IST 687 INTRODUCTION TO DATA SCIENCE

DATA ANALYSIS FOR AMAZON

SEPTEMBER 14, 2023

SYRACUSE UNIVERSITY

Cheromaine Smith | Vic Millar | Thomas Lento | Jordan Epstein

Table of Contents

Description (Background)

If we were to compare the retail landscape of today to one of the past, we would be amazed. Over the last decade, the retail industry has experienced a massive overhaul. Amazon, a multinational technology company and a behemoth of online retail, has completely transformed the way we shop. From its humble beginnings as an online bookstore, to an auction site, to becoming a global powerhouse, Amazon's growth and influence are undeniable. Founded in July 1994, Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking (Amazon,2023).

Project Scope and Objective

For this project we'll explore Amazon sales data from 2020. We'll take a deep look at the numbers and trends to gain insights into the performance of Amazon's sales. By examining factors like product categories that drives sales, customer purchasing behavior and inventory pricing by categories. We believe that by analyzing this data, we can gain valuable insights into Amazon's success and understand how it has become a dominant force in the ecommerce industry. The objective is to be able to provide Amazon executives with a synopsis of the areas needing improvement in terms of customer satisfaction and what categories, if any, are leading to less satisfied customers. So, buckle up and get ready to dive into the world of Amazon.

Project Deliverables

To present our findings in a comprehensive manner, we have defined a set of deliverables. These deliverables will include, Data Munging, a detailed analysis report, amazing visualizations, and actionable recommendations. We believe, they will provide a holistic understanding of the sales data and offer practical strategies to enhance performance on the Amazon website. Let's get started by reviewing what to look forward to with each deliverable. Through:

- **Data Cleaning:** We plan to prepare the data for further analysis by removing NAs, missing information, unique characters and formatting invalid fields.
- **Comprehensive Analysis Report:** This report will provide a detailed overview of the analysis conducted on the Amazon Sales data. It will include key findings, insights and trends discovered during the analysis process.

- **Visualizations:** TO showcase engaging visual representations, such as charts, graphs, word clouds, models and tables created to present the sales data. These visualizations will help stakeholders easily interpret and understand the patterns and trends in the data.
- **Recommendations & Interpretations:** We will suggest actionable recommendations based on the interpretation of the analysis. These recommendations will focus on improving sales performance, optimizing pricing, and enhancing overall success on the Amazon platform.

Data Acquisition

To gather our dataset of at least 10 variables and 1000 observations, we searched multiple sites including FiveThirtyEight, Kaggle, GitHub and Datahub.io. In the end, we decided that we wanted to get our data from Kaggle as they had a wide range of data available, and there was also reviews as to whether a dataset was helpful and effective to analyze. The next requirement for our data was whatever dataset we chose we wanted to take a consulting agency point-of-view. Where should that company be focusing? What are the complaints? What do they do well? Which category/department/team make them the most money? Which ones are they losing more than their making? Which category correlates with another? What should be marketed together etc. These requirements narrowed down our dataset to a few different files: the NFL yearly data by player, Walmart Sales Data, Cost Data for US Airlines, Trends and Insights of Global Tourism, Amazon Sales Data, European Soccer and Airbnb Prices in European Cities.

The issue with majority of the data we reviewed is that they did not meet our first requirement of 10 variables and 1000 observations. For example, the cost data for US airlines dataset, Walmart dataset and the trends and Insights of Global Tourism had around 4, 8 and 7 variables respectively, which means we had to throw those out of our pool of choices. Next up was the review of all our sports datasets: NFL yearly data and European Soccer, while they did meet the basic requirements, they were not really business oriented our review would be more focused on player improvement than company improvement, so those got thrown out as well. Finally, was the review of all the sales type data, Amazon Sales Data and Airbnb Prices in European Cities. The Airbnb Prices dataset was not chosen because it required combining 20 different files and had multiple variables with true or false values instead of numerical ones. In the end, we decided that the amazon dataset was the only one that met our requirements; it was from Kaggle, had at least 10 variables with 1000 observations, was business focused and not overtaken by true/false fields.

Link to Amazon Dataset:

https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset?resource=download

Preview of Dataset:

product_id	product_name	category	discounted_price	actual_price	discount_percenta	rating	rating_count	about_product	user_id	user_name	review_id	review_title	review_content
B07JW9H4J1	Wayona Nylon Bra	Computers&Accessories|Ac	â‚¹399	â‚¹1,099	64%	4.2	24,269	High Compatibility	AG3D6O4STAQKAY2	Manav,Adarsh gup	R3HXWT0LRP0NMF	Satisfied,Charging	Looks durable Char
B098NS6PVG	Ambrane Unbreaka	Computers&Accessories|Ac	â‚¹199	â‚¹349	43%	4	43,994	Compatible with al	AECPFYFQVRUWC3I	ArdKn,Nirbhay kum	RGIQEG07R9HS2,R1	A Good Braided Cal	I ordered this cable
B096MSW6CT	Sounce Fast Phone	Computers&Accessories|Ac	â‚¹199	â‚¹1,899	90%	3.9	7,928	ã€ Fast Charger& D	AGU3BBQ2V2DDAN	Kunal,Himanshu,vi	R3J3EQQ9TZI5ZJ,R3	Good speed for ear	Not quite durable a
B08HDJ86NZ	boAt Deuce USB 30	Computers&Accessories|Ac	â‚¹329	â‚¹699	53%	4.2	94,363	The boAt Deuce USE	AEWAZDZZJLQUYVC	Omkar dhale,JD,HE	R3EEUZKKK9J36I,R3	Good product,Good	Good product,long
B08CF3B7N1	Portronics Konnect	Computers&Accessories|Ac	â‚¹154	â‚¹399	61%	4.2	16,905	[CHARGE & SYNC FU	AE3Q6KSUK5P75D5	rahuls6099,Swasat	R1BP4L2HH9TFUP,F	As good as original	Bought this instead
B08Y1TFSP6	pTron Solero TB301	Computers&Accessories|Ac	â‚¹149	â‚¹1,000	85%	3.9	24,871	Fast Charging & Da	AEQ2YMXSZWEOHK	Jayesh,Rajesh k.,So	R7S8ANNSDPR40,R3	It's pretty good,Ave	It's a good product
B08WRWPM22	boAt Micro USB 55	Computers&Accessories|Ac	â‚¹176.63	â‚¹499	65%	4.1	15,188	It Ensures High Spe	AG7C6DAADCTRQJG	Vivek kumar,Amazc	R8E73K2KWJRDS,R5	Long durable.,good	Build quality is goo
B08DDRGWTJ	MI Usb Type-C Cabl	Computers&Accessories|Ac	â‚¹229	â‚¹299	23%	4.3	30,411	1m long Type-C USE	AHW6E5LQ2BDYOI	Pavan A H,Jayesh b	R2X090D1YHACKR,I	Worth for money -	Worth for money -
B008IFXQFU	TP-Link USB WiFi A	Computers&Accessories|Ne	â‚¹499	â‚¹999	50%	4.2	1,79,691	USB WiFi Adapter â	AGV3IEFANZCKECFG	Azhar JuMan,Aniru	R1LW6NWSVTVZ2H	Works on linux for	I use this to connec
B082LZGK39	Ambrane Unbreaka	Computers&Accessories|Ac	â‚¹199	â‚¹299	33%	4	43,994	Universal Compatil	AECPFYFQVRUWC3I	ArdKn,Nirbhay kum	RGIQEG07R9HS2,R1	A Good Braided Cal	I ordered this cable
B08CF3D7QR	Portronics Konnect	Computers&Accessories|Ac	â‚¹154	â‚¹339	55%	4.3	13,391	[CHARGE & SYNC FU	AGYLPKPZHVYKKZH	Tanya,Anu,Akshay,V	R11MQS7WD9C3IC	Good for fast charg	The cable is efficie
B0789LZTCJ	boAt Rugged v3 Ext	Computers&Accessories|Ac	â‚¹299	â‚¹799	63%	4.2	94,363	The boAt rugged ca	AEWAZDZZJLQUYVC	Omkar dhale,JD,HE	R3EEUZKKK9J36I,R3	Good product,Good	Good product,long
B07KSMBL2H	AmazonBasics Flex	Electronics|HomeTheater,T	â‚¹219	â‚¹700	69%	4.4	4,26,973	Flexible, lightweigh	AEYJ5I6JZZPOJB6M	Rishav Gossain,Sha	R1FKOKZ3HHKJBZ,F	It's quite good and	I am using it for 14
B085DTN6R2	Portronics Konnect	Computers&Accessories|Ac	â‚¹350	â‚¹899	61%	4.2	2,262	[20W PD FAST CHAF	AGUAYQHARAKR2V2	Priya,Mansi,Plabai	R1QETDIPRCX4S0,R	Works,Nice Produc	Definitely isnâ€™t
B09KLVMZ3B	Portronics Konnect	Computers&Accessories|Ac	â‚¹159	â‚¹399	60%	4.1	4,768	[CHARGE & SYNC FU	AF2XXVO7JUBUVAC	Deepaak Singh,siva	R20XIOU25HEX80,F	Great but,Worked v	Loosing charging ca
B083342NKJ	MI Braided USB Typ	Computers&Accessories|Ac	â‚¹349	â‚¹399	13%	4.4	18,757	1M Long Cable. Usb	AGSGSRTEZBQY64W	Birendra ku Dash,A	R2JPQNKCOE10UK,	Good product,usin	I like it ðŸ'ðŸ',Best
B0B6F7LX4C	MI 80 cm (32 inche	Electronics|HomeTheater,T	â‚¹13,999	â‚¹24,999	44%	4.2	32,840	Note : The brands, I	AHEVOQADJSSRX7D	Manoj maddheshiy	R13UTIA6KOF6QV,F	It is the best tv if yc	Pros- xiomi 5a is b
B082LSVT4B	Ambrane Unbreaka	Computers&Accessories|Ac	â‚¹249	â‚¹399	38%	4	43,994	Compatible with al	AECPFYFQVRUWC3I	ArdKn,Nirbhay kum	RGIQEG07R9HS2,R1	A Good Braided Cal	I ordered this cable
B08WRBG3XW	boAt Type C A325 T	Computers&Accessories|Ac	â‚¹199	â‚¹499	60%	4.1	13,045	Type C A 325 Cable	AFB5KJR4Q5FICAHE	Rohan Narkar,JAGV	R2BP8Y5OJXKJLF,R2	Good for charging a	Check for offera be
B08DPLCM6T	LG 80 cm (32 inche	Electronics|HomeTheater,T	â‚¹13,490	â‚¹21,990	39%	4.3	11,976	Resolution: HD Rea	AHBNKB74LGTYUOK	NIRMAL.N,Manoj ku	R2PNR69G0BQG2F,	Sound quality,Very	LG was always Goo
B09C6HXFC1	Duracell USB Lightr	Computers&Accessories|Ac	â‚¹970	â‚¹1,799	46%	4.5	815	Supports Ios Device	AFNYIBWKJLJQKY4E	Prasannavijayarag	R12D1BZF9MU8TN,	Good cable for car,	I trust this product
B085194JFL	tizum HDMI to VGA	Electronics|HomeTheater,T	â‚¹279	â‚¹499	44%	3.7	10,962	Superior Stability: I	AEO5FHWNOSFBT5	aditya d.,Paranthai	R1GYK05NN6747O,	Good product ; Ave	This connector has
B09F6S8BT6	Samsung 80 cm (32	Electronics|HomeTheater,T	â‚¹13,490	â‚¹22,900	41%	4.3	16,299	Resolution: HD Rea	AHEVO4Q5NM4YXN	Rahman Ali,MARIYA	R1SN0D4DFBKAZI,R	Good,Sound is very	Overall good.,TV pi
B09NHVCHS9	Flix Micro Usb Cab	Computers&Accessories|Ac	â‚¹59	â‚¹199	70%	4	9,378	Micro usb cable is	AHIKJUDTVJ4T6DV6	$@\|\\|TO$|-|,Sethu	R3F4T5TRYPTMIG,R	Worked on iPhone	Worked on iPhone
B0B1YVCJ2Y	Acer 80 cm (32 incl	Electronics|HomeTheater,T	â‚¹11,499	â‚¹19,990	42%	4.3	4,703	Resolution : HD Rea	AFSMISGEYDYIP3Z4	Ayush,ROHIT A.,Kec	R1EBS3566VCSCG,F	Wonderful TV and A	About the TV - Won
B01M4GGIVU	Tizum High Speed H	Electronics|HomeTheater,T	â‚¹199	â‚¹699	72%	4.2	12,153	Latest Standard HD	AGVUE2NFN2MQEC	Yashpreet Singh,Ab	R2DIHMHOPYEASB,	Cheap product and	The signal is too ur
B08B42LWKN	OnePlus 80 cm (32	Electronics|HomeTheater,T	â‚¹14,999	â‚¹19,999	25%	4.2	34,899	Resolution: HD Rea	AFUT7ANZTZYGLXU6	ATHARVA BONDRE,S	R3COVVOP2R7Z28,	Worthy and most a	This OnePlus TV is
B094JNXNPV	Ambrane Unbreaka	Computers&Accessories|Ac	â‚¹299	â‚¹399	25%	4	2,766	Blazing Charging - A	AFYR53OTBUX2RNA	Anand sarma,lokes	R249YCZVKYR5XD,F	Ok cable,three pin	The product seems
B09W5XR9RT	Duracell USB C To L	Computers&Accessories|Ac	â‚¹970	â‚¹1,999	51%	4.4	184	1.2M Tangle Free d	AHZWJCVEIEI76H2V	Amazon Customer,C	R1Y30KU04V3QF4,I	Very good product.	Fast charging.,Cabl
B077Z65HSD	boAt A400 USB Typ	Computers&Accessories|Ac	â‚¹299	â‚¹999	70%	4.3	20,850	2 meter special rev	AFA332YHUPB6I7KI	GHOST,Amazon Cus	R1G4I5FLAHM16P,F	Just buy it dont eve	One amazing cable
B00NH11PEY	AmazonBasics USB	Computers&Accessories|Ac	â‚¹199	â‚¹750	73%	4.5	74,976	One 9.8-foot-long (.	AGBX233C7B7D7YZ	Pravin Kumar,Maes	R1C8MVU3EIX56Y,F	Nice,good,Paisa va	Sufficient length,ex
B09CMM3VGK	Ambrane 60W / 3A	Computers&Accessories|Ac	â‚¹179	â‚¹499	64%	4	1,934	Stay ahead and nev	AGHYCMV7RJ5D76U	Rishabh,Amazon Cu	R223OIZPTZ994S,R	Good product,Good	The cable build qua
B08QSC1XY8	Zoul USB C 60W Fa	Computers&Accessories|Ac	â‚¹389	â‚¹1,099	65%	4.3	974	{3A/QC 3.0 FAST CH	AHMKXORT3VNMB1	Pratyush Pahuja,Th	R2SOAYWUV349HP	Great Cable, Chargi	Not charging as fas
B008FWZGSG	Samsung Original T	Computers&Accessories|Ac	â‚¹599	â‚¹599	0%	4.3	355	USB Type-C to Type-	AEQWVGESA7TDGK	Verified Buyer,Avis	R2Z9ENI1BK4EAB,R	Good,Genuine proc	Buy it,Received in g
B0B4HJNPV4	pTron Solero T351 .	Computers&Accessories|Ac	â‚¹199	â‚¹999	80%	3.9	1,075	Universal Compatil	AF477BP57JM7Z4J	Placeholder,à¤¶à¥	R1Q323BB35OP30,	The metal pin is los	It's a good data cat
B08Y1SJVV5	pTron Solero MB30	Computers&Accessories|Ac	â‚¹99	â‚¹666.66	85%	3.9	24,871	Fast Charging & Da	AEQ2YMXSZWEOHK	Jayesh,Rajesh k.,So	R7S8ANNSDPR40,R3	It's pretty good,Ave	It's a good product
B07XLCFSSN	Amazonbasics Nylc	Computers&Accessories|Ac	â‚¹899	â‚¹1,900	53%	4.4	13,552	Fast Charge: When	AF2IRSQZKMBGX44	Wraith,Krishna Eng	R213ILI3XNVHQ0,R	Good,Worth to buy	Good budget mfi ce
B09RZS1NQT	Sounce 65W OnePl	Computers&Accessories|Ac	â‚¹199	â‚¹999	80%	4	576	[USB C To USB C Co	AHUH7OYN3LAUATI	Anmol,Vani,Tejas J	RW294SCHB5QTK,F	Worth it!,Good one	It does the job real
B0B3MMYHYW	OnePlus 126 cm (5	Electronics|HomeTheater,T	â‚¹32,999	â‚¹45,999	28%	4.2	7,298	Resolution: 4K Ultr	AGDOVGWZKEQ3M	Abhishek Kumar,Ar	R2J3Q3BUHJ2S7E,R	Decent product. Va	I am posting this af
B09C6HWG18	Duracell Type C To	Computers&Accessories|Ac	â‚¹970	â‚¹1,999	51%	4.2	462	Up To 10,000+ Benc	AHRUMHBZ7IAQPLI	Koushal K Jain,Mat	R32JZC43P990BL,R3	Product is as expec	Same type is availa
B00NH11KIK	AmazonBasics USB	Computers&Accessories|Ac	â‚¹209	â‚¹695	70%	4.5	1,07,687	One 6-foot-long (1.	AEYHTCWWZYU3JQ	Shiva,Uzef,kottala s	R2AE3BN2Y58N55,F	Functionality as de	Using it and satisfa
B09JPC82QC	Mi 108 cm (43 inch	Electronics|HomeTheater,T	â‚¹19,999	â‚¹34,999	43%	4.3	27,151	Resolution : Full HD	AHB43CZ4RHLJ5S6	Sameer Patil,Techn	R1VOXBV87EI37W,	DETAILED REVIEW a	NOTE:@ If you sele
B07JW1Y6XV	Wayona Nylon Bra	Computers&Accessories|Ac	â‚¹399	â‚¹1,099	64%	4.2	24,269	[High Compatibility	AG3D6O4STAQKAY2	Manav,Adarsh gup	R3HXWT0LRP0NMF	Satisfied,Charging	Looks durable Char
B07KRCW6LZ	TP-Link Nano AC60	Computers&Accessories|Ne	â‚¹999	â‚¹1,599	38%	4.3	12,093	High Speed WiFi â€	AEM356PVXFHAXW	Paul Joe,Simon Rex	RSNHWPVLK9SAQ,F	Dual Bandwidth,It's	Easy to use,It's goo
B09NJN8L25	FLiX (Beetel USB to	Computers&Accessories|Ac	â‚¹59	â‚¹199	70%	4	9,378	Micro USB charging	AHIKJUDTVJ4T6DV6	$@\|\\|TO$|-|,Sethu	R3F4T5TRYPTMIG,R	Worked on iPhone	Worked on iPhone
B07XJYYH7L	Wecool Nylon Brai	Computers&Accessories|Ac	â‚¹333	â‚¹999	67%	3.3	9,792	Special Features Of	AE47XF2766XJOEO	Amazon Customer,I	RWSHFGBE1WU3I,F	Its slow in charging	Charging power is v

Data Preprocessing (Munging/Cleaning)

Before any data munging, the Amazon dataset consisted of 16 variables and 1465 observations. The 16 variables were Product ID, Product Name, Category, Discounted Price, Actual Price, Discount Percentage, Rating, Rating Count, About Product, UserID, Username, ReviewID, Review Title, Review Content, Image Link and Product Link. After we library in the necessary packages and read in the csv file, we did a large-scale view of our dataset.

During the first quick glance at the dataset, we realized that all the currency information was in Indian Rupees. As we did not know how to fix currency issues, we decided to move on to cleaning the rest of the dataset first. We used the sum function to find any NA's or missing information in our datasets, which R advised there was at least 3. Further investigation into the variables determined that the NA's were in the rating and rating count fields since there were very few NA values. We decided to just remove the 3 observations, as those 3 observations in theory should not affect a dataset of over 1400.

After researching currencies in R and monetary symbols, the decision was made to clean up the currency information by removing the rupee symbol and converting the currency amount to USD. This process did not work being combined into one, so we decided to do each part in steps. We removed the currency symbol first, then removed any commas from the prices, converted the prices and rating field to numeric then created a function that converted Indian Rupee to United States Dollars. To ensure this function would work, we created a test formula to try it before using it in our dataset. With the test being successful, we were able to create two new variables with prices in USD: the actual price and discounted price. We were then left with 18 variables and 1462 observations.

After completing the preliminary cleaning of this dataset, we were very intrigued by the category column. There were so many different combinations of products and items for sales. Each row had a category, with multiple subcategories on top of subcategories, all separated by the "|" symbol. For example, one cell would be "Computer & Accessories| Accessories & Peripherals| Cables & Accessories| Cables| USB Cables". The first thing we did was create subsets of the data by the main categories; "Car & Motorbikes", "Computer & Accessories", "Electronics" , "Health & Personal Care", "Home & Kitchen", "Home Improvement"," Musical Instruments", "Office Products" and "Toys & Games". Then we separated the subset datasets category column into multiple different columns based on the subcategories. We were left 0 observations in Cars & Motorbikes, 451 observations in Computer & Accessories, 526 observations in Electronics, 1 observation in Health & Personal Care, 447 observations in Home & Kitchen, 2 observations in Home Improvement, 2 observations in Musical Instruments, 31 observations in Office Products, 1 observations in the Toys & Games subset.

