Correlation and Regression

Minerva Schools at KGI

CS51

Prof. Terrana

January 31, 2020

Correlation and Regression

Introduction

Video gaming is one of the fastest-growing industries in the world. In 2018, a new record of videogame sales was recorded, exceeding the revenue of $43.4 billion worldwide. In the US alone, over 164 million adults play videogames, and three-quarters of the population have at least one gamer in their household (ESA, 2019). Thus, this report is motivated to explore the increasing popularity of videogames worldwide and evaluate the extent to which the sales in North America can predict global sales variance. Ultimately, using the Video Games Sales data, generated by a scrape of VGChartz, the paper aims to analyze the sales data from more than 16500 games and perform a correlation and regression analysis (Smith, 2017).

Dataset

The data includes videogame names, platform, release years, genre, publisher, and global and distributed sales in millions across four regions: North America, Japan, European Union, and other areas. The data was initially processed by Formal Analysis professors to make the data import easier. The validity of such manipulation is worth considering in the evaluation.

The report aims to answer the research question “Is the sales figure in North America a good predictor of global sales?” by building and evaluating a regression model. Thus, the report explores how helpful the North American videogame revenue is in predicting global sales. Given the true population regression line expressed as $y = \alpha + \beta \times x$, where α is an

intercept and β is a slope, the following hypothesises will be tested and critically examined based on statistical (confidence intervals) and practical significance (Person's r, R-squared).

Null hypothesis: The true linear model has slope zero. There is not enough evidence to conclude a linear relationship in the population between North American and global sales.

$$\beta = 0$$

Alternative hypothesis: The true linear model has a slope nonzero. There is evidence to conclude a linear relationship in the population between North American and global sales.

$$\beta \neq 0$$

1. *Videogame sales in North America, million dollars.* It is a quantitative discrete variable, as it is countable in a finite amount of time. It is an independent variable or predictor since its variations do not depend on another variable within this system. In the regression model, it is expected to predict the variance of another variable.
2. *Global Sales, million dollars.* It is quantitative discrete: it has a countable number of values between any two values. It is a dependent variable, or response, as it is the effect of the change. The paper analyzes its response to a change in the predictor.

The Pearson's correlation and t-test were used to assess the potential regression model's practical significance since they are used to test quantitative data. Study results can be affected by confounding variables that influence both the sales in North America and the globe, including the game's publisher, genre, and media promotion. The game might not be

popularized in one region (e.g., Assassin's Creed in Japan), or games with specific genres and publishers might not enter the global market (e.g., games by Namco Bandai are mainly popular in Japan). Since the variables were not selected within a specific publisher or genre, such confounding variables might decrease the analysis's strength and validity, introducing biases and affecting the regression model statistics[1].

Methods

To answer the research question, several Python packages were used, including pandas, stats, matplotlib.pyplot and numpy. Firstly, to summarize the data, the bar charts were drawn (the code is available in Appendix A).

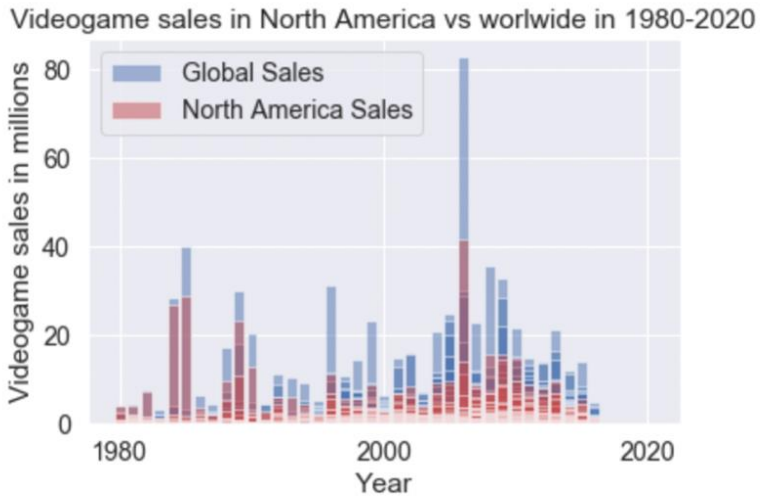


Figure 1. The bar chart showing a comparison of global and North American sales

[1] **#variables:** Before evaluating the data, I identified and classified the variables' types and examined the relationship between them. To take the analysis further, I described dependent and independent variables, classified them either as quantitative or qualitative, and defined the units. Describing variables was a necessary test because it influences the choice of an appropriate type of visualization and statistical analysis. I also commented on the consideration of confounding variables and how they might affect the study results.

Figure 1 helps describe an approximate share of North American sales in the global market as time progressed. They are accounted for roughly half of the global values. The chart gives some hints on potential associations: during high global sales, North American sales tend to be high too and vice versa; nevertheless, it does not provide any quantitative strength of the relationship. Since 2004, more video games were published per year that made less than $5 million in revenue, which created the overlap of values and colors in the graph. The graph is unhelpful in quantifying and understanding the associations when the global gain is small. Therefore, the scatter plots between multiple variables were drawn[2].

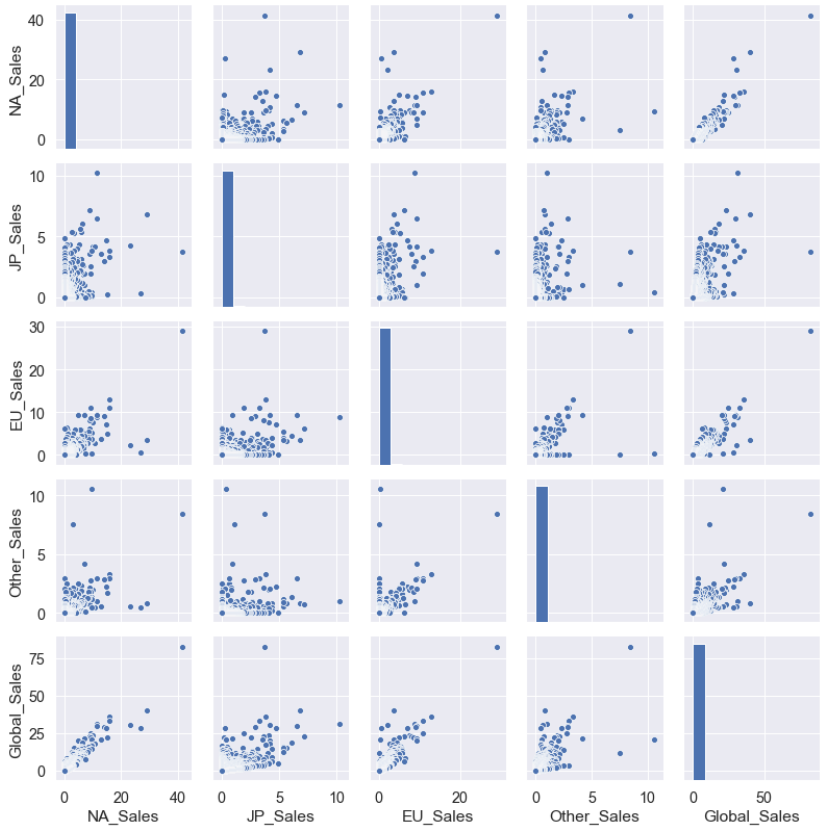


Figure 2. Scatter plots between all variables within the dataset

[2] **#dataviz:** The bar chart is necessary for future inferences of the data. It helps understand possible linear or nonlinear associations between the variables and check necessary conditions later in practical significance (Figure 2). It shows a distribution of data points and performs a comparison of values across two sales in 40 years. Here, I generated data visualization for the summary purpose, analyzed and interpreted the chart in its caption, and provided justification. Finally, I critiqued the graph to explain why scatter plots should be drawn to interpret this dataset better.

The Python sns.pairplot function was used (see the code in Appendix B and C) to draw and evaluate associations between all possible variables. It is observed that the relationship between North American and global sales is a strong linear. Also, relatively strong linear associations are found between international and EU sales, global and Japan sales. Both combinations can be used in the regression model as the plots are not curved and have no significant outliers. Nevertheless, if all predictors are plot, it goes against multicollinearity, as in Figure 3, sales in the EU and Japan show a moderate correlation with North American sales. It worth avoiding multiple regression models as the results are likely to be influenced by these confounding variables. Ultimately, a single variable regression is expected to better predict global sales compared to the multiple regression model.

Figure 3. The matrix of Pearson's Bivariate Correlation among all independent and the dependent variables

Practical significance

Pearson's r was calculated using the corr function in Python (code in Appendix D) to describe the strength of the linear relationship between variables. The formula for the calculation is given below, where n is the number of samples; $(x_i;\ y_i)$ is individual data points; $\bar{x}$ and $\bar{y}$ are sample means; s_x and s_y are standard deviations.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

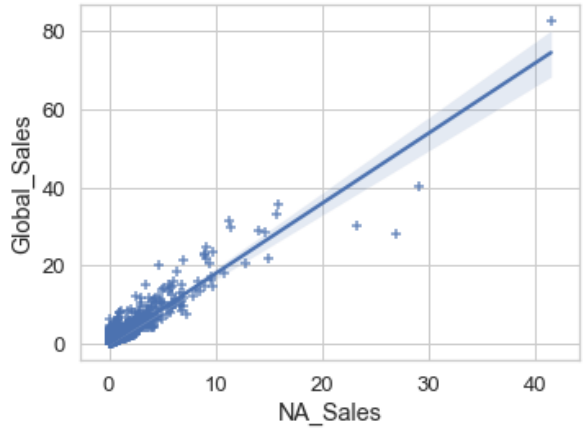


Figure 4. The regression line

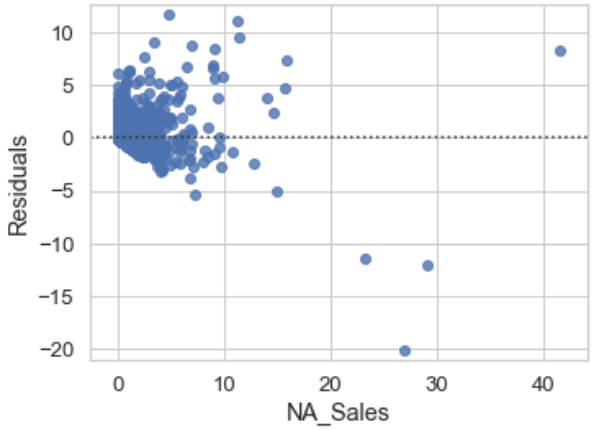


Figure 5. The scatter plot of residuals relative to the regression line

In Figure 4, the points mainly do not lie exactly on a line but are scattered more-or-less evenly around the regression line, creating a football-shaped graph with the Pearson's r value of 0.941. Given a large sample size, it proves a strong positive correlation. High figures of worldwide sales strongly correspond to high values of sales in North America and vice versa.

However, before interpretating and assessing the plausibility, the data needs to satisfy the assumption of linear regression models:

1. *Linearity*. This assumption is met since Figure 4 depicts a linear trend between the predictor and response, not curvilinear.
2. *Independence*. We collect observations from the same publisher over time (longitudinal dataset). Thus, the first game's success in a sequel might affect subsequent games' sales in a series. Since the dataset is not cross-sectional, where the data collected on entities only once and thus, assumed to be independent of each other, this assumption is not met.

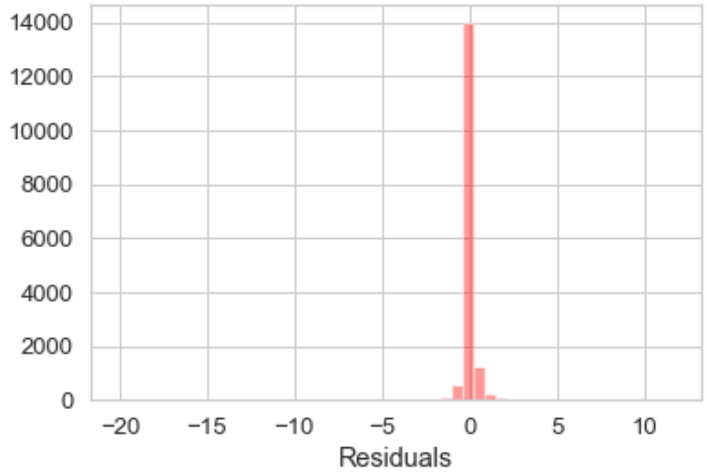


Figure 6. The distribution of residuals

3. *Normality*. Figure 6 illustrates a positively skewed histogram of residuals (the difference between observed and expected. In Figure 8, the skewness is equal to 1.093; it has to be within [-0.5;0.5] to be symmetric. It is skewed due to a few large outliers that exert a disproportionate influence on parameter estimates. Another violation is observed in the QQ Plot in Figure 7, where the plot is s-shaped. If it is normal, the plot should fall close to the diagonal reference line. An s-shaped pattern

of deviations proves that the residuals have excessive kurtosis of 241.131 (Figure 8). Thus, this condition is not met[3].

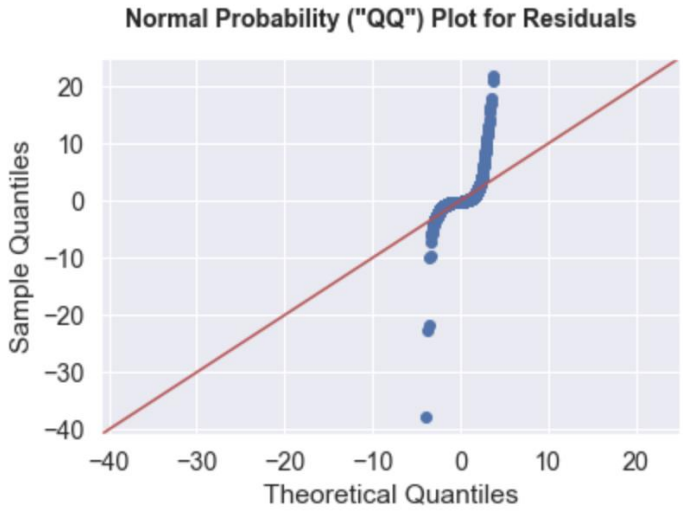


Figure 7. Normal distribution check with the QQ Plot

4. *Equal variance.* This assumption is met since the data is equally spread out around the regression model and has a relatively small standard error of 0.005 (Figure 8). The dataset is homoscedastic, and the residuals are mostly within $8 million distance from the predicted values with less than ten outliers for thousands of data entities[4].

[3] **#distribution:** One of the four assumptions (LINE – linearity, independence, normality, and equal variance) in building and interpreting a regression model is imposed on normality: the errors should all have a normal distribution with a mean zero. In linear regression, normality is not required from the predictor and response but only from residuals. All residual error of regression has to have the same variance and are identically distributed. However, this condition is not met. Ultimately, the violation of the assumption poses some challenges in calculating confidence intervals and whether the correlation coefficient is significantly different from zero. Since the calculation of confidence intervals is based on the assumptions of normality, if the error distribution is significantly skewed, confidence intervals may be too wide or too narrow.

[4] **#correlation:** The scatter plots in Figure 2 represent bivariate data where both measurements are taken per one variable (videogame). Here, the scatterplots are a way to visualize multivariate data to help classify and understand the relationships among the variables. The scatter plots are homoscedastic since the scatter is about the same in different vertical slices through the plots. Before interpreting regression (R-squared), I had to compute and interpret the correlation coefficient. The variables are strongly positively correlated, but one does not cause the other. There might be extraneous variables causing this correlation, including the money spend for marketing and social media promotion. Figures 2 and 3 give initial evidence for a correlation, but before concluding end results Figure 4,5 and 6 were drawn to check the scatterplot and residuals. Correlation is about the linear association. If the variables are associated in a nonlinear curve, have significant outliers, and are not quantitative, the Pearson's r value is invalid and should be reexamined.

