BRIGHT TV VIEWERSHIP

METHODOLOGY

Contents

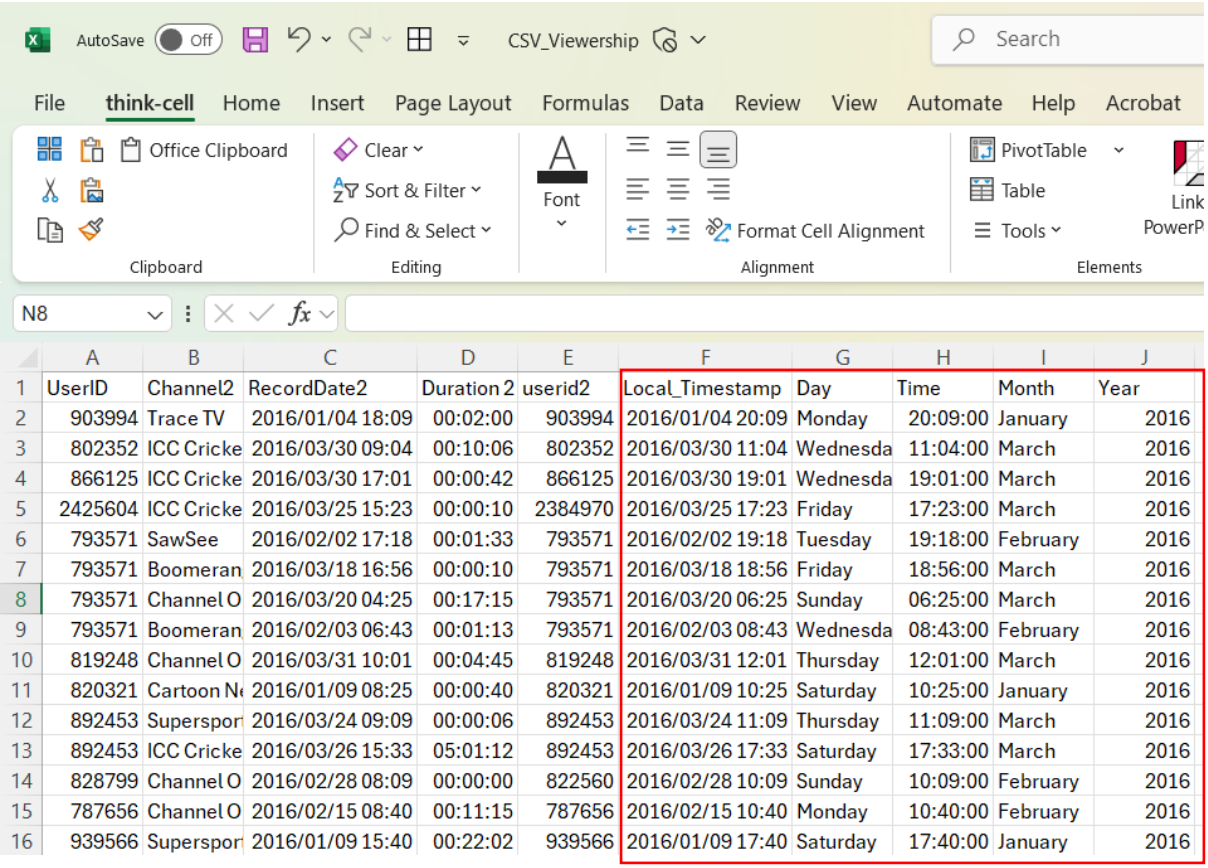
1. Background and Introduction

BrightTV ‘s CEO has an objective to grow the company’s subscription base for this financial year. He has approached you to provide insights that would assist CVM (Customer Value Management) team in meeting this year’s objective.

The dataset provided (User_Profiles and Viewership) contained information on the user profiles and viewer transactions for BrightTV respectively. The two files were loaded on Snowflakes (SQL) for analysis.

2. Date Manipulation

From the viewership file, time stamp for each record was provided. Times and dates in the dataset were supplied in UTC and have been converted to SA time. The Day, Time, Month, and the Year were extracted from the Timestamp using excel before loading the file onto Snowflake. New columns were added as shown below:



	A	B	C	D	E	F	G	H	I	J
1	UserID	Channel2	RecordDate2	Duration 2	userid2	Local_Timestamp	Day	Time	Month	Year
2	903994	Trace TV	2016/01/04 18:09	00:02:00	903994	2016/01/04 20:09	Monday	20:09:00	January	2016
3	802352	ICC Cricke	2016/03/30 09:04	00:10:06	802352	2016/03/30 11:04	Wednesda	11:04:00	March	2016
4	866125	ICC Cricke	2016/03/30 17:01	00:00:42	866125	2016/03/30 19:01	Wednesda	19:01:00	March	2016
5	2425604	ICC Cricke	2016/03/25 15:23	00:00:10	2384970	2016/03/25 17:23	Friday	17:23:00	March	2016
6	793571	SawSee	2016/02/02 17:18	00:01:33	793571	2016/02/02 19:18	Tuesday	19:18:00	February	2016
7	793571	Boomeran	2016/03/18 16:56	00:00:10	793571	2016/03/18 18:56	Friday	18:56:00	March	2016
8	793571	Channel O	2016/03/20 04:25	00:17:15	793571	2016/03/20 06:25	Sunday	06:25:00	March	2016
9	793571	Boomeran	2016/02/03 06:43	00:01:13	793571	2016/02/03 08:43	Wednesda	08:43:00	February	2016
10	819248	Channel O	2016/03/31 10:01	00:04:45	819248	2016/03/31 12:01	Thursday	12:01:00	March	2016
11	820321	Cartoon N	2016/01/09 08:25	00:00:40	820321	2016/01/09 10:25	Saturday	10:25:00	January	2016
12	892453	Superspor	2016/03/24 09:09	00:00:06	892453	2016/03/24 11:09	Thursday	11:09:00	March	2016
13	892453	ICC Cricke	2016/03/26 15:33	05:01:12	892453	2016/03/26 17:33	Saturday	17:33:00	March	2016
14	828799	Channel O	2016/02/28 08:09	00:00:00	822560	2016/02/28 10:09	Sunday	10:09:00	February	2016
15	787656	Channel O	2016/02/15 08:40	00:11:15	787656	2016/02/15 10:40	Monday	10:40:00	February	2016
16	939566	Superspor	2016/01/09 15:40	00:22:02	939566	2016/01/09 17:40	Saturday	17:40:00	January	2016

The highlighted columns have been added and they are in SA standard time.

3. Completeness of Data

The number of records in each file was extracted. Data cleaning was performed in case of duplicates, empty rows or missing files.

3.1 Checking the number of records

The number of records from user_profiles was **5375** which was the total number of unique users. The total number of records from Viewership was **10000.** Each time a user logs in, there is a separate record. This was computed using the following query.

```sql
-- Checking number of records

SELECT COUNT(*)
FROM user_profiles;

SELECT COUNT(DISTINCT userid)
FROM user_profiles;

SELECT COUNT(*)
FROM viewership;

SELECT COUNT(userid)
FROM viewership;
```

3.2 Checking the Duplicates

The following query has been done to check rows that were repeating, i.e. having the same records across all columns.

```sql
-- Checking for completely duplicate rows

SELECT *,
    COUNT(*)
FROM user_profiles
GROUP BY ALL
HAVING COUNT(*) > 1;

SELECT *,
    COUNT(*)
FROM viewership
GROUP BY ALL
HAVING COUNT(*) > 1; --(5 recodrs have duplicates)
```

The user_profiles file had no duplicates, while the Viewership file had **5** rows that duplicated. A new temporary table called Viewership_new that has removed the duplicate rows. The new file has **9995** unique rows.

```sql
-- Creating a temporary table with no duplicates as viewership_new

SELECT DISTINCT *
FROM viewership;

CREATE OR REPLACE TEMPORARY TABLE viewership_new AS (
    SELECT DISTINCT *
    FROM viewership
);
```

3.3 Checking & Replacing Missing Values

Missing values were checked using the following query:

```sql
-- Checking for missing values in the tables

SELECT * FROM user_profiles
WHERE userid IS NULL OR NAME IS NULL OR surname IS NULL OR email IS NULL OR gender IS NULL OR RACE IS NULL OR AGE IS NULL OR
PROVINCE IS NULL OR SOCIAL_MEDIA_HANDLE IS NULL;

SELECT * FROM viewership_new
WHERE userid IS NULL OR channel2 IS NULL OR recorddate2 IS NULL OR duration_2 IS NULL OR userid2 IS NULL OR local_timestamp IS NULL
OR month IS NULL OR year IS NULL;
```

The Viewership_new file did not contain any null values. User_profiles file had a couple of missing records. These records were replaced with "**None**" values as displayed. A new temporary table was created to account for this change. This table also bucketed age into different age groups.

```sql
-- Replacing missing records with 'None' and creating a temp table

CREATE OR REPLACE TEMP TABLE user_profiles_new AS (
    SELECT
        userid,
        age,
        IFNULL(name, 'None') AS Name,
        IFNULL(surname, 'None') AS Surname,
        IFNULL(email, 'None') AS email,
        IFNULL(gender, 'None') AS Gender,
        IFNULL(race, 'None') AS Race,
        IFNULL(province, 'None') AS Province,
        IFNULL(social_media_handle, 'None') AS social_media_handle,
        CASE
            WHEN age BETWEEN 1 AND 12 THEN 'Younger than 13'
            WHEN age BETWEEN 13 AND 25 THEN '13 to 25'
            WHEN age BETWEEN 26 AND 44 THEN '26 to 44'
            WHEN age >= 45 THEN '45 and older'
            ELSE 'Not Specified'
        END AS Age_group
    FROM user_profiles
);
```

