



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





„obyś modelował ciekawe czasy”

~adaptacja chińskiego przekleństwa



brainstorm – jakie czynniki wpływają na sprzedaż piwa butelkowanego



┐ w skład bazy wchodzi dane z kilkunastu różnych źródeł

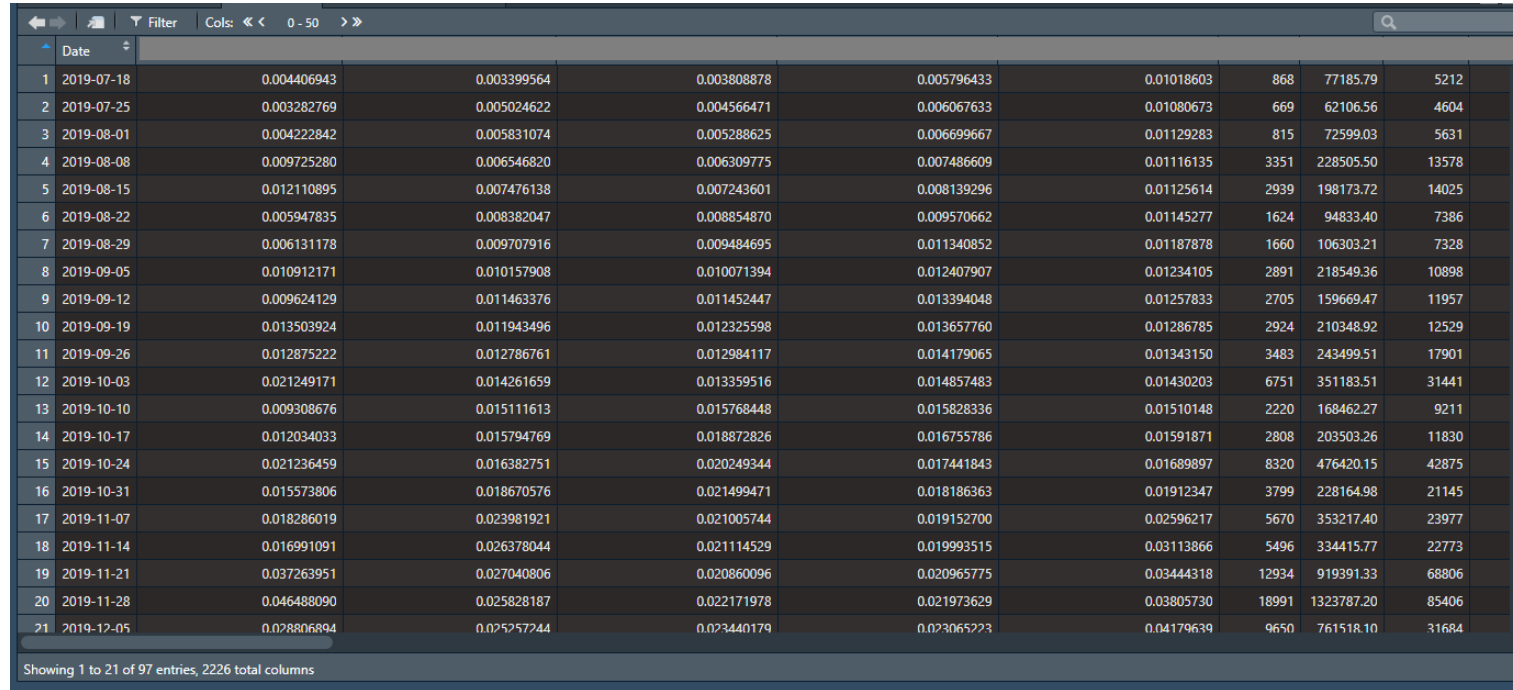
Obszar	Zmienne	Częstotliwość	Źródło
Dane sprzedażowe	Wartość, wolumen, dystrybucja, cena	Dz./tyg.	Klient (retail), Nielsen (FMCG), IQVIA (pharma), GfK (FMCG)
Aktywności mediowe	Telewizja, radio, social, search, display, VOD	Dz./tyg.	Nielsen, Google, Facebook, TikTok, Radio Track, inni dostawcy
Aktywności tradeowe	Standy, faceing, płachty, katalogi, sampling, ulotki	Tyg./mies.	Klient, Nielsen, FOCUS
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,	Dz./tyg.	Google Trends, Google Mobility, GfK, agencje badawcze, dane rządowe



↳ baza danych do modelowania składa się z kilkuset lub kilku tysięcy zmiennych

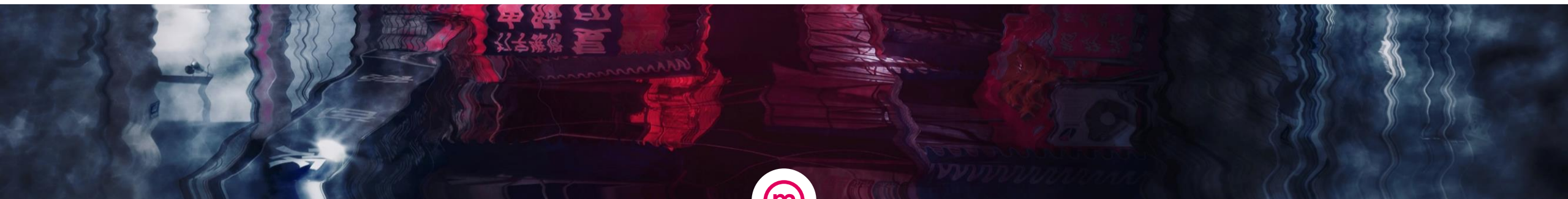
zmienne

obserwacje



	Date								
1	2019-07-18	0.004406943	0.003399564	0.003808878	0.005796433	0.01018603	868	77185.79	5212
2	2019-07-25	0.003282769	0.005024622	0.004566471	0.006067633	0.01080673	669	62106.56	4604
3	2019-08-01	0.004222842	0.005831074	0.005288625	0.006699667	0.01129283	815	72599.03	5631
4	2019-08-08	0.009725280	0.006546820	0.006309775	0.007486609	0.01116135	3351	228505.50	13578
5	2019-08-15	0.012110895	0.007476138	0.007243601	0.008139296	0.01125614	2939	198173.72	14025
6	2019-08-22	0.005947835	0.008382047	0.008854870	0.009570662	0.01145277	1624	94833.40	7386
7	2019-08-29	0.006131178	0.009707916	0.009484695	0.011340852	0.01187878	1660	106303.21	7328
8	2019-09-05	0.010912171	0.010157908	0.010071394	0.012407907	0.01234105	2891	218549.36	10898
9	2019-09-12	0.009624129	0.011463376	0.011452447	0.013394048	0.01257833	2705	159669.47	11957
10	2019-09-19	0.013503924	0.011943496	0.012325598	0.013657760	0.01286785	2924	210348.92	12529
11	2019-09-26	0.012875222	0.012786761	0.012984117	0.014179065	0.01343150	3483	243499.51	17901
12	2019-10-03	0.021249171	0.014261659	0.013359516	0.014857483	0.01430203	6751	351183.51	31441
13	2019-10-10	0.009308676	0.015111613	0.015768448	0.015828336	0.01510148	2220	168462.27	9211
14	2019-10-17	0.012034033	0.015794769	0.018872826	0.016755786	0.01591871	2808	203503.26	11830
15	2019-10-24	0.021236459	0.016382751	0.020249344	0.017441843	0.01689897	8320	476420.15	42875
16	2019-10-31	0.015573806	0.018670576	0.021499471	0.018186363	0.01912347	3799	228164.98	21145
17	2019-11-07	0.018286019	0.023981921	0.021005744	0.019152700	0.02596217	5670	353217.40	23977
18	2019-11-14	0.016991091	0.026378044	0.021114529	0.019993515	0.03113866	5496	334415.77	22773
19	2019-11-21	0.037263951	0.027040806	0.020860096	0.020965775	0.03444318	12934	919391.33	68806
20	2019-11-28	0.046488090	0.025828187	0.022171978	0.021973629	0.03805730	18991	1323787.20	85406
21	2019-12-05	0.028806894	0.025257744	0.023440179	0.023065223	0.04179639	9650	761518.10	31684

Showing 1 to 21 of 97 entries, 2226 total columns





02

dane nie-mediowe

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



browar



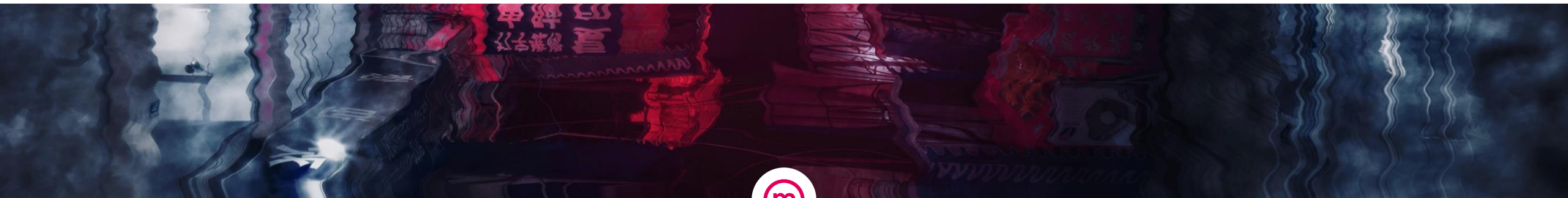
dyskont



konsument

Opis case study:

- Marka zakupiła jeden model ekonometryczny
- Sprzedaż w jednym łańcuchu dyskontów (50 sklepów)
- Dane zagregowane pomiędzy sklepami (szereg czasowy)
- 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).



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

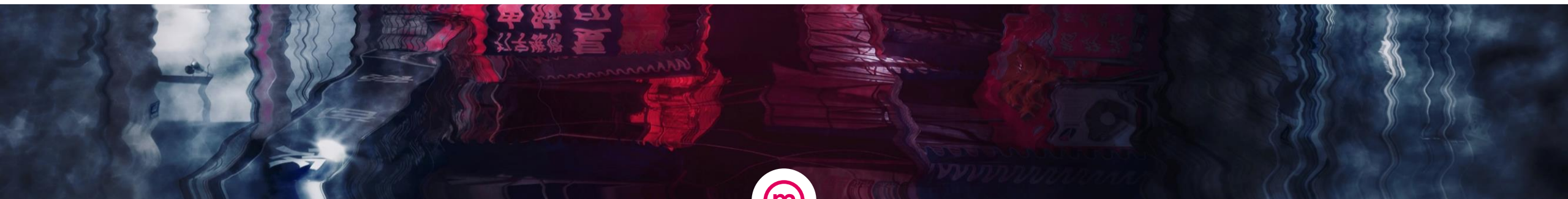
Możliwe kanały transakcyjne:

- Browar -> dyskont
- Dyskont -> konsument

Możliwe metryki:

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

**Odp: wolumen sprzedaży
w sklepach ujęty w litrach**



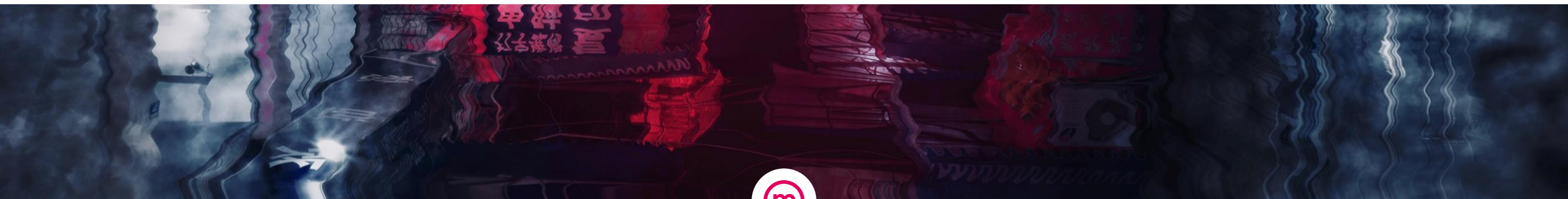
wartość i wolumen

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

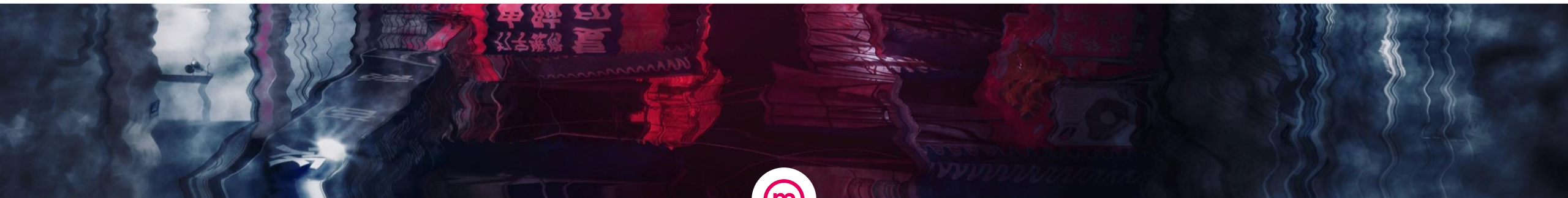
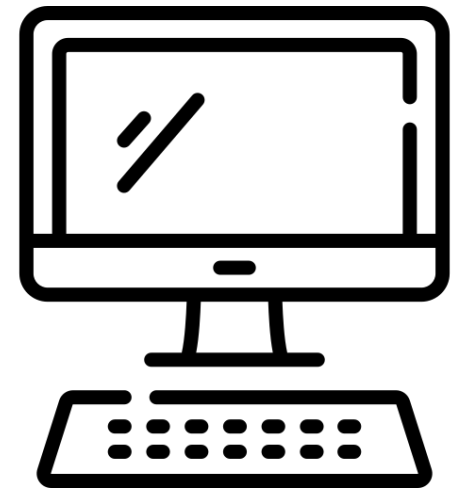
przygotowanie zmiennych

- 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 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



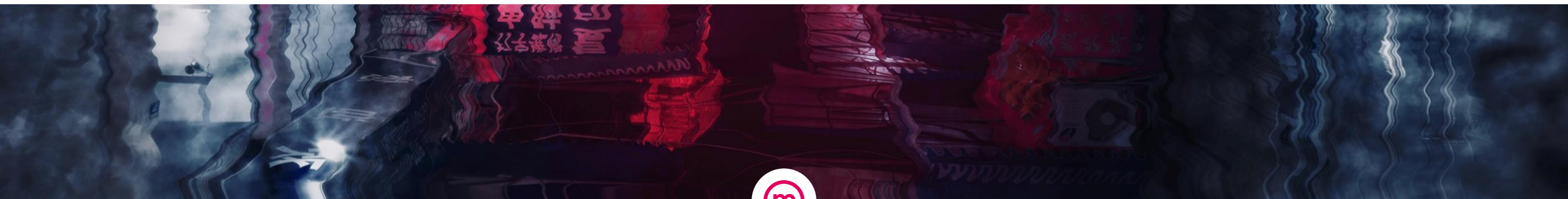
cena długo- i krótko-okresowa

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:

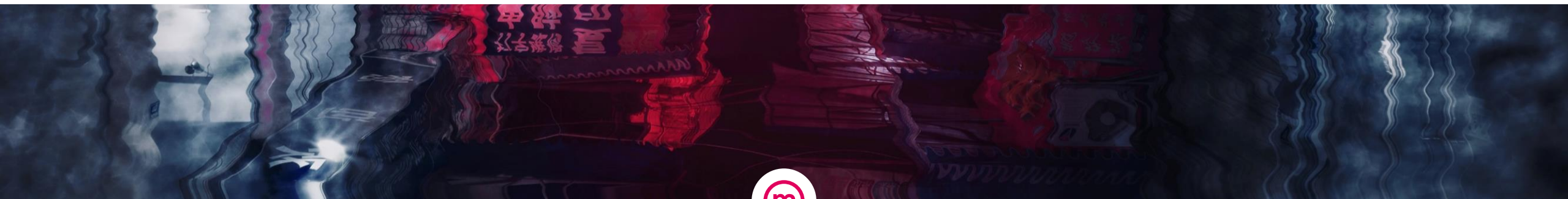
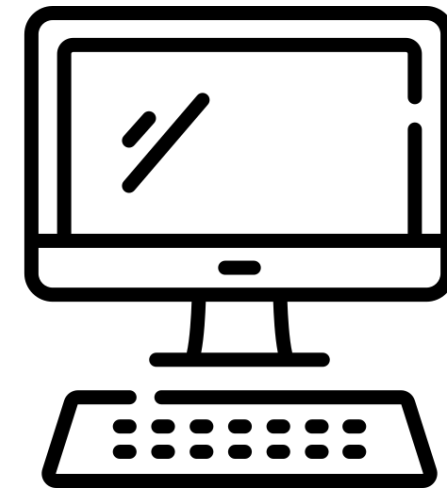
- braki danych dla ceny powstają gdy wolumen sprzedaży (mianownik) jest zerowy
- cenę długookresową najlepiej imputować medianą/wartością maksymalną z pozostałych dni, a cenę krótkookresową – zerem (slajd 16 doskonale obrazuje dlaczego).



