Missing Values & Duplicated Value

```
Daily Time Spent on Site     13
Age                           0
Area Income                  13
Daily Internet Usage         11
Male                          3
Timestamp                     0
Clicked on Ad                 0
city                          0
province                      0
category                      0
```

- The above shows 4 columns with missing values which are **'Daily Time Spent on Site'**, **'Area Income'**, **'Daily Internet Usage'** and **'Male'** along with the total number of missing values for each column
- **Median** of each numerical column is used to fill in missing values
- Meanwhile, **mode** of categorical column (‘Male’) is used to fill in missing values
- There is **no duplicate value** in this data set

For further details, please refer to this Jupyter Notebook

Extract datetime data

Four new columns are extracted from column ‘Timestamp’

- Column 'Date_Day' shows the day represented by 0 as Monday, 1 as Tuesday, 2 as Wednesday, 3 as Thursday, 4 as Friday, 5 as Saturday, and 6 as Sunday
- Column 'Date_Month' shows the month based on the given timestamp represented by numerical value
- Column 'Date_Year’ shows the year in numerical value from the given timestamp
- Column 'Date_Week' shows the week in a year from the given timestamp
- Column ‘Timestamp’ then dropped s it has represented by the extracted data

For further details, please refer to this Jupyter Notebook

Feature Encoding

- **Ordinal Encoding** is done for column **‘Clicked_on_Ad’** and **‘Male’**
- **One Hot Encoding** is done for column **‘city’**, **‘province’** and **‘category’**
- These 5 columns are then **dropped** as it has represented by the encoded columns

For further details, please refer to this Jupyter Notebook

Feature Selection

Feature selection is done to drop irrelevant and redundant columns

- Drop features that has **high correlation r>0.70**

- Encoded **province columns that has high correlation with encoded city columns are dropped** and the city columns are retained as city has more detailed information

For further details, please refer to this Jupyter Notebook

Split data

- X represents the features data
  - This includes all the data after cleaning and preprocessing is done except for column ‘clicked_on_ad’
  - It has a total of 51 columns

- y represents the target data from column ‘clicked_on_ad’

For further details, please refer to this Jupyter Notebook

Data Cleaning & Preprocessing

- Pada tahap ***cleaning data***, tunjukan ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, serta cara penyelesaiannya.

- Tulislah pula proses **extract datetime data** sebelum dilakukan model machine learning.

- Tunjukan **Split Data** sebelum melakukan model machine learning

- Tulislah proses **feature encoding** pada tahap ini (gunakan get_dummy)

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini