DATA MANGEMENT PROJECT REPORT

(Project Semester August-December 2018)

VIDEO GAME ANALYSIS

Submitted by

Amit Singh Sansoya

Registration No. 11615144

Programme and Section KEM45

Course Code INT-217

Under the Guidance of

Hargobind Singh

Discipline of CSE/IT

Lovely School of <u>Computer Science and Engineering</u>

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Amit Singh Sansoya bearing Registration no.11615144 has completed INT 217 project titled, **“Video Game Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Hargobind Singh
Assitant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date:

DECLARATION

I, Amit Singh Sansoya student of P132-B.Tech Computer Science and Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:

Signature

Registration No. 11615144

Amit Singh Sansoya
Name of the student

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of LFTS and our sir Hargobind Singh which helped me in successfully completing my project work. Also, I would like to extend our sincere esteems to all staff in laboratory for their timely support.

TABLE OF CONTENT

INTRODUCTION

This report contains about the various objectives and works which are performed on the basis of the project. The project has the topic of the Video Games Analysis and the analysis is done on Microsoft Excel and we would be looking at the various factors and other trends on which we would be performing the analysis. The dataset has the ranking of the games, their publisher, genre and the sales data available all over the globe. This dataset contains data of more than 12000 games published by the different companies on the various platforms. We would be trying to read the data and will try to checkout if there exists a trend which can be seen in the development of games and affecting the popularity of the games. We will be discussing our major objectives in this report further. The project will basically go from the various fundamental steps like Cleaning, Visualizing and Analysing. The data is large and has some of the anomalies which can affect the analysis and hence we would be trying to remove these anomalies. We would discuss about the data here

- Rank – Rank of the games
- Name – Name of the games
- Platform – Platform for which the games were developed like PS, PSP
- Year – The year on which the game was launched
- Genre – Main theme on which game are developed like Action, Sports
- Publisher – Company which developed the game
- NA_Sales – America Sales of Game
- EU_Sales – Europe Sales of Game
- JP_Salses – Japan Sales of Game
- Other_Sales – Sales of the game in other regions
- Global_Sales – Sales of the game globally

Above are the field included in the dataset and on the basis of these columns we would be analyzing the data and would be discovering the trend if there exists in the dataset for the sales and popularity of the game. We would be trying to draw out some of the facts if that could be obtained from the datasets. Games has been popular and it would be amazing to look after the popular genres of the games which are popular among the people and which platform for the period of the 1980-2016.

We will be trying to understand the dataset through various graphs and the trend line. We would also try to focus on the finding out the popularity of the genre or publisher which we would be justifying using the use of co-relation parameter. We will consider some of the assumptions like:

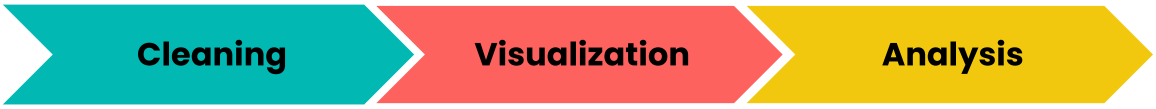
1. If the genre count and the global sales are both showing positive value, i.e they are showing positive correlation we can say that the genre is the popular one.
2. If the genre count and the global sales are both showing negative value i.e they are inversely proportional then we can say that the genre is not popular.
3. We would be studying the years and would be considering the popular rise or drop trend of the particular genre.

We would be dealing with the pivot tables and tables, charts we would be doing the conditional formatting and will be applying the filters. Noticing the change in the data over the time is the key observation to notice and mark the trend in the data.

We will be talking about the games and trends in this analysis

SCOPE OF THE ANALYSIS

After cleaning the dataset we can perform the analysis and can draw the visualizations to support the analysis. In the current chosen dataset we would look for the most popular platforms and genres for the games used by people varying over the time. We would try to find out if there is any key factor which actually affects the datasets and favors a particular genre of the game. We will be comparing sales and would be comparing and predicting the sales for the future and we would be trying to get the future scope of the popular genre.



Cleaning

Visualization

Analysis

In the analysis we would be trying to predict the future of the sales and popularity of genres by choosing appropriate model and factors.

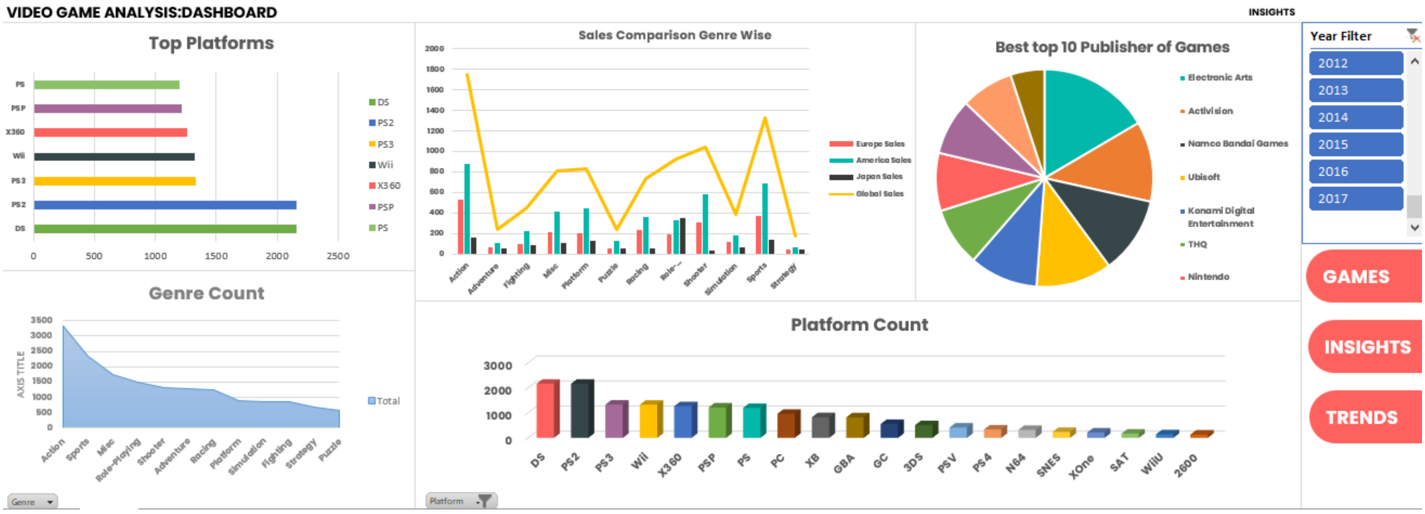
Correlation would be basic rule which one needs to understand as they would tell and prove if the hypothesis is correct and then only we can get the results. For that we have to make sure that the data is clean and we can perform visualization and then we would create model and then through that model we would do analysis and find out the trend and correlation and the other trends which are required. We would be using the visualization and correlation as our major rule to justify the objectives we would be establishing or explaining.

Let us talk about the analysis and we would discuss about the objectives on which my whole analysis is done.

- <u>To know and find out that which platform was most famous and popular among people</u>
    - Analysis on the platforms of the games. The platform which was most used to play the game by the people, we will be looking at that moreover we will be seeing the trend how the platform uses have changed as the time passed. We would look after the correlation of the platform and sales so that we can say that if there exists any such relation where we can say that if they are correlated.
- <u>Finding out which genre of game is most famous and finding out the relation among genre with sales</u>
    - Analysis on the genre column for knowing the most popular types of the games among the people. We would study how the time varied with the genre.

Moreover I have used it for the justification for the popularity of the types of game by plotting the Regression Graph for it.

- Comparing the sales of the games and discussing its trend if it exists and future Prediction
    - This field focuses on the classification of the games by the sales. We would be predicting the values and then we would see those data in the graph. We would be focusing on the various type of dependency factors which can be seen through sales like the genre, publications and etc.
- Analysis on the publication to find out the successful publisher of the game.
    - Providing the top publisher by the sales and providing the top publisher by the count of the published games. For this we would be keeping the correlation as one of the judging factor of the analysis.



Dashboard is one of the most important part when presenting the objectives

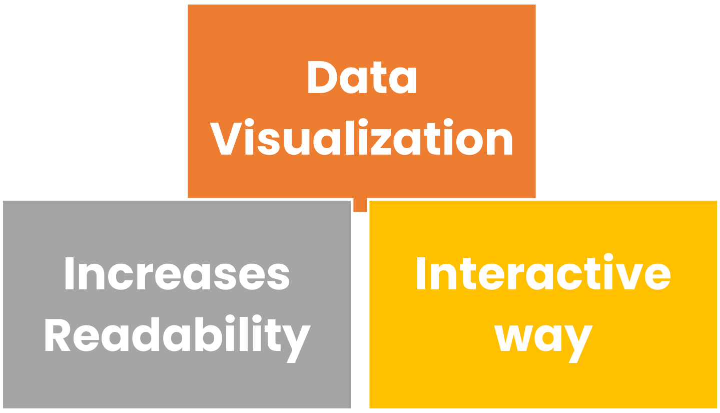
EXISTING SYSTEM

Existing system is the collection of the vast data without formatting or anything else. It's just raw data and there is not much information that can be extracted from that. Though that raw data is important for drawing out the major results and for finding out the major factor. The current existing system needs to be processed to form information. By applying proper formatting and cleaning we can represent the data in an impressive way which would actually help in analysis.

Existing system is the older way to represent the data with the new technology and the new tools available we can create the model and represent the data in new form which is more interactive and seems to be more informative then the existing one. The data needs to be processed fast and needs to provide the detailed view which has been made possible with the help of the new system available to us.

Disadvantage of Current Existing System

- Data is not in the form that can be represented
- Data is not clean and contains lot of anomalies and empty values
- Data needs to be processed further to make it more presentable
- The data lacks visualization and data is huge and tedious making it difficult to study.
- There can't be any results drawn by just looking at the data.



