COSC3000 Project 1
Data Analysis and Visualization Report
Discovering Trends in Video Games: Platforms, Genres, Publishers, and Profitable Regions

Name : Abimanyu Yuda Dewa

Student Number : 47292903

Contents

Introduction

In recent years, the video game industry has seen explosive growth. As more and more people turn to video games as a source of entertainment, it's becoming increasingly important for game developers and publishers to stay on top of the latest trends in the industry.

One effective way to gain insights into these trends is through data visualization. By analysing data related to video game sales, player demographics, and other key metrics, it's possible to identify patterns and make informed decisions about game development and marketing strategies.

In this report, we'll explore some of the most significant trends in the video game industry using data visualization techniques. We'll look at data related to game sales, player demographics, and other factors that can influence the success of a video game. By examining this data, we hope to provide valuable insights that can help game developers and publishers make more informed decisions about their products.

Aims

This report aims to find trends in the dataset and discover the popular genres and platforms which yield high sales and review scores. Various methods of data analysis and visualization will be used to discover hidden trends in the data. These could help developers decide which types of games are favoured across the world and with it, they can maximize their profits.

Data

The data was sourced from Kaggle.com, available in .csv format with numerical and categorical entries. The data is mostly clean, so no extensive pre-processing was required. The data includes video games from years 1983 to 2012 and was last updated three years ago. The original dataset by Andy Bramwell can be found here.

Methods

Jupyter Notebook was used for exploratory data analysis. Univariate and Bivariate analysis was done on the columns of interest, such as publisher, genre, platform, review, etc. Upon exploring, there were several problems with the data. First, the sales columns did not have any information on how the numbers were represented, there were discussions regarding the dataset on this matter, however, no response has come from the owner. So, for the sake of simplicity, we will take that as the total sales, and we will use that to measure how popular it is based on the demographics in the data. Second, the values for the 'Year' column were in floating point and some were missing. During preparation, the type was converted to an integer but the rows with missing values were kept since the data is already small. In the end, the data with the missing year values were excluded from the plots.

For plotting the data, simple plots were made with matplotlib in Jupyter Notebook, but the plots included in this report are made with Tableau. This is because some of the plots that were planned to be used in this report prove to be difficult to make with Python and matplotlib, however, they were relatively easier to build with Tableau.

Data Analysis

Univariate

For univariate exploratory data analysis, we used the line and sorted bar chart for plotting the data. This was the simplest and most understandable plot type for the data. The following figures are the number of game titles based on year, genre, and platform.

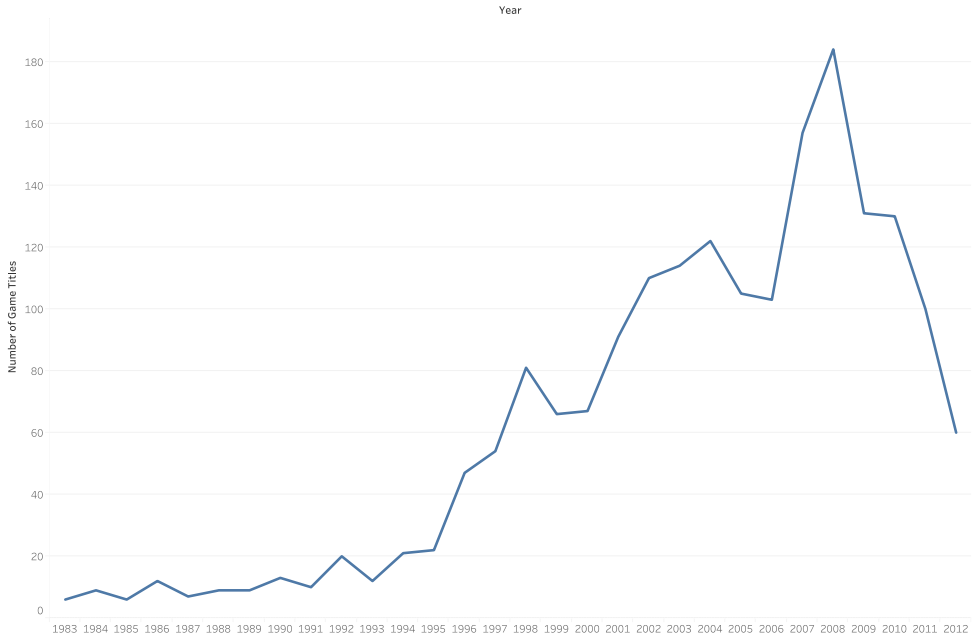

Figure 1 Total number of game titles per year

Figure 1 depicts the number of game titles per year from the data. We can see it rising from 1993 to 2008, after that, the number decreases. The year 2008 has the highest number of games. This could also indicate the point where the video game industry started to grow, along with the release of popular titles and consoles. There are less than 20 video games that were released before 1992, which can lead to overfitting when we calculate the average reviews per year.

