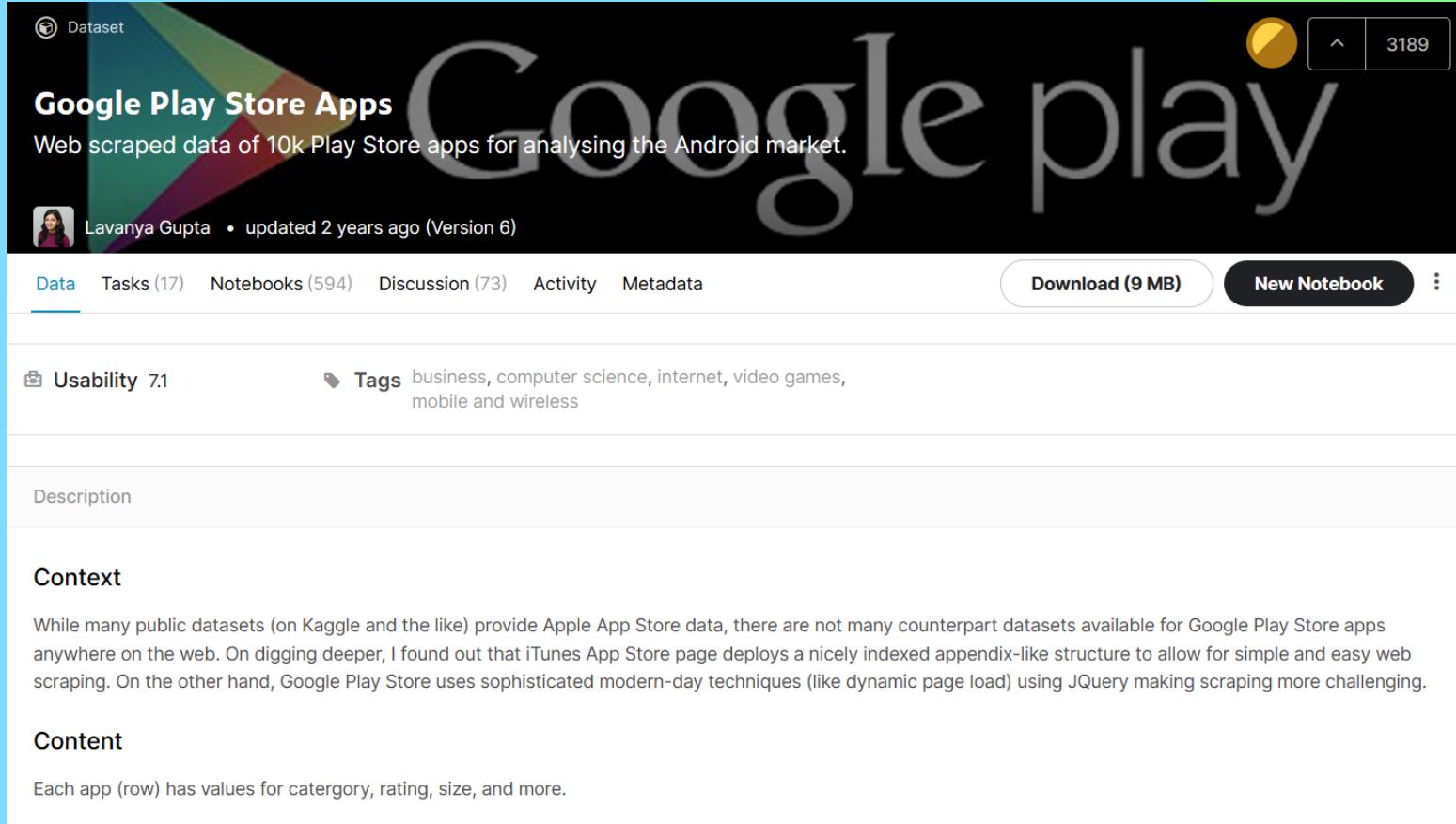


Google Play

Store Apps Rating



Dataset Presentation & Introduction



The screenshot shows a Kaggle dataset page titled "Google Play Store Apps". The page header includes a "Dataset" icon, the title "Google Play Store Apps", a subtitle "Web scraped data of 10k Play Store apps for analysing the Android market.", and a profile picture of Lavanya Gupta with the text "Lavanya Gupta • updated 2 years ago (Version 6)". Below the header are navigation links for "Data", "Tasks (17)", "Notebooks (594)", "Discussion (73)", "Activity", and "Metadata". On the right side of the header are "Download (9 MB)" and "New Notebook" buttons. A "3189" badge is visible in the top right corner. The main content area features a large "Google play" logo. Below it, there are sections for "Usability 7.1", "Tags business, computer science, internet, video games, mobile and wireless", and "Description". The "Context" section contains a detailed paragraph about the availability of datasets for Google Play Store apps compared to the Apple App Store. The "Content" section states that each app has values for category, rating, size, and more.

<https://www.kaggle.com/lava18/google-play-store-apps>

The provided dataset collects app data from Google Play Store that users can download and use on their Google devices.



Data Understanding

Sequential Path



Data Semantics



Data Semantics



1		App <i>(string)</i> <i>Application name</i>	8		Price <i>(string)</i> <i>Cost of the app</i>
2		Category <i>(string)</i> <i>Category the app belongs to</i>	9		Content Rating <i>(string)</i> <i>Age group of users</i>
3		Rating <i>(float)</i> <i>App evaluation</i>	10		Genres <i>(string)</i> <i>Any multiple genres app belongs to</i>
4		Reviews <i>(string)</i> <i>Number of reviews per app</i>	11		Last Updated <i>(string)</i> <i>Date when the app was last updated</i>
5		Size <i>(string)</i> <i>App dimension</i>	12		Current Ver <i>(string)</i> <i>Current version of the app</i>
6		Installs <i>(string)</i> <i>Number of downloads per app</i>	13		Android Ver <i>(string)</i> <i>Minimum required Android version to download the app</i>
7		Type <i>(string)</i> <i>If the app is free or not</i>			

Data Quality Evaluation

Values investigation

- The 10472nd record was one tab shifted, therefore we regularized it.

