

LAVANYA · UPDATED 7 YEARS AGO

5099

<> Code

Download

Google Play Store Apps

Data of 10k Play Store apps for analysing the Android market.

Google

Data Card

Code (1178)

Discussion (80)

Suggestions (0)

About Dataset

[ADVISORY] IMPORTANT

Instructions for citation:

If you use this dataset anywhere in your work, kindly cite as the below:
L. Gupta, "Google Play Store Apps," Feb 2019. [Online]. Available: <https://www.kaggle.com/lava18/google-play-store-apps>

Context

While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. On digging deeper, I found out that iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

Content

Each app (row) has values for category, rating, size, and more.

Acknowledgements

This information is scraped from the Google Play Store. This app information would not be available without it.

Inspiration

Usability

8.24

License

CC BY-SA 4.0

Expected update frequency

Not specified

Tags

Computer Science

Instructions for citation:

If you use this dataset anywhere in your work, kindly cite as the below:
L. Gupta, "Google Play Store Apps," Feb 2019. [Online]. Available: <https://www.kaggle.com/lava18/google-play-store-apps>

Context

While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. On digging deeper, I found out that iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

Content

Each app (row) has values for category, rating, size, and more.

Acknowledgements

This information is scraped from the Google Play Store. This app information would not be available without it.

Inspiration

Data Card

Code (1178)

Discussion (80)

Suggestions (0)

This information is scraped from the Google Play Store. This app information would not be available without it.

Inspiration

View more

googleplaystore.csv (1.36 MB)

Download

Columns

More

Detail

Compact

Column

10 of 13 columns

About this file

Suggest Edits

details of the applications on Google Play. There are 13 features that describe a given app.. Explo. Ed

App	Category	Rating	Reviews	Size	Installs
Application name	Category the app belongs to	Overall user rating of the app (as when scraped)	Number of user reviews for the app (as when scraped)	Size of the app (as when scraped)	Number of installs of the app

Data Explorer

Version 6 (9.03 MB)

googleplaystore.csv

googleplaystore_user_review: license.txt

Summary

3 files

18 columns

From the screenshots below I have applied column quality, and column distribution from the view tab on my dataset.

The first data quality mistake was that in the googleplaystore_user_reviews table, the column names were counted as row and the columns were named as numbers. I simply renamed the columns and removed the first row.

`fx` `= table.TransformColumnTypes(source,{{"Column1", type text}}, {"Column2", type text}}, {"Column3", type text}}, {"Column4", type text}},`

A ^B _C Column1	A ^B _C Column2	A ^B _C Column3	A ^B _C Column4	A ^B _C Column5
<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% 	<ul style="list-style-type: none"> Valid 99% Error 0% Empty < 1% 	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% 	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% 	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%
1 App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity

The second data quality issue was on the Translated_Review column that had some missing values that were removed. To fix this I removed I unchecked the (blank) on the filter.

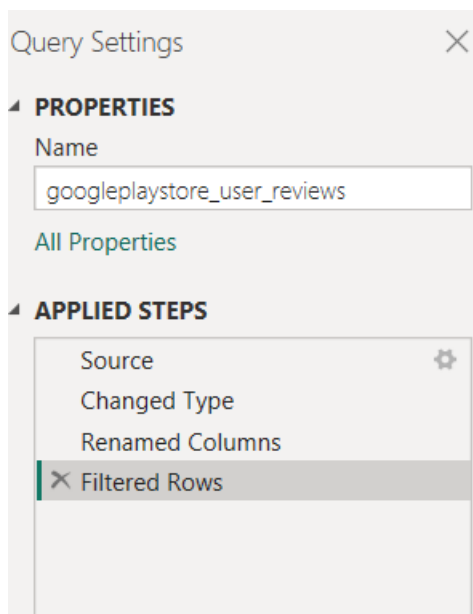
`SelectRows("#Renamed Columns", each ([APP] <> "App"))`

A ^B _C Translated_Review
<ul style="list-style-type: none"> Valid 99% Error 0% Empty < 1%