Squaring the correlation coefficient gives R-squared, the coefficient of determination for a single variable regression model. In the formula below, SSE stands for error sum of squares, SSTO is the total sum of squares and $\hat{y}_i$ is the predictor value of the dependent variable.

$$R = r^2 = 0.941^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

It is equal to 0.886. Thus, 88.6% of variance in Global Sales can be explained by the regression model with sales in North American as a predictor.

There are not outstanding outliers, and the residuals are relatively small (Figure 5). The equation of regression line is $y = 1.794 \times x + 0.064$. The slope is positive (1.794), which means that for every unit (million) increase of videogame sales in North America, there is a $1.794 million increase in world revenue. The intercept (0.064) shows that it is expected to have $64000 world revenue when there is no revenue in North America. Nevertheless, we cannot say that sales in North America are solely causing the change in the world revenue. According to Figure 4, it is a possible outcome given that there are data points around the origin. The sales might come from the EU, Japan or other regions. For example, in 2009, Friend Collection by Nintendo had zero revenue in North America but had $3 million revenue solely from Japan (Smith, 2017)[5].

[5] **#regression:** I computed the regression line, interpreting the slope and intercept within the system along with identification of units and graphical representation. I explained the relation between dependent and independent variables. The calculated the coefficient of determination say a lot about of the model. In the given context of over 16000 data points, it allows to predict that in future North American market will be influential in global revenue of videogames. Nevertheless, in spite of globalization and interconnected supply and demand chain, some games might not enter a wider market and get popularity within smaller communities in Japan, EU and the rest of the world. Ultimately, 88.6% of the variation in global sales is reduced by taking into account predictor of sales in North America. The slope of the line is not equal to zero, which challenges the null hypothesis. However, it is too early to conclude the linear relationship unless the statistical significance is assessed.

The Pearson's r comparing North American and Global sales	$r = 0.941$
R-squared	$R^2 = 0.886$
The equation of the regression line	$y = 1.794 \times x + 0.064$

Table 1. The summary for the Linear Regression Model from the code in Appendix D

Statistical significance

A confidence interval was computed and interpreted to assess the statistical significance of the slope of the regression line. Using a built-in Python library function for Ordinary Least Squares, the linear regression model's summary statistics were calculated (see Appendix E). According to Figure 8, the calculated 95% confidence interval is [1.784; 1.804]. Thus, we are 95% confident that the population slope is within this range.

OLS Regression Results

Dep. Variable:	Global_Sales	**R-squared:**	0.886
Model:	OLS	**Adj. R-squared:**	0.886
Method:	Least Squares	**F-statistic:**	1.266e+05
Date:	Sat, 30 Jan 2021	**Prob (F-statistic):**	0.00
Time:	23:47:20	**Log-Likelihood:**	-12749.
No. Observations:	16291	**AIC:**	2.550e+04
Df Residuals:	16289	**BIC:**	2.552e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0644	0.004	14.777	0.000	0.056	0.073
NA_Sales	1.7938	0.005	355.783	0.000	1.784	1.804

Omnibus:	9587.887	**Durbin-Watson:**	1.928
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	38495037.040
Skew:	1.093	**Prob(JB):**	0.00
Kurtosis:	241.131	**Cond. No.**	1.42

Figure 8. The summary statistics

To estimate a confidence interval manually, the formula below is used (refer to Appendix F for the code):

$$point\ estimate\ \pm t^* \times SE = 1.794 \pm 1.960 \times 0.005 = 1.794 \pm 0.0098$$

It gives the same results as Python. Thus, we are 95% confident that for every additional one-degree increase in North American videogame sales, the global revenue increases between \$1.784 and \$1.804 million. The interval is narrow and does not contain zero. Nevertheless, due to the failure of satisfying assumptions earlier, we fail to reject the null hypothesis and conclude that there is no evidence of a linear relationship between North America sales and global sales in the population[6].

Results and Conclusion

The paper explored and answered the research question "Is the sales figure in North America a good predictor of global sales numbers?". From the practical significance, it is concluded that there is a strong correlation between independent and dependent variables ($r = 0.941$). The coefficient of determination is $R^2 = 0.886$: 88.6% of variance in global sales can be

[6] **#confidenceinterval:** We are interested in drawing conclusions about the population, not particular sampling. Python calculated the confidence interval for the true population slope. The formula for manual calculation was also used to explain the derivation and the variables that contributed to the final results. The point estimate is derived from the regression model. The t score was calculated by using stats.t.ppf function for the t-distribution with 95% confidence (adding 2.5% on tails). The degree of freedom is n-2 or 16289. The derived t score is 1.96, which is the same value we get for the z-score in a normal distribution. It is reasonable given the large sample size since for large sample t distributions are similar to normal ones. The standard error is derived from Figure 8. Thus, both calculations give the same confidence intervals. I accurately interpreted the meaning of the range and connected it back to the regression line and hypothesis test. The narrow width of the confidence intervals is achieved by having spread out values of the predictor and large sample size. Because of these factors, the t-multiplier is smaller (large degrees of freedom n-2, closer to a normal distribution), and standard error is larger (more terms to add up in the squares of the difference of the mean predictor and individual independent values), increasing the denominator and narrowing the interval. Nevertheless, since the residuals are not normally distributed, the confidence results are invalid. We cannot use a t distribution if we do not satisfy the independence and normality conditions.

explained by the model. The equation of the regression line is $y = 1.794 \times x + 0.064$. The statistical significance was assessed by confidence intervals. It was derived that the true population slope lies within the [1.784; 1.804] range. Nevertheless, there are limitations in the model as the regression failed to satisfy independence and normality conditions. It challenges underlying assumptions and the plausibility of results. Thus, the analysis fails to reject the null hypothesis and conclude that there is no linear relationship between videogame sales in North America and the globe.

I came to this conclusion using the eliminative induction by presenting a variety of different pieces of evidence to support the conclusion. The argument is built on generalization, where sample observations are used to draw a conclusion about a larger population and follows the Wilkin's first factor of the relationship of necessity. If we accept the premises of the inductive argument, the conclusion is likely to be true.

The argument is strong and reliable. Although practical and statistical significance provides enough evidence to reject the null hypothesis, one cannot trust the nonzero slope due to the violation of regression models' assumptions. The quantitative measures of skewness are true, and plots were rechecked several times. To make the argument stronger, one can increase the quantity of evidence supporting the conclusion or increase the variety of evidence. For example, as further research, I can process data beforehand to delete outliers and cleanse the data of repetitive videogame entries. Also, to satisfy conditions, one can interpret Durbin-Watson and Jarque-Bera tests that give more context for normality. Finally, I can reduce my own biases as I might have used the availability heuristic to correlate North American sales with global. Both the US and Canada have a strong foundation in technology that allows

video game businesses to flourish globally. To mitigate the bias, I should consider alternatives and extend my research[7].

Word count: 1384[8]

Reflection

Significance and confidence intervals help evaluate the statistical significance, which assesses the existence of the relationship. In contrast, correlation and regression measure the practical significance, which refers to the effect's magnitude. It tells whether the effect is large enough to make plausible conclusions consistent with the real world. Both significance measures complement each other to make more plausible and generalizable conclusions based on #induction and #plausability.

[7] **#induction:** I analyzed and applied inductive reasoning using evidence (statistical and practical significance) to support my conclusion and reasoning from a set of instances to a generalization about them. The conclusion of the analysis goes beyond the content of premises since I create a general claim about the whole population. Thus, the truth value of the argument cannot be guaranteed. I evaluated the results' strength and reliability, assessing how likely the conclusion is given that the premises are true. I also considered biases and the ways that might strengthen the argument further. I named the relevant Wilkins factors and explained how it specifically applies to this argument.

[8] **#professionalism:** I followed established guidelines for presenting my work professionally. To communicate effectively, I considered my tone and forms of address. I proofread the final written work for errors several times and adequately attributed quotations, ideas, and other sources. I used APA citations for citations and formatting of my bibliography. Overall, I ensured that my approach to communication meets the expectations relevant to the context.

Reference

ESA. (2019). *2019 Essential Facts.* Retrieved from https://www.theesa.com/wp-content/uploads/2019/05/ESA_Essential_facts_2019_final.pdf

Smith, G. (2017). *Video Games Sale*. Retrieved from kaggle: https://www.kaggle.com/gregorut/videogamesales

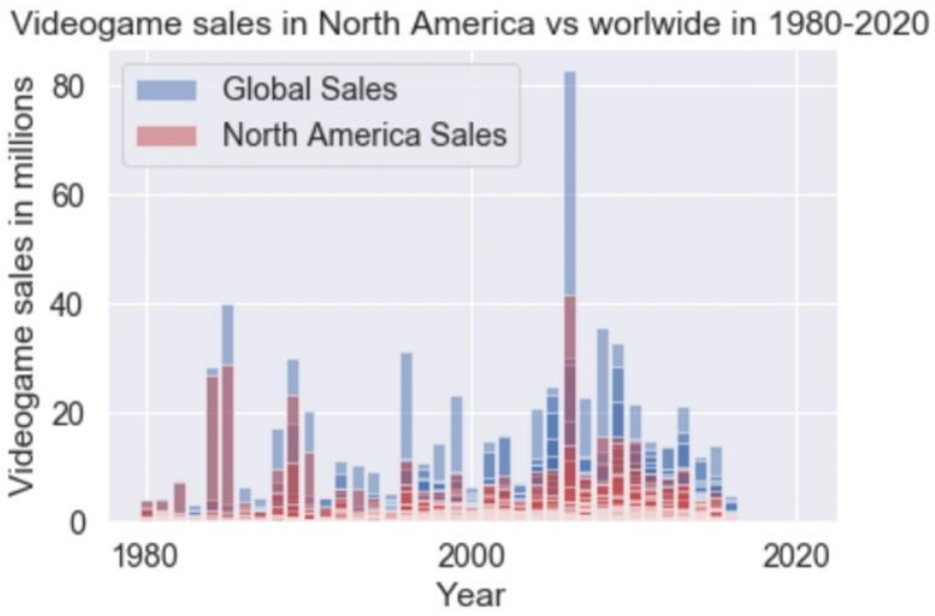
Appendix[9]

Appendix A: #dataviz. The bar chart showing the North American sales relative to the global figures.

Input:

```
plt.bar(data["Year"], data["Global_Sales"], color='b', alpha=0.5, label = "Global Sales") #the bar chart for Global Sal
plt.bar(data["Year"], data["NA_Sales"], color='r', alpha=0.5, label = "North America Sales") #the bar chart for North A
plt.title("Videogame sales in North America vs worlwide in 1980-2020") #creating the title
plt.xlabel("Year")
plt.ylabel("Videogame sales in millions")
txt="The bar chart shows how the share of North America videogame sales changed in the global market since 1980. Till a
plt.figtext(0.5, 0.01, txt, ha='center', wrap=True) #create the caption under the graph and wrap it
plt.xticks(np.arange(1980, 2025, step=20))
plt.legend()
plt.show()
```

Output:



The bar chart shows how the share of North America videogame sales changed in the global market since 1980. Till about 1990, North American sales took a significant ratio of the international sales. However, as time progressed, the videogame market expanded, gaining popularity in the rest of the world. As a result, the ratio of videogame revenue in North America relative to the global income decreased.

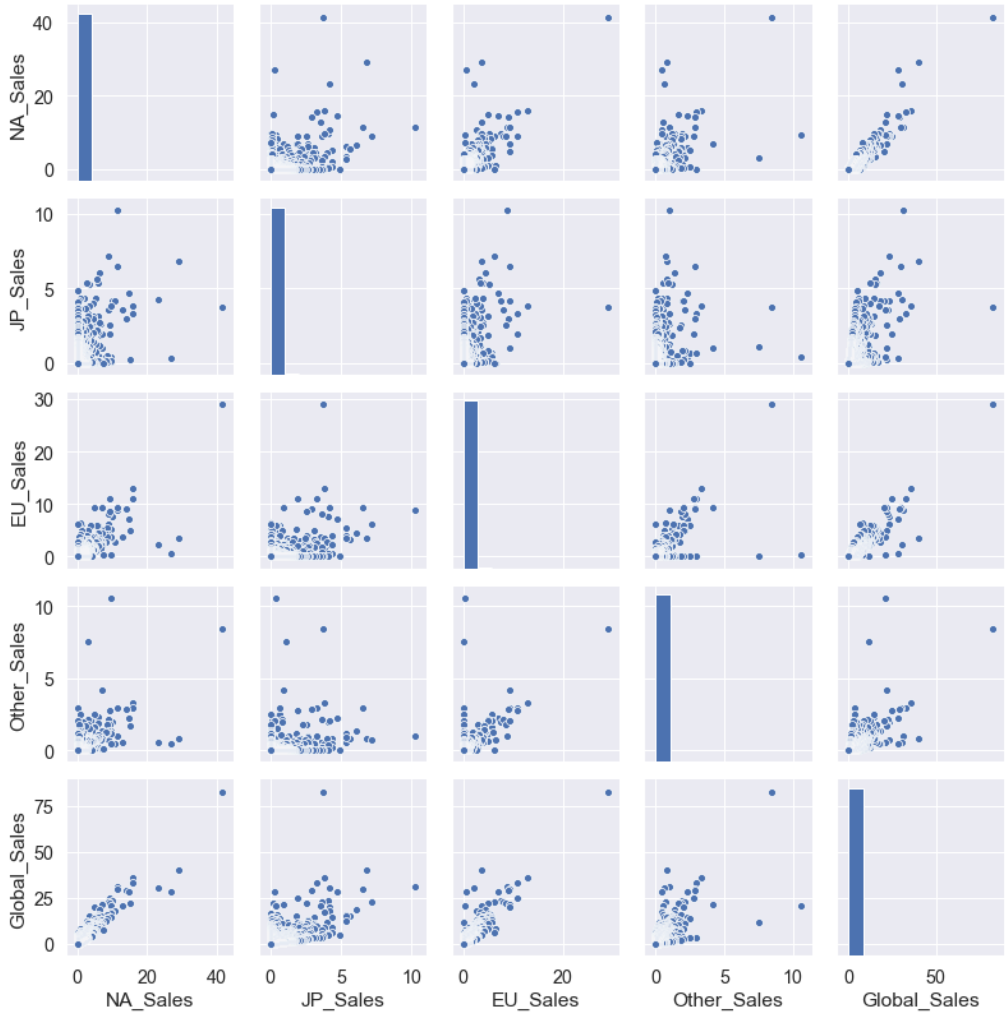
[9] The code was adapted from CS51 classes. The comments are self-created.

Appendix B: All possible scatter plots on bivariate data

Input:

```
column_x = ["NA_Sales", "JP_Sales", "EU_Sales", "Other_Sales"] #predictors
column_y = 'Global_Sales' #response
columnstoplot = column_x + [column_y] #every predictor is correlated with the response

sns.pairplot(data[columnstoplot], x_vars=columnstoplot, y_vars=columnstoplot, height=2.2) #creating 5*5 scatter plots
```

Output:

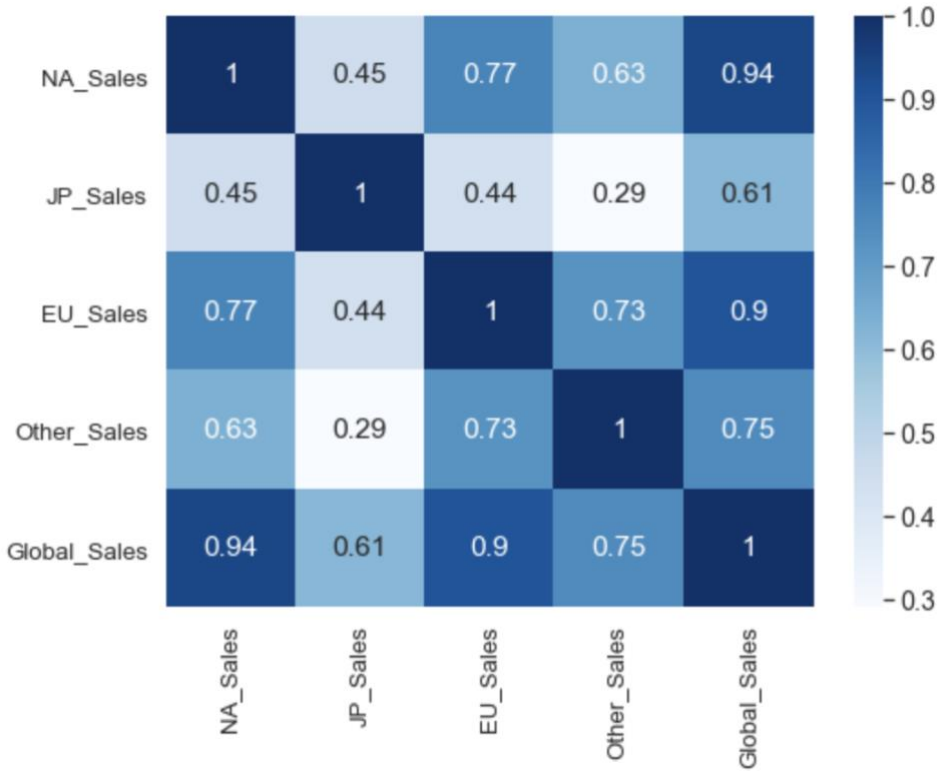


Appendix C: Pearson's correlation coefficient matrix for the plots in Appendix B

Input:

```
corrMatrix = data[columnstoplot].corr() #for eveyr grid in the 5*5 matrix the correlation coefficient was calculated
f, ax = plt.subplots(figsize=(8, 6))
sns.set(font_scale=1.3)
sns.heatmap(corrMatrix, annot=True, cmap='Blues') #the coefficient was written down in every square
plt.show()
```

Output:



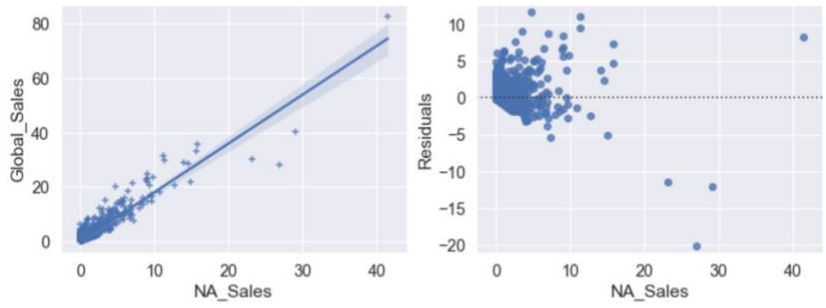
Appendix D: The calculation of the Pearson's r, R-squared and the equation of the regression line

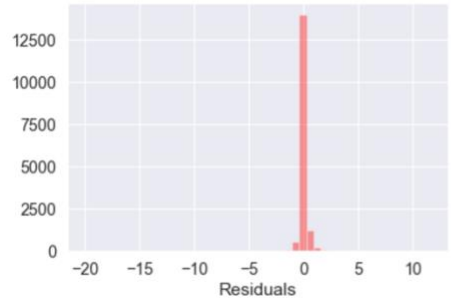
Input:

```
def pcorr(column_a, column_b): #the function accepts two variables and returns the Person's r value
    print("The pearson's r value comparing", column_a , "to", column_b , "is:", round(syndata[column_a].corr(syndata[cc
    print("")

def regression_model(column_x, column_y): #the function builds three plots and gives the equation of the regression lir
    # the function uses existing library functions to create a scatter plot with the predictor and response varibales,
    # plots of the residuals in a scatter plot and histogram, compute coefficient of determination,
    # and display the regression equation in the form y=bx+a

    # fit the regression line using "statsmodels" library
    X = statsmodels.add_constant(data[column_x])
    Y = data[column_y]
    regressionmodel = statsmodels.OLS(Y,X).fit() #here, OLS stands for "ordinary least squares"

    # extract regression parameters from OLS model above:
    Rsquared = round(regressionmodel.rsquared,3) #the values are rounded to the third decimal points
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    # make plots:
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=data, marker="+", ax=ax1) # a scatter plot
    sns.residplot(x=column_x, y=column_y, data=data, ax=ax2) # residual plot in a scatter plot
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.figure() # histogram with the distribution of residuals
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram

    pcorr(column_x, column_y)
    # print the computed results:
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)

regression_model('NA_Sales','Global_Sales')
```

Output:

```
The pearson's r value comparing NA_Sales to Global_Sales is: 0.941

R-squared =  0.886
Regression equation: Global_Sales =  1.794 * NA_Sales +  0.064
```