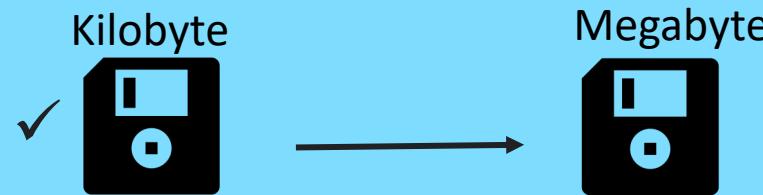
Missing values detection

- **Rating:** 1474 missing values removed.
- **Category & Genres:** only one missing value replaced using the mode by Content Rating and Size.
- **Current Ver:** 4 missing values replaced using the mode by Last Updated and Android Ver.
- **Android Ver:** 2 missing values replaced using the mode by Last Updated.

Variables Transformation

Size

- ✓ "Varies with device" → 0.0



- ✓ *String* → *Float*

Installs

- ✓ *String* → *Intervals*

1+ (1 - 5)
5+ (6 - 10)
10+ (11 - 50)
50+ (51 - 100)
100+ (101 - 500)
500+ (501 - 1000)

. . .

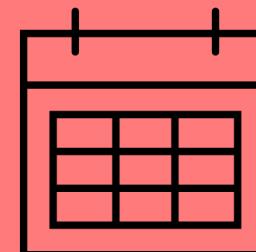
Price

- ✓ *String* → *Float*



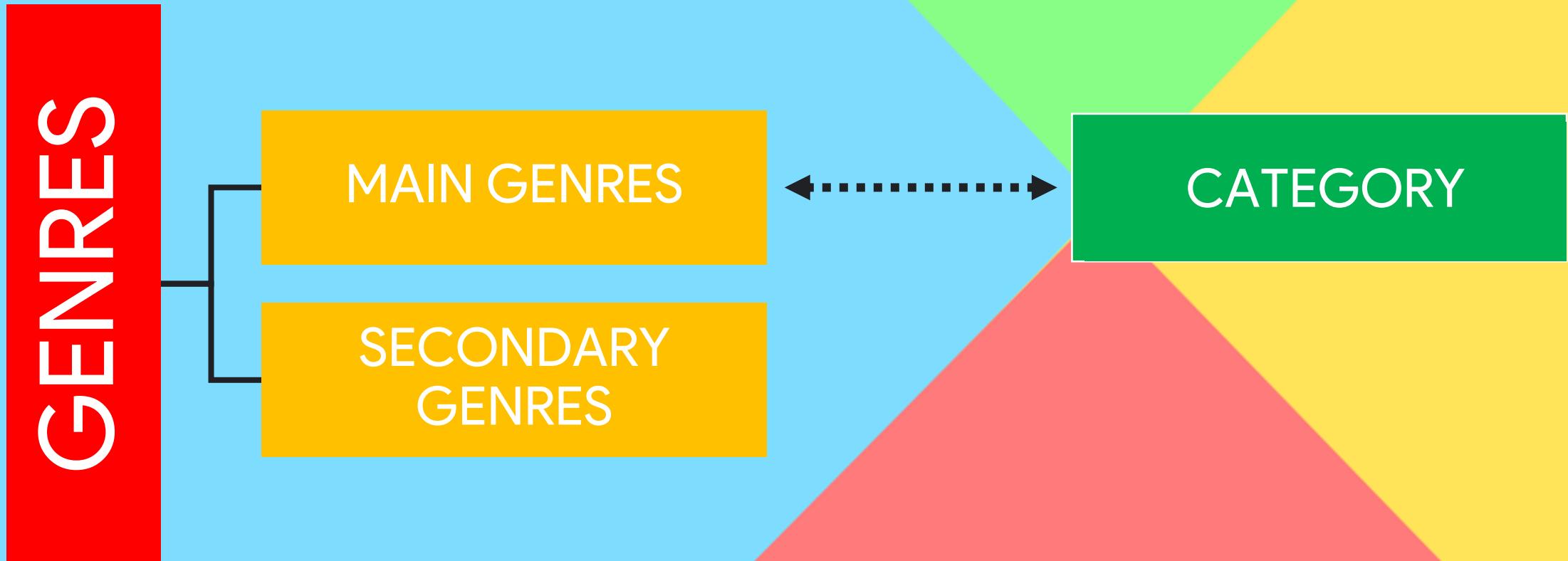
Last Update

- ✓ *String* → *Datetime*
- ✓ '%m-%Y'



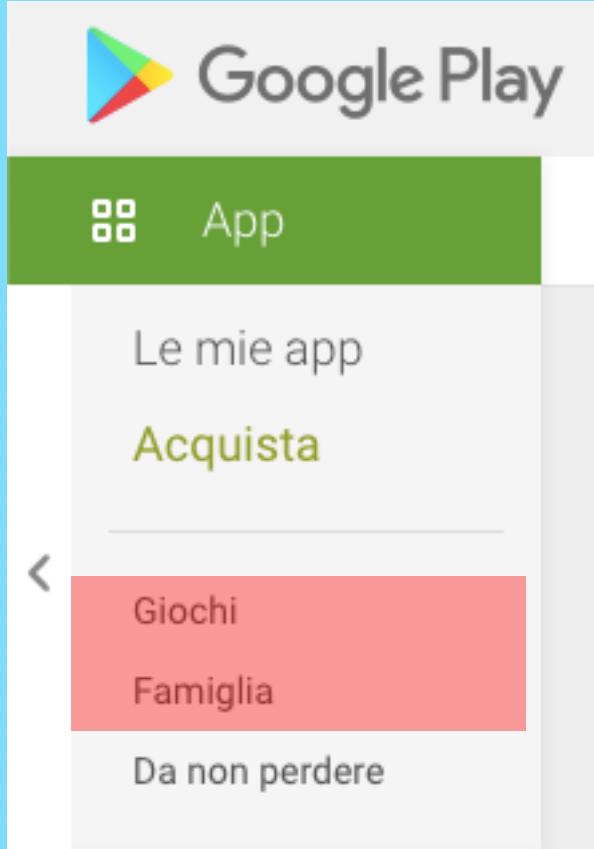
New variables creation

Main Genres and Secondary Genres



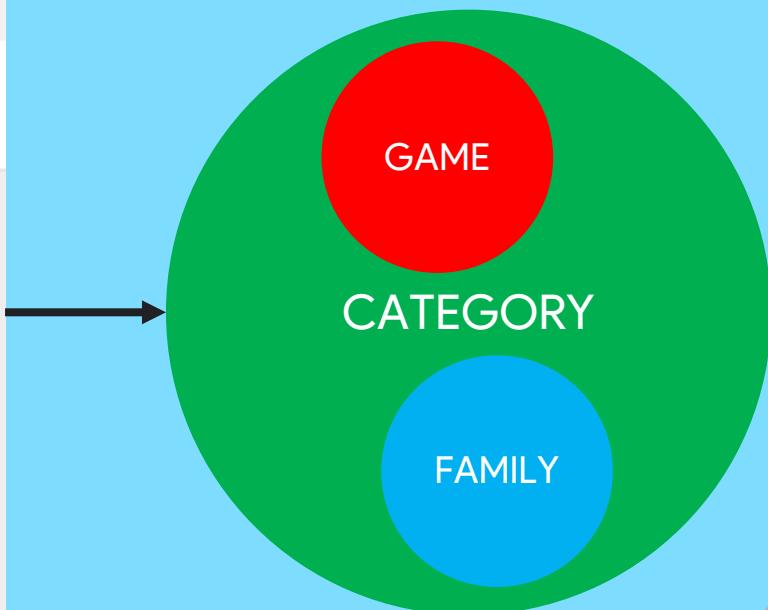
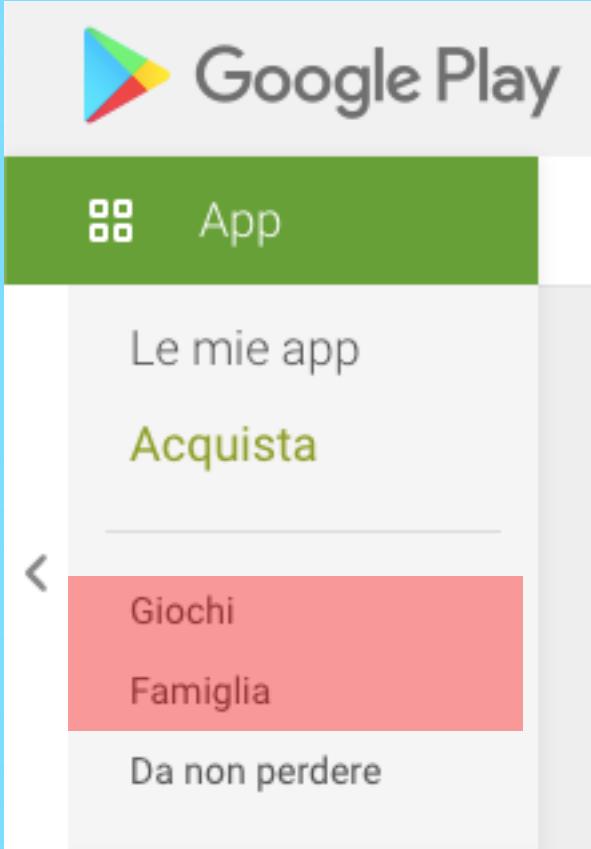
New variables creation