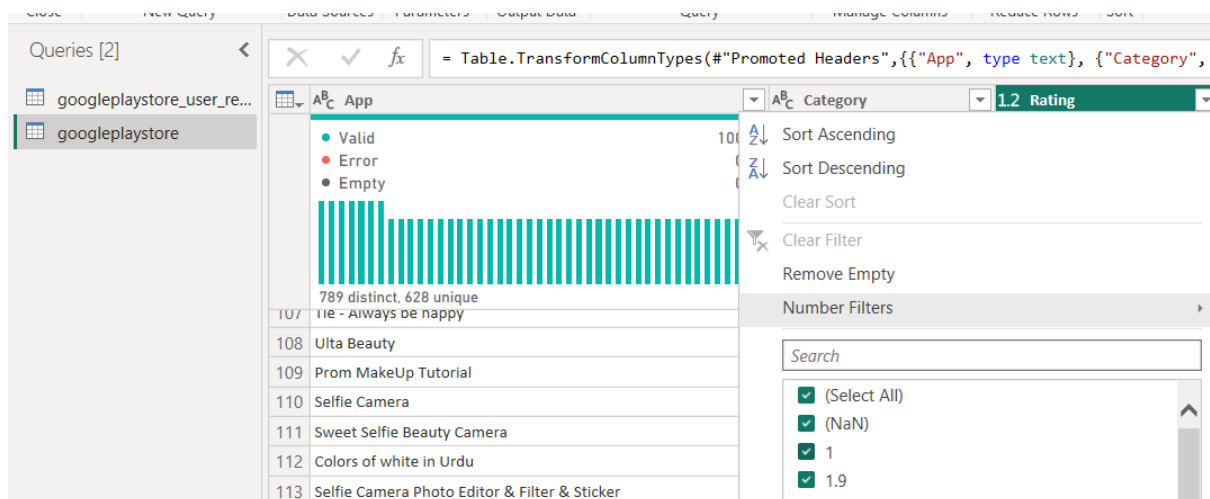
The third data quality issue was that there are rows marked as nan in the same table that had no use. To fix this I also unchecked nan value from one of the columns filter.

<ul style="list-style-type: none"> Error 0% Empty 0% 	<ul style="list-style-type: none"> Error 0% Empty 0% 	<ul style="list-style-type: none"> Error < 1% Empty 0% 	<ul style="list-style-type: none"> Error 0% Empty 0% 	<ul style="list-style-type: none"> Error 0% Empty 0% 	
1	10 Best Foods for You	I like eat delicious food. That's I'm cooking food myself, case "10 Best ...	Positive	1.0	0.5333333333333333
2	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.28846153846153844
3	10 Best Foods for You	nan	nan	nan	nan
4	10 Best Foods for You	Works great especially going grocery store	Positive	0.4	0.875
5	10 Best Foods for You	Best idea us	Positive	1.0	0.3
6	10 Best Foods for You	Best way	Positive	1.0	0.3
7	10 Best Foods for You	Amazing	Positive	0.6000000000000001	0.9
8	10 Best Foods for You	nan	nan	nan	nan
9	10 Best Foods for You	Looking forward app,	Neutral	0.0	0.0
10	10 Best Foods for You	It helpful site ! It help foods get !	Neutral	0.0	0.0
11	10 Best Foods for You	good you.	Positive	0.7	0.6000000000000001
12	10 Best Foods for You	Useful information The amount spelling errors questions validity infor...	Positive	0.2	0.1
13	10 Best Foods for You	Thank you! Great app!! Add arthritis, eyes, immunity, kidney/liver det...	Positive	0.75	0.875
14	10 Best Foods for You	Greatest ever Completely awesome maintain health.... This must ppl t...	Positive	0.9921875	0.8666666666666667
15	10 Best Foods for You	Good health..... Good health first priority.....	Positive	0.5499999999999999	0.5111111111111112
16	10 Best Foods for You	nan	nan	nan	nan
17	10 Best Foods for You	Health It's important world either life . think? :)	Positive	0.45	1.0

This is the applied steps for the issues above



Another data quality issue is on the table googleplaystore column Rating, where there are nan values as well. To fix this I did the same by unselecting (NaN).



