



LAVANYA · UPDATED 7 YEARS AGO

▲ 5099

Code

Download



Google Play Store Apps

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



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

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.

Expected update frequency
Not specified

Tags

Computer Science
Internet
Video Games
Mobile and Wireless

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

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)	Numb downl the ap

Data Explorer

Version 6 (9.03 MB)

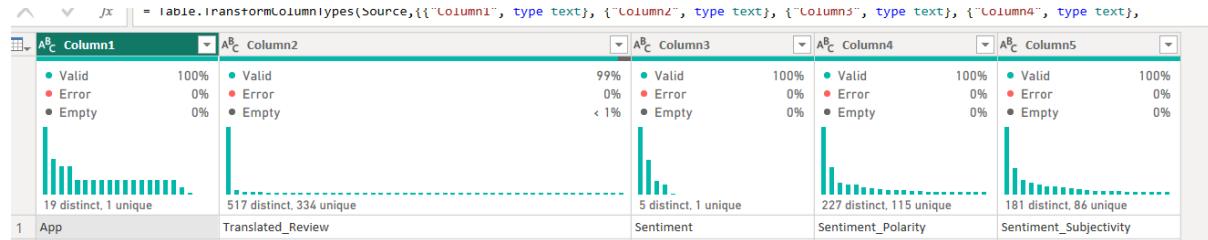
googleplaystore.csv
googleplaystore_user_reviews.csv
license.txt

Summary

3 files
18 columns

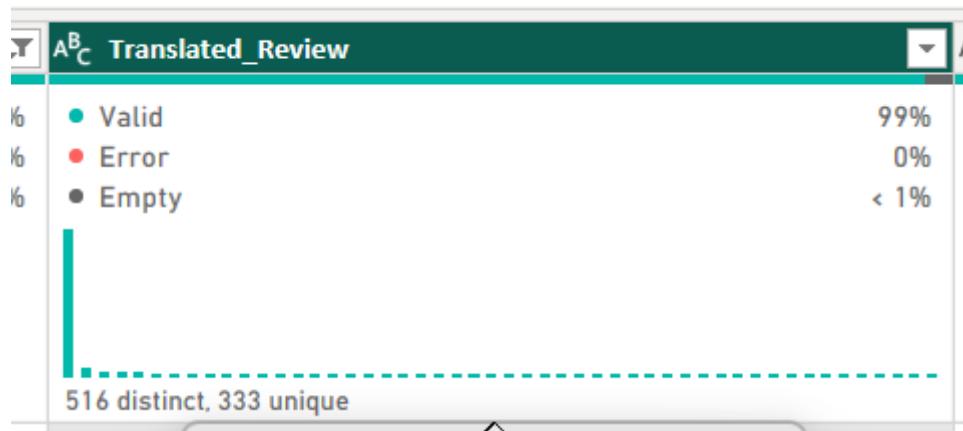
From the screenshots below I have applied column quality, and column distribution from the view tab on my dataset.

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.

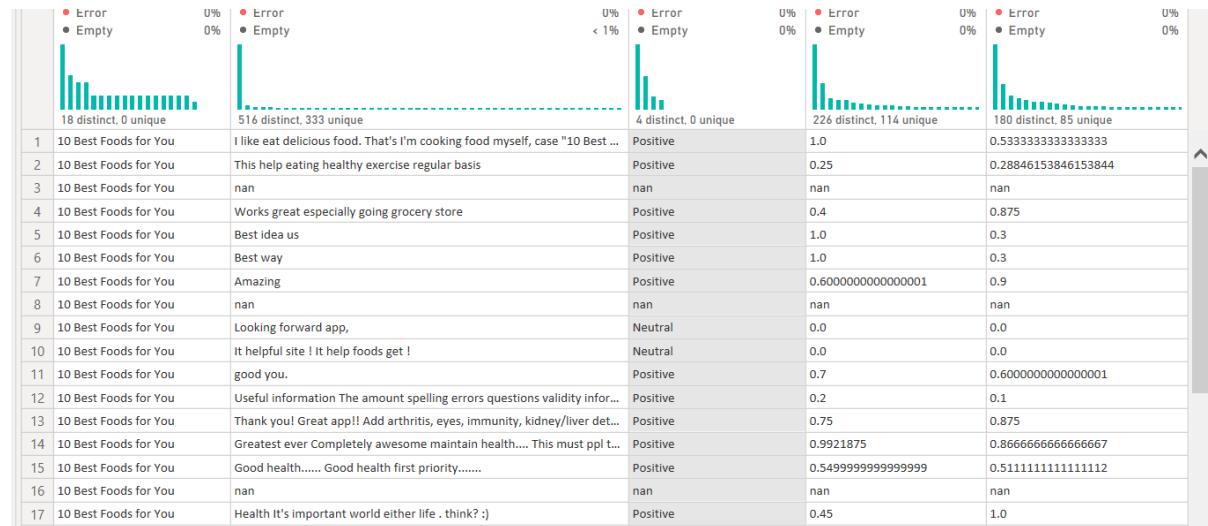


The second data quality issue was on the Translated_Review column that had some missing values that were removed. To fix this I removed I unchecked the (blank) on the filter.

