UC San Diego

Google Play Store App Popularity Analysis & Prediction

ECE 143 (FA25) Project: Group 5

Abdumannon Yovkochov, Aditi Pagey, Ke Liu, Shaurya Chopra, Zijun Zhang

December 5, 2025

Agenda

1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

Agenda

1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

About the Dataset

- L. Gupta, "Google Play Store Apps," Feb 2019. [Online] https://www.kaggle.com/lava18/google-play-store-apps
- `googleplaystore.csv`:
    - list of apps, with various features - category, rating, #reviews, size, price, content rating, last updated, android version, and #installs
    - #installs is our 'target' variable
- `googleplaystore_user_reviews.csv`:
    - user review texts, sentiment scores, sentiment subjectivity scores

Why care about no. of installs?

- How much money an app makes is directly related to the number of installs
- Marketers wish to know what makes an app ‘successful’
  - is it better to make an app free or paid?
  - does having 5 stars make an app successful?
  - whom should you target your app to?
  - does it matter how many reviews an app has? what kind of reviews?
  - …

What does our dataset look like?

App	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Last Updated	Current Ver	Android Ver
Photo Editor & Candy Camera & Gri	ART_AND_DESIGN	4.1	159	19M	10,000	0	Everyone	January 7, 2018	1.0.0	4.0.3 and up
3D Color Pixel by Number - Sandbox	ART_AND_DESIGN	4.4	1518	37M	100,000	0	Everyone	August 3, 2018	1.2.3	2.3 and up
Used cars for sale - Trovit	AUTO_AND_VEHICLES	4.2	52530	7.0M	5,000,000	0	Everyone	July 16, 2018	4.47.3	4.0.3 and up
Fines of the State Traffic Safety Insp	AUTO_AND_VEHICLES	4.8	116986	35M	5,000,000	0	Everyone	August 2, 2018	1.9.7	4.0.3 and up
SK Enca Direct Malls - Used Cars Se	AUTO_AND_VEHICLES	3.6	1379	16M	500,000	0	Everyone	August 2, 2018	2.2.21	4.2 and up
Selfie Camera Photo Editor & Filter &	BEAUTY	4.1	187	30M	50,000	0	Teen	July 24, 2018	3.0.1	4.0.3 and up
Eyes Makeup Beauty Tips	BEAUTY	4.2	30	2.9M	10,000	0	Everyone	April 9, 2018	3.3.9	4.0.3 and up

App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
10 Best Foods for You	Works great especially going grocery store	Positive	0.4	0.875
10 Best Foods for You	Best idea us	Positive	1.0	0.3
10 Best Foods for You	Best way	Positive	1.0	0.3
1800 Contacts - Lens S	Great hassle free way order contacts. Got call remin	Positive	0.6000000000000000	0.775
1800 Contacts - Lens S	It's expensive I expected (I thought I'd saving money	Negative	-0.3	0.5
1800 Contacts - Lens S	Super fast navigation brand fast/easy checkout add	Positive	0.26666666666666700	0.6333333333333330

Agenda

1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

Dataset Cleanup

googleplaystore.csv cleanup:

- Remove exact duplicate rows.
- For apps with the same name, keep the newest by Last Updated. If dates are the same, keep the one with more Reviews. If both dates are missing, keep the first one seen.
- Force Rating to be between 0 and 5.
  Clean Installs by removing commas and plus, then convert to an integer
- Clean Price by turning Free into 0, remove the dollar sign, then convert to a number.
- Converted the app size to MB ,filled missing values using the median.
- Bucketed Android Ver columns into integer ranges.

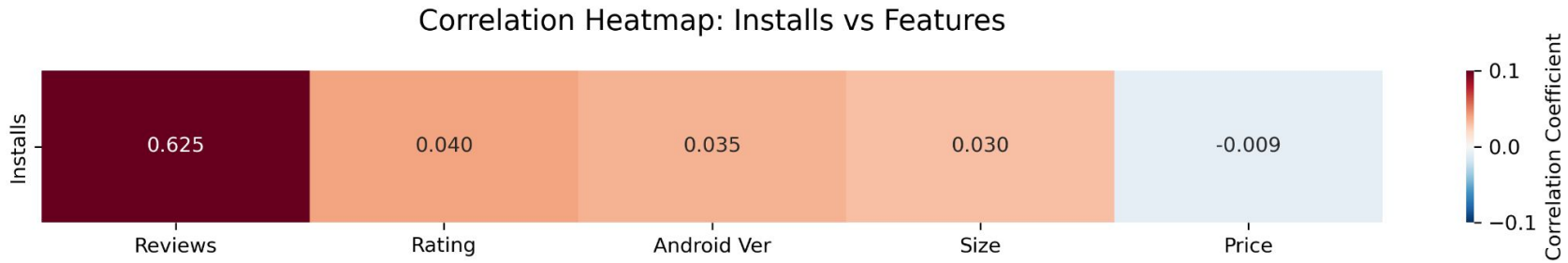
googleplaystore_user_reviews.csv:

- Normalize all of the columns by removing leading/trailing spaces from all string columns
- Converts "Sentiment_Polarity" and "Sentiment_Subjectivity" columns to numeric, coercing invalid values to NaN.
- Count number of NaN’s and number of reviews for each app
- Merge all of the comments to one row for each app by separating them with “|” sign
- Calculate the percentage of positive reviews, mean, median and std of the polarity and subjectivity

Agenda

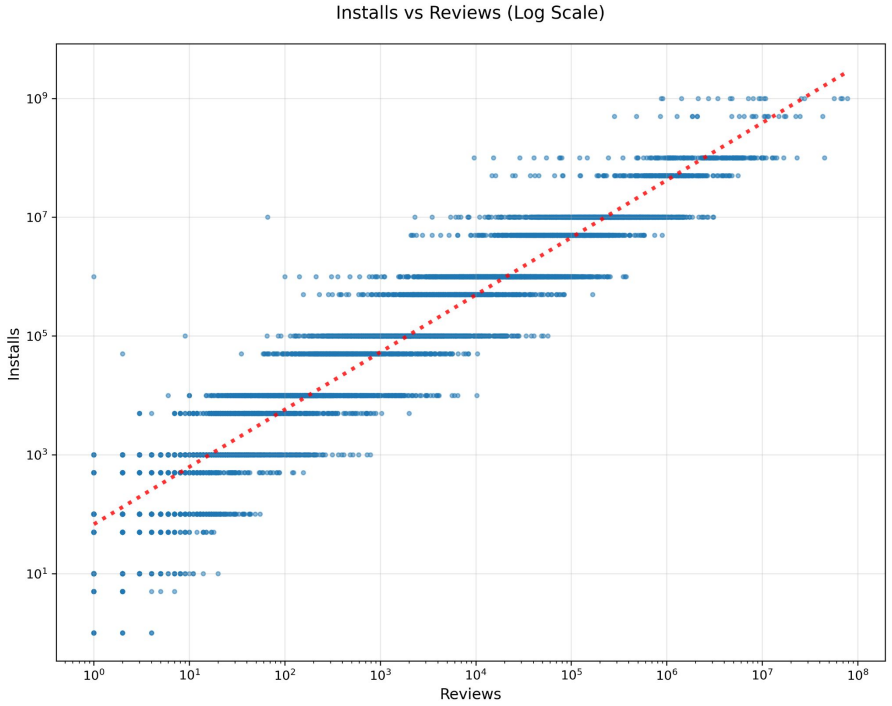
1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

The Big Picture: Correlation Heatmap



- heatmap of correlations of ‘numerical’ columns
- no. of reviews has a strong positive correlation
- price has a weak negative correlation; why weak?
- why are the others weakly positively correlated? is there something more going on?

