Marketing Mix Modeling

data processing in RStudio

2022-10-26

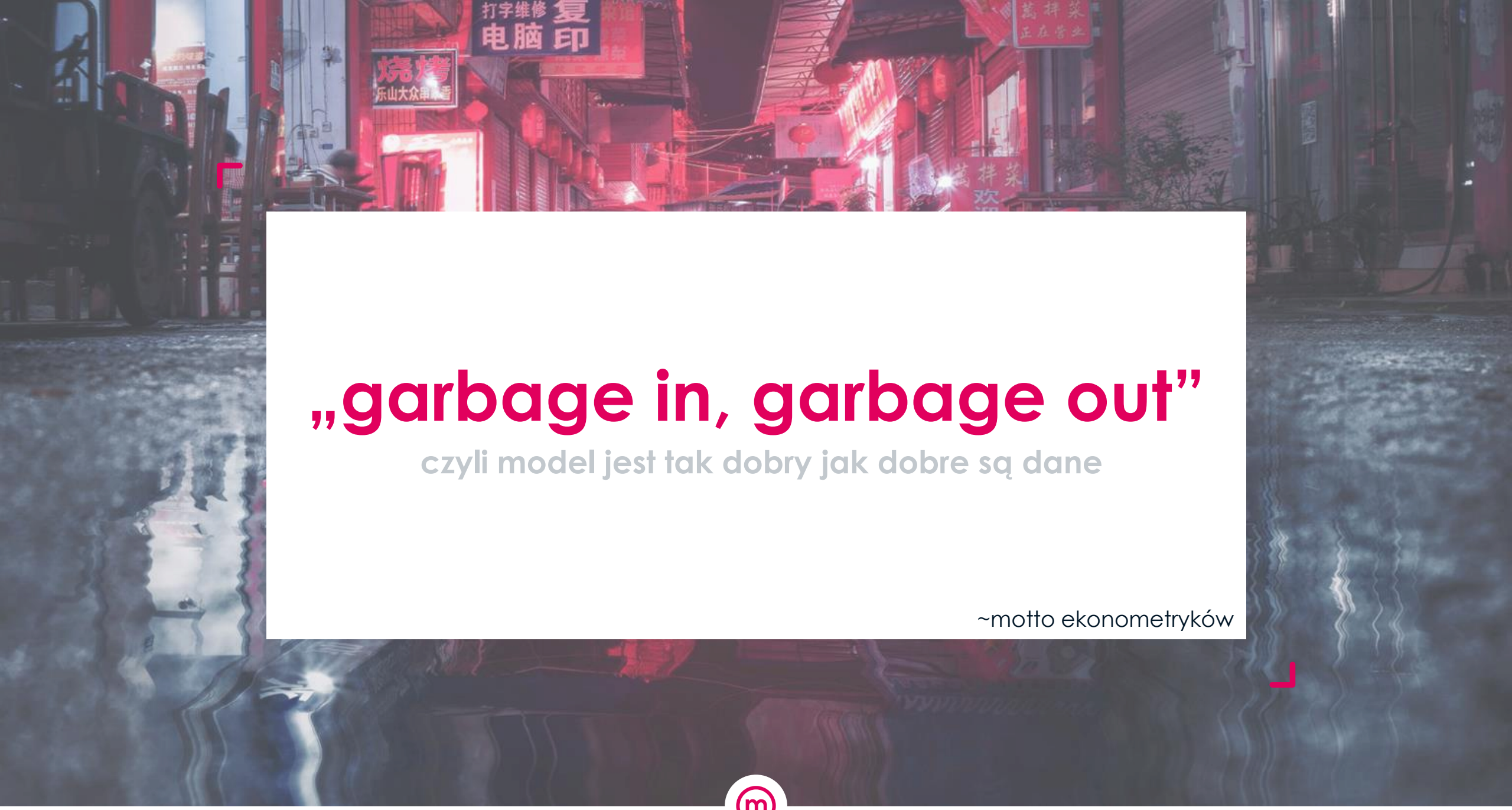
01 źródła danych i postać bazy

02 przetwarzanie danych nie-mediowych

03 przetwarzanie danych mediowych

01

źródła danych i postać bazy

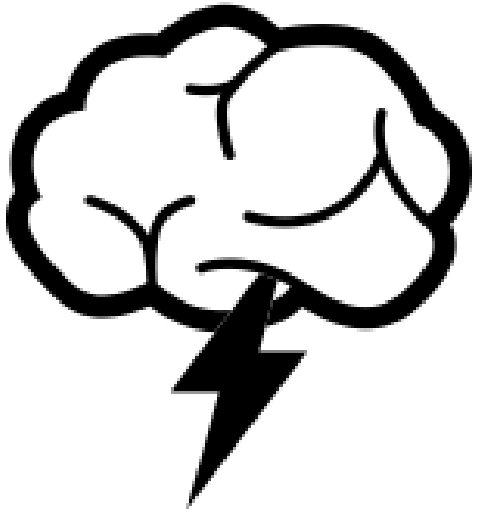


„garbage in, garbage out”

