DATA ANALYSIS PORTFOLIO

MY PORTFOLIO

I am a dedicated and committed individual who is very keen on solving problems. I am always ready to take up any challenge and very good at working under pressure.
Above all I love to draw insights from data. It is so fascinating how data can help in a lot of ways.

......I am passionate about everything I do......

Portfolio Outline

Portfolio Outline

Portfolio Outline

Portfolio Outline

Udemy Project Description

The dataset for the education tech company Udemy contained four different files of four different courses; Web development, Business, Music and Design courses and the main objective of the analysis was to analyze their revenue.

- To also understand why "Web Development" courses are believed to be the most popular courses and why a recommendation to increase its charge would increase revenue.

- The Udemy Course Dataset was imported into excel, merged, appended, cleaning was done. Visualization was done using excel, and tableau.

- An increase in the price of web design and graphic design would result in a substantial increase in the revenue generated by Udemy. However, it is suggested that more free courses be made available to beginners in these courses also in other to drive their interest, especially in this new era of tech dominance.

The Problem

The aims of this analysis were;

- To understand how best to increase the next quarterly revenue for the education tech company Udemy and also to track the performance of courses in Udemy

- To also understand why "Web Development" courses are believed to be the most popular courses and why a recommendation to increase its charge would increase revenue.

Data Design

- The Udemy Course Dataset was imported into excel, merged, appended, cleaning was done using excel by first removing duplicates & blank cells, headers were checked and the consistency of each data was checked such as the "Web development" sub-category in "Subject category".

- Vlookup Function was used to get data of the first top 20 courses into a new spreadsheet.

- Visualization was done using pivot tables of the representative 20 courses in excel, tableau and insights were drawn from each representative visual. A root cause analysis was then generated to understand our insights better.

Findings
Findings 1

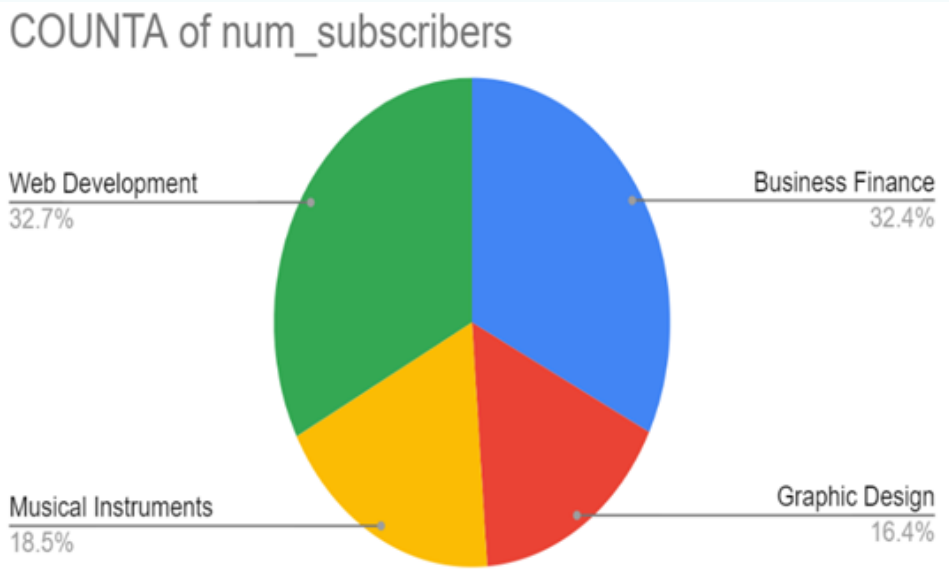### Figure 1: Total Subscriptions by Category

COUNTA of num_subscribers

Web Development
32.7%

Business Finance
32.4%

Musical Instruments
18.5%

Graphic Design
16.4%

Figure 2: Average Content duration by Category

AVERAGE of content_duration vs. subject
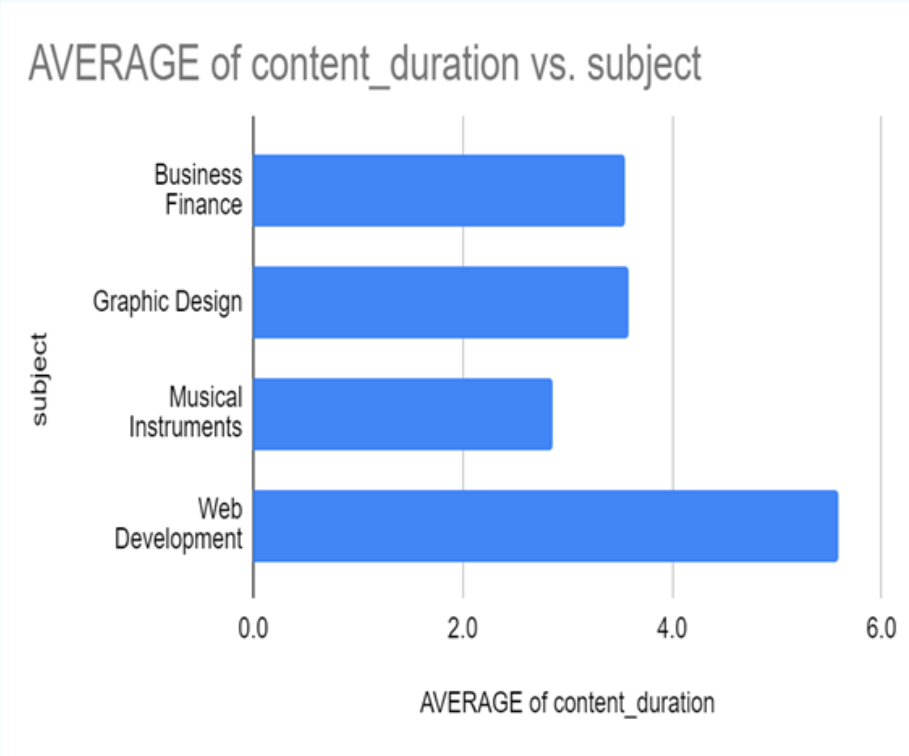
Findings 2

