Exploratory Data Anaylsis based on video game sales

BOSSGANABATHI A/L VELLOW (B032110418)

VIMALRAJ A/L GANESON (B032110053)

VIKNESH A/L MANIMARAN (B032010195)

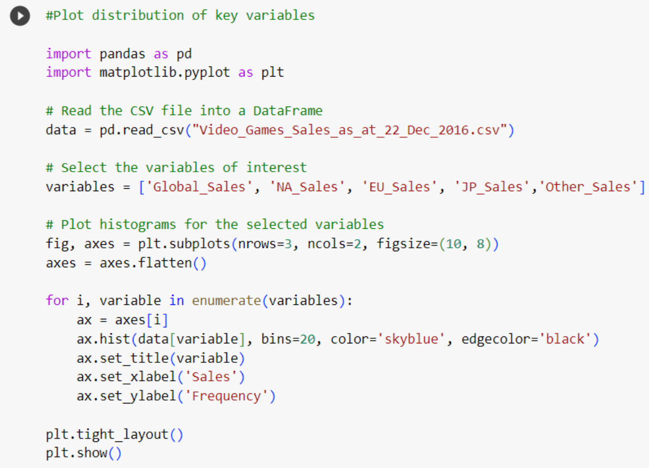
***Abstract*—The video game industry has witnessed remarkable growth and transformation in recent years. In this data analysis study, we explore a comprehensive dataset of video game sales to uncover valuable insights and identify significant trends. Our objectives include understanding the market dynamics, exploring regional variations, and examining the performance of different game genres. We utilized a dataset encompassing various attributes such as game title, platform, genre, release year, and global sales. Through descriptive analysis and visualization techniques, we examined key metrics such as total sales, regional sales distribution, and genre popularity. Our findings revealed intriguing patterns and trends within the video game industry. We observed the dominance of certain platforms and genres, as well as variations in sales across different regions. Additionally, we explored the relationship between game release year and sales performance, shedding light on the impact of technological advancements and evolving consumer preferences. This analysis provides valuable insights for stakeholders in the gaming industry, including game developers, publishers, and market researchers. The results can aid in strategic decision-making, game development planning, and identifying market opportunities.**

Keywords*—video games, data analysis, sales trends, regional variations, genre popularity

I. Introduction

The video game industry has experienced rapid growth and significant changes in recent years, becoming a major player in the entertainment sector. As the popularity of video games continues to rise, understanding the underlying trends and dynamics within the industry becomes crucial for stakeholders such as game developers, publishers, and market researchers. Data analysis provides a powerful tool to uncover valuable insights and identify patterns that can inform decision-making and strategy development. In this study, we delve into a comprehensive dataset of video game sales to perform a thorough analysis of the industry. The dataset includes information on game titles, platforms, genres, release years, and global sales. By exploring this rich dataset, we aim to gain a deeper understanding of the video game market, identify significant trends, and unveil key factors that contribute to the success of video games. Our analysis encompasses several important aspects of the industry. Firstly, we examine the overall sales performance of video games, analyzing total sales figures and identifying top-selling games. Additionally, we investigate regional variations in game sales, uncovering differences in consumer preferences and market dynamics across different regions. Furthermore, we explore the popularity of different game genres, providing insights into the types of games that resonate most with consumers. By utilizing data visualization techniques and statistical analysis, we aim to present a comprehensive picture of the video game industry's landscape. Our findings will shed light on the evolving trends, patterns, and opportunities within the industry, enabling stakeholders to make informed decisions and adapt their strategies accordingly. Through this study, we contribute to the existing body of knowledge in the field of video game analysis and provide valuable insights for industry professionals. The analysis not only helps in understanding the current state of the video game market but also serves as a foundation for future research and exploration of emerging trends and dynamics. Overall, this data-driven analysis offers a comprehensive and detailed examination of the video game industry, facilitating a deeper understanding of its key drivers, trends, and opportunities.