Number of Reviews

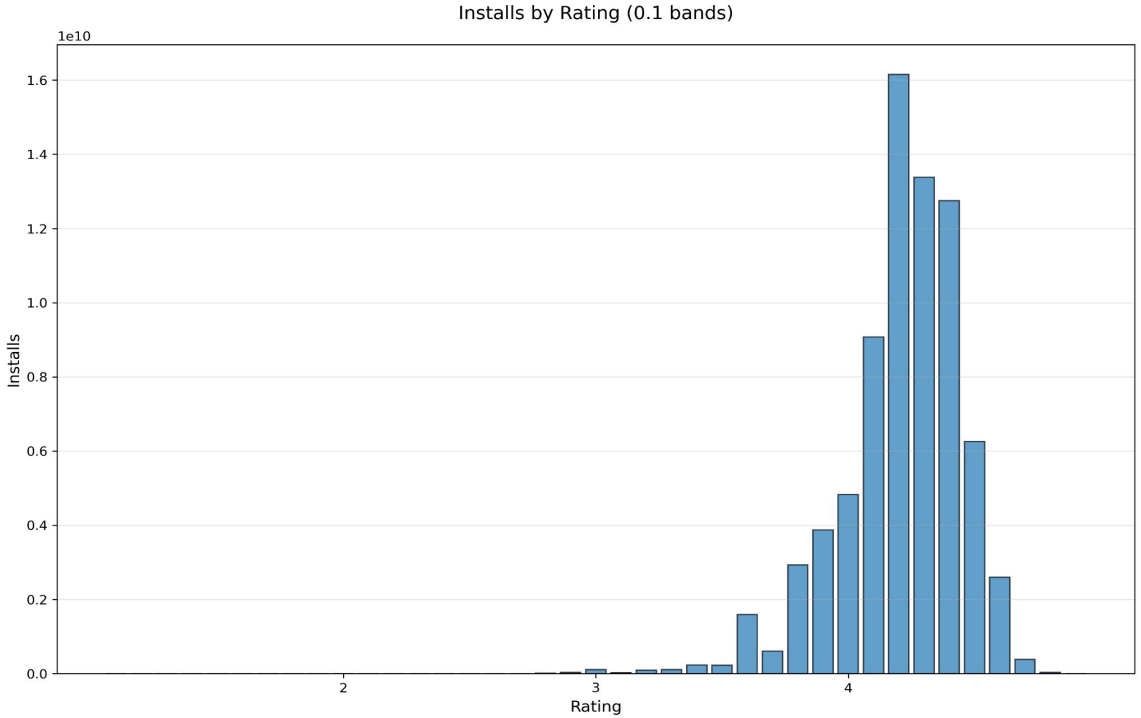


- log-log plot ⇒ exponential relationship
- note: flat lines, not cluster of dots, is a consequence of the dataset

Insights

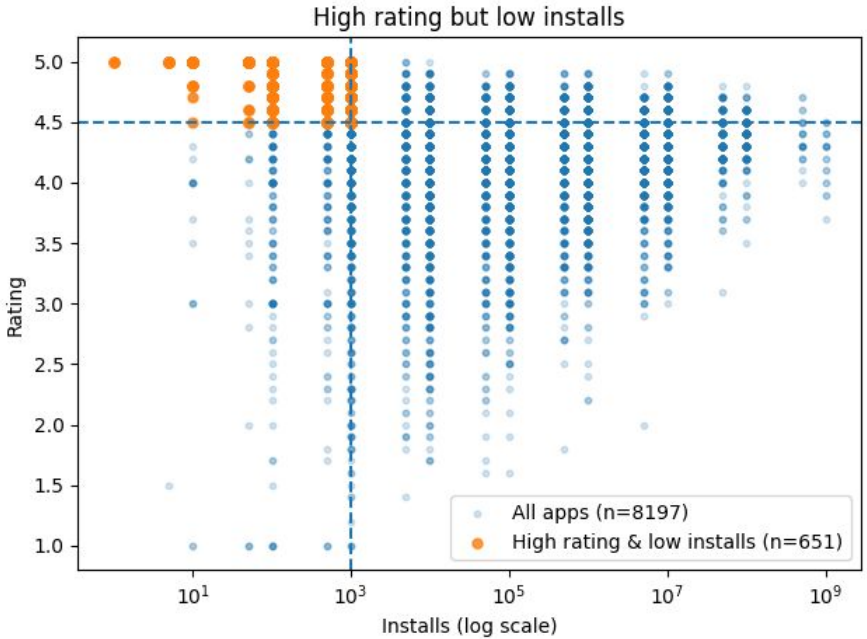
- no. of reviews has an exponential relationship with no. of installs.
  - social proof! people trust apps that others have reviewed

Star-Rating



the higher the installs, the less likely a 5 star rating!

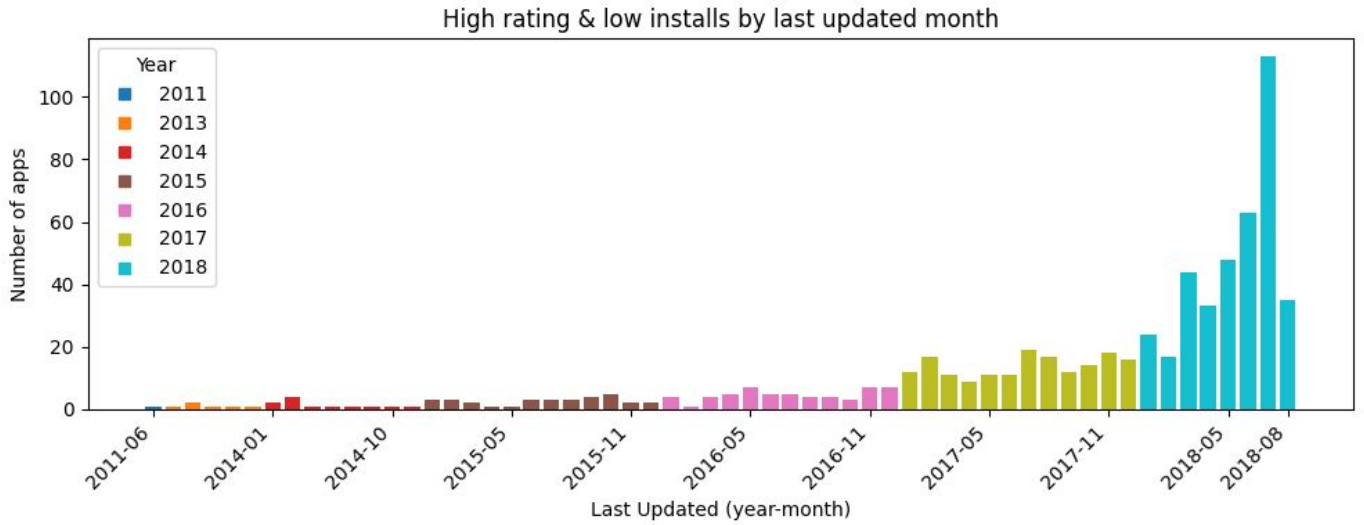
Star-Rating



4 quadrants : high/low installs and high/low ratings

the orange quadrant raises the question: if rating is high but installs are low, maybe it is because the app is new?