<u>Data Summary before Munging:</u>

```
> summary(AmazonData)
  product_id         product_name         category         discounted_price    actual_price      
 Length:1465        Length:1465        Length:1465        Length:1465        Length:1465       
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  



 discount_percentage     rating       rating_count     about_product         user_id         
 Length:1465         Min.   :2.000   Min.   :     2   Length:1465        Length:1465       
 Class :character    1st Qu.:4.000   1st Qu.:  1186   Class :character   Class :character  
 Mode  :character    Median :4.100   Median :  5179   Mode  :character   Mode  :character  
                     Mean   :4.097   Mean   : 18296                                        
                     3rd Qu.:4.300   3rd Qu.: 17336                                        
                     Max.   :5.000   Max.   :426973                                        
                     NA's   :1       NA's   :2                                             
  user_name          review_id         review_title       review_content        img_link        
 Length:1465        Length:1465        Length:1465        Length:1465        Length:1465       
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  



 product_link      
 Length:1465       
 Class :character  
 Mode  :character  
```

Data Summary after Munging:

```
  product_id         product_name         category         discounted_price  actual_price     discount_percentage
 Length:1462        Length:1462        Length:1462        Min.   :   39    Min.   :    39    Length:1462        
 Class :character   Class :character   Class :character   1st Qu.:  325    1st Qu.:   800    Class :character   
 Mode  :character   Mode  :character   Mode  :character   Median :  799    Median :  1670    Mode  :character   
                                                          Mean   : 3130    Mean   :  5453                       
                                                          3rd Qu.: 1999    3rd Qu.:  4321                       
                                                          Max.   :77990    Max.   :139900                       
     rating       rating_count     about_product         user_id           user_name          review_id        
 Min.   :2.000   Min.   :     2   Length:1462        Length:1462        Length:1462        Length:1462       
 1st Qu.:4.000   1st Qu.:  1192   Class :character   Class :character   Class :character   Class :character  
 Median :4.100   Median :  5179   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :4.097   Mean   : 18307                                                                              
 3rd Qu.:4.300   3rd Qu.: 17342                                                                              
 Max.   :5.000   Max.   :426973                                                                              
 review_title       review_content       img_link         product_link       discounted_price_usd actual_price_usd
 Length:1462        Length:1462        Length:1462        Length:1462        Min.   :  0.47       Min.   :   0.47 
 Class :character   Class :character   Class :character   Class :character   1st Qu.:  3.90       1st Qu.:   9.60 
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :  9.59       Median :  20.04 
                                                                             Mean   : 37.56       Mean   :  65.44 
                                                                             3rd Qu.: 23.99       3rd Qu.:  51.85 
                                                                             Max.   :935.88       Max.   :1678.80 
```

R Code:

```
setwd("/Users/victormillar/downloads")
AmazonData <- read_csv("amazon.csv")

sum(is.na(AmazonData))

glimpse(AmazonData)
```

```
na_counts <- colSums(is.na(AmazonData))
na_counts

AmazonData <- na.omit(AmazonData)
na_counts <- colSums(is.na(AmazonData))
na_counts

# 3 rows removed, no more NAs
# Need to remove Indian Rupee currency symbol before converting to
# Numeric
AmazonData$discounted_price <- substring(AmazonData$discounted_price, 2)
AmazonData$actual_price <- substring(AmazonData$actual_price, 2)
# symbols are removed, will try numeric again now
# now need to remove the commas from the prices
AmazonData$actual_price <- as.numeric(gsub("[^0-9.]", "", AmazonData$actual_price))
AmazonData$discounted_price <- as.numeric(gsub("[^0-9.]", "", AmazonData$discounted_price))
# prices are now numeric and have no comma
AmazonData$rating <- as.numeric(AmazonData$rating)
# Rating is now also numeric

AmazonData2 <- AmazonData
# Creating a Save Point so I don't mess up previous work

# I am going to write a function that converts Indian Rupee to USD
convert_usd <- function(rupee_price) {
exchange_rate <- 0.012 # Replace this if exchange rate changes
usd_price <- rupee_price * exchange_rate
return(usd_price)
}
# Will now test the function to confirm it works
test_data1 <- data.frame(rupee_test = c(100, 200, 300, 400))
test_data1$USD_price <- convert_usd(test_data1$rupee_test)
test_data1 # The test was successful

AmazonData2 <- AmazonData2 %>%
mutate(discounted_price_usd = round(convert_usd(discounted_price),
2), actual_price_usd = round(convert_usd(actual_price), 2))
# view(AmazonData2) Now have two new columns with prices in USD

AmazonData3 <- AmazonData2
# Creating another Save Point
# I want to separate out the Category column into 7 separate columns,
# one for each tier of the category.
categories <- strsplit(AmazonData3$category, "\\|")
AmazonData3$Cat1 <- sapply(categories, `[`, 1)
AmazonData3$Cat2 <- sapply(categories, `[`, 2)
AmazonData3$Cat3 <- sapply(categories, `[`, 3)
AmazonData3$Cat4 <- sapply(categories, `[`, 4)
AmazonData3$Cat5 <- sapply(categories, `[`, 5)
AmazonData3$Cat6 <- sapply(categories, `[`, 6)
AmazonData3$Cat7 <- sapply(categories, `[`, 7)
AmazonData3 <- subset(AmazonData3, select = -category) #delete original column
view(AmazonData3)

AmazonData4 <- AmazonData3 # New Save Point
```

 # Now that the data is cleaned up, I can start working through the
 # business questions.

Descriptive & Inferential Statistics

It was now time to start analyzing our data beginning with descriptive statistics. Descriptive statistics provides us with a snapshot of the main characteristics of the dataset. By calculating measures like the mean, median and standard deviation of any given dataset, we can gain insights into the central tendency, spread and distribution of the data. These statistics will help us understand the overall performance and patterns within the Amazon sales data. Let's begin to explore the various types of descriptive statistics performed in this project, starting with:

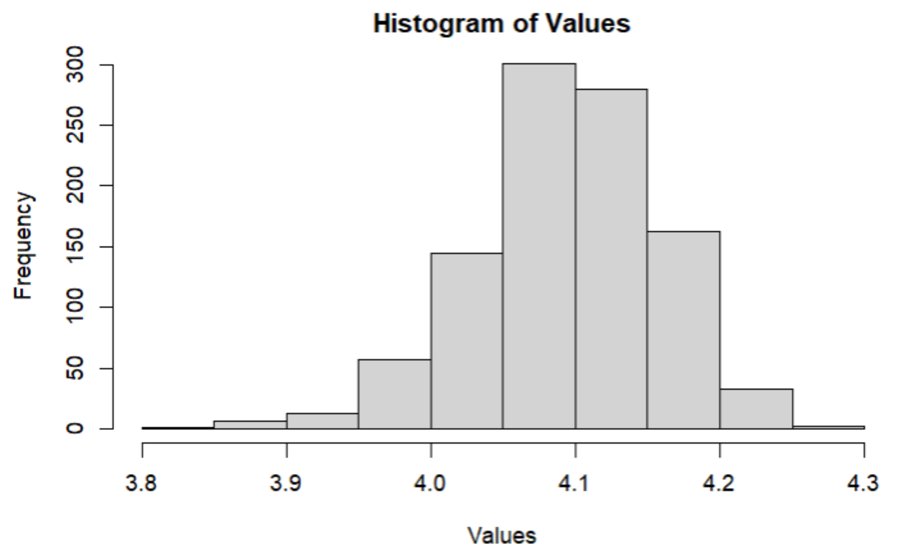
<u>Discount & Actual Price</u>

```
 discounted_price_usd actual_price_usd
 Min.   :  0.47       Min.   :   0.47
 1st Qu.:  3.90       1st Qu.:   9.60
 Median :  9.59       Median :  20.04
 Mean   : 37.56       Mean   :  65.44
 3rd Qu.: 23.99       3rd Qu.:  51.85
 Max.   :935.88       Max.   :1678.80
```

Now that we have dived into the numbers, lets dig in a little deeper. With a deeper dive, we can uncover insights that go beyond just the numbers and make meaningful inferences about the larger population. From understanding customer behavior and preferences to predicting future trends, inferential statistics can help us make data-driven decisions and take our Amazon sales strategy to the next level. Below we will review a sampling we did on the Amazon data. As you can see in the histogram, while the frequency shape looks like a bell curve (normal distribution), this data shows some skewness. So, having more information, would help us to smooth out the discrepancies.

<u>Sampling & Replication</u>

```
Values<-replicate(1000,mean(sample(AmazonData3$rating,22,TRUE)))
hist(Values)
```



Modeling Techniques

While investigating this unique dataset, we decided to complete a few different models to help gather an accurate representation of this dataset. The first modeling we introduced to this dataset was linear regression, both simple and multiple. In the simple linear regression, we were able to summarize and study the relationship between variables in our data set like Price, Category, Rating etc. Below is our investigating into the models:

Simple Linear Regression

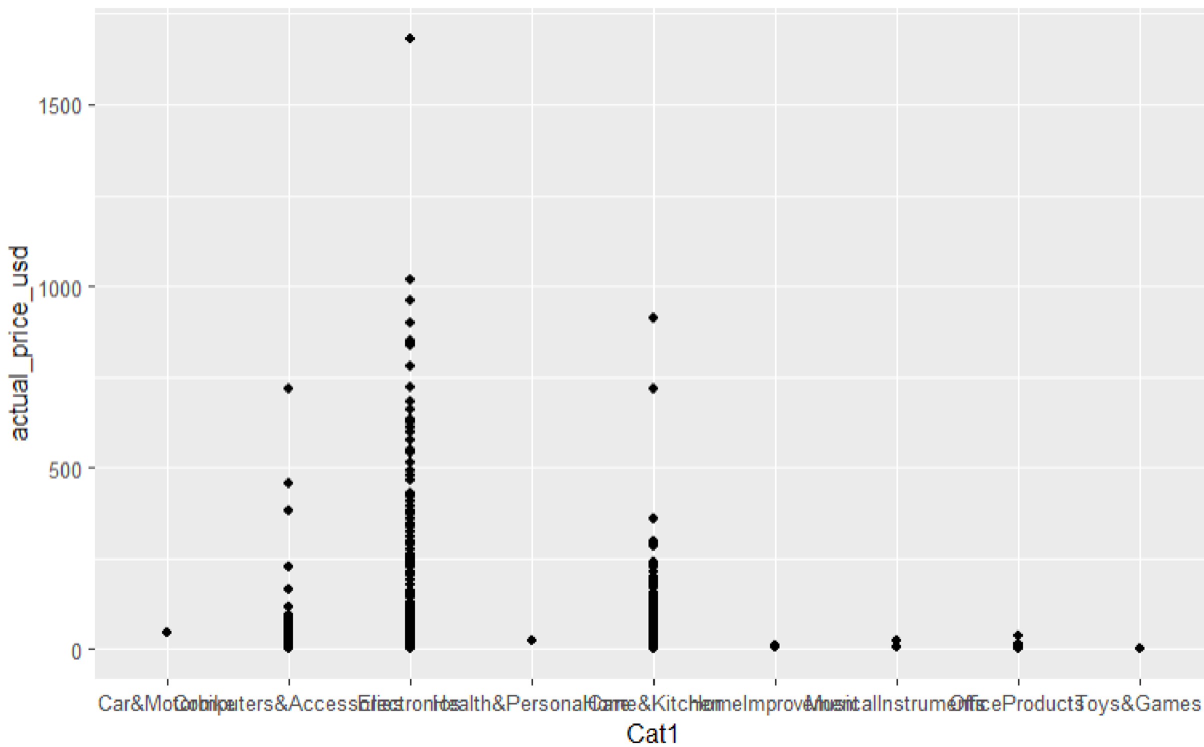
To see if it was possible to obtain a line that best fits the data and if any of these variables had a relationship. Upon looking at these variables, we went through the 3 steps to determine significance; checked the P-Value of the F-Statistic, then the R^2 value, and finally the P-Value of the Coefficient.

In out Simple Linear Model between Price & Category, right away we were able to interpret the equation as the P-value of the F-Statistic was statistically significant. However, when it came to the P-value of the coefficient, none of our variables were statistically significant and we could not interpret. When it came to the linear Model between Price & Rating, we were able to find a line of best fit. The Linear Model below also shows that the P-Value was significant. With further analysis, we were able to determine that there is a positive relationship between price and rating. This makes sense in the business term because higher priced items are in higher demand leading to an increase in price. Where the opposite could also be true, lower reviews, lead to no demand for an item, leading to the seller to drop their prices.

Price & Category

```
ggplot(AmazonData4,aes(x=Cat1,y=actual_price_usd))+geom_point()+stat_smooth(method="lm", col="red")
Ama_Cat= lm(actual_price_usd~Cat1,AmazonData4)
summary(Ama_Cat)
```

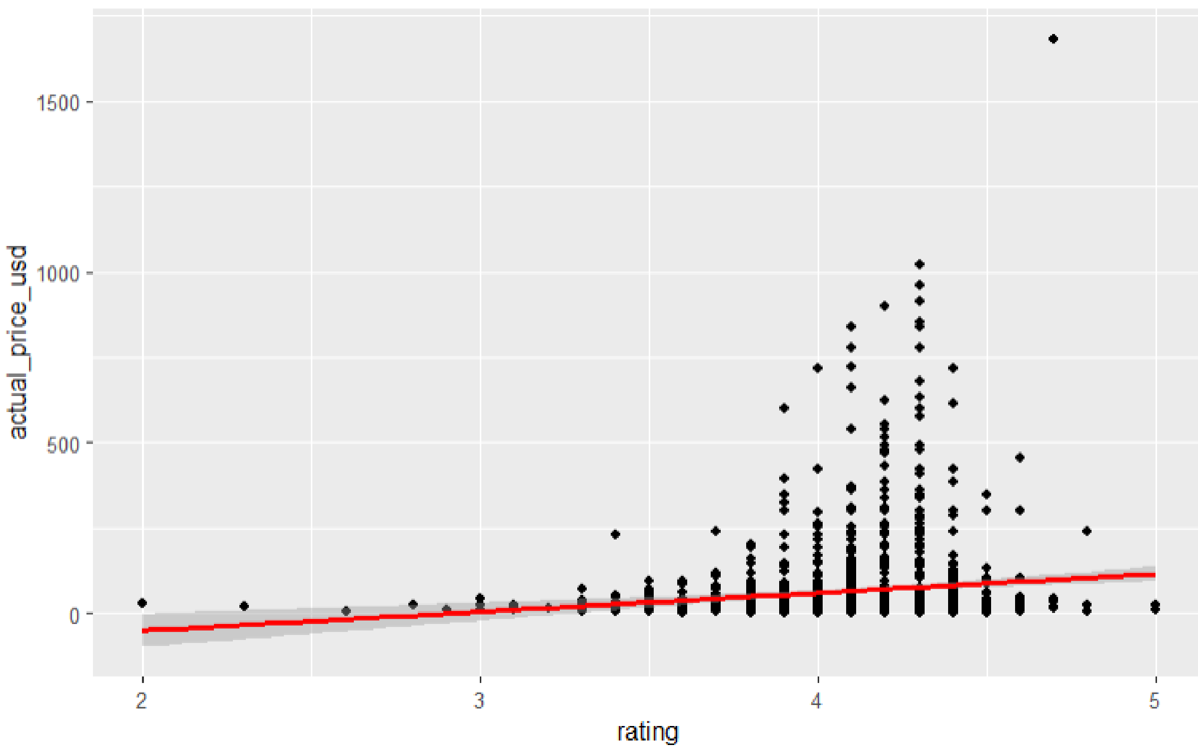
```
Call:
lm(formula = actual_price_usd ~ Cat1, data = AmazonData4)

Residuals:
    Min      1Q  Median      3Q     Max 
-119.48  -44.00  -14.25    1.79 1557.27 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  48.00     123.36   0.389    0.697
Cat1Computers&Accessories   -27.76     123.49  -0.225    0.822
Cat1Electronics              73.53     123.47   0.596    0.552
Cat1Health&PersonalCare     -25.20     174.45  -0.144    0.885
Cat1Home&Kitchen              1.99     123.50   0.016    0.987
Cat1HomeImprovement         -38.41     151.08  -0.254    0.799
Cat1MusicalInstruments      -31.84     151.08  -0.211    0.833
Cat1OfficeProducts          -43.23     125.33  -0.345    0.730
Cat1Toys&Games              -46.20     174.45  -0.265    0.791

Residual standard error: 123.4 on 1453 degrees of freedom
Multiple R-squared:  0.1129,	Adjusted R-squared:  0.108 
F-statistic: 23.12 on 8 and 1453 DF,  p-value: < 2.2e-16
```



Price & Rating

```
ggplot(AmazonData4,aes(x=rating,y=actual_price_usd))+geom_point()+stat_smooth(method="lm", col="red")
Ama_Rat= lm(actual_price_usd~rating,AmazonData4)
summary(Ama_Rat)
```

```
Call:
lm(formula = actual_price_usd ~ rating, data = AmazonData4)

Residuals:
    Min      1Q  Median      3Q     Max 
-103.35  -58.43  -42.58   -1.65 1580.03 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -160.92      48.13  -3.344 0.000848 ***
rating         55.25      11.72   4.715 2.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.7 on 1460 degrees of freedom
Multiple R-squared:  0.015,	Adjusted R-squared:  0.01432 
F-statistic: 22.23 on 1 and 1460 DF,  p-value: 2.649e-06
```



Multiple Linear Regression

The other modeling technique used in our analysis is a multiple linear regression model. We wanted to see the result of three different categories on our prices. The three variables we decided to investigate was discounted usd price, rating, and rating counts. Per the below chart, we can see that while the equation is significant. Rating and Rating Count in relation to actual price cannot be interpreted. Finally, the relationship between actual price and discount price is very minimal. So, there are not many assumptions or insights that can be made from the multiple linear regression.

```
Call:
lm(formula = actual_price_usd ~ discounted_price_usd + rating +
    rating_count, data = AmazonData4)

Residuals:
    Min      1Q  Median      3Q     Max
-288.67   -9.59   -5.09    1.39  448.44

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.757e+00  1.334e+01  -0.282    0.778
discounted_price_usd  1.505e+00  1.129e-02 133.235   <2e-16 ***
rating                3.242e+00  3.269e+00   0.992    0.321
rating_count         -3.274e-05  2.198e-05  -1.490    0.137
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.71 on 1458 degrees of freedom
Multiple R-squared:  0.9254,    Adjusted R-squared:  0.9253
F-statistic:  6031 on 3 and 1458 DF,  p-value: < 2.2e-16
```

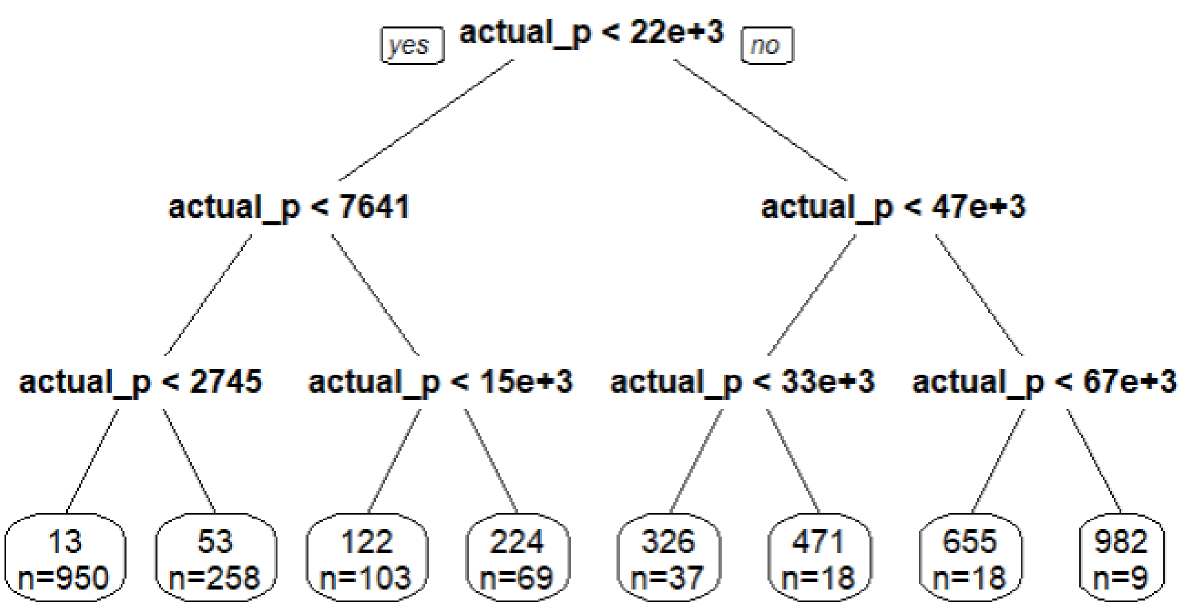
Support Vector Machine

Next, lets explore how Support Vector Machines(SVM) can be a powerful tool for analyzing Amazon sales data. SVM is a machine learning algorithm that can help us classify and predict various aspects of sales performance. By training the SVM model on historical sales data, we can uncover patterns and trends that can assist in predicting future sales, identify key factors that influence sales and even separating customers based on their purchasing behaviors. With SVM, we can harness the power of machine learning to gain valuable insights and optimize our sales strategy.

Below we used SVM modeling techniques, to predict the actual usd price of many products by using a few significant variables from our models. Per the code, you will see that we divided our dataset into training and testing so that we could check and validate our results. We needed to keep in mind that seeing as we had NULL values and some categories had more observation than others, we needed to be mindful with these predictions. While our findings could be an accurate representation of the population, there was also an increasing likelihood that we should not be overtly confident. With all this in mind, these were our findings:

```r
library(tidyverse)
library(caret)
library(rpart)
library(rpart.plot)
library(kernlab)
AmazonData5<-data.frame(discounted_price=AmazonData3$discounted_price,
                        actual_price=AmazonData3$actual_price,
                        rating=AmazonData3$rating,
                        rating_count=AmazonData3$rating_count,
                        discount_price_usd=AmazonData3$discounted_price_usd,
                        actual_price_usd=AmazonData3$actual_price_usd)
cartTree<-rpart(actual_price_usd~.,data=AmazonData5)
prp(cartTree, extra=1)
t<-varImp(cartTree)
trainList<-createDataPartition(y=AmazonData5$actual_price_usd, p=.60, list = FALSE)
training<-AmazonData5[trainList,]
testing<-AmazonData5[-trainList,]
model.rpart<-train(rating~., data=training, method="rpart",
                   preProc=c("center","scale"))
model.rpart
t<-varImp(cartTree)
t%>%arrange(desc(Overall))%>%slice(1:5)
```

	Overall <dbl>
actual_price	4.8406624
discount_price_usd	3.7262290
discounted_price	3.7262290
rating_count	0.3816159
rating	0.1507718



```
CART

878 samples
  5 predictor

Pre-processing: centered (5), scaled (5)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 878, 878, 878, 878, 878, 878, ...
Resampling results across tuning parameters:

  cp          RMSE       Rsquared     MAE
  0.01230660  0.3134381  0.029970315  0.2266818
  0.02295032  0.3087114  0.020138992  0.2241195
  0.03545476  0.3095838  0.009191025  0.2244098

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.02295032.
```

Business Questions

The business questions that have been recognized and answered through the projects are as follows:

1. Which top level category brought in the most revenue for Amazon?

R Code:

```
# Group the data by the Cat1 column and calculate the total revenue
# for each category
category_revenue <- AmazonData4 %>%
    mutate(total_revenue = discounted_price_usd * rating_count) %>%
    group_by(Cat1) %>%
    summarize(total_revenue = sum(total_revenue, na.rm = TRUE))

# Sort results to find top category
top_category <- category_revenue %>%
    arrange(desc(total_revenue)) %>%
    head(1)

top_category
```

Answer:

```
## # A tibble: 1 x 2
##   Cat1         total_revenue
##   <chr>                <dbl>
## 1 Electronics     710185543.
```

```
# Based on a rough estimate using discounted price x rating count, we
# believe teh top selling category in this dataset is Electronics.
# This is based on the assumption that everyone who gave a product a
# rating also purchased a product, but it is only useful for trends
# and not an exact revenue count, since there will be customers who
# bought a product but did not give it a rating.
```

2. What is the average rating by broad category (Tier 1, ex. Electronics)?

R Code:

```
average_rating_by_cat <- AmazonData4 %>%
    group_by(Cat1) %>%
    summarize(average_rating = mean(rating)) %>%
    arrange(desc(average_rating))

average_rating_by_cat
```

Answer:

```
## # A tibble: 9 x 2
##   Cat1                    average_rating
##   <chr>                            <dbl>
## 1 OfficeProducts                    4.31
## 2 Toys&Games                        4.3
## 3 HomeImprovement                   4.25
## 4 Computers&Accessories             4.16
## 5 Electronics                       4.08
## 6 Home&Kitchen                      4.04
## 7 Health&PersonalCare               4
## 8 MusicalInstruments                3.9
## 9 Car&Motorbike                     3.8
```

```
# Office Products has the highest average rating with 4.31/5 stars.
# Car & Motorbike has the worst average rating with 3.8
```

a) What is the average rating by price?