Table 1: Average Course Rating by Skill Level

AVERAGE of rating level subject	All Levels	Beginner Level	Expert Level	Intermediate Level
Business Financ	0.69	0.69	0.70	0.70
Graphic Design	0.73	0.73	0.88	0.72
Musical Instrume	0.31	0.31	0.30	0.28
Web Developme	0.65	0.64	0.50	0.67
Grand Total	0.6193714286	0.5973170732	0.5901724138	0.6157345972

Figure 3:Average Course Rating by Skill Level
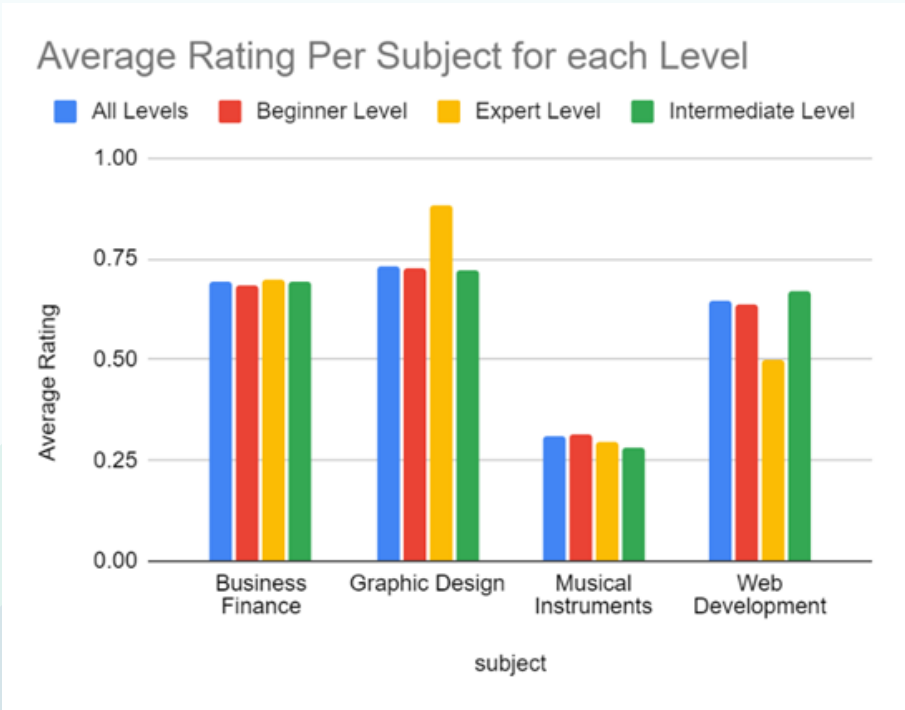
Finding 3

Table 2: Count of Free and Paid by Subject

COUNTA of Cou Course_fee_category		
Subject	Free	Paid
Business Financ	1	
Musical Instrume	1	1
Web Developme	10	7
Grand Total	12	8

Figure 4: Sum of Lectures by Course

Sum of Number of Lectures per Subject
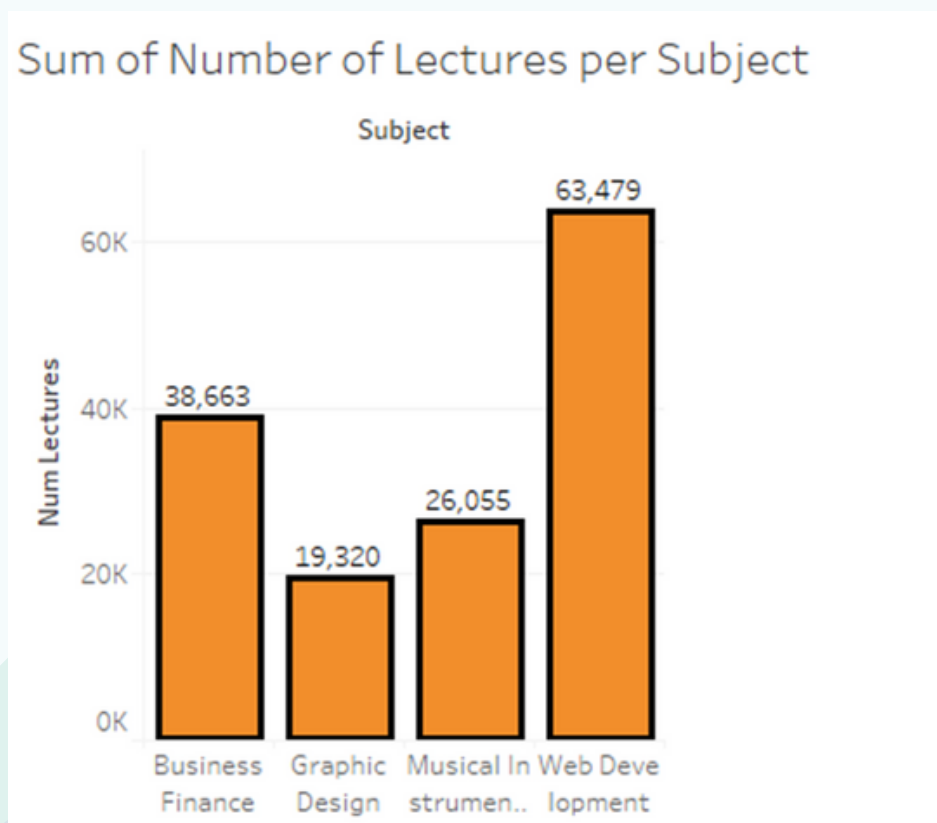
Analysis

