UNSTRUCTURED DATA ANALYSIS

ANALYSIS BY -

Raloue R Kapoor
Akshiti Parashar
Ansh Tiwari
Srihari Ananthan

INDEX

1. Introduction
    a. Understanding the Data
2. Data Cleaning
    a. How did we prepare it for training ?
    b. Data count vs Country
    c. Correlation heatmap
3. Anomaly Detection
    a. Applying LOF and HBOS
    b. Understanding the results
4. Performing EDA
    a. EDA Plots
    b. Deriving conclusions

Intro: What’s Unstructured data?

Unsupervised or undirected data science uncovers hidden patterns in unlabeled data. In unsupervised data science, there are no output variables to predict.

How to work about it?

We use Unsupervised Methods. These methods also estimate that any malicious data would be different statistically from normal data.

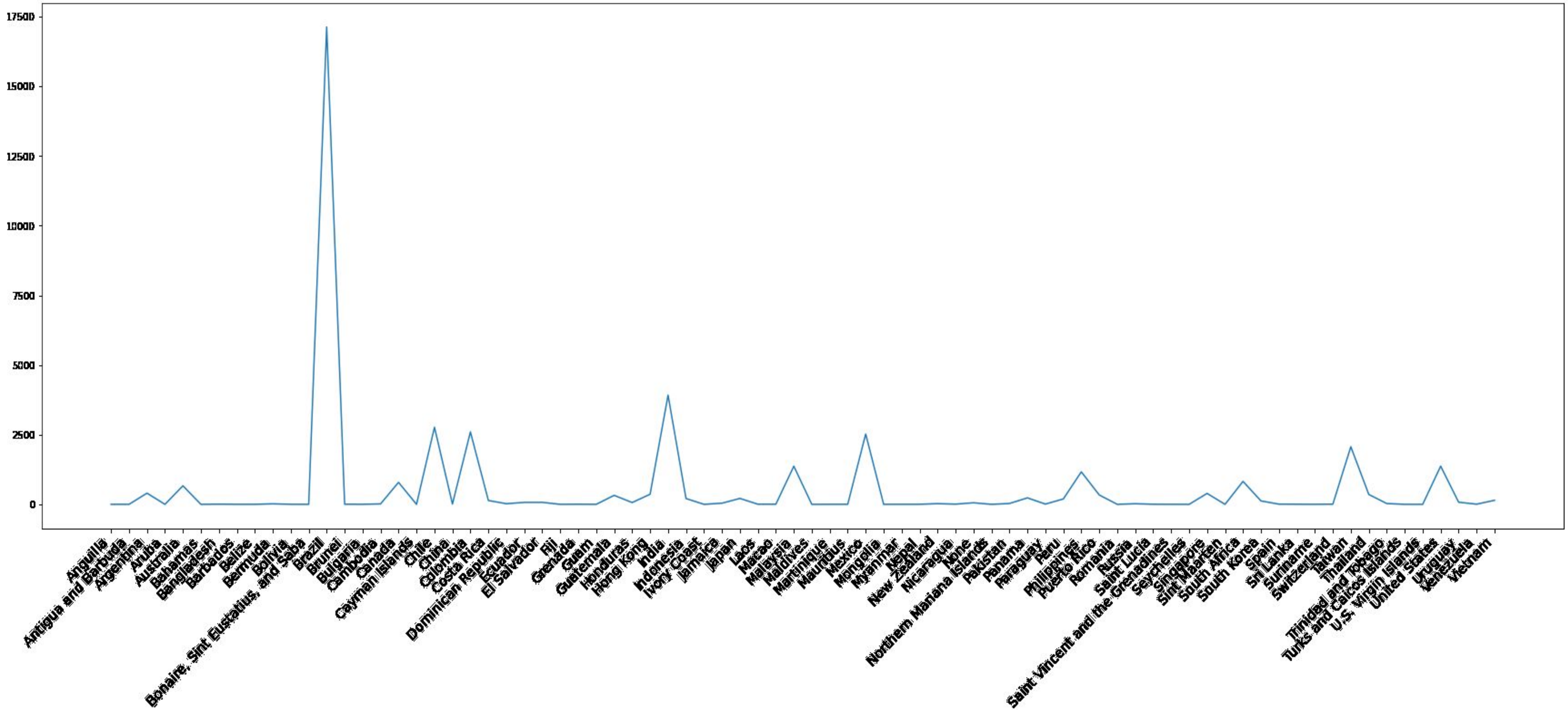
Here, we have performed both HBOS and LOF for a comparative analysis.

DATA CLEANING

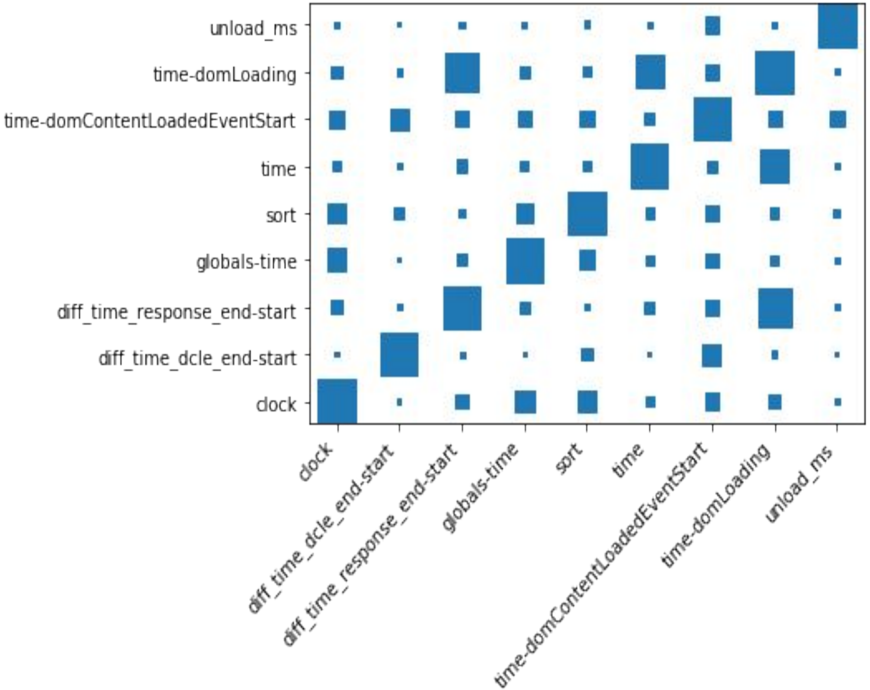
How did we prepare it for training?

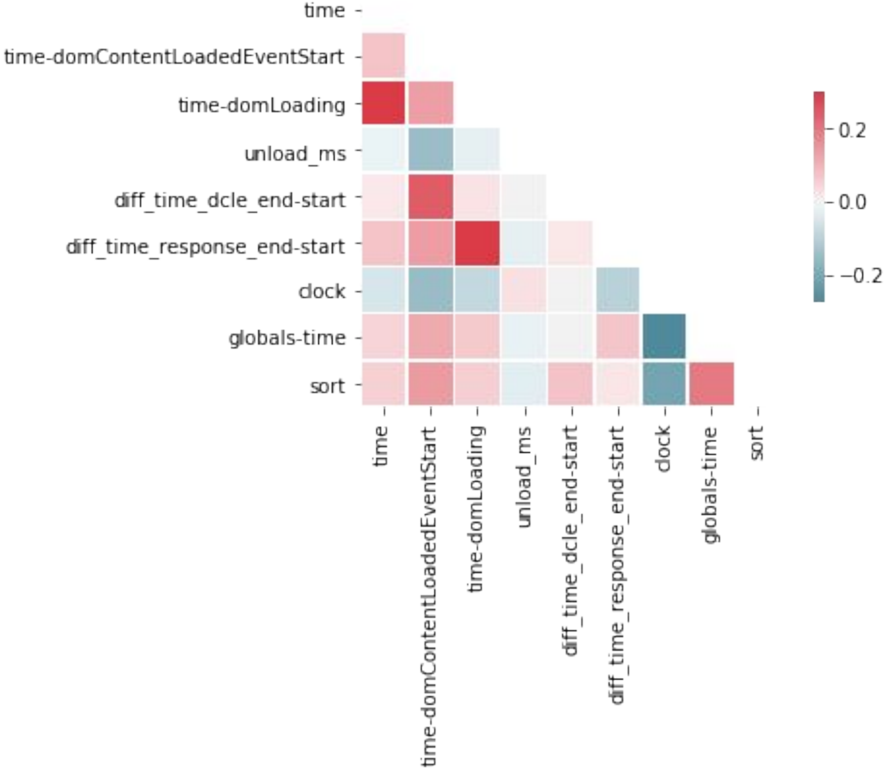
1. Filtered columns like GEO_AS_(#), GEO_CITY_(#), GEO_LAT_(#), the time columns, etc. by combining/combining them for columns with useful values. Eg. Eliminated the city, country, region, etc. based on if it matches with IP address, latitude and longitude.
2. Divided motion and orientation columns into 3 parts (x, y, z).
3. Added isMotionTrue and difference_time columns.
4. Extracted columns for Android version and Phone model from appVersion column.
5. Eliminated the redundant columns.
6. Categorized the remaining columns into Features and TVs.
7. Eliminated rows with >10 NaN values.

Where is our data from ?



Heatmap - Correlation between features





Anomaly Detection

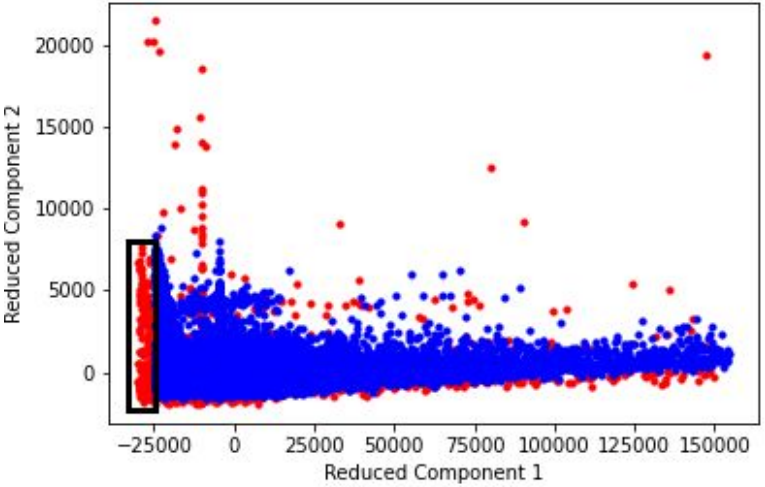
We had applied both LOF and HBOS on this dataset.

Let's have a comparison of both algorithms and understand

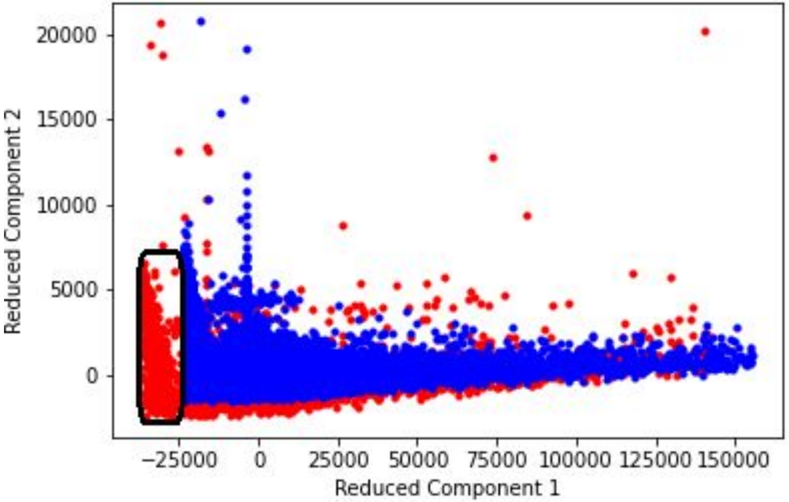
	LOF	**HBOS**
n_neighbors / n_bins	20	20
Number of outliers	2846	4105
Accuracy	Better than HBOS	-
Detection	Local	Global
Speed	-	Better than LOF

Understand it better with PCA PLOT

Local Outlier Factor



Histogram Based Outlier Scores



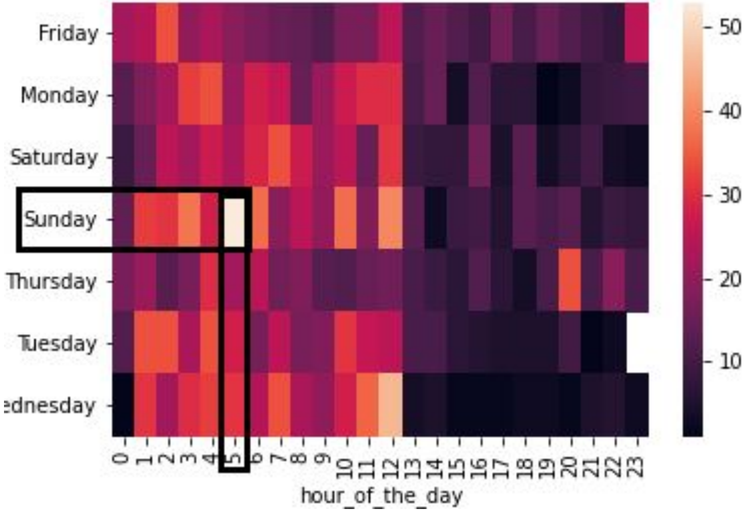
Performing EDA

We started to plot data by taking different columns together and tried to come up with a good insight from it. These are the few plots we drew -

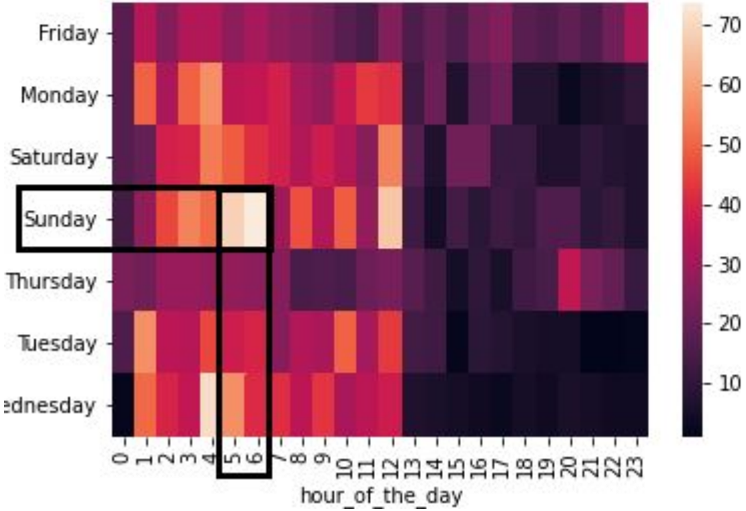
1. Heatmap of Day of the Week vs Hour of the day
2. Provider vs Memory
3. Android Version vs Providers
4. Android version vs Labels
5. Service Provider vs Labels
6. Plotting outliers on map
7. Analyzing features ,TV with outliers to obtain visual and workable results

