CCT College Dublin

Assessment Cover Page

Module Title:	Statistical techniques for Data Analysis
Assessment Title:	Integrated CA
Lecturer Name:	Aldana Louzan
Student Full Name:	Raphael Fernandes Gomes
Student Number:	2022091
Assessment Due Date:	27/05/2022
Date of Submission:	27/05/2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Index

Introduction

This report is the continuation of CA1 and gives a small overview of Hypothesis Test through research and practical test, application of correlation analysis and construction of a Linear Regression model with a data set in JupyterNotebook.

The purpose of the three sections is to better understand Hypothesis test, Correlation/Causation, and the behaviour of variables with the algorithm Linear Regression and apply them to the data set to perform analysis.

Introduction of the Data set

The data set that I choose to deal with is a data set containing thousands of games with their names, the number of Sales in North America (NA) Europe (EU), Japan (JP), Others (Rest of the World), and Global Sales (Total worldwide), Year of release, Genre, Platforms were the games released, Publisher of the Game and Review.

I chose this data set as a fan of games.

Data Dictionary

Col	Full Name of Variables	Definition of Variables	Type of Variables	
			Qualitative/Quantitative	Categorical Disc/Contin
A	Rank	Rank of overall sales	Qualitative	Categorical
B	Game_Title	The games name	Qualitative	Categorical
C	Platform	Platform of the games release (i.e. PC, PS4, Xbox, etc.)	Qualitative	Categorical
D	Year	Year of the game's release	Quantitative	Continuous
E	Genre	Genre of the games (i.e. Racing, Sports, Action, etc.)	Qualitative	Categorical
F	Publisher	Publisher of the game	Qualitative	Categorical
G	North_America	Sales in North America	Quantitative	Continuous
H	Europe	Sales in Europe	Quantitative	Continuous
I	Japan	Sales in Japan	Quantitative	Continuous
J	Rest_of_World	Sales in the rest of the world	Quantitative	Continuous
K	Global_Sales	Total Worldwide sales	Quantitative	Continuous
L	Review	Total Review	Quantitative	Continuous

Importing Data and Data Review

Importing and checking all the information about the content in my dataset, (Data type) of each of columns, memory that the Data frame occupies, number of rows, columns, memory usage.

```
games_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1907 entries, 0 to 1906
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Rank           1907 non-null   int64
 1   Game_Title     1907 non-null   object
 2   Platform       1907 non-null   object
 3   Year           1878 non-null   float64
 4   Genre          1907 non-null   object
 5   Publisher      1905 non-null   object
 6   North_America  1907 non-null   float64
 7   Europe         1907 non-null   float64
 8   Japan          1907 non-null   float64
 9   Rest_of_World  1907 non-null   float64
 10  Global_Sales   1907 non-null   float64
 11  Review         1907 non-null   float64
dtypes: float64(7), int64(1), object(4)
memory usage: 178.9+ KB
```

```
games_df.head()
```

	Rank	Game_Title	Platform	Year	Genre	Publisher	North_America	Europe	Japan	Rest_of_World	Global_Sales	Review
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	40.43	28.39	3.77	8.54	81.12	76.28
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	91.00
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	14.50	12.22	3.63	3.21	33.55	82.07
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	14.82	10.51	3.18	3.01	31.52	82.65
4	5	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26	88.00

```
games_df.tail()
```

	Rank	Game_Title	Platform	Year	Genre	Publisher	North_America	Europe	Japan	Rest_of_World	Global_Sales	Review
1902	1903	Lizzie McGuire 2: Lizzie Diaries	GBA	2004.0	Action	Disney Interactive Studios	0.60	0.22	0.00	0.01	0.83	55.00
1903	1904	Xenoblade Chronicles	Wii	2010.0	Role-Playing	Nintendo	0.39	0.22	0.16	0.07	0.83	91.74
1904	1905	SingStar Abba	PS3	2008.0	Misc	Sony Computer Entertainment	0.25	0.44	0.00	0.14	0.83	73.00
1905	1906	FIFA Soccer World Championship	PS2	2000.0	Sports	Electronic Arts	0.27	0.21	0.28	0.07	0.83	73.00
1906	1907	WWE SmackDown vs. Raw 2011	X360	2010.0	Fighting	THQ	0.42	0.32	0.00	0.09	0.83	82.00

```
games_df.nunique()
```

```
Rank             1907
Game_Title       1519
Platform           22
Year               30
Genre              12
Publisher          94
North_America     375
Europe            273
Japan             218
Rest_of_World     129
Global_Sales      479
Review            734
dtype: int64
```

```
games_df.describe()
```