A. Plot distribution of key variables



```
#Plot distribution of key variables

import pandas as pd
import matplotlib.pyplot as plt

# Read the CSV file into a DataFrame
data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

# Select the variables of interest
variables = ['Global_Sales', 'NA_Sales', 'EU_Sales', 'JP_Sales','Other_Sales']

# Plot histograms for the selected variables
fig, axes = plt.subplots(nrows=3, ncols=2, figsize=(10, 8))
axes = axes.flatten()

for i, variable in enumerate(variables):
    ax = axes[i]
    ax.hist(data[variable], bins=20, color='skyblue', edgecolor='black')
    ax.set_title(variable)
    ax.set_xlabel('Sales')
    ax.set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```

Figure 1: plot distribution code

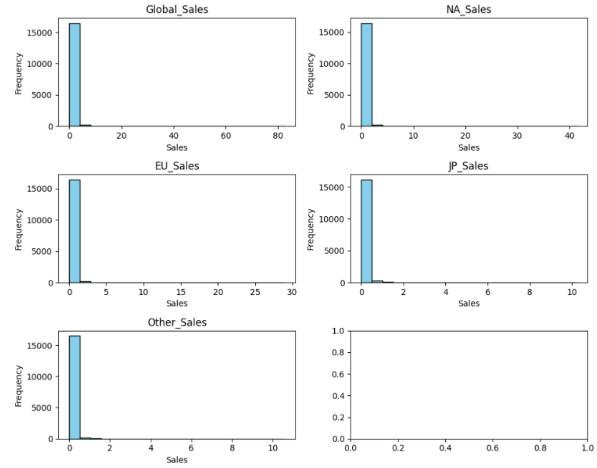
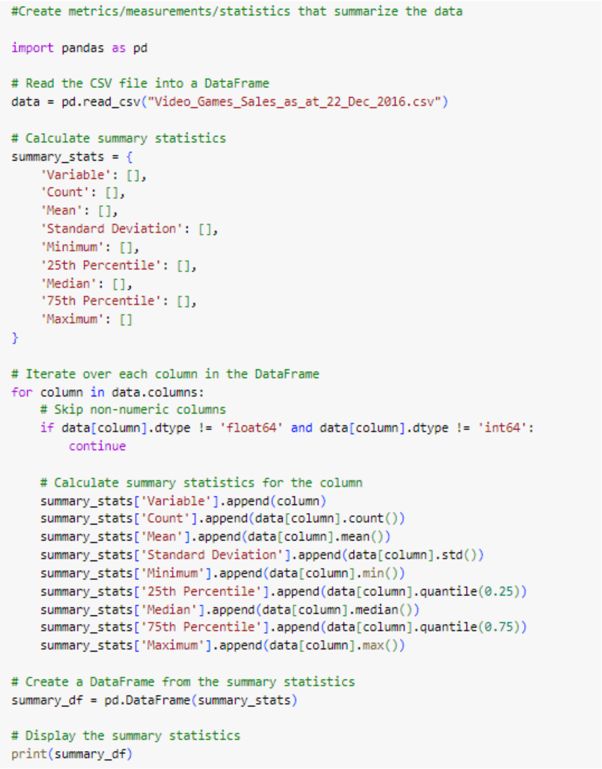


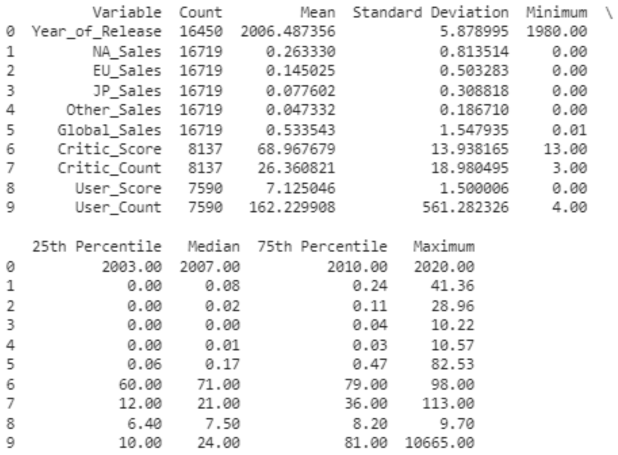
Figure 2: plot distribution graph

B. *Create metrics/measurements/statistics that summarize the data*