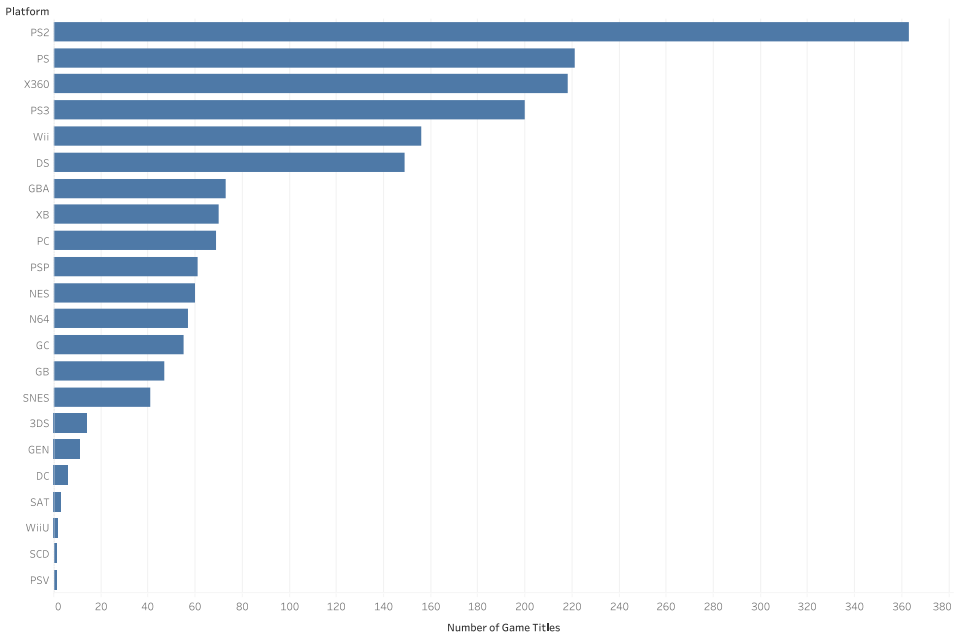

Figure 2 Total number of game titles for each platform

Figure 2 shows the number of games per platform. We can see that the PlayStation 2 has the most video game titles, which is no surprise. PlayStation 2 had great video games and a good price point. Iconic video game titles like the second and third Metal Gear Solid and God of War got their start on this platform. After that came the original PlayStation, which has popular titles such as the original Metal Gear Solid and Final Fantasy IX. Next, comes the PlayStation 3 and Xbox 360. These two consoles have great exclusives, namely Uncharted, Little Big Planet, Gears of War, and Halo.

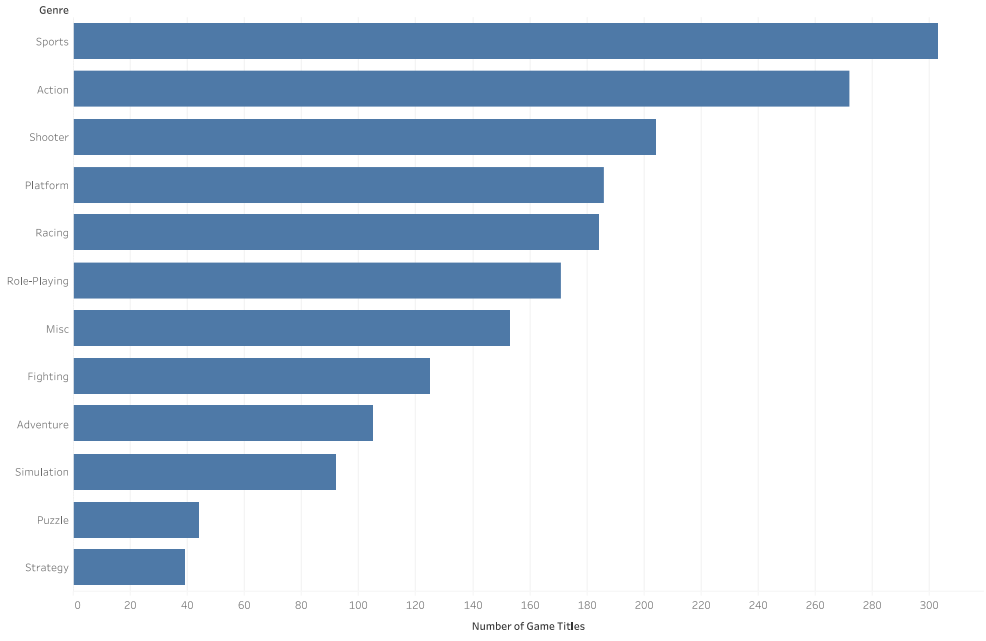

Figure 3 Total number of game titles per genre

From Figure 3, we can see that most of the games in the data are sports games. This is most likely due to sports titles like FIFA and PES being released each year. We can see that the strategy and puzzle genres have the least number of titles. This is likely due to them not having a big community of players or bad marketing strategies, so developers steer clear of such titles.

Bivariate

Now we will explore genres and publishers to see which has video games with a high average review score. We will also explore genre and platform popularity based on demographics. Bar charts and scatterplots will be used as they are simple to understand.