Here are the applied steps for this tables issue

Query Settings

×

▲

PROPERTIES

Name

googleplaystore

All Properties

▲

APPLIED STEPS

Source

⚙

Promoted Headers

⚙

✕ Filtered Rows

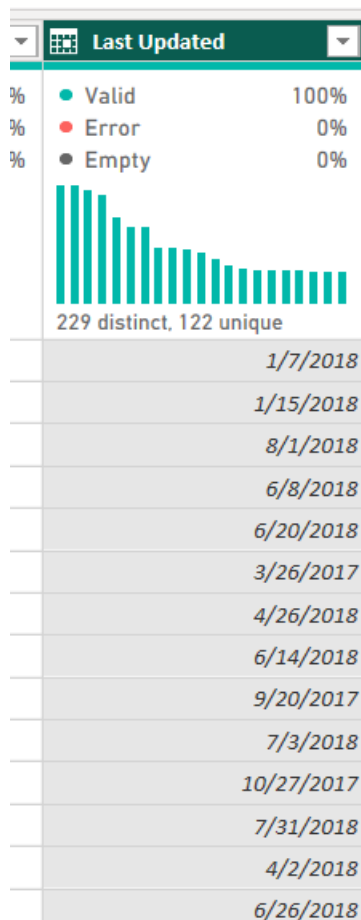
⚙

For q2

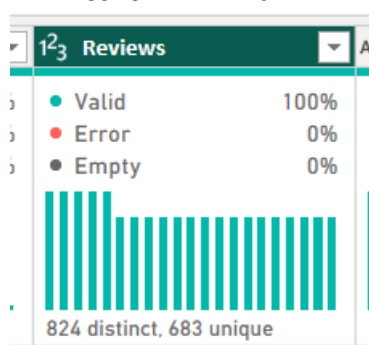
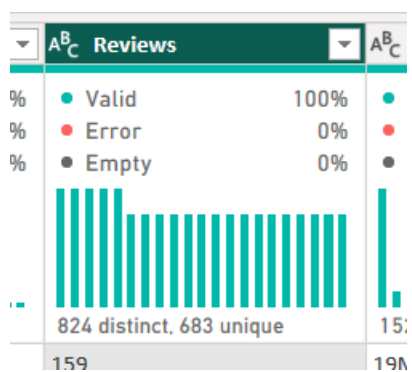
The date type was text and I changed it to date format



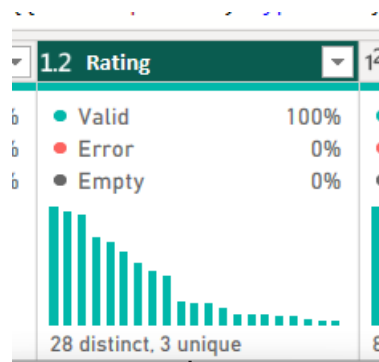
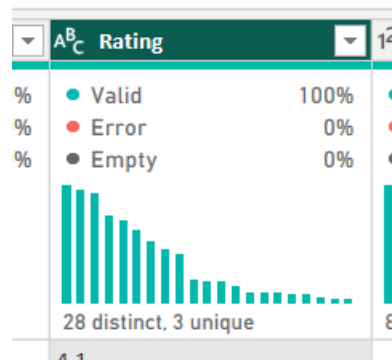
And here is after.



The column reviews was set as text as well and I changed it to a whole number, as the number of reviews made cannot be words. Here are the before and after.



The values on rating column were also set to text as there were some values that were labeled as NaN, I changed this to decimal number as it fits rating more. Here are the before and after.



The values on the price column have the currency sign \$ making them have text as their type. To fix this I used locale using the English(United States) to parse the \$ sign. Here are the before and after.

100

AB Type AB Price

Sort Ascending

Sort Descending

Clear Sort

Clear Filter

Remove Empty

Text Filters

Search

- ☒ (Select All)
- ☒ \$2.99
- ☒ \$3.99
- ☒ \$4.99
- ☒ \$5.99
- ☒ \$6.99
- ☒ \$7.99
- ☒ 0

List may be incomplete. Load more

AB Type 1.2 Price

Sort Ascending

Sort Descending

Clear Sort

Clear Filter

Remove Empty

Number Filters

Search

- ☒ (Select All)
- ☒ 0
- ☒ 2.99
- ☒ 3.99
- ☒ 4.99
- ☒ 5.99
- ☒ 6.99
- ☒ 7.99

List may be incomplete. Load more

OK Cancel

Here are the applied steps for question 2

Query Settings

×

PROPERTIES

Name

googleplaystore

All Properties

APPLIED STEPS

Source

⚙

Promoted Headers

⚙

Filtered Rows

⚙

Changed Type

⚙

✕ Changed Type with Locale

⚙

For q3 I trimmed, cleaning, capitalized, and split the android ver column. I also renamed the new columns and also filtered out and replaced values that were obvious in this column's case.

