BrightTV Viewership Methodology

Contents

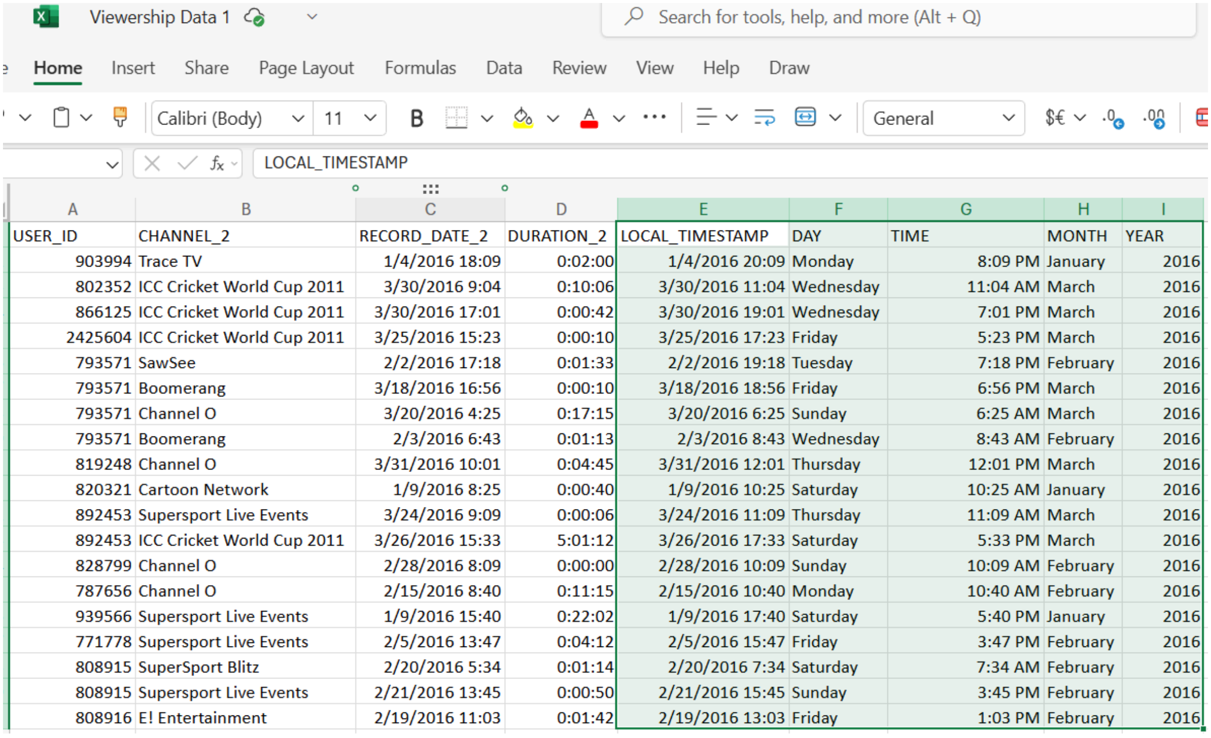
1.Background and Introduction

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) contains information on the user profiles and viewer transactions for BrightTV. The two files were loaded onto Databricks (SQL) for further analysis.

2. Date Manipulation

The viewership file contains a time stamp for each record. 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 Databricks. New columns were added as shown below:



USER_ID	CHANNEL_2	RECORD_DATE_2	DURATION_2	LOCAL_TIMESTAMP	DAY	TIME	MONTH	YEAR
903994	Trace TV	1/4/2016 18:09	0:02:00	1/4/2016 20:09	Monday	8:09 PM	January	2016
802352	ICC Cricket World Cup 2011	3/30/2016 9:04	0:10:06	3/30/2016 11:04	Wednesday	11:04 AM	March	2016
866125	ICC Cricket World Cup 2011	3/30/2016 17:01	0:00:42	3/30/2016 19:01	Wednesday	7:01 PM	March	2016
2425604	ICC Cricket World Cup 2011	3/25/2016 15:23	0:00:10	3/25/2016 17:23	Friday	5:23 PM	March	2016
793571	SawSee	2/2/2016 17:18	0:01:33	2/2/2016 19:18	Tuesday	7:18 PM	February	2016
793571	Boomerang	3/18/2016 16:56	0:00:10	3/18/2016 18:56	Friday	6:56 PM	March	2016
793571	Channel O	3/20/2016 4:25	0:17:15	3/20/2016 6:25	Sunday	6:25 AM	March	2016
793571	Boomerang	2/3/2016 6:43	0:01:13	2/3/2016 8:43	Wednesday	8:43 AM	February	2016
819248	Channel O	3/31/2016 10:01	0:04:45	3/31/2016 12:01	Thursday	12:01 PM	March	2016
820321	Cartoon Network	1/9/2016 8:25	0:00:40	1/9/2016 10:25	Saturday	10:25 AM	January	2016
892453	Supersport Live Events	3/24/2016 9:09	0:00:06	3/24/2016 11:09	Thursday	11:09 AM	March	2016
892453	ICC Cricket World Cup 2011	3/26/2016 15:33	5:01:12	3/26/2016 17:33	Saturday	5:33 PM	March	2016
828799	Channel O	2/28/2016 8:09	0:00:00	2/28/2016 10:09	Sunday	10:09 AM	February	2016
787656	Channel O	2/15/2016 8:40	0:11:15	2/15/2016 10:40	Monday	10:40 AM	February	2016
939566	Supersport Live Events	1/9/2016 15:40	0:22:02	1/9/2016 17:40	Saturday	5:40 PM	January	2016
771778	Supersport Live Events	2/5/2016 13:47	0:04:12	2/5/2016 15:47	Friday	3:47 PM	February	2016
808915	SuperSport Blitz	2/20/2016 5:34	0:01:14	2/20/2016 7:34	Saturday	7:34 AM	February	2016
808915	Supersport Live Events	2/21/2016 13:45	0:00:50	2/21/2016 15:45	Sunday	3:45 PM	February	2016
808916	E! Entertainment	2/19/2016 11:03	0:01:42	2/19/2016 13:03	Friday	1:03 PM	February	2016

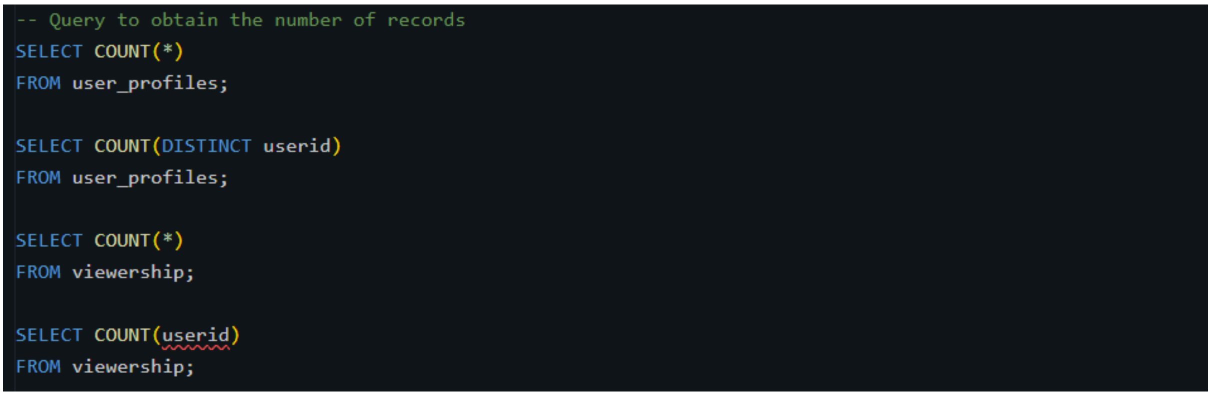
Highlighted columns feature day, time,month and year converted to SA standard time.

3. Completeness of Data

Following the ingestion of the transformed Viewership table as well as the User_Profiles, using SQL queries the number of records in each file was extracted. Data cleaning was performed in case of duplicates, empty rows or missing files.