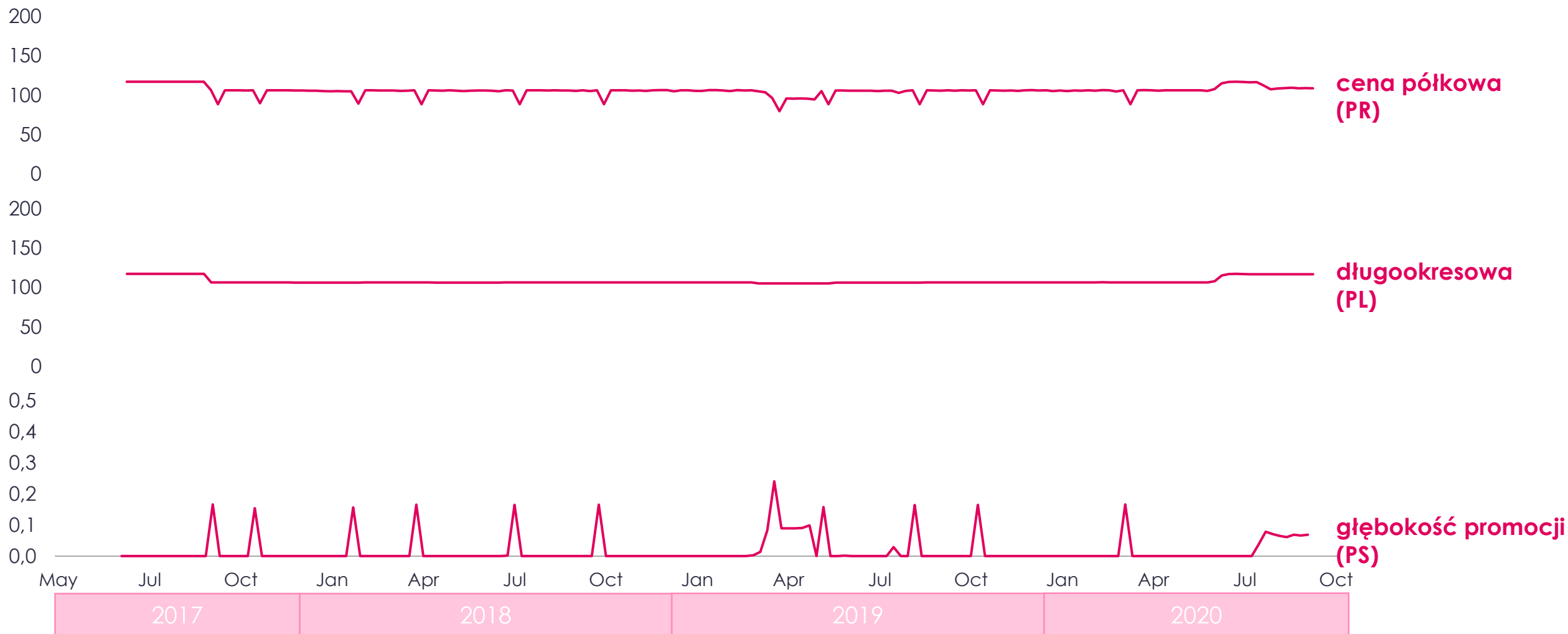
zadanie 2

(1 pkt za rozwiązanie manualne lub 3 pkt za uniwersalne)

- stwórz zmienne zawierające cenę półkową wszystkich SKU piwa
- przykładowa nazwa zmiennej zawierającej cenę: **PR_MILK_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

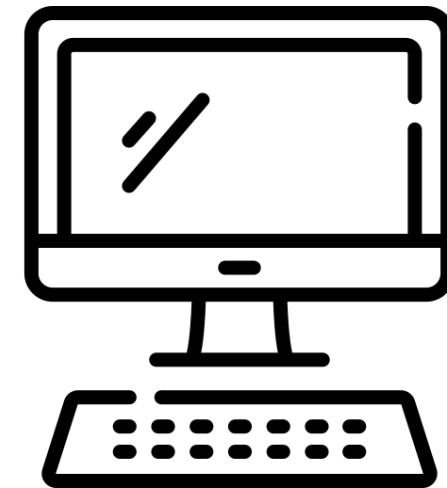


cena długo- i krótko-okresowa

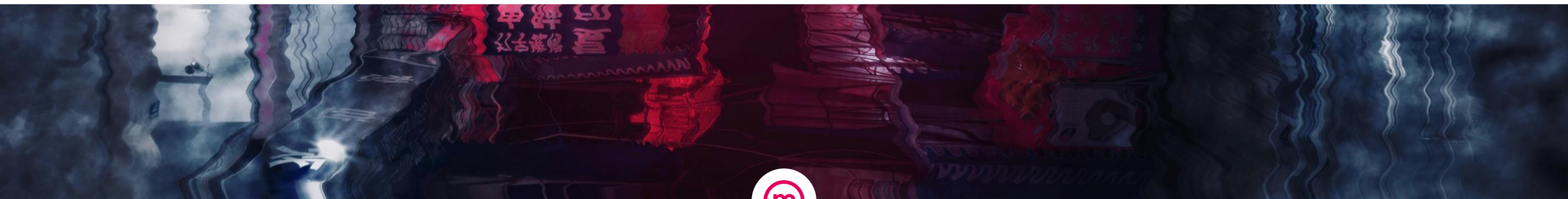


zadanie 3

(2 pkt)

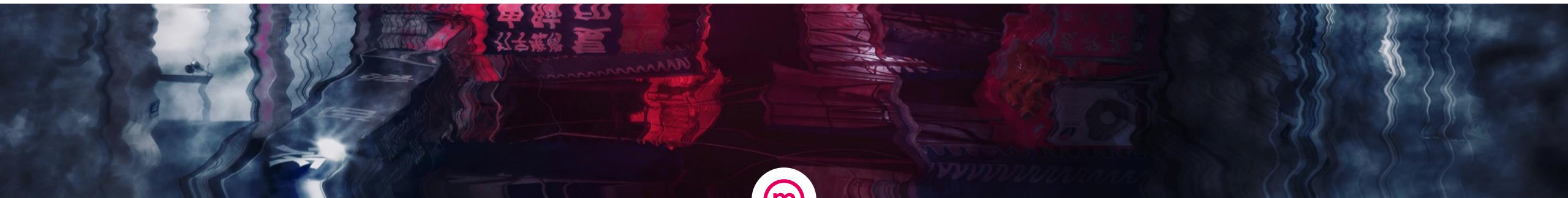
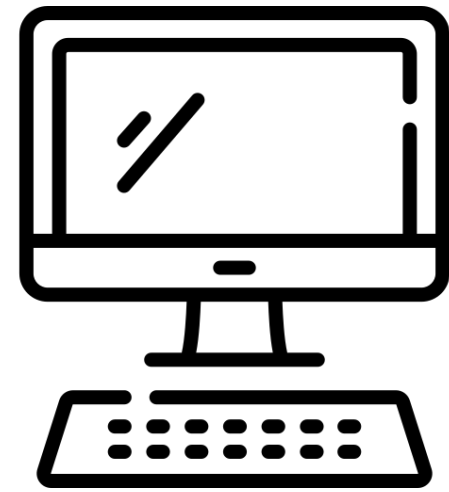


- obejrzyj zmienne PR stworzone w poprzednim zadaniu na wykresie.
- na podstawie PR i PL stwórz zmienne zawierające cenę krótkookresową (PS).
 - **TIP1:** Zmienne PS, jako głębokość promocji powinny być wyrażone w procentach (wartość 30% mówi, że w danym tygodniu cena PR spadła o 30% w stosunku do ceny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



zadanie dodatkowe

- Stwórz uniwersalny algorytm tworzący zmienne PL i PS na podstawie zmiennych PR
- Uniwersalny algorytm powinien działać dla dowolnej liczby inputowych zmiennych PR
- Output funkcji może bazować na założeniach i przybliżeniach, nie musi być w 100% tak dokładny jak zmienne powstałe w przypadku ręcznego przetwarzania na podstawie oglądania zmiennych na wykresie.
- **UWAGA:** Napisanie uniwersalnej i poprawnie działającej funkcji/algorytmu zostanie nagrodzone podniesieniem oceny końcowej przedmiotu o 0.5.