czyli model jest tak dobry jak dobre są dane

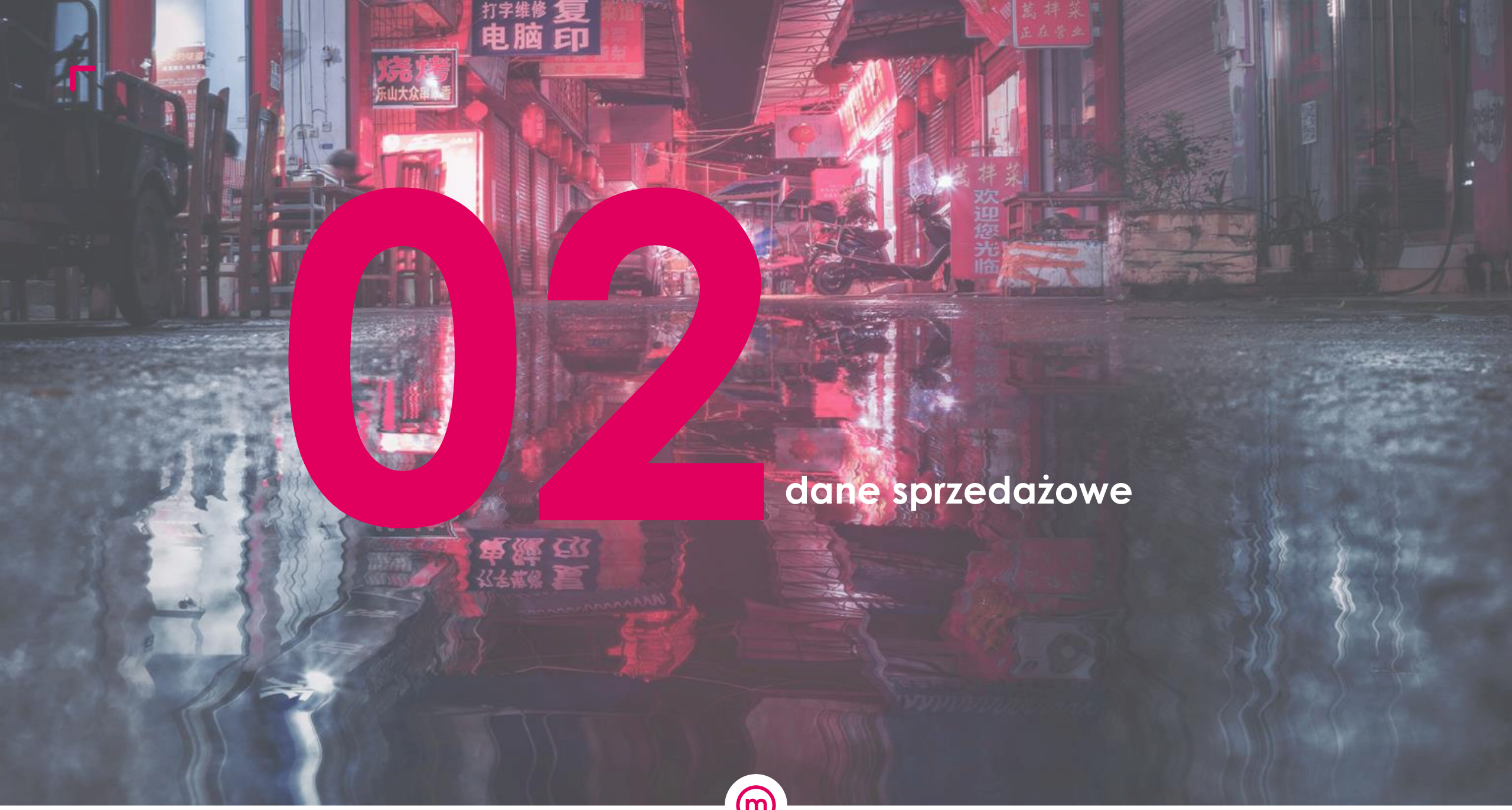
~motto ekonometryków

brainstorm – jakie czynniki wpływają na sprzedaż piwa butelkowanego



w skład bazy wchodzą dane z kilkunastu różnych źródeł

Obszar	Zmienne	Częstotliwość	Źródło
Dane sprzedażowe	Wartość, wolumen, dystrybucja, cena, liczba sklepów	Dz./tyg.	Klient (retail), Nielsen (FMCG), IQVIA (pharma), GfK (FMCG)
Aktywności tradeowe	Standy, faceing, płachty, katalogi, sampling, ulotki	Tyg./mies.	Klient, Nielsen, FOCUS
Aktywności mediowe	Telewizja, radio, social, search, display, VOD	Dz./tyg.	Nielsen, Google, Facebook, TikTok, Radio Track, inni dostawcy
Dane ekonomiczne	CPI, Konsumpcja, urodzenia, CCI	Mies./kw.	GUS, OECD, strony rządowe
Święta i sezonowość	Święta, dni handlowe, cykl sezonowości	Dz./tyg.	Kalendarz, strony rządowe
Pogoda	Opady, temperatura, nasłonecznienie	Dz./tyg.	IMGW, strony rządowe, Dark Sky
Czynniki zewnętrzne	Trendy konsumenckie, COVID i inne czarne łabędzie	Dz./tyg.	Google Trends, Google Mobility, GfK, agencje badawcze, dane rządowe



02 dane sprzedażowe

case study: marka na rynku FMCG posiadająca 3 SKU*

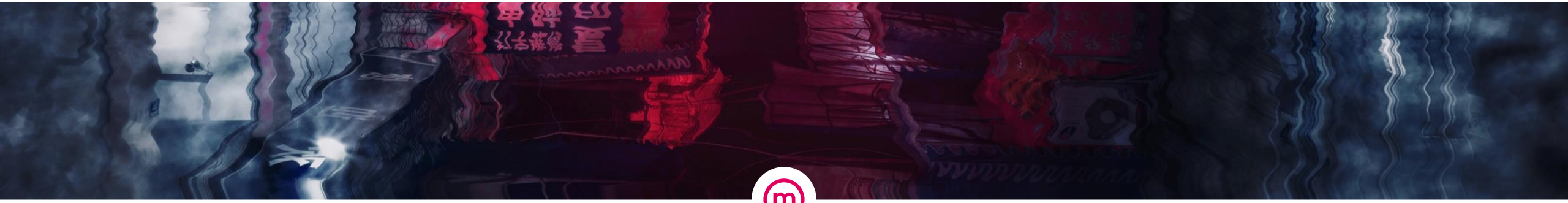
Oferowane przez markę produkty (SKU):

- Piwo puszka – 0.5l
- Piwo butelka – 0.5l
- Piwo butelka – 0.33l

Opis case study:

- Browar zakupił jeden model ekonometryczny
- Sprzedaż w jednym łańcuchu dyskontów (50 sklepów)
- Dane zagregowane pomiędzy sklepami (szereg czasowy) lub w każdym sklepie osobno (model panelowy)
- Okres modelowany: 2 lata (104 tygodnie)
- Model został zakupiony przez producenta/browar (a nie dyskont):
  - Browar sprzedaje dyskontowi piwo po stałej cenie, czyli operuje na stałej marży. Cena do konsumenta jest ustalana przez dyskont (może on ale nie musi kierować się rekomendacjami cenowymi browaru).

*stock keeping unit



case study: co powinno być zmienną modelowaną by dostarczyć klientowi (browarowi) jak najbardziej wartościowe wnioski?

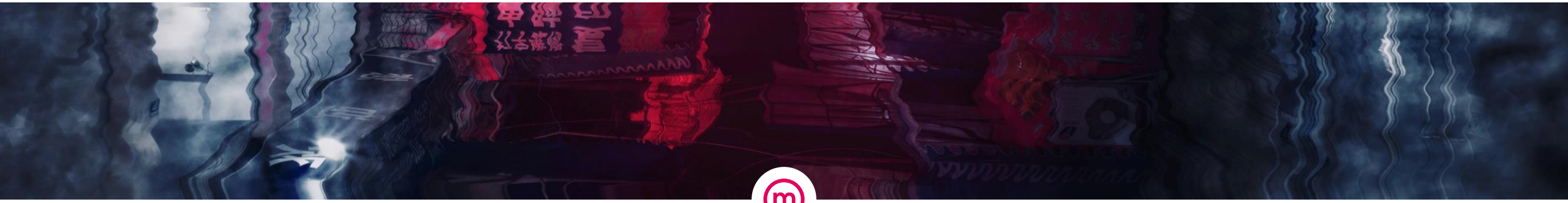
Możliwe kanały transakcyjne:

- Browar -> dyskont
- Dyskont -> konsument

Możliwe metryki:

- Liczba dokonanych transakcji (paragonów)
- Liczba sprzedanych butelek
- Litry sprzedanego piwa
- Wartość w PLN sprzedanego piwa

Odp: wolumen sprzedaży w sklepach ujęty w litrach



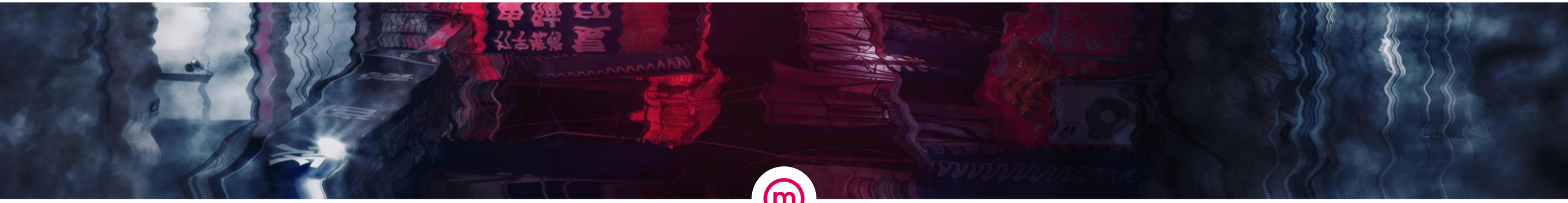
wartość i wolumen

wybór zmiennej modelowanej:

- metryka pozwalająca na wyciąganie wniosków wartościowych dla producenta piwa (a nie dla dyskontu)
- metryka na zmienność której bezpośredni wpływ mają zachowania konsumentów (pamiętajmy, że głównym celem projektu MMM jest zbadanie efektywności mediów. Media w zamyśle oddziałują na konsumentów końcowych a nie na właściciela dyskontu.)
- metryka pozwalająca zagregować wszystkie SKU do jednej zmiennej

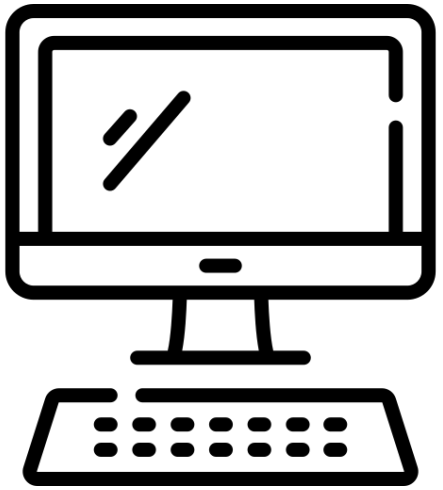
przygotowanie zmiennych

- wartości i wolumeny (przy wybraniu uniwersalnej metryki takiej jak wolumen w litrach lub wartość metryki) mogą być agregowane (sumowane)
- zmienne przygotowujemy na różnych poziomach agregacji (**SKU, podtyp, cała marka**), natomiast dążymy do uwzględniania w modelu jak najbardziej szczegółowych zmiennych (by dostarczyć szczegółowe wnioski i uniknąć paradoksu agregacji, o którym będzie mowa w dalszej części prezentacji). Ograniczeniami w przypadku dużej szczegółowości zmiennych są:
  - współliniowość
  - liczba stopni swobody

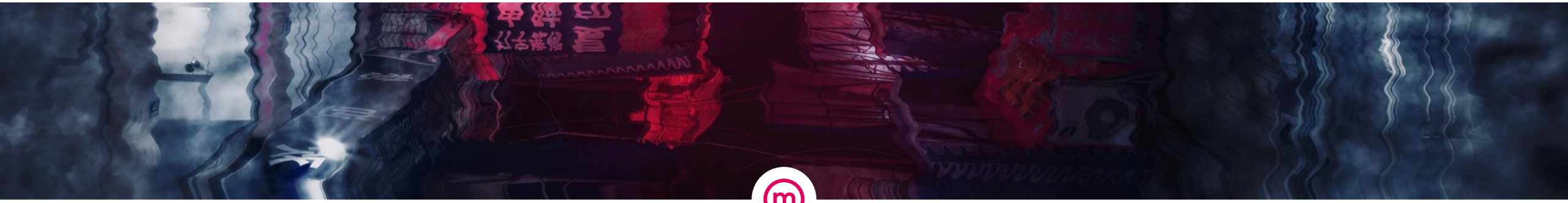


zadanie 1

(1 pkt za sposób z dodawaniem kolumn, 2 pkt za sposób, który zadziałałby przy dowolnej liczbie kolumn VO_)



- za pomocą biblioteki ***readxl*** wczytaj bazę ***data_processing.xlsx***
- zapoznaj się z bazą ***data.df***
- stwórz finalną zmienną objaśnianą, nazwij ją ***ZM_MOD***
  - Zmienna ***ZM_MOD*** jest sumą wolumenów wszystkich SKU piwa ***(VO_)***
  - Zlogarytmuj zmienną ***ZM_MOD*** (logarytm naturalny)
- wynikiem zadania 1 powinna być niezależna ramka danych