3.1. Check the number of records

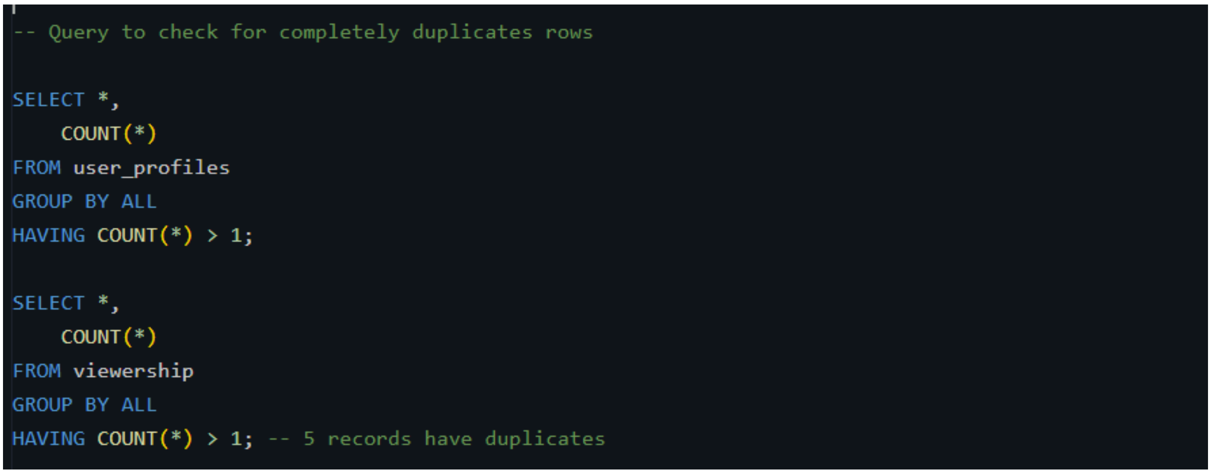
The number of total records from the Viewership table is **10000**, including duplicates. The number of total records from User_Profiles is **5375.** The SQL queries were ran on Databricks to extract this data:



```sql
-- Query to obtain the number of records
SELECT COUNT(*)
FROM user_profiles;

SELECT COUNT(DISTINCT userid)
FROM user_profiles;

SELECT COUNT(*)
FROM viewership;

SELECT COUNT(userid)
FROM viewership;
```

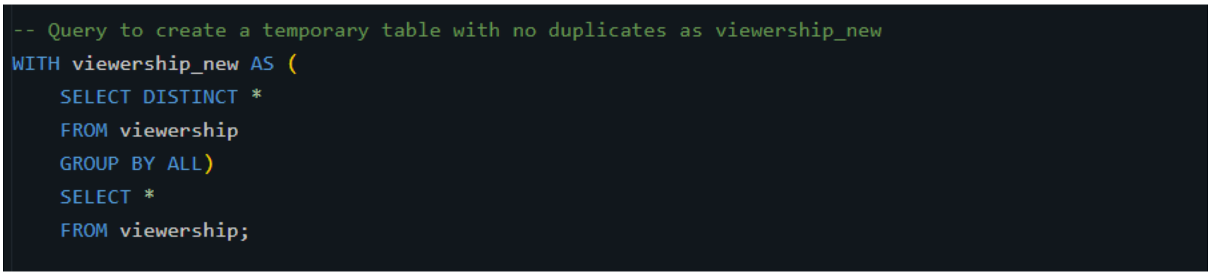
3.2. Checking for Duplicates

The following query ran on Databricks to check rows which contained duplicated data.



```sql
-- Query to check for completely duplicates rows

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

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

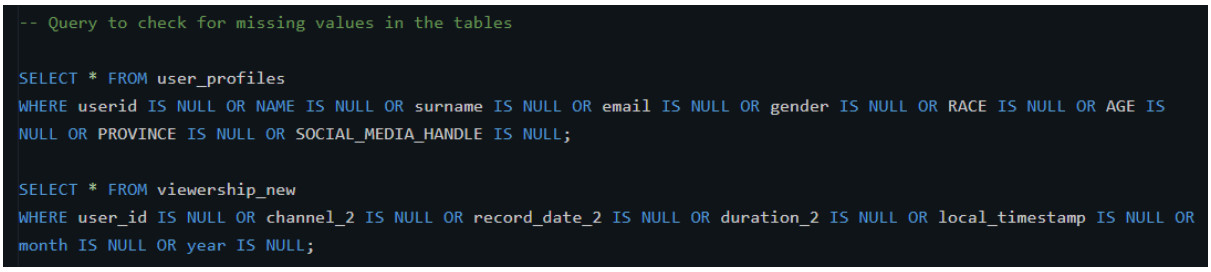
The Use_Profile table does not contain duplicates. However, Viewership contained **5 duplicated rows**. A new temporary table, **Viewership_new** has been created using the query below to retrieve a new table without duplication. The new file has **9995 unique rows**.



```sql
-- Query to create a temporary table with no duplicates as viewership_new
WITH viewership_new AS (
    SELECT DISTINCT *
    FROM viewership
    GROUP BY ALL)
    SELECT *
    FROM viewership;
```

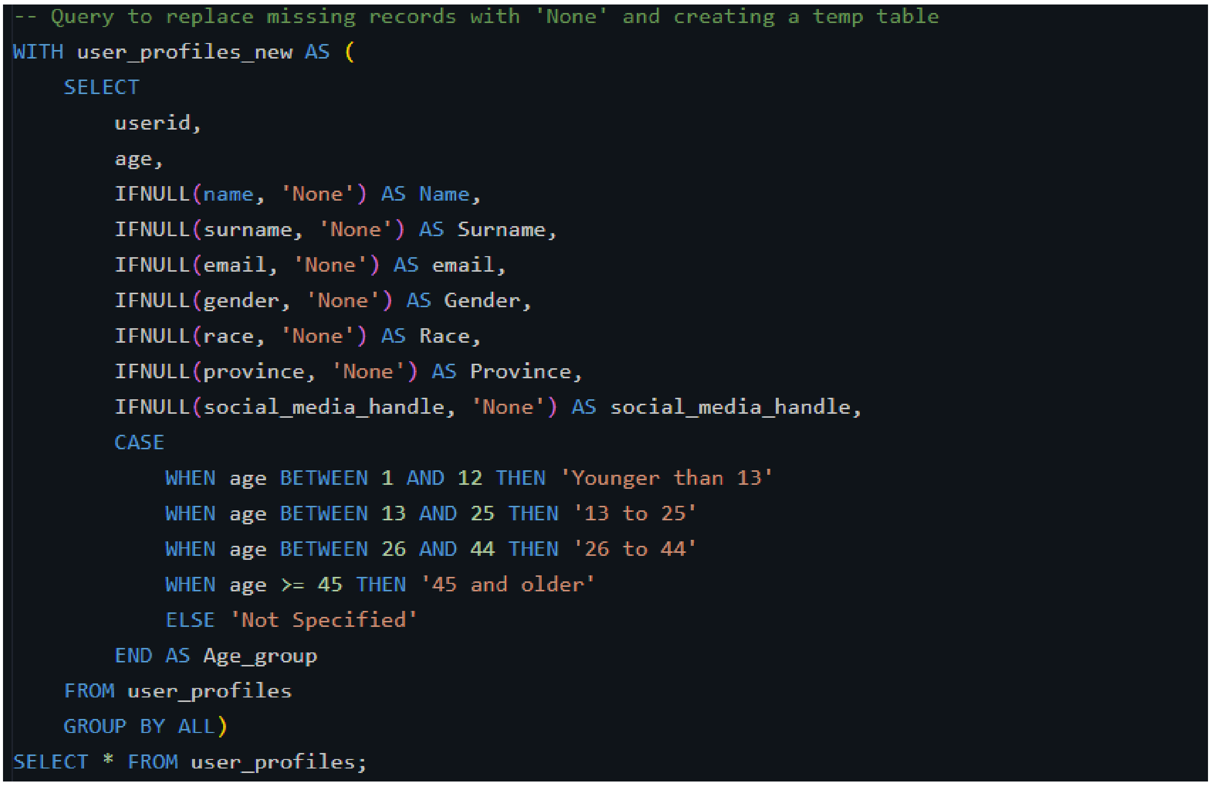
3.3 Checking and Replacing Missing Values

The query below was used to check for missing values:



```sql
-- Query to check 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 user_id IS NULL OR channel_2 IS NULL OR record_date_2 IS NULL OR duration_2 IS NULL OR local_timestamp IS NULL OR
month IS NULL OR year IS NULL;
```

The viewership file did not contain any missing values. However, User_Profiles contained missing values. These records were replaced with “None” using the query shown below. A new table was created to account for this transformation. Using the CASE statement, this table was bucketed into distinct age groups.

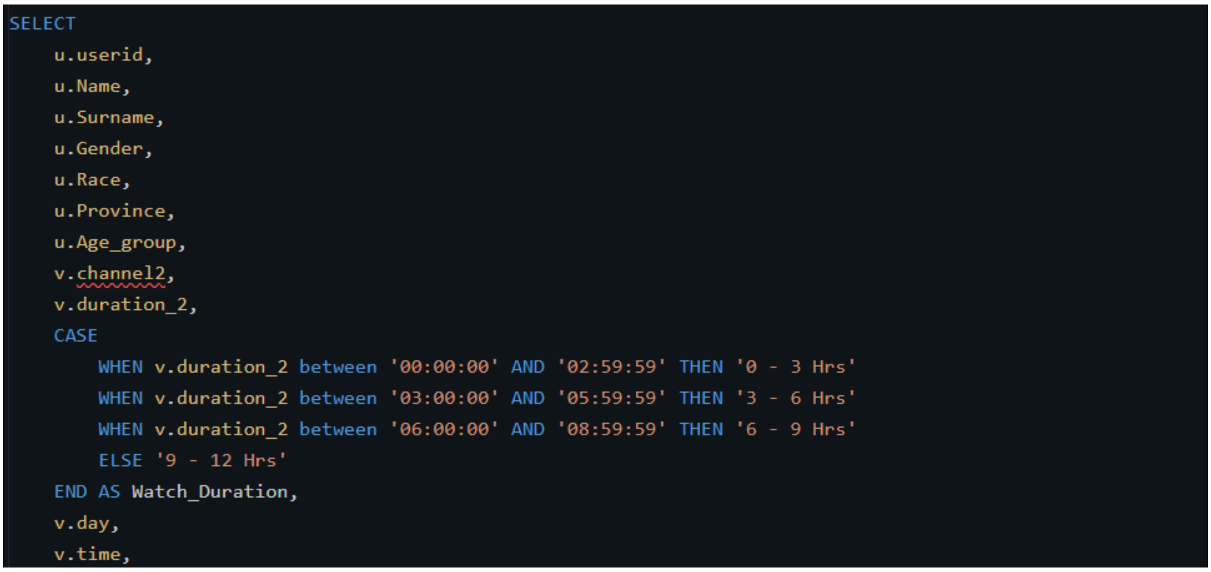


```sql
-- Query to replace missing records with 'None' and creating a temp table
WITH user_profiles_new AS (
    SELECT
        userid,
        age,
        IFNULL(name, 'None') AS Name,
        IFNULL(surname, 'None') AS Surname,
        IFNULL(email, 'None') AS email,
        IFNULL(gender, 'None') 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
    GROUP BY ALL)
SELECT * FROM user_profiles;
```

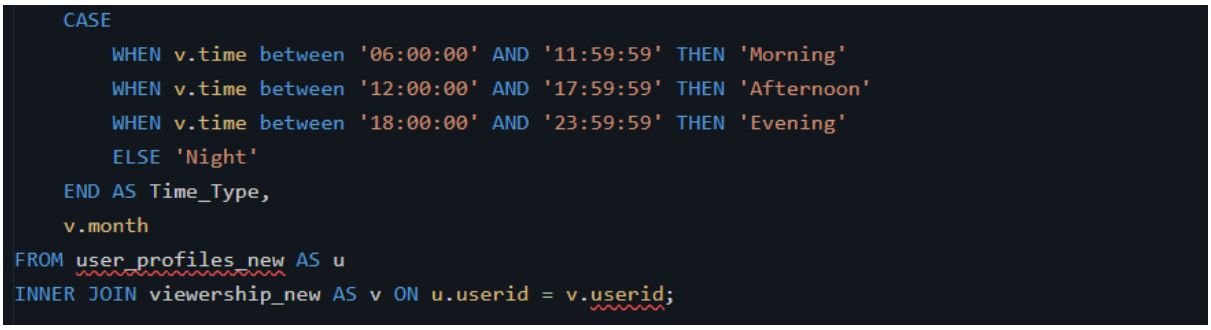
3.4. Joining The Two Working Tables

Next, we join the two tables since data completeness is completed, the tables are joined using **INNER JOIN**. This was done to create a new comprehensive file that displays the users that watched the channels as well as the programs watched. The tables User_Profiles and Viewership_New were joined using **UserID as a common column**.

Viewership Duration and Time of Day were bucketed into new columns upon the joining of the two tables, as indicated by the query below.



