



Huddersfield business school

Student Name and ID	Md Junayed Hossain ; 2251090321
Unit Name and Code	Managing Big Data; BMD004-QGF-2122
Module Instructor	Mr. Munshi Ahmed (Razib Ahmed)
No. of Assignment	Assignment 01
Due Date	11 March, 2022
Date Submitted	10 March, 2022

Declaration:

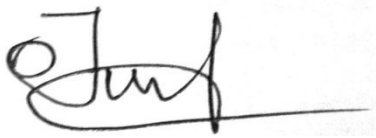
☐ I hold a copy of this assignment if the original is lost or damaged.

☐ I hereby certify that no part of this assignment or product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.

☐ I hereby certify that no part of this assignment has been submitted by me for any other assessment or if it has that appropriate referencing has been provided.

☐ No part of the assignment/product has been written/produced for me by any other person except where collaboration has been authorized by the subject lecturer/session concerned

☐ I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism *(which may retain a copy on its database for future plagiarism checking)*



Signature:

Contents

1. Introduction.....	3.
2. Influence of Big Data Analytics to Business Operations and Performances.....	3
3. Important of Big Data to Netflix.....	4
4. Big Data Analytics Best Practices Within Entertainment Industry	4
5. Data brief	5
6. Data cleansing and preparation	5
7. Key metrics and relationships	7
1. Sum of Sales Gender-wise	7
2. Dominant Payment Methods and Consumer	8
3. Profit Earned in States	9
4. The Relation Amongst Referrals, Sales and Profit	10
5. Age Group VS Segments	11
8. Strategic Insight	11
9. Conclusion	13
10. Appendices.....	14
Appendix 1	14
Appendix 2.....	14
Appendix 3.....	15
Appendix 4.....	15
Appendix 5.....	16
Appendix 6.....	15
Appendix 7.....	16
Appendix 8.....	17
11. References:.....	17

1. Introduction

This report consists of a couple of parts, Part A(1) and Part B(2). The first part will exhibit the impact of big data analytics (BDA) on business operations and organization performances along with the significance of BDA to Netflix and the best implementation in the entertainment industry.

The latter part will talk about the dataset(s) and will explore some relations amongst the features showing some pictorial data views and their deep insights with the aid of the Business Intelligence tool, Tableau.

1.1. Influence of Big Data Analytics to Business Operations and Performances

“In God we trust, all others bring data.” — W. Edwards Deming illustrates the significance of the big data analysis and measurement while going through the business process (Hansen, 2019). It’s nearly impossible to grasp this open market without proper data-based knowledge. Generally, data analytics techniques and technologies provide a pathway to assess and analyse datasets to gather insights about business operations and performances (Chai et al., 2021). They also mention that big data analytics is paramount for organizations to end up with data-driven decisions on the business operations—refers to everything an organization does every day to keep running and earning revenue—with effective strategies. Nowadays, companies go through data analytics to identify targeted markets so that the money spent on the campaigns don’t go in vain. According to a recruiting analytics blog post of Pierpoint (n.d.), the success of an organization depend on its workforce and data analytics seems to be the new hero in recruiting. It also states that hiring managers can rely on data science to select the fittest candidates for the organization. Overall, the influence of data analytics is omnipresent in any industry since we all go through the decision-making process before embarking upon something.

1.2. Importance of Big Data to Netflix

Netflix, an American streaming company, has been selected for portraying the significance of big data. Big data and thereby its analysis has evolved the scenario of Netflix dramatically. Using big data Netflix has brought a revolution to the automation process, for instance, it goes through a Machine Learning Algorithm utilizing the data to comprehend consumers' understanding of what they might watch next according to their preference and taste (Hayden, 2018). For example, if you watched a sci-fi genre movie, you will get recommendation of that type rated movies (DataFlair, 2019). This feature reduces customer searching efforts and increases retention. This recommendation feature, is the hook for its massive success in this industry, provides the best customer services that they need not get perplexed about their watching stack. afterwards, Netflix has epitomized the value of big data in the operational optimization in the business process such as it uses predictive algorithms to estimate the cost for filming in one location vs. another and also predicts post-production things that mitigate bottlenecks—congestion in a production system and streamlining workflows (Mixon, 2021) and also exhibits customers buying patterns and behaviour (Dixon, 2019) which help the company improve its supply chain management to the end customers. Overall, Netflix uses big data for production planning to identify perfect shoot location, day and time of shooting to reduce expenses. Moreover, it is taking full advantage of big data for margin enhancement. For example, "Stranger Things" was predicted to have a view of 100 million people and thus they were eager to invest 500 million in that series. The company reported that they pocketed \$665 million profits on 6.2 billion revenue with 6.8 million new customers after this series (Lee, 2019). Simply, using big data they bring out the right-size investment (Biddle, 2022). And consequently, according to Iqbal (2022), Netflix had revenue of \$3.1 billion in 2011 whereas it stood at almost 25 billion in 2020 with the aid of big data and DA. The figure is proof of how much is big data significant for Netflix.

1.3. Big Data Analytics Best Practices within the Entertainment Industry

The best practice within this entertainment industry such as Amazon Prime Video, BBC iPlayer, Disney+ and HBO Max etc. is numberless. Apart from making personalized recommendation shows, the media companies must invest in the content that they will be commercially successful where these companies use big data analytics to gain prior

knowledge about the customers' views. For example, Netflix has gone through this process for an American version of the British show, House of Cards (Shah, n.d.). Moreover, the major earning source of these media and entertainment companies is advertising. These media streaming companies have been left with no other option other than using BDA to ensure relevant ads to the viewers as per their geographic locations such as media streaming company, NBC has been doing these according to Emma (2019). Thereabouts all the entertainment companies do pricing research using Big Data Analytics before streaming any shows to ascertain that price is in customer range and thereby they earn money from customers. To infer, not only the entertainment industry but also others have adopted this new miracle, Big Data Analytics

2. Data Brief

We are living in an era of Data Science where data are a million-dollar market. Yet, there are several platforms where we can use secondary data for free such as Kaggle, Data World, Data Bank, and so on. It is imperative to have credible data sources for market analysis, and Kaggle has more than 100000 data currently (Joyce, 2021).

This analysis will utilize the secondary dataset "Online Store Customer Data" ([Data Link](#)) of the USA and 'US E-commerce records 2020" ([Data Link](#)) both retrieved from Kaggle. Both of them have similarities with missing fields hence, they are used in this analysis report to avoid ambiguity. The first dataset has some missing fields such as sales, profit, quantity, category etc. whereas the latter dataset has missing demographic information, for instance, marital status, age group, employment status. Thus, these two have been joined to retrieve a bird's eye view of the USA online platform. This dataset will be very conducive to finding out some crucial strategic insights about online business namely which group of people are making more purchases, how they prefer making payments, which location has fewer purchases and such things.

2.1. Data Cleansing and Preparation

Datasets found on the Internet are not cleansed have blank cells, superfluous attributes and so on usually. Hence, it has to go with some sort of optimization so that it gets fit to

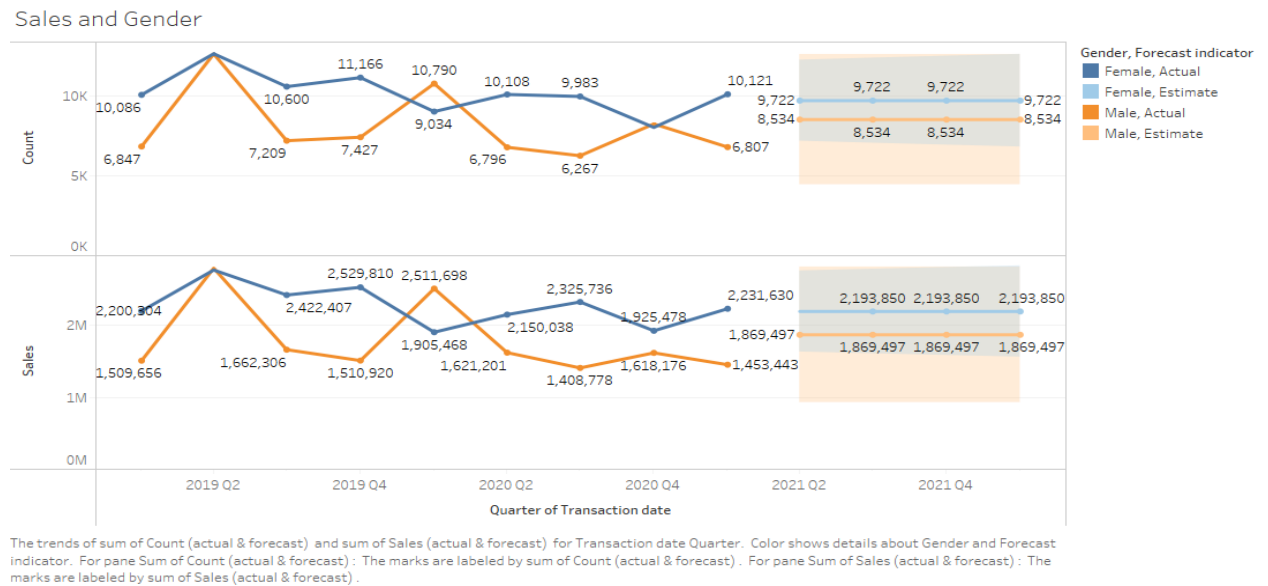
Tableau for bringing out better analytical insights about the consumers and e-commerce. Data cleansing—often challenging and extravagant—is a procedure of finding and taking off the inaccuracy from the data warehouse keeping clean and consistent data otherwise data mining and data warehousing will not produce the expected results (Deshmukh and Wangiukar, 2011). It is an expeditious way to find out errors in the big data despite being a time-taking process, per Ridzuan, et al. (2019). Moreover, Handel, et al. (2020) have established an algorithm for data cleansing purposes that rectify the generated errors and ameliorate the preservation of the data.

Before getting into Tableau, I have modified the dataset a bit despite containing lots of redundant information such as blank cells that produce null values in the BI tools. Some IDs such as ‘Transaction ID’, ‘Row ID’ have been removed from the dataset forasmuch as they do not provide any meaningful relationships and insights about the project. Moreover, no data has been deducted other than those blank cells owing to having a larger view of USA online consumers’ behaviour mostly based on demography, the amount spent on purchasing, and geographic location, quantity, sales and profits. Furthermore, a new column, ‘Count’ has been added with a value of 1 to each (**Appendix 8**), thereby this analysis will be able to identify the total number, for instance number of single and married, the portion of unemployment, number of payments methods—PayPal or card. Afterwards, a new cell has been added to the ‘Online Store Customer Data’ which is ‘Age Group’ (**Appendix 8**). If it goes by categorically, it will produce unambiguous insight such as how many teens, young, mid and olds as per this formula: **IF (C2<20,"Teen", IF (AND(C2>=20, C2<30), "Young", IF(AND(C2>=30,C2<45), "Mid", "Old")))**. Here C2 stands for a value of a cell means if the value is lower than 20, that customer will fall under the “Teen” group and such a thing will happen according to the formula.

Appendix 1 shows the procedure—these two datasets have common attributes called ‘State Name’ that has been used for full outer join purpose so that it is able to grasp a greater view about this online consumer market therein. Thus it will try to find lots of why what and how throughout the report and gaps, issues regarding sales declining as well.

2.2. Key metrics and relationships

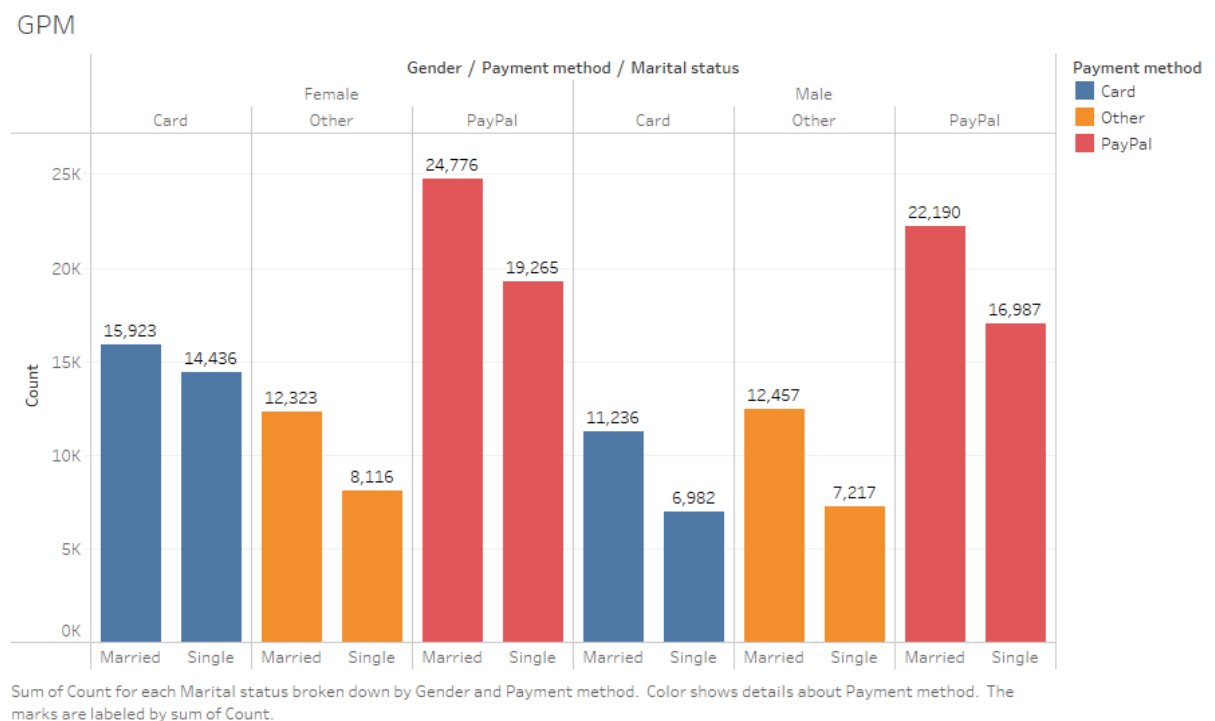
2.2.1 Sum of sales gender-wise



This line graph shows a parallel trend for gross sales and the number of consumers followed by their gender from 2019 to the first quarter of 2021 with anticipation of both features from Q2, 2021 to Q1, 2022. It is evident that the consumer number has a direct impact on the sales. Throughout the periods, the female has almost always been the graph leader for both features. On the contrary, in the commencement of 2020, there was a fall for women consumers and so the sales whereas dramatically male stood at their peak, above 2.5 million sales and approximately 11 thousand customers. It can be inferred that

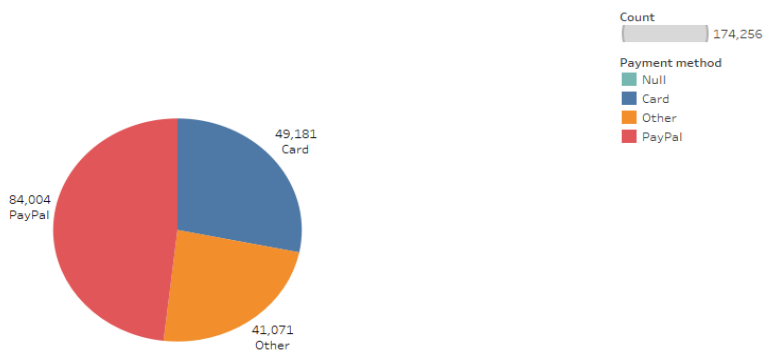
owing to the COVID-19, there was a fall in trend at the beginning of 2020 compared to the prior. By the end of the periods mentioned in the datasets, the trend was going up that perhaps because of the market stability started coming gradually. Surprisingly, the estimation for both cases portrays a steadiness up until quarter 1 of 2022. Money capacity has become curtailed to the consumer might be a reason.

2.2.2. Dominant Payment Methods and Consumers



The bar chart above illustrates the number of different payment method users followed by their marital status and sex. It seems that PayPal is dominant in this market for both cases where the others two such as card is slightly higher than the other yet PayPal is far ahead. The pie chart below will give us a complete numeric value for these three payment system.

Payment

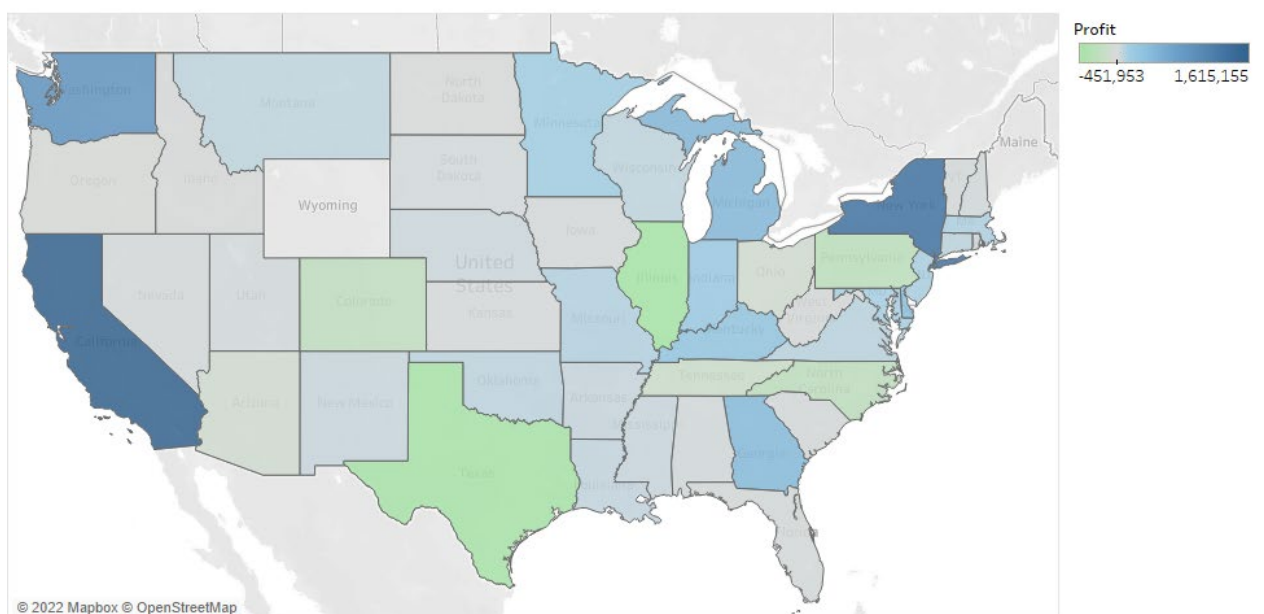


Sum of Count and Payment method. Color shows details about Payment method. Size shows sum of Count and Payment method. The marks are labeled by sum of Count and Payment method.

It can be presumed that PayPal might provide more opportunities and an easier transaction policy which have made them the most popular in the USA market.

2.2.3. Profit Earned in States

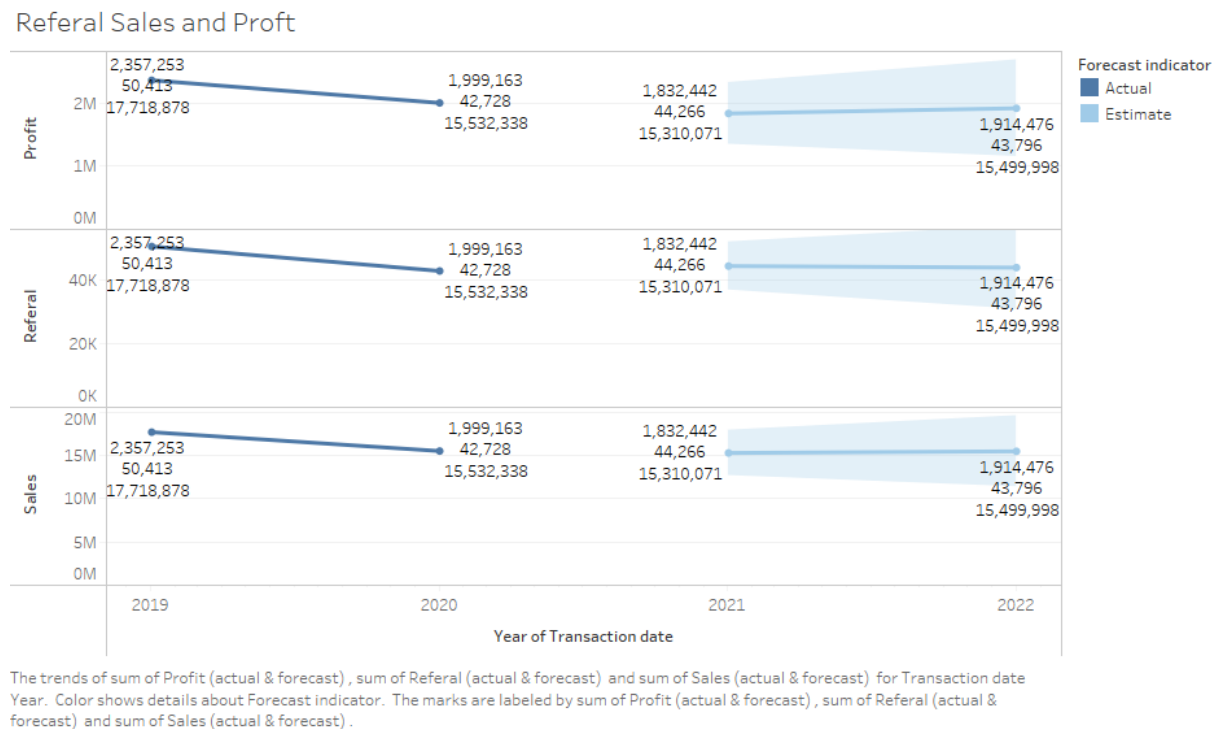
Profit



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Profit. Details are shown for State names. The view is filtered on State names, which excludes Null.

The pictorial view of the map shows that prime cities such as New York, California, and Washington of the USA have gone through profit coloured with blue while there are still some cities making a loss in this buzzing market. The reason might be that those aforementioned cities are the most populous therein hence, the number of consumers is high along with the quantity sold. Yet there are many WHYS behind those cities going through negative profit that will be brought to the light in the insights section.

2.2.4. The Relation amongst Referrals, Sales and Profit



This line graph proves that referral is a strong tool in this market. It is already mentioned that referral works better than actual advertisements. Here it is evident that if referral goes down then sales and profit both will deteriorate their portion. The referral was about 50

thousand in 2019. By the end of 2020, it had a decrease of almost 10 thousand which lead to a loss of more than 35 thousand within these timeframes. The reason behind a decrease in the referral might be products quality, failing to identify the appropriate targeted group for particular categorized products and failing to reach them, I think. Though COVID-19 is a fact yet there was a boom in the online shopping market during this pandemic since people choose to make purchases staying at the home.

2.2.5. Age Group VS. Segments

Age and Segment

Age Group	Segment					Count
	Basic	Gold	Missing	Platinum	Silver	
Old	40,496	6,971	11,076	14,351	17,560	866 40,496
Mid	19,609	5,551	3,412	5,586	10,592	
Young	7,928	1,854	2,957	4,520	6,124	
Teen	8,144	1,176	866	2,839	2,644	

Sum of Count broken down by Segment vs. Age Group. Color shows sum of Count. The marks are labeled by sum of Count.

This table portrays the information about the age group and their chosen service segment. The people, aged 45 or above have covered the lion's share of the consumer total number in five different categories where mid-aged people stood at second and the teen group contributed the least. This dataset has 3 years of data starting from 2019 therefore the period falls under the pandemic. As a result, it is assumed that the older group was in the most vulnerable situation thus they could not come out for making purchases. Consequently, they hold the dominant number in this market taking all the categories into account.

2.3. Strategic Insights

Making a profit is the ultimate cornerstone for companies to sustain themselves in this free market. Thereby, this report has gone into a deep analysis of how big cities are making big bucks from the process such as New York, Washington, and California (**Appendix 2**). These cities demonstrate a dependency amongst the features—**Profit, Sales and Referrals** which are higher correspondent to each city respectively. Thus it can be deduced that companies' referral campaign therein was successfully produced a better result than actual advertisements. Moreover, these are all densely populated cities where understanding consumer behaviour is a must topic to be covered by companies. Assuming that companies have done their analysis on consumer behaviour to grab this big market in the big cities.

On the other hand, there might have several reasons behind the loss despite the high sales of some cities (**Appendix 3**). Probably, lower referrals might be the ground behind the lower repeat customer that is reducing dividends. Perhaps the most significant reason behind this downfall is the **Cost of Production**. As we all know

$$\text{Profit} = \text{Revenue} - \text{Cost of Production}$$

Since their profit is negative, their cost of production is much higher than the revenue in those aforementioned areas along with probably maintenance cost. Overall, their COP, maintenance cost and other fringe costs might be the probable reasons behind their downward trend of profit.

In the last year, sales were dramatically lower compared to a previous couple of years. As per **Appendix 4**, online business was least affected by the pandemic perhaps in the first two years while the last year was a nightmare for the online platform. The buying capacity of the consumers shrank and consequently, the last year's sales stood at about 14%, 30% and 25 % less than the years 2019 and 2020 respectively.

Approximately 24% customers are not using PayPal or Card. Assuming cash is the option opted by those customers (**Appendix 7**). Perhaps, the procedure is easier to them compared to others. In such case, Companies should educate them about the process where they can make any alternative easier card systems to reduce the procedural time because cash calculation takes time. Thus this problem can be solved.

Finding out the target audience is another hook for getting sustainability in every market. As datasets exhibit that people aged 45 and onwards, old contributed around 50% while teen stood at a mere 10% amongst other three groups (**Appendix 5**). The first probable reason behind this huge number might be these aged people were at higher risk and didn't prefer going out during pandemic. Secondly, the savings and buying capacity is higher for them compared to others. Their contribution to the chart thereby is astronomical. Meanwhile, rather than cutting off the fewer contributors-- "Teen" and "Young", companies might utilise these groups using data analytics on pricing predictability and products as their spending capacity and tastes. Moreover, around 90% amount spent on making the purchase was from the somehow employed group (**Appendix 6**). Since it is the dominant group, marketers should focus their products according to the taste of these groups namely old and mid and their status and bringing variations—adding a new colour, including, excluding that will help to retain the customers.

2.4.Conclusion

To infer, after going through the analysis process, Tableau has found out some issues regarding Sales declining in some states and profit as well and some probable reasons, grounds behind them where it has suggested some possible solution to the issues. Overall, it has tried to touch on relevant fields regarding online business in the USA.

2.5. Appendices

Appendix 1

The screenshot shows a data tool interface with a sidebar on the left containing 'Connections' and 'Sheets'. The main workspace displays a join operation titled 'online_store_customer_data+ (Multiple Connections)'. A dialog box titled 'Join' is open, showing a Venn diagram with two overlapping circles. The 'Left' circle is selected, and the join type is set to 'Inner'. The 'Data Source' is 'US E-commerce records 2020.csv', and the 'State names' field is being joined with the 'State' field. Below the dialog, a table preview shows the resulting data with columns: Category, Product Name, Sales, Quantity, Discount, and Profit. The table has 2500 rows.

Category	Product Name	Sales	Quantity	Discount	Profit
ishings	Linden 10" Round Wall Cl...	48.90	4	0.200000	8.56
ishings	Linden 10" Round Wall Cl...	48.90	4	0.200000	8.56
ishings	Linden 10" Round Wall Cl...	48.90	4	0.200000	8.56
ishings	Linden 10" Round Wall Cl...	48.90	4	0.200000	8.56

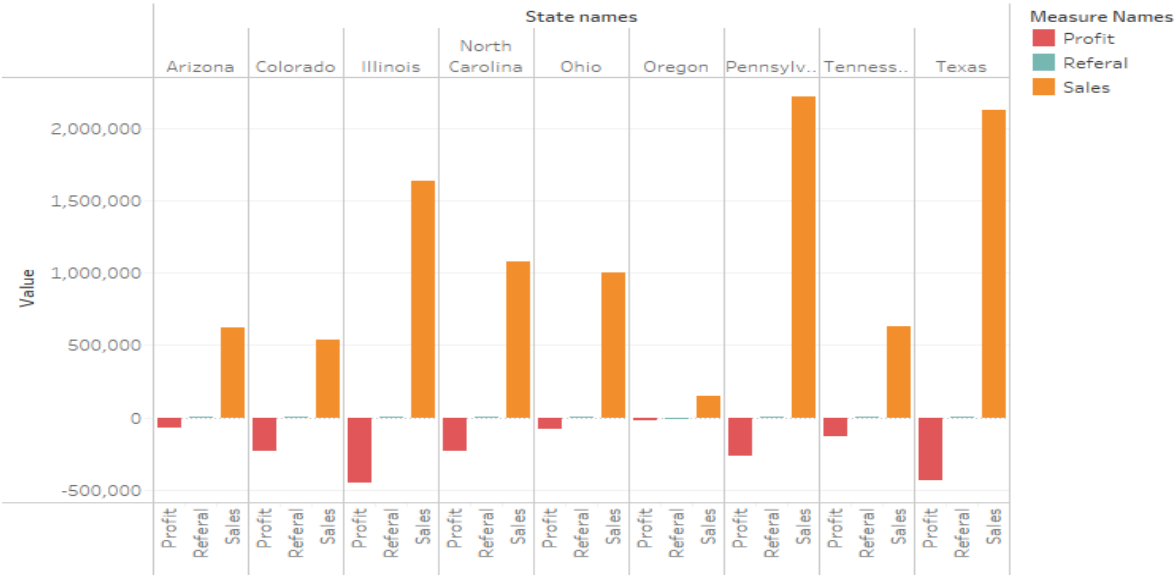
Appendix 2

Sheet 11

State names	Profit	Referral	Sales
California	1,615,155	25,194	8,051,359
New York	1,339,639	10,912	5,165,765
Washington	897,353	7,095	3,408,075
Michigan	441,364	1,876	1,343,350
Georgia	412,671	2,852	1,226,241
Delaware	338,979	627	770,279
Kentucky	280,352	1,428	916,556

Appendix 3

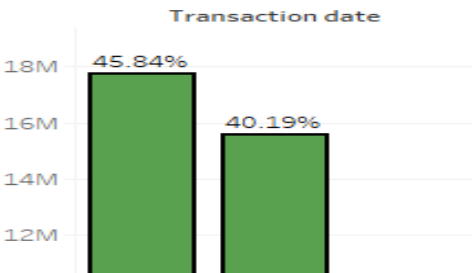
Negative Profit



Profit, Referral and Sales for each State names. Color shows details about Profit, Referral and Sales. The view is filtered on State names, which keeps 9 of 51 members.

Appendix 4

Sales Per Year



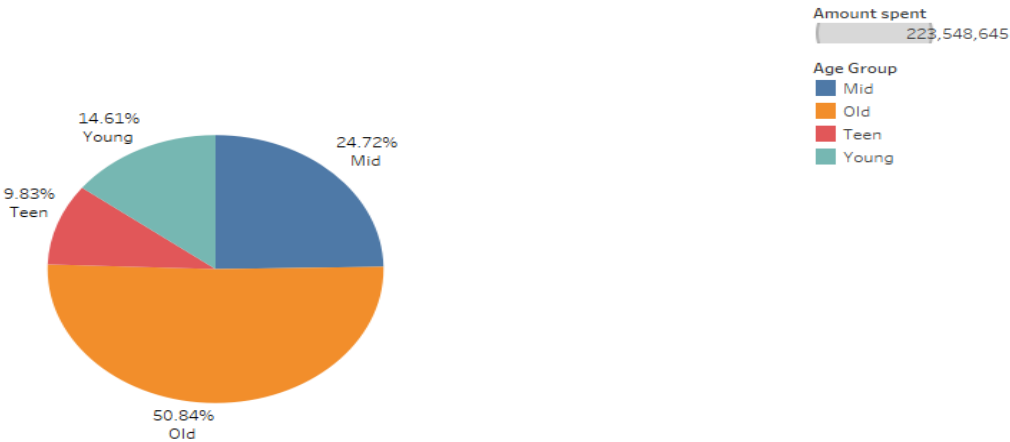
Appendix 6

Sheet 12

Age Group	Employees status				% of Total Amount spent	
	Employee..	self-emp..	Unemplo..	workers		
Old	46.02%	14.06%	10.76%	28.79%	4.54%	46.02%
Mid	42.05%	19.06%	9.28%	29.08%		
Young	33.41%	24.98%	4.54%	36.45%		
Teen	40.23%	22.68%	7.20%	29.81%		

Appendix 5

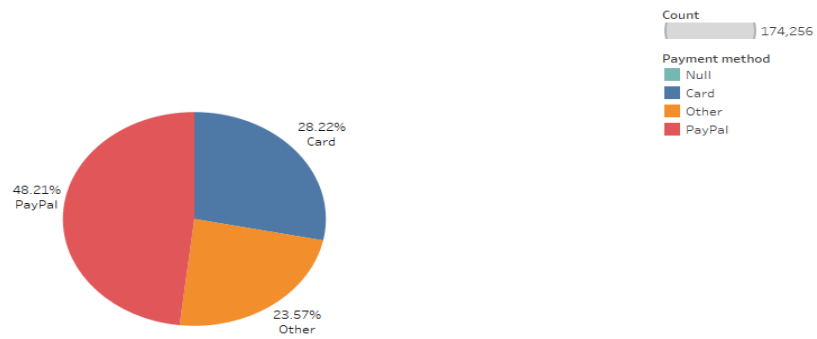
Sheet 12



% of Total Amount spent and Age Group. Color shows details about Age Group. Size shows sum of Amount spent. The marks are labeled by % of Total Amount spent and Age Group.

Appendix 7

Payment



% of Total Count and Payment method. Color shows details about Payment method. Size shows sum of Count. The marks are labeled by % of Total Count and Payment method.

Appendix 8

F	G	H	I	J	K	L	M
Segment	Employees_status	Payment_method	Referral	Amount_spent	Count	Age Group	
Basic	Unemployment	Other	1	2051.36	1	Teen	
Basic	self-employed	Card	0	544.04	1	Old	
Basic	workers	PayPal	1	1572.6	1	Old	
Platinum	workers	Card	1	1199.79	1	Teen	
Basic	self-employed	Card	0		1	Young	
Basic	Employees	PayPal	1	2922.66	1	Old	
Platinum	Employees	PayPal	1	1481.42	1	Mid	
Basic	workers	PayPal	1	1149.55	1	Mid	
Silver	Employees	Card	0	1046.2	1	Old	
Gold	Unemployment	Card	1	2730.6	1	Mid	
Basic	Employees	PayPal	0	1712.82	1	Old	
Basic	workers	Other	1	154.31	1	Old	
Basic	self-employed	Card	1	819.08	1	Old	
Basic	Employees	PayPal	1	1719.83	1	Old	
Platinum	self-employed	Other	1	954.12	1	Young	
Platinum	workers	Other	1	1005.92	1	Teen	
Silver	self-employed	PayPal	1	2882.77	1	Old	
Platinum	workers	Card	0	2999.98	1	Old	
Basic	self-employed	Other	0	1902.73	1	Mid	
Basic	Employees	Card	1	2968.95	1	Old	

References

1. Hansen, H. L. (2019). *In God we trust. All others must bring data.* IBM. <https://www.ibm.com/blogs/nordic-msp/in-god-we-trust-all-others-must-bring-data/>

2. Chai, W., Labbe, M., & Stedman, C. (2021). *Big Data Analytics*. TechTarget. <https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics>
3. Pierpoint. (n.d.). The Growing Role of Data Analytics in Recruiting the Perfect Candidate. *PierPoint*. <https://pierpoint.com/blog/data-analytics/>
4. Iqbal, M. (2022). *Netflix Revenue and Usage Statistics (2022)*. BusinessofApps. <https://www.businessofapps.com/data/netflix-statistics/>
5. Mixson, E. (2021). *Data Science at Netflix: How Advanced Data & Analytics Helps Netflix Generate Billions*. AI Data & Data Analytics Network. <https://tinyurl.com/5ehn32ch>
6. Shah, R. (n.d.). *How Big Data Analytics aids Media & Entertainment*. Phrazor. <https://tinyurl.com/2p87k7f4>
7. Emma. (2019, May 21). Big Data: Shaping the Future of the Media Advertising Industry. *Advendio*. <https://www.advendio.com/big-data-shaping-future-media-advertising-industry>
8. Joyce, J. (2021). *20 Awesome Sources of Free Data*. Search Engine Journal. <https://www.searchenginejournal.com/free-data-sources/302601/#close>
9. Deshmukh, R. R., & Wangikar, V. (2011). *Data Cleaning: Current Approaches and Issues*. https://www.researchgate.net/publication/278301609_Data_Cleaning_Current_Approaches_and_Issues
10. Woolley, C. S. C., Handel, I. G., Bronsvoort, B. M., Schoenebeck, J. J., & Clements, D. N. (2020). Is it time to stop sweeping data cleaning under the carpet? A novel algorithm for outlier management in growth data. *PloS ONE*, 15(1), <https://doi.org/10.1371/journal.pone.0228154>
11. Ridzuan, F., & Wan Zainon, Wan Mohd Nazmee. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, 161, 731-738. doi:10.1016/j.procs.2019.11.177
12. Hayden, J. (2018). *4 Qualities That Make Netflix a Good Model for Supply Chain Visibility*. Savi. <https://www.savi.com/supply-chain-visibility-netflix-model/>
13. Biddle, G. (2022). *Netflix's 2020 Product Strategy*. ProductLed. <https://productled.com/blog/netflixs-2020-product-strategy/>
14. Dixon, M. (2019). *How Netflix used big data and analytics to generate billions*. Selerity. <https://tinyurl.com/mrpwx5fp>

15. DataFlair , (2019). *Data Science at Netflix – A Must Read Case Study for Aspiring Data Scientists*. <https://data-flair.training/blogs/data-science-at-netflix/>
16. Lee, E. (2019, October 16). 'Stranger Things' Helps Netflix Increase Subscribers. *The New York Times*. <https://www.nytimes.com/2019/10/16/business/media/netflix-q3-earnings.html>