

Sandbox Data Summary

Wandy Halim

Table of Contents

| | |
|--|----|
| Data Exploration..... | 2 |
| General Summary | 2 |
| Benefits of Analysis. | 2 |
| Attributes..... | 3 |
| user_pseudo_id..... | 3 |
| sku..... | 4 |
| app_version | 5 |
| geo | 6 |
| install_source..... | 7 |
| ua | 8 |
| device | 10 |
| idfa | 15 |
| idfv | 15 |
| table_date | 15 |
| is_returning_user and session_id | 15 |
| Methodology | 16 |
| Data Cleaning | 16 |
| Remove Unnecessary Columns..... | 16 |
| Replacing Error Inputs..... | 16 |
| Data Preprocessing | 17 |
| timestamp_raw..... | 17 |
| device_language..... | 17 |
| is_limited_ad_tracking..... | 18 |
| Missing Value Handling | 19 |
| Additional Insight | 21 |
| Limited Ad Tracking in Several Countries..... | 21 |
| Download Popularity in Certain Hour | 23 |
| Most Downloaded Language Version of the App (Based on os)..... | 24 |
| Conclusion | 24 |

Data Exploration

General Summary

Generally, the dataset consists of the application installment information across different devices, application store source, location, time zone, etc. The installment information is based on the user who installed the application, represented by their pseudo ids. **The data records the related information of users who downloaded the application from 30th of September (afternoon, 2 pm) to 31st of September (midnight, 11:59 pm).** For instance, one of the data rows explains that the user is downloading the app from google play. The app version is 1.20.1 and the user is downloading it in India. The source of installation is from android vending. The device on which the user installed the apps is a Sony mobile device running on android with os version 5.1.1. The timezone offset is -18000 and the user is downloading it at 3 pm on 30th of December. That is an example of how the data describes the downloader/installer information.

The numbers of rows in the dataset are 16000 with 25 columns/attributes. Numerical data (float data) does not exist in this database, thus statistical analysis and correlation value will not be applied in the analysis processes. (Figure 1)

```
print('Number of rows in the dataset: ',df.shape[0])
print('Number of columns in the dataset: ',df.shape[1])
```

```
Number of rows in the dataset: 16000
Number of columns in the dataset: 25
```

Figure 1 Number of rows in dataset – Jupyter Notebook

Benefits of Analysis.

From the dataset, the author believes that by further analysing the dataset (applying data preprocessing and analysis methodology), It will gain several benefits such as:

- Understanding of user's download pattern behavior and enabling a clearer insight of potential users
- Determine regional-based popularity.
- Determine the effective time for maintenance or update of the app.
- Evaluate marketing strategy based on the data. (which marketing strategy works the best)
- Better utilised idfa for targeted audience ads

Attributes

The dataset contains 25 attributes; however, several attribute columns are irrelevant for further analysis (will be explained in a later section). Below is the description of all the attributes with their insights:

user_pseudo_id

user_pseudo_id is the user id who download the app from the store. **Most of the ids are unique to every row in this dataset, and there is no null value in the columns. However, during the finding, there are two reoccurring/duplicated pseudo ids** (as shown in *Figure 2*)

```
df[df.duplicated(subset=['user_pseudo_id'], keep='first')]
```

| | user_pseudo_id | sku | app_version | geo_country | geo_region | geo_city | install_source | ua_name | ua_medium | ua_source |
|-------|----------------------------------|-------------|-------------|-------------|------------------|-------------|---------------------|----------|-----------|-------------|
| 6129 | 1129502c52fb26add2b858e4bca7f12e | Google Play | 1.20.1 | India | Gujarat | Ahmedabad | com.android.vending | (direct) | (none) | (direct) |
| 15692 | e0d0b3b9ef0ab35ecae96e6476194159 | Google Play | 1.20.1 | Brazil | State of Paraiba | Joao Pessoa | com.android.vending | NaN | organic | google-play |

2 rows × 25 columns

Figure 2 Duplicated Pseudo Ids – Jupyter Notebook

There are no significant differences in the duplicated user_pseudo_id. For instance, data on row 6124 and 6131 have the same id; however, the differences are just on the region and the city where the user installed it, having the rest of the column with the same values (device, source, etc.). The same case goes to data on rows 15683 and 15694, both having almost the same data with a different download region.

sku

Count by sku

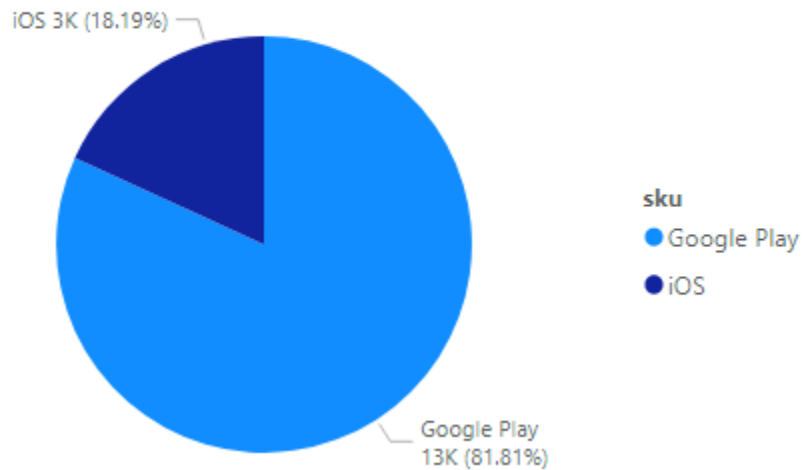


Figure 3 sku Distribution

sku describes the application store used for the installment. The value is categorical, ranging only from google play and ios. Figure 2 shows the distribution of sku in the dataset. As shown in *Figure 3*, **most of the users downloading the application are doing it from google play store**, populating the majority of the dataset (81.8 percent), and the rest, 18.19 percent, are doing it on the ios app store.

app_version

app_version column represents the version of the application which the user downloaded. The range of the value is from the latest version, 1.20.1, to the oldest version of 1.12.10. There is no null value in this column.

Row Count by app_version

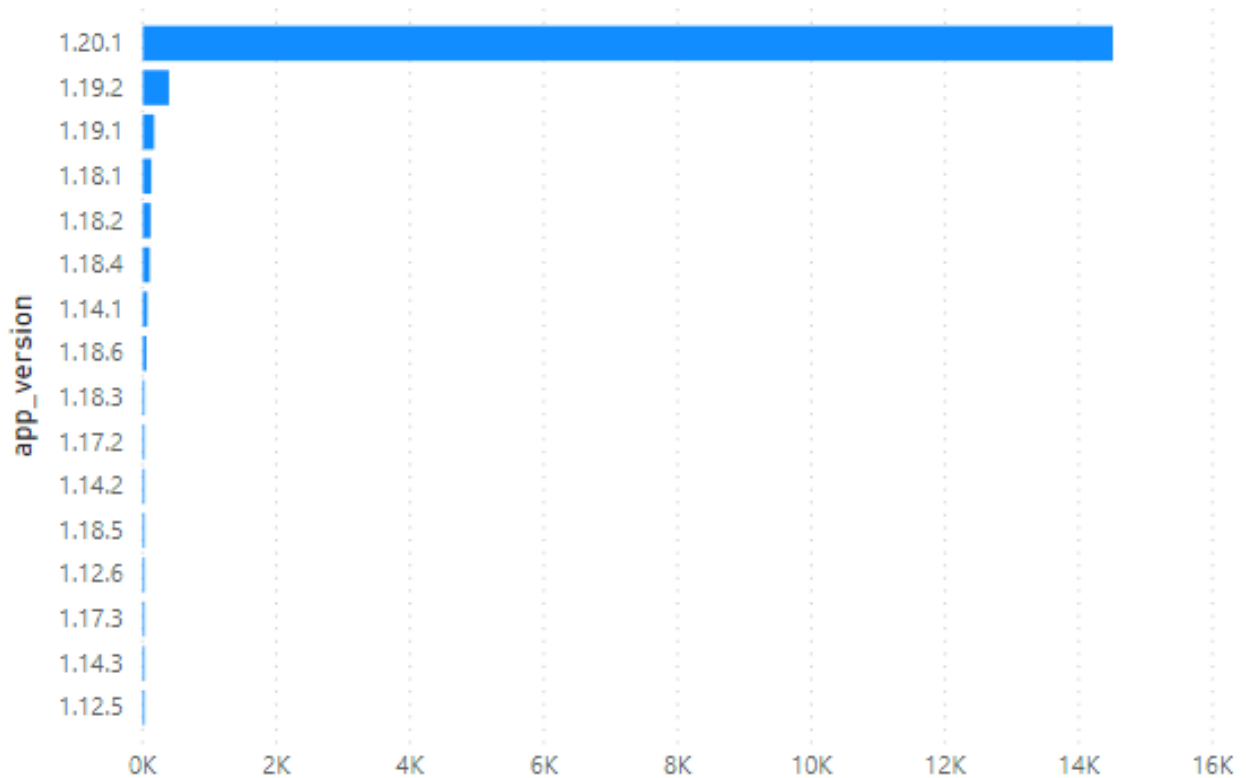


Figure 4 app version count

The barchart (*Figure 4*) shows that **the majority of user is downloading the latest version of the app (1.20.1) reaching the 90 percent portion of the dataset and decreasing as the version gets older.** The findings also describe that the timestamp of the download is independent of the version of the application.

geo

Describes the location information where the user installs the application. Those attributes are:

geo_country

geo_region

geo_city

Several errors of input are found in *geo_city* (as shown in *Figure 5*). Date are mistakenly inputted as *geo_city*. The row of this value needs to be preprocessed during the analysis methodology

| <i>geo_city</i> | Count |
|---------------------|-------|
| | 3233 |
| 6th of October City | 9 |
| A Coruna | 7 |
| Aba | 1 |
| Abano Terme | 1 |
| Acailandia | 1 |

Figure 5 geo_city , input error

Based on the map graph in *Figure 6* and treemap in *Figure 7*, **country such as India, China, Brazil and Italy have the highest amount of app download**. The blue circle at *Figure 6* represents a more specific city location of the user.



Figure 6 geo_city Distribution



Figure 7 geo_country Treemap

install_source

install_source is the source of installation of the application. Based on the dataset, **every user who downloads the app from the ios app store will always use iTunes or manual install as their only source of installation (Figure 8)**. On the other hand, if the user is downloading from google play. **There are many installation sources varying from the official android vending to third party installation source.** Most installations are from android official vending

sku ● Google Play ● iOS

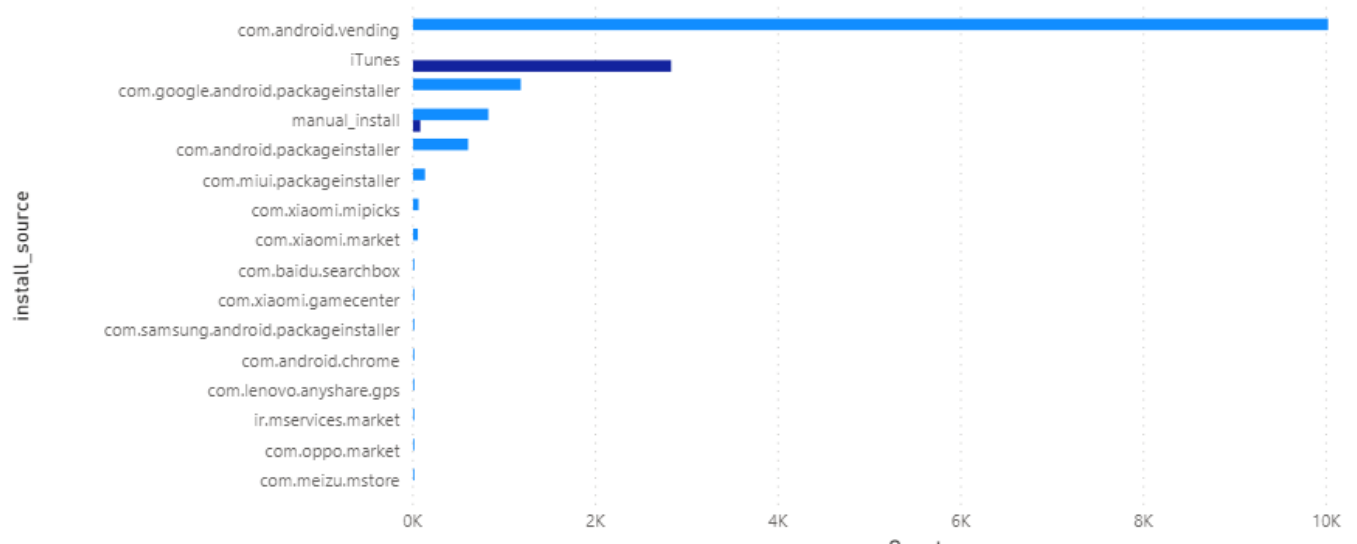


Figure 8 sku and install_source

ua

describes the information about the user's ua. There are three attributes of the ua:

ua_name

ua_medium

ua_source

As shown in the bar chart below,

- Every user who downloads from the ios store will always have no medium with a direct ua name and ua source. (refer to *Figure 9*)
- Dynamic link and organic medium are used on the majority of google play download.
- Most of the downloads by using google play are not using any ua_name (blank) (refer to *Figure 10*)