R Code:

```
price_bins <- c(0, 25, 50, 100, Inf) # Creating price bins

AmazonData4 <- AmazonData4 %>%
  mutate(price_group = cut(discounted_price_usd, breaks = price_bins, labels = c("0-25","25-50","50-100","100+")))

# New column 'price_group' now exists

average_rating_by_price <- AmazonData4 %>%
  group_by(price_group) %>%
  summarize(average_rating = mean(rating)) %>%
  arrange(desc(average_rating))

average_rating_by_price
view(AmazonData4)
# Rating seems to go up as the price goes up
cor(AmazonData4[,c("discounted_price_usd","rating")], use = "complete")
# there is a positive, but weak, correlation between discount_price_usd and rating
```

```
## # A tibble: 4 x 2
##   price_group average_rating
##   <fct>                <dbl>
## 1 100+                  4.18
## 2 50-100                4.16
## 3 0-25                  4.08
## 4 25-50                 4.08
```

```
view(AmazonData4)
# Rating seems to go up as the price goes up
cor(AmazonData4[, c("discounted_price_usd", "rating")], use = "complete")
```

```
##                      discounted_price_usd    rating
## discounted_price_usd            1.0000000 0.1211309
## rating                          0.1211309 1.0000000
```

Answer:

```
# there is a positive, but weak, correlation between
# discount_price_usd and rating
```

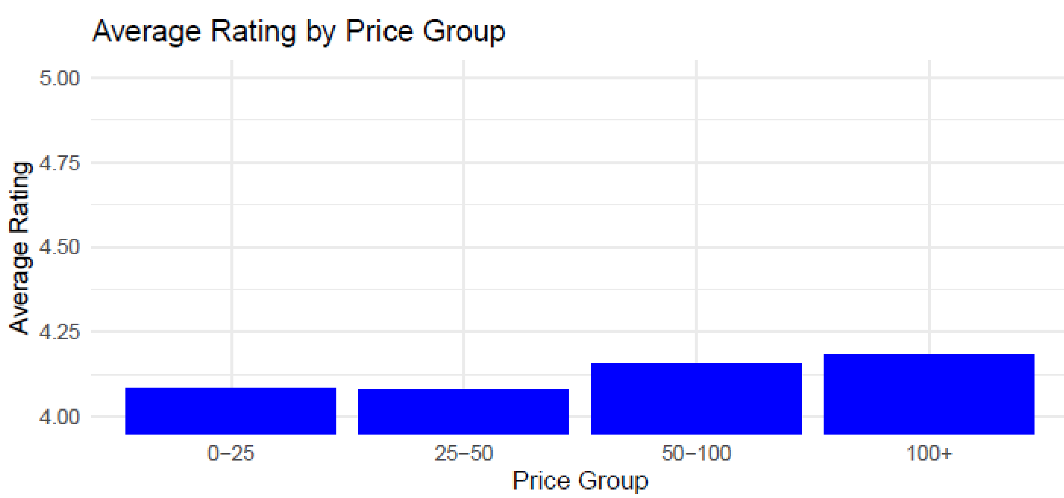
b) Can both be visualized in the same graph?

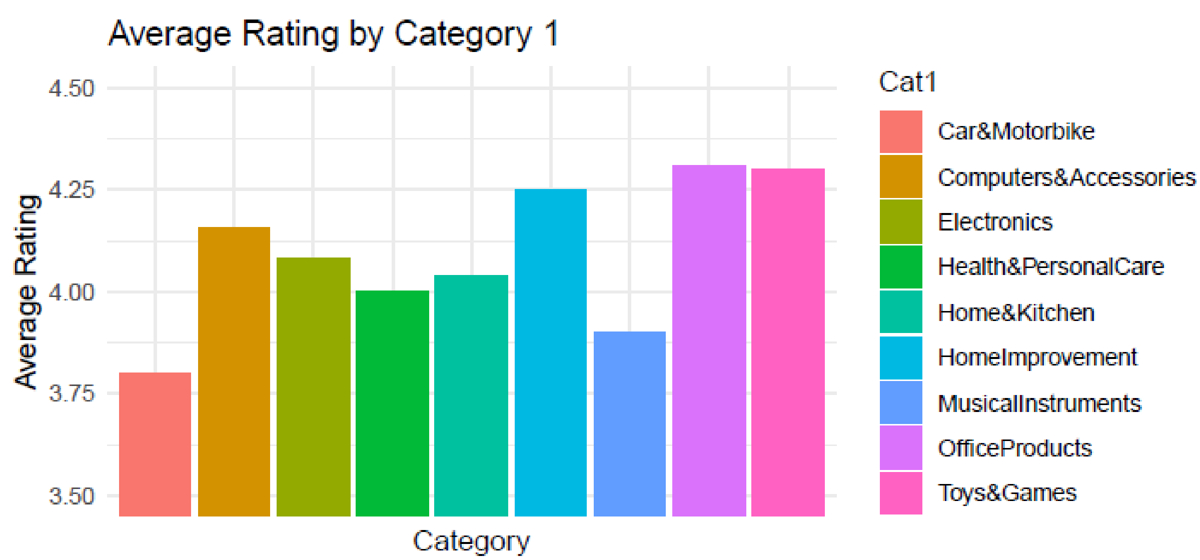
R Code:

```
library(ggplot2)

# Create a bar chart for Average Rating by Price Group
ggplot(average_rating_by_price, aes(x = price_group, y = average_rating)) +
    geom_bar(stat = "identity", fill = "blue") + labs(title = "Average Rating by Price Group",
    x = "Price Group", y = "Average Rating") + theme_minimal() + coord_cartesian(ylim = c(4,
    5))
```

```
# Create a bar chart for Average Rating by Category 1
ggplot(average_rating_by_cat, aes(x = Cat1, y = average_rating, fill = Cat1)) +
    geom_bar(stat = "identity") + labs(title = "Average Rating by Category 1",
    x = "Category", y = "Average Rating") + theme_minimal() + coord_cartesian(ylim = c(3.5,
    4.5)) + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```

Answer:





3. What are the top 5 selling products (based on number of ratings)?

R Code:

```
top_rated_products <- AmazonData4 %>%
    arrange(desc(rating_count)) %>%
    head(5) %>%
    select(product_name, rating_count)

top_rated_products
```

Answer:

```
## # A tibble: 5 x 2
##   product_name                                                   rating_count
##   <chr>                                                                 <dbl>
## 1 AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps~       426973
## 2 Amazon Basics High-Speed HDMI Cable, 6 Feet - Supports Ethernet,~       426973
## 3 Amazon Basics High-Speed HDMI Cable, 6 Feet (2-Pack),Black              426973
## 4 AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps~       426972
## 5 boAt Bassheads 100 in Ear Wired Earphones with Mic(Taffy Pink)          363713
```

```
# The 4 products with highest rating counts are all HDMI cables
# Number 5 is a set of wired headphones
```

a) What are the top 5 products based on revenue (rating count x discounted price)?

How many products overlap of each set of 5?

R Code:

```
top_revenue_products <- AmazonData4 %>%
    mutate(revenue = rating_count * discounted_price_usd) %>%
    arrange(desc(revenue)) %>%
```

```
    head(5) %>%
    select(product_name, revenue)

top_revenue_products
```

Answer:

```
## # A tibble: 5 x 2
##   product_name                                                           revenue
##   <chr>                                                                    <dbl>
## 1 Redmi 9 Activ (Carbon Black, 4GB RAM, 64GB Storage) | Octa-core Helio~  3.20e7
## 2 Redmi 9A Sport (Coral Green, 3GB RAM, 32GB Storage) | 2GHz Octa-core ~  2.82e7
## 3 Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Storage) | 2GHz Octa-core ~  2.45e7
## 4 Redmi 9A Sport (Carbon Black, 2GB RAM, 32GB Storage) | 2GHz Octa-core~  2.45e7
## 5 Redmi 126 cm (50 inches) 4K Ultra HD Android Smart LED TV X50 | L50M6~  1.79e7
```

```
# The products with the most revenue are cellphones and tvs There is
# NO OVERLAP between these two groups of 5 products.
```

b) What is the top selling product in each category?

R Code:

```
categories <- AmazonData4 %>%
    select("Cat1") %>%
    distinct()

conflicts_prefer(dplyr::filter)
```

```
top_revenue_products_by_cat <- AmazonData4 %>%
    filter(Cat1 %in% categories$Cat1) %>%
    mutate(revenue = rating_count * discounted_price_usd) %>%
    group_by(Cat1) %>%
    arrange(desc(revenue)) %>%
    top_n(1, wt = revenue) %>%
    ungroup() %>%
    select(product_name, revenue, Cat1)

view(top_revenue_products_by_cat)
# These are the top selling products in each category
```

Answer:

	product_name	revenue	Cat1
1	Redmi 9 Activ (Carbon Black, 4GB RAM, 64GB Storage) | Oct...	32008133.64	Electronics
2	SanDisk 1TB Extreme Portable SSD 1050MB/s R, 1000MB/s ...	5161088.66	Computers&Accessories
3	Aquaguard Aura RO+UV+UF+Taste Adjuster(MTDS) with Ac...	2151439.94	Home&Kitchen
4	Boya ByM1 Auxiliary Omnidirectional Lavalier Condenser Mi...	657801.12	MusicalInstruments
5	Casio FX-991ES Plus-2nd Edition Scientific Calculator, Black	89510.40	OfficeProducts
6	Dr Trust Electronic Kitchen Digital Scale Weighing Machine (...	39523.77	Health&PersonalCare
7	Reffair AX30 [MAX] Portable Air Purifier for Car, Home & Off...	31382.26	Car&Motorbike
8	Faber-Castell Connector Pen Set - Pack of 25 (Assorted)	28560.60	Toys&Games
9	Gizga Essentials Cable Organiser, Cord Management System...	17895.15	HomeImprovement

4. What is the average price discount by Tier 1 category?

R Code:

```
# Remove the % symbol and convert 'discount_percentage' to numeric
AmazonData4$discount_percentage <- as.numeric(sub("%", "", AmazonData4$discount_percentage))