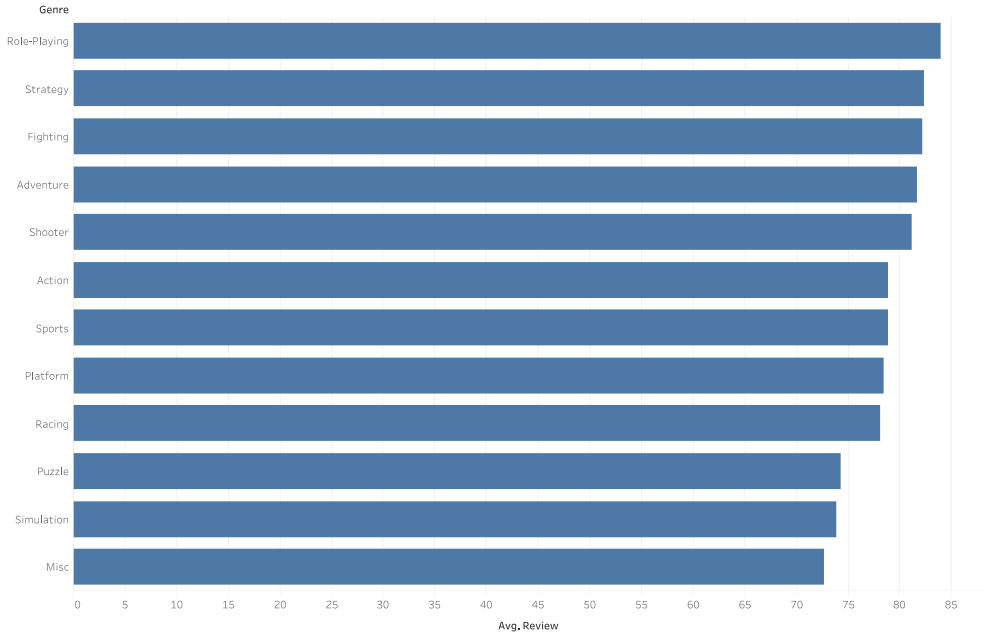

Figure 4 Average review of each genre

Figure 4 shows the average review of video games grouped by each genre. Most of the genre has a high average review score. We can see the top 3 genres with the highest review scores are role-playing, strategy, and fighting. There are also some sort of clusters in the data. Strategy, Fighting, Adventure, and Shooter games have an average review between 80-85. Action, Sports, Platform, and Racing games all have an average review between 75-80. The rest below them have an average review between 70-75.

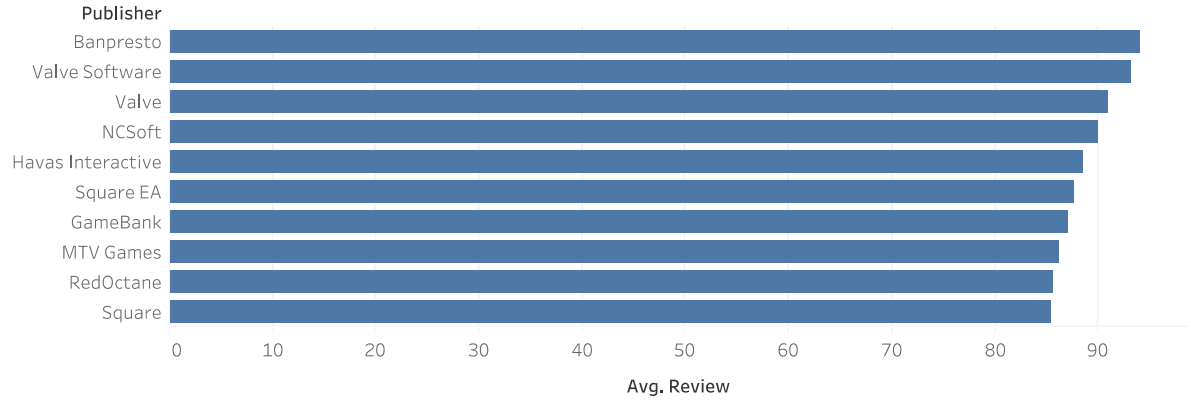

Figure 5 Top 10 Publishers based on average review

From Figure 5, we can see the top ten publishers which released video games with the highest average review score. The top three publishers have video games with an average review score that exceeds 90. One of them is Valve which released notable titles such as Counter-Strike, Half-Life, Dota, Left 4 Dead, and Team Fortress.

Now we will discuss the popularity of the video game genre based on demographics. This will give us insight into which genre is well-liked in certain countries. The first three plots below depict which genre has the highest sum of sales each year in North America, Europe, and Japan.

