**Dokumentacja projektu**

**Przetwarzanie dużych zbiorów danych**

*Rynek osobowych samochodów używanych oraz nowych  
(OtoMoto)*

Autorzy:

Bartosz Kustra 249494

Tomasz Nadrowski 243343

Michał Rudawski 249045

Bartosz Żurakowski 249083

Wrocław, 2021

**Spis treści:**

1. Wstęp
2. Opis projektu
3. Wykorzystywane technologie
4. Pobrane dane
5. Analiza danych
6. Bibliografia
7. **Wstęp**

Spośród oferowanych w obrębie kursu Przetwarzanie Dużych Zbiorów Danych, możliwych do realizacji ścieżek projektowych, omawiana w widocznym projekcie jest ta, oznaczona numerem cztery.

Wyróżniającą ją cechą na tle pozostałych możliwości, była opcja poruszenia wybranego przez grupę projektową faktycznego zagadnienia, które – jak w tym przypadku – powiązane byłoby z życiem codziennym, każdej z osób współtworzących pracę. Dzięki wykonanemu projektowi, możliwe było pozyskanie nie tylko niezwykle istotnej wiedzy praktycznej z zakresu przetwarzania danych, ale również zaspokojenie swojej ciekawości związanej z danym obszarem gospodarczym w Polsce (lub też oczywiście w skali globalnej – uwzględniając pozostałe kraje globu).

Wybierając spośród wielu obszarów życia codziennego, spoglądając na pokrewne między członkami grupy projektowej zainteresowania, oczywistą tematyką, nasuwającą się w pierwszej myśli była motoryzacja. Posiadając doświadczenie w obrębie poszukiwań wymarzonych środków transportu, oraz ogólnorozumiane obeznanie z ową sferą życia, zaskoczeniem nie było wybranie serwisu [www.otomoto.pl](http://www.otomoto.pl), będącego prawdopodobnie największym w Polsce, dedykowanym w pełni motoryzacji, portalem aukcyjnym. Zapewnia on niezliczoną możliwość konfiguracji, personalizacji własnych potrzeb oraz oczekiwań względem wyszukiwania, a co za tym idzie udostępnia pokaźną listę różnego rodzaju filtrów. Każdy z nich może stanowić osobne pole do analizy, ponieważ udostępnia przeważnie kilka wariantów odpowiedzi na filtrujące „zapytanie”.

Co oczywiste, z uwagi na pragnienie stworzenia jak najbardziej rozbudowanego projektu, pozwalającego na jak najbardziej dogłębną analizę zagadnienia, wykorzystywane były dane pozyskane z filtrów podlegających pod nazwane na stronie internetowej: filtrowanie zaawansowane. Opcja ta przekierowuje użytkownika na podstronę dedykowaną samochodom osobowym, w obrębie których prowadzona będzie analiza.

Widocznym w tytule jest, iż przeprowadzone badania dotyczyły zarówno samochodów używanych, jak i nowych. Pomimo faktu, iż na OtoMoto zamieszczane są wyłącznie ogłoszenia, mające na celu sprzedaż samochodów (lub części samochodowych, lecz to nie jest przedmiotem badań), ogłoszenia ów, mogą zostać zamieszczone zarówno przez osoby prywatne oraz dealerów należących do polskich sieci danych marek.

Niezwykle istotnym aspektem, wpływającym również w znaczącym stopniu, na dokonany wybór ścieżki projektowej oraz poruszonej weń tematyki, była potężna i mocno rozwinięta społeczność serwisu. Wszakże działalność OtoMoto nie ogranicza się jedynie do odpowiedniego zarządzania i utrzymywania strony internetowej. Istotnym jest również kontakt z ludźmi bezpośrednio lub pośrednio zaangażowanymi w ten niebanalny projekt. Pomijając liczne publikacje wideo, dostępne na serwisach YouTube oraz Facebook, które nie są przedmiotem analizy projektowej; istotnym jest udostępnianie szerokorozumianej publice odpowiednich statystyk, dotyczących serwisu, aukcji nań wystawianych oraz wielu innych aspektów funkcjonowania rynku samochodów używanych oraz nowych w Polsce. Wszystko to zawierane jest w formie raportów, których analizy dotyczą okresu jednego roku kalendarzowego.



Kierując się ciekawością, odnoszącą się do wspomnianych już raportów, grupa projektowa postanowiła stworzyć niejako odpowiedź, w postaci własnoręcznie sporządzonego raportu. Co oczywiste, nieunikniona jest znacząca różnica w jakości wspomnianych publikacji, względem pracy studenckiej, co determinowane jest wieloma czynnikami. Pierwszym z nich jest brak możliwości pozyskania informacji z przestrzeni całego roku kalendarzowego – wpływ mają na to ograniczenia sprzętowe oraz, co być może bardziej istotne, możliwość wykrycia przez serwis OtoMoto niepożądanego ruchu sieciowego, który miałby charakter cykliczny i niezwykle częsty (scraping danych). Kolejnym z aspektów uniemożliwiających dokładne odtworzenie „konkurencji” prezentowanych w oficjalnym raporcie, jest fakt, iż przeciętny użytkownik, jakim jest każdy z realizujących zadanie studentów, nie posiada dostępu do informacji czysto administracyjnych, takich jak choćby: ilość użytkowników zarejestrowanych na stronie czy też zlecane wyszukiwania, spełniające postawione przez użytkowników kryteria.

Analiza zatem ogranicza się do powszechnie dostępnych informacji, pozyskanych bezpośrednio ze strony, na której operować i wprowadzać zmiany może przeciętny użytkownik portalu, nie administrator.

1. **Opis projektu**

Wspomnianą inspiracją do stworzenia projektu były udostępniane przez OtoMoto, cyklicznie – rocznie, raporty zdradzające wiele interesujących informacji odnoszących się zarówno do tendencji widocznych wśród samych konsumentów – potencjalnych lub faktycznych kupujących samochody – oraz także w świecie motoryzacji, dostępne na portalu są bowiem, wspomniane już, ogłoszenia samochodów nowych.

Dostęp do opisywanych publikacji jest całkowicie bezpłatny oraz powszechny. Pobieranie jest możliwe zarówno bezpośrednio ze strony tytułowego portalu internetowego oraz ze źródeł współpracujących i współdziałających w obrębie realizacji raportów: Samar, Carsmile – uzależnione jest to od roku publikacji dzieła.

Głównym celem widocznego projektu, jest zatem skonfrontowanie pozyskanych informacji oraz próba wykonania zbliżonego ideą raportu, którego główna tematyka dotyczyłaby właśnie statystyk z obszaru samochodów osobowych. Jak już zostało wspomniane, pominięte zostały obszary sprzedaży części samochodowych, motoryzacji jednośladowej oraz pojazdów „ciężkich”.

Prawdopodobnie największym wzywaniem, powiązanym z wykonywaniem tegoż projektu, było odpowiednie pobranie oraz obrobienie surowych danych, w taki sposób, by później w przejrzysty sposób móc wykorzystywać zasoby programistyczne, w celu dokonywania jak najbardziej wnikliwej analizy. Obszar motoryzacji jest niezwykle przepastny, czego najlepszym przykładem jest skala poddanych analizie 212 149 rekordów, w którym na każdy z nich składały się wartości 32 kolumn.

Po pobraniu wspomnianej – potężnej ilości danych, musiałyby one zostać odpowiednio usystematyzowane w tabelę, wedle przytoczonych wcześniej kolumn, tak by każda dana odpowiadała dla siebie bliźniaczym, w pionowej analizie. Szczególną uwagę należało poświęcić występującym – nielicznym, lecz jednak duplikatom. Obecność ogłoszeń zawierających te same parametry (oczywiście w pełnym zakresie) wymagała pozostawienia zaledwie jednej „kopii” takiegoż anonsu. W przypadku niewystosowania wspomnianego procesu, możliwe byłoby zachwianie i zaburzenie precyzji przeprowadzonej analizy, gdyż niemożliwe byłoby pozostanie rzetelnym wobec odbiorcy.

Zważywszy na ciągle rozwijającą się bazę ogłoszeń na portalu OtoMoto – zarówno pod względem ich liczby oraz dostępnych do ustawiania parametrów (przykładem może się stać dodanie w obrębie możliwości kreacji nowych marek, a tym samym modeli), należy przyjąć, iż prezentowane statystyki mają charakter poglądowy, wszelkie pobrane dane pochodzą z dnia 14 kwietnia 2021 roku oraz ówcześnie były najbardziej aktualną z oferowanych wersji, wszakże OtoMoto udostępnia możliwość edytowania dodanych wcześniej ogłoszeń.

1. **Wykorzystywane technologie**

Jak zostało to już wspomniane we wcześniejszej części projektu: potężna baza danych uzupełniana o pola odpowiadające konkretnym filtrom OtoMoto, sprawiła, iż przetwarzanie tychże pól byłoby niemożliwe do wykonania w akceptowalnym czasie pracy, połączonych mocy obliczeniowych komputerów wszystkich osób biorących udział w projekcie. Niezbędne było zatem wykorzystanie rozwiązania chmurowego. Dzięki temu, wszystkie zlecone procesy mogły zostać wykonywane, nie zabierając – wspomnianej już – mocy obliczeniowej urządzeń studenckich – niejako pracując w tle. Dodatkowo, przetwarzanie danych, z wykorzystaniem operacji przeprowadzanych w chmurze, następowały szybciej, niż miałoby to miejsce w przypadku konwencjonalnego podejścia.

Wybór systemu chmurowego był dokonywany pomiędzy dwoma, prawdopodobnie najbardziej popularnymi obecnie rozwiązaniami: AWS oferowane przez Amazon lub Google Cloud. Każdy ze studentów biorących udział we współtworzeniu projektu, na innych zajęciach – w tym także laboratoryjnych – miał już styczność z AWS, zatem wybór padła na rozwiązanie oferowane przez Google. Celem tegoż działania było stworzenie możliwości i sposobności do zapoznania się z nowym środowiskiem oraz poszerzenia swoich horyzontów w obszarze rozwiązań rozproszonych.

Następnie wyekstrahowane już dane, należało w odpowiedni sposób oczyścić, tak by wyeliminować wszelkie niedopatrzenia, poczynione przez osoby wystawiające ogłoszenia, w trakcie ich publikowania. Najczęstsze błędy dotyczyły formatu zapisu ceny, który był niejednoznaczny oraz niespójny z innymi udostępnianymi. Zagadnienie to zostało ułatwienie dzięki oferowanym przez Pandas. Jest to biblioteka stworzona do współpracy z aplikacjami tworzonymi w języku programowania Python. Nazwa wspomnianej biblioteki, została zaproponowana w oparciu o połączenie słów we frazie: „panel data”, które jest pojęciem odnoszącym się bezpośrednio do analizy danych. Dodatkowo, niektóre źródła, podają, iż nazwa powiązana jest również z: „Python data analysis”. Niezależnie jednak od etymologii nazwy samej biblioteki, jest ona niezwykle przydatna w projektach, przypominających lub będących w większej części spokrewnionych z prezentowanym. Ogromna liczba funkcjonalności pozwala między innymi na: indeksowanie danych, dokonywanie zmian w obrębie poszczególnych kolumn (ich dodawanie lub też usuwanie), łączenie poszczególnych porcji danych czy też najważniejsze z perspektywy tegoż projektu: odpowiednie filtrowanie danych. Po wspomnianym „oczyszczeniu” pobranych danych, należało wykorzystać ostatnią z wymienionych właściwości Pandas, by móc zobrazować w przyszłości pożądane wyniki przeprowadzanej analizy.

Oprócz biblioteki Pandas, w części programowej projektu, zaimplementowane zostały również biblioteki: Numpy – pozwalająca na analizę wielowymiarowych wolumenów danych oraz Matplotlib, dzięki której możliwe było kreślenie odpowiednich wykresów, które w „namacalny” sposób przedstawiałyby analizy przeprowadzone w obrębie pobranych danych.

1. **Pobrane dane**

We wcześniejszej części projektu, wspomniana została potężna ilość niezbędnych do przeanalizowania danych. Należałoby zatem pochylić się nad tym, w jaki sposób, informacje, które widoczne są w ogłoszeniach zamieszczonych w tytułowym portalu motoryzacyjnym, znalazły swoje miejsce w zapisie, umożlwiającym ich dalszą analizę.

Całość części programistycznej, opierała się zatem na dwóch skryptach, które były ze sobą w odpowiedni sposób połączone, tak by zapewnić w ogólnym rozrachunku pełne pokrycie danych, pobieranych z serwisu internetowego. Pierwszy z nich, determinował podwalinę pod resztę pobieranych danych, klasyfikując i porządkując marki samochodów, które w momencie scrappingu były oferowane, jako możliwości wyboru w trakcie kreacji ogłoszeń. Obecnie, stale analizując tytułowy serwis internetowy, widocznym jest, iż poczynione zostały kolejne ekspansje, a zatem przybyło kolejnych marek o raz – idąc w sukurs – również i modeli.

Drugi natomiast skrypt, pozwalał na przyporządkowanie do konkretnych marek, podlegających podeń modeli oraz opisanego w ogłoszeniu:

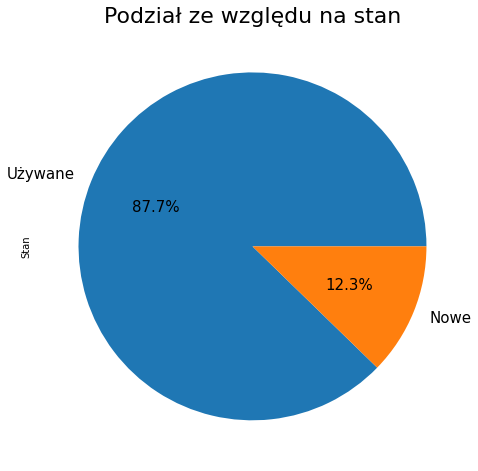
* Parametry powiązane z lokalizacją oferowanego pojazdu,
* Parametry powiązane z silnikiem,
* Cena,
* Kraj pochodzenia,
* Status pojazdu:
  + Czy jest zarejestrowany w Polsce,
  + Czy jest bezwypadkowy,
  + Czy jest zarejestrowany jako zabytek,
  + Czy posiada kierownicę po prawej stronie (Anglik),
  + Czy był serwisowany w Autoryzowanym Serwisie,
* Czy posiada filtr DPF,
* Rok produkcji,
* Data pierwszej rejestracji pojazdu,
* Przebieg,
* Rodzaj oferty i finansowania,
* Parametry powiązane z napędem,
* Typ nadwozia,
* Kolor lakieru,
* Stan pojazdu (nowy lub używany),
* Czy dany pojazd jest obecnie w rękach swojego pierwszego właściciela,
* Data dodania ogłoszenia w serwisie,
* Link do ogłoszenia.

Tak pobrane i powiązane ze sobą dane, zostały umieszczone w tabelach: z podziałem ze względu na markę oraz w zbiorczej. Te pierwsze pomagały w weryfikacji poprawności funkcjonowania oprogramowania, druga natomiast stała się bazą pozwalającą na dokonywanie wszystkich analiz zamieszczonych w poniższym rozdziale.

1. **Analiza danych**

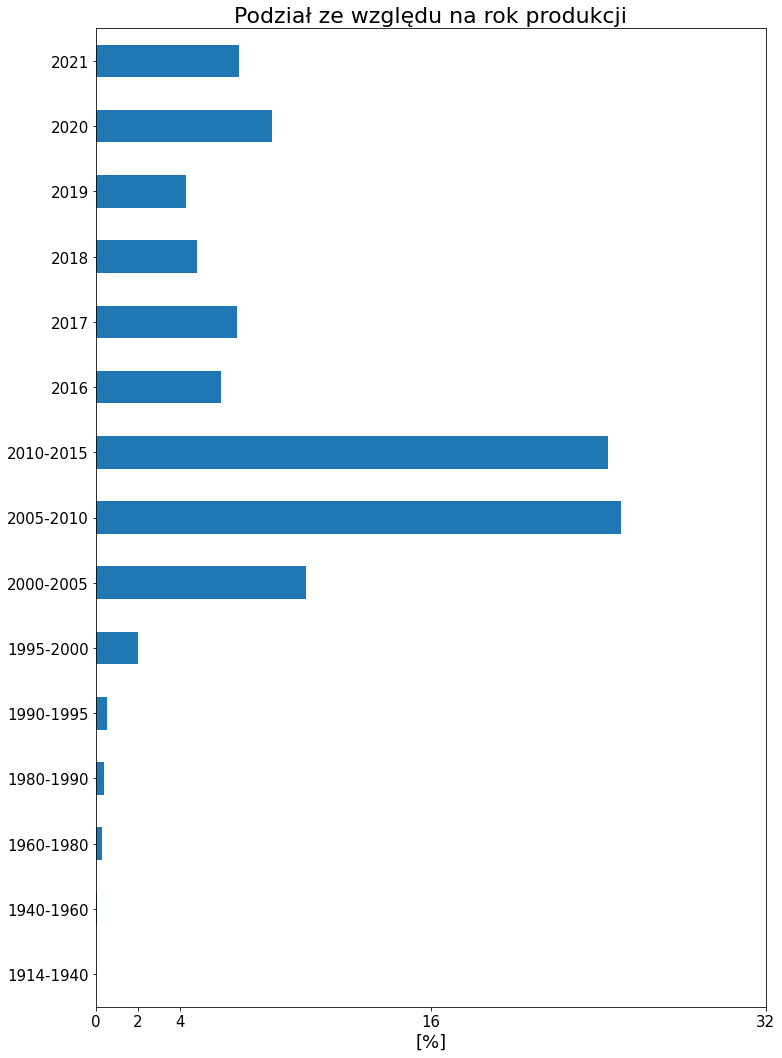
* ~~Najmocniejszy samochód (KM) – top 10,~~
* ~~Najdroższe samochody top 10,~~
* Najdroższe zabytkowe,
* ~~Podział % ze względu na stan: nowe lub używane~~,
* ~~Podział % ze względu na paliwa,~~
* ~~Kraj pochodzenia – jaki % polskich,~~
* ~~Podział % ze względu na segmenty,~~
* ~~Podział % ze względu na przekładnię zmiany biegów,~~
* ~~Ile samochodów z jakiego rocznika (może być % ale nie musi),~~
* ~~Podział % ze względu na cenę:~~
  + ~~0-9 tys~~
  + ~~10-19~~
  + ~~Itd…~~
  + ~~100-149~~
  + ~~150-199~~
  + ~~200+~~
* ~~Topka marek~~
* ~~(opcjonalnie) topka modeli~~
* ~~Podział % ze względu na kolor~~

Pierwszym z zagadnień, które należałoby poruszyć jest oczywiście kwestia stanu samochodów, których dotyczą ogłoszenia na OtoMoto – jak rozkłada się liczba aut nowych oraz używanych. Wykres zaprezentowany został poniżej:



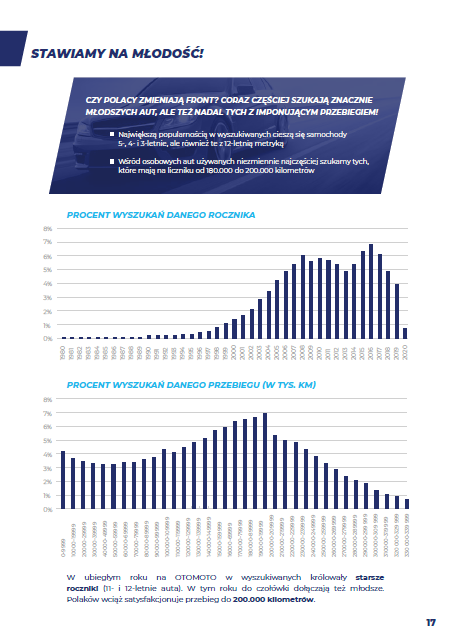
Co jest więc widoczne: pomimo panujących trendów, w których to coraz więcej autoryzowanych serwisów zaczyna wdrażać się w rynek internetowy, za pośrednictwem portalu OtoMoto, wciąż znacznie mocniej dominują na rynku samochody używane. Wiele informacji z branży motoryzacyjnej wskazuje, iż stosunek tych dwóch wielkości będzie dążył do redukcji panującej między nimi różnicy, są to jednak prognozy rozplanowane na wiele najbliższych lat, nie należy zatem traktować ich, jako pewnych.

Wiedząc już, że zdecydowanie bardziej popularne w serwisie są ogłoszenia, których przedmiotem są samochody używane, interesującym zagadnieniem byłoby, dowiedzieć się, jak prezentuje się rozkład roczników produkcji danych maszyn:



Oprócz najbardziej oczywistych wniosków, jakoby samochody, które zostały wyprodukowane w latach 2005- 2015 zdominowały rynek (wszakże jest to też największe pole czasowe), interesującym jest, że dostępnych jest więcej samochodów z obecnego roku (pomimo faktu, że raport sporządzany był w kwietniu), niż z lat: 2019 czy 2018. Najprawdopodobniej zatem, gdy rok kalendarzowy będzie się chylił ku końcowi, przebita może zostać wartość modeli wyprodukowanych w roku 2020.

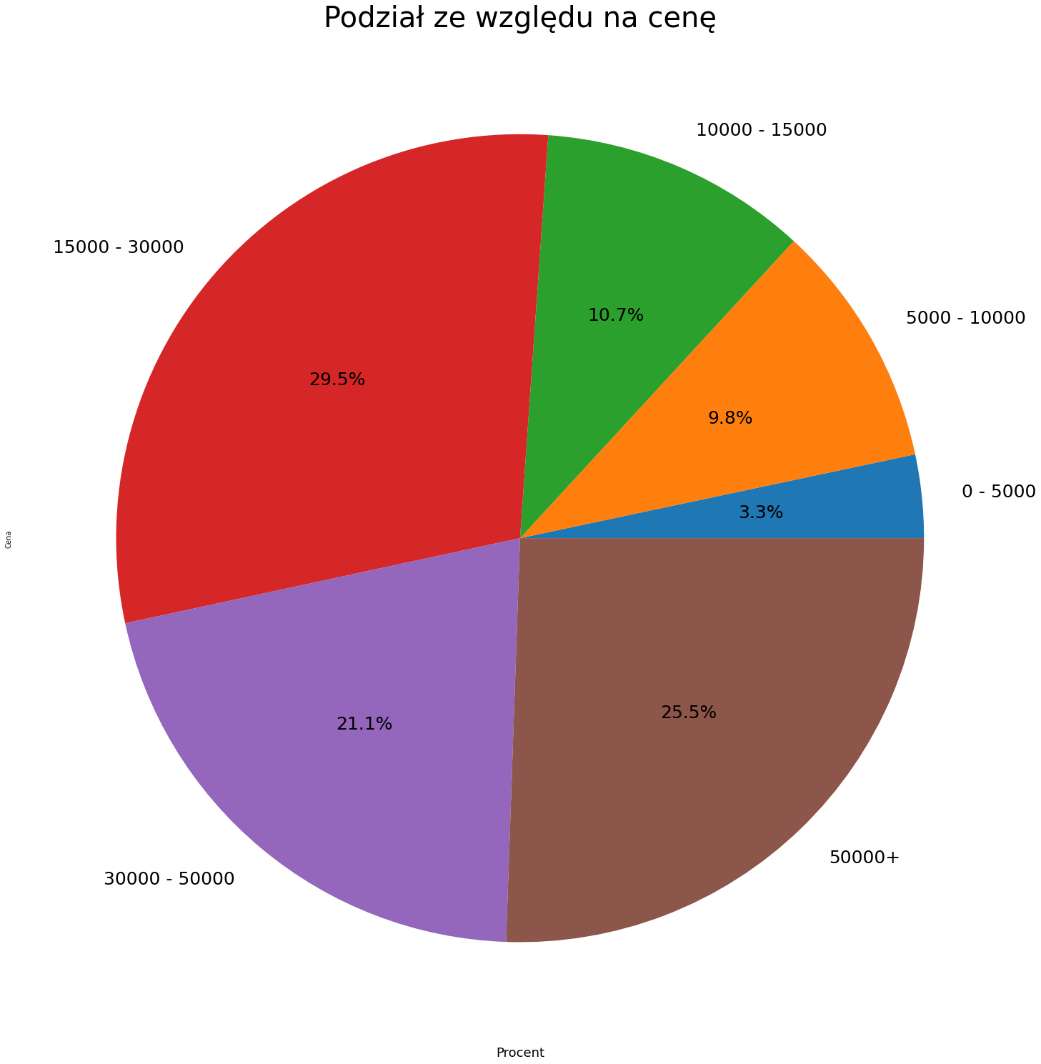
Opisany powyżej proces, może stanowić niejako odpowiedź na rosnące zapotrzebowanie na samochody młodsze, czego dowodzić mogą statystyki sporządzone przez OtoMoto i udostępnione w najnowszym raporcie corocznym (2020), które dotyczą wyszukiwań użytkowników serwisu:



Obraz zawierający tekst

Opis wygenerowany automatycznie

Posiadając wiedzę dotyczącą roczników wystawionych w ogłoszeniach samochodów, zastanawiające mogą być najbardziej popularne przedziały cenowe, w których oferowane są pojazdy:







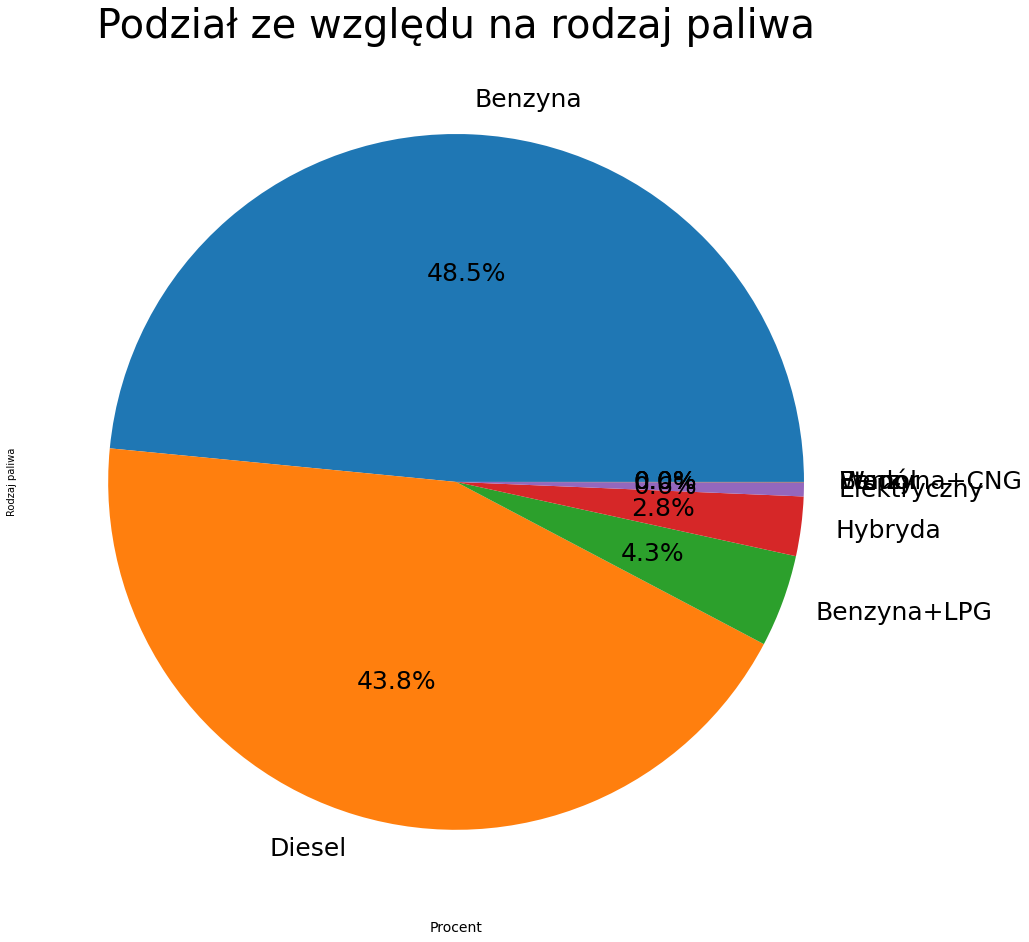
KRAJ POCHODZENIA:

Obraz zawierający kwadrat

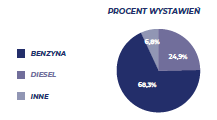
Opis wygenerowany automatycznie

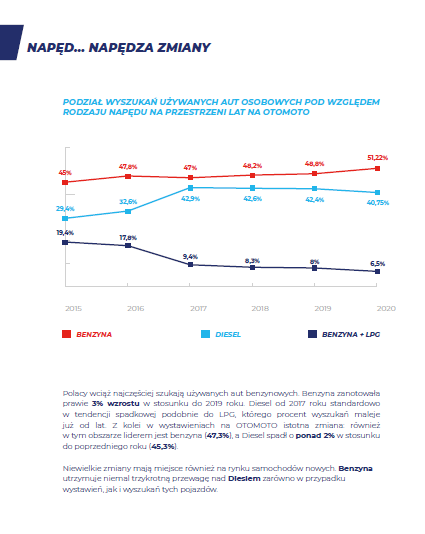


PALIWO:

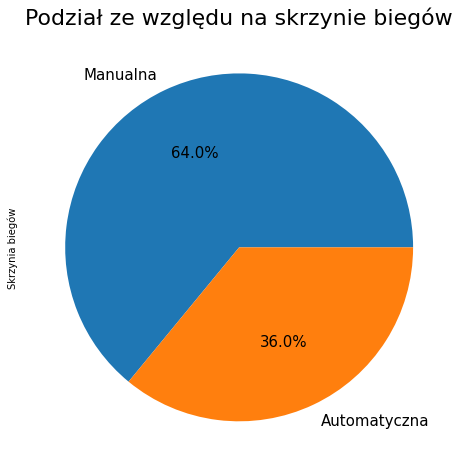






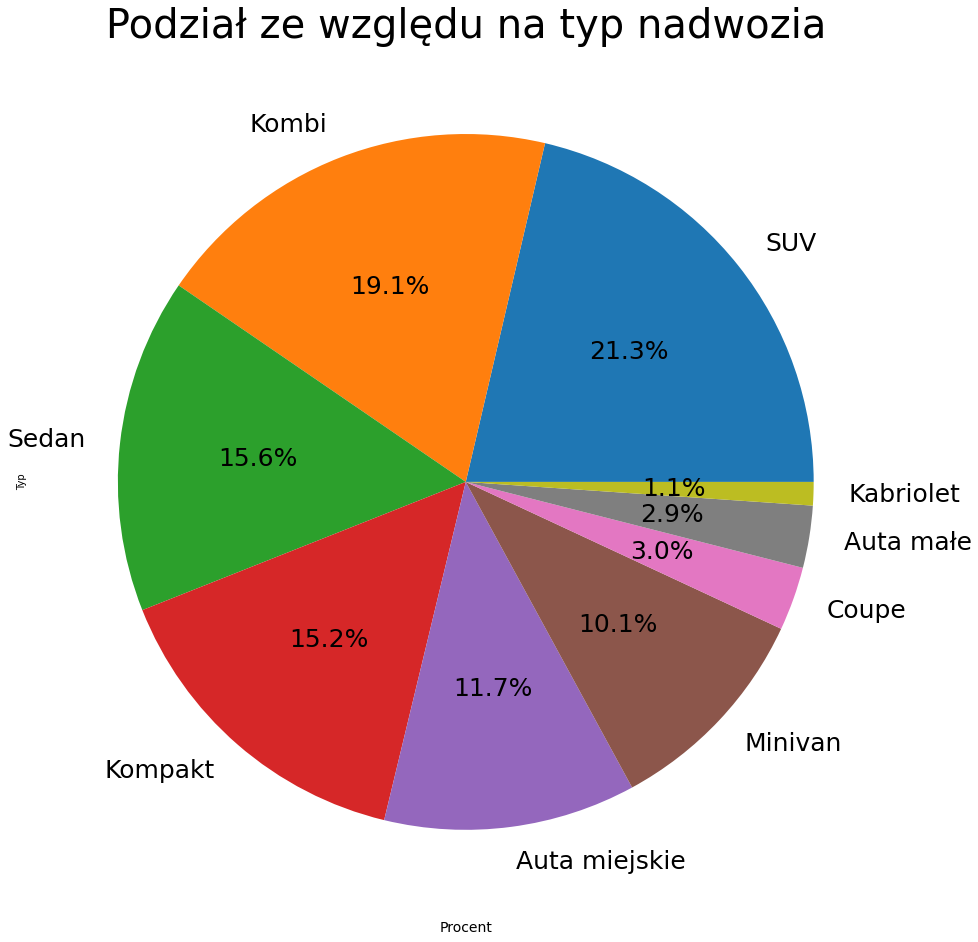


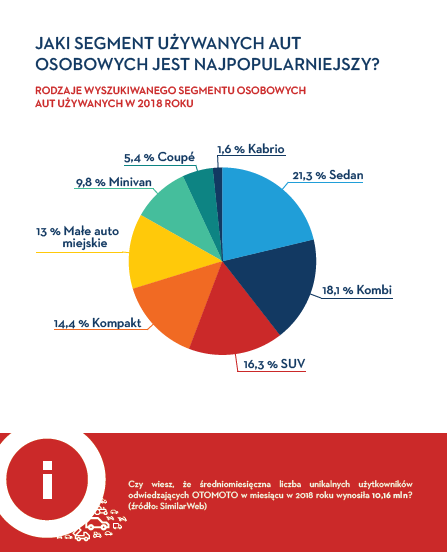
SKRZYNIA BIEGÓW:



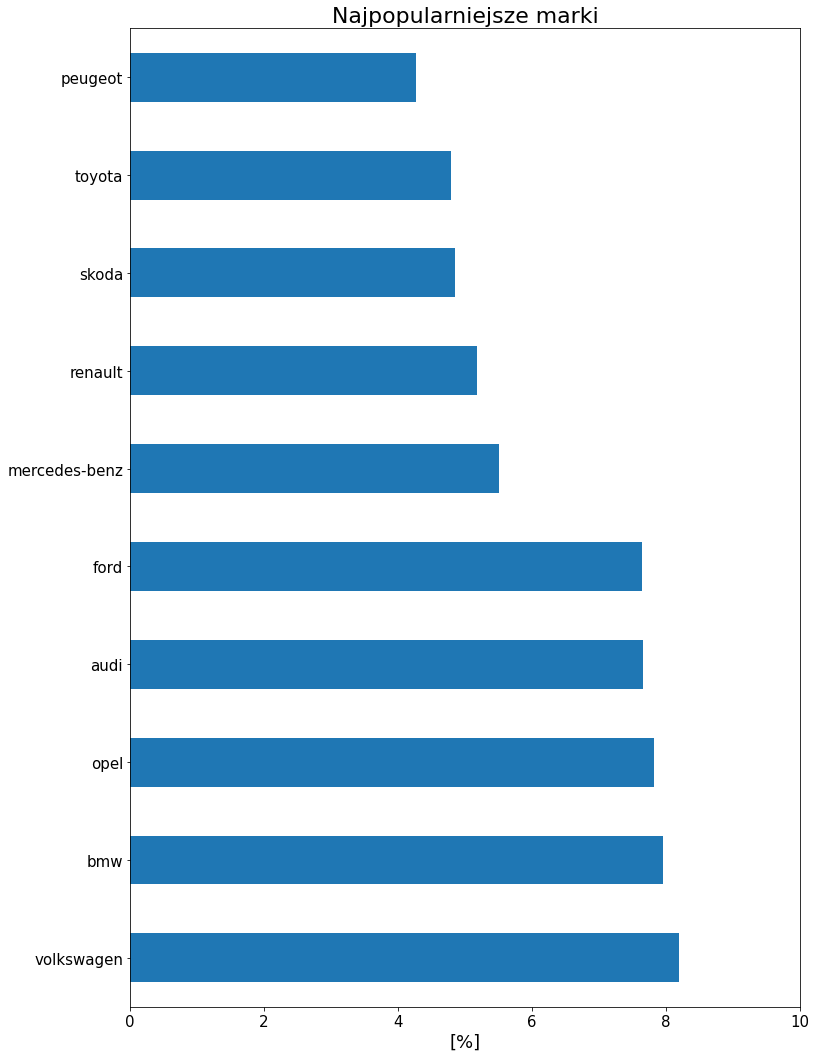


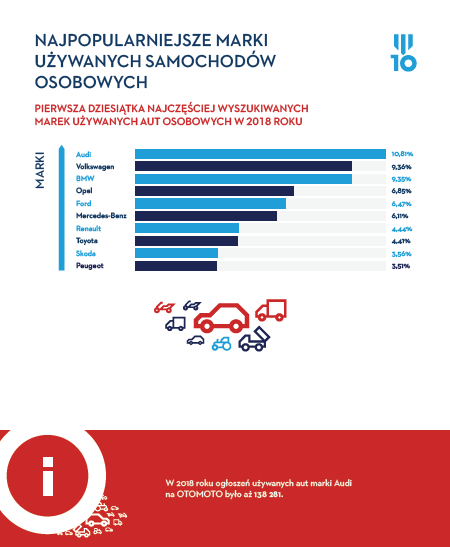
TYP NADWOZIA:





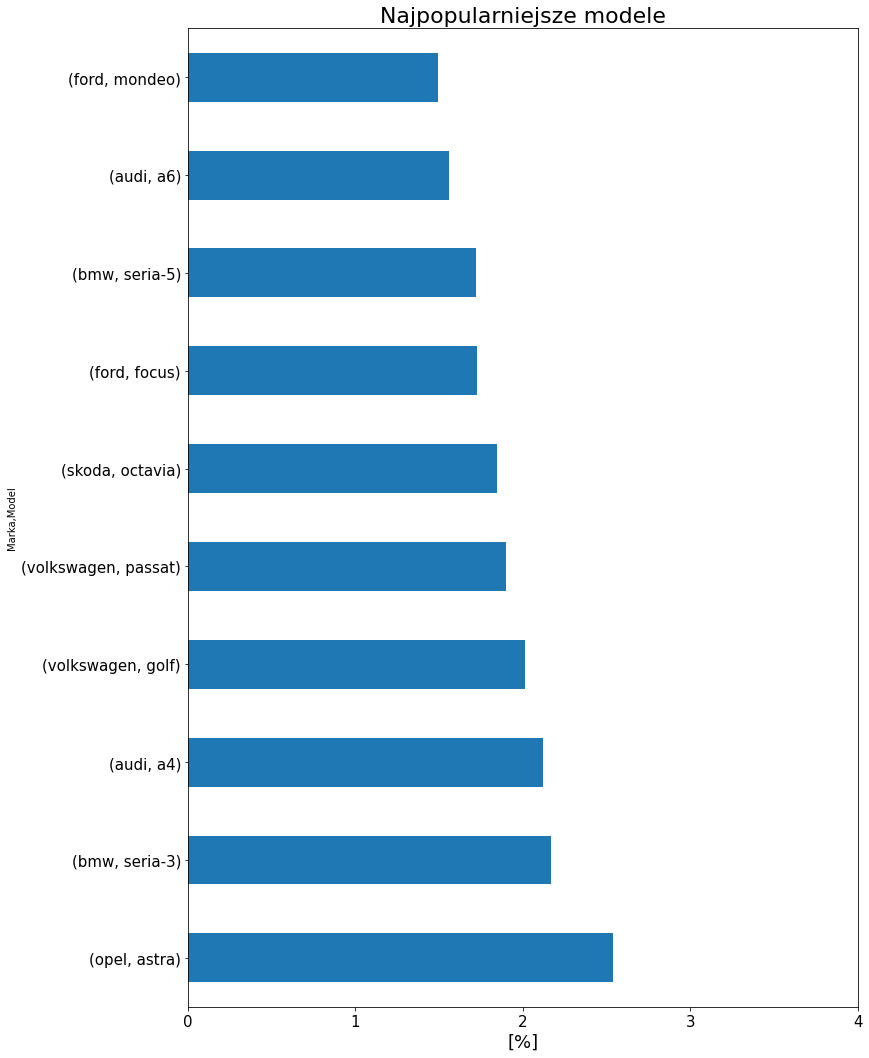
NAJPOPULARNIEJSZE MARKI:



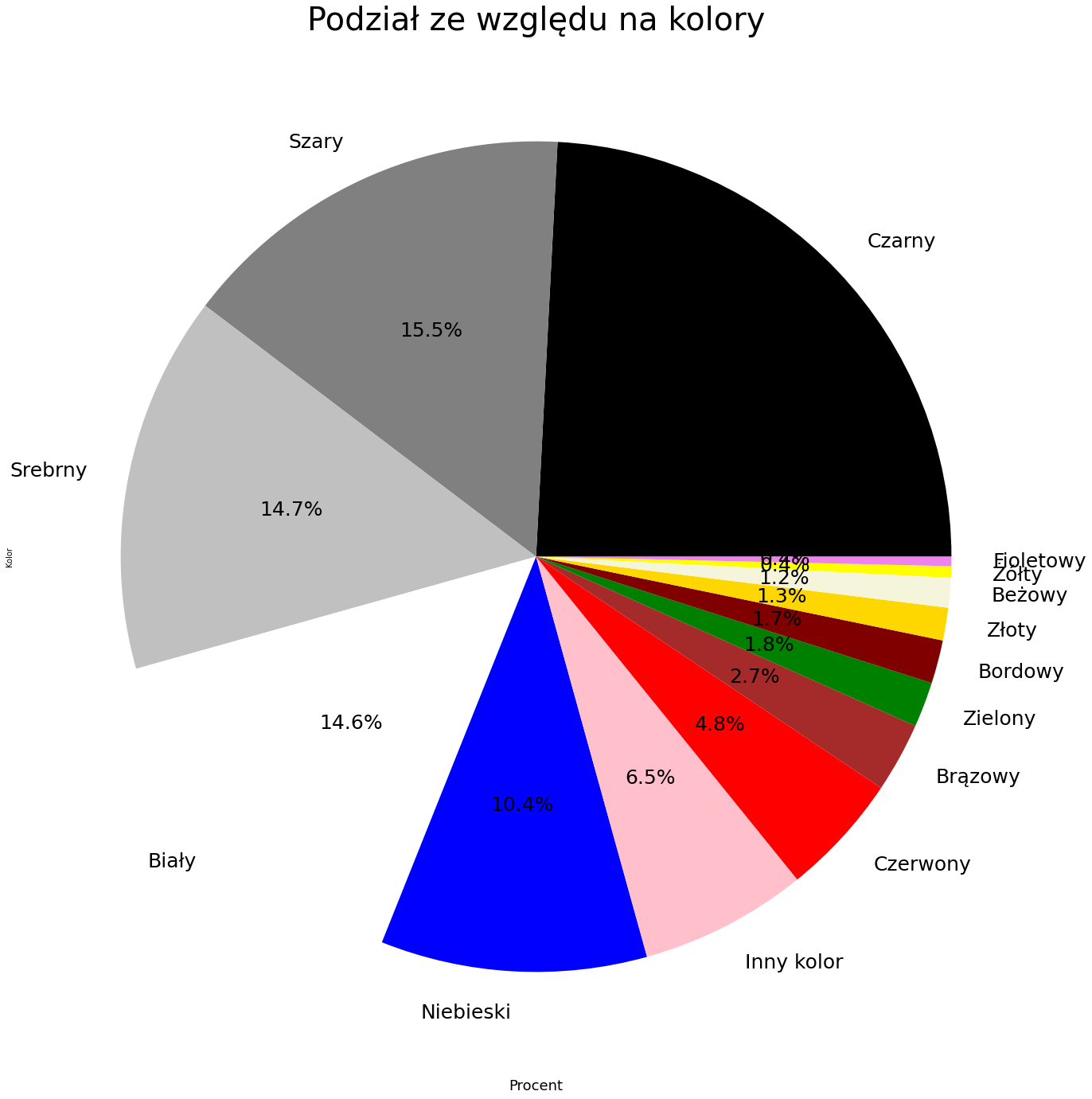


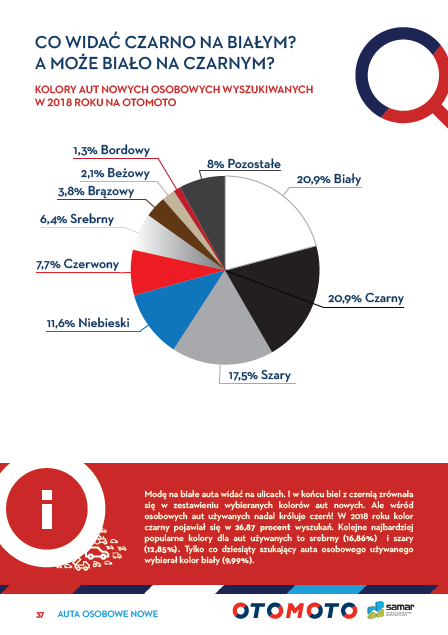


NAJPOPULARNIEJSZE MODELE:



KOLOR:

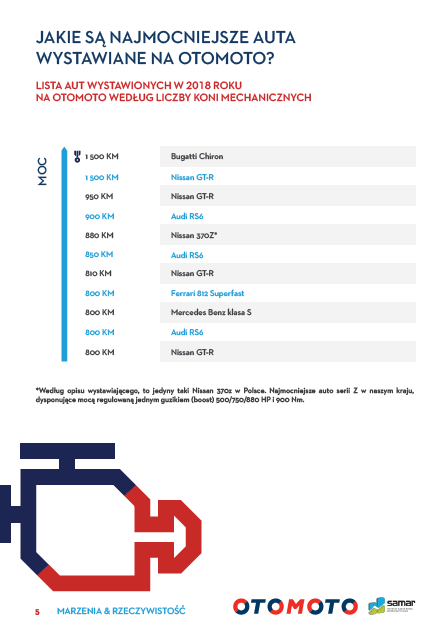




NAJMOCNIEJSZE:

Obraz zawierający stół

Opis wygenerowany automatycznie



NAJDROŻSZE:

Obraz zawierający stół

Opis wygenerowany automatycznie

Obraz zawierający tekst

Opis wygenerowany automatycznie

Obraz zawierający stół

Opis wygenerowany automatycznie

NAJDROŻSZY KLASYK:

Obraz zawierający stół

Opis wygenerowany automatycznie

Obraz zawierający stół

Opis wygenerowany automatycznie

1. **Bibliografia**