	Rank	Year	North_America	Europe	Japan	Rest_of_World	Global_Sales	Review
count	1907.0000	1878.000000	1907.000000	1907.000000	1907.000000	1907.000000	1907.000000	1907.000000
mean	954.0000	2003.766773	1.258789	0.706675	0.317493	0.206471	2.489240	79.038977
std	550.6478	5.895369	1.956560	1.148904	0.724945	0.343093	3.563159	10.616899
min	1.0000	1983.000000	0.000000	0.000000	0.000000	0.000000	0.830000	30.500000
25%	477.5000	2000.000000	0.510000	0.230000	0.000000	0.060000	1.110000	74.000000
50%	954.0000	2005.000000	0.810000	0.440000	0.020000	0.130000	1.530000	81.000000
75%	1430.5000	2008.000000	1.375000	0.810000	0.300000	0.220000	2.540000	86.230000
max	1907.0000	2012.000000	40.430000	28.390000	7.200000	8.540000	81.120000	97.000000

First section: Hypothesis Test & Research

Metacritic is a website that aggregates reviews of films, TV shows, music albums, video games and formerly, books. For each product, the scores from each review are averaged.

The Site has a certain relevance in terms of Video Games Review, as it has sometimes been questioned that the Review given by the site would be or could affect the result of games sales.

One of the biggest video game platforms is Nintendo, a Japanese game producer.

As a video game lover, I conducted a survey on the site to find out if Nintendo games are highly rated and if Nintendo games are worth buying.

According to Metacritic, Nintendo Platform games have an average rating of 76% (Generally Favourable), which means that Nintendo games are well received by players and are worth buying.

However, many players do not trust the data presented by Metacritic, as it often has a big difference compared to other review sites like IGN or Eurogamer, which tend to follow the same proportion in the ratings results.

So, I decided to get from another website that aggregates reviews, to realize if the Metacritic average score can be reliable or not.

I will conduct my hypothesis test to find out if the average Nintendo games rating by the Metacritic website is reliable if the average Nintendo games rating from the other website is the same as what was presented by the Metacritic.

To carry out this test, I will take a sample of 20 Nintendo games review from the site "VGChartz" where the dataset that I am using comes from to know if the average of the games is really 76% or if it is different from 76%.[5][6]

Executing the Hypothesis Test

```
Nintendo_Total = games_df[games_df["Publisher"]=="Nintendo"]
Nintendo_Total.head(5)
```

	Rank	Game Title	Platform	Year	Genre	Publisher	North America	Europe	Japan	Rest of World	Global	Review
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	40.43	28.39	3.77	8.54	81.12	76.28
1	2	Super Mario Bros.	Nintendo NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	91.00
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	14.50	12.22	3.63	3.21	33.55	82.07
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	14.82	10.51	3.18	3.01	31.52	82.65
4	5	Tetris	GameBoy Color	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26	88.00

```
NS = list(Nintendo_Total['Review'])
NS
```

```
[76.28,
 91.0,
 82.07,
 82.65,
 88.0,
 90.0,
 61.64,
 84.0,
 88.18,
 85.0,
 89.0,
 81.2,
 91.34,
 80.83,
 94.0,
 78.05,
 82.0,
 90.0,
 93.0,
```

Sample

```
random.seed(20)

Nintendo_Sample = random.sample(NS, 20)
Nintendo_Sample
```

```
[81.0,
 86.0,
 86.0,
 77.0,
 75.0,
 89.0,
 80.83,
 83.0,
 85.0,
 94.59,
 88.0,
 77.0,
 69.0,
 81.0,
 91.0,
 85.0,
 76.0,
 84.0,
 85.0,
 79.0]
```

As I do not know the standard deviation of the population, I will apply the T-Test and calculate the mean and standard deviation of the sample that will be taken to execute the test.

