IUM Projekt – Etap 1 (zadanie 2)

Wojciech Nowicki, nr indeksu 304088

Gustaw Daczkowski, nr indeksu 304031

1. Definicja problemu biznesowego

“Są osoby, które wchodzą na naszą stronę i nie mogą się zdecydować, którym produktom przyjrzeć się nieco lepiej. Może dałoby się im coś polecić?"

Problemem biznesowym jest stworzenie systemu, który dla użytkownika odwiedzającego sklep internetowy będzie w stanie wygenerować rekomendacje produktów, którymi to potencjalny kupujący byłby zainteresowany. Ma to na celu zwiększenie dziennej liczby wyświetleń produktów, co może zwiększyć sprzedaż przekładając się na wzrost przychodów sklepu. Obecnie średnia dzienna liczba wyświetleń produktów w sklepie to 1066 odsłon. Planowane jest zwiększenie tej liczby o 15%, czyli do około 1226 wyświetleń na dzień.

2. Zadanie modelowania oraz kryteria sukcesu

Zadaniem modelowania jest przygotowanie modelu rekomendacyjnego, który na podstawie dostarczonych danych (sesje użytkowników, aktualny katalog produktów, informacje o użytkownikach) będzie w stanie wybrać takie produkty z aktualnego katalogu, które użytkownik będzie skłonny zakupić. Analityczny kryterium sukcesu będzie współczynnik:

$$\frac{rekomendacje\ kliknięte\ przez\ użytkownika}{wszystkie\ rekomendacje}$$

którego wartość powinna być większa bądź równa:

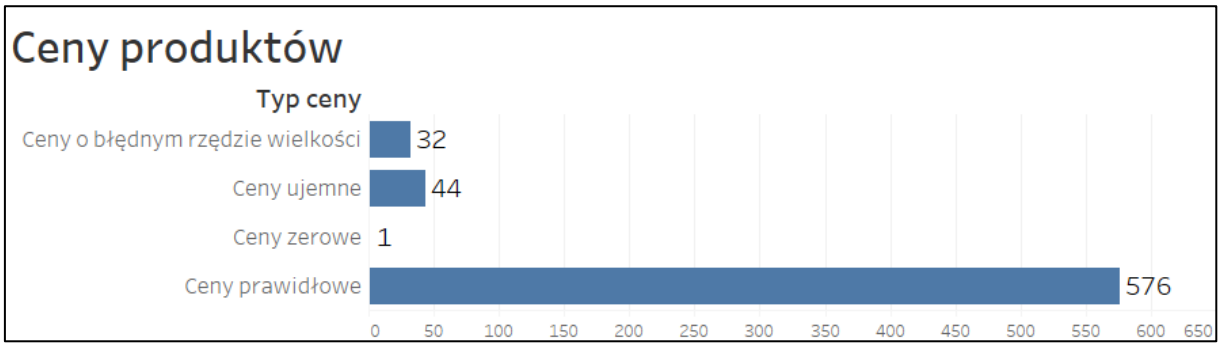
$$\frac{160}{dzienna\ liczba\ odsłon\ strony\ sklepy}$$

W ramach jednego zestawu rekomendacji będzie pojawiać się co najwyżej 5 przedmiotów sugerowanych dla danego użytkownika.

3. Analiza danych

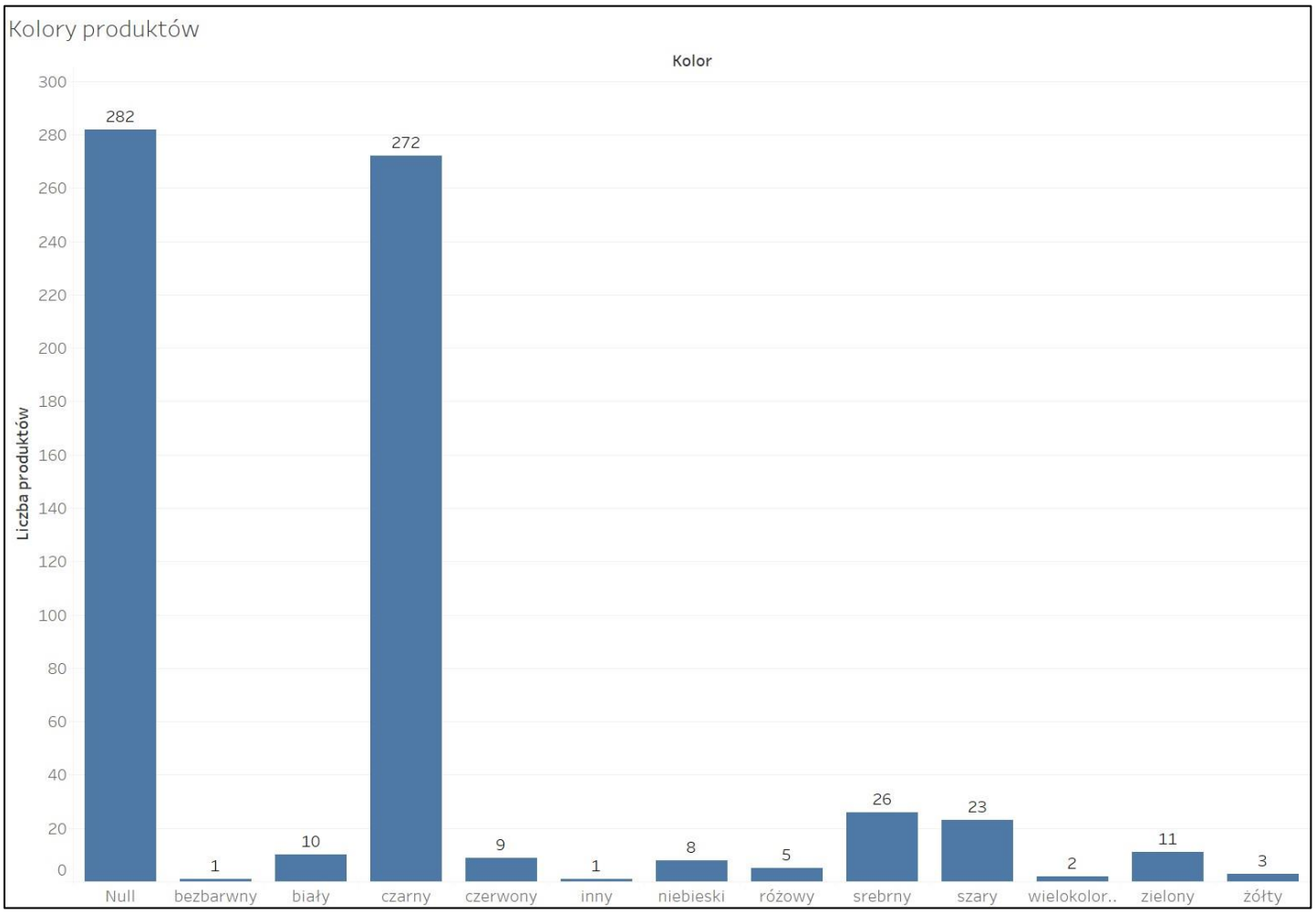
Produkty:

- Ceny produktów rozkładają się następująco:



Do cen ujemnych zostanie użyty moduł. Produkty z cenami zerowymi będą usunięte z katalogu, a produkty z cenami o złym rzędzie wielkości zostaną naprawione ręcznie - o ile kontekst (czyli inne podobne produkty o właściwych cenach) na to pozwala.

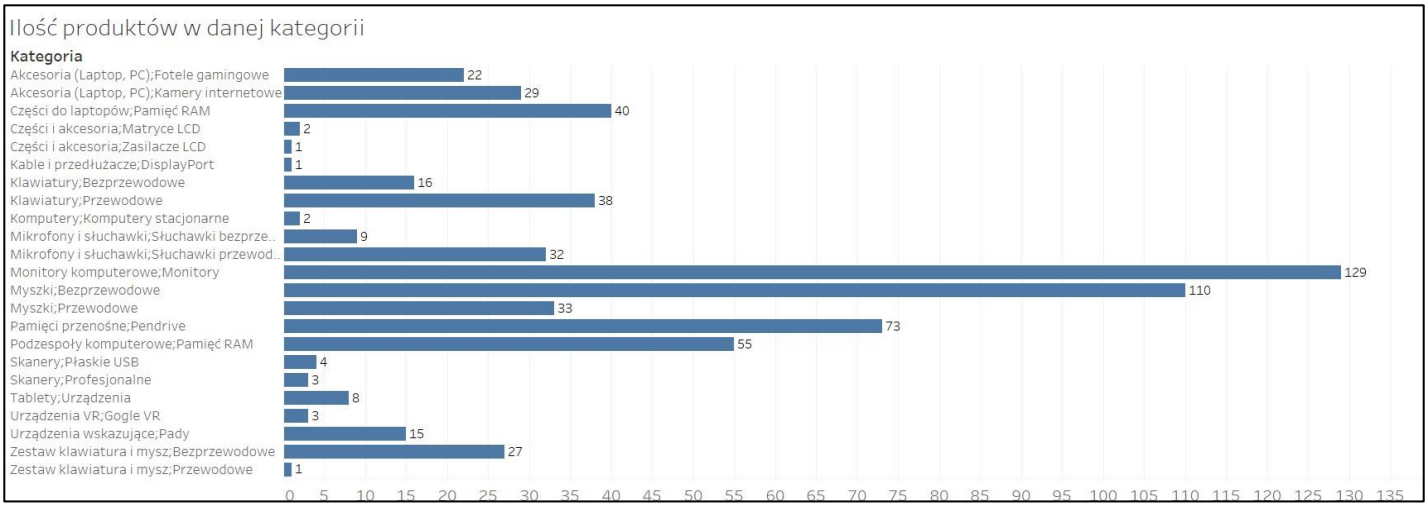
- Rozkład kolorów wśród produktów:



Dla wielu produktów (z tych, które mają wartość Null) kolor nie ma znaczenia lecz znajdują się tam też takie, dla których kolor miałby sens i realny wpływ (tutaj prośba o więcej danych). Ponadto są przypadki, dla których kolor nie znajduje się w *optional_attributes*, a widnieje w nazwie – w takich sytuacja kolor zostanie wyciągnięty z nazwy produktu i przekazany do *optional_attributes* w celu ułatwienia analizy.

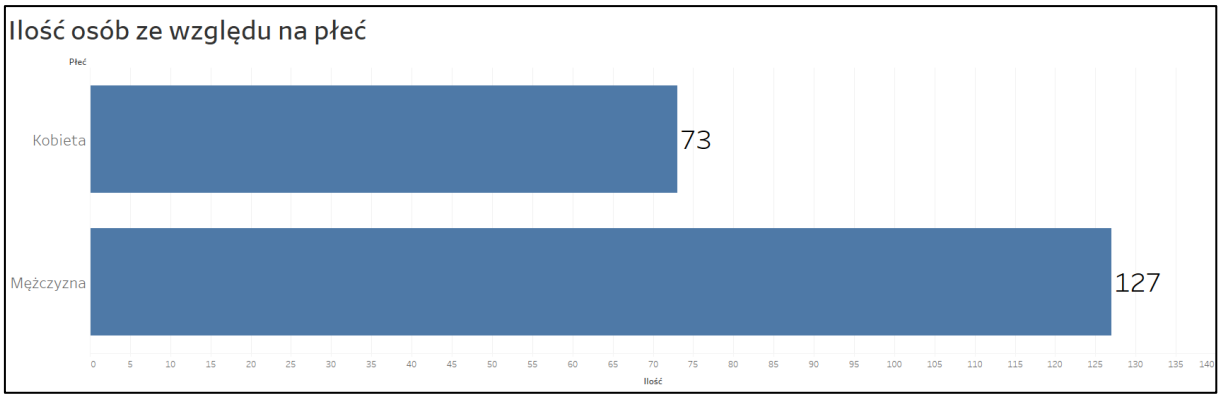
- Przydatną informacją poza oceną użytkowników, byłaby także liczba tych ocen – pozwalałoby to określić rzetelność oceny produktu. (prośba o udzielenie tych danych)
- Rekordy z polskimi literami zakodowanymi według konwencji “unicode escape” zostaną przekonwertowane na odpowiedniki literowe
- Ścieżka do produktu (kategorie) zostanie rozbita – obecna forma nie jest "przyjazna" do przetwarzania.

- [opcjonalnie] Rozszerzenie oferty sklepu, gdyż w niektórych kategoriach znajduje się tylko jeden produkt:



Użytkownicy:

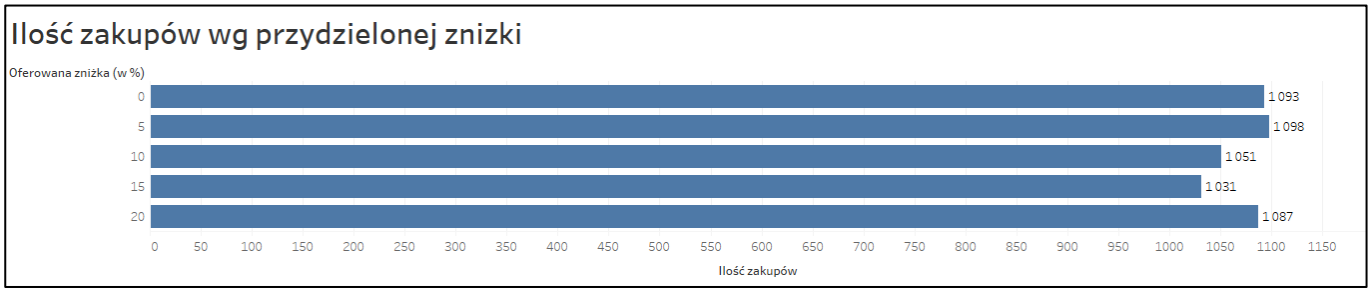
- Rekordy z polskimi literami zakodowanymi według konwencji “unicode escape” zostaną przekonwertowane na odpowiedniki literowe
- Może wystąpić potrzeba zbalansowania danych użytkowników, gdyż mężczyzn jest znacznie więcej niż kobiet (przyjmując, że mam do czynienia z polskimi imionami)



Sesje:

- Dane będą służyć do trenowania i walidacji modelu - najważniejsze z punktu widzenia całego zadania.
- Średni czas trwania sesji to obecnie 12 minut 49 sekund (czas mierzony był od pierwszego do ostatniego eventu w sesji, przy czym pominięto sesje z jednym eventem, gdyż nie da się zweryfikować jej długości)
- Średnia liczba wyświetleń na sesję to obecnie 5 odsłon (zaokrąglone do całości)
- Średnia dzienna liczba wyświetleń to obecnie 1066 odsłon (zaokrąglone do całości)
- Średnia dzienna liczba zakupów to obecnie 64 (zaokrąglone do całości)

- Na podstawie dwóch powyższych podpunktów otrzymuje współczynnik średnio 17 wyświetleń na zakup, który w ramach działania modelu powinien ulec pomniejszeniu.
- Wysokość przydzielonej zniżki nie ma wpływu na finalny zakup (analiza zniżki przy zakupie per produkt wykazuje podobne rezultaty jak poniżej)



- Niektóre z rekordów zostaną usunięte, gdyż nie posiadają (wartość null) najważniejszej informacji z perspektywy trenowania modelu – identyfikatora produktu.
- W niektórych przypadkach brakuje identyfikatora użytkownika – jeśli będzie to możliwe zostanie on uzupełniony na podstawie identyfikatora sesji (poprzedzające / następujące logi). Jeśli nie będzie to możliwe, tak jak w poprzednim przypadku, rekordy zostaną usunięte.

Dostawy (opcjonalnie):

- Na podstawie dat zamówienia i doręczenia można określić średni czas dostawy danego produktu – może to być przydatny atrybut przy tworzeniu modelu o ile użytkownicy w sklepie internetowym mają wgląd do takiej statystyki.
- Obecny średni czas dostawy wynosi 3,5 dnia. (pominięto rekordy, w których brakowało czasu doręczenia)

4. Podsumowanie

- Dane poprawne / do przygotowania, który zostaną wykorzystane w procesie modelowania:
  - Informacje o użytkowniku:
    - Identyfikator użytkownika
    - Płeć
    - Miejsce zamieszkania (z agregacją do miasta)
  - Sesje użytkownika:
    - Identyfikator sesji
    - Data
    - Długość sesji
    - Zakres czasowy sesji (od, do)
    - Przeglądane produkty (identyfikatory)
    - Zakupione produkty (jeśli były)

    - Produkty
        - Nazwa
        - Kategoria
        - Kolor (niepełny)
        - Cena
        - Marka
        - Ocena
        - Ilość ocen
        - Średni czas dostawy (wyliczony na podstawie *Dostaw*)

- Prośba o dodatkowe dane:
    - Liczba ocen danego produktu (dodatkowo obok średniej oceny)
    - Dodanie kolorów do produktów, w których są braki (brak jako *optional_attributes* i brak w nazwie przedmiotu) jeśli ma to sens
    - [opcjonalnie] Rozszerzenie oferty sklepu, w kategoriach gdzie występuje jeden produkt