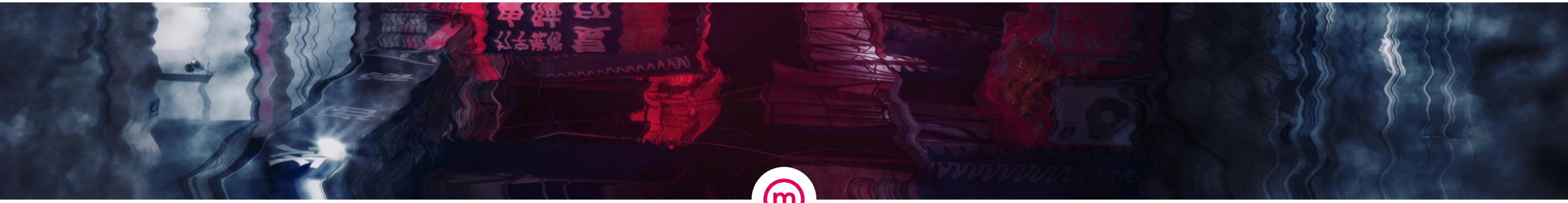
cena długo- i krótko-okresowa

cenę półkową należy rozbić na cenę długookresową oraz obniżki cenowe:

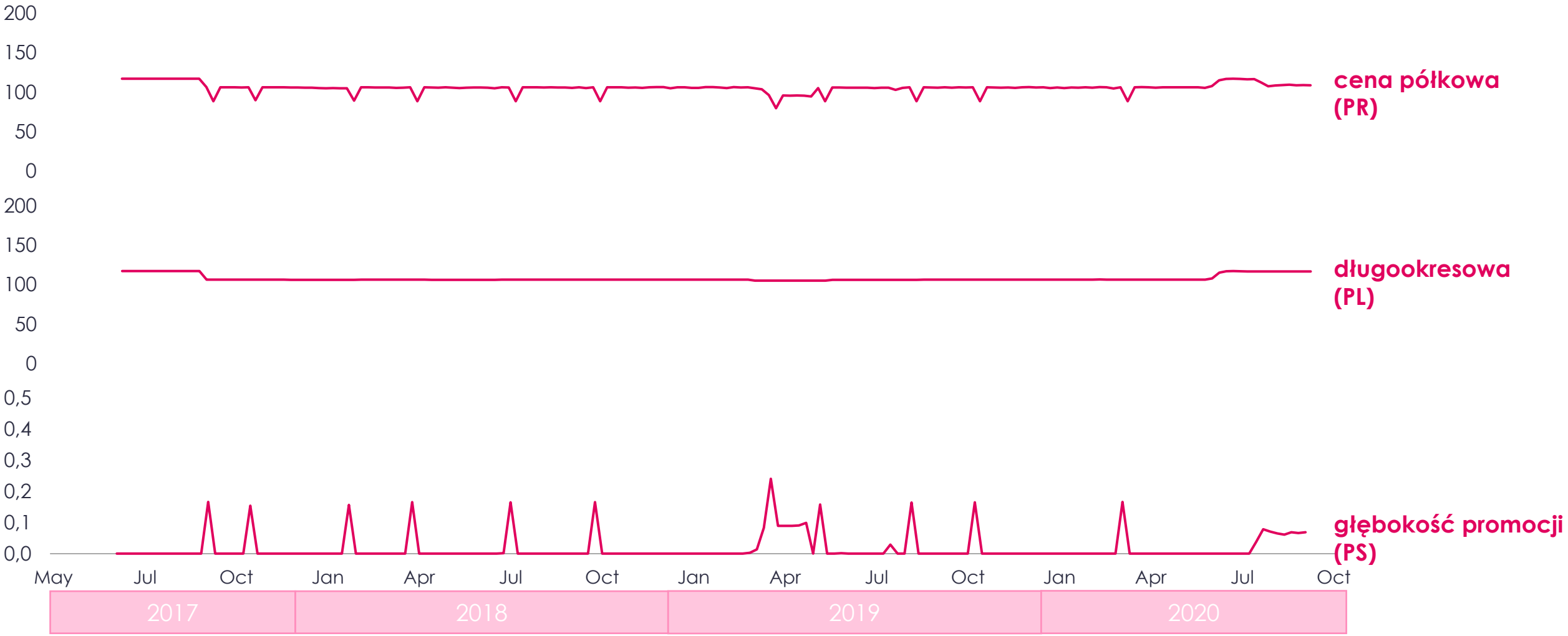
- dla każdego SKU dzielimy wartość przez wolumen
- otrzymana cena to cena półkowa (PR) – zawiera ona w sobie jednocześnie cenę długookresową (PL - względnie stały poziom) oraz obniżki cenowe (PS - okresowe promocje)
- identyfikujemy cenę długookresową
- od ceny długookresowej odejmujemy cenę półkową. wynik odejmowania dzielimy przez cenę długookresową
- wynikiem tego dzielenia jest „głębokość promocji" czyli okresowe obniżki cenowe

imputacja braków danych:

- braki danych dla ceny powstają gdy wolumen sprzedaży (mianownik) jest zerowy
- cenę długookresową najlepiej imputować medianą/wartością maksymalną z pozostałych dni, a cenę krótkookresową – zerem (slajd 16 doskonale obrazuje dlaczego).

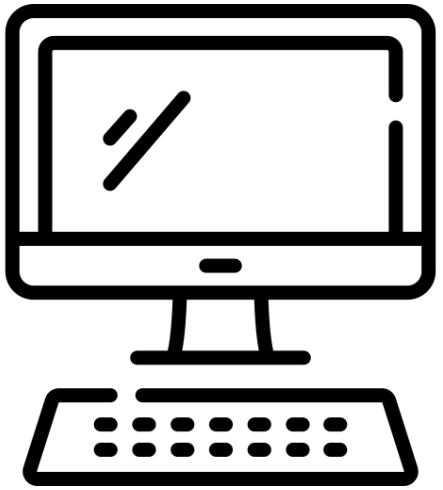


cena długo- i krótko-okresowa

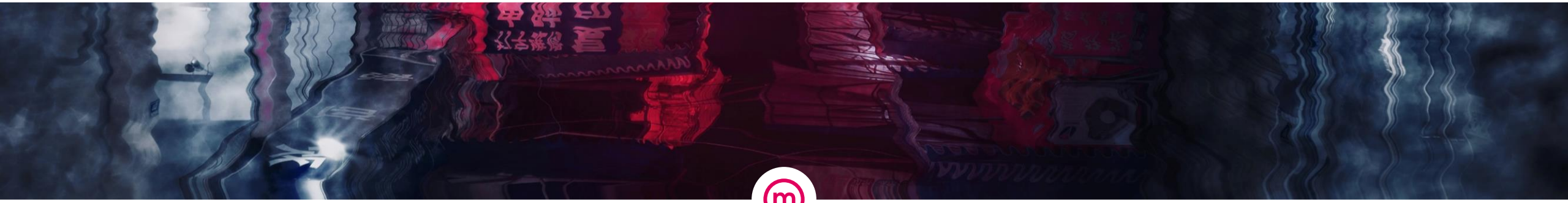


zadanie 2

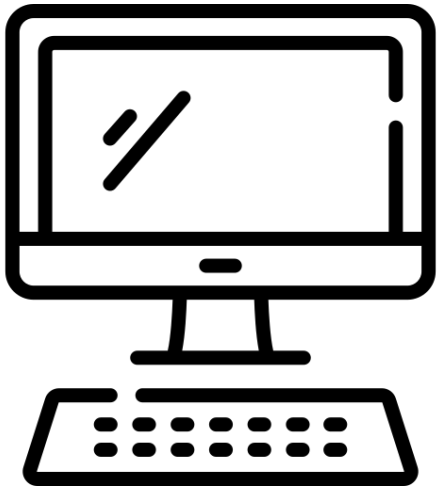
(1 pkt za rozwiązanie manualne lub 2 pkt za uniwersalne)



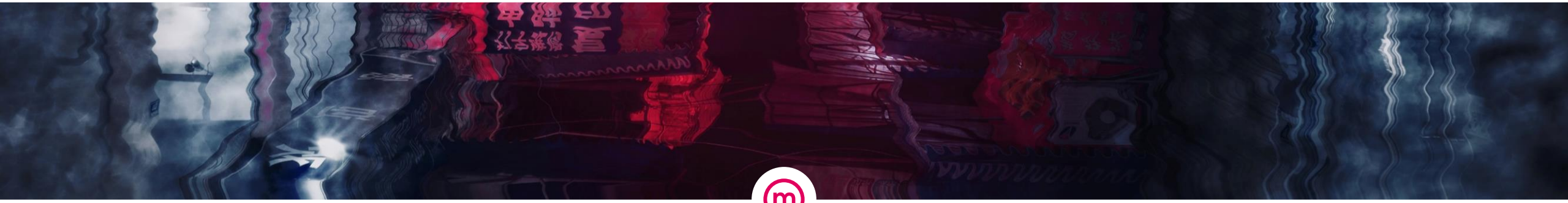
- stwórz zmienne zawierające cenę półkową wszystkich SKU piwa
- przykładowa nazwa zmiennej zawierającej cenę: ***PR_BEER_SKU1***
- ***UWAGA:*** Wykorzystaj bibliotekę ***tidyverse***, by napisany kod był uniwersalny tzn. by mógł w niezmienionej formie posłużyć do stworzenia nie 3, lecz dowolnej liczby zmiennych ***PR_***
  - ***TIP1:*** Wykorzystaj funkcję pivot_longer() by móc operować na wszystkich SKU jednocześnie
  - ***TIP2:*** Wykorzystaj funkcję substr() by wyodrębnić z nazwy zmiennej jej metrykę (VA i VO) i SKU
- wynikiem zadania 2 powinna być niezależna ramka danych



zadanie 3
(2 pkt)



- obejrzyj zmienne PR stworzone w poprzednim zadaniu na wykresie.
- na podstawie PR i PL stwórz zmienne zawierające cenę krótkookresową (PS).
  - ***TIP1:*** Zmienne PS, jako głębokość promocji powinny być wyrażone w procentach (wartość 30% mówi, że w danym tygodniu cena PR spadła o 30% w stosunku do ceny PL).
- obejrzyj gotowe zmienne na wykresach.
- wynikiem zadania 3 powinna być niezależna ramka danych

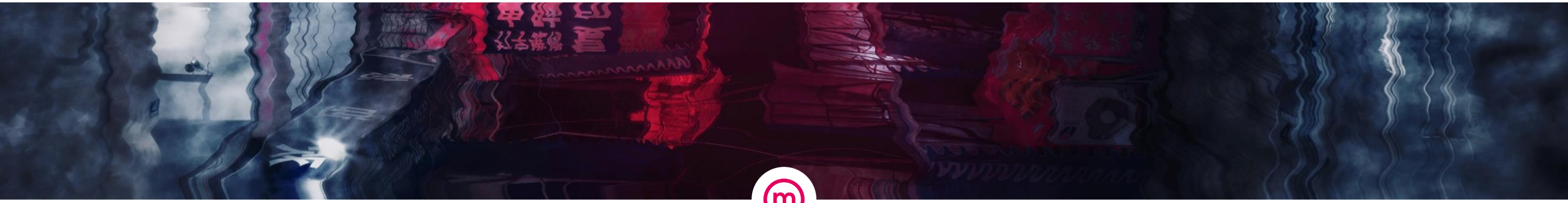


paradoks agregacji cen

uwzględnianie cen na poziomie całej marki może prowadzić do tzw. *„price aggregation paradox”*:

- czyli sytuacji, w której cena za litr każdego SKU rośnie, natomiast cena za litr całej marki spada
- dzieje się tak, ze względu na zmiany nawyków konsumentów, którzy w przypadku nieproporcjonalnych zmian cen zaczynają kupować inne produkty tej samej marki
- przykład:

SKU	Cena za litr	Tyg. wolumen	Cena za litr na całej marce	Cena za litr	Tyg. wolumen	Cena za litr na całej marce
But. – 0.5l	4.0 PLN	1000 l	3.83 PLN	4.4 PLN	200 l	2.92 PLN
Puszka – 0.5l	3.0 PLN	200 l		3.3 PLN	1000 l	



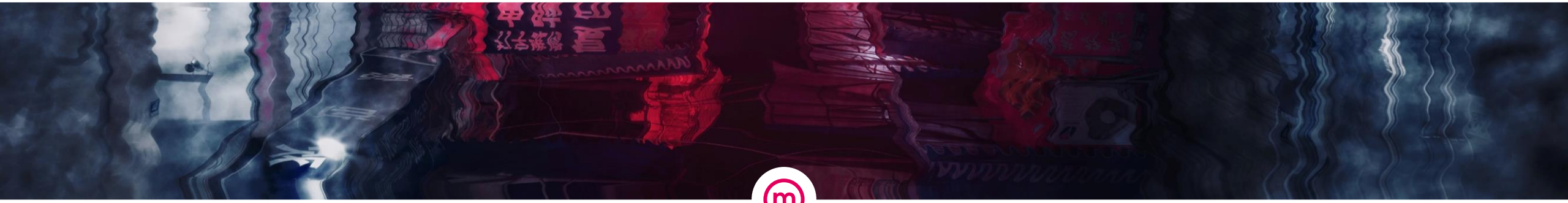
dystrybucja numeryczna i ważona na przykładzie pojedynczego SKU

numeryczna

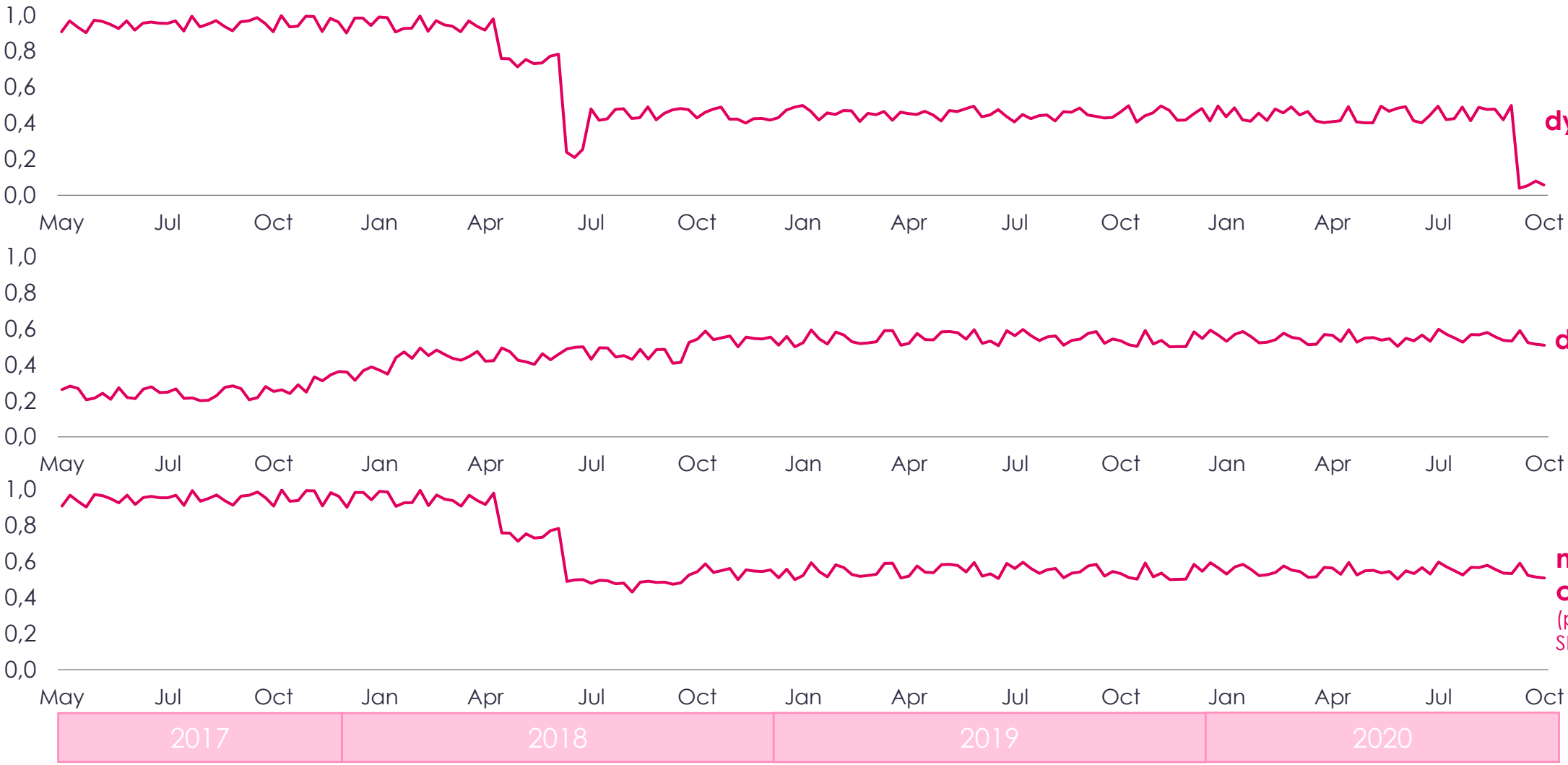
- odsetek sklepów w których dany SKU był dostępny (przykład: jeżeli dla SKU-1 dystrybucja w pierwszym tygodniu 2022 roku wynosi 68%, znaczy to, że ten produkt był dostępny w 68% sklepów).
- zawiera się w przedziale od 0 do 1
- traktuje każdy sklep jednakowo

ważona

- średnia dostępność SKU ważona sprzedażą danego SKU w każdym z rozważanych sklepach (przykład: jeżeli mamy 3 sklepy o rocznym obrocie 1mln PLN i czwarty sklep o rocznym obrocie 2mln, a produkt był w danym tygodniu dostępny tylko w tych trzech mniejszych sklepach, to **dystrybucja numeryczna wyniesie 75% ale ważona tylko 60%**).
- zawiera się w przedziale od 0 do 1
- lepiej odzwierciedla rzeczywistość, ponieważ uwzględnia różnice w wielkości sklepu i klienteli
- wymaga dostępności danych z podziałem na sklep



dystrybucja SKU versus całej marki

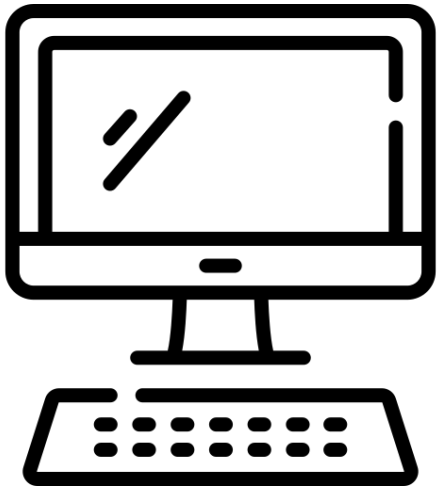


dystrybucja SKU1

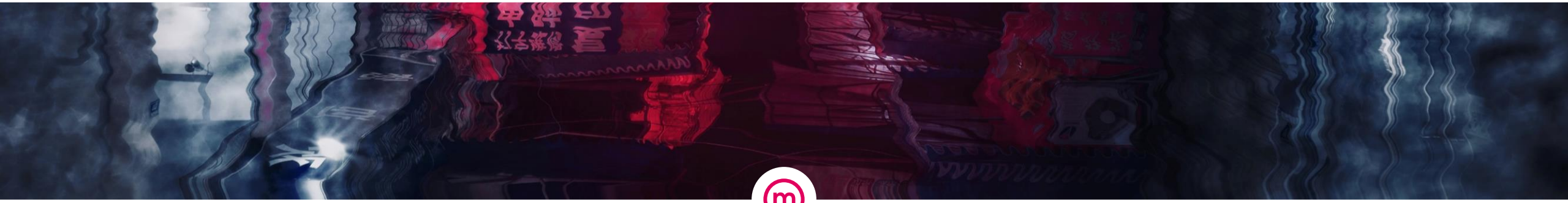
dystrybucja SKU2

minimalna dyst. całej marki
(przybliżona jako MAX z SKU1 i SKU2)

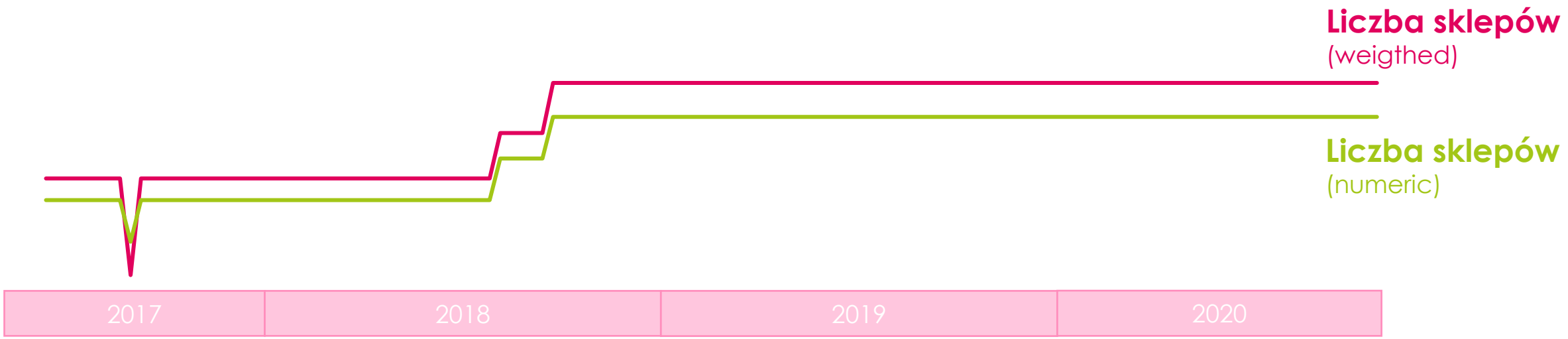
zadanie 4
(1 pkt)



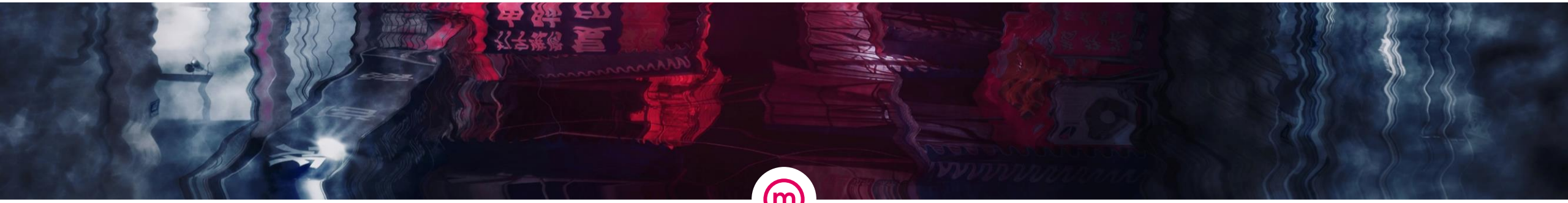
- na podstawie zmiennych dotyczących dystrybucji SKU stwórz zmienną mówiącą o przybliżonej dystrybucji całej marki ***DW_BEER***
  - ***TIP1:*** wykorzystaj funkcję ***pmax()***
- wynikiem zadania 4 powinna być niezależna ramka danych



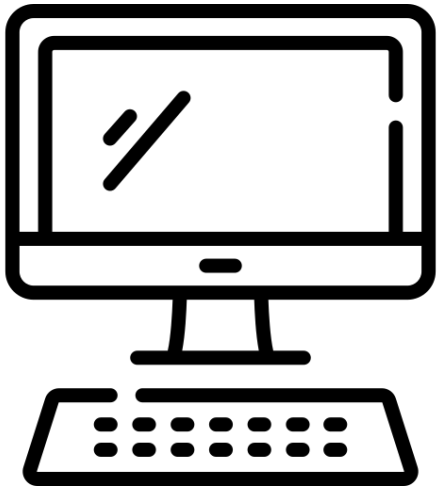
Jak zaadresować zmieniającą się liczbę sklepów gdy każdy sklep jest innej „wielkości”?



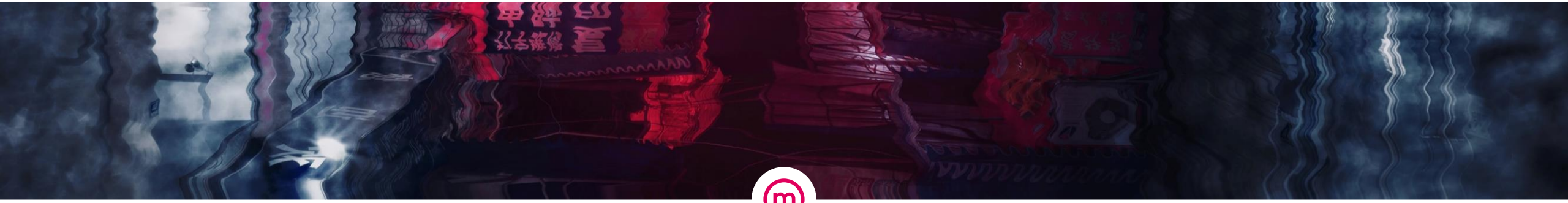
Preferowanym podejściem jest przeważenie liczby sklepów przez ich przeciętny tygodniowy obrót w okresie modelowanym. Dzięki temu odzwierciedlimy fakt, że zamknięcia/otwarcie „dużych” sklepów silniej oddziałują na zmiany wartości sprzedaży piwa w całej Polsce niż zamknięcia/otwarcia „małych sklepów”. Przykładowo, na wykresie powyżej w 2017 jeden sklep został zamknięty tymczasowo – widzimy, że był to ponadprzeciętnie „duży” sklep. Relatywny spadek w poziomie zmiennej ważonej (czerwonej) jest głębszy niż spadek w przypadku surowej numerycznej liczby sklepów.

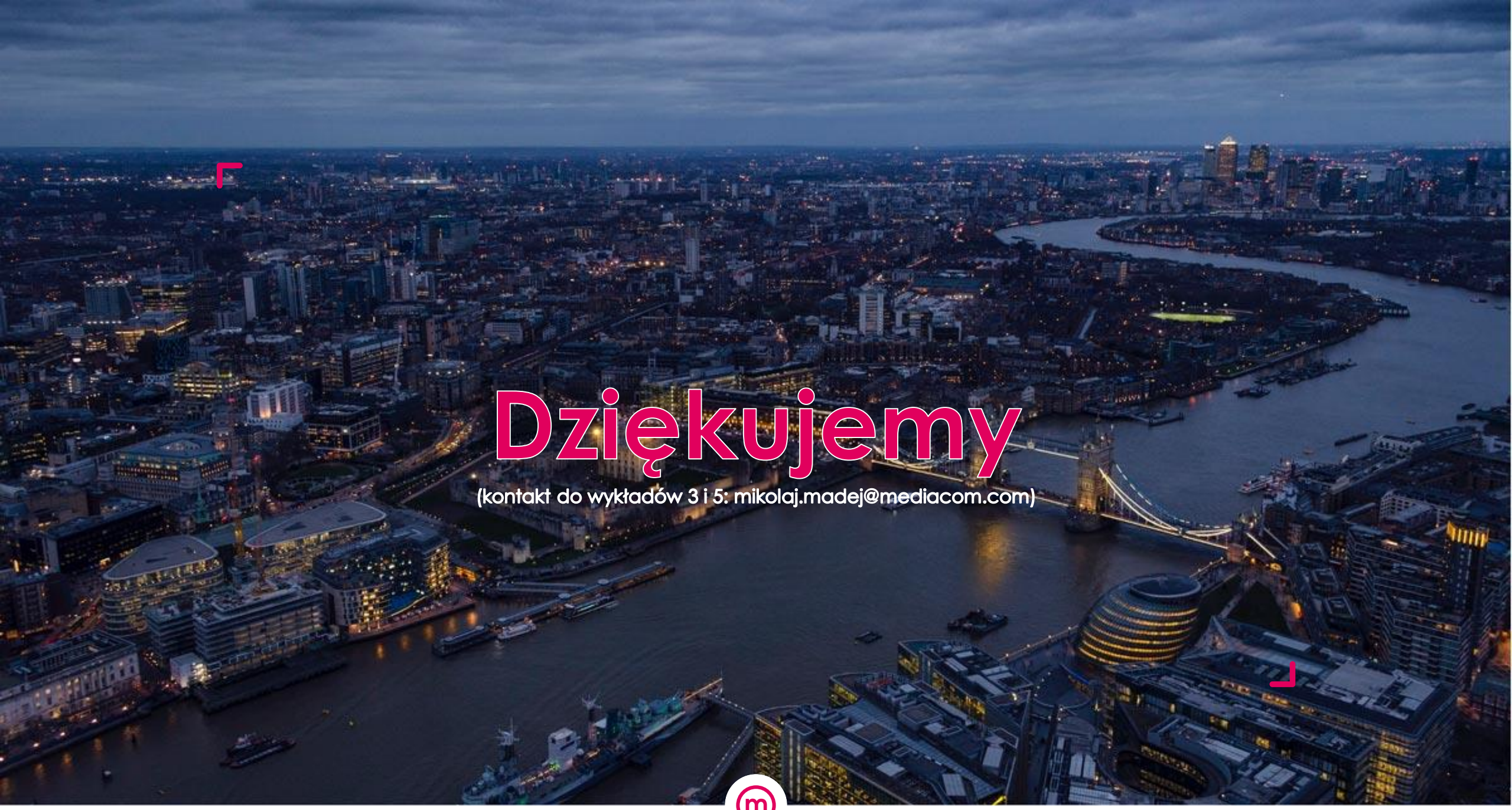


zadanie 5
(2 pkt)



- przygotuj zmienną odzwierciedlającą ważoną liczbę sklepów, korzystając ze zmiennych SHOP_1, SHOP_2 i SHOP_3
- wynikiem zadania 5 powinna być niezależna ramka danych




Dziękujemy
(kontakt do wykładów 3 i 5: mikolaj.madej@mediacom.com)