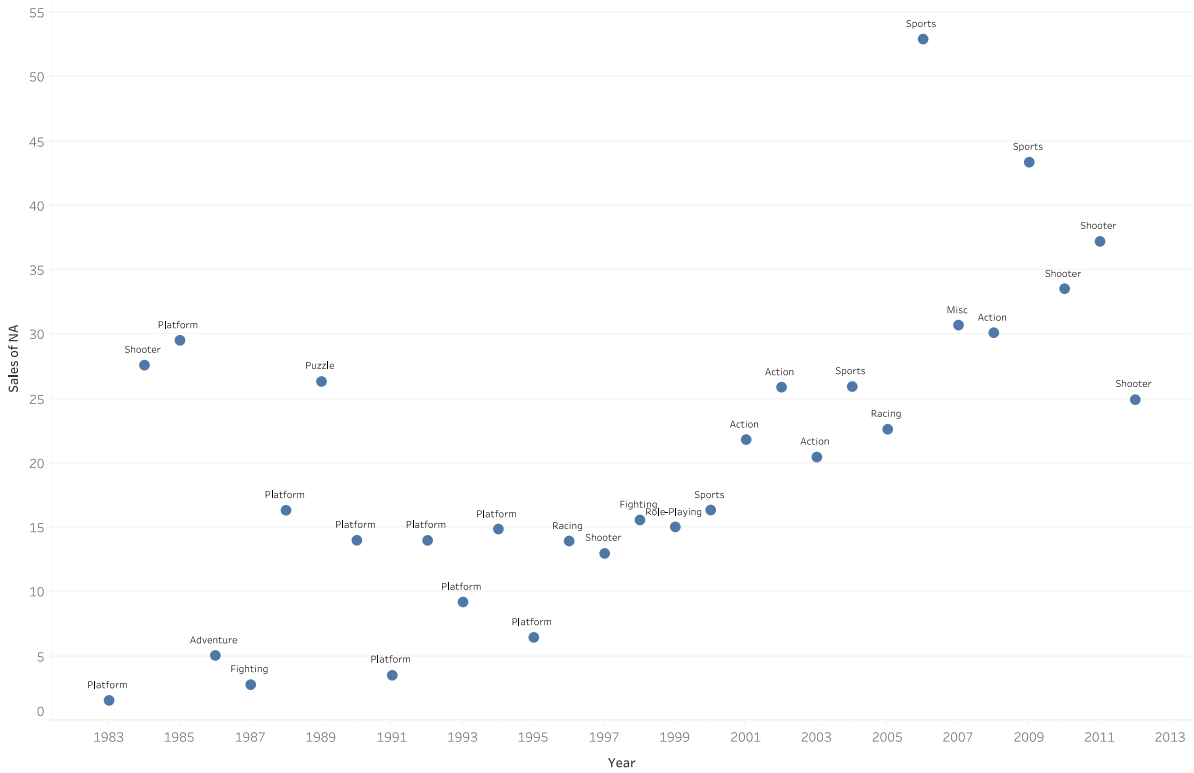

Figure 6 Genre with the max sum of sales per year in North America

Figure 6 shows which genre has the highest sum of sales each year in North America. We can see that in the late 80s and early 90s, platform games dominate the market. After that, action and sports games were popular in the early 2000s. Lastly, in 2009 – 2012, shooters have the highest sales.

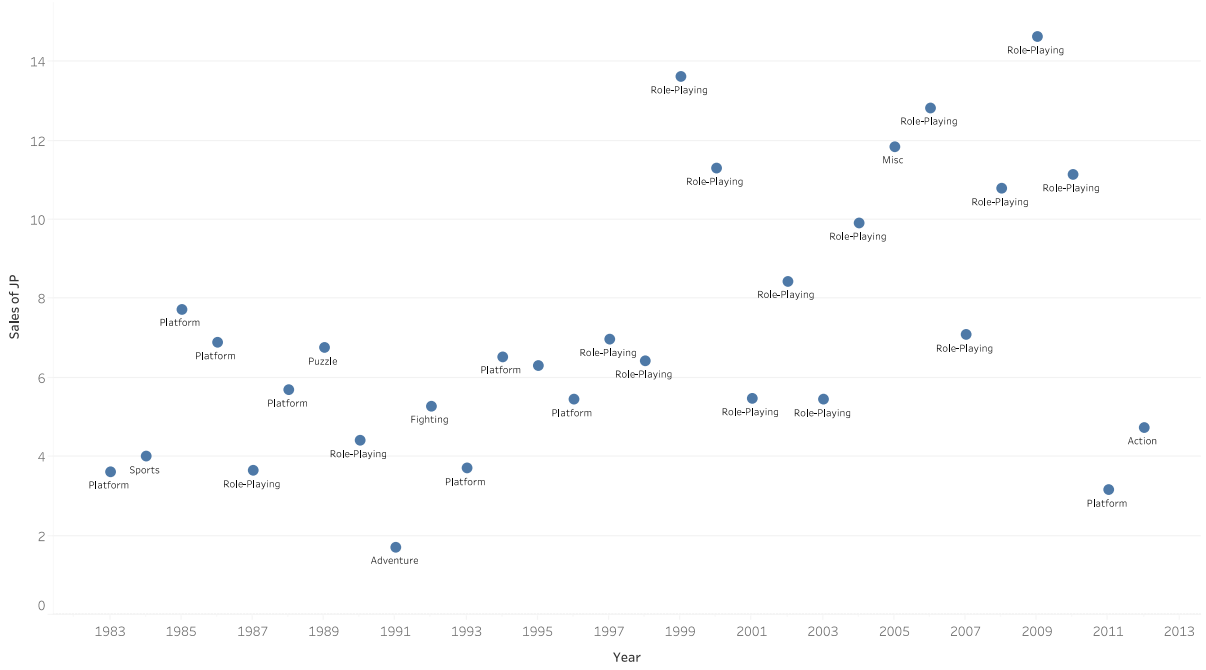

Figure 7 Genre with the max sum of sales per year in Japan

Figure 7 shows the highest-selling video game genres per year in Japan. We can still see that platform games were popular in the 80s and 90s. However, starting from 1997, role-playing games have the highest sales until 2010. It seems role-playing games are popular in Japan.