Star-Rating

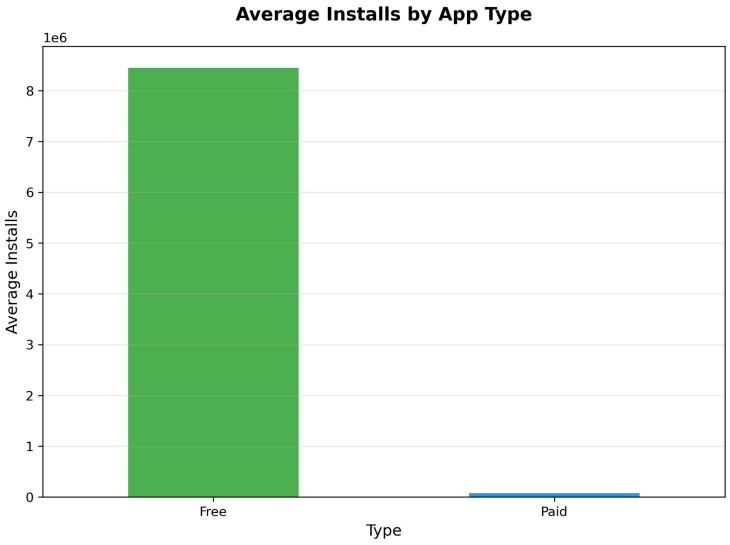


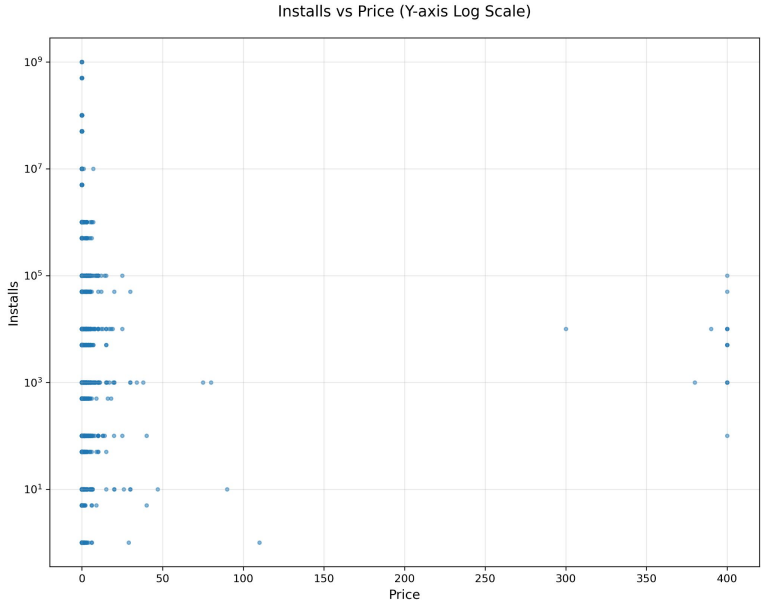
answer: yes!

Insights

- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it’s been around for a while, 4.2-4.3 is the sweet spot
  - as an app gets more and more installs, there will always be some haters

Price





people like free apps, for the most part …

Price

Top 5 Priciest Apps (Price > $200)

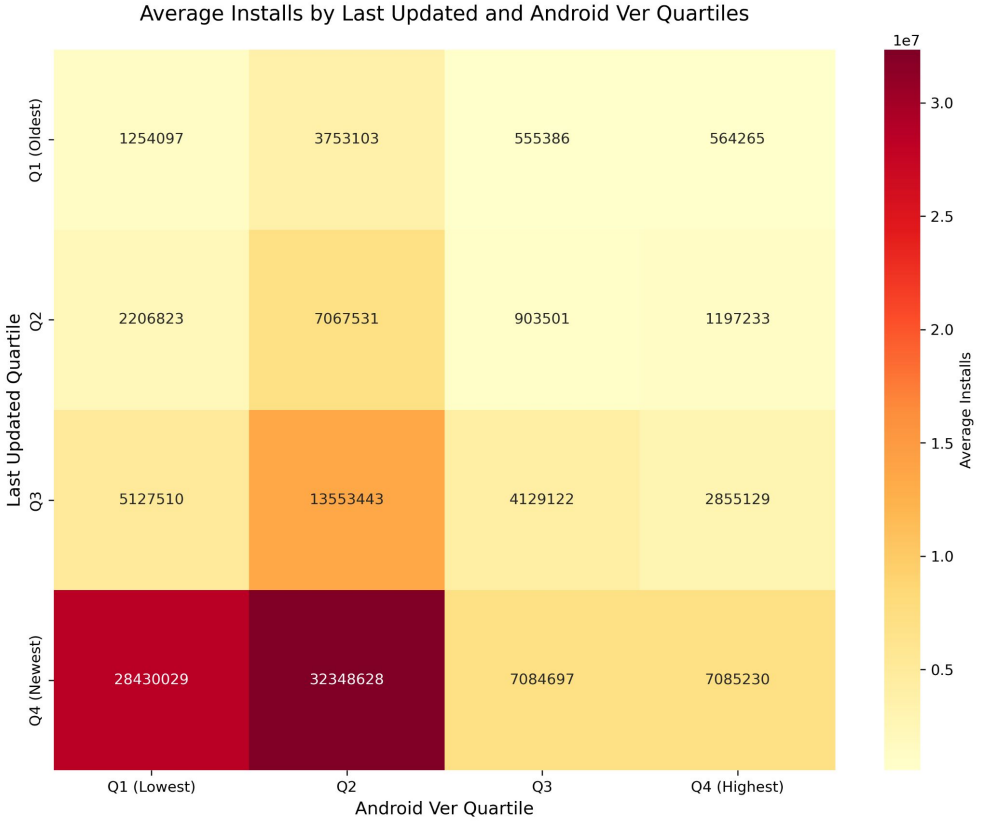
Price	App	Category	Rating	Reviews	Installs
$400.00	I'm Rich - Trump Edition	LIFESTYLE	3.6	275	10000
$399.99	most expensive app (H)	FAMILY	4.3	6	100
$399.99	□ I'm rich	LIFESTYLE	3.8	718	10000
$399.99	I AM RICH PRO PLUS	FINANCE	4.0	36	1000
$399.99	I am Rich	FINANCE	4.3	180	5000

… but with some interesting exceptions!

Insights

- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it has been around for a while, 4.2-4.3 is the sweet spot
- free apps clearly do much better (with a few interesting exceptions).
  - how does that help marketers trying to get an app to make money? likely, ads
  - outlier: gambling & the associated poor judgement

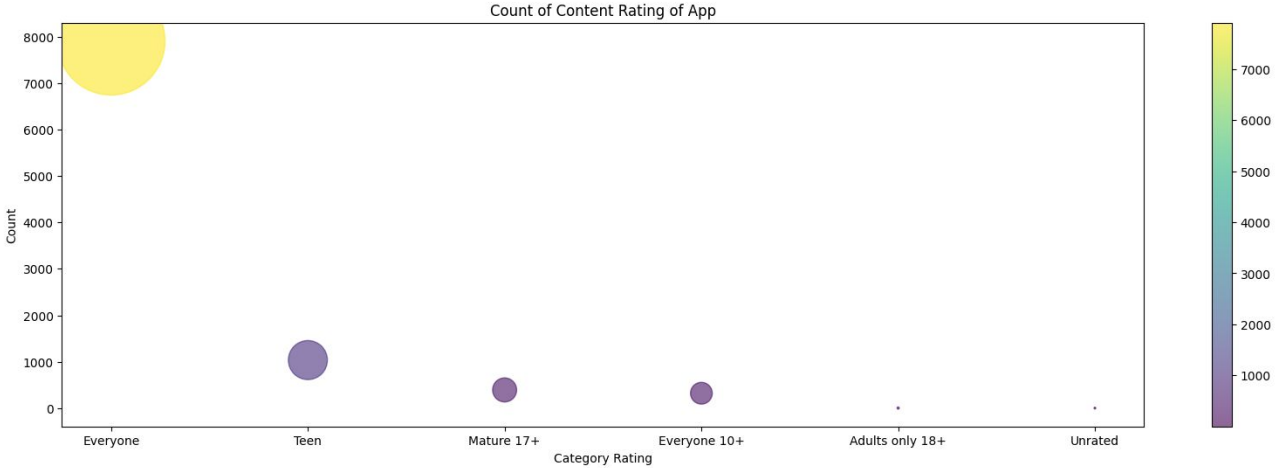
Developer Activity



Insights

- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it has been around for a while, 4.2-4.3 is the sweet spot
- free apps clearly do much better (with a few interesting exceptions).
- apps well maintained by developers tend to do much better.
    - incorporating user feedback (gained from reviews)

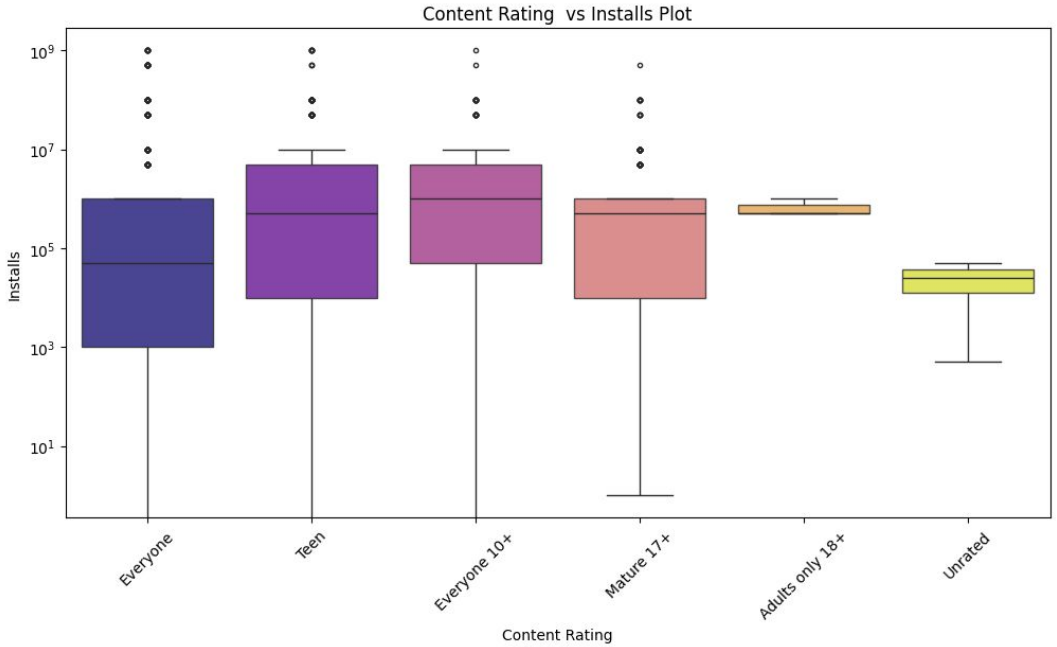
Content Rating



most apps are rated for Everyone

but…

Content Rating

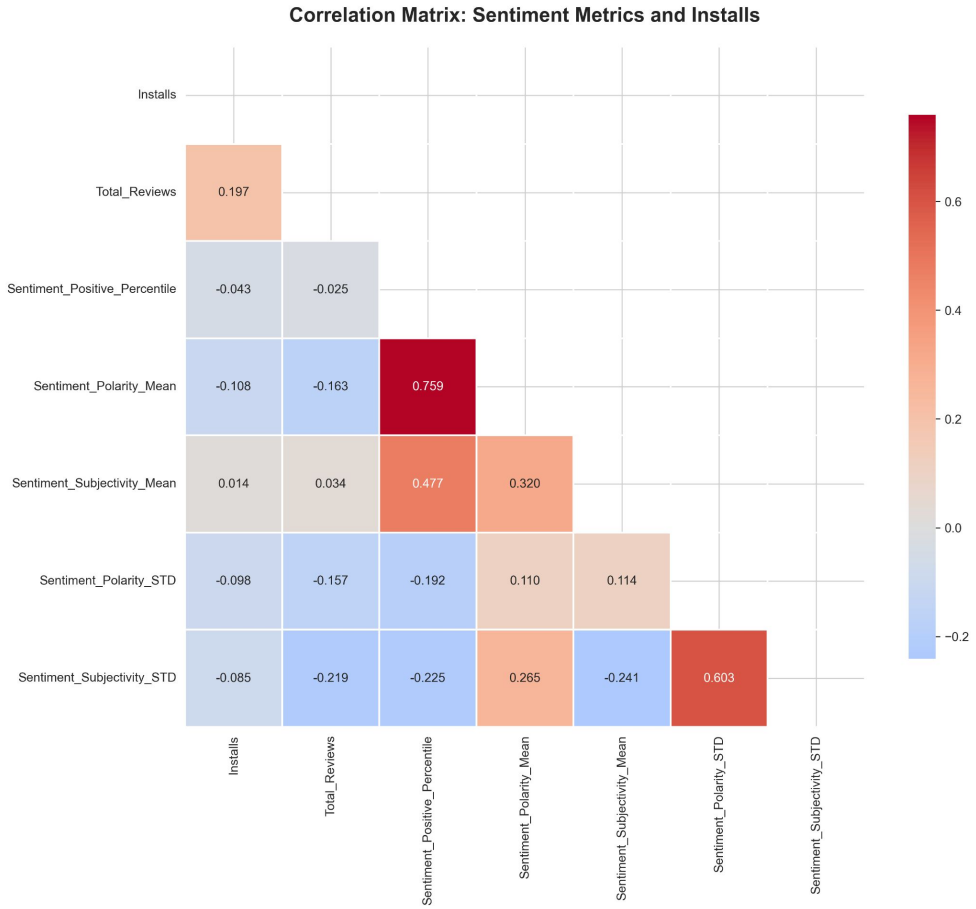


… targeting a demographic is a better strategy!

Insights

- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it has been around for a while, 4.2-4.3 is the sweet spot
- free apps clearly do much better (with a few interesting exceptions).
- apps well maintained by developers tend to do much better.
- it is better to target, in your marketing, a particular demographic

Sentiment


Correlation Matrix: Sentiment Metrics and Installs
Installs
Total_Reviews
Sentiment_Positive_Percentile
Sentiment_Polarity_Mean
Sentiment_Subjectivity_Mean
Sentiment_Polarity_STD
Sentiment_Subjectivity_STD
0.197
-0.043
-0.025
-0.108
-0.163
0.759
0.014
0.034
0.477
0.320
-0.098
-0.157
-0.192
0.110
0.114
-0.085
-0.219
-0.225
0.265
-0.241
0.603
0.6
0.4
0.2
0.0
−0.2

Insights

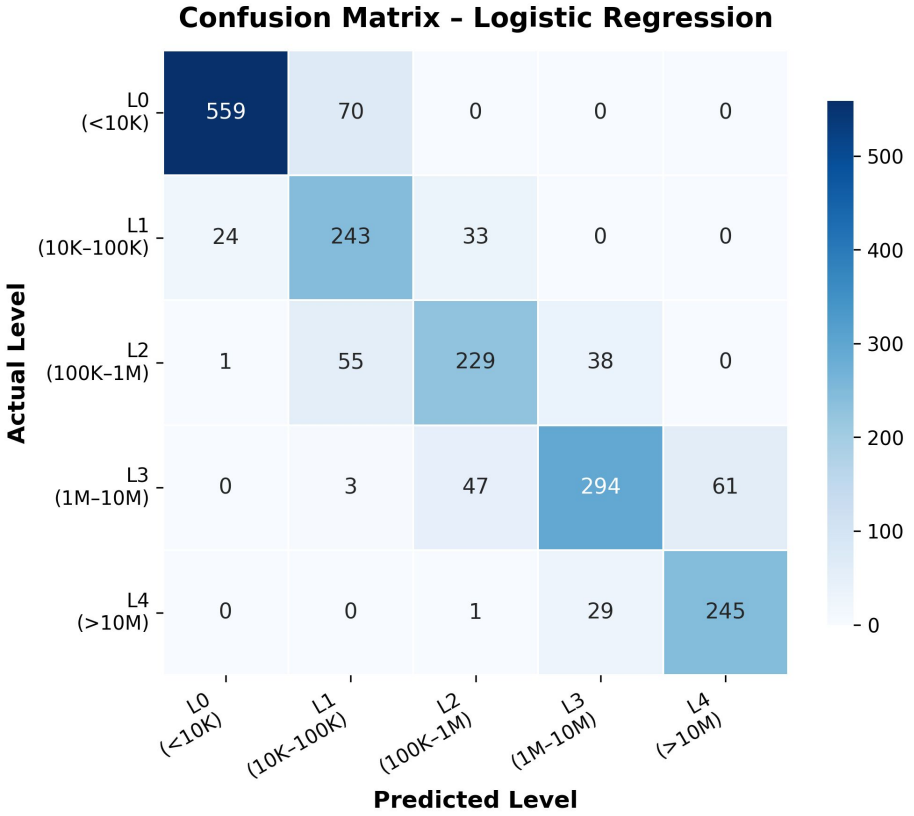
- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it has been around for a while, 4.2-4.3 is the sweet spot
- free apps clearly do much better (with a few interesting exceptions).
- apps well maintained by developers tend to do much better.
- it is better to target, in your marketing, a particular demographic
- just because a lot of apps of a certain type are being made, doesn't mean you'll do well
- sentiment doesn't correlate much with number of installs
  - likely reason: this dataset is not time series, and further, have much lower no. of sentiment samples. which could mean that an app which initially got negative reviews improved over time by incorporating feedback.

Agenda

1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

Prediction Model

Accuracy	0.813
Balanced accuracy	0.805



Agenda

1. About the dataset
2. Cleanup
3. Dataset Analysis deep dive
4. Prediction Model
5. Conclusion

Conclusions

Things that you should focus on if you want your app to do well:

- No. of reviews
- price (free apps)
- targeted demographic
- developer activity
- star rating (depends on app age)

Logistic regression based model with ~80% accuracy

UC San Diego

Thank you!

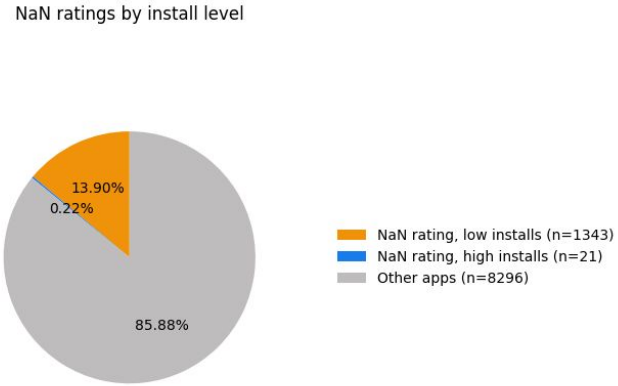
UC San Diego

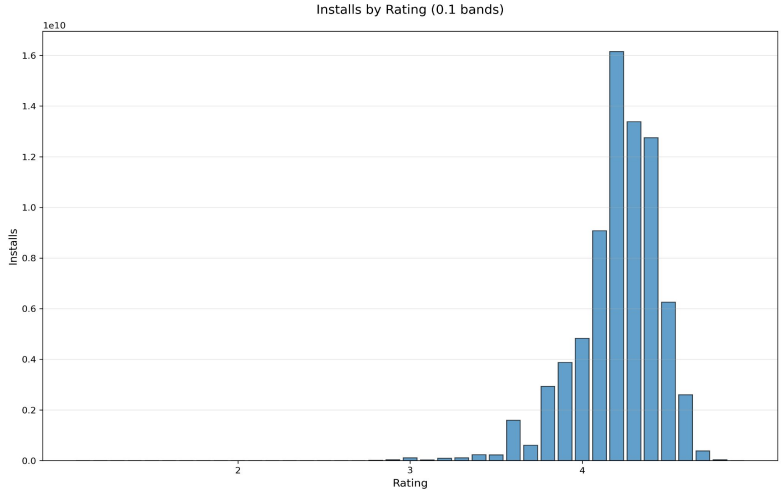
Backup

Size

some categories just have large-sized apps (e.g. games)

Star-Rating

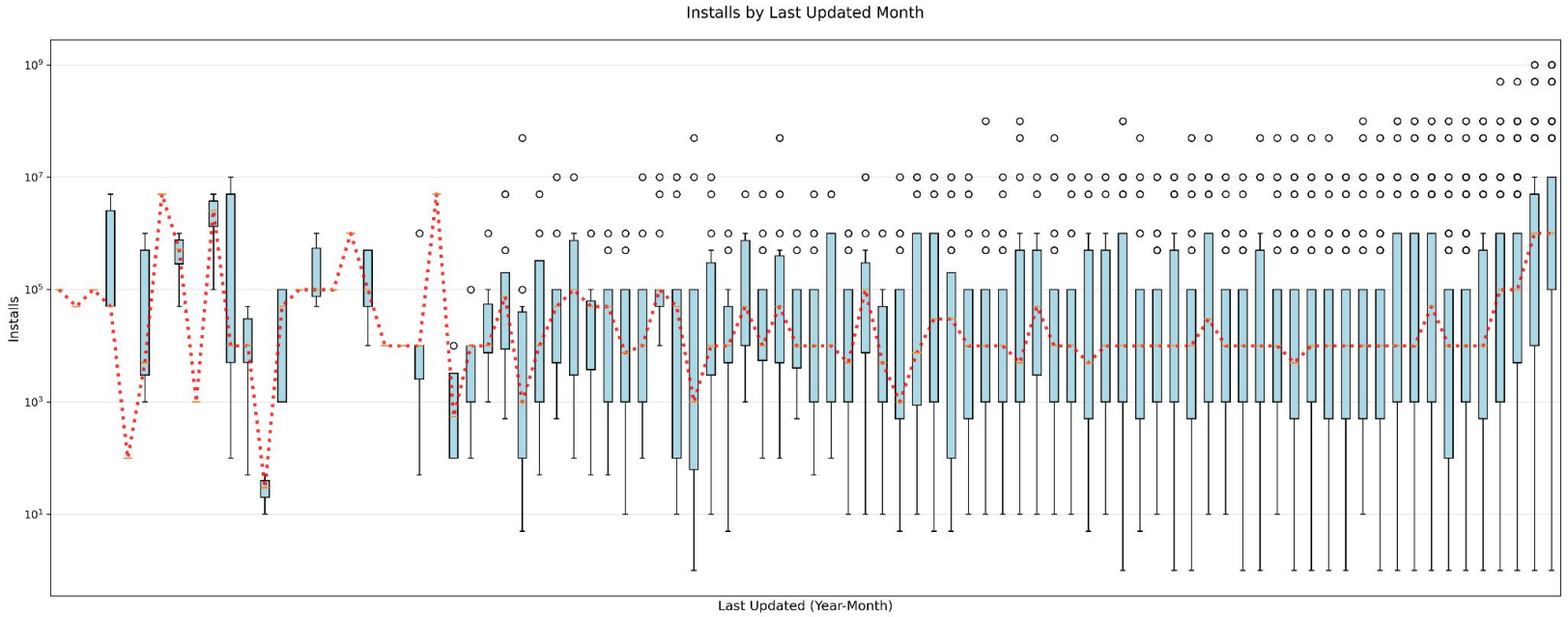




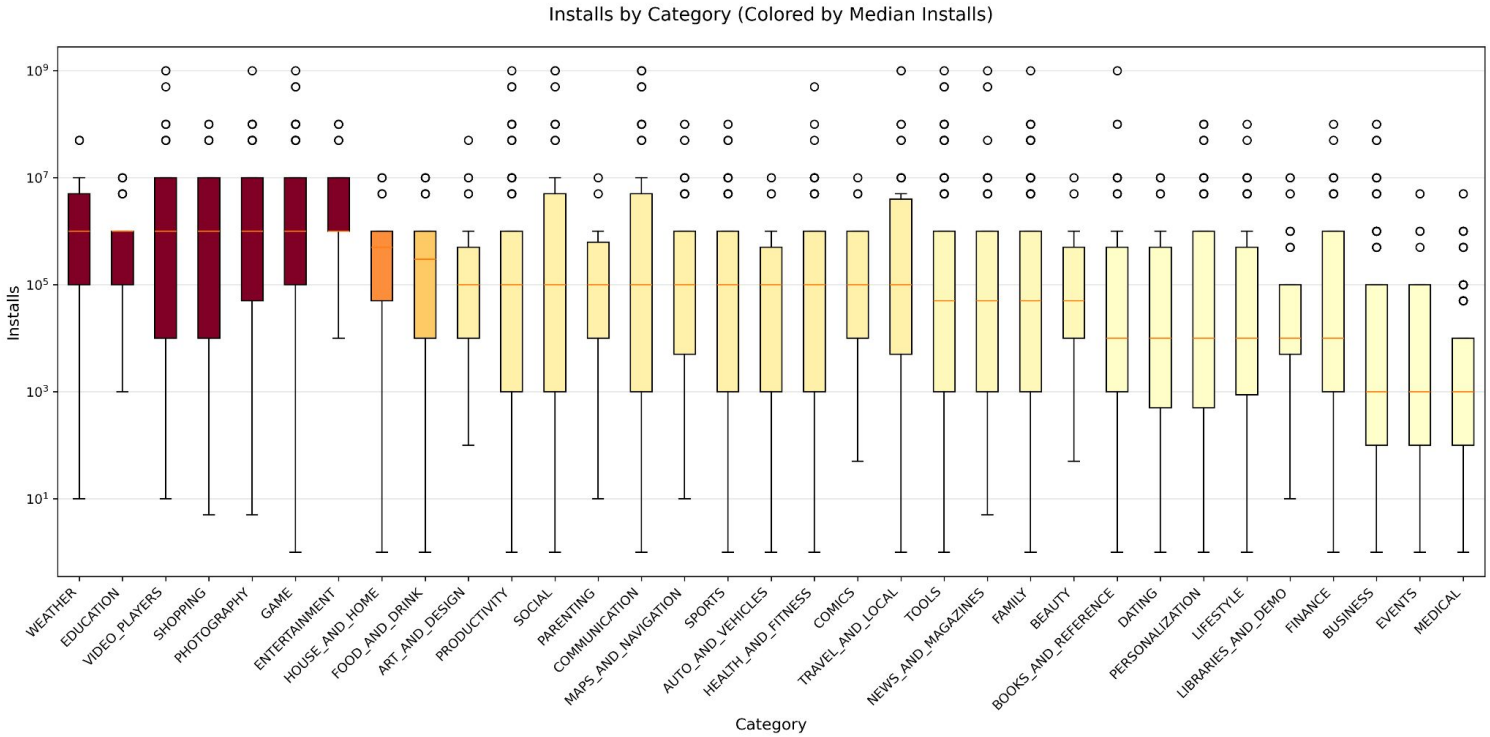
some ratings in the data are NaN (treat as zero). most are for low installs (expected). NaN for high installs are outliers; ignore

the higher the installs, the less likely a 5 star rating!

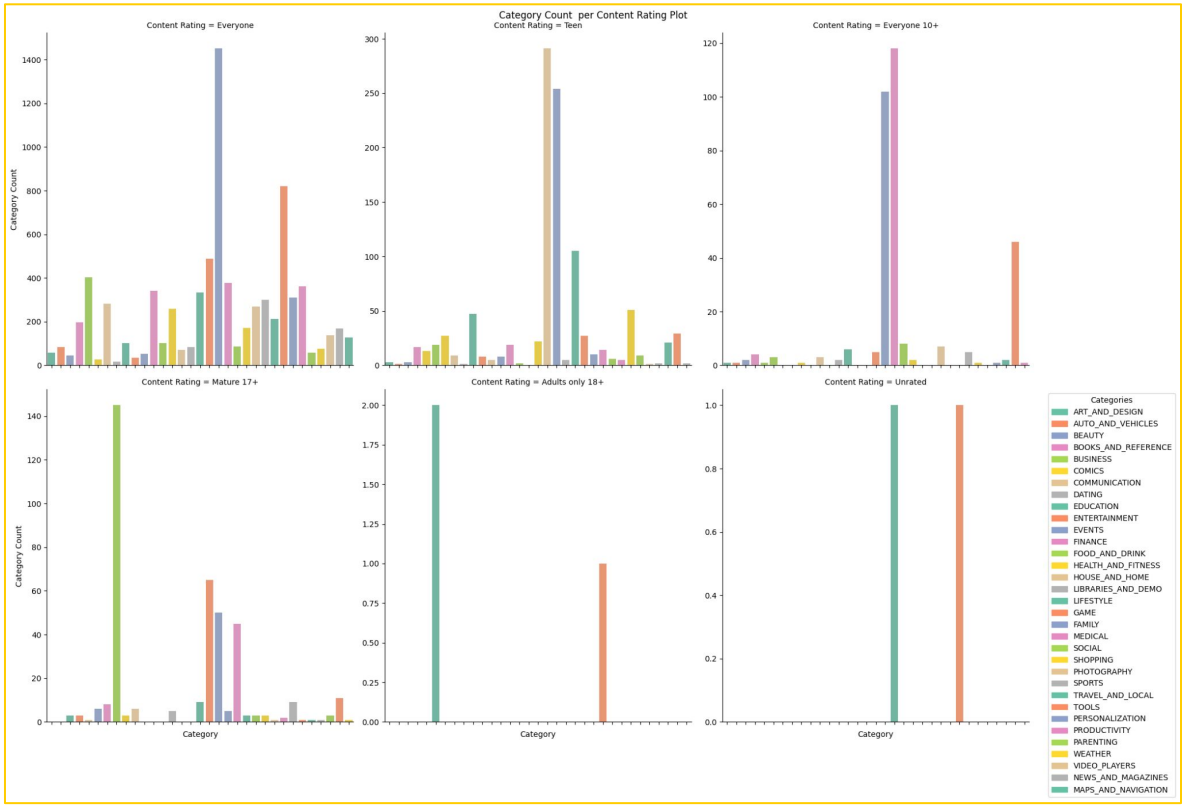
Developer Activity



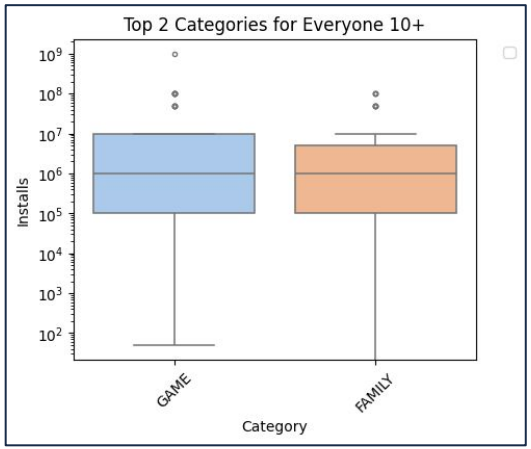
Category

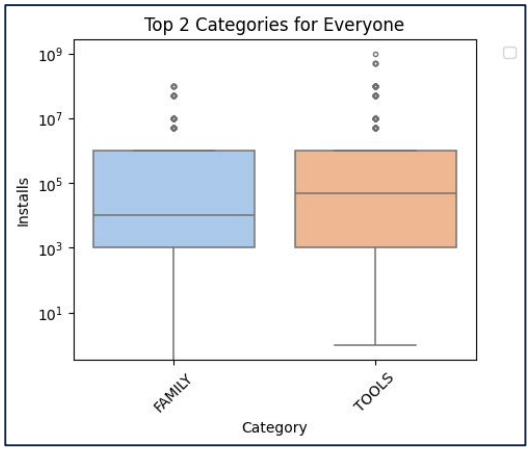

Installs by Category (Colored by Median Installs)
Installs
10^9
10^7
10^5
10^3
10^1
WEATHER
EDUCATION
VIDEO_PLAYERS
SHOPPING
PHOTOGRAPHY
GAME
ENTERTAINMENT
HOUSE_AND_HOME
FOOD_AND_DRINK
ART_AND_DESIGN
PRODUCTIVITY
SOCIAL
PARENTING
COMMUNICATION
MAPS_AND_NAVIGATION
SPORTS
AUTO_AND_VEHICLES
HEALTH_AND_FITNESS
COMICS
TRAVEL_AND_LOCAL
TOOLS
NEWS_AND_MAGAZINES
FAMILY
BEAUTY
BOOKS_AND_REFERENCE
DATING
PERSONALIZATION
LIFESTYLE
LIBRARIES_AND_DEMO
FINANCE
BUSINESS
EVENTS
MEDICAL
Category

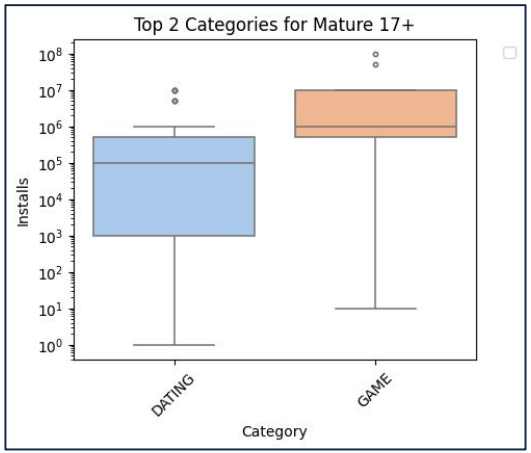
Content Rating



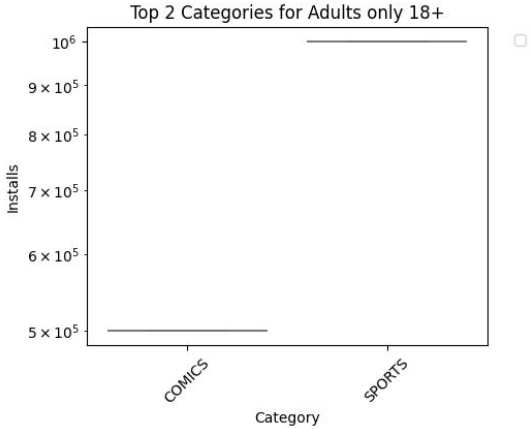
Content Rating

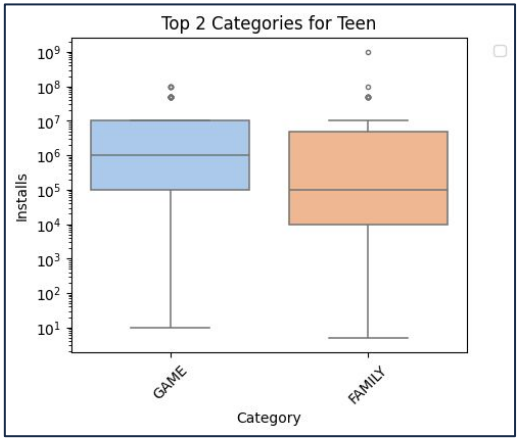




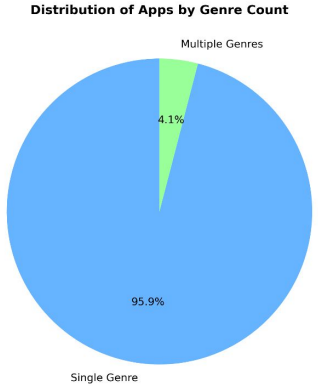


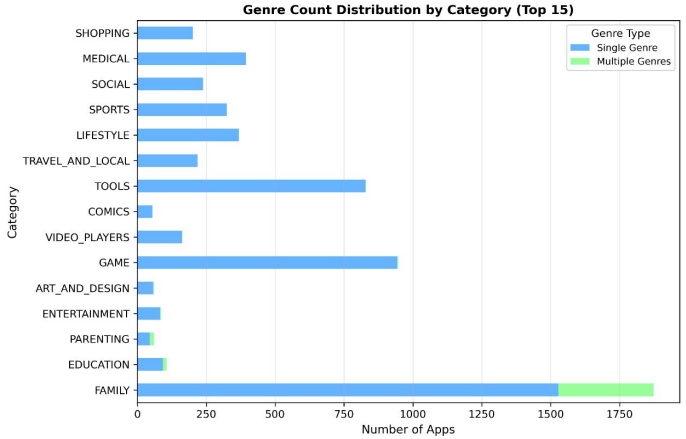
Content Rating





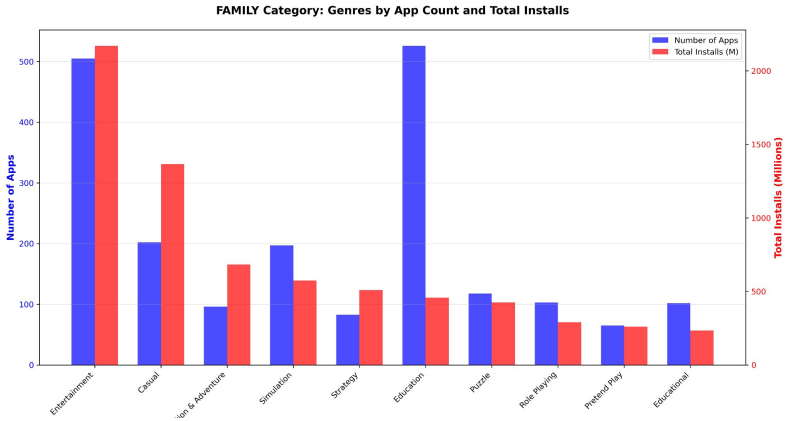
Genres

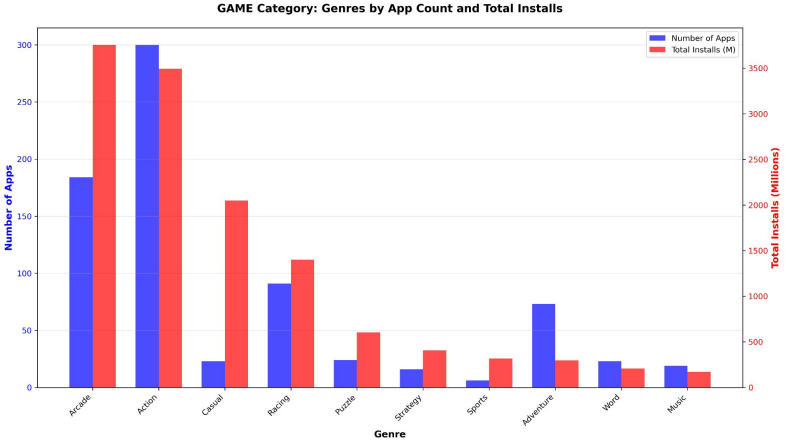




some categories/genres have a lot of apps…

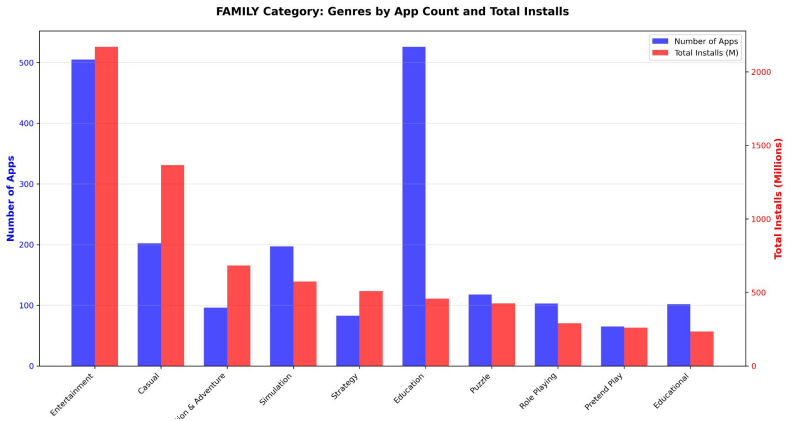
Genres

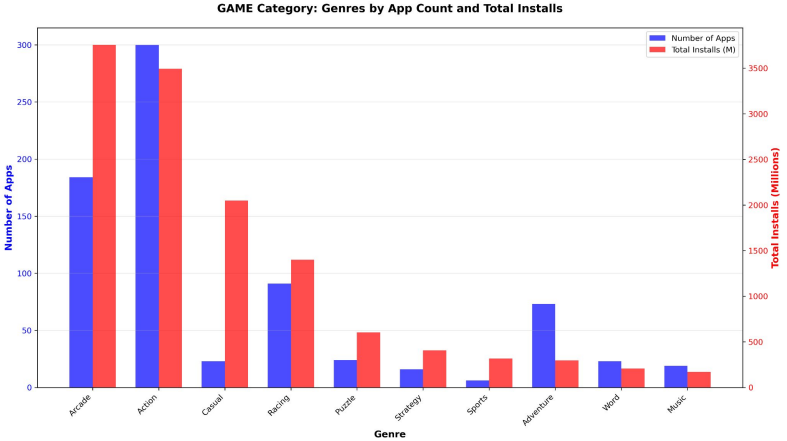




… but that doesn't really reflect in the number of installs!

Genres





genre having a lot of apps doesn’t reflect in the number of installs!

Insights

- no. of reviews has an exponential relationship with no. of installs.
- if the app is new, you want a 5-stars. if it has been around for a while, 4.2-4.3 is the sweet spot
- free apps clearly do much better (with a few interesting exceptions).
- apps well maintained by developers tend to do much better.
- it is better to target, in your marketing, a particular demographic
- just because a lot of apps of a certain type are being made, doesn't mean you'll do well
  - don't follow the rat race/try to break into an existing market. a lot of app markets are winner-take-all

Prediction Model

Table 1: Feature groups used in the logistic regression model

Feature group	# Features
Reviews	5
Rating × last updated	1
Price (free vs paid)	1
Developer activity score	1
Content rating targeting score	1
Category	32
Total	41

Table 2: Install level definitions

Level	Label	Install range
L0	Low	Installs $< 10\,000$
L1	Medium-Low	$10\,000 \leq$ Installs $< 100\,000$
L2	Medium	$100\,000 \leq$ Installs $< 1\,000\,000$
L3	Medium-High	$1\,000\,000 \leq$ Installs $< 10\,000\,000$
L4	High	Installs $\geq 10\,000\,000$

Prediction Model