Data Analysis Portfolio

Prepared By:-
Satyabrata Sahoo

Professional Background

Currently in my final year pursuing B.Tech-CSE. I have secured 8.54 CGPA (till 7th sem) and have several skills including Machine Learning, Data Analysis, Python, Cloud Computing.

I have worked with several companies as an intern like Technocolabs, Internshala etc, as a Machine Learning Engineer. I have also worked as a Project Manager with workskills where I have worked on their Data Analytics course from scratch and managed different teams.

I have also published a research paper titled- "APPLICATION OF MACHINE LEARNING ALGORITHMS IN BANKING SECTOR FOR LOAN PREDICTION" in a journal and have worked on several projects related to web development, data analysis, machine learning.

As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work. Being a fresher, I think I am very flexible and adaptive to learn new things. I have theoretical knowledge. But I am waiting to use my theoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

Table Of Contents

Udemy Project Description

The dataset used for the analysis is Udemy Course Data of various courses containing various columns and tuples. This report shows the findings and the insights derived after cleaning the data and plotting various plots to find the relationship among various items. The main aim of this project was to find insights among the data provided.

The Problem

You’re a Data Analyst working for the education tech company Udemy. You have been asked by your manager, Head of Curriculum at Udemy, to present the data on course revenue, and you have been provided with data on courses from different topics to understand where opportunities to increase revenue may lie, and track the performance of courses.

Your manager has suggested encouraging Web Development courses to charge more because she believes that these are the most popular courses. She needs to send a report to the CEO in the next three weeks on how they will increase their next quarterly earnings.

Design

Steps taken to clean the data:

- First importing the datasets provided and merging them.
- Then removing the duplicates and the blank cells.
- Improving the headers of each column with proper values.

Finding and replacing the data with correct values.

Tools used for visualization:

- Excel
- Tableau

Finding-1

subject	SUM of num_su
Business Financ	1868711
Graphic Design	1063148
Musical Instrume	846689
Web Developme	7981935
Grand Total	**11760483**

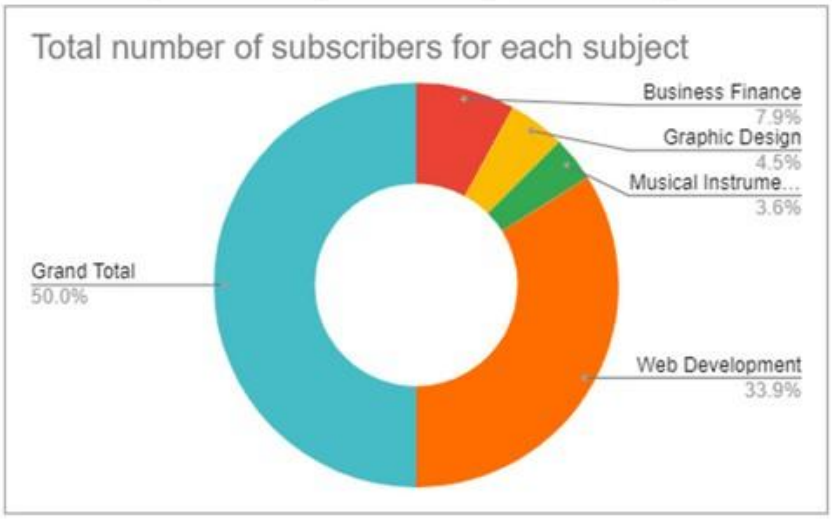


Fig. 1: Total No. of Subscribers for each subject

When plotting total number of subscribers for each subject (Excel) as in Fig. 1 , I found out that "Web Development" was the most subscribed course while "Musical Instruments" the least subscribed.

Finding-2

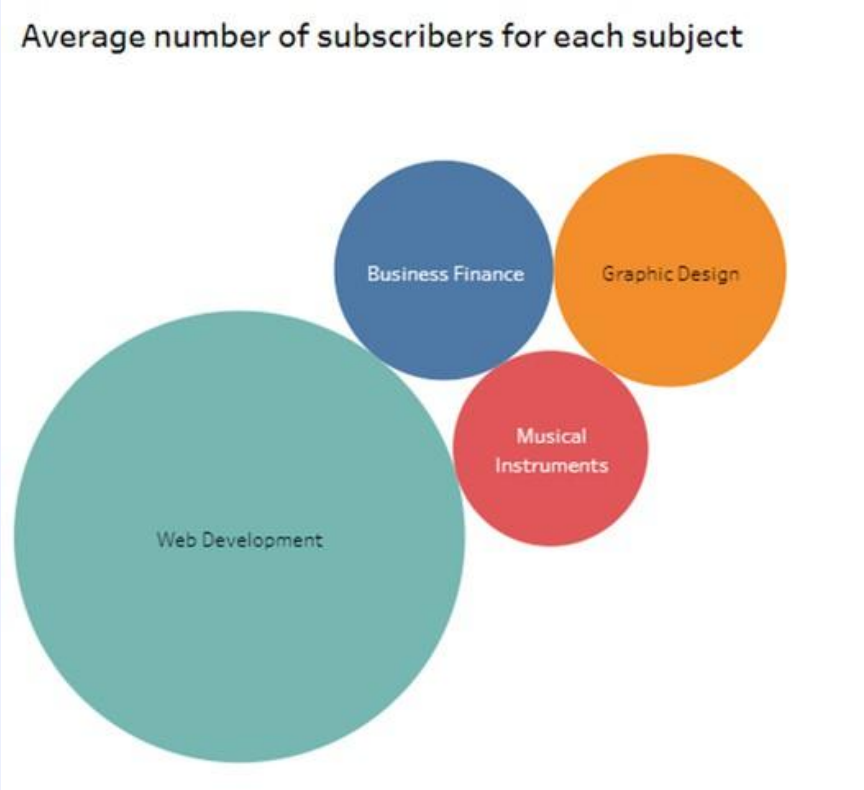


Fig. 2: Avg. No. of subscribers for each subject

From this plot (tableau) Fig. 2 I can clearly infer that the "Web Development" course has the most number of average subscribers.

Finding-3

SUM of price	level				
subject	All Levels	Beginner Level	Expert Level	Intermediate Lev	Grand Total
Web Developme	47190	33145	940	11400	92675
Musical Instrume	16065	13030	580	4025	33700
Graphic Design	20810	9325	200	4515	34850
Business Financ	44435	27425	1645	8310	81815
Grand Total	**128500**	**82925**	**3365**	**28250**	**243040**

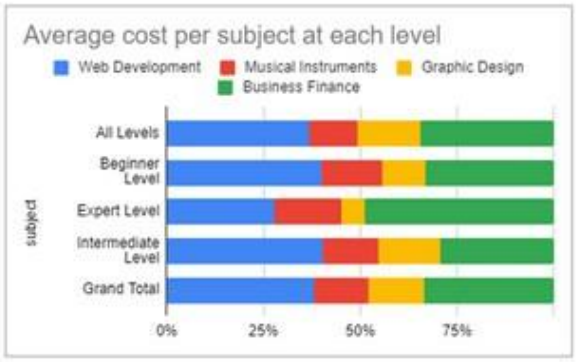


Fig. 3: Avg. cost per subject at each level

While creating the pivot table as in Fig. 3 to find average cost per subject at each level, I found out that at all levels- "Web Development" cost was the highest and "Musical Instruments" costed the lowest. Similarly for each levels the costliest course was- Beginner level it is Business Finance, Expert level it is again Business Finance, Intermediate level it is Web Development. Like that I can see for the cheapest ones also. Plotting the column chart also helps show the distinction among them.

Finding-4

subject	AVERAGE of co
Web Developme	5.59
Musical Instrume	2.85
Graphic Design	3.59
Business Financ	3.56
Grand Total	**4.10**

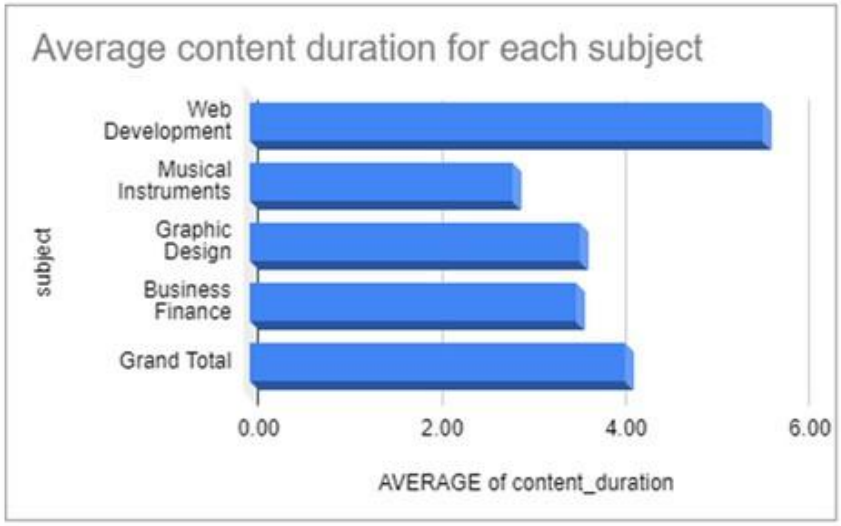


Fig. 4: Average Content Duration

From the above plot (Fig. 4), we found out that "Web Development" course was the longest in time and "Musical Instruments" the shortest.

Finding-5

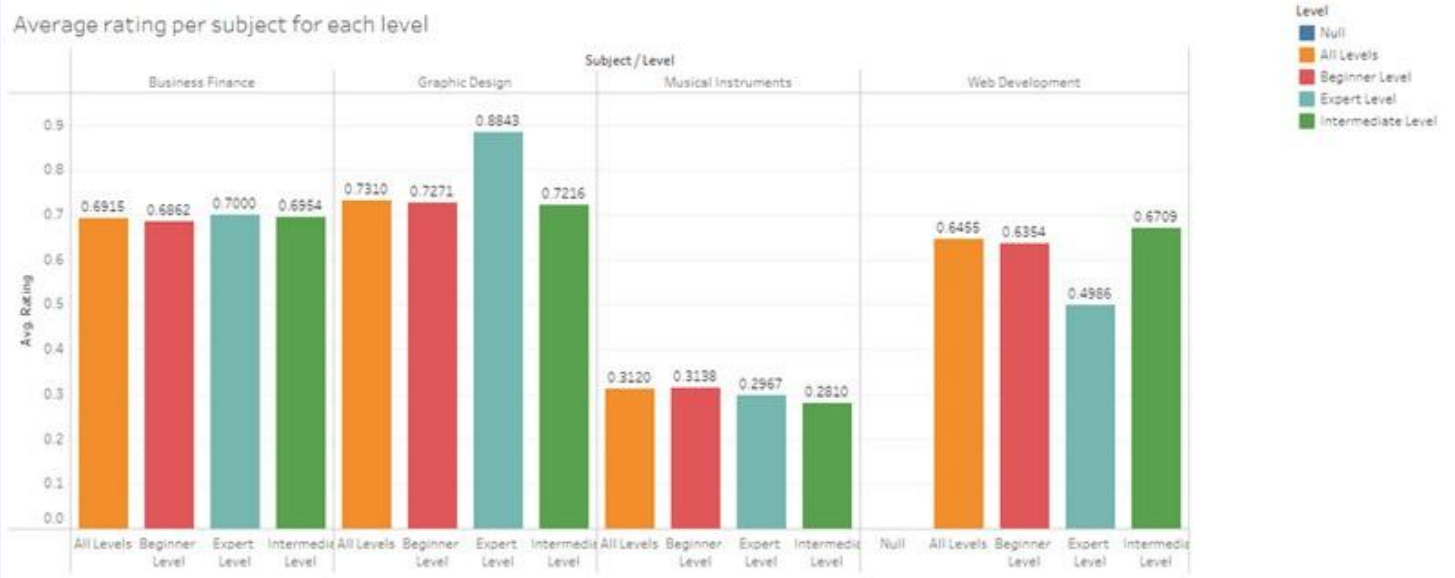


Fig. 5: Avg. rating per subject at each level

After plotting the average rating per subject at each level (Fig. 5) I inferred that "Graphic Design" had the highest rating at each level, "Business Finance" had consistent ratings at each level, "Musical Instruments" had the lowest rating overall.

Analysis

After the analysis, I found out that "Musical Instruments" was the least subscribed course domain among all with the least rating and least course duration.

Using the 5 Whys approach I am finding the root cause:-
"Why is it least subscribed?"
Maybe the marketing done is not up to the mark.

"Why its duration is low?"
-Trainers are not readily available.

"Why it has least rating?"
-The quality of the course is not good.

"Why its price is low?"
-Because of low subscription.

"Why are we continuing with the course?"
-To diversify the content available.

Conclusion

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out the courses that were a burden for Udemy to continue providing. Analysing the data proved helpful in finding various issues among the courses.

Capstone Project Description

The dataset that I chose for the project is from- “kaggle.com” and the title of the dataset is “GPUs Specs” which has different columns dealing with the specifications of different GPUs produced over the years. The problem statement to be considered is- “You are provided a dataset containing different manufactures of GPUs over the years and with the specifications of their GPUs. There was a period when the clock speed of GPUs was stagnant and increased at a very slow pace. Identify that period and derive insights from the data.”

Design

Steps taken to clean the data:

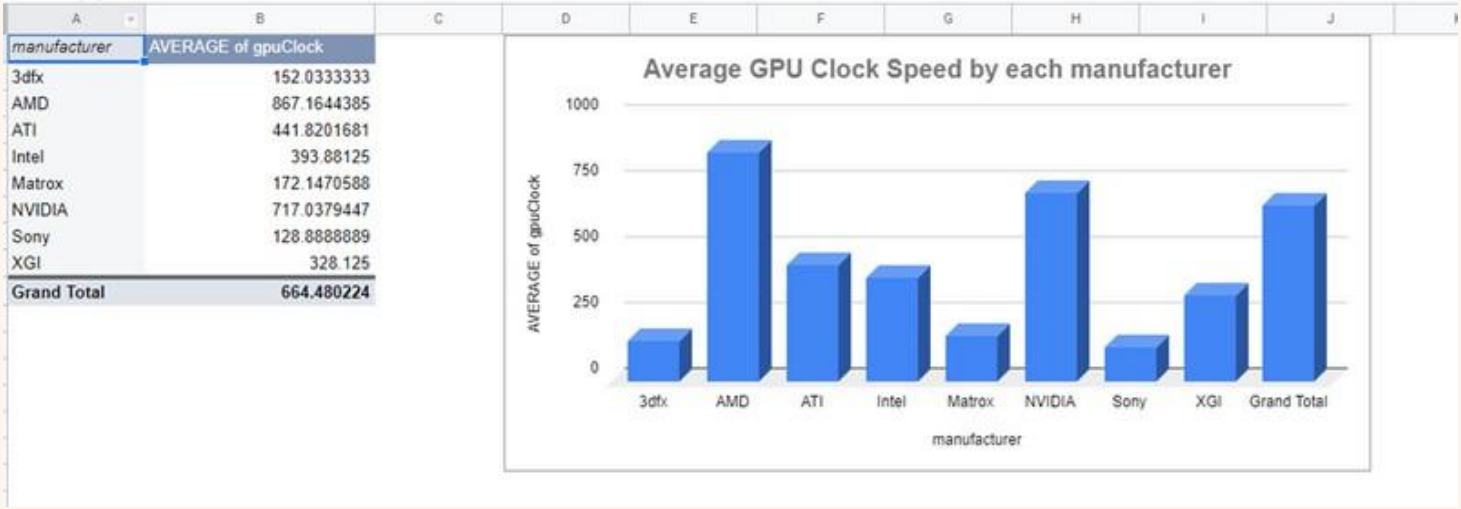
- First importing the dataset provided.
- Then removing the duplicates and the blank cells.
- Improving the headers of each column with proper values.

Finding and replacing the data with correct values.

Tools used for visualization:

- Excel
- Tableau

Finding-1



manufacturer	AVERAGE of gpuClock
3dfx	152.0333333
AMD	867.1644385
ATI	441.8201681
Intel	393.88125
Matrox	172.1470588
NVIDIA	717.0379447
Sony	128.8888889
XGI	328.125
Grand Total	664.480224

Fig. 1: Average GPU Clock Speed by each manufacturer

When plotting average gpu clock speed by each manufacturer (Fig. 1), we found out that AMD and NVIDIA are the leading competitors with the highest speeds. SONY and 3dfx had the lowest speeds of their respective gpus.

Finding-2

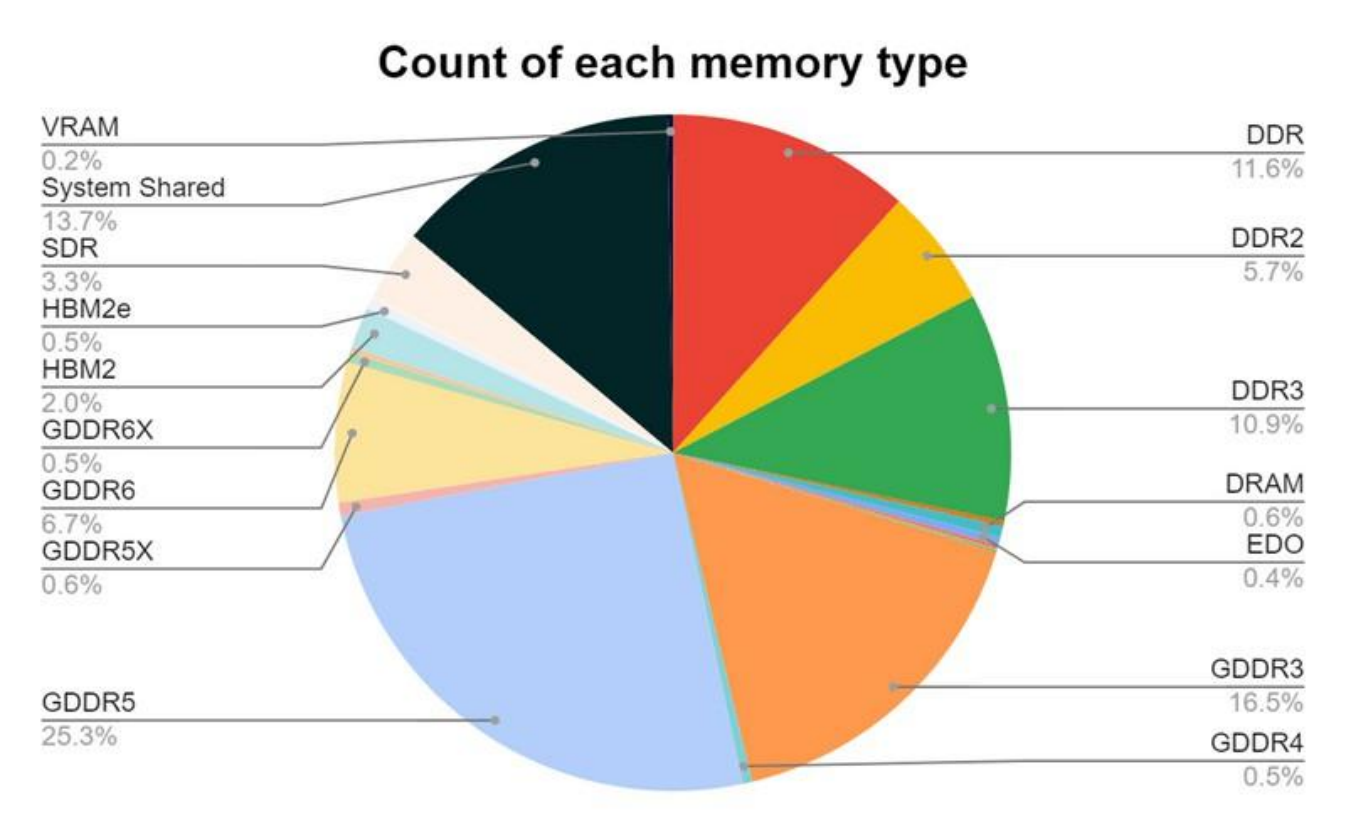


Fig. 2: Count of each memory type

From this plot (Fig. 2) we can infer that GDDR5 was the most sold memory type as it accounted for 25.3% among all the memory types manufactured. From this we can infer that maybe this was the period when the growth speed of clock speeds was stagnant.

Finding-3

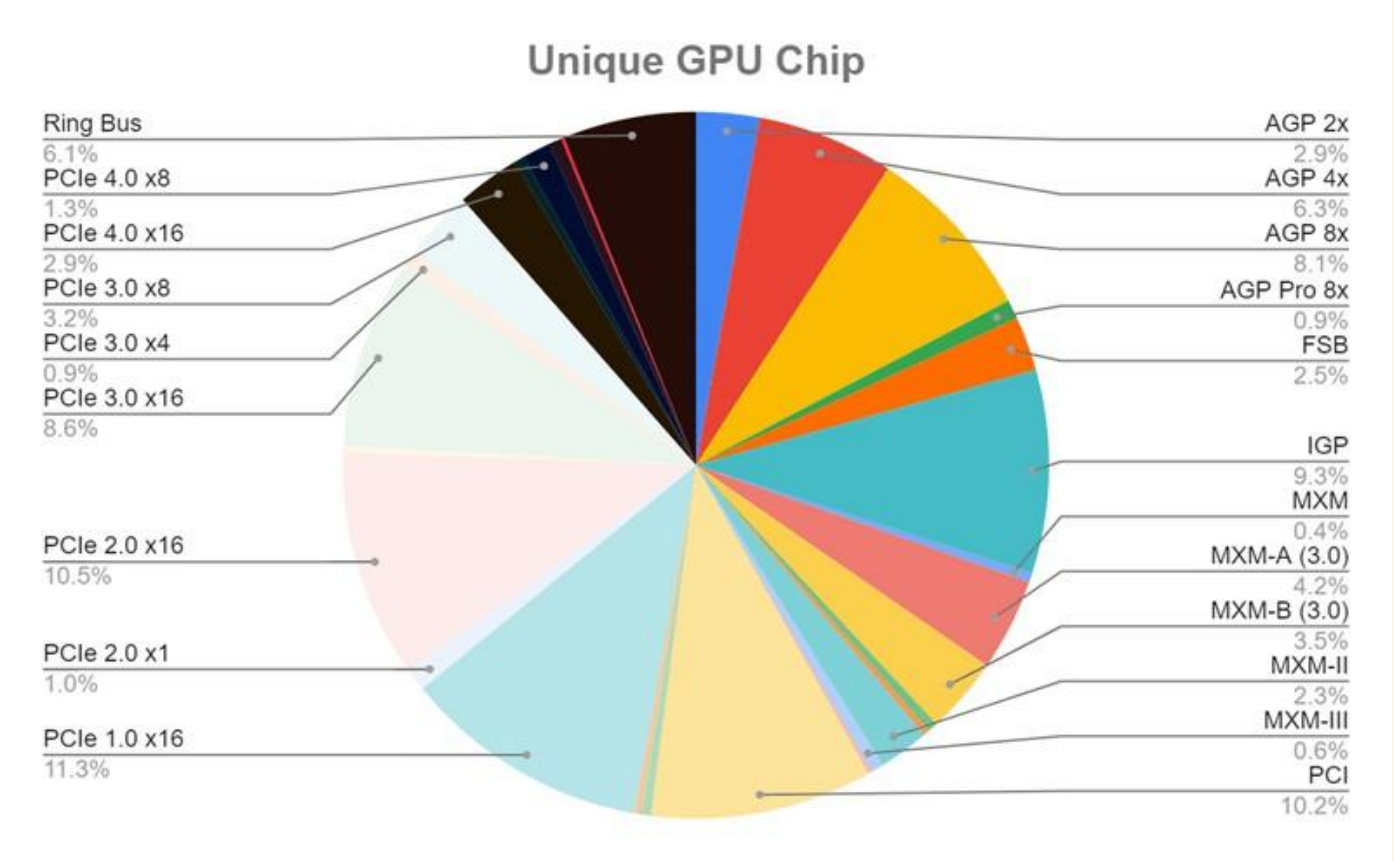


Fig. 3: Unique GPU Chip

From the above plot (Fig. 3) we can clearly see that PCIe 1.0 x16 and PCIe 2.0 x16 are the most produced chip type. PCI generation is most widely produced chip.

Finding-4

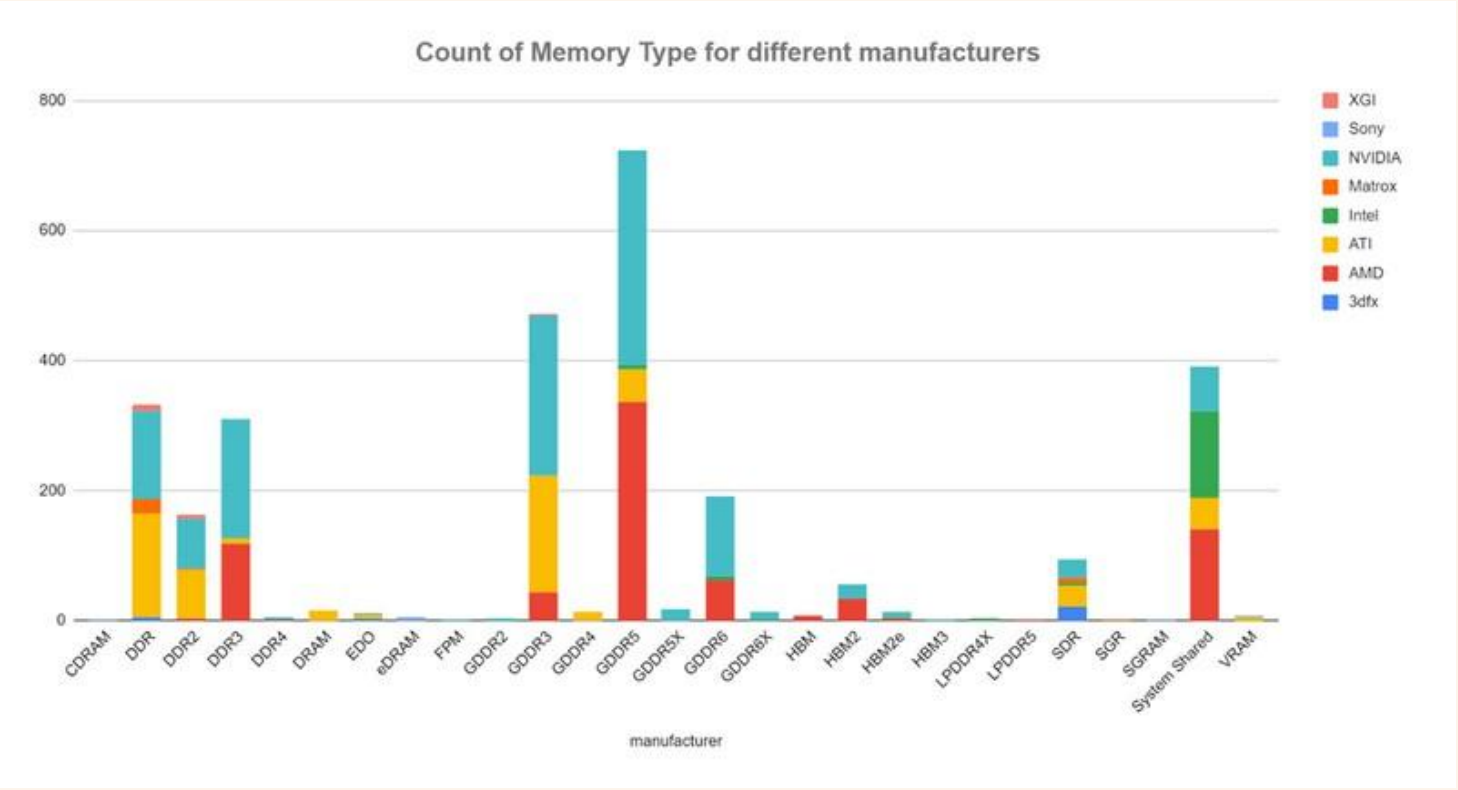


Fig. 4: Count of Memory Type for different manufacturers

From the above plot (Fig. 4), we found out that GDDR5 was the most produced chip by all manufacturers, NVIDIA and AMD the leading manufacturers. GDDR3 was also widely produced and used.