AB C Min Android Ver

AB C Max Android Ver

Valid 100%

Error 0%

Empty 0%

Valid 100%

Error 0%

Empty 0%

21 distinct, 3 unique

2 distinct, 0 unique

4.0.3

Up

4.0.3

Up

4.0.3

Up

4.2

Up

4.4

Up

2.3

Up

4.0.3

Up

4.2

Up

3.0

Up

4.0.3

Up

4.1

Up

4.0

Up

4.1

Up

PROPERTIES

Name

googleplaystore

All Properties

APPLIED STEPS

Source

⚙

Promoted Headers

⚙

Filtered Rows

⚙

Changed Type

⚙

Changed Type with Locale

⚙

Trimmed Text

⚙

Cleaned Text

⚙

Capitalized Each Word

⚙

Filtered Rows1

⚙

Split Column by Delimiter

⚙

Changed Type1

⚙

Renamed Columns

⚙

Replaced Value

⚙

✕ Replaced Value1

⚙

×

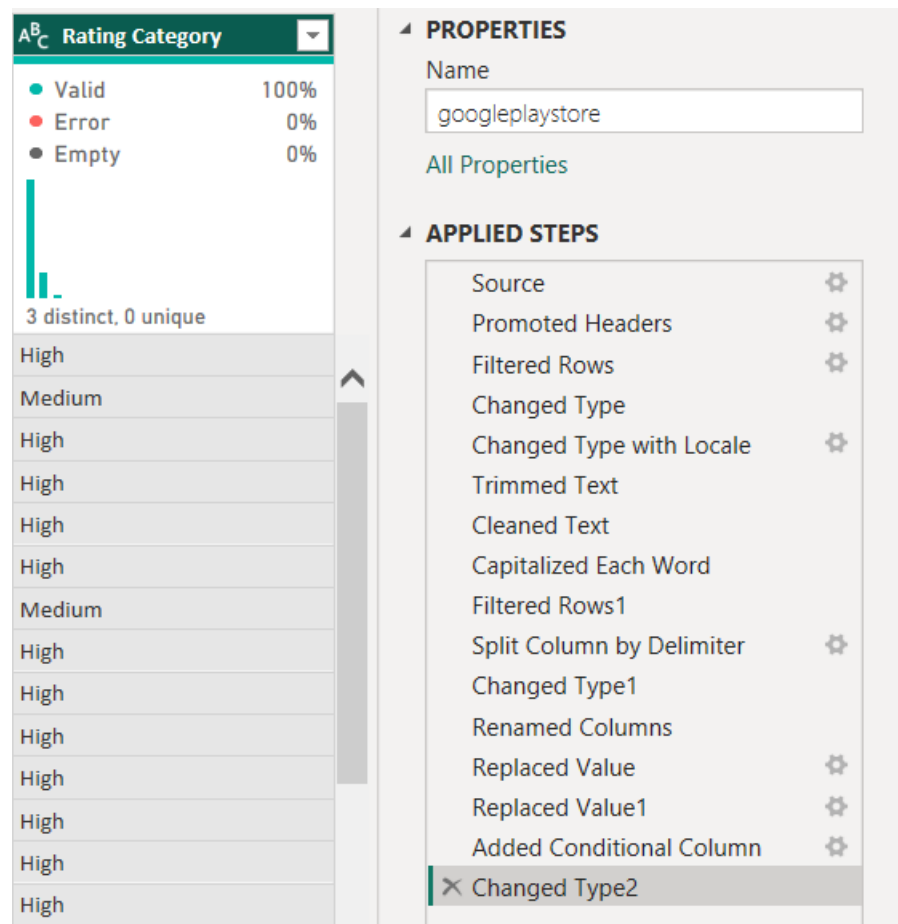
Add a conditional column that is computed from the other columns or values.

Rating Category

Add Clause

ABC
123 ▾ Low

Cancel



	\$ Price With Tax
Valid	100%
Error	0%
Empty	0%
7 distinct, 4 unique	
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00
	0.00

PROPERTIES

Name
googleplaystore

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Filtered Rows
- Changed Type
- Changed Type with Locale
- Trimmed Text
- Cleaned Text
- Capitalized Each Word
- Filtered Rows1
- Split Column by Delimiter
- Changed Type1
- Renamed Columns
- Replaced Value
- Replaced Value1
- Added Conditional Column
- Changed Type2
- Added Custom
- Changed Type3

For q5 I went with group by. I duplicated the table googleplaystore and renamed it category stats and grouped it by the below conditions

Group By

Specify the columns to group by and one or more outputs.

☐ Basic ☒ Advanced

Category

Add grouping

New column name

Total Apps

Operation

Count Rows

Column

Average Rating

Average

Rating

Add aggregation

OK

Cancel

Queries [3]

googleplaystore_user_re...
googleplaystore
Category Stats

Query Settings

Properties

Name
Category Stats

All Properties

Applied Steps

Source
Promoted Headers
Filtered Rows
Changed Type
Changed Type with Locale
Trimmed Text
Cleaned Text
Capitalized Each Word
Filtered Rows1
Split Column by Delimiter
Changed Type1
Renamed Columns
Replaced Value
Replaced Value1
Added Conditional Column
Changed Type2
Added Custom
Changed Type3
Grouped Rows

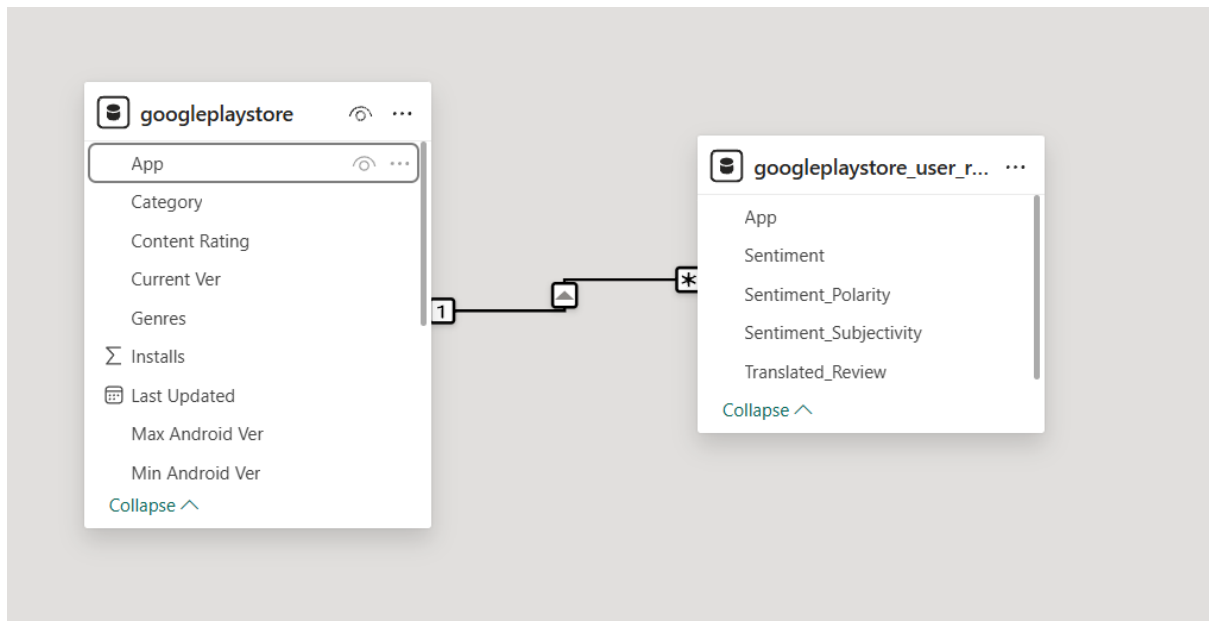
Table:

Category	Total Apps	Average Rating
ART_AND_DESIGN	59	4.376271186
AUTO_AND_VEHICLES	66	4.165151515
BEAUTY	39	4.297435897
BOOKS_AND_REFERENCE	147	4.325170068
BUSINESS	264	4.124621212
COMICS	49	4.159183673
COMMUNICATION	232	4.120689655
DATING	183	3.96557377
EDUCATION	114	4.394738842
ENTERTAINMENT	97	4.131958763
EVENTS	38	4.45
FINANCE	270	4.111111111
FOOD_AND_DRINK	87	4.095402299
HEALTH_AND_FITNESS	227	4.225991189
HOUSE_AND_HOME	56	4.1625
LIBRARIES_AND_DEMO	62	4.203225806
LIFESTYLE	285	4.08245614
GAME	1045	4.283157895
FAMILY	1673	4.191273162

Q6

googleplaystore_user_reviews is the fact table because it contains the events, reviews submitted by users and changes frequently. While googleplaystore is the dimension table it contains the descriptive attributes about the apps and serves as a lookup table with unique app names.

