AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Analiza danych Ward's Automotive Yearbook z 1985r. dotycząca samochodów

Autor: Majka Mieziańko
Kierunek studiów: Informatyka i Ekonometria
Opiekun pracy: mgr Aneta Piechaczek

Kraków, 2021r.

Spis treści

1. Wstęp

Jak pisze Jean Baudrillard: „Ściśle rzecz ujmując, człowiek epoki dostatku nie egzystuje już, jak uprzednio i odwiecznie, w środowisku innych ludzi, lecz otoczony jest przez przedmioty”[1]. Mimo iż teza francuskiego uczonego może wydawać się zbyt radykalna, to niezaprzeczalnie w jego słowach tkwi też prawda. A jednym z najważniejszych przedmiotów spotykanych w otoczeniu człowieka jest samochód. Według danych zebranych przez branżowy portal Ward’s Auto liczba samochodów na świecie w 2010 roku przekroczyła miliard. W Polsce pod koniec 2017 roku było zarejestrowanych ponad 29,5 miliona pojazdów mechanicznych, z czego 22,5 miliona stanowiły samochody osobowe[2]. Dzięki wprowadzeniu na początku XX wieku linii montażowej przez Henry’ego Forda samochód, najpierw w Stanach Zjednoczonych, a potem w innych krajach stał się przedmiotem dostępnym dla mas. W wielu współczesnych rodzinach auto jest głównym po mieszkaniu artykułem konsumpcji, często spełniającym wiele funkcji i generującym sporą liczbę znaczeń. Samochód dawno przestał być postrzegany, jeśli kiedykolwiek jedynie tę funkcję mu przypisywano, jako przedmiot służący mobilności ludzi.

Choć w swoim badaniu skupię się na nieco bardziej przyziemnych i oddalonych od patetycznej semantyki parametrach, chciałabym zwrócić uwagę na fakt, w jak dużym stopniu jesteśmy kulturowo, ale i też czysto praktycznie związani ze stanem posiadania auta.

[1] J. Baudrillard, *Społeczeństwo konsumpcyjne. Jego mity i struktury*, Wydawnictwo Sic!,

[2] Rocznik Statystyczny Rzeczypospolitej Polskiej 2018, Główny Urząd Statystyczny, Warszawa 2018, s. 551, tabl. 24(465).

2. Cel pracy i hipotezy

W niniejszej pracy chciałabym się skupić na analizie danych pochodzących z Ward's Automotive Yearbook z 1985r. Jako, że interesuję się motoryzacją, chciałabym odkryć jak kształtują się pewne zależności dotyczące parametrów samochodów oraz ich ceny. Ponadto dzięki statystyce opisowej mam zamiar uzyskać wstępne informacje na temat skośności danych czy charakterystyki ilościowej danych zmiennych, dzięki czemu będziemy mieć szerszy ogląd podczas przeprowadzania konkretnych testów statystycznych. Ostatecznie każde pytanie postaram się również zilustrować odpowiednimi wykresami w celu wizualnzacji danego problemu lub wyniku badania.

W swojej pracy dokonam analizy danych dotyczących samochodów pod kątem następujących pytań badawczych:

1. **Czy większość aut ma silniki wolnossące (std)? Czy wiekszość samochodów napędzana jest dieslem? Jakich marek samochodowych jest najwięcej? Które marki samochodów są najdroższe?**
   Metoda: statystyka opisowa.

Zestaw hipotez:

H_0: większość aut ma silniki wolnossące

H_1: większość aut ma silniki turbo

H_0: większość samochodów napędzana jest dieslem

H_1: większość silników napędzana jest LPG

2. **Jak skorelowane są dane? Czy wielkość silnika jest powiązana z wielkością auta?**
   Metoda: współczynnik korelacji Pearsona.

Zestaw hipotez dla cor.test():

H_0: wielkość silnika nie jest skorelowana z wielkością auta, $r = 0$

H_1: wielkość silnika jest skorelowana z wielkością auta, $r \neq 0$

3. **Czy auta napędzane LPG odznaczają się mniejszym spalaniem w mieście niż pojazdy napędzane dieslem?**
   Metoda: test U-Mann-Whitney'a-Wilcoxona.

