UNIVERSITY OF
TEXAS
ARLINGTON

INSY–5337 WEB AND SOCIAL ANALYTICS

PROJECT REPORT

TABLE OF CONTENTS

INTRODUCTION:

We are a brand-new company planning to launch a new energy drink. **The Business goal is to make the brand name reach a large population as a first step.** Many celebrities do paid partnerships with companies to promote products on social media platforms such as Facebook, Instagram, and Twitter. We are in search of such a celebrity **to launch and promote the product.**

APPROACH:

In the present world, it is easy to introduce a product, but the major problem lies in getting the product to customers' hands and making people all over the world made aware of the product. So, to reach a wide variety of audiences, we will be taking an athlete with a huge fan following as well as a good reach for their posts on Facebook.

We tried mining the required data from Facebook. But, when we attempted to mine data it gave us a permission error, even after providing permissions for all the things which were mentioned on Facebook. Turns out, we can only access data on our pages. Data extraction from all other pages has been restricted from version 2.8 and the API doesn't allow us to use prior versions.

We overcame the above problem with the help of an online analytics tool called **'popsters'**. In short, popsters is a tool that helps to get compare stats and measure the efficiency of posts on any social media page that we are interested in.

DATA DESCRIPTION & SOURCE:

DATA SOURCE:

The below dataset was created by our **team manually**. We have filtered the top 50 cricket players who were all **active on social media**. And for each player, using their profile link and with the help of the popsters tool, we filled the values for all this column. We have 22 columns of data. Out of which, 3 are Qualitative variables and the remaining 19 are Quantitative variable

We have fixed a Period of **6 Months, starting from October 2021 to April 20222**. Though most of the columns were directly imported from the online tools, a few other columns were calculated by us. For examples fields such as Promotional post, Promotional Post Percent, and Average Post Frequency. A detailed description of every column with its uses and formula was given below the dataset in a separate table.

Players	Node ID	Profile Link	Followers	Posts	Likes	Shares	Comments	Avg. Likes	Avg. Shares	Avg. Comments	Likes Ratio	Shares ratio
DAVID MILLER	100044343763446	https://www.facebook.com/100044343763446	5493502	6	369803	317	2389	61633	52	398	1.12190%	5.77045E-05
YUZVENDRA SINGH CHAHAL	100044284813471	https://www.facebook.com/100044284813471	1281491	2	3383	9	103	1691	4	51	0.26399%	7.02307E-06
MOAHMMED SIRAJ	100044794445847	https://www.facebook.com/100044794445847	1358079	1	36731	202	1090	36731	202	1090	2.70460%	0.00014874
PRITHVI SHAW	100044395660694	https://www.facebook.com/100044395660694	2242883	2	177277	2097	371	88638	1049	185	7.90398%	0.00093457
BHUVENSHWAR KUMAR	100044317335294	https://www.facebook.com/100044317335294	10376635	2	32981	103	2536	16490	51	1268	0.15890%	9.92615E-06
YUVARAJ SINGH	100044286886682	https://www.facebook.com/100044286886682	19064947	5	81059	3124	1767	16211	624	353	0.42517%	0.000163861
SAM CURRAN	100057200719768	https://www.facebook.com/100057200719768	2547182	2	117275	250	2195	58637	125	1097	4.60411%	9.81477E-05
AJINKYA RAHANE	100044289256775	https://www.facebook.com/100044289256775	8476878	4	13591	43	378	3397	10	94	0.04010%	5.07262E-06
DAVID WARNER	100044542354906	https://www.facebook.com/100044542354906	7695287	2	4681	33	501	2340	16	250	0.06083%	4.28834E-06
RAHUL TRIPATHI	100057597272933	https://www.facebook.com/100057597272933	553257	8	204413	1546	4004	25551	193	500	36.94721%	0.002794361
RAHUL TEWATIA	100046826631970	https://www.facebook.com/100046826631970	1221964	16	129640	95	1155	8102	5	72	10.60915%	7.77437E-05
SANJU SAMSON	100044281486786	https://www.facebook.com/100044281486786	2935248	29	813791	1565	15553	28061	53	536	27.72478%	0.000533175
RAVICHANDRAN ASWIN	100044160351587	https://www.facebook.com/100044160351587	4074691	78	203638	2441	4131	2611	31	52	4.99763%	0.000599064
UMESH YADAV	100044509476307	https://www.facebook.com/100044509476307	6211135	8	39665	70	1273	4958	8	159	0.63861%	1.12701E-05
ROBIN UTHAPPA	100044339533916	https://www.facebook.com/100044339533916	2889016	86	997335	1476	5711	11596	17	66	0.40140%	0.000510901
JASPRIT BUMRAH	100044444169040	https://www.facebook.com/100044444169040	9395089	37	1757146	5870	20066	47490	158	542	18.70281%	0.000624795
KULDEEP YADAV	100044197371116	https://www.facebook.com/100044197371116	1947454	23	153799	345	3736	6686	15	162	7.89744%	0.000177154
ROHIT SHARMA	100044240029060	https://www.facebook.com/100044240029060	20931904	20	1332768	5011	40543	66638	250	2027	6.36716%	0.000239395
AB De	100044338843815	https://www.facebook.com/ABdeVilliers17	13216242	43	4768558	53539	105897	110896	1245	2462	36.08104%	0.004051
MS DHONI	100044481937493	https://www.facebook.com/100044481937493	27214451	20	647915	20816	24248	32395	1040	1212	2.38078%	0.000764888
RASHID KHAN	100044281426914	https://www.facebook.com/100044281426914	2729178	25	2063184	10649	146942	82527	425	5877	75.59727%	0.003901907
CHRIS GAYLE	100044317545336	https://www.facebook.com/100044317545336	13851349	10	93404	380	3544	9340	38	354	0.67433%	2.74342E-05
HARDIK PANDYA	100044386001022	https://www.facebook.com/100044386001022	11907722	23	1241491	4456	23454	53977	193	1019	10.42593%	0.000374211
Sachin Tendulkar	100044226919512	https://www.facebook.com/100044226919512	37575565	111	6552020	177044	170306	59027	1594	1534	17.43692%	0.004711679
RISHABH PANT	100044527715691	https://www.facebook.com/100044527715691	5582288	42	1003576	3096	18924	23894	73	450	17.97786%	0.000554611
SHIKHAR DHAWAN	100044278427035	https://www.facebook.com/100044278427035	17140774	54	1802822	13689	35332	33385	253	654	10.51774%	0.000798622
KRUNAL PANDYA	100044260937684	https://www.facebook.com/100044260937684	2262069	19	161828	428	2115	8517	22	111	7.15398%	0.000189207
AKSHAR PATEL	100044547304498	https://www.facebook.com/100044547304498	2360426	35	102415	155	2213	2926	4	63	4.33884%	6.56661E-05
RAHUL CHAHAR	100018555806290	https://www.facebook.com/100018555806290	525181	38	849996	878	6916	22368	23	182	161.84820%	0.001671805
SHREYAS GOPAL	100044447798813	https://www.facebook.com/100044447798813	224754	21	148088	157	1069	7051	7	50	3.13760%	0.000698542
KL RAHUL	100044235379214	https://www.facebook.com/100044235379214	10284388	54	3027838	11552	71043	56071	213	1315	29.44111%	0.001123256
Virat Kohli	100044473084573	https://www.facebook.com/100044473084573	49145016	97	24339754	166883	981214	250925	1720	10115	49.52639%	0.003395726
Suresh Raina	100044298046085	https://www.facebook.com/SureshKumarRainaOfficial	6043506	86	2347646	10220	32244	27298	118	374	38.84576%	0.001691071
RAVINDRA JADEJA	100045146962 13	https://www.facebook.com/100045146962 13	10844998	30	1671536	9211	33529	55717	307	1117	15.41297%	0.000849332
MAYANK AGARWAL	100044437192 33	https://www.facebook.com/100044437192 33	2384770	67	469282	600	5037	7004	8	75	19.67829%	0.000251597
SHREYAS IYER	100044290546403	https://www.facebook.com/100044290546403	2688432	27	572847	2345	14048	21216	86	520	21.30785%	0.000872256
ISHAN KISHAN	100044148800557	https://www.facebook.com/100044148800557	3727674	21	632491	1752	15440	30118	83	735	16.96744%	0.000469998
LASITH MALINGA	100043925989321	https://www.facebook.com/100043925989321	3592181	77	618885	11017	55279	8037	143	717	0.22370%	0.003066939
HARBHAJAN SINGH	100044143374060	https://www.facebook.com/100044143374060	10224052	92	835480	4876	35922	9081	53	390	0.08880%	0.000476915
WASHINGTON SUNDAR	100051414284787	https://www.facebook.com/100051414284787	509723	72	794749	1920	9223	11038	26	128	155.91782%	0.003766752

Players	Comments Ratio	Promotional Posts	Promotional Post percent	Average Post Frequency	Engagement Rate / post	Engagement Rate / day	Highest Active Day	Highest day active percent	Time of the day
DAVID MILLER	0.000434877	2/20	10	0.032967033	1.13%	0.04%	Sunday	40:60%	02:00
YUZVENDRA SINGH CHAHAL	8.03751E-05	2/2	100	0.010989011	0.14%	0.00%	Saturday	85.15	08:00
MOAHMMED SIRAJ	0.000802604	0/30	0	0.005494505	2.80%	2.80%	Tuesday	100%	03:00
PRITHVI SHAW	0.000165412	15/20	75	0.010989011	4.01%	1.00%	Wednesday	100%	02:00
BHUVENSHWAR KUMAR	0.000244395	1/20	5	0.010989011	0.07%	0.00%	Wednesday	96.02%	10:00
YUVARAJ SINGH	9.26832E-05	8/20	40	0.027472527	0.09%	0.04%	Friday	72.13%	23:00
SAM CURRAN	0.000861737	0/2	0	0.010989011	2.35%	1.57%	Thursday	64.36%	04:00
AJINKYA RAHANE	4.45919E-05	2/20	10	0.021978022	0.04%	0.00%	Wednesday	50.95%	06:00
DAVID WARNER	6.51048E-05	2/2	100	0.010989011	0.03%	0.02%	Thursday	50.58%	13:00
RAHUL TRIPATHI	0.007237143	4/8	50	0.043956044	4.73%	0.22%	Friday	47.35%	05:00
RAHUL TEWATIA	0.0009452	2/20	10	0.087912088	0.37%	0.67%	Tuesday	39.64%	05:00
SANJU SAMSON	0.0052987	2/15	13.33333333	0.159340659	0.98%	0.17%	wednesday	38.45%	00:00
RAVICHANDRAN ASWIN	0.001013819	20/20	100	0.428571429	0.07%	0.03%	Saturday	37.62%	09:00
UMESH YADAV	0.000204954	1/20	5	0.043956044	0.08%	0.03%	Friday	37.55%	14:00
ROBIN UTHAPPA	0.001976798	2/20	10	0.472527473	0.40%	0.21%	Thursday	34.31%	14:00
JASPRIT BUMRAH	0.002135797	1/20	5	0.203296703	0.51%	0.11%	Tuesday	32.64%	05:00
KULDEEP YADAV	0.001918402	2/20	10	0.126373626	0.35%	0.30%	Tuesday	31.57%	09:00
ROHIT SHARMA	0.0019369	8/20	40	0.10989011	0.33%	0.04%	Sunday	30.73%	05:00
AB De	0.008012641	8/10	80	0.236263736	0.87%	0.22%	Thursday	30.72%	13:00
MS DHONI	0.000890997	14/15	93.33333333	0.10989011	0.13%	0.02%	wednesday	28.64%	05:00
RASHID KHAN	0.05384112	3/15	20	0.137362637	3.25%	0.52%	sunday	27.87%	13:00
CHRIS GAYLE	0.00025586	11/20	55	0.054945055	0.07%	0.00%	Wednesday	26.53%	09:00
HARDIK PANDYA	0.001969646	3/10	30	0.126373626	0.46%	0.06%	Friday	26.45%	21:00
Sachin Tendulkar	0.00453236	4/20	20	0.60989011	0.17%	0.10%	Thursday	26.38%	09:00
RISHABH PANT	0.003390008	10/20	50	0.230769231	0.44%	0.11%	Sunday	25.21%	03:00
SHIKHAR DHAWAN	0.002061284	8/20	40	0.296703297	0.20%	0.06%	Sunday	24.94%	09:00
KRUNAL PANDYA	0.000934985	5/20	25	0.104395604	0.38%	0.04%	Tuesday	24.84%	02:00
AKSHAR PATEL	0.000937543	7/20	35	0.192307692	0.13%	0.15%	Saturday	24.12%	06:00
RAHUL CHAHAR	0.013168793	8/20	40	0.208791209	4.30%	0.00%	Monday	23.64%	03:00
SHREYAS GOPAL	0.004756311	3/20	15	0.115384615	3.16%	0.38%	Thursday	23.52%	01:00
KL RAHUL	0.006907849	6/20	30	0.296703297	0.56%	0.17%	Friday	23.19%	13:00
Virat Kohli	0.019965687	6/20	30	0.532967033	0.53%	0.29%	Thursday	21.66%	12:00
Suresh Raina	0.005335314	2/10	20	0.472527473	0.46%	0.22%	Friday	20.32%	16:00
RAVINDRA JADEJA	0.003091656	4/20	20	0.164835165	0.53%	0.10%	Sunday	20.06%	09:00
MAYANK AGARWAL	0.002112153	0/20	0	0.368131868	0.30%	0.11%	saturday	19.79%	04:00
SHREYAS IYER	0.005225351	0/20	0	0.148351648	0.81%	0.12%	Monday	19.38%	08:00
ISHAN KISHAN	0.004141993	1/20	5	0.115384615	0.83%	0.11%	Friday	19.14%	05:00
LASITH MALINGA	0.015388701	9/20	45	0.423076923	0.25%	0.11%	Sunday	17.88%	10:00
HARBHAJAN SINGH	0.00351348	3/20	15	0.505494505	0.09%	0.05%	Thursday	17.83%	15:00
WASHINGTON SUNDAR	0.018094141	6/20	30	0.395604396	2.20%	0.87%	Thursday	16.25%	07:00

DATA DESCRIPTION:

Column Name	Description
Players	Name of the Cricket Players
Node ID	Unique Node ID which represents the Player
Profile Link	The Players Facebook Profile Link
Followers	The Total Number of followers that the player has on their Facebook Profile as of April 2022
Posts	The Total Number of posts published on the Player's profile as of April 2022
Likes	The Total Count of like for all the posts combined as of April 2022. Likes Include all the reactions from the Audience
Shares	The Total Count of Shares for all the posts published as of April 2022. Shares of Shares will also come under this category.
Comments	The Total Count of Comments for all the posts combined as of April 2022.
Average Likes	Average Likes are calculated by the formula Total number of likes / Total number of posts
Average Shares	Average Likes are calculated by the formula Total number of Shares / Total number of posts
Average Comments	Average Likes are calculated by the formula Total number of Comments / Total number of posts
Likes Ratio	The likes Ratio is calculated by the formula Total Number of Likes / Total Number of Followers
Shares Ratio	The shares Ratio is calculated by the formula Total Number of Likes / Total Number of Followers
Comments Ratio	Comments Ratio is calculated by the formula Total Number of Likes / Total Number of Followers
Promotional Post	The number of Promotional Posts to the Number of Total Posts. Total Post include both personal posts and promotional posts
Promotional Post Percentage	The Promotional Post is converted into the form of a Percentage
Average Post Frequency	To Find How Frequently player posts on their social media handle. Calculated by the formula Total Number of Post/ Number of Days
Engagement Rate / Post	Engagement Rate is a metric that shows how many followers were engaged with the published posts. ERpost = (likes+shares+comments) / count of followers
Engagement Rate / Day	ERday= (total likes + total shares + total comments for the day) / count of followers

Highest Active Day	The Day on which the Player is highly Active. It helps in finding out the pattern.
Highest Day Active Percent	The Percentage of Activity of the most active day of the week
Time of the Day	The Time of the Day when there is a sudden burst of users in the account

RESEARCH QUESTIONS:

To ensure that the athlete meets our requirements, we have framed the following set of questions. The questions are framed in such a way that it will not only help us in finding the exact match but also will help us in finding the outliers.

- What are the average Likes, Comments, and shares per post for the last 6 months period?
- What is the Number of views to several follower's ratio for all the players?
- What are the Like Ratio, Comment Ratio, and Share Ratio per post?
- What is the Number of Promotional Posts that the players have posted?
- Identifying the type of products promoted and frequency of posts for that product
- What is the total number of posts saved by the users?
- Which type of products reach a wider audience? (Clothing, accessories, food, etc.)
- At what point in time there is an increase in users on the player's Facebook Profile?
- At what day there is an increase in users on the player's Facebook Profile?
- How frequently do the players posts?

Considering all the above questions, we have decided to have the columns presented in the dataset. All the columns such as likes, comments, shares, followers, or engagement focuses on **measuring the reach** of a particular post by a player. Analysis and modelling are done based on the collected data to determine the appropriate brand ambassador for our product – Energy Drink.

METHODOLOGY

DATA CLEANING:

The dataset which we used is **created by our team manually** so, we made sure that there are **no missing values** and while collecting data of 3 players we observed that they have not posted any promotional related posts almost for the past one year. Therefore, we have removed those players as they will not provide any value to the product promotion. We have collected enough information for each player, so we analyze various factors contributing to the reach of a post.

CALCULATED FIELDS:

In the dataset, a few columns were collected directly from the tool. Excel computations are used for the following columns:

- **Likes / Comments / Share ratio** - computed with their respective field and follower's count.
- **promotional post percentage** - promotional posts /total posts
- **Average post frequency** - Average of Posts for the stipulated time.

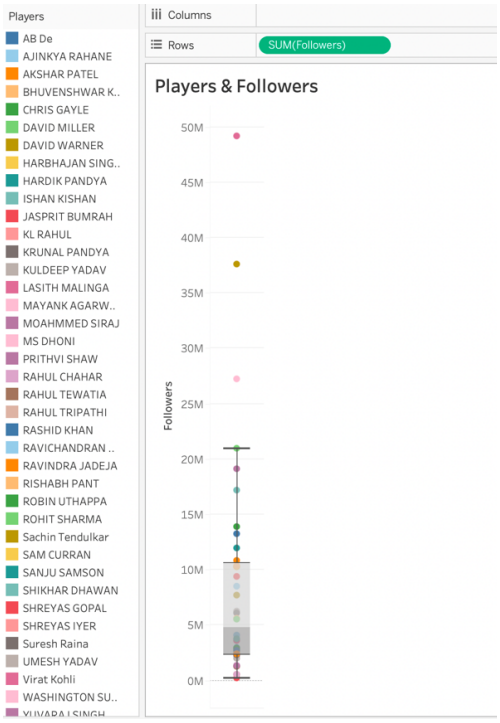
EXPLORATORY ANALYSIS:

We have performed many types of exploratory analysis **for better understanding of the data.** Based on our analysis we were able to get an idea of who might be our **potential brand ambassadors.** The following depicts the different visualization used.

FOLLOWER DISTRIBUTION:

We use box plot to determine the relationship between the players and the follower distribution. The box plot also known as a box-and-whisker plots shows the distribution of

values along an axis. The boxes indicate the middle 50% percent of the data which is the middle quartiles of the data's distribution. It also portrays the distribution of our data, outliers and median. The box plot here is mainly used to determine the players and follower distribution. We distribute various players representing different countries based on the follower count.



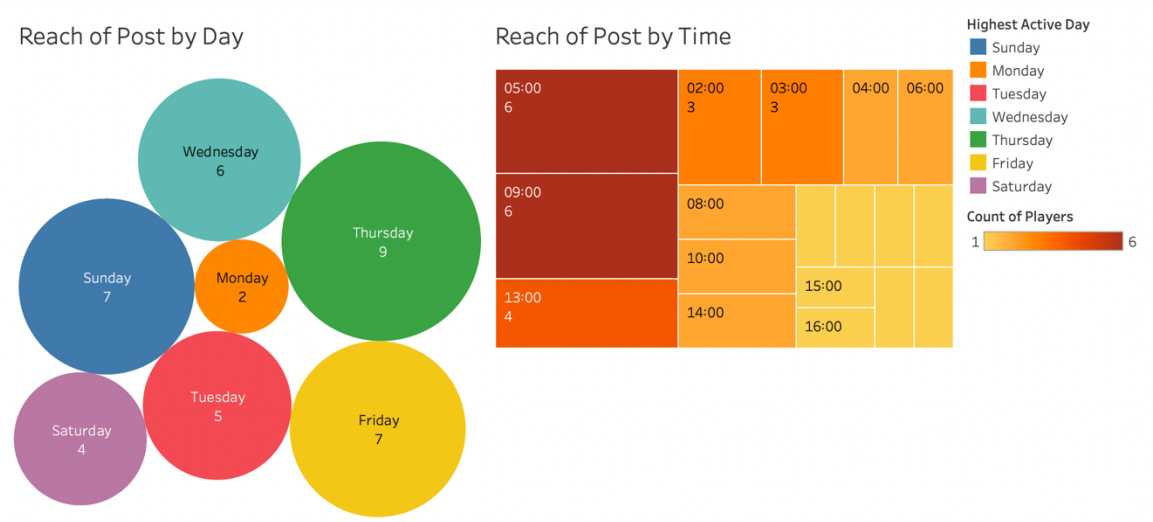
In the image we can find the players with huge follower count viz **Virat Kohli, Sachin Tendulkar and MS Dhoni would be depicted as outliers** because the primary goal of this project is to pick a celebrity who can be cost efficient and someone who can reach a wider spectrum of audiences. It is difficult to pick someone from the mentioned players as they would charge more than our intended budget as they are the superstars of their respective fields. It is evident that we can't chose someone from the outliers.

Hence, we decide to into **account the upper 25% band from the box plot**. The players we concentrate mainly here are Rohit Sharma, AB de Villiers, Jadeja, Chris Gayle, Yuvraj Singh etc. We must consider that there are no negative outliers in the box plot because we take the followers count into consideration.

POST REACH DISTINGUISHED BY DAY & TIME:

The purpose of bubble chart is to display data as a cluster of circles. Each of the value in the dimension field represents a circle whereas the value of measure represents the size of those circles.

Tree map is used to visualize data that can be hierarchical for comparative analysis. Tree map plays an important part to analyze the anomalies in the data set.



We chose the bubble graph and the tree map to show the distribution of engagement of posts on social media and the maximum reach it can have in terms of days and time.

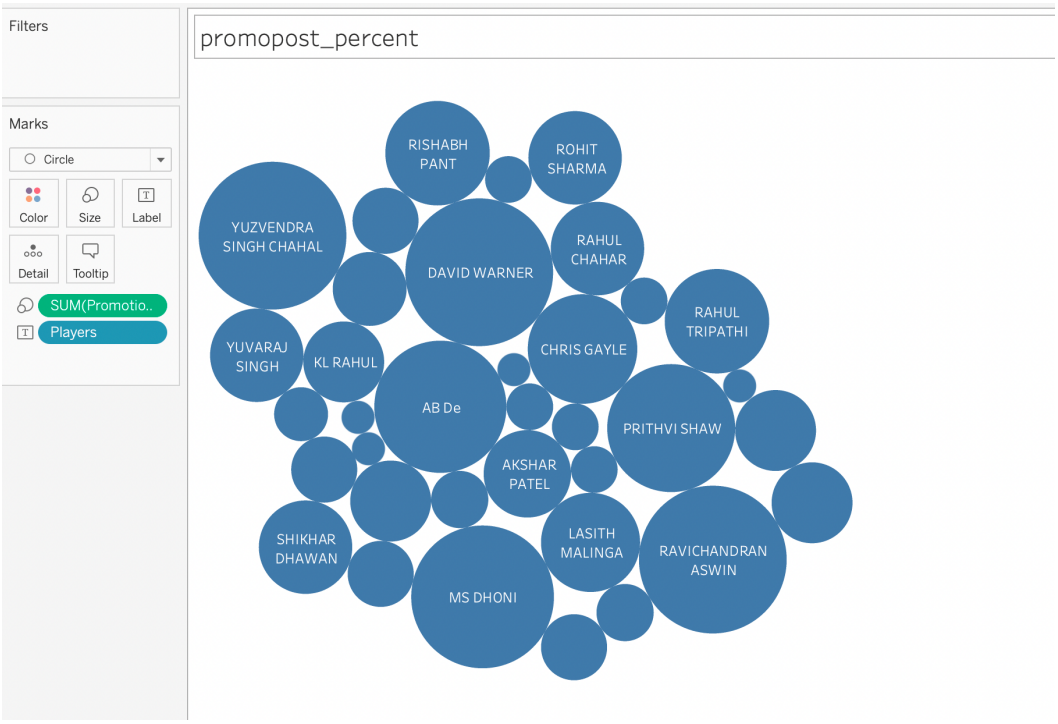
The bubble graph here displays seven days of the week, we use bubble graph here to find out which day of the week has the highest reach of posts on social media.

The graph indicates that **Thursday, Friday, and Sunday** the reach for the posts on social media are high compared to other days of the week. It is understood that many people use social media during weekends more that may be the reason why reach is pretty high.

The tree map implies that the reach ratio is higher between **05:00 AM and 09:00 AM**. It is because people wake up generally at time. As the current generation access social media, the first thing they wake up, that is the reason why the reach of social posts is comparatively high during that time frame.

