UNIWERSYTET GDAŃSKI – WYDZIAŁ EKONOMICZNY

Błażej Stelmarski

Numer albumu: **286236**

Kierunek studiów: **EKONOMIA**

ANALIZA RYNKU AUT OSOBOWYCH W POLSCE Z WYKORZYSTANIEM NARZĘDZI BUSINESS INTELLIGENCE

Praca magisterska wykonana
w Katedrze Ekonomii
Międzynarodowej i Rozwoju
Gospodarczego
pod kierunkiem
prof. UG dr hab. Stanisława
Umińskiego

Sopot 2024

Spis treści

Streszczenie pracy

Celem niniejszej pracy jest dostarczenie odpowiedzi na kluczowe pytania dotyczące rynku motoryzacyjnego w Polsce, koncentrując się w szczególności na segmencie aut osobowych. Analiza ta opiera się na danych z maja 2021 roku, co jest istotne, ponieważ ten okres przypadał na czas jeszcze wciąż trwającej pandemii Covid-19. Pandemia miała znaczący wpływ na wiele aspektów życia społeczno-gospodarczego, w tym na rynek motoryzacyjny. W związku z tym, wnioski i wyniki przedstawione w tej pracy mogą nie odzwierciedlać obecnej sytuacji rynkowej w 2024 roku.

Pandemia Covid-19 wpłynęła na zakłócenia w łańcuchach dostaw, zmiany w zachowaniach konsumentów oraz na różnorodne strategie marketingowe i produkcyjne firm motoryzacyjnych. Warto podkreślić, że w czasie pandemii wiele osób unikało korzystania z transportu publicznego, co mogło zwiększyć zainteresowanie zakupem samochodów osobowych. Z drugiej strony, niepewność ekonomiczna mogła skłonić część konsumentów do odłożenia planów zakupu nowych pojazdów.

Chociaż dane z 2021 roku są cennym źródłem informacji, ich analiza może nie być w pełni miarodajna dla obecnej sytuacji na rynku. Niemniej jednak, praca ta stanowi ważny wkład w zrozumienie dynamiki rynku motoryzacyjnego w czasie pandemii i może służyć jako punkt odniesienia do porównań z danymi z lat późniejszych. Porównanie stanu rynku motoryzacyjnego z maja 2021 roku ze stanem obecnym w 2024 roku mogłoby dostarczyć cennych informacji na temat kierunków rozwoju oraz ewentualnych zmian w preferencjach konsumentów i strategiach rynkowych. Taka analiza byłaby z pewnością ciekawa i inspirująca, otwierając nowe perspektywy dla badań nad rynkiem motoryzacyjnym w Polsce.

Master‘s thesis summary

This study aims to provide answers to key questions regarding the automotive market in Poland, with a particular focus on the passenger car segment. The analysis is based on data from May 2021, which is significant as this period coincided with the ongoing Covid-19 pandemic. The pandemic had a substantial impact on many aspects of socio-economic life, including the automotive market. Therefore, the conclusions and results presented in this study may not reflect the current market situation in 2024

The Covid-19 pandemic led to disruptions in supply chains, changes in consumer behavior, and various marketing and production strategies of automotive companies. It is worth noting that during the pandemic, many people avoided using public transportation, which may have increased interest in purchasing passenger cars. On the other hand, economic uncertainty may have prompted some consumers to postpone their plans to buy new vehicles.

Although the data from 2021 is a valuable source of information, its analysis may not be fully indicative of the current situation in the market. Nonetheless, this study provides an important contribution to understanding the dynamics of the automotive market during the pandemic and can serve as a reference point for comparisons with data from later years. Comparing the state of the automotive market in May 2021 with the current state in 2024 could provide valuable insights into development trends and potential changes in consumer preferences and market strategies. Such an analysis would undoubtedly be interesting and inspiring, opening new perspectives for research on the automotive market in Poland.

Wstęp

Niniejsza praca ma na celu zbadanie struktury rynku aut osobowych w Polsce, według stanu z maja 2021 roku, ponieważ wykorzystana w niej baza danych jest fotografią tego czasu. Jest to temat ciekawy, choć z pozoru dość oczywisty, mimo to warto poświęcić mu więcej czasu. W dzisiejszych czasach kwestie motoryzacji, jej przyszłości, tak od strony technologicznej, jak i makroekonomicznej czy geopolitycznej, są niezmiernie istotnym elementem debaty publicznej w krajach wszystkich szerokości geograficznych.

Metodą wykorzystaną w pracy jest analiza przedstawiona w formie dashboardu managerskiego, czyli narzędzia wykorzystującego wizualizacje danych do ustrukturyzowanej opowieści o danych wykorzystanych w jej trakcie. Ponieważ poruszane tematy są w gruncie rzeczy badaniami regionalnymi, zatem analizą przestrzenną rozkładu popularności aut, korelacji ich ceny bądź liczby z innymi czynnikami, itp. w poszczególnych województwach, to należy pamiętać o tym charakterze pracy przy weryfikacji hipotez. W toku prac zweryfikowano kilka z nich, postawionych jeszcze na początku analizy. Aby lepiej wytłumaczyć logikę stojącą zarówno za tą bazą, jak i jej analizą, można wyobrazić sobie siebie w roli np. importera aut. Wtedy hipotezy, będą przypuszczalnymi odpowiedziami na pytania, które ów importer musi sobie zadać, aby zdecydować czy warto podjąć ekonomiczne ryzyko pewnych działań, czy też skierować swe wysiłki w innych kierunkach. Jako że rozpatrywane są sytuacje w podziale na województwa, należy najpierw ustalić określone zależności. W pracy postawiono następujące hipotezy:

- popularność kolorów oferowanych aut w poszczególnych województwach różni się między sobą w sposób zasadniczy,
- napęd auta zależy bardziej od ceny auta go posiadającego, a nie od województw, w których najlepiej jest takowe auto użytkować z racji dużej ilości piaszczystych dróg,
- gusta Polaków w zakresie typów pojazdów również różnią się przestrzennie, w podziale na wschód i zachód kraju,
- im bliżej niemieckiej granicy, tym większa popularność niemieckich marek aut,

- im bogatszy region, co znajduje odzwierciedlenie między innymi w średnim wynagrodzeniu, bądź też dochodzie rozporządzalnym, tym droższe auta są tam oferowane,
- im więcej ludności w jednym miejscu, tym większa koncentracja kapitału, tak ludzkiego, jak i pieniężnego, co też wpływa na wzrost średniej ceny auta,
- z drobnymi wyjątkami, im nowsze, mniej zużyte i lepsze technologicznie auto, tym wyższa jego cena,
- zwiększony dostęp do kolei, jak również jakość jej oferty skierowanej do mieszkańców danego regionu, wpływa na zmniejszenie liczby zarejestrowanych aut,
- im lepszy jest dostęp do dobrej jakości dróg, tym większa liczba aut.

Rozdział pierwszy opowiada o sytuacji na współczesnym rynku automotive w Polsce, jest streszczeniem obecnych trendów na rynku oraz wyzwań jakie przed nim stały i stoją nadal.

Drugi rozdział poświęcono wybranym zagadnieniom związanym z bazami danych i analityką danych, temu jak wyglądają czy też powinny wyglądać modele baz danych, jak kształtuje się ich historia na przestrzeni lat oraz jaka czeka ich przyszłość. Można zapoznać się z ich możliwościami i rozwiązaniami technicznymi rewolucjonizującymi analizę danych, już na zawsze zmienioną dzięki wynalezieniu relacyjnego modelu baz danych.

Trzeci rozdział łączy wcześniej streszczoną teorię z praktyką bazodanową, zatem opisuje jak przebiegał proces przygotowania danych do analizy z wykorzystaniem narzędzi takich jak Excel, SQL Server, Power Query oraz Power BI. Z racji wykorzystania w pierwszej kolejności plików płaskich, tj. takich, które nie są powiązane kluczami z innymi plikami, nie posiadają żadnych relacji z innymi plikami, na wstępnym etapie pracy Excel towarzyszył niemal nieustannie autorowi pracy. Jednak następnie po wyczerpaniu się możliwości dalszego korzystania z niego, wykorzystano do pracy SQL Server. Jest to niezwykle wydajne narzędzie, jednocześnie proste w początkowej obsłudze, a komplikujące się dopiero na kolejnych etapach pracy. Na koniec, kiedy już zasadnicze poprawki zostały wdrożone w życie, nastąpiło umieszczenie odpowiednich baz danych w Power BI, ostateczne poprawki w Power Query i dalej wykorzystywane były do wizualizacji danych.

Na tym właśnie skupia się rozdział czwarty, jest on mianowicie opisem powstałego w toku prac, dashboardu managerskiego, który w atrakcyjny sposób służy końcowemu użytkownikowi zarówno w eksploracji danych samemu, jak też dostarcza przetworzone już dane, zamienione w konkretną informację. Przygląda się on z bliska postawionym hipotezom i umożliwia ich weryfikację. Pozostałe konkluzje z pracy zostały podsumowane w zakończeniu.

Rozdział I. Rynek motoryzacji osobowej w Polsce

Rynek automotive w Polsce mimo wielu zachodzących zmian nie kurczy się, a raczej przyrasta w tempie podobnym do ogólnego wzrostu gospodarczego. Nadal rośnie liczba zarejestrowanych pojazdów, zwiększa się także udział nowych pojazdów w zarejestrowaniach. Jeszcze kilka lat temu wprowadzane były na rynek w dużo mniejszej liczbie. Coroczna analiza Polskiego Związku Przemysłu Motoryzacyjnego, wskazuje, że w kategorii pojazdów osobowych w maju 2023 odnotowano 38 575 rejestracji, czyli aż o 7,5 % więcej niż rok wcześniej, co daje wzrost 2 678 szt. r/r. Wspominane nowe pojazdy to odzwierciedlenie ogólnego, europejskiego trendu. Przykładowo, w marcu 2023 r. zarejestrowano prawie 30% więcej nowych pojazdów, w porównaniu do tego samego miesiąca poprzedniego roku. Te trendy są możliwe do określenia, bez wnikliwych analiz rynku, tak jak rosnąca liczba pojazdów elektrycznych i hybrydowych. Dane te potwierdza Europejskie Stowarzyszenie Producentów Samochodów. Zmieniają się także wybory konsumenckie. Typowy klient coraz częściej sięga po auto bardzo dobrze wyposażone, które można określić mianem klasy premium. Analiza IBRM Samar wskazuje, że w listopadzie 2022 roku zarejestrowano 7 451 nowych pojazdów klasy premium, gdzie porównując dane z listopada 2021, liczba ta jest większa o prawie 14 %. Porównując pierwsze cztery miesiące 2023 roku do analogicznego okresu w poprzednim roku, wzrost również jest bardzo widoczny i wyniósł ponad 12 %.

Należy uwzględnić te statystyki, by pokazać punkt odniesienia, w którym znajduje się branża motoryzacyjna, czy też szerzej, jak wygląda struktura gustów konsumenckich na tym rynku w Polsce. Pośrednio może to też zobrazować kondycję gospodarczą Polski. Według danych przedstawianych przez firmę KPMG, coraz więcej Polaków staje się konsumentami dóbr luksusowych. Tę grupę społeczną możemy sklasyfikować jako osoby, których zarobki przekraczają miesięcznie 50 tys. złotych. Ich liczba wzrosła między 2019 rokiem do 2020 aż o 11, 6 %, do 77 tys. osób [1].

Ważnym aspektem tej pracy jest analiza branży samochodów osobowych w Polsce, dlatego nie sposób pominąć pewnych trendów, które w krytyczny sposób wpływają na otaczającą nas rzeczywistość. Do głównych wydarzeń kształtujących rynek, z pewnością można zaliczyć inwazję Rosji na Ukrainę, która spowodowała dodatnie saldo migracji, a także spowodowała wzrost liczby konsumentów na rynku. Bardzo wysoka inflacja

spowodowana m.in. wzrostem cen energii elektrycznej, a także niedobory surowców na rynku, zachwianie globalnym łańcuchem dostaw spowodowały istotne wydłużenie czasu oczekiwania na dostawę aut, do nawet kilku miesięcy.

1.1. Elektromobilność

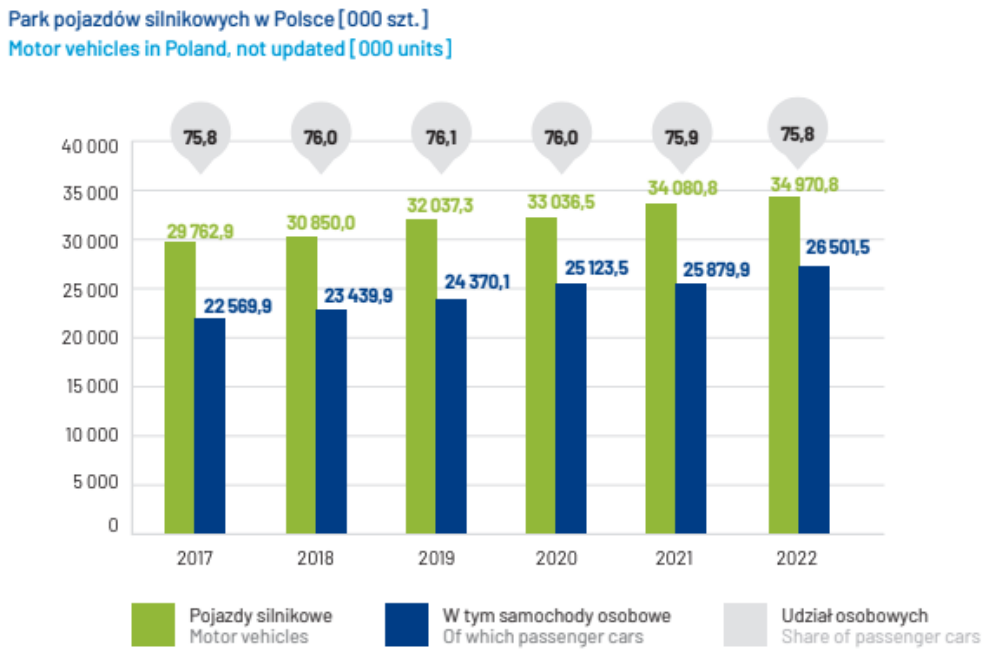
Dużo na rynku zaczyna się zmieniać, ze względu na wzrost popularności samochodów elektrycznych. Jak wskazują analizy przedstawione w „Liczniku Elektromobilności" Polskiego Związku Przemysłu Motoryzacyjnego oraz Polskiego Stowarzyszenia Nowej Mobilności, w maju 2023 zarejestrowanych było ponad 42 tys. samochodów całkowicie elektrycznych, co stanowi wzrost o 71 % w stosunku do okresu pierwszych pięciu miesięcy, porównując rok 2023 do 2022. Powyższa dynamika zmian, może w przyszłości zmienić układ sił na szeroko pojętym rynku automotive w Europie. Już teraz widzimy zaostrzającą się walkę chińskich producentów, którzy szukają sposobów na skuteczne wejście na rynek Unii Europejskiej. Niestety, spóźniona transformacja niemieckich marek, jak również podważenie wiarygodności grupy Volkswagen w 2015 roku aferą dieselgate, już teraz powoduje upadek znanych firm, które od dziesiątek lat produkują podzespoły do aut pojazdów takich jak Skoda, Volkswagen czy Mercedes-Benz. W 2023 roku została złamana pewna reguła, porządek panujący od lat na światowym rynku. Chińskie koncerny wyeksportowały 3 mln samochodów, a niemieckie koncerny 2,6 mln. Ta zmiana może wywołać wiele negatywnych skutków dla firm z regionu Europy Środkowo- Wschodniej.

Co sprawia jednak, że konsumenci wybierają coraz częściej pojazdy z alternatywnym źródłem napędu? Według raportu Deloitte, Polska. Global Automotive Consumer Study 2024 [2], największą motywacją do zakupu pojazdów elektrycznych, jest pogląd, że koszty paliwa, okażą się znacząco niższe. Kolejnymi kwestami najczęściej poruszanymi przez respondentów badania, była troska o środowisko naturalne, warunki jazdy, czy potencjalny zakaz sprzedaży nowych pojazdów spalinowych. Dopiero na kolejnych miejscach znalazły się rządowe zachęty czy programy stymulacyjne, a także możliwość wprowadzenia dodatkowego opodatkowania/opłat od pojazdów spalinowych. Konsumenci są także świadomi pewnych ograniczeń czy wad, jakie mogą charakteryzować pojazdy całkowicie elektryczne. W tym samym badaniu ankietowani odpowiedzieli, że najbardziej martwią się o czas potrzebny na ładowanie baterii, koszt

zakupu, zasięg jazdy oraz obawy związane z bezpieczeństwem eksploatowanej baterii. Nie powoduje to jednak zmniejszenia „elektryków” na rynku, a wręcz przeciwnie - rozpędzanie się trendu kupowania samochodów stricte elektrycznych.

1.2. Park samochodowy

Najbardziej dokładną analizę w kontekście struktury samochodów osobowych można odnaleźć w „Raporcie Branży Motoryzacyjnej Polskiego Związku Przemysłu Motoryzacyjnego 2023/2024”. Bazuje on na danych Centralnej Ewidencji Pojazdów [3]. W 2022 roku w Polsce zarejestrowanych było najwięcej od lat pojazdów silnikowych, bo aż 34 970,8 tys., z czego aż 26 501,5 tys. to pojazdy osobowe. Trend wzrostowy utrzymuje się przynajmniej od 2017 roku, co wskazano na rys.1.



Rys.1. Park pojazdów silnikowych w Polsce w latach 2017-2022, źródło: Raport Branży Motoryzacyjnej Polskiego Związku Przemysłu Motoryzacyjnego 2023/2024

Statystki potwierdzają tezę, że Polacy w znacznej mierze korzystają z aut używanych. Średni wiek samochodu osobowego w 2022 roku wyniósł prawie 15 lat i był o 0,4 roku wyższy niż w 2021 roku. Największymi grupami w przedziale wiekowym są więc te obecne na rynku od jedenastu do dwudziestu lat (49 % wszystkich pojazdów), oraz te które mają ponad 20 lat (23% wszystkich pojazdów). Niestety, negatywne skutki eksploatacji takich pojazdów są ogólnie odczuwane, także przez właścicieli aut nowszych, bądź osób które auta w ogóle nie posiadają. Zwiększony poziom emitowanego

hałasu, większa ilość emitowanych do atmosfery spalin czy substancji toksycznych, a także mniejsza liczba systemów bezpieczeństwa, mogą istotnie wpływać na poziom zanieczyszczenia w miastach czy szeroko pojęte bezpieczeństwo ruchu drogowego[3].

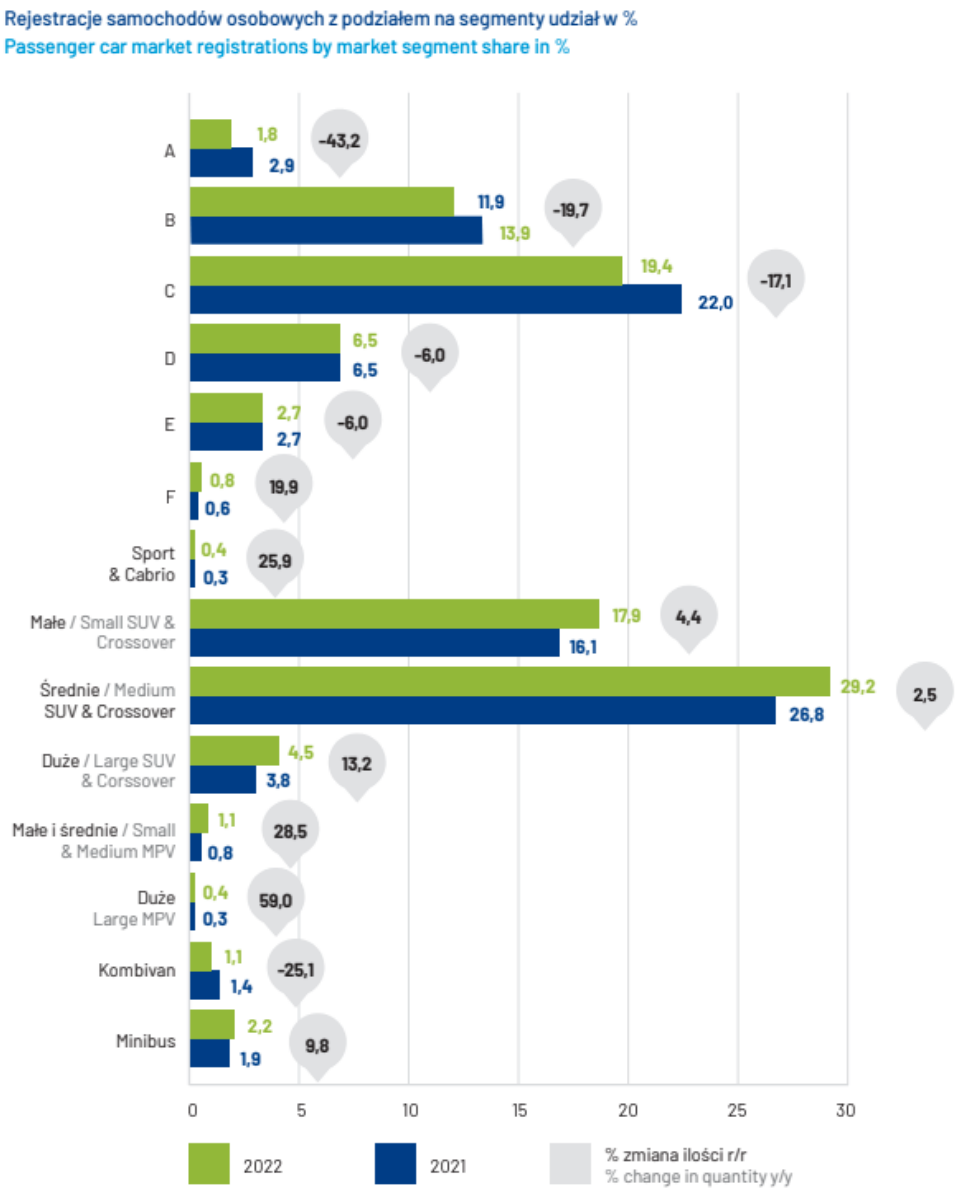
Przeglądając statystyki dla 2022 roku, można zauważyć podział aut ze względu na rodzaj paliwa. I tak, benzyna zasilała 45 % wszystkich aut, diesel 39 % pojazdów, a LPG aż 13 %. W zestawieniu zabrakło pojazdów elektrycznych, natomiast 2 % to pojazdy hybrydowe. W zestawieniu wzięto pod uwagę jednak wszystkie pojazdy na rynku. Struktura rejestracji w 2022 roku wskazała, że pojazdy z silnikami benzynowymi nadal dominują (48 %), jednak tuż za nimi uplasowały się hybrydy (32,6 %). Jeszcze dalej znajdziemy pojazdy z silnikiem diesla (11 %). Około 3 procent rejestracji osiągnęły auta elektryczne oraz napędzane gazem LPG. Pod kątem struktury właścicielskiej, ponad 90 % pojazdów należało do osób fizycznych, a dokładnie 9,5 % do przedsiębiorstw. To firmy inwestują najczęściej w auta nowe, w tym pojazdy z alternatywnym źródłem napędu, takim jak hybryda, gdzie udział w strukturze napędu wyniósł 16 %. Auta w wieku do 4 lat posiada tylko 45 % osób prywatnych, resztę, czyli 55 % należy do podmiotów prawnych posiadających numer REGON. W segmencie aut w wieku do czterech lat najpopularniejszą marką była Toyota (13,8 %), przed Skodą (10,5 %) i Volkswagenem (7,9 %). W segmencie od pięciu do dziesięciu lat prym wiódł Volkswagen (8,6 %) przed Oplem (8,3 % i Fordem (8,2 %). Na liście najpopularniejszych marek w grupie aut osobowych w 2022 roku czołową pozycję utrzymała Toyota. Rejestracja 73,9 tys. aut pozwoliła japońskiej marce uzyskać 17,6 proc. rynku. Drugie miejsce zajęła Skoda z udziałem 10 % rynku. Trzecie miejsce przypadło KIA z wynikiem 8 % udziału w rynku. Na kolejnych lokatach znalazł się Volkswagen oraz Hyundai. Co ciekawe, na ostatniej klasyfikowanej pozycji (35) znalazła się chińska marka Seres, gdzie polskie tablice rejestracyjne w 2021 uzyskała tylko jedna sztuka, a w 2022 już 153 szt., co dało wzrost w wysokości 15 200 % [3].

Dane dotyczące współczynnika zmotoryzowania ściśle zależą m.in. od poziomu zamożności w danych regionach kraju, dostępności do dróg czy alternatywnych środków transportu. Pierwszym kryterium, które wydawać by się mogło dominujące podczas analizy, to liczba pojazdów na 1000 mieszkańców. W 2023 roku prym wiedzie województwo mazowieckie z 576 autami w 2022 roku, następnie woj. wielkopolskie z 559 samochodami, a kolejne jest woj. lubuskie – 535 pojazdów. Na drugim końcu znalazły się województwa: warmińsko-mazurskie (456), zachodniopomorskie (469) i

podlaskie - 474 pojazdów na 1000 mieszkańców. Dla samochodów osobowych w Polsce współczynnik ten wyniósł 517 aut na tysiąc mieszkańców, co jest wynikiem wyższym o 15, w stosunku do 2022 roku. Co ciekawe, w Polsce w 2022 roku wyjechało z taśm produkcyjnych ponad 420 tys. aut osobowych oraz lekkich dostawczych. Najpopularniejszą fabryką został Volkswagen Poznań z udziałem w rynku 53% krajowej produkcji.

Jakiej klasy pojazdami jeżdżą natomiast Polacy? Otóż w 2022 roku już kolejny rok z rzędu dominowały SUV-y (29,2 % udziału w rynku). Modele tzw. klasy C, czyli auta kompaktowe znalazły się tuż za nimi z udziałem 19,4 % w rynku. Podium zamykają małe SUV-y i crossovery z udziałem 17,9 %. Trend rezygnacji z samochodów klasy C na rzez SUV-ów staje się bardzo wyraźny, jak również coraz bardziej niepokojący ze względu na coraz większą ingerencję w ograniczoną przestrzeń miasta w kontekście np. miejsc do parkowania czy bezpieczeństwo niechronionych uczestników ruchu drogowego. Szczegółowe dane dot. poszczególnych segmentów pojazdów zostały przedstawione na rys.2.

Najpopularniejszym modelem auta już kolejny rok z rzędu została Toyota Corolla z wynikiem 21,4 tys. rejestracji, na drugim miejscu znalazła się Toyota Yaris (12,5 tys. aut), podium zamyka natomiast KIA Sportage (11,2 tys. aut.) [3].



Rys.2. Rejestracje samochodów osobowych z podziałem na segmenty udział w %, źródło: Raport Branży Motoryzacyjnej Polskiego Związku Przemysłu Motoryzacyjnego 2023/2024

1.3. Wybrane aspekty prawne

Branża pojazdów osobowych w Polsce nie jest objęta szczególnie dużą liczbą regulacji prawnych. Strefy czystego transportu praktycznie nie istnieją, wobec tego nie ma nałożonych pewnych standardów jakościowych na posiadaczy czy użytkowników aut. Nie istnieją strefy ograniczonego dostępu, a strefy uspokojonego ruchu to nadal niezrealizowana, acz zapowiadana przez włodarzy wielu miast wizja. Wszystko wskazuje na to, że czasy taniego i stosunkowo bezproblemowego posiadania auta zaczynają się

powoli kończyć, a to za sprawą planów objęcia sektorów transportu drogowego systemem EU ETS, stopniowym zakazem sprzedaży samochodów spalinowych czy intensywnym rozwojem infrastruktury alternatywnych źródeł energii. Celem postawionym przez kraje członkowskie Unii Europejskie, jest ograniczenie do 90 % emisji gazów cieplarnianych pochodzących z transportu w UE. Niewiele na ten temat wiadomo do tej pory, gdyż więcej czasu w przestrzeni medialnej poświęca się na dyskusje nad ustandaryzowaniem i zapewnieniem odpowiedniej dostępności do ładowarek pojazdom elektrycznym czy stacjom tankowania wodoru.

Oprócz wyzwań stojących przed przyszłymi właścicielami aut, pojawiają się także szanse na wzmocnienie rodzimego rynku branży automotive. W tym celu oferuje się start-upom czy grupom badawczym instrumenty dotacyjne i dłużne zapewniane przez Narodowy Fundusz Ochrony Środowiska i Gospodarki Wodnej, Wojewódzkie Fundusze Ochrony Środowiska i Gospodarki Wodnej, Narodowe Centrum Badań i Rozwoju czy Ministerstwo Funduszy i Polityki Regionalnej. Mają one na celu rozwój transportu niskoemisyjnego oraz elektromobilność.

1.4. Wybrane aspekty funkcjonowania rynku

Motoryzacja już od wielu lat jest kluczowym sektorem polskiej gospodarki narodowej. Niedawne inwestycje zagraniczne w przemyśle motoryzacyjnym powodują znaczny wzrost udziału w zatrudnieniu w tej branży, wzrost produkcji przemysłowej, nakładów inwestycyjnych, jak też wzrost eksportu. Mimo kryzysu covidowego, polska gospodarka notuje dobre wskaźniki gospodarcze, np. wskaźnik PKB, którego wzrost w 2021 roku wyniósł 5,9 %, natomiast w 2022 roku 5,1 %. W tym przypadku Polska zdecydowanie wyróżnia się na tle pozostałych krajów europejskich, które przeżywają dalsze skutki tegoż kryzysu, jak np. polscy sąsiedzi z południa i zachodu, czyli Czesi i Niemcy, gdyż omawiany wskaźnik dla Unii Europejskiej i strefy euro wyniósł w obu tych przypadkach jedynie 3,5%.

Mocnymi stronami branży samochodowej są duże moce produkcyjne, które jednocześnie konsumowane są przez popyt wewnętrzny na marki masowe, przy jednoczesnym wzroście popularności marek premium. Wspomniana w poprzednich podrozdziałach inflacja należała do barier ograniczających wzrost, a eksport na poziomie 90 % wywołał zachwianie, spowodowane wahaniem kursów walutowych czy wzrostem cen energii elektrycznej. Biorąc jednak pod uwagę zarówno produkcję pojazdów samochodowych,

jak i części oraz akcesoriów motoryzacyjnych, a także handel oraz usługi związane z motoryzacją, polska branża motoryzacyjna jest największą w Europie Środkowo-Wschodniej. Raport Polskiego Związku Przemysłu Motoryzacyjnego [3], wskazuje jednoznacznie, że branża ta, pozostaje jednym z najważniejszych sektorów polskiej gospodarki, odpowiadając za 8,2% wartości produkcji sprzedanej w przemyśle w 2022 roku. Jest to wzrost o 22 % w ujęciu r/r. W tej statystyce pod względem wartości musi ustąpić jedynie produkcji artykułów spożywczych. Pod kątem produkcji pojazdów samochodowych, przyczep i naczep - nakładów inwestycyjnych, rynek był wart aż 7,5 mld zł, co stanowiło 6,9 % udziału w nakładach przemysłu ogółem. Zatrudnionych było ponad 196 tys. osób, co stanowi ponad 7% w zatrudnieniu ogólnym. Przytoczone statystyki pozwalają stwierdzić, że branża automotive w Polsce znaczy stosunkowo dużo dla gospodarki krajowej, zarówno w kontekście finansowym, jak i zapewnienia miejsc pracy dla tysięcy osób.

1.5. Wnioski na przyszłość

Przedstawione w poprzedniej części pracy dane przedstawiły jednoznacznie, że rynek motoryzacyjny w Polsce ulega wielu zmianom na przestrzeni czasu. Istotna rola dla gospodarki kraju czy jego obywateli nakłada na nas obowiązek bacznej obserwacji oraz weryfikacji jakim trendom ulegają konsumenci, czyli jak zmienia się rynek. Pozwoli to utrzymać Polsce swój status jako wytwórcy, a wykorzystując słabości innych krajów, da możliwość budowania swoich przewag komparatywnych.

W weryfikacji trendów ważne są badania naukowców z Uniwersytetu im. Adama Mickiewicza w Poznaniu, który przeprowadzili analizę rynku nowych samochodów osobowych w Polsce, pod kątem czasowym oraz geograficznym. Wynika z niej, że największy wpływ warunkujący popyt mają takie aspekty jak poziom dochodów konsumentów, uwarunkowania demograficzne, poziom rozwoju lokalnego czy poziom nasycenia rynku. Nie da się jednak przedstawić jednoznacznego wniosku określonego np. wzorem matematycznym, gdyż wspomniane czynniki oddziałują w różny sposób w różnych układach przestrzennych [4]. Wyniki wielu badań wskazują także, że im wyższy przyrost rzeczywisty, czyli suma przyrostu naturalnego ludności oraz salda migracji wewnętrznych i zagranicznych (stałych i czasowych) , tym wyższy popyt na samochody. Podobną korelację obserwuje się w dochodach konsumentów, gdzie ich wzrost, oznacza także wzrost popytu na auta. To samo tyczyć się może poziomu urbanizacji czy zaspokojenia potrzeb motoryzacyjnych. Według przytoczonych badań (Train, Winston,

2007), największy wpływ na użytkowanie samochodu ma cena benzyny. W literaturze specjalistycznej można wyróżnić także referaty pn. „Przegląd ekonomiczno-przestrzennych badań rynku samochodów osobowych" [5] oraz „Czynniki wpływające na liczbę rejestracji nowych samochodów w Polsce, a model regresji liniowej". W ostatnim badaniu model ten miał objaśnić jaki wpływ mają czynniki ekonomiczne, które zostały wybrane w drodze eliminacji spośród 12, których przykładem może być długość dróg ekspresowych i autostrad na 1000 m2, PKB per capita, dochód gospodarstw domowych czy stopa referencyjna. Niestety nie udało się wyciągnąć jednoznacznych wniosków, ze względu na niski poziom istotności zmiennych objaśniających.

W Pracach Naukowych Uniwersytetu Ekonomicznego we Wrocławiu opublikowany został raport dotyczący trendów konsumenckich wśród pokolenia Z w obszarze dobór luksusowych [6]. Wobec stopniowego wejścia pokolenia Z na rynek pracy, jak również osiągnięciem przez nich wieku pozwalającego na kierowanie samochodami osobowymi, trend ten może być kluczowy dla rynku w Polsce w perspektywie najbliższych kilku czy kilkunastu lat. To pokolenie Z staje się coraz wyższym odsetkiem osób mających wpływ na siłę nabywczą i w najbliższych latach.

Dlaczego autor opracowania stawia tezę, że to właśnie rynek pojazdów premium będzie się dynamicznie rozwijał i stanowił pierwszy wybór dla pokolenia Z? Otóż opierając się na raporcie firmy KPMG, można stwierdzić, że to najmłodsi konsumenci nabywają dobra luksusowe znacznie częściej niż poprzednie pokolenie, a więc osoby w wieku powyżej 51 lat. Nabywcy w przedziale wiekowym od 18 do 35 lat częściej niż pozostałe grupy wiekowe wybierają luksusową odzież i obuwie. Trend ten już teraz jest bardzo widoczny, a podczas stopniowego dorastania i bogacenia się, możemy przewidzieć, że to pokolenie stanie się już niedługo kluczowym konsumentem luksusowych nieruchomości, biżuterii, usług hotelarskich, jachtów czy samochodów. Raport wskazuje, że produkty i usługi powinny być jak najbardziej personalizowane, tak aby dopasować je do potrzeb wymagającego konsumenta i podkreślić jego indywidualność. Marka, by być odbieraną pozytywnie i zyskać status cenionej na rynku, powinna dobrze rezonować z wybranymi grupami społecznymi, odpowiadać ich przekonaniom i wartościom. W tym przypadku promocja oparta na znanych osobach może nie pasować do oczekiwań grupy docelowej, w przeciwieństwie do podkreślania etyczności marki, zaangażowania społecznego czy działaniach CSR. Sam proces zakupowy musi być prosty i przejrzysty. Powierzchowna obsługa czy długi czas

oczekiwania to elementy zdecydowanie zniechęcające „pokolenie Z". Respondenci zostali poproszeni również o wskazanie wartości, którymi powinna kierować się organizacja, aby najchętniej korzystać z marek premium. Najwięcej osób wskazało (do wyboru 3 z 9 opcji) „wiarygodność marki" (86 osób), „innowacyjność i nieszablonowość" (74 osoby) oraz odpowiedzialność (72 osoby). Niewiele mniej badanych wskazało jako istotny „szacunek" (53 osoby) oraz „działania proekologiczne firmy" (44 osoby) [6].

Analiza literatury pozwoliła także dostrzec coraz istotniejsze trendy, dzięki którym klienci nabywają pojazdy, a mianowicie preferowanie automatycznych skrzyń biegów czy też budowanie intuicyjnych samochodów, które to w maksymalnie prosty sposób zaoferują maksymalnie dużo opcji. Mowa tu o spersonalizowanej nawigacji, automatycznych ustawieniach do kierowcy np. lusterek czy foteli, czy też obecność wirtualnego asystenta, rozpoznającego komendy głosowe. Nie sposób zapomnieć także o indywidualizacji wyposażenia. Coraz częściej wiodące marki udostępniają klientom konfiguratory, w których mogą wybrać rozwiązania i systemy, które chcieliby zastosować w swoim aucie dostosowując je do własnych potrzeb i co równie ważne – możliwości finansowych. Ostatnim trendem, na który być może będziemy musieli jednak jeszcze chwilę poczekać, są pojazdy autonomiczne. W tym momencie rozwiązanie może wydawać się stosunkowo abstrakcyjne. Jednak to nie technologia, a brak regulacji czy poligonów testowych i odcinków do doskonalenia oprogramowania uniemożliwia rozwój tej gałęzi motoryzacji. W tej sprawie słychać już pierwsze głosy dobiegające z Unii Europejskiej, jednak nadal brak konkretnych decyzji, do rychłej liberalizacji.

Podsumowując, już wkrótce branża automotive może stanąć przed szansą, ale też wyzwaniem, jakim jest sprostanie oczekiwaniom konsumentów z „pokolenia Z". Przed branżą wiele wyzwań, takich jak postawienie na wiarygodność marki i rzetelność w prowadzeniu biznesu czy też wiele wyzwań technologicznych, takich jak szeroka implementacja wydajnych pojazdów elektrycznych czy wodorowych, szczególnie w klasie premium. Być może to koncerny motoryzacyjne będą miały swój udział w budowie infrastruktury do ładowania, a może skupią się na rozwijaniu relacji z klientami i nie tylko sprzedawaniu pojazdów, ale nawiązywania pewnych więzi, w celu świadczenia usług przez wiele lat.

1.6. Konkluzje

Gromadzenie danych ilościowych przez Główny Urząd Statystyczny, Centralną Ewidencję Pojazdów czy przemysł motoryzacyjny, umożliwia analizę istotnych trendów czy zależności, które zachodzą na polskim rynku. Mimo wyzwań gospodarczych na świecie, w przypadku rynku motoryzacyjnego, w znacznej mierze wynikających z problemów geopolitycznych, prowadzonej transformacji energetycznej, rynek automotive radzi sobie stosunkowo dobrze. Jest on silnym filarem polskiej gospodarki, szczególnie w produkcji poszczególnych elementów czy podzespołów. Dzięki działaniom stymulującym ze strony Unii Europejskiej czy też coraz bardziej wymagającego „pokolenia Z", z całą pewnością można stwierdzić, że rozwój technologiczny będzie następował w stosunkowo szybkim tempie. Branżę czekać będzie także zmiana nastawienia do klienta, a być może przejście z modelu sprzedaży towaru na rzecz świadczenia długoterminowych usług. Przedstawiona analiza to wyłącznie fragment danych o branży, aktualnych na rok 2024. Szczegółowe omówienie na przykładzie roku 2018, zostanie pokazane w kolejnych rozdziałach. Za pomocą specjalistycznych narzędzi wizualizacji i baz danych, uda się wyciągnąć jednoznaczne wnioski z informacji zebranych w postaci rozbudowanej bazy danych, wymagającej pierwotnie m.in uprzątnięcia rekordów.

Rozdział II. Przegląd wybranych pojęć i uwarunkowań metodycznych analizy danych

Baza danych została stworzona na podstawie pobranej bazy danych z serwisu internetowego Kaggle. Tworzył ją jeden plik płaski o rozszerzeniu .csv, ze wszystkimi danymi w jednej tabeli. Aby umożliwić dalszą pracę należało rozdzielić dane, tak aby powstał bazodanowy model według schematu gwiazdy (*ang. star schema*). W każdej kolumnie policzono wartości unikatowe i na ich podstawie stworzono kolejne tabele oraz nadano im w nich klucze, z którymi łączą się z bazą główną. Dzięki temu tabela faktów (ang. fact table) składa się w głównej mierze z kluczy obcych pozostałych tabel z modelu. Poza nimi, w głównej tabeli modelu znajdują się fakty mierzalne, ilościowe, dające się zapisać w formie liczb, w tym przypadku cena auta oraz jego przebieg, pozostałe dane są w tabelach stanowiących kolejne ramiona gwiazdy. Dwie kolumny były szczególnie problematyczne w przekształceniu w następne tabele, mianowicie kolumna z miejscami ofert oraz z kolumna z wyszególnionymi elementami wyposażenia jakie dane auto posiada. Przygotowanie tych danych do analizy zostało opisane w następnym rozdziale. Kiedy już wszystkie dane zostały w odpowiedni sposób ułożone i zaimportowane do środowiska Power BI, zostały poddane analizie i procesowi tworzenia dashboardu managerskiego. W kolejnym rozdziale zostaną omówione poszczególne jego strony, cel ich powstania, sens działania oraz techniczne kwestie stojące za przyjętymi rozwiązaniami. W tym rozdziale natomiast uwaga będzie poświęcona metodologii tworzenia i działania hurtowni danych.

2.1. Informacje a dane

Aby przejść do pojęcia systemu bazodanowego, należy przyjrzeć się dwóm innym pojęciom, czym są informacje, a czym są dane. Według Teresy M. Ostrowskiej [7] „informacja jest jednym ze składników zarówno procesów zarządzania, jak i procesów produkcyjnych. […] jest podstawą sterowania i kontroli, konieczna do podejmowania decyzji. […] traktuje się jako zasób, kategorię filozoficzną". Oznacza to, że jest to nie do końca sprecyzowana materia, ale jednak w jakiś sposób mierzalna i interpretowalna – „kamień leśny porośnięty mchem informuje o stronach świata – konieczna jest tylko umiejętność interpretacji tego zjawiska". Czym w takim razie są dane? Autorka wskazuje,

iż w kontekście bazodanowym są to „ciągi symboli uporządkowane zgodnie z przyjętymi regułami i zapisane na nośniku danych”. Informacja zatem jest efektem interpretacji tego ciągu, na który wpływają czynniki zarówno wewnętrzne, jak jego struktura, ale także zewnętrzne, np. wiedza i umiejętności interpretatora czy czas i miejsce analizy. W celu zachowania rzetelności tej analizy wymagane jest to, aby dane były odpowiednie zarówno statystycznie (np. były w odpowiedniej ilości), syntaktycznie (kwestie strukturalne) oraz semantycznie (odpowiednia reprezentacja zjawiska, do którego się odnoszą). Ponadto muszą być możliwe do poddania weryfikacji, aktualizacji, ochrony i oraz analizy. I temu właśnie służą bazy danych [7].

2.2. System bazodanowy

W systemach informatycznych dane są zasobem, które dostarczają informacji w procesie przetwarzania. Ich organizacja w rekordy, które składają się z pól zawierających wartości danych to baza danych. Struktura tej organizacji to model logiczny, zatem baza danych to zbiór wystąpień rekordów w tymże modelu. Obie te materie, wraz z aplikacjami użytkowników, tworzą system bazodanowy. Pakiety oprogramowania, które tworzą i umożliwiają jego eksploatację to systemy zarządzania bazami danych [7].

2.3. Formalizacja zapisu modelu konceptualnego

Aby przejść do relacyjnego modelu bazy danych należy najpierw wyjaśnić kwestię formalizacji zapisu modelu konceptualnego. Jest jakiś obszar modelowania, czyli wycinek rzeczywistości bądź abstrakcji, który podlega analizie. Obiekty są ze sobą powiązane w jeden ze sposobów: 1:1 (jeden do jednego), 1:n (jeden do wielu), n:m (wiele do wielu). W modelu obiekt jest reprezentowany przez encję (ang. entity – byt, istnienie)., czyli „model informacyjny obiektu fizycznego bądź abstrakcyjnego”. Zbiory tych samych atrybutów są z kolei nazywane klasami encji. Poszczególne właściwości obiektów są natomiast opisane atrybutami, z których każdy z nich musi mieć unikatową nazwę w ramach jednej encji oraz zdefiniowane właściwości, jak typ danych, dozwolone wartości, czy maksymalny rozmiar. Atrybuty są zadaniowane, albo identyfikują albo opisują stan encji. Każda encja posiada swój unikatowy identyfikator [7]. W zależności od typu relacji, w innych tabelach, będzie on kluczem obcym. W związkach typu 1:n oznacza to, że jeden obiekt z encji A może być w relacji z wieloma obiektami z encji B, ale każdy obiekt z encji B może być w relacji tylko z jednym obiektem z encji A. W związkach typu n:m każdy obiekt z encji A może być w wielu relacjach z obiektami z

encji B i odwrotnie [8]. W informatyce związek wiele do wielu jest w praktyce nie do zaimplementowania, toteż w modelach starać należy się do zachowania któregoś ze związków prostych. Powyższy opis dotyczy modelu logicznego, niezależnego od implementacji, która nadaje mu fizyczny, istniejący charakter. Transformacja ta odbywa się według ustalonych reguł, jak to że każdej encji odpowiada jedna tabela, atrybut encji to kolumn tabeli, zaś identyfikator encji to klucz główny tabeli. Tak samo jak encje, transformowane są też ich związki [9].

2.4. Relacyjny model baz danych

Zaimplementowany model logiczny stanowi relacyjny model baz danych. Po raz pierwszy sformułowania tego użył w 1970 roku brytyjski naukowiec Edgar Codd. W matematycznym spojrzeniu baza danych to zbiór relacji traktowanej jako podzbiór iloczynu kartezjańskiego zbiorów wartości. W przypadku tej bazy danych jej reprezentantem jest dwuwymiarowa tabela złożona z kolumn i wierszy, oraz kluczy głównych i kluczy obcych [10]. Relacyjny model danych jest zatem zbiorem konstrukcji opisowych struktury danych oraz operacji na nich, oparty na przedstawianiu danych w postaci powiązanych ze sobą tabel [11]. Tabele są z kolei rozumiane jako uporządkowane listy danych przechowujące dane tylko jednego typu[12].

Należy jednak wspomnieć w tym miejscu, że relacyjny model danych należy do drugiej generacji modelów danych. Samo pojęcie powstało właśnie wraz z jego opracowaniem, jednak pierwszą generację stanowią proste modele danych, czyli pliki płaskie, np. zwykły plik csv w Excelu. Drugą generacją są klasyczne modele danych, np. te relacyjne, chociaż także hierarchiczne i sieciowe. Dostarczają one danych, ale nie metod ich interpretacji. Trzecią generacją są zaś semantyczne modele danych, które dopiero się tworzą i w przeciwieństwie do poprzedników, zawierają także semantykę danych, przykładem, choć nieidealnym, może być obiektowy model danych [13].

Warto się zastanowić chwilę, jak ogromnym postępem było stworzenie relacyjnego modelu danych, jakie przewagi ma on nad zwykłym plikiem płaskim. Dane nie są wyizolowane od siebie, ale powiązane ze sobą, nie tak jak w systemie plików. Nie są także zdublowane, co jest skutkiem izolacji plików i braku integralności danych. Wspólne wykorzystanie zasobów prowadzi do optymalizacji procesów użytkowania baz danych [14] .

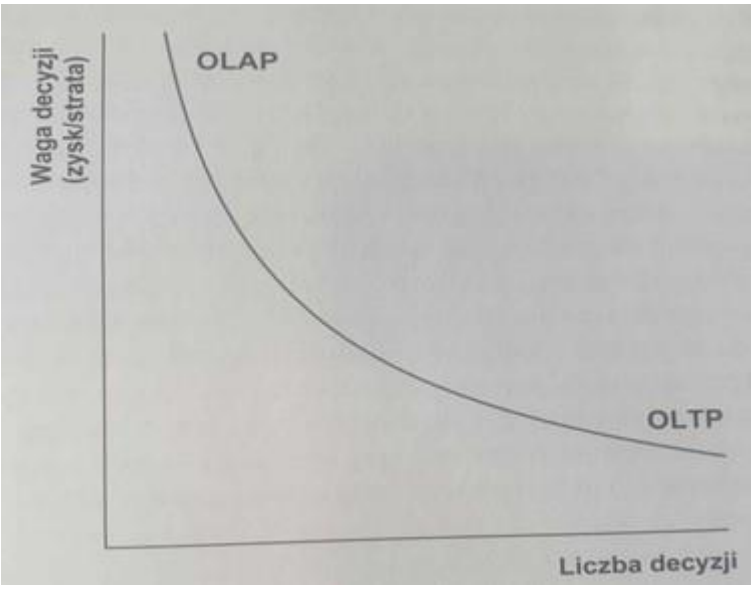
2.5. Języki bazodanowe

Aby komunikować się skutecznie z bazami danych należało stworzyć specjalne języki programowania, służące tylko do tego, czyli do wykonywania operacji na bazach danych. Porównując je do innych języków programowania są one nieco ograniczone, ponieważ liczy się w nich głównie efektywność pisanego kodu. Głównym językiem bazodanowym, który spośród wielu przetrwał w dość niezmienionej formie po dziś dzień jest język SQL, czyli Structured Query Language [15]. Nie jest to miejsce na dokładny opis jego możliwości i struktury, jest dużo podręczników do jego nauki, z których autor pracy także korzystał, inaczej niemożliwym byłoby stworzenie relacyjnej bazy danych, jaka została tu wykorzystana.

2.6. OLTP i OLAP

Bazy danych mają różnorakie biznesowe zastosowania i są grupowane w zależności od nich. Pierwszą grupę baz danych stanowią bazy operacyjne, których używanie ma nazwę OLTP (ang. On-line Transaction Processing), czyli bezpośrednie przetwarzanie transakcyjne. Służą głównie do utrzymania spójności danych, umożliwiały także rejestrowanie faktów z podstawowej operacyjnej działalności firmy. Drugą grupą są bazy strategiczne, których używanie ma nazwę OLAP (ang. On-Line Analytical Processing), czyli bezpośrednie przetwarzanie analityczne [11] . Służą one w znacznej mierze do czegoś więcej niż tylko do bieżącej działalności. Ich głównym celem jest wspomaganie decyzji poprzez dostarczanie odpowiednim osobom informacji o wychwyconych biznesowych wzorcach, czy trendach. Z tego powodu są to bazy, w których znajdują się także historyczne dane, a nie tylko te najnowsze. W celu zwiększenia ich czytelności i zrozumiałości zawierają już przetworzone dane, które z reguły nie są już dalej modyfikowane, a jedynie uzupełniane o nowe dane, ze źródłowych baz operacyjnych. Takie bazy nazywane są hurtowniami danych (ang. data warehouse) bądź też magazynami danych lub bazami analitycznymi.

Bardzo ważnym zagadnieniem jest także decyzyjność, bowiem od niej zależy miejsce systemów transakcyjnych i analitycznych. Im większa waga podejmowanych decyzji, tym większa rola OLAP, im mniejsza, tym większa rola OLTP [16], co pokazuje rys.3:



Rys.3. Miejsce przetwarzania analitycznego i transakcyjnego w zależności od wagi i liczby decyzji, źródło: Adam Pelikant, Hurtownie danych. Od przetwarzania analitycznego do raportowania. Wydanie II. Helion, 2021

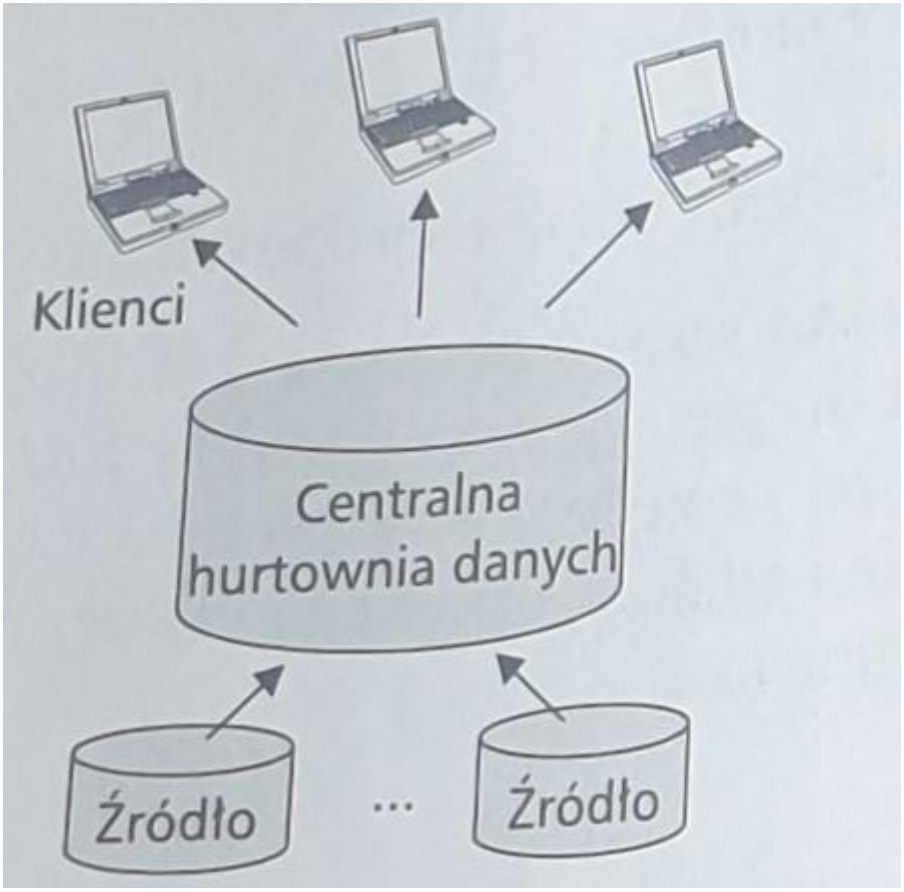
2.7. Hurtownie danych

Typowym przedstawicielem hurtowni danych jest baza sprzedażowa lub ofertowa, tak jak w modelu wybranym w tej pracy [10]. Często używane jest też specjalistyczne oprogramowanie umożliwiające data mining, czyli eksplorację danych w celu znalezienia ciekawych trendów i zjawisk. Eksploracja (ang. data drilling) w przeciwieństwie do dogłębnej analizy, jest obsługiwaniem wymiarów w dwojaki sposób, albo uszczegółowienia (drill down) albo agregacji (drill up) [7].

Zaletą hurtowni danych w praktyce jest fizyczne odseparowanie przetwarzania analitycznego od przetwarzania transakcyjnego, dlatego że zazwyczaj oba typy baz danych znajdują się na różnych serwerach. Co więcej postać danych, jak i ich organizacja jest już od razu przygotowana do analiz, także model przechowywania danych został temu podporządkowany [16].

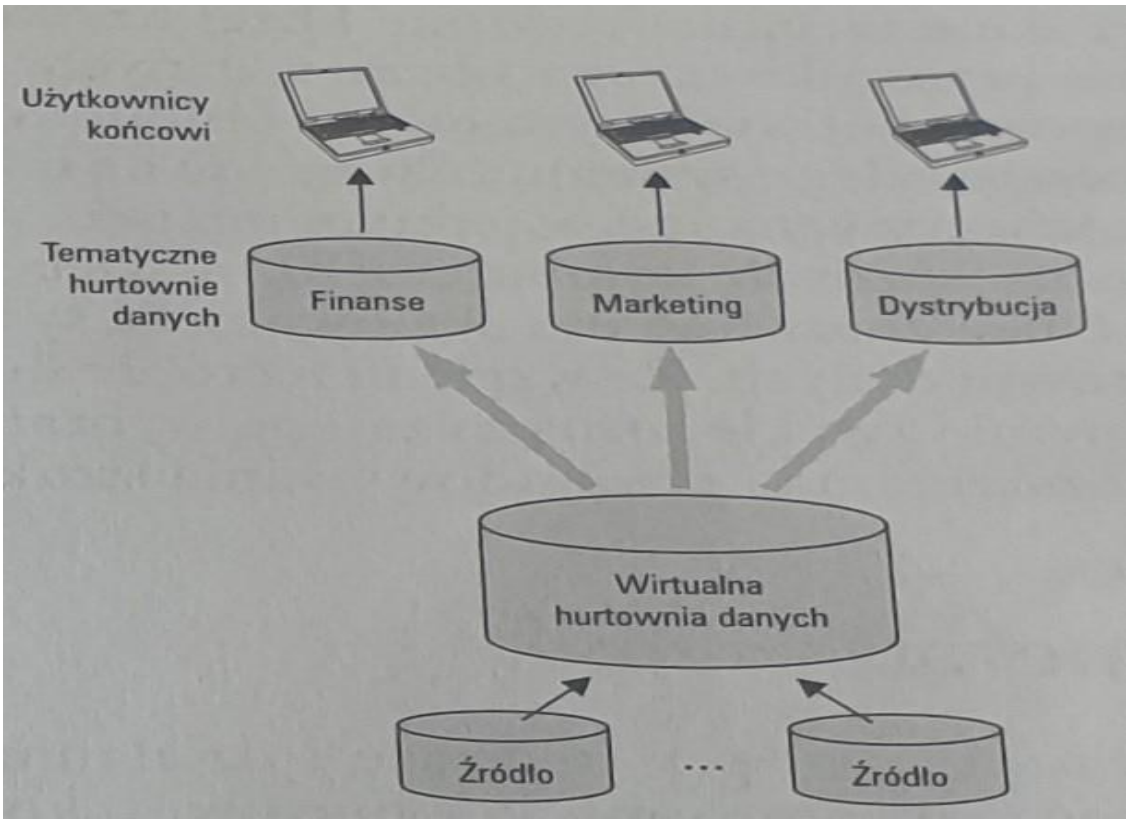
2.8. Architektura hurtowni danych

Wyróżnia się trzy kształty architektury hurtowni danych: scentralizowaną, federacyjną oraz warstwową. Pierwszy z nich, jak nazwa wskazuje, zawiera jedno centrum, jedną centralną hurtownię danych. Ich zaletą na pewno będzie ujednolicenie modelu oraz fakt, że nadaje się do instytucji, które same w sobie też są scentralizowane, ale jest to okupione kosztem zmniejszenia ich wydajności.

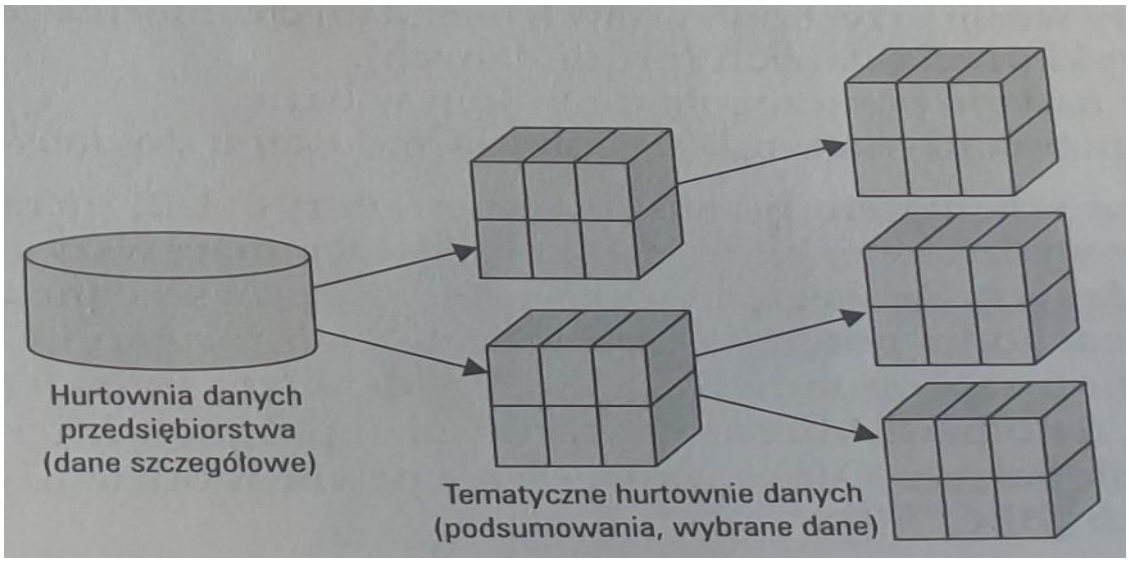


Rys.4. Architektura scentralizowana, źródło: Matthias Jarke, Hurtownie danych. WSIP Wydawnictwa Szkolne i Pedagogiczne, 2003

Aby jednak poprawić największe wady tego typu architektury, można hurtownie danych zdecentralizować, na dwa sposoby. Jednym z nich jest federacja, czyli kilka różnych tematycznych hurtowni danych, które wirtualnie łączą się w jedną globalną, albo na architekturę warstwową, która różni się od federacji tym, że globalna hurtownia istnieje nie tylko wirtualnie [17].



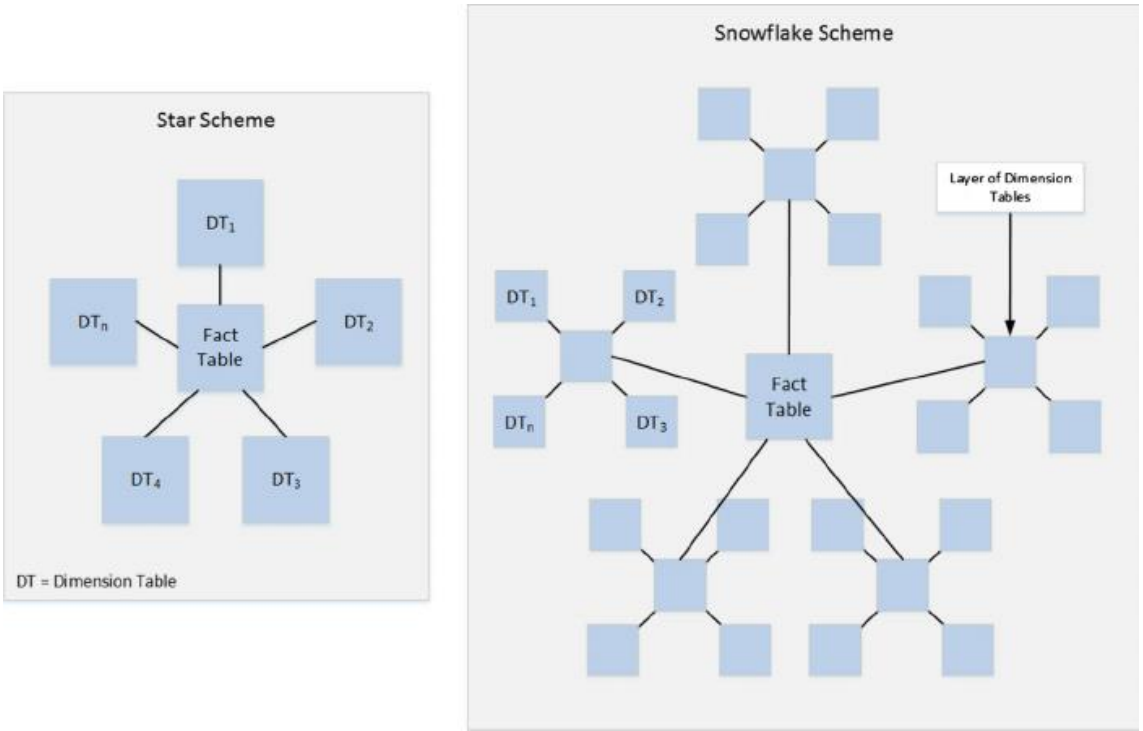
Rys.5. Architektura federacyjna, źródło: Matthias Jarke, Hurtownie danych. WSIP Wydawnictwa Szkolne i Pedagogiczne, 2003



Rys.6. Architektura warstwowa, źródło: Matthias Jarke, Hurtownie danych. WSIP Wydawnictwa Szkolne i Pedagogiczne, 2003

2.9. Struktura hurtowni danych

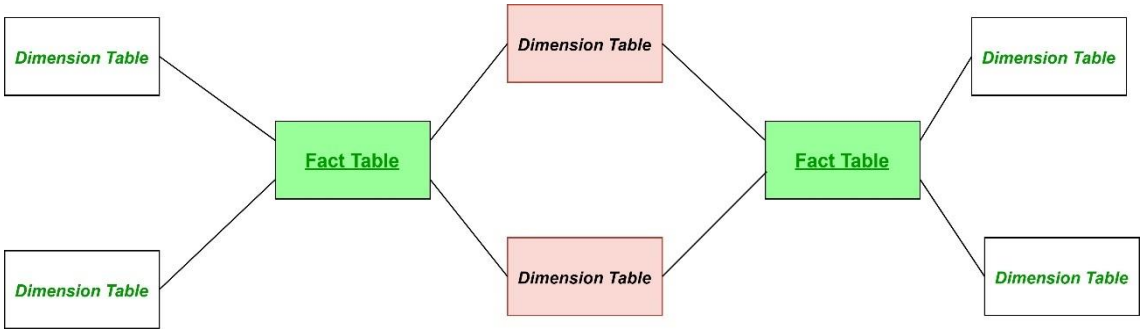
Struktura hurtowni danych różni się od struktury baz transakcyjnych. Wyróżnia się zazwyczaj trzy ich typy: model MOLAP (ang. Multidimensional OLAP), ROLAP (ang. Relational OLAP) i hybrydę dwóch poprzednich [18]. Łączy je tabelaryczność, ale rozumiana inaczej w obu przypadkach. W tym drugim bowiem wszystkie tabele są równoprawne, zaś w pierwszym jest jedna tabela centralna, zwana też tabelą faktów, zawierającą mierzalne dane, fakty, np. sprzedaż czy obroty, oraz atrybuty pozostałych tabel – klucze obce, oraz tabele wymiarów, z danymi opisującymi obiekty z tabeli faktów, według okoliczności ich zaistnienia. Wyróżnia się trzy podstawowe modele danych: gwiazdy, płatka śniegu oraz burzy śniegowej. Model gwiazdy (ag. Star schema) charakteryzuje się jedną tabelą faktów i wieloma tabelami wymiarów. Model płatka śniegu (ang. snowflake schema) charakteryzuje się tym, że część tabel wymiarów posiada swoje hierarchiczne odgałęzienia, które jeszcze bardziej doprecyzowują temat i agregują więcej informacji.



Rys. 7. Porównanie modelu gwiazdy z modelem płatka śniegu, źródło: Research Gate, “ Conceptual view of star and snowflake schemas”

Model burzy śniegowej charakteryzuje się tym, że zawiera wiele tabel o tych samych wymiarach, np. w jednocześnie są zawarte tabele sprzedażowa, jak i produkcyjna.

Niektórzy badacze nazywają go modelem konstelacji gwiazd – gwiazdozbioru (ang. galaxy schema), który polegać ma na wspólnych tabelach wymiarów dla poszczególnych tabel faktów [18].



Rys.8. Model burzy śniegowej, zwany też modelem konstelacji gwiazd, źródło: Geek for geeks, " Fact Constellation in Data Warehouse modelling."

Istnieje także kostka OLAP. Jest to rozwinięcie struktury MOLAP, wielowymiarowy model danych w formie kostki przypominającej kostkę Rubika, której krawędzie są określone wymiarami, czyli reprezentacją zbiorów obiektów i ich wartościami. Jest to interpretacja geometryczna danych, którą można sprowadzić do wielowymiarowej tablicy [7], [10].

2.10. Proces ETL

Proces ETL to proces integracji danych, ładowania ich do bazy. Nazwa to skrót z angielskiego „Extract – Transform - Load", czyli pobieranie – przekształcanie – ładowanie. Pierwszy etap oznacza sczytywanie baz operacyjnych, które stanowią źródło dla hurtowni, drugi etap to oczyszczanie danych tak, aby doprowadzić je do stanu używalnego w nowej bazie, a trzeci to zapisywanie ich tam. Co ważne i co zaznacza Agnieszka Chodkowska-Gyurics, ETL jest procesem, a nie narzędziem i porównała odwrotność takiego twierdzenia do zdania, że „programowanie to komputer". Jest to zatem mylne postrzeganie ETL i należy to zaznaczyć, zwłaszcza w dobie coraz częstszego popełniania tego błędu [19].

2.11. Typy wymiarów ze względu na zawartość

„Wymiar to obiekt w hurtowni danych, implementowany w postaci tabeli” [19]. Rozróżnia się kilka ich rodzajów ze względu na zawartość: uzgodnione (conformed dimensions), abstrakcyjne (junk dimensions), zdegenerowane (degenerate dimensions) i wielokrotnego zastosowania (role-playing dimensions). Wymiary mogą się pokrywać, tzn. być kilkoma rodzajami naraz, jak też nie należeć do żadnej z tych kategorii. To ważne, bo brak klasyfikacji nie oznacza bynajmniej popełnienia błędów, po prostu teoria nie zawsze nadąża za praktyką i potrzebuje czasu, aby ustandaryzować pojęcia.

Wymiar uzgodniony to inaczej uniwersalny, czyli taki, który niezależnie od tabeli faktów zawsze posiada to samo znaczenie. Przykładem może być data, niezależnie od tego czy mówimy o sprzedaży, czy o produkcji, data zawsze oznacza to samo. Dopiero kontekst dodaje dodatkowych znaczeń i o tym należy pamiętać, gdyż nie zawsze rok kalendarzowy to to samo co rok podatkowy, rok budżetowy, bądź rok operacyjny – tu należy uważać, ponieważ dwa wymiary muszą być takie same, bądź jeden musi zawierać się w drugim, aby należeć do tego typu wymiarów. Innymi przykładami mogą być klienci, pracownicy, placówki.

Wymiar abstrakcyjny jest bardzo niesprecyzowanym, za to łatwym do modyfikowania wymiarem. Mogą to być dane o bardzo małej liczności, np. dane o płci – wtedy będą tylko dwie wartości klucza głównego, 1 oznaczający kobietę i 2 oznaczający mężczyznę. Można jednak takie tabele rozbudowywać o kolejne atrybuty, np. o wiek. Wtedy tabela będzie składać się z grup wiekowych w podziale na wiek. Jak doda się do tego kolejne atrybuty, np. którego banku ktoś jest klientem, to tabela poszerza się o kolejny wymiar i o kolejne rekordy, co zwiększa wartości klucza głównego. Można dodawać kolejne atrybuty, zawsze jednak podstawowe pytanie brzmi czy należy to tak rozbudowywać czy też nie. Odpowiedź zależy stricte od struktury bazy i od tego, co analityk chce osiągnąć, od celu biznesowego. Ralf Kimball np. nie zaleca tworzenia więcej niż 26 wymiarów, badacze jednak wskazują na arbitralność tej liczby i doradzają zdrowy rozsądek w ich tworzeniu, nie da się bowiem przypisać jednej odgórnej uniwersalnej liczby dla każdego przypadku. Należy pamiętać jednak o tym, że najlepsze do budowy tego typu wymiaru są atrybuty, których wartości znane są uprzednio, gdyż znacząco upraszcza to proces ETL, jeśli zaplanowane to zostało z góry. W innym przypadku ładowanie danych stałoby się mocno uciążliwe, system bowiem za każdym razem sprawdzałby czy dana kombinacja istnieje, w przywołanym wyżej przypadku np. kobieta w wieku 18-25 lat

będąca klientem banku X, a jeśli nie, to samemu taki rekord utworzyć i nadać wartość klucza głównego. Sprawdzanie po kolei każdego rekordu z bazy w celu odszukania danej kombinacji znacząco obniża wydajność kodu i stanowi podstawę do poszukiwania jego optymalizacji.

Wymiary zdegenerowane to obiekty o bardzo dużej liczności (ang. cardinality), które są w rzeczywistości bazodanowej wymiarem, ale znajdować się będą jednak w tabeli faktów. Dobrym przykładem będzie numer zamówienia, tzw. klucz biznesowy. Jeśli rozpatrywać ten przypadek, to w tabeli faktów znalazłyby się informacje np. o dacie zamówienia, placówce, w której go dokonano, pracowniku i kliencie którzy dokonali transakcji, i oczywiście koszt zamówienia. Numer zamówienia w tym wypadku byłby jedyną kolumną tabeli wymiarów, ale nie ma potrzeby, żeby ona istniała, skoro tę jedną kolumnę można przypisać do tabeli faktów. Należy jednak zaznaczyć, iż nie ma zastosowania przekształcanie atrybutów specjalnie w wymiary zdegenerowane, aby je umieszczać w tabeli faktów. Jest to mylny obraz, ponieważ głównym kryterium dla tego typu wymiarów jest brak atrybutów i to tym należy się kierować w kwestii umieszczania ich w tabeli faktów.

Ostatnim typem wymiarów są te wielokrotnego zastosowania, czyli takie, które występują pod różnymi nazwami, ale znaczą to samo, np. daty w tabeli sprzedażowej. Może w niej istnieć zarówno data złożenia zamówienia, data opłacenia faktury i data wydania towaru z magazynu. Trzy różne daty, ale wymiar ten sam, więc wszystkie znajdą się w jednej tabeli faktów, powiązanej z jednym kalendarzem, a nie z trzema różnymi [19].

2.12. Typy wymiarów ze względu na obsługę zmian

Ze względu na to, jak obsługiwane są zmiany danych, można wyróżnić trzy typy wymiarów: stałe (fixed dimensions), wolnozmienne (slowly changing dimensions) i szybkozmienne (rapidly changing dimensions). W przeciwieństwie do poprzedniej klasyfikacji, ta jest ostra i każdy wymiar jest tylko jednej kategorii.

Wymiary stałe to takie, których zarówno wartość, jak i liczba wierszy jest niezmienna w hurtowni danych, np. data, której liczba jest znana, tak samo jak wartość, od najstarszej do najmłodszej, bądź odwrotnie. Należy jednak ustrzec się błędu poznawczego, polegającego na uznaniu bardzo wolno zachodzących zmian, czasem niedostrzegalnych gołym okiem za wymiar stały. Klasycznym przykładem może być kategoria kodu

pocztowego. Fakt, dla istniejących miejscowości jest wciąż taki sam, ale co z miejscowościami, które właśnie powstają i mają nadany nowy kod? Proces ETL nie przypisze żadnego kodu pocztowego takiej miejscowości, bo nie było go w słowniku. Nie musi być takich błędów dużo, ale należy pamiętać co wskazuje definicja wymiaru stałego – są nim tylko te dane, których liczba i zawartość jest niezmienna w czasie. Toteż nie należy brać każdego słownika od razu za wymiar stały.

Wymiar wolnozmienny to taki, który raczej często się nie zmienia, ale jest to dopuszczalne, aby zilustrować to przykładem z życia, np. miejsce zamieszkania klienta, bądź stan cywilny. Istnieje klika typów podstawowych tego wymiaru: typ 0, 1, 2, 3 i 4.

Typ podstawowy 0, to taki typ wymiaru, w której pierwotną wartość atrybutu wiąże fakt, niezależnie od jego bieżącej wartości, dzieli się też na dwa podtypy, wymiary wywiedzione (ang. derived dimension) i wymiar zdarzeniowy (event dimension). O pierwszym z nich mówi się, gdy nie ma tabeli źródłowej, która posłużyłaby za bazę dla pożądanego wymiaru. Jako że modyfikowanie wierszy jest niemożliwe w tym typie, to w przypadku niepasujących analitykom wierszy należy dodać nowy. Drugi z nich rośnie z upływem czasu, ale, jak w życiu, nie ma możliwości cofnięcia zaistniałych zdarzeń, a każdy fakt biznesowy zostaje zapisany w nowym wierszu. Przykładem podawanym przez [19], jest wystawienie recepty. Zdarzenie o określonych atrybutach, ale o nieznanej przyszłości, w końcu jest to już własność pacjenta, a nie przychodni, w związku z czym dalszy los tej recepty nie jest zazwyczaj znany.

Typ podstawowy 1 to typ, w którym przechowywane są te wartości atrybutów, które są najnowsze i albo ich wartości są czasem wynikiem pomyłki i należy je nadpisać, albo są nieistotne z punktu widzenia biznesowego, a jedynie operacyjnego. Jest swego rodzaju odwrotnością typu 0, ponieważ, jest powiązany ściśle z bieżącą wersją atrybutu, niezależnie od historycznych wartości.

Typ podstawowy 2 to typ, w którym przechowywane są dane o zmianach, ponieważ wartość atrybutu z chwili wystąpienia jest tą, która jest powiązana z faktem. Innymi słowy, w momencie zmiany jakiegoś atrybutu, w tabeli faktów pojawia się nowy rekord, z nowym kluczem głównym, przy niezmienności istniejących już relacjach między tabelami. Aby zachować odpowiednią strukturę tabeli, potrzebne są informacje o tym, od kiedy, do kiedy i jak długo dany wiersz był ważny, oraz flagę ważności.

Typ podstawowy 3 to typ, w którym zmodyfikowane wiersze się nadpisuje nowymi danymi, a poprzednie są umieszczane w kolumnie pomocniczej. Z tego powodu też nie są zbyt często stosowanym rozwiązaniem, bo liczy się głównie bieżąca i przedostatnia wartość atrybutu w kwestii powiązania z faktem.

Typ podstawowy 4 to typ, który stosuje się w tabelach historycznych, w których umieszcza się historię zmian atrybutów, w tabeli podstawowej znajdują się tylko bieżące rekordy. Struktura obu tabel jest taka sama, jedynie w tabeli historycznej znajduje się informacja o dacie i czasie zmiany rekordu.

Ponadto wyróżnia się także kilka podtypów hybrydowych. Wymiary szybkozmienne zaś, są bardzo nieprecyzyjne, bowiem nieprecyzyjne jest samo stwierdzenie, iż jakiś wymiar zmienia się szybko. Dla każdego będzie to znaczyło co innego. Wobec tego, należy zdać się na intuicję i mówić o tym wymiarze tylko po uprzednim zanalizowaniu tematu i wykryciu szybkich zmian w obrębie jednego atrybutu [19].

Tab.1. Podsumowanie cech typów wymiarów w formie tabeli, źródło: opracowanie własne, na podstawie tabeli z: Agnieszka Chodkowska-Gyurics, Hurtownie danych. Teoria i praktyka. Wydawnictwo Naukowe PWN, 2017

Typ	Dodawanie nowych wierszy	Modyfikowanie istniejących wierszy	Historia zmian wartości atrybutu
Stały	Nie	Nie	Nie
0	Tak	Nie	Nie
1	Tak	Tak	Nie
2	Tak	Tylko kolumny robocze	Tak
3	Nie	Tak	Niepełna, tylko dla wybranego atrybutu
4	Tak, w tabeli historycznej	Tak	Tak

2.13. Tabela faktów

Modele wielowymiarowe zawierają fakty, które są zdefiniowane przez wymiary i wartości miar. Tabele faktów mogą mieć różną ziarnistość, czyli poziom szczegółowości danych. Jest ona dobierana w zależności od analizy, którą chce się wykonać, która byłaby z biznesowego punktu widzenia najbardziej adekwatna. Sama tabela faktów powinna składać się z klucza głównego oraz kluczy obcych, które wskazują na relacje z

pozostałymi tabelami z modelu. Było to już wskazywane w poprzednich akapitach. Ważne do odnotowania jest natomiast pojęcie klucza kandydującego, czyli takiej kolumny bądź ich grupy, która pozwala na identyfikację rekordu, tak jak klucz główny, ale nie to było celem ich umieszczenia w tabeli. Dla przykładu można podać dane osobowe, które można rozumieć na kilka sposobów, bowiem zarówno PESEL, jak i nr dowodu tożsamości, jak i imię i nazwisko z datą urodzenia pozwalają na identyfikację konkretnej osoby. Jeden z tych sposobów będzie kluczem głównym, pozostałe - odrzucone, kluczami kandydującymi właśnie. Jest to istotne zadanie przed twórcą hurtowni, ponieważ wybór możliwie najlepszego, naturalnego klucza głównego determinuje możliwość zaistnienia zduplikowanych rekordów w procesie ETL [19].

W kwestii miar istotne jest zaś pamiętanie o celowej nadmiarowości, tzn. o sytuacjach, które występują, gdy jedna miara zależy od drugiej, np. zapłacony podatek VAT zależy od kwoty transakcji. Mimo że można łatwo ten procent obliczyć, to tego się nie robi, ponieważ przy dużej ilości danych jest to czaso- i pamięciochłonne, po prostu nieoptymalne. Do tego nieprzydatne w analizach, gdzie potrzebujemy już nominalnych kwot, a nie formuł na obliczanie rekordów w kolumnie [19].

Wyróżnia się trzy typy miar: addytywne, semiaddytywne i nieaddytywne. Pierwsze z nich to takie, które są sumowalne dla każdego wymiaru, np. zysk brutto. Drugie z nich to taki rodzaj miar, który jest sumowalny tylko dla niektórych wymiarów, np. stan magazynu. Trzecie z nich to takie, które byłyby dodane wbrew logice tabeli i nie niosą ze sobą żadnego sensu [19]. Warto zaznaczyć, że istnieją także tabele faktów bez miar, czyli tabele ewidencyjne oraz zdarzeniowe [19].

Wyróżnia się trzy typy tabel faktów, których to typów nie należy ze sobą łączyć: tabele transakcyjne (transactional), migawki okresowe (periodic snapshot) oraz migawki przyrostowe (accumulating snapshot). Pierwszy z nich, to najczęściej spotykany typ, polegający na tym, że jedno wydarzenie biznesowe, jedna transakcja to jeden fakt z tabeli. Drugi z nich, to typ, w którym przechowuje się dane opisujące powtarzający się okres, jak dzień, miesiąc, rok. Trzeci z nich, to najrzadziej wykorzystywany typ tabeli faktów, ale nieraz potrzebny, bowiem konieczny do jego zastosowania jest ściśle określony proces, typu badanie analityczne czy przetwarzanie zamówienia, gdzie dokładnie znany jest każdy etap, co po czym następuje itp. [19].

2.14. Hurtownie danych w biznesie

Należy zaznaczyć, iż gromadzenie danych w ramach hurtowni wymaga tego, aby na początku już wiedzieć jakich danych się potrzebuje, przewidzieć potencjalne ryzyka i zaplanować jej strukturę. Oznacza to, że jest to działanie trudniejsze niż po prostu bieżąca działalność operacyjna. Podstawowym problemem biznesu jest bowiem fakt, że bardzo często nie wie, czego tak naprawdę chce i rolą analityka jest wiedzieć za niego, co może mu się przydać. W wielkich korporacjach projektant hurtowni, programista czy analityk mają jasne określone role, ale w pozostałych firmach różnie to wygląda. Jak wskazuje Agnieszka Chodkowska-Gyurics [19], „rolę analityka można obrazowo określić jako tłumacza między światem biznesu a światem techniki. Dobry tłumacz […] zna nie tylko oba języki, lecz musi mieć pojęcie o kontekście kulturowym zarówno twórcy, jak i czytelnika". Wymaga się też od niego aktywnego uczestnictwa w definiowaniu zarówno wymagań biznesowych, jak i technicznych, obie fazy powinny także zawierać dialog z projektantem hurtowni, zwłaszcza w drugiej. W pierwszej bowiem to co istotne, to zrozumienie, że nie da się dla każdego przedsiębiorstwa stworzyć uniwersalnej hurtowni. Trzeba zrozumieć kontekst działalności oraz otoczenie konkurencyjne, aby wiedzieć co można danemu biznesowi zaproponować. W porozumieniu z firmą należy zatem ustalić strategię analizy danych i dostosować strukturę hurtowni do niej. Bez wątpienia jednym z najtrudniejszych etapów jest przełożenie zdefiniowanych już wymagań biznesowych na język techniki, kiedy powstaje model logiczny bazy danych. I to ten etap i rzetelność jego realizacji determinuje ilość błędów w przyszłości oraz odpowiada za kształt fizyczny danej hurtowni danych [19].

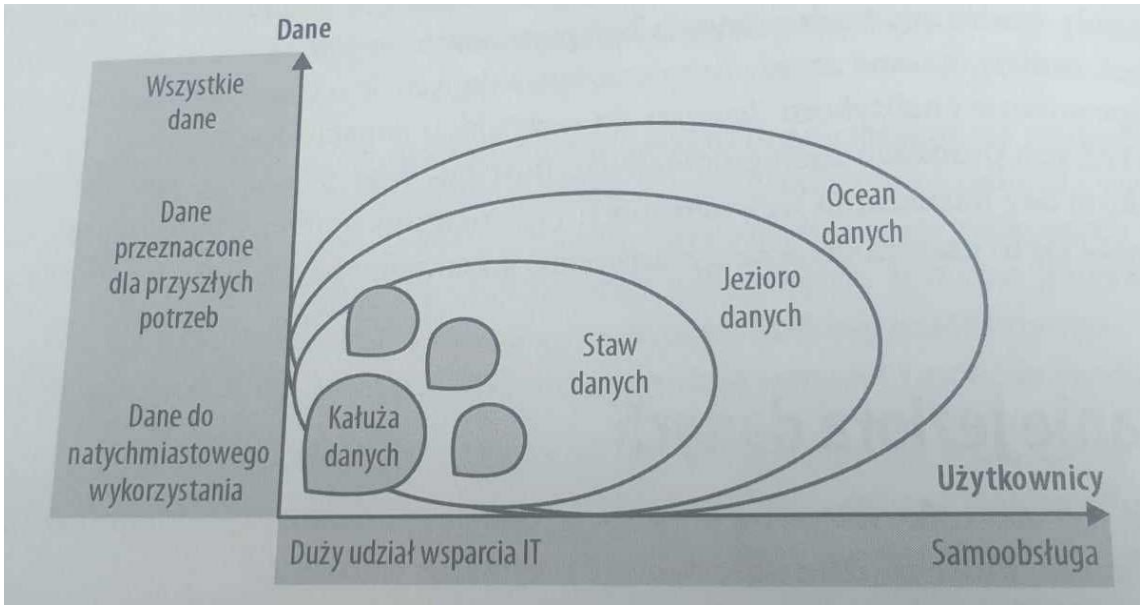
2.15. Big Data

Big Data to termin określający mnogość danych, z jakimi do czynienia firmy mają dzisiaj. Jest określana jako jedna z najważniejszych technologii do dalszego rozwoju branży. Można potraktować je jako ogromny strumień danych w czasie rzeczywistym, który przez swą objętość nie niesie żadnych wartości, dopiero umiejętność łączenia różnych typów danych pozwala na jej wyciągnięcie [20].

2.16. Jeziora danych

Z powodu przeniknięcia ogromu danych do życia, tak ludzi, jak i wielkich firm, należy sprostać wyzwaniom jakie to zjawisko stawia. Przede wszystkim stawia to rozwiązania opisane w poprzednich podrozdziałach jako archaiczne i nieprzystające do obecnych

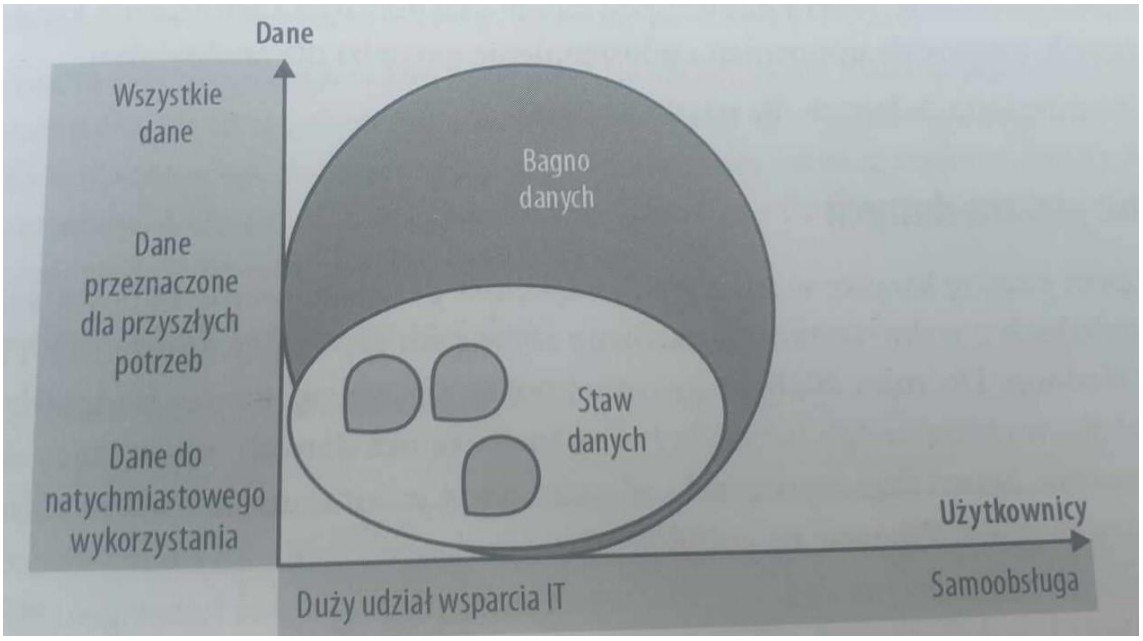
czasów, także z powodu niewydolności systemów ręcznie stworzonych przez człowieka wobec takiego ogromu danych. Jednym ze sposobów odpowiedzi na to zapotrzebowanie zastąpienia hurtowni danych przy wykorzystaniu Big Data jest jezioro danych. Aby jednak wyjaśnić czym ono jest, najpierw należy wyjaśnić czym jest kałuża danych (ang. data puddle) i staw danych (ang. data pond). Pierwsza z koncepcji to skład danych dla pojedynczego celu, projektu, zespołu, a głównym powodem wykorzystania Big Data zamiast hurtowni danych to po prostu troska o obniżenie kosztów i zwiększenie wydajności. Druga koncepcja to zbiór kałuż danych, co choć czasem może wyglądać jak źle zaprojektowana hurtownia danych, to może wpływać podobnie jak pojedyncza kałuża, przy okazji będąc tańszym zamiennikiem hurtowni. Problemem jest niestety duże zaangażowanie działu IT w utrzymanie tegoż zbioru. Z tego powodu wymyślono koncepcję jezior danych, które różnią się od stawów większą samoobsługą, co pozwala ograniczyć konieczność korzystania z wiedzy działu IT, a także szerszym dostępem do danych, także tymi, które nie są w danej chwili potrzebne końcowym użytkownikom. Jeszcze szerszym podejściem są oceany danych, które zwiększają skalę zalet jezior danych [21], co widać na rysunku 9:



Rys.9. Cztery etapy dojrzałości, źródło: Alex Gorelik, Korporacyjne jezioro danych. Wykorzystaj potencjał big data w swojej organizacji. Helion, 2019

Należy wspomnieć jednak o ryzyku bagna danych (ang. data swamp), czyli stawów danych, które stały się jeziorami, ale z powodu braku odpowiedniej liczby analityków

odstraszonych niską samoobsługą, większość zawartych danych pozostaje nieudokumentowana, niejako w „szarej strefie” [21], co przedstawia rysunek 10.



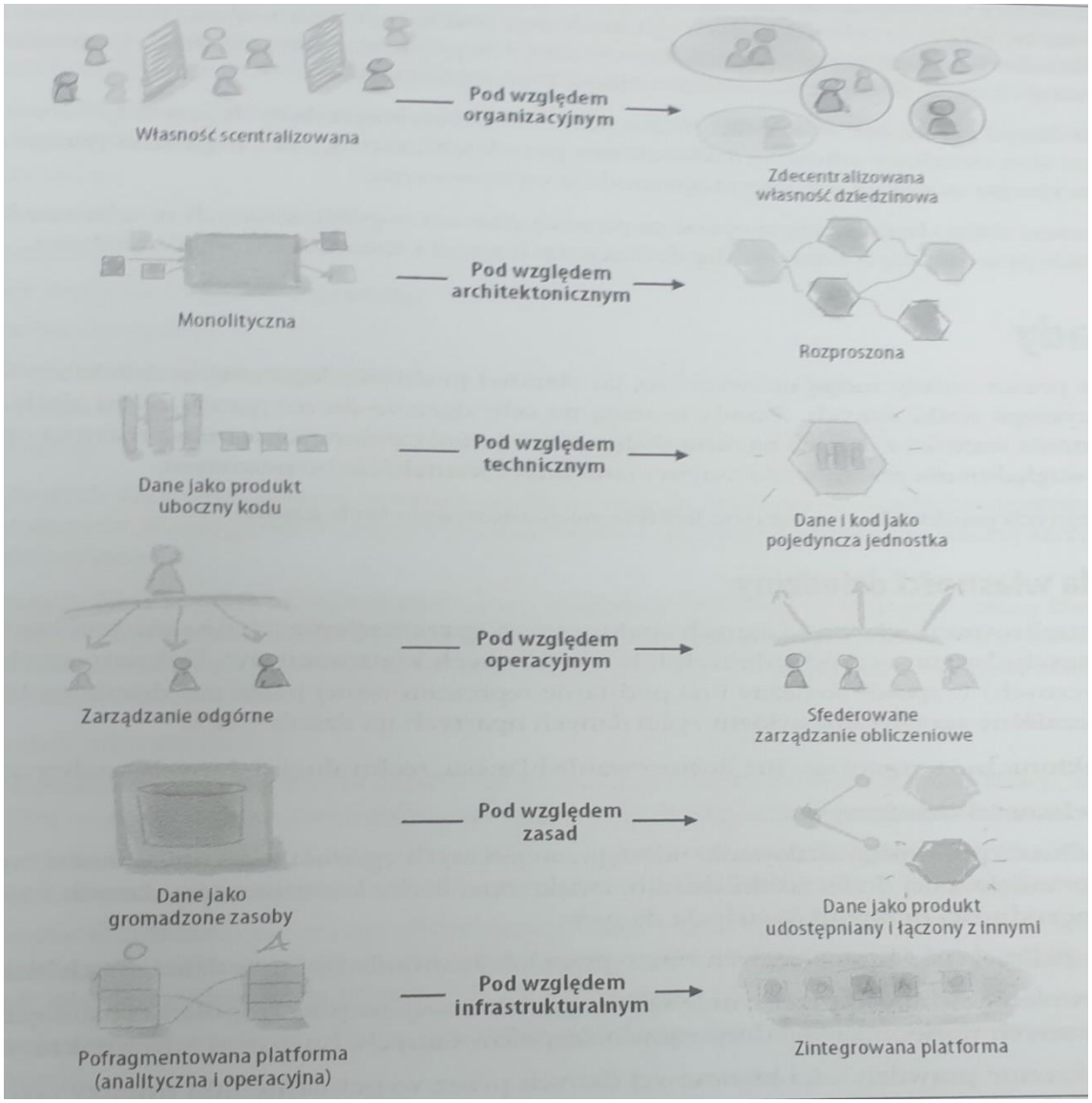
Rys.10. Bagno danych, źródło: Alex Gorelik, Korporacyjne jezioro danych. Wykorzystaj potencjał big data w swojej organizacji. Helion, 2019

Proces przekształcania się hurtowni danych w jeziora danych, z etapem przejściowym w postaci w stawów danych można opisać następująco. Stawy danych służą świetnie zarówno do przechowywania danych, czyli spełniają podstawową funkcję hurtowni, ale dzięki wykorzystaniu technologii Big Data, takich jak np. Hadoop, są równie dobrym miejscem do analizy i przekształcania danych. To już odróżnia je od hurtowni, względem których ta transformacja odbywa się zewnętrznie (ETL) lub wewnętrznie (ELT), ale nadal jako część szerszego procesu ETL, który jest odciążany właśnie przez stworzenie stawów. Wyewoluowane w stawy hurtownie danych zawierają do analizy zarówno dane surowe, jak i przekształcone, a potem rozszerzają się dalej poprzez zasysanie kolejnych danych, niepochodzących z oryginalnej hurtowni. W takiej formie uprawnionym jest nazywanie ich w pełni jeziorami danych [21].

2.17. Siatka danych

Jak wcześniej zostało to wspomniane, jeziora danych to nowoczesne podejście, które miało spowodować zastąpienie hurtowni danych lepszymi rozwiązaniami. Okazuje się jednak, że opracowywane są jeszcze nowsze koncepcje i jedną z nich jest właśnie siatka danych. Hurtownie danych w toku spływania coraz większej ilości różnorodnych danych

okazywały się być zbyt scentralizowanymi narzędziami do skutecznej analizy. Jeziora danych, które miały je zastąpić z kolei zbyt często w praktyce biznesowej przekształcają się w bagna danych [22]. Z tymi problemami, przynajmniej do wynalezienia jeszcze lepszego narzędzia, mają uporać się siatki danych (ang. data mesh), czyli „zdecentralizowane socjotechniczne podejście do współdzielenia i udostępniania danych analitycznych i zarządzania nimi w złożonych i dużych środowiskach – wewnątrz jednej organizacji lub w ramach wielu" [22]. Zmiany jakie siatka danych wprowadza w stosunku do poprzednich rozwiązań zarządzania danymi analitycznymi pokazuje rysunek 11.



Rys.11. Wymiar zmian wprowadzonych przez siatkę danych, źródło: Zhamak Dehghani, Siatka danych. Nowoczesna koncepcja samoobsługowej infrastruktury danych. Helion, 2023

Co można wywnioskować z kolejnych podrozdziałów historia modeli danych jest bardzo bogata, nowsze wypierają starsze, niewiele jest takich narzędzi które zostały tylko nieznacznie zmienione wraz z upływem czasu. Coraz to nowsze podejścia wymagają od analityków dostosowywania się do nich i uczenia się nowych umiejętności, ale ciekawość i złożoność tej materii musi w pewnym sensie wynagradzać włożone w to zaangażowanie.

2.18. Power BI, DAX i Power Query

Zanim opisany zostanie proces i cel przygotowania dashboardu managerskiego, należy uzupełnić pewne wprowadzenie o programie, który to umożliwia, czyli Power BI. Z początku był to jedynie dodatek do MS Office, od 2015 roku jest to osobna usługa, cały czas doskonalona. Tam właśnie jest wykorzystany język DAX, czyli Data Analysis eXpressions, który służy pisaniu miar w Power BI, czyli obliczaniu wartości na podstawie danych z kolumn w tabelach [23]. DAX istotnie różni się od innych języków programowania, podobnie jak SQL, który jest językiem stricte bazodanowym i deklaratywnym, tak „DAX jest językiem zaprojektowanym specjalnie w celu przetwarzania formuł biznesowych w modelu danych" [24]. Co więcej należy zauważyć, że DAX jest językiem stricte funkcyjnym, w którym wszystko jest wyrażeniem mającym przywołać daną funkcję, w przeciwieństwie innych języków programowania np. do Pythona, nie zawiera on koncepcji pętli czy poleceń [24]. Różnicą w stosunku do SQL z kolei, jest przede wszystkim obsługiwanie relacji, które w SQL należy określać w zapytaniach, w DAX z kolei nie ma takiej potrzeby, z racji wbudowania relacji w model, który DAX rozumie i automatycznie do niego się odnosi [24].

Power Query jest z kolei narzędziem pozwalającym wprowadzać istotne poprawki już po wgraniu baz danych do Power BI. Jest także dostępny w Excelu, jednak pełnię możliwości języka M można wykorzystać właśnie przy pomocy Power BI.

Podobnie jak w przypadku SQL, należy wspomnieć o konieczności nauki podstaw narzędzi, aby w ogóle móc zacząć takie dashboardy tworzyć.

Rozdział III. Metodologia – praktyka bazodanowa – przygotowanie danych za pomocą SQL Server, Excel, Power Query, Power BI i DAX

Podczas pracy nad obróbką danych w pobranej bazie, najważniejsze było przygotowanie do analizy dwóch kolumn, ponieważ pozostałe kolumny były już oczyszczone i nie wymagały większych zmian. Pierwszą z nich była kolumna Offer_location, zawierająca jak nazwa wskazuje, miejsca wystawienia oferty sprzedaży danego auta. Jej pierwsze wiersze wyglądały następująco:

Offer_location
ul. Jubilerska 6 - 04-190 Warszawa, Mazowieckie (Polska)
kanonierska12 - 04-425 Warszawa, RembertÄłw (Polska)
Warszawa, Mazowieckie, BiaĹ‚oĹ‚Ä™ka
Jaworzno, ĹšlÄ…skie
ul. GorzysĹ‚awa 9 - 61-057 PoznaĹ„, Nowe Miasto (Polska)
ModliĹ„ska 157 - 03-186 Warszawa, BiaĹ‚oĹ‚Ä™ka (Polska)
Ĺ»ary, ĹĽarski, Lubuskie
Ojcowska 2 - 02-918 Warszawa, MokotÄłw (Polska)
ul. GorzysĹ‚awa 9 - 61-057 PoznaĹ„, Nowe Miasto (Polska)
ul. GorzysĹ‚awa 9 - 61-057 PoznaĹ„, Nowe Miasto (Polska)
GdaĹ„sk, Pomorskie, ĹšrÄ…dmieĹ›cie
Bielsko-BiaĹ‚a, ĹšlÄ…skie
RzeszÄłw, Podkarpackie
SulechÄłw, zielonogÄłrski, Lubuskie
Koszalin, Zachodniopomorskie
Zgorzelec, zgorzelecki, DolnoĹ›lÄ…skie
ul. Jubilerska 6 - 04-190 Warszawa, Mazowieckie (Polska)

Rys.12. Pierwsza, „brudna" wersja kolumny z lokalizacjami w bazie danych, źródło: opracowanie własne

Jak można zauważyć, problemem była ogromna nieregularność danych w komórkach – czasem pojawiała się tylko nazwa miasta, wraz z województwem, w którym się znajduje, czasem do tego występowała nazwa ulicy, adres, kod pocztowy, nazwa powiatu, bądź gminy, a czasem zupełnie niepowiązane z samą miejscowością informacje lub ciągi znaków. Nie sposób ustalić także samą regułę pojawiania się tych informacji, ponieważ wszystkie pojawiają się bez wyznaczonego schematu. Ponadto brak zachowania polskich znaków diakretycznych utrudniał odczytanie nazw i to również wymagało naprawy.

Na początek aby było w ogóle możliwe zaimportowanie pliku płaskiego csv do SQL Servera, należało podzielić tę kolumnę, ponieważ część komórek w niej była zwyczajnie za długa aby mogła stanowić jedną wartość – nie zgadzał się typ danych. Aby tego dokonać, zaimportowano ów plik do programu Power BI. Tam zastosowano polecenie

Split column by delimiter w Power Query, zaś tym według czego ją rozdzielano były przecinki, a potem także spacje. Chodziło o to, aby maksymalnie w jednej kolumnie znajdowało się w komórkach jedno słowo, tak aby możliwe było zaimportowanie bazy do programu SQL Server. Efekt końcowy tegoż podziału wyglądał następująco:

Offer_location 1 1 1	Offer_location 1 ...	Offer_locatio...	Offer_locatio...	Offer_locat...	Offer...	Offer_l...	Offer_loc...	Offer...	Offer_...	Offer_location 2 1	Offer_location 3
ZamoĹ›Ä‡										Lubelskie	
LubaczÃłw										lubaczowski	Podkarpackie
Szadkowska	65	-	98-220	ZduĹ„ska	Wola					zduĹ„skowolski	ĹÃłdzkie (Polska)
Aleja	MARSZAĹKA	JÃ“ZEFA	PIĹSUDSKI...	1	-	33-300	Nowy	SÄ…cz		MaĹ‚opolskie	
ChodzieĹĽ										chodzieski	Wielkopolskie
Skierniewice										ĹÃłdzkie	
Aleja	Prymasa	TysiÄ…clecia	64	-	01-4...	Warsz...				Wola	
LÄ™bork										lÄ™borski	Pomorskie
Osobnica										jasielski	Podkarpackie
Przebieczany										wielicki	MaĹ‚opolskie

Rys.13. Efekt działania „split column by delimiter”, źródło: opracowanie własne

3.1. Zmiana znaków na polskie

Można dostrzec że podobnie jak wcześniej, nie ma żadnej reguły pojawiania się danych w kolumnach – raz może być to nazwa miasta, innym razem nazwa ulicy, raz nazwa powiatu. W tym miejscu należy zacząć przekodowanie znaków na polskie znaki diakretyczne:

```
UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Äąâ€š', 'ł')
WHERE Nazwa1 LIKE '%Äąâ€š%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Ă„â€¦', 'ą')
WHERE Nazwa1 LIKE N'%Ă„â€¦%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Ä‚Ĺ‚', 'ó')
WHERE Nazwa1 LIKE N'%Ä‚Ĺ‚%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Ă„â„˘', 'ę')
WHERE Nazwa1 LIKE N'%Ă„â„˘%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Äąâ€ž', 'ń')
WHERE Nazwa1 LIKE N'%Äąâ€ž%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Äąâ€ş', 'ś')
WHERE Nazwa1 LIKE N'%Äąâ€ş%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'ÄąĹş', 'ź')
WHERE Nazwa1 LIKE N'%ÄąĹş%';

UPDATE Carsales1
```

```
SET Nazwa1 = REPLACE(Nazwa1, 'ÄąÂ', 'Ł')
WHERE Nazwa1 LIKE N'%ÄąÂ%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'ÄąÄ˝', 'ż')
WHERE Nazwa1 LIKE N'%ÄąÄ˝%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'ÄąĹ˜', 'Ś')
WHERE Nazwa1 LIKE N'%ÄąĹ˜%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Ł»', 'Ż')
WHERE Nazwa1 LIKE N'%Ł»%';

UPDATE Carsales1
SET Nazwa1 = REPLACE(Nazwa1, 'Ă„â€˜', 'ć')
WHERE Nazwa1 LIKE N'%Ă„â€˜%';
```

3.2. Kolumny z miejscowościami

Aby stworzyć kolumnę województwa, która zawierałaby tylko nazwy miejscowości z których pochodzi oferta, potrzebna jest baza rozkodowująca. W tym wypadku będzie to baza wszystkich polskich miejscowości w Polsce, pobrana ze stron Głównego Urzędu Statystycznego. Poza nazwą samej miejscowości, zawiera ona także kod TERYT województwa, powiatu oraz gminy, co ułatwia identyfikację miejscowości o tych samych nazwach, ale znajdujących się w różnych jednostkach samorządu terytorialnego. Ponadto jest także kolumna JPT_KOD_JE, która jest połączeniem kodu TERYT województwa i powiatu, a także kolumna z rodzajem gminy, co jest o tyle ważne, że baza ta zawierała także nazwy dzielnic dużych miast, co przy pierwszej próbie rozkodowywania sprawiło duże kłopoty. Jeśli nazwy dzielnic występowały także w którejś kolumnie bazy sprzedażowej, to także one pojawiały się jako rezultat rozkodowywania, a były jednak elementem niepożądanym, bo zbyt szczegółowym dla przedmiotu analizy. Z bazy rozkodowującej zatem zostały usunięte wszystkie nazwy dzielnic, osiedli bądź części miast. Następnie baza rozkodowująca została zaimportowana do środowiska SQL Server i tam połączona z bazą sprzedażową, w następujący sposób:

```
SELECT DISTINCT
    C.ID,
    M1.NAZWA AS Nazwa1,
    M2.NAZWA AS Nazwa2,
    M3.NAZWA AS Nazwa3,
    M4.NAZWA AS Nazwa4,
    M5.NAZWA AS Nazwa5,
    M6.NAZWA AS Nazwa6,
    M7.NAZWA AS Nazwa7,
    M8.NAZWA AS Nazwa8,
```

```
    M9.NAZWA AS Nazwa9,
    M10.NAZWA AS Nazwa10,
    M11.NAZWA AS Nazwa11,
    M12.NAZWA AS Nazwa12
INTO Magisterka.dbo.Locat
FROM Magisterka.dbo.Car_sale1 C
LEFT JOIN Magisterka.dbo.Miasta M1 ON C.[Offer_location 1 1 1] = M1.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M2 ON C.[Offer_location 1 1 2] = M2.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M3 ON C.[Offer_location 1 1 3] = M3.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M4 ON C.[Offer_location 1 1 4] = M4.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M5 ON C.[Offer_location 1 1 5] = M5.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M6 ON C.[Offer_location 1 1 6] = M6.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M7 ON C.[Offer_location 1 1 7] = M7.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M8 ON C.[Offer_location 1 1 8] = M8.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M9 ON C.[Offer_location 1 1 9] = M9.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M10 ON C.[Offer_location 1 1 10] = M10.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M11 ON C.[Offer_location 2 1] = M11.NAZWA
LEFT JOIN Magisterka.dbo.Miasta M12 ON C.[Offer_location 3] = M12.NAZWA
ORDER BY ID;
```