Zestaw hipotez

H_0: F(x) = G(x) dla wszystkich x

H_1: H1: F(x) $\neq$ G(x) dla niektórych x,

4. **Czy ceny pojazdów podlegają rozkładowi Gaussa? Jaki ma to wpływ na dalsze badania? Czy logarytm zmiennej cena lepiej odwzorowuje rozkład normalny?**
   Metoda: test Kolmogorova-Smirnova.

Zestaw hipotez:

H_0: ceny mają rozkład normalny

H_1: ceny nie mają rozkładu normalnego

5. **Czy ceny aut z silnikami turbo są większe od cen aut z silnikami wolnossącymi?**
   Metoda: test istotności t-Studenta dla dwóch średnich – jednostronny obszar krytyczny.

Zestaw hipotez:

H_0: $m_1 = m_2$.

H_1: $m_1 > m_2$,

gdzie m_1 oznacza średnią cenę w populacji aut z silnikiem turbo, a m_2 średnią cenę w populacji aut z silnikiem turbo

6. **Czy rodzaj nadwozia auta wpływa na jego cenę? Z jakim rodzajem nadwozia ceny aut najbardziej różnią się od pozostałych?**
   Metoda: Anova oraz test Kruskala-Wallisa.

Zestaw hipotez:

H_0: wartości oczekiwane w podgrupach są sobie równe,

H_1: przynajmniej jedna wartość oczekiwana w podgrupie jest różna

3. Część teoretyczna

Wykorzystywane w pracy metody

3.1 Statystyki opisowe

Statystyki opisowe posłużą mi do opisania najważniejszych informacji na temat analizowanych w badaniu zmiennych i grup osób badanych. Za ich pomocą określamy liczbę obserwacji, średnie wyniki, zróżnicowanie obserwacji i inne. Do statystyk opisowych możemy zaliczyć:

- miary występowania, np. liczba obserwacji, procent skumulowany

- miary położenia np. średnia, mediana, modalna

- miary zmienności np. odchylenie standardowe, wariancja

- miary asymetrii np. skośność

- miary położenia np. kurtoza

Głównym celem przedstawienia statystyk opisowych w pracach naukowych jest opisanie najważniejszych właściwości zmiennych, bądź grup badanych osób pod względem danych zmiennych.

3.2 Współczynnik korelacji

Współczynnik korelacji jest wielkością niemianowaną, przyjmującą wartość z przedziału (-1,1). Jeśli r = 0, zmienne X i Y są nieskorelowane. Jeśli pr≠ 0, to zmienne określamy jako skorelowane. W przypadku, gdy r > 0 proste regresji mają dodatnie nachylenie i mówimy o dodatnim skorelowaniu obu zmiennych. W przypadku gdy r < 0 proste regresji mają ujemne nachylenie i mówimy o ujemnym skorelowaniu zmiennych.[3] Moduł

[3] Introductory Statistics (Third Edition), Sheldon M. Ross, 2010

współczynnika korelacji równy jest jedności wtedy i tylko wtedy, gdy między cechami zachodzi funkcyjny związek liniowy. Bezwzględna wartość współczynnika korelacji wskazuje na siłę liniowego skorelowania cech, wyrażającą stopień determinacji wartości jednej cechy przez wartości drugiej cechy.

Miarą siły korelacji między zmiennymi w przypadku wielowymiarowym, gdy liczba zmiennych przekracza 2, jest współczynnik korelacji wielorakiej oraz współczynnik korelacji cząstkowej.

Zauważmy, że z definicji współczynnika korelacji wynika, iż współczynnik ten – w przeciwieństwie do wskaźników korelacyjnych – przyjmuje taką samą wartość niezależnie od tego, którą z cech przyjmujemy za zależną, a którą za niezależną. Ponadto, współczynnik korelacji może być obliczany zarówno na podstawie danych indywidualnych, jak i pogrupowanych w tablicy korelacyjnej, pod warunkiem, że obie rozpatrywane cechy są mierzalne.[4]

3.3 Test U Manna-Whitneya-Wilcoxona

Za pomocą tego testu można sprawdzić hipotezę, że dwie niezależne próby pochodzą z identycznych populacji. Zastępuje on klasyczny parametryczny test t-Studenta dla dwóch średnich, gdy nie są spełnione założenia wymagane do jego stosowania.