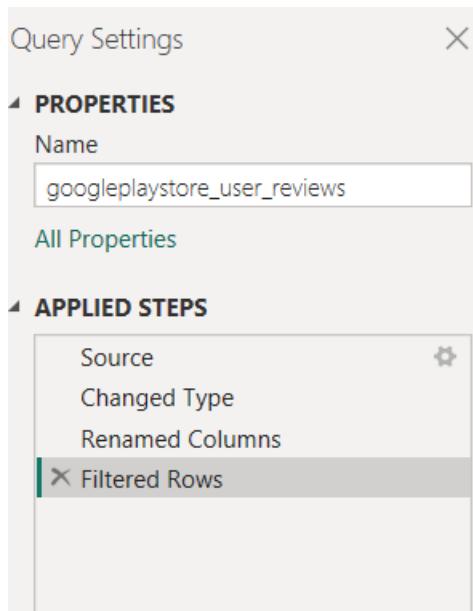
```
SelectRows(#"Renamed Columns", each ([APP] <> "App"))
```



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.



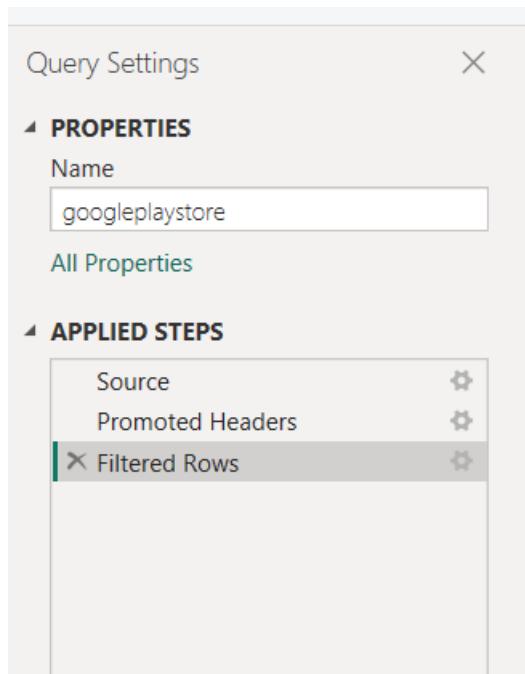
This is the applied steps for the issues above



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).

The screenshot shows the Power BI Data View. A table named 'googleplaystore' is displayed with columns 'App', 'Category', and 'Rating'. The 'Rating' column has a dropdown menu open, showing filter options: '(Select All)', '(NaN)', '1', '1.9', and '2'. The table data includes rows such as '108 Ulta Beauty', '109 Prom MakeUp Tutorial', '110 Selfie Camera', '111 Sweet Selfie Beauty Camera', '112 Colors of white in Urdu', and '113 Selfie Camera Photo Editor & Filter & Sticker'. A tooltip indicates '789 distinct, 628 unique' entries.

Here are the applied steps for this tables issue

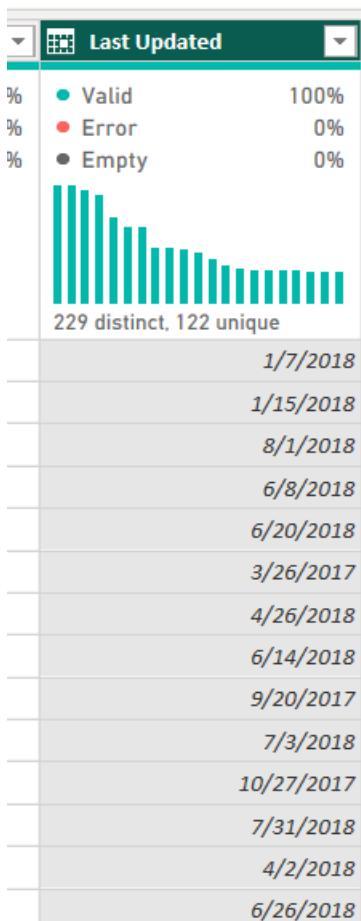


For q2

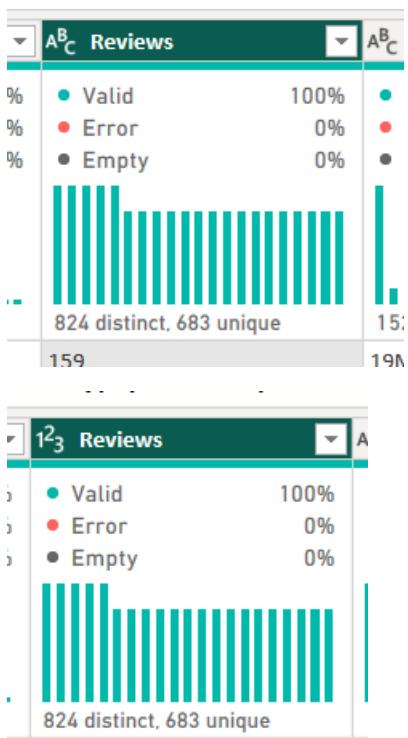
The date type was text and I changed it to date formart



