Project 1 – Part 1 – dry answers

In this code, we added several new columns to help identify malicious TV program titles. First, we created date-related columns:

- **air_date_parsed** converts the air date into a real date format
- **day** gives the day of the month
- **weekday** shows the day of the week (like Mon or Fri)

These columns are used to check if a program aired on a specific day, like Friday the 13th.

```
daily_prog_df = daily_prog_df.withColumn("air_date_parsed", to_date(col("air_date"), "yyyyMMdd")) \
    .withColumn("day", dayofmonth(col("air_date_parsed"))) \
    .withColumn("weekday", date_format(col("air_date_parsed"), "E"))
```

Then, we added 7 condition flags (cond_1 to cond_7), each representing one of the 7 conditions that were given:

- **cond_1**: True if a program's duration is greater than the average duration.
- **cond_2**: True if the program was watched by households where the vehicle_make is '91' (Toyota). We created a column by joining the viewing data with demographic info and filtering for Toyota owners, then marking the relevant programs.
- **cond_3**: True if the program was watched by households with exactly 2 adults whose age difference is 6 years or less. This was done by filtering the demographic data for households with 2 adults and calculating the age difference using helper column age_diff.
- **cond_4**: True if the program aired on a Friday the 13th, using the day and weekday columns.
- **cond_5**: True if the program was watched by a household that has more than 3 devices and an income lower than the average. We computed average income, counted devices per household, joined both, and filtered for households meeting both conditions.
- **cond_6**: True if the program's genre contains keywords like "Music", "Animated", etc.
- **cond_7**: True if the program's title contains at least two words from a given list. We split the title into words, created helper column words_array that contains to flag each word, and then summed the matches to a new helper column title_flag_count to check if the count is ≥ 2.

All of these columns helped us quickly check each program against the required conditions, saving time and making the data easier to work with in later steps.