Powyższy kod skutkuje poniższym rezultatem:

Results Messages

	ID	Nazwa1	Nazwa2	Nazwa3	Nazwa4	Nazwa5	Nazwa6	Nazwa7	Nazwa8	Nazwa9	Nazwa10	Nazwa11	Nazwa12
1	1	NULL	Józefa	NULL	NULL	NULL	NULL	Stargard	NULL	NULL	NULL	NULL	NULL
2	2	Graniczna	NULL	NULL	NULL	Płock	NULL	NULL	NULL	NULL	NULL	NULL	NULL
3	3	Bielsko-Biała	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
4	4	Janów	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
5	5	NULL	NULL	NULL	NULL	Piaseczno	NULL	NULL	NULL	NULL	NULL	NULL	NULL
6	6	Kraków	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
7	7	NULL	NULL	NULL	NULL	Warszawa	NULL	NULL	NULL	NULL	NULL	NULL	NULL
8	8	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Zabrze	NULL	NULL	NULL
9	9	Częstochowa	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
10	10	NULL	NULL	NULL	NULL	NULL	Mysłowice	NULL	NULL	NULL	NULL	NULL	NULL
11	11	Płock	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Rys.14. Efekt działania kodu łączącego z bazą rozkodowującą, źródło: opracowanie własne

3.3. Scalenie kolumn

Teraz należy dokonać połączenia wszystkich kolumn w jedną:

```
-- Dodanie nowej kolumny do przechowywania zbiorczych wartości
ALTER TABLE Magisterka.dbo.Locat
ADD Miejscowosc NVARCHAR(MAX);

-- Aktualizacja nowej kolumny, łącząc wszystkie wartości z jednego rekordu
UPDATE Magisterka.dbo.Locat
SET Miejscowosc = CONCAT_WS(', ',
                            ISNULL(Nazwa1, ''),
                            ISNULL(Nazwa2, ''),
                            ISNULL(Nazwa3, ''),
                            ISNULL(Nazwa4, ''),
                            ISNULL(Nazwa5, ''),
                            ISNULL(Nazwa6, ''),
                            ISNULL(Nazwa7, ''),
```

```
                ISNULL(Nazwa8, ''),
                ISNULL(Nazwa9, ''),
                ISNULL(Nazwa10, ''),
                ISNULL(Nazwa11, ''),
                ISNULL(Nazwa12, ''));
```

Results Messages

	Nazwa3	Nazwa4	Nazwa5	Nazwa6	Nazwa7	Nazwa8	Nazwa9	Nazwa10	Nazwa11	Nazwa12	Miejscowosc
1	NULL	NULL	NULL	NULL	Stargard	NULL	NULL	NULL	NULL	NULL	, Józefa, , , , , Stargard, , , , ,
2	NULL	NULL	Płock	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Graniczna, , , , Płock, , , , , , ,
3	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Bielsko-Biała, , , , , , , , , , ,
4	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Janów, , , , , , , , , , ,
5	NULL	NULL	Piaseczno	NULL	NULL	NULL	NULL	NULL	NULL	NULL	, , , , Piaseczno, , , , , , ,
6	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Kraków, , , , , , , , , , ,
7	NULL	NULL	Warszawa	NULL	NULL	NULL	NULL	NULL	NULL	NULL	, , , , Warszawa, , , , , , ,
8	NULL	NULL	NULL	NULL	NULL	NULL	Zabrze	NULL	NULL	NULL	, , , , , , , , Zabrze, , ,
9	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Częstochowa, , , , , , , , , , ,
10	NULL	NULL	NULL	Mysłowice	NULL	NULL	NULL	NULL	NULL	NULL	, , , , , Mysłowice, , , , , ,
11	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Płock

Rys.15. Efekt działania kodu łączącego wszystkie dane w jednej kolumnie, źródło: opracowanie własne

Poniższa metoda nie zapewniła jednak dostatecznych rezultatów – widać, że już w pierwszych rekordach znalazło się słowo, które nie powinno się tam znaleźć. Tak jednak się stało z powodu tego, że JOIN połączył obie tabele w taki sposób, że jeśli jakiekolwiek słowo znajdujące się w odpowiedniej kolumnie bazy rozkodowującej, to przypisuje to słowo w bazie sprzedażowej, bez kontekstu. To znaczy, że jeżeli w bazie rozkodowującej znalazła się jakaś miejscowość która w swojej nazwie posiada przymiotnik „Graniczna", a w którejś z kolumn bazy sprzedażowej również to słowo wystąpiło, np. w charakterze ulicy – to także zostało to przypisane w wyniku połączenia. Aby zapobiec tego typu błędom, usprawniono tę metodę. W bazie rozkodowującej nazwy miejscowości zostały ujęte w nawias kwadratowy, po to, aby JOIN brał pod uwagę jedynie te wartości, które są dokładnie takie same. Pozwoliło to na uniknięcie sytuacji, w której wystąpienie tylko jednego z członów nazwy powodowało pozytywny rezultat.

```
UPDATE Magisterka.dbo.Miasta
SET Magisterka.dbo.Miasta.NAZWA = '[' + ISNULL(Magisterka.dbo.Miasta.NAZWA, '') +
']';
```

W bazie sprzedażowej natomiast, należy pozbyć się przecinków i zamienić je na spacje:

```
UPDATE Magisterka.dbo.Locat
SET Miejscowosc = REPLACE(Miejscowosc, ', ', ' ')
WHERE Miejscowosc LIKE '%, %';
```

Aby mieć wszystkie nazwy w bazie sprzedażowej w nawiasach kwadratowych:

```
-- Aktualizacja wartości w kolumnie Miejscowosc, okalając nazwy nawiasami
kwadratowymi, z wyjątkiem tych oddzielonych tylko jedną spacją
UPDATE Magisterka.dbo.Locat
SET Miejscowosc = '[' + TRIM(Miejscowosc) + ']'
WHERE Miejscowosc LIKE '% %';
```

Można zauważyć, że miejscowości dwuczłonowe, bez myślnika w nazwie, np. Zielona Góra czy Gorzów Wielkopolski, są oddzielone tylko jedną spacją. Z kolei inne wartości tekstowe, które nie są miejscowościami o wieloczłonowych nazwach, są oddzielone wieloma spacjami. Aby je łatwiej zidentyfikować zamieniamy je na ukośniki:

```
-- Aktualizacja wartości w kolumnie Miejscowosc, zamieniając dwie spacje na slash
UPDATE Magisterka.dbo.Locat
SET Miejscowosc = REPLACE(TRIM(Miejscowosc), '  ', '/');
```

Skutkuje to poniższym rezultatem:

Results | Messages

	Nazwa3	Nazwa4	Nazwa5	Nazwa6	Nazwa7	Nazwa8	Nazwa9	Nazwa10	Nazwa11	Nazwa12	Miejscowosc
283	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Opole]
284	NULL	NULL	Gniezno	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Gniezno]
285	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Biestrzyków]
286	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Grota]
287	NULL	NULL	Kowala	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Kotarwice//Kowala]
288	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Warszawa]
289	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Radom]
290	NULL	NULL	Warszawa	NULL	NULL	NULL	NULL	NULL	Bemowo	NULL	[Połczyńska//Warszawa///Bemowo]
291	NULL	NULL	Zielona	Góra	NULL	NULL	NULL	NULL	NULL	NULL	[Zielona Góra]
292	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[]
293	NULL	NULL	Szczecin	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Szczecin]
294	NULL	NULL	Masłowo	NULL	NULL	NULL	NULL	NULL	NULL	NULL	[Masłowo]

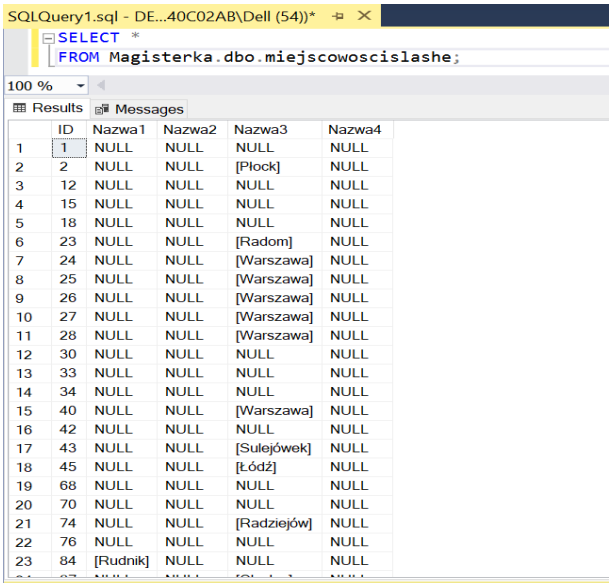
Rys.16. Efekt działania kodu zmieniającego spacje na ukośniki oraz dodającego nawiasy kwadratowe w celu identyfikacji nazw dwuczłonowych, źródło: opracowanie własne

3.4. Łączenie tabel z bazą rozkodowującą

Kolejnym krokiem jest rozbicie kolumny Miejscowosc po ukośnikach, a następnie powtórzenie mechanizmu łączenia tabel z bazą rozkodowującą. Najpierw jednak należy przenieść do osobnej tabeli nazwy miast, w których nie występują ukośniki – są one gotowe do połączenia. Potem to samo należy powtórzyć z miejscowościami do dalszej obróbki, tych w których występują ukośniki.

```
UPDATE Magisterka.dbo.Miejscowosci
SET [Column2 1 1 1] = '[' + ISNULL([Column2 1 1 1], '') + ']',
    [Column2 1 1 2] = '[' + ISNULL([Column2 1 1 2], '') + ']',
    [Column2 1 1 3] = '[' + ISNULL([Column2 1 1 3], '') + ']',
    [Column2 1 2] = '[' + ISNULL([Column2 1 2], '') + ']',
    [Column2 2] = '[' + ISNULL([Column2 2], '') + ']';

SELECT DISTINCT C.ID, P1.NAZWA, P2.NAZWA, P3.NAZWA, P4.NAZWA, P5.NAZWA
INTO Magisterka.dbo.miejscowoscislashe
FROM Magisterka.dbo.Miejscowosci C
LEFT JOIN Magisterka.dbo.Miasta P1 ON C.[Column2 1 1 2]=P1.NAZWA
LEFT JOIN Magisterka.dbo.Miasta P2 ON C.[Column2 1 1 2]=P2.NAZWA
LEFT JOIN Magisterka.dbo.Miasta P3 ON C.[Column2 1 1 3]=P3.NAZWA
LEFT JOIN Magisterka.dbo.Miasta P4 ON C.[Column2 1 2]=P4.NAZWA
LEFT JOIN Magisterka.dbo.Miasta P5 ON C.[Column2 2]=P5.NAZWA
ORDER BY ID;
```



	ID	Nazwa1	Nazwa2	Nazwa3	Nazwa4
1	1	NULL	NULL	NULL	NULL
2	2	NULL	NULL	[Płock]	NULL
3	12	NULL	NULL	NULL	NULL
4	15	NULL	NULL	NULL	NULL
5	18	NULL	NULL	NULL	NULL
6	23	NULL	NULL	[Radom]	NULL
7	24	NULL	NULL	[Warszawa]	NULL
8	25	NULL	NULL	[Warszawa]	NULL
9	26	NULL	NULL	[Warszawa]	NULL
10	27	NULL	NULL	[Warszawa]	NULL
11	28	NULL	NULL	[Warszawa]	NULL
12	30	NULL	NULL	NULL	NULL
13	33	NULL	NULL	NULL	NULL
14	34	NULL	NULL	NULL	NULL
15	40	NULL	NULL	[Warszawa]	NULL
16	42	NULL	NULL	NULL	NULL
17	43	NULL	NULL	[Sulejówek]	NULL
18	45	NULL	NULL	[Łódź]	NULL
19	68	NULL	NULL	NULL	NULL
20	70	NULL	NULL	NULL	NULL
21	74	NULL	NULL	[Radziejów]	NULL
22	76	NULL	NULL	NULL	NULL
23	84	[Rudnik]	NULL	NULL	NULL

Rys.17. Efekt działania kodu rozbijającego dane po ukośnikach, źródło: opracowanie własne

Następnie tworzy się połączenie tych wszystkich wierszy w jedną kolumnę, łącznie z pozbyciem się nawiasów kwadratowych, które po połączeniu dwóch tabel, nie są już potrzebne:

```
;WITH Grupa AS(
SELECT
    ID,
    CONCAT(
        COALESCE(Nazwa1, ''),
        COALESCE(Nazwa2, ''),
        COALESCE(Nazwa3, ''),
        COALESCE(Nazwa4, '')
    ) AS Miejscowosc
FROM Magisterka.dbo.miejscowoscislashe)
SELECT
```

```
    ID,
    NULLIF(
        REPLACE(
            REPLACE(
                REPLACE(
                    REPLACE(
                        REPLACE(Miejscowosc, '[', ''),
                        ']', ''),
                    'NULL, ', ','),
                'NULL', ''),
            ', ,', ','),
        '') AS Miejscowosc
INTO Miejscowoscizeslashemdocs1
FROM
    Grupa;
```

Tworzy się w ten sposób jedną zbiorczą kolumnę, która ze wszystkich rekordów, w których pojawiały się miejscowości z ukośnikami i po połączeniu ich z bazą rozkodowującą, zwraca te rekordy, w których nazwy zostały rozkodowane.

Results Messages

	ID	Miejscowosc
1	1	NULL
2	2	Płock
3	12	NULL
4	15	NULL
5	18	NULL
6	23	Radom
7	24	Warszawa
8	25	Warszawa
9	26	Warszawa
10	27	Warszawa

Rys.18. Efekt działania kodu łączącego wszystkie dane w jednej kolumnie, po uprzednich działaniach oczyszczających, źródło: opracowanie własne

3.5. Poprawa nazw miejscowości

Następnie trzeba pozbyć się nawiasów kwadratowych z tabeli rozkodowującej, jak i z tabeli miejscowości bez ukośników:

```
;WITH Gruppen AS(
       SELECT ID, Miejscowosc
       FROM MiejscowosciBezslashy)
SELECT
    ID,
    NULLIF(REPLACE(REPLACE(Miejscowosc, '[', ''), ']', ''), '') AS Miejscowosc
```

```
INTO Miejscowoscibezslashydocs1
FROM
    Gruppen
ORDER BY ID;
```

Gdy przygotowane w ten sposób są już dwie tabele miejscowości, tj. zarówno tabela miejscowości, w których pierwotnie występowały ukośniki, jak i te które już za pierwszym połączeniem z tabelą rozkodowującą zostały oczyszczone, należy połączyć je już bezpośrednio z bazą sprzedażową:

```
SELECT *
FROM Magisterka.dbo.Car_sale1 AS CS
LEFT JOIN Miejscowoscizeslashemdocs1 AS M1 ON CS.ID = M1.ID
LEFT JOIN MiejscowosciBezslashydocs3 AS M2 ON M2.ID_Car = CS.ID;
```

Results | Messages

	Features 47	Features 48	Features 49	Features 50	Features 51	Features 52	Features 53	ID	ID	Miejscowosc	ID_Car	Miejscowosc1
1								1	1	NULL	NULL	NULL
2								2	2	Plock	NULL	NULL
3								3	NULL	NULL	3	Bielsko-Biała
4								4	NULL	NULL	4	Janów
5								5	NULL	NULL	5	Piaseczno
6								6	NULL	NULL	6	Kraków
7								7	NULL	NULL	7	Warszawa
8								8	NULL	NULL	8	Zabrze
9								9	NULL	NULL	9	Częstochowa
10								10	NULL	NULL	10	Mysłowice
11								11	NULL	NULL	11	Plock

Rys.19. Efekt połączenia oczyszczonych nazw miejscowości z bazą sprzedażową, źródło: opracowanie własne

Potem należy połączyć dwie kolumny ID i dwie kolumny Miejscowości ze sobą:

```
;WITH Bla AS(

      SELECT
      COALESCE(ID_Car, ID_auto) AS ID,
    COALESCE(Miejscowosc1, Miejscowosc2) AS Place
FROM
    Magisterka.dbo.Cars3001)
SELECT *
INTO Magisterka.dbo.MiejscowosciOgulem
FROM Bla
WHERE Place IS NOT NULL
ORDER BY ID;
```

Otrzymamy wówczas następujący efekt:

Results Messages

	ID	Place
1	2	Płock
2	3	Bielsko-Biała
3	4	Janów
4	5	Piaseczno
5	6	Kraków
6	7	Warszawa
7	8	Zabrze
8	9	Częstochowa
9	10	Mysłowice
10	11	Płock
11	13	Katowice

Rys.20. Tabela wymiaru z nazwami miejscowości, źródło: opracowanie własne

3.6. Łączenie z tabelą faktów

Tę tabelę należy połączyć z tabelą faktów i w ten sposób uzyskany został pożądany już na samym początku rezultat, tj. otrzymana została kolumna z samą nazwą miejscowości, bez żadnych innych informacji towarzyszących.

SQLQuery3.sql - DE...40C02AB\Dell (51))* SQLQuery1.sql - DE...40C02AB\Dell (54))*

```
SELECT *
FROM Magisterka.dbo.MiejscowosciOgulem
JOIN Magisterka.dbo.Car_sale1 ON Magisterka.dbo.Car_sale1.ID = Magisterka.dbo.MiejscowosciOgulem.ID;
```

100 %

Results Messages

	ID	Place	Price	Condition	Vehicle_brand	Vehicle_model	Vehicle_generation	Production_year	Mileage_km	Power_HP	Displaceme
1	2	Płock	22300	Used	Alfa Romeo	159		2011	190000	120	1900
2	3	Bielsko-Biała	7200	Used	Alfa Romeo	156		2004	234000	140	1910
3	4	Janów	16900	Used	Alfa Romeo	159		2008	199666	150	1910
4	5	Piaseczno	32000	Used	Alfa Romeo	Giulietta		2011	162750	170	1368
5	6	Kraków	9300	Used	Alfa Romeo	147		2005	218000	120	1598
6	7	Warszawa	67500	Used	Alfa Romeo	Stelvio		2020	19000	280	1995
7	8	Zabrze	19500	Used	Alfa Romeo	159		2007	193000	150	1910
8	9	Częstochowa	29800	Used	Alfa Romeo	Brera		2006	283500	260	3195
9	10	Mysłowice	161800	Used	Alfa Romeo	Stelvio		2018	24000	280	1995
10	11	Płock	18900	Used	Alfa Romeo	159		2010	233000	120	1910
11	13	Katowice	13500	Used	Alfa Romeo	159		2008	246000	120	1910

Query executed successfully. DESKTOP-40C02AB (15.0 RTM) DESKTOP-40C02AB\Dell (54) master 00:00:13 162 460 rows

Rys.21. Tabela wymiarów połączona kluczem głównym z tabelą faktów, źródło: opracowanie własne

Aby było możliwe przypisanie do każdej miejscowości wspomnianych na początku kodów TERYT województwa, powiatu i gminy, należało połączyć tabelę z miejscowościami ogółem z tabelą kodów TERYT. Było to przydatne zarówno w analizie danych, jak i przyszłych wizualizacjach. Na tym etapie pojawia się jednak problem, mianowicie część miejscowości o tych samych nazwach pojawia się w wielu województwach. Oznacza to, że połączenie tabeli sprzedaży z tabelą kodów TERYT, zwielokrotni oferty w taki sposób, że ich ID przestanie być unikatowe wskutek przypisania jednej oferty z danego miejsca, do wszystkich miejsc o tej nazwie w Polsce, np. jeśli w kolumnie Place znajduje się miejscowość Wola, która znajduje się w np.7 województwach, to przy takim połączeniu, pojawi się 7 zwielokrotnionych rekordów z tym samym ID, różniących się między sobą jedynie kodem TERYT województwa bądź powiatu tejże miejscowości. Jest to problem nie do uniknięcia na tym etapie i takie połączenie zostało zastosowane. Aby jednak nie manipulować danymi i nie duplikować ich bez potrzeby, zostało przyjęte założenie, że w pierwszej kolejności liczą się gminy miejskie, potem miejsko-wiejskie, a na końcu gminy wiejskie. Jest bowiem problem, polegający na tym, że część miast, a więc gmin miejskich, jest także siedzibami gmin wiejskich, o takiej samej nazwie. Dlatego rekord np. z miejscowością Braniewo jest przypisany w takim połączeniu dwojako, zarówno do gminy miejskiej, jak i wiejskiej, o tej samej nazwie. W tej pracy przyjęto także założenie o tym, że w takich przypadkach jak ten, oferta auta zawsze pochodzi z gminy miejskiej. Dopiero kiedy nazwa miejscowości jest powiązana jedynie z gminami wiejskimi bądź miejsko-wiejskimi i nie dubluje się ona z żadnymi innymi, poniższy kod przypisze je odpowiednio. Jest to także pewne uproszczenie, ponieważ gmin wiejskich o tej samej nazwie może być w Polsce kilka lub kilkanaście. Aby nie dublować jednak ID sprzedaży, zostały wzięte pod uwagę jedynie gminy, które zostały wyświetlone poniższym kodem jako pierwsze w grupie.

```
;WITH Miejsca AS(
	SELECT ID, Place, WOJ, POW, JPT_KOD_JE, GMI, RODZ_GMI
	FROM Magisterka.dbo.Zapasik
	JOIN  Magisterka.dbo.MiejscowosciBaza  ON  Magisterka.dbo.Zapasik.Place  =
Magisterka.dbo.MiejscowosciBaza.NAZWA),
Miejsca2 AS(
		SELECT ID, Place, WOJ, POW, JPT_KOD_JE, GMI, RODZ_GMI,
		DENSE_RANK() OVER (PARTITION BY ID ORDER BY RODZ_GMI) AS DenseRanking,
		ROW_NUMBER () OVER (PARTITION BY ID ORDER BY RODZ_GMI) AS Ranking
FROM Miejsca),
Miejsca3 AS (
	SELECT *
	FROM Miejsca2
```

```
      WHERE DenseRanking = 1 AND Ranking = 1)
SELECT ID AS ID_Place, Place, WOJ, POW, JPT_KOD_JE, GMI, RODZ_GMI
INTO Magisterka.dbo.Miejscowosci0302
FROM Miejsca3;
```

Następnie przyłączamy tę kolumnę do całej tabeli, żeby sprawdzić ile wierszy w tabeli ofert sprzedaży pozostało:

```
SELECT *
INTO Magisterka.dbo.Wersja0302
FROM Magisterka.dbo.Miejscowosci0402
JOIN Magisterka.dbo.Zapasik ON ID_Place = ID
```

Następnie tworzymy tabelę zawierającą unikatowe wartości, tak aby móc nadać ID_Place i w ten sposób stworzyć dimension table, które będzie się po wspomnianym kluczu łączyć z tabelą ofert sprzedaży:

```
SELECT DISTINCT Place1, WOJ, POW, JPT_KOD_JE, GMI, RODZ_GMI
INTO Magisterka.dbo.DistinctPlaces
FROM Magisterka.dbo.Miejscowosci0402
```

Results | Messages

	ID_Place	Place1	WOJ	POW	JPT_KOD_JE	GMI	RODZ_GMI
1	1	Abramowice	6	63	663	1	1
2	2	Abramowice Kościelne	06	09	0609	05	2
3	3	Abramów	06	08	0608	02	2
4	4	Adamin	30	09	3009	09	2
5	5	Adamowa Góra	14	28	1428	04	2
6	6	Adamowizna	20	12	2012	08	2
7	7	Adamowo	28	04	2804	01	2
8	8	Adamów	30	10	3010	06	2
9	9	Adamówka	14	28	1428	01	1
10	10	Adolfowo	14	13	1413	04	2

Rys.22. Pełna tabela wymiaru z nazwami miejscowości, wraz z kodami TERYT ułatwiającymi nanoszenie punktów na mapie i identyfikację miejscowości, źródło: opracowanie własne

3.7. Kolumna z elementami wyposażenia

Drugą wymagającą szczególnej pracy była kolumna Features, zawierająca elementy wyposażenia każdego auta z osobna. Są to dane dość oczyszczone, pod tym kątem że nie ma tutaj żadnych niepożądanych informacji, takich, które wymagałyby dalszej obróbki. Problem jaki występuje z tą kolumną jest następujący – są to ciągi wypisanych po przecinku elementów wyposażenia jakie posiada dane auto. Nie da się przeprowadzić żadnej sensownej analizy na takich danych, ponieważ cechy te, nie są rozpatrywane

osobno. Oznacza to, że każde auto analizowane byłoby w takim przypadku w sposób nietransparentny, ponieważ informacja o tym, że auto posiada przypuśćmy ABS, ASR, ESP i przyciemniane szyby, byłaby znacząco inna od informacji że inne auto posiada ABS, ASR, ESP i klimatyzację, a przecież trzy na cztery elementy wyposażenia w obu autach są takie same i się pokrywają. Jednak kiedy wszystkie one są wymienione po przecinku jako jeden ciąg znaków, nie ma możliwości wychwycenia takich niuansów. Dlatego też potrzebny był rozdział tej kolumny według przecinków. W ten sposób powstało tyle kolumn, ile maksymalnie elementów wyposażenia posiadało jedno z aut, a więc 53. Rozdzielenie tych kolumn, jakie nastąpiło w programie PowerBI , było zresztą warunkiem wgrania tabeli z tymi danymi do SQL Servera, ponieważ ciągi znaków były po prostu za długie.