sku and ua_medium

ua_medium ● (none) ● dynamic_link ● organic

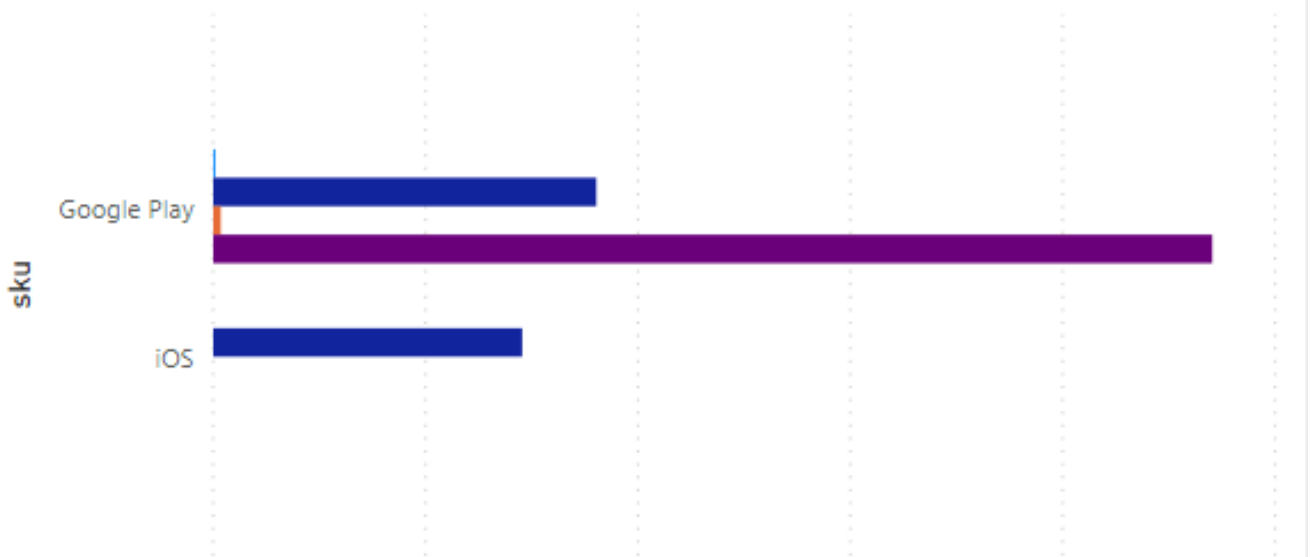


Figure 9 sku and ua medium

sku and ua_name

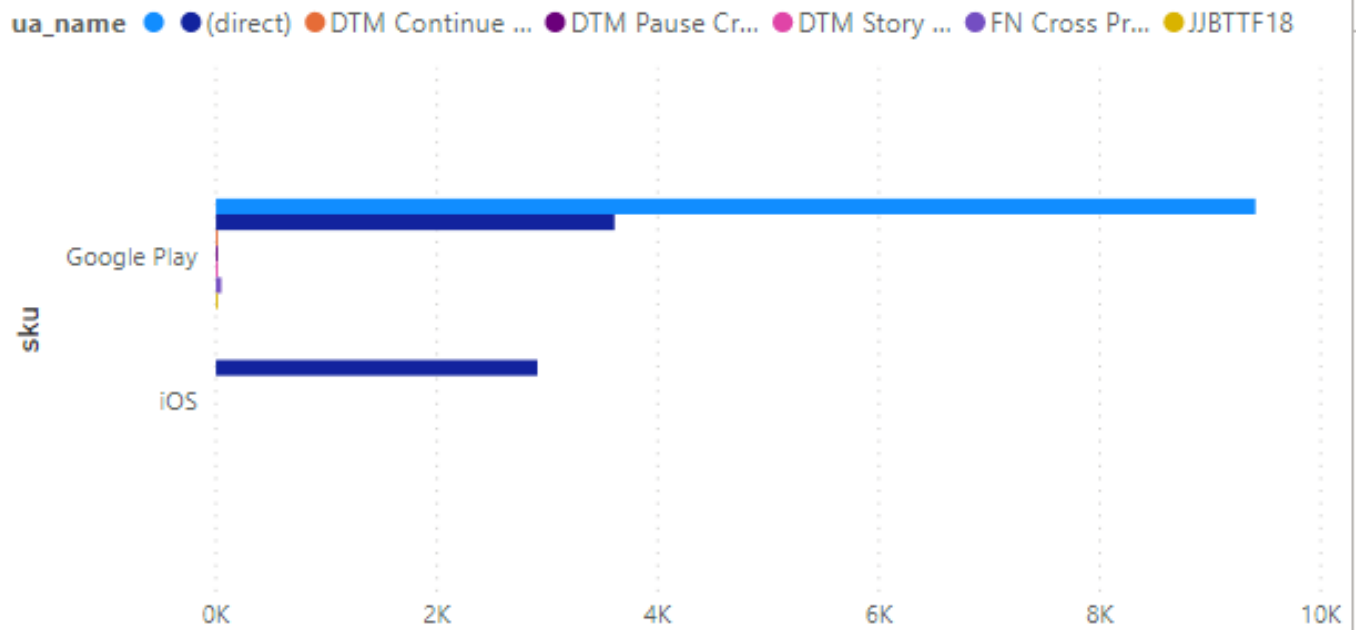


Figure 10 sku and ua_name

sku and ua_source

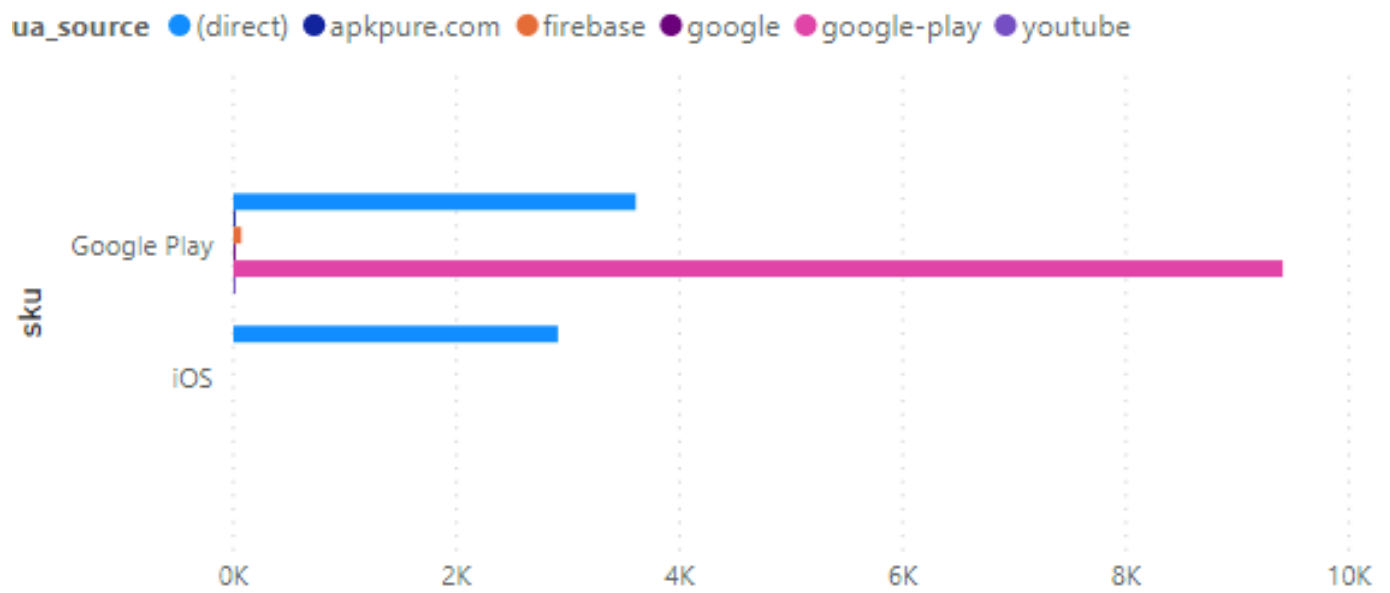


Figure 11 sku and ua_source

device

Describe the information of the device of the user. The attributes are:

device_category

describe the category of the device (i.e., mobile, tablet). **Figure 12 shows that most downloads are from a mobile phone (88.45 %)**

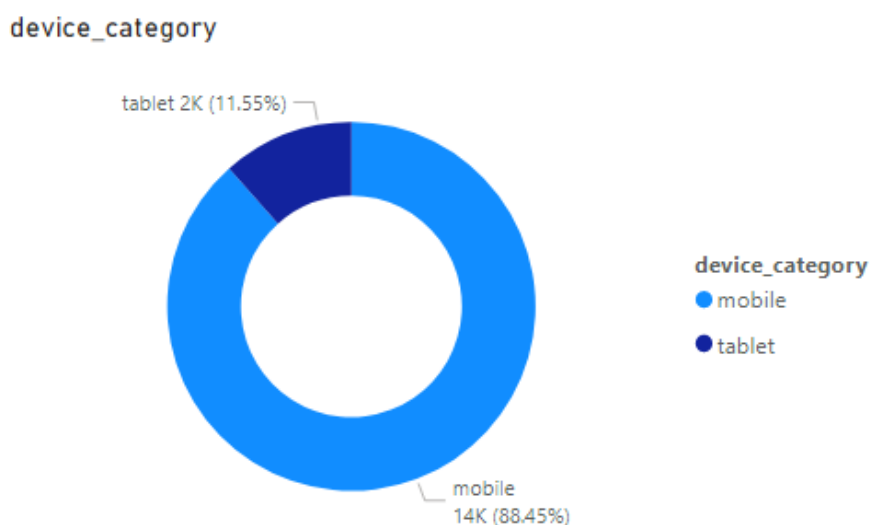


Figure 12 device_category

device_model_name

describes the specific model/series of the device (i.e., iPhone 5, Samsung SM-A105JH). Although most downloads are by users using Samsung as their devices, **the highest amount of models goes to iPhone 7, followed by iPhone 6 and iPhone 6s**. However, there are 649 rows with an empty device model name, which populate the largest portion of the dataset. All of these missing model names are from a device with an unlisted model name running on android OS.

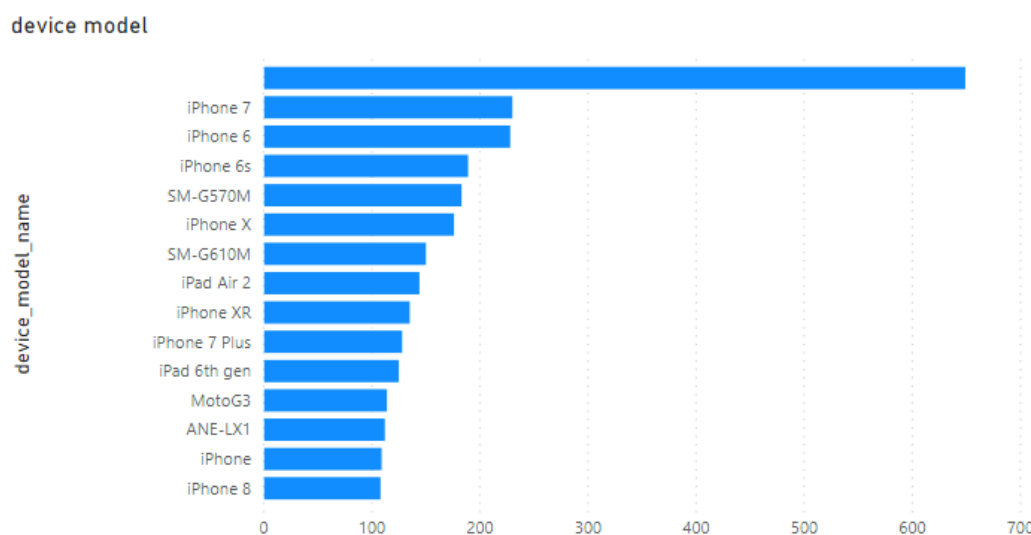


Figure 13 Device_model_name Bar Chart

device_brand_name

describe the brand of the device the users are using. (Sony, Samsung, Apple, etc). **As shown in figure 14, Most users' downloads are using Samsung reaching almost 5k total download, followed by nearly half of it on apple devices (2.9k), 2K on Huawei, and around 1.5k on Xiaomi.**

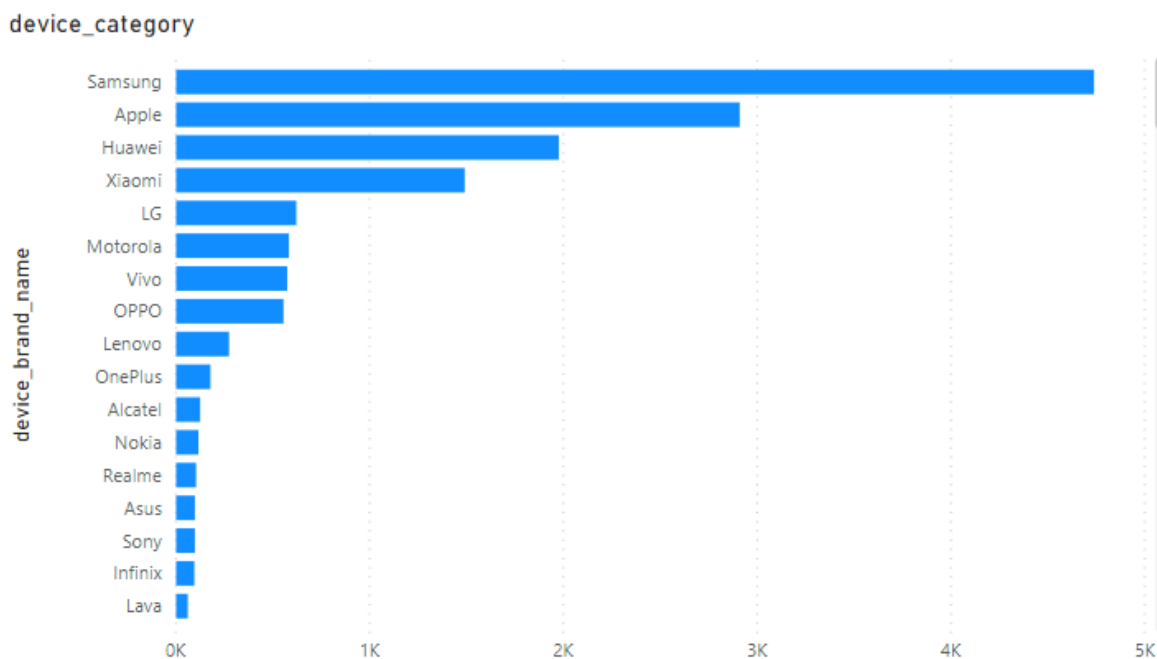


Figure 14 device_brand_name Bar Chart

device_os_hardware_model

describe the hardware model of the device. Most of the hardware model of the devices are the same as their model name.

device_os

describe the os running on the device (i.e., ios, Android). *Figure 15* shows that **the majority of users are using Android as their device operating system.**

