effect in player in- game among video-games to certain extent. In other words, while looking at the "Steam Playtime vs COVID-19 New Cases" graph it is clear that there is one spike in player count at first while there is a gradual increase after the first spike when COVID-19 cases increases. This means that while new cases were reported the more players were seen to join gaming communities. On the other hand, when comparing game discounts with quarantine length, looking at the graph it is clear that before COVID-19 there were discounts towards the end of 2019 and seems to be a trend regarding holidays. Moreover, when looking at the result between number of games released vs COVID-19 confirmed cases, there seems to have no effect between the two. Though we reject two of the three initial hypothesis which put us in favor of the position that the virus does not affect the video gaming communities, we still stand with our position in the hypothesis that COVID-19 does have a positive impact on the game

It is crucial to consider the limitation these analysis can extend. For example, we need to factor in the importance of the player counts among games. As more player participated in many games, there will be more in- game purchases being made at the same time, which is a factor when deciding whether virus is contributing such behavior. However, in- game purchases and game sales are private information that are held by the Steam platform. Without these, we cannot determine confidently whether our hypothesis is valid to the question. Also, we assume the first spike in player count is caused by U.S national quarantine, but we did not take in the fact that other nations also slowly begin the quarantine mandate, therefore, it is debatable that the second spikes are caused by the other nations quarantine or local quarantine. This assumption might post an error when we conducted our analysis.

Team Contributions

- **Brandon Vazquez**
  - formulated hypothesis
  - edited and formatted ethics and privacy
  - did the conclusion & overview
  - edited and formulated the research questions
- **Ernesto Escusa**
  - edited formulated the research questions
  - edited background & prior works
  - edited and formatted ethics and privacy
  - formulated hypothesis
  - acquired dataset
- **Chung En Pan**
  - search for datasets
  - did overview
  - narrowed down the hypothesis
  - cleaned steam dataset, convert string format to datetime format and groupby the player count by day for 2000+ files and merged them into a info dataframe. clean 2000+ price history files and merged to the info dataframe
  - EDA (Average player count across top 100 games), (Release Date for games frequency), (Games discount frequecy)
  - EDA (Geo- spatial analysis for COVID-19 infection rate using Folium)
  - Analyzed result
  - Discussion
- **Manuel Rodriguez Nunez**
  - helped to ideate project hypothesis
  - edited and updated research questions
  - wrote ethics & privacy report
  - formulated team presentation
- **Eric Estabaya**
  - Support and clarifications on programming aspects
  - Generation and cleaning of part of data set
  - Aid in handling EDA
  - Assisted in analyzing figures and data

In [ ]: