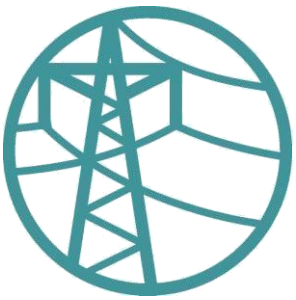


WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Maciej Paluch

Usługi Sieciowe w Biznesie

Business Intelligence z zastosowaniem narzędzi KNIME, R, Python i Tableau

Rzeszów, 2022

Spis treści

1. Założenia projektu

Celem projektu jest zastosowanie elementów Business Intelligence na zestawie danych historycznych firmy zajmującej się sprzedażą sprzętu elektronicznego. Za pomocą narzędzi analitycznych takich jak KNIME, R, Python, oraz Tableau została przeprowadzona analiza biznesowa, która miała za zadanie przekształcić surowe dane w użyteczne informacje biznesowe. Wynikiem projektu są odpowiedzi na niektóre pytania biznesowe, oraz stworzone wizualizacje, tzw. dashboardy – dla graficznej prezentacji danych.

Zestaw danych został pobrany z portalu Kaggle.com.
Link: https://www.kaggle.com/datasets/knightbearr/sales-product-data

2. Opis użytych narzędzi

2.1. KNIME

KNIME jest darmowym oprogramowaniem wspierającym proces przetwarzania i przekształcania danych (w tym procesów ETL). Proces – (ang. workflow) jest zbiorem graficznych bloków funkcjonalnych (ang. nodes), które wykonują określone akcje takie jak np. dostęp do danych, utworzenie tabeli, dodanie kolumny itp. Węzły połączone ze sobą łączone tworzą spójny proces przetwarzania i analizy danych. W projekcie wykorzystany do wstępnego przetworzenia danych do dalszej analizy.

2.2. R

R jest językiem programowania stosowanym do obliczeń statystycznych, analizy i wizualizacji danych. Jest oparte o przetwarzanie wektorów, dlatego operacje na dużych zbiorach danych są szybsze niż w przypadku używania skomplikowanych pętli. W projekcie wykorzystany do sprawdzania poprawności operacji na danych i prostej klasteryzacji.

2.3. Python

Python jest jednym z najpopularniejszych języków programowania. Jest prosty w zrozumieniu, oraz posiada szeroki zasób bibliotek do analizy danych, czy machine learningu – dlatego jest wykorzystywany przez analityków danych na całym świecie. W projekcie wykorzystany do zaimplementowania algorytmu APRIORI – analizy koszykowej polegającej na znalezieniu grup wspólnie kupowanych produktów.

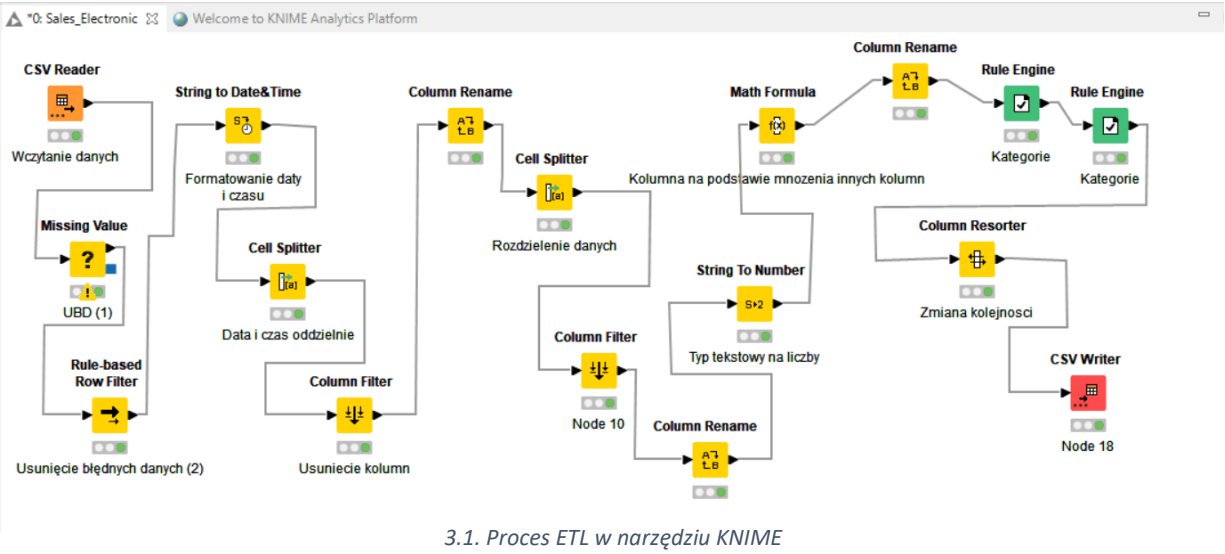
2.4. Tableau

Tableau jest narzędziem do tworzenia interaktywnych, przejrzystych wizualizacji danych. Wykorzystywany jest jako rozwiązanie skierowane dla biznesu – pozwala na podejmowanie decyzji w oparciu o dane, które dzięki przedstawieniu wizualnemu stają się bardziej czytelne. W projekcie wykorzystany do prezentacji wyników analizy.

3. Przygotowanie danych

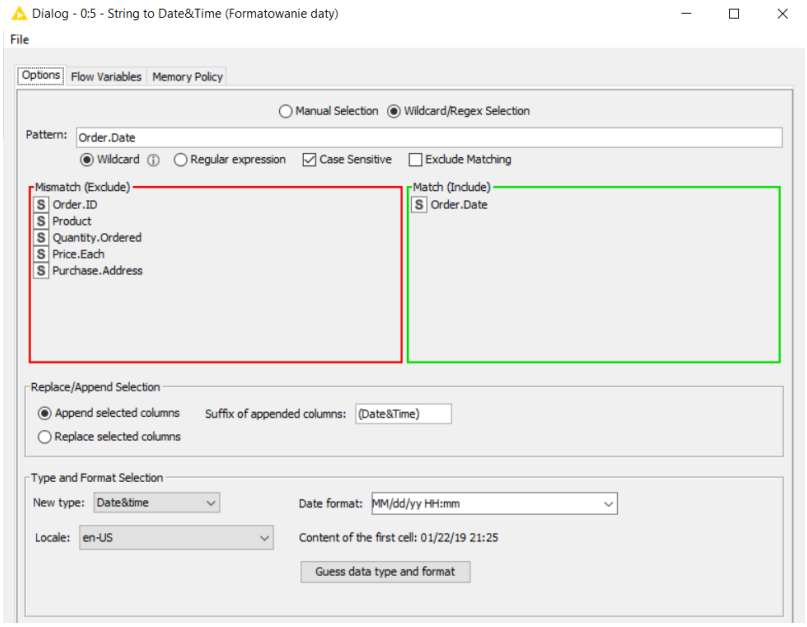
Projekt rozpoczął się od wczytania danych do programu KNIME. Został utworzony proces, dzięki któremu surowe dane zostały odpowiednio przygotowane do dalszej analizy. Na tym etapie projektu zostały wykonane takie zagadnienia jak usunięcie nieprawidłowych danych (w tym danych brakujących), poprawienie formatowania daty i czasu, zamiana typu tekstowego na liczbowy, dodanie odpowiednich kolumn. Tak przygotowany zestaw danych został zapisany do formatu *csv*.

Proces wykorzystany w projekcie wygląda następująco:



3.1. Proces ETL w narzędziu KNIME

Przykładowy węzeł – formatowanie daty i czasu:



3.2. Formatowanie daty i czasu w KNIME

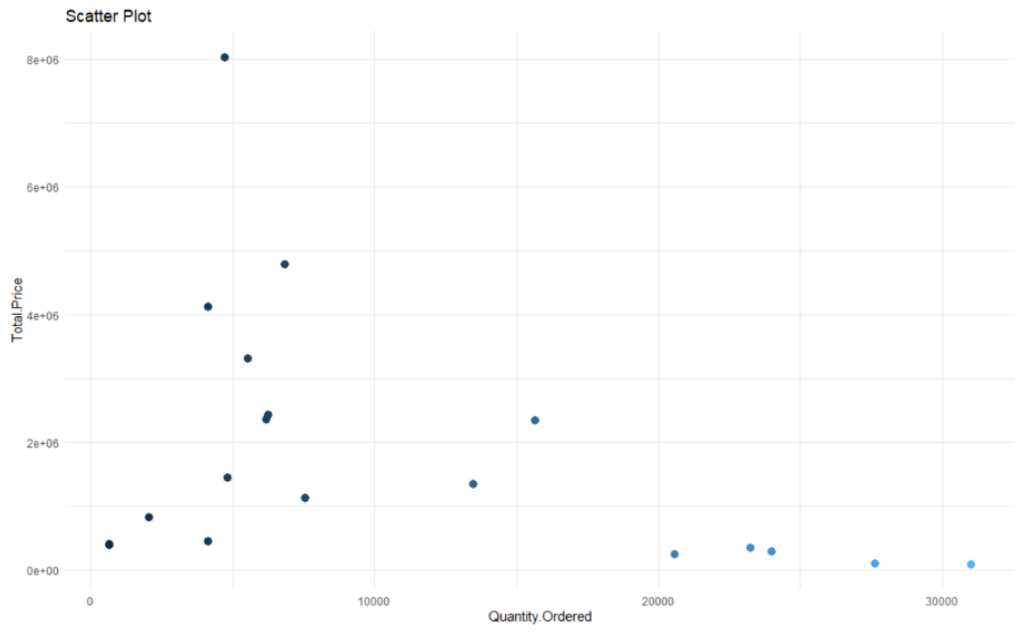
4. Poprawność danych i prosta klasteryzacja

4.1. Sprawdzenie poprawności danych

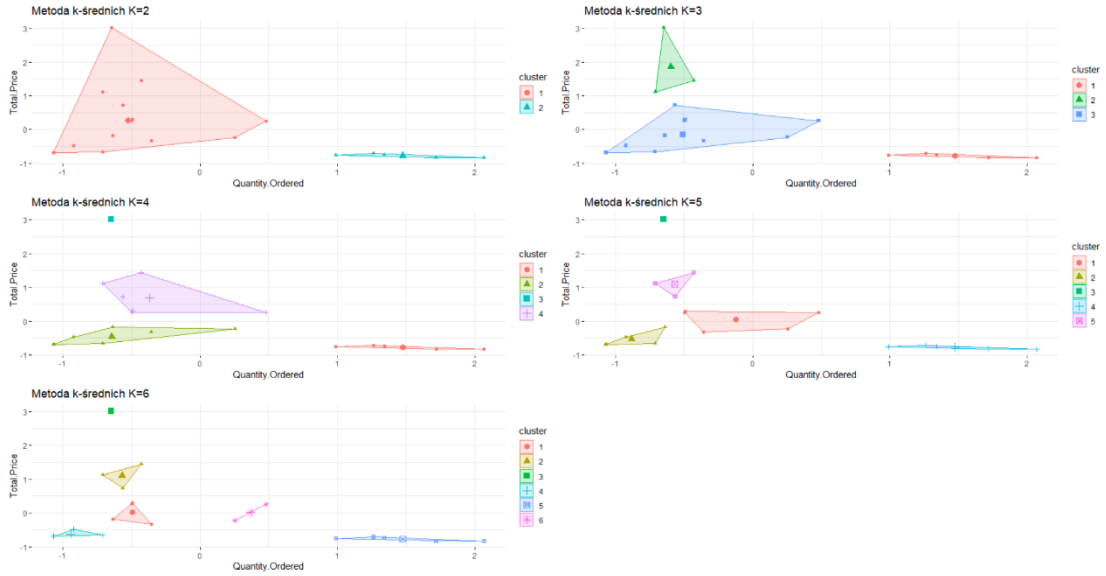
Przygotowane wcześniej dane zostały wczytane do środowiska RStudio i sprawdzone pod kątem poprawności danych – okazało się, że kolumna z całkowitą zapłatą za produkt została wczytana jako typ *character*, dlatego nastąpiło przeformatowanie na typ liczbowy. Z takiego zestawu zostały wybrane dane do klasteryzacji – celem było sprawdzenie jak grupowały się produkty ze względu na ilość zamówień a cenę, jaką przyszło zapłacić klientom za te właśnie produkty. Nastąpiły tutaj agregacje na podstawie powyższych cech, a następnie wybranie tylko danych numerycznych i ich standaryzacja.

4.2. Metoda k-means i wybór optymalnego k

Metoda klasteryzacji to metoda k-średnich (ang k-means) – dzięki temu algorytmowi zostały utworzone odmienne skupienia, które podzieliły zbiór danych na podobne sobie produkty. Na początku dla ustalenia optymalnej liczby skupień zostały wyświetlone wykresy zarówno rozkładu danych „surowych", jak i dla modeli klastrowania przy różnej wartości k (od k=2 do k=6).



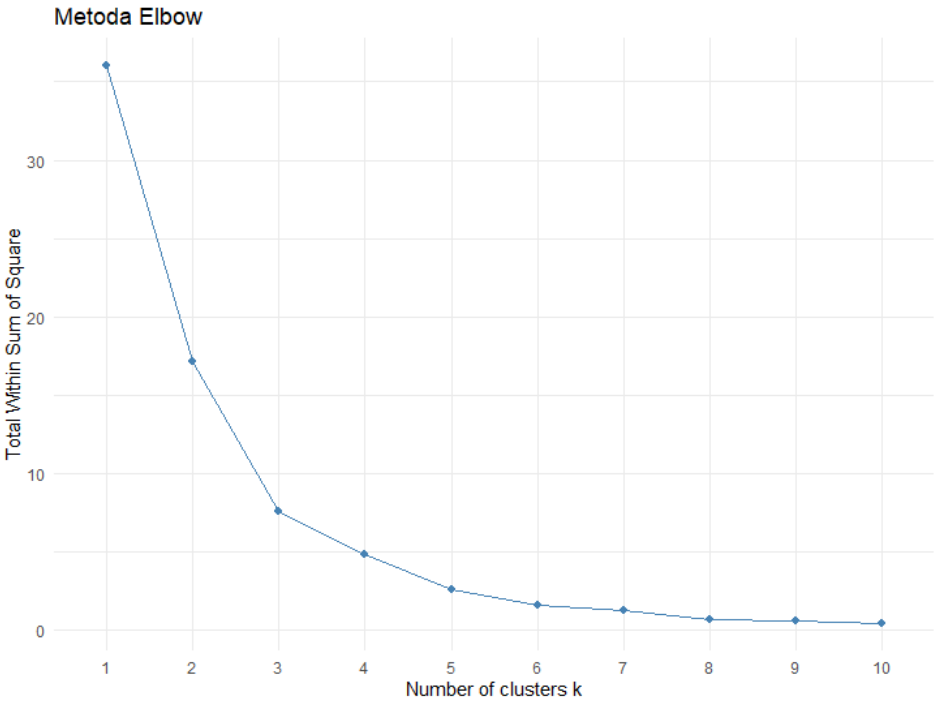
4.1. Wykres rozrzutu łącznej ceny od ilości zamówień



4.2. Klastry dla różnego k

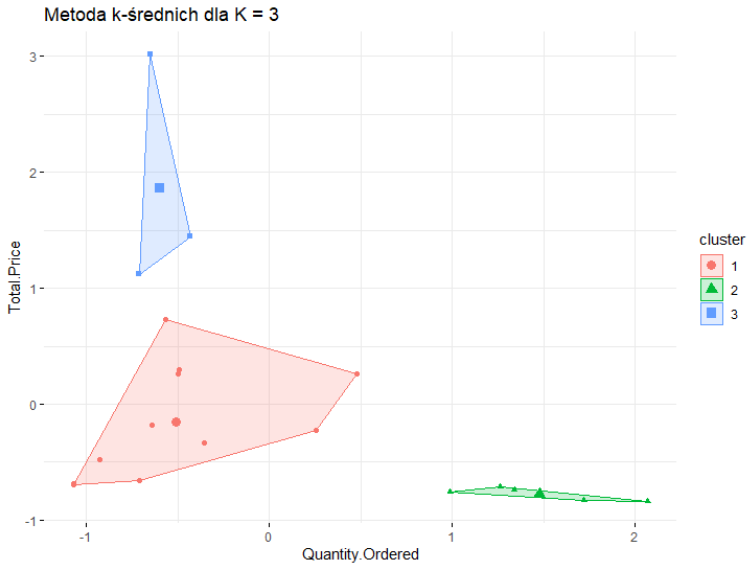
W metodzie k-średnich chodzi o takie dobranie liczby klastrów, aby łączna ich wariancja była minimalna. Wariancja ta jest przedstawiana jako suma kwadratów odległości pomiędzy każdą z obserwacji, a środkiem danego klastra.

Została wybrane metoda „łokcia” (ang. Elbow method) do identyfikacji optymalnej liczby klastrów. Metoda ta sprawdza, dla jakiej liczby klastrów suma wariancji przestaje gwałtownie maleć, a zwiększanie liczby klastrów nie wprowadza już widocznej poprawy. Jeśli chodzi o wykres to szukany jest punkt pierwszego widocznego „zagięcia”. W przypadku wybranych danych pierwsze zagięcie widoczne jest dla trzech klastrów.



4.3. Metoda "zagiętego łokcia" w klasteryzacji danych

Został utworzony model z liczbą klastrów równą 3. Widoczne są oddzielne skupiska.



4.4. Podział na skupiska dla k=3

4.3. Interpretacja wyników

Po spojrzeniu na wykres można dojść do pewnych wniosków:

- Do klastra 1 należą przedmioty o stosunkowo niewielkiej ilości zamówień i generujące stosunkowo niski zarobek
- Do klastra 2 należą przedmioty o dużej ilości zamówień łącznie, ale też generujące bardzo mały zarobek
- Do klastra 3 należą przedmioty o małej ilości zamówień łącznie, ale generujące bardzo duży zarobek

Dobrym pomysłem byłoby skupienie się na klastrze 3 pod względem reklamowania produktów – przedmioty te generują bardzo duży zarobek dla firmy, więc skuteczne reklamy mogą w przyszłości zaowocować jeszcze większą sprzedażą, co przenosi się na jeszcze większe dochody. Inną ideą jest przyglądnięcie się produktom z klastra 2 – są to bardzo tanie przedmioty, lecz zamawiane bardzo często, dlatego można je proponować klientom, którzy już mają w koszyku inny przedmiot z pozostałych klastrów. Dodatkowo te przedmioty są pasujące do innych – przykładowo do zamówienia iPhone'a można zaproponować kable do ładowarki.

Kod jest dostępny na platformie GitHub:
https://github.com/MPaluxch/Uslugi-Sieciowe-w-Biznesie/blob/main/Clusters.R

5. Algorytm APRIORI

5.1. Teoria

Algorytm APRIORI jest stosowany w procesie rekomendacji produktów. Należy do reguł asocjacyjnych. Mechanizm opiera się na analizie wcześniejszych zakupów – dokładniej mówiąc sprawdzenie, które produkty najczęściej były kupowane razem. W ten sposób można dowiedzieć się, które produkty warto zarekomendować, aby zwiększyć prawdopodobieństwo dodania ich do koszyka – w ten sposób sklep zarobi więcej na sprzedaży.

Algorytm APRIORI wskazuje dwa bardzo ważne pojęcia:

- Support – częstość występowania

$$Support(Produkt\ 1, Produkt\ 2) = \frac{liczba\ transakcji\ w\ której\ Produkt\ 1\ i\ 2\ występują\ razem}{wszystkie\ transakcje}$$

Jest to miara, która wskazuje jak często występują zakupy z Produktem 1 i Produktem 2. Im większa wartość, tym częstsze zjawisko.

- Confidence – pewność reguły

$$Confidence = \frac{Support(Produkt\ 1, Produkt\ 2)}{Support(Produkt\ 1)}$$

Jest to miara, która wskazuje, czy kupno Produktu 2, w przypadku kiedy był kupiony również Produkt 1, było przypadkiem. Im większa wartość, tym częściej spełniony jest warunek, gdzie ludzie, którzy kupili Produkt 1, kupują również Produkt 2. Jeżeli wartość jest niska, oznacza to że klienci kupowali Produkt 1, ale bez Produktu 2.

W obydwu przypadkach wartości są z zakresu (0, 1).

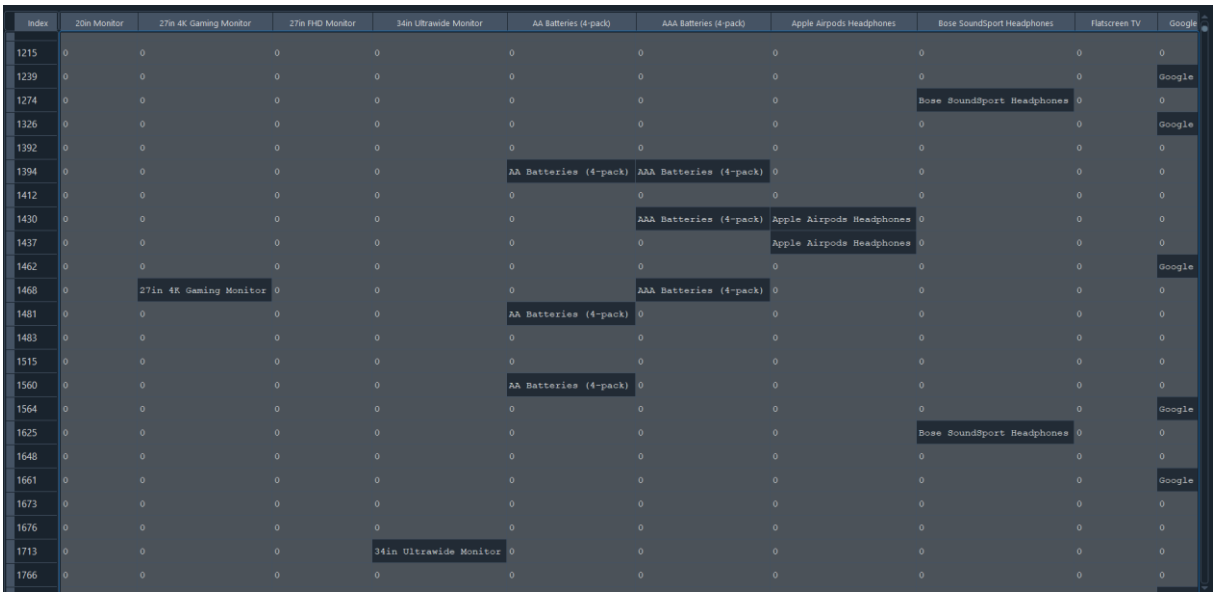
W rekomendacjach wybiera się reguły, które spełniają warunek wysokiej wartości confidence.

W dużych zbiorach danych mogą występować oczywiście informacje o zakupach większej ilości produktów. Zasada działania jest taka sama. Jeżeli klient dodał do koszyka dwa produkty, to dzięki analizie koszykowej, można sprawdzić który trzeci produkt często pojawiał się w innych, wcześniejszych transakcjach – w ten sposób można go zarekomendować.

5.2. Implementacja algorytmu w Pythonie

Na początku, zaraz po wczytaniu danych, wyselekcjonowane zostały tylko te numery transakcji, które się powtarzały – w ten sposób można było połączyć wszystkie produkty w jednej, tej samej transakcji. Została utworzona tablica 0/1 (tzw. „dummy table"), która wskazuje, czy dana zmienna kategoryczna występuje w danym rekordzie. Jeśli tak to przypisana wartość 1, w przeciwnym wypadku 0. Dla usprawnienia algorytmu tablica ta została zmieniona tak, aby zamiast 1 występowała nazwa danej zmiennej kategorycznej. Kolumny to nazwy danego produktu, wiersze – numer indeksu danej transakcji. W poniższym screenie można zauważyć, że w transakcji nr 1468 zostały zakupione razem produkty:

- 27 calowy monitor gamingowy 4K
- 4-pak baterii AAA



Index	20in Monitor	27in 4K Gaming Monitor	27in FHD Monitor	34in Ultrawide Monitor	AA Batteries (4-pack)	AAA Batteries (4-pack)	Apple Airpods Headphones	Bose SoundSport Headphones	Flatscreen TV	Google
1215	0	0	0	0	0	0	0	0	0	0
1239	0	0	0	0	0	0	0	0	0	Google
1274	0	0	0	0	0	0	0	Bose SoundSport Headphones	0	0
1326	0	0	0	0	0	0	0	0	0	Google
1392	0	0	0	0	0	0	0	0	0	0
1394	0	0	0	0	AA Batteries (4-pack)	AAA Batteries (4-pack)	0	0	0	0
1412	0	0	0	0	0	0	0	0	0	0
1430	0	0	0	0	0	AAA Batteries (4-pack)	Apple Airpods Headphones	0	0	0
1437	0	0	0	0	0	0	Apple Airpods Headphones	0	0	0
1462	0	0	0	0	0	0	0	0	0	Google
1468	0	27in 4K Gaming Monitor	0	0	0	AAA Batteries (4-pack)	0	0	0	0
1481	0	0	0	0	AA Batteries (4-pack)	0	0	0	0	0
1483	0	0	0	0	0	0	0	0	0	0
1515	0	0	0	0	0	0	0	0	0	0
1560	0	0	0	0	AA Batteries (4-pack)	0	0	0	0	0
1564	0	0	0	0	0	0	0	0	0	Google
1625	0	0	0	0	0	0	0	Bose SoundSport Headphones	0	0
1648	0	0	0	0	0	0	0	0	0	0
1661	0	0	0	0	0	0	0	0	0	Google
1673	0	0	0	0	0	0	0	0	0	0
1676	0	0	0	0	0	0	0	0	0	0
1713	0	0	0	34in Ultrawide Monitor	0	0	0	0	0	0
1766	0	0	0	0	0	0	0	0	0	0

5.1. Tabela z występowaniem produktów w danej transakcji

Następnie, na tak przygotowanej tabeli z danymi został wykonany algorytm APRIORI z biblioteki *efficient_apriori*. Stworzone zostały listy kombinacji wszystkich występujących ze sobą produktów. Sprawdzone zostały zestawy produktów, które zaufanie było większe niż 90%. Ostatnim elementem analizy było podzielenie zestawów produktów na takie, gdzie występowały produkty w liczbie 2, 3 i 4. Z tych zestawów zostało wyliczone TOP10 takich kombinacji (ten element został wykonany tylko ze względów przejrzystości).

Dla zestawu z dwoma elementami:

Index	0	1
144	('Lightning Charging Cable', 'iPhone')	1011
126	('Google Phone', 'USB-C Charging Cable')	997
158	('Wired Headphones', 'iPhone')	462
128	('Google Phone', 'Wired Headphones')	422
101	('Apple Airpods Headphones', 'iPhone')	373
153	('USB-C Charging Cable', 'Vareebadd Phone')	368
103	('Bose SoundSport Headphones', 'Google Phone')	228
154	('USB-C Charging Cable', 'Wired Headphones')	203
156	('Vareebadd Phone', 'Wired Headphones')	149
143	('Lightning Charging Cable', 'Wired Headphones')	129

5.2. Produkty najlepiej sprzedające się razem – zestaw dwóch produktów

Dla zestawu z trzema elementami:

Index	0	1
44	('Google Phone', 'USB-C Charging Cable', 'Wired Headphones')	87
50	('Lightning Charging Cable', 'Wired Headphones', 'iPhone')	63
26	('Apple Airpods Headphones', 'Lightning Charging Cable', 'iPhone')	47
33	('Bose SoundSport Headphones', 'Google Phone', 'USB-C Charging Cable')	35
51	('USB-C Charging Cable', 'Vareebadd Phone', 'Wired Headphones')	33
30	('Apple Airpods Headphones', 'Wired Headphones', 'iPhone')	27
34	('Bose SoundSport Headphones', 'Google Phone', 'Wired Headphones')	24
35	('Bose SoundSport Headphones', 'USB-C Charging Cable', 'Vareebadd Phone')	16
36	('Bose SoundSport Headphones', 'USB-C Charging Cable', 'Wired Headphones')	5
37	('Bose SoundSport Headphones', 'Vareebadd Phone', 'Wired Headphones')	5

5.3. Produkty najlepiej sprzedające się razem - zestaw trzech produktów

Dla zestawu z czterema elementami:

Index	0	1
8	('Apple Airpods Headphones', 'Lightning Charging Cable', 'Wired Headphones', 'iPhone')	4
9	('Bose SoundSport Headphones', 'Google Phone', 'USB-C Charging Cable', 'Wired Headphones')	3
10	('Bose SoundSport Headphones', 'USB-C Charging Cable', 'Vareebadd Phone', 'Wired Headphones')	2
0	('27in FHD Monitor', 'Google Phone', 'USB-C Charging Cable', 'Wired Headphones')	1
1	('34in Ultrawide Monitor', 'Bose SoundSport Headphones', 'Google Phone', 'USB-C Charging Cable')	1
2	('AA Batteries (4-pack)', 'Google Phone', 'USB-C Charging Cable', 'Wired Headphones')	1
3	('AA Batteries (4-pack)', 'Lightning Charging Cable', 'Wired Headphones', 'iPhone')	1
4	('Apple Airpods Headphones', 'Google Phone', 'Lightning Charging Cable', 'Wired Headphones')	1
5	('Apple Airpods Headphones', 'Google Phone', 'Lightning Charging Cable', 'iPhone')	1
6	('Apple Airpods Headphones', 'Google Phone', 'USB-C Charging Cable', 'Wired Headphones')	1

5.4. Produkty najlepiej sprzedające się razem - zestaw czterech produktów

Tak więc z algorytmu wynika, że jeśli klient dodał do swojego koszyka np. telefon to bardzo rozsądne będzie zarekomendowanie produktów z takich kategorii jak ładowarki, czy słuchawki. Co można zauważyć, to to, że klienci bardzo często kupowali razem telefon danej marki i pasującą do niego ładowarkę. Widać tu także zasadę, że rekomendacje powinny być tańsze niż produkt już oczekujący w koszyku, oraz ta rekomendacja powinna być w innej kategorii – nie powinno się rekomendować innego telefonu, gdzie klient już jeden taki zamówił – co jest raczej oczywiste, jednak wychodzi to również dzięki przeprowadzonej analizie.

Kod jest dostępny na platformie GitHub:
https://github.com/MPaluxch/Uslugi-Sieciowe-w-Biznesie/blob/main/MostComItems.py

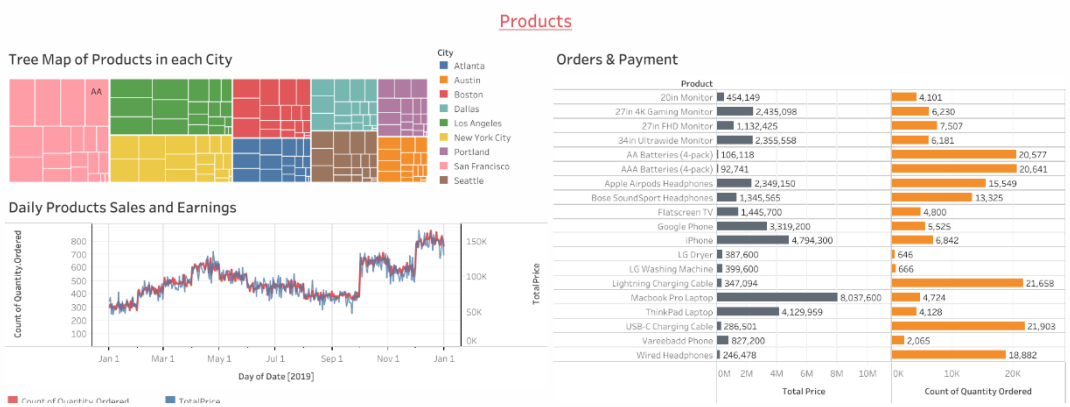
6. Przygotowanie wizualizacji w Tableau

Dzięki wizualizacjom możliwe jest zaprezentowanie danych w sposób przejrzysty, bardziej zrozumiały dla przeciętnego człowieka. Dobry dashboard jest niezbędny do skutecznej analizy danych i podejmowania decyzji. Pozwala ludziom szybko i łatwo zrozumieć zachodzące wzorce i dostrzec trendy – w przypadku danych tabelarycznych mogłoby być to utrudnione.

Tableau jest doskonałym narzędziem do wizualizacji danych. Oferuje bogatą listę wykresów, czy animacji, ale także pozwala na tworzenie interaktywnych dashboardów. Dzięki temu każdy może „wyklikać" interesujące go kategorie, przedziały czasowe itp.

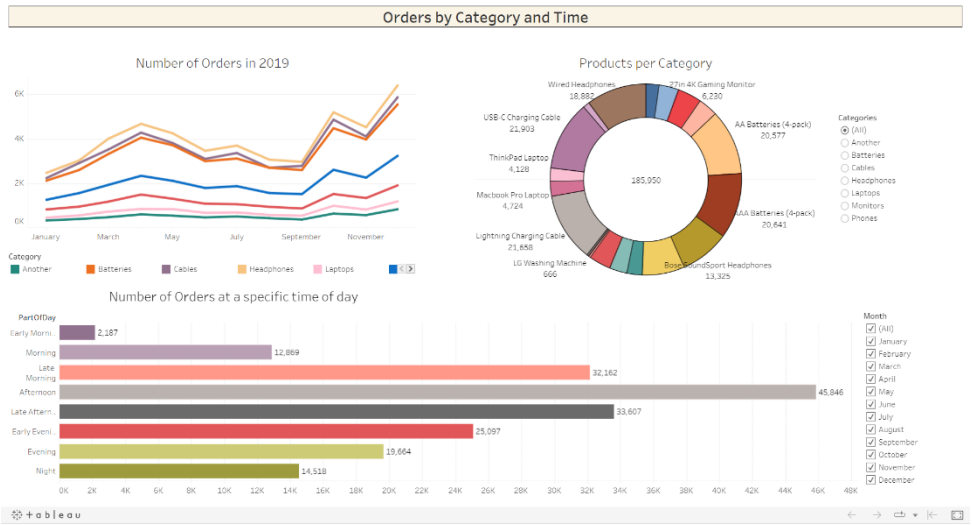
W ramach projektu zostały utworzone dwa dashboardy – jeden statyczny i jeden dynamiczny, zmieniający swoją strukturę, ze względu na wybranie kategorii produktu, lub przedziału czasowego.

Dashboard statyczny:



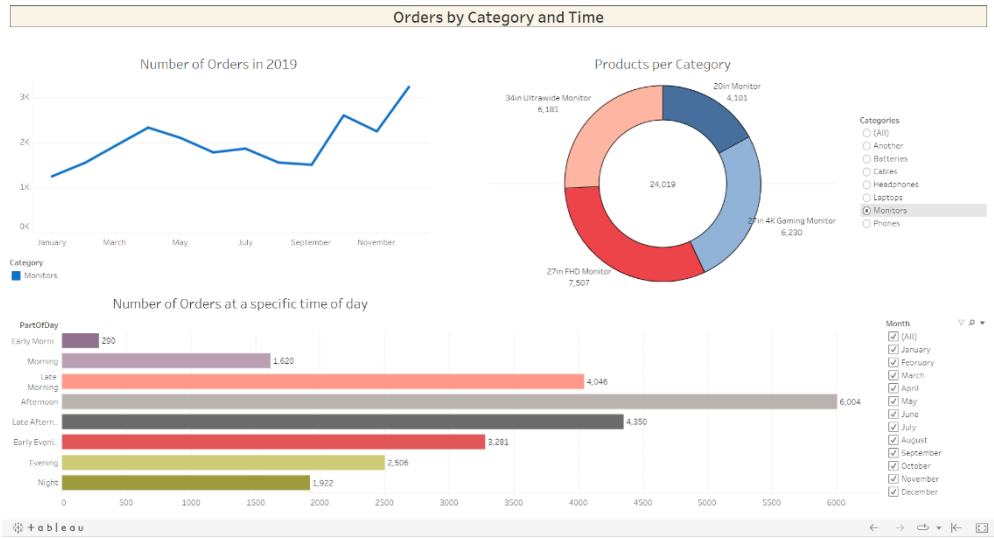
6.1. Dashbard statyczny

Dashboard dynamiczny:



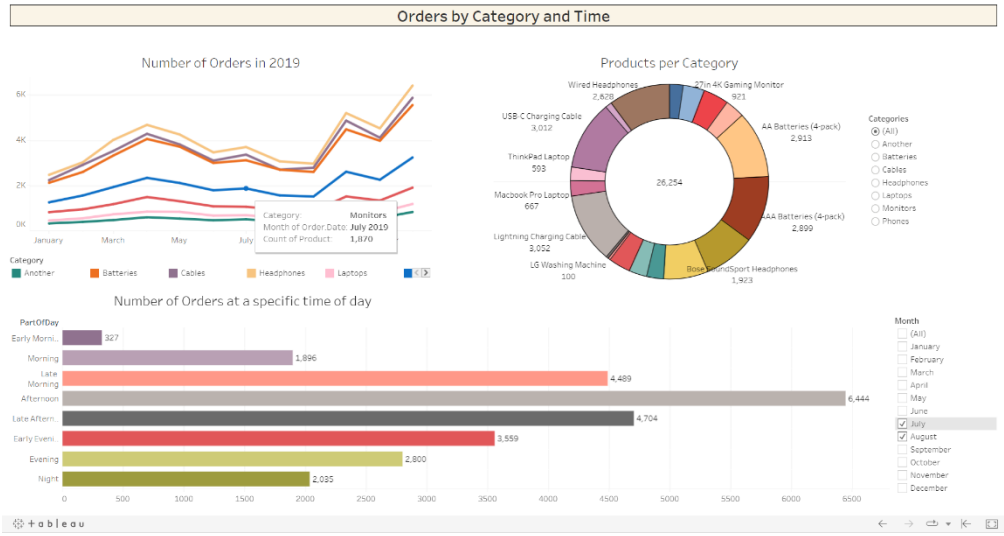
6.2. Dashboard dynamiczny - "część główna"

Dashboard dynamiczny po zmianie kategorii na „Monitory”:



6.3. Dashboard dynamiczny - "kategoria: Monitory"

Dashboard dynamiczny po zmianie przedziału czasowego na okres wakacyjny:



6.4. Dashboard dynamiczny - "Miesiące: Lipiec i Sierpień"

Tableau oferuje przechowywanie swoich prac na stronie Tableau Public, gdzie można również oglądać wizualizacje innych użytkowników z całego świata i czerpać inspiracje.

Wizualizacje można znaleźć na stronie:
https://public.tableau.com/app/profile/maciej.paluch

7. Źródła

[1] „How to Use the String to Date&Time Node”
Strona: KNIME Tutorials - YouTube
(Data dostępu: 20.04.2022)
Link: https://www.youtube.com/watch?v=hXecMGvrJFo

[2] „Partitional Clustering in R: The Essentials”
Strona: Datanovia.com
(Data dostępu: 21.04.2022)
Link: https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

[3] „Analiza koszykowa – Algorytm APRIORI”
Strona: WPDesk.pl
(Data dostępu: 02.05.2022)
Link: https://www.wpdesk.pl/blog/analiza-koszykowa-algorytm-apriori/

[4] „Efficient-Apriori”
Strona: Efficient-Apriori – Read the Docs.io
(Data dostępu: 03.05.2022)
Link: https://efficient-apriori.readthedocs.io/en/latest/