Appendix E: Confidence intervals and summary statistics for the regression model

Input:

```
def mult_regression(column_x, column_y):

    if len(column_x)==1: #since it is a single variable regression model, the function takes only one predictor
        plt.figure()
        sns.regplot(x=column_x[0], y=column_y, data=data, marker="+",fit_reg=True,color='orange') #drawing a regressior

    # assign predictor X and response Y:
    X = data[column_x]
    X = statsmodels.add_constant(X)
    Y = data[column_y]

    # construct a linear regression model:
    global regressionmodel
    regressionmodel = statsmodels.OLS(Y,X).fit() # here, OLS stands for ordinary least squares

    # creating a residual plot:
    plt.figure()
    residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='green')
    residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
    residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)

    # QQ plot to check normality of residuals:
    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45')
    qqplot.suptitle("Normal Probability (\"QQ\") Plot for Residuals",fontweight='bold',fontsize=14)

mult_regression(['NA_Sales'],'Global_Sales')
regressionmodel.summary()
```

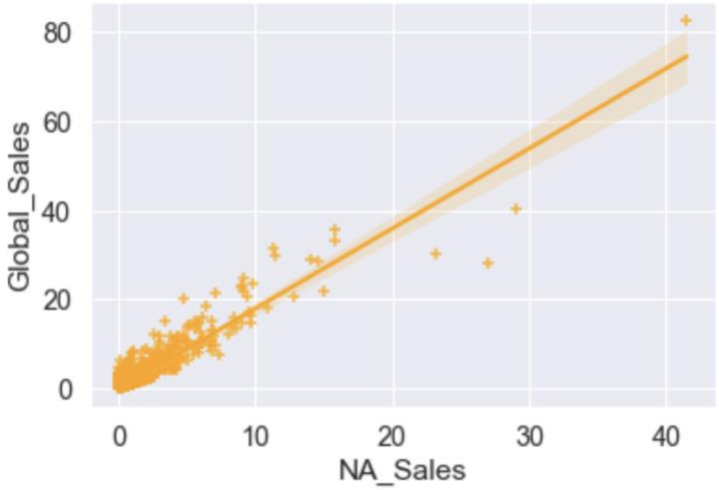
Output:

OLS Regression Results

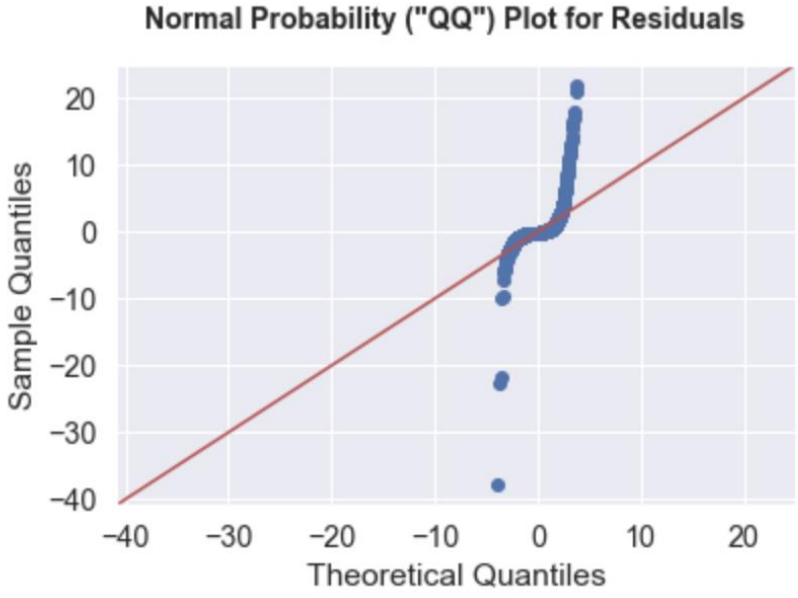
Dep. Variable:	Global_Sales	**R-squared:**	0.886
Model:	OLS	**Adj. R-squared:**	0.886
Method:	Least Squares	**F-statistic:**	1.266e+05
Date:	Sat, 30 Jan 2021	**Prob (F-statistic):**	0.00
Time:	23:47:20	**Log-Likelihood:**	-12749.
No. Observations:	16291	**AIC:**	2.550e+04
Df Residuals:	16289	**BIC:**	2.552e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0644	0.004	14.777	0.000	0.056	0.073
NA_Sales	1.7938	0.005	355.783	0.000	1.784	1.804

Omnibus:	9587.887	**Durbin-Watson:**	1.928
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	38495037.040
Skew:	1.093	**Prob(JB):**	0.00
Kurtosis:	241.131	**Cond. No.**	1.42


Global_Sales
NA_Sales
0
20
40
60
80
0
10
20
30
40


Residuals vs Fitted values
Residuals
Fitted values for Global_Sales
10
5
0
−5
−10
−15
−20
0
20
40
60


Normal Probability ("QQ") Plot for Residuals
Sample Quantiles
Theoretical Quantiles
20
10
0
−10
−20
−30
−40
−40
−30
−20
−10
0
10
20

Appendix F: Calculation of confidence intervals:

Input:

```
#confidence intervals

point_estimate = 1.794 #the slope of the regression line from a sample
t = stats.t.ppf(0.975, 16289) # n-2 for degress of freedom
SE = 0.005 #from summary statistics

print("Confidence intervals: [", round( mean - t * SE, 3), ",", round( mean + t * SE, 3), "]")
```

Output:

```
Confidence intervals: [ 1.784 , 1.804 ]
```