3.4 Joining the two working tables

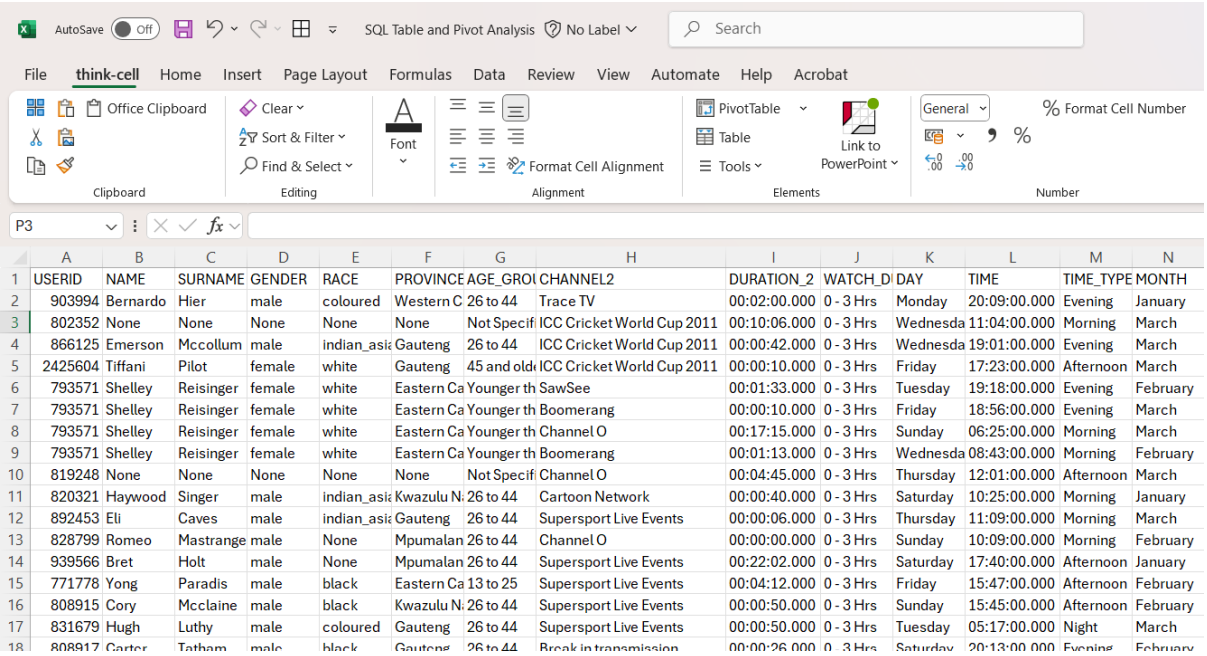
Now that the data completeness have been accomplished, the two tables were joined together using **Inner Join.** This was done to display the users that have watched the channel, together with the programs they have watched. The two tables, user_profiles_new and Viewership_new were joined on **UserID** as the common column. The watching duration and the time of the day were bucked into new columns when the tables were joined. The following query shows how the tables were joined.

```sql
SELECT
    u.userid,
    u.Name,
    u.Surname,
    u.Gender,
    u.Race,
    u.Province,
    u.Age_group,
    v.channel2,
    v.duration_2,
    CASE
        WHEN v.duration_2 between '00:00:00' AND '02:59:59' THEN '0 - 3 Hrs'
        WHEN v.duration_2 between '03:00:00' AND '05:59:59' THEN '3 - 6 Hrs'
        WHEN v.duration_2 between '06:00:00' AND '08:59:59' THEN '6 - 9 Hrs'
        ELSE '9 - 12 Hrs'
    END AS Watch_Duration,
    v.day,
    v.time,
    CASE
        WHEN v.time between '06:00:00' AND '11:59:59' THEN 'Morning'
        WHEN v.time between '12:00:00' AND '17:59:59' THEN 'Afternoon'
        WHEN v.time between '18:00:00' AND '23:59:59' THEN 'Evening'
        ELSE 'Night'
    END AS Time_Type,
    v.month
FROM user_profiles_new AS u
INNER JOIN viewership_new AS v ON u.userid = v.userid;
```

This table has all the column that were used in analysis. The table was exported as CSV file for further analysis.