Is_Family Is_Game



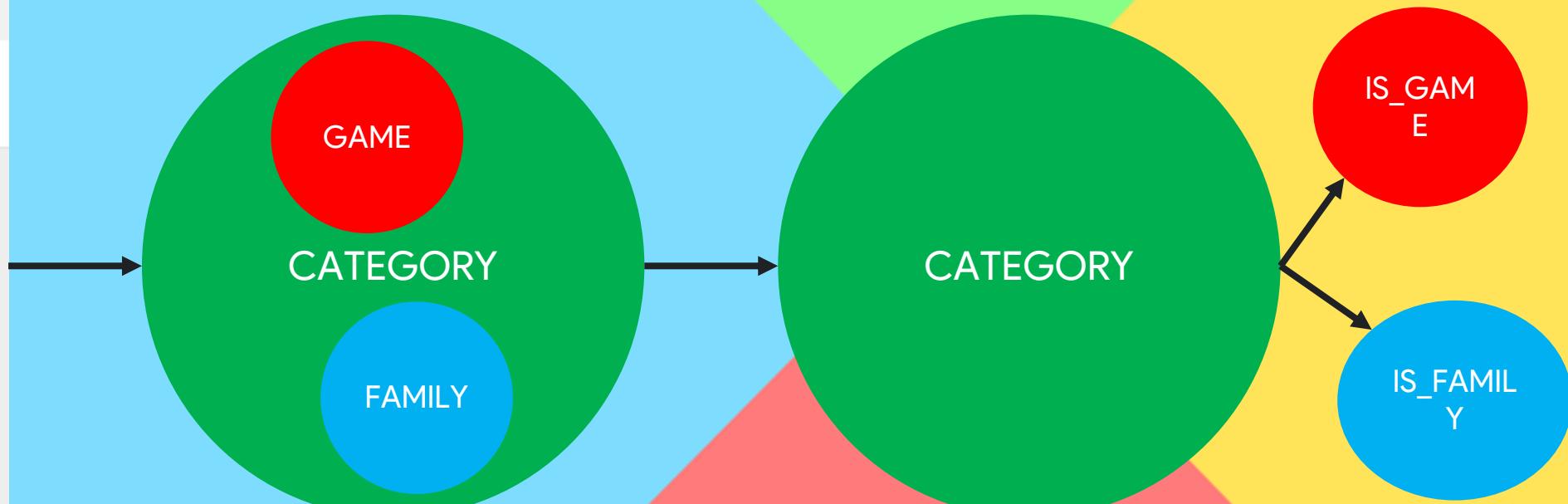
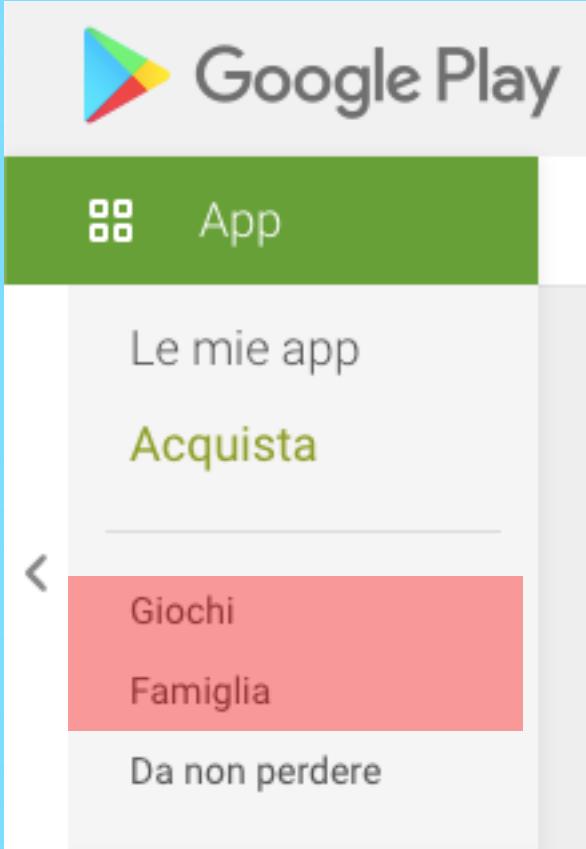
New variables creation

