Name: **DSI_kickstarterscrape_dataset..csv**
Author: Unknown
Description:
As of 2020, over 500,000 kickstarter campaigns have been launched. With a population of nearly 500,000 and a sample size of 50,000, the confidence level possible with this dataset can be as high as 99% with a margin of error less than 1%. This is pretty excellent in terms of statistical significance. The latest project appears to be in 2012, which means the data isn't particularly current and, otherwise, information about the source and reliability of this assumedly third party data is lacking. The dataset is none-the-less useful for preliminary analysis.

Cleaning Change Log
kickstarter _casestudy - DSI_kickstarterscrape_dataset.csv
February 10, 2022

1. Opened **DSI_kickstarterscrape_dataset..csv** file in Google Sheets for preliminary cleaning and reformatting. The limited number of rows in this dataset made this possible.
2. I checked for NULL values in all columns and none were found except under *pledged*, *location*, and *renewed levels* columns. The 12 NULL values in the *pledged* column were replaced with values calculated from the *goal* and *funded percentage* columns. The NULL values of the *reward levels* column all only corresponded to rows with 0 in the *levels* column. *Location* column NULLS could not be replaced or justified without greater inquiry.
3. Three values under *location* contained non-sense Chinese characters. These characters were removed so that values read only as "Sweden" and "Egypt" respectively.
4. Some values in the *category* and *subcategory* columns contained erroneous inclusions of "amp". These were found and replaced with appropriate values without these extra letters.
5. The "Remove Duplicates" add-on was used to identify and delete 162 duplicate rows.
6. *Pledged* and *goal* columns were converted to currency data types.
7. Using values in the *funded date* and *duration* columns, a *launch_date_time* column was produced and converted to a date/time data type.
8. *Month, day*, and *time* columns were then produced from this column and converted to date and time data types respectively.
9. Not all project ids were the same character length, but I assume this is acceptable for now.
10. Some values in the *name* column appeared to have extraneous strings in them like "amp" and "quot", but these names will not be used to run analysis, so they will be unaltered for now.
11. *Funded percentage* was renamed *funded_percentage_asdec* and duplicated into another row as *funded_percentage_asper* and converted to a percentage data type.
12. *project id* column was renamed *project_id.*
13. *reward levels* column was renamed *reward_levels*.
14. *funded date* was renamed *Funded Date/Time (OG)*.
15. Whitespace was trimmed from 993 cells in the *name* column and one cell in the *location* column.
16. This dataset was downloaded to a local folder "CLEANED", and renamed **kickstarter _casestudy - DSI_kickstarterscrape_dataset.csv**
17. Given the size of the file and the chance that loading error may have caused some of the duplication and glitches in cleaning, this file was uploaded to BigQuery to double check for duplicates and NULL values with the following queries:

```
SELECT *
FROM ks_casestudy.ks_casestudy
WHERE
project_id IS NULL or name IS NULL or url IS NULL or category IS NULL or subcategory IS NULL
or status IS NULL or goal IS NULL or funded_percentage_asper IS NULL
or funded_percentage_asdec IS NULL or backers IS NULL or day IS NULL or month IS NULL or time IS NULL
or launch_date_time IS NULL or funded_date_time IS NULL or Funded_Date_Time__OG_ IS NULL or levels IS NULL
or updates IS NULL or comments IS NULL or duration IS NULL
```

18. Running the following queries confirmed that there were 0 NULL values in the *pledged* column, 58 null values in the *reward levels* column, and 1317 NULL values in the *location* column of the respective datasets:

```
SELECT *
FROM ks_casestudy.ks_casestudy
WHERE
pledged IS NULL
```

```
SELECT *
FROM ks_casestudy.ks_casestudy
WHERE
reward_levels IS NULL
```

```
SELECT *
FROM ks_casestudy.ks_casestudy
WHERE
location IS NULL
```

19. Running the following query confirmed that there were no longer any duplicate values under *project_id*:

```
SELECT project_id, COUNT(*)
FROM ks_casestudy.ks_casestudy
GROUP BY project_id
HAVING COUNT(*) > 1
```