```
#Create metrics/measurements/statistics that summarize the data

import pandas as pd

# Read the CSV file into a DataFrame
data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

# Calculate summary statistics
summary_stats = {
    'Variable': [],
    'Count': [],
    'Mean': [],
    'Standard Deviation': [],
    'Minimum': [],
    '25th Percentile': [],
    'Median': [],
    '75th Percentile': [],
    'Maximum': []
}

# Iterate over each column in the DataFrame
for column in data.columns:
    # Skip non-numeric columns
    if data[column].dtype != 'float64' and data[column].dtype != 'int64':
        continue

    # Calculate summary statistics for the column
    summary_stats['Variable'].append(column)
    summary_stats['Count'].append(data[column].count())
    summary_stats['Mean'].append(data[column].mean())
    summary_stats['Standard Deviation'].append(data[column].std())
    summary_stats['Minimum'].append(data[column].min())
    summary_stats['25th Percentile'].append(data[column].quantile(0.25))
    summary_stats['Median'].append(data[column].median())
    summary_stats['75th Percentile'].append(data[column].quantile(0.75))
    summary_stats['Maximum'].append(data[column].max())

# Create a DataFrame from the summary statistics
summary_df = pd.DataFrame(summary_stats)

# Display the summary statistics
print(summary_df)
```

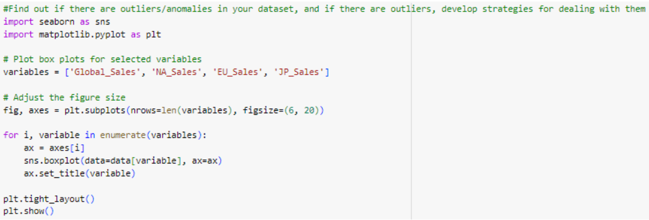
Figure 3: statistic summary code



```
          Variable  Count         Mean  Standard Deviation  Minimum  \
0  Year_of_Release  16450  2006.487356            5.878995  1980.00   
1         NA_Sales  16719     0.263330            0.813514     0.00   
2         EU_Sales  16719     0.145025            0.503283     0.00   
3         JP_Sales  16719     0.077602            0.308818     0.00   
4      Other_Sales  16719     0.047332            0.186710     0.00   
5     Global_Sales  16719     0.533543            1.547935     0.01   
6     Critic_Score   8137    68.967679           13.938165    13.00   
7     Critic_Count   8137    26.360821           18.980495     3.00   
8       User_Score   7590     7.125046            1.500006     0.00   
9       User_Count   7590   162.229908          561.282326     4.00   

   25th Percentile   Median  75th Percentile   Maximum  
0          2003.00  2007.00          2010.00   2020.00  
1             0.00     0.08             0.24     41.36  
2             0.00     0.02             0.11     28.96  
3             0.00     0.00             0.04     10.22  
4             0.00     0.01             0.03     10.57  
5             0.06     0.17             0.47     82.53  
6            60.00    71.00            79.00     98.00  
7            12.00    21.00            36.00    113.00  
8             6.40     7.50             8.20      9.70  
9            10.00    24.00            81.00  10665.00  
```

Figure 4: statistic summary output

C. *Find out if there are outliers/anomalies in your dataset, and if there are outliers, develop strategies for dealing with them*



```
#Find out if there are outliers/anomalies in your dataset, and if there are outliers, develop strategies for dealing with them
import seaborn as sns
import matplotlib.pyplot as plt

# Plot box plots for selected variables
variables = ['Global_Sales', 'NA_Sales', 'EU_Sales', 'JP_Sales']

# Adjust the figure size
fig, axes = plt.subplots(nrows=len(variables), figsize=(6, 20))

for i, variable in enumerate(variables):
    ax = axes[i]
    sns.boxplot(data=data[variable], ax=ax)
    ax.set_title(variable)