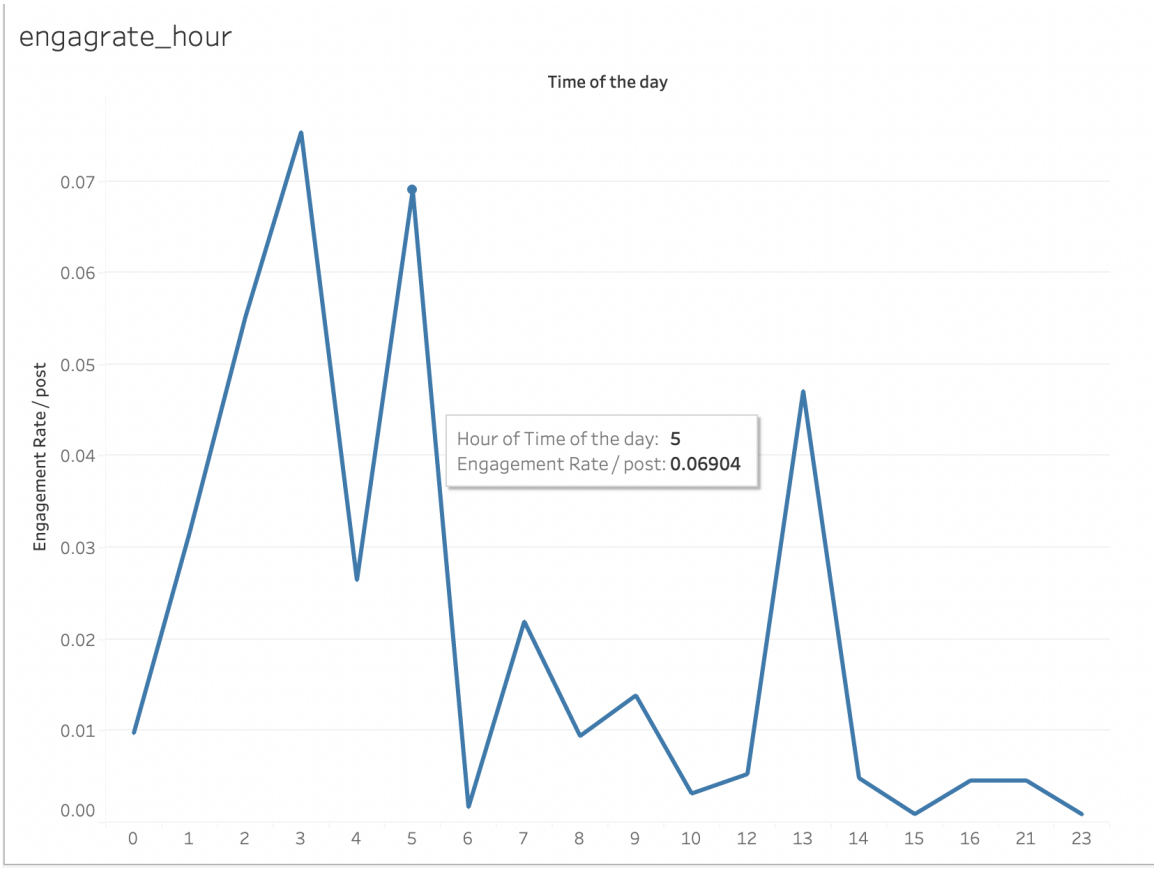
PLAYERS BY PROMOTION:

We use bubble chart again to analyze which players does more in comparison to others. This to get a basic idea of how the **individual players fare in endorsements**. By the bubble chart below we can see that David Warner, Chahal, Dhoni and Ashwin endorse more comparted to their contemporaries.



ENGAGEMENT RATE:

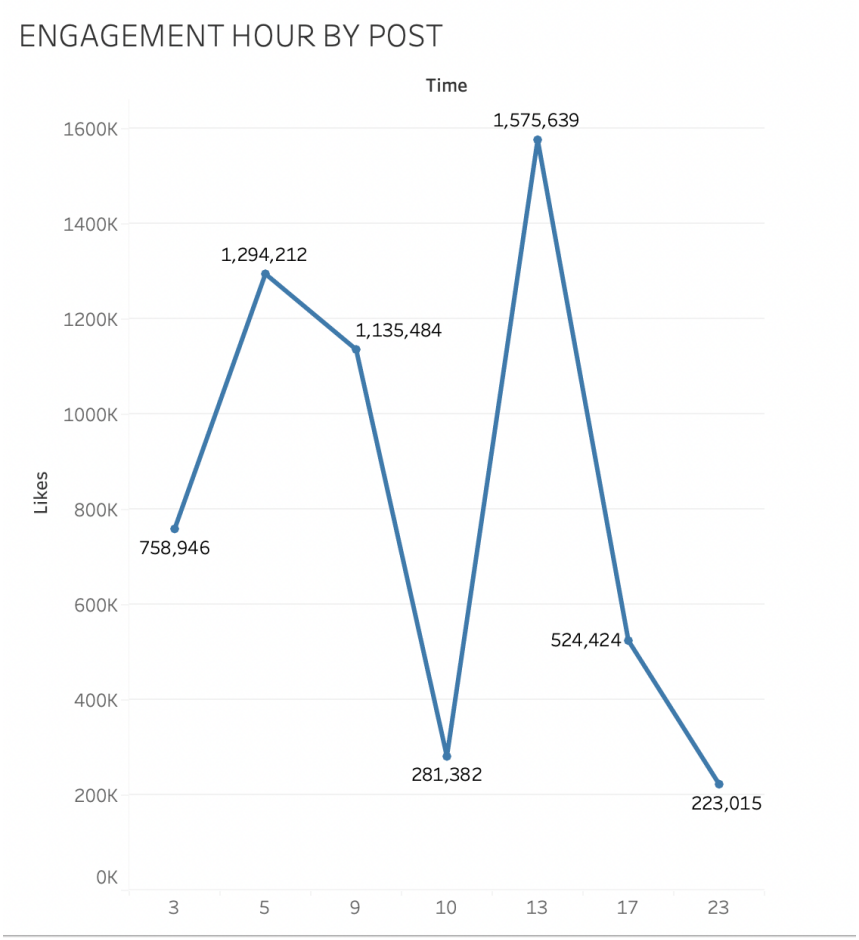
When comparing the engagement rate with hours of the day, we can find that the maximum reach of posts peak at **3:00 AM in the morning**. Intuitively, we can say that 3:00 AM posts are seen by the **Gen Z generation**. The posts at morning 5**:00 AM are accessed by the millennials.** The rest of people access social media between 5:00 AM and 13:00 PM. The thirteenth hour marks the hour after lunch time or the lunch time.



PROOF:

To confirm whether our hunch is right, we took a post of AB de Villiers for a period of 5 days and cumulated the count of likes for the timings - 3:00,5:00, 10:00,13:00,17:00,23:00.

This analysis confirmed that the likes are highest at 13:00 AM and the second highest being 5:00AM as shown in the graph.



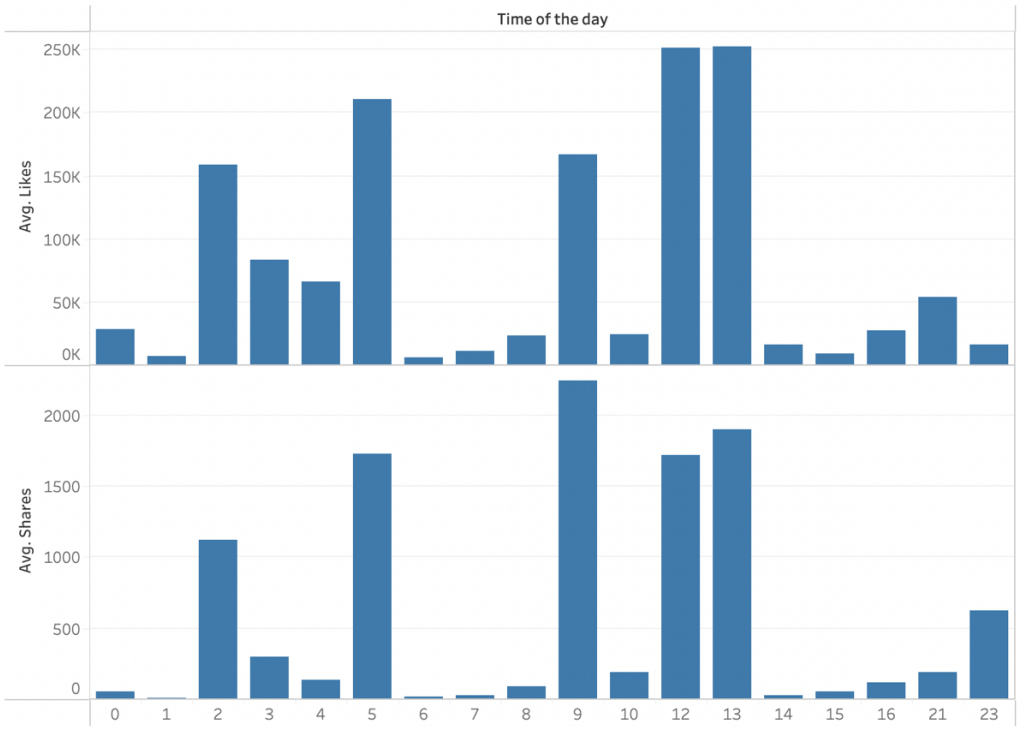
The graph also confirms that the highest engagement rates are 3:00, 5:00, 9:00 and 13:00, the pattern is same as analyzed in our whole dataset.

