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

# Initial Dataset

## We came across two different datasets that will complement each other to enhance the visualization:

## 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.

## The first dataset is available in [Back4App](https://www.back4app.com/database/paul-datasets/cell-phone-dataset) and can be freely downloaded for further use and can be accessed either by the raw file or by their API. The second dataset will need to be treated because the information is stored on a table of a [Wikipedia Web Page](https://en.wikipedia.org/wiki/List_of_best-selling_mobile_phones#Annual_sales_by_manufacturer). Furthermore, the last dataset will complement the first dataset, which doesn't contain the brands sales.

We are going to use all the attributes from both datasets, since all help answer the following questions.

(from “Dataset\_Cell\_Phones\_Model\_Brand.json”)

Brand; Model; Announced; Audio\_jack; Battery; Bluetooth; CPU; Chipset; Colors; Dimensions; Display\_resolution; Display\_size; Display\_type; EDGE; FourG; GPRS; GPS; GPU; Internal\_memory; Loud\_speaker; Memory\_card; NFC; Network; Network\_Speed; Operating\_System; Primary\_camera; RAM; Radio; SIM; Secondary\_camera; Sensors; Status; ThreeG; TwoG; USB; WLAN

{

"Model": "\_3",

"Brand": "Nokia",

"Network": "GSM / HSPA / LTE",

"TwoG": "GSM 850 / 900 / 1800 / 1900 - SIM 1 & SIM 2 (dual-SIM model only)",

"ThreeG": "HSDPA 850 / 900 / 1900 / 2100",

(...)

"Radio": "FM radio with RDS",

"USB": "microUSB 2.0| USB On-The-Go",

"Sensors": "Accelerometer| gyro| proximity| compass",

"Battery": "Non-removable Li-Ion 2630 mAh battery",

"Colors": "Silver White| Matte Black| Tempered Blue| Copper White"

}

(from “List of best-selling mobile phones - Annual sales by manufacturer”)

Manufacturer; 1992; 1993; 1994; 1995; 1996; 1997; 1998; 1999; 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019

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

## Description of:

## Which were selected from your dataset?

## Which derived measures did you consider and why (based on your tasks and questions)?

# Data Abstraction

Data abstraction description:

* Description of the dataset type (spatial, table, field, etc.);
* Description of each item and attribute (nominal/ordinal/etc., diverging/sequential scale, etc.);
* Semantics (what does each attribute and item stand for).

# Data Processing

Description of how the dataset was processed (cleaned, problems found and solutions, how did you fix missing values, cross-referenced different tables/datasets, etc.).

# Mapping (Data sample/Questions)

Some examples that show that with your data sample you will be able to provide the answers to the questions you formulated.