Calculating Mean and Standard Deviation of the sample.

```
np.mean(Nintendo_Sample)
```

```
82.62100000000001
```

```
np.std(Nintendo_Sample)
```

```
5.9040840949295434
```

```
#Mean of the Sample
x̄ = np.mean(Nintendo_Sample)
```

```
#Mean of Population
μ = 76
```

```
#Standart Deviation of the Sample
std = np.std(Nintendo_Sample)
```

```
#Number of the Sample
n = 20
```

```
raiz = np.sqrt(n)
```

Variables: μ = 76; $\bar{x}$ = 82.621; s = 5.904; n = 20

Hypothesis

HT0: It is believed that Nintendo Games has a 76% average score.

HT1: It is believed that Nintendo Games has a average score different of 76%.

HT0: μ = 76%

HT1: μ ≠ 76%

Is there enough evidence from this sample that the average score of Nintendo Games have an average score different from Metacritic?

Two Tailed T-test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

A t-test is a hypothesis-testing technique for analysing the means of one or two populations. A t-test can be used to see if a single group differs from a known value (a one-sample t-test), if two groups differ from each other (an independent two-sample t-test), or if there is a significant difference in paired measurements (an independent two-sample t-test) (a paired, or dependent samples t-test).[3][4]

For my hypothesis, I will apply the two tailed T-test, once my HT1 is different from the HT0.

```
t = (x-μ)/(std/raiz)
t
```

```
5.015174526982345
```

Confidence interval: 95% and 99,9%

Degrees of Freedom: n-1 = 20-1 = 19

t value = 5.015

Checking the table (Two tailed test): 2.064 -> range goes from -2.064 to 2.064

TABLE A.2

t Distribution: Critical Values of t

Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	*Significance level* 10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883

Conclusion

As -2.093 < 5.015 > 2.093 it means that the value falls outside the range, so I have enough evidence at the 95% level of confidence to reject the hypothesis. That means that the average score is different to 76%. Then I can believe that the Average score from Metacritic is not totally reliable as other site with the same parameters shows a different average score.

As -3.883 < 5.015 > 3.883 it means that the value falls

outside the range, so I have enough evidence at the 99.9% level of confidence

to reject the hypothesis. That means that the average score is

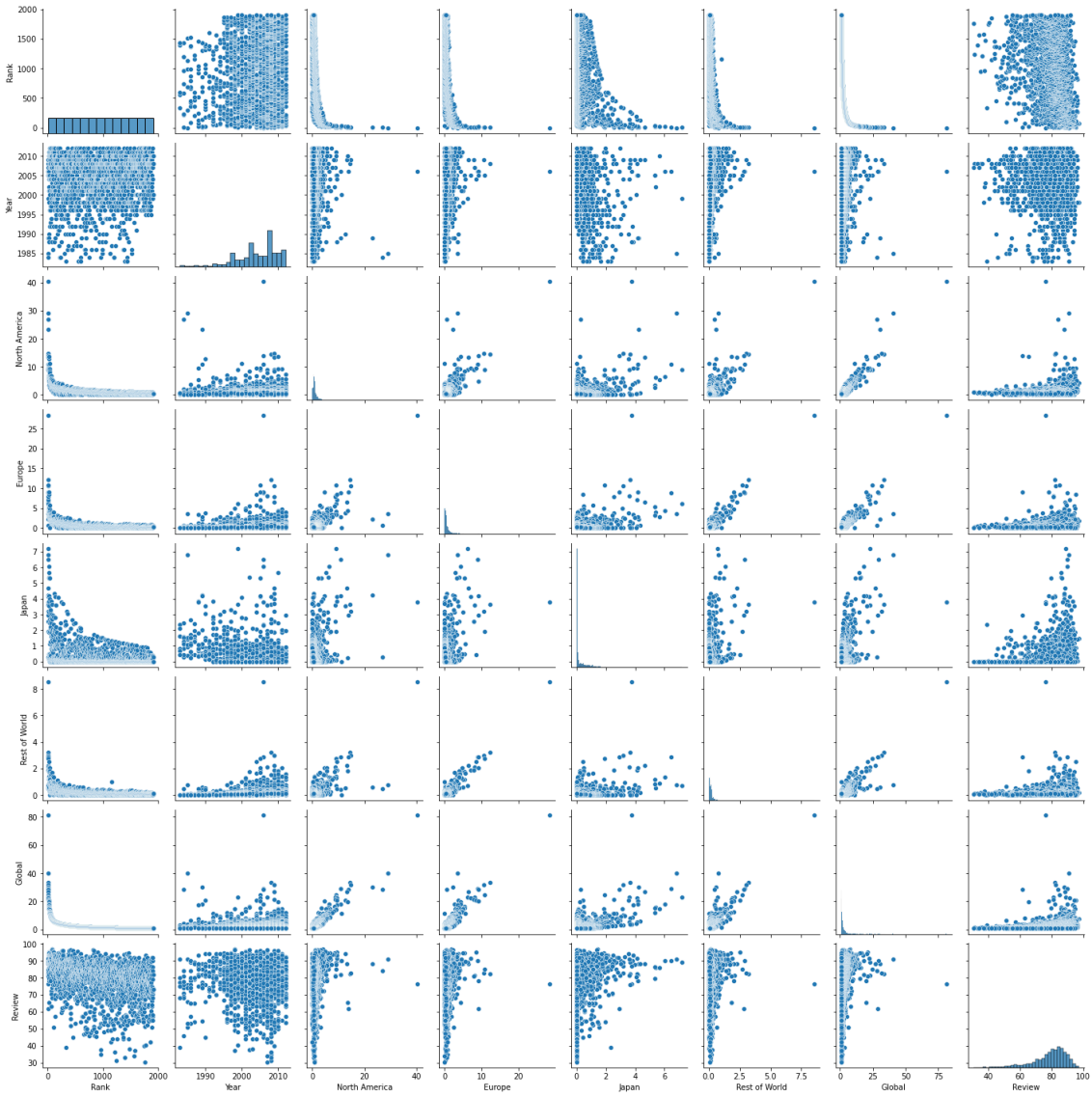
different to 76%. Then I can believe that the Average score from Metacritic is not totally reliable as other site with the same parameters shows a different average score.

Second section: Correlation analysis between 2 variables

Correlation

A correlation coefficient is an example of a descriptive statistic.

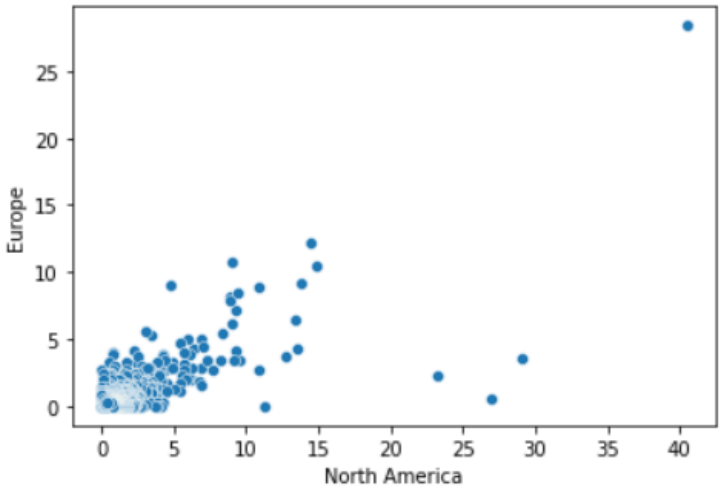
That is, it summarizes sample data without allowing you to draw conclusions about the population. [1][7]



I will use Correlation to analyse whether video game units sold in North America are related to game units sold in Europe.

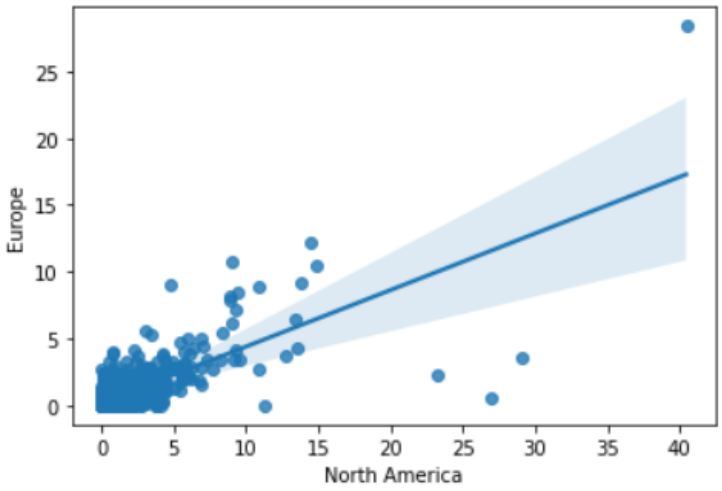
```
sns.scatterplot(data=games_df, x="North America", y="Europe")
```

```
<AxesSubplot:xlabel='North America', ylabel='Europe'>
```




```
sns.regplot(data = games_df, x="North America", y="Europe")
```

```
<AxesSubplot:xlabel='North America', ylabel='Europe'>
```



Through this plot, I can clearly see that the variables have a higher positive Linear correlation, but they do not have a perfect correlation, as the points are closer to the line than perfectly on the line.