ANALYSIS BY LIKES AND SHARES:

When endorsing a brand or product, we realized that just the promotional posts alone doesn't suffice to reach a large chunk of people. The reach is best measured by the shares

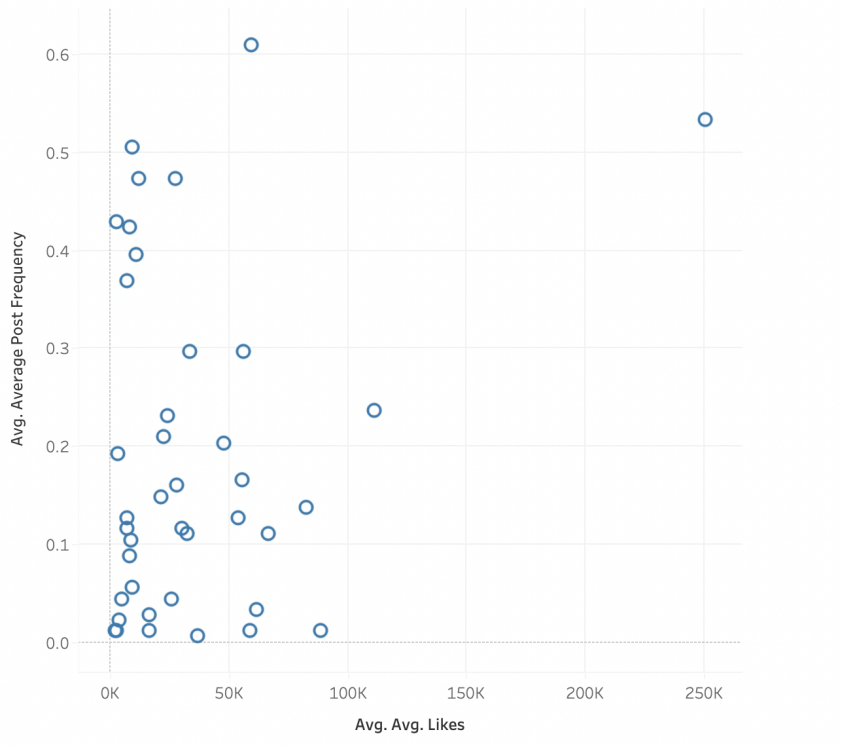
made on social media. Each Post shared would compound the reach made to a large group of people indicating an exponential growth of a product or a brand.

As we have seen in the engagement rate, we noticed that there are a higher number of likes and shares at 3:00 AM, 5:00 AM, 9:00 AM, 12:00 AM and 13:00 PM respectively. **Therefore, as a part of our strategy, we believe that we should promote our product during this time frame.**

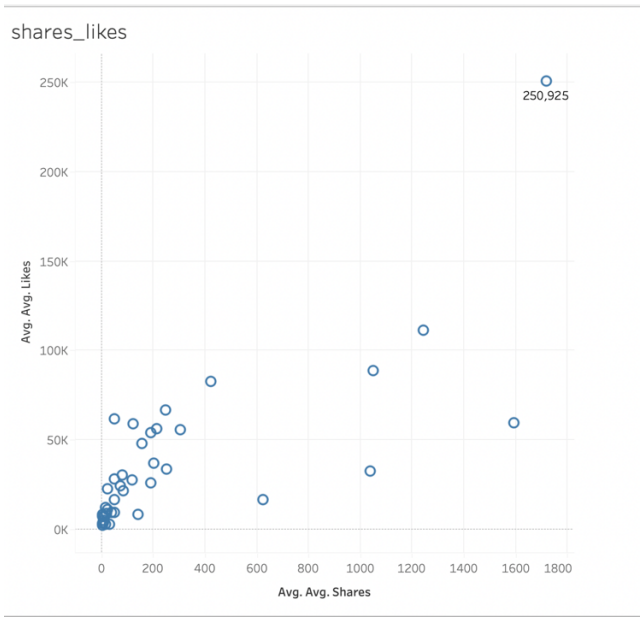


We use scatter plot to visualize the relationship between numerical variables. You create a scatter plot by placing at least one measure on the columns shelf and one measure on the rows shelf. We use the scatter plot her to determine the relationship between the likes to the post frequency.

In the scatter plot below, we can find that **Virat Kohli** dominates the entire players in all the major categories of **likes/post frequency and has the highest number of reaches**. AB De Villiers, from the scatter plot, has a moderate number of posts and has a better entanglement of posts with the reach compared to the other contemporary players.



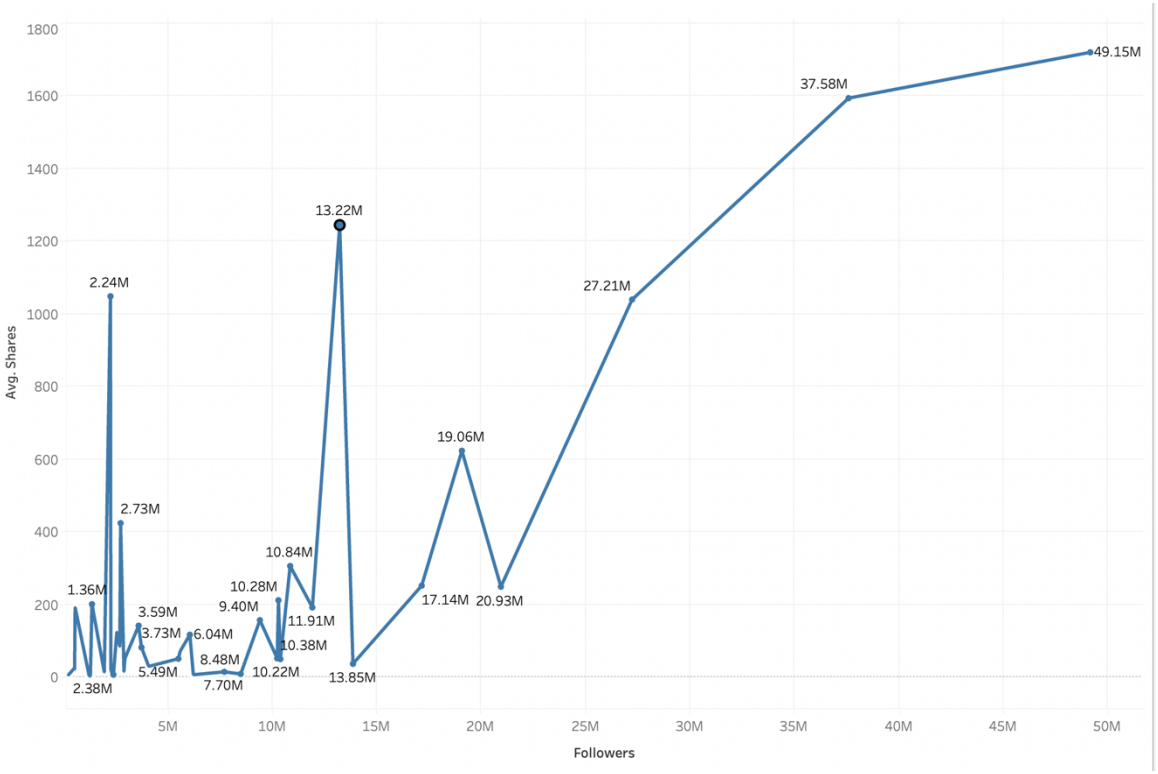
When comparing the reach of endorsement by considering the number of likes and shares with respect to the players. We identify that Virat Kohli and Sachin Tendulkar has the highest ratio. We also realized that apart from AB DE Villiers; MS Dhoni, Prithivi Shaw and Yuvraj Singh have a great reach for their posts. We concluded that the rest of the players kind of have the same effect in terms of reach.



Having a huge number of followers doesn't necessarily infer that it can guarantee reach. **Share is considered as a good form of measurement to understand a post's reach.** Therefore, followers and posts are analyzed together on how far a post about a product can reach to different section of audiences.

We found out that apart from the obvious highest,

- AB DE Villiers(13.2 Million) has a great reach for his posts even though the followers are comparatively less than the top few players.
- Prithivi Shaw (2.24 Million) has a decent number of reach for his promotional posts on social media with a relatively less number of followers.

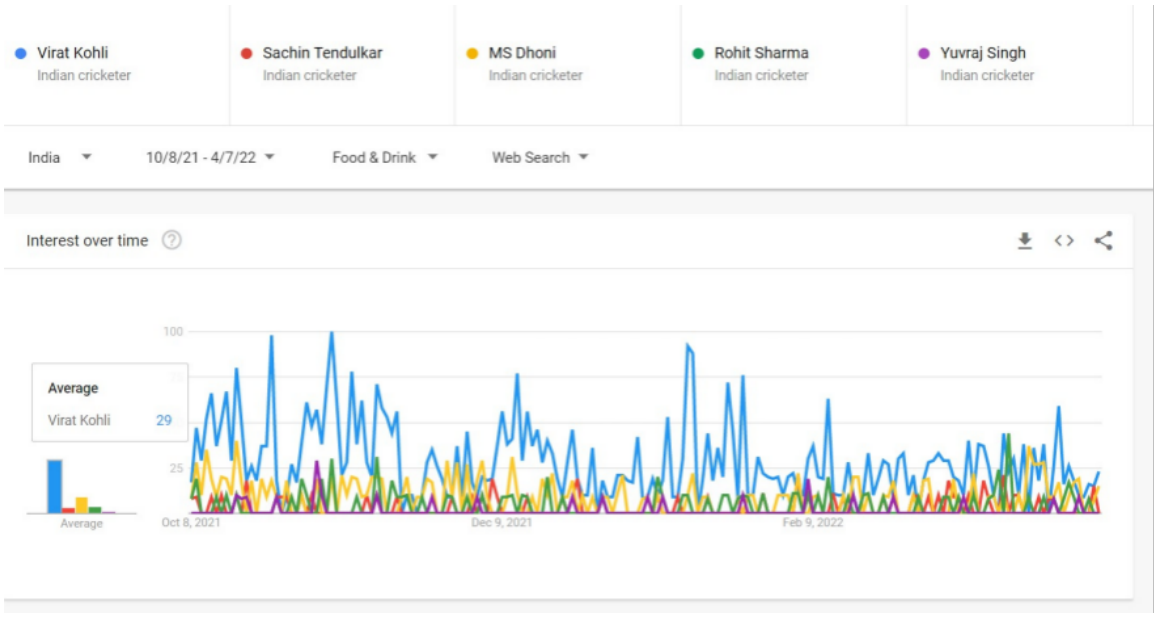


GOOGLE TRENDS:

To extend the exploratory analysis we have used google trends and through this website, we will be able to analyze, the popularity of the top search queries in google search across various regions and languages and it uses graphs to compare the search volume of different queries over time.

Here, we have taken the **top 10 players** based on their follower count, we have searched the trend of their name for the period of 6 months i.e., from 8th October 2021 to 7th April 2022 which is the same data as our dataset, and we can select a specific category of search as our product is an energy drink, **we chose food & drinks category for analysis.**

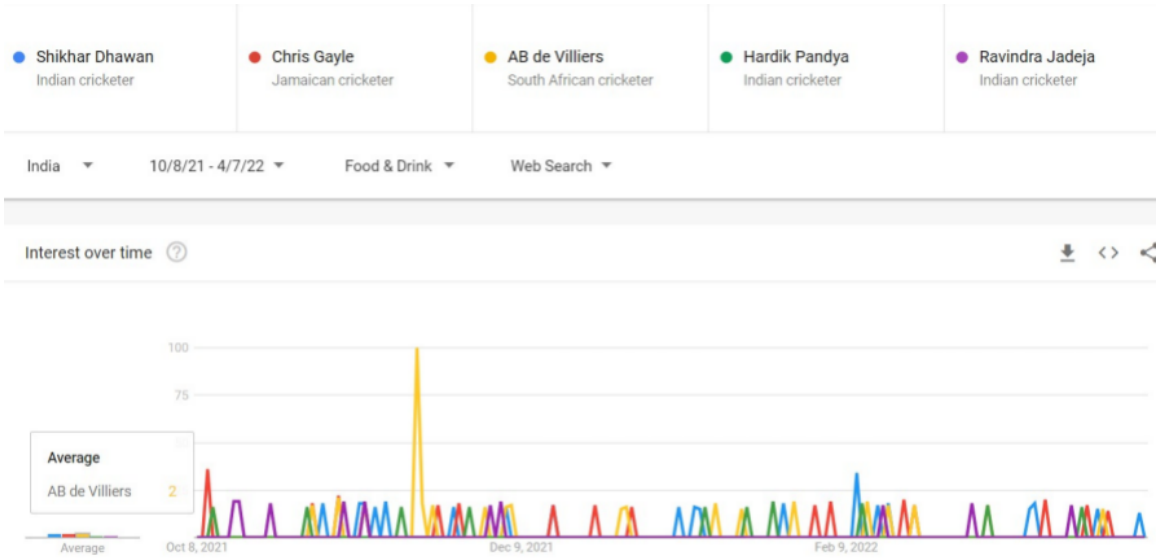
We have started our trend analysis for the first top 5 players they are Virat Kohli, Sachin Tendulkar, MS Dhoni, Rohit Sharma, and Yuvraj Singh & we observed that Virat Kohli is dominating as well among those players.



Screenshot of comparison graph of first top 5 players from google trends:

Here the colour blue indicates Virat Kohli and we can observe that he has high spikes overall when compared with other players and an **average of 29** which is the highest and that we can be able to analyze from the bar graph on left.

Next, we continued trend analysis with the next top 5 players' comparison they are Shikar Dhawan, Chris Gayle, AB de Villiers, Hardik Pandya, and Ravindra Jadeja & we observed that AB de Villiers has an edge over other players.



Screenshot of comparison graph of next top 5 players from google trends:

Here, the yellow colour represents AB de Villiers and we can see that he has little high spikes than other players with an average of 2 which is slightly higher than and that we can figure out from the bar graph on left and we have searched google what caused his high spike and we figured that he announced his retirement during that time.

As now we have compared the top 10 players, we continued with the top 2 players from both comparisons and analyzed them for better decision making and those players are Virat Kohli and AB de Villiers

Screenshot of comparison graph of Virat Kohli and AB de Villiers:

Here, the red represents Virat Kohli and blue AB de Villiers, and we can observe that a single spike of AB de Villiers is matching with Virat Kohli.

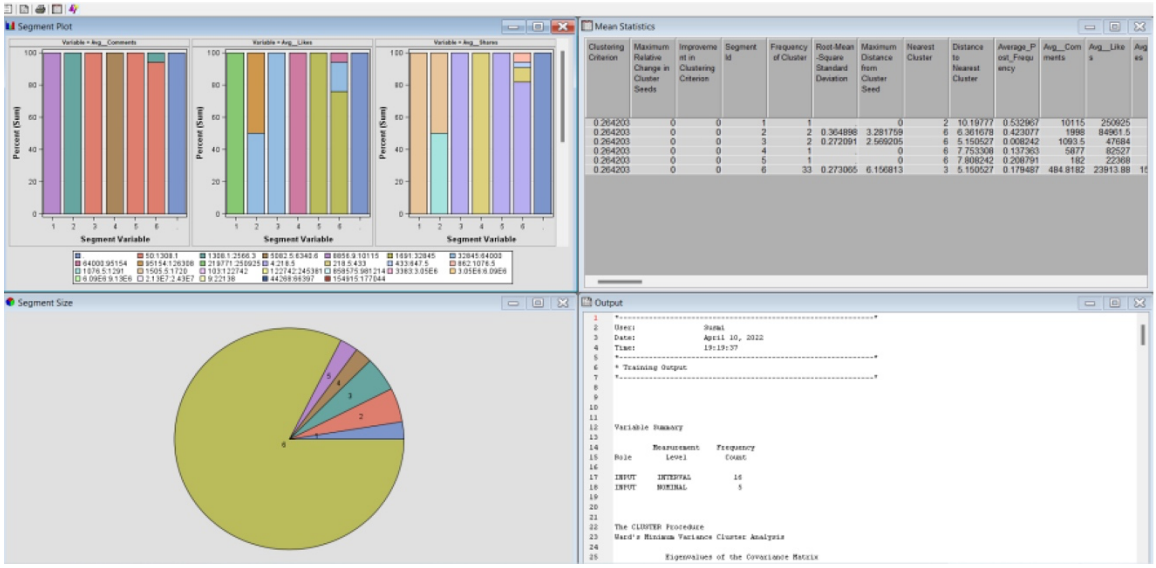
MODEL BUILDING

The major objective of our business problem is to find a cricket player who would be effective in promoting a product to the public. This is done based on the collected dataset of the top cricket players and by considering the various factors affecting the reach of a promotional post made by them.