1. Heatmap - Day of the week vs Hour of the day

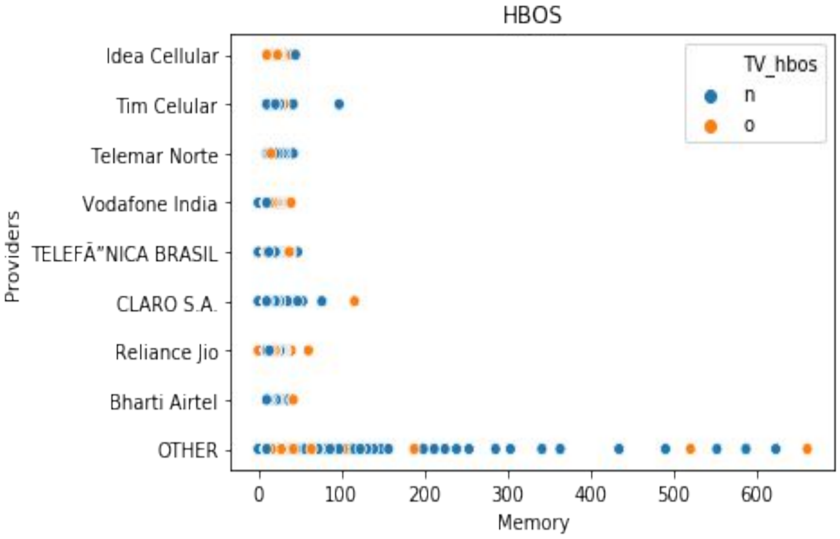
Local Outlier Factor

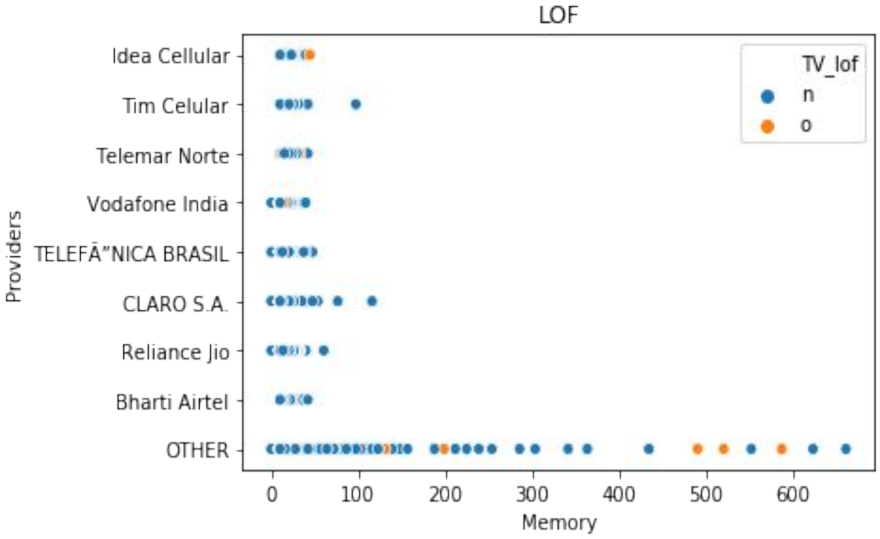


Histogram Based Outlier Scores

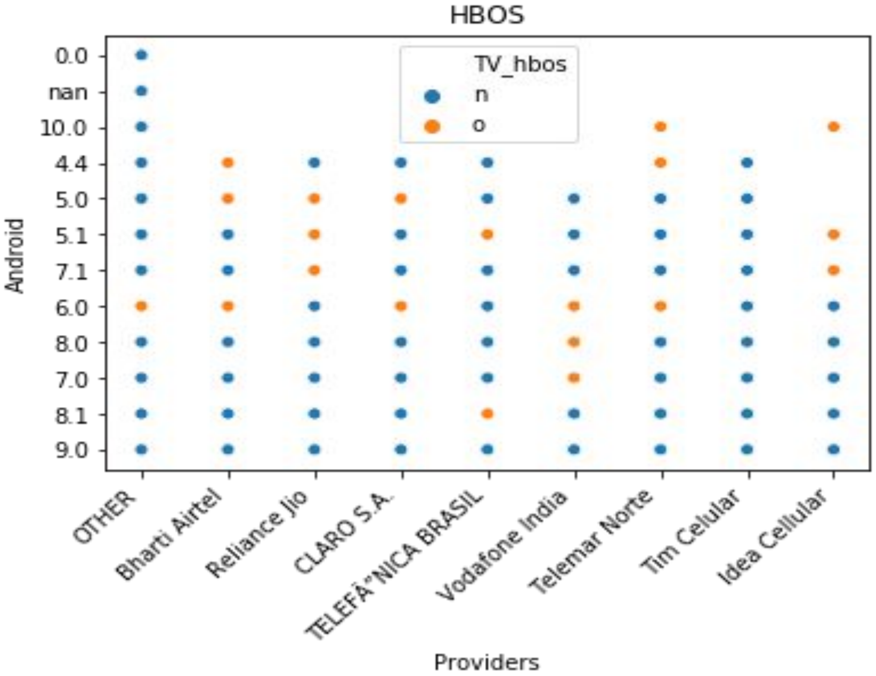


2. Service Provider vs Memory Comparison

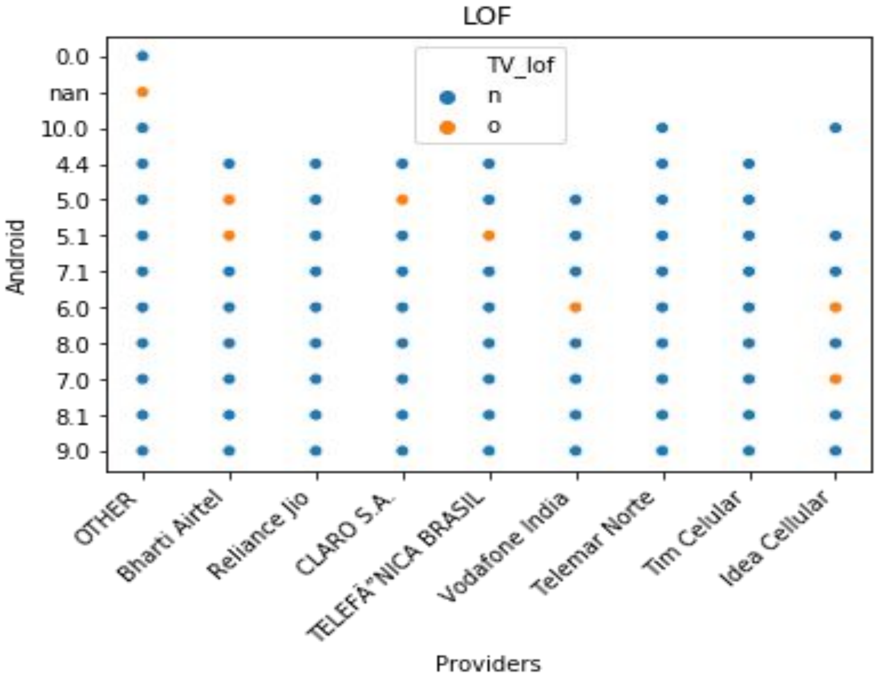




3. Android Version vs Service Providers

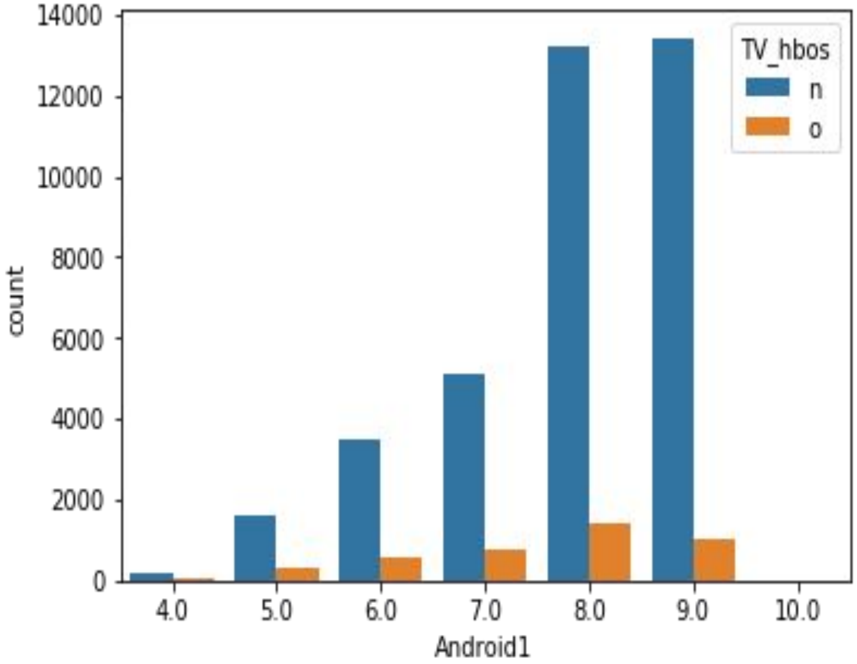


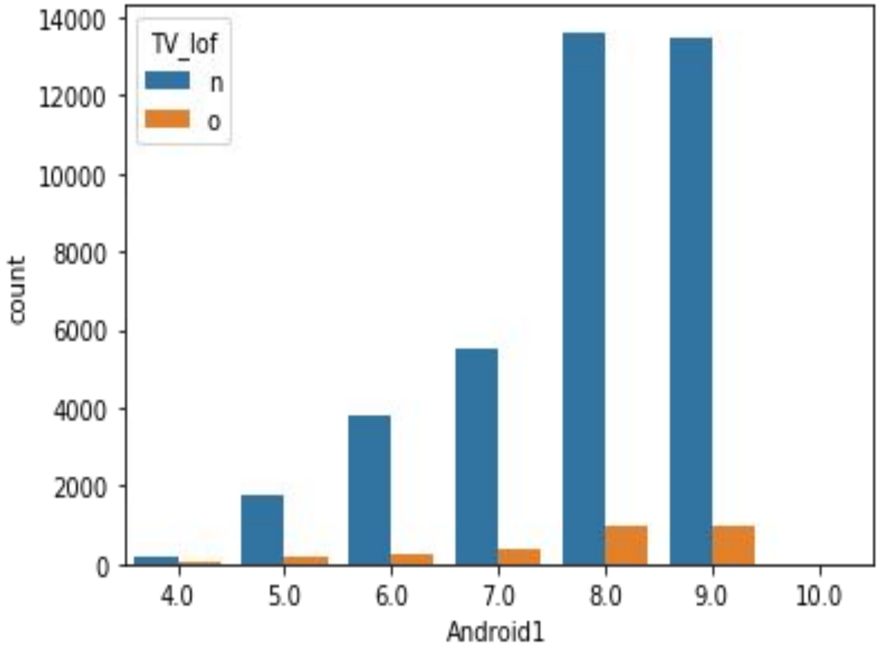
HBOS



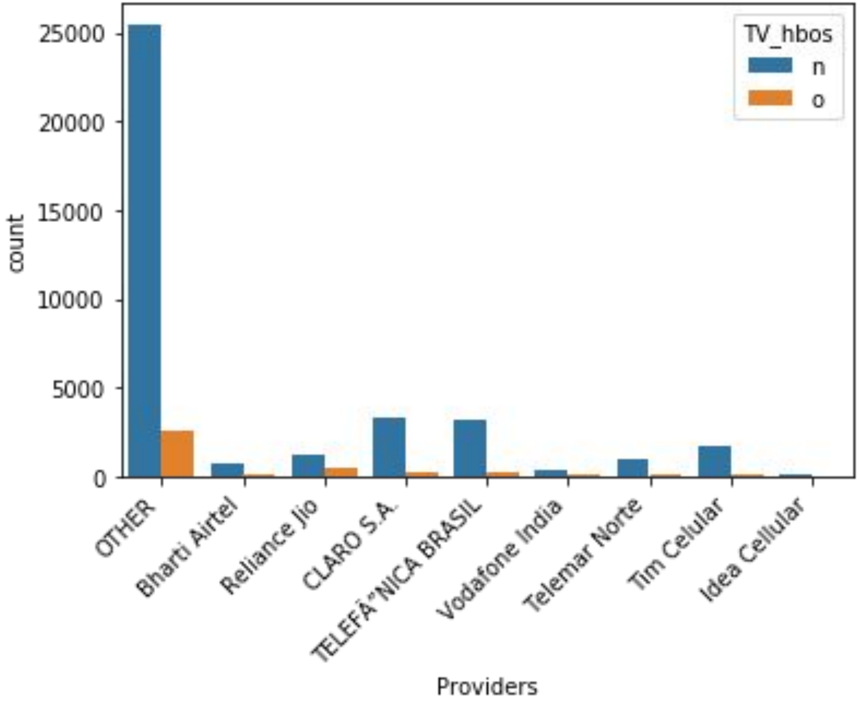
LOF

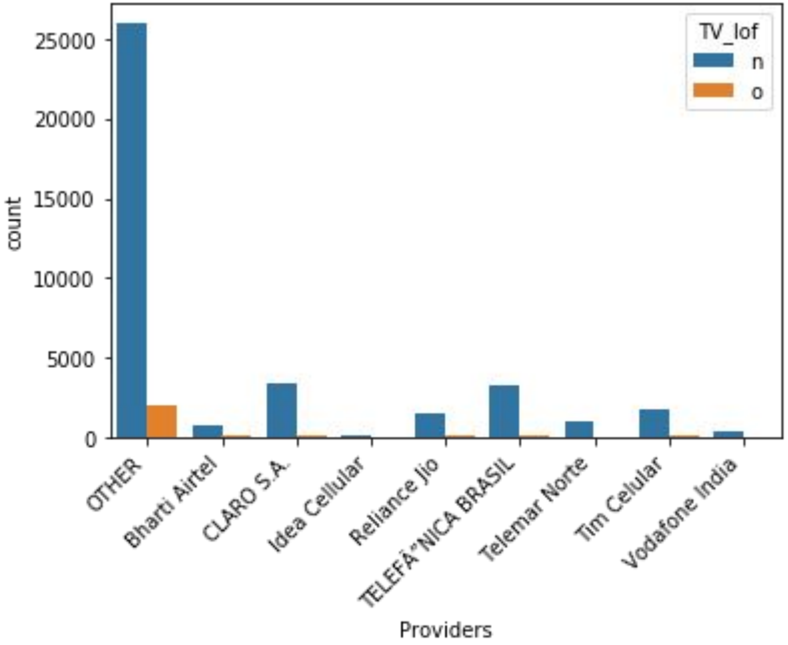
4. Android Version vs Labels



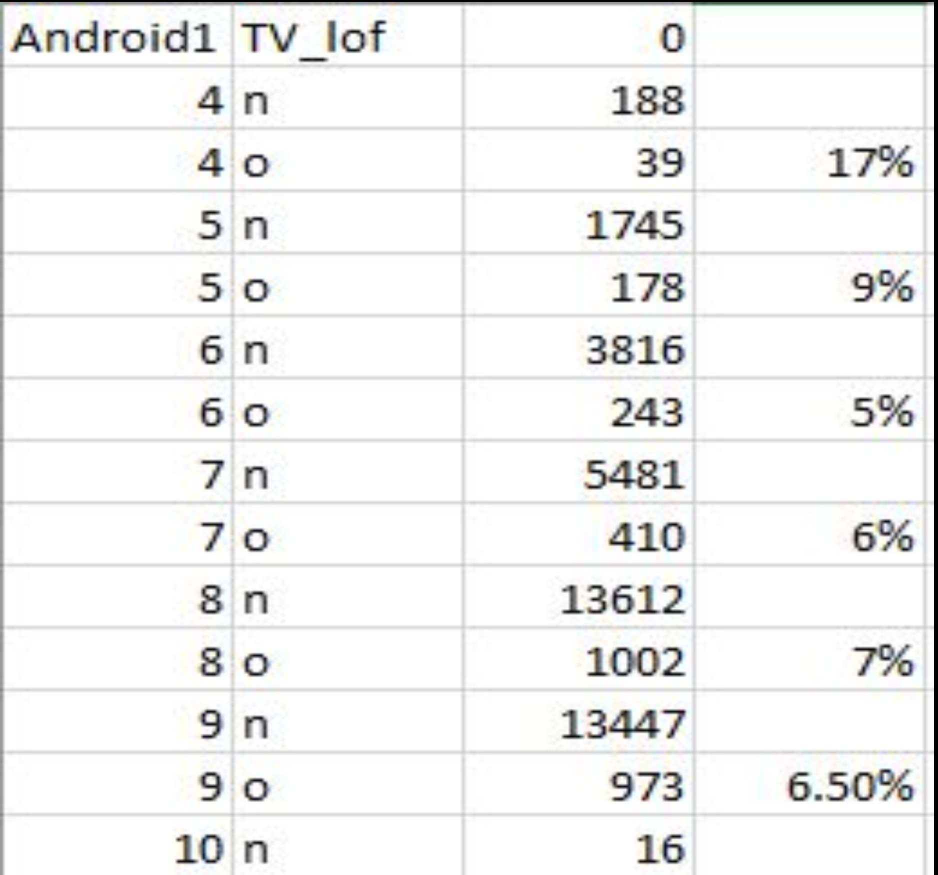


5. Service Provider vs Labels

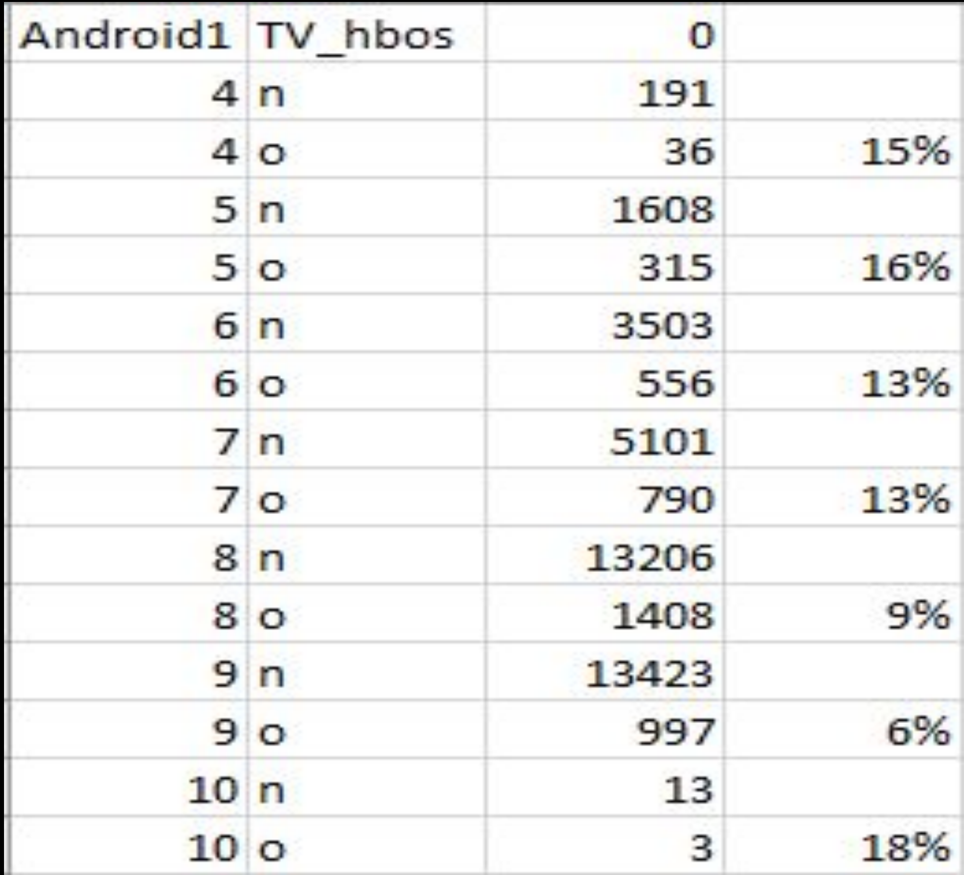




Percentage outlier for ANDROID VERSION

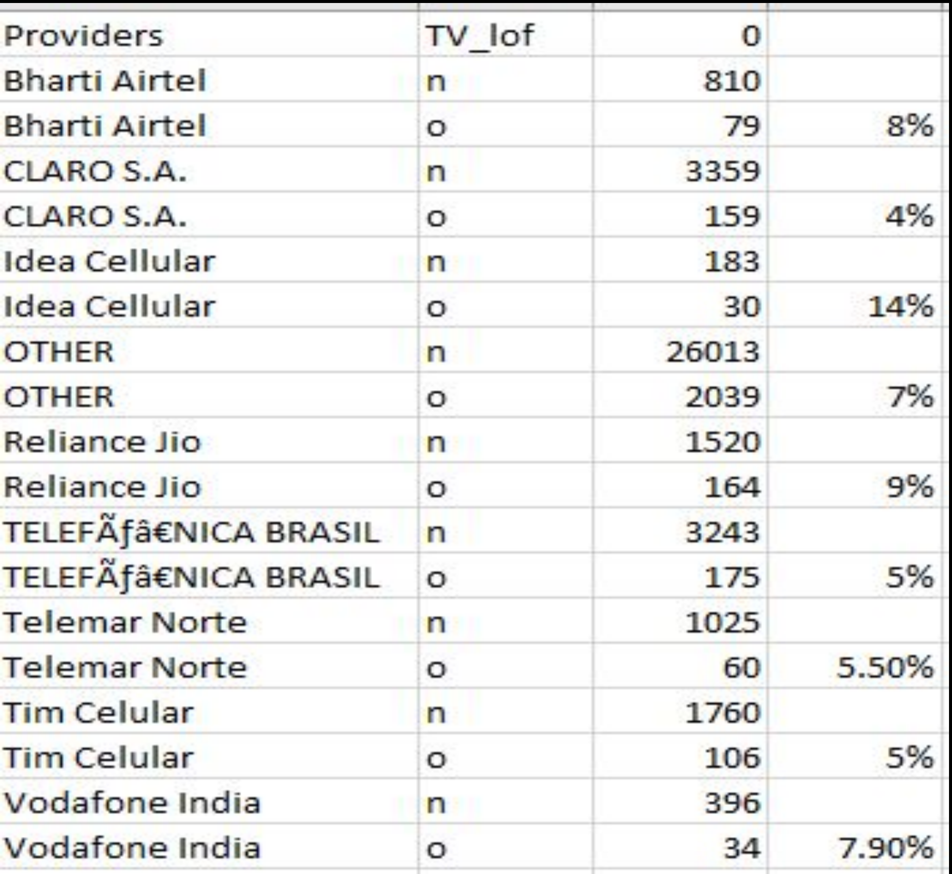


Android1	TV_lof	0	
4	n	188	
4	o	39	17%
5	n	1745	
5	o	178	9%
6	n	3816	
6	o	243	5%
7	n	5481	
7	o	410	6%
8	n	13612	
8	o	1002	7%
9	n	13447	
9	o	973	6.50%
10	n	16	

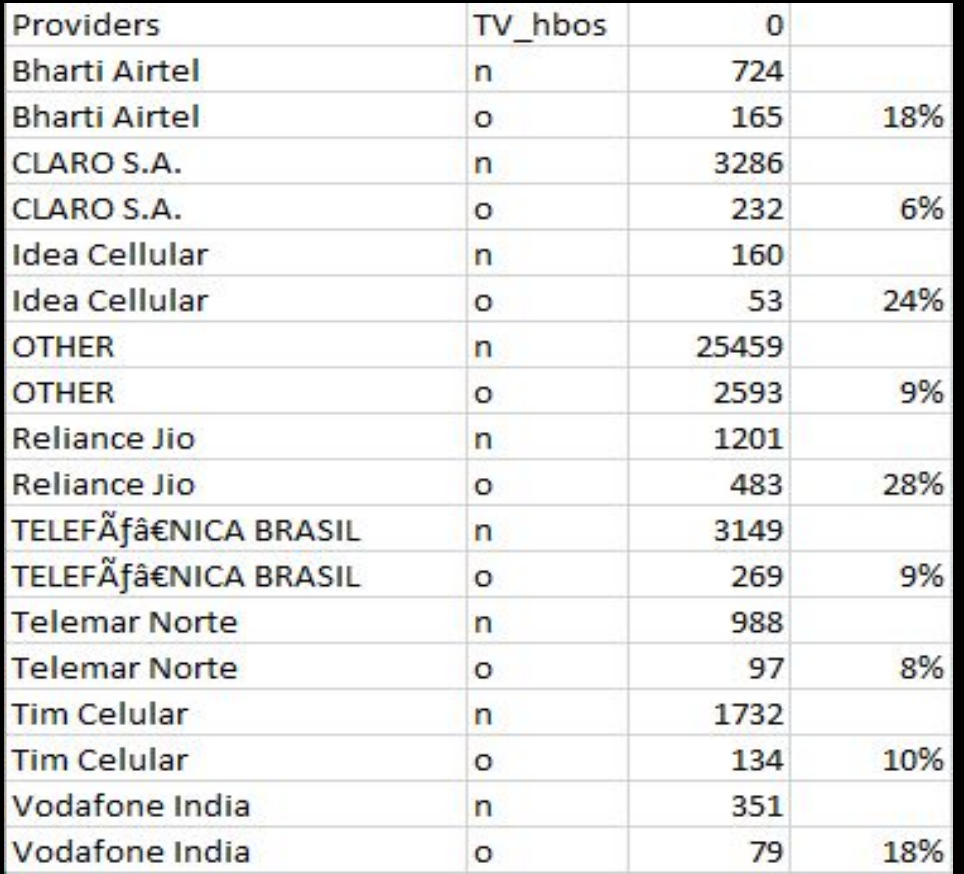


Android1	TV_hbos	0	
4	n	191	
4	o	36	15%
5	n	1608	
5	o	315	16%
6	n	3503	
6	o	556	13%
7	n	5101	
7	o	790	13%
8	n	13206	
8	o	1408	9%
9	n	13423	
9	o	997	6%
10	n	13	
10	o	3	18%

Percentage outlier for service providers

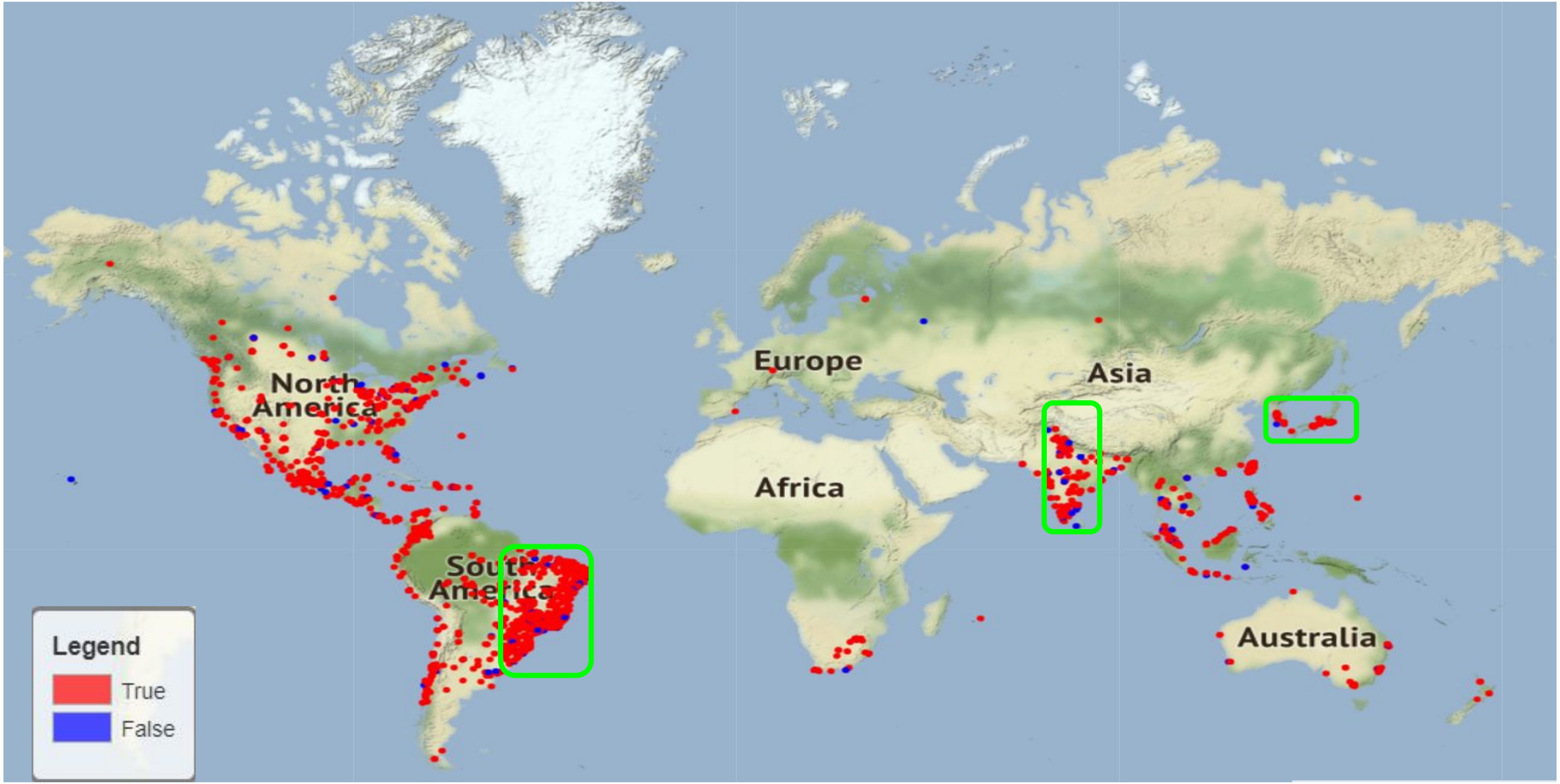


Providers	TV_lof	0	
Bharti Airtel	n	810	
Bharti Airtel	o	79	8%
CLARO S.A.	n	3359	
CLARO S.A.	o	159	4%
Idea Cellular	n	183	
Idea Cellular	o	30	14%
OTHER	n	26013	
OTHER	o	2039	7%
Reliance Jio	n	1520	
Reliance Jio	o	164	9%
TELEFÃƒâ€NICA BRASIL	n	3243	
TELEFÃƒâ€NICA BRASIL	o	175	5%
Telemar Norte	n	1025	
Telemar Norte	o	60	5.50%
Tim Celular	n	1760	
Tim Celular	o	106	5%
Vodafone India	n	396	
Vodafone India	o	34	7.90%

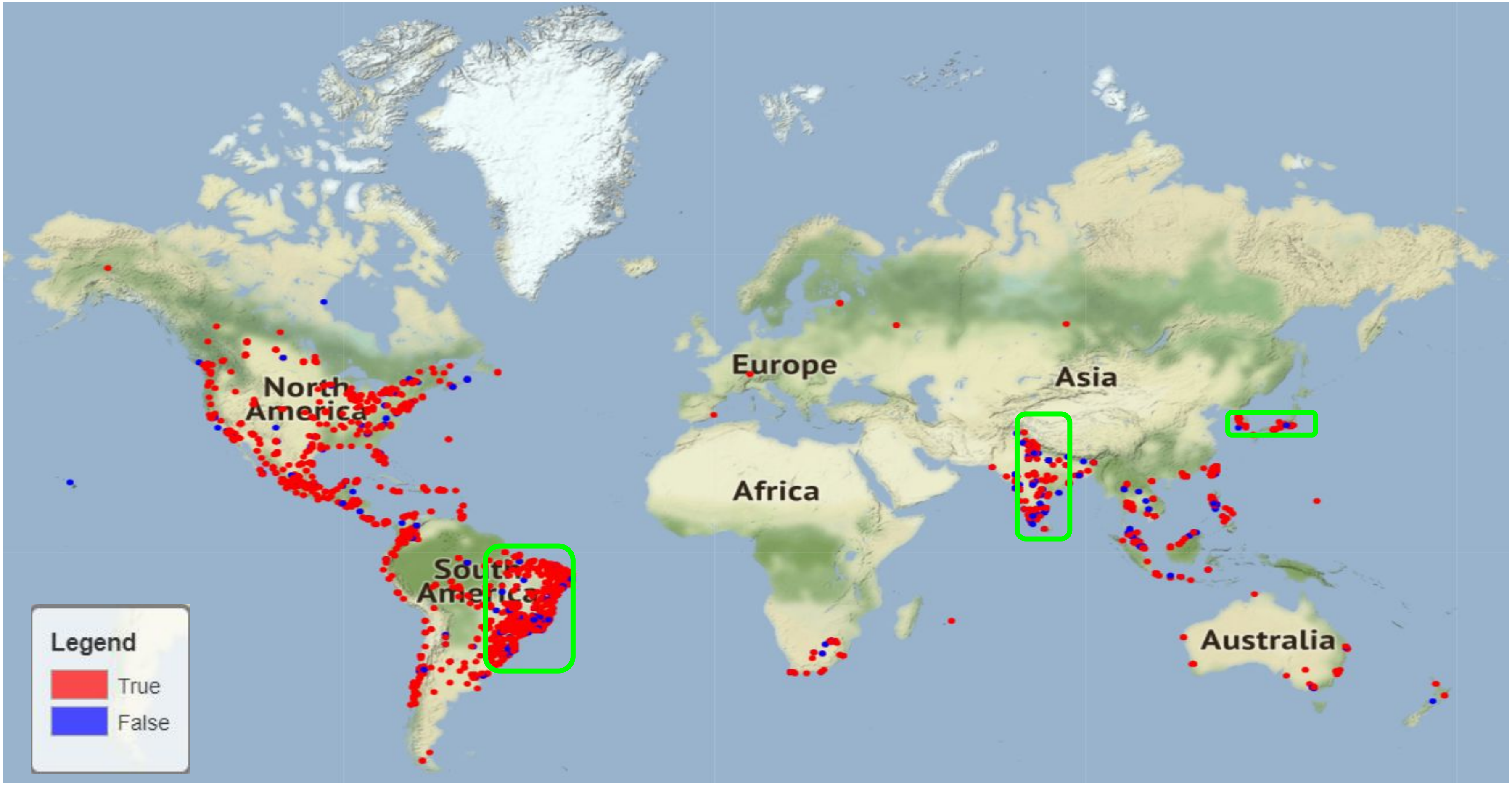


Providers	TV_hbos	0	
Bharti Airtel	n	724	
Bharti Airtel	o	165	18%
CLARO S.A.	n	3286	
CLARO S.A.	o	232	6%
Idea Cellular	n	160	
Idea Cellular	o	53	24%
OTHER	n	25459	
OTHER	o	2593	9%
Reliance Jio	n	1201	
Reliance Jio	o	483	28%
TELEFÃƒâ€NICA BRASIL	n	3149	
TELEFÃƒâ€NICA BRASIL	o	269	9%
Telemar Norte	n	988	
Telemar Norte	o	97	8%
Tim Celular	n	1732	
Tim Celular	o	134	10%
Vodafone India	n	351	
Vodafone India	o	79	18%

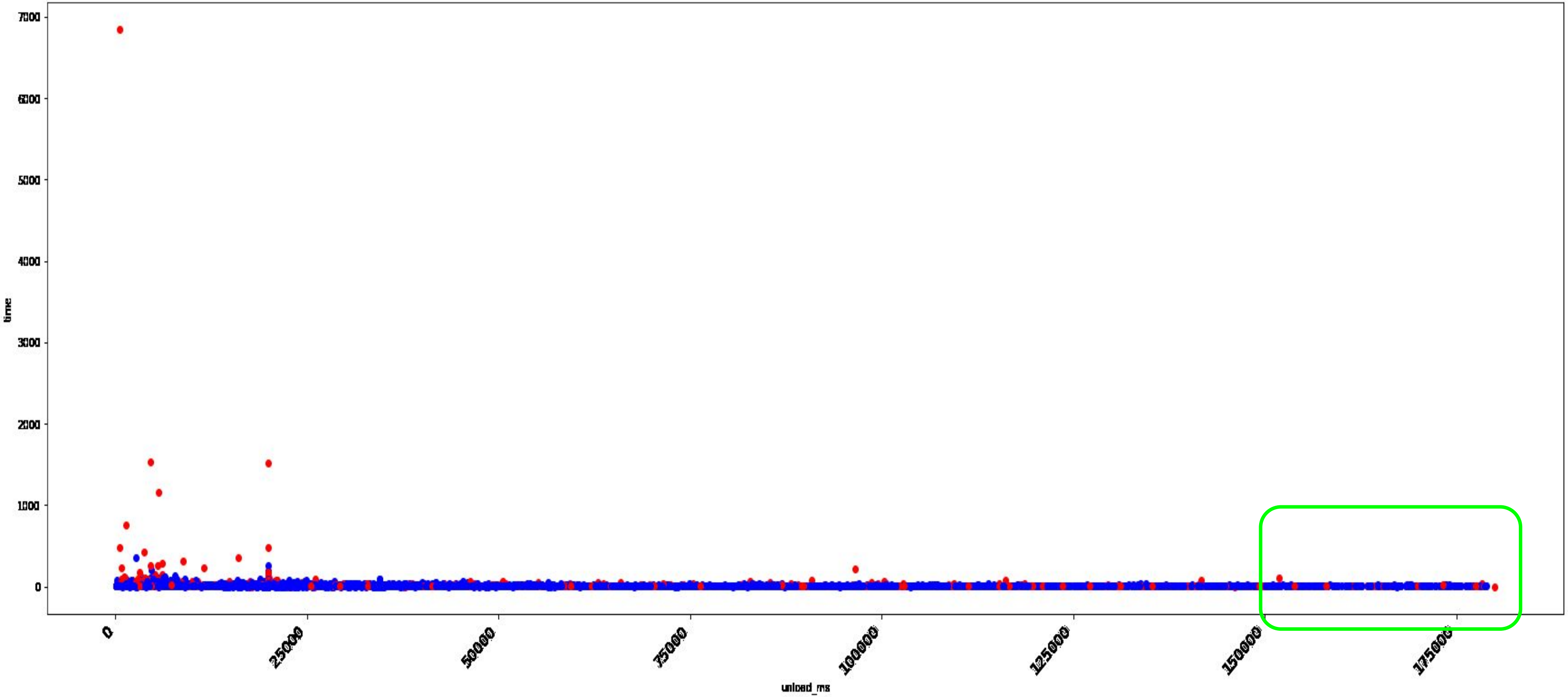
True and false callbacks on map - lof



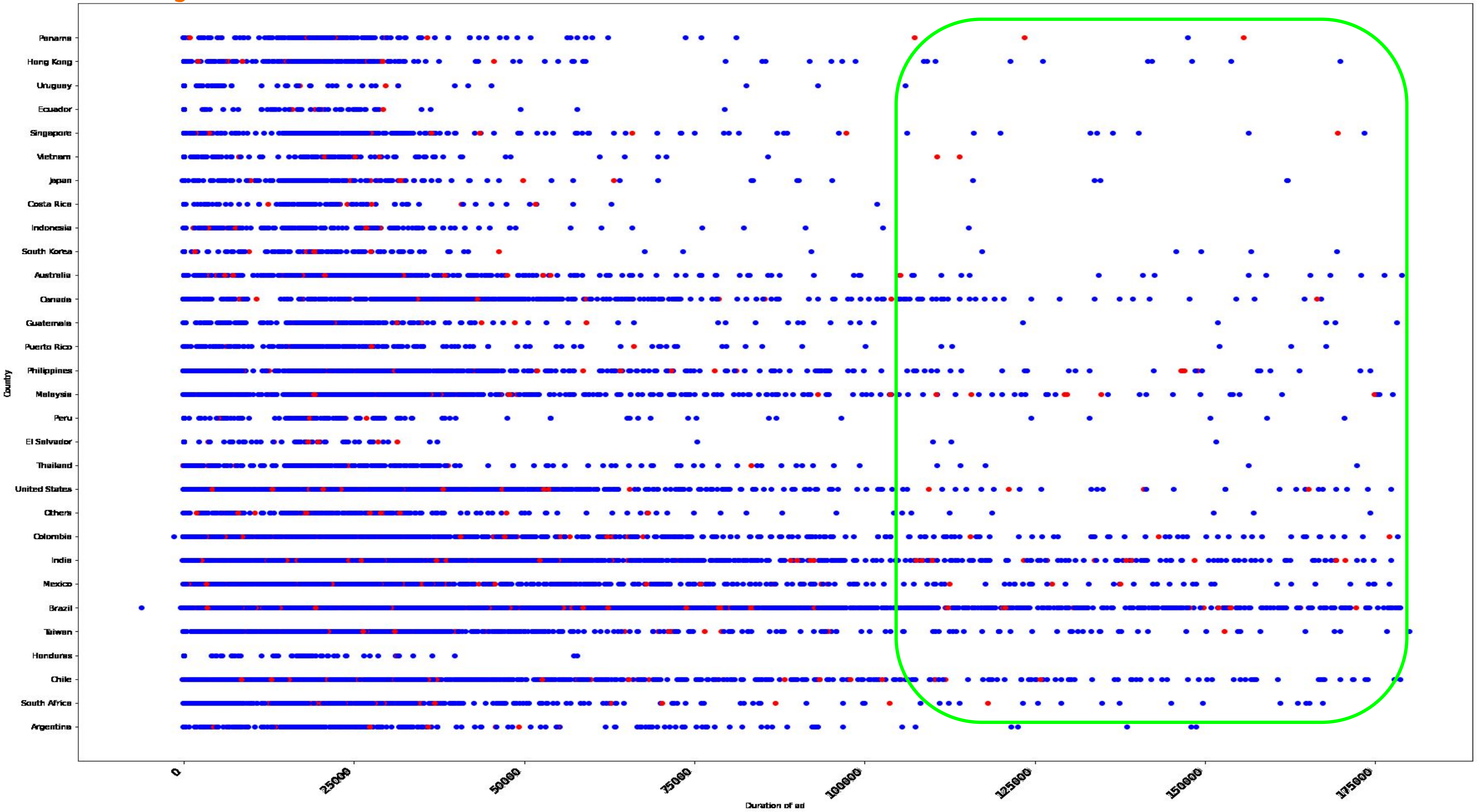
True and false callbacks on map - hbos



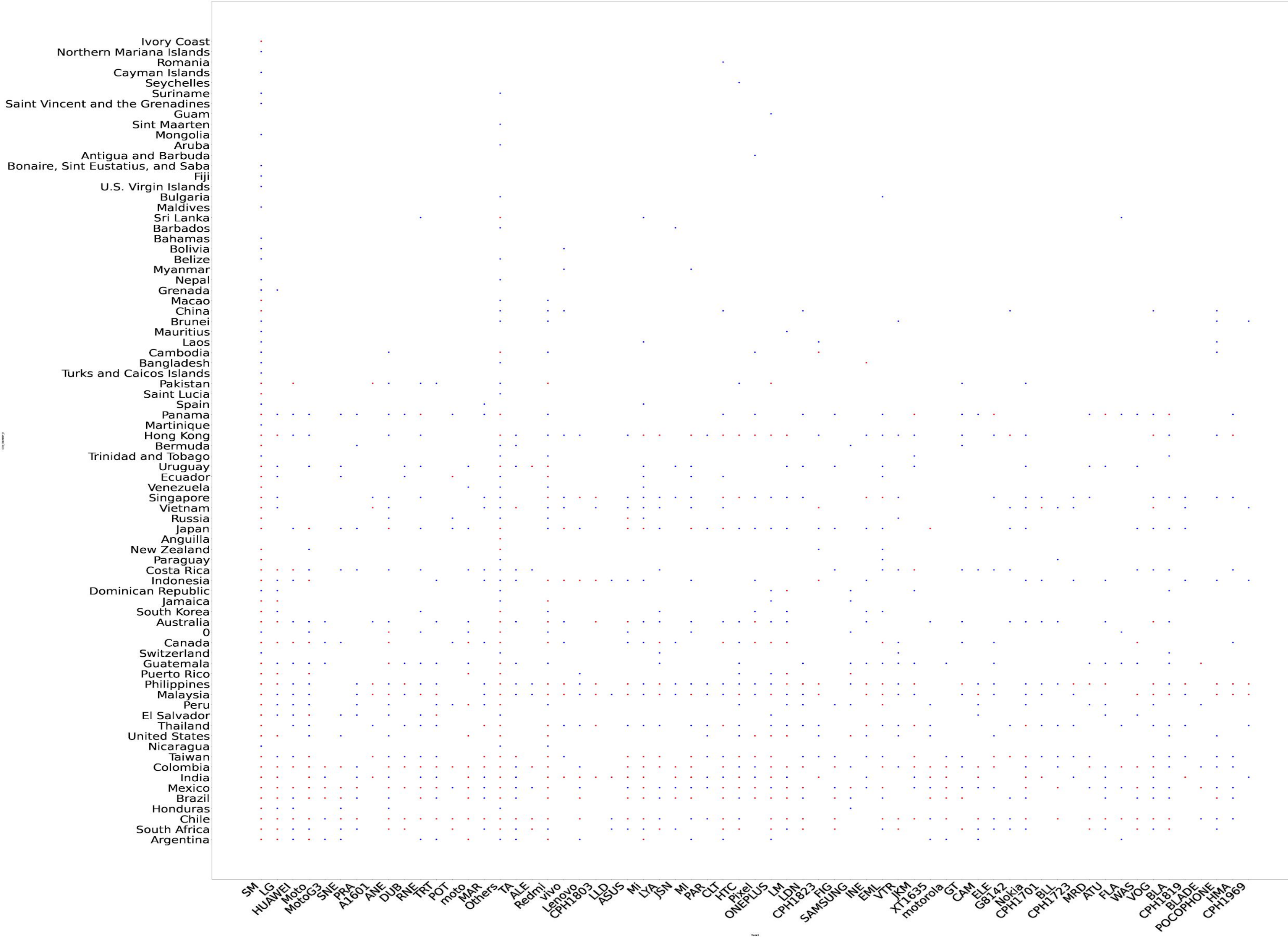
Start time v/s unload_ms v/s label



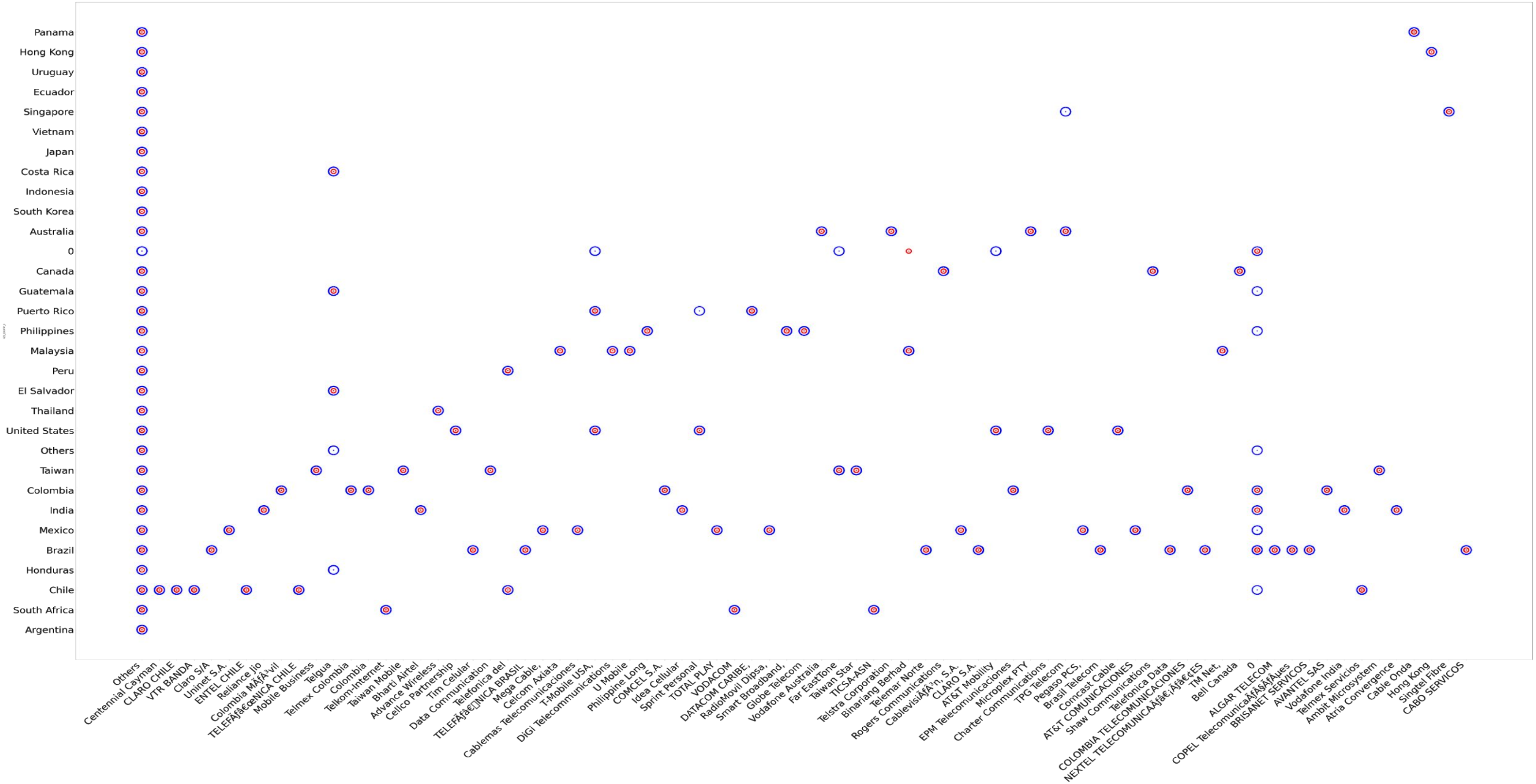
Country v/s duration of ad v/s label



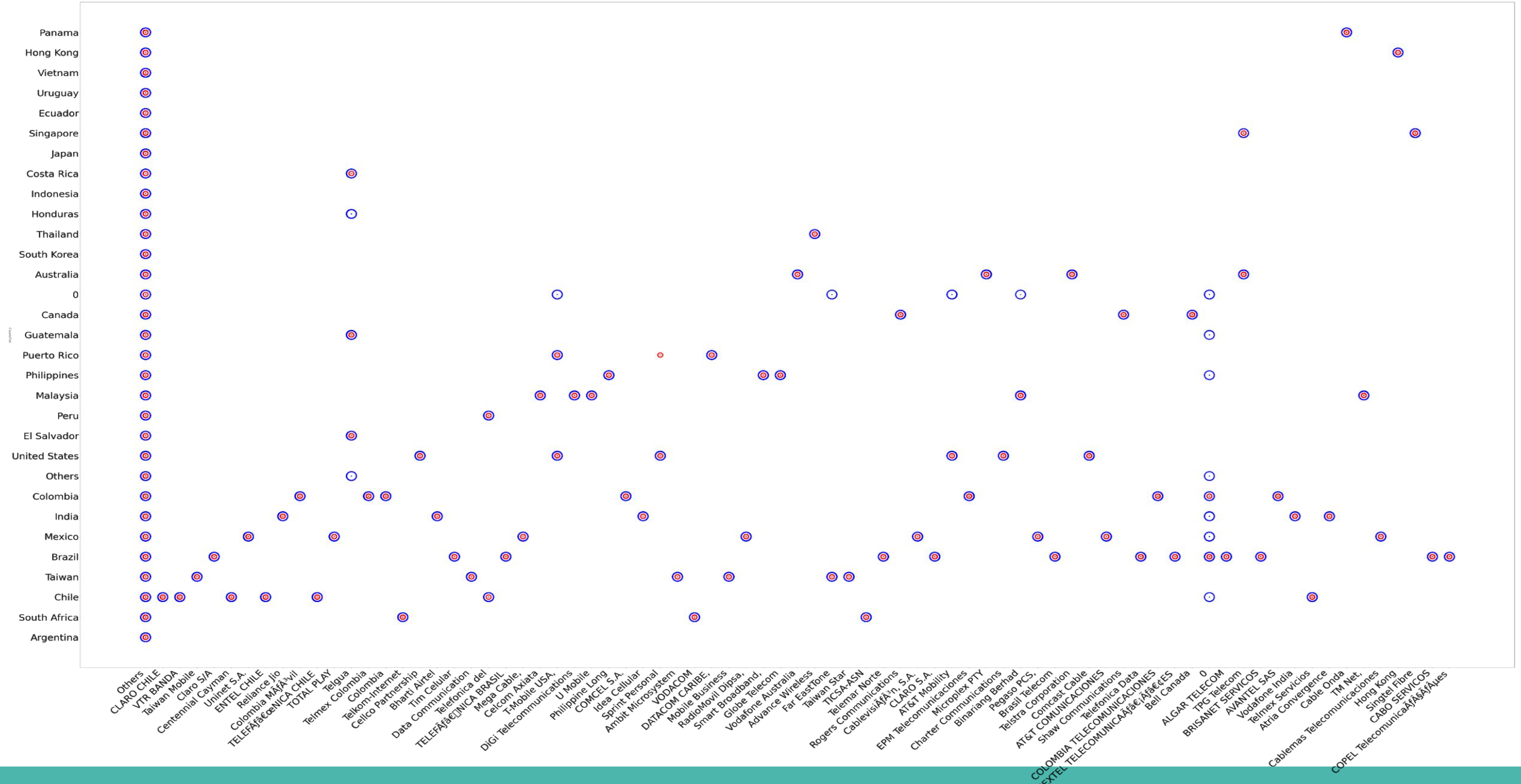
Country v/s models v/s label



Country v/s provider v/s label - lof



Country v/s provider v/s label - hbos



Bad bots vs good bots

- Bots are autonomous programs on a network (especially the Internet) which can interact with systems or users.
- Some Bots are especially designed to behave like a person on the network.
- A **good** bot is any bot that performs useful or **helpful tasks that aren't detrimental to a user's experience** on the Internet.
- **Bad** bots **scrape data from sites without permission** in order to reuse it (e.g., pricing, inventory levels) and gain a competitive edge. The truly nefarious ones undertake criminal activities, such as fraud and outright theft.
- Because good bots can share similar characteristics with malicious bots, the challenge is ensuring good bots aren't blocked when putting together a bot management strategy.
- Bad bots generally **spend more time, and occupy more memory on the site and servers.**
- With enough data one can differentiate between good and bad bots.

Finding bad bots / fake users

1. **Set duration of ad threshold to 60% of max duration (0.12 seconds)**