Person Correlation

To find out how strong the correlation of these two variables is I will apply the Pearson method.

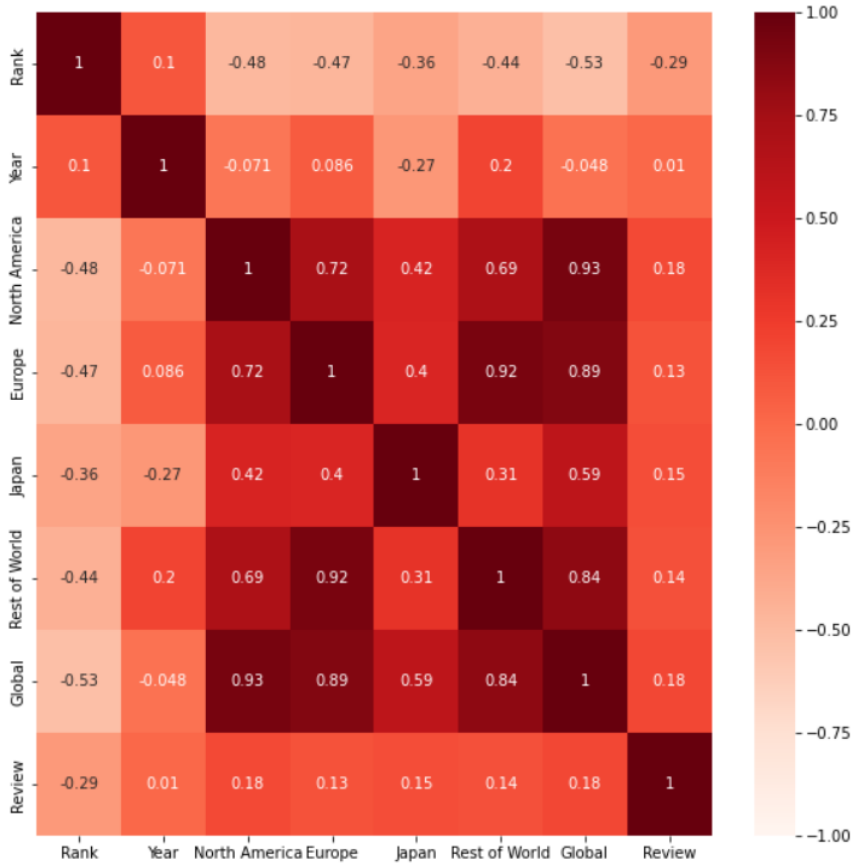
The Pearson correlation coefficient, also known as Pearson's r or the bivariate correlation in statistics, is a statistic that measures the linear correlation between two variables X and Y. Its value ranges from +1 to 1. A value of +1 indicates a total positive linear correlation, a value of 0 indicates no linear correlation, and a value of 1 indicates a total negative linear correlation.

```
games_df.corr(method='pearson')
```

	Rank	Year	North America	Europe	Japan	Rest of World	Global	Review
Rank	1.000000	0.101943	-0.480582	-0.466451	-0.358849	-0.436750	-0.529373	-0.292892
Year	0.101943	1.000000	-0.071347	0.085549	-0.274221	0.201768	-0.047886	0.010387
North America	-0.480582	-0.071347	1.000000	0.720766	0.416743	0.693662	0.933073	0.175684
Europe	-0.466451	0.085549	0.720766	1.000000	0.402289	0.922623	0.888902	0.129741
Japan	-0.358849	-0.274221	0.416743	0.402289	1.000000	0.308785	0.591751	0.148584
Rest of World	-0.436750	0.201768	0.693662	0.922623	0.308785	1.000000	0.837469	0.138467
Global	-0.529373	-0.047886	0.933073	0.888902	0.591751	0.837469	1.000000	0.181881
Review	-0.292892	0.010387	0.175684	0.129741	0.148584	0.138467	0.181881	1.000000

```
correlation = games_df.corr("pearson")
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(correlation, cmap= 'Reds',vmin=-1, vmax=1, annot = True)
```

```
<AxesSubplot:>
```



Through this Heatmap Correlation with the pearson method, I was able to confirm that the relationship between the chosen variables has a strong positive correlation as represented in the heatmap as “0.72”.[1][7]

Conclusion

Even with those information I can not assume that the units of the same games sold in North America caused the amount of games sold for the same games in Europe, as the correlation does not imply the cause, since to determine the cause it is necessary to analyze three more factors.