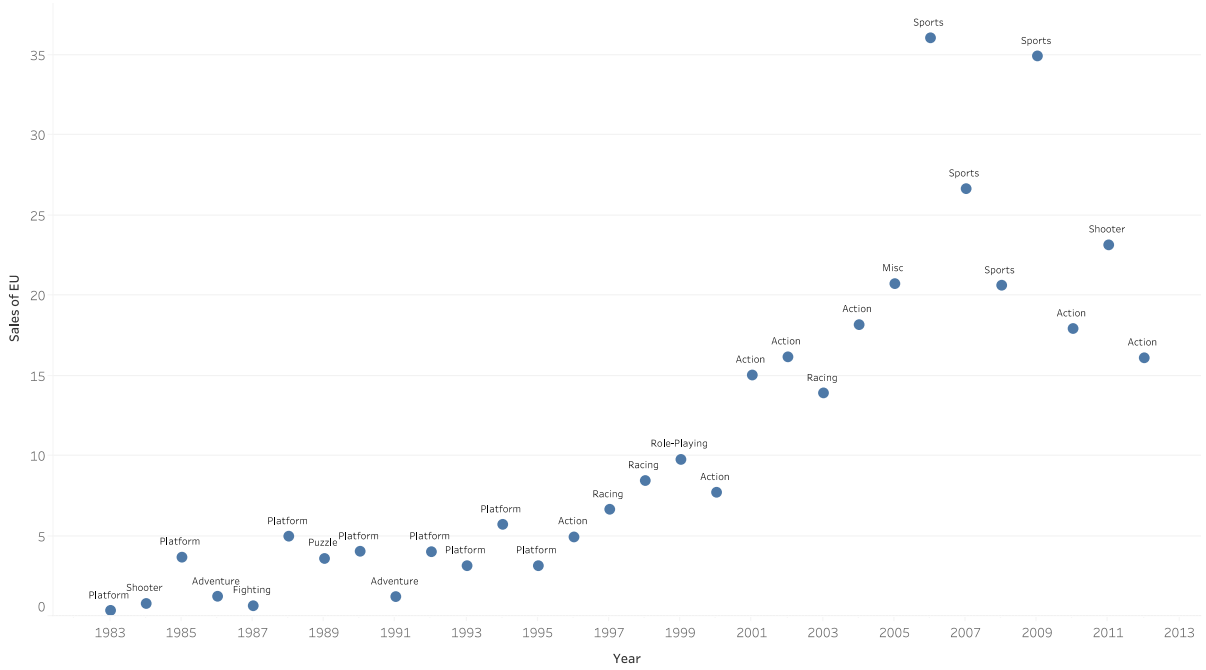

Figure 8 Genre with the max sum of sales per year in Europe

Figure 8 shows the genre with the highest sales in Europe. We can see a similar pattern in the late 1980s and 1990s with platform games. In contrast, from the year 1996 onwards, action games top the sales charts along with sports games.

We will take a closer look at the platforms that have the highest sales per year for video games. This will give us an idea of which console people prefer to use for playing games. Additionally, it will reveal whether people prefer playing games while on the move or in the comfort of their homes. The type of games that can be played are different based on the console, as games on portable platforms tend to be less hardware intensive. Moreover, this information can provide us with insights into the types of games that they prefer playing.

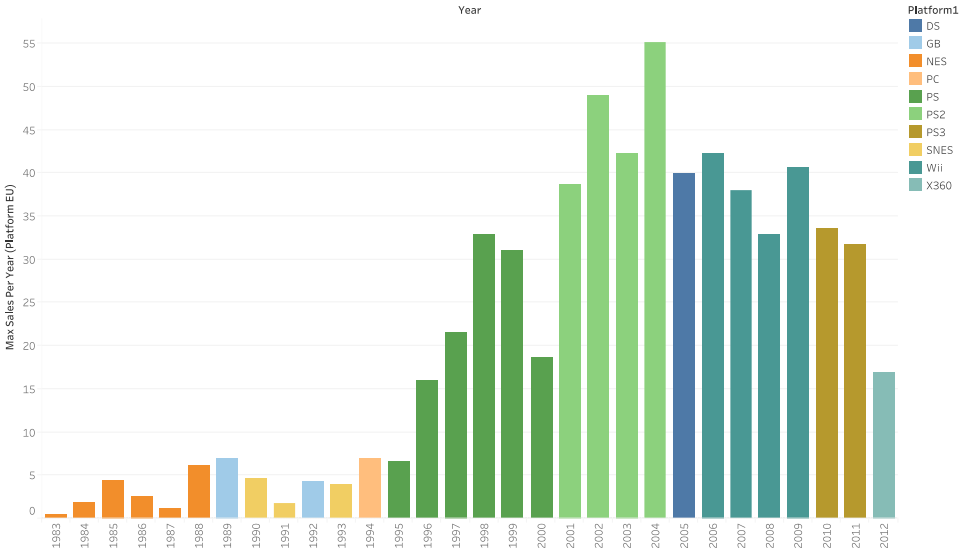

Figure 9 Platforms which have the max sum of sales of games in EU

From Figure 9, we can observe that NES games were quite popular during the late 1980s, while PC gaming was a hit in 1994. However, from the late 1990s up to the early 2000s, games on PlayStation and PlayStation 2 gained a huge following. The period from 2006 to 2009 saw the highest sales in video games on Wii, while the last three years were dominated by PlayStation 3 and Xbox 360, which were the current generation of consoles at that time.

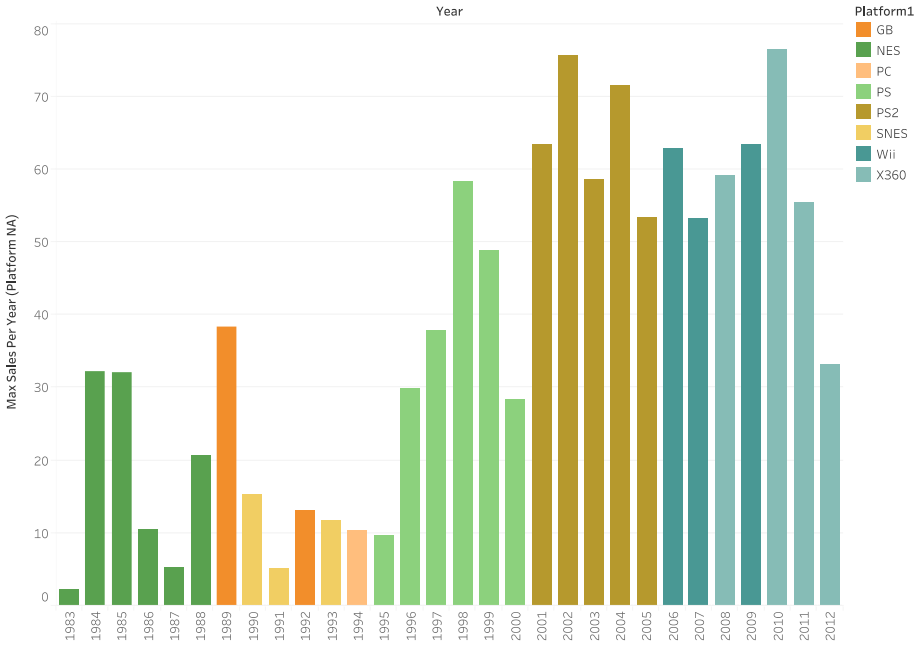

Figure 10 Platforms which have the max sum of sales of games in NA

We can see some similarities with Europe in Figure 10. NES games were also popular in the late 1980s here. PC gaming was also popular in 1994. The same phenomenon happened in the late 1990s and in the early 2000s. The difference lies on the right side of the plot. We do not see PlayStation 3 taking a spot anywhere, instead, we just see Wii and Xbox 360. So, games on Xbox 360 and Wii are popular in the later years.

