**Badanie wpływu pogody na zamówienia codziennych zakupów online**

## Zespół Danonkowych Żółwi Ninja:

## *Katarzyna Solawa i Mateusz Sperkowski*

Nasz projekt obejmuje dwa zbiory danych (<https://www.kaggle.com/c/instacart-market-basket-analysis/data>, <https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events>). Pierwszy to dane o dostawach zakupów do mieszkańców USA z aplikacji Instacart z roku 2017, a drugi mówi o wydarzeniach pogodowych w różnych miastach USA. Celem hurtowni jest zbadanie zależności kupowanych produktów od pogody w danym dniu.

1. Opis biznesowy [2pkt]
   1. Why - Użytkownik projektu będzie mógł zbadać jak na zamówienia ludzi w danym dniu wpływa pogoda. W ten sposób może on potencjalnie wykorzystać tą wiedzę do obrania nowej/ulepszenia strategii w dni z gorszą lub lepszą pogodą, by zwiększyć dochody przedsiębiorstwa. Przykładowo może być tak, że w deszczowe dni jest więcej zamówień, ponieważ ludzie nie chcą wychodzić z domu i częściej zamawiają zakupy kategorii X. Taką wiedze można wykorzystać i zoptymalizować działanie firmy.
   2. How - Końcowemu użytkownikowi pokazywany będzie raport wizualizujący zależności i wykresy dotyczące różnic w zakupach ludzi zależnie od pogody. Może to pozwolić odkryć trudne do zidentyfikowania zależności w zakupach ludzi.
   3. What - Pierwszy zbiór danych to kolejne zamówienia użytkowników na zakupy z aplikacji Instacart. Obsługuje ona mieszkańców USA i przedstawia co było kupione, z jakiej półki/kategorii i czy to kolejny zakup użytkownika. Drugi zbiór mówi o wydarzeniach pogodowych w różnych miastach USA i przedstawia rodzaj wydarzenia, date i jego intensywność.
2. Opis danych [2pkt]
   1. opis schemy wejściowej danych

**Instacart -**

aisles.csv -

aisle\_id – klucz główny, int;

aisle – opis alejki, string;

### Departments.csv -

department\_id - klucz główny, int;

department - oddział sklepu;

### order\_products\_\_\*.csv

order\_id – klucz główny, int

product\_id – klucz główny, klucz obcy, int

add\_to\_cart\_order - kolejność dodania do zamówienia, int

Reordered – czy wcześniej klient zamawiał przedmiot, int

### orders.csv

order\_id - klucz główny, int

user\_id - klucz główny, int

eval\_set – typ danych, test train prior, string

order\_number – numer orderu uzytkownika, int

order\_dow – dzien tygodnia, int

order\_hour\_of\_day – godzina dnia, int

days\_since\_prior\_order - dni od ostatniego zamowienia, float

### products.csv

product\_id – klucz główny, int

product\_name – nazwa produktu, string, int

aisle\_id – nazwa alejki, klucz obcy, int

department\_id – klucz obcy, przynaleznosc do oddziału, int

### sample\_submission.csv - tabela nam niepotrzebna, ignorowana

order\_id – klucz obcy, nr. zamowienia

Products – produkty, int

**us-weather-events**

us-weather-events.csv

EventId – klucz główny, string

Type – string, typ opadów/wydarzenia pogodowego

Severity – string, poziomowy opis powagi wydarzenia

StartTime(UTC) - data, RRRR-MM-DD HH:MM:SS

EndTime(UTC) - data, RRRR-MM-DD HH:MM:SS

Precipitation(in) - wielkość opadów w calach

TimeZone – string, strefa czasowa

AirportCode – string, lokalizacja najbliższego lotniska

LocationLat - szerokość geograiczna, float

LocationLng - długość geografczna, float

City – miasto, string

County – hrabstwo, string

State – stan, string

ZipCode – kod pocztowy, int

**Calendar** – ten sam co korzystaliśmy na zajęciach

* 1. typ źródła - pliki .csv z tabelami + generowanie danych uzupełniających/potrzebnych do połączenia

Sposób generowania - Python:

* + - Dodanie dat zamówień na bazie: order\_number, order\_dow i days\_since\_prior\_order
    - Potencjalnie dodanie miast do zamówień

1. Opis architektury
   1. Podział na warstwy / moduły - **zadanie każdej warstwy, co jest wystawiane na wyjściu + najlepiej diagram architektury [1 pkt]**



* 1. ETL - **lista transformacji poszczególnych pól, sposób ewentualnego łączenia danych [2pkt]**

Usunięcie kolumny eval\_set, order\_dow, days\_since\_prior\_order z Orders.csv.

Usunięcie kolumny airportCode z us-weather-events

Złączenie tabel - order\_products\_\_\*.csv, Orders.csv, po order\_id,

Złączenie tabel - Aisles.csv, Departments.csv, products.csv - po department\_id i aisle\_id

Ustawienie klucza obcego w datach w order\_products i us-weather-events z Calendar – tu w pogodzie godziny wyrzucamy do nowej kolumny z godzinami. Jednego dnia/jedna data będzie w danej lokalizacji mogła mieć wtedy więcej wydarzeń niż jedno.

* 1. Model hurtowni - **podział na tabele faktów oraz wymiary, typy kolumn, relacje pomiędzy tabelami, diagram hurtowni [2pkt]**

