# Domácí úkol na BI-BIG

Vypracoval Matyáš Skalický

## Dataset

Dataset byl pro účel zadání upraven a rozdělen na 3 datasety za pomocí frameworku Pandas. JupyterNotebook obsahující tuto transformaci je popsán v souboru <DatasetTransform.ipynb>.

### Zdroj datasetu:

<https://www.kaggle.com/mehdidag/black-friday>

### Popis datasetu

Vzhledem k faktu, že pro agregace používám jeden a ten samý dataset rozdělený na 3 části, popíšu nejprve všechny sloupce datasetu a následně jejich rozdělení do nových menších datasetů.

Dataset obsahuje 550 000 záznamů z Black Friday z amerického obchodního domu. Obsahuje jak numerické, tak kategorické proměnné. Obsahuje chybějící hodnoty.

### Sloupce

Popis sloupců původního datasetu.

#### User\_ID 537577 non-null int64

* Unikátní identifikátor uživatele.

#### Product\_ID 537577 non-null int64

* Unikátní identifikátor produktu.

#### Gender 537577 non-null object

* Pohlaví zákazníka (M/F).

#### Age 537577 non-null object

* Věková skupina zákazníka.

#### Occupation 537577 non-null int64

* Kategorie dle typu zaměstnání (0-20).

#### City\_Category 537577 non-null object

* Kategorie města (nejspíš dle velikosti).

#### Stay\_In\_Current\_City\_Years 537577 non-null object

* Délka pobytu v současném městě v letech.

#### Marital\_Status 537577 non-null int64

* Manželský status (1=ženatý, 0=jinak).

#### Product\_Category\_1 537577 non-null int64

* Počet produktů, které uživatel zakoupil z cenové kategorie 1.

#### Product\_Category\_2 370591 non-null float64

* Počet produktů, které uživatel zakoupil z cenové kategorie 2.

#### Product\_Category\_3 164278 non-null float64

* Počet produktů, které uživatel zakoupil z cenové kategorie 3.

#### Purchase 537577 non-null int64

* Částka, kterou uživatel utratil v USD.

### Ukázka datasetu (10 řádků)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| User\_ID | Product\_ID | Gender | Age | Occupation | City\_Category | Stay\_In\_Current\_City\_Years | Marital\_Status | Product\_Category\_1 | Product\_Category\_2 | Product\_Category\_3 | Purchase |
| 1000001 | 69042 | F | 0-17 | 10 | A | 2 | 0 | 3 | nan | nan | 8370 |
| 1000001 | 248942 | F | 0-17 | 10 | A | 2 | 0 | 1 | 6 | 14 | 15200 |
| 1000001 | 87842 | F | 0-17 | 10 | A | 2 | 0 | 12 | nan | nan | 1422 |
| 1000001 | 85442 | F | 0-17 | 10 | A | 2 | 0 | 12 | 14 | nan | 1057 |
| 1000002 | 285442 | M | 55+ | 16 | C | 4+ | 0 | 8 | nan | nan | 7969 |
| 1000003 | 193542 | M | 26-35 | 15 | A | 3 | 0 | 1 | 2 | nan | 15227 |
| 1000004 | 184942 | M | 46-50 | 7 | B | 2 | 1 | 1 | 8 | 17 | 19215 |
| 1000004 | 346142 | M | 46-50 | 7 | B | 2 | 1 | 1 | 15 | nan | 15854 |
| 1000004 | 97242 | M | 46-50 | 7 | B | 2 | 1 | 1 | 16 | nan | 15686 |
| 1000005 | 274942 | M | 26-35 | 20 | A | 1 | 1 | 8 | nan | nan | 7871 |

### Rozdělení datasetu na 3 nové datasety.

TODO

# Import dat do databáze

Databáze běží v docker kontejneru. Vyžaduje nainstalovaný a plně funkční [docker-compose](https://docs.docker.com/compose/).

## Build image pro spark

Bude vytvořen image *spark* který bude využit dle parametrů při spuštění jak pro worker, tak pro master node.

docker build -f spark.df -t spark .

## Spuštění master a worker node

docker-compose up

## Spuštění spark-shell

docker run -it -p 8088:8088 -p 8042:8042 -p 4041:4040 --name driver -h driver spark:latest bash