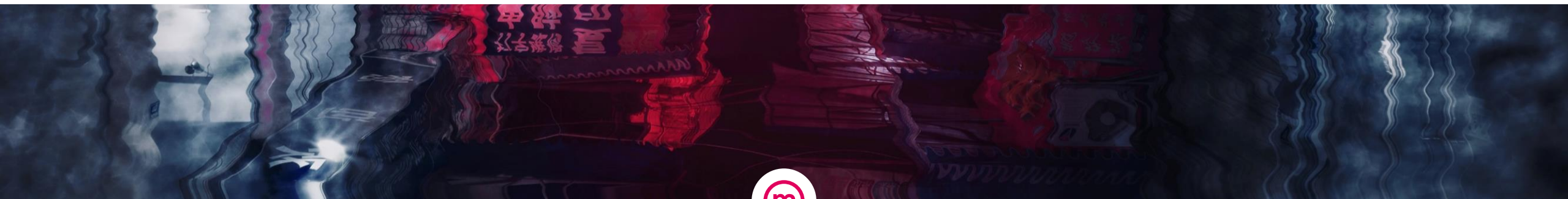


paradoks agregacji cen

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
UHT – 1l	4.0 PLN	1000 l	3.83 PLN	4.4 PLN	200 l	2.92 PLN
UHT – 2l	3.0 PLN	200 l		3.3 PLN	1000 l	



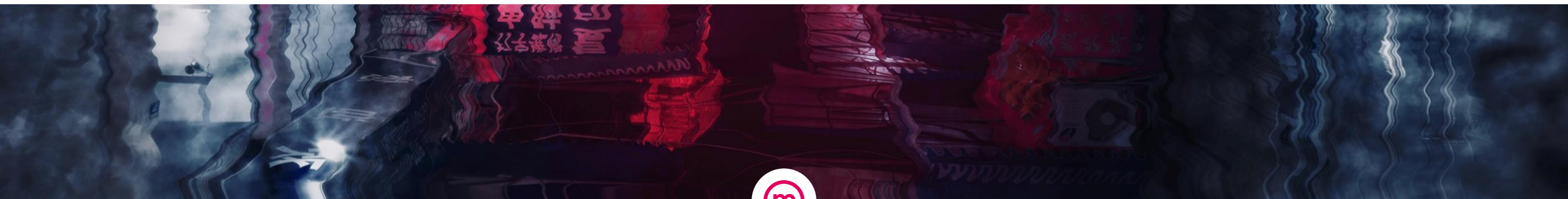
dystrybucja numeryczna i ważona na przykładzie pojedynczego SKU

numeryczna

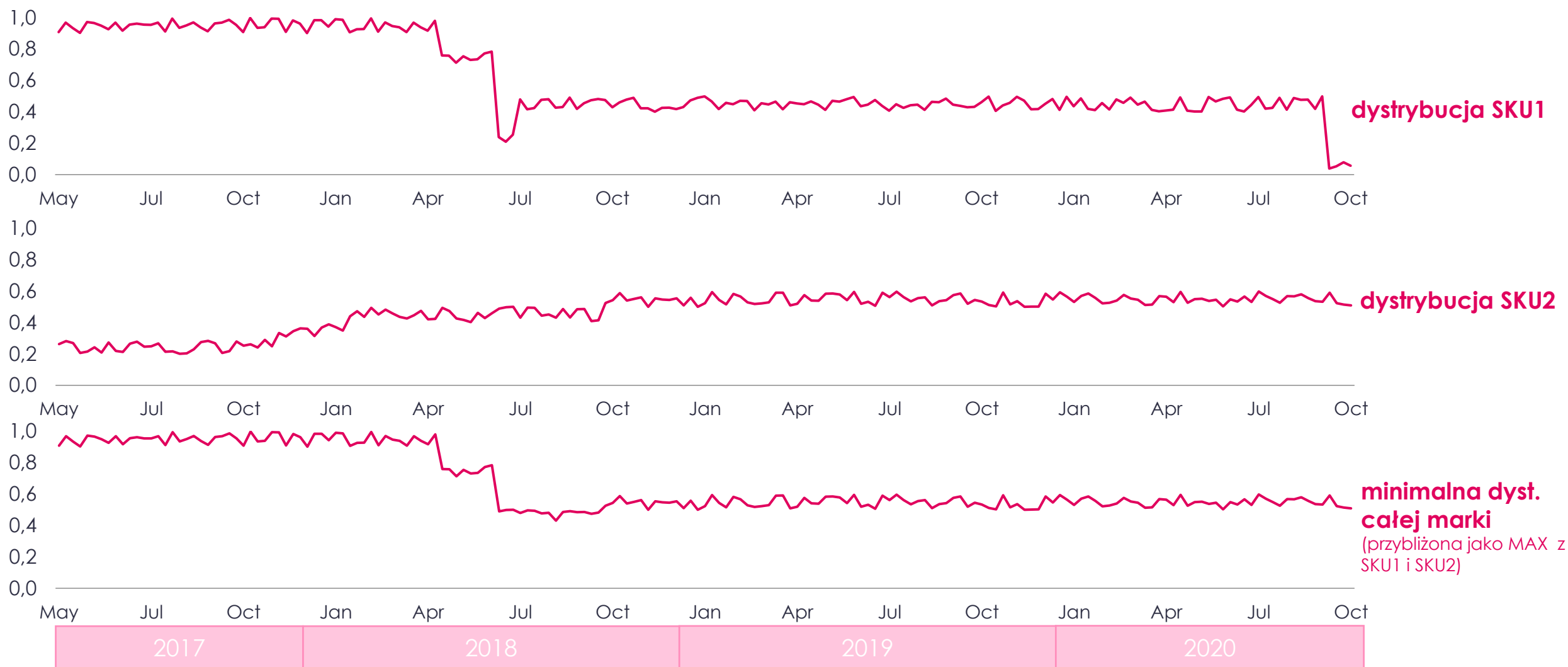
- odsetek sklepów w których dany SKU był dostępny (przykład: jeżeli dla SKU-1 dystrybucja w pierwszym tygodniu 2022 roku wynosi 68%, znaczy to, że ten produkt był dostępny w 68% sklepów).
- zawiera się w przedziale od 0 do 1
- traktuje każdy sklep jednakowo

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

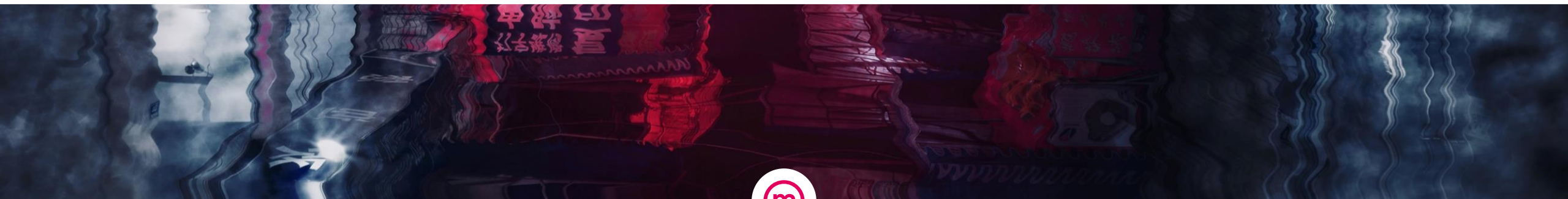
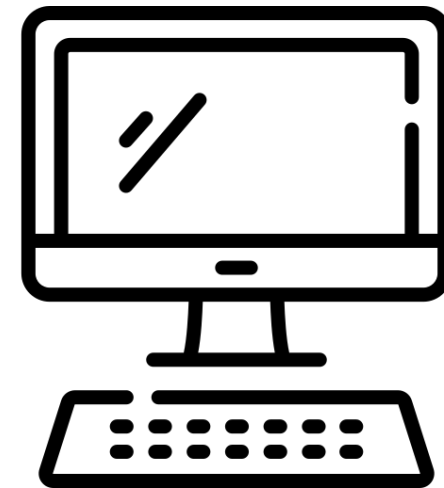


dystrybucja SKU versus całej marki



zadanie 4 (1 pkt)

- na podstawie zmiennych dotyczących dystrybucji SKU stwórz zmienną mówiącą o przybliżonej dystrybucji całej marki **DW_MILK**
 - **TIP1:** wykorzystaj funkcję **pmax()**
- wynikiem zadania 5 powinna być niezależna ramka danych



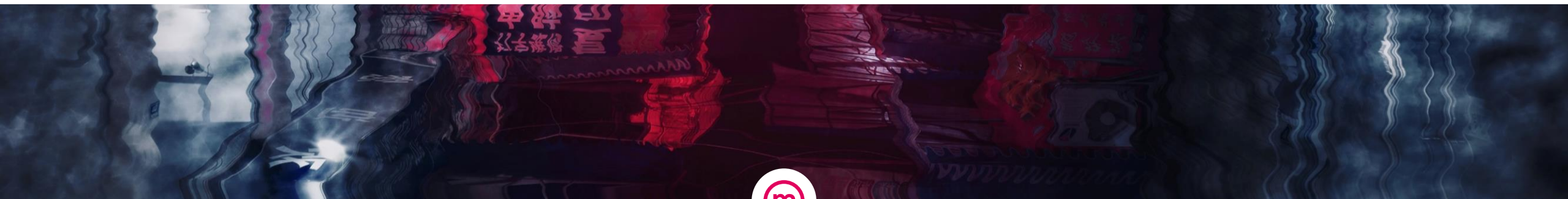
inflacja

jak przygotować zmienną?

- potrzebujemy współczynnika inflacji month-on-month
- jako punkt odniesienia możemy wybrać miesiąc przed rozpoczęciem okresu modelowanego
- konieczna będzie transformacja danych miesięcznych (GUS) na tygodniowe
- na końcu wskazane jest wygładzenie przetransformowanej zmiennej średnią ruchomą

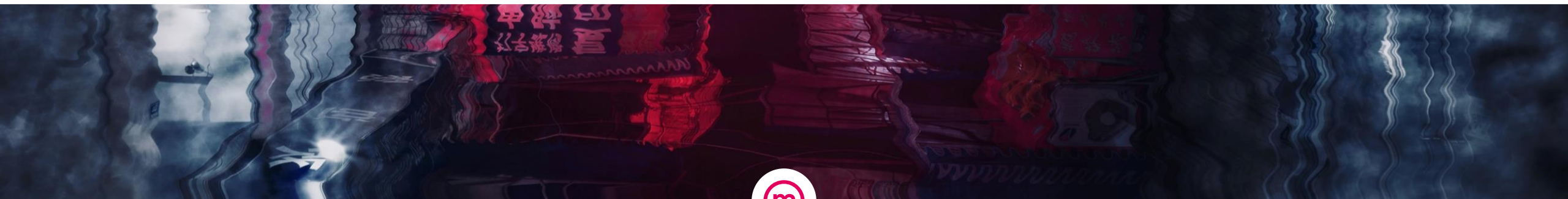
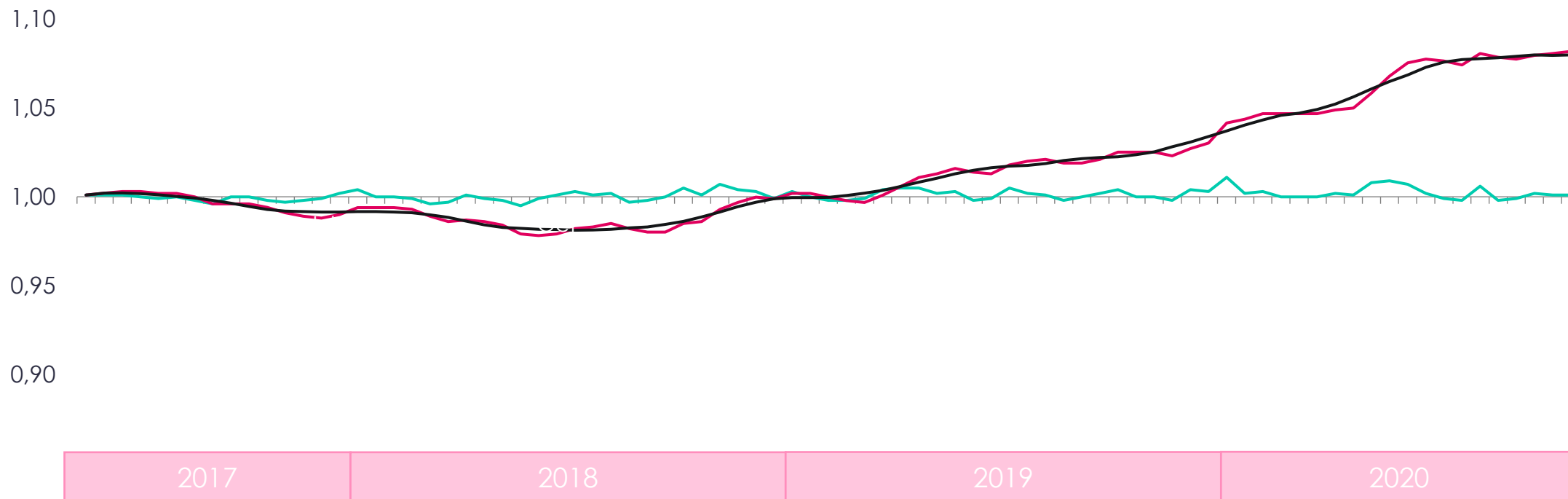
jak używać inflacji w modelach?

- gdy modelujemy **wartość** sprzedaży:
 - podzielenie zmiennej objaśnianej przez inflację
 - uwzględnienie inflacji jako jednej ze zmiennych objaśniających
- gdy modelujemy **wolumen** sprzedaży:
 - pominięcie inflacji (np. w przypadku dóbr pierwszej potrzeby)
 - dzielenie wszystkich uwzględnianych w modelu cen przez współczynnik inflacji



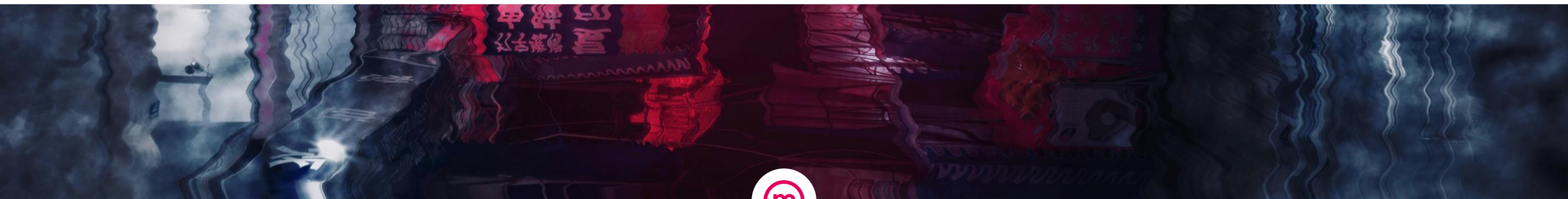
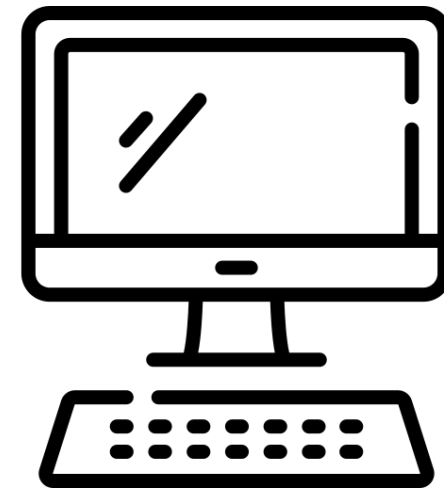
inflacja

month-on-month → indeks do pierwszego miesiąca → średnia ruchoma



zadanie 5 (2 pkt)

- na podstawie zmiennej **INFLACJA_WOW** (współczynnik inflacji tydzień do tygodnia, czyli np. 1.002 w tygodniu T6 znaczy, że poziom cen w T6 wzrósł o 0.2% w porównaniu do T5) przygotuj zmienne:
 - **INFLACJA_BASE**, która będzie zawierała skumulowany współczynnik inflacji w odniesieniu do pierwszego tygodnia przed okresem modelowanym.
 - **INFLACJA_SMOOTH**, która będzie scentrowaną średnią ruchomą z 5 tygodni policzoną na podstawie zmiennej INFLACJA_BASE.
- Wynikiem zadania 5 powinna być niezależna ramka danych



COVID

wybór zmiennej odzwierciedlającej pandemię. Musi to być:

- metryka „czysta” czyli niezawierająca w sobie efektów niezwiązanych z COVIDem (np. Consumer Confidence Index jest złą zmienną, bo zawiera w sobie też efekty sytuacji gospodarczej i geopolitycznej).
- metryka mająca bezpośredni wpływ na zmienną objaśnianą (np. liczba zajętych respiratorów nie ma prawie żadnego powiązania z ilością sprzedanego piwa).
- metryka mająca interpretację biznesową oraz zdroworozsądkową

Dane

Liczba chorych

Liczba nowych przyp.

Liczba szczepień

Liczba zaj. respiratorów

Mobilność ludności

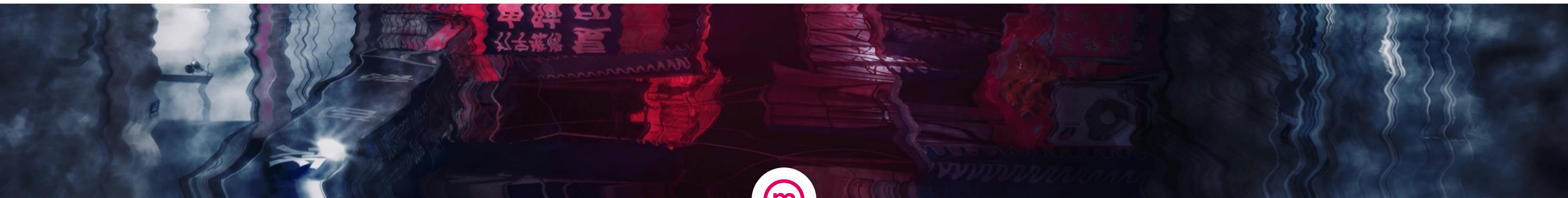
Wyszukania słów
kluczowych w Google

Lockdowny (0-1)

Zamknięcia sklepów (0-1)

Cons. Confidence Idx.