Features.1	Features.2	Features.3	Features.4	Features.5	Features.6	Features.7	Features.8	Features.9	Features.10	Features.11
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Bluetooth'	'Electrochromic	'AUX socket'	'Manual air cor	'MP3'	'Heated side mir	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Lane assistant'	'Twilight senso	'ESP(stabilizatic	'Isofix'	'On-board com	'Speed limiter'	'Rear side airba
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Twilight senso	'AUX socket'	'Automatic air	'Air curtains'	'Speed limiter'	'Front side airba	'Start-Stop syste
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Rain sensor'	'Rear parking s	'USB socket'	'Automatic air	'Air curtains'	'Speed limiter'	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Rain sensor'	'Rear parking s	'USB socket'	'Dual zone air	'MP3'	'Heated side mir	'Start-Stop syste
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Rain sensor'	'Rear parking s	'AUX socket'	'Rear view cam	'Air curtains'	'Speed limiter'	'Tinted windows
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'ASR (traction c	'Front parking	'Electric rear wi	'Isofix'	'On-board com	'Speed limiter'	'Rear side airba
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Bluetooth'	'Electric rear wi	'USB socket'	'Dual zone air	'MP3'	'Heated side mir	'Start-Stop syste
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'ASR (traction c	'Twilight senso	'Electrochromic	'AUX socket'	'Dual zone air	'MP3'	'Front side airba
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'ASR (traction c	'Rear parking s	'ESP(stabilizatic	'USB socket'	'Dual zone air	'MP3'	'Front side airba
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'Rain sensor'	'Electric rear wi	'Isofix'	'Air curtains'	'Heated front s	'Tinted windows	'Daytime runnin
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'ASR (traction c	'Twilight senso	'Electrochromic	'Electrically adj	'Isofix'	'On-board comp	'Airbag protecti
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Bluetooth'	'Front parking	'ESP(stabilizatic	'Rear view cam	'On-board com	'Heated front se	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Bluetooth'	'Front parking	'ESP(stabilizatic	'Rear view cam	'Air curtains'	'Heated side mir	'Tinted windows
['ABS'	'Electric front w	'Drivers airbag'	'Power steering	'ASR (traction c	'Rear parking s	'Electric rear wi	'Automatic air	'Air curtains'	'Heated front se	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alarm'	'Lane assistant'	'Blind spot sen	'ESP(stabilizatic	'Isofix'	'On-board com	'Speed limiter'	'Airbag protecti
['ABS'	'Electrically adj	'Passengers air	'Alarm'	'Bluetooth'	'Electrochromic	'SD socket'	'Rear view cam	'On-board com	'Heated front se	'Roof rails'
['ABS'	'Electrically adj	'Passengers air	'Alloy wheels'	'Twilight senso	'SD socket'	'Rear view cam	'On-board com	'Speed limiter'	'Airbag protectir	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alarm'	'Lane assistant'	'Blind spot sen	'ESP(stabilizatic	'USB socket'	'Automatic air	'MP3'	'Tinted windows
['ABS'	'Electrically adj	'Passengers air	'Alarm'	'Lane assistant'	'Blind spot sen	'ESP(stabilizatic	'Isofix'	'MP3'	'Front side airba	'Start-Stop syste

Rys.23. Wygląd kolumn z elementami wyposażenia po ich rozbiciu „split by delimiter", źródło: opracowanie własne

Kiedy już wgrano tabelę ofert sprzedaży z rozdzielonymi kolumnami, pierwszym krokiem było sprawdzenie ile unikatowych cech w ogóle w tych pięćdziesięciu trzech kolumnach występuje, poleceniem Select discount oraz Union all. Kod ten wskazał listę unikatowych wartości w tych kolumnach, jednak ponieważ ich liczba sięgała 70, w toku prac uznano, że należy wprowadzić cechy agregatowe i zbyt szczegółowe dane połączyć w jeden element wyposażenia, np. wyszczególnione cztery rodzaje klimatyzacji w jedną zbiorczą cechę. Ponadto wśród cech znalazły się przypadkowo także nazwy dwóch powiatów, zostały one również usunięte z tabeli ofert sprzedaży. W wyniku tych prac ustalono listę poszczególnych cech i w oparciu o nią, stworzono kolumny, każda

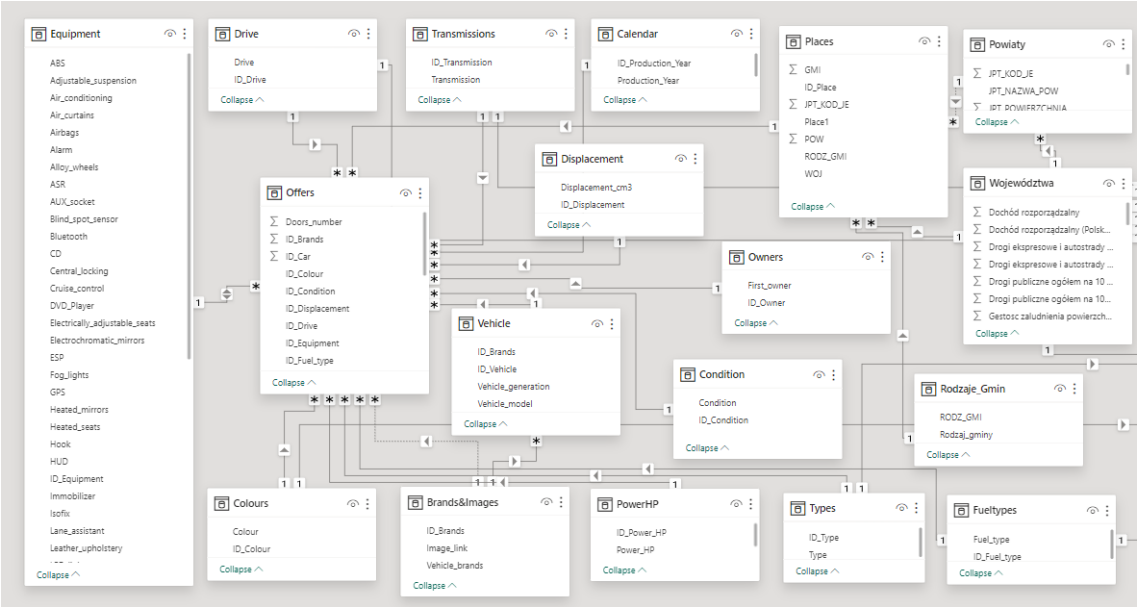
odpowiadająca danemu elementowi wyposażenia, w których to ‘Yes’ jest oznaką że dane auto posiada to wyposażenie, a NULL - że nie posiada. Kolejny kod sprawdza w każdej z 53 kolumn czy w jakiejkolwiek komórce w tym zakresie pojawia się wyposażenie ‘ABS’. Jeśli tak, w tabeli ofert sprzedaży, we wcześniej utworzonej nowej kolumnie o nazwie ABS, danemu rekordowi jest przypisywana wartość ‘Yes’. Jeśli w danym rekordzie wartość ‘ABS’ nie pojawia się w żadnej z 53 kolumn, w kolumnie ABS jest przypisywany NULL. Na podstawie tego kroku zostało przeprowadzone analogiczne działanie ze wszystkimi pozostałymi elementami wyposażenia aut z listy unikatowych wartości tychże. Następnie zostały usunięte kolumny od Features1 do Features53, zaś w bazie pozostały tylko nowe kolumny z nazwami poszczególnych elementów. Następnie aby wyeksportować te dane do tabeli wymiarów, każdemu autu został przypisany kod, taki sam jak kod ID auta oraz wszystkie elementy wyposażenia. W ten sposób powstała tabela Equipment.

3.8. Tworzenie tabeli wymiarów

Kolejnym krokiem w przygotowaniu tych danych do analizy, było stworzenie tabel wymiarów (ang. dimension table), które zawierałyby listę unikatowych wartości z danej kolumny, a każdej z nich nadany zostałby osobny kod ID. Taki proces został przeprowadzony dla pozostałych kolumn, poza tymi z unikatowymi wartościami, jak przebieg (mileage_km), czy cena (price), a także liczba drzwi, ponieważ jest ona i tak cyfrą od 3 do 7, tworzenie ID nie ma tu wielkiej celowości. Tak wyglądał przykładowy proces dla tabeli wymiarów pod nazwą Vehicles:

```
SELECT *
FROM Magisterka.dbo.Vehicle
JOIN Magisterka.dbo.Wersja0302 ON Magisterka.dbo.Vehicle.Vehicle_brand =
Magisterka.dbo.Wersja0302.Vehicle_brand
AND Magisterka.dbo.Vehicle.Vehicle_model = Magisterka.dbo.Wersja0302.Vehicle_model
AND Magisterka.dbo.Vehicle.Vehicle_generation =
Magisterka.dbo.Wersja0302.Vehicle_generation;
```

Podobny kod napisano dla pozostałych tabeli wymiarów. Pozostałe tabele stworzyły dimension tables dla fact table jaką jest tabela Offers. Poniżej można zobaczyć jak układają się relacje między tabelami w programie Power BI.



Rys.24. Relacyjny model danych stworzony w tej pracy, źródło: opracowanie własne

Tak przygotowane dane będą podlegały dalszej analizie w programie Power BI.

Rozdział IV. Kokpit managerski (dashboard) – wizualizacja danych w Power BI (data vis)

Kokpit managerski to narzędzie, służące do prezentacji danych kluczowych dla organizacji, a także właściwych wskaźników o danej działalności. Jest jednym z narzędzi systemów Business Intelligence, które przetwarzają dane w informacje, a informacje w wiedzę. Dzięki niemu w jednym miejscu ma się dostęp i możliwość monitorowania najważniejszych wskaźników wydajności (KPI) w firmie, co znacząco ułatwia ich analizę. Także w tej pracy, autor starał się stworzyć taki dashboard, który przypominałby te znane biznesowi, chociaż z oczywistymi korektami, ze względu na taką, a nie inną strukturę wykorzystanej po temu bazy danych.

4.1. Struktura

Rys.25. Pierwsza strona dashboardu managerskiego, źródło: opracowanie własne

Pierwsza strona dashboardu służy analizie struktury bazy, którą chcemy zobaczyć zarówno w wymiarze przestrzennym, to znaczy jak rozkłada się liczba aut w Polsce w podziale na województwa, oraz w wymiarze marek aut, czyli które marki aut są najczęściej oferowanymi w bazie. Co ważne, obie wizualizacje, tak tabela, jak i mapa, na siebie wzajemnie oddziałują, co oznacza, że można zobaczyć wymiar przestrzenny tylko

dla jednej marki, jak też i rozkład marek tylko w jednym województwie. Niektóre zależności są zauważalne bez wykorzystania specjalistycznych narzędzi analitycznych, przede wszystkim to, że najwięcej aut oferowanych jest w najbardziej ludnych, ale też najbogatszych regionach kraju, przede wszystkim na Mazowszu i na Śląsku, a najmniej w najmniej ludnych, jak Opolszczyzna, Lubuskie, czy Warmia i Mazury, oraz najbiedniejszych, jak Podlasie czy ziemia świętokrzyska.

Co do rozkładu marek, dominują zdecydowanie marki niemieckie, Volkswagen, Audi, Opel, BMW i Mercedes oraz amerykański Ford. Ale też jest to zmienne w zależności od regionu, w większości regionów, tak jak w skali ogólnopolskiej, liderem jest Volkswagen, ale w niektórych regionach, np. na Podlasiu, liderem jest BMW. Jedno jest niezmienne, w każdym województwie w rankingu z największą ilością ofert na pierwszych miejscach są niemieckie marki aut.

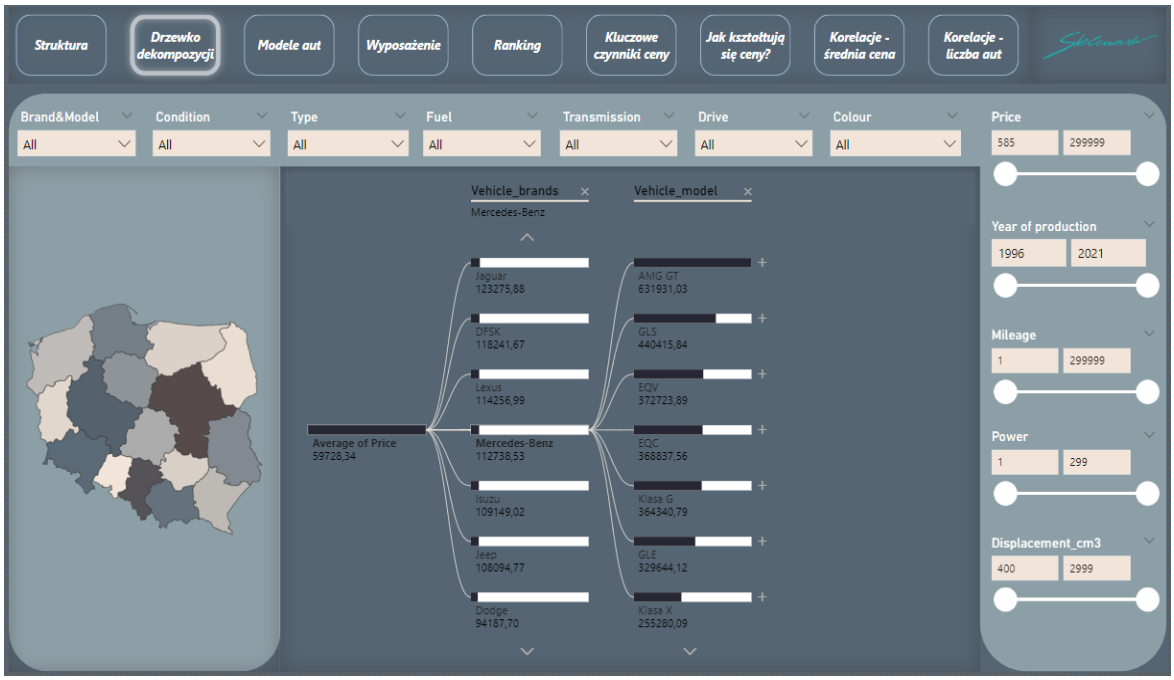
Pod tabelą z udziałem poszczególnych marek została umieszona druga tabela, wskazująca rozwarstwienie cenowe marek. Jest to istotne, gdyż chcemy się dowiedzieć w jakich przedziałach cenowych, jakie marki są najczęściej oferowanymi. Przyjęto że im ciemniejszy kolor, tym więcej aut w danym przedziale cenowym. Jest to o tyle ciekawe, że dostarcza wielu ciekawych wniosków, głównie dlatego że różni się to w zależności od regionu. Na przykład w województwie mazowieckim Audi najlepiej sprzedaje się w przedziale cenowym 125 tysięcy złotych do 250 tysięcy złotych, ale w województwie wielkopolskim już w przedziale 20 tysięcy złotych do 30 tysięcy złotych – najwyraźniej Audi nie jest tam zbyt cenioną marką, ponieważ w drogich autach króluje BMW, najbardziej w przedziale ponad 250 tysięcy złotych. W biednych województwach zaś, najczęściej królują oferty aut tanich i starych. Podobnych zależności można zauważyć więcej, po odpowiednio długim spędzeniu czasu z bazą.

Na stronie znalazł się także wykres pokazujący liczbę aut w zależności od roku produkcji – jest to jedyny wymiar czasowy w tej bazie, co nastręczało pewnych problemów związanych z tzw. storytellingiem, jak też wykorzystaniem potencjału drzemiącego w możliwościach Power BI w tym zakresie. Wykres także zmienia się w zależności od warunków z innych wizualizacji. Zauważyć można pewną tendencję wzrostową, tzn. liczba ofert rosła proporcjonalnie ze wzrostem roczników oferowanych aut – najwięcej ofert dotyczyło stosunkowo młodych aut. Podobnie rzecz się ma w niemal wszystkich województwach, ale ciekawe zjawisko wystąpiło na Śląsku, mianowicie zauważalnie

wzrosła tam liczba aut oferowanych z najnowszego ówcześnie rocznika, 2021, ponad trzykrotnie więcej niż ofert z rocznikiem 2019. W innych województwach taki wzrost ofert nie nastąpił, raczej był zgodny z dość liniowym trendem.

Co więcej, jak można zauważyć w całości kompozycji strony, nacisk został postawiony bardziej na samodzielną eksplorację danych przez użytkownika dashboardu niż na twarde analizy. Wynika to z założenia o daniu wolności w poszukiwaniu użytkownikowi na początku, tak aby zaznajomił się z bazą sam, a dopiero następnie podawać coraz więcej analiz. Wynika to po części z natury bazy, a konkretniej jej powstania. Zwebscrappowano oferty sprzedaży aut z serwisu internetowego w Polsce w maju 2021. Z tego powodu, baza przypomina nieco tenże serwis i znalazło to odzwierciedlenie w wyglądzie dashboardu. W miejscach, w których użytkownikowi pozostawiono swobodę w eksploracji, dodano także narzędzia, aby mógł sam manipulować wynikami, tak by pokazywały to co on chce. Stąd możliwość wyboru marki i modelu, stanu auta, rodzaju auta, rodzaju paliwa, rodzaju skrzyni biegów, rodzaju napędu i koloru, a także przedziału cenowego, roku produkcji, przebiegu, mocy w koniach mechanicznych i pojemności w centymetrach sześciennych.

4.2. Drzewko dekompozycji



Rys.26. Druga strona dashboardu managerskiego, źródło: opracowanie własne

Podobnie do pierwszej strony wygląda druga, czyli drzewko dekompozycyjne. Jest to analogiczne narzędzie do poprzedniego i także dające swobodę eksploracji danych. Jego zaletą jest wizualizacja przeprowadzana przez sztuczną inteligencję (AI), co umożliwia znalezienie następnego wymiaru, na podstawie określonych kryteriów. Jest to „bardzo przydatne w przypadku eksploracji ad hoc i przeprowadzania analizy głównej przyczyny.” – jak można przeczytać na oficjalnej stronie Microsoftu [25].

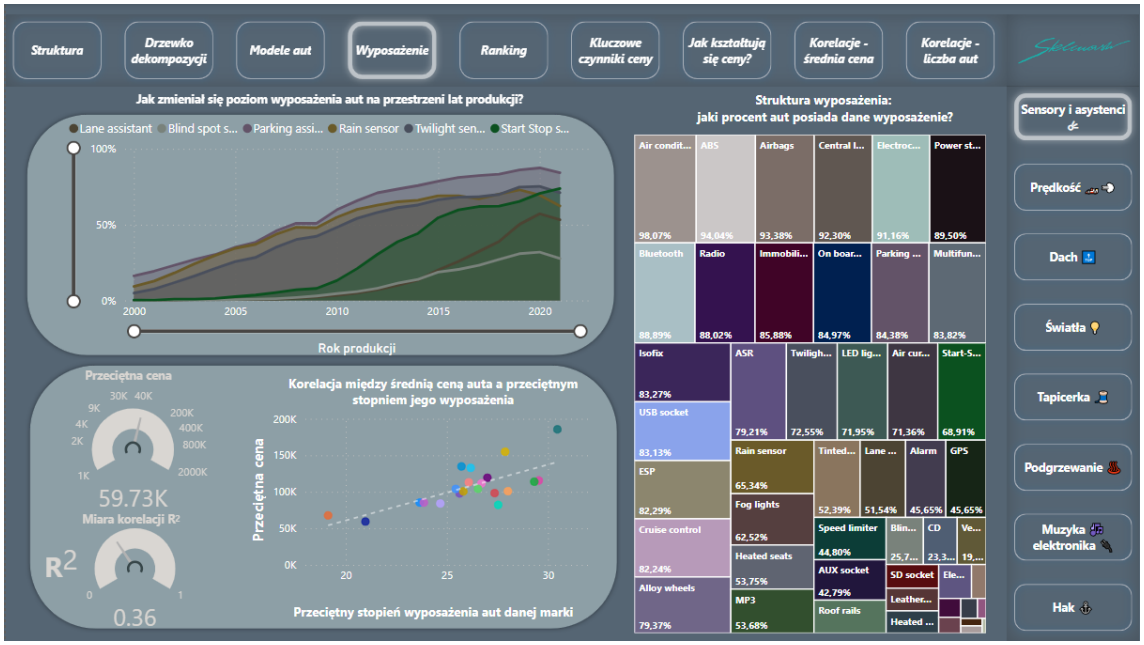
4.3. Modele aut – deska rozdzielcza

Rys.27. Trzecia strona dashboardu managerskiego, źródło: opracowanie własne

Na trzeciej stronie dashboardu znajduje się autorski pomysł połączenia koncepcji z tematyką. Podobnie jak w poprzednich przypadkach tak i tu użytkownik ma swobodę ruchów. Głównym jej celem jest analiza struktury, ale już nie na poziomie przestrzennym, czy poziomie marek aut, a bardziej modeli. Umożliwia sprawdzenie średnich cen i przebiegów poszczególnych marek i modeli aut, średnią moc silnika, średnią pojemność baku oraz średni rok produkcji (Tu zastosowano jednak ograniczenie, aby był to średni rok produkcji po roku 1995. Wynika to z występowania aut o roku produkcji z lat 60., 70., czy 80. ubiegłego wieku. Zakłamuje to niejako wyniki, bo są pojedyncze przypadki, ale jednak brane pod uwagę przy liczeniu standardowej średniej), a także, na mniejszych wykresach, strukturę:

- miejsca gdzie najczęściej są one wystawiane, czy jest to gmina miejska, czy wiejska,
- typu pojazdów, w obrębie jednej marki czy modelu, czy są to sedany, kombi, czy może jeszcze inne typy,
- stanu pojazdów, czy są to pojazdy nowe, czy używane,
- rodzaju skrzyni biegów, czy częściej są to auta z automatyczną, czy manualną skrzynią biegów,
- typu napędu, czy najczęściej są to auta z napędem na przednie koła, czy na tyle, czy 4x4 w różnych konfiguracjach,
- typu paliwa, czy są to pojazdy dieslowe, czy na benzynę, a może elektryczne.

4.4. Wyposażenie



Rys.28. Czwarta strona dashboardu managerskiego, źródło: opracowanie własne

Na czwartej stronie znajdują się pierwsze analizy, tu analiza wyposażenia jakie auta miały zadeklarowane w swoich ofertach. Większość poszczególnych elementów wyposażenia zostało pogrupowane i nazwane zbiorczo:

- sensory i asystenci, czyli asystent pasa ruchu, system informacji o martwym polu, asystent parkowania, sensor deszczu, sensor zmierzchu i system Start-Stop,
- prędkość, czyli tempomat i ogranicznik prędkości,
- dach, czyli szyberdach, dach panoramiczny i relingi,
- światła, czyli światła ksenonowe, LEDowe i przeciwmgielne,
- tapicerka, skórzana bądź welurowa,
- podgrzewanie, siedzeń oraz szyb,
- muzyka i elektronika, Bluetooth, CD, DVD, MP3, HUD, SD i USB
- hak.

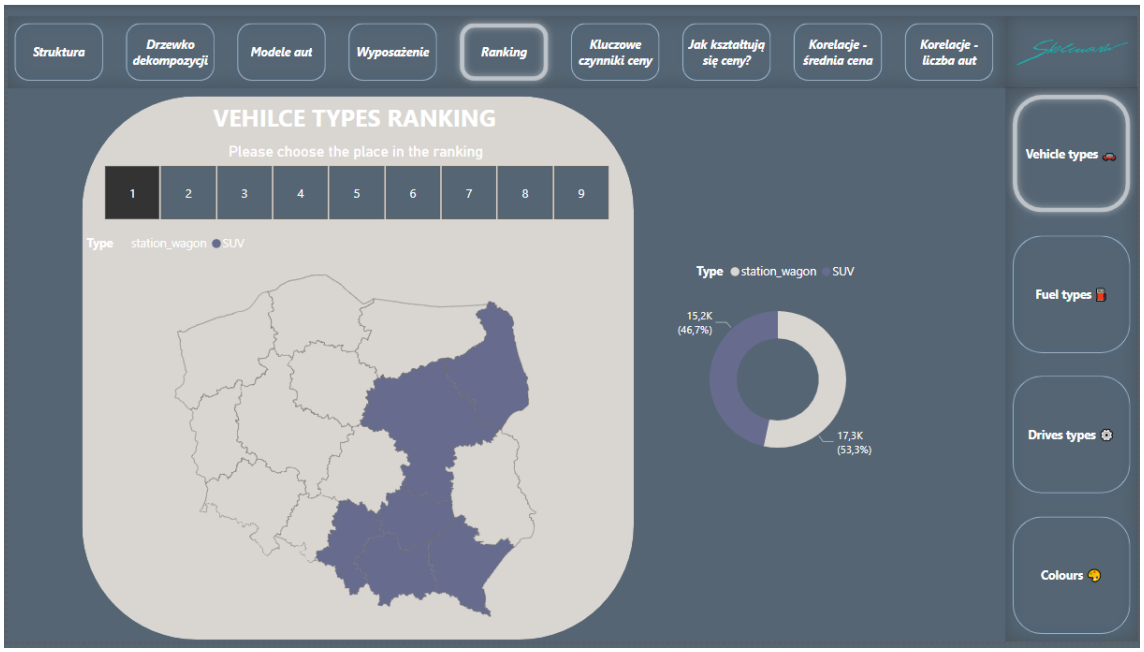
Grupy te zostały stworzone, aby obrazowały na wykresach zmienność wyposażenia aut na przestrzeni lat produkcji. Chodzi o to, aby zobaczyć, które elementy zyskują na popularności, a które wprost odwrotnie, i w jakim tempie to się odbywa. O ile np. wzrost popularności wszelkiego rodzaju sensorów i elektronicznych asystentów kierowcy nie dziwi, podobnie jak to, że schyłkowe z punktu widzenia 2024 roku technologie, jak DVD, CD, czy SD, są coraz rzadziej montowane w nowych autach, o tyle jednak w pewnym sensie zaskakujący może być fakt coraz mniejszej popularności montowania haka, czy też tego, że większą popularnością cieszą się dachy panoramiczne, kosztem klasycznych szyberdachów.

Ważnym aspektem jest także określenie jaki procent aut z bazy posiada dany rodzaj wyposażenia i temu służy wykres po lewej stronie.

W dolnym prawym rogu zaś, została przedstawiona korelacja między średnim stopniem wyposażenia aut danej marki a jego średnią ceną. Miara korelacji R^2 równa 0,36 pokazuje, że jest to korelacja niezbyt silna, acz widoczna, co sugeruje, iż po prostu nie jest to jedyny czynnik wpływający na średnią cenę auta. Aby miara była możliwie najbardziej dopasowana, z wizualizacji usunięto auta droższe niż 200 tysięcy złotych (oczywiste jest to, że na cenę nie wpływają najbardziej poszczególne elementy wyposażenia, a ogólna marka premium i jej konsekwencje), a także auta z rokiem produkcji poniżej 2000 roku, marki aut, których liczebność w bazie jest mniejsza od 100 (problem z reprezentatywnością) oraz przebieg mniejszy niż 25 tysięcy km (kiedy

przebieg jest duży, to to jest jedną z determinant ceny, nie można porównywać aut o małym i dużym przebiegu i na tej podstawie wnioskować o korelacji między ceną a średnim stopniem wyposażenia – oczywistym jest to, że nie będzie to miarodajne). Wśród tych marek aut, które powyższe warunki spełniają, najtańsze, ale też średnio najsłabiej wyposażone są Fiat i Dacia, z kolei wśród najlepiej wyposażonych, ale też najdroższych są Lexusy i Jeepy. Warto jednak zauważyć że auta o podobnym średnim stopniu wyposażenia co one, ale o niższej średniej cenie, to Hondy, Mazdy i Nissany.

4.5. Rankingi



Rys.29. Piąta strona dashboardu managerskiego, źródło: opracowanie własne

Na piątej stronie dashboardu znalazło się miejsce na podsumowanie popularności pewnych rozwiązań technicznych na mapach. Zostały stworzone rankingi, które sklasyfikowały jak często pojawiały te rozwiązania w zależności od polskich województw. Można wybrać miejsce rankingowe oraz aspekt, który chce się sprawdzić – typ pojazdów, rodzaj paliwa, rodzaj napędu oraz kolor aut.

