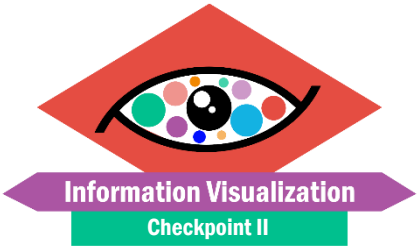

Checkpoint II: Data Cleaning & Processing

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 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	Ordinal	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	Continuous	AmpsH of the battery
battery_type , display_type	Nominal	String describing both types
aspect_ratio, screen_body_ratio	Ratio	Ratio of screen and % screen to body ratio
ram_MB, im_MB	Ratio	MegaBites of RAM and Internal Memory

primary_camera_MP	Ratio	Megapixels of primary camera
primary_camera_autofocus,primary_camera_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 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	Ordinal	Year of the record
# Models, Sales	Ratio	Number of models produced by brand in year 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 the IQR formula) 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 over other specs?"* - comparing which brands have more models with higher battery life in a given year. Attributes: Brand, Model, battery_mAh, Year.
- *"What cell phone brands had a peak in sales? When?"* - comparing the Sales values for a given brand. Attributes: Brand, Sales, Year.
- *"How many models did each brand develop in a given time period?"* - the # Models attribute in BrandsParsed.csv (only for a given year) or extracting, from ModelsParsed.csv, all models of each brand released in a time interval (comparing Year/Quarter/Month). Attributes: Brand, Model, #Models, Year, Quarter, Month.
- *"Is there a correlation between the number of models of a brand and that brand's revenue?"*- comparing the number of models released in a year and that year's sales. Attributes: Brand, # Models, Sales.
- *"Is there a cyclic period of releases of phone models? Do the peaks occur every year? Every six months?"* - graphing the releases of models month by month over some years and calculating where peaks are (if they exist). Attributes: Model, Year, Month.
- *"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 have a certain attribute over time. Attributes: Model, Year, Month, any component attribute.
- *"Is there a relationship between the sudden usage of a new component (like Bluetooth, DUAL SIM, etc. …) by a brand and the change in revenue of that brand?"* - checking if the increase in use of an attribute over the years coincides with an increase in sales. Attributes: Brand, Year, Sales, any component attribute.