Konsumpcja prywatna



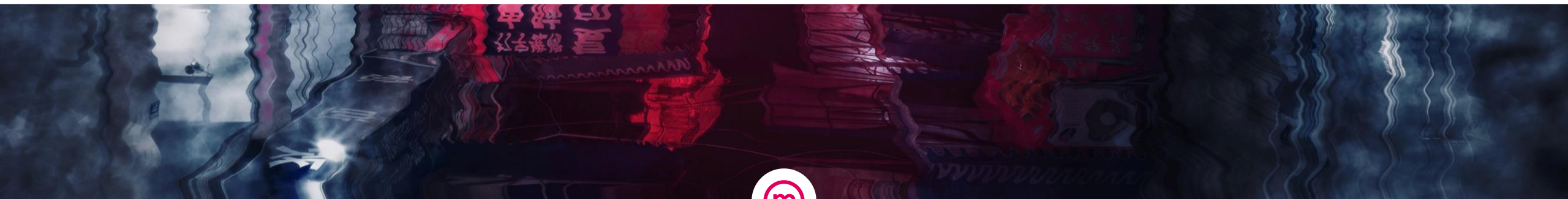
zmienne binarne

przykłady

- model na danych tygodniowych: zmienna przyjmująca wartość 1 **w każdym** tygodniu wielkanocnym
- model na danych tygodniowych: zmienna przyjmująca wartość 1 w tygodniach obowiązywania narodowej kwarantanny
- model na danych dziennych: zmienna przyjmująca wartość 1 w każdy piątek

best practise

- zmienne binarne (i zero-jedynkowe) bardzo silnie oddziałują na statystyki modelu, trzeba więc być ostrożnym i nie używać ich liberalnie
- nigdy nie wykorzystujemy w modelu zmiennych **z wartością nierówną zero w pojedynczym tygodniu**
- zmienna binarna ZAWSZE musi mieć interpretację w modelu (niedopuszczalna jest sytuacja, w której wyjątkowo wysoką resztę w pierwszym tygodniu lutego 2021 roku „wyciągamy” zmienną binarną X_2021_02_01 nie mając uzasadnienia/interpretacji dla tej zmiennej).



03

dane mediowe

wybór metryki jest kluczowy w prawidłowym uwzględnieniu zmiennej mediowej w modelu

metryka powinna:

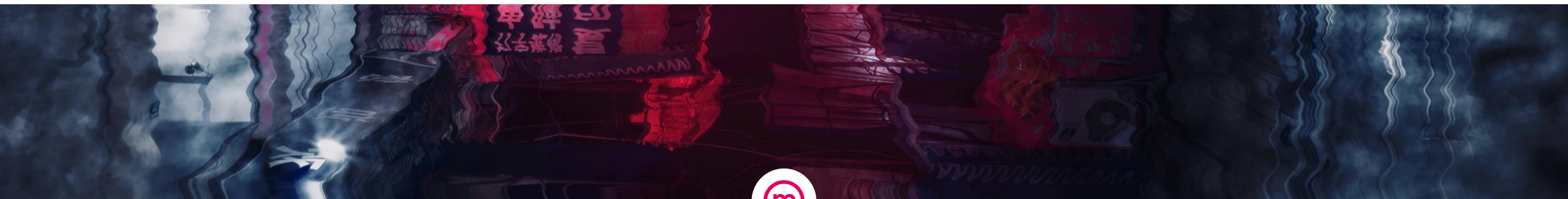
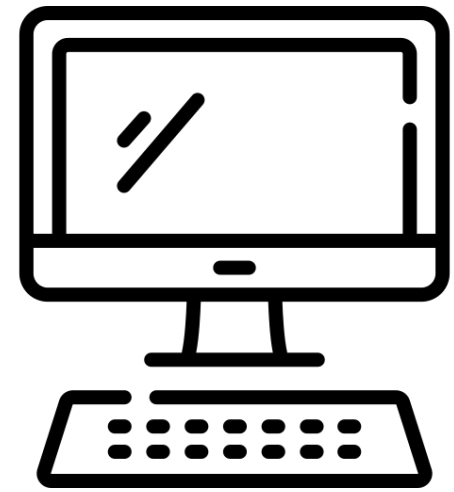
- odzwierciedlać jak najbliżej i najdokładniej kontakt odbiorcy z reklamą:
 - koszt kampanii nie jest dobrą metryką, ponieważ nie reprezentuje natężenia kontaktu z odbiorcą. Droga kampania mogła zawieść na etapie egzekucji i nie być widoczna dla odbiorcy.
- odzwierciedlać/zliczać wszystkie kontakty:
 - kliknięcia w reklamę zwykle są gorszą metryką niż impresje, ponieważ widząc liczbę klików nie potrafimy wskazać prawdziwej liczby kontaktów reklamy z odbiorcą. Widzimy wtedy jedynie kontakty, w których doszło do interakcji.
- być możliwie porównywalna pomiędzy mediami
- być dostosowana do formy działania konkretnego medium (np. liczba wysłanych maili jest intuicyjnie gorszym miernikiem kampanii niż liczba maili otwartych)

Medium	Preferowana metryka
TV	GRP/TRP
Radio	GRP/TRP
OOH	GRP/TRP
Kino	GRP/TRP
social (Meta)	impresje
video online	impresje
display online	impresje
YouTube	impresje
search	impresje
afiliacje	kliknięcia
artykuły sponsorowane	kliknięcia
mailing	otwarte emaille



zadanie 6 (1 pkt)

- obejrzyj zmienne zaczynające się od TV_ na wykresie
- zainputuj braki danych wartością 0 (zakładamy, że w okresie dla którego nie mamy danych, dane aktywności nie wystąpiły).
- wynikiem zadania 6 powinna być niezależna ramka danych



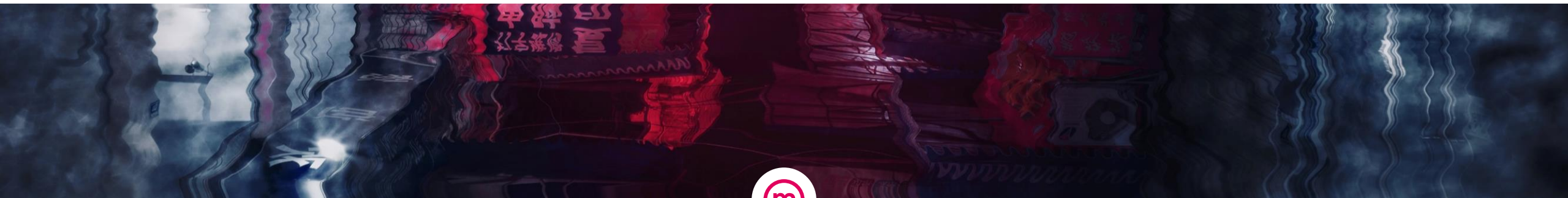
definicje impresji

różni dostawcy/platformy w różny sposób definiują impresje:

- **Meta:** „Ad Manager counts an impression each time a creative is downloaded in the user's device and has begun to load. Note that the impression is counted before the creative is fully downloaded and viewed by the end user.”
- **Google:** „An impression is counted when a list item is visible in the short list view, and also when the list view is expanded, whether or not the item is actually scrolled into view. Therefore, if an item is visible in the short view and the user then clicks to expand the list, two impressions are counted.”

sytuacja jest bardziej skomplikowana w przypadku video:

- **YouTube:**
 - „Someone watches a complete ad that's 11–30 seconds long.”
 - „Someone watches at least 30 seconds of an ad that's more than 30 seconds long.”
 - „Someone interacts with the ad.”