Is_Family Is_Game



New variables creation

Is_Family Is_Game



New variables creation

Compatibility



Varies with device

High compatibility

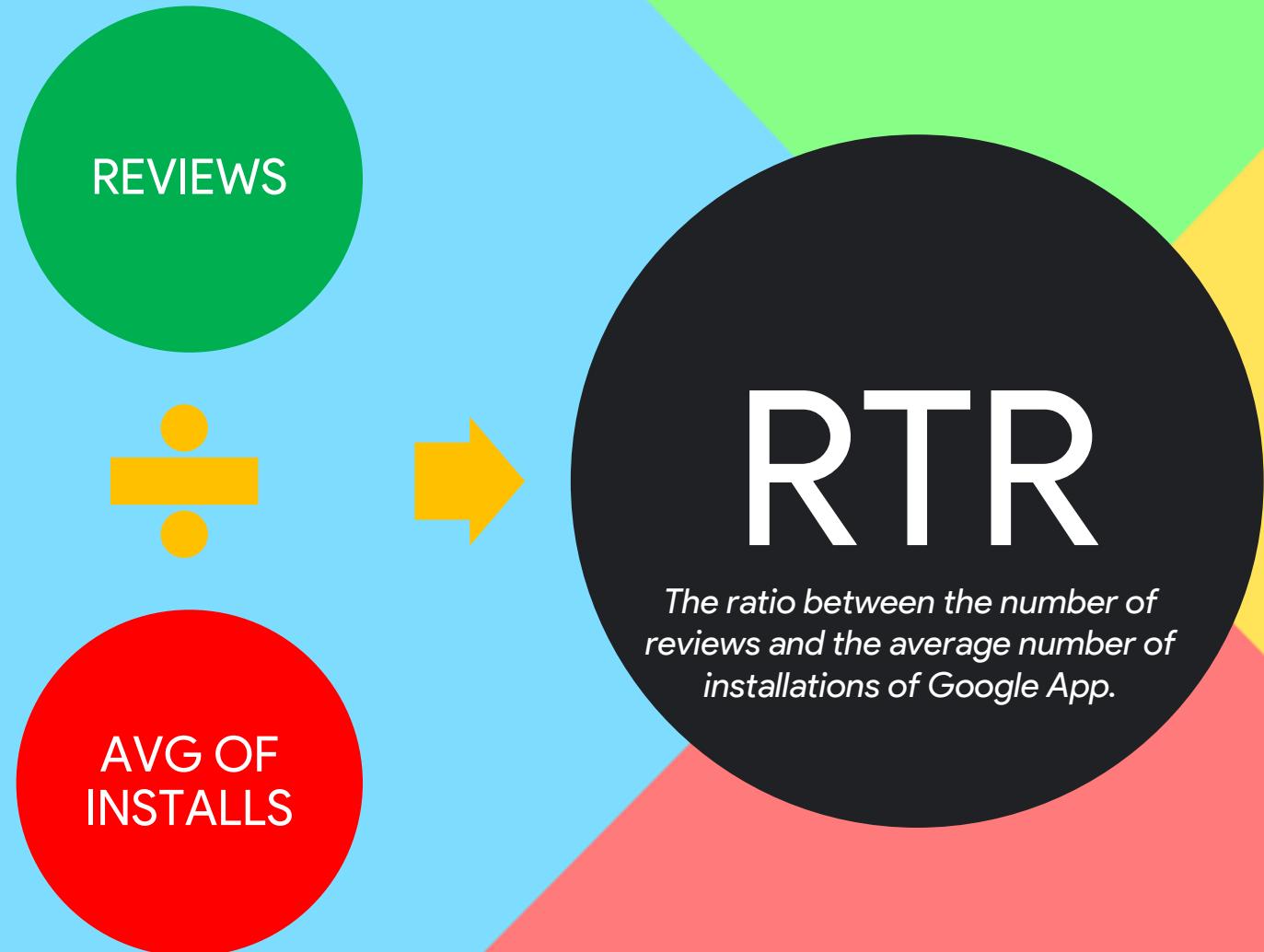
Medium compatibility

Low compatibility

- "Varies with device"
- «1.0 and up»
- «2.3.3 and up»
- «4.4W and up»
- All the others versions
- «7.0 and up»
- «8.0 and up»
- «5.0 – 6.0»

New variables creation

Reviews - Through Rate



G Data Transformation



1	App	8	Price	14	Current Ver
2	Category	9	Content Rating	15	Android Ver
3	Rating	10	Genres	16	Is_Family
4	Reviews	11	Main_Genres	17	Is_Game
5	Size	12	Secondary_Genres	18	Compatibility
6	Installs	13	Last Updated	19	RTR
7	Type				

Giochi

App

Film

Libri

Data Cleaning



Data Cleaning



1	App	8	Price	14	Current Ver
2	Category	9	Content Rating	15	Android Ver
3	Rating	10	Genres	16	Is_Family
4	Reviews	11	Main_Genres	17	Is_Game
5	Size	12	Secondary_Genres	18	Compatibility
6	Installs	13	Last Updated	19	RTR
7	Type				

Giochi

App

Film

Libri



Data Cleaning



1	App	8	Price	14	Current Ver
2		9	Content Rating	15	Android Ver
3	Rating	10		16	Is_Family
4	Reviews	11	Main_Genres	17	Is_Game
5	Size	12	Secondary_Genres	18	Compatibility
6	Installs	13	Last Updated	19	RTR
7					