W pierwszym rankingu, dotyczącym typów pojazdów można wyciągnąć wnioski, iż we wschodniej Polsce najbardziej popularne są SUVy, a na ścianie zachodniej – auta kombi. Odwrotnie sytuacja wygląda na drugim miejscu. Na trzeciej lokacie na ścianie zachodniej i południowej umiejscowiły się auta kompaktowe, a na ścianie wschodniej i północnej sedany. Na dalszych miejscach w różnych konfiguracjach pojawiły się też auta miejskie,

minivany i auta typu coupe. Na ostatnim miejscu we wszystkich polskich województwach w rankingu popularności typów aut znalazły się kabriolety, co zważywszy na polski klimat nie może zaskakiwać.

W drugim rankingu, dotyczącym typów paliwa, jakimi napędzane są auta, można zauważyć, iż silniki Diesla są najpopularniejsze w stosunkowo bogatszych regionach, jak Pomorze, Wielkopolska, Małopolska, Dolny Śląsk, ale też Podkarpacie i Kielecczyzna, zaś silniki benzynowe w pozostałych regionach. Na kolejnych miejscach niemal we wszystkich województwach były: benzyna + gaz LPG, auta hybrydowe, auta elektryczne, benzyna + gaz CNG oraz na ostatnim miejscu w każdym województwie (poza opolskim, w którym taka oferta nie wystąpiła) auta na wodór.

W trzecim rankingu, dotyczącym rodzajów napędu aut, najpopularniejszym typem napędu w każdym województwie jest napęd na przednie koła. Na kolejnych lokatach jest już ciekawszy rozkład geograficzny, mianowicie napęd na tylne koła, występuje głównie w centralnej Polsce, od Bugu do Odry, zaś różne rodzaje napędu 4x4 są popularniejsze na północy i na południu. Dowodzi to tezie, iż mimo wyższej ceny, auta 4x4 są najczęściej oferowane nie w województwach bogatych, a tych które mają warunki naturalne, które umożliwiają korzystanie z tegoż napędu w sposób zgodny z założeniami.

W czwartym rankingu, dotyczącym kolorów aut, pierwsze miejsce bezapelacyjnie wygrywa kolor czarny. Jest on najpopularniejszym kolorem w każdym województwie. Na kolejnych miejscach są kolory szary, siwy i biały. Na piątym miejscu w każdym województwie znajduje się kolor niebieski. Za nim inne kolory, kolor czerwony, brązowy, zielony, burgundowy, beżowy i złoty. Najmniej popularne w każdym województwie są kolory fioletowy i żółty. Jest to cenna wskazówka, między innymi w doborze przyszłego auta, jeśli właściciel będzie miał plan je sprzedać dalej po jakimś czasie. Należy wziąć wtedy pod uwagę fakt, iż niektóre kolory zwyczajnie się ludziom nie podobają, jak również to, że jest to czasem zróżnicowane przestrzennie w skali kraju.

4.6. Kluczowe czynniki wpływające na cenę

Rys.30. Szósta strona dashboardu managerskiego, key infuencers – increse, źródło: opracowanie własne

Na szóstej stronie dashboardu znajdują się kluczowe czynniki wpływające na cenę (ang. Key Influencers Factors). Jest to wizualizacja wykorzystująca techniki sztucznej inteligencji (ang. AI – artificial intelligence), która pomaga zidentyfikować czynniki mające największy wpływ na badane zjawisko lub wynik. Ta funkcja umożliwia szybkie zrozumienie, które zmienne lub czynniki mają największy wpływ na wybraną zmienną, co może być niezwykle cenne przy podejmowaniu decyzji biznesowych. W tym przypadku bada się wpływ różnych czynników na średnią cenę auta. Na wizualizację zostały nałożone dwa filtry, które mają na celu zwiększenie rzetelności badania, po pierwsze cena ma być niższa niż 200 tysięcy złotych (wyłączamy drogie auta, które zachowują się jak outliery w tym modelu), a po drugie rok produkcji ma być wyższy od 2000 (inaczej starsze, zabytkowe auta jakie także znajdują się w bazie, powodowałyby anomalie, jak między innymi ta, że cena rośnie, wraz ze wzrostem wieku auta, co oczywiste jest nieprawdą w przypadku nowych, czysto użytkowych aut).

Jak można odczytać z danych, średnia cena rośnie ona o:

- 71,59 tysięcy złotych, jeśli auto jest nowe, a nie używane,
- 70,14 tysięcy złotych, jeśli auto jest elektryczne, a także o 57,73 tysiące złotych, jeśli auto jest hybrydowe,
- 49,6 tysięcy złotych, gdy skrzynia biegów jest automatyczna,
- 47,87 tysięcy złotych, gdy moc silnika w koniach mechanicznych jest większa od 177,
- 39,15 tysięcy złotych, gdy napęd auta to stały napęd 4x4,
- 33,9 tysięcy złotych, gdy typem pojazdu jest SUV, 21,21 tysięcy złotych, gdy typem pojazdu jest coupe oraz 12,39 tysięcy złotych, gdy typem pojazdu jest sedan,
- 24,34 tysięcy złotych, gdy moc silnika znajduje się w przedziale 147-154 KM,
- 21,51 tysięcy złotych, gdy kolor auta to kolor biały,
- 21,34 tysięcy złotych, gdy pojemność baku w centymetrach sześciennych jest większa od 1910 cm^3,
- 17,42 tysięcy złotych, gdy rok produkcji jest młodszy od przeciętnego o 5,42 lat,
- 10,44 tysięcy złotych, gdy przebieg jest mniejszy od przeciętnego o 93 tysiące kilometrów,
- 10,34 tysięcy złotych, gdy gmina jest gminą miejską.

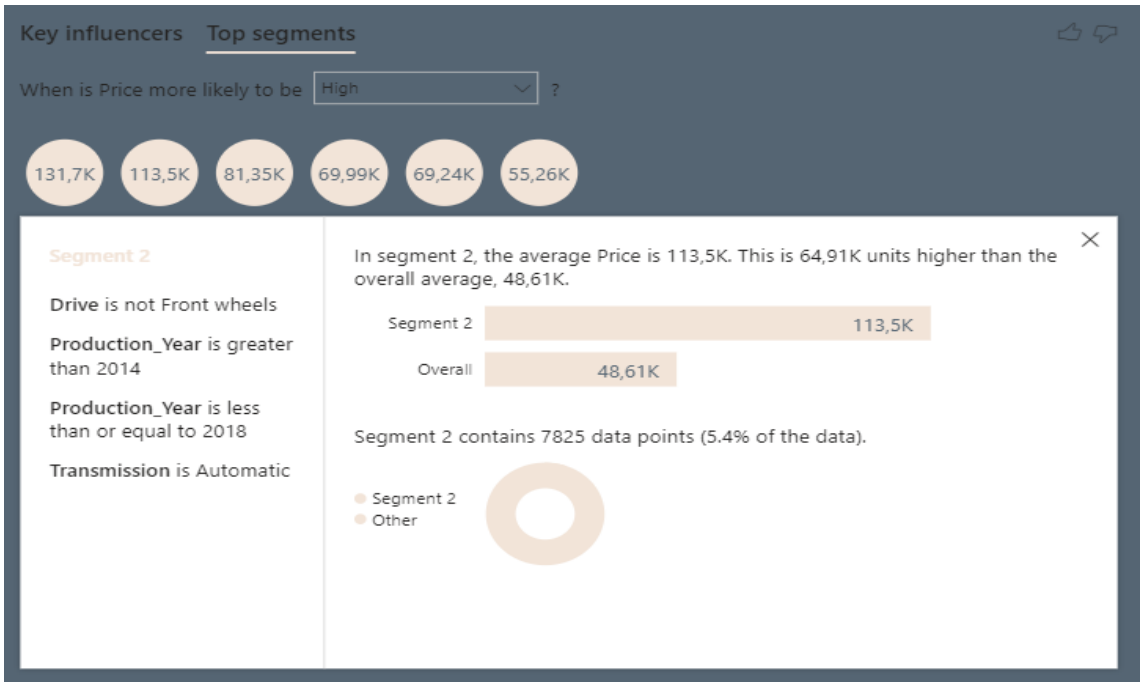
Rys.31. Szósta strona dashboardu managerskiego, key influencers - decrease, źródło: opracowanie własne Z kolei średnia cena auta spada o:

- 71,59 tysięcy złotych, gdy auto jest używane, a nie nowe,
- 49,6 tysięcy złotych, gdy auto posiada manualną skrzynię biegów,
- 37,85 tysięcy złotych, gdy pojemność baku znajduje się w przedziale między 1800 a 1910 cm^3, 17,43 tysięcy złotych, gdy pojemność baku znajduje się w przedziale między 1500 a 1800 cm^3 oraz 12,11 tysięcy złotych, gdy pojemność baku jest mniejsza niż 1461 cm^3,
- 34,23 tysięcy złotych, gdy moc silnika jest mniejsza bądź równa 94 KM oraz o 21,35 tys. zł, gdy moc silnika jest w przedziale 94-147 KM,
- 31,91 tys. zł, gdy typem pojazdu jest auto małe, 26,77 tys. zł, gdy typem pojazdu jest auto miejskie, 12,26 tys. zł, gdy typ pojazdu to minivan oraz 11,1 tys. zł, gdy typem pojazdu jest auto kompaktowe,
- 29,21 tys. zł, gdy napęd auta jest na przednie koła,
- 23,85 tys. zł, gdy rodzajem paliwa jest kombinacja benzyna + gaz LPG,
- 19,06 tys. zł, gdy kolorem auta jest kolor srebrny,
- 17,42 tysięcy złotych, gdy rok produkcji jest starszy od przeciętnego o 5,42 lat,

- 10,44 tysięcy złotych, gdy przebieg jest większy od przeciętnego o 93 tysiące kilometrów.

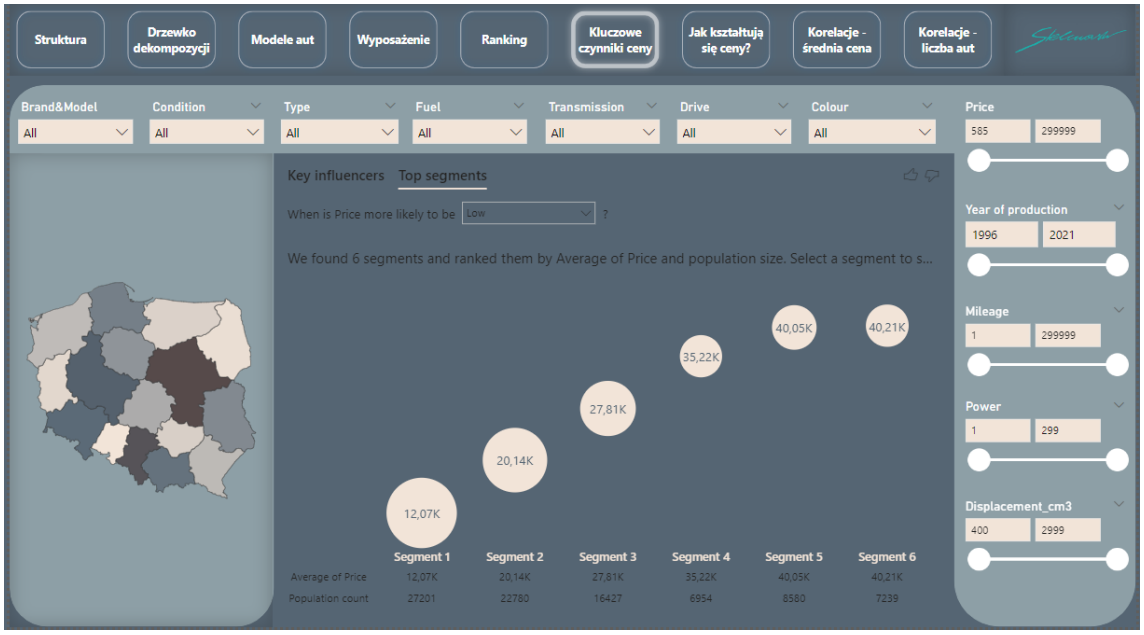
Rys.32. Szósta strona dashboardu managerskiego, top segments, źródło: opracowanie własne

Ponadto ta wizualizacja wspomagana przez AI, stwarza możliwość pewnego rodzaju analizy skupień – wyznacza bowiem tzw. główne segmenty, które grupuje według rozkładu badanej cechy. W tym przypadku kiedy cena będzie prawdopodobnie bardziej wysoka, znalezione zostało 6 głównych segmentów, jak na obrazie powyżej. W każdy segment można wejść i zobaczyć na jakich podstawach został on wydzielony z ogółu, np. na poniższym obrazie, można dostrzec, iż segment drugi składa się aut, których napęd nie jest napędem na przednie koła, rok produkcji znajduje się w przedziale od 2014 do 2018 roku, zaś skrzynia biegów jest automatyczna. Na bazie takich założeń, wyodrębniona została ta grupa, która charakteryzuje się średnią ceną wewnątrz niej o 64,91 tys. zł wyższą od średniej ogółu. Grupa ta stanowi także 5,4% całości liczebności.



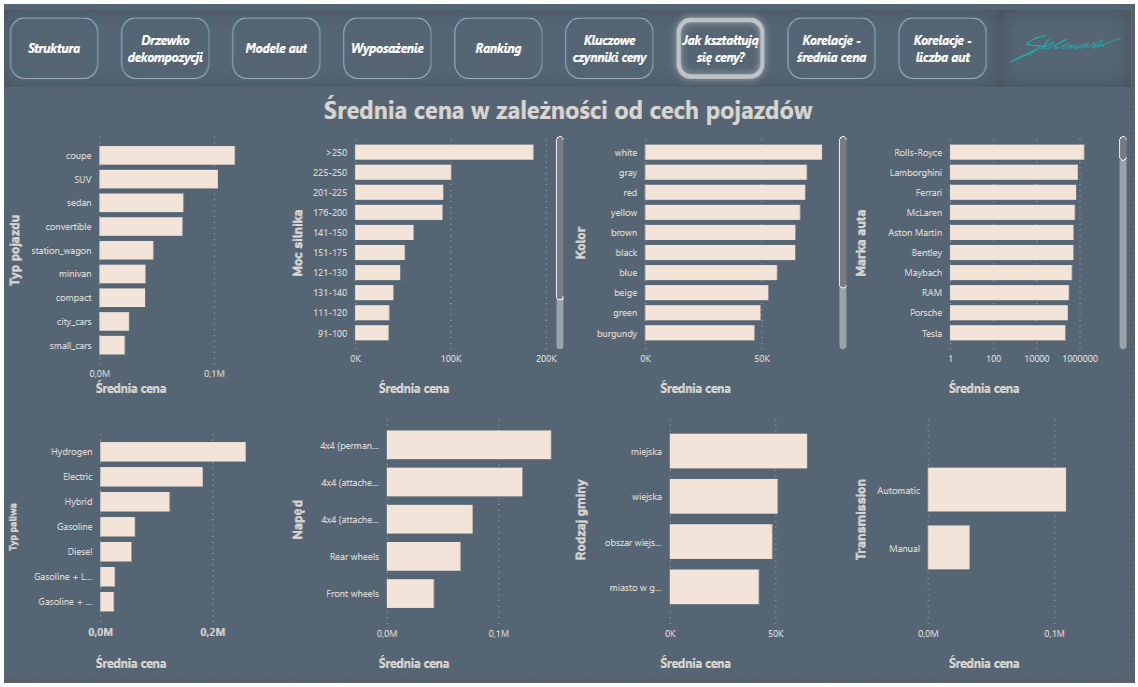
Rys.33. Szósta strona dashboardu managerskiego, top segments – szczegóły, źródło: opracowanie własne

Z kolei kiedy cena będzie prawdopodobnie bardziej niska, segmenty te wyglądają już inaczej, jak na obrazie poniżej. Również w tym przypadku jest możliwe zobaczenie szczegółów każdego z wyodrębnionych segmentów.



Rys.34. Szósta strona dashboardu managerskiego, top segments - ranking, źródło: opracowanie własne

4.7. Jak kształtują się ceny?



Rys.35. Siódma strona dashboardu managerskiego, źródło: opracowanie własne

Na siódmej stronie dashboardu została umiejscowiona dalsza analiza zależności wysokości ceny od różnych czynników. Jak można dostrzec, najdroższymi autami w bazie, są SUVy i coupe, a najtańszymi auta małe i miejskie. Najwięcej kosztują także auta o dużej mocy silnika, przekraczającej 250 koni mechanicznych, a najmniej auta o słabej mocy, poniżej 80 KM. Ciekawą zależność można odczytać w kwestii koloru, mianowicie, najdroższymi autami są auta białe, szare, czerwone i żółte, a najtańszymi srebrne i fioletowe. Kiedy porówna się to z rozkładem przestrzennym popularności kolorów, można zauważyć, że najdroższymi autami są zarówno auta o kolorach dość popularnych, jak biały czy siwy, ale też tych niekoniecznie popularnych, jak kolor żółty. Podobnie rzecz się ma w przypadku najtańszych aut – są nimi zarówno bardzo popularne auta o kolorze srebrnym, jak i te o kolorze mało popularnym jak fioletowy, czy złoty. Trudno zatem o wskazanie jednoznacznie jakiejś zależności.

W przypadku zależności ceny od marki auta przyjęto założenia mające na celu zwiększenie rzetelności badań, a więc wykluczenie outlierów, takich jak auta bardzo drogie, o wartości powyżej 200 tys. zł, czy auta stare, o roczniku produkcji poniżej 2000 roku oraz wykluczenie marek aut mało liczebnych, poniżej 100 wystąpień w bazie. Przy

takich założeniach, najdroższymi markami aut okazały się Maserati, Jaguar i Lexus, zaś najtańszymi Lancia, Saab i Daihatsu.

W kwestii rodzaju paliwa zdecydowanie najdroższe są auta wodorowe i elektryczne, zaś najtańsze na benzynę i wybrany rodzaj gazu, LPG lub CNG. Rozpatrując zaś rodzaj napędu, różne typy napędu 4x4 są zdecydowanie droższe od tych aut, które posiadają napęd na koła przednie lub tylne. Cena różni się także w zależności od rodzaju gminy, w którym była wystawiona dana oferta – jeśli była to gmina miejska, to cena była znacznie wyższa od cen w pozostałych rodzajach gmin, średnio prawie 15 tysięcy zł droższa niż w gminach wiejskich i przeszło 20 tysięcy droższa niż w miastach w gminach miejsko-wiejskich, w których oferowane auta były średnio najtańsze. Co do skrzyni biegów zaś, auta z automatyczną są średnio trzy razy droższe od tych z manualną skrzynią biegów.

4.8. Korelacje – średnia cena aut

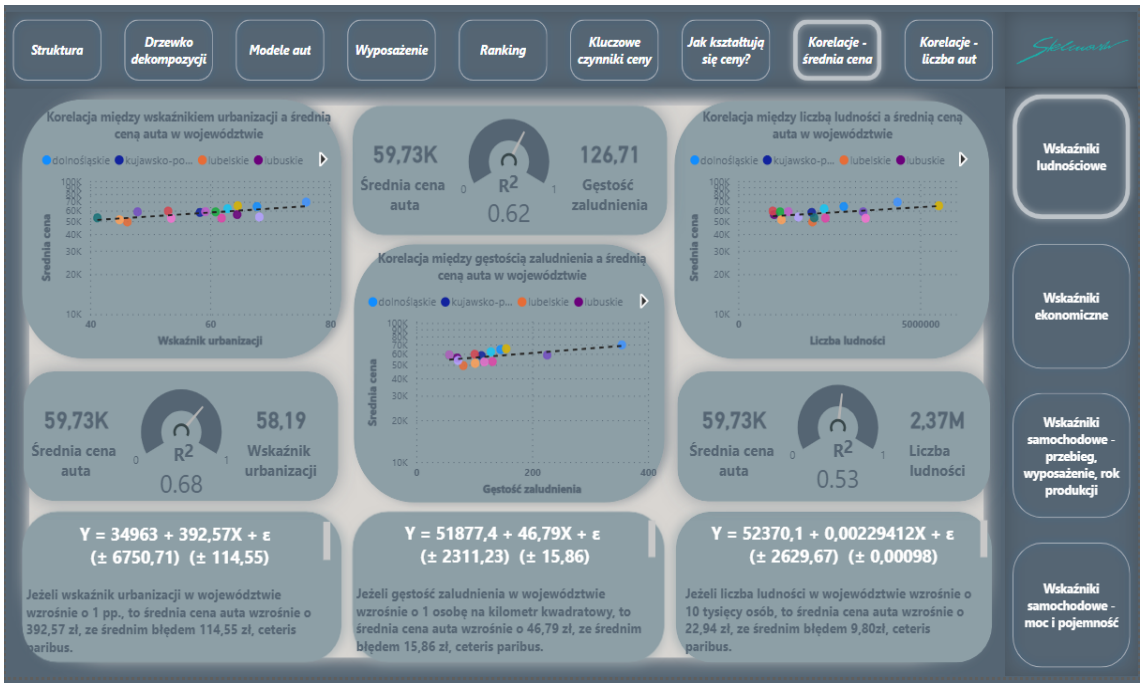
Na przedostatniej stronie znajdują się korelacje jakie zbadano między średnią ceną aut a wskaźnikami:

- ludnościowymi:
    - wskaźnikiem urbanizacji,
    - gęstością zaludnienia,
    - liczbą ludności,
- ekonomicznymi:
    - stopą bezrobocia,
    - średnim wynagrodzeniem,
    - dochodem rozporządzalnym,
- samochodowymi, tj. dotyczącymi bezpośrednio aut z bazy danych:
    - przebiegiem,
    - rokiem produkcji,
    - przeciętnym stopniem dodatkowego zadeklarowanego wyposażenia,
    - mocą silnika,

- pojemnością baku.

Ostatnia grupa została z przyczyn zachowania przejrzystości wizualnej dashboardu rozbita na dwie zakładki, pierwsze trzy cechy znalazły się w pierwszej z nich (w tym korelacja między średnią ceną auta a przeciętnym stopniem dodatkowego zadeklarowanego wyposażenia, co zostało już opisane na jednej z poprzednich kart, jednak dla zachowania spójności kompozycyjnej zostało umieszczone także na tej stronie), moc i pojemność zostały umieszczone w drugiej. Ponadto dołączono do korelacji średniej mocy silnika ze średnią ceną oraz średniej pojemności baku ze średnią ceną, także korelację niezwiązaną ze średnią ceną, mianowicie korelację między pojemnością baku a mocą silnika.

Problemy z badaniami przestrzennymi, takimi jak wykorzystane w tej pracy, jest taki, że bada się intensywność danego zjawiska. Ciężko jednak jest wyodrębnić wpływ jednego tylko czynnika na zmienną objaśnianą, ponieważ wszystkie niejako są skorelowane z jedną zmienną – liczbą ludności. Ona nie w pełni, ale w istotnej części, determinuje większość z pozostałych zmiennych objaśniających. Z tego też powodu nie został stworzony jeden model, który zawierałby wszystkie zmienne objaśniające i wyjaśniał tym samym wpływ wszystkich z nich naraz. Próba zbudowania takiego modelu została podjęta przez autora, niemniej udowodniła tylko wcześniejsze przypuszczenia – następował problem z weryfikacją tego modelu, gdyż zmienne objaśniające wykazywały zdecydowanie zbyt wysoką współliniowość. Objawiała się ona w tym, że pozytywnie wyszedł test na globalną istotność modelu, a więc sprawdzenie, czy chociaż jedna zmienna istotnie statystycznie wpływa na poziom zmiennej objaśniającej, przy jednoczesnym fiasku testowania istotności pojedynczych zmiennych objaśniających – wszystkie okazały się nieistotne. Jest to charakterystyczne właśnie dla modeli, które wykorzystały wysoko skorelowane ze sobą zmienne objaśniające. Wobec tego korelacje pozostawiono w formie takiej jaką można dostrzec w dashboardzie, czyli badanie wpływu pojedynczych zmiennych, bez używania do tego całego modelu ekonometrycznego.



Rys.36. Ósma strona dashboardu managerskiego, korelacje – wskaźniki ludnościowe, źródło: opracowanie własne

R^2 - miara jakości dopasowania modelu do danych, inaczej współczynnik determinacji, przedstawia się następująco:

- w 68% zmienność średniej ceny została wyjaśniona przez zmienność wskaźnika urbanizacji,
- w 62% zmienność średniej ceny została wyjaśniona przez zmienność gęstości zaludnienia,
- w 53% zmienność średniej ceny została wyjaśniona przez zmienność liczby ludności,
- w 49% zmienność średniej ceny została wyjaśniona przez zmienność stopy bezrobocia, z nachyleniem ujemnym,
- w 66% zmienność średniej ceny została wyjaśniona przez zmienność średniego wynagrodzenia,
- w 53% zmienność średniej ceny została wyjaśniona przez zmienność dochodu rozporządzalnego,

- w 36% zmienność średniej ceny została wyjaśniona przez zmienność przeciętnego stopnia dodatkowego wyposażenia,
- w 33% zmienność średniej ceny została wyjaśniona przez zmienność średniego przebiegu auta, z nachyleniem ujemnym,
- w 85% zmienność średniej ceny została wyjaśniona przez zmienność średniego roku produkcji (powyżej 1995 roku),
- w 73% zmienność średniej ceny została wyjaśniona przez zmienność średniej mocy silnika,
- w 57% zmienność średniej ceny została wyjaśniona przez zmienność średniej pojemności baku.

Ponadto w 77% zmienność średniej mocy silnika została wyjaśniona przez zmienność średniej pojemności baku.

Oznacza to, że np. w przypadku równania modelu korelacji wskaźnika urbanizacji i średniej ceny auta, które jest następujące: $Y = 34963 + 392{,}57X + \varepsilon$, $(\pm 114{,}55)$, interpretuje się to jak poniżej:

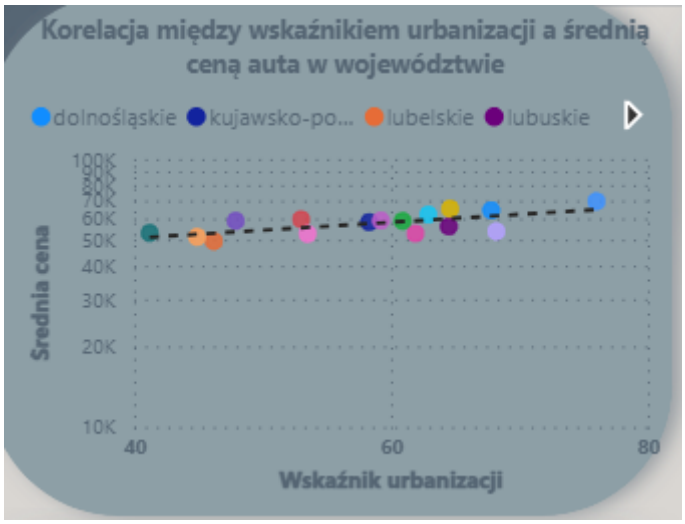
„Jeżeli wskaźnik urbanizacji w województwie wzrośnie o 1 pp., to średnia cena auta wzrośnie o 392,57 zł, ze średnim błędem 114,55 zł, ceteris paribus."

Analogicznie rzecz się ma w przypadku ujemnej korelacji, np. stopy bezrobocia ze średnią ceną auta. Równanie wygląda następująco: $Y = 66779{,}7 - 1374{,}06X + \varepsilon$, $(\pm 650{,}59)$. Interpretacja analogiczna jak w poprzednim przykładzie, ale z racji ujemnego nachylenia krzywej, oznacza to spadek, a nie wzrost:

„Jeżeli stopa bezrobocia w województwie wzrośnie o 1pp., to średnia cena auta spadnie o 1374,06 zł, ze średnim błędem 650,59 zł, ceteris paribus."

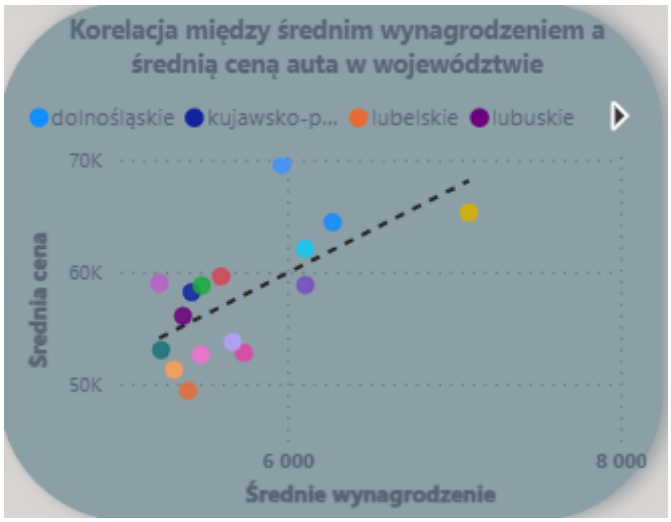
Można zatem zauważyć, które z korelacji są najsilniejsze:

- wśród wskaźników ludnościowych zdecydowanie najsilniejszy wpływ na średnią cenę przejawia wskaźnik urbanizacji – poszlaką ku temu jest fakt, iż, jak wiemy z poprzedniej strony dashboardu, najdroższe auta są oferowane właśnie w miastach; można przypuszczać zatem, że w tych województwach, w których miast jest najwięcej, będą także najdroższe auta,



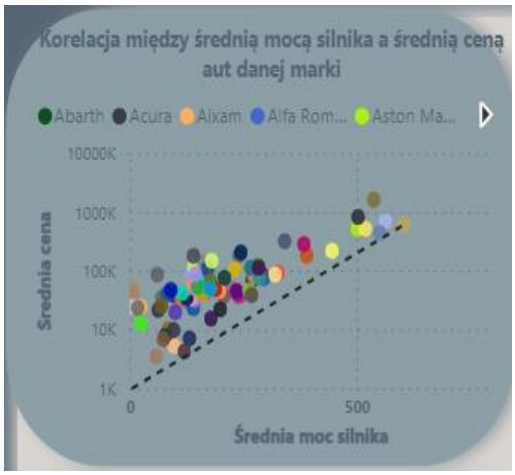
Rys.37. Korelacja między wskaźnikiem urbanizacji a średnią ceną auta w województwie, źródło: opracowanie własne

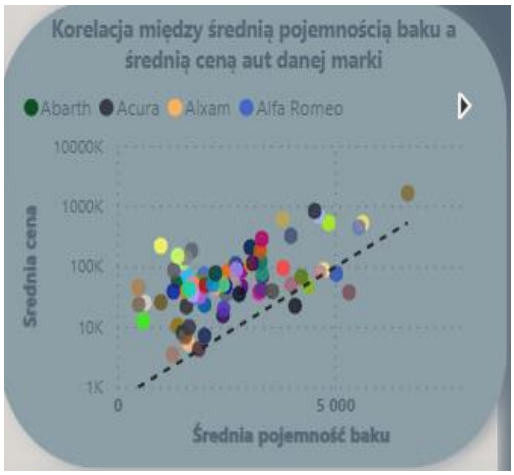
- wśród wskaźników ekonomicznych wpływającym w największym stopniu czynnikiem jest średnie wynagrodzenie – jest to dość logiczne, ponieważ im bogatszy region, tym na więcej stać jego mieszkańców,



Rys.38. Korelacja między średnim wynagrodzeniem a średnią ceną auta w województwie, źródło: opracowanie własne

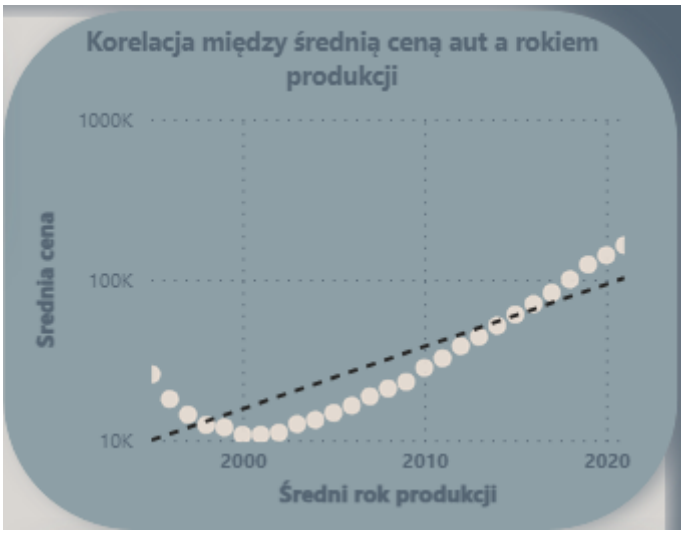
- wśród wskaźników samochodowych widać zależność między średnią ceną auta a średnią mocą silnika i pojemnością baku – jest to logiczne, gdyż im większe te liczby są, tym większe jest samo auto, jak również lepsze wyniki, jakie osiąga, co przekłada się na cenę,





Rys.38 i rys.39. Korelacje między średnią mocą silnika oraz średnią pojemnością baku a średnią ceną auta w województwie, źródło: opracowanie własne

- najciekawszą i zdecydowanie najbardziej widoczną korelacją jest jednak korelacja średniej ceny auta z rokiem produkcji; wzięto pod uwagę jedynie roczniki powyżej 1995 roku, aby zabytkowe auta sprzed lat nie powodowały występowania anomalii, że im starsze auta, tym są droższe; jak już było to wyjaśniane, nie sprawdza się to w przypadku aut czysto użytkowych.



Rys.40. Korelacja między średnim wiekiem auta a średnią ceną auta w województwie, źródło: opracowanie własne

4.9. Korelacje – liczba aut na 10 tysięcy mieszkańców

Na ostatniej stronie znajdują się korelacje jakie zbadano między liczbą aut na 10 tysięcy mieszkańców a wskaźnikami:

- ludnościowymi:
    - wskaźnikiem urbanizacji,
    - gęstością zaludnienia,
    - liczbą ludności,
- ekonomicznymi:
    - stopą bezrobocia,
    - średnim wynagrodzeniem,
    - dochodem rozporządzalnym,
- drogowymi:
    - długością dróg publicznych ogółem na 10 tysięcy ludności w województwie,
    - średnim przebiegiem aut z bazy,
    - liczbą ofiar śmiertelnych na 10 tysięcy ludności w województwie,
- kolejowymi:
    - liczbą stacji kolejowych o wymianie pasażerskiej między 100 a 1000 osób na dobę w województwie,
    - długością linii kolejowych na 10 tysięcy ludności w województwie,
    - średnią liczbą zatrzymań pociągów na dobę w województwie.

Rys.41. Dziewiąta strona dashboardu managerskiego, korelacje – wskaźniki ekonomiczne, źródło: opracowanie własne

R^2 - miara jakości dopasowania modelu do danych, inaczej współczynnik determinacji, przedstawia się następująco:

- w 26% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność wskaźnika urbanizacji,
- w 35% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność gęstości zaludnienia,
- w 48% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność liczby ludności,
- w 36% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność stopy bezrobocia, z nachyleniem ujemnym,
- w 36% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność średniego wynagrodzenia,
- w 28% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność dochodu rozporządzalnego,

- w 26% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność długości dróg publicznych ogółem na 10 tysięcy ludności w województwie, z nachyleniem ujemnym,
- w 31% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność średniego przebiegu auta,
- w 48% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność liczby ofiar śmiertelnych na 10 tysięcy ludności w województwie, z nachyleniem ujemnym,
- w 48% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność liczby stacji kolejowych o wymianie pasażerskiej między 100 a 1000 osób na dobę w województwie,
- w 32% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność długości linii kolejowych na 10 tysięcy ludności w województwie, z nachyleniem ujemnym,
- w 46% zmienność liczby aut na 10 tysięcy mieszkańców została wyjaśniona przez zmienność średniej liczby zatrzymań pociągów na dobę w województwie.

Oznacza to, że np. w przypadku równania modelu korelacji liczby aut na 10 tysięcy mieszkańców i liczy ofiar śmiertelnych wypadków drogowych na 10 tysięcy mieszkańców, które jest następujące: $Y = 58,56 - 2,81X + \varepsilon$, ($\pm$ 1,36) interpretuje się to jak poniżej:

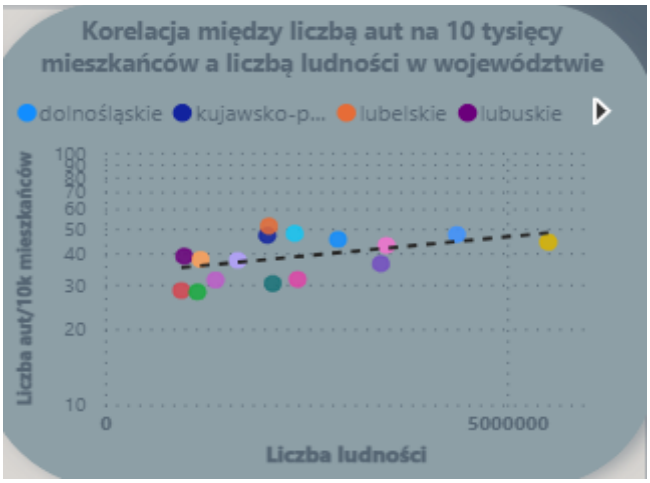
„Jeżeli liczba ofiar śmiertelnych na 10 tysięcy ludności w województwie wzrośnie o 1 osobę, to liczba aut na 10 tysięcy mieszkańców spadnie o 2,81 auta, ze średnim błędem 1,36 auta, ceteris paribus.”

Analogicznie rzecz się ma w przypadku ujemnej korelacji, np. liczby aut na 10 tysięcy mieszkańców z długością linii kolejowych na 10 tysięcy ludności w województwie. Równanie wygląda następująco: $Y = 47,19 - 1,40X + \varepsilon$, ($\pm$ 1,11). Interpretacja analogiczna jak w poprzednim przykładzie, ale z racji ujemnego nachylenia krzywej, oznacza to spadek, a nie wzrost:

„Jeżeli długość linii kolejowych na 10 tysięcy ludności w województwie wzrośnie o 1 km, to liczba aut na 10 tysięcy mieszkańców spadnie o 1,40 auta, ze średnim błędem 1,11 auta, ceteris paribus.”

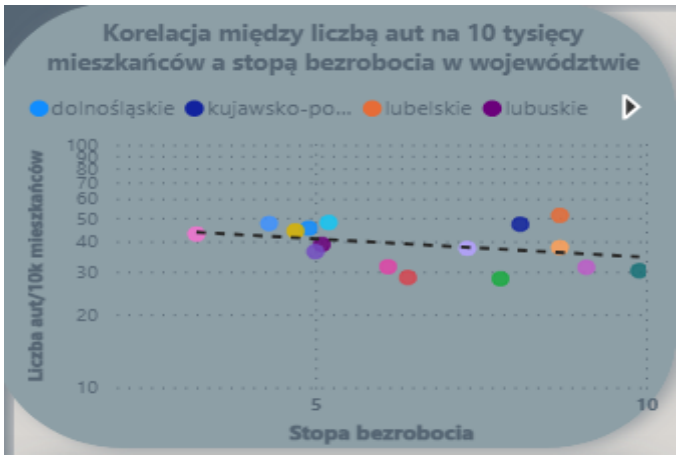
Do najsilniejszych korelacji należą:

- z czynników ludnościowych zdecydowanie korelacja z liczbą ludności, im większa ludność w danym województwie, tym większa liczba aut na 10 tysięcy mieszkańców,



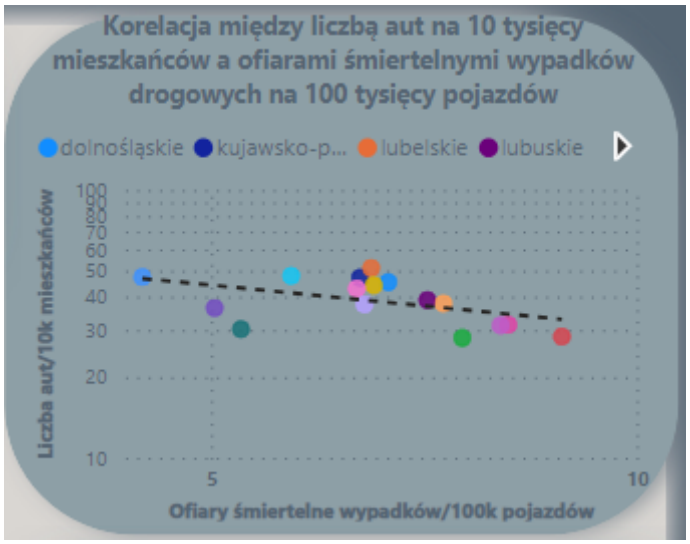
Rys.42. Korelacja między liczbą aut na 10 tysięcy mieszkańców a liczbą ludności w województwie, źródło: opracowanie własne

- z czynników ekonomicznych korelacja ze stopą bezrobocia, im jest ona większa, tym bardziej spada liczba aut na 10 tysięcy mieszkańców,



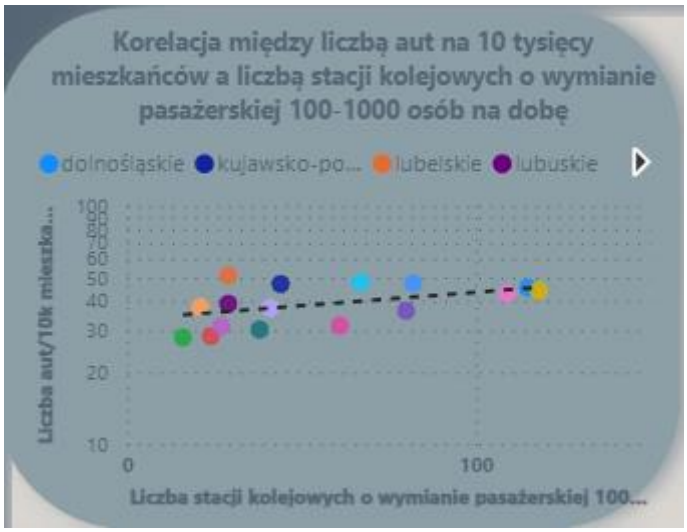
Rys.43. Korelacja między liczbą aut na 10 tysięcy mieszkańców a stopą bezrobocia w województwie, źródło: opracowanie własne

- z czynników drogowych korelacja z liczbą ofiar śmiertelnych wypadków drogowych na 100 tysięcy pojazdów, im jest ta liczba większa, tym bardziej spada liczba aut na 10 tysięcy mieszkańców – może to sporo mówić o tzw. kulturze jazdy, czymś niematerialnym i ciężkim do zdefiniowania, acz dla każdego użytkownika dróg z doświadczeniem ogólnokrajowym dość oczywistym, że w różnych miastach i regionach jeździ się po prostu inaczej,



Rys.44. Korelacja między liczbą aut na 10 tysięcy mieszkańców a liczbą ofiar śmiertelnych wypadków drogowych na 100 tysięcy pojazdów w województwie, źródło: opracowanie własne

- z czynników kolejowych korelacja z liczbą stacji kolejowych o wymianie pasażerskiej między 100 a 1000 osób na dobę w województwie, jeśli liczba ta rośnie, to rośnie także liczba aut na 10 tysięcy mieszkańców. Początkowa hipoteza brzmiała odwrotnie, zakładano że im lepszy dostęp do kolei oraz im lepsza oferta i większa popularność tegoż środka lokomocji, tym liczba aut na 10 tysięcy mieszkańców mniejsza. Okazało się odwrotnie i przypuszczenie z jakiego powodu jest następujące – otóż, kwestią nadrzędną jest ogólne bogactwo danego regionu; im jest on bogatszy, tym większą mobilność wykazują jego mieszkańcy, zaś sam region ma więcej pieniędzy na infrastrukturę wszelkiego rodzaju, tak kolejową, jak i drogową. Kondycja ekonomiczna regionu determinuje zatem ogólnie to jak wygląda mobilność jego mieszkańców, a mając możliwości utrzymać infrastrukturę różnego rodzaju, daje im większy wybór, co jeszcze bardziej sprzyja mobilności.



Rys.45. Korelacja między liczbą aut na 10 tysięcy mieszkańców a liczbą stacji kolejowych o wymianie pasażerskiej między 100 a 1000 osób na dobę w województwie, źródło: opracowanie własne

Zakończenie

Głównym celem pracy było zbadanie struktury rynku aut osobowych w Polsce oraz określenie zróżnicowania przestrzennych zależności i różnic w preferencjach konsumenckich Polaków. To się w znacznej mierze udało. Rozdział pierwszy miał na celu wprowadzenie w tematykę ekonomiczną owej pracy, zaś drugi i trzeci pokazywał proces przygotowywania danych do analizy ekonomicznej, tak od strony teoretycznej, jak i praktycznej. Dzięki uprzedniemu wytłumaczeniu jak działają relacyjne modele baz danych, prościej jest odnaleźć się później w obliczeniach i kodach użytych do stworzenia takowej bazy na podstawie posiadanego materiału z pliku csv. Ostatni, czwarty rozdział poświęcony był właśnie opisowi dashboardu managerskiego oraz temu jak analiza danych została przygotowana i w nim przedstawiona. To tam były też przedstawione cząstkowe wnioski z pracy, badane hipotezy postawione we wstępie, na bieżąco wymieniane wraz z następnymi stronami dashboardu. W tej części pracy natomiast zostaną one niejako podsumowane oraz zebrane, również jako wnioski nasuwające się już po zakończeniu badań.

Jak zostało to wspomniane we wstępie, „aby lepiej wytłumaczyć logikę stojącą zarówno za tą bazą, jak i jej analizą, można wyobrazić sobie siebie w roli np. importera aut.” Jest to bardzo pomocne w zrozumieniu aspektu ekonomicznego tejże pracy. Wśród postawionych hipotez, były między innymi te o zróżnicowaniu przestrzennym poszczególnych cech pojazdów, czyli o tym, że gusta konsumenckie Polaków w zakresie wyboru aut różnią się geograficznie – co innego będzie się najwięcej sprzedawać na Podlasiu, a co innego w Wielkopolsce. To zostało udowodnione, rzeczywiście występują takie zależności, np. inne typy pojazdów, a także ich kolory, czy rodzaje napędu występują na ścianie wschodniej, a inne na ścianie zachodniej kraju. Potwierdziło się także, że napęd 4x4 jest najpopularniejszy w miejscach, gdzie są możliwości wykorzystać go w pełni, bardziej niż w miejscach, które z racji relatywnego bogactwa na zakup aut z takim rodzajem napędu stać. Podobnie potwierdzono korelacje między średnią ceną auta a średnim wynagrodzeniem oraz dochodem rozporządzalnym, w momencie wzrostu zmiennej objaśniającej, rośnie także zmienna objaśniana. Potwierdza to hipotezę o tym, że im bogatszy jest region, tym więcej i tym droższe są tam auta. Podobnie rzecz się ma z liczbą ludności, im większa koncentracja ludności, tym ceny rosną.

Negatywnie zweryfikowane zostały z kolei hipotezy między innymi o tym, że popularność niemieckich marek aut jest tym większa, im bliżej granicy niemieckiej znajduje się województwo. Okazuje się bowiem że i tak najpopularniejszymi markami aut w każdym województwie są te pochodzące z Niemiec, różnią się jedynie konkretnymi markami, a nie krajem pochodzenia. Podobnie nie potwierdzono zależności między większym dostępem do kolei i jej bogatszą ofertą, a malejącą liczbą aut na 10 tysięcy mieszkańców. Jest dokładnie odwrotnie, co można interpretować jako fakt, iż wzrost mobilności mieszkańców jako takiej wpływa pozytywnie na podróżowanie każdym rodzajem transportu i nie występuje zjawisko zmiany jednego rodzaju na inny, pod wpływem lepszego dostępu do niego. Być może sytuacja ulegnie zmianie po rozpowszechnieniu się i tak droższych od aut spalinowych, aut elektrycznych i wtedy alternatywą dla podróży autem będą podróże pociągami, ale póki co, kolej nie stanowi konkurencji dla aut osobowych w takim stopniu, aby ludzie rezygnowali z ich posiadania. Z punktu widzenia pomysłów deglomeracji i „zasypywania różnic" między poszczególnymi obszarami kraju, jest to ciekawy punkt wyjścia do dalszego prowadzenia badań, należy jednak do nich przystąpić za kilka, kilkanaście lat.

Bibliografia

[1] "Santander Consumer Multirent." Accessed: May 10, 2024. [Online]. Available: https://www.scmultirent.pl/

[2] "Global Automotive Consumer Study 2024 ." Accessed: May 10, 2024. [Online].Available: https://www2.deloitte.com/pl/pl/pages/technology/articles/Global-Automotive-Consumer-Study-2024.html

[3] "Raport Branży Motoryzacyjnej Polskiego Związku Przemysłu Motoryzacyjnego 2023/2024." Accessed: May 10, 2024. [Online]. Available: https://www.pzpm.org.pl/pl/Publikacje/Raporty/Rocznik-Raport-Branzy-Motoryzacyjnej-PZPM-2023-2024

[4] T. Stryjakiewicz, R. Kudłak, J. Gadziński, B. Kołsut, W. Dyba, and W. Kisiała, "Czasoprzestrzenna analiza rynku nowych samochodów osobowych w Polsce," *Studies of the Industrial Geography Commission of the Polish Geographical Society*, vol. 31, no. 3, pp. 64–79, Sep. 2017, doi: 10.24917/20801653.313.5.

[5] T. Stryjakiewicz *et al.*, "Przegląd ekonomiczno-przestrzennych badań rynku samochodów osobowych = A review of economic and spatial research on the market for passenger cars," *Przegląd Geograficzny*, vol. 93, no. 2, pp. 249–268, Jul. 2021, doi: 10.7163/PrzG.2021.2.6.

[6] J. Dębski, "Trendy konsumenckie pokolenia Z w obszarze dóbr luksusowych. Gen Z jako przyszły kluczowy klient motoryzacyjnych marek premium," *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, vol. 68, no. 1, pp. 14–25, 2024, doi: 10.15611/pn.2024.1.02.

[7] Ostrowska Teresa M., *Relacyjne systemy bazodanowe. Podstawy projektowania i eksploatacji*. Politechnika Warszawska, 2002.

[8] Leonard Grinke, *Relacyjne bazy danych*. Bielsko-Biała: Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej, 2015.

[9] Krystyna Czapla, *Bazy danych. Podstawy projektowania i języka SQL*. Helion, 2015.

[10] Elżbieta Mrówka-Matejewska, Krzysztof Stencel, and Lech Banachowski, *Systemy baz danych. Wykłady i ćwiczenia*. PJWSTK, 2004.

[11] David Hand, *Eksploracja danych*. Wydawnictwo WNT, 2005.

[12] Ben Forta, *SQL w mgnieniu oka. Opanuj język zapytań w 10 minut dziennie. Wydanie V*. Helion, 2020.

[13] Jacek Bartman, *Bazy Danych*. Rzeszów, 2013.

[14] Włodzimierz Khadzhynov and Piotr Ratuszniak., *Bazy danych*. Wydawnictwo Politechniki Koszalińskiej, 2005.

[15] Lausen Georg and Vossen Gottfried, *Obiektowe bazy danych*. Helion, 2000.

[16] Adam Pelikant, *Hurtownie danych. Od przetwarzania analitycznego do raportowania. Wydanie II*. Helion, 2021.

[17] Matthias Jarke, *Hurtownie danych*. WSIP Wydawnictwa Szkolne i Pedagogiczne, 2003.

[18] Jacek Rumiński, *Wprowadzenie do hurtowni i eksploracji danych*. Wydawnictwo Politechniki Gdańskiej, 2015.

[19] Agnieszka Chodkowska-Gyurics, *Hurtownie danych. Teoria i praktyka*. Wydawnictwo Naukowe PWN, 2017.

[20] Lee Hyunjoung and Sohn Il, *BigData w przemyśle*. Wydawnictwo Naukowe PWN, 2016.

[21] Alex Gorelik, *Korporacyjne jezioro danych. Wykorzystaj potencjał big data w swojej organizacji*. Helion, 2019.

[22] Zhamak Dehghani, *Siatka danych. Nowoczesna koncepcja samoobsługowej infrastruktury danych*. Helion, 2023.

[23] Michiel Rozema and Henk Vlootman, *DAX i Power BI w analizie danych. Tworzenie zaawansowanych i efektywnych analiz dla biznesu*. Helion, 2019.

[24] Russo Marco and Ferrari Alberto, *Kompletny przewodnik po DAX. Analiza biznesowa przy użyciu Microsoft Power BI, SQL Server Analysis Services i Excel*. APN Promise, 2019.

[25] Microsoft Learn, “Tworzenie i wyświetlanie wizualizacji drzewa dekompozycji w usłudze Power BI,.” Accessed: May 10, 2024. [Online]. Available: https://learn.microsoft.com

[26] Research Gate, “ Conceptual view of star and snowflake schemas” Accessed: May 10, 2024. [Online]. Available: https://www.researchgate.net/figure/Conceptual-view-of-star-and-snowflake-schemas_fig2_281587025

[27] Geek for geeks, “ Fact Constellation in Data Warehouse modelling.” Accessed: May 10, 2024. [Online]. Available: https://www.geeksforgeeks.org/fact-constellation-in-data-warehouse-modelling/

Spis ilustracji

Spis tabel

Załączniki

Załącznik nr 1 do zarządzenia Rektora UG nr 70/R/15 ze zm.

OŚWIADCZENIE

Oświadczam, że przedłożona praca dyplomowa została przygotowana przeze mnie samodzielnie, nie narusza praw autorskich, interesów prawnych i materialnych innych osób oraz wykorzystanie materiałów wytworzonych przez generatywne narzędzia sztucznej inteligencji odbyło się w zakresie uzgodnionym z promotorem.

10.06.2024

....

data *podpis*

Załącznik nr 3 do zarządzenia Rektora UG nr 70/R/15

OŚWIADCZENIE

Wyrażam zgodę ~~/ nie wyrażam zgody~~* na udostępnienie osobom zainteresowanym mojej pracy dyplomowej dla celów naukowo-badawczych.
Zgoda na udostępnienie pracy dyplomowej nie oznacza wyrażenia zgody na kopiowanie pracy dyplomowej w całości lub w części.

* *niepotrzebne skreślić*

10.06.2024

....

data *podpis*