2. **Set memory used to threshold of 60% of max.**
3. **Any callbacks that uses higher time on site and memory than the thresholds, is considered a bad bot/user.**
4. **List out countries, devices and provider that have been used by said bots.**

Countries and their frequency of bad bots:

'Brazil': 18, 'India': 7, 'Taiwan': 6, 'Colombia': 4, 'Canada': 3, 'Chile': 3, 'Mexico': 3, 'Others': 2, 'Australia': 2, 'Philippines': 1, 'South Africa': 1, 'Hong Kong': 1, 'Malaysia': 1, 'Panama': 1, 'Japan': 1, 'Singapore': 1, 'United States': 1

Service providers and their frequency of bad bots:

'Taiwan Mobile': 1, 'Mobile Business': 1, 'Philippine Long': 1, 'Telstra Corporation': 1, 'COMCEL S.A.': 1, 'Others': 8, 'TELEFÃƒâ€œNICA, CHILE': 1, 'DiGi Telecommunications': 1, 'Reliance Jio': 1, 'TELEFÃƒâ€\x9dNICA BRASIL': 1, 'CLARO S.A.': 1, 'Tim Celular': 1, 'CentennialCayman': 1, 'Data Communication': 1, 'Colombia MÃƒÂ³vil': 1, 'TPG Telecom': 1, 'Vodafone India': 1, 'CLARO CHILE': 1, 'TOTALPLAY': 1, 'Uninet S.A.': 1, 'Far EastTone': 1, 'Cable Onda': 1, 'Telmex Colombia': 1, 'Brasil Telecom': 1, 'Bharti Airtel': 1, 'Mega Cable,':1, 'Bell Canada': 1, 'Telemar Norte': 1

Devices and their frequency of bad bots:

'SM': 1, 'Others': 2, 'INE': 1, 'ANE': 1, 'ONEPLUS': 1, 'MI': 1, 'vivo': 1, 'moto': 1, 'ASUS': 1, 'G8142': 1, 'JSN': 1, 'Redmi': 1, 'Pixel': 1, 'MAR': 1, 'Moto': 1, 'LG': 1, 'HTC': 1, 'motorola': 1, 'LM': 1, 'Mi': 1