Finding-5

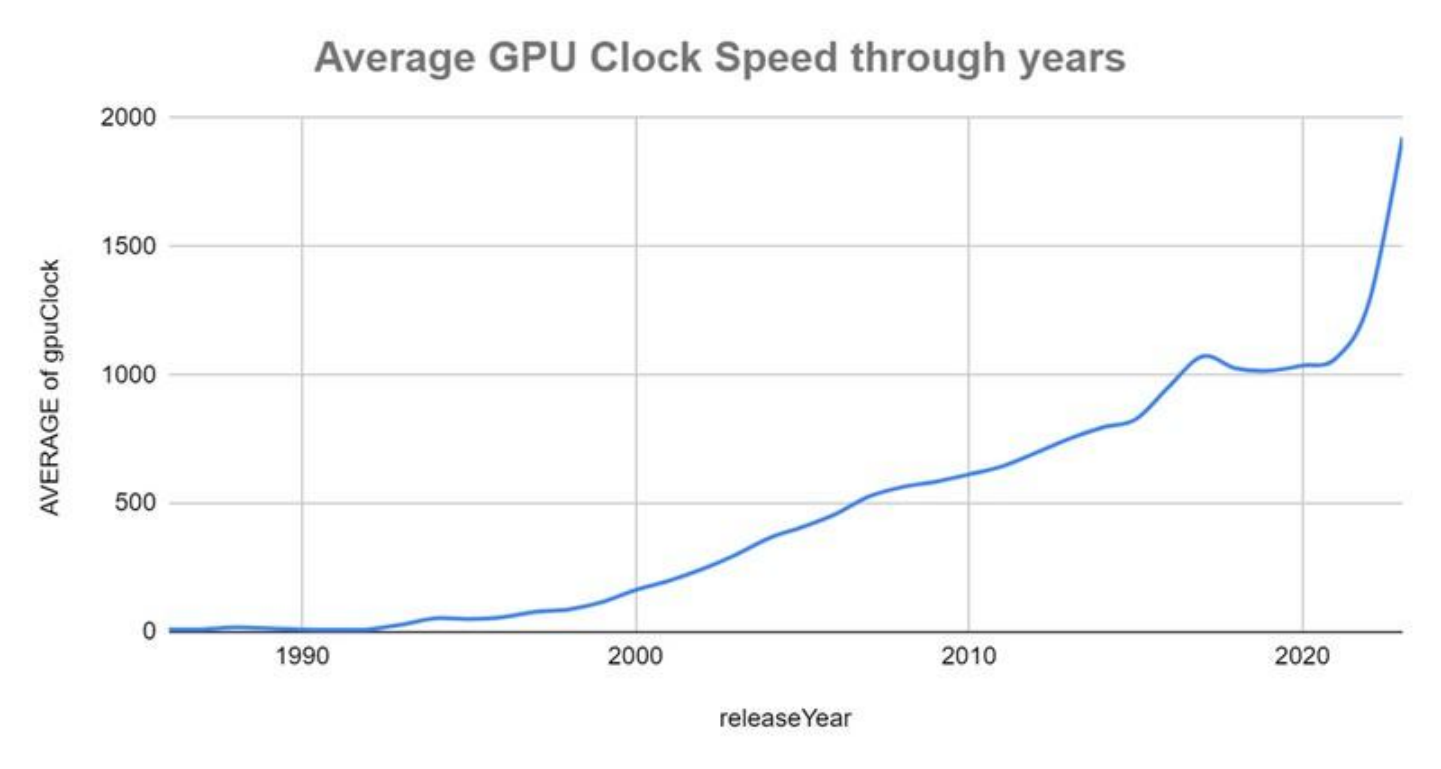


Fig. 5: Avg. GPU Clock Speed through years

As we can see from Fig. 5 the clock speed of GPUs constantly increased through the years but there was a significant increase or uptick since the year 2020. From 1990 to 2000 there was no significant improvement but from 2000 to 2010 there was improvement as these were the years in which the gaming industry was properly established and after 2020 there has been a huge demand of graphic cards for various purposes.

Finding-6

Fig. 6: Count of products produced by each manufacturer

We can clearly see that NVIDIA and AMD are the leading producers, ATI and AMD merged into one company in later years.

Analysis

After the analysis, I found out that "GDDR5" was most produced GPU type and AMD, NVIDIA are the leading manufacturers.

Using the 5 Whys approach I am finding the root cause:-
"Why is it most produced?"

-Consumers liked it the most.

"Why other manufacturers are lagging?"

-Maybe the marketing done is better than the rest.

"Why other GPU types didn't work out?"

-The quality of others were not good.

"Why its production is high?"

-Because of high demand

"Why AMD and NVIDIA are market leaders?"

-Because of better quality.

Conclusion

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data which once looked useless gave some very useful insights.

Appendix

Google Sheets Link for Udemy Project:-

https://docs.google.com/spreadsheets/d/1x5BsZ1q5Uz9mDIW373t0TN2EVQaOUDHkr5giL4Epuac/edit?usp=sharing

Tableau Link for Udemy Project:-

https://public.tableau.com/views/Entry-Level_Satyabrata/Sheet2?:language=en-US&:display_count=n&:origin=viz_share_link

Google Sheets Link for GPU Specs Project:-

https://docs.google.com/spreadsheets/d/1J-z4ROdrV7N0spnjs7Ihq8nhUUwjQ-GPHFl6V--HOKk/edit?usp=sharing

Tableau Link for Udemy GPU Specs Project:-

https://public.tableau.com/views/CapstoneProject_Satyabrata/S heet1?:language=en-US&:display_count=n&:origin=viz_share_link