SOURCE OF DATASET

Kaggle is an online community of data scientists and machine learners, owned by Google, Inc. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

Source: https://www.kaggle.com/gregorut/videogamesales

This dataset contains a list of video games with sales greater than 100,000 copies.

Fields include

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC, PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

ETL PROCESS

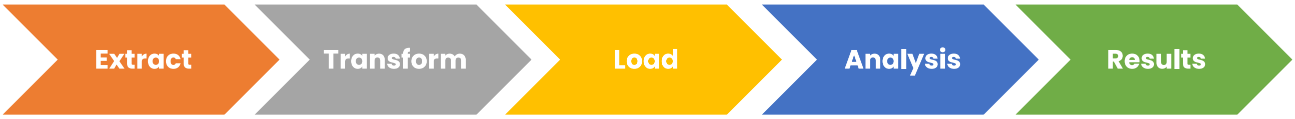
ETL process stands for the Extract, Transform and Load. ETL plays the major role in this analysis also because at first the data obtained from the excel sheet was taken cleaned that is what we call the extraction and transformation. The transformation contains the cleaning of data, formation of tables, formation of pivot tables accordingly and then loading it on to excel again. So the process goes from the excel to excel in this, However the main purpose of ETL in this was to make the data available for the analysis and making it clean.

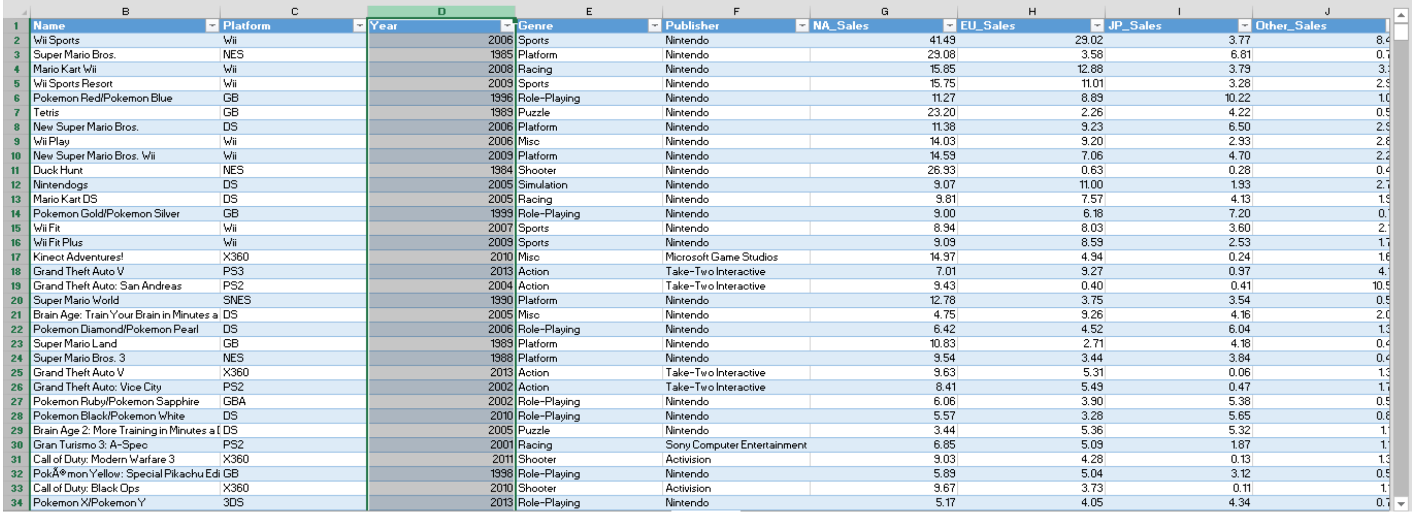
Extraction

Extraction is the process of extracting the data from somewhere but in my case the data was taken from the kaggle and was dumped on to excel. It was in the CSV (comma separated values) format which was later changed to xlsx which is standard excel format the data was then skimmed and then we changed it into the table. This whole process of extracting the data from the web to excel was the process of the extraction.

Transform

Transform allows the user to modify, clean and make the data appropriate so that it can be used for the purpose of the analysis and that is why this is considered as the most important part of the data analysis. In my case also the data was not that clean, I had to clean the data where I had to look for the rows which were missing values or were having N/A values. The values were required to be filled with the appropriate values to get the correct result and at the same there were rows which were required to drop to get results. The data cleaning is not just about cleaning the values and there is more to it in which we have to look after the format of the data. We have to classify rows into the strings, numbers, dates and currencies and so on. In order to maintain the proper data format so that the analysis would not be affected.



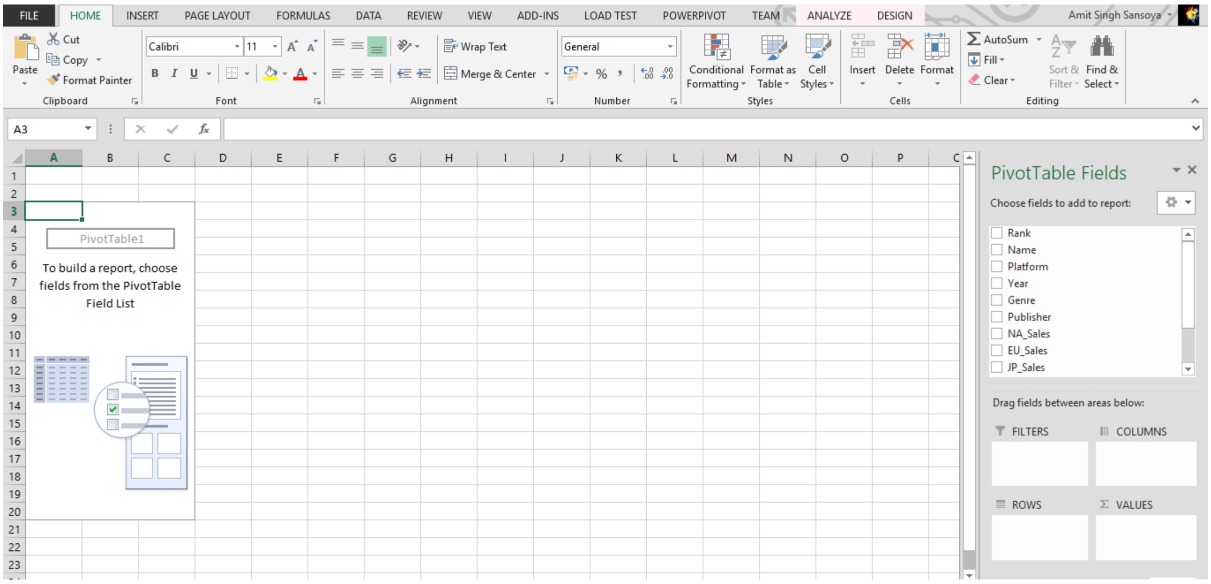


	B	C	D	E	F	G	H	I	J
1	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales
2	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.4
3	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.7
4	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.3
5	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.9
6	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.0
7	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.5
8	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.50	2.9
9	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.20	2.93	2.8
10	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.70	2.2
11	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.4
12	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11.00	1.93	2.7
13	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.9
14	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.7
15	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.60	2.1
16	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.7
17	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.6
18	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.1
19	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.40	0.41	10.5
20	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.5
21	Brain Age: Train Your Brain in Minutes a	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.0
22	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.42	4.52	6.04	1.3
23	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.4
24	Super Mario Bros. 3	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.4
25	Grand Theft Auto V	X360	2013	Action	Take-Two Interactive	9.63	5.31	0.06	1.3
26	Grand Theft Auto: Vice City	PS2	2002	Action	Take-Two Interactive	8.41	5.49	0.47	1.7
27	Pokemon Ruby/Pokemon Sapphire	GBA	2002	Role-Playing	Nintendo	6.06	3.90	5.38	0.5
28	Pokemon Black/Pokemon White	DS	2010	Role-Playing	Nintendo	5.57	3.28	5.65	0.8
29	Brain Age 2: More Training in Minutes a [	DS	2005	Puzzle	Nintendo	3.44	5.36	5.32	1.1
30	Gran Turismo 3: A-Spec	PS2	2001	Racing	Sony Computer Entertainment	6.85	5.09	1.87	1.1
31	Call of Duty: Modern Warfare 3	X360	2011	Shooter	Activision	9.03	4.28	0.13	1.3
32	PokÃ®mon Yellow: Special Pikachu Edi	GB	1998	Role-Playing	Nintendo	5.89	5.04	3.12	0.5
33	Call of Duty: Black Ops	X360	2010	Shooter	Activision	9.67	3.73	0.11	1.1
34	Pokemon X/Pokemon Y	3DS	2013	Role-Playing	Nintendo	5.17	4.05	4.34	0.7

Glimpse of the process while cleaning the data

Loading

Loading means once the data is transformed then it is ready to be processed for the task of analysis. In the excel once the data is cleaned then the data is loaded on to form the pivot table, tables, and power pivot in order to get the desired outputs. Once these results are done then the data is loaded and the next form which are visualization and the other things start taking place. Once the data is cleaned and loaded one can proceed on to with the process of creating visualizing and creating formulas to do analysis and justify it and this also contains the part of the correlation for my project.



PivotTable1

To build a report, choose fields from the PivotTable Field List

PivotTable Fields

Choose fields to add to report:

□ Rank
□ Name
□ Platform
□ Year
□ Genre
□ Publisher
□ NA_Sales
□ EU_Sales
□ JP_Sales

Drag fields between areas below:

FILTERS

COLUMNS

ROWS

VALUES

Cleaning the data and loading it to the pivot table.

Analysis

Let us look at the objectives which we will be doing in this analysis and then we would jump into the detailed view of it.

Objectives

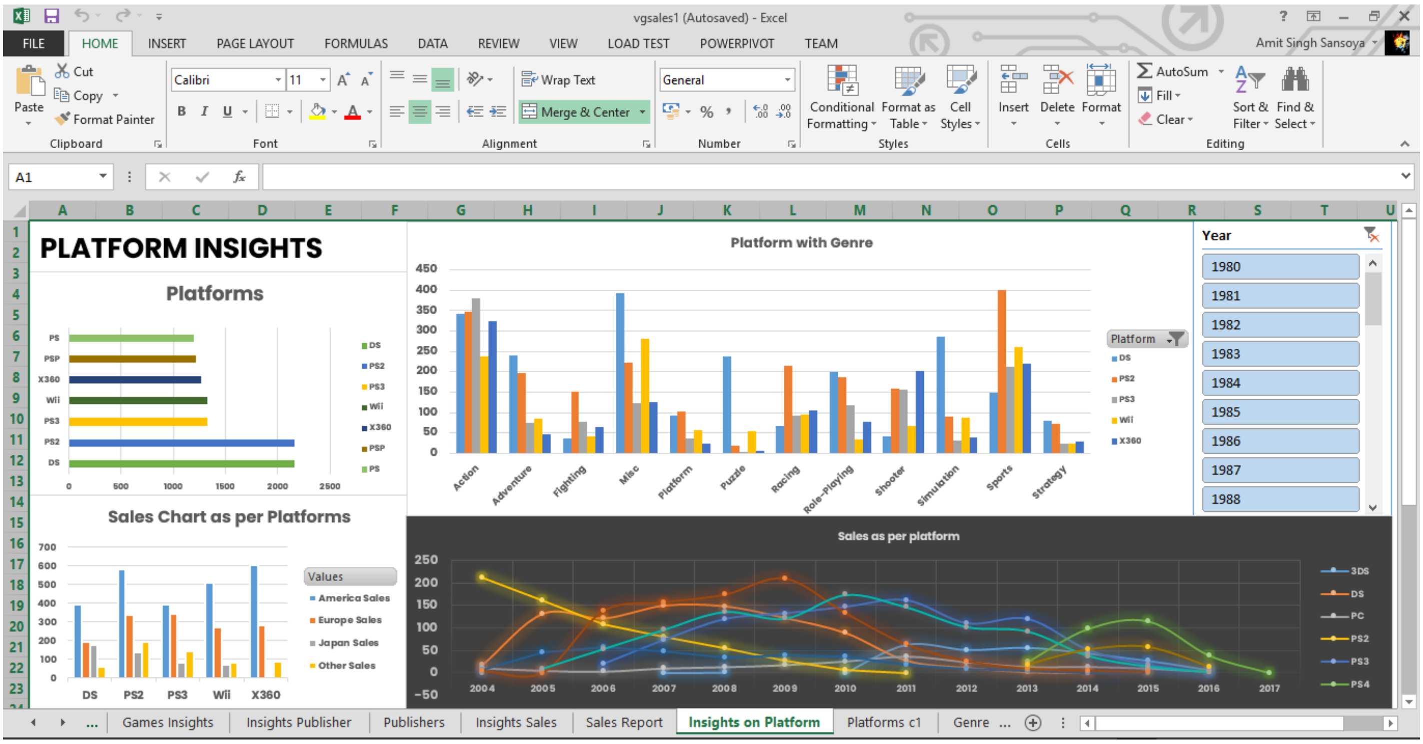
1. **Platform Analysis – Studying platform column of the dataset**
   a. To know and find out that which platform was most famous and popular among people. To provide the insights and other related factors with the platform publication.
2. **Genre Analysis –Studying genre column of the dataset**
   a. To study the genre and treating it as one of the major factor for finding out the popularity of the games. Treating it against time and studying the trend.
   b. Genre Regression – Justification factor for the popularity of games
   Comparing the genre and making it most important part for the analysis to support the popularity factor by plotting the Scatter plot and finding the correct graph line which would satisfy these trend
3. **Comparing Sales – Year wise sales record of the genre and publications**
   a. Showing top 10 genre and companies record with the respective years and studying about the various sales pattern.
4. **Sales Prediction – Prediction of sales**
   a. Using python and excel tool we will create the model the model will be developed in excel directly will be loaded on to python and by running required algorithm we would do the analysis.
5. **Publications – Listing out the most famous publications**
   a. Providing the detailed view on best games publications by comparing on the various different fields through graphs.

Objective 1: Studying various platforms where the game is played

Introduction

This field is about how the various platforms are been famous in the people for playing the games. The various gaming consoles like Nintendo DS, PS, PS-2, PSP and others. There are many other things like reading about the platforms through the sales and genres.

I have tried to analyze the data by keeping Platform as an independent factor and making all other factors as dependent arguments. By this I was trying to state that how platform can be taken as a parameter such that other platforms depend on the platform and how platform can be used to showcase the sales, popular genre and the sales per platform.



Platforms Insights and the other graphs

Description

Here we will be seeing the close look up at the platforms. We would be observing top 7 platforms by count as we can consider that these top 7 platform are the one which were actually sold most and have contributed most to the sales part of it. I have also analyzed the platform v/s sales chart in order to support the best platforms and hence in this way we can prove that they are correlated to large extend. Similarly to compare further there are other supports provided by the comparison of platform with genres and there year basis comparison. The result outcome should be positive if the hypothesis which I have thought is correct is considered as true. The hypothesis is that if the particular platform is popular by count then it must have the most sales contribution.

Special Requirements, Formula's

There is a requirement of multiple pivot tables between many things and the co-relation function as it would be also proving the fact of the hypothesis.

Required Stuff

1. Sum of all the count of the platforms categorized by the various platform of the game
2. Sum of all America's, Europe's, Japan and other sales by the sum of their individual sales categorized by the platform of games.
3. Sum of all the count of the platforms categorized by the genre.
4. Similarly sum of all the count of the platforms on the year basis categorized by the year.
5. Knowledge of Correlation

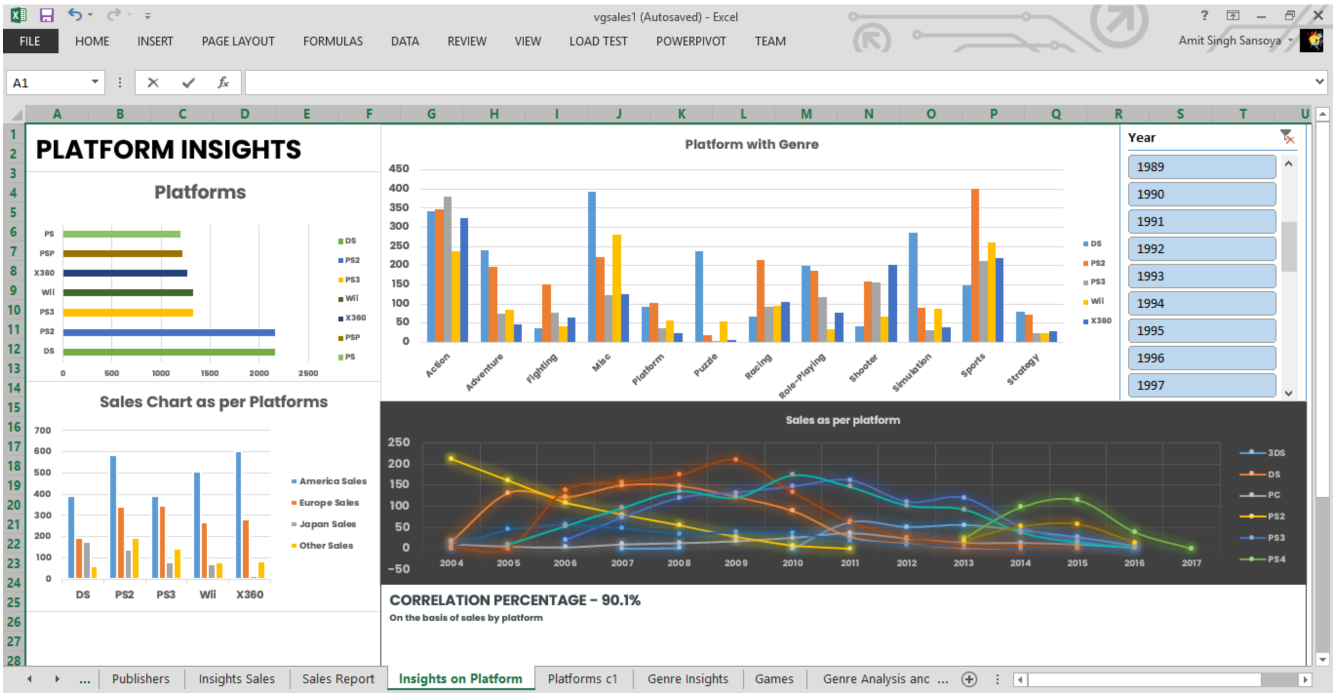
$$f(x) = \sum_{n=1}^{n} (col[n])$$

Correlation in Excel =Correl(array1,array2)

Analysis Results

After creating all the pivot tables and plotting graphs we can clearly see that the hypothesis we considered where correct and the correlation between the taken arguments where suitable to prove the fact. The correlation result between the sales by the platform is 90.81% so we can say that the hypothesis is correct. The sample of portion was also tested and when performed on top 10 the correlation was positive hence supporting the hypothesis.

Visualizations



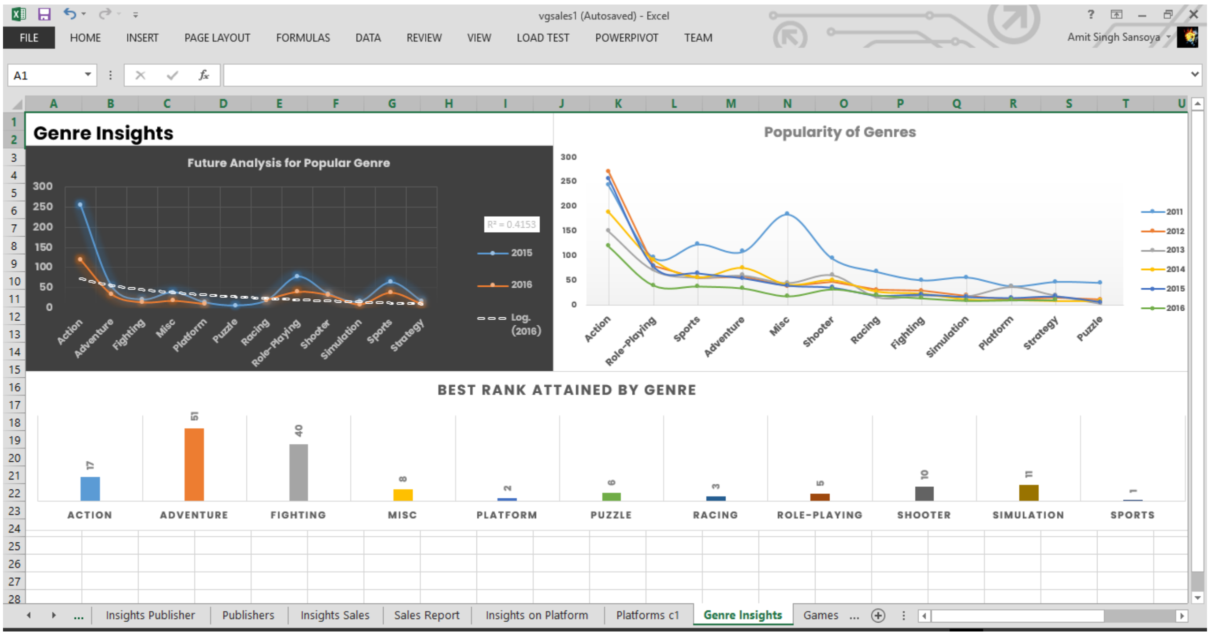
Visualization supports the hypothesis and we can directly see the trend and also can observe the best way which I can support the hypothesis.

Objective 2: Studying the Genre of the games

Introduction

This field is about how the various genres has been famous in the people for playing the games. The various genres are action, shooting, racing and others. There are many other things like reading about the genres through the sales and publications.

I have tried to analyze the data by keeping genre as an independent factor and making all other factors as dependent arguments. By this I was trying to state that how platform can be taken as a parameter such that other platforms depend on the platform and how platform can be used to showcase the sales, popular publication and the sales per platform.



Genre Insights and the other graphs

Description

Genre has been treated as one of the most important factor by me as it would tell how the population has liked the games and what type of games were liked by them the most and we will try to provide the correlation of the genre and sales. We will study genre by sales and will also study about the best rank and will try to do the future analysis using the trend line in excel. We will study the logarithmic trend line instead of the linear line which we all know. Focusing on the regression we will try to obtain the results.

Special Requirements, Formula’s

There is a requirement of multiple pivot tables between many things and the co-relation function as it would be also proving the fact of the hypothesis.

Required Stuff

1. Sum of all the count of the genres categorized by the various platform of the genres
2. Sum of all genres by the sales
3. Minimum Genre Rank obtained
4. Knowledge of Correlation

$$f(x) = \sum_{n=1}^{n} (col[n])$$

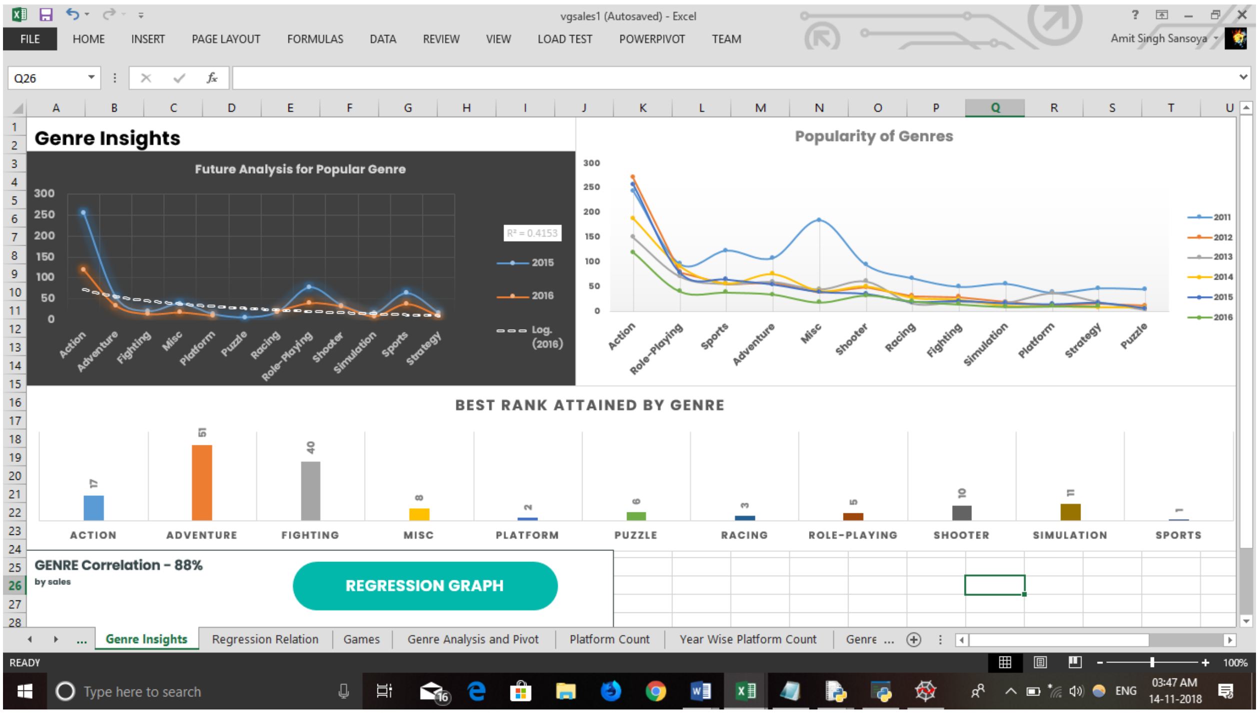
Correlation in Excel =Correl(array1,array2)

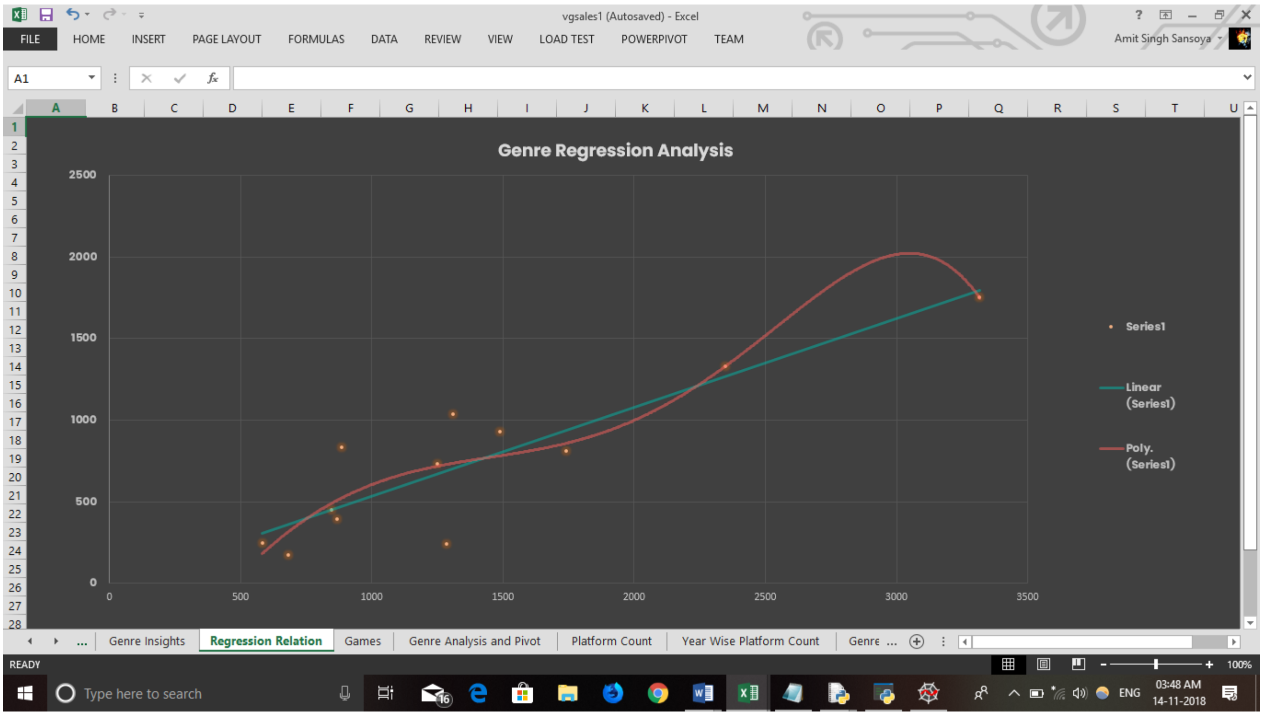
Analysis Results

After creating all the pivot tables and plotting graphs we can clearly see that the hypothesis we considered where correct and the correlation between the taken arguments where suitable to prove the fact. The correlation result between the sales by the platform is 88.1% so we can say that the hypothesis is correct. The sample of portion was also tested and when performed on top 10 the correlation was positive hence supporting the hypothesis.

Moreover this result is also obtained that the most liked genre by the people was the sports followed by platforms based games. Action games were famous at most of the early stages.

Visualizations


vgsales1 (Autosaved) - Excel
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW LOAD TEST POWERPIVOT TEAM
Amit Singh Sansoya
Q26
Genre Insights
Future Analysis for Popular Genre
300 250 200 150 100 50 0
Action Adventure Fighting Misc Platform Puzzle Racing Role-Playing Shooter Simulation Sports Strategy
R² = 0.4153
2015
2016
Log. (2016)
Popularity of Genres
300 250 200 150 100 50 0
Action Role-Playing Sports Adventure Misc Shooter Racing Fighting Simulation Platform Strategy Puzzle
2011 2012 2013 2014 2015 2016
BEST RANK ATTAINED BY GENRE
17 51 40 8 2 6 3 5 10 11 1
ACTION ADVENTURE FIGHTING MISC PLATFORM PUZZLE RACING ROLE-PLAYING SHOOTER SIMULATION SPORTS
GENRE Correlation – 88%
by sales
REGRESSION GRAPH
Genre Insights | Regression Relation | Games | Genre Analysis and Pivot | Platform Count | Year Wise Platform Count | Genre ...
READY
Type here to search
100%
ENG 03:47 AM 14-11-2018


vgsales1 (Autosaved) - Excel
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW LOAD TEST POWERPIVOT TEAM
Amit Singh Sansoya
A1
Genre Regression Analysis
2500 2000 1500 1000 500 0
0 500 1000 1500 2000 2500 3000 3500
Series1
Linear (Series1)
Poly. (Series1)
Genre Insights | Regression Relation | Games | Genre Analysis and Pivot | Platform Count | Year Wise Platform Count | Genre ...
READY
Type here to search
100%
ENG 03:48 AM 14-11-2018

Visualization supports the hypothesis and we can directly see the trend and also can observe best way which I can support the hypothesis.

Objective 3: <u>Comparing Sales</u>

Introduction

This field is about the sales of the games over the years we would be analyzing the sales with the genres, publications and time. The each sales comparison is done in their respective sheet and domain and can be explained while going to those domains for now while going inside the sales we can see the trend from the year from 2009 to 2016. The data was fully observed from the period of 1980 to 2016 and then we have drawn the results out of it such that we can see the trend. There we can see the sales record of America, Europe and Japan along with the global sales.

Description

Sales is the most important factor on which this whole data is existing because this can be seen as the mutual parameter which is available for every things analysis. The basis of the popularity of the games and the judgment of it depends on it by a large extend and this is how this plays the most important role. The correlation factor can be only tested and judged if the record of sales is available and that's it. The data is classified as the America's sale, Europe Sales, Japan's Sales and their sum is the global sales.

Special Requirements, Formula's

There is a requirement of multiple pivot tables between many things and the co-relation function as it would be also proving the fact of the hypothesis.

Required Stuff

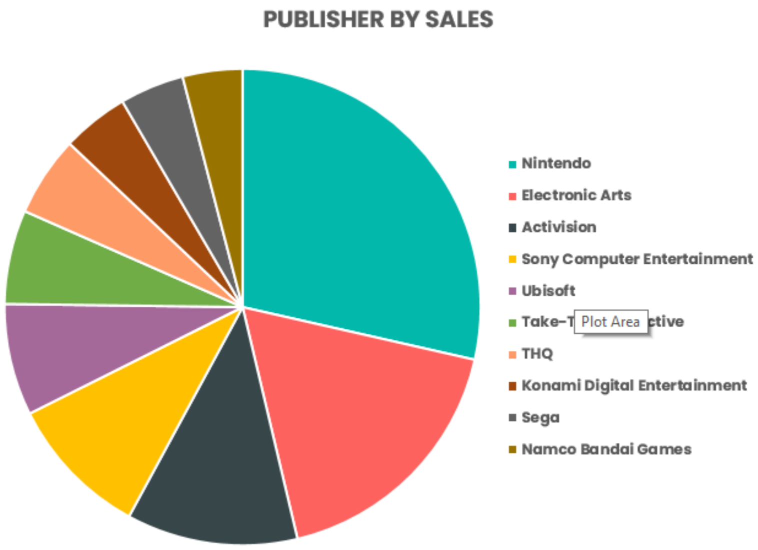
1. Sum of the NA sales
2. Sum of the EU sales
3. Sum of the JP sales
4. Sum of the Global sales

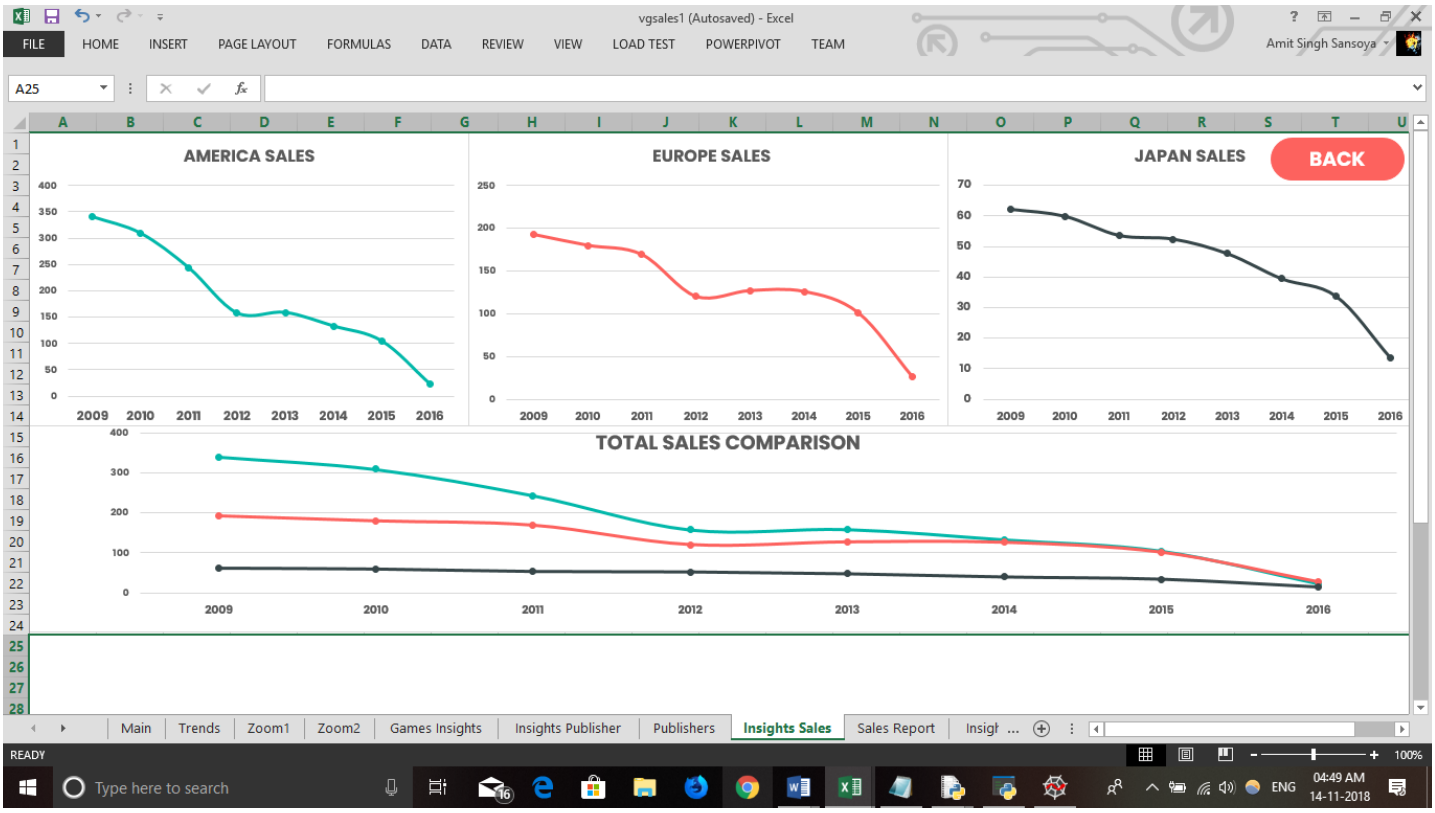
All categorized by the years.

$$f(x) = \sum_{n=1}^{n} (col[n])$$

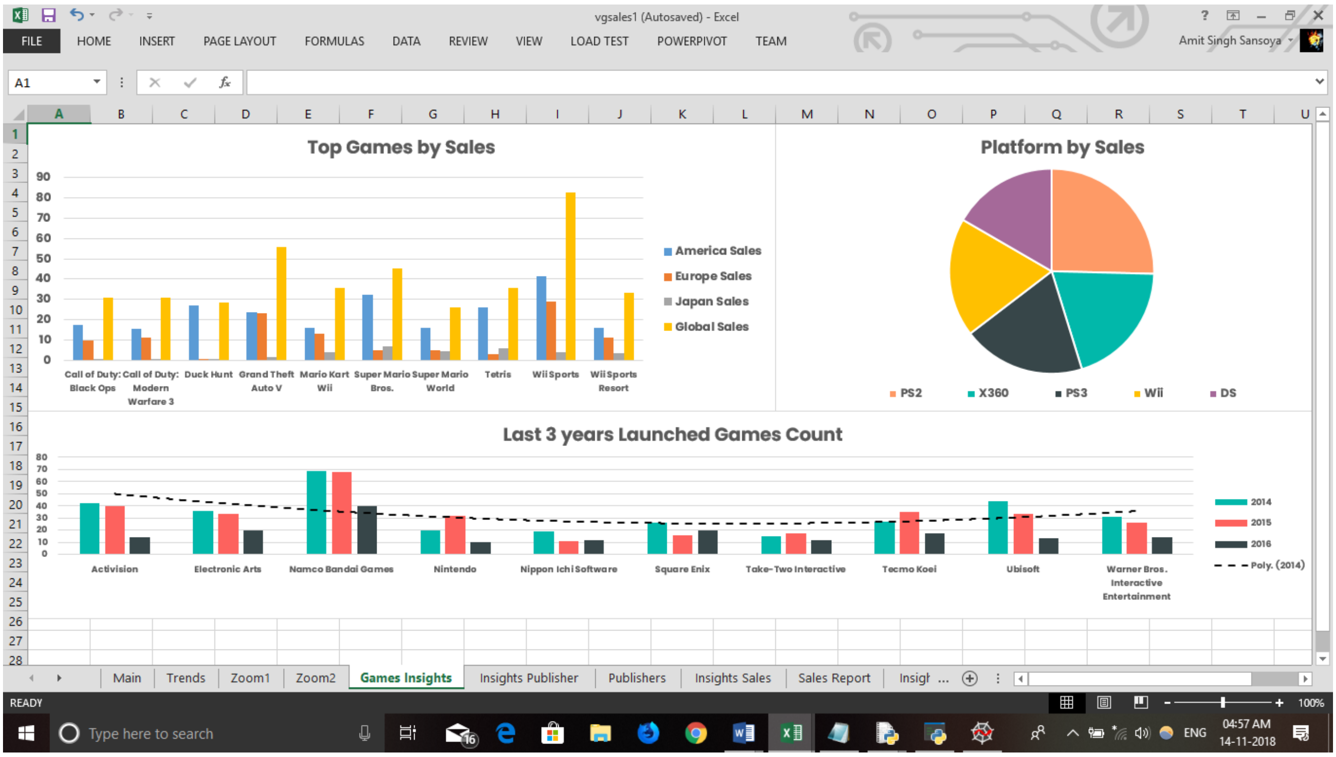
Analysis Results

After the plotting of the graph we can see the declining trend of the sales in all the places all over the world. We can see that the Nintendo and Electronic Arts are the best in earning sales conquering almost 47% of the total sales.