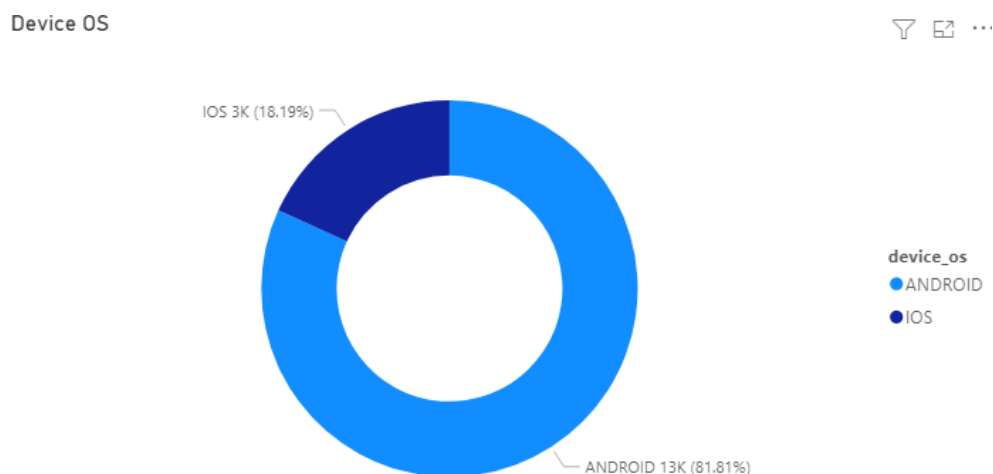


Figure 15 device_os Pie Chart

device_os_version

describe the version of the os on the device. (i.e. 5.1.1, 4.4.2). The chart below (*Figure 16*) shows the ios version on which user downloaded the app on. **The grouping/binning is done to the ios version for easier visualisation and understanding. The highest users on ios are using ios 12 in their device (not the latest ios 13 version).** The minority goes to the rest of the version, ios 11, 10, and so on.

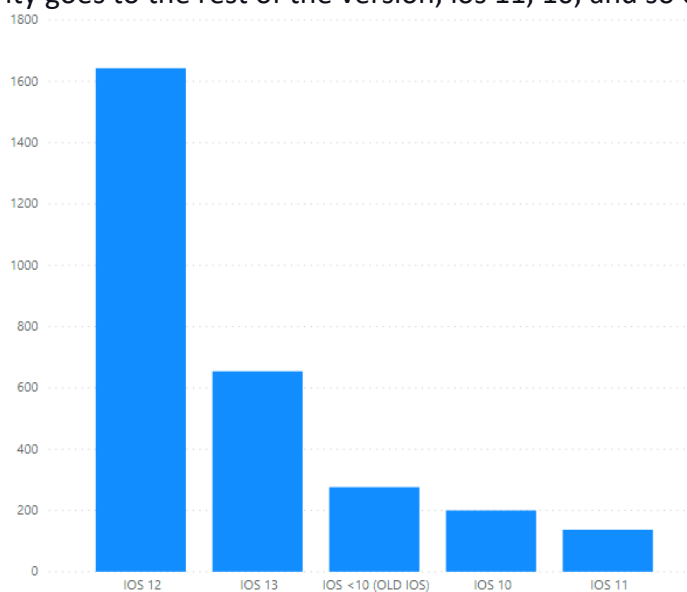


Figure 16 iOS version Bar Chart

On the other hand, As shown in *Figure 17*, The **highest amount of android user is using android os version 9 (not the latest android 10 version).** A small portion of users are still using the old version (4 and marshmallow)

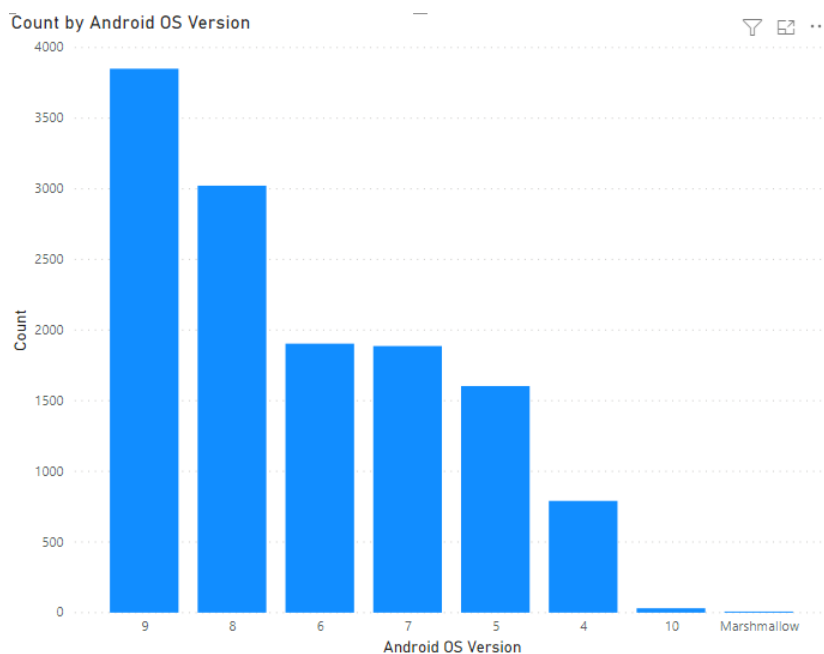


Figure 17 Android os version Bar Chart

device_language

describe the language code of the device. (i.e. pt-br,es-mx,en-au) . **Figure 18** shows that the most downloaded language version is Brazilian Portuguese (pt-br), United States English(en-us) and United Kingdom English (en-gb) . From the previous geographical data, most of the downloads are from the region speaking these languages.

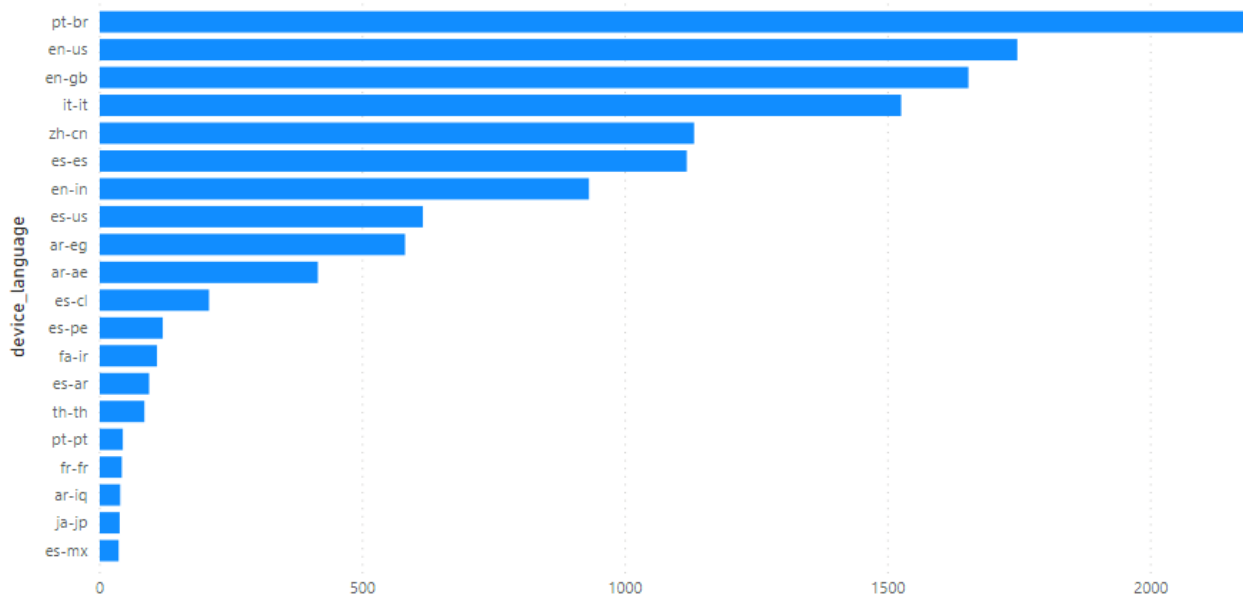


Figure 18 device_language Bar Chart

device_timezone_offset

describe the amount of timezone offset in seconds from or to the universal timezone and, in this case, UTC (i.e -18000,28800).As shown in Figure 19, **The highest device timezone offset is at 7200, followed by 19800 and -10800.** The timezone offset column can provide geographical data based on their timezone. For example, +7200 is the timezone of several countries in Africa and Europe.

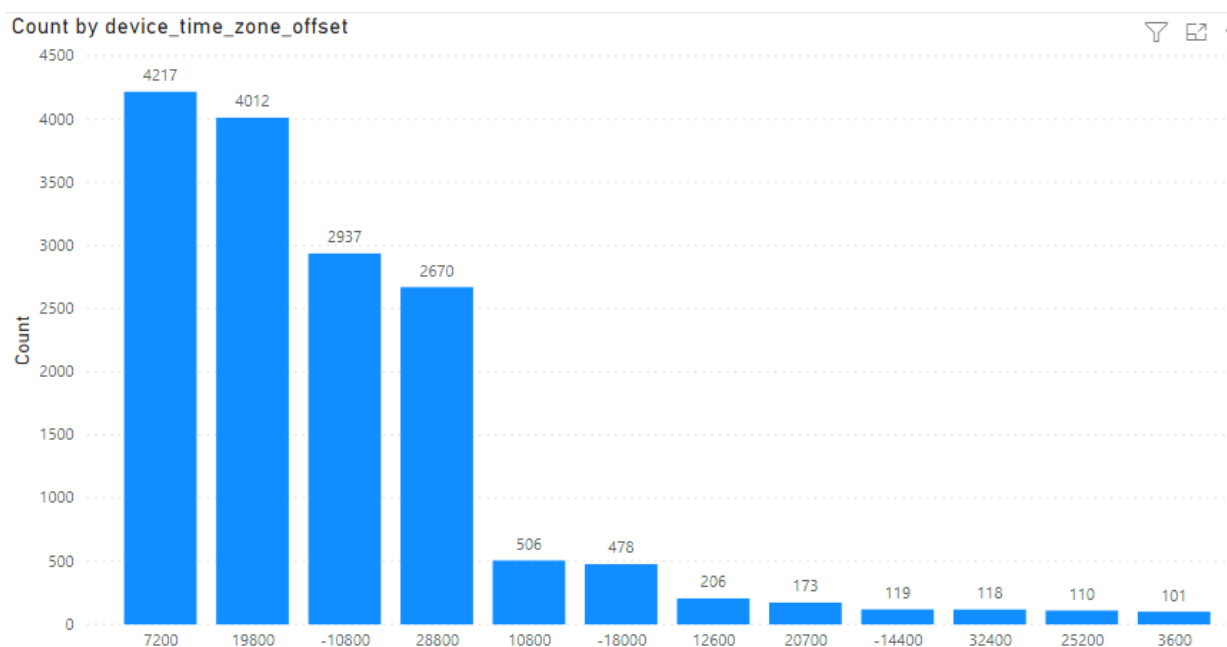


Figure 19 device_timezone_offset

is_limited_ad_tracking

Describes whether the users are turning on limited ad tracking for masking their IDFA. **Most users are having their limited ad tracking turned off.**

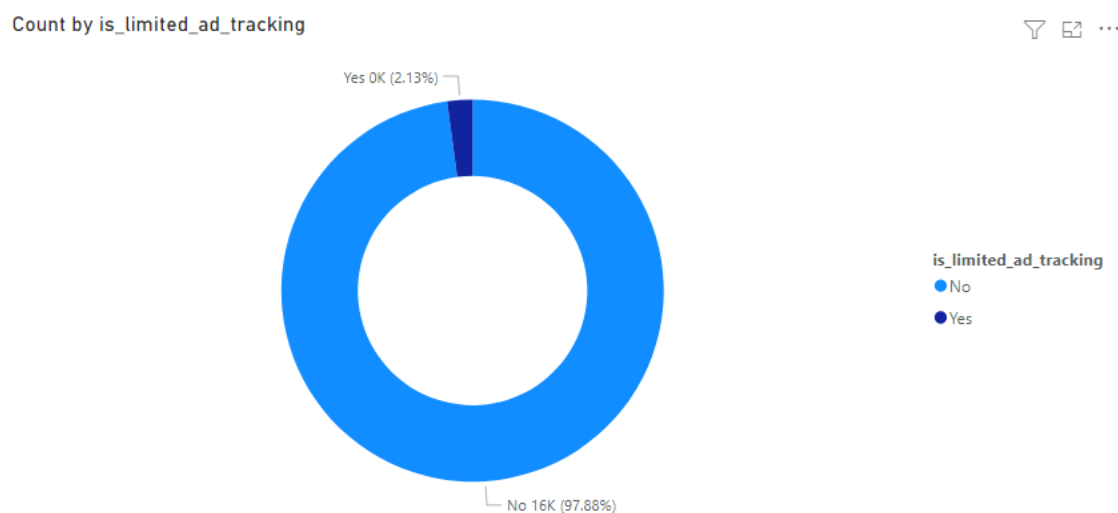


Figure 20 is_limited_ad_tracking Pie Chart

idfa

describes the identifier for customized advertising. This column consists of randomly generated id for tracking purposes and several empty value.

idfv

describes the identifier all apps on the same vendor in the same device. **This column has several empty values. No both of idfa and idfv values exist in the same row. If idfv exists, idfa will be empty, vice versa.**

table_date

The column consists of datetime datatype. **This column will be removed from further analysis since it only contains one value, 2019-10-01 00:00:00 UTC**

is_returning_user and session_id

The column does not have any value and will not be used for analysis purposes.

Methodology

Data Cleaning

Remove Unnecessary Columns

Several columns are dropped. Those columns are: (Figure 21)

table_date: having only one value => (2019-10-01) 00:00:00 UTC

is_returning_user : empty column

session_id : empty column

These attributes are dropped since they are not useful for further analysis.

| table_date | is_returning_user | session_id |
|-------------------------|-------------------|------------|
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |
| 2019-10-01 00:00:00 UTC | | |

Figure 21 Unnecessary column

Replacing Error Inputs

There are 9 rows of input error on geo_city attribute (6th of October City), the value will be replaced into empty value. (Figure 22)

```
df.groupby('geo_city').count()
```

| geo_city | user_pseudo_id | sku | app_version | geo_country |
|---------------------|----------------|-----|-------------|-------------|
| 6th of October City | 9 | 9 | 9 | 9 |
| A Coruna | 7 | 7 | 7 | 7 |
| Aba | 1 | 1 | 1 | 1 |
| Abano Terme | 1 | 1 | 1 | 1 |

Figure 22 error of inputs (geo_city)

Data Preprocessing

timestamp_raw Column Transformation

Since all of the timestamp_raw rows having the same consistent value “Date Time UTC”, UTC can be removed from the timestamp_raw value in order to change the datatype into datetime. The column will be then splitted into hour and date to ease the aggregation method , as shown in *Figure 23*

| timestamp_raw |
|-----------------------------|
| 2019-09-30 20:30:39.340 UTC |
| 2019-10-01 01:00:10.586 UTC |
| 2019-09-30 20:01:08.768 UTC |

| timestamp_raw |
|----------------------|
| 9/30/2019 8:30:39 PM |
| 10/1/2019 1:00:11 AM |
| 9/30/2019 8:01:09 PM |

| hour | date |
|------|-----------|
| 20 | 9/30/2019 |
| 1 | 10/1/2019 |
| 20 | 9/30/2019 |
| 15 | 9/30/2019 |
| 12 | 10/1/2019 |

