**Outlogic Data**

**Data transfer for Outlogic:**

1. We uploaded the files from December 2020 to July 2021 onto our internal S3 bucket from the S3 bucket shared by Outlogic (via the World Bank DDP team).
   1. While we had initially decided to look at June 2020-July 2021, we had to reduce our time period due to missing months (July, August, September, November 2020).
2. We verified the file count since each day’s data was split into 1000 parquet files labelled part000 to 999)
   1. This was done via the S3 console and corrected for dates with less than 1000 files.
   2. We ended up with 188GB for 8 months of data.
3. Next, we performed data integrity checks (in case of corrupted data) for each parquet file through matching them with S3 checksums.

**Table Creation**

1. We created a table of Outlogic data with all columns as string since we did not require data types immediately
2. We ran MSCK for Athena to recognise and load the data and partioned the table by date as provided by source as showcased in ‘table\_creation.sql’
3. Issues run into:
   1. The data type proved problematic and was resolved via trial and error. The Parquet files and DDL from Data Partnerships note field data type were not an exact match leading to errors. A sample such error:
   * **Error**

Query ID e9017f56-67f9-4f74-a6de-d5721b16b01e

* + Error details: HIVE\_BAD\_DATA: Field background's type DOUBLE in parquet file s3://outlogic-data-bucket/Outlogic/location\_dt=2021-05-04/part.167.parquet is incompatible with type integer defined in table schema. You may need to manually clean the data at location 's3://outlogic-data-bucket/results\_athena/CREATE%20-%20v2/2022/01/16/tables/e9017f56-67f9-4f74-a6de-d5721b16b01e' before retrying. Athena will not delete data in your account. This query ran against the "outlogic\_db" database, unless qualified by the query. Please post the error message on our [forum](https://forums.aws.amazon.com/forum.jspa?forumID=242&start=0)  or contact [customer support](https://console.aws.amazon.com/support/home?#/case/create?issueType=technical&serviceCode=amazon-athena&categoryCode=query-related-issue)  with query id.
  1. Ultimately, we skipped the background filed from the dataset since the issue could not be resolved. It was showing up as DOUBLE or INT32 in different fields and Athena could not merge/coalesce them. Hence, our dataset did not have them in our analysis.

1. Once our India wide table was created, we moved Maharashtra-specific data into a separate table (as showcased in mah\_table.sql) Please note:
   1. We verified with 3 random dates to check proper separation.
   2. Only ‘IN-MH’ records were considered, hence even if the latitude and longitude was within Maharashtra (according to OpenStreetMap), we ignored it if the location field from Outlogic was not ‘IN-MH’.
   3. We created the table using data for one month and used INSERT for the rest. This was because Athena was running into a partition limit if we selected all seven months into a single query.

**Transformation and Additions to the Data**

After we had created the Maharashtra only table of Outlogic data, we next had to reverse geocode the latitude and longitude provided to the district labels required for our analysis. To do so, we did the following:

1. We got shapes from Overpass using an Overpass query as outlined in mah\_geocode.txt
2. Using the exported geojson from this query, we converted it to a file containing only the name and coordinates using the script outlined in ‘mah\_geojson\_to\_parquet\_for\_athena.py’. The script also contains a link to the overpass query export and the output parquet file with the name and coordinates is also available.
3. A ‘boundaries’ table was created with district names and shapes/coordinates columns to act as a reference.
4. We joined the Outlogic coordinates from the Maharashtra table with the boundaries table to get the district names for every Outlogic entry as showcased in ‘geocoded\_mh\_outlogic.sql’
5. We checked if Outlogic data marked as IN-MH matches our reference boundaries table. We find that:

|  |  |  |
| --- | --- | --- |
| Number of records | ISO\_REGION | % of total records |
| 26 | IN-GA | 0.00% |
| 2 | IN-CT | 0.00% |
| 340 | IN-AP | 0.00% |
| 12841596 | IN-MH | 98.46% |
| 194434 |  | 1.49% |
| 153 | IN-DN | 0.00% |
| 121 | IN-MP | 0.00% |
| 1368 | IN-KA | 0.01% |
| 2740 | null | 0.02% |
| 1473 | IN-GJ | 0.01% |

* 1. Without checking for IN-MH, when we converted the raw Outlogic latitude and longitude data to district, we saw that 98.5% of what Outlogic marks as MH matches our OSM reference. Our OSM map additionally includes a few spots from surrounding states as Karnataka (IN-KA), hence the presence of other states in this list.
  2. We limited ourselves to records marked as IN-MH, since those are in the raw files from Outlogic

1. We also computed new columns from the existing data such Date from timestamp for which the parquet date partition was used directly.

**Analysis and Summaries**

Since we were interested in inter-district changes, we filtered by locations and not devices, aggregating the pings and unique devices from each Maharashtra district for each day to obtain our final dataset. We performed two kinds of filters (one by pings and one by unique devices).

1. Date – district - pings: Each ad-ID can have multiple pings per day (from the same or different publisher ID i.e., apps). We extract daily counts and then compute monthly/weekly summaries as shown in ‘mah\_geocoded\_district\_pings.sql’.
2. Date - district - Ad IDs**:** Used as a broader indication of activity - since the number of Ad-IDs that show pings per day can be an indicator of the number of devices/individuals engaging in activity. This is shown in ‘mah\_geocoded\_district\_devices.sql’

After filtering the dataset by location (i.e., district), we tracked the percentage increase and decrease month-on-month in the number of unique devices for each district within the study. Using this process, we get consistent pings over time and avoid the hassle of obtaining origin-destination matrices for specific devices.

Steps of specific analysis of Outlogic Data compared to baseline, monthly changes and peak analysis is outlined in ‘mah\_final\_analysis.xlsx’ or in the paper itself in the methodology section.

**Spaceknow Indices**

No cleaning was required for SpaceKnow data, only selection of the correct indices.

To map out the impact of the second wave on general economic activity, we looked at ‘Aggregated Industrial’, ‘Aggregated Manufacturing’, and ‘Aggregated Construction’ SpaceKnow indices. For Maharashtra-specific industries, we chose the following indexes: ‘Car Factory’ and ‘Car Distribution’ (Automobiles), ‘Electronics’ (Electronics), and ‘Food’ (Food Processing). While industries such as textiles are critical to Maharashtra, due to the lack of a SpaceKnow index for the sector, it was not possible to include it in the analysis. Further, since the 30-day daily aggregated indices had the highest coverage in terms of satellite imagery and confidence, we used them to track changes in day-to-day economic activity.

Details of specific steps taken are outlined in ‘mah\_final\_analysis.xlsx’ or in the paper itself in the methodology section.