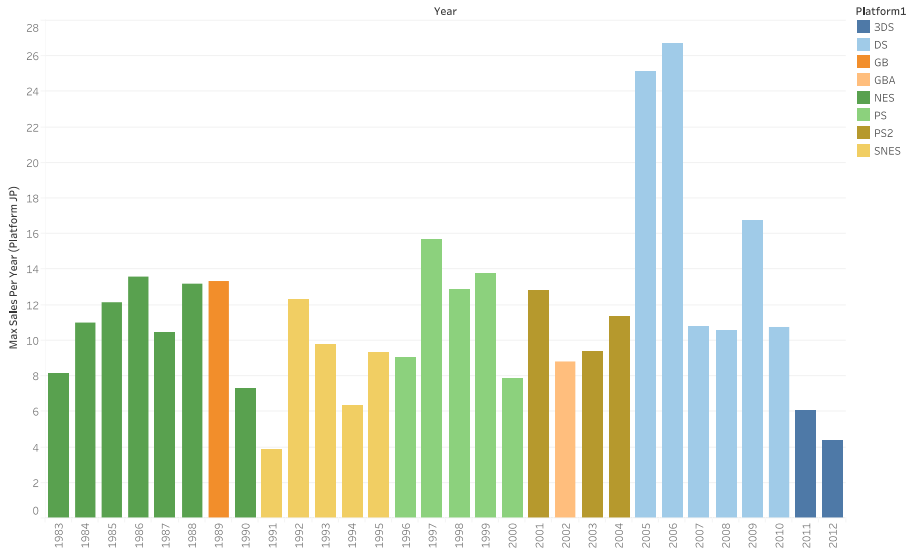

Figure 11 Platforms which have the max sum of sales of games in JP

Lastly, this is the Figure 11 for Japan. We see a similar pattern with NES but PC games are not popular very popular in Japan. Instead, Super Nintendo Entertainment System (SNES) is the one that is popular in the early 1990s. PlayStation and PlayStation 2 games still dominated in the same years as in EU and NA. However, the right of the plot is very different, games on Nintendo DS and Nintendo 3DS are taking the top spot in sales.

It would be interesting to see if the high number of sales corresponds to a high average review. However, the data does not have review scores per region, therefore the review scores are considered to be global. Nevertheless, the next six plots will depict the video game genres and platforms with the highest average review score per year.

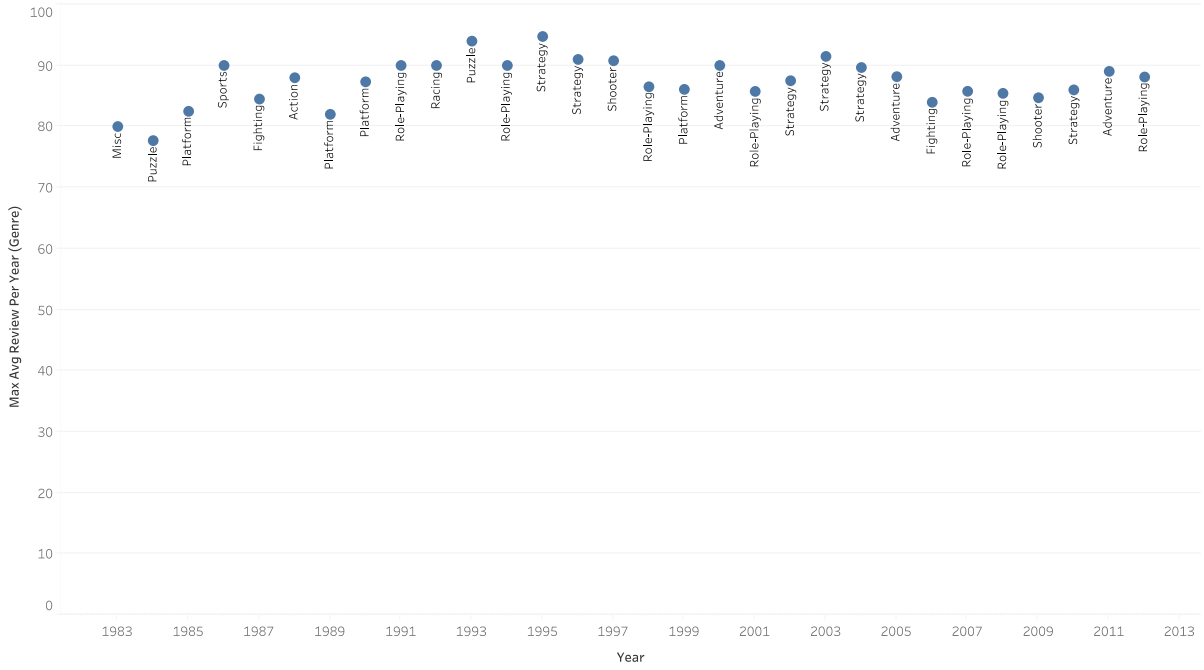

Figure 12 Maximum Avg. Review of video game genre per year

Figure 12 depicts the genres with the highest average review score per year. As we can see, it does not seem to correspond to any of the maximum sums of sales plots. It looks like a high sum of sales does not mean a high average review for video game genres.

