|  |  |  |
| --- | --- | --- |
| Checkpoint II | Checkpoint II: Data Cleaning & Processing | |
| Group: | G16 |
| Date: | 2020/10/16 |
|  |  |

# Initial Dataset

The datasets we’ll be using are “*Cell Phones Brands and Models*”, a dataset containing over 8000 models and 100 brands, each model along with its hardware specifications; and “*List of best-selling mobile phones - Annual sales by manufacturer*”, which has information about the revenue of each of the major brands by year.

**(from “Dataset\_Cell\_Phones\_Model\_Brand.json”)** { "Model": "\_3", "Brand": "Nokia", "Battery": "Non-removable Li-Ion 2630 mAh battery", "Sensors": "Accelerometer| gyro| proximity| compass", "Announced": "2017  February", "Audio\_jack": "Yes", "Bluetooth": "4.0| A2DP| LE", (...) "GPS": "Yes with A-GPS", "Radio": "FM radio with RDS", "Display\_type": "IPS LCD capacitive touchscreen  16M colors", "Display\_resolution": "5.0 inches (~67.3% screen-to-body ratio)", "Display\_size": "720 x 1280 pixels (~294 ppi pixel density)", "RAM": "2 GB RAM", "Internal\_memory": "16 GB", "Primary\_camera": "8 MP| f/2.0| autofocus| LED flash|"}

**(from “List of best-selling mobile phones - Annual sales by manufacturer”)** Nokia; 3; 5; 9; 13; 8; 20.593; 37.374; 76.335; 126.369; 139.672; 151.422; 180.672; 207.231; 265.615; 344.916; 435.453; 472.315; 440.8816; 461.3182; 422.4783; 333.938; 250.7931; ; ; ; ; ;

# Selected/Derived Data

The **selected** **attributes** from the first dataset are *Model*, *Brand*, Sensors, Audio\_jack, Bluetooth, GPS, Radio, Display\_type, Display\_resolution, Display\_size, RAM, Internal\_memory, Primary\_camera. From the second dataset, we selected the Brands, Years and Sales. The **derived measures** are Aspect\_ratio (dimension1 / dimension2, both extracted from Display\_resolution), ram\_MB and im\_MB (both converted to MB from the attributes *RAM* and *Internal\_memory*, respectively and Year, Quarter and Month (parsed from Announced, months were sometimes converted to respective quarter)  and # Models (derived from models dataset, separated by brand and year).

# Data Abstraction

The first dataset *ModelsParsed.csv* is of Table type and static with 8186 items each with 28 attributes that describe it. Each item of this dataset represents a phone model produced.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Semantic** |
| Model, Brand | Nominal | Name of the model and brand |
| Year, quarter, month | Continuous Sequential | Date the model was announced |
| Audio\_jack, Bluetooth, GPS, Radio | Nominal | Model has the technology (Boolean) |
| battery\_removable | Nominal | Battery is removable (Boolean) |
| battery\_amps | Ratio Sequential | AmpsH of the battery |
| battery\_type , display\_type | Nominal | String describing both types |
| aspect\_ratio, screen\_body\_ratio | Ratio Sequential | Ratio of screen and % screen to body |
| ram\_MB, im\_MB | Ratio Sequential | MB of RAM and Internal Memory |
| primary\_camera\_MP | Ratio Sequential | Megapixels of primary camera |
| primary\_camera\_autofocus,primary\_cam era\_LED\_flash,primary\_camera\_VGA | Nominal | Model has the camera spec (Boolean) |
| sensor\_accelerometer,sensor\_fingerprint, sensor\_heart\_rate,sensor\_iris\_scanner, sensor\_proximity,sensor\_temperature | Nominal | Model has the sensor (Boolean) |
| sensor\_fingerprint\_mounted | Nominal | Where fingerprint is mounted (String) |

The second dataset *BrandsParsed.csv* is of Table type, static, and time-based with 1239 items each with 4 attributes that describe it. Each item of this dataset represents a record of a given brand in a given year.

|  |  |  |
| --- | --- | --- |
| Brand | Nominal | Brand of record |
| Year | Continuous Sequential | Year of the record |
| # Models, Sales | Ratio Sequential | Nr of models produced and Sales in Millions of $ |

# Data Processing

The processing for the **first final dataset** was done mostly by parsing string attributes from the first original dataset and converting it into another type for the final dataset. For the **second final dataset**, we took the original data of the second original dataset (a table of Brand by Year, with the sales as values) and converted it into a table with columns Brand, Year, Sales and # Models (from the first original dataset). Some of the main problems were: extracting relevant data from the first dataset (like the camera attributes), where we had to use *Regex* and *Filters*, excluding *outliers* (using *Filters* to remove these values) and assigning a *sentinel value* of null for *missing values*.

# Mapping (Data sample/Questions)

* *“What are the* ***brands*** *that manufacture* ***models*** *that prioritize* ***battery life (mAh)*** *over other specs?”* - comparing which **brands (Acer)** have more **models (\_X960)** with higher **battery life (1530)** in a given **year (2009)**.
* *“What cell phone* ***brands*** *had a peak in* ***sales****?* ***When (year (1997))****?”* - comparing the **Sales (2631)** values for a given **brand (Alcatel)**.
* *“How many* ***models*** *did each* ***brand*** *develop in a given* ***time period****?”* - the **# Models (1)** attribute in BrandsParsed.csv (only for a given year) or extracting, from ModelsParsed.csv, all models of each **brand (Alcatel)** released in a time interval(comparing **Year/Quarter/Month (1997)**).
* *“Is there a correlation between the* ***number of models*** *of a* ***brand (Alcatel)*** *and that brand’s* ***revenue****?”*- comparing the **number of models (1)** released in a year and that year’s **sales(2631)**.
* *“Is there a cyclic period of releases of phone* ***models****? Do the peaks occur every* ***year****? Every six* ***months****?”* - graphing the releases of ***models* (\_X960) *months (February)*** by month over some ***years* (2009)**and calculating where peaks are (if they exist).
* *“When did a certain specification /* ***hardware component*** *start to be implemented on phones? What was its prevalence in phone* ***models*** *across the* ***years****?”* - graphing which **models (\_X960)** have a **certain attribute (for instance bluetooth)** over **time (2009/1/February)**.
* *“Is there a relationship between the sudden usage of a* ***new component*** *(like Bluetooth, DUAL SIM, etc. …) by a* ***brand (Apple)*** *and the change in* ***revenue*** *of that brand?”* - checking if the increase in use of an **attribute (Audio\_jack)** over the **years (2014)** coincides with an increase in **sales (191426)**.