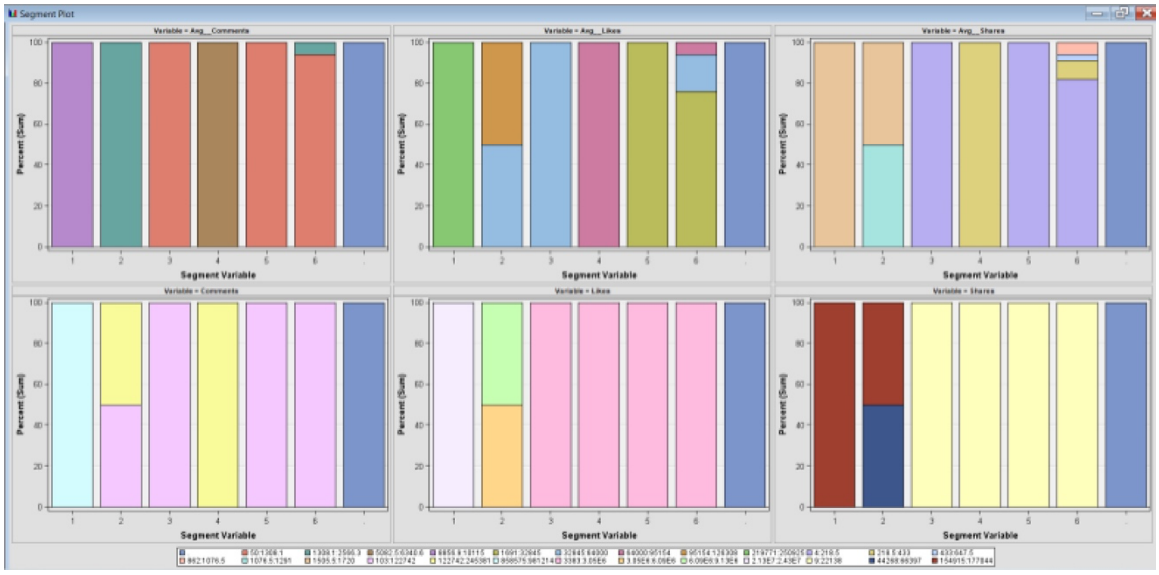
Though **exploratory analysis is sufficient** to find the best player with high promotional posts, **model building is done for our future purposes** to bring out intuitive results. Here, as a part of model building, K-means clustering, and K-Nearest Neighbor are performed for the required analysis using SAS enterprise miner. K-means clustering is performed to group the players with similar characteristics, and K-NN is attempted to check the categorization of the players.

Regression is not performed using the data since our goal is not to predict using the data.

K-MEANS CLUSTERING:



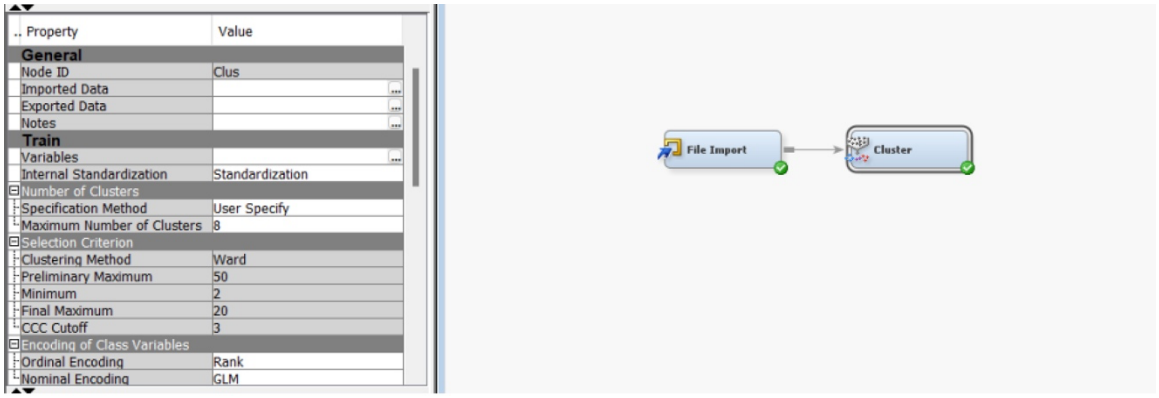
The Excel data file was used to perform clustering, where players are clustered into groups with similar characteristics. The default clustering resulted in six clusters at the beginning, with cluster 2 containing only one player, Virat Kohli, because of his remarkable reach of posts and following on the social media that made him stand out among all the other players and belong in a separate cluster.



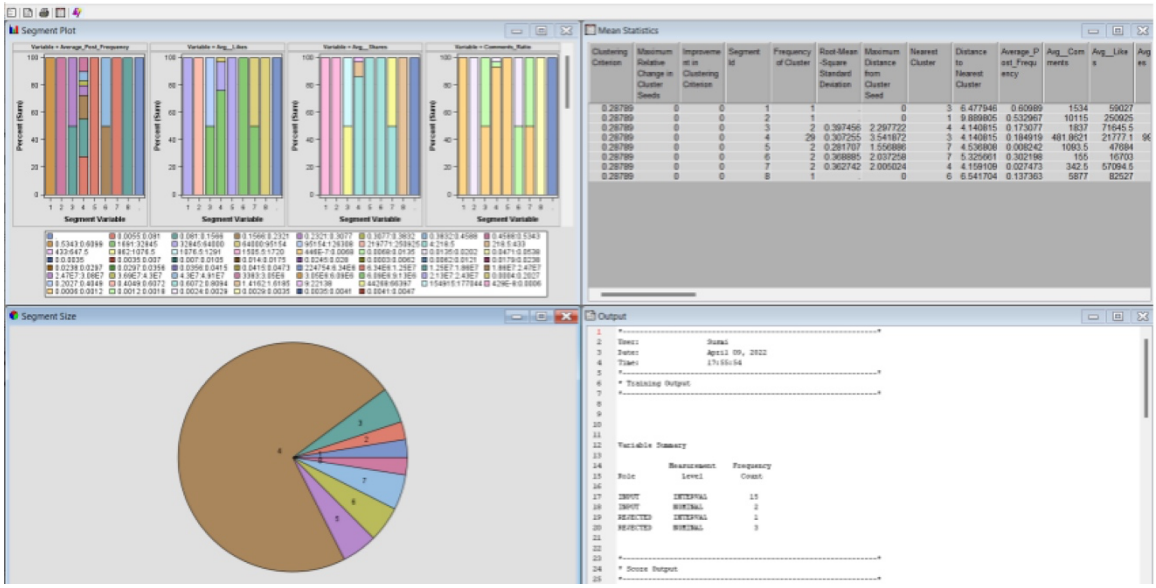
The above figure represents the segmentation of the cluster concerning the various variables. With this as a reference, new data could be categorized into any of these clusters with similar characteristics correspondingly. **The variable Importance shows the variables that are been considered to form the cluster breakdown**.

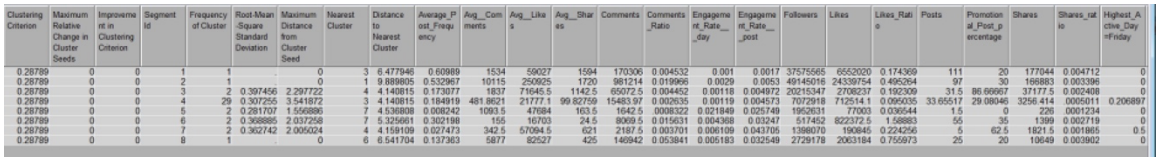
Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Avg__Comments		1	0	1.00000
Comments	Comments	0	1	0.98742
Avg__Likes		0	1	0.97468
Likes	Likes	0	1	0.97468
Avg__Shares		0	1	0.97468
Shares	Shares	0	1	0.97468



In the case of User Specified Clustering, when the number of clusters has been set to **8 Clusters, most of the players converge towards one as like in the default clustering**. While the other players have some metrics that sets them apart from the group. This User specified clustering helped in analyzing the similarities among the players that lie in between the extremes.





Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Followers	Followers	0	2	1.00000
Engagement_Rate___post		1	0	0.77247
Engagement_Rate___day		0	1	0.73863
Likes_Ratio		0	1	0.72699
Comments_Ratio		0	1	0.72699
Shares_ratio		0	1	0.71517
Avg__Shares		1	0	0.69163
Shares	Shares	0	1	0.68293
Likes	Likes	0	1	0.67411
Avg__Likes		0	1	0.67411
Average_Post_Frequency		0	1	0.66518

This data using the user-defined clustering with 8 clusters is like the box plot distribution analyzed during the exploratory study earlier. This result is in accordance with the output of the exploratory analysis of the data. Here, the major clusters considered are:

- Cluster 1 – Virat Kohli
- Cluster 2 – Sachin Tendulkar
- Cluster 3 – AB de Villiers and Prithivi Shaw

Apart from considering only the basic common factors likes, comments, and shares, the other factors like followers, engagement rate, and Post Frequency were also considered, as it was because of these variables, unique cluster characteristics were obtained, and the **clusters are found to be more realistic and intuitive than the default clustering.**

K NEAREST NEIGHBOR:

Nearest neighbor	Nearest neighbor	Nearest neighbor	Nearest neighbor	Players	Followers	Posts	Likes	Shares	Comments	Node_ID	Profile_Link	Avg__Likes	Avg__Shares
22.0	14.0	6.0	1.0	Sachin Tendulkar	37575565	111	6552020	177044	170306	1.00044226919512E14	https://www.facebook.com/100...	59027.0	1594.0
6.0	14.0	1.0	2.0	Virat Kohli	49145016	97	24339754	166893	981214	1.00044473084573E14	https://www.facebook.com/100...	250925.0	1720.0
39.0	12.0	4.0	3.0	KL RAHUL	10284388	54	3027838	11552	71043	1.00044235379214E14	https://www.facebook.com/100...	56071.0	213.0
3.0	39.0	8.0	4.0	RAVINDRA JADEJA	10844998	30	1671536	9211	33529	1.00044514696213E14	https://www.facebook.com/100...	55717.0	307.0
16.0	37.0	19.0	5.0	RISHABH PANT	5582288	42	1003576	3096	18924	1.00044527715691E14	https://www.facebook.com/100...	23894.0	73.0
14.0	28.0	6.0	22.0	ROHIT SHARMA	20931904	20	1332768	5011	40543	1.00044240029O6E14	https://www.facebook.com/100...	66638.0	250.0
10.0	18.0	38.0	7.0	ISHAN KISHAN	3727674	21	632491	1752	15440	1.00044148800557E14	https://www.facebook.com/100...	30118.0	83.0
33.0	4.0	39.0	8.0	HARDIK PANDYA	11907722	23	1241491	4456	23454	1.00043860010222E14	https://www.facebook.com/100...	53977.0	193.0
11.0	34.0	10.0	9.0	RASHID KHAN	2729178	25	2063184	10649	146942	1.00044281426914E14	https://www.facebook.com/100...	82527.0	425.0
13.0	11.0	10.0	34.0	SANJU SAMSON	2935248	29	813791	1565	15553	1.00044281406786E14	https://www.facebook.com/100...	28061.0	53.0
34.0	10.0	13.0	11.0	SHREYAS IYER	2688432	27	572847	2345	14048	1.00044290546403E14	https://www.facebook.com/100...	21216.0	86.0
3.0	4.0	12.0	39.0	JASPRIT BUMRAH	9395089	37	1757146	5870	20066	1.00044441690 4E14	https://www.facebook.com/100...	47490.0	158.0
17.0	27.0	11.0	13.0	MAYANK AGARWAL	2384770	67	469282	600	5037	1.00044443719233E14	https://www.facebook.com/100...	7004.0	8.0
28.0	22.0	6.0	14.0	MS DHONI	27214451	20	647915	20816	24248	1.00044819374936E14	https://www.facebook.com/100...	32395.0	1040.0
4.0	3.0	9.0	15.0	AB De	13216242	43	4768558	53539	105897	1.00044338043815E14	https://www.facebook.com/ABd...	110896.0	1245.0
19.0	37.0	5.0	16.0	Suresh Raina	6043506	86	2347646	10220	32244	1.00044298046085E14	https://www.facebook.com/Sur...	27298.0	118.0
20.0	23.0	17.0	27.0	PRITHVI SHAW	2242883	2	177277	2097	371	1.00043955606994E14	https://www.facebook.com/100...	88638.0	1049.0
10.0	7.0	38.0	18.0	RAVICHANDRAN ASWIN	4074691	78	203638	2441	4131	1.00044160351587E14	https://www.facebook.com/100...	2611.0	31.0
32.0	5.0	19.0	37.0	UMESH YADAV	6211135	8	39665	70	1273	1.00044509476307E14	https://www.facebook.com/100...	4958.0	8.0
23.0	17.0	27.0	20.0	KULDEEP YADAV	1947454	23	153799	345	3736	1.00044197371116E14	https://www.facebook.com/100...	6686.0	15.0
33.0	4.0	21.0	8.0	CHRIS GAYLE	13851349	10	93404	380	3544	1.00044317545336E14	https://www.facebook.com/100...	9340.0	38.0
21.0	28.0	6.0	22.0	YUVARAJ SINGH	19064947	5	81059	3124	1767	1.00044266886682E14	https://www.facebook.com/100...	16211.0	624.0
29.0	27.0	17.0	23.0	AKSHAR PATEL	2360426	35	102415	155	2213	1.00044547304498E14	https://www.facebook.com/100...	2926.0	4.0
26.0	35.0	31.0	24.0	RAHUL TEWATIA	1221964	16	129640	95	1155	1.0004682663197E14	https://www.facebook.com/100...	8102.0	5.0
36.0	26.0	30.0	25.0	WASHINGTON SUNDAR	509723	72	794749	1920	9223	1.00051414284787E14	https://www.facebook.com/100...	11038.0	26.0
30.0	36.0	25.0	26.0	RAHUL TRIPATHI	553257	8	204413	1546	4004	1.00057597272933E14	https://www.facebook.com/100...	25551.0	193.0
29.0	23.0	17.0	27.0	KRUNAL PANDYA	2262069	19	161828	428	2115	1.00044260937684E14	https://www.facebook.com/100...	8517.0	22.0
6.0	22.0	21.0	28.0	SHIKHAR DHAWAN	17140774	54	1802822	13689	35332	1.00044278427035E14	https://www.facebook.com/100...	33385.0	253.0
17.0	27.0	29.0	23.0	SAM CURRAN	2547182	2	117275	250	2195	1.00057200719768E14	https://www.facebook.com/100...	58637.0	125.0
36.0	26.0	30.0	25.0	RAHUL CHAHAR	525181	38	849996	878	6916	1.00018555806529E14	https://www.facebook.com/100...	22368.0	23.0
20.0	24.0	31.0	35.0	YUZVENDRA SINGH C...	1281491	2	3383	9	103	1.00044284813471E14	https://www.facebook.com/100...	1691.0	4.0
37.0	40.0	19.0	32.0	DAVID WARNER	7695287	2	4681	33	501	1.00044542354906E14	https://www.facebook.com/100...	2340.0	16.0
40.0	4.0	39.0	33.0	BHUVENSHWAR KUMAR	10376635	2	32981	103	2536	1.00044317335294E14	https://www.facebook.com/100...	16490.0	51.0
13.0	11.0	10.0	34.0	ROBIN UTHAPPA	2889016	86	997335	1476	5711	1.00044339533916E14	https://www.facebook.com/100...	11596.0	17.0
20.0	24.0	31.0	35.0	MOAHMMED SIRAJ	1358079	1	36731	202	1090	1.00044794445847E14	https://www.facebook.com/100...	36731.0	202.0
30.0	26.0	25.0	36.0	SHREYAS GOPAL	224754	21	148088	157	1069	1.00044477988136E14	https://www.facebook.com/100...	7051.0	7.0
18.0	5.0	19.0	37.0	DAVID MILLER	5493502	6	369803	317	2389	1.00044343763446E14	https://www.facebook.com/100...	61633.0	52.0
10.0	18.0	38.0	7.0	LASITH MALINGA	3592181	77	618885	11017	55279	1.00043925989321E14	https://www.facebook.com/100...	8037.0	143.0
12.0	4.0	39.0	33.0	HARBHAJAN SINGH	10224052	92	835480	4876	35922	1.00044143374O6E14	https://www.facebook.com/100...	9081.0	53.0
33.0	32.0	12.0	40.0	AJINKYA RAHANE	8476878	4	13591	43	378	1.00044289256775E14	https://www.facebook.com/100...	3397.0	10.0
39.0	32.0	33.0	40.0		.	.	.	.	.	.		.	.

The K Nearest Neighbor is used to determine the players with similar attributes in the dataset. This algorithm groups the data points based on the closeness to the other data points. 4 nearest neighbors are used to see which players go along together in terms of the variables with respect to the promotions. **The value of K is chosen as 4 since it has the lowest misclassification error rate.** In 4NN, as one of the neighbors is itself on calculating

the shortest distance, the other three points are considered the actual neighbors. From the results, it could be inferred that Sachin, Dhoni, Yuvaraj, and Rohit are neighbors with similar traits. **When new data comes in, we can match them with their neighbors and find out their characteristics easily.**

INTERPRETATION

From both exploratory analysis and modelling, it could be assured that Virat Kohli is the best player for any kind of Facebook Promotion. But Virat Kohli, Sachin Tendulkar, and MS Dhoni are the top cricketers and are considered highly influential among the public on the social media platforms with high number of followers and their charges for a promotional post range between **1 – 5 crore rupees.**

On the other hand, it was also found that **AB De Villiers and Prithvi Shaw** have great reach for their posts, be it promotional or normal, with a good post engagement rate. For a promotional post, **AB De Villiers charges nominally around 30 - 50 Lakhs with 13.2 million engagements,** whereas Prithivi Shaw's post reach is good, but he can reach only 2.24M.

Therefore, Prithvi Shaw would be a good fit for smaller promotions and AB De Villiers would be considered a good cricket player for making effective promotions. These players do not have as many followers as Virat Kohli or MS Dhoni, but their posts have met the desirable reach with the public and attained a good engagement rate.

RESULTS & CONCLUSION:

With all the analysis and clustering techniques, we can conclude that choosing **AB de Villiers as the brand ambassador would be ideal, as it would be economical cutting costs for the company and promotes the product to a larger audience.** In addition, the post must be published in the analyzed peak timings and days drawn from the exploratory analysis to meet the expected reach among the public.

This would help the brand spend a decent amount of charge to the cricket player for the promotional pots and get the desired reach and exposure to the product as well as the brand among the public.