Introduction and Setup

👋 Welcome a data analysis of Google Play market!

📈 Here, we delve into a comprehensive comparison of thousands of Android apps. By exploring this dataset, we aim to provide you with valuable insights on various aspects, including:

- App Category Competitiveness: Gain a deeper understanding of the competitiveness across different app categories such as Games, Lifestyle, Weather, etc.
- Identifying Lucrative App Categories: Discover app categories that present enticing opportunities based on their popularity and user engagement.
- Evaluating Pricing Strategies: Determine the potential impact of making your app paid versus free by analyzing the number of downloads you may gain or lose.
- Establishing Optimal Pricing Points: Discover the optimal price range for a paid app, considering factors such as user behavior and market demand.
- Maximizing Revenue: Identify the highest-grossing paid apps in the dataset, enabling you to learn from their success and apply similar strategies to your own app.
- Cost Recovery Analysis: Understand the likelihood of recouping the development costs of a paid app based on its sales revenue, helping you make informed decisions regarding your app's profitability.
- ...and Much More: Explore additional insightful findings and revelations within the dataset to further enhance your understanding of the Google Play Store landscape.

About the Dataset of Google Play Store Apps & Reviews

Data Source:
App and review data was scraped from the Google Play Store by Lavanya Gupta in 2018. Original files listed here.

Import Statements

Notebook Presentation

Read the Dataset

```
Loading the data ⏳...Done! ✅
The shape of the data: (10841, 12)

🔎 Let's check 5 random rows of data:
```

App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Ge

6187	Dating for 50 plus Mature Singles – FINALLY	DATING	4.60	13049	13.00	500,000	Free	0	Mature 17+	Da
2312	AE Garage	AUTO_AND_VEHICLES	4.40	64	66.00	1,000	Free	0	Everyone	Au Veh
8431	LightInTheBox Online Shopping	SHOPPING	4.00	41986	26.00	5,000,000	Free	0	Teen	Shop
7602	My Cycles Period and Ovulation	HEALTH_AND_FITNESS	4.30	26652	41.00	1,000,000	Free	0	Everyone	Heal Fit
6899	Acorns - Invest Spare Change	FINANCE	4.30	45962	9.15	1,000,000	Free	0	Everyone	Fina

Data Cleaning

Drop Unused Columns

❌ The columns `Last_Updated` and `Android_Version` will not be used. Let's drop them:

```
✅ Columns after dropping:

 ['App' 'Category' 'Rating' 'Reviews' 'Size_MBs' 'Installs' 'Type' 'Price'
 'Content_Rating' 'Genres']

✅ New shape: (10841, 10)
```

Find and Remove NaN values in Ratings

❓ The next step is to find out how many rows have a NaN value in the `Rating` column.

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	(
0	Ak Parti Yardım Toplama	SOCIAL	NaN	0	8.70	0	Paid	$13.99	Teen	
1	Ain Arabic Kids Alif Ba ta	FAMILY	NaN	0	33.00	0	Paid	$2.99	Everyone	Edu
2	Popsicle Launcher for Android P 9.0 launcher	PERSONALIZATION	NaN	0	5.50	0	Paid	$1.49	Everyone	Persona
3	Command & Conquer: Rivals	FAMILY	NaN	0	19.00	0	NaN	0	Everyone 10+	S
4	CX Network	BUSINESS	NaN	0	10.00	0	Free	0	Everyone	Bu

...	...	...	...	...	...	...	...	...	...	
5840	Em Fuga Brasil	FAMILY	NaN	1317	60.00	100,000	Free	0	Everyone	Sim
5862	Voice Tables - no internet	PARENTING	NaN	970	71.00	100,000	Free	0	Everyone	Pa
6141	Young Speeches	LIBRARIES_AND_DEMO	NaN	2221	2.40	500,000	Free	0	Everyone	Libr
7035	SD card backup	TOOLS	NaN	142	3.40	1,000,000	Free	0	Everyone	
7175	Android TV Remote Service	TOOLS	NaN	1	3.70	1,000,000	Free	0	Everyone	

1474 rows × 10 columns

```
✅ New shape of the data: (9367, 10)
```

Find and Remove Duplicates

👯 Let's check for duplicates. How many entries can you find for the "Instagram" app?

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Ratin
946	420 BZ Budeze Delivery	MEDICAL	5.00	2	11.00	100	Free	0	Mature 17
1133	MouseMingle	DATING	2.70	3	3.90	100	Free	0	Mature 17
1196	Cardiac diagnosis (heart rate, arrhythmia)	MEDICAL	4.40	8	6.50	100	Paid	$12.99	Everyor
1231	Sway Medical	MEDICAL	5.00	3	22.00	100	Free	0	Everyor
1247	Chat Kids - Chat Room For Kids	DATING	4.70	6	4.90	100	Free	0	Mature 17
...	...	...	...	...	...	...	...	...	.
10802	Skype - free IM & video calls	COMMUNICATION	4.10	10484169	3.50	1,000,000,000	Free	0	Everyor
10809	Instagram	SOCIAL	4.50	66577313	5.30	1,000,000,000	Free	0	Tee
10826	Google Drive	PRODUCTIVITY	4.40	2731211	4.00	1,000,000,000	Free	0	Everyor
10832	Google News	NEWS_AND_MAGAZINES	3.90	877635	13.00	1,000,000,000	Free	0	Tee
10839	Subway Surfers	GAME	4.50	27725352	76.00	1,000,000,000	Free	0	Everyone 10

476 rows × 10 columns

```
⬇️ Below are the first 15 duplicated app:
CBS Sports App - Scores, News, Stats & Watch Live     5
ESPN                                                  4
Google Keep                                           3
Nick                                                  3
Bleacher Report: sports news, scores, & highlights    3
WatchESPN                                             3
theScore: Live Sports Scores, News, Stats & Videos    3
Quizlet: Learn Languages & Vocab with Flashcards      3
eBay: Buy & Sell this Summer - Discover Deals Now!    3
Skyscanner                                            3
Udemy - Online Courses                                2
Target - now with Cartwheel                           2
Extreme Coupon Finder                                 2
CNN Breaking US & World News                          2
Expedia Hotels, Flights & Car Rental Travel Deals     2
Name: App, dtype: int64
```

❗ After considering some rows of the duplicated apps we can notice that it have to be specified, which rows should be checked for duplicates:

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Genres
10806	Instagram	SOCIAL	4.50	66577313	5.30	1,000,000,000	Free	0	Teen	Social
10808	Instagram	SOCIAL	4.50	66577446	5.30	1,000,000,000	Free	0	Teen	Social
10809	Instagram	SOCIAL	4.50	66577313	5.30	1,000,000,000	Free	0	Teen	Social
10810	Instagram	SOCIAL	4.50	66509917	5.30	1,000,000,000	Free	0	Teen	Social

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Genres
10806	Instagram	SOCIAL	4.50	66577313	5.30	1,000,000,000	Free	0	Teen	Social

```
✅ One more checking of the data shape: (8199, 10)
```

🚨 Our data observation reveals several important aspects regarding the dataset sourced from the Google Play Store:

1. Sampled Features: The dataset comprises information extracted from 13 different features obtained through web scraping of the Google Play Store.

2. Representativeness: While we assume the sampled data to be representative of the entire Google Play store, it is essential to acknowledge that this assumption may not hold true. The sample was obtained based on geographical location and user behavior, specifically by Lavanya Gupta.

3. Data Compilation: The dataset was compiled around 2017/2018, meaning it may not encompass the most recent information. It's important to note that pricing data reflects the USD Dollar amounts at the time of scraping. Developers have the flexibility to modify app pricing and offer promotions, potentially resulting in changes over time.

4. Size Conversion: To ensure uniformity, the app sizes have been converted into floating-point numbers measured in megabytes (MB). In cases where data was missing, the average size for the respective category was used as a replacement.

5. Installations Reporting: The reported number of installs is not precise. For example, if an app has 245,239 installs, Google may display an approximate range such as 100,000+. In our analysis, we have removed the '+'

symbol and assumed the exact number of installs for simplicity.

Despite these considerations, the dataset still provides valuable insights into the Google Play Store ecosystem and can be leveraged to uncover meaningful patterns and trends. By recognizing the limitations and context surrounding the data, we can draw informed conclusions and make informed decisions based on our analysis.

Data Analysing

The Highest Rated Apps

🔝 Let's identify which apps are the highest rated. We don't rely exclusively on ratings alone to determine the quality of an app, but also on the number of reviews. 🗣️

```
The highest rated app is: Ríos de Fe 🥇

The other top 10 apps 💪 are listed below:
```

	App	Rating	Reviews	Installs
2095	Ríos de Fe	5.00	141	1,000
2438	FD Calculator (EMI, SIP, RD & Loan Eligilibility)	5.00	104	1,000
3115	Oración CX	5.00	103	5,000
2107	Barisal University App-BU Face	5.00	100	1,000
2069	Master E.K	5.00	90	1,000
1968	CL REPL	5.00	47	1,000
790	AJ Cam	5.00	44	100
1275	AI Today : Artificial Intelligence News & AI 101	5.00	43	100
2544	CS & IT Interview Questions	5.00	43	1,000
1789	Ek Vote	5.00	43	500

Top 5 Largest Apps in terms of Size

What's the size in megabytes (MB) of the largest Android apps in the Google Play Store. Based on the data, we can find out, what is the limit on the Google Play store:

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Genre
1795	Navi Radiography Pro	MEDICAL	4.70	11	100.00	500	Paid	$15.99	Everyone	Medic
3144	Vi Trainer	HEALTH_AND_FITNESS	3.60	124	100.00	5,000	Free	0	Everyone	Heal Fitne
4176	Car Crash III Beam DH Real Damage Simulator 2018	GAME	3.60	151	100.00	10,000	Free	0	Everyone	Racir

7926	Post Bank	FINANCE	4.50	60449	100.00	1,000,000	Free	0	Everyone	Financ
7927	The Walking Dead: Our World	GAME	4.00	22435	100.00	1,000,000	Free	0	Teen	Actic

🏋️‍♂️ It is evident from our observations that there exists a noticeable upper limit of 100 MB for the size of apps. This limitation is not only apparent in our dataset but is also corroborated by a simple Google search, which indicates that the Google Play Store imposes this constraint.

The 5 App with Most Reviews

📢 The next metric we can figure out using the data is the highest number of reviews:

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating
10805	Facebook	SOCIAL	4.10	78158306	5.30	1,000,000,000	Free	0	Teen
10785	WhatsApp Messenger	COMMUNICATION	4.40	69119316	3.50	1,000,000,000	Free	0	Everyone
10806	Instagram	SOCIAL	4.50	66577313	5.30	1,000,000,000	Free	0	Teen
10784	Messenger – Text and Video Chat for Free	COMMUNICATION	4.00	56642847	3.50	1,000,000,000	Free	0	Everyone
10650	Clash of Clans	GAME	4.60	44891723	98.00	100,000,000	Free	0	Everyone 10+
10744	Clean Master- Space Cleaner & Antivirus	TOOLS	4.70	42916526	3.40	500,000,000	Free	0	Everyone
10835	Subway Surfers	GAME	4.50	27722264	76.00	1,000,000,000	Free	0	Everyone 10+
10828	YouTube	VIDEO_PLAYERS	4.30	25655305	4.65	1,000,000,000	Free	0	Teen
10746	Security Master - Antivirus, VPN, AppLock, Boo...	TOOLS	4.70	24900999	3.40	500,000,000	Free	0	Everyone
10584	Clash Royale	GAME	4.60	23133508	97.00	100,000,000	Free	0	Everyone 10+
10763	Candy Crush Saga	GAME	4.40	22426677	74.00	500,000,000	Free	0	Everyone
10770	UC Browser - Fast Download Private & Secure	COMMUNICATION	4.50	17712922	40.00	500,000,000	Free	0	Teen
10735	Snapchat	SOCIAL	4.00	17014787	5.30	500,000,000	Free	0	Teen
10489	360	TOOLS	4.60	16771865	3.40	100,000,000	Free	0	Everyone

	Security - Free Antivirus, Booster, Cleaner								
10731	My Talking Tom	GAME	4.50	14891223	36.00	500,000,000	Free	0	Everyone
10594	8 Ball Pool	GAME	4.50	14198297	52.00	100,000,000	Free	0	Everyone
10302	DU Battery Saver - Battery Charger & Battery Life	TOOLS	4.50	13479633	14.00	100,000,000	Free	0	Everyone
10354	BBM - Free Calls & Messages	COMMUNICATION	4.30	12842860	3.50	100,000,000	Free	0	Everyone
10549	Cache Cleaner-DU Speed Booster (booster & clea...	TOOLS	4.50	12759663	15.00	100,000,000	Free	0	Everyone
10757	Twitter	NEWS_AND_MAGAZINES	4.30	11667403	6.30	500,000,000	Free	0	Mature 17+
10721	Viber Messenger	COMMUNICATION	4.30	11334799	3.50	500,000,000	Free	0	Everyone
10578	Shadow Fight 2	GAME	4.60	10979062	88.00	100,000,000	Free	0	Everyone 10+
10813	Google Photos	PHOTOGRAPHY	4.50	10858556	6.90	1,000,000,000	Free	0	Everyone
10724	LINE: Free Calls & Messages	COMMUNICATION	4.20	10790289	3.50	500,000,000	Free	0	Everyone
10717	Pou	GAME	4.30	10485308	24.00	500,000,000	Free	0	Everyone
10792	Skype - free IM & video calls	COMMUNICATION	4.10	10484169	3.50	1,000,000,000	Free	0	Everyone
10628	Pokémon GO	GAME	4.10	10424925	85.00	100,000,000	Free	0	Everyone
10388	Minion Rush: Despicable Me Official Game	GAME	4.50	10216538	36.00	100,000,000	Free	0	Everyone 10+
10694	Yes day	GAME	4.50	10055521	94.00	100,000,000	Free	0	Everyone
10695	Hay Day	FAMILY	4.50	10053186	94.00	100,000,000	Free	0	Everyone
10644	Dream League Soccer 2018	GAME	4.60	9882639	74.00	100,000,000	Free	0	Everyone
10696	My Talking Angela	GAME	4.50	9881829	99.00	100,000,000	Free	0	Everyone
10660	VivaVideo	VIDEO_PLAYERS	4.60	9879473	40.00	100,000,000	Free	0	Teen

	- Video Editor & Photo Movie								
10786	Google Chrome: Fast & Secure	COMMUNICATION	4.30	9642995	3.50	1,000,000,000	Free	0	Everyone
10817	Maps - Navigate & Explore	TRAVEL_AND_LOCAL	4.30	9235155	12.00	1,000,000,000	Free	0	Everyone
10672	Hill Climb Racing	GAME	4.40	8923587	63.00	100,000,000	Free	0	Everyone
10734	Facebook Lite	SOCIAL	4.30	8606259	5.30	500,000,000	Free	0	Teen
10649	Asphalt 8: Airborne	GAME	4.50	8389714	92.00	100,000,000	Free	0	Teen
10699	Mobile Legends: Bang Bang	GAME	4.40	8219586	99.00	100,000,000	Free	0	Teen
10322	Battery Doctor-Battery Life Saver & Battery Co...	TOOLS	4.50	8190074	17.00	100,000,000	Free	0	Everyone
10396	Piano Tiles 2™	GAME	4.70	8118880	36.00	100,000,000	Free	0	Everyone
10777	Temple Run 2	GAME	4.30	8118609	62.00	500,000,000	Free	0	Everyone
10822	Google	TOOLS	4.40	8033493	3.40	1,000,000,000	Free	0	Everyone
10359	Truecaller: Caller ID, SMS spam blocking & Dialer	COMMUNICATION	4.50	7820209	3.50	100,000,000	Free	0	Everyone
10711	SHAREit - Transfer & Share	TOOLS	4.60	7790693	17.00	500,000,000	Free	0	Everyone
10389	Sniper 3D Gun Shooter: Free Shooting Games - FPS	GAME	4.60	7671249	36.00	100,000,000	Free	0	Mature 17+
10676	Farm Heroes Saga	GAME	4.40	7614130	70.00	100,000,000	Free	0	Everyone
10576	PicsArt Photo Studio: Collage Maker & Pic Editor	PHOTOGRAPHY	4.50	7594559	34.00	100,000,000	Free	0	Teen
10461	PhotoGrid: Video &	PHOTOGRAPHY	4.60	7529865	6.90	100,000,000	Free	0	Everyone

	Pic Collage Maker, Photo Ed...								
10502	GO Launcher - 3D parallax Themes & HD Wallpapers	PERSONALIZATION	4.50	7464996	6.15	100,000,000	Free	0	Everyone

🆓 From the output above we can conclude there are only free apps among the top 50.

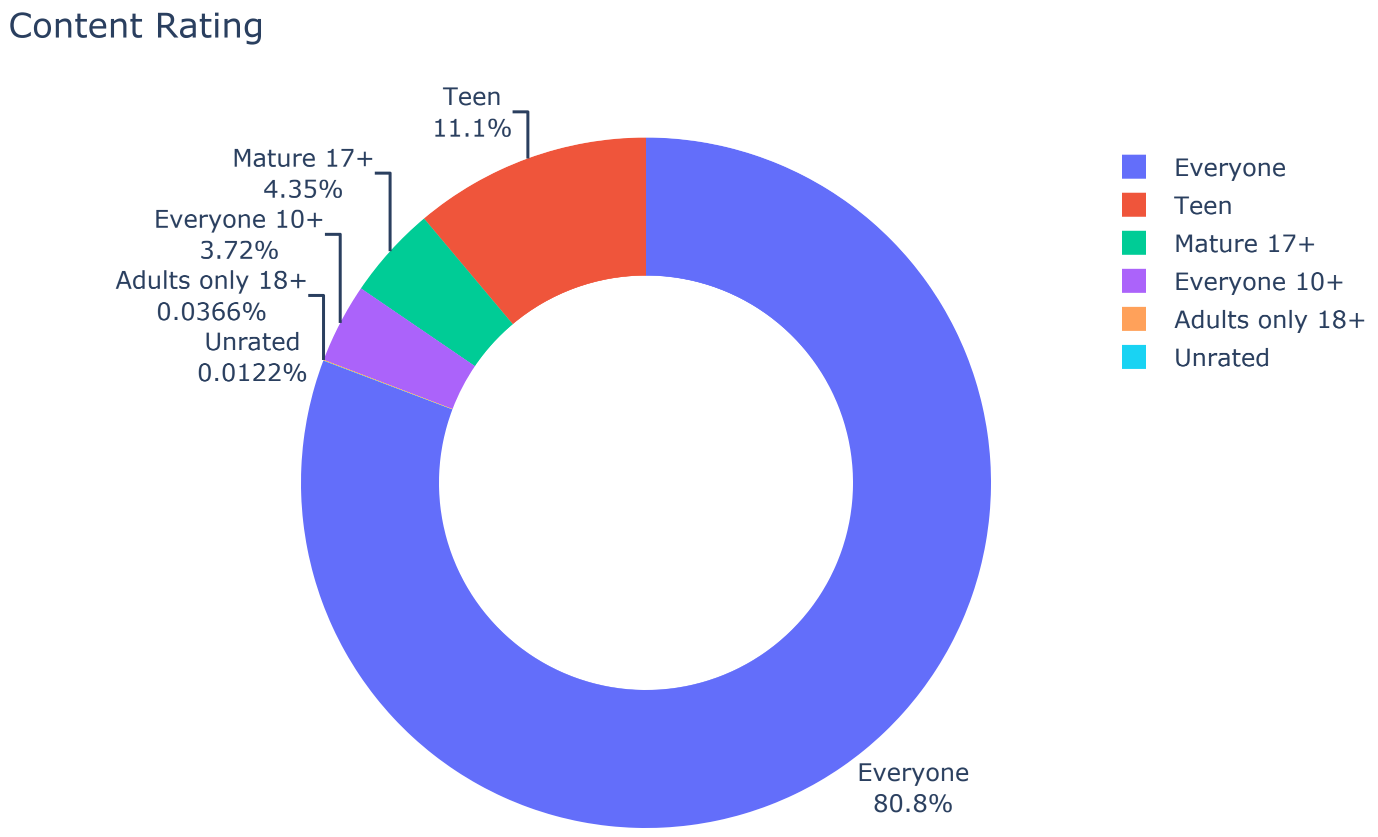
```
Free    50
Name: Type, dtype: int64
```

Visualise Categorical Data: Content Ratings

🕵♂ In our dataset, each Android app is assigned a content rating, such as "Everyone," "Teen," or "Mature 17+".

📊 Now, let's explore the distribution of these content ratings and explore various visualization methods to gain further insights.

```
Everyone           6621
Teen                912
Mature 17+          357
Everyone 10+        305
Adults only 18+       3
Unrated               1
Name: Content_Rating, dtype: int64
```



Examination the Number of Installs

Now let's find how many apps had over 1 BILLION installations 🤯 and how many apps just had a single install 🙈.

Let's start with the checking of the data type and (if needed) the converting to Numeric Type.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8199 entries, 21 to 10835
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             8199 non-null   object 
 1   Category        8199 non-null   object 
 2   Rating          8199 non-null   float64
 3   Reviews         8199 non-null   int64  
 4   Size_MBs        8199 non-null   float64
 5   Installs        8199 non-null   object 
 6   Type            8199 non-null   object 
 7   Price           8199 non-null   object 
 8   Content_Rating  8199 non-null   object 
 9   Genres          8199 non-null   object 
dtypes: float64(2), int64(1), object(7)
memory usage: 704.6+ KB
```

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	
10731	My Talking Tom	GAME	4.50	14891223	36.00	500,000,000	Free	0	Everyone	
10746	Security Master - Antivirus, VPN, AppLock, Boo...	TOOLS	4.70	24900999	3.40	500,000,000	Free	0	Everyone	
10711	SHAREit - Transfer & Share	TOOLS	4.60	7790693	17.00	500,000,000	Free	0	Everyone	
10713	imo free video calls and chat	COMMUNICATION	4.30	4785892	11.00	500,000,000	Free	0	Everyone	Commun
10717	Pou	GAME	4.30	10485308	24.00	500,000,000	Free	0	Everyone	
...	...	...	...	...	...	...	...	...	...	
2403	CI Screwed - Icon Pack	PERSONALIZATION	4.70	19	6.40	1,000	Free	0	Everyone	Persona
2402	EG Movi	TOOLS	4.20	40	7.40	1,000	Free	0	Everyone	
28	Ra Ga Ba	GAME	5.00	2	20.00	1	Paid	$1.49	Everyone	
47	Mu.F.O.	GAME	5.00	2	16.00	1	Paid	$0.99	Everyone	
21	KBA-EZ	MEDICAL	5.00	4	25.00	1	Free	0	Everyone	I

Health
Guide

8199 rows × 10 columns

After the observation the `Installs` column we can notice, that the values are grouped or rounded to the nice round numbers.

	App
Installs	
1	3
1,000	698
1,000,000	1417
1,000,000,000	20
10	69
10,000	988
10,000,000	933
100	303
100,000	1096
100,000,000	189
5	9
5,000	425
5,000,000	607
50	56
50,000	457
50,000,000	202
500	199
500,000	504
500,000,000	24

🧮 Let's process and convert the numbers to be able to estimate the numeric data

```
App                object
Category           object
Rating            float64
Reviews             int64
Size_MBs          float64
Installs            int64
Type               object
Price              object
Content_Rating     object
Genres             object
dtype: object
```

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating
10835	Subway	GAME	4.50	27722264	76.00	1000000000	Free	0	Everyone 10+

	Surfers									
10806	Instagram	SOCIAL	4.50	66577313	5.30	1000000000	Free	0	Teen
10783	Google Play Books	BOOKS_AND_REFERENCE	3.90	1433233	5.70	1000000000	Free	0	Teen
10784	Messenger – Text and Video Chat for Free	COMMUNICATION	4.00	56642847	3.50	1000000000	Free	0	Everyone
10785	WhatsApp Messenger	COMMUNICATION	4.40	69119316	3.50	1000000000	Free	0	Everyone
...	...	...	...	...	...	...	...	...	...
99	Anatomy & Physiology Vocabulary Exam Review App	MEDICAL	5.00	1	4.60	5	Free	0	Everyone
82	Brick Breaker BR	GAME	5.00	7	19.00	5	Free	0	Everyone
47	Mu.F.O.	GAME	5.00	2	16.00	1	Paid	$0.99	Everyone
28	Ra Ga Ba	GAME	5.00	2	20.00	1	Paid	$1.49	Everyone
21	KBA-EZ Health Guide	MEDICAL	5.00	4	25.00	1	Free	0	Everyone

8199 rows × 10 columns

```
🤯 Apps with 1B and more installations: (20, 10)
```

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating
10835	Subway Surfers	GAME	4.50	27722264	76.00	1000000000	Free	0	Everyone 10+
10806	Instagram	SOCIAL	4.50	66577313	5.30	1000000000	Free	0	Teen
10783	Google Play Books	BOOKS_AND_REFERENCE	3.90	1433233	5.70	1000000000	Free	0	Teen
10784	Messenger – Text and Video Chat for Free	COMMUNICATION	4.00	56642847	3.50	1000000000	Free	0	Everyone
10785	WhatsApp Messenger	COMMUNICATION	4.40	69119316	3.50	1000000000	Free	0	Everyone
10786	Google Chrome: Fast & Secure	COMMUNICATION	4.30	9642995	3.50	1000000000	Free	0	Everyone
10788	Hangouts	COMMUNICATION	4.00	3419249	3.50	1000000000	Free	0	Everyone
10792	Skype - free IM & video calls	COMMUNICATION	4.10	10484169	3.50	1000000000	Free	0	Everyone
10803	Google Play Games	ENTERTAINMENT	4.30	7165362	9.35	1000000000	Free	0	Teen
10805	Facebook	SOCIAL	4.10	78158306	5.30	1000000000	Free	0	Teen

10787	Gmail	COMMUNICATION	4.30	4604324	3.50	1000000000	Free	0	Everyone
10807	Google+	SOCIAL	4.20	4831125	5.30	1000000000	Free	0	Teen
10817	Maps - Navigate & Explore	TRAVEL_AND_LOCAL	4.30	9235155	12.00	1000000000	Free	0	Everyone
10818	Google Street View	TRAVEL_AND_LOCAL	4.20	2129689	12.00	1000000000	Free	0	Everyone
10822	Google	TOOLS	4.40	8033493	3.40	1000000000	Free	0	Everyone
10824	Google Drive	PRODUCTIVITY	4.40	2731171	4.00	1000000000	Free	0	Everyone
10828	YouTube	VIDEO_PLAYERS	4.30	25655305	4.65	1000000000	Free	0	Teen
10829	Google Play Movies & TV	VIDEO_PLAYERS	3.70	906384	4.65	1000000000	Free	0	Teen
10831	Google News	NEWS_AND_MAGAZINES	3.90	877635	13.00	1000000000	Free	0	Teen
10813	Google Photos	PHOTOGRAPHY	4.50	10858556	6.90	1000000000	Free	0	Everyone

The Most Expensive Apps, Filter out the Junk, and Calculate a Sales Revenue Estimate

🧐 Let's examine the Price column more closely.

We have to convert the price column to numeric data again.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8199 entries, 21 to 10835
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             8199 non-null   object 
 1   Category        8199 non-null   object 
 2   Rating          8199 non-null   float64
 3   Reviews         8199 non-null   int64  
 4   Size_MBs        8199 non-null   float64
 5   Installs        8199 non-null   int64  
 6   Type            8199 non-null   object 
 7   Price           8199 non-null   float64
 8   Content_Rating  8199 non-null   object 
 9   Genres          8199 non-null   object 
dtypes: float64(3), int64(2), object(5)
memory usage: 704.6+ KB
```

The most expensive apps sub $250

Then we investigate the top 20 most expensive apps in the dataset. It seems the dataset has some junk data with the price over 250$.

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Genres

3946	I'm Rich - Trump Edition	LIFESTYLE	3.60	275	7.30	10000	Paid	400.00	Everyone	Lifestyle
2461	I AM RICH PRO PLUS	FINANCE	4.00	36	41.00	1000	Paid	399.99	Everyone	Finance
4606	I Am Rich Premium	FINANCE	4.10	1867	4.70	50000	Paid	399.99	Everyone	Finance
3145	I am rich(premium)	FINANCE	3.50	472	0.94	5000	Paid	399.99	Everyone	Finance
3554	💎 I'm rich	LIFESTYLE	3.80	718	26.00	10000	Paid	399.99	Everyone	Lifestyle
5765	I am rich	LIFESTYLE	3.80	3547	1.80	100000	Paid	399.99	Everyone	Lifestyle
1946	I am rich (Most expensive app)	FINANCE	4.10	129	2.70	1000	Paid	399.99	Teen	Finance
2775	I Am Rich Pro	FAMILY	4.40	201	2.70	5000	Paid	399.99	Everyone	Entertainment
3221	I am Rich Plus	FAMILY	4.00	856	8.70	10000	Paid	399.99	Everyone	Entertainment
3114	I am Rich	FINANCE	4.30	180	3.80	5000	Paid	399.99	Everyone	Finance
1331	most expensive app (H)	FAMILY	4.30	6	1.50	100	Paid	399.99	Everyone	Entertainment
2394	I am Rich!	FINANCE	3.80	93	22.00	1000	Paid	399.99	Everyone	Finance
3897	I Am Rich	FAMILY	3.60	217	4.90	10000	Paid	389.99	Everyone	Entertainment
2193	I am extremely Rich	LIFESTYLE	2.90	41	2.90	1000	Paid	379.99	Everyone	Lifestyle
3856	I am rich VIP	LIFESTYLE	3.80	411	2.60	10000	Paid	299.99	Everyone	Lifestyle
2281	Vargo Anesthesia Mega App	MEDICAL	4.60	92	32.00	1000	Paid	79.99	Everyone	Medical
1407	LTC AS Legal	MEDICAL	4.00	6	1.30	100	Paid	39.99	Everyone	Medical
2629	I am Rich Person	LIFESTYLE	4.20	134	1.80	1000	Paid	37.99	Everyone	Lifestyle
2481	A Manual of Acupuncture	MEDICAL	3.50	214	68.00	1000	Paid	33.99	Everyone	Medical
4264	Golfshot Plus: Golf GPS	SPORTS	4.10	3387	25.00	50000	Paid	29.99	Everyone	Sports

	App	Category	Rating	Reviews	Size_MBs	Installs	Type	Price	Content_Rating	Genres
2281	Vargo Anesthesia Mega App	MEDICAL	4.60	92	32.00	1000	Paid	79.99	Everyone	Medical
1407	LTC AS Legal	MEDICAL	4.00	6	1.30	100	Paid	39.99	Everyone	Medical
2629	I am Rich Person	LIFESTYLE	4.20	134	1.80	1000	Paid	37.99	Everyone	Lifestyle
2481	A Manual of Acupuncture	MEDICAL	3.50	214	68.00	1000	Paid	33.99	Everyone	Medical
2463	PTA Content Master	MEDICAL	4.20	64	41.00	1000	Paid	29.99	Everyone	Medical

2207	EMT PASS	MEDICAL	3.40	51	2.40	1000	Paid	29.99	Everyone	Medical
4264	Golfshot Plus: Golf GPS	SPORTS	4.10	3387	25.00	50000	Paid	29.99	Everyone	Sports
504	AP Art History Flashcards	FAMILY	5.00	1	96.00	10	Paid	29.99	Mature 17+	Education
4772	Human Anatomy Atlas 2018: Complete 3D Human Body	MEDICAL	4.50	2921	25.00	100000	Paid	24.99	Everyone	Medical
3241	Muscle Premium - Human Anatomy, Kinesiology, B...	MEDICAL	4.20	168	25.00	10000	Paid	24.99	Everyone	Medical
2119	NewTek NDI	PHOTOGRAPHY	3.50	77	1.20	1000	Paid	19.99	Everyone	Photography
4470	DRAGON QUEST VIII	FAMILY	4.50	7812	27.00	50000	Paid	19.99	Everyone 10+	Role Playing
2293	Hospitalist Handbook	MEDICAL	4.80	12	18.00	1000	Paid	19.99	Everyone	Medical
526	USMLE Step 2 CK Flashcards	FAMILY	5.00	1	40.00	10	Paid	19.99	Everyone	Education
2473	boattheory.ch Full 2018	FAMILY	4.70	54	50.00	1000	Paid	19.40	Everyone	Education
4090	I am Rich Premium Plus	FINANCE	4.60	459	2.00	10000	Paid	18.99	Everyone	Finance
1508	SkyTest BU/GU Lite	BUSINESS	2.90	28	20.00	500	Paid	17.99	Everyone	Business
3778	The World Ends With You	GAME	4.60	4108	13.00	10000	Paid	17.99	Everyone 10+	Arcade
2603	2017 EMRA Antibiotic Guide	MEDICAL	4.40	12	3.80	1000	Paid	16.99	Everyone	Medical
3439	Trine 2: Complete Story	GAME	3.80	252	11.00	10000	Paid	16.99	Teen	Action

Highest Grossing Paid Apps (ballpark estimate)

💰 Let's add a column called 'Revenue_Estimate' to the DataFrame. This column should hold the price of the app times the number of installs. What are the top 10 highest grossing paid apps according to this estimate? Out of the top 10 highest grossing paid apps, how many are games?

	App	Category	Revenue_Estimate
9220	Minecraft	FAMILY	69,900,000.00
8825	Hitman Sniper	GAME	9,900,000.00
7151	Grand Theft Auto: San Andreas	GAME	6,990,000.00

7477	Facetune - For Free	PHOTOGRAPHY	5,990,000.00
7977	Sleep as Android Unlock	LIFESTYLE	5,990,000.00
6594	DraStic DS Emulator	GAME	4,990,000.00
6082	Weather Live	WEATHER	2,995,000.00
7954	Bloons TD 5	FAMILY	2,990,000.00
7633	Five Nights at Freddy's	GAME	2,990,000.00
6746	Card Wars - Adventure Time	FAMILY	2,990,000.00

Bar Charts & Scatter Plots: Analysing App Categories

🤔 When choosing which app category you want to release, should you go for a competitive one or a popular one with many downloads? Alternatively, you can aim for a category that balances popularity and wider app distribution.

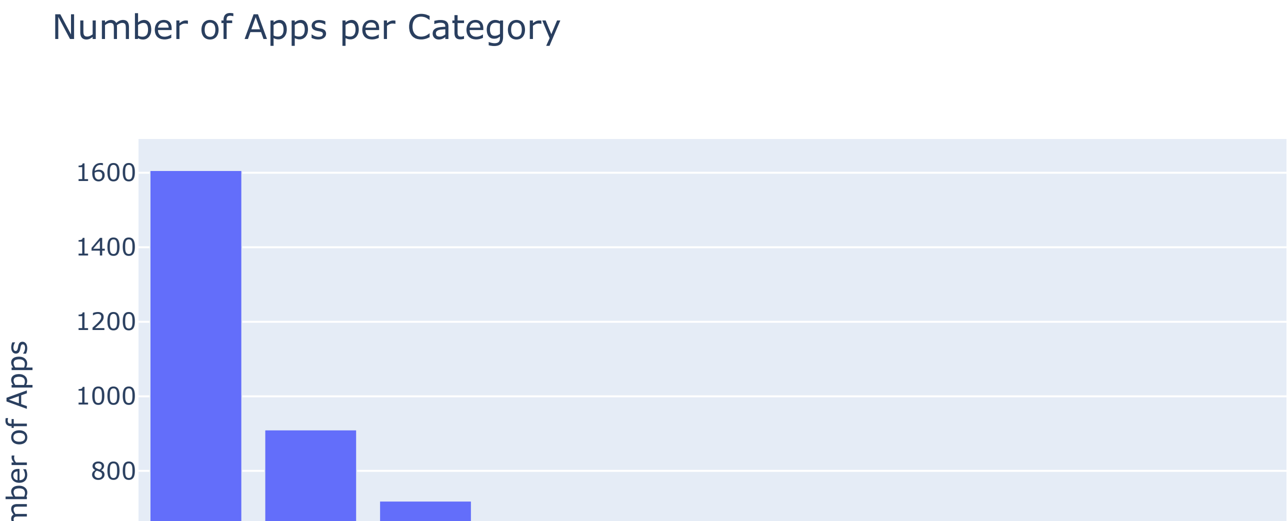
💹 Analyzing this using bar charts and scatter plots can reveal the dominant market categories.

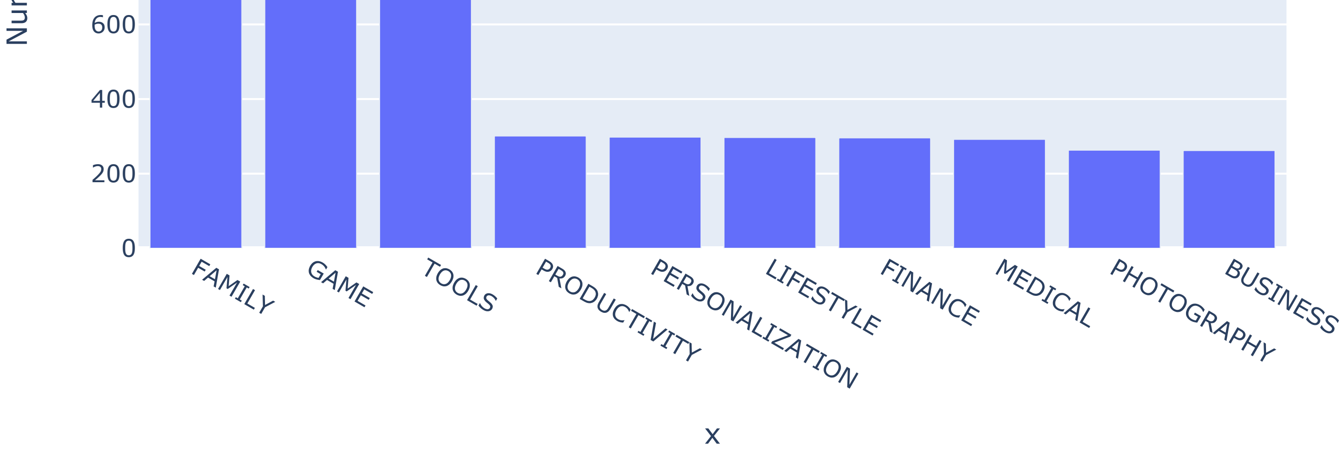
```
['MEDICAL' 'GAME' 'SPORTS' 'BUSINESS' 'BOOKS_AND_REFERENCE' 'SOCIAL'
 'TOOLS' 'FAMILY' 'COMMUNICATION' 'PRODUCTIVITY' 'LIFESTYLE' 'DATING'
 'EVENTS' 'MAPS_AND_NAVIGATION' 'SHOPPING' 'PERSONALIZATION' 'PARENTING'
 'PHOTOGRAPHY' 'HEALTH_AND_FITNESS' 'FOOD_AND_DRINK' 'NEWS_AND_MAGAZINES'
 'FINANCE' 'TRAVEL_AND_LOCAL' 'AUTO_AND_VEHICLES' 'ART_AND_DESIGN'
 'BEAUTY' 'VIDEO_PLAYERS' 'COMICS' 'WEATHER' 'HOUSE_AND_HOME'
 'LIBRARIES_AND_DEMO' 'EDUCATION' 'ENTERTAINMENT']
```

```
➡️ Total Number of Categories: 33
```

```
FAMILY             1606
GAME                910
TOOLS               719
PRODUCTIVITY        301
PERSONALIZATION     298
LIFESTYLE           297
FINANCE             296
MEDICAL             292
PHOTOGRAPHY         263
BUSINESS            262
Name: Category, dtype: int64
```

Highest Competition (Number of Apps)

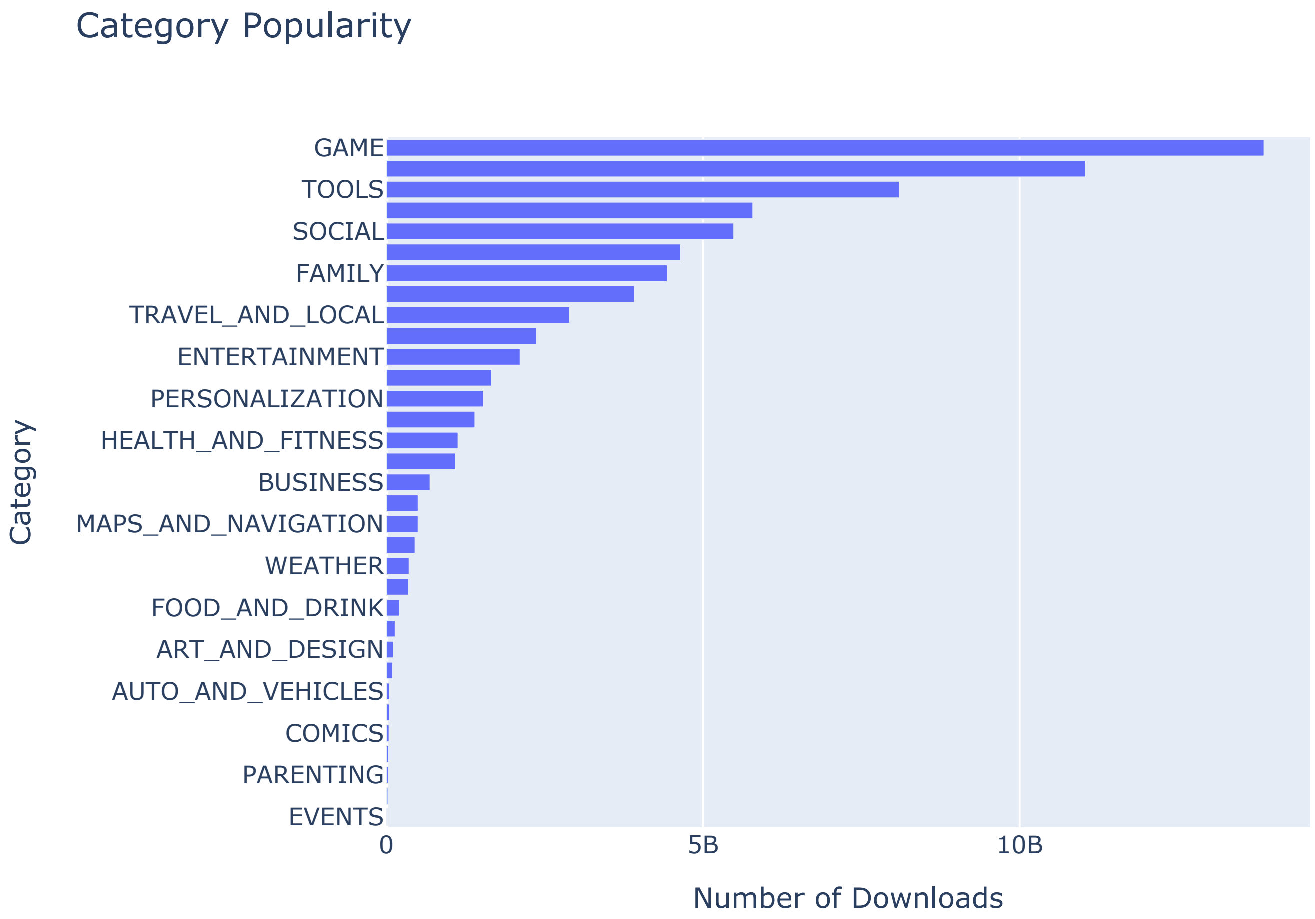




Most Popular Categories (Highest Downloads)

	Installs
Category	
EVENTS	15949410
BEAUTY	26916200
PARENTING	31116110
MEDICAL	39162676
COMICS	44931100
LIBRARIES_AND_DEMO	52083000
AUTO_AND_VEHICLES	53129800
HOUSE_AND_HOME	97082000
ART_AND_DESIGN	114233100
DATING	140912410
FOOD_AND_DRINK	211677750
EDUCATION	352852000
WEATHER	361096500
FINANCE	455249400
MAPS_AND_NAVIGATION	503267560
LIFESTYLE	503611120
BUSINESS	692018120
SPORTS	1096431465
HEALTH_AND_FITNESS	1134006220
SHOPPING	1400331540
PERSONALIZATION	1532352930
BOOKS_AND_REFERENCE	1665791655
ENTERTAINMENT	2113660000
NEWS_AND_MAGAZINES	2369110650
TRAVEL_AND_LOCAL	2894859300

VIDEO_PLAYERS	3916897200
FAMILY	4437554490
PHOTOGRAPHY	4649143130
SOCIAL	5487841475
PRODUCTIVITY	5788070180
TOOLS	8099724500
COMMUNICATION	11039241530
GAME	13858762717

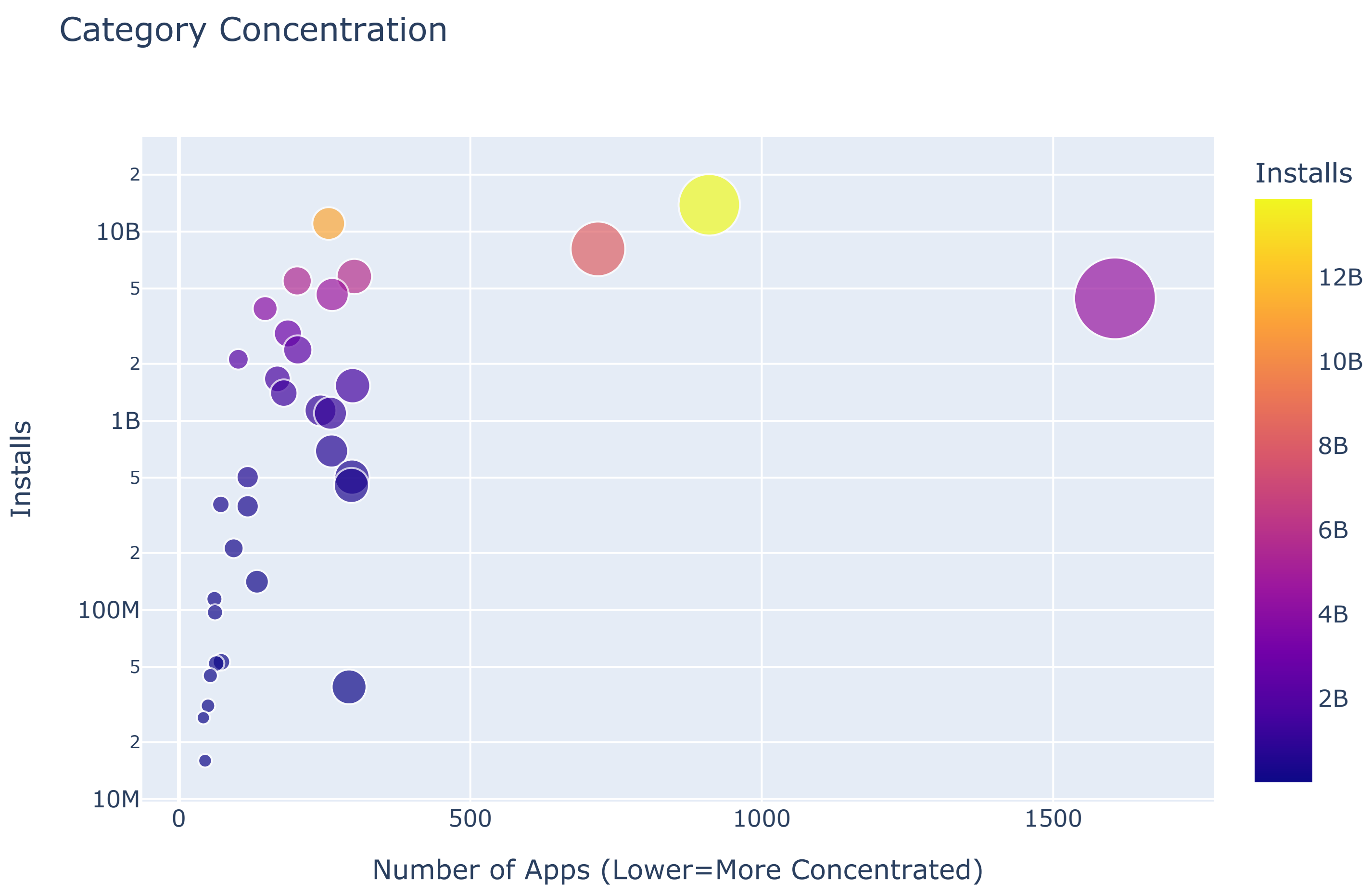


Now we see that 🎮 Games and 🔧 Tools are actually the most popular categories. If we plot the popularity of a category next to the number of apps in that category we can get an idea of how concentrated a category is. Do few apps have most of the downloads or are the downloads spread out over many apps?

Category Concentration - Downloads vs. Competition

	App	**Installs**
Category		
GAME	910	13858762717
COMMUNICATION	257	11039241530
TOOLS	719	8099724500

PRODUCTIVITY	301	5788070180
SOCIAL	203	5487841475
PHOTOGRAPHY	263	4649143130
FAMILY	1606	4437554490
VIDEO_PLAYERS	148	3916897200
TRAVEL_AND_LOCAL	187	2894859300
NEWS_AND_MAGAZINES	204	2369110650



What we see is that the categories like Family, Tools, and Game have many different apps sharing a high number of downloads. But for the categories like video players and entertainment, all the downloads are concentrated in very few apps.

Extracting Nested Data from a Column

🆎 The next step could be investigating how many different types of genres are there. Here below we can see that an app can belong to more than one genre (i.e. "Adventure;Brain Games", "Lifestyle;Pretend Play"). Let's re-structure the dataset.

```
Tools                                     718
Entertainment                             467
Education                                 429
Productivity                              301
```

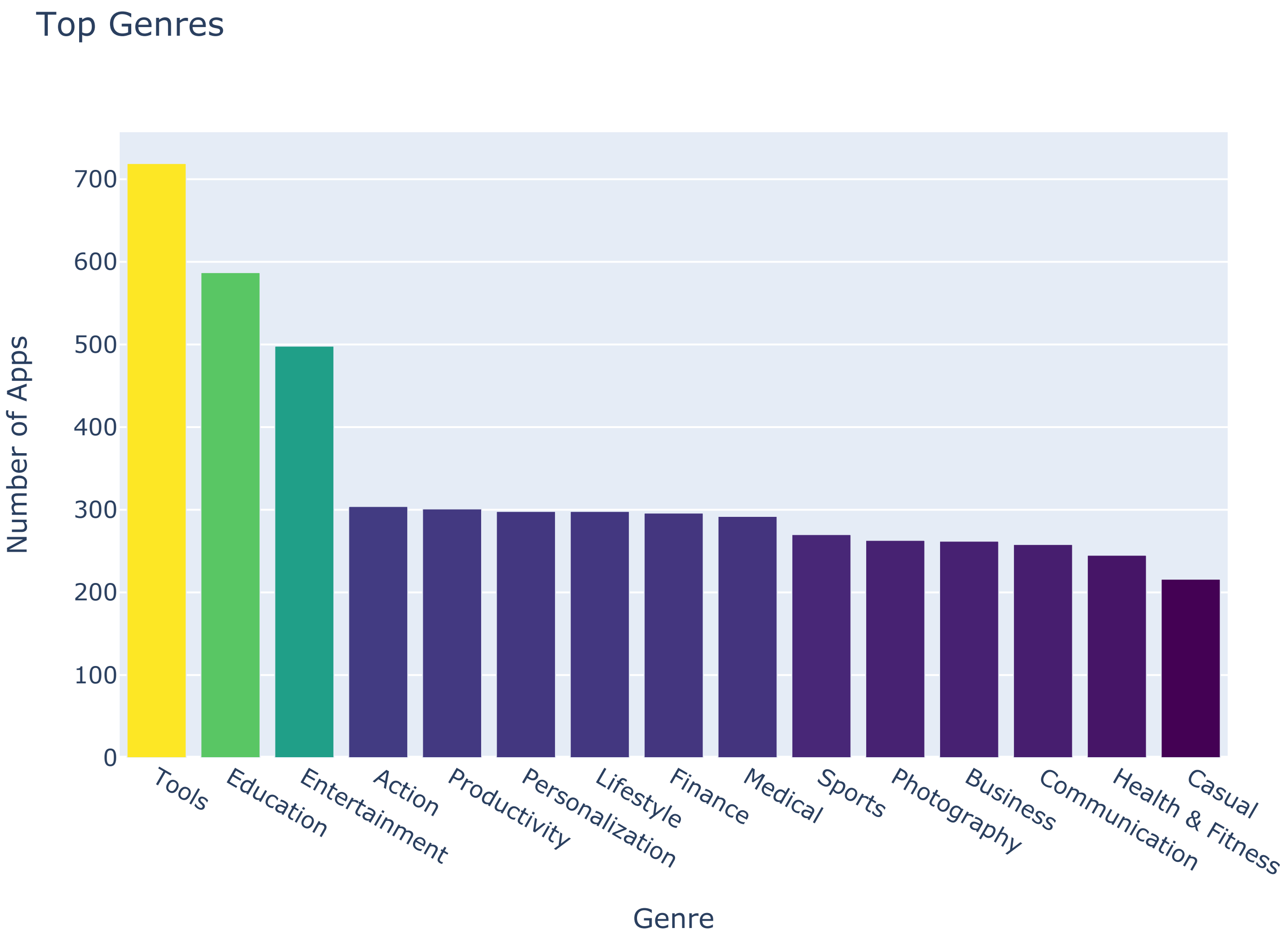
```
Personalization                        298
                                      ...
Adventure;Brain Games                    1
Travel & Local;Action & Adventure        1
Art & Design;Pretend Play                1
Music & Audio;Music & Video              1
Lifestyle;Pretend Play                   1
Name: Genres, Length: 114, dtype: int64

We now have a single column with shape: (8564,)
Number of genres: 53

Index(['Tools', 'Education', 'Entertainment', 'Action', 'Productivity',
       'Personalization', 'Lifestyle', 'Finance', 'Medical', 'Sports',
       'Photography', 'Business', 'Communication', 'Health & Fitness',
       'Casual', 'News & Magazines', 'Social', 'Simulation', 'Travel & Local',
       'Arcade', 'Shopping', 'Books & Reference', 'Video Players & Editors',
       'Dating', 'Puzzle', 'Maps & Navigation', 'Role Playing', 'Racing',
       'Action & Adventure', 'Strategy', 'Food & Drink', 'Educational',
       'Adventure', 'Auto & Vehicles', 'Weather', 'Pretend Play',
       'Brain Games', 'Libraries & Demo', 'Art & Design', 'House & Home',
       'Board', 'Comics', 'Parenting', 'Card', 'Events', 'Beauty', 'Casino',
       'Music & Video', 'Creativity', 'Trivia', 'Word', 'Music',
       'Music & Audio'],
      dtype='object')
```

Colour Scales in Plotly Charts - Competition in Genres

Now let's create a chart with the Series containing the genre data:

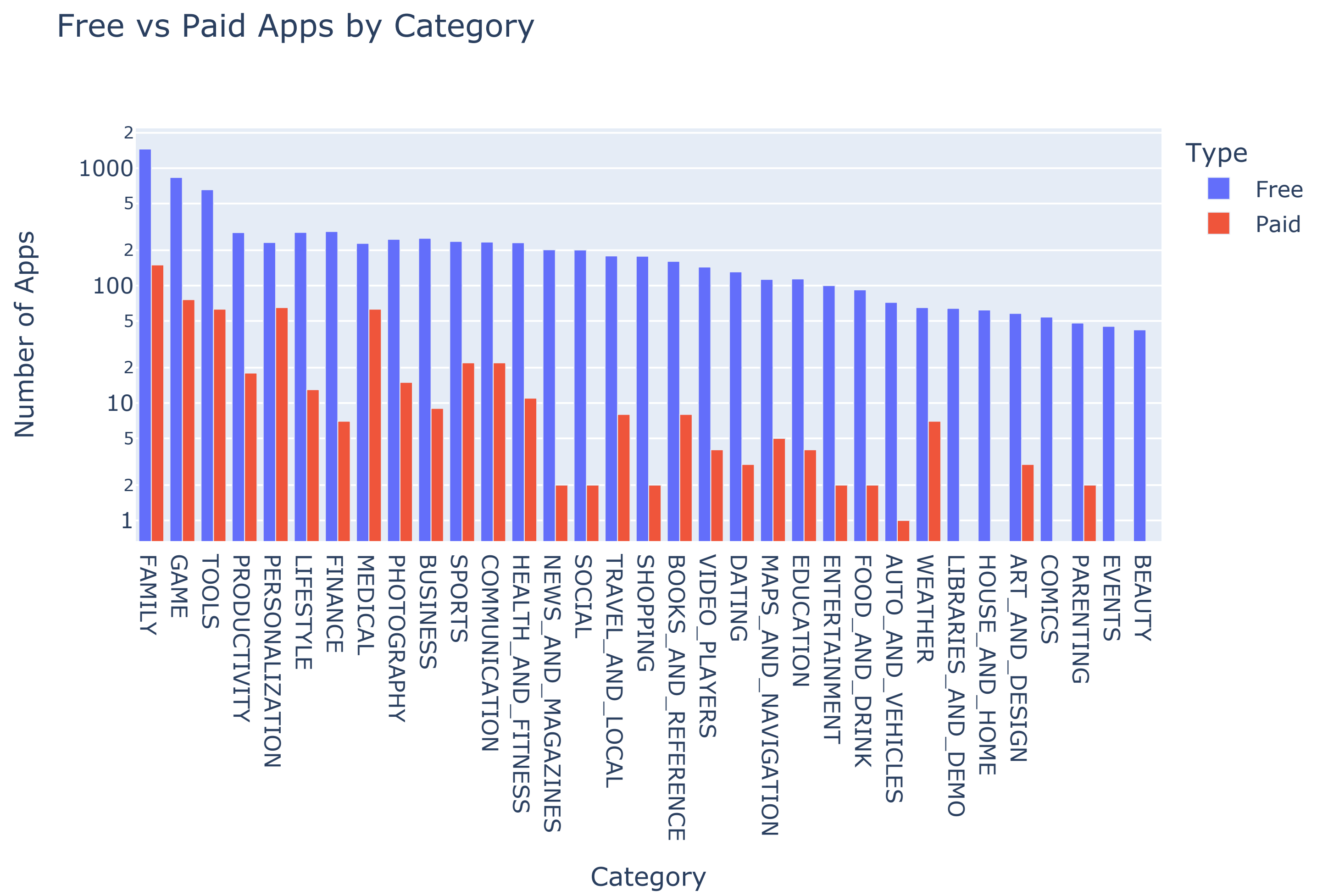


Grouped Bar Charts: Free vs. Paid Apps per Category

💸 Now that we've looked at the total number of apps per category and the total number of apps per genre, let's see what the split is between free and paid apps.

```
Free    7595
Paid     589
Name: Type, dtype: int64
```

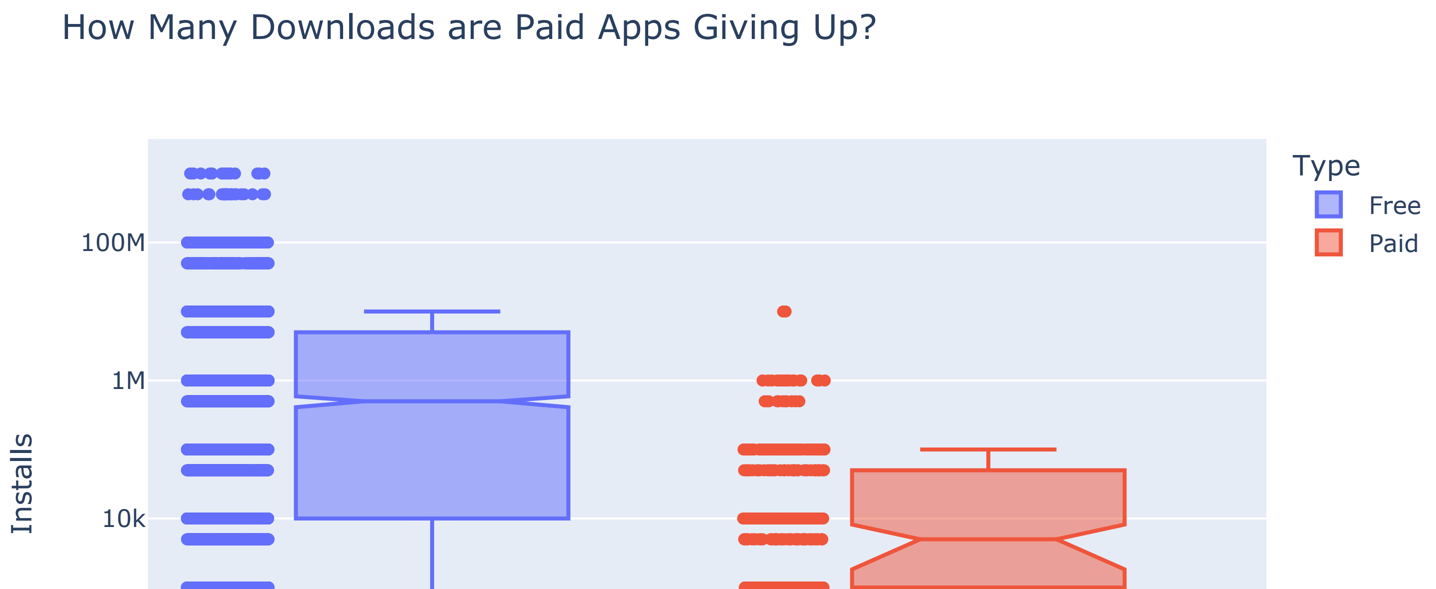
	Category	Type	App
0	ART_AND_DESIGN	Free	58
1	ART_AND_DESIGN	Paid	3
2	AUTO_AND_VEHICLES	Free	72
3	AUTO_AND_VEHICLES	Paid	1
4	BEAUTY	Free	42
5	BOOKS_AND_REFERENCE	Free	161
6	BOOKS_AND_REFERENCE	Paid	8
7	BUSINESS	Free	253
8	BUSINESS	Paid	9
9	COMICS	Free	54
10	COMMUNICATION	Free	235
11	COMMUNICATION	Paid	22
12	DATING	Free	131
13	DATING	Paid	3
14	EDUCATION	Free	114
15	EDUCATION	Paid	4
16	ENTERTAINMENT	Free	100
17	ENTERTAINMENT	Paid	2
18	EVENTS	Free	45
19	FAMILY	Free	1456
20	FAMILY	Paid	150
21	FINANCE	Free	289
22	FINANCE	Paid	7
23	FOOD_AND_DRINK	Free	92
24	FOOD_AND_DRINK	Paid	2
25	GAME	Free	834
26	GAME	Paid	76
27	HEALTH_AND_FITNESS	Free	232
28	HEALTH_AND_FITNESS	Paid	11
29	HOUSE_AND_HOME	Free	62

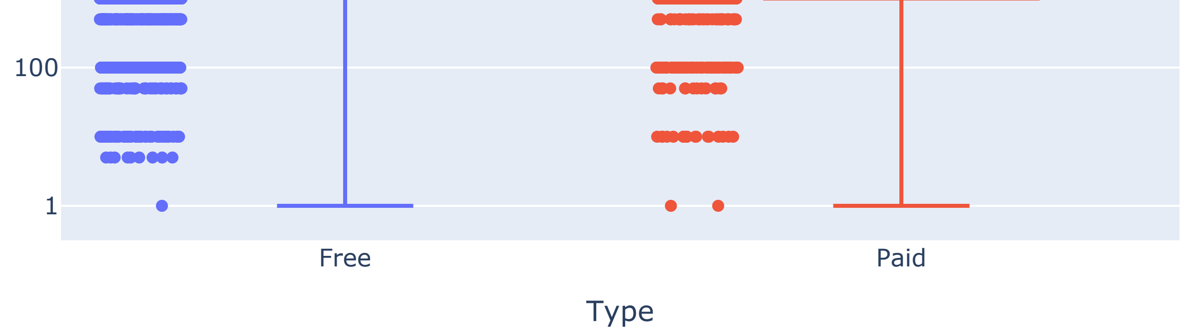


The data reveals that the Google Play Store has a limited number of paid apps. However, certain categories, such as Personalization, Medical, and Weather, have a comparatively higher proportion of paid apps. Considering the category you intend to target, it could be advantageous to release a paid app in those categories.

Lost Downloads for Paid Apps

Now let's plot some box plots with the number of Installs for free versus paid apps, and let's compare median number of installations.

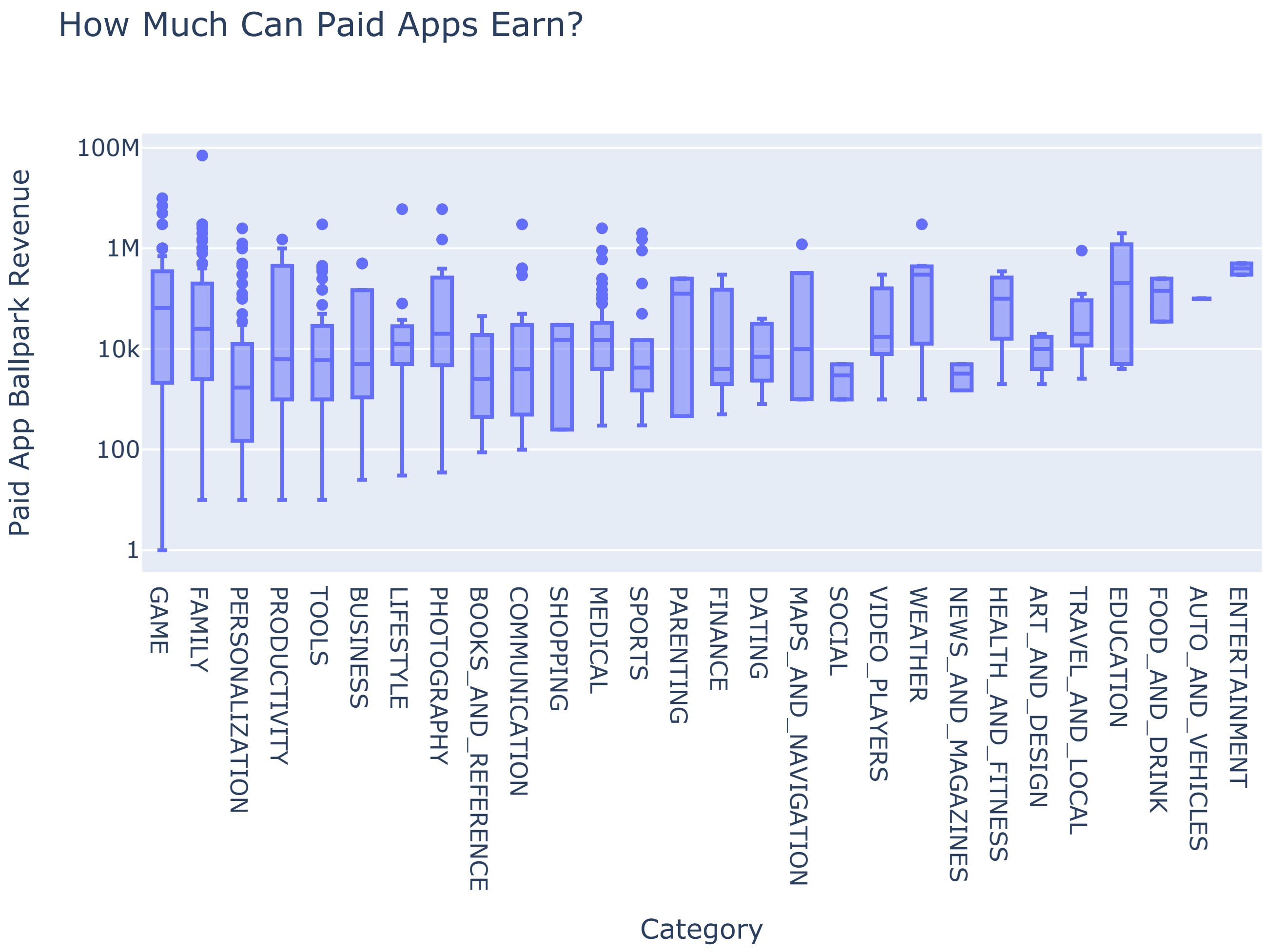




Based on the hover text in the chart, we observe a substantial difference in the median number of downloads between free and paid apps. Free apps have a median of 500,000 downloads, whereas paid apps have a significantly lower median of approximately 5,000 downloads.

Revenue by App Category

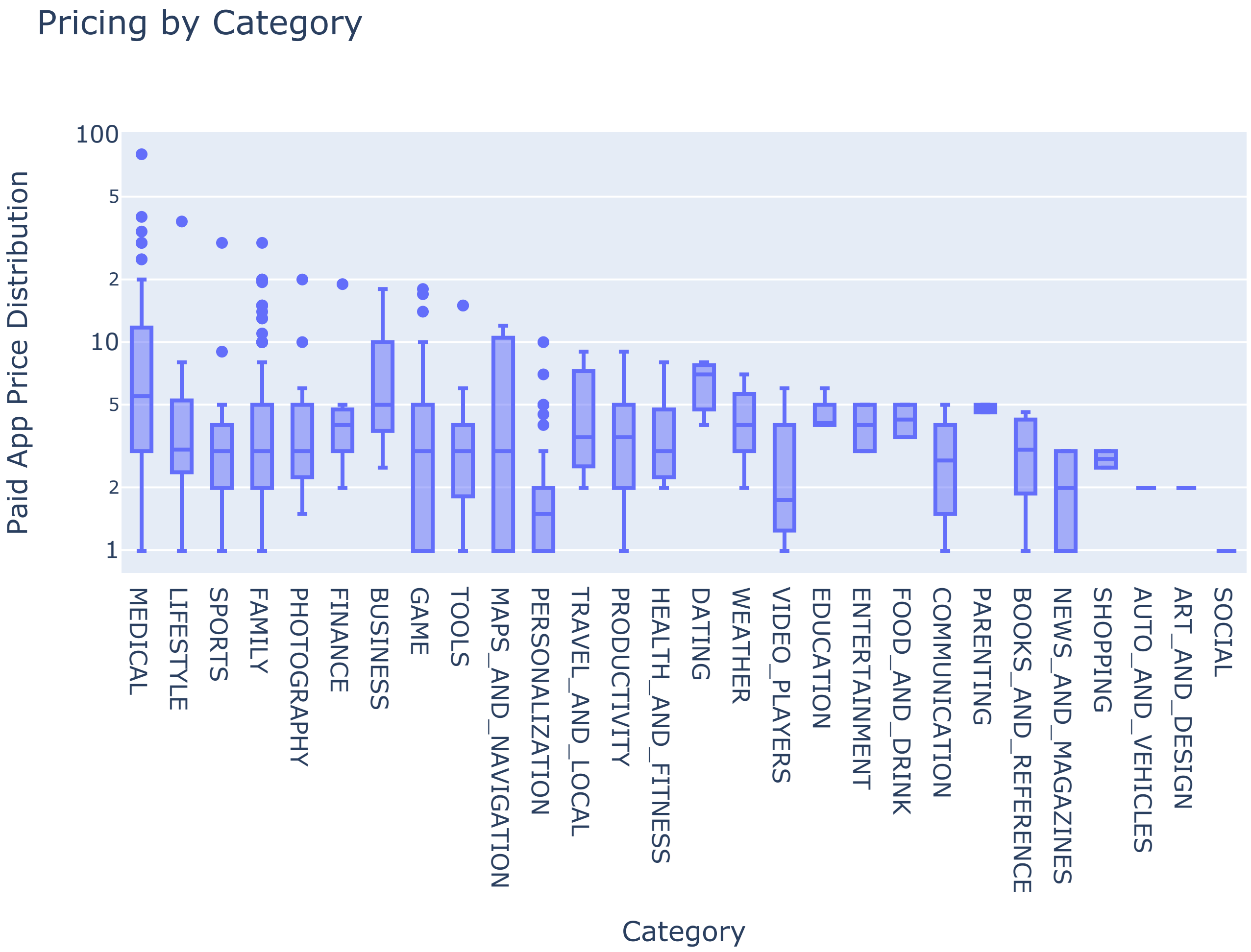
If we want to release a paid app, we need to understand how much does the median app earn in the different categories.



Assuming the development cost of 30,000 USD for an Android app, only a limited number of app categories have an average revenue that can cover this expense. The median revenue for paid photography apps is around 20,000 USD, indicating that many other app categories generate even lower revenues. To compensate for the development costs, these apps would require additional revenue sources such as advertising or in-app purchases. However, certain app categories stand out with a significant number of outliers that exhibit much higher estimated revenue. Examples include Medical, Personalization, Tools, Games, and Family categories, which display a notable presence of apps with substantial revenue potential.

How Much Can We Charge - Paid App Pricing Strategies by Category

💵 Let's investigate what is the median price price for a paid app.



Among the different categories, there are variations in median prices. Notably, Medical apps stand out with the highest median price of 5.49 USD, indicating that they tend to be more expensive. On the other hand, Personalization apps have a comparatively lower average price of 1.49 USD. Additional categories with higher median prices include Business (4.99 USD) and Dating (6.99 USD). These findings suggest that customers in these categories are relatively less hesitant to pay a slightly higher price for their apps, emphasizing their willingness to invest in quality and specific functionalities.