Figure 23 timestamp transformation

device_language Column Transformation

device_language column consists of the value “Main Language – Local” by splitting this into the two different columns; it will ease the binning of language installed in users’ device (*Figure 24*). After the columns are split, the main column is dropped. The purpose of doing this is to get a clearer insight of the most popular version of language download.


| device_language |
|-----------------|
| pt-br |
| pt-br |
| es-mx |

| device_language | language_region |
|-----------------|-----------------|
| pt | br |
| pt | br |
| es | mx |

Figure 24 device_language transformation

is_limited_ad_tracking Column Transformation

is_limited_ad_tracking consists of the value “yes” or “no.” Aggregation is done to count the total of “yes” and “no” values in the column, grouped by the country. By having the value aggregated, the attribute can be pivoted for easier and more accurate visualisation. (As shown in *Figure 25*)



| A ^B _C is_limited_ad_tracking |
|--|
| No |
| Yes |
| No |
| No |
| No |
| No |
| No |
| No |
| No |
| No |
| No |
| No |

| A ^B _C geo_country | 1 ² ₃ No | 1 ² ₃ Yes |
|---|--------------------------------|---------------------------------|
| Angola | 17 | 0 |
| Aruba | 2 | 0 |
| Belize | 14 | 0 |
| Benin | 5 | 0 |
| Bhutan | 2 | 0 |

Figure 25 is_limited_ad_tracking aggregation and pivoting

Missing Value Handling

```
import matplotlib.pyplot as plt
%matplotlib inline
df.isnull().sum()

user_pseudo_id      0
sku                  0
app_version          0
geo_country          40
geo_region          1708
geo_city            3233
install_source       0
ua_name             9409
ua_medium            2
ua_source            0
device_category      0
device_brand_name    649
device_model_name    649
device_os_hardware_model 0
device_os            0
device_os_version    0
idfa                1365
idfv                15552
is_limited_ad_tracking 0
device_language      0
device_time_zone_offset 0
timestamp_raw        0
dtype: int64
```

Figure 26 Missing Value Counts - Jupyter Notebook

As shown in Figure 26, Several columns on the dataset have missing values. Since the column is not a numerical value, ignoring or putting modes is the only way we can do to solve most of the missing values.

geo_country

40 data rows of geo_country have empty value; however, several rows have their geo region; we can get the country name from this information, for example. geo_region of Crimea has Russia as the country; thus geo_country (which are empty) are replaced into Russia. (As shown in Figure 27)

| A ^B _C user_pseudo_id | A ^B _C sku | A ^B _C app_version | A ^B _C geo_country | A ^B _C geo_region |
|--|---------------------------------|---|---|--|
| 8c3790fe777ba8136986b4ae1771a7... | Google Play | 1.20.1 | | Crimea |
| e2003265f39391d29f221cd8951ad0... | Google Play | 1.20.1 | | Crimea |
| 1124d12856324d9b925c5b3f6e46e... | Google Play | 1.20.1 | | Crimea |
| f62cd88233a2f8348a2f0bbb4de3f2fb | Google Play | 1.20.1 | | Crimea |

```
df.loc[df['geo_region']=="Crimea", "geo_country"]="Russia"
df[df['geo_country']=="Russia"]
```

| | user_pseudo_id | sku | app_version | geo_country | geo_region |
|----|----------------------------------|-------------|-------------|-------------|------------|
| 7 | 8c3790fe777ba8136986b4ae1771a763 | Google Play | 1.20.1 | Russia | Crimea |
| 9 | e2003265f39391d29f221cd8951ad037 | Google Play | 1.20.1 | Russia | Crimea |
| 19 | 1124d12856324d9b925c5b3f6e46ebb4 | Google Play | 1.20.1 | Russia | Crimea |
| 25 | f62cd88233a2f8348a2f0bbb4de3f2fb | Google Play | 1.20.1 | Russia | Crimea |
| 28 | 44D699BE272D430288C95F8C7E221E25 | iOS | 1.20.1 | Russia | Crimea |

replaced value

Figure 27 geo_country , missing value handling

For the rest of the missing country, there are two options:

- Delete row: since the missing rows are only 40 rows out of 16000 rows, Ignoring the data will not have a significant impact, and if the data are used for predictive analysis, it will provide better accuracy.
- Ignore/put modes: since the column is not a numerical value, ignoring or putting modes is the only way we can do to solve most of the missing values.

idfa and idfv

Regarding the missing value in idfa or idfv, the columns are left as they are since for every missing idfa, there is always an existing idfv in the rows and vice versa.

Additional Insight

The following are additional insights from the preprocessed data:

Limited Ad Tracking in Several Countries

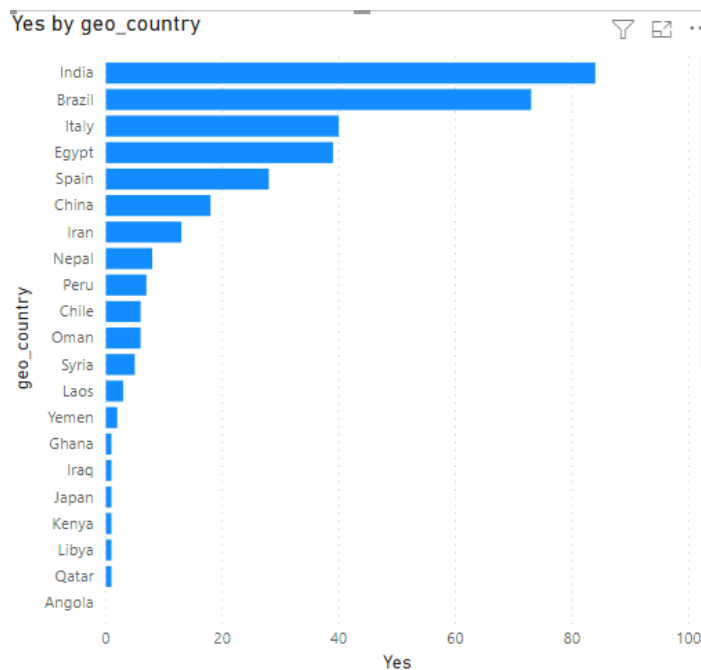


Figure 28 User enabling limited_ad_tracking amount

Although **India has the most amount of their users turning on limited ad tracking** (refer to Figure 28), by percentage, the amount is merely 2.5 percent of all of the downloads in India, which is relatively low compared to other countries. (Figure 29)

Percent of user with limited-ad-tracking enabled

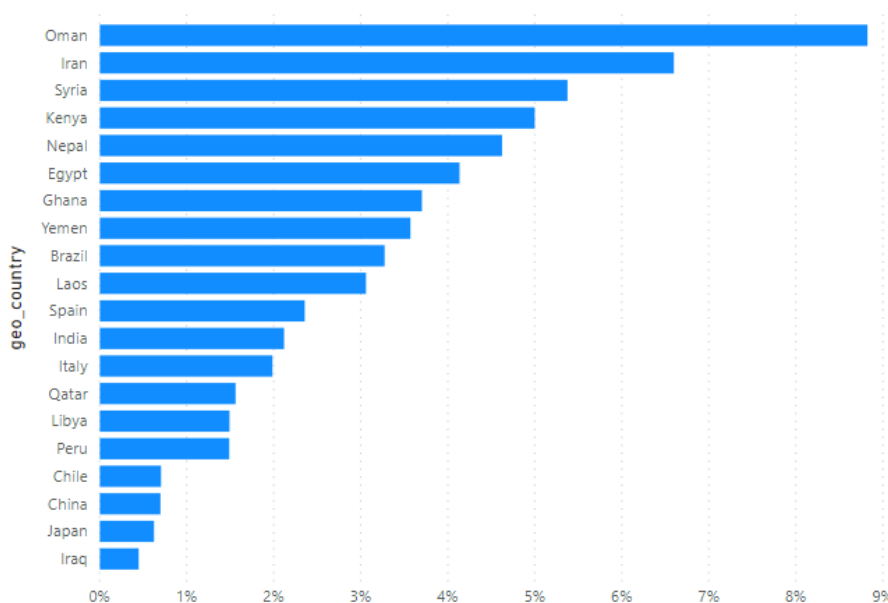


Figure 29 Percentage of Downloader with limited ad tracking

On the other hand, as shown in *Figure 29*, **Oman has the highest percentage of downloaders turning on their limited ad tracking, reaching almost 9 percent of all of their downloads (6 out of 68 downloaders)**. 6.6% of Iran downloaders have it enabled, followed by Syria at 5.38%. (More details are shown in Table 1)

| geo_country | No | Yes | Custom plus Yes % difference from No |
|-------------|------|-----|--------------------------------------|
| Oman | 68 | 6 | 8.82% |
| Iran | 197 | 13 | 6.60% |
| Syria | 93 | 5 | 5.38% |
| Kenya | 20 | 1 | 5.00% |
| Nepal | 173 | 8 | 4.62% |
| Egypt | 943 | 39 | 4.14% |
| Ghana | 27 | 1 | 3.70% |
| Yemen | 56 | 2 | 3.57% |
| Brazil | 2229 | 73 | 3.28% |
| Laos | 98 | 3 | 3.06% |
| Spain | 1188 | 28 | 2.36% |
| India | 3963 | 84 | 2.12% |
| Italy | 2014 | 40 | 1.99% |
| Qatar | 64 | 1 | 1.56% |
| Libya | 67 | 1 | 1.49% |
| Peru | 470 | 7 | 1.49% |
| Chile | 850 | 6 | 0.71% |
| China | 2575 | 18 | 0.70% |
| Japan | 160 | 1 | 0.62% |
| Iraq | 223 | 1 | 0.45% |
| Angola | 17 | 0 | 0.00% |

Figure 30 Percentage of downloader with limited ad tracking (Table)

Download Popularity in Certain Hour

Figure 31 shows the total download by hours. **Based on the chart, the highest amount of download is at around 2–3 pm in the afternoon.** The download amount plummeted after 3 pm, reaching the lowest at around 420 totals of download. Based on the line chart, the least amount of downloads happen from midnight to the morning at around 9 am.

Downloader Count by hour



Figure 31 Download Popularity by Hour

Additionally, If more days are put in analysis, a significant amount of information can be gathered, such as the most effective advertisement/any marketing strategy which persuades the installation of the application, customer up time , pattern of download.

Most Downloaded Language Version of the App (Based on os)

After splitting the columns, the popularity of the language can be determined. **As shown in Figure 32, most google play downloads are using English as their language, followed by Portuguese and Spanish. In contrast, downloaders on iOS are mostly using Chinese rather than English (second place) and Italian.** This chart's language popularity is equivalent to the geographic language of the dataset; refer to Figure 6.

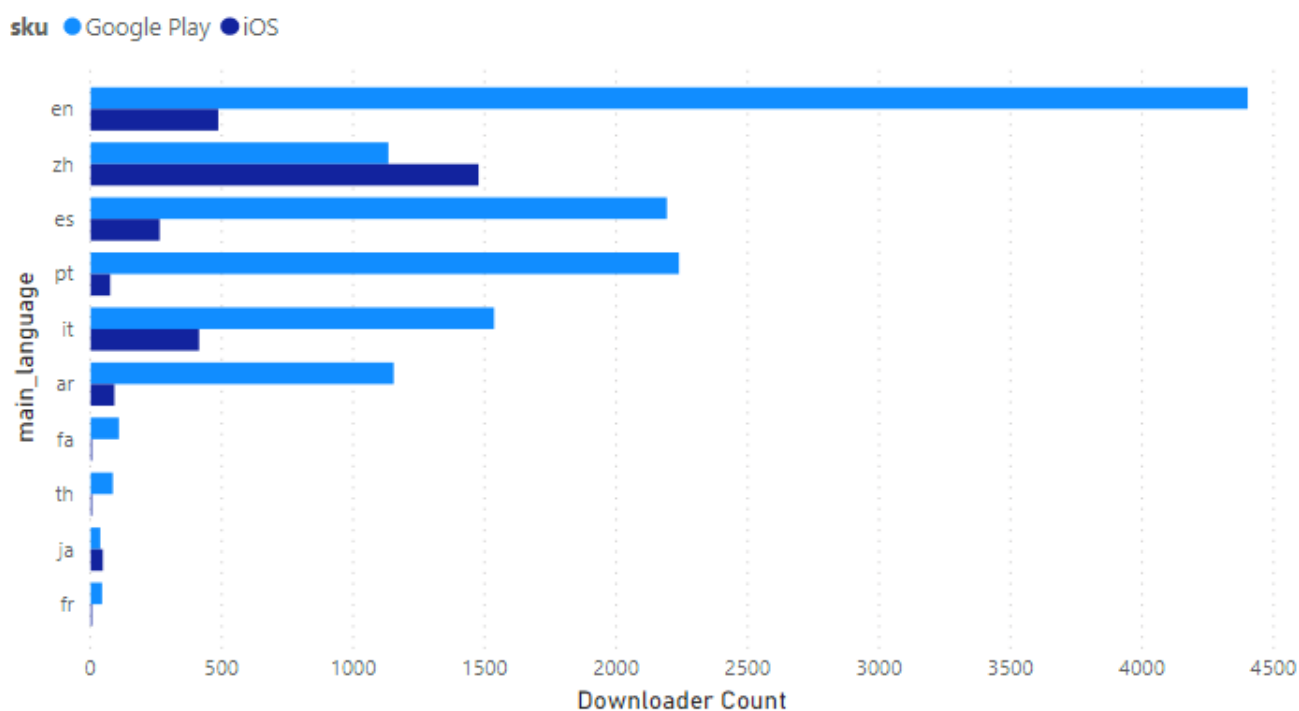


Figure 32 main_language and sku Bar Chart

Conclusion

Overall, many insights can be produced from this versatile dataset, depends on the business needs and the goals of the analysis. The author strongly believed that having a larger dataset (more column, more timestamp, more numerical/statistical related column) is tremendously beneficial for future analysis. Having a target column with more relevant information/attributes will also introduce a predictive analysis that can detect users' download patterns by algorithm, resulting in more insights gathered from the dataset.