Outliers and Noise

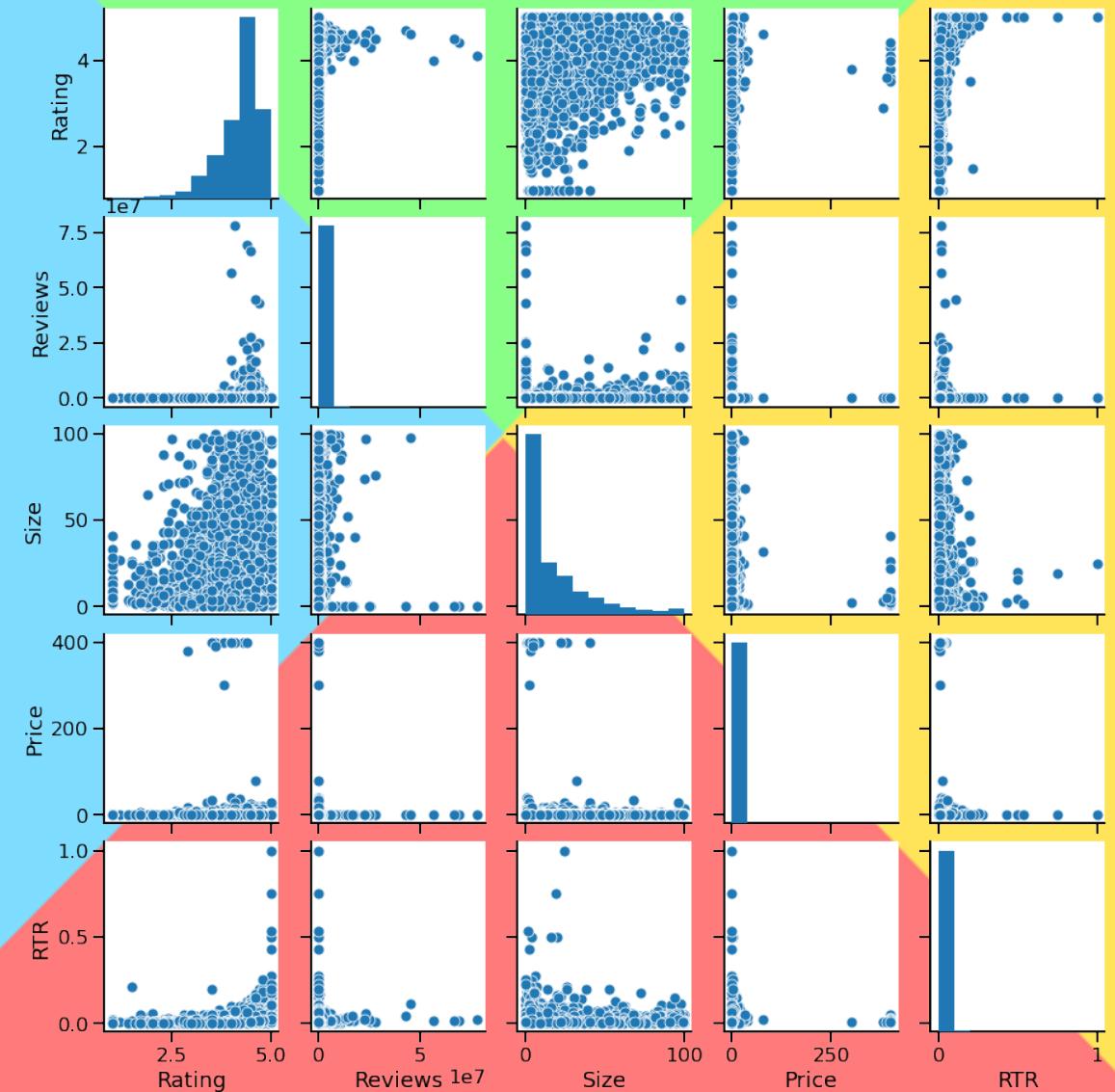
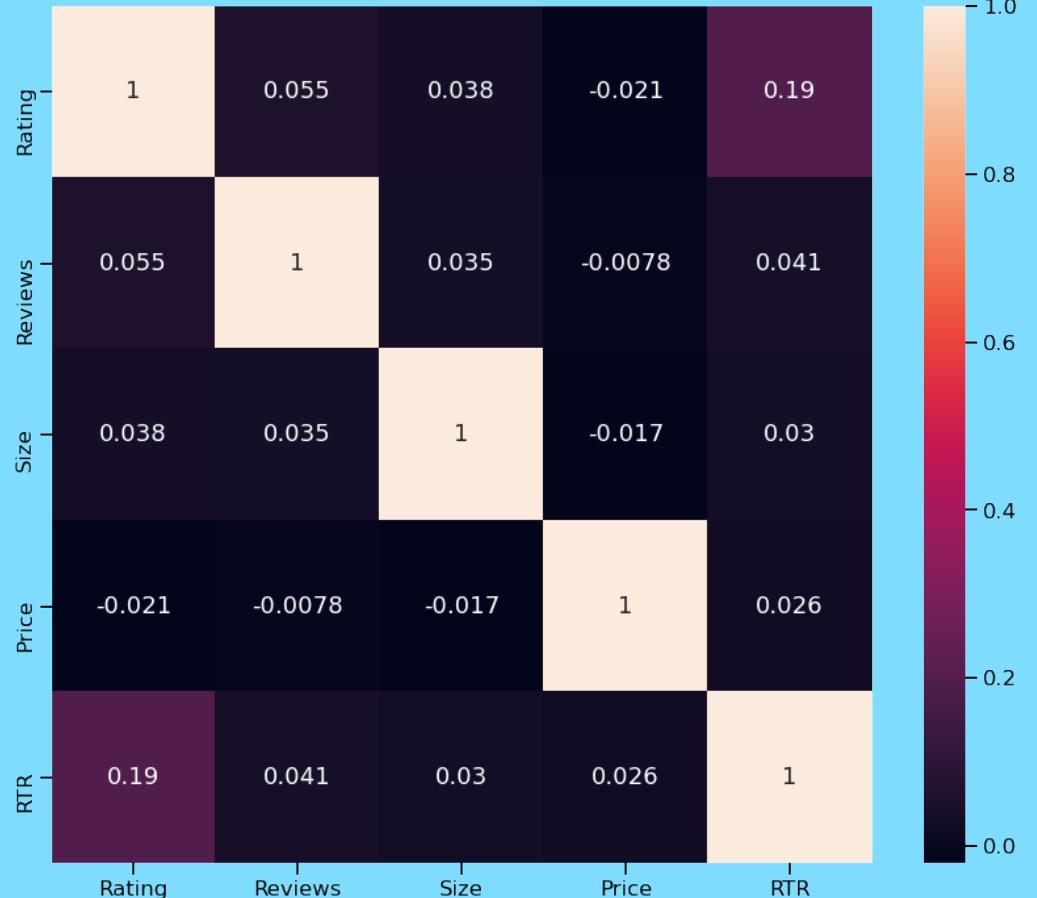
- ✓ Status symbol apps ("I am Rich")

Price	Category	Frequency	
299.99	Lifestyle	1	
379.99	Lifestyle	1	
389.99	Family	1	
399.99	Finance	6	
	Family	3	
	Lifestyle	2	
400.00	Lifestyle	1	

- ✓ 'Unrated' value removal from Content Rating

Pairwise correlation

Pairwise Correlation



Data Integration

Integration from external sources

Mobile App Store (7200 apps)
Analytics for Mobile Apps

Ramanathan • updated 2 years ago (Version 7)

Data Tasks (1) Notebooks (84) Discussion (22) Activity Metadata

Download (13 MB) New Notebook :

Usability 8.2 License GPL 2 Tags business, internet, mobile and wireless

Description

Mobile App Statistics (Apple iOS app store)

The ever-changing mobile landscape is a challenging space to navigate. . The percentage of mobile over desktop is only increasing. Android holds about 53.2% of the smartphone market, while iOS is 43%. To get more people to download your app, you need to make sure they can easily find your app. Mobile app analytics is a great way to understand the existing strategy to drive growth and retention of future user.

With million of apps around nowadays, the following data set has become very key to getting top trending apps in iOS app store. This data set contains more than 7000 Apple iOS mobile application details. The data was extracted from the iTunes Search API at the Apple Inc website. R and linux web scraping tools were used for this study.

Data collection date (from API);

A Data Collection Date (From API): July 2017 (UPDATED 2018)

A Dimension of Dataset: 7197 rows and 16 columns (*id, track_name, size_bytes, currency, price, ratingcounttot, ratingcountver, user_rating, userratingver, ver, cont_rating, prime_genre, sup_devices.num, ipadSc_urls.num, lang.num, vpp_lic*).

https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps?select=appleStore_description.csv

Data Integration

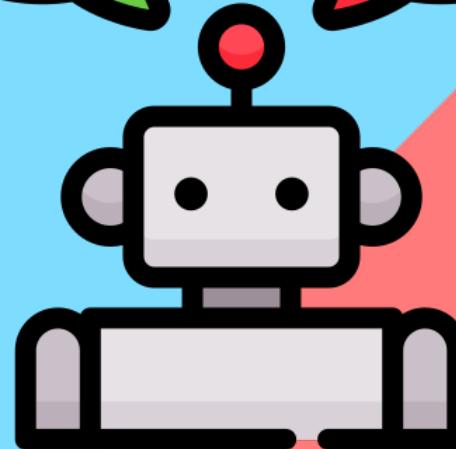
Both_Stores variable creation



Download on the
App Store



GET IT ON
Google Play



Download on the
App Store



GET IT ON
Google Play



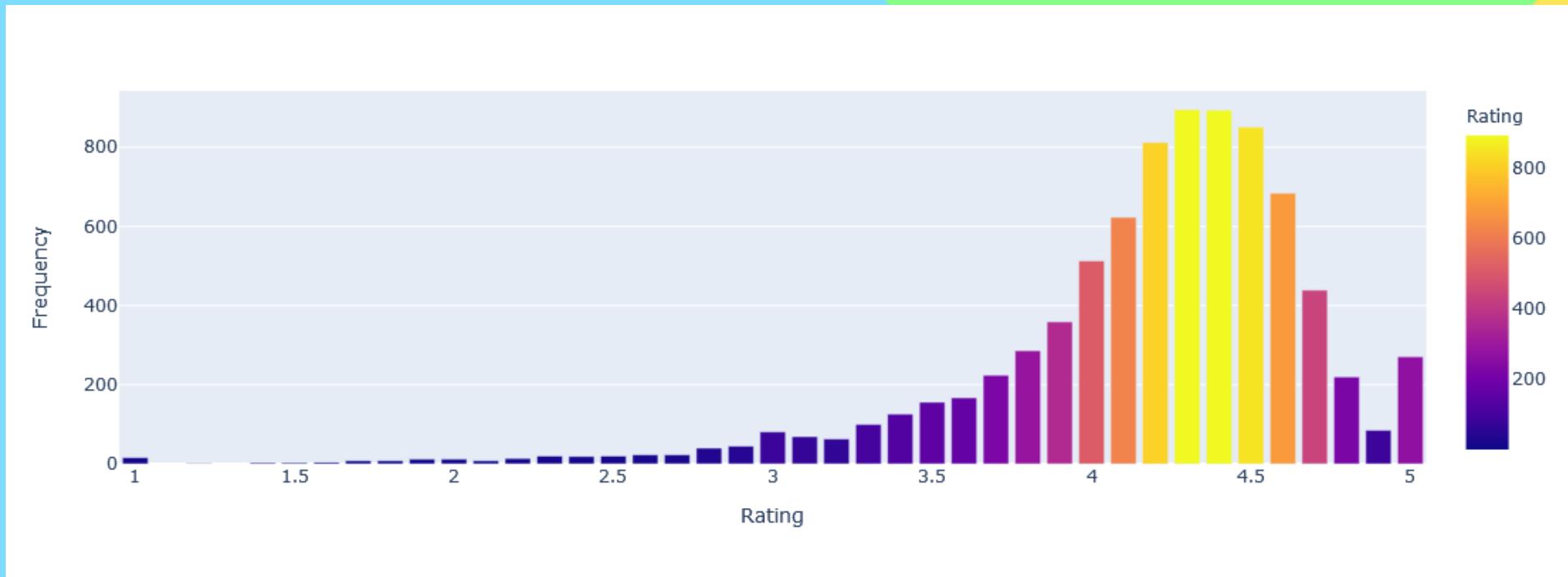
Data Integration



1		App	8		Last Updated	14		Is_Game
2		Rating	9		Android Ver	15		Compatibility
3		Reviews	10		Current Ver	16		RTR
4		Size	11		Main_Genres	17		Both_Stores
5		Installs	12		Secondary_Genres			
6		Price	13		Is_Family			
7		Content Rating						