plt.tight_layout()
plt.show()
```

Figure 5: Outlier graph code

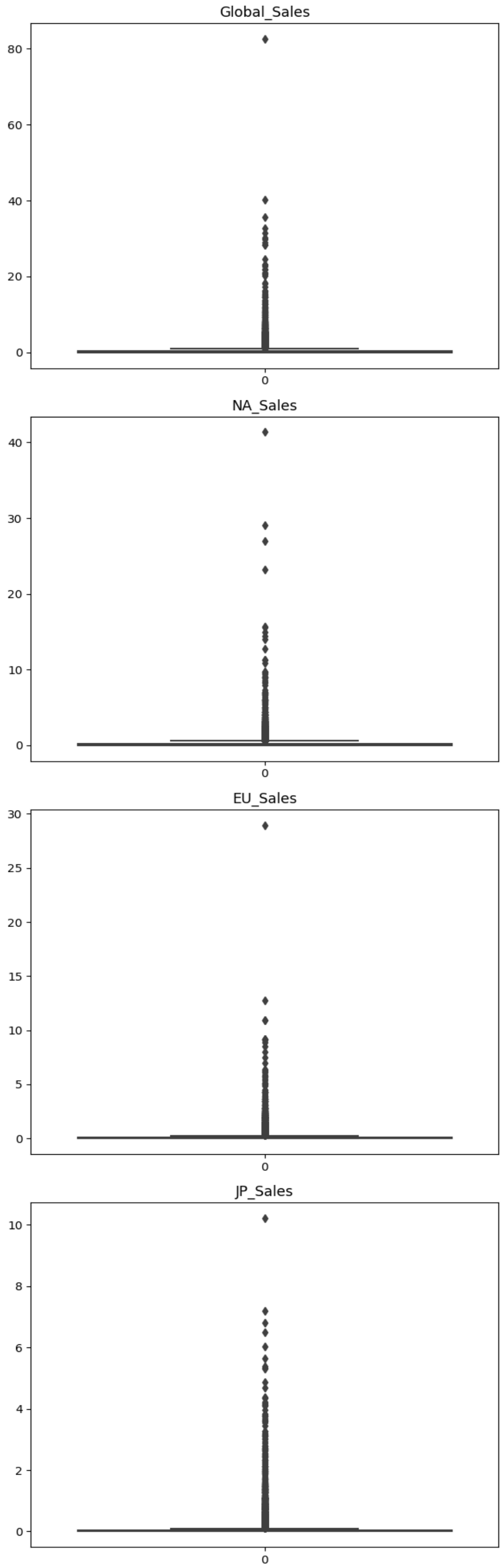
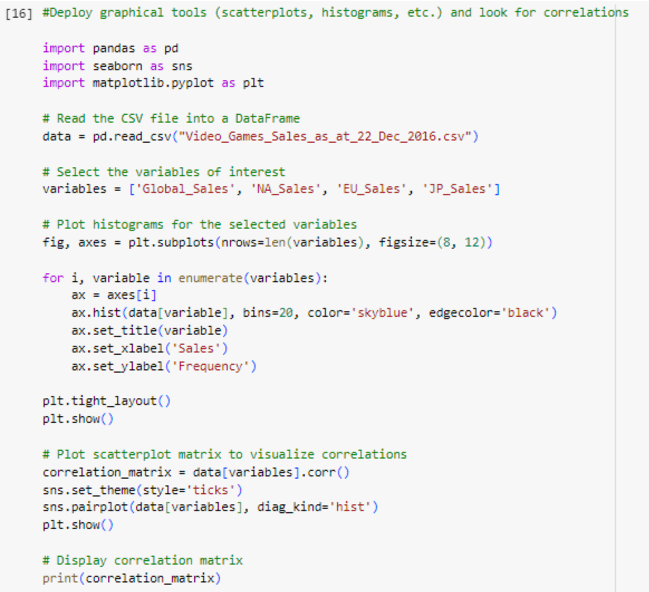


Figure 6: Outlier graph

D. *Deploy graphical tools (scatterplots, histograms, etc.) and look for correlations*



```
[16] #Deploy graphical tools (scatterplots, histograms, etc.) and look for correlations

     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Read the CSV file into a DataFrame
     data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

     # Select the variables of interest
     variables = ['Global_Sales', 'NA_Sales', 'EU_Sales', 'JP_Sales']

     # Plot histograms for the selected variables
     fig, axes = plt.subplots(nrows=len(variables), figsize=(8, 12))

     for i, variable in enumerate(variables):
         ax = axes[i]
         ax.hist(data[variable], bins=20, color='skyblue', edgecolor='black')
         ax.set_title(variable)
         ax.set_xlabel('Sales')
         ax.set_ylabel('Frequency')

     plt.tight_layout()
     plt.show()