```
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,
```

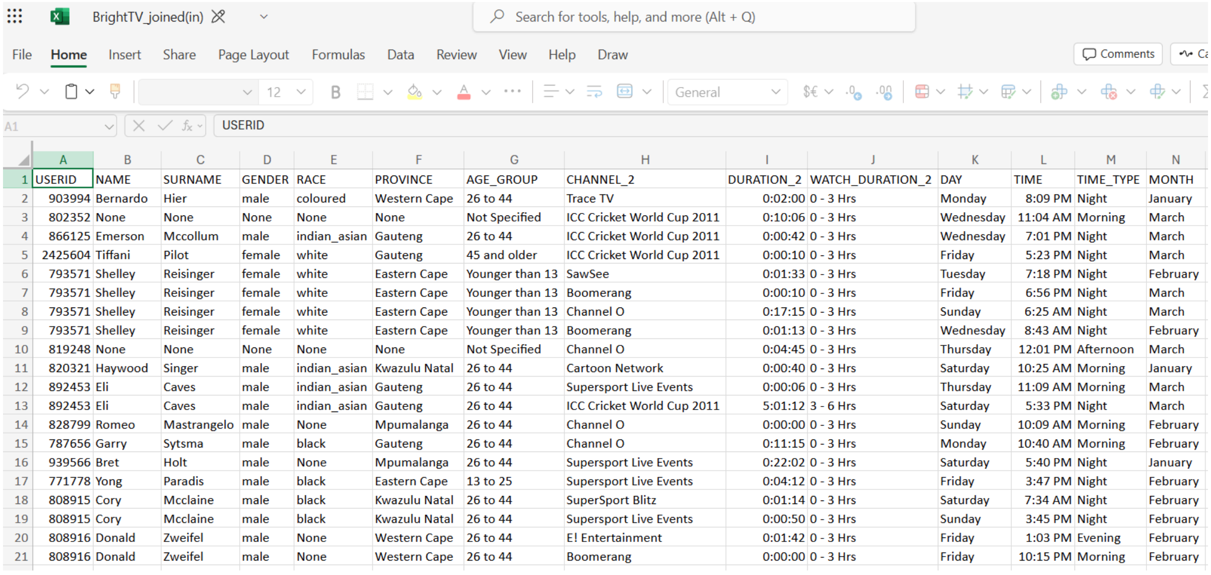


