Analysis Report

1. WhatsApp Chat

- Clearly there is **no mention** for **Dabur band** in the chats.

- However there are products, in which **Dabur deal like: massage oil**(coconut, olive and mustard oil) which are talked about quite often with frequencies 28, 14 and 8 respectively.

- Also there is clear mentions of brands like **Vicks**, **Cerelac** and **Himalaya** especially(**Bonnison**)with frequencies 16, 15, 7 and 12 respectively as market-leader in their specific products which according to dabur lack.

- **Dabur** lal tel is a prominent product but there are discussions in group to prepare **home-made remedies** to counter this product.

- I did sequential **cleaning**, **tokenizing** and represented words as a **vector**. The following notable things came out of clusters and word frequencies.

Link for codes: https://github.com/yashchoubey/BabyDestination/blob/master/whatsApp%20.ipynb

The following tuples of word and its frequency tell us that groups **theme** is about mommies and their babies. ('baby', 640), ('babies', 128), ('mommies', 100), ('mom', 27)

The following cluster prove the **major concern** is breastfeeding:
('milk', 195), ('breast', 48), ('breastmilk', 7), ('feed', 70), ('feeding', 67), ('breastfeeding', 29)

Also the clusters of ('oil', 91), ('coconut', 28), ('massage', 26), ('mustard', 8), ('garlic', 14), ('olive', 14), suggests that **massage** was the prominent point for discussion.

The following clusters and respective frequencies helps us understand that **cold, cough and congestion** is most frequent disease in babies. ('cold', 74), ('cough', 41), ('doctor', 39), ('chest', 25), ('congestion', 4)**.**

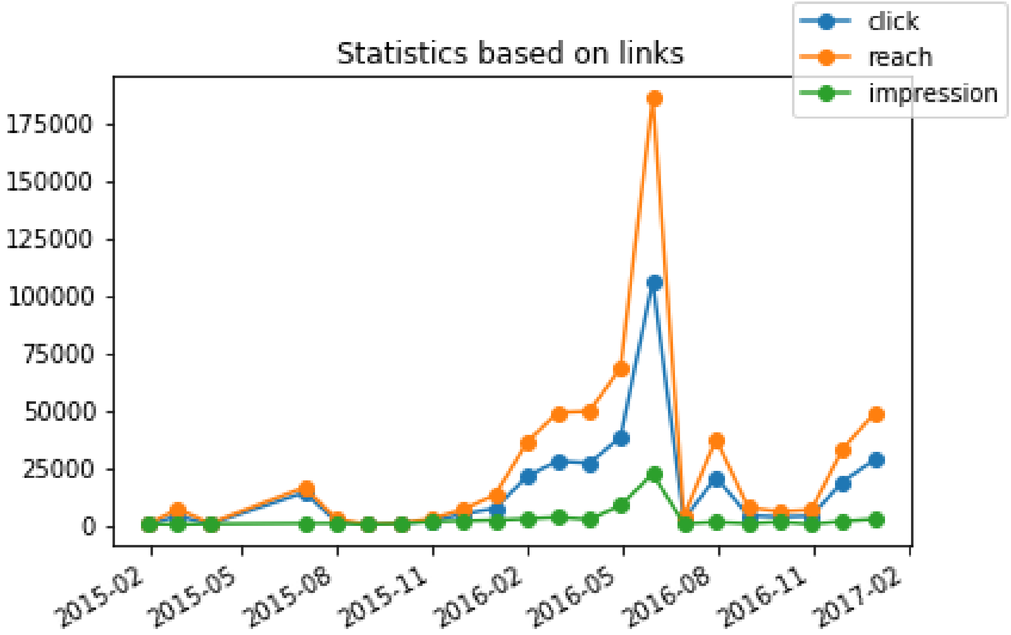
Also for **feeding** the mentioned items were most discussed:
('ragi', 57), ('porridge', 29), ('powder', 47), ('rice', 36), ('juice', 31), ('fruits', 29),('puree', 28), ('banana', 23), ('ghee', 20), ('egg', 19),('apple', 18), ('cerelac', 15),('honey', 14),('potato', 12),('ajwain', 11), ('jaggery', 10),('rasam', 8), ('grapes', 10)

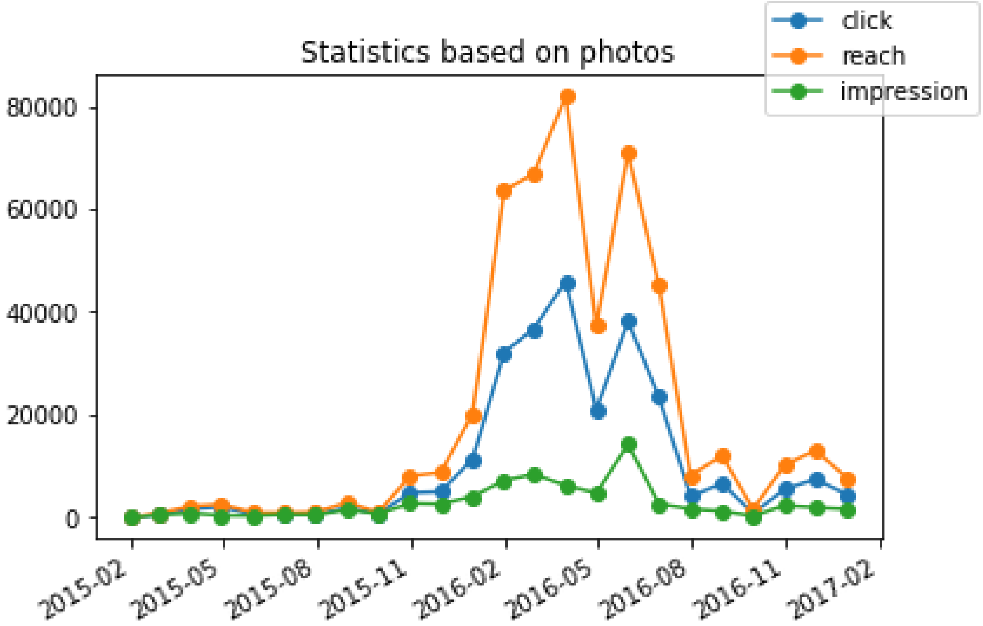
We can easily find clusters of medical related terms and can suggest related products: ('pain', 55), ('fever', 42), ('infection', 20), ('vaccine', 19), ('medicines', 18), ('blood', 13), ('pregnancy', 13), ('marks', 10), ('stool', 10), ('diseases', 10), ('constipation', 10), ('digestion', 8), ('polio', 11)

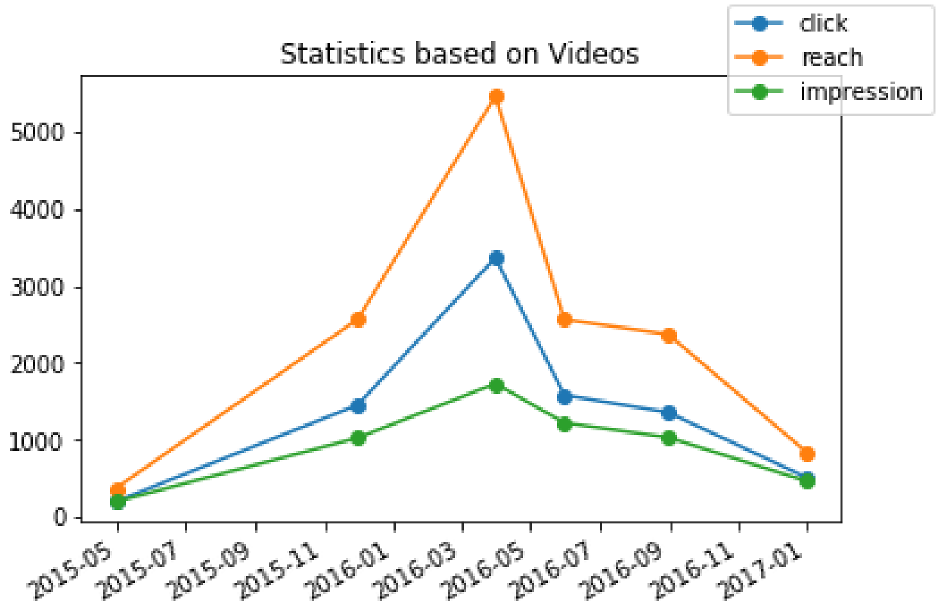
Also I find a lot of discussion on **hair care** after pregnancy which is suggested by ('hair', 29)

2. Data Analysis

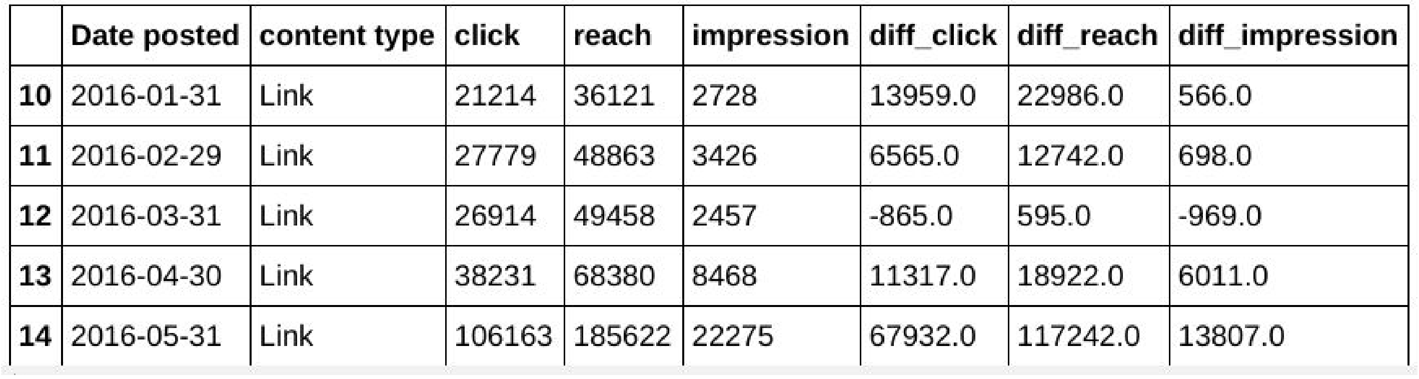
I segregated the data in links, photos and videos and then add them monthly to get the desired output.
Link for codes: https://github.com/yashchoubey/BabyDestination/blob/master/DataAnalysis.ipynb



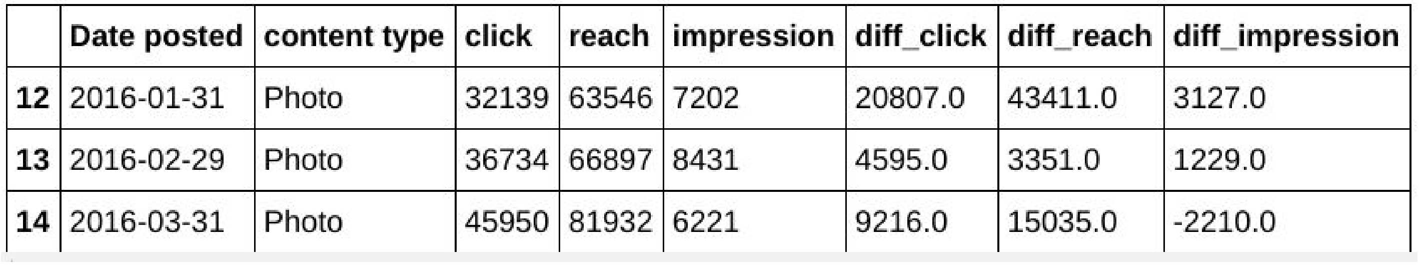




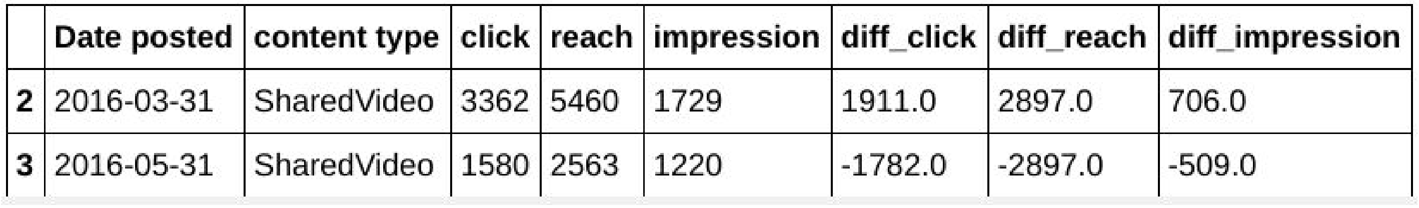
Monthly changes for reach in given months are as follows:



	Date posted	content type	click	reach	impression	diff_click	diff_reach	diff_impression
10	2016-01-31	Link	21214	36121	2728	13959.0	22986.0	566.0
11	2016-02-29	Link	27779	48863	3426	6565.0	12742.0	698.0
12	2016-03-31	Link	26914	49458	2457	-865.0	595.0	-969.0
13	2016-04-30	Link	38231	68380	8468	11317.0	18922.0	6011.0
14	2016-05-31	Link	106163	185622	22275	67932.0	117242.0	13807.0



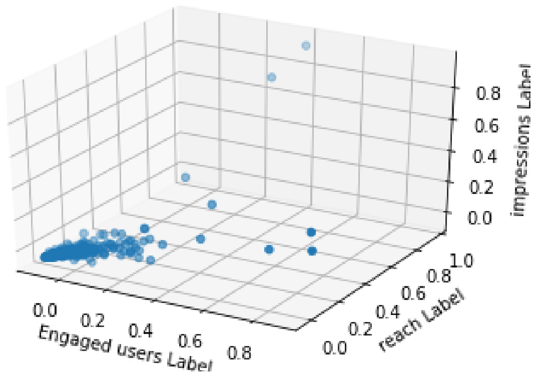
	Date posted	content type	click	reach	impression	diff_click	diff_reach	diff_impression
12	2016-01-31	Photo	32139	63546	7202	20807.0	43411.0	3127.0
13	2016-02-29	Photo	36734	66897	8431	4595.0	3351.0	1229.0
14	2016-03-31	Photo	45950	81932	6221	9216.0	15035.0	-2210.0

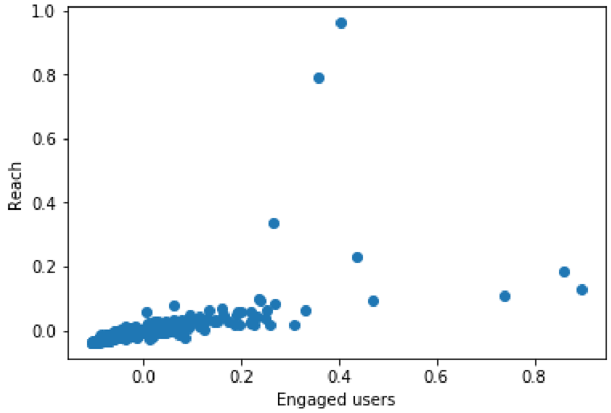


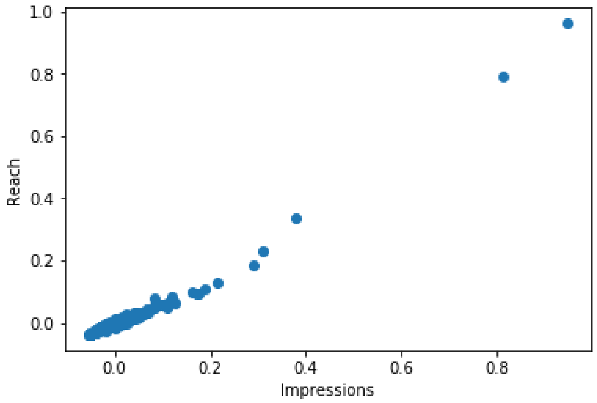
	Date posted	content type	click	reach	impression	diff_click	diff_reach	diff_impression
2	2016-03-31	SharedVideo	3362	5460	1729	1911.0	2897.0	706.0
3	2016-05-31	SharedVideo	1580	2563	1220	-1782.0	-2897.0	-509.0

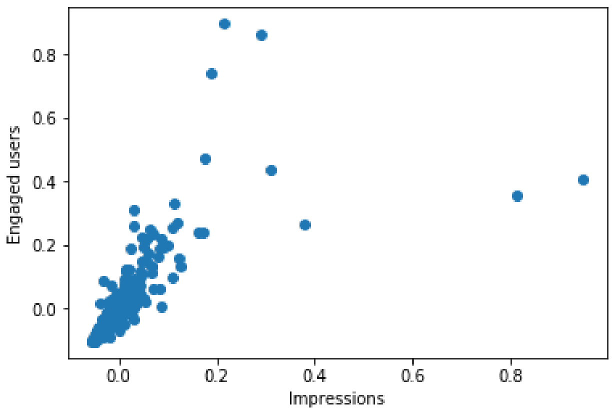
Regression model:

The data was normalized on range 0 to 1. The following plots clearly shows the scatteredness of data.









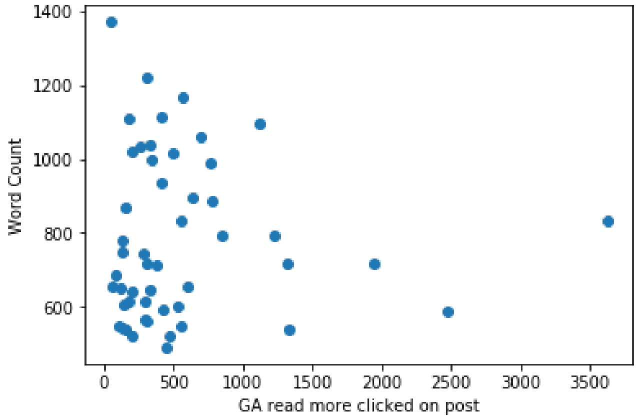
I used the ridge regression with default settings to train the model.
Link for codes: https://github.com/yashchoubey/BabyDestination/blob/master/DataAnalysis.ipynb

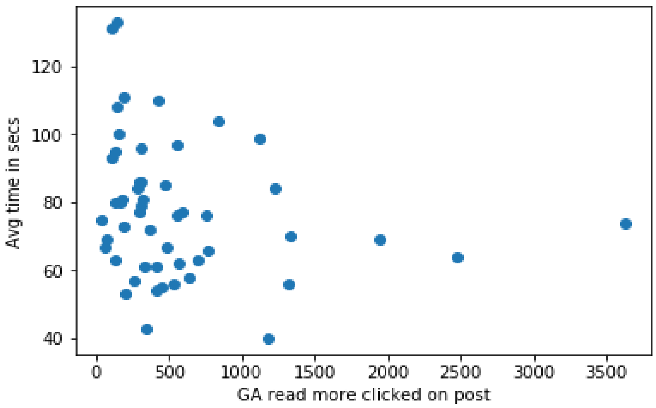
The following are the key findings:

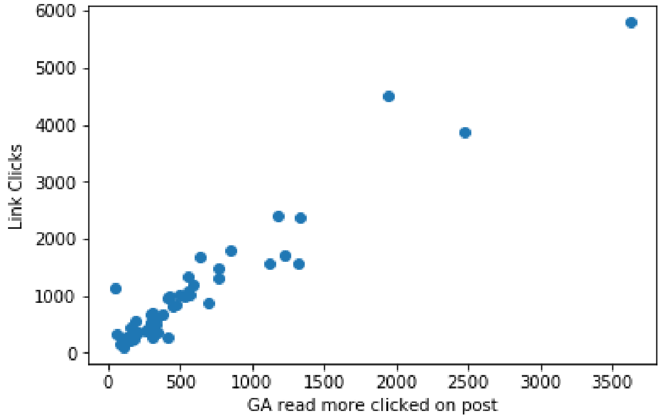
1. On cross-validation using ‘neg_mean_absolute_error’ the results are : Mean -45.576 with Standard Deviation:1.427
2. Such a high deviation from mean is due to the fact of scatteredness of data.
3. Simple using Reach and Impressions as feature we can get lower mean absolute error.

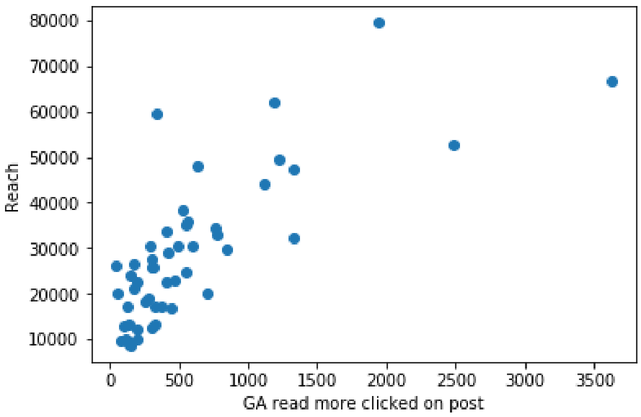
Rolling-mean for 30 days: https://github.com/yashchoubey/BabyDestination/blob/master/rolling_mean.ipynb

Improve read more click rate







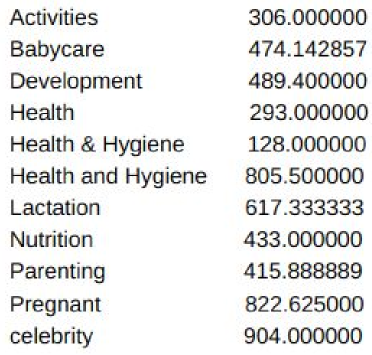


The following can be inferred from the data:

- Read more click increases with increase on Reach approximately .
- Read more click increases linearly with Link Clicks.
- Read more click is independent of word count.
- Read more click is somewhat inversely dependent on Avg time spent.

Furthermore based on the mean read more clicks category wise data shows:

- Posts related to celebrity have max read more clicks.
- Categories like Pregnant, Health and Hygiene and Lactation also have a good probability of read more clicks.



```
Activities              306.000000
Babycare                474.142857
Development             489.400000
Health                  293.000000
Health & Hygiene        128.000000
Health and Hygiene      805.500000
Lactation               617.333333
Nutrition               433.000000
Parenting               415.888889
Pregnant                822.625000
celebrity               904.000000
```