     # Plot scatterplot matrix to visualize correlations
     correlation_matrix = data[variables].corr()
     sns.set_theme(style='ticks')
     sns.pairplot(data[variables], diag_kind='hist')
     plt.show()

     # Display correlation matrix
     print(correlation_matrix)
```

Figure 7: Scatter plot and histogram code for the correlation

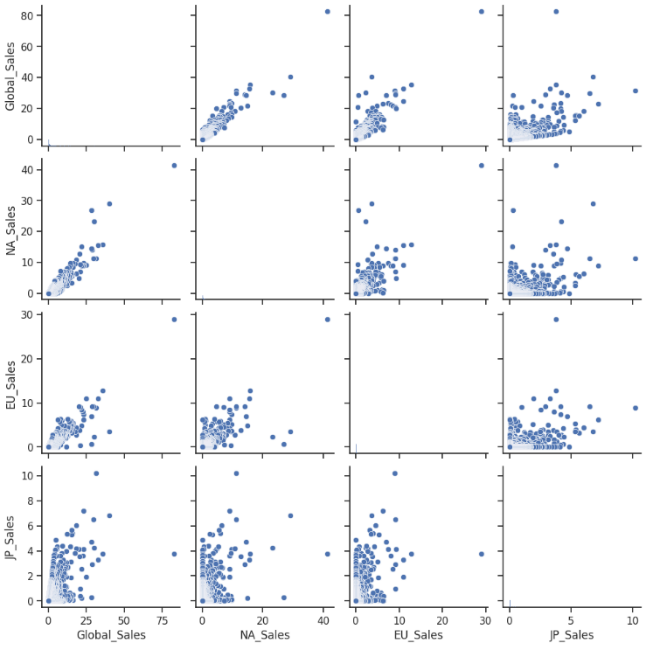


Figure 8: Scatter plot

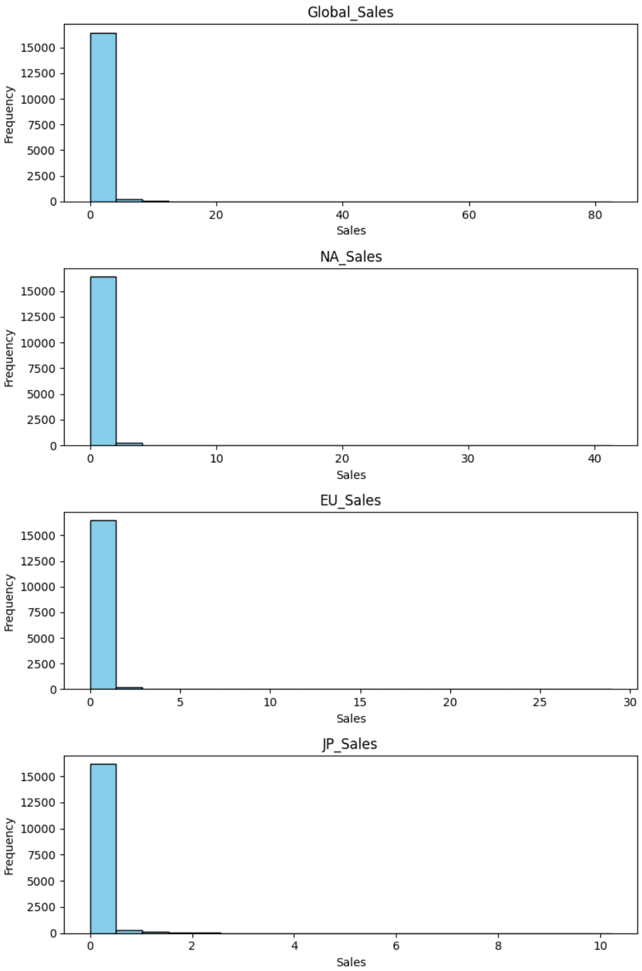
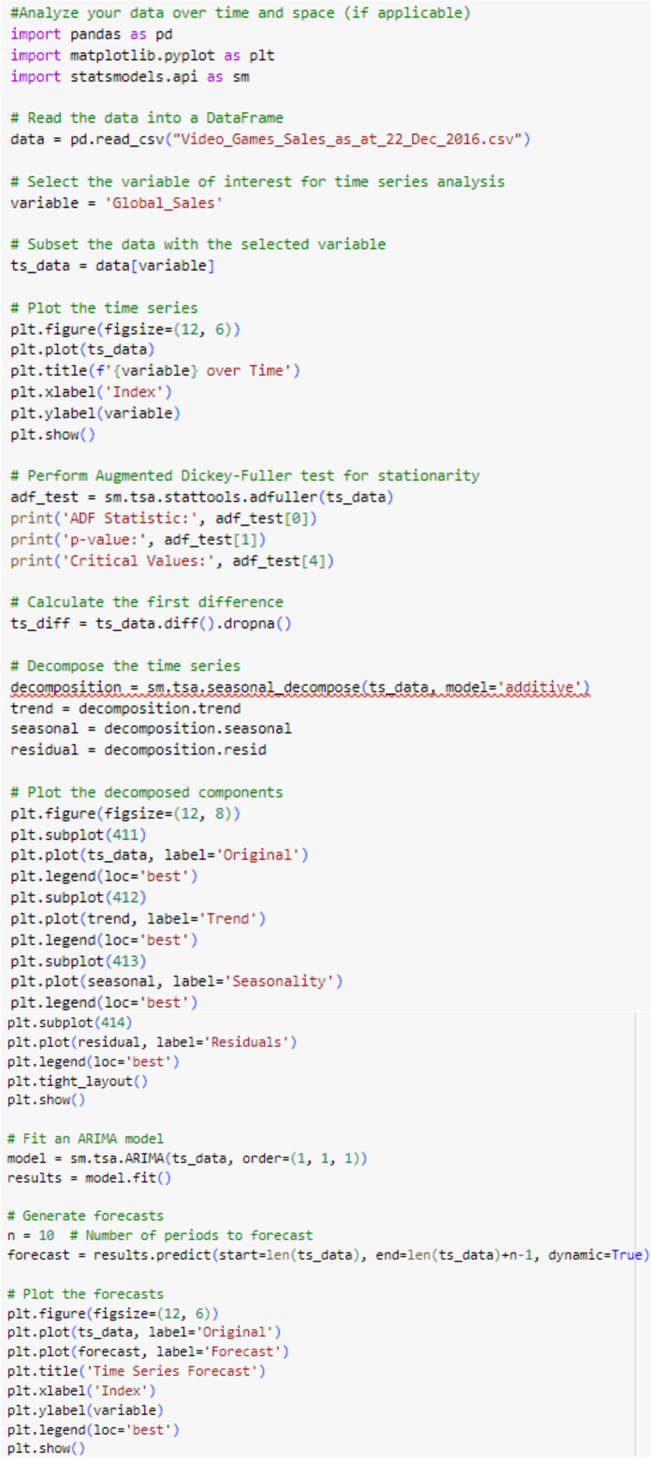


Figure 9: Histogram plot for the correlation

E. Analyze your data over time and space (if applicable)



```
#Analyze your data over time and space (if applicable)
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Read the data into a DataFrame
data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

# Select the variable of interest for time series analysis
variable = 'Global_Sales'

# Subset the data with the selected variable
ts_data = data[variable]

# Plot the time series
plt.figure(figsize=(12, 6))
plt.plot(ts_data)
plt.title(f'{variable} over Time')
plt.xlabel('Index')
plt.ylabel(variable)
plt.show()

# Perform Augmented Dickey-Fuller test for stationarity
adf_test = sm.tsa.stattools.adfuller(ts_data)
print('ADF Statistic:', adf_test[0])
print('p-value:', adf_test[1])
print('Critical Values:', adf_test[4])

# Calculate the first difference
ts_diff = ts_data.diff().dropna()

# Decompose the time series
decomposition = sm.tsa.seasonal_decompose(ts_data, model='additive')
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

# Plot the decomposed components
plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(ts_data, label='Original')
plt.legend(loc='best')
plt.subplot(412)
plt.plot(trend, label='Trend')
plt.legend(loc='best')
plt.subplot(413)
plt.plot(seasonal, label='Seasonality')
plt.legend(loc='best')
plt.subplot(414)
plt.plot(residual, label='Residuals')
plt.legend(loc='best')
plt.tight_layout()
plt.show()

# Fit an ARIMA model
model = sm.tsa.ARIMA(ts_data, order=(1, 1, 1))
results = model.fit()

# Generate forecasts
n = 10  # Number of periods to forecast
forecast = results.predict(start=len(ts_data), end=len(ts_data)+n-1, dynamic=True)

# Plot the forecasts
plt.figure(figsize=(12, 6))
plt.plot(ts_data, label='Original')
plt.plot(forecast, label='Forecast')
plt.title('Time Series Forecast')
plt.xlabel('Index')
plt.ylabel(variable)
plt.legend(loc='best')
plt.show()
```

Figure 10: analyze the data over time and space.

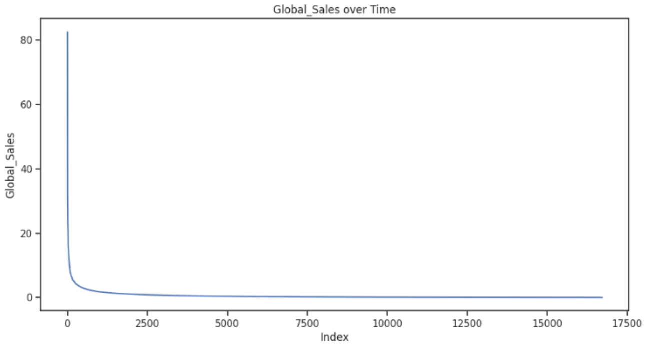
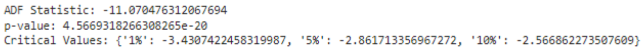


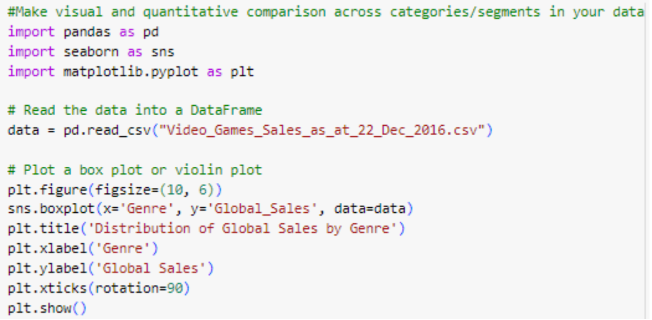
Figure 11: analyze the global sales over time.



```
ADF Statistic: -11.070476312067694
p-value: 4.5669318266308265e-20
Critical Values: {'1%': -3.4307422458319987, '5%': -2.861713356967272, '10%': -2.566862273507609}
```

Figure 12: Showing the statistical values p-values.

F. Make visual and quantitative comparison across categories/segments in your data



```
#Make visual and quantitative comparison across categories/segments in your data
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Read the data into a DataFrame
data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