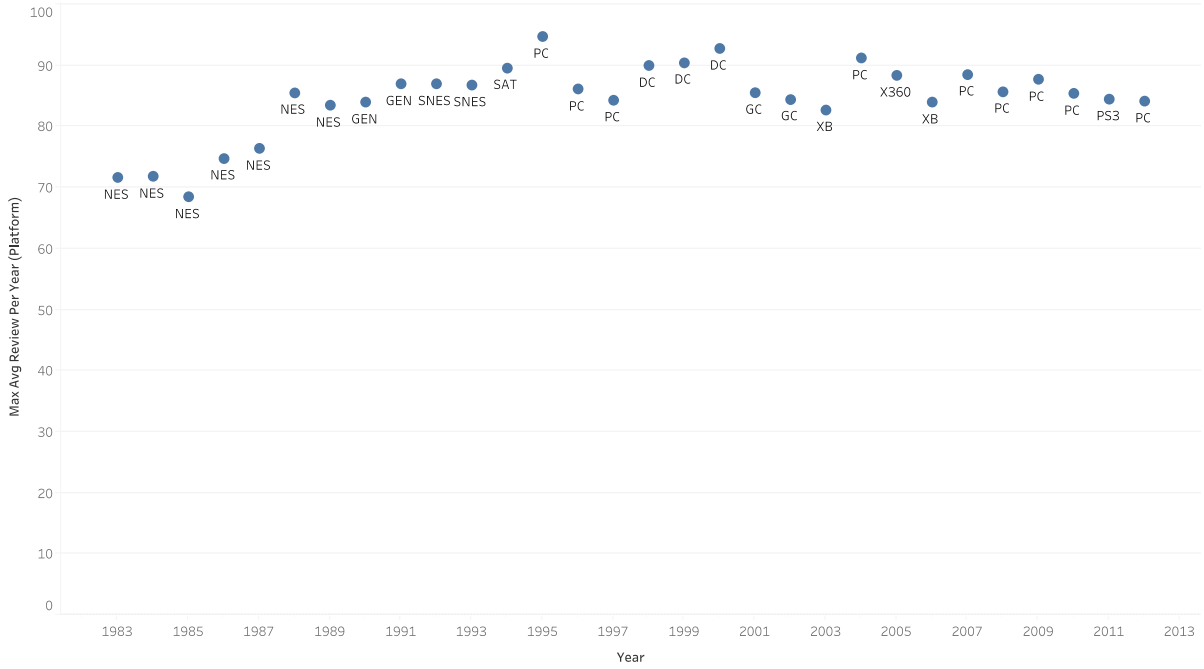

Figure 13 Video game platforms with the highest Avg. review score

Figure 13 shows the video game platforms with the highest average review score. We do not see little correspondence here. NES has the highest sales in the 1980s, so it makes sense to see that the games on the same platforms earned maximum average review scores as well. However, we can see that PC games have the highest average review score in the most recent years, but PC games did not achieve maximum sales anywhere in that timeframe. This could mean that high sales do not mean a high average review.

Conclusion

The data and visualizations presented a clear picture of trends in platforms and genres from 1983 – 2012. What was found interesting is how different the gaming trend was in Japan compared to North America and Europe in the year 2005 and later. The popularity of video games on PS3 and Xbox 360 was expected in recent years, but I was taken by surprise that games on Nintendo DS and 3DS were popular in Japan in that timeframe. Based on this discovery, people in Japan most likely play lighter games on the go, not AAA titles, just as a way to pass the time.

I was a little disappointed by the visualizations I had to use in this project because I felt like there was a lack of variety. For the design aspects, I tried to use plots that were suitable for a general audience, since the topic is about video games, therefore I used line graphs, scatter plots, and bar charts. These basic visualizations should be effortless to understand since we see this type of data visualization quite often. On limitations, I feel like I ended up relying too heavily on bar charts, simply because I could not find a more suitable plot to visualize the data. I still felt like I could have used other ways to visualize them. Given more time to learn other visualization techniques, I would certainly include a wider range of plots, but most of my time was spent learning how to use Tableau to plot my data. In the end, I was satisfied with the plots I was able to make with Tableau, especially the plot about demographics. I was given a chance to learn a new tool for data analysis, so I do not have to rely solely on Jupyter Notebook anymore.

Appendix A: Python Scripts

Data Visualization Project.ipnyb which is used for exploratory data analysis

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


import pandas as pd
import numpy as np


# In[2]:


df = pd.read_csv("Video Games Sales.csv")
df.head()


# In[64]:


df[df['Year'] == 2006]


# In[3]:


df.shape


# The 'Rank' column identifies each game's ranking according to its global sales (highest to lowest). This can help you identify which games are most popular globally.<br>
# The 'Game Title' column contains the name of each video game, which allows you to easily discern one entry from another.<br>
# The 'Platform' column lists the type of platform on which each game was released, e.g., PlayStation 4 or Xbox One, so that you can make comparisons between platforms as well as specific games for each platform.<br>
# The 'Year' column provides an additional way of making year-on-year comparisons and tracking changes over time in global video game sales.<br>
```

```
# In addition, this dataset also contains metadata such as genre ('Genre'), publisher ('Publisher'), and review score ('Review') that add context when considering a particular title's performance in terms of global sales rankings.

# Lastly but no less important are the three variables dedicated exclusively for geographic breakdowns: North America ('North America'), Europe (Europe), Japan (Japan), Rest of World (Rest of World), and Global (Global). This allows us to see how certain regions contribute individually or collectively towards a given title's overall sales figures; by comparing these metrics regionally or collectively an interesting picture arises -- from which inferences about consumer preferences and supplier priorities emerge!