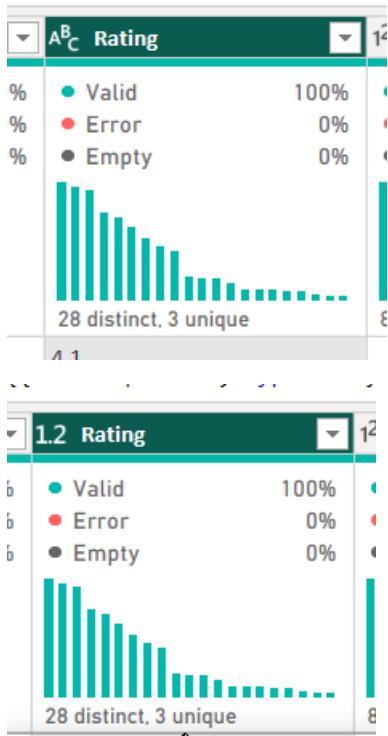
And here is after.



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 rows filtered from 100 total rows , last updated , type values ,

A_B_C Type A_B_C 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

A_B_C 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 X

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.

	Min Android Ver	Max Android Ver
6	Valid 100%	Valid 100%
6	Error 0%	Error 0%
6	Empty 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**

For q4 I created a column called Ratings category and here are the conditions and results.

Add Conditional Column

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

New column name
Rating Category

	Column Name	Operator	Value ⓘ	Output ⓘ
If	Rating	is greater than or...	ABC 123 ⓘ 4	Then ABC 123 ⓘ High
Else If	Rating	is greater than or...	ABC 123 ⓘ 2.5	Then ABC 123 ⓘ Medium
...				

Add Clause

Else ⓘ ABC 123 ⓘ Low

OK **Cancel**

For the custom column here is the formula and results.

Custom Column

Add a column that is computed from the other columns.

New column name

Custom column formula ⓘ

```
= [Price]*1.08
```

Available columns

- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content_Rating

<< Insert

[Learn about Power Query formulas](#)

✓ No syntax errors have been detected.

OK Cancel

Price With Tax

	Valid	100%
Valid	0.00	100%
Error	0.00	0%
Empty	0.00	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
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

X

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

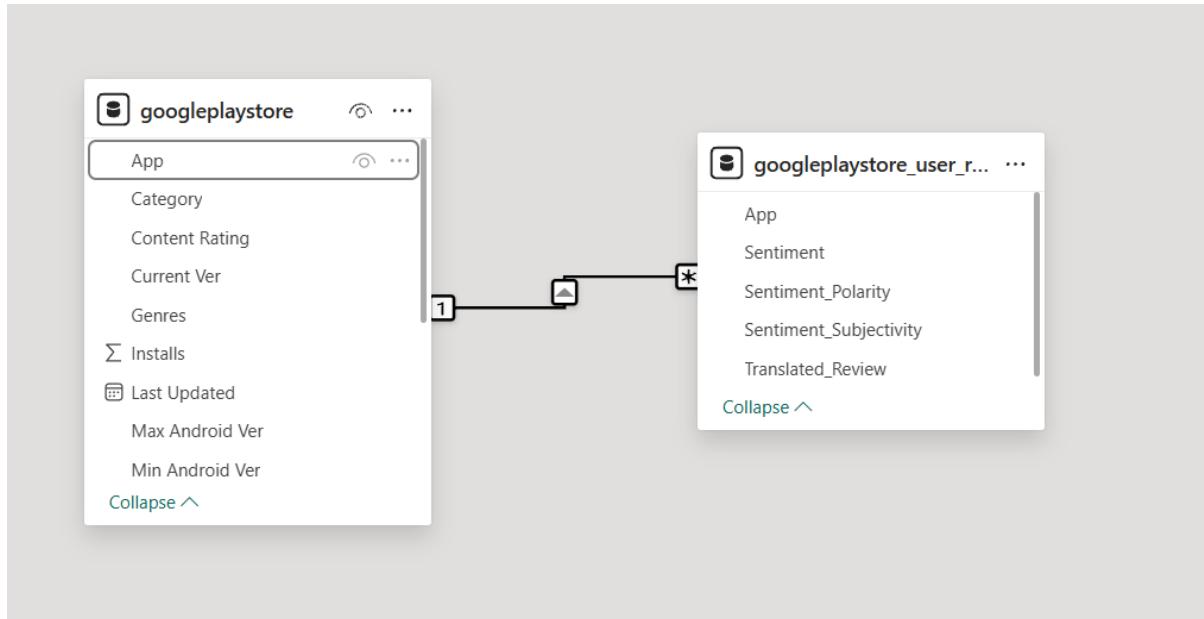
The screenshot shows the Power BI Query Editor interface. On the left, there's a tree view of queries: 'googleplaystore_user_re...' (selected), 'googleplaystore', and 'Category Stats'. The main area displays a grouped table with three columns: 'Category' (with a dropdown menu showing 'ART_AND DESIGN', 'AUTO_AND_VEHICLES', etc.), 'Total Apps' (with a dropdown menu showing '34 distinct, 34 unique'), and 'Average Rating' (with a dropdown menu showing '34 distinct, 34 unique'). Below the table, there are three bar charts showing the distribution of 'Valid', 'Error', and 'Empty' values for each category. The right side of the screen has a 'Query Settings' pane with 'Properties' (Name: Category Stats) and 'Applied Steps' (listing various data transformation steps like Promoted Headers, Filtered Rows, etc., with 'Grouped Rows' highlighted).

Category	Total Apps	Average Rating
ART_AND DESIGN	34 distinct, 34 unique	34 distinct, 34 unique
AUTO_AND VEHICLES	32 distinct, 30 unique	32 distinct, 30 unique
BEAUTY	59	59
BOOKS_AND_REFERENCE	66	66
BUSINESS	39	39
COMICS	147	147
COMMUNICATION	264	264
DATING	49	49
EDUCATION	232	232
ENTERTAINMENT	183	183
EVENTS	114	114
FINANCE	4.394738842	4.394738842
FOOD_AND_DRINK	97	97
HEALTH_AND_FITNESS	38	38
HOUSE_AND_HOME	4.131958763	4.131958763
LITERATURE_AND_DMO	38	38
LIFESTYLE	270	270
GAME	87	87
FAMILY	4.095402299	4.095402299
	227	227
	56	56
	62	62
	285	285
	1045	1045
	1673	1673

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

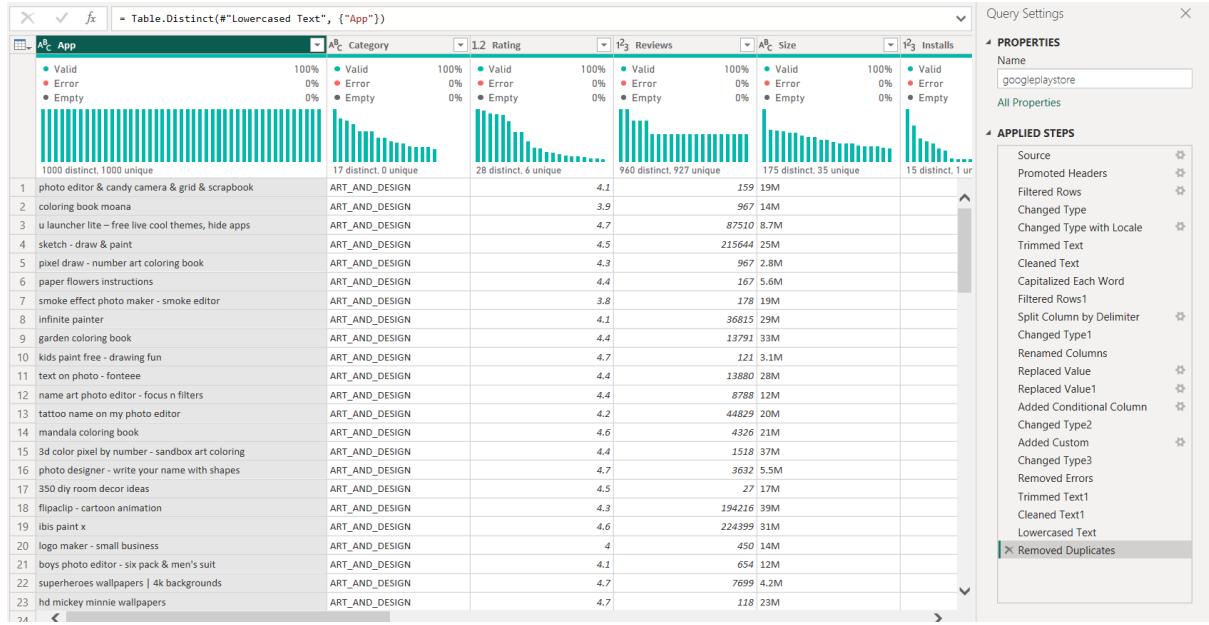
Active/inactive: Active

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)		