Temporal Sequence where units sold in North America should have happened before units sold in Europe. But the vast majority of games are released simultaneously.

Non-spurious relationship where the relationship between North America and Europe cannot just happen by chance. However, through the Heatmap and the interpretation of the variables, it is clear that the correlation between them is by chance and both only have a strong correlation because both cause changes in another variable to "Global Sales", showing that they follow in the same direction, but not causing any change between they.

Elimination of alternative causes where there is no other intervening or unexplained variable that is responsible for the relationship between North America and Europe, but in this scenario, we have the "Global Sales" variable that explains the relationship between them.

Third section: Linear Regression Model

In this section I will create a Linear regression model using numpy, but in order to make sure my prediction is correct, I will build a model using SciKit Learn only to compare the predictions.

As seen in the previous section, the two variables have a strong correlation between them, so I will use the polynomial function (polyfit) to find the best line that fits my dataset.

```
import numpy as np
gamesSales_fit = np.polyfit(games_df.North_America, games_df.Europe, 1)

gamesSales_fit
```

```
array([0.4232382 , 0.17390796])
```

As my values is in quantiti in Milions I do not need all the decimals.
The output gives me a = 0.42 and b = 0.17. With this two values I can now present my Lineal Regression Model:

North_America: NAM

Europe: EU

The function for my Lineal Regression Model will be: EU= 0.42*NAM + 0.17

Interpretation: EU= 0.42*NAM + 0.17

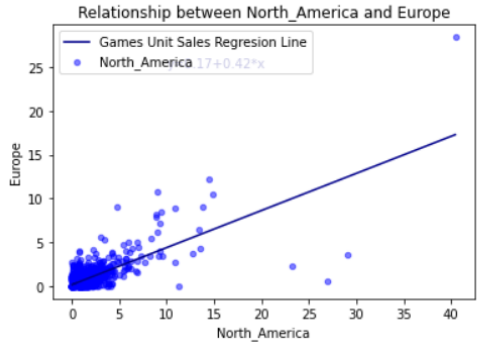
For every 1 milion unit game sold in North America, I can expect that Unit game sold in Europe increase by 0.42 (420 thousand).

```
ax1 = games_df.plot(kind='scatter', x='North_America', y='Europe', color='blue', alpha=0.5)

plt.plot(games_df.North_America, gamesSales_fit[0] * games_df.North_America + gamesSales_fit[1], color='darkblue')


plt.text(10, 25,'y={:.2f}+{:.2f}*x'.format( gamesSales_fit[1],  gamesSales_fit[0]), color='darkblue')

plt.legend(labels=['Games Unit Sales Regresion Line', 'North_America', 'Europe'])
plt.title('Relationship between North_America and Europe')
plt.xlabel('North_America')
plt.ylabel('Europe');
```



As I can see through this plot, this is the best fit line for my dataset.

In order to validate my model, I will compare the prediction between the model build by numpy with model build by the sciKit Learn.

```
from sklearn.linear_model import LinearRegression


# create linear regression object
lr_games = LinearRegression()

# fit linear regression
lr_games.fit(games_df[['North_America']], games_df['Europe'])

# get the slope and intercept of the line best fit
print(lr_games.intercept_)

print(lr_games.coef_)
```

```
0.17390795879585885
[0.4232382]
```

As I already can see, the model from sciKit Learn gave me the same output from the numpy.

Getting a predicted value with Numpy and Scikit Learn:

```
# predictions using numpy
print(np.polyval(gamesSales_fit, [23.20]))

# predictions using scikit learn
print(lr_games.predict([[23.20]]))
```

```
[9.99303411]
[9.99303411]
```

As shoewed the prediction are equals, that means both models are predicting in the same way.