# Calculate the average price discount by category
average_discount_by_category <- AmazonData4 %>%
```

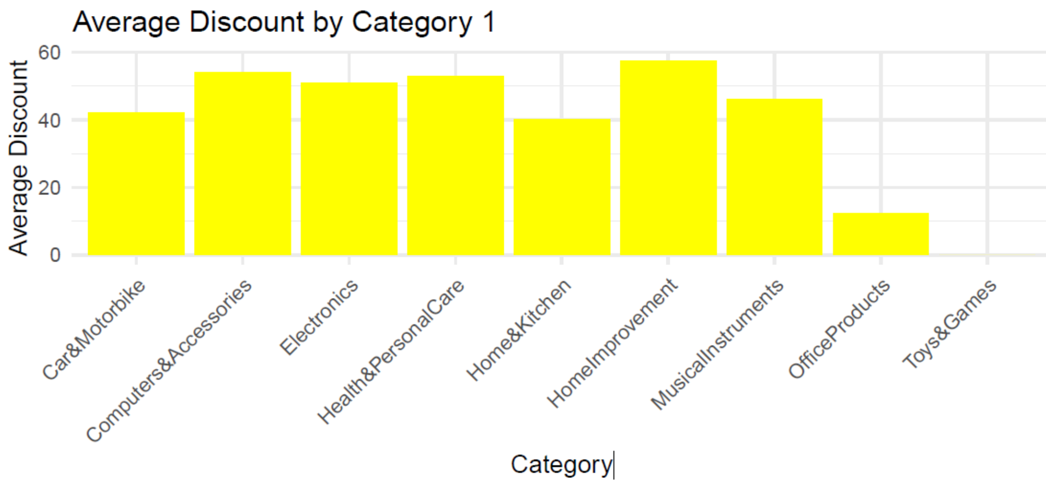
```
    group_by(Cat1) %>%
    summarize(average_discount_percentage = mean(discount_percentage, na.rm = TRUE)) %>%
    arrange(desc(average_discount_percentage))

average_discount_by_category
```

Answer:

```
## # A tibble: 9 x 2
##   Cat1                   average_discount_percentage
##   <chr>                                        <dbl>
## 1 HomeImprovement                               57.5
## 2 Computers&Accessories                         53.9
## 3 Health&PersonalCare                           53
## 4 Electronics                                   50.8
## 5 MusicalInstruments                            46
## 6 Car&Motorbike                                 42
## 7 Home&Kitchen                                  40.2
## 8 OfficeProducts                                12.4
## 9 Toys&Games                                     0
```

```
# Home Improvement has the highest discount percentage with 57.5%
```

```
# Q 4.2 Create a bar chart for Average Discount % by Category 1
ggplot(average_discount_by_category, aes(x = Cat1, y = average_discount_percentage)) +
    geom_bar(stat = "identity", fill = "yellow") + labs(title = "Average Discount by Category 1",
    x = "Category", y = "Average Discount") + theme_minimal() + theme(axis.text.x = element_text(angle =
    hjust = 1))
```



5. Key Word Analysis: What words appear the most frequently in the About, Review Title and User Reviews section? Visualize this sentiment analysis.

R Code:

```
AmazonData4$doc_id <- 1:nrow(AmazonData4)   #Adding unique id for every row

review_title_corpus <- corpus(AmazonData4$review_title, docnames = AmazonData4$doc_id)
review_title_dfm <- dfm(review_title_corpus, remove_punct = TRUE, remove = stopwords("english"),
    )
textplot_wordcloud(review_title_dfm, min_count = 2)
```

```
review_content_corpus <- corpus(AmazonData4$review_content, docnames = AmazonData4$doc_id)
review_content_dfm <- dfm(review_content_corpus, remove_punct = TRUE, remove = stopwords("english"),
    )
textplot_wordcloud(review_content_dfm, min_count = 3)
```

Answer:

Q 5.1: Review Title Wordcloud

```
# Wordcloud based on Review Title, words used at least twice Top
# words: good product, nice, quality, money, price
```

Q 5.2: Review Content Wordcloud

```
# Wordcloud based on Review Content, words used at least 3 times
# Larger word cloud, but similar top words: good, product, quality,
# price, easy, phone, batter
```

6. Are we able to accurately predict the user rating based on key words and price discount percentage?

R Code:

```
wordCloudFromDataFrame <- function(df_, max_words = 50) {
    df1_ <- df_[df_$review_content != "", ]
```

```
    review_content <- as.vector(df1_$review_content)
    review_content <- iconv(review_content, from = "UTF-8", to = "UTF-8",
        sub = "")
    words.vec <- VectorSource(review_content)
    words.corpus <- Corpus(words.vec)
    words.corpus <- tm_map(words.corpus, removePunctuation)
    words.corpus <- tm_map(words.corpus, removeNumbers)
    words.corpus <- tm_map(words.corpus, content_transformer(tolower))
    conflicts_prefer(tm::stopwords)
    words.corpus <- tm_map(words.corpus, removeWords, stopwords("english"))

    tdm <- TermDocumentMatrix(words.corpus)
    tdm

    m <- as.matrix(tdm)

    wordCounts <- rowSums(m)

    totalWords <- sum(wordCounts)
    totalWords
    words <- names(wordCounts)
    head(words)

    wordCounts <- sort(wordCounts, decreasing = TRUE)
    length(wordCounts)
    wordCounts <- head(wordCounts, max_words)
    length(wordCounts)
    head(wordCounts)

    cloudFrame <- data.frame(word = names(wordCounts), freq = wordCounts)
    suppressWarnings(wordcloud(cloudFrame$word, cloudFrame$freq))
}

wordCloudFromDataFrame(azm, 200)
```

```r
wordCountsVector <- function(df_, max_words = 50) {
    df1_ <- df_[df_$review_content != "", ]

    review_content <- as.vector(df1_$review_content)
    review_content <- iconv(review_content, from = "UTF-8", to = "UTF-8",
        sub = "")
    words.vec <- VectorSource(review_content)
    words.corpus <- Corpus(words.vec)
    words.corpus <- tm_map(words.corpus, removePunctuation)
    words.corpus <- tm_map(words.corpus, removeNumbers)
    words.corpus <- tm_map(words.corpus, content_transformer(tolower))
    conflicts_prefer(tm::stopwords)
    words.corpus <- tm_map(words.corpus, removeWords, stopwords("english"))

    tdm <- TermDocumentMatrix(words.corpus)
    tdm

    m <- as.matrix(tdm)

    wordCounts <- rowSums(m)

    totalWords <- sum(wordCounts)
    totalWords
    words <- names(wordCounts)
    head(words)

    wordCounts <- sort(wordCounts, decreasing = TRUE)
    length(wordCounts)
    wordCounts <- head(wordCounts, max_words)
    return(wordCounts)

    # cloudFrame<-data.frame(word=names(wordCounts),freq=wordCounts)
    # return(cloudFrame)
}

wcv <- wordCountsVector(azm)
```

Answer:


```
##      good  product  quality      use      can      one    cable     like
##      4485     2800     2080     1492     1426     1269     1233     1155
##     price     will     also    using    phone charging  battery     easy
##      1147     1141     1138      951      948      871      772      759
##      time     just     well  working      buy    watch    sound      get
##       748      747      737      728      683      669      668      649
##      used     even   better    works    great   really     dont     best
##       637      635      593      591      567      565      560      559
##       now     fast      got     much    water     nice   camera     need
##       515      507      488      474      470      451      450      449
##    amazon    money    power  overall     fine   screen     work      bit
##       448      440      431      427      426      426      419      417
##    little     long
##       412      407
```

```r
nrow(azm[azm$rating > 3, ])
```

```
## [1] 1455
```

```r
azm_pos <- azm[azm$rating > 4, ]
nrow(azm_pos)
```

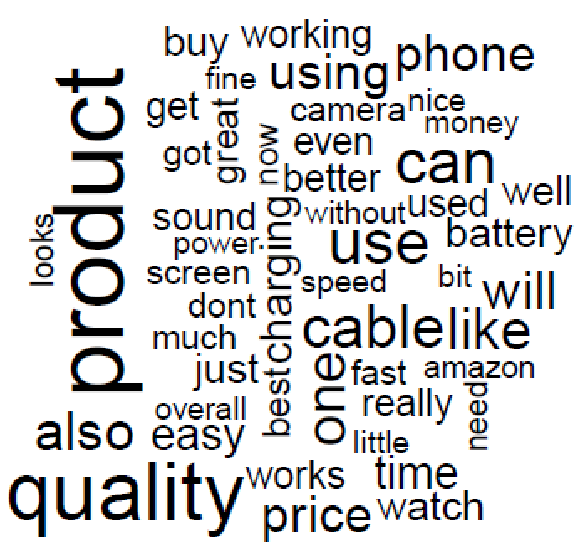
```
## [1] 930
```

```r
azm_neg <- azm[azm$rating < 2.5, ]
nrow(azm_neg)
```

```
## [1] 3
```

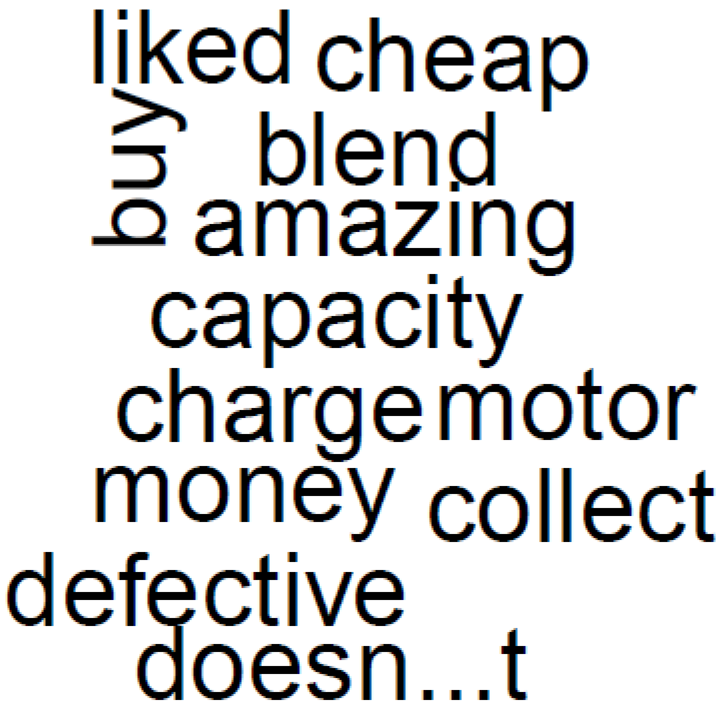
```r
# The data Amazon provides in this is strongly biased, with only 7
# reviews below 3, but 1455 reviews above 3
```

```r
wordCloudFromDataFrame(azm_pos, 50)
```



```
wordCloudFromDataFrame(azm_neg, 20)
```

```
## [conflicted] Removing existing prefer
## [conflicted] Will prefer tm::stopword
```



7. Are good or bad user feelings about a product more likely to generate a high volume or ratings and reviews? Are users more motivated to write a good product review or a bad product review?

<u>**R Code & Answer:**</u>

```
azm1 <- azm %>%
    mutate(word_present = as.numeric(str_detect(review_content, fixed("good"))))

azm1 <- azm1 %>%
    filter(!is.na(rating), !is.na(word_present))

correlation <- cor(azm1$rating, azm1$word_present)
correlation
```

```
## [1] 0.1193504
```

```
# check for correlation between word and rating
correlate <- function(word) {
    azm1 <- azm %>%
        mutate(word_present = as.numeric(str_detect(review_content, fixed(word))))

    azm1 <- azm1 %>%
        filter(!is.na(rating), !is.na(word_present))

    correlation <- cor(azm1$rating, azm1$word_present)
    return(correlation)
}

# find the word with the strongest correlation for rating
max(c(correlate("good"), correlate("product"), correlate("quality"), correlate("use"),
    correlate("can"), correlate("one"), correlate("cable"), correlate("like"),
    correlate("price"), correlate("will"), correlate("also"), correlate("using"),
    correlate("phone"), correlate("charging"), correlate("battery"), correlate("easy"),
    correlate("time"), correlate("just"), correlate("well"), correlate("working"),
    correlate("buy"), correlate("watch")))
```

```
## [1] 0.1292382
```

```
# the word 'good' has the strongest correlation with a value of 0.129
```

```
cor(AmazonData4$rating, AmazonData4$rating_count)
```

```
## [1] 0.1022348
```

```
# 0.1022348

# Yes, there is a weak positive correlation between rating and
# rating_count of 0.1022348.  The higher the product rating, the more
# likely the buyer is to rate that product, which means a higher
# rating count for that product.
```

8. How are ratings distributed based on quantity of ratings? Bucket all ratings (0 - 1, 1 - 2, 2 - 3, etc) and visualize this distribution.

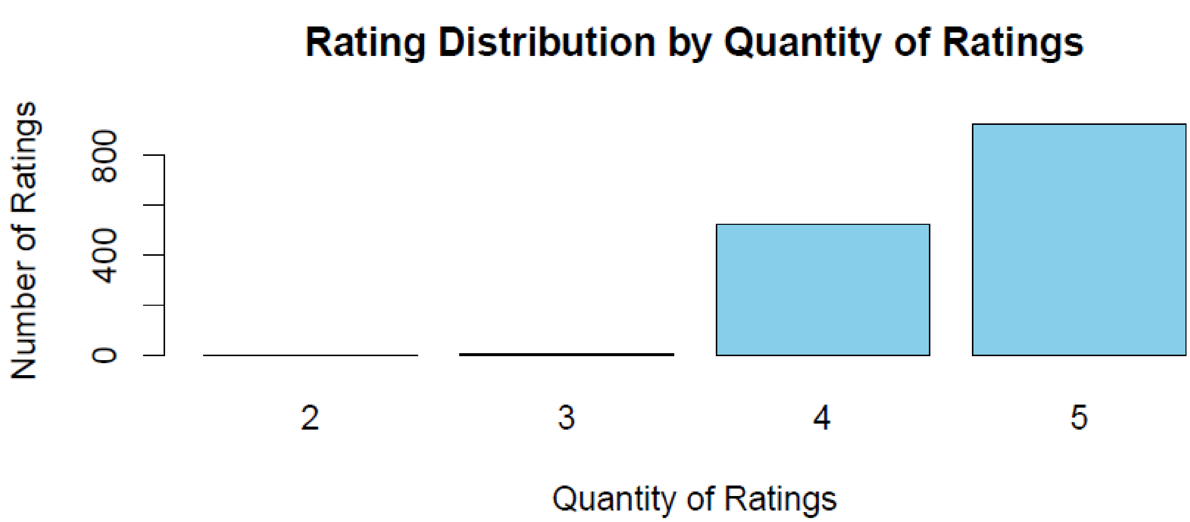
R Code:

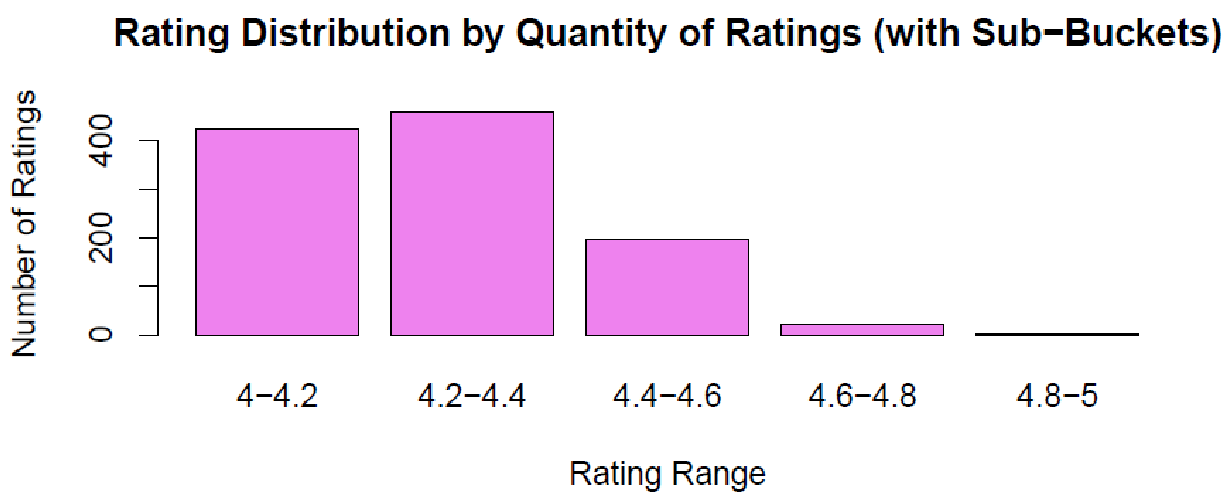
```r
# Step 1: Create Rating Buckets
AmazonData4$rating_bucket <- cut(AmazonData4$rating, breaks = seq(0, 5,
    by = 1), labels = FALSE)

# Step 2: Count the Number of Ratings in Each Bucket
rating_counts <- table(AmazonData4$rating_bucket)
```

```r
# Step 3: Visualize the Distribution
barplot(rating_counts, main = "Rating Distribution by Quantity of Ratings",
    xlab = "Quantity of Ratings", ylab = "Number of Ratings", col = "skyblue")
```

```r
# Step 4: Deeper dive into ratings 4 - 5

# Step 4.1: Create Rating Buckets with Sub-Buckets
breaks <- seq(4, 5, by = 0.2)
AmazonData4$rating_bucket <- cut(AmazonData4$rating, breaks = breaks, labels = FALSE,
    right = FALSE)

# Step 4.2: Count the Number of Ratings in Each Bucket
rating_counts <- table(AmazonData4$rating_bucket)

# Step 4.3: Visualize the Distribution
barplot(rating_counts, main = "Rating Distribution by Quantity of Ratings (with Sub-Buckets)",
    xlab = "Rating Range", ylab = "Number of Ratings", col = "violet",
    names.arg = paste(breaks[-length(breaks)], breaks[-1], sep = "-"))
```

Answer:






```
# Most ratings fall between 4.2 and 4.4.
```

9. Is there a way to estimate the actual number of sales based on the available data here?

<u>**RCode & Answer:**</u> We found that the best we could do with this data was to use the ratings count * discounted_price_usd formula which we've used in previous questions. While there are lightly positive correlations present in the data, there isn't enough to make a confident guess into the actual number of sales based on the data available in this dataset.

```
cor(AmazonData4$discount_percentage, AmazonData4$rating_count)
```

```
## [1] 0.01129439
```

```
# 0.01129439
cor(AmazonData4$discount_percentage, AmazonData4$rating)
```

```
## [1] -0.155679
```

```
#-0.155679

# We expected that the greater the discount percentage, the higher
# the rating would be and the higher the rating count would be, ie we
# expected a positive and stronger correlation between discount
# percentage and rating, as well as discount percentage and rating
# count. Contrary to what we expected, the resulting correlation was
# actually very weak and negative


cor(AmazonData4$rating, AmazonData4$rating_count)
```

```
## [1] 0.1022348
```

```
# 0.1022348

# We though that the higher ratings would be conducive to higher
# rating counts, that is to say we expected a positive and stronger
# correlation between rating and rating_count We were correct that
# there was a positive correlation, but the correlation was much
# weaker than we expected


cor(AmazonData4$discounted_price_usd, AmazonData4$rating_count)
```

```
## [1] -0.02730249
```

```
#-0.02730249
cor(AmazonData4$actual_price_usd, AmazonData4$rating_count)
```

```
## [1] -0.03621571
```

```
#-0.03621571

# The correlation between discounted_price_usd and rating_count, as
# well as actual_price_usd and rating_count were both what we
# expected. We thought online shoppers would expect different prices
# for different items, so We did not expect price alone be a
# significant factor in rating or rating_count. If there would be a
# correlation at all, it would probably be negative, because nobody
# wants to pay more.


cor(AmazonData4$discounted_price_usd, AmazonData4$rating)
```

```
## [1] 0.1211309
```

```
# 0.1211309
cor(AmazonData4$actual_price_usd, AmazonData4$rating)
```

```
## [1] 0.1224666
```

```
# 0.1224666

# We expected these results to be the same as the above, but there
# was actually a positive correlation between discounted_price_usd
# and rating as well as actual_price_usd and rating the correlation
# was weak as we expected, but we didn't expect it to be positive for
# the same reason mentioned above.


# The last 4 results are very weak, but surprisingly consistent.
```

10. Based on the answers to questions 1, 2 and 3, pick the Tier 1 category with the most user activity. Now continue that analysis down from every subcategory, tier 2 - tier 6. What new takeaways are there from this detailed analysis? Are there any outliers in the data that can be identified?

R Code & Answer:

Q 10.1: Electronics is the Tier 1 category we have selected

```
electronics <- AmazonData4 %>%
    filter(Cat1 == "Electronics")

electronics_revenue <- electronics %>%
    group_by(Cat2) %>%
    summarize(total_revenue = sum(discounted_price_usd, na.rm = TRUE)) %>%
    arrange(-total_revenue)

electronics_revenue
```

Q 10.2: Finding subcategories

```
## # A tibble: 9 x 2
##   Cat2                                       total_revenue
##   <chr>                                              <dbl>
## 1 HomeTheater,TV&Video                             20232.
## 2 Mobiles&Accessories                              13783.
## 3 WearableTechnology                                2134.
## 4 Headphones,Earbuds&Accessories                     751.
## 5 HomeAudio                                          297.
## 6 Cameras&Photography                                244.
## 7 Accessories                                        136.
## 8 GeneralPurposeBatteries&BatteryChargers             64.5
## 9 PowerAccessories                                    15.5
```

```
# Home Theater, TV and Video is the Electronics subcategory that
# earns the most revenue

homeTheater <- electronics %>%
    filter(Cat2 == "HomeTheater,TV&Video")

homeTheater_revenue <- homeTheater %>%
    group_by(Cat3) %>%
    summarize(total_revenue = sum(discounted_price_usd, na.rm = TRUE)) %>%
    arrange(-total_revenue)

homeTheater_revenue
```

```
## # A tibble: 5 x 2
##   Cat3                         total_revenue
##   <chr>                                <dbl>
## 1 Televisions                        19296.
## 2 Accessories                          510.
## 3 Projectors                           360.
## 4 SatelliteEquipment                    41.6
## 5 AVReceivers&Amplifiers                23.9
```

```
# Televisions is the Home Theater category that earns the most
# revenue

televisions <- homeTheater %>%
    filter(Cat3 == "Televisions")

tv_revenue <- televisions %>%
    group_by(Cat4) %>%
    summarize(total_revenue = sum(discounted_price_usd, na.rm = TRUE)) %>%
    arrange(-total_revenue)

tv_revenue
```

```
## # A tibble: 2 x 2
##   Cat4                      total_revenue
##   <chr>                             <dbl>
## 1 SmartTelevisions                 18779.
## 2 StandardTelevisions                517.
```

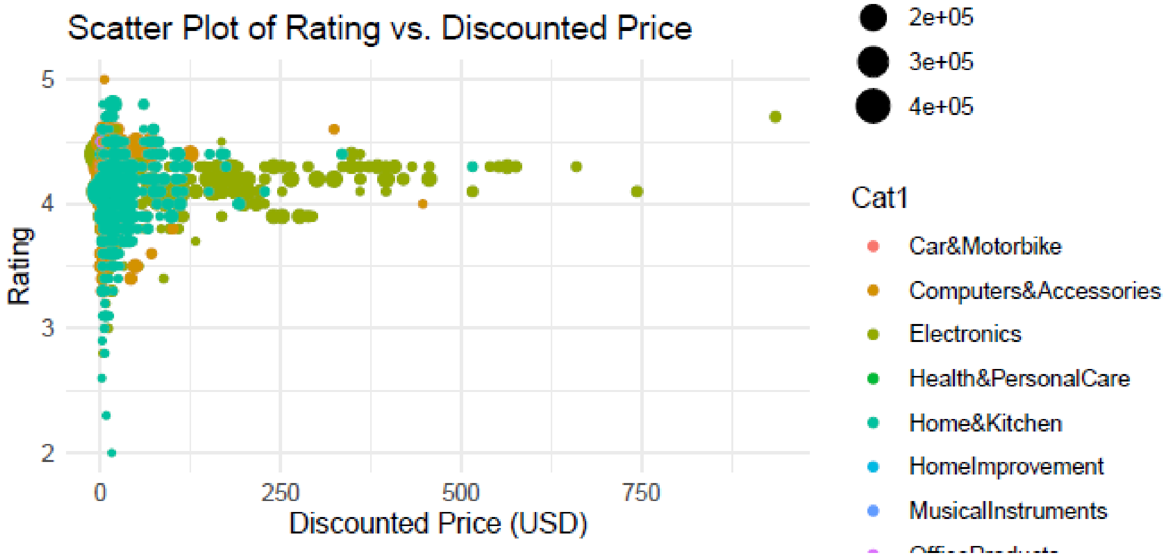
```
# Smart TV is the Televisions category that earns the most revenue
# This is the end of the category tier for this line of products
```

```
unique(AmazonData4$Cat1)
```

```
## [1] "Computers&Accessories" "Electronics"           "MusicalInstruments"
## [4] "OfficeProducts"        "Home&Kitchen"          "HomeImprovement"
## [7] "Toys&Games"            "Car&Motorbike"         "Health&PersonalCare"
```

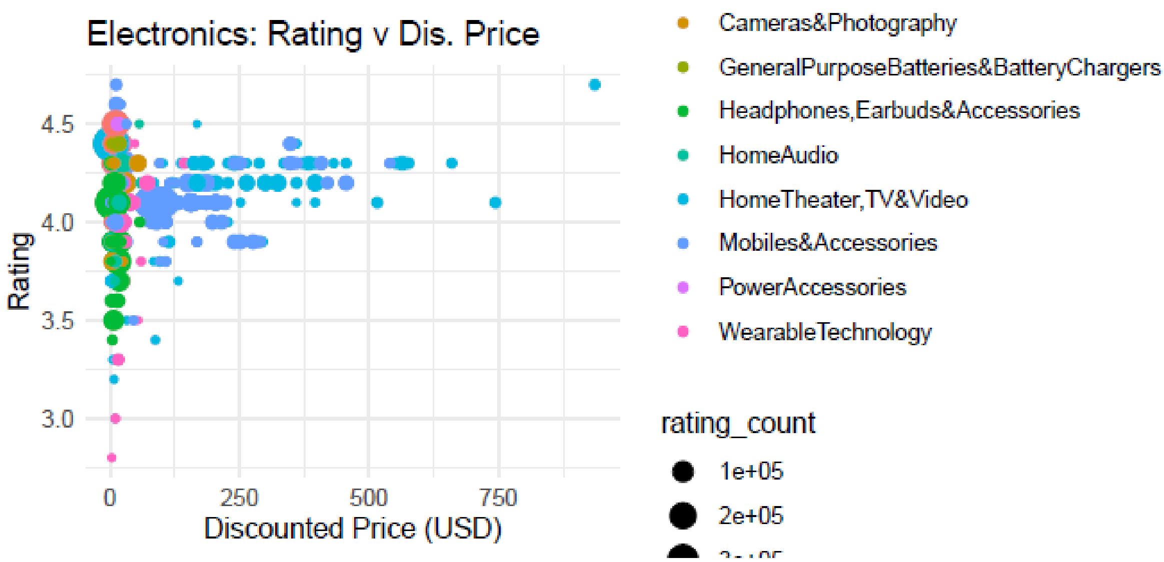
```
ComputersAccessories <- AmazonData4 %>%
    filter(Cat1 == "Computers&Accessories")

MusicalInstruments <- AmazonData4 %>%
    filter(Cat1 == "MusicalInstruments")

OfficeProducts <- AmazonData4 %>%
    filter(Cat1 == "OfficeProducts")

HomeKitchen <- AmazonData4 %>%
    filter(Cat1 == "Home&Kitchen")

HomeImprovement <- AmazonData4 %>%
    filter(Cat1 == "HomeImprovement")

ToysGames <- AmazonData4 %>%
    filter(Cat1 == "Toys&Games")

CarMotorbike <- AmazonData4 %>%
    filter(Cat1 == "Car&Motorbike")

HealthPersonalCare <- AmazonData4 %>%
    filter(Cat1 == "Health&PersonalCare")
```

```
ggplot(data = AmazonData4, aes(x = discounted_price_usd, y = rating, color = Cat1,
    size = rating_count)) + geom_point() + labs(title = "Scatter Plot of Rating vs. Discounted Price",
    x = "Discounted Price (USD)", y = "Rating") + theme_minimal()
```



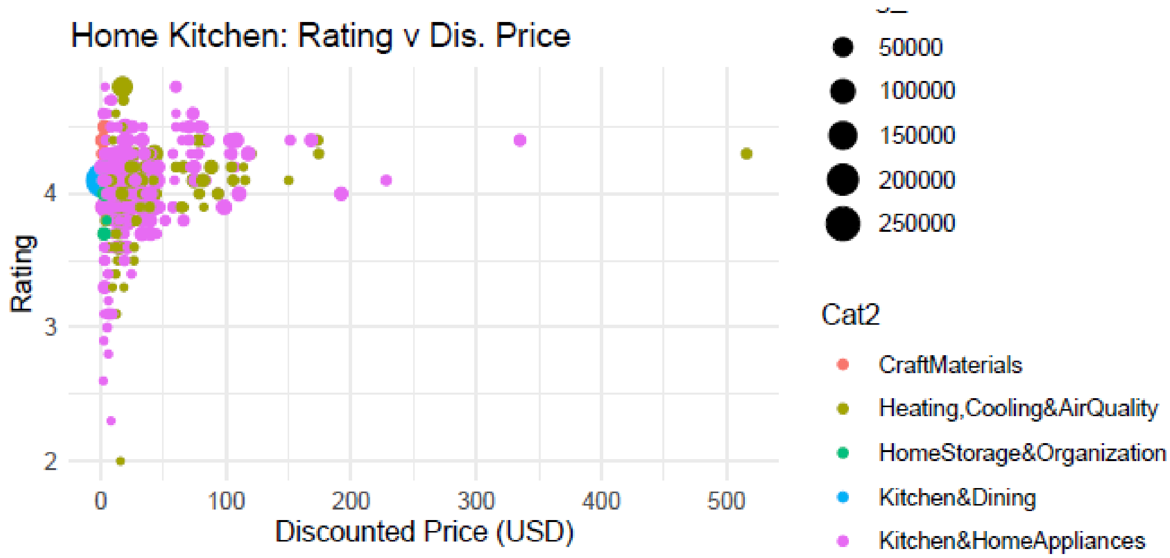

```
# Takeaways: Electronics has the highest priced products. The cheaper
# the product, the higer the ratings count. The average rating for
# most products falls between 4 and 4.5.
```

```
ggplot(data = electronics, aes(x = discounted_price_usd, y = rating, color = Cat2,
    size = rating_count)) + geom_point() + labs(title = "Electronics: Rating v Dis. Price",
    x = "Discounted Price (USD)", y = "Rating") + theme_minimal()
```




```
# Takeaways: Home Theater has the most expensive products. Mobiles
# get the highest volume of ratings. Most ratings fall between 4 and
# 4.4.
```

```
ggplot(data = HomeKitchen, aes(x = discounted_price_usd, y = rating, color = Cat2,
    size = rating_count)) + geom_point() + labs(title = "Home Kitchen: Rating v Dis. Price",
    x = "Discounted Price (USD)", y = "Rating") + theme_minimal()
```



Interpretations

After conducting a detailed analysis of Amazon data, some interesting insights have emerged. By examining various metrics and trends, we can gain valuable information about Amazon's performance and market position. These insights were able to shed light not only on Amazon's growth, but also their customer behaviors. Through this thorough analysis we were able to confirm somethings and find others that were a little surprising.

One thing that was more of a confirmation was the relationship between price and rating. As an Amazon Customer myself, I have been on the other side of top-rated products rising in sales price due to thousands of top tier ratings. On the other side, one thing that was surprising to me was that Office Supplies had the highest rating. I believed it would be electronics due to Black Friday, Prime Day and the devices Amazon has used to change the market, but I was also pleasantly surprised it was not. This got me thinking that it would be interesting to see if these questions would reflect the same answers when used on Amazon US Sales Data. Are office supplies popular because there is a multitude of remote positions that are outsourced from the US? Does it only seem popular because there are only 30 observations, compared to a bigger subset like Electronics that has over 500?

Based on our final linear regression model, we confirmed a couple of ways we can predict an Amazon product's final rating:

1. As the discount percentage increases, the rating tends to decrease.
2. As the actual (non-discounted) price increases, the rating also tends to increase.
3. As the discounted price increases, the rating tends to decrease.
4. Overall (based on P Values), discount percentage and actual price have the strongest impact on an Amazon product's final rating compared to the discounted price.

We believe that consumers trust that expensive products are worth their investment, and the court of public opinion would never steer them wrong. Expensive products validate their monetary decisions as a consumer, and therefore they like to give those products high ratings. However, if they bought a product that was heavily discounted – especially a product that was heavily discounted but remains a pricey product – shoppers are more likely to give the product a harsher rating, as they fear they were tricked by a shiny discount into buying a product that wasn't worth it. The reason a product was so heavily discounted in the first place may be because people stopped buying it due to its low quality. The second synopsis is other customers opinions matter a great deal; people are more likely to purchase an item with high ratings than one with low ratings.

Actionable Steps & Insights

Our analysis provides valuable insights, but we would have developed a deeper understanding with more granular data. Variables such as actual revenue per category, product view counts, and calendrical sales data (including special days like Black Friday or Prime sales) would provide better context. There were 1455 reviews with ratings above 3, but a mere 7 reviews below 3. Our analysis would be better balanced if the dataset included far more negative reviews.

Our word cloud analysis demonstrates the importance consumers place on product quality. Reviews including words such as 'good', 'quality', and 'product' correlate positively with higher ratings and sales, whereas words like 'cheap' and 'defective' indicate decreased ratings and sales. Additionally, we identified a weakly negative correlation of -0.155679 between discount percentage and product rating. Like the insights we gleaned from the word clouds, this indicates that offering greater discounts on lower quality items is not an effective strategy. To prioritize customer satisfaction, we need to offer good quality products, even at a higher price.

The data identifies that home theater, TV (particularly Smart TVs), and video yield the highest revenue within the electronics category. Therefore, these products are worthiest of greater marketing budgets. We can promote items consumers are more likely to purchase alongside Smart TVs, such as sound systems, streaming devices, TV furniture, etc., employing similar marketing strategies by prioritizing higher quality and higher prices over lower quality and higher discounts.

The data proved that quality reigns supreme. So, one insight we would offer is to make sure whatever you are selling is a good quality product, because if it is not the customers will know and they will alert everyone else to this fact.

References

1. https://www.aboutamazon.com/about-us
2. https://www.aboutamazon.com/impact
3. Facts about Amazon