Sales Insight



Sales by Games

We can observe a the trend and can see that Wiisports has the most sales all over the globe while on the other hand the top 5 platform has equal distribution of the sales. On other hand we have a trend line with the order of 5 polynomial and equation near to $6n^5$. All are facing loss if we compare it with the other years.

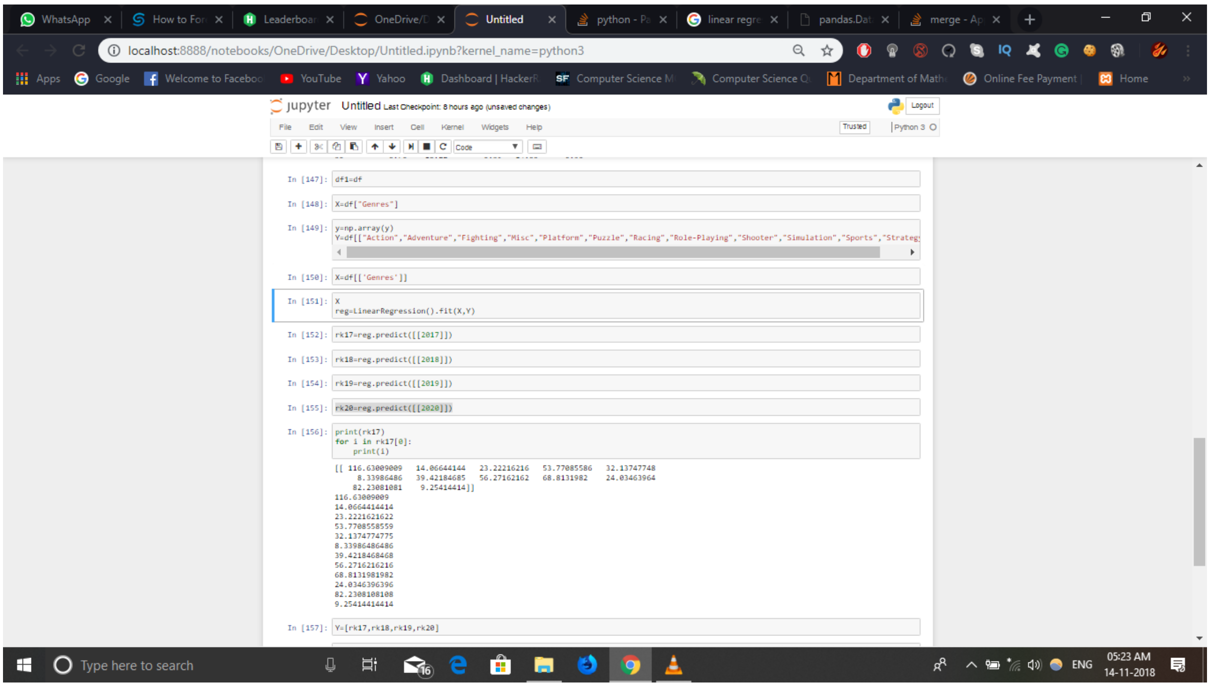
Objective 4: Predicting the Sales for the Upcoming year

Introduction

For this work I have assumed the positive trend because in 2017 lot of games were introduced and the market was positive for the games all over the globe so I have added the positive factor to the prediction and using the dataset from 1980 to 2016 I will be judging the sales for the year for the year of 2017, 2018, 2019, 2020. The prediction will be done on the basis of Genre that is which genre is most likely to get that amount of sales.

Description

In this module I have used the python for performing linear regression though we can perform it using excel but since my excel analysis toolkit is not working I have used it. But instead of creating model there I have created it all in excel so that I have to apply the directly algorithm over there. This is the best way of ETL process which we have discussed above.



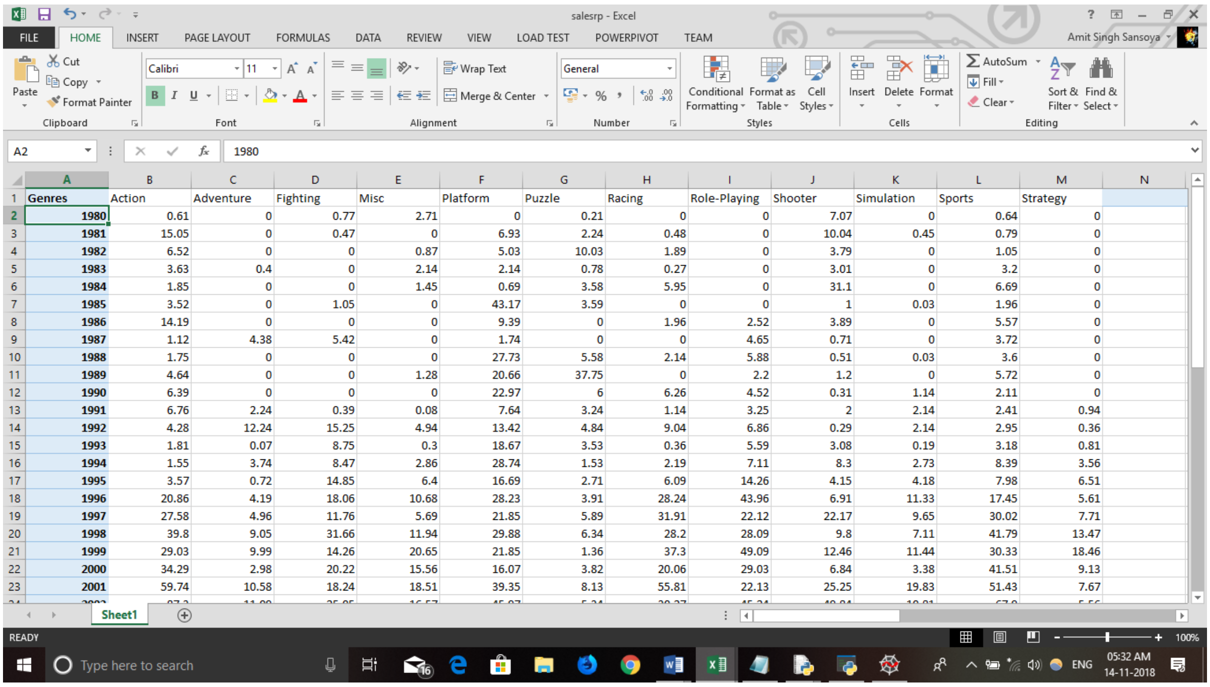
```
In [147]: df1=df

In [148]: X=df["Genres"]

In [149]: y=np.array(y)
          Y=df[["Action","Adventure","Fighting","Misc","Platform","Puzzle","Racing","Role-Playing","Shooter","Simulation","Sports","Strateg

In [150]: X=df[['Genres']]

In [151]: X
          reg=LinearRegression().fit(X,Y)

In [152]: rk17=reg.predict([[2017]])

In [153]: rk18=reg.predict([[2018]])

In [154]: rk19=reg.predict([[2019]])

In [155]: rk20=reg.predict([[2020]])

In [156]: print(rk17)
          for i in rk17[0]:
              print(i)

          [[ 116.63009009   14.06644144   23.22216216   53.77085586   32.13747748
               8.33986486   39.42184685   56.27162162   68.8131982    24.03463964
              82.23081081    9.25414414]]
          116.63009009
          14.0664414414
          23.2221621622
          53.7708558559
          32.1374774775
          8.33986486486
          39.4218468468
          56.2716216216
          68.8131981982
          24.0346396396
          82.2308108108
          9.25414414414

In [157]: Y=[rk17,rk18,rk19,rk20]
```

Implementation of Algorithm

Special Requirements, Formula’s

In this once require the knowledge of python and basic formatting of table, pivot tables in excel. The work needs to be done in such a way that we create the required model from the pivot table and that pivot table is changed to the table and passed in to other excel file or sheet in my case it is file and then Linear Regression is applied on it.

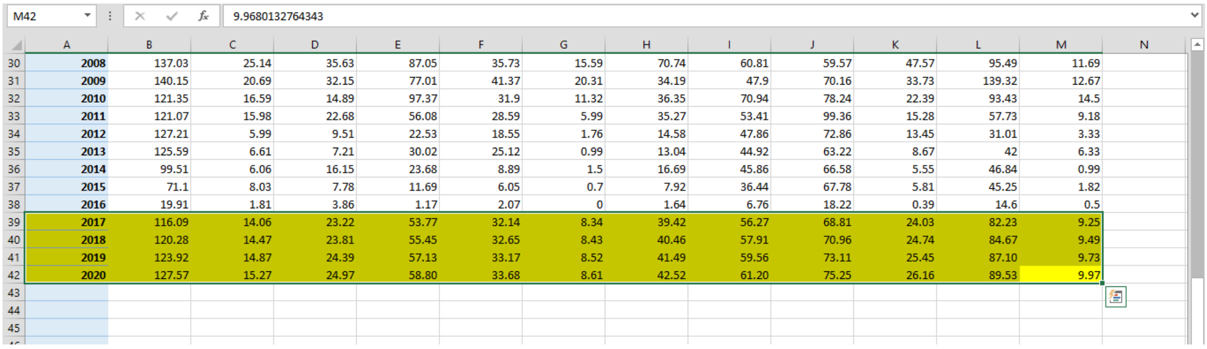
- Basic knowledge of Linear Regression
- Y=mx+C , basic graph reading line skill
- Sum of all the sales categorized by the genres and sorted out in the year



A2 1980

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Genres	Action	Adventure	Fighting	Misc	Platform	Puzzle	Racing	Role-Playing	Shooter	Simulation	Sports	Strategy
2	1980	0.61	0	0.77	2.71	0	0.21	0	0	7.07	0	0.64	0
3	1981	15.05	0	0.47	0	6.93	2.24	0.48	0	10.04	0.45	0.79	0
4	1982	6.52	0	0	0.87	5.03	10.03	1.89	0	3.79	0	1.05	0
5	1983	3.63	0.4	0	2.14	2.14	0.78	0.27	0	3.01	0	3.2	0
6	1984	1.85	0	0	1.45	0.69	3.58	5.95	0	31.1	0	6.69	0
7	1985	3.52	0	1.05	0	43.17	3.59	0	0	1	0.03	1.96	0
8	1986	14.19	0	0	0	9.39	0	1.96	2.52	3.89	0	5.57	0
9	1987	1.12	4.38	5.42	0	1.74	0	0	4.65	0.71	0	3.72	0
10	1988	1.75	0	0	0	27.73	5.58	2.14	5.88	0.51	0.03	3.6	0
11	1989	4.64	0	0	1.28	20.66	37.75	0	2.2	1.2	0	5.72	0
12	1990	6.39	0	0	0	22.97	6	6.26	4.52	0.31	1.14	2.11	0
13	1991	6.76	2.24	0.39	0.08	7.64	3.24	1.14	3.25	2	2.14	2.41	0.94
14	1992	4.28	12.24	15.25	4.94	13.42	4.84	9.04	6.86	0.29	2.14	2.95	0.36
15	1993	1.81	0.07	8.75	0.3	18.67	3.53	0.36	5.59	3.08	0.19	3.18	0.81
16	1994	1.55	3.74	8.47	2.86	28.74	1.53	2.19	7.11	8.3	2.73	8.39	3.56
17	1995	3.57	0.72	14.85	6.4	16.69	2.71	6.09	14.26	4.15	4.18	7.98	6.51
18	1996	20.86	4.19	18.06	10.68	28.23	3.91	28.24	43.96	6.91	11.33	17.45	5.61
19	1997	27.58	4.96	11.76	5.69	21.85	5.89	31.91	22.12	22.17	9.65	30.02	7.71
20	1998	39.8	9.05	31.66	11.94	29.88	6.34	28.2	28.09	9.8	7.11	41.79	13.47
21	1999	29.03	9.99	14.26	20.65	21.85	1.36	37.3	49.09	12.46	11.44	30.33	18.46
22	2000	34.29	2.98	20.22	15.56	16.07	3.82	20.06	29.03	6.84	3.38	41.51	9.13
23	2001	59.74	10.58	18.24	18.51	39.35	8.13	55.81	22.13	25.25	19.83	51.43	7.67

Created model for the analysis/prediction

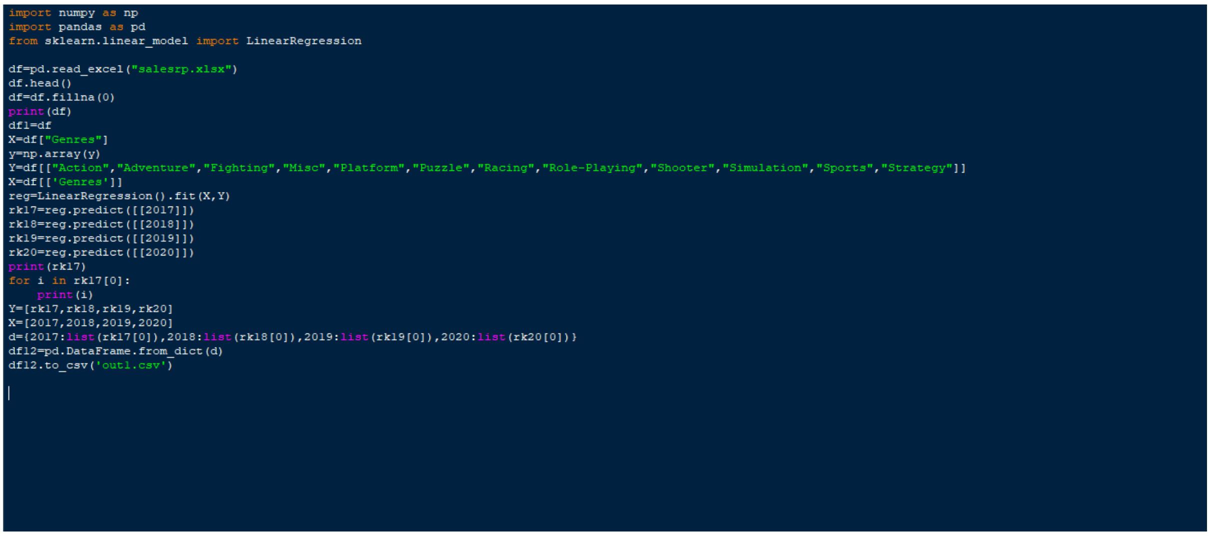


M42 9.9680132764343

	A	B	C	D	E	F	G	H	I	J	K	L	M
30	2008	137.03	25.14	35.63	87.05	35.73	15.59	70.74	60.81	59.57	47.57	95.49	11.69
31	2009	140.15	20.69	32.15	77.01	41.37	20.31	34.19	47.9	70.16	33.73	139.32	12.67
32	2010	121.35	16.59	14.89	97.37	31.9	11.32	36.35	70.94	78.24	22.39	93.43	14.5
33	2011	121.07	15.98	22.68	56.08	28.59	5.99	35.27	53.41	99.36	15.28	57.73	9.18
34	2012	127.21	5.99	9.51	22.53	18.55	1.76	14.58	47.86	72.86	13.45	31.01	3.33
35	2013	125.59	6.61	7.21	30.02	25.12	0.99	13.04	44.92	63.22	8.67	42	6.33
36	2014	99.51	6.06	16.15	23.68	8.89	1.5	16.69	45.86	66.58	5.55	46.84	0.99
37	2015	71.1	8.03	7.78	11.69	6.05	0.7	7.92	36.44	67.78	5.81	45.25	1.82
38	2016	19.91	1.81	3.86	1.17	2.07	0	1.64	6.76	18.22	0.39	14.6	0.5
39	2017	116.09	14.06	23.22	53.77	32.14	8.34	39.42	56.27	68.81	24.03	82.23	9.25
40	2018	120.28	14.47	23.81	55.45	32.65	8.43	40.46	57.91	70.96	24.74	84.67	9.49
41	2019	123.92	14.87	24.39	57.13	33.17	8.52	41.49	59.56	73.11	25.45	87.10	9.73
42	2020	127.57	15.27	24.97	58.80	33.68	8.61	42.52	61.20	75.25	26.16	89.53	9.97

Predicted Values from the analysis/prediction

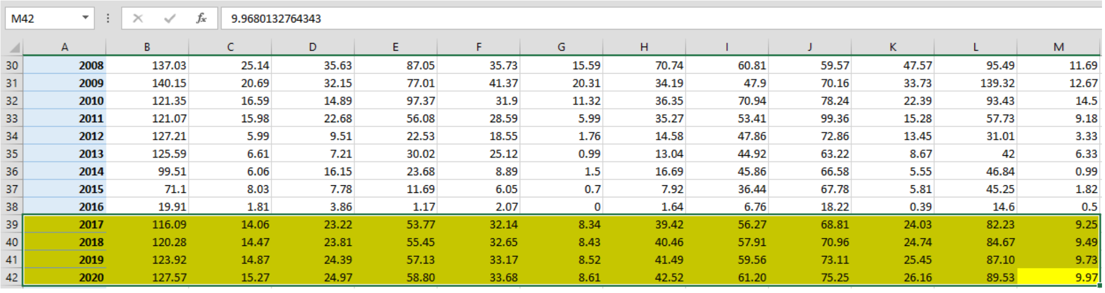
Basic Code Written In python



```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

df=pd.read_excel("salesrp.xlsx")
df.head()
df=df.fillna(0)
print(df)
df1=df
X=df["Genres"]
y=np.array(y)
Y=df[["Action","Adventure","Fighting","Misc","Platform","Puzzle","Racing","Role-Playing","Shooter","Simulation","Sports","Strategy"]]
X=df[['Genres']]
reg=LinearRegression().fit(X,Y)
rk17=reg.predict([[2017]])
rk18=reg.predict([[2018]])
rk19=reg.predict([[2019]])
rk20=reg.predict([[2020]])
print(rk17)
for i in rk17[0]:
    print(i)
Y=[rk17,rk18,rk19,rk20]
X=[2017,2018,2019,2020]
d={2017:list(rk17[0]),2018:list(rk18[0]),2019:list(rk19[0]),2020:list(rk20[0])}
df12=pd.DataFrame.from_dict(d)
df12.to_csv('out1.csv')
```

Analysis Results

The data was successfully loaded and predicted and loaded back to the excel sheet and the new prediction line was successfully generated on the basis of predicted values which pointed towards the most of the positive trend. Since the data has only sales record as the input the features for determining were less. So this may shall not be taken as serious values these are just the values which were generated by applying the linear regression on the model i.e on the previous sales so these may not be correct.



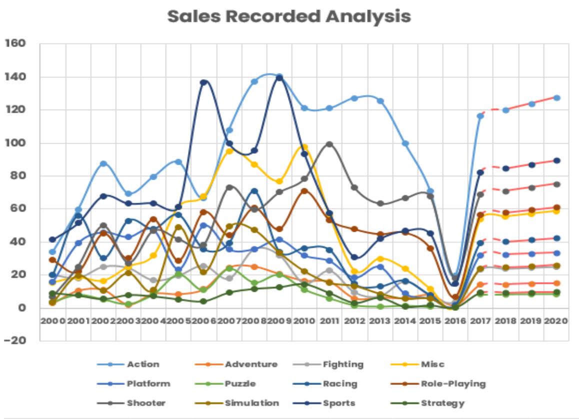
M42 | 9.9680132764343

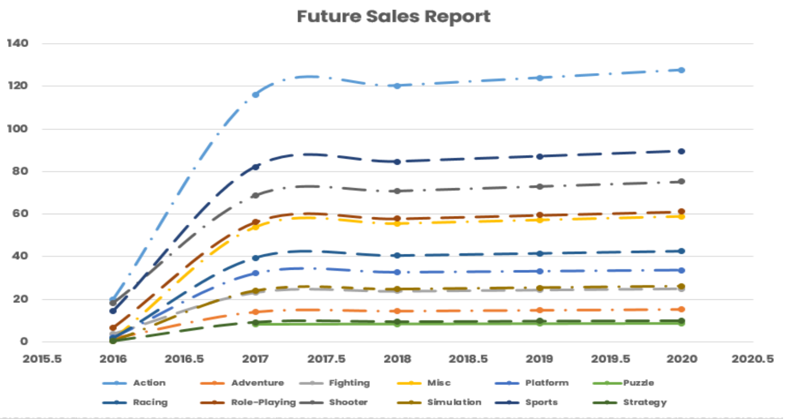
	A	B	C	D	E	F	G	H	I	J	K	L	M
30	**2008**	137.03	25.14	35.63	87.05	35.73	15.59	70.74	60.81	59.57	47.57	95.49	11.69
31	**2009**	140.15	20.69	32.15	77.01	41.37	20.31	34.19	47.9	70.16	33.73	139.32	12.67
32	**2010**	121.35	16.59	14.89	97.37	31.9	11.32	36.35	70.94	78.24	22.39	93.43	14.5
33	**2011**	121.07	15.98	22.68	56.08	28.59	5.99	35.27	53.41	99.36	15.28	57.73	9.18
34	**2012**	127.21	5.99	9.51	22.53	18.55	1.76	14.58	47.86	72.86	13.45	31.01	3.33
35	**2013**	125.59	6.61	7.21	30.02	25.12	0.99	13.04	44.92	63.22	8.67	42	6.33
36	**2014**	99.51	6.06	16.15	23.68	8.89	1.5	16.69	45.86	66.58	5.55	46.84	0.99
37	**2015**	71.1	8.03	7.78	11.69	6.05	0.7	7.92	36.44	67.78	5.81	45.25	1.82
38	**2016**	19.91	1.81	3.86	1.17	2.07	0	1.64	6.76	18.22	0.39	14.6	0.5
39	**2017**	116.09	14.06	23.22	53.77	32.14	8.34	39.42	56.27	68.81	24.03	82.23	9.25
40	**2018**	120.28	14.47	23.81	55.45	32.65	8.43	40.46	57.91	70.96	24.74	84.67	9.49
41	**2019**	123.92	14.87	24.39	57.13	33.17	8.52	41.49	59.56	73.11	25.45	87.10	9.73
42	**2020**	127.57	15.27	24.97	58.80	33.68	8.61	42.52	61.20	75.25	26.16	89.53	9.97

Predicted Values from the analysis/prediction

Visualization

Once the analysis and predicted results were obtained we checked for the visualization and since the code as we can see that I predicted on the large output scale providing data from 1980 to 2016 we got the positive to negative impact so the games now on would try to show the repetitive nature but with less amplitude. Here are the below observation obtained.





The left one is record from the 2000 to 2020 the dotted lines in that suggests the starting of predicted line and we can see the positive growth.

The right one is zoomed version of different line graph on it so that we can clearly understand the growth from 2017 to 2020.

Objective 5: Listing out the famous publications/Game developing companies

Introduction

This field is about the study of the publications here we would look after the fact that how the publications have been affected or have affected the data structure. Finding out which company has most number of games as the result will tell us the popular publication just like the genre analysis and we would look after top earning game publishers.

Description

We would look at the values and plot the top 10 or 13 companies’ best by count and sales as well as best by the launches as per last three years of the time. We would find the correlation and would see if the hypothesis is valid just as we did for the previous one.

Special Requirements, Formula’s

There is a requirement of multiple pivot tables between many things and the co-relation function as it would be also proving the fact of the hypothesis.

Required Stuff

1. Sum of all the count of the publication categorized by the various platform of the publications
2. Sum of publications by sales
3. Knowledge of Correlation

$$f(x) = \sum_{n=1}^{n} (col[n])$$

Correlation in Excel =Correl(array1,array2)

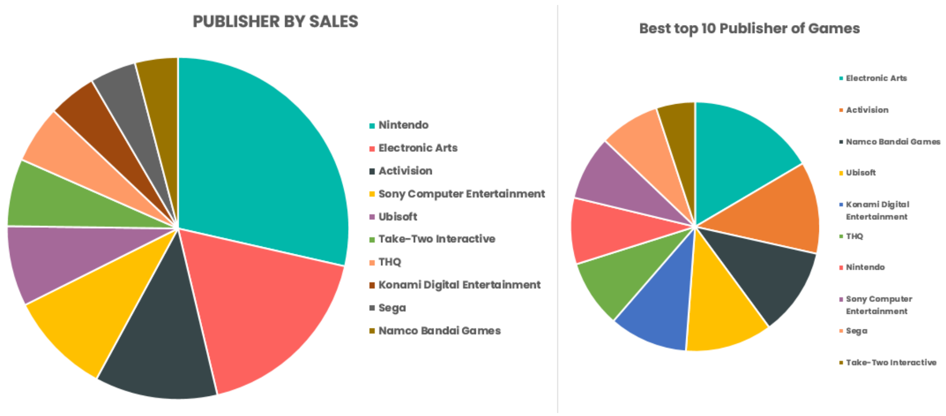
Analysis Results

The result can be seen that in sales Nintendo is leading followed by the EA.

The count of the games published by EA.

Correlation Result -90.1% sales vs count of the publication

Visualization



Pie Chart Regarding Sales vs Publisher and Launched Games

List of Analysis and Results

- **Analysis on Genre**
  - Correlation between Genre and Global Sales to Justify the popularity of games
  - Result Correlation Exists – Positive Trend
  - Correlation % - 88.1%
- **Analysis on Platform**
  - Correlation between Platform and Global Sales to justify that the popularity of games were more due to platforms
  - Result Correlation Exists – Positive Trend
  - Correlation % - 91%
- **Prediction on the Sales**
  - Predicting the sales on the basis of previous year sales
  - Created positive trend as 2017 passed
  - Linear Regression Applied
  - Successfully obtained the required results
- **Analysis on the Sales**
  - Obtained the declining nature in the selling of games becoming less popular, the prediction says they are more likely to go up
- **Analysis on Publications**
  - Finding out the top companies which produces games on the basis of the count and sales
  - Result – EA, Nintendo
  - Correlation % – Positive 90%
- **Other Analysis – Basic**
  - Best Genre – Sports Rank 1
  - Most Spread Genre – Action
  - Most Developed Games – EA
  - Most Earned Game – WII Sports
  - Most Used Platform – Nintendo DS
  - Most Sales done by Genre – Action
  - And many more

Future Scope

As talking about the future scope there will be time when the sales would increase when the sales would increase again as the prediction says so. As per the available data we have found that the sales are declining as the year passes this is mostly because of the non-availability of one most important platform from the dataset which is mobile phones. Mobile phones game have become very famous and are earning lot of money if added it to the dataset the accuracy would increase for the prediction and we would get the new factor. However we get to know that in near future the game which would be most famous genre would be the action. Moreover Nintendo was holding the market for large time but it seems that it would be now taken by new platforms like smartphones. Mobile games are the future and seeing the data I feel that was the only reason why the sales of the games went falling down form the period of 2005 onwards. On thinking more deep we can say that publisher of game matters and in the near future the company with the better image and reputation can with stand in the world and people like to play the games from the trusted vendors like EA only. Moreover then comes the platform which affects the sales for now we know it has to be mobile phones but literally talking about the gaming console it has to be the PS4 or XBOX ONE.

Concerning about the games if we consider the future the sales are more likely to go up and raise the sales. The data had linear progression we would go for the pattern like structure and we can find the boosts in sale in between.

Further Future Use:

- The analysis will help us to know the future of the gaming types as we can suggest the most upcoming types of games which can be helpful the game development industry.
- The analysis is important for the game development companies as they would be focused on grabbing the most sales and best rank which they can attain.
- The analysis will help the companies to keep the track of the games sales and would also help them to know the best way where they would know which platform had helped them to reach that rank or would help them to gain the popularity.

References

- Mr Hargobind Singh
  (Assistant Professor)
  Lovely Professional University

- Pradeep Kumar
  (Assistant Professor)
  Lovely Professional University

BIBLIOGRAPHY

- https://www.kaggle.com/
  Kernels and DataSets

- https://stackoverflow.com/
  Doubts and basic programming stuff

- https://math.stackexchange.com/
  For mathematical help and other