definicja GRP



grupa docelowa

pierwszy spot



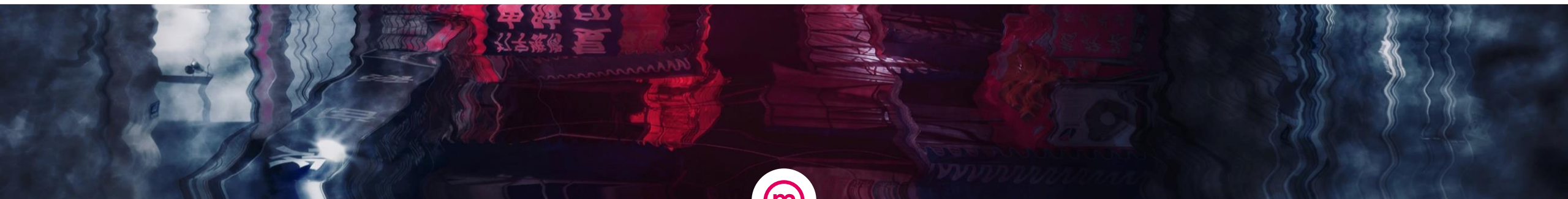
Rating I spotu:
3 osoby
25% - rating 25

drugi spot



Rating II spotu:
4 osoby
33% - rating 33

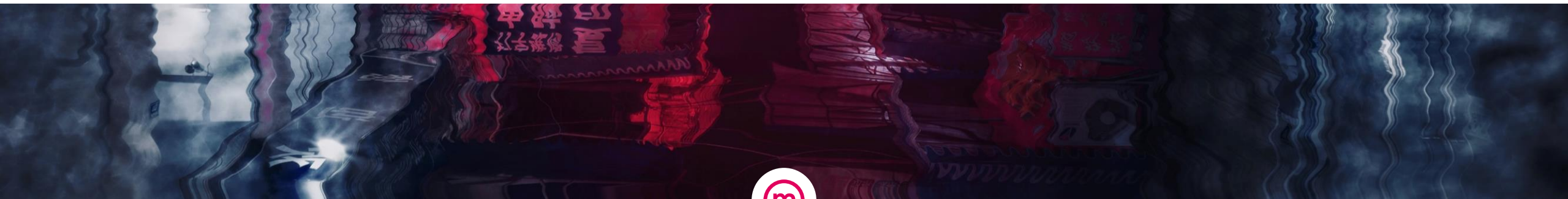
SUMA:
58 GRP



definicja GRP – jak uwzględnić fakt, że spoty mają różną długość?

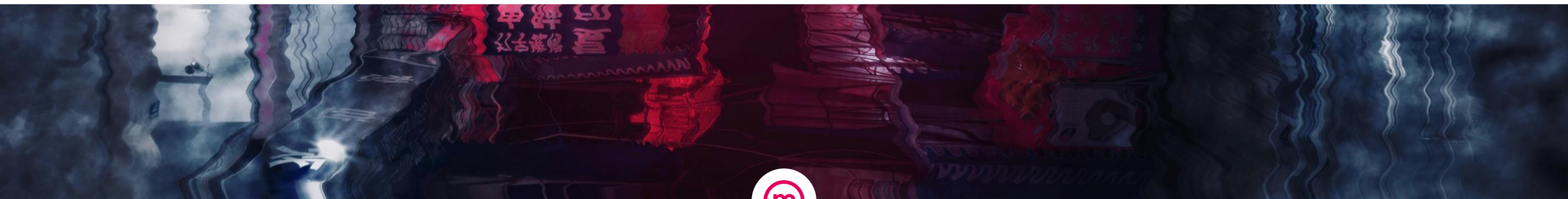
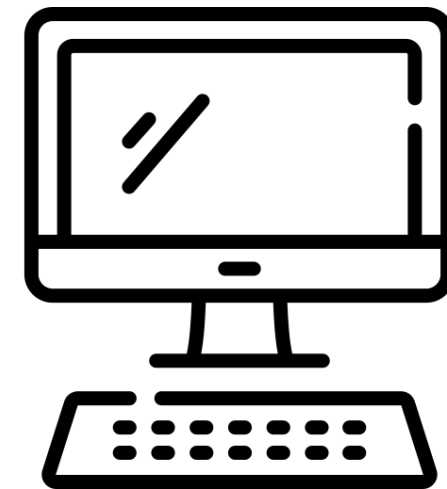
po 30'' spocie, który obejrzało 40% grupy docelowej spodziewamy się silniejszego wpływu na sprzedaż niż po spocie 15'', który również obejrzało 40% grupy

- przygotowując zmienne do modelowania ekonometrycznego należy więc uwzględnić różnice w długości spotu
- Najpopularniejszą długością kopii na polskim rynku jest 30 sekund
- GRPsy wszystkich spotów przeliczamy więc w taki sposób, by były ekwiwalentami spotu 30'', według przeliczników:
 - kopia 45'' – 1.50
 - kopia 30'' – 1.00
 - kopia 20'' – 0.85
 - kopia 15'' – 0.70
 - kopia 10'' – 0.55



zadanie 7 (2 pkt)

- korzystając z widocznego w udostępnionym skrypcie słownika przelicz GRP na eqGRP (czyli ekwiwalenty kopii 30'')
- **TIP1:** możesz skorzystać z `pivot_longer()` i `pivot_wider()` w analogiczny sposób co w zadaniu 2
- **TIP2:** słownik dołącz do danych komendą **`mutate(idx = tv.dict[XXX])`** gdzie XXX to nazwa kolumny, do której chcesz dołączyć słownik
- zsumuj wszystkie zmienne **TV_** do jednej zmiennej **TV**
- wynikiem zadania 7 powinna być niezależna ramka danych



AdStock czyli model odpowiedzi konsumentów na przekaz marketingowy

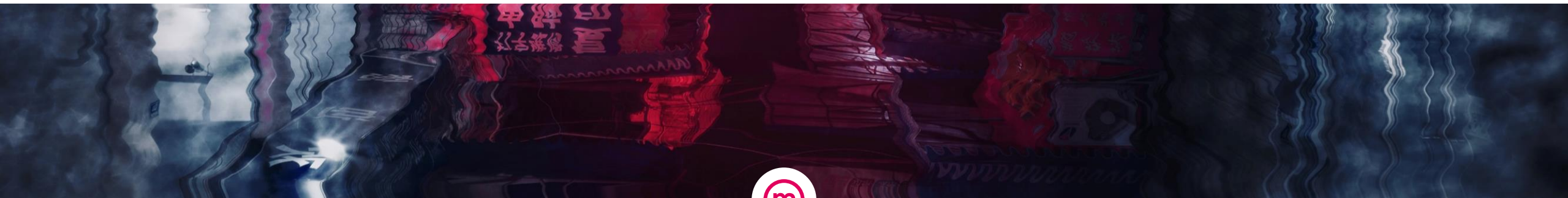
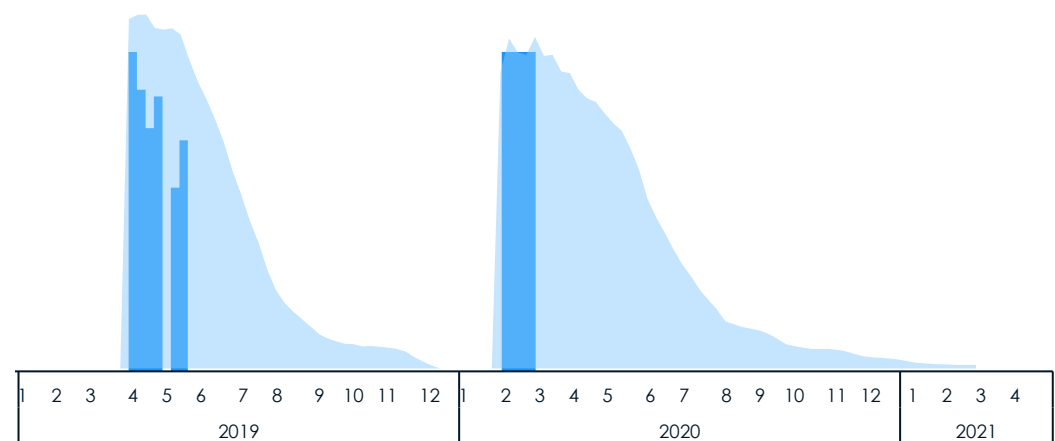
Wyemitowane GRP lub impresje nie oddziałują na konsumenta jedynie w momencie emisji reklamy. Przekaz marketingowy zwykle zostaje zapamiętany i realizuje się jako zakup produktu po kilku dniach lub nawet tygodniach lub miesiącach.

Model, który opisuje w jaki sposób przekaz zostaje „przesunięty” czy też „przedłużony” w czasie nazywany jest AdStockiem.

AdStockiem nazywamy również potocznie przekształcenie matematyczne nakładane na dane impresyjne/GRP.