```
    CASE
        WHEN v.time between '06:00:00' AND '11:59:59' THEN 'Morning'
        WHEN v.time between '12:00:00' AND '17:59:59' THEN 'Afternoon'
        WHEN v.time between '18:00:00' AND '23:59:59' THEN 'Evening'
        ELSE 'Night'
    END AS Time_Type,
    v.month
FROM user_profiles_new AS u
INNER JOIN viewership_new AS v ON u.userid = v.userid;
```

The file has been converted to CSV format for further analysis, contains all columns needed for analysis.

4. Analysis

Display of the final exported table:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	USERID	NAME	SURNAME	GENDER	RACE	PROVINCE	AGE_GROUP	CHANNEL_2	DURATION_2	WATCH_DURATION_2	DAY	TIME	TIME_TYPE	MONTH
2	903994	Bernardo	Hier	male	coloured	Western Cape	26 to 44	Trace TV	0:02:00	0 - 3 Hrs	Monday	8:09 PM	Night	January
3	802352	None	None	None	None	None	Not Specified	ICC Cricket World Cup 2011	0:10:06	0 - 3 Hrs	Wednesday	11:04 AM	Morning	March
4	866125	Emerson	Mccollum	male	indian_asian	Gauteng	26 to 44	ICC Cricket World Cup 2011	0:00:42	0 - 3 Hrs	Wednesday	7:01 PM	Night	March
5	2425604	Tiffani	Pilot	female	white	Gauteng	45 and older	ICC Cricket World Cup 2011	0:00:10	0 - 3 Hrs	Friday	5:23 PM	Night	March
6	793571	Shelley	Reisinger	female	white	Eastern Cape	Younger than 13	SawSee	0:01:33	0 - 3 Hrs	Tuesday	7:18 PM	Night	February
7	793571	Shelley	Reisinger	female	white	Eastern Cape	Younger than 13	Boomerang	0:00:10	0 - 3 Hrs	Friday	6:56 PM	Night	March
8	793571	Shelley	Reisinger	female	white	Eastern Cape	Younger than 13	Channel O	0:17:15	0 - 3 Hrs	Sunday	6:25 AM	Night	March
9	793571	Shelley	Reisinger	female	white	Eastern Cape	Younger than 13	Boomerang	0:01:13	0 - 3 Hrs	Wednesday	8:43 AM	Night	February
10	819248	None	None	None	None	None	Not Specified	Channel O	0:04:45	0 - 3 Hrs	Thursday	12:01 PM	Afternoon	March
11	820321	Haywood	Singer	male	indian_asian	Kwazulu Natal	26 to 44	Cartoon Network	0:00:40	0 - 3 Hrs	Saturday	10:25 AM	Morning	January
12	892453	Eli	Caves	male	indian_asian	Gauteng	26 to 44	Supersport Live Events	0:00:06	0 - 3 Hrs	Thursday	11:09 AM	Morning	March