4. Analysis

The final table was exported and it looked as follows



USERID	NAME	SURNAME	GENDER	RACE	PROVINCE	AGE_GROU	CHANNEL2	DURATION_2	WATCH_DI	DAY	TIME	TIME_TYPE	MONTH
903994	Bernardo	Hier	male	coloured	Western C	26 to 44	Trace TV	00:02:00.000	0 - 3 Hrs	Monday	20:09:00.000	Evening	January
802352	None	None	None	None	None	Not Specif	ICC Cricket World Cup 2011	00:10:06.000	0 - 3 Hrs	Wednesda	11:04:00.000	Morning	March
866125	Emerson	Mccollum	male	indian_asia	Gauteng	26 to 44	ICC Cricket World Cup 2011	00:00:42.000	0 - 3 Hrs	Wednesda	19:01:00.000	Evening	March
2425604	Tiffani	Pilot	female	white	Gauteng	45 and old	ICC Cricket World Cup 2011	00:00:10.000	0 - 3 Hrs	Friday	17:23:00.000	Afternoon	March
793571	Shelley	Reisinger	female	white	Eastern Ca	Younger th	SawSee	00:01:33.000	0 - 3 Hrs	Tuesday	19:18:00.000	Evening	February
793571	Shelley	Reisinger	female	white	Eastern Ca	Younger th	Boomerang	00:00:10.000	0 - 3 Hrs	Friday	18:56:00.000	Evening	March
793571	Shelley	Reisinger	female	white	Eastern Ca	Younger th	Channel O	00:17:15.000	0 - 3 Hrs	Sunday	06:25:00.000	Morning	March
793571	Shelley	Reisinger	female	white	Eastern Ca	Younger th	Boomerang	00:01:13.000	0 - 3 Hrs	Wednesda	08:43:00.000	Morning	February
819248	None	None	None	None	None	Not Specif	Channel O	00:04:45.000	0 - 3 Hrs	Thursday	12:01:00.000	Afternoon	March
820321	Haywood	Singer	male	indian_asia	Kwazulu N	26 to 44	Cartoon Network	00:00:40.000	0 - 3 Hrs	Saturday	10:25:00.000	Morning	January
892453	Eli	Caves	male	indian_asia	Gauteng	26 to 44	Supersport Live Events	00:00:06.000	0 - 3 Hrs	Thursday	11:09:00.000	Morning	March
828799	Romeo	Mastrange	male	None	Mpumalan	26 to 44	Channel O	00:00:00.000	0 - 3 Hrs	Sunday	10:09:00.000	Morning	February
939566	Bret	Holt	male	None	Mpumalan	26 to 44	Supersport Live Events	00:22:02.000	0 - 3 Hrs	Saturday	17:40:00.000	Afternoon	January
771778	Yong	Paradis	male	black	Eastern Ca	13 to 25	Supersport Live Events	00:04:12.000	0 - 3 Hrs	Friday	15:47:00.000	Afternoon	February
808915	Cory	Mcclaine	male	black	Kwazulu N	26 to 44	Supersport Live Events	00:00:50.000	0 - 3 Hrs	Sunday	15:45:00.000	Afternoon	February
831679	Hugh	Luthy	male	coloured	Gauteng	26 to 44	Supersport Live Events	00:00:50.000	0 - 3 Hrs	Tuesday	05:17:00.000	Night	March
808917	Carter	Tatham	male	black	Gautens	26 to 44	Break in transmission	00:00:26.000	0 - 3 Hrs	Saturday	20:13:00.000	Evening	February

From the table above, the following analysis have been performed in pivot tables in which visuals were developed:

- Demographic view
    - Viewership per Gender

    - Viewership per Race
    - Viewership per Province
    - Viewership per Age Group
- Trend analysis
    - Viewership per Month
    - Viewership per Weekday
- Channel analysis
    - Viewership per Channel
    - Viewership per Watch Duration

5. Sample of Pivot Table

	A	B	C
1	**Demographic Analysis**		
2			
3			
4	**Viewership per Gender**		
5			
6	**Row Labels**	**Count of USERID**	**Count of USERID2**
7	female	976	9,8%
8	male	8757	87,6%
9	None	262	2,6%
10	(blank)		0,0%
11	**Grand Total**	**9995**	**100,0%**
12			
13			
14	**Viewership per Race**		
15			
16	**Row Labels**	**Count of USERID**	**Count of USERID2**
17	black	4329	43,3%
18	coloured	1631	16,3%
19	indian_asian	1575	15,8%
20	white	1291	12,9%
21	None	1067	10,7%
22	other	102	1,0%
23	(blank)		0,0%
24	**Grand Total**	**9995**	**100,0%**

From these tables, visuals were developed.