After understanding and framing the problem with the Udemy Dataset and gathering initial insights from the data, The 5 Whys technique was used to determine its root cause.

- Why is Web Development course the most subscribed subject than other subjects? It has more free courses for beginners.
- Why does it have more free courses for beginners? To encourage subscribers who want to start a career in tech.
- Why only are subscribers who want to start a career in tech? This is because the resources for web development courses are easily available.
- Why are the courses for Web design easily available? This is because the internet is almost available globally

.

Looking at the count of subscribers in the charts above Web Development course had the highest number accompanied by "Business Finance", "Graphic Design" and then "Musical Instrument". This same order was also seen in a comparison of the number of reviews and average price per subject. However, a different order when analyzing the Udemy Course data in relation to the average duration of the course, the average count of subscribers, and the average course rating indicated that the highest for each of these categories was "Web Development", "Graphic Design" "Business Finance", and then "Musical Instrument". Since Average is a better measure of a true value it can be said that "Graphic Design" had a higher count of subscribers, rating, and duration of content than "Business Finance".

Analysis 2

The root cause analysis into why Web Development is the most subscribed subject indicated that this might be because it had certain course titles that are free for beginners, which would give subscribers a feel of the course thereby drawing their interest to pay and learn further.

However, the highest rating was observed in "Graphic Design" this can be attributed to the fact that today the tech industry is growing at an exponential rate, especially with experts trying to keep up with new innovations and designs to better improve their services.

Conclusion

In conclusion, an increase in the price of web design and graphic design would result in a substantial increase in the revenue generated by Udemy. However, it is suggested that more free courses be made available to beginners in these courses also in other to drive their interest, especially in this new era of tech dominance. Web Development has the highest number of lectures among all the 20 most subscribed courses thus an increase in the course content of other subjects might also increase the number of subscribers thereby adding to the revenue generated by the "web Development" course.

Google Play Store Apps Project Description

The Google Play Store apps data was gotten from Kaggle. The Dataset were in two different files one with data size 140,943 and size 64296. The aims of this analysis were;

- To understand the google play store and also to track the performance of the different Apps category in Google Play Store.

- To know why medical Apps has the highest number of installs and how this data can be used to improve Google Play store Apps.

- To also understand how sentiments affect people's choice of Apps.

- To also understand how the Play Store Data can be used to drive the Apps to success

Data Design

The Google Play Store Apps Dataset were in two different files one with data size 140,943 and the other on Google Play Store reviews with dataset size 64296 both were imported merging, appending, and cleaning of the data was done using excel first by removing duplicates and blank cells in both datasets and "NaN" was replaced with "0". Right/Left Function was used to separate the year last updated.

Datasets were also filtered and sorted using various values, ascending to descending order of values and descending to ascending order of values. Sorting of items in the dataset by rating from highest to lowest top 50 and vice versa. The filtering and sorting conditions in excel were;

- Sorting of items in Google Play Store reviews by "Sentiment_Polarity" from the highest to the lowest, the "Average Sentiment_Subjectivity" of the top 50 and the bottom 50 data after sorting was analyzed.
- Sorting of items in Google Play Store reviews by "Sentiment_Subjectivity" from the highest to lowest, the "Average Sentiment_Polarity" of the first 50 and the bottom 50 after sorting was analyzed.

Data Design 2

- Sorting of items in Google Play Store reviews by both "Sentiment_Subjectivity" and "Sentiment_Polarity" by the value "0" after sorting the count of the Apps with both neutral remarks was analyzed.

During the sorting of items in the Google Play Store Dataset, it was noticed that the App 'Life Made Wi-fi' was missing a category and genres column leading to wrong entries in other columns a quick search on the Google Play Store on my phone and saw that this app was under "LIFESTYLE" category and did the input myself. Excel "Right Function" was used to bring out the year from the column with the header "Last_updated". Excel Vlookup Function was used to get the category of each App in the Google Play Store reviews data.

Visualization was done using pivot tables in excel, tableau, and insights were drawn from each representative visual. A root cause analysis was then generated to understand our insights better.