# # Data Exploration

# ## Univariate EDA

# ### Column "Rank"

# In[4]:


df['Rank'].describe()


# In[5]:


df['Rank'].unique()


# All of the values are unique

# ### Column "Game Title"

# In[6]:


df['Game Title'].describe()


# In[7]:


df['Game Title'].nunique()
```

```
# The column 'Game Title' seems to have duplicate values

# In[8]:

df['Game Title'].value_counts()

# In[9]:

df[df['Game Title'] == 'LEGO Batman: The Videogame']

# In[10]:

df[df['Game Title'] == 'FIFA Soccer 08']

# Some video game titles are released for more than one platform

# ### Column 'Platform'

# In[11]:

df['Platform'].describe()

# In[12]:

df['Platform'].unique()

# ### Column 'Year'

# In[13]:
```

```
df['Year'].describe()


# The data type of the column is float64, won't be a problem for analytics.

# In[14]:


# df['Year'] = pd.to_datetime(df['Year'].fillna(0).astype(int), format='%Y')
# df['Year'] = pd.to_datetime(pd.to_numeric(df['Year'], errors='coerce'), errors='coerce',
format='%Y').fillna(df['Year'])
# df.head()


# In[15]:


df.groupby('Year')['Game Title'].count()


# In[61]:


temp1 = pd.DataFrame(df.groupby(['Year', 'Genre'])['North America'].sum())
temp1.to_csv(r'/Users/abimanyu/Desktop/Notes and Docs/COSC/temp1.csv', index=True)


# ### Column 'Genre'

# In[16]:


df['Genre'].describe()


# In[17]:


df['Genre'].unique()


# ### Column 'Publisher'
```

```
# In[18]:


df['Publisher'].describe()


# In[19]:


df['Publisher'].unique()


# ### Null values

# In[20]:


df.isnull().sum()


# We have a small number of null values in the data, we'll remove them for now.

# In[21]:


df1 = df.dropna()
df1.isnull().sum()


# ## Bivariate EDA

# In[22]:


import matplotlib.pyplot as plt
import seaborn as sns


# ### Columns 'Genre' and 'Review'

# In[23]:
```

```
genre_reviews = df1[['Genre', 'Review']].groupby('Genre').mean()
genre_reviews


# In[24]:


plt.bar(genre_reviews.index, genre_reviews['Review'])
plt.xticks(rotation = 45)
plt.show()


# We see that the role-playing genre has the highest average review score.There is not much difference between the average review scores of the different genres.

# ### Column 'Platform' and 'Review'

# In[25]:


platform_reviews = df1[['Platform', 'Review']].groupby('Platform').mean()
platform_reviews


# In[26]:


plt.bar(platform_reviews.index, platform_reviews['Review'])
plt.xticks(rotation = 60)
plt.show()


# Video games on Sega Dreamcast has the highest average review score

# ### Column 'Publisher' and 'Review'

# In[27]:


publisher_reviews = df1[['Publisher', 'Review']].groupby('Publisher').mean().sort_values('Review', axis=0,
ascending=False)
publisher_reviews
```

```
# In[28]:


plt.bar(publisher_reviews.head().index, publisher_reviews.head()['Review'])
plt.xticks(rotation = 45)
plt.show()


# Here are the top 5 publishers of video games with the highest average review score.

# ### Column 'Genre' and 'North America'

# Here, we are assuming that the sales are not percentages, but profits in dollars.

# In[29]:


genre_NA = df1[['Genre', 'North America']].groupby('Genre').sum().sort_values('North America', ascending=False)
genre_NA


# In[30]:


plt.bar(genre_NA.head().index, genre_NA.head()['North America'])
plt.xticks(rotation = 45)
plt.show()


# Here are the top 5 most profitable video game genres in North America

# ### Colum 'Genre' and 'Europe'

# In[31]:


genre_EU = df1[['Genre', 'Europe']].groupby('Genre').sum().sort_values('Europe', ascending=False)
genre_EU
```

```
# In[32]:


plt.bar(genre_EU.head().index, genre_EU.head()['Europe'])
plt.xticks(rotation = 45)
plt.show()


# The top 5 most profitable genre in Europe are the same as in North America

# ### Column 'Genre' and 'Japan'

# In[33]:


genre_JP = df1[['Genre', 'Japan']].groupby('Genre').sum().sort_values('Japan', ascending=False)
genre_JP


# In[34]:


plt.bar(genre_JP.head().index, genre_JP.head()['Japan'])
plt.xticks(rotation = 45)
plt.show()


# We see that the most popular genre of video games in Japan are different from NA and EU. Here, Role-Playing games takes the top spot.

# ## Exploring Trends in Video Games

# ### Popular Video Games Each Year In NA

# In[35]:


yearly_NA = df1.loc[df1.groupby('Year')['North America'].idxmax()]
yearly_NA


# In[ ]:
```

```
years = yearly_NA['Year'].tolist()
sales = yearly_NA['North America'].tolist()
titles = yearly_NA['Game Title'].tolist()

plt.scatter(yearly_NA['Year'], yearly_NA['North America'])

for i, txt in enumerate(sales):
    plt.annotate(titles, (years[i], sales[i]))


# In[ ]:


df[df['Year'] == 1983]


# ### Popular Video Games Each Year In EU

# In[38]:


df1.loc[df1.groupby('Year')['Europe'].idxmax()]


# ### Popular Video Games Each Year In JP

# In[39]:


df1.loc[df1.groupby('Year')['Japan'].idxmax()]
```