**Hurtownie danych – Projekt**

PWr. Wydział Informatyki i Telekomunikacji Data: 6.05.2025

|  |  |
| --- | --- |
| Student | ------------------------------------------------------------- |
| Imię | Zlata |
| Nazwisko | Ranchukova |

# Tytuł projektu (wybranego lub zmienionego w celu realizacji)

Hurtownia Danych dla Analizy Rynku Nieruchomości w Wielkiej Brytanii w oparciu o Dane Price Paid Data oraz dane o inflacji i PKB

* **Uzasadnienie wyboru lub zmiany**

Mi się wydaje, że dane są ciekawsze i lepszej jakości. Są 2 pliki, które są przydatne w analizie oraz robią process budowania projektu ciekawszym

# Charakterystyka dziedziny problemowej

# Opis obszaru analizy (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

Projekt koncentruje się na analizie rynku nieruchomości w Wielkiej Brytanii, w szczególności na danych transakcyjnych z bazy "UK Price Paid Data" (PPD), zarządzanej przez HM Land Registry. Dane obejmują transakcje kupna-sprzedaży nieruchomości od 1995 roku (dla kategorii standardowych) oraz od 2013 roku (dla transakcji dodatkowych, takich jak nieruchomości komercyjne czy przejęcia bankowe).

**Kluczowe aspekty analizy:**

* Trendy cenowe w podziale na regiony, miasta i kody pocztowe.
* Porównanie typów nieruchomości (domy wolnostojące, szeregowe, mieszkania).
* Wpływ wieku nieruchomości (nowe vs. używane) na cenę.
* Wpływ zmiany PKB oraz inflacji na liczbę sprzedawanych nieruchomości oraz na ich cenę
* Rozpodział cen nieruchomości na mapie WB
* Analiza form własności (freehold vs. leasehold).

# Problemy

* **P01 – Brak scentralizowanego narzędzia do analizy trendów**

Obecnie dane PPD są dostępne jako pliki CSV bez zaawansowanych narzędzi do wizualizacji i analizy przestrzennej.

* **P02 – Trudność w identyfikacji atrakcyjnych inwestycyjnie lokalizacji**

Dane wymagają agregacji i porównań międzyregionowych, aby wskazać obszary o największej aprecjacji cen.

* **P03 – Niejednoznaczność danych adresowych**

Niektóre pola (np. PAON, SAON, Locality) mogą zawierać niepełne lub zmieniające się dane.

* **P04 – Duża objętość danych historycznych**

Zbiór PPD zawiera miliony rekordów, co wymaga wydajnych metod przetwarzania i przechowywania.

# Cel przedsięwzięcia

# Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

Hurtownia danych ma umożliwić użytkownikom podejmowanie decyzji w oparciu o:

* **Lokalizację** – Gdzie ceny rosną najszybciej? Które obszary są stabilne?
* **Typ nieruchomości** – Czy lepiej inwestować w mieszkania czy domy jednorodzinne?
* **Rynek pierwotny vs. wtórny** – Czy nowe budownictwo przynosi wyższe zyski?
* **Dynamikę rynku** – Jak sezonowość wpływa na ceny?

# Pytania badawcze - - Zakres analizy

1. Q01 - Jak / Czy średnia cena nieruchomości zależy od inflacji i PKB?
2. Q02 – Jak / Czy wolumen transakcji (liczba sprzedaży) zależy od inflacji i PKB?
3. Q03 – Wpływ wskaźników makroekonomicznych na średnią cenę i wolumen dla poszczególnych charakterystyk nieruchomości?
4. Q04 – Jakie są najdroższe lokalizację?
5. Q05 – Jakie typy nieruchomości są najdroższe i najtańsze oraz jakie są najpopularniejsze?

# Potencjalni użytkownicy

* **Analitycy rynku nieruchomości** – Tworzenie raportów i prognoz.
* **Deweloperzy i inwestorzy** – Identyfikacja najlepszych lokalizacji pod nowe inwestycje.
* **Banki i instytucje finansowe** – Ocena ryzyka kredytowego i wycena zabezpieczeń.
* **Rząd i samorządy** – Monitorowanie polityki mieszkaniowej i cenowej.
* **Indywidualni kupujący** – Porównanie cen przed zakupem nieruchomości.

# Dane źródłowe

# Źródła danych

Charakterystyka pliku zawierający danę źródłowe przeznaczone do stworzenia tematycznej hurtowni danych jest przedstawiona w tab. 1.

Tabela 1. Zbiory danych źródłowych

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lp.** | **Plik** | **Typ** | **Liczba rek.** | **Rozmiar[MB]** | **Opis** |
|  | pp-complete | csv | 27mln | 5145 | Pełny zbiór danych transakcyjnych od 1995 r. Zawiera ID, cenę, datę, typ nieruchomości, status, adres, kod pocztowy |
|  | series-030525 | csv | 624 | 0.011 | Inflacja w Wielkiej Brytaniji |
|  | series-260625 | csv | 288 | 0.05 | PKB w Wielkiej Brytaniji |

# Lokalizacja, dostępność danych źródłowych

Dane są dostępne pod linkami:

* [*https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads*](https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads)
* [*https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/l55o/mm23*](https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/l55o/mm23)
* [*https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ihyq/pn2*](https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ihyq/pn2)

# Słownik danych – interpretacja

Interpretacja oraz wyjaśnienie znaczeń pojęć dziedzinowych zostały zawarte w tab.2.

Tabela 2. Słownik atrybutów

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: pp-complete.csv | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Znaczenie** | **Uwagi** |
| 1. | county | Tekst | Okręg / Powiat | - |
| 2. | date\_of\_transfer | Data | Data zawarcia transakcji | YYYY-MM-DD |
| 3. | district | Tekst | Dystrykt | - |
| 4. | locality | Tekst | Dzielnica | Może powtarzać nazwę miasta |
| 5. | new\_build | Znak | Czy dana nieruchomość jest nowobudową. Y/N | Dopuszczalne znaczenia: Y, N |
| 6. | paon (Primary Addressable Object Name) | Tekst | Numer domu lub nazwa | Może być numeryczny lub tekstowy |
| 7. | postcode | Tekst | Kod pocztowy lokalizacji nieruchomości | Może zawierać literówki lub być niepełny |
| 8. | ppd\_type | Znak | Wskazuje typ transakcji. A = Standardowy wpis dotyczący zapłaconej ceny, obejmuje pojedyncze nieruchomości mieszkalne sprzedane za wartość. B = Dodatkowy wpis Price Paid, w tym przeniesienia na mocy prawa sprzedaży/odkupu, transakcje typu buy-to-lets (jeśli można je zidentyfikować na podstawie hipoteki), przeniesienia na rzecz osób fizycznych i sprzedaż, w przypadku której typ nieruchomości jest sklasyfikowany jako „Inny”. | Należy zauważyć, że kategoria B nie identyfikuje oddzielnie podanych typów transakcji.  HM Land Registry gromadzi informacje na temat transakcji kategorii A od stycznia 1995 roku. Transakcje kategorii B zostały zidentyfikowane w październiku 2013 roku. |
| 9. | price | Liczba | Cena sprzedaży nieruchomości w funtach | >0, możliwe wartości odstające |
| 10. | property\_type | Znak | Typ nieruchomości: D, S, T, F, O. D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other | Dopuszczalne znaczenia: D, S, T, F, O |
| 11. | record\_status | Znak | A = Addition C = Change D = Delete | Nie dotyczy pełnego pliku, ale ta kolumna jest i zawiera tylko „A”. Podczas ETL ta kolumna zostanie usunięta. |
| 12. | saon (Secondary Addressable Object Name) | Tekst | Numer mieszkania lub analogiczna jednostka | Dużo pustych pól (jeżeli nieruchmość = cały dom, a nie, na przykład, mieszkanie w domie) |
| 13. | street | Tekst | Ulica | Może zawierać literówki |
| 14. | tenure | Znak | Rodzaj własności: F – Freehold, L – Leasehold | Dopuszczalne znaczenia: F, L |
| 15. | town | Tekst | Miasto | - |
| Plik: series-030525.csv | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Znaczenie** | **Uwagi** |
| 1. | Date | Data | Okres, którego dotyczy procent inflacji | Plik zawiera dane z datą w postaci roku, roku + krawtalu, rok + miesiąc |
| 2. | percentage | Liczba | Znaczenie inflacji w procentach | - |
| Plik: series-260625.csv | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Znaczenie** | **Uwagi** |
| 1. | Date | Data | Okres, którego dotyczy procent GDP | Plik zawiera dane z datą w postaci roku + krawtału |
| 2. | percentage | Liczba | Znaczenie inflacji w procentach | - |

# Ocena jakościowa danych

Wynik analizy jakościowej przeprowadzonej za pomocą programu Tableau oraz profilu danych SSIS został przedstawiony w tab. 3.

Tabela 3. Ocena jakościowa danych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: pp-complete.csv. Plik nie zawiera nagłówków kolumn | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Zakres wartości** | **Uwagi - ocena jakości danych** |
|  | transaction\_id | Tekst | Unikalne wartości | Brak brakujących wartości; Wszystkie wartości są unikalne |
|  | price | Liczba | >0 | Brak anomalii, Brak brakujących wartości |
|  | date\_of\_transfer | Data | 01.01.2000 do 31.12.2024 | Wszystkie rekordy są w poprawnym formacie, nie ma anomalii, Brak brakujących wartości |
|  | postcode | Tekst | - | Są w róznych formatach. Możliwe nieaktualne dane (zmiany kodów pocztowych) |
| 5. | property\_type | Znak | D, S, T, F, O | Brak nieprawidłowych wartości |
| 6. | new\_build | Znak | Y, N | Brak nieprawidłowych wartości |
| 7. | tenure | Znak | F, L | Brak nieprawidłowych wartości |
| 8. | paon (Primary Addressable Object Name) | Tekst | Numeryczne (np. "12") lub tekstowe (np. "Rose Cottage") | - |
| 9. | saon (Secondary Addressable Object Name) | Tekst | Głównie puste (dla domów) lub numery mieszkań (np. "3") | Większość rekordów pusta, ale to jest normalne i nie wymaga naprawienia |
| 10. | street | Tekst | - | - |
| 11. | locality | Tekst | - | Mogą powtarzać nazwę miasta |
| 12. | town | Tekst | - | Brak błędów |
| 13. | district | Tekst | - | Możliwe nieaktualne dane (zmiany granic administracyjnych) |
| 14. | county | Tekst | - | - |
| 15. | ppd\_type | Znak | A (standardowe), B (specjalne, np. buy-to-let) | Brak błędów |
| Plik: series-030525.csv | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Zakres wartości** | **Uwagi - ocena jakości danych** |
| 1. | Date | Data | - | Różne formaty (rok, rok+kwartał, rok+miesiąc) |
| 2. | Percentage | Liczba | Wartości procentowe (np. 2.5, 3.1) | - |
| Plik: series-260625.csv | | | | |
| **Lp.** | **Atrybut** | **Typ danych** | **Zakres wartości** | **Uwagi - ocena jakości danych** |
| 1. | Date | Data | - | rok+kwartał |
| 2. | Percentage | Liczba | Wartości procentowe (np. 2.5, 3.1) | - |

# Analityczne modele wielowymiarowe

# Fakty **podlegające analizie oraz ich miary**

Analizie będzie podlegał zbiór zarejestrowanych zdarzeń (tab. 4.)

Tabela 4. Fakty podlegające analizie

|  |  |  |  |
| --- | --- | --- | --- |
| **Lp.** | **Fakty** | **Miary** | **Uwagi** |
|  | Transakcje nieruchomości | Sale Price – cena z PPD  Count of Transactions  Average/Min/Max Price (agregaty) | Wszystkie miary są wyrażone w GBP; |

# Kontekst analizy faktów

Ustalony kontekst analizy faktów został przedstawiony w tab. 5.

Tabela 5. Wymiary analizy faktów

|  |  |  |  |
| --- | --- | --- | --- |
| Lp. | Wymiar | Atrybuty | Uwagi |
| 1. | Lokalizacja | County → District → Town → Locality → Postcode → Street / PAON / SAON (DimLocation) | Pozwala na analizy regionalne i przestrzenne |
| 2. | Czas | Rok → Kwartał → Miesiąc → Dzień (DimDate) | Wymiar czasowy do wykrywania trendów i sezonowości |
| 3. | Typ nieruchomości | Property Type (D, S, T, F, O) + New Build (Y/N) – atrybut wielowartościowy (DimPropertyType) | Porównania domów vs. mieszkań oraz rynku pierwotnego i wtórnego |
| 4. | Własność | Tenure (Freehold / Leasehold) — osobny wymiar (DimTenure) | Ocena wpływu prawa własności na cenę |
| 7. | Inflacja | Rok → Kwartał → Miesiąc, CPI % (DimInflation) | Korelacja inflacji z cenami nieruchomości |
| 8. | GDP | Rok → Kwartał, GDP % QoQ (DimGdp) | Korelacja koniunktury gospodarczej z cenami nieruchomości |

# Modele wielowymiarowe (UML)

Po przeanalizowaniu atrybutów źródła danych oraz ustalonego faktu/faktów i kontekstu analizy zaproponowano wielowymiarowy model konceptualny (rys. 1.). Składa się on z faktu FactSell oraz 6 wymiarów (DimDate, DimLocation, DimPropertyType, DimTenure, DimInflation, DimGdp) . Model ten reprezentowany jest w postaci schematu na rys. 1. ?.

Uwaga: własności klas nie zawierają kluczy obcych (jest to mechanizm implementacyjny!)

Зображення, що містить текст, ряд, схема, Графік

Вміст на основі ШІ може бути неправильним.

**Rysunek 1**. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

Na rysunku 1 przedstawiono wielowymiarowy model analityczny hurtowni danych w ujęciu konceptualnym – centralną tabelę faktów FactSell wraz z powiązanymi wymiarami czasu (DimDate), lokalizacji (DimLocation), typu nieruchomości (DimPropertyType), statusu nowości (DimNewBuild), formy własności (DimTenure) oraz makroekonomicznych wskaźników inflacji (DimInflation) i PKB (DimGdp).

# Implementacja modeli wielowymiarowych (6*.4 – opcjonalnie*)

# Schemat bazy danych HD (skrypt SQL z zachowaniem kolorowania)

CREATE TABLE dbo.DimDate (

date\_key INT IDENTITY(1,1) PRIMARY KEY,

full\_date DATE NOT NULL UNIQUE,

[year] INT,

[quarter] INT,

quarter\_name NVARCHAR(2),

[month] INT,

month\_name NVARCHAR(10),

[day] INT

);

CREATE TABLE dbo.DimLocation (

location\_id INT IDENTITY(1,1) PRIMARY KEY,

postcode NVARCHAR(20),

county NVARCHAR(100),

district NVARCHAR(100),

locality NVARCHAR(100),

town NVARCHAR(100),

street NVARCHAR(100),

paon NVARCHAR(100),

saon NVARCHAR(100),

CONSTRAINT UQ\_DimLocation UNIQUE

(postcode, county, district, locality, town, street, paon, saon)

);

CREATE TABLE dbo.DimPropertyType (

property\_type\_id INT IDENTITY(1,1) PRIMARY KEY,

property\_type\_code NCHAR(1) NOT NULL UNIQUE,

property\_type\_name NVARCHAR(50) NOT NULL

);

CREATE TABLE dbo.DimNewBuild (

new\_build\_id INT IDENTITY(1,1) PRIMARY KEY,

new\_build\_code NCHAR(1) NOT NULL UNIQUE,

new\_build\_name NVARCHAR(20) NOT NULL

);

CREATE TABLE dbo.DimTenure (

tenure\_id INT IDENTITY(1,1) PRIMARY KEY,

tenure\_code NCHAR(1) NOT NULL UNIQUE,

tenure\_name NVARCHAR(50) NOT NULL

);

CREATE TABLE dbo.DimInflation (

inflation\_id INT IDENTITY(1,1) PRIMARY KEY,

[year] INT NOT NULL,

[month] INT NOT NULL,

month\_name NVARCHAR(10),

[quarter] INT,

quarter\_name NVARCHAR(2),

percentage DECIMAL(5,2),

CONSTRAINT UQ\_DimInfl UNIQUE ([year],[month])

);

CREATE TABLE dbo.DimGdp (

gdp\_id INT IDENTITY(1,1) PRIMARY KEY,

[year] INT NOT NULL,

[quarter] INT NOT NULL,

quarter\_name NVARCHAR(2),

percentage DECIMAL(5,2),

CONSTRAINT UQ\_DimGdp UNIQUE ([year],[quarter])

);

CREATE TABLE dbo.FactSell (

transaction\_id NVARCHAR(38) PRIMARY KEY,

price DECIMAL(12,2) NOT NULL,

date\_key INT NOT NULL, -- DimDate

location\_id INT NOT NULL, -- DimLocation

property\_type\_id INT NOT NULL, -- DimPropertyType

new\_build\_id INT NOT NULL, -- DimNewBuild

tenure\_id INT NOT NULL, -- DimTenure

inflation\_id INT NOT NULL, -- DimInflation

gdp\_id INT NOT NULL, -- DimGdp

CONSTRAINT FK\_FactSell\_Date FOREIGN KEY (date\_key)

REFERENCES dbo.DimDate(date\_key),

CONSTRAINT FK\_FactSell\_Location FOREIGN KEY (location\_id)

REFERENCES dbo.DimLocation(location\_id),

CONSTRAINT FK\_FactSell\_PropType FOREIGN KEY (property\_type\_id)

REFERENCES dbo.DimPropertyType(property\_type\_id),

CONSTRAINT FK\_FactSell\_NewBuild FOREIGN KEY (new\_build\_id)

REFERENCES dbo.DimNewBuild(new\_build\_id),

CONSTRAINT FK\_FactSell\_Tenure FOREIGN KEY (tenure\_id)

REFERENCES dbo.DimTenure(tenure\_id),

CONSTRAINT FK\_FactSell\_Inflation FOREIGN KEY (inflation\_id)

REFERENCES dbo.DimInflation(inflation\_id),

CONSTRAINT FK\_FactSell\_Gdp FOREIGN KEY (gdp\_id)

REFERENCES dbo.DimGdp(gdp\_id)

);

Zupdatowałam skrypt HD – rozdzieliłam DimPropertyType na DimPropertyType i DimNewBuild, ponieważ takie grupowanie było bardzo niewygodne podczas analizy i budowy wykresów. Większość wymiarów ma pola „\*\_name”, czyli tekstowy zrozumiały opis znaczenia z plików źródłowych i te kolumny są wykorzystywane w NameColumn w properties wymiarów w SSAS (jak to było pokazywane na laboratoriach).

Dodatkowo, wszędzie zostały użyte klucze sztuczne.

* 1. Widok danych (Data Source View – kopia fragment ekranu)

Зображення, що містить текст, схема, План, Креслення

Вміст на основі ШІ може бути неправильним.

**Rysunek 2**. Widok danych

Rysunek 2 przedstawia widok danych (DSV) po wprowadzeniu kluczy obcych (sztucznych), schemat jest zgodny ze skryptem SQL z punktu 6.1. Fakt FactSell łączy się tu z wymiarami czasu, lokalizacji, typu nieruchomości, nowości, formy własności oraz inflacji i PKB.

* 1. *Wymiary (kopia ekranu: Atrybuty + Hierarchie)*

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 3**. DimDate

Na rysunku 3 pokazano konfigurację wymiaru **DimDate** w projekcie SSAS — dostępne atrybuty oraz hierarchię **Year → Quarter → Month**.

Зображення, що містить текст, знімок екрана, Шрифт, число

Вміст на основі ШІ може бути неправильним.

**Rysunek 4**. DimLocation

**Rysunek 4** prezentuje konfigurację wymiaru **DimLocation**: po lewej widać dostępne atrybuty (County, District, Locality, Street, Postcode oraz klucz Location Id), a po prawej zbudowaną hierarchię administracyjną **County → District → Town → Locality → Street → Postcode**. Taka kolejność odpowiada naturalnemu zagłębieniu adresu w danych HMLR i pozwala wygodnie „drill-downować” analizy od poziomu hrabstwa do konkretnej ulicy lub kodu pocztowego.

Zrezygnowano z dodawania atrybutów **PAON** i **SAON** (numery/nazwy budynków i lokali), ponieważ pełny kod pocztowy w Wielkiej Brytanii najczęściej identyfikuje pojedynczy blok lub niewielką grupę budynków, co jest wystarczające dla analiz najpopularniejszych lub najdroższych kompleksów mieszkaniowych. Pozostawienie tych pól wyłącznie w warstwie źródłowej ogranicza rozmiar hierarchii i przyspiesza procesowanie kubu, a w razie potrzeby umożliwia późniejsze rozszerzenie modelu o poziom pojedynczego lokalu.

Зображення, що містить текст, Шрифт, знімок екрана, ряд

Вміст на основі ШІ може бути неправильним. 

**Rysunek 5-6**. DimTenure

Rysunek 5-6 przedstawia konfigurację wymiaru **DimTenure**. Kluczem atrybutu jest kolumna tenure\_id (surrogate-key), natomiast w polu **NameColumn** wskazano tenure\_name, dzięki czemu w raportach zamiast kodów technicznych wyświetlają się zrozumiałe etykiety *Freehold* i *Leasehold*. Analogiczne ustawienie **NameColumn** zastosowano w większości wymiarów zawierających skróty lub kody (np. typ nieruchomości, atrybuty quarter i month), aby użytkownik otrzymywał czytelne nazwy podczas analizy danych.

Зображення, що містить текст, Шрифт, знімок екрана, білий

Вміст на основі ШІ може бути неправильним.

**Rysunek 7**. DimPropertyType

Rysunek 7 ilustruje wymiar **DimPropertyType** po rozdzieleniu go od atrybutu „nowości” (New Build). Kluczem jest kolumna property\_type\_id, natomiast w **NameColumn** (niewidocznym na zrzucie) ustawiono property\_type\_name, co pozwala prezentować w raportach czytelne nazwy typów nieruchomości (Detached, Semi-detached, Terraced, Flat, Other) zamiast pojedynczych liter-kodów.

Зображення, що містить текст, знімок екрана, Шрифт, Прямокутник

Вміст на основі ШІ може бути неправильним.

**Rysunek 8**. DimNewBuild

Rysunek 7 przedstawia wymiar **DimNewBuild** po wydzieleniu go z typu nieruchomości. Jedynym atrybutem jest klucz zastępczy new\_build\_id, którego **NameColumn** ustawiono na new\_build\_name, dzięki czemu w raportach zamiast kodów „Y/N” pojawiają się czytelne etykiety *New build* i *Resale*. Mała krotność wymiaru (tylko dwa wiersze) sprawia, że pełni on głównie rolę filtra binarnego w analizach średniej ceny i struktury sprzedaży.

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 9**. DimGdp

Rysunek 11 przedstawia wymiar **DimGdp**. Atrybuty zawierają:

* **Gdp Id** – klucz zastępczy (KeyColumns),
* **Year** i **Quarter** – poziomy kalendarza wykorzystywane w hierarchii,
* **Percentage** – kwartalny wskaźnik wzrostu PKB.

Po prawej zdefiniowano hierarchię **Year → Quarter**, co pozwala analizować dynamikę PKB na poziomie roku i kwartału. W kolejnych etapach kategoryzacja (np. przedziały wzrostu PKB) zostanie dodana już w kubie jako miara kalkulowana, bez rozszerzania struktury wymiaru.

Зображення, що містить текст, знімок екрана, Шрифт, число

Вміст на основі ШІ може бути неправильним.

**Rysunek 10**. DimInflation

Rysunek 10 przedstawia konfigurację wymiaru **DimInflation**.

* **Inflation Id** – klucz zastępczy (kolumna KeyColumns).
* **Year → Quarter → Month** – trzy poziomy kalendarza używane w hierarchii.
* **Percentage** – miesięczna wartość wskaźnika CPI.

Po prawej zdefiniowano hierarchię **Year → Quarter → Month**, co będzie pozwalało agregować CPI od miesiąca do kwartału i roku.

Dla wymiarów DimDate, DimGdp oraz Din Inflation wybrałam Typ „Time” oraz odpowiednie typy (Years, Quarters, Months) dla artybutów, co na etapie koskti pozwoliło używać agregację „Average over time”.

* 1. *Modele wielowymiarowe – Kostki (kopia ekranu: Miary + Wymiary + Diagram)*

Зображення, що містить текст, схема, знімок екрана, План

Вміст на основі ШІ може бути неправильним.

**Rysunek 11**. Kostka

Na etapie budowania kostki dodałam nowe pola kalkulacyjne dla DimGdp („GDP %” – dla akumulowania procentu zmiany za rok, jest wyliczany na podstawie atrybutu percentage, gdzie znaczenie opisuje zmianę za kwartał; „”), DimInflation oraz FactSell.

Dla DimGdp:

* Dodałam wymiar jeszcze jak Measure Group.
* Gdp Percentage (Measures -> DimGdp -> New Measure -> Usage = Sum -> Source Column = percentage) – także automatycznie akumuluję się w Gdp dla całego roku, bo Gdp całego roku jest równe sumię gdp kwartałow.
* Gdp Category (Calculations -> New Calculated Member -> Case Expression) – rozdziela Gdp Percentage na kategorie (Severe Recession, Recession, Stagnation / Weak Growth, Moderate Growth, Rapid Expansion)

Dla DimInflation:

* Dodałam wymiar jeszcze jak Measure Group.
* Inflation Percentage (Measures -> DimInflation -> New Measure -> Usage = Average over time (co stało możliwe dzięki ostatniemu kroku w punkcie 6.3) -> Source Column = percentage) – także automatycznie akumuluję się w Inflation dla całego roku lub kwartału, bo Inflacja całego roku / kwartału jest równe średniej inflacji każdego miesiąca.
* Inflation Category (Calculations -> New Calculated Member -> Case Expression) – rozdziela Inflation na kategorie (Deflation, Low (<1%), Target (1–3%), Moderate (3–5%), High (≥5%))

Dla FactSell:

* Sales Price Sum (Measures -> FactSell -> New Measure -> Usage = Sum -> Source Column = price).
* Sales Count (Measures -> FactSell -> New Measure -> Usage = Count of rows)
* Average Price – średnia cena:

IIF(

[Measures].[Sales Count] = 0,

NULL,

[Measures].[Sales Price Sum] / [Measures].[Sales Count]

)

* AvgPrice Category (Calculations -> New Calculated Member -> Case Expression) – rozdziela Average Price na kategorie (Ultra-Low (<50k), Budget (50-200k), Mid-Range (200-500k), Upper-Mid (500k-1M), High-End (1-5M), Luxury (≥5M))

# Projekt procesu ETL

# Specyfikacja procesów ETL (Control Flow + Data Flow)

Зображення, що містить схема, текст, ряд, знімок екрана

Вміст на основі ШІ може бути неправильним.

**Rysunek 12**. ControlFlow

Na rusynku 12 widzimy cały ControlFlow. Poszczwgólne etapy:

* czyszczenie poczekalni i tabeli błędów
* załadowanie danych do poczekalni
* budowanie wymiarów z danych z poczekalni lub zwykłym skryptem SQL (dla stałych kategorii new\_build, property\_type, tenure)
* załadowanie tabeli faktu (z walidacją oraz sprawdzeniem czy jest rekord duplikatem)

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 13**. DataFlow dla plików sell.csv. Dla gdp.csv i inflation.csv DataFlow wygląda analogicznie.

Na rusynku 13 jest pokazany DataFlow dla plików sell\_raw / inflation\_raw / gdp\_raw. Po prostu importujemy rekordy z csv i wrzucamy do poczekalni bez czyszczenia, konwertacji typów i walidacji.

Зображення, що містить текст, знімок екрана, Шрифт, число

Вміст на основі ШІ може бути неправильним. Зображення, що містить текст, знімок екрана, Шрифт, дизайн

Вміст на основі ШІ може бути неправильним.

**Rysunek 14-15**. DataFlow dla DimDate i DimLocation.

Poniżej jest opis DataFlow z rysunku 14 i 15.

DimDate:

* importujemy dane z poczekalni
* konwertacja string na db\_date
* dodanie kolumn year, month, day, quarter i quarter\_name
* dodanie kolumny month\_name przez skrypt C#
* zachowanie tylko unikalnuch dat przed lookupem
* lookup – zprawdzenie czy jest taka data w DimDate
* załadowanie unikalnych nowych dat do DimDate

DimLocation:

* importujemy dane z poczekalni
* Zamieniamy puste znaczenia na UNKNOWN COŚ (lub „-” dla locality i saon, ponieważ brak tych wartości jest brdziej oczekiwany i normalny)
* lookup – sprawdzenie czy jest taki adres w DimLocation
* załadowanie unikalnych nowych adresów do DimLocation
* miałam problem, że nawet unikalne adresy (przeszli lookup) nie dodawało do DimLocation, dlatego dodałam wyjście do sell\_error. Okazało się (po 8 godzinach męczenia ☹), że problem jest nie we flow, a full cache / patrial cache. Zostawiłam ten blok na przyszłość, jeżeli znowu będą błędy.

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Вміст на основі ШІ може бути неправильним. Зображення, що містить текст, знімок екрана, Шрифт, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 16-17**. DataFlow dla DimInflation i DimGdp

Poniżej jest opis DataFlow z rysunku 16 i 17.

DimInflation:

* załadowanie danych
* plik zawiera dane w postaciach rok, rok + kwartał i rok + miesiąc, z tego wybieramy tylko rok + miesiąc
* „2000 JAN” -> year = 2000, month\_abbr = „JAN”
* dodanie kolumny month (w postaci numeru)
* dodanie kolumny quarter i quarter\_name
* dodanie month\_name (skrypt C#)
* dookup
* dodanie unikalnych rekordów do DimInflation

DimGdp:

* załadowanie danych
* dodajemy kolumny year, quarter, quarter\_name
* lookup
* dodanie unikalnych rekordów do DimGdp

Зображення, що містить схема, План, Креслення, схематичний

Вміст на основі ШІ може бути неправильним.

**Rysunek 18**. DataFlow dla FactSell

Opis poszczególnych kroków zawartych na rysunku 18:

* załadowanie głównej tabeli z poczekalni
* zamiana typu price i date\_of\_transfer (string -> numeric / date)
* walidacja kolumn transaction\_id, price\_numeric, date\_clean, property\_type, new\_build, tenure; odrzucanie rekordów, gdzie ppd\_type != "A" (ponieważ są to transakcje niestandardowe, które nie odzwierciedlają normalne wzory działalności rynku nieruchomości i nie potrzebujemy ich do analizy.
* wszystkie niewalidne rekordy otrzymują kolumne error\_desc z opisem błądu (np., "InvalidTransactionId" lub "InvalidDate"
* we wszystkich zwalidowanych rekordach zamieniamy puste znaczenia w kolumnach adresu, jak robiliśmy to podczas ładowania adresów do DimLocation
* dalej są lookupy do każdego wymiaru, które sprawdzają czy jest taka data / adres / typ nieruchomości itd w tabelach wymiarów, w taki sposób otrzymujemy też klucze obce (wszędzie są to klucze naturalne, oprócz adresu, ponieważ klucz naturalny zawierałby za dużo kolumn i SQL Server wyświetlił warning:  
  *Warning! The maximum key length for a clustered index is 900 bytes. The index 'PK\_DimLocation' has maximum length of 1240 bytes. For some combination of large values, the insert/update operation will fail.*
* Wszystkie rekordy, które nie przeszli lookupy otrzymują opis błędu i też idą do tabeli sell\_error
* Pozostałe rekordy są ładowane do tabeli FactSell w HD.

# Analiza wizualna danych w kontekście pytań badawczych

* 1. Realizacja procesów analitycznych

Chcę zacząć od analizy średniej ceny na nieruchomości w poszczgólnych latach i jej zależności od wskaźników makroekonomicznych (inflacja i GDP).

Зображення, що містить текст, знімок екрана, Графік, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 19**. Average Price and Inflation Percentage by Year

Jeżeli by cena na nieruchomość była bardzo zależna od inflacji, to granatowa linija (inflacja) powtarzała by liniję zmiany średniej ceny (niebieskie słupki). Widzimy z wykresu, że to się nie sprawdza i wzrost cen nie jest bardzo zależny od inflacji, a rośnie o wiele szybciej niż inflacja.

Sprawdziłam też, czy może średnia cena bardziej zależy od jakichś zdarzeń, ale też na wielorocznej skali tego nie widać. Przykłady:

* 2007-2009 - Globalny kryzys finansowy, upadek Northern Rock – można powiedzieć, że widzimy to z wykresu, ponieważ w tych latach cena przestała rosnąć i nawet trochę spadła, ale też nie są to znaczące zmiany.
* 04.2013 - Uruchomienie programu Help to Buy (rządowe dopłaty do kredytów hipotecznych) – w 2013 i później nie widać zmiany trendu – cena równomiernie rośnie.
* 03.2020 - Pierwszy lockdown związany z COVID – i też globalnego wpływu nie widać

Może te i dużo innych zdarzeń wpłyneły na wzrost cen lub liczbę sprzedaży na jakieś konkretne typy nieruchomości lub lokaliacje, wrócimy do tego później.

Też możemy popatrzeć na ten sam wykres (Rysunek 20) , ale uszczególnony o kwartały.

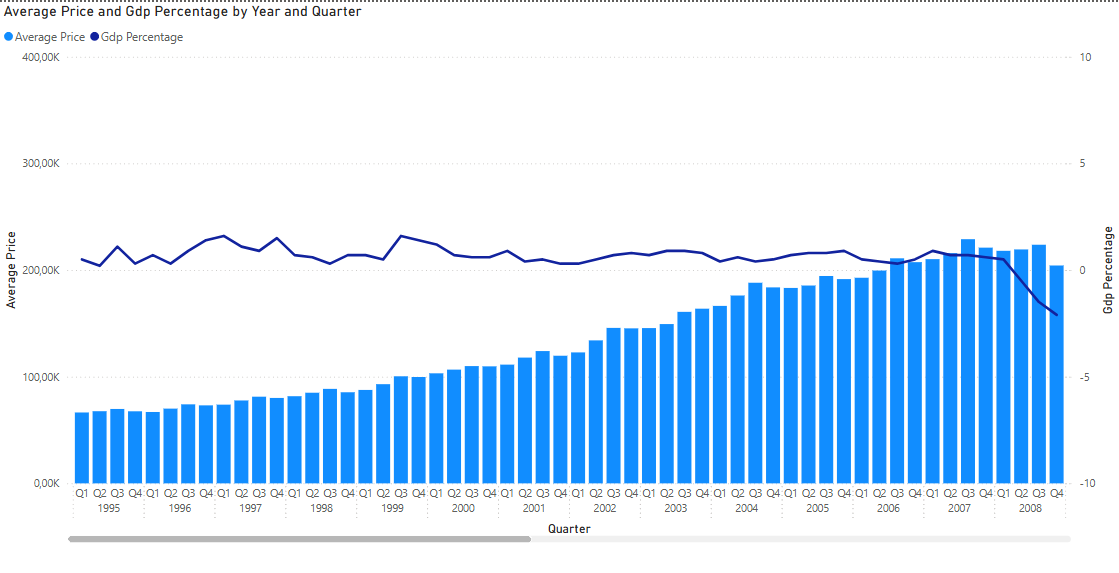
Зображення, що містить текст, Графік, знімок екрана, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 20**. Average Price and Inflation Percentage by Year and Quarter (2015-2025)

A teraz na rysunku 20 już widzimy większą zależność cen od inflacji na skali kwartałow. To znaczy, że wpływ inflacji na ceny ujawnia się przede wszystkim w krótkim horyzoncie: skoki CPIH z jednego–dwóch kolejnych kwartałów zaczynają przekładać się na dynamikę średnich cen z opóźnieniem rzędu właśnie 1-2 kwartałów. Innymi słowami – im bardziej zagęszczamy oś czasu, tym wyraźniej widać, że wzrost inflacji „podbija” tempo wzrostu cen niemal w czasie rzeczywistym, podczas gdy na wykresie rocznym efekt rozmywa się w uśrednianiu.

Dalej przechodzimy do analogicznej analizy zależności cen od GDP.



Зображення, що містить знімок екрана, Графік, ряд, текст

Вміст на основі ШІ може бути неправильним.

Зображення, що містить Графік, знімок екрана, ряд, схема

Вміст на основі ШІ може бути неправильним.

**Rysunek 21-23**. Average Price and GDP Percentage by Year and Quarter (1995-2008, 2008-2022, 2011-2025)

Zestawienie pokazuje, że:

* **Korelacja jest luźna.** Linie PKB (granatowa) i słupki średniej ceny nie poruszają się synchronicznie. Wzrosty PKB nie przekładają się natychmiast na ceny, a spadki gospodarcze – nawet głębokie – powodują jedynie krótkie „pauzy” w trendzie wzrostowym cen.
* **Reakcja z opóźnieniem.** Gwałtowne zmiany PKB odbijają się na rynku mieszkaniowym dopiero po 1-2 kwartałach. Widać to zarówno w kryzysie 2008-2009, jak i w szoku covidowym 2020.
* **Ceny rosną szybciej niż gospodarka.** Nawet przy umiarkowanym lub zerowym wzroście PKB (lata 2010-2019) średnia cena kontynuuje marsz w górę. Wskazuje to, że długoterminowo ważniejsza jest dostępność kredytu i chroniczny niedobór podaży, a nie bieżący cykl koniunkturalny.

**Wniosek:** PKB wpływa na ceny mieszkań, ale efekt jest łagodny i opóźniony; czynniki finansowe i strukturalne mają większe znaczenie niż krótkookresowe wahania gospodarki.

P.S.: GDP w 2020 mógł być lagiem, ale sprawdziłam informacje i rzeczywiście tak było. Zasadniczo rok 2020 był rokiem spadku gospodarczego dla Wielkiej Brytanii, podczas gdy w 2021 r. nastąpiło znaczne ożywienie, choć z pewnymi miesięcznymi wahaniami.

Teraz chcę przeanalizować liczbę sprzedaży (nie koncentrując na cenach) i ich zależność od wskaźników makroekonomicznych.

Зображення, що містить текст, знімок екрана, Графік, схема

Вміст на основі ШІ може бути неправильним.

**Rysunek 24**. Sales Count and Inflation Percentage by Year (ujęcie roczne, 1995-2025)

**Co widać na wykresie**

* **1995-2007 – długi boom hipoteczny**  
  Liczba sprzedaży rośnie z ~0,8 mln do rekordowych **1,4 mln w 2006-2007**, mimo że inflacja systematycznie spada (z 4-5 % do 2 %). Wniosek: **podaż taniego kredytu i programy wspierające zakup były ważniejsze niż CPI**.
* **2008-2009 – kryzys finansowy**  
  Słupki spadają prawie o połowę do ~0,65 mln; inflacja podbija do 4-5 %, ale **rynek zamiera przede wszystkim z powodu niepewności i restrykcyjnej polityki banków**, nie przez wzrost cen towarów.
* **2010-2019 – płaskowyż ze słabnącą inflacją**  
  Obroty wahają się w przedziale 0,65-0,95 mln, a inflacja z każdym rokiem niższa; relacja między obiema seriami pozostaje luźna – niższe CPI nie pociąga za sobą proporcjonalnie większej liczby transakcji.
* **2020 – COVID-19**  
  Lockdown wycina liczbę sprzedaży (<0,2 mln), mimo że CPI spada niemal do zera. **Szok podażowo-transakcyjny, nie inflacyjny.**
* **2021-2022 – odbicie i pik inflacyjny**  
  Transakcje skaczą do 1,1 mln (stamp-duty holiday), inflacja z opóźnieniem eksploduje do ~8 %. Tu widzimy, że **wzrost CPI przychodzi PO skoku sprzedaży**, a sam wzrost cen towarów szybko studzi popyt – w 2023 wracamy do ~0,7 mln transakcji.

**Wnioski**

1. **Zależność odwrotna** – rynek reaguje silniej na czynniki kredytowe i regulacyjne niż na poziom inflacji; wysokie CPI w 2022 koreluje z spadkiem, a nie wzrostem sprzedaży.
2. **Opóźnienie efektów** – każde makro-uderzenie (kryzys 2008, COVID 2020, inflacja 2022) przenosi się na liczbę transakcji niemal natychmiast, natomiast ceny i wolumen wracają na ścieżkę trendu po 4-6 kwartałach.
3. **Do dalszej analizy** – które segmenty (new-build vs resale, London vs regiony) najbardziej hamują przy wysokiej inflacji i jak zmienia się struktura typu nieruchomości podczas takich szoków.

Зображення, що містить текст, знімок екрана, Графік, схема

Вміст на основі ШІ може бути неправильним.

**Rysunek 25**. Sales Count and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **New Build**)

Зображення, що містить текст, знімок екрана, Графік, схема

Вміст на основі ШІ може бути неправильним.

**Rysunek 26**. Sales Count and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **Resale**)

Tabela 6. Sprzedaż new-build vs resale na tle CPIH (1995-2025)

|  |  |  |
| --- | --- | --- |
| **Segment** | **Kluczowe obserwacje** | **Wniosek** |
| **New-build (Rys. 25)** | • Skala 70-140 k rocznie (≈ 7-10 % rynku). • Silny wzrost 2001-2007 → tani kredyt, nie inflacja. • Help-to-Buy 2013 podniósł wolumen +35 %. • Inflacja 2022-23 (8 %) i drogie materiały ↓ sprzedaż ~-30 %. | Budownictwo pierwotne **mocno reaguje na koszty i programy rządowe**. |
| **Resale (Rys. 26)** | • Dominuje: 0,8-1,3 mln transakcji. • 2010-2019 stabilnie mimo wahań CPI. • Lockdown 2020 duży spadek, ale szybkie odbicie 2021 (SDLT). • Inflacja 2022 ↓ wolumen, lecz płycej niż w new-build | Rynek wtórny **bardziej odporny** – właściciele mogą przeczekać drogi kredyt. |

**Podsumowanie**

* Inflacja wpływa na sprzedaż **z opóźnieniem ~1 roku**; szoki kosztowe uderzają w new-build szybciej i mocniej.
* Czynniki polityki kredytowej (Help-to-Buy, SDLT holiday) napędzają wolumeny bardziej niż sama CPIH.

Зображення, що містить текст, знімок екрана, Графік, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 27**. Average Price and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **New Build**)

Зображення, що містить текст, знімок екрана, Графік, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 28**. Average Price and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **Resale**)

Tabela 7. Średnia cena new-build vs resale na tle CPIH (1995-2025)

|  |  |  |
| --- | --- | --- |
| **Segment** | **Kluczowe obserwacje** | **Wniosek** |
| **New-build (Rys. 27)** | • **1995-2007 :** ceny startują nisko (~70 k £) i rosną szybciej niż inflacja – do >200 k £; boom napędzany taniym kredytem i popytem na nowoczesne domy.  • **2008-2009 :** spadek ~-10 %, ale krótszy niż spadek sprzedaży – deweloperzy częściej wstrzymują podaż niż obniżają ceny.  • **Po 2013 (Help-to-Buy)** – tempo wzrostu przyspiesza; w 2021 przebija 350 k £.  • **Inflacja 2022** – ceny nadal zwyżkują, choć wolniej; realny koszt budowy przerzucany na kupujących. | Ceny pierwotne są **bardziej dynamiczne i wrażliwe na politykę** (subwencje, koszty materiałów) niż na sam CPIH; deweloperzy dostosowują wolumen, a nie poziom cen. |
| **Resale (Rys. 28)** | • **1995-2007 :** łagodniejsza ścieżka – z 70 k £ do ~210 k £, rosnące równolegle z new-build, lecz bez „skoków”.  • **Kryzys 2008-09** – korekta ~-15 % rozłożona na 2 lata; inflacja 3-4 % nie zatrzymuje spadku realnej wartości.  • **2010-2019** – stabilny trend +4-6 % rocznie, niezależnie od niskiego CPIH.  • **2020-2023** – ceny nadal idą w górę mimo pandemii; pik inflacji 2022 spowalnia dynamikę, ale nie odwraca trendu. | Rynek wtórny **wolno koryguje**; właściciele wstrzymują sprzedaż zamiast obniżać cenę. Inflacja działa głównie przez stopy procentowe i koszty kredytu, z opóźnieniem 1-2 lat. |

**Podsumowanie**

* **New-build** drożeje szybciej i reaguje bezpośrednio na koszty budowy oraz programy wsparcia; inflacja przechodzi tu w ceny niemal od razu.
* **Resale** zachowuje się bardziej inercyjnie – ceny spadają tylko przy silnych kryzysach finansowych, a wysokie CPIH przekłada się raczej na mniejszy wolumen niż na niższe ceny.

Зображення, що містить текст, знімок екрана, ряд, Графік

Вміст на основі ШІ може бути неправильним.

**Rysunek 29**. Average Price and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **Freehold**)

Зображення, що містить текст, знімок екрана, Графік, ряд

Вміст на основі ШІ може бути неправильним.

**Rysunek 30**. Average Price and Inflation Percentage by Year (ujęcie roczne, 1995-2025, **Leasehold**)

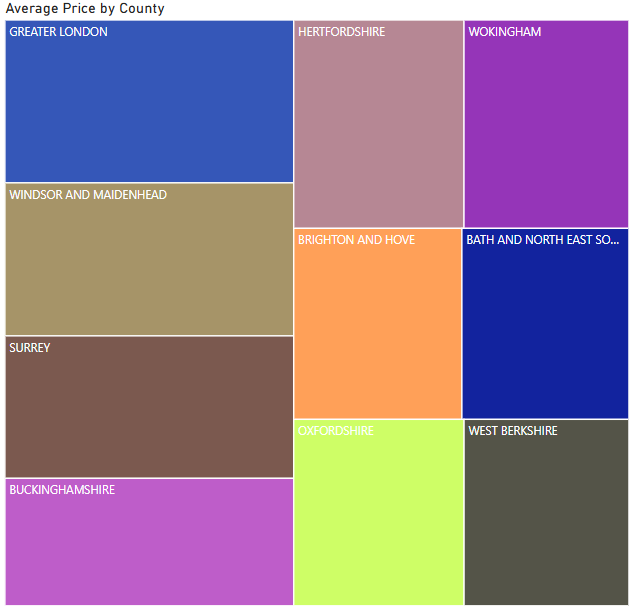
Zmiany obu serii (rysunki 29 i 30) są prawie równoległe:

* **Poziom** – Freehold jest konsekwentnie droższy (ok. +20-25 %) od Leasehold, ale kształt słupków pozostaje ten sam.
* **Trend 1995-2021** – płynna, niemal liniowa ścieżka wzrostu; ani reformy ground-rent, ani wahania CPIH nie powodują odchyłek między segmentami.
* **Inflacja 2022-23** – skok CPIH odbija się na obu segmentach identycznie: dynamika cen lekko wyhamowuje, lecz Freehold i Leasehold zachowują dotychczasową różnicę poziomów.

**Wniosek:** Forma własności wpływa na *poziom* ceny, natomiast **sam trend i reakcja na inflację są praktycznie identyczne** – regulacje leasehold nie zaburzają ogólnego ruchu rynku.

Ponieważ już na wielu wykresach było widać, że zależność rynku od wskaźników makroekonomicznych jest luźna, chcę przejść do kolejnych tematów.

Dalej chciałabym popatrzeć, gdzie w WB są najwyższe ceny, gdzie najwięcej sprzedaży i tak dalej. Tak otrzymamy wstępną analizę, w jaką nieruchomość i w jakiej lokalizacji warto inwestować. Do bardziej szczegółowej analizy potrzebowałabym dodatkowe wyliczone miary, niestety nie mam ich i nie mam możliwości już dodać



**Rysunek 31**. Average Price by County (2018-2025, **Top-10**)

Widzimy 10 hrabstw z najwyższą średnią ceną za 2018-2025 lata.

1. **Greater London** dominuje – największy kafel, wyraźnie odstaje cenowo od reszty kraju.
2. Za nim ścisły „pas premium” południowego wschodu: **Windsor & Maidenhead**, **Surrey**, **Buckinghamshire**, **Hertfordshire** – wszystkie w zasięgu dojazdu do stolicy.
3. **Brighton and Hove** i **Bath and North East Somerset** – przykłady drogiej urbanistyki historycznej poza Londynem.
4. **Wokingham**, **Oxfordshire**, **West Berkshire** – kompletują pierwszą dziesiątkę, potwierdzając przewagę regionu South East.

Układ kafli pokazuje, że **9 z 10 najdroższych hrabstw leży w promieniu ~100 km od Londynu**; wyjątkiem pozostaje Bath, co podkreśla koncentrację wartości nieruchomości wokół stolicy i głównych ośrodków akademickich / turystycznych.

Початок форми

Зображення, що містить текст, знімок екрана, Електрик синій, Прямокутник

Вміст на основі ШІ може бути неправильним.

**Rysunek 32**. Average Price by County, District, Town, Locality and Street (2018-2025)

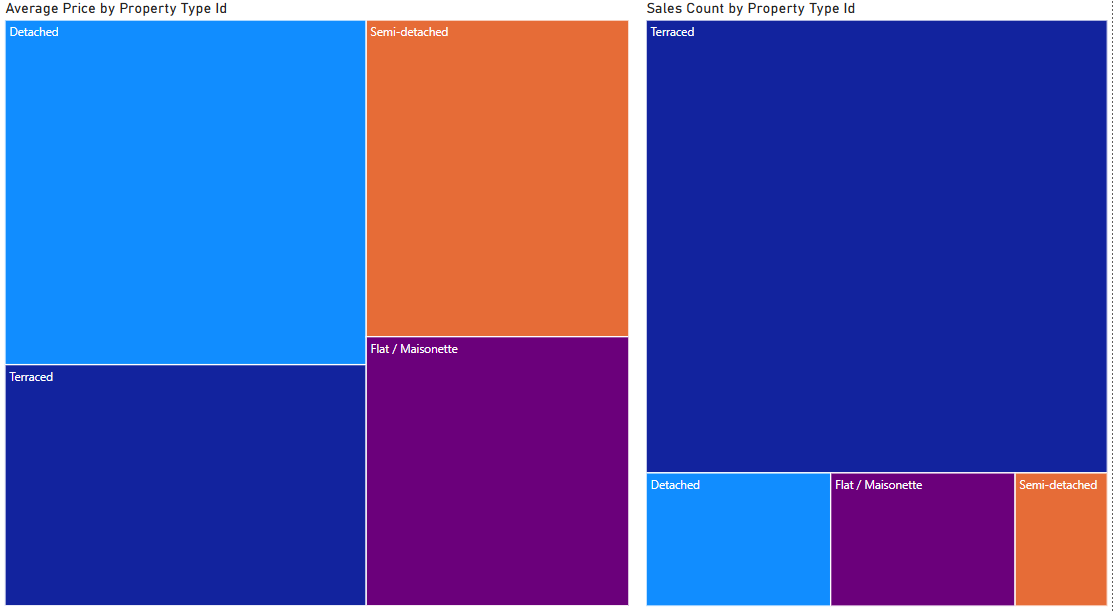
Po zejściu na najniższy poziom agregacji widać, że rynek „rozpada się” na pojedyncze, ekstremalnie drogie mikrolokalizacje:

* **THE VALE (Kensington & Chelsea)** zajmuje prawie cały obszar wykresu ­— to ulica z portfelem kilku rezydencji powyżej 20 mln £ każda.
* **MARKHAM STREET** to kolejny adres w tej samej dzielnicy, którego średnia cena również znacząco przewyższa nawet londyńską medianę.

Wniosek: **im głębiej wchodzimy w hierarchię lokalizacji, tym większe dysproporcje cenowe** — pojedyncze ulice w prime Central London potrafią „przykryć” całe hrabstwa o przeciętnych cenach. Taki drill-down dobrze pokazuje, że analiza inwestycyjna musi uwzględniać nie tylko region czy miasto, ale często konkretny kod pocztowy, a nawet pojedynczy ciąg zabudowy.

Link do strony Wikipedii o ulicy The Vale:  
<https://en.wikipedia.org/wiki/The_Vale,_Chelsea>

Został niewykorzystany w analizie wymiar typu nieruchomości i na rysunku 34 jest podział średniej ceny i wolumen na typy nieruchomości.



**Rysunek 33 a, b**. Average Price by Property Type (a) i Sales Count by Property Type (b) (2018-2025)

**Obserwacje:**

• **Detached** – zdecydowanie najdroższy segment (≈ 1,5 × średniej rynku).   
• **Semi-detached i Terraced** układają się schodkowo niżej; bloki (Flat/Maisonette) najtańsze.

**Terraced** dominuje pod względem liczby sprzedaży – ponad połowa rynku wtórnego.   
• **Detached** mimo najwyższej ceny odpowiada za ułamek transakcji.   
• **Flats i Semi-detached** plasują się pomiędzy.

**Szybkie podsumowanie**

* **Cena rośnie wraz z metrażem/ogrodem (Detached > Semi > Terraced > Flat)** – oczywisty, ale tutaj jasno wizualny gradient.
* **Popyt skupia się na zabudowie szeregowej (Terraced)** – kompromis między kosztem a powierzchnią; ulice z taką zabudową są motorem rynku.
* **Segment Detached to nisza premium** – najmniej transakcji, ale największy udział wartościowo, co czyni go atrakcyjnym dla inwestorów szukających wysokiego kapitału per nieruchomość, lecz z niższą płynnością.
  1. Podsumowanie - wnioski z analizy (Pytania badawcze + odpowiedzi)

**Q01 - Średnia cena vs inflacja/PKB**

*Rys. 20, 24* pokazują, że:   
▪ korelacja z CPIH i PKB w skali **rocznej** jest słaba; ceny rosną szybciej niż wskaźniki makro.  
▪ w skali **kwartału** widać umiarkowany, opóźniony (1-2 kw.) wpływ – szok inflacyjny 2022 spowalnia wzrost cen dopiero w 2023.  
▪ PKB oddziałuje jeszcze słabiej; kryzysy (2009, 2020) wywołują tylko krótkie spłaszczenie trendu.

**Q02 - Wolumen transakcji vs inflacja/PKB**

*Rys. 22* ujawnił, że **sprzedaż reaguje szybciej i mocniej** niż ceny:   
▪ kryzys 2008 (-45 % transakcji) i lockdown 2020 (>-70 %) ‒ od razu po spadku PKB.  
▪ wysoka inflacja 2022 redukuje wolumen już w 2023, choć ceny wciąż rosną.   
→ wolumen jest dobrym wczesnym wskaźnikiem koniunktury; CPIH działa głównie przez stopy i dostępność kredytu.

**Q03 - Wpływ makro na segmenty**

*Rys. 23, 24, 28-29*   
▪ **New-build**: najmocniej odczuwa koszty (inflacja materiałów) i politykę (Help-to-Buy) – spadek wolumenu -30 % przy CPIH 8 %.  
▪ **Resale**: wolniejsza reakcja; właściciele wstrzymują sprzedaż, nie obniżają cen.  
▪ **Detached vs Flat**: droższe segmenty rosną szybciej po COVID („race for space”).  
▪ Leasehold i Freehold reagują podobnie – różnią się jedynie poziomem cen.

**Q04 - Najdroższe lokalizacje**

*Rys. 31 & 32*   
▪ TOP-10 hrabstw cenowych to prawie wyłącznie **Greater London + South-East** (Surrey, Windsor & Maidenhead, Bucks…).  
▪ Drill-down do ulic wskazuje ultra-premium hot-spoty jak **The Vale** i **Markham Street** w Kensington & Chelsea, gdzie pojedyncze transakcje przekraczają 20 mln £.

**Q05 - Najdroższe / najtańsze i najpopularniejsze typy**

*Rys. 33*   
▪ **Najdroższe**: Detached (> 1,5× średniej).  
▪ **Najtańsze**: Flat/Maisonette.  
▪ **Najpopularniejsze (wolumen)**: Terraced (ponad połowa wszystkich sprzedaży).  
→ rynek masowy skupia się na zabudowie szeregowej; segment premium to domy wolnostojące o niskiej płynności.

# Wnioski końcowe z realizacji projektu

* 1. Problemy związane z realizacją projektu HD

Im dalej – tym łatwiej, na czszęście. Na początku źle rozumiałam się na danych, które z nich będą przydatnie i nie rozumiałam dobrze, jak ma wyglądać stuktura mojej HD. Teraz to rozumiem i mam gotową hurtownie o dobrzej strukturze z Dimentions i FactTransaction.

Było dużo błędów z kodowaniem (1250, 1251, UTF-8 itd), dlatego w końcu wszędzie korzystałam z Unicode.

Proces ETL w SSIS był prostrzy, niż oczekiwałam, ale i tak zajął ~5 razy więcej czasu. Kategoryzacja (niska/wysoka inflacja i gdp, niska/średnia/wysoka cena itd) zostanie zrobiona na etapie kostki, ponieważ planuję dopasowywać kategorie do rozkładu danych.

Stworzenie kostki też było łatwe, ale trzeba zrobić na nowo, ponieważ zmieniłam nazwy tabel i trochę strukturę wymiarów (rozbudowałam).

Na laboratorium robiłam dużo analiz wirtualnych i oczekuję, że zrobię to szybko i dobrze.

* 1. Pozyskana wiedza i doświadczenie

W trakcie realizacji projektu zdobyłam bardzo dużo praktycznej wiedzy dotyczącej budowy hurtowni danych i całego procesu ETL. Po raz pierwszy miałam okazję pracować z dużym, prawdziwym zbiorem danych (ponad 30 milionów rekordów) i zrozumiałam, jak wygląda rzeczywista praca z taką skalą – nie tylko pod względem analizy, ale również pod względem technicznym (wydajność, pamięć, błędy w SSIS, kodowanie znaków itd.).

Nauczyłam się:

* tworzyć i normalizować strukturę hurtowni danych zgodnie z zasadami modelu gwiazdy (fakt + wymiary),
* tworzyć oraz zarządzać stagingiem („poczekalnią”) i surowymi danymi (\*\_raw),
* rozdzielać dane na poprawne i błędne z automatyczną klasyfikacją błędów,
* wykorzystywać SSIS do tworzenia dynamicznych i wieloetapowych przepływów danych (Control Flow, Data Flow),
* pracować z Lookupami, Derived Columnami, warunkami błędów i wieloma formatami danych (CSV, XLS),
* efektywnie walczyć z problemami kodowania i formatów dat.

# Doświadczenia i rekomendacje dotyczące korzystania z AI na zajęciach z Hurtowni Danych

ChatGPT bardzo pomagał w naprawieniu błędów i z opisaniem etapów ETL w SSIS.  
A potem też bardzo dużo szkodził. Czasami jak nie zna odpowiedzi, godzinami chodzi w kółko.

# Źródła informacji użyte w etapie analizy danych

ChatGPT

https://www.ons.gov.uk/economy