Data Visualization

Distribution of Target variable



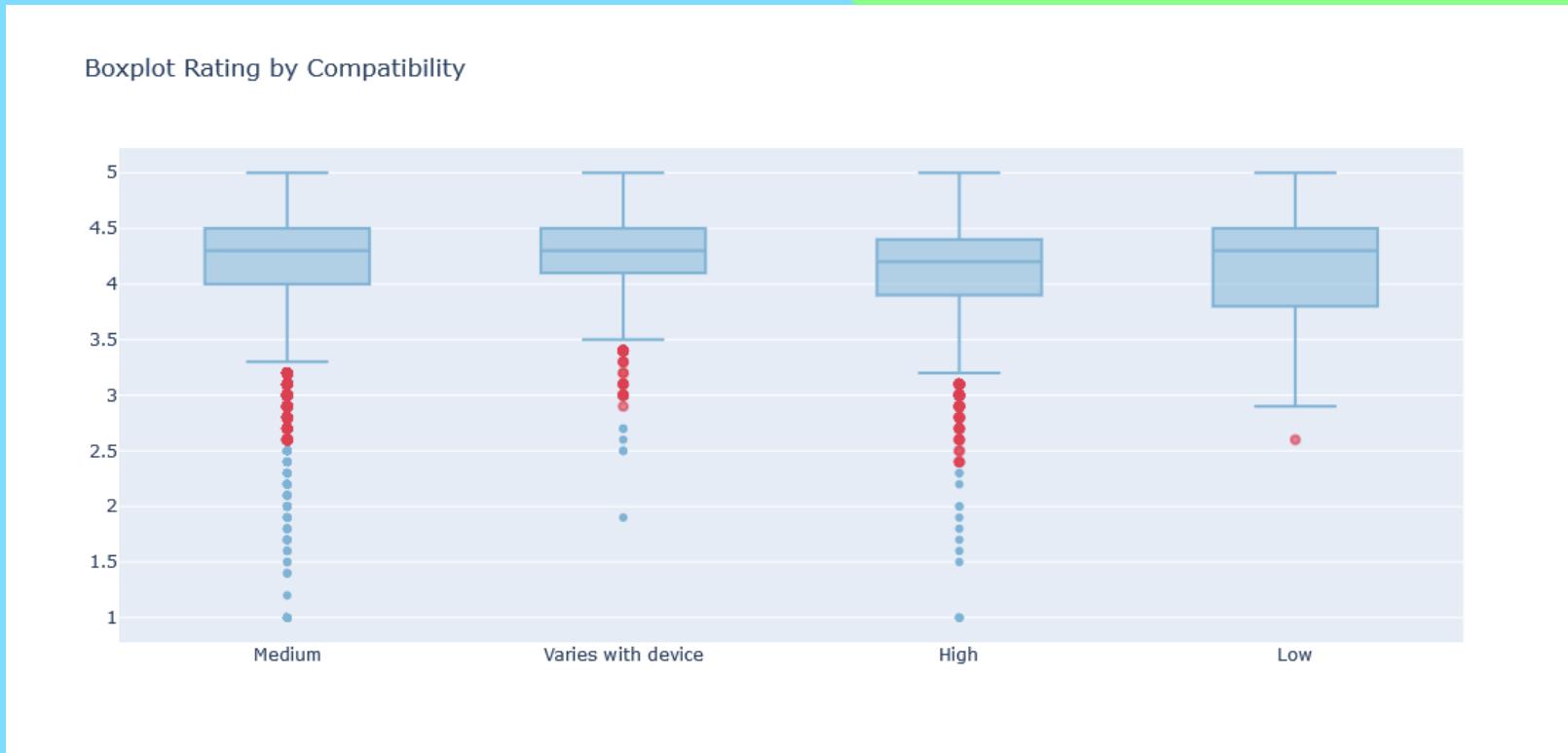
- ✓ Negative asymmetric distribution
- ✓ Frequency spike around 4.3
- ✓ Increasing in rating 1.0

Distribution of boolean variables



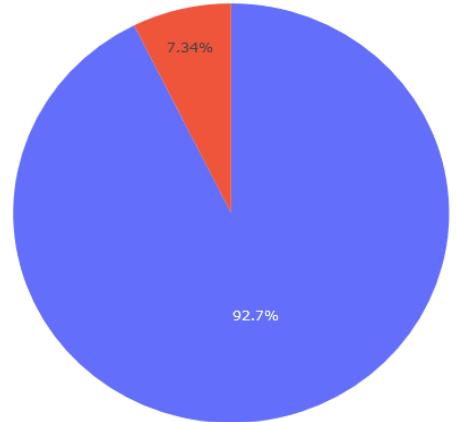
- ✓ No difference in the median rating if boolean variables are 'No'
- ✓ Main difference observed when the app is in both stores (Apple & PlayStore)
- ✓ No ratings ≤ 3.0 when the app is in both stores (Apple & PlayStore)

Distribution of Compatibility



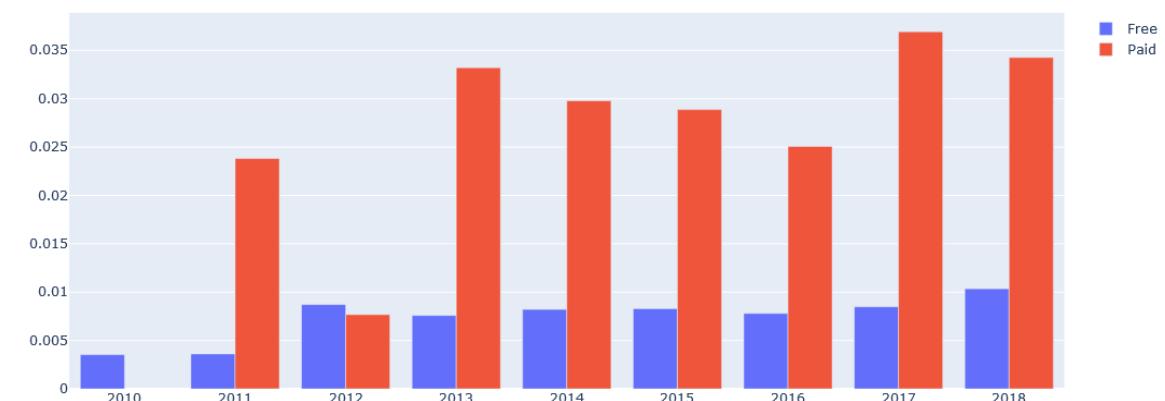
- ✓ "Varies with device" with the highest median rating and symmetric distribution
- ✓ "Medium" and "High" with the largest range of rating and more potential outliers
- ✓ "Low" with lowest values that may not be considered as potential outliers

Distribution of RTR by Last Updated and Price



✓ Unbalanced distribution for Price

✓ Slight increasing in the most recent updates (in 2018) for free apps



Task Proposal