Findings

Findings 1

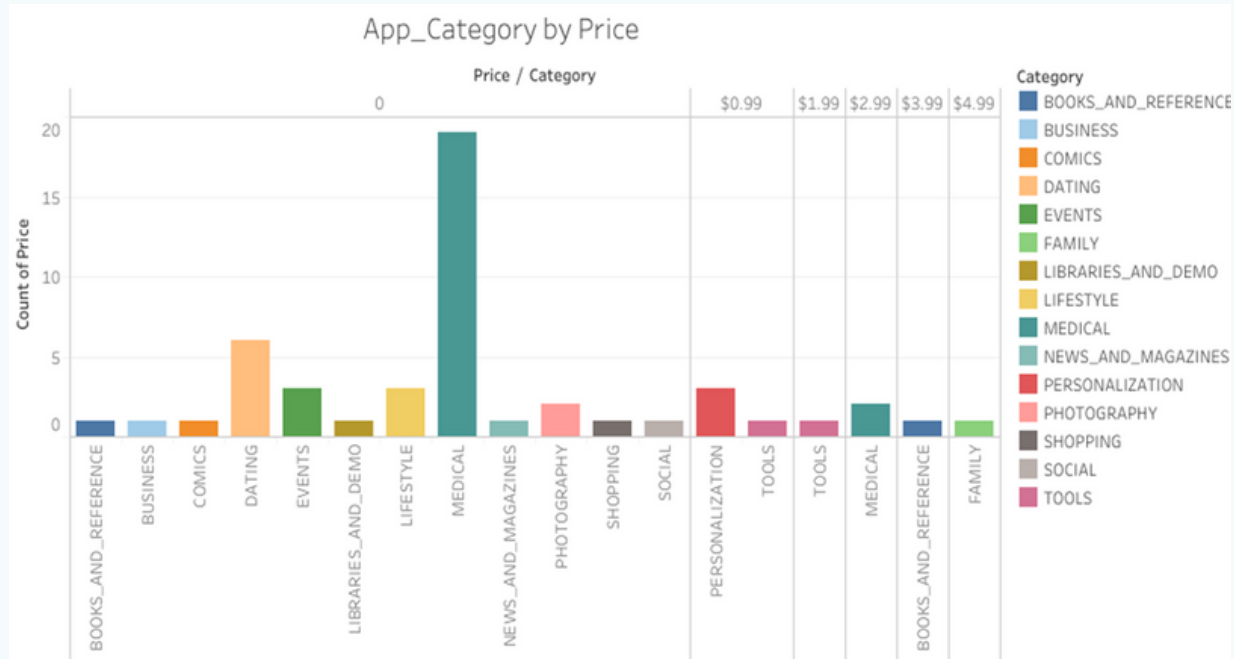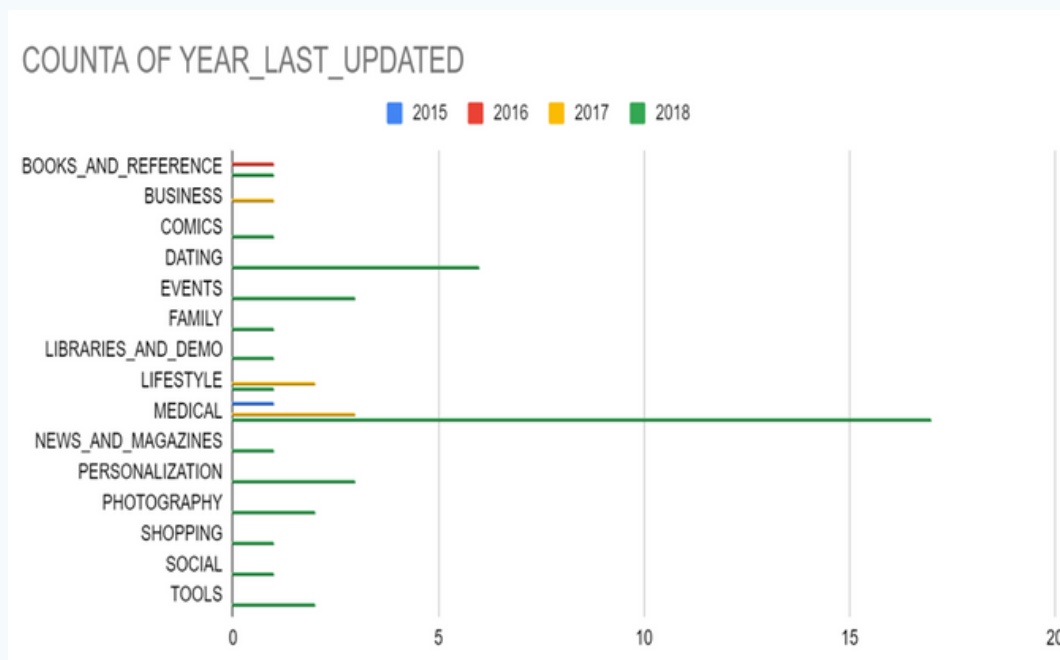### Figure 1: Price by Category

Figure 2: Last update by Category

Findings 2

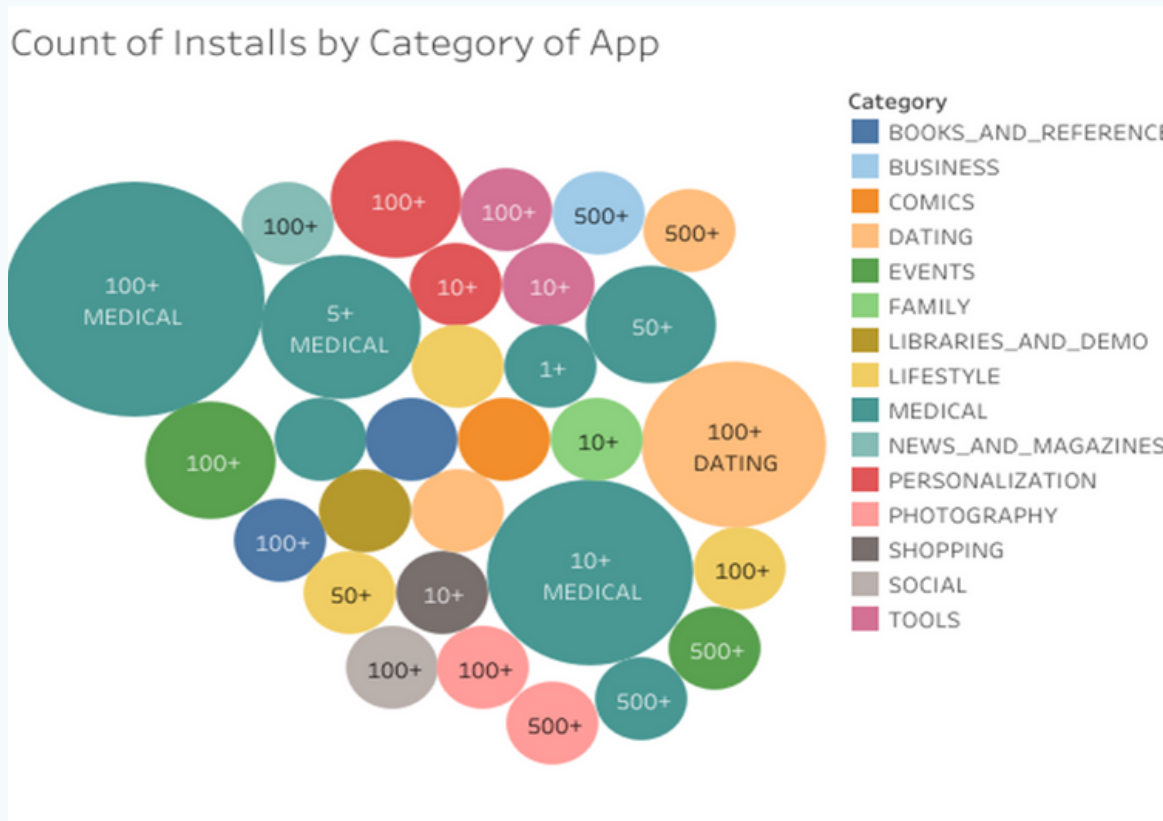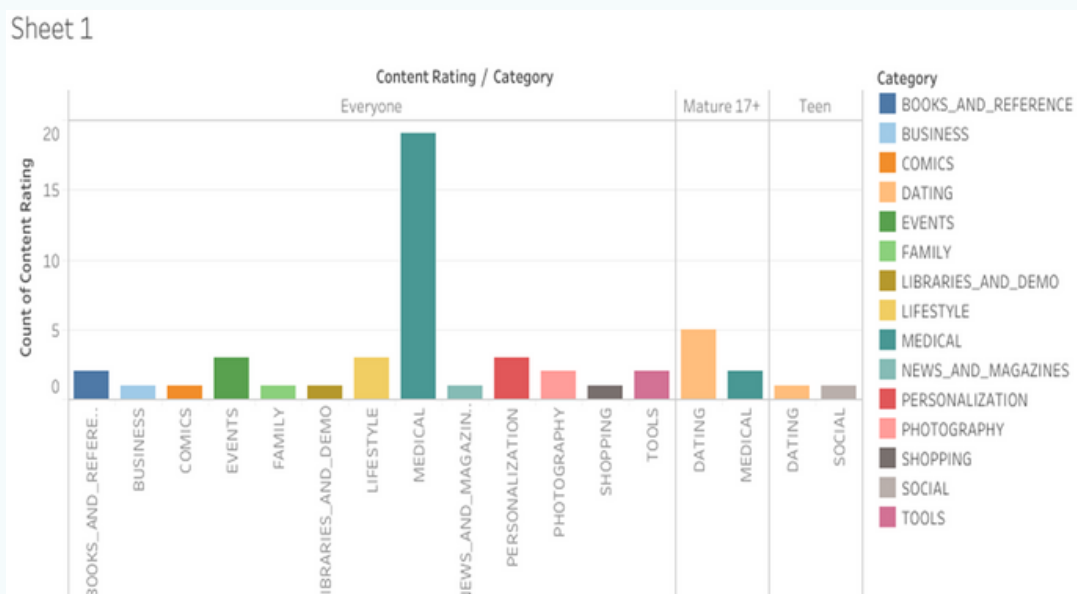## Figure 3: Total installs by Category and Size

Count of Installs by Category of App

Figure 4: Content Rating by Category

Sheet 1

Findings 3

Figure 5: Neutral values by Category

COUNTA of Sentiment_Polarity and COUNTA of Sentiment_Subjectivity with NEUTRAL VALUES

#N/A — 7.8%
TRAVEL_AND_LOCAL — 5.0%
TOOLS — 7.6%
SPORTS — 3.5%
SOCIAL — 2.3%
SHOPPING — 2.2%
PRODUCTIVITY — 3.7%
PHOTOGRAPHY — 3.0%
PERSONALIZATION — 3.7%
MEDICAL — 4.3%

BOOKS_AND_REFERE... — 2.6%
BUSINESS — 4.0%
COMMUNICATION — 3.3%
DATING — 4.3%
ENTERTAINMENT — 2.7%
FAMILY — 4.3%
FINANCE — 4.2%
GAME — 4.7%
HEALTH_AND_FITNESS — 4.1%
LIFESTYLE — 4.0%
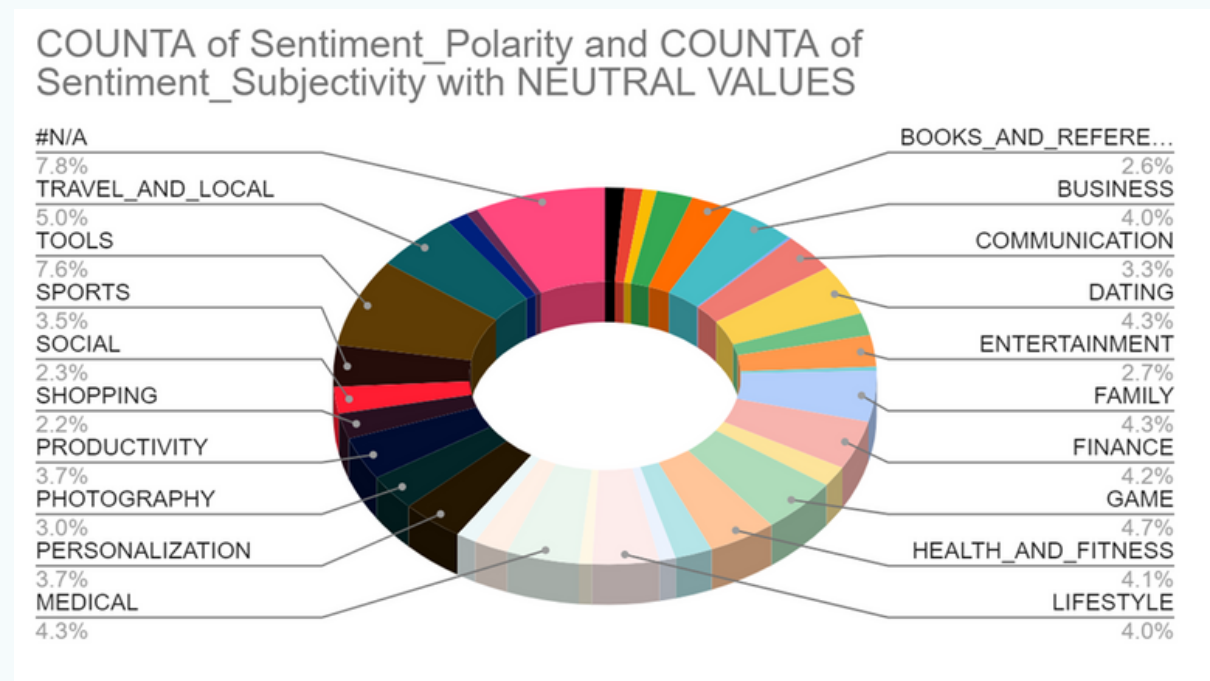
Figure 6: Total Polarity by Category

COUNTA of Sentiment_Polarity

#N/A — 4.1%
TRAVEL_AND_LOCAL — 2.0%
SPORTS — 10.2%
SHOPPING — 10.2%
PERSONALIZATION — 6.1%
MEDICAL — 6.1%
HEALTH_AND_FITN... — 20.4%

COMMUNICATION — 4.1%
DATING — 24.5%
GAME — 12.2%
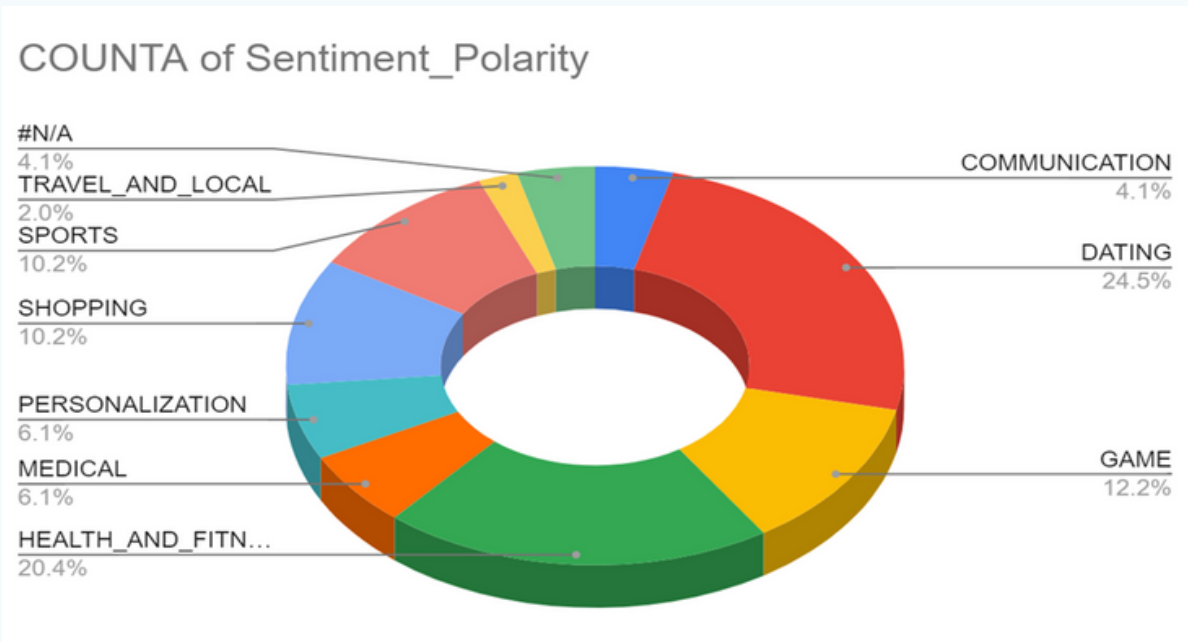
Findings 4

Figure 7: Bottom Average Polarity by Category

Bottom 50 of AVERAGE of Sentiment_Polarity vs. Category
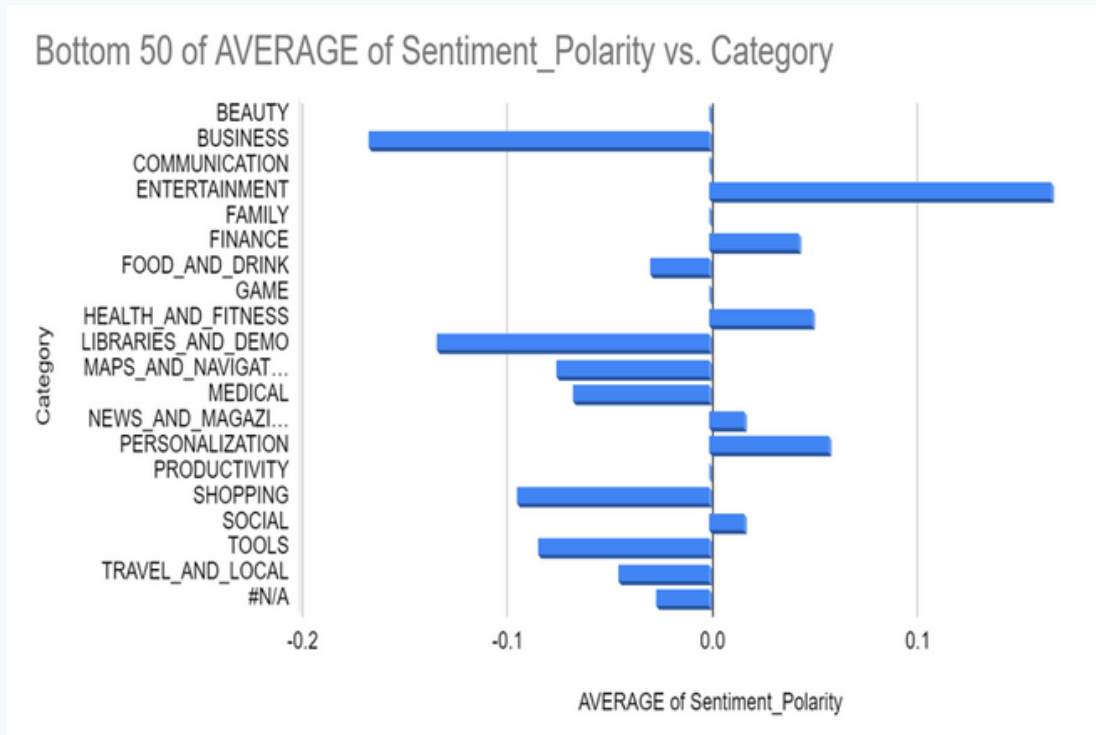
Figure 8: Top Subjectivity by Category

Top 50 of COUNTA of Sentiment_Subjectivity vs. Category
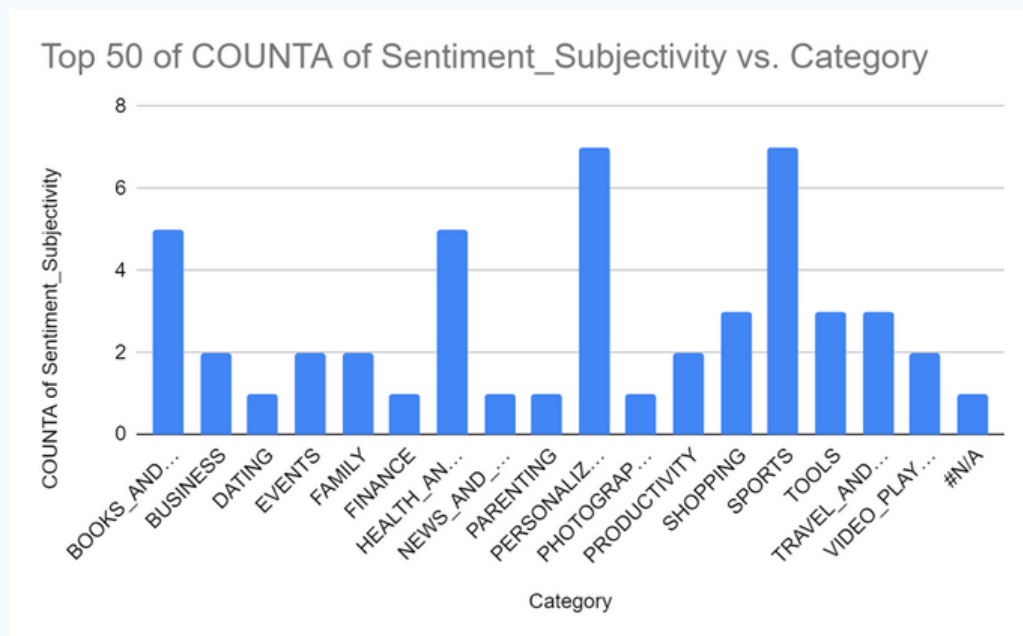
Findings 5

Figure 9: Bottom Subjectivity by Category

Bottom 50 of AVERAGE of Sentiment_Subjectivity vs. Category
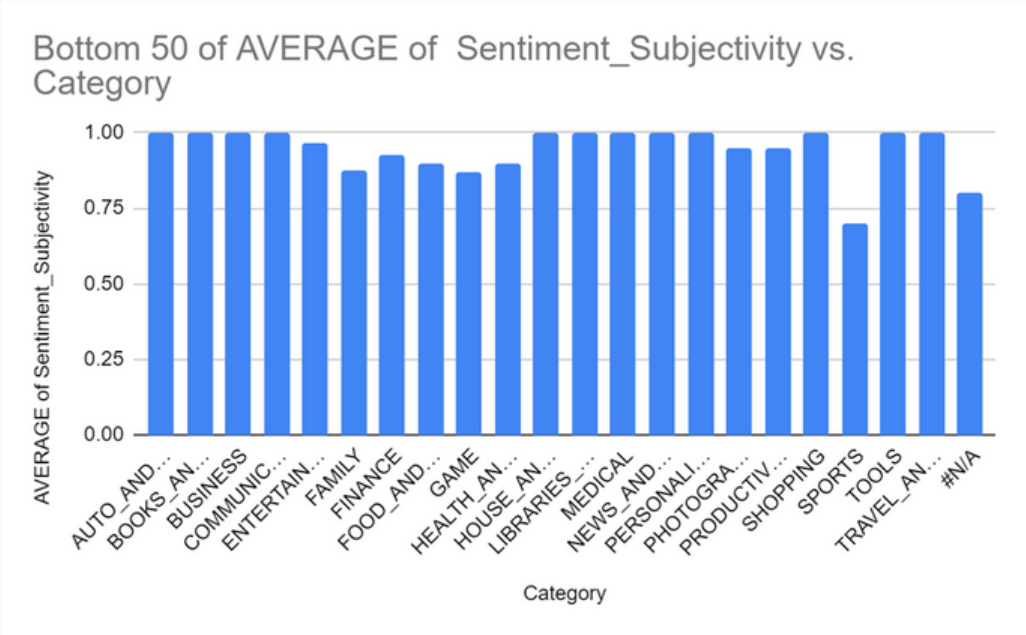
Analysis

After understanding and framing the problem with the Google Play Store Dataset and gathering initial insights from the data, The 5 Whys technique was used to determine its root cause.

- Why is Medical Apps the most installed App Category? It has more free Apps whose content are suitable for all age brackets.
- Why does it have more free Apps whose content are more suitable for everyone? This is because health issues are a thing of general concern.

. Looking at the count of installs of each App category in the charts above the Apps under "Medicals" had the highest number of installs showing that people are more engaged in Apps that would give them medical advice. We could say that this was a result of the Apps in the medical Category having the highest number of Apps that were recently updated last between 2015 and 2018. Our root cause analysis into why the Apps in the medical Category were the most installed indicated that most of the Apps under this category had content that was suitable for everyone i.e all age brackets and also had more Apps that were free.

Analysis 2

Sentiment analysis is the use Of Natural Processing Language(NLP) and is widely used, especially as a part of social media analysis for any domain, be it a business, a recent movie, or a product launch, to understand its reception by the people and what they think of it based on their opinions or sentiment. Polarity refers to the strength of an opinion. It could be positive or negative.

From sorting the data based on subjectivity, the highest 50 were all selected and their polarity was found to be one with the Apps in the Dating category having the highest count of Apps with positive polarity of 1. This implies that people enjoy the idea of meeting someone on a dating App. Also, after sorting the data based on subjectivity, the lowest 50 were all selected and those Apps under the entertainment category had the lowest average polarity indicating there is a need to improve the Apps in the entertainment category.

Analysis 3

Further sorting of the data based on polarity, the highest 50 were all selected and their subjectivity was found to be 0.3 all through so a count was done to know the App category with the highest number, the Apps in the personalization category had the highest count. Hence to drive the interest of android users more Apps in the personalization category with the present features and more would drive the interest of users into using android phones.

Subjectivity refers to the degree to which a person is personally involved in an object. What matters the most here are personal connections and individual experiences with that object, which may or may not differ from someone else's point of view. While polarity refers to the overall positivity or negativity of a statement, subjectivity focuses on how much someone values an object and expresses his/her opinion about it.

Analysis 4

From the charts above after sorting the data based on category of Apps that have both subjectivity and polarity to be neutral the Apps that didn't fall into any category were the highest followed by the Apps in the Tools category from our root cause analysis this may be due to people not really having interest in the Apps in that category hence there is a need to improve the Apps under the tools category. Further sorting the data based on polarity, the lowest 50 were all selected and those Apps under the sports category had the lowest average subjectivity which implies a need for improvement in that category.

Conclusion

From the trends above the Google Play Store Apps can be improved by first updating Apps as often as possible so that users can have a feel of new features. Also Apps which can be used by all age bracket and free can be used to capture more users into using android phones. Also, creating more Medical and Dating Apps can be used to capture the android market.

This report recommends the creation of more Apps in the medical category whose content are free and have content that are for all age bracket. Also, the creation of more Apps with features like that of the Apps in the Dating category.

Appendix

Google Sheets Data Set for Udemy Project
https://docs.google.com/spreadsheets/d/1rN2H6s3go90
TujF-MZM9ojEST_XqgdQpLGiB7mTFWu4/edit?usp=sharing

Tableau Data Visualization For Google Play Store Apps Project
https://public.tableau.com/views/EntryCapstoneProject/
Dashboard1?:language=en-
US&:display_count=n&:origin=viz_share_link.

https://public.tableau.com/views/EntryCapstoneProject/
Dashboard2?:language=en-
US&:display_count=n&:origin=viz_share_link