Now to finish I want to calculate the sum of squares through the Coefficient of determination.

```
from sklearn.metrics import r2_score
actual_unitsoldEU = [28.39, 3.58, 12.22, 10.51,2.26, 2.26]
predicted_unitsoldEU = [17.29, 12.48, 6.31, 6.45,9.57, 9.99]
R_square = r2_score(actual_unitsoldEU,predicted_unitsoldEU)
print('Coefficient of Determination', R_square)
```

```
Coefficient of Determination 0.2722331149759236
```

I can conclude that 27.22% of the total sum of squares can be explained by using the estimated regression equation to predict the EU Unit games Sold. The remainder is error (72.78%).

Conclusion

Games	Total North America Units Sold	Total Europe Unit Sold	Y=0.42*x+0.17	y(predict games unit sold)
	x	y		
1	40.43	28.39	Y=0.42(40.43)x+0.17	17.29
2	29.08	3.58	Y=0.42(29.08)x+0.17	12.48
3	14.50	12.22	Y=0.42(14.50)x+0.17	6.31
4	14.82	10.51	Y=0.42(14.82)x+0.17	6.45
5	22.20	2.26	Y=0.42(22.20)x+0.17	9.57
6	23.20	2.26	Y=0.42(23.20)x+0.17	9.99
			Difference	0.42

I can conclude that the prediction works well as I can confirm that for every 1 milion unit game sold in North America, I can expect that Unit game sold in Europe increase by 0.42 (420 thousand) as showing in the table above, more clear between 5 and 6.

But in the other hand, by the Coefficient of determination that is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor, as the value get was 27.22%, indicate that the calculation did not fails to accurately model the data at all, but suggests that 27.22% of the dependent variable is predicted by the independent variable. But unfortunately with this result the model is a lower reliable model for future forecasts.[2]

Conclusion

As a result of this report, in the First session I were able to conclude that the average score from website Metacritic is not totally reliable as other site with the same parameters shows a different average score. After the Two tailed T-Test I had enough evidence at 95% of confidence to reject the hypothesis.

In the second session, after analysing the correlation between the two picked variable (North_America and Europe), I could conclude that, even if this two variable showed a strong correlation "0.72" there are not causation between them, where units sold in North America should have happened before units sold in Europe. But the vast majority of games are released simultaneously, and through the Heatmap and the interpretation of the variables, it is clear that the correlation between them is by chance and both only have a strong correlation because both cause changes in another variable to "Global Sales", and there is no

other intervening or unexplained variable that is responsible for the relationship between North America and Europe, but in this scenario, we have the "Global Sales" variable that explains the relationship between them.

In the third section I could conclude that the prediction is correct since I can confirm that for every 1 million unit game sold in North America, I can anticipate a 0.42 (420 thousand) rise in unit game sales in Europe, as shown in the table above, with a clearer increase between 5 and 6. However, the Coefficient of determination, which is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor, was 27.22 percent, indicating that the calculation did not fail to accurately model the data at all, but rather that the independent variable predicts 27.22 percent of the dependent variable. However, because of this discovery, the model is less dependable for future forecasts.

References

Works Cited

[1] Bhandari, Pritha. “A Guide to Correlation Coefficients.” *Scribbr*, 2 Aug. 2021, www.scribbr.com/statistics/correlation-coefficient/. Accessed 18 May 2022.

[2] Bloomenthal, Andrew. “How the Coefficient of Determination Works.” *Investopedia*, 10 Oct. 2021, www.investopedia.com/terms/c/coefficient-of-determination.asp. Accessed 19 May 2022.

[3] Fernandez, Javier. “The Statistical Analysis T-Test Explained for Beginners and Experts.” *Medium*, 8 June 2020, towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62. Accessed 17 May 2022.

[4] Kenton, Will. “T-Test Definition.” *Investopedia*, 22 Mar. 2020, www.investopedia.com/terms/t/t-test.asp#:~:text=Key%20Takeaways-. Accessed 17 May 2022.

[5] Metacritic. “About Us - Metacritic.” *Www.metacritic.com*, www.metacritic.com/about-metacritic. Accessed 15 May 2022.

[6] “Nintendo.” *Metacritic*, www.metacritic.com/company/nintendo?dist=positive. Accessed 14 May 2022.

[7] Nickolas, Steven. “What Does It Mean If the Correlation Coefficient Is Positive, Negative, or Zero?” *Investopedia,* 31 May 2021, www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp. Accessed 18 May 2022.