# Plot a box plot or violin plot
plt.figure(figsize=(10, 6))
sns.boxplot(x='Genre', y='Global_Sales', data=data)
plt.title('Distribution of Global Sales by Genre')
plt.xlabel('Genre')
plt.ylabel('Global Sales')
plt.xticks(rotation=90)
plt.show()
```

Figure 13: distribution of global sales by genre code

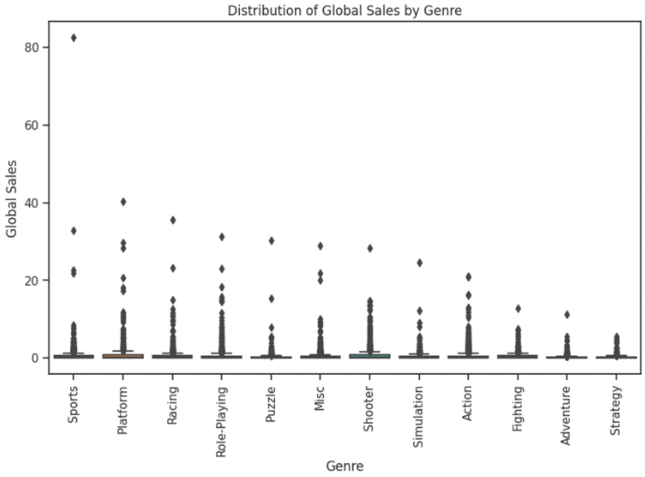
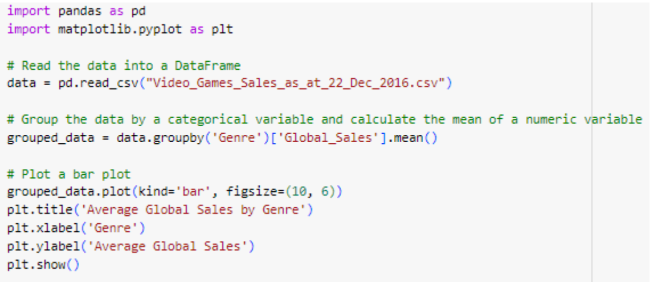


Figure 14: distribution of global sales plot

G. Look for, describe and interpret any patterns you find



```
import pandas as pd
import matplotlib.pyplot as plt

# Read the data into a DataFrame
data = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")

# Group the data by a categorical variable and calculate the mean of a numeric variable
grouped_data = data.groupby('Genre')['Global_Sales'].mean()

# Plot a bar plot
grouped_data.plot(kind='bar', figsize=(10, 6))
plt.title('Average Global Sales by Genre')
plt.xlabel('Genre')
plt.ylabel('Average Global Sales')
plt.show()
```

Figure 15: Average global sales by genre code

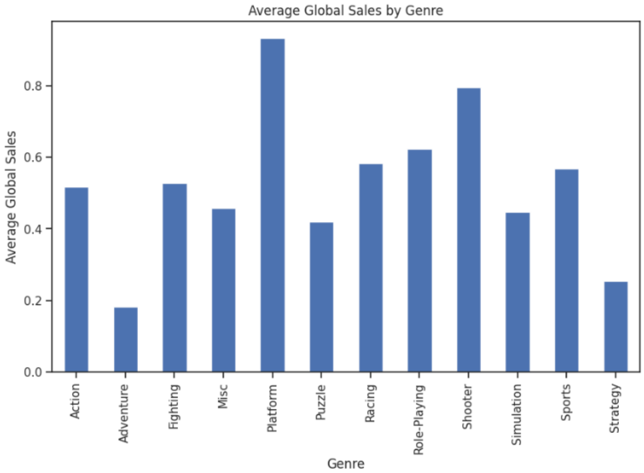


Figure 16: Plot average global sales based on genre.

II. Explanation

In figure 16 shows the relationship between genre and average global sales obviously platform games are highest selling games. Adventure games have the lowest sales recorded. From this figure 16 the upcoming companies or new game developers can visualize the current demand in the marketplace.

III. Discusssion

The problem we came across during data cleaning and data reshaping. We used preprocessing methods to remove all the null values from the datasets. Make sure the datasets are clean and precise.

IV. Acknowledgments

We would like to express our gratitude to the contributors and organizations who made this video game data analysis possible. We acknowledge the efforts of the data providers and the creators of the original dataset used in this study. Their diligent data collection and documentation have been instrumental in enabling us to conduct our analysis. We would also like to extend our appreciation to the developers and researchers in the field of data analysis and visualization. Their valuable contributions, innovative tools, and open-source libraries have greatly facilitated our analysis process and enhanced the quality of our findings. Furthermore, we would like to thank our team members and colleagues who provided support and guidance throughout the project. Their valuable insights, discussions, and feedback have been instrumental in shaping the direction of our analysis and improving the overall quality of our work. Finally, we express our heartfelt thanks to the academic and research community for their continuous efforts in advancing knowledge in the field of data analysis and its application to various domains. Their dedication to research and their willingness to share knowledge have been invaluable in our journey. We acknowledge that this work is built upon the contributions of many individuals and organizations, and we are grateful for their contributions and support.

References

[1] Prasad Patil "What is exploratory data analysis" published by Toward Data Science, Mar 24, 2015, https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15