Q7



Tables: googleplaystore to googleplaystore_user_reviews

Cardinality: 1-to-Many

Cross filter direction: Single

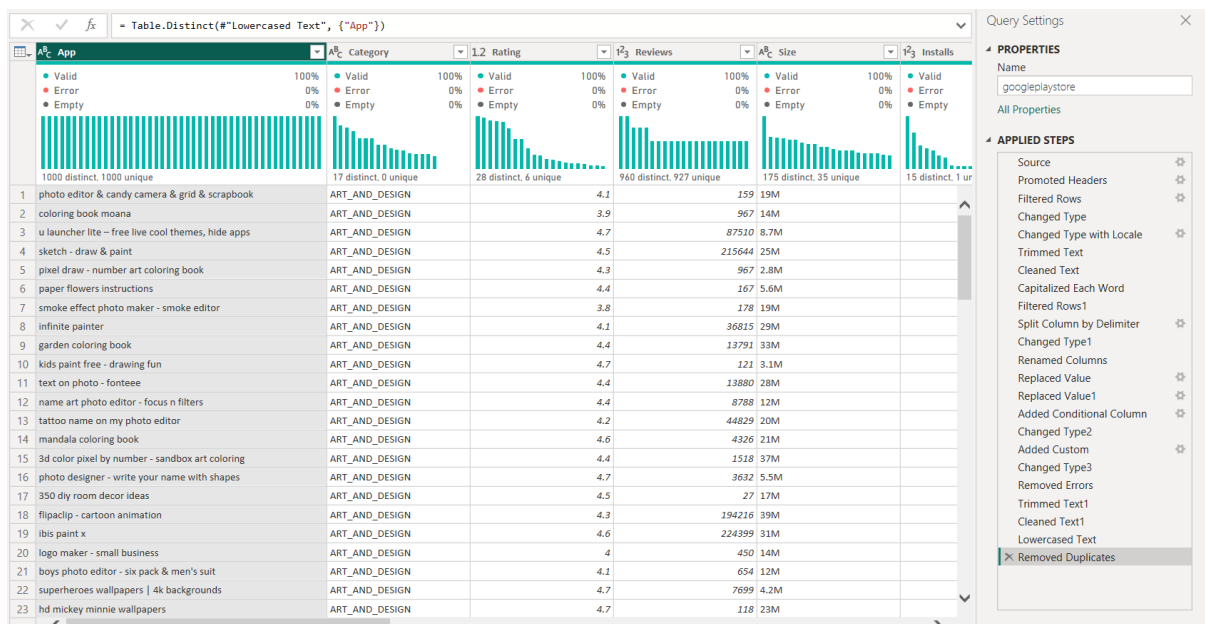
Active/inactive: Active

Why correct: This is correct because one single app in the dimension table can have many individual user reviews in the fact table.

A modeling mistake that would break totals is setting a many-to-many relationship instead of a 1-to-many relationship. If both tables had duplicate app names, the relationship would become many-to-many, which causes filter ambiguity and results in incorrect, inflated totals when visualizing the data.

Q8

I verified the dimension key (App) was unique by checking the model relationship. Power BI initially flagged a many-to-many cardinality, indicating duplicates existed. I checked for duplicates and resolved them in Power Query by Trimming the text, converting it to lowercase, and applying the 'Remove Duplicates' function to ensure strictly unique keys for a valid 1-to-many relationship.



Q9

Yes. It should exist.

1. It allows you to use Power BI's built-in Time Intelligence DAX functions like Year-to-Date or Month-over-Month calculations.
2. It acts as a central filter, allowing you to slice multiple fact tables using one continuous, standardized calendar without missing dates.

It should contain: Date, Year, and Month Name

Bonus points

Table	Column	Meaning	Type	Key/Attribute
googleplaystore	App	Unique name of the application	Text	Primary Key
googleplaystore	Price	Cost of the app (after locale conversion)	Currency	Attribute
googleplaystore	Rating Category	Custom conditional grouping (High/Medium/Low)	Text	Attribute
googleplaystore	Min Android Version	Minimum OS required (from split column step)	Text	Attribute
googleplaystore	Last Updated	The date the application was last updated	Date	Attribute
googleplaystore_user_reviews	App	Name of the app being reviewed	Text	Foreign Key
googleplaystore_user_reviews	Translated_Review	The user's written feedback in English	Text	Attribute
googleplaystore_user_reviews	Sentiment	The emotional tone of the review	Text	Attribute

Table	Column	Meaning	Type	Key/Attribute
		(Positive/Negative)		