13	892453	Eli	Caves	male	indian_asian	Gauteng	26 to 44	ICC Cricket World Cup 2011	5:01:12	3 - 6 Hrs	Saturday	5:33 PM	Night	March
14	828799	Romeo	Mastrangelo	male	None	Mpumalanga	26 to 44	Channel O	0:00:00	0 - 3 Hrs	Sunday	10:09 AM	Morning	February
15	787656	Garry	Sytsma	male	black	Gauteng	26 to 44	Channel O	0:11:15	0 - 3 Hrs	Monday	10:40 AM	Morning	February
16	939566	Bret	Holt	male	None	Mpumalanga	26 to 44	Supersport Live Events	0:22:02	0 - 3 Hrs	Saturday	5:40 PM	Night	January
17	771778	Yong	Paradis	male	black	Eastern Cape	13 to 25	Supersport Live Events	0:04:12	0 - 3 Hrs	Friday	3:47 PM	Night	February
18	808915	Cory	Mcclaine	male	black	Kwazulu Natal	26 to 44	SuperSport Blitz	0:01:14	0 - 3 Hrs	Saturday	7:34 AM	Night	February
19	808915	Cory	Mcclaine	male	black	Kwazulu Natal	26 to 44	Supersport Live Events	0:00:50	0 - 3 Hrs	Sunday	3:45 PM	Night	February
20	808916	Donald	Zweifel	male	None	Western Cape	26 to 44	E! Entertainment	0:01:42	0 - 3 Hrs	Friday	1:03 PM	Evening	February
21	808916	Donald	Zweifel	male	None	Western Cape	26 to 44	Boomerang	0:00:00	0 - 3 Hrs	Friday	10:15 PM	Morning	February

From the above table, pivot tables and visuals were developed to perform the following analysis:

Demographic Analysis

- ★ Viewership by Race
- ★ Viewership by Gender
- ★ Viewership by Age
- ★ Viewership by Province

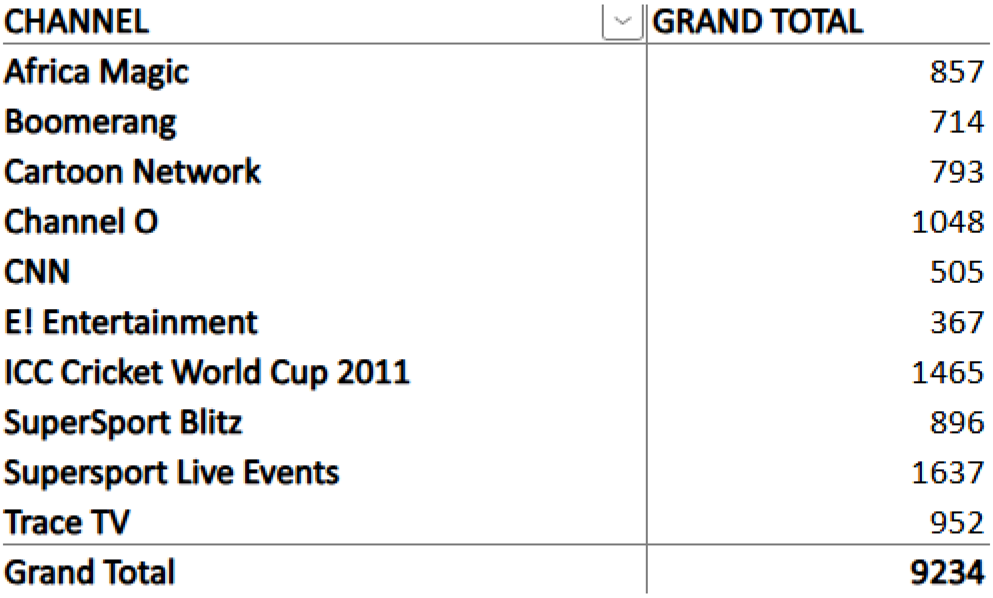
Trend Analysis

- ★ Total Viewers by Weekday
- ★ Total Views Over Time

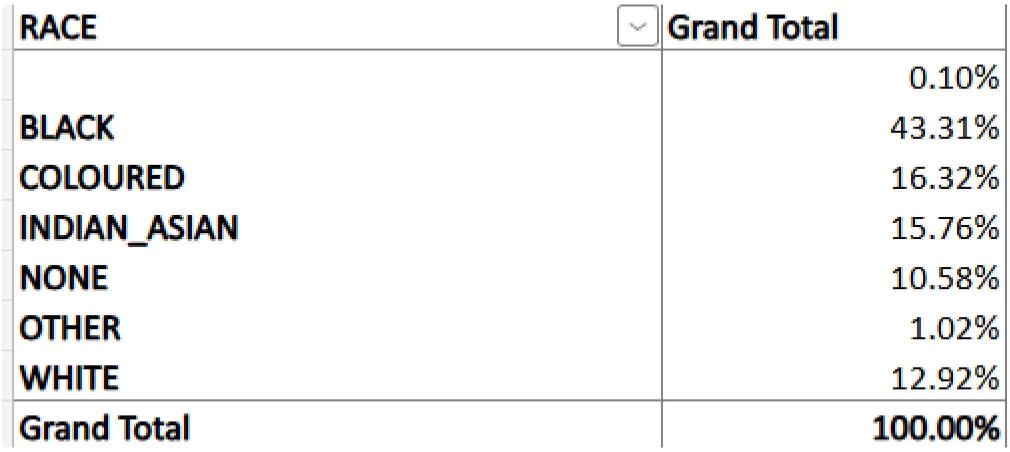
Channel Analysis

- ★ Top 10 most watched channels
- ★ Viewership by Day and Duration

5. Pivot Table Samples



CHANNEL	GRAND TOTAL
Africa Magic	857
Boomerang	714
Cartoon Network	793
Channel O	1048
CNN	505
E! Entertainment	367
ICC Cricket World Cup 2011	1465
SuperSport Blitz	896
Supersport Live Events	1637
Trace TV	952
Grand Total	**9234**



RACE	Grand Total
	0.10%
BLACK	43.31%
COLOURED	16.32%
INDIAN_ASIAN	15.76%
NONE	10.58%
OTHER	1.02%
WHITE	12.92%
Grand Total	**100.00%**