Poziomy AdStocku zwykle podlegają następującym zależnościom:

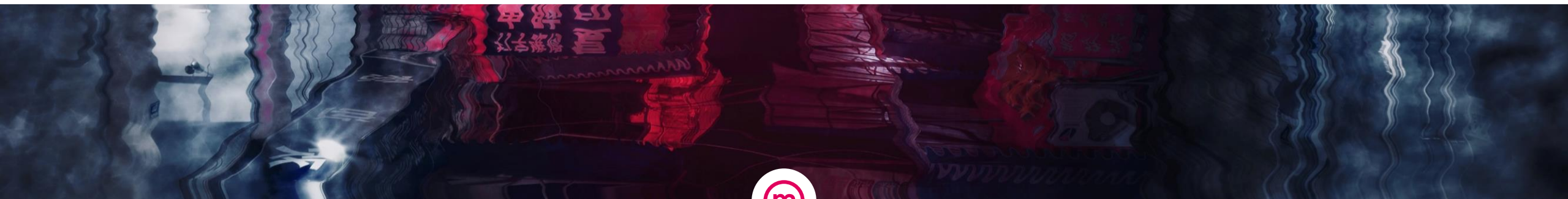
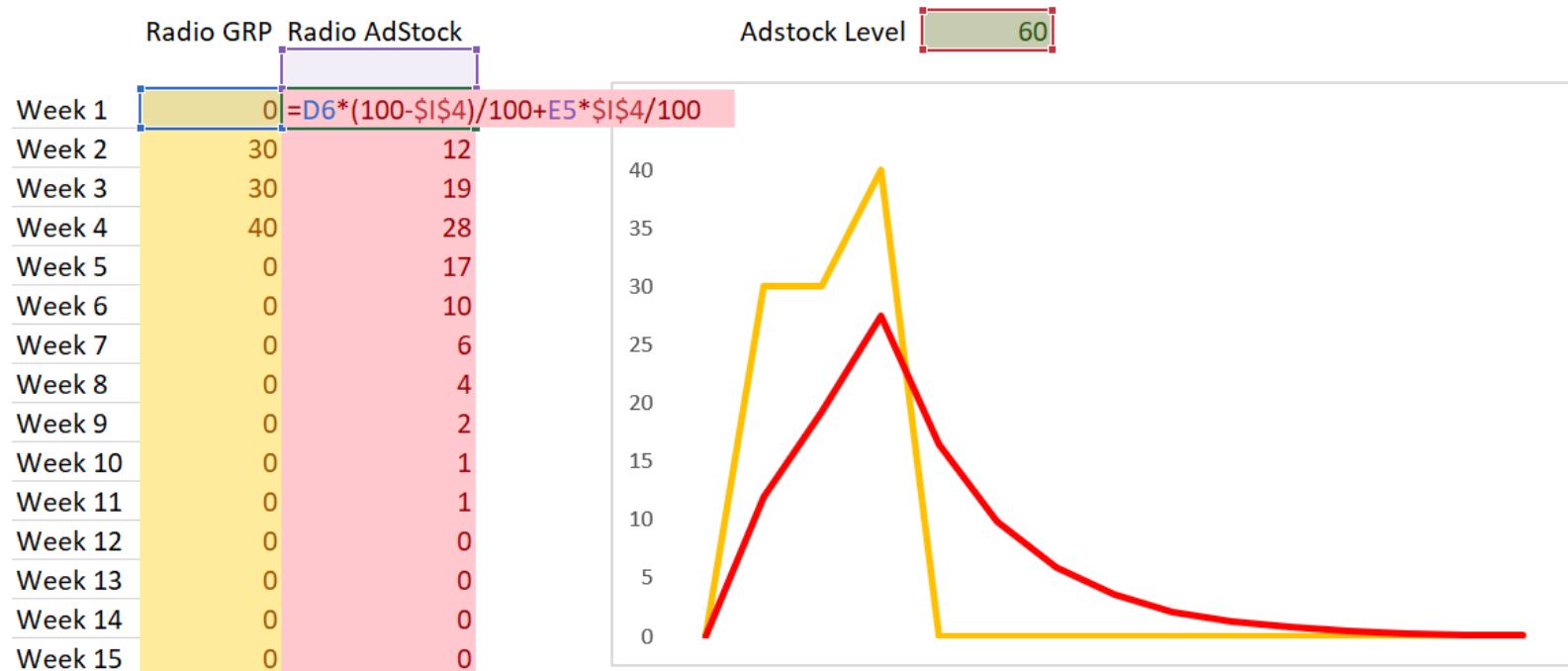
- Spot promocyjny < Spot opowiadający historię
- Search < Display < Online Video < TV



wzór na AdStock

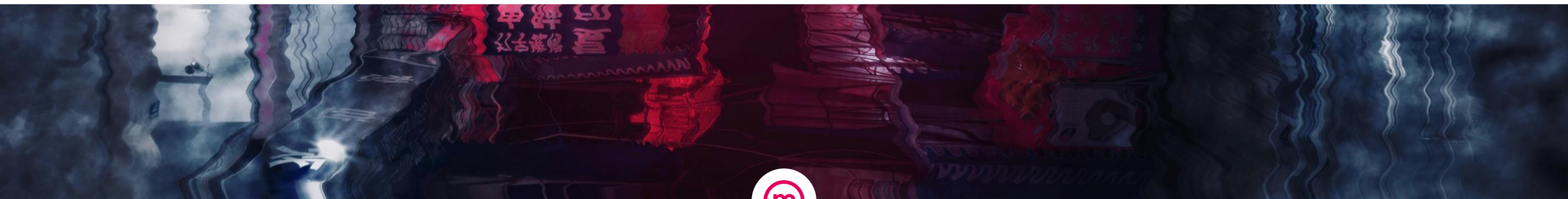
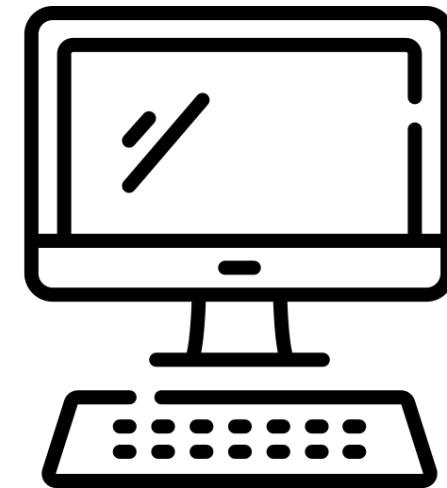
założmy AdStock na poziomie 60%

- w danym tygodniu wpływ radia na sprzedaż składa się z:
 - 60% wartości wpływu z zeszłego tygodnia
 - 40% wartości GRP w aktualnym tygodniu



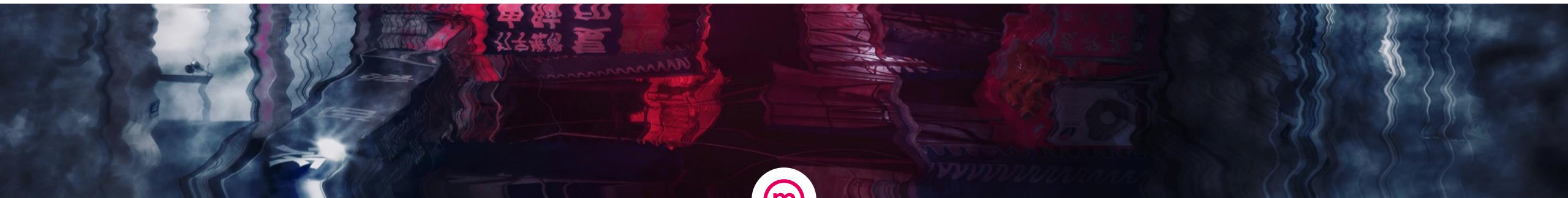
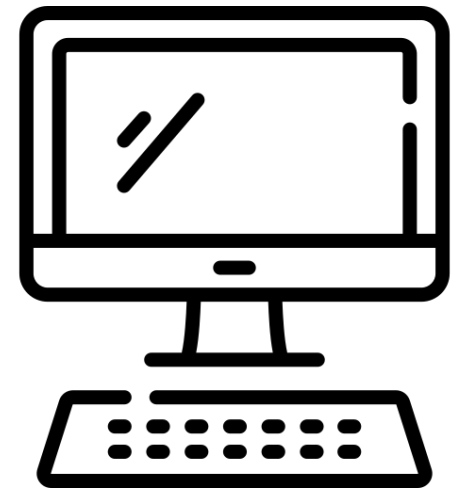
zadanie 8 (3 pkt)

- na przygotowaną w poprzednim zadaniu zmienną **TV** nałóż przekształcenie $\text{AdStock} = 80\%$. Nową zmienną nazwij **TV_ADS80**.
 - **TIP1:** możesz skorzystać z pętli for
 - **TIP2:** żeby zainicjować pętlę wymagane będzie ręczne przypisanie wartości w pierwszym tygodniu zmiennej **TV_ADS80**. Załóżmy, że $\text{TV_ADS80}' = \text{TV}'$.
- obejrzyj obie zmienne **TV** i **TV_ADS80** na wykresie
- wynikiem zadania 8 powinna być niezależna ramka danych



zadanie 9 (1 pkt)

- połącz wszystkie stworzone w zadaniach 1-8 ramki danych w finalną bazę nazwaną **database.df**
- wyczyść środowisko kasując z niego wszystkie obiekty oprócz finalnej ramki danych
- wynikiem zadania 9 powinna być niezależna ramka danych





Dziękujemy

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