Sprawdzaną hipotezę zapisujemy jako:

H_0: F(x) = G(x) dla wszystkich x

wobec dwustronnej alternatywy

H_1: F(x) $\neq$ G(x) dla niektórych x,

gdzie F(x) i G(x) są dystrybuantami odpowiednio w populacjach A i B.

Jeżeli założymy, że ewentualne różnice pomiędzy populacjami dotyczą tylko lokalizacji rozkładów (czyli krzywe rozkładu mają ten sam kształt, a jedynie są przesunięte względem siebie wzdłuż osi odciętych), to powyższe hipotezy będą dotyczyć także średnich w odpowiednich rozkładach i mogą być zapisane jako:

H_0: E(X) = E(Y)

H_1: E(X) $\neq$ E(Y)

[4] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str. 347-349

Jeśli hipoteza zerowa jest prawdziwa, to obserwacje z obu prób powinny być losowo rozmieszczone w ramach uzyskanego rankingu i w konsekwencji poziom rang, mierzony ich. Sumą lub wartością średnią powinien być zbliżony. Jeśli próby pochodzą z populacji o różnych średnich, to dane pochodzące z populacji o niższej średniej będą miały na ogół niższe rangi, a obserwacje z populacji o wyższej średniej będą miały wyższe rangi.[5]

3.4 Test zgodności Kołmogorowa-Smirnowa

Test ten, opracowany przez Smirnowa, służy do weryfikacji hipotezy, że dwie populacje mają jednakowy rozkład lub (co na jedno wychodzi), że dwie niezależne próby pochodzą z tej samej populacji.

Załóżmy, że badane są dwie populacjie, które mają ciągłe rozkłady opisane dystrybuantami $F_1(x)$ oraz $F_2(x)$. Sformułowaną hipotezą jest:

H_0: $F_1(x) = F_2(x)$

wobec

H_1: $F_1(x) \neq F_2(x)$

W celu zweryfikowania tej hipotezy należy wylosować niezależnie z każdej populacji próbę i następnie określić dystrybuanty empityczne $Fn_1(x)$ oraz $Fn_2(x)$, gdzie symbole n_1 i n_2 oznaczają liczebności obu prób.

Jeśli prawdziwa jest H_0 to róznica między wartościami tych dystrybuant nie powinna być zbyt duża. Miarą zgodności tych dwóch rozkładów empirycznych jest statystyka:

$D\ n_1, n_2 = \sup | Fn_1(x) - Fn_2(x) |$.[6]

3.5 Test istotności t-Studenta dla dwóch średnich

Przyjmijmy, że obie badane populacje mają rozkłady normalne $N(m_1, \sigma_1)$ oraz $N(m_2, \sigma_2)$, ale żaden z tych parametrów nie jest znany. Jeśli chcemy zweryfikować hipotezę:

H_0: $m_1 = m_2$.

H_1: $m_1 \neq m_2$,

[5] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str.280-281

[6] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str.289

to nie możemy w tym celu wykorzystać statystyki U, ponieważ nie znamy wartości σ_1, σ_2. Jeśli natomiast, nie znając tych wartości, będziemy wiedzieli, że $\sigma_1 = \sigma_2 = \sigma$ tzn. odchylenia standardowe w obu populacjach są identyczne, to będziemy mogli skorzystać z tego, że statystyka t ma rozkład t-Studenta z $v = n_1 + n_2 - 2$ stopniami swobody.

Ponadto, jeśli prawdziwa jest H_0 to róznica między średnimi arytetycznymi nie powinna być zbyt duża, a obszar krytyczny określony jest równością:

$$P(|t| >= t_{a,\, n1 + n2 - 2}) = a$$

gdzie a jest poziomem istotności.

Jest oczywiste, że dla hipotez dotyczących dwóch wartości oczekiwanych można budować również jednostronne obszary krytyczne (w zależności od postaci hipotezy alternatywnej) dokładnie w taki sam sposób, jak dla hipotez dotyczących jednej wartości oczekiwanej.[7]

3.6 Analiza wariancji ANOVA oraz test Kruskala-Wallisa

Metoda statystyczna zwana analizą wariancji została opracowana i upowszechniona w latach dwudziestych XX wieku przez R.A. Fishera. Wprowadzono tę metodę najpierw do doświadczalnictwie rolnym, później znalazła ona zostosowanie w wielu innych dziedzinach badań.

Ogólnie mówiąc, analiza wariancji jest techniką badania wyników (obserwacji), które zależą od jednego lub więcej czynników działających równocześnie. Za pomocą tej techniki określa się, czy wyodrębnione czynniki wywierają wpływ na obserwowane wyniki. Zmienną, która takiej obserwacji podlega nazywamy zmienną objaśnianą.

Jeśli w badaniu uwzględnia się jeden czynnik, to mamy do czynienia z analizą wariancji z klasyfikacją pojedynczą (jednokierunkową analizą wariancji). Możliwe jest także badanie wpływu dwóch (lub więcej) czynników na zmienną objaśnianą. Mówimy wtedy o analizie wariancji z klasyfikacją podwójną lub wielowymiarowej analizie wariancji.

W celu weryfikacji hipotezy o identyczności średnich w r populacjach przeprowadza się postępowanie polegające na dekompozycji sumy kwadratów odchyleń od średniej z próby na dwa składniki. Jeden z nich mierzy stopień zróżnicowania wartości zmiennej objaśnianej Y między wyróżnionymi grupami lub też, inaczej mówiąc, stopień zróżnicowania tej zmiennej

[7] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str.224-225

wywołanego działaniem uwzględnionego czynnika. Drugi składnik wyraża zróżnicowanie zmiennej wewnątrz każdej z grup, wynikające z losowego charakteru zmiennej Y.

Podstawą do oceny istotności różnic między średnimi z prób jest ocena wkładu zróżnicowania międzygrupowego i wewnątrzgrupowego do ogólnego zróżnicowania zmiennej. Im wyższy jest udział tego pierwszego, tym wyraźniejszy wpływ czynnika na poziom zmiennej.

W modelu analizy wariancji przyjmuje się założenia o normalności zmiennej objaśnianej oraz o jednorodności jej wariancji (identyczności wariancji) w badanych populacjach.[8]

Analiza wariancji dla rang jest nieparametryczną alternatywą testu ANOVA. Test Kruskala-Wallisa jest uogólnieniem testu U Manna_whitneya na większą niż 2 liczbę populacji. Służy on do sprawdzania – na podstawie niezależnych prób – hipotezy, że rozkłady cechy są w kilku populacjach jednakowe. Podobnie jak w klasycznym teście F, poszczególne populacje mogą być wyodrębnione na podstawie jakościowych kategorii zmiennej niezależnej, podczas gdy porównywane są rozkłady rang dla zmiennej zależnej, mierzonej w skali porządkowej.

Test Kruskala-Wallisa wykorzystuje się także dla zmiennych interwałowych, gdy założenia klasycznego modelu analizy wariancji, a więc założenie o normalnym rozkładzie i (lub) założenie o jednorodnej wariancji cechy w grupach, nie są spełnione. Jego względna efektywność w stosunku do klasycznego testu F wynosi ok. 95%, zaletą są natomiast mniejsze wymagania co do założeń. W przypadku stosowania testu Kruskala-Wallisa wymaga się jedynie, aby:

- obserwacje mogły być rangowane dla wszystkich grup łącznie,
- próby z poszczególnych populacji były od siebie niezależne
- najniższą skalą pomiaru zmiennej zależnej była skala porządkowa.[9]

[8] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str.300-309
[9] Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski, str.323-324

4. Część praktyczna

4.1 Informacje na temat zbioru danych

Wykorzystany przeze mnie zbiór danych składa się z trzech typów podmiotów:

(a) specyfikacja samochodu pod względem różnych cech,

(b) przypisana mu ocena ryzyka ubezpieczeniowego,

(c) znormalizowane straty w użytkowaniu w porównaniu z innymi samochodami. Podpunkt (b) odnosi się do stopnia, w jakim auto jest bardziej ryzykowne, niż wskazuje na to jego cena. Samochody mają początkowo przypisany symbol czynnika ryzyka związany z ich ceną. Następnie, jeśli jest ono bardziej ryzykowne (lub mniej), ten symbol jest korygowany, poprzez przesunięcie go w górę (lub w dół) skali. Matematycy zajmujący się kalkulacją ubezpieczeniową nazywają ten proces „symbolizowaniem". Wartość +3 oznacza, że auto jest ryzykowne, -3 że prawdopodobnie jest całkiem bezpieczne. Trzeci czynnik to względna średnia wysokość odszkodowania za rok ubezpieczenia pojazdu. Wartość ta jest znormalizowana dla wszystkich samochodów w ramach określonej klasyfikacji wielkości (dwudrzwiowe małe, kombi, sportowe / specjalistyczne itp.) i przedstawia średnią stratę na samochód rocznie. Podpunkty (b) i (c) nie są przeze mnie używane w badaniu, zatem usunęłam je w procesie preprocessingu.

Źródłami danych Ward's Automotive Yearbook z 1985r są:

1) 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.

2) Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038

3) Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

W badaniu zajmuję się następującymi zmiennymi:

fuel-type = rodzaj paliwa
aspiration = parametr silnika (wolnossący czyli *std* oraz *turbo*)
num-of-doors = liczba drzwi

body-style = rodzaj nadwozia

drive-wheels = napęd (na oś)

engine-location = umiejscowienie silnika

wheel-base = rozstaw osi

width = szerokość auta

height = wysokość auta

curb-weight = masa własna auta

engine-type = typ silnika

num-of-cylinders = liczba cylindrów

engine-size = wielkość silnika

fuel-system = system paliwowy

bore = parametry

stroke = parameter cylindra

compression-ratio = parameter silnika

horsepower = konie mechaniczne

peak-rpm = maksymalne obroty silnika

city-mpgsort = spalanie w mieście (a raczej miles per galon – czyli ilość mil na galon benzyny)

highway-mpg = mile na gallon na autostradzie

price = cena

4.2 Preprocessing danych

Mówi się, że kluczem do przygotowania wiarygodnych wyników analiz są dane. Mówiąc o przygotowaniu danych (data preprocessing) – należy mieć na uwadze zarówno aspekt techniczny – poprawne załadowanie danych do środowiska programistycznego, jak i aspekt analityczny. Często może okazać się, że dane będą zawierały niekompletne dane – będziemy musieli wówczas arbitralnie podjąć decyzję co z takimi danymi zrobić.

Pierwszym krokiem, który podjęłam po wczytaniu danych, była zmiana typów niektórych wartości, które widniały jako liczby zapisane w formacie string. Postanowiłam zmienić ich typ na numeryczny. Następnie usunęłam rzędy z brakującymi danymi, jako że chciałam doprowadzić dane do takiego stanu, żeby wszystkie wartości były reprezentatywne – niosły jakąś informację. Operacje te – wraz z wczytaniem danych – zamknęłam w jednej funkcji *func_zbior_danych*, a następnie utworzyłam zmienną *automobil_dane,* która uruchamia tę funkcję.

Ostatnim etapem było usunięcie kolumn, z których nie zamierzałam korzystać w większości badań: *normalized.losses* oraz *symboling*. Są to dane wprowadzone na potrzeby oceny ryzyka ubezpieczeniowego, tak więc w początkowych etapach pracy były dla mnie zbędne.

Podsumowując, stworzyłam dwie istotne zmienne:

-automobil_dane – która zawiera wszystkie dane, po wstępnym preprocessingu

- automobil_clean – która zawiera dane po preprocessingu bez zmiennych dotyczących ryzyka ubezpieczeniowego (*normalized.losses* oraz *symboling)*

Tabela 1 Struktura danych automobile_dane

```
>  str(automobil_dane)
'data.frame':   195 obs. of  26 variables:
 $ symboling        : int  3 3 1 2 2 2 1 1 1 2 ...
 $ normalized.losses: chr  "?" "?" "?" "164" ...
 $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
 $ fuel.type        : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration       : chr  "std" "std" "std" "std" ...
 $ num.of.doors     : chr  "two" "two" "two" "four" ...
 $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ engine.location  : chr  "front" "front" "front" "front" ...
 $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ length           : num  169 169 171 177 177 ...
 $ width            : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 64.8 ...
 $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 54.3 ...
 $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 2395 ...
 $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ num.of.cylinders : chr  "four" "four" "six" "four" ...
 $ engine.size      : int  130 130 152 109 136 136 136 136 131 108 ...
 $ fuel.system      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
 $ bore             : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.5 ...
 $ stroke           : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 2.8 ...
 $ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 8.8 ...
 $ horsepower       : num  111 111 154 102 115 110 110 110 140 101 ...
 $ peak.rpm         : num  5000 5000 5000 5500 5500 5500 5500 5500 5500 5800 ...
 $ city.mpg         : int  21 21 19 24 18 19 19 19 17 23 ...
 $ highway.mpg      : int  27 27 26 30 22 25 25 25 20 29 ...
 $ price            : num  13495 16500 16500 13950 17450 ...
```

Struktura danych *automobil_dane*

Tabela 2 Struktura danych automobile_clean

```
>   str(automobil_clean)
'data.frame':   195 obs. of  24 variables:
 $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
 $ fuel.type        : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration       : chr  "std" "std" "std" "std" ...
 $ num.of.doors     : chr  "two" "two" "two" "four" ...
 $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ engine.location  : chr  "front" "front" "front" "front" ...
 $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ length           : num  169 169 171 177 177 ...
 $ width            : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 64.8 ...
 $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 54.3 ...
 $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 2395 ...
 $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ num.of.cylinders : chr  "four" "four" "six" "four" ...
 $ engine.size      : int  130 130 152 109 136 136 136 136 131 108 ...
 $ fuel.system      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
 $ bore             : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.5 ...
 $ stroke           : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 2.8 ...
 $ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 8.8 ...
 $ horsepower       : num  111 111 154 102 115 110 110 110 140 101 ...
 $ peak.rpm         : num  5000 5000 5000 5500 5500 5500 5500 5500 5500 5800 ...
 $ city.mpg         : int  21 21 19 24 18 19 19 19 17 23 ...
 $ highway.mpg      : int  27 27 26 30 22 25 25 25 20 29 ...
 $ price            : num  13495 16500 16500 13950 17450 ...
>
```

Struktura danych *automobil_clean*

4.3 Wstępna statystyka opisowa

R pozwala na podsumowanie danego zbioru danych za pomocą funkcji *summary*[10]. Dla kolumn z zawartością liczbową wypisywane są minimum, maksimum, średnia, mediana i wartości pierwszego i trzeciego kwartyla. By to wykonać na początku tworzę nową zmienną *auto_summary*, która zawiera wektor kolumn o typie numerycznym.

Tabela 3 Statystyka opisowa dla danych za pomocą funkcji summary()

```
>   summary(auto_summary)
   wheel.base        length          width           height       curb.weight    engine.size
 Min.   : 86.6   Min.   :141.1   Min.   :60.30   Min.   :47.80   Min.   :1488   Min.   : 61.0
 1st Qu.: 94.5   1st Qu.:166.3   1st Qu.:64.05   1st Qu.:52.00   1st Qu.:2145   1st Qu.: 98.0
 Median : 97.0   Median :173.2   Median :65.40   Median :54.10   Median :2414   Median :120.0
 Mean   : 98.9   Mean   :174.3   Mean   :65.89   Mean   :53.86   Mean   :2559   Mean   :127.9
 3rd Qu.:102.4   3rd Qu.:184.1   3rd Qu.:66.90   3rd Qu.:55.65   3rd Qu.:2944   3rd Qu.:145.5
 Max.   :120.9   Max.   :208.1   Max.   :72.00   Max.   :59.80   Max.   :4066   Max.   :326.0
      bore           stroke     compression.ratio   horsepower       peak.rpm       city.mpg
 Min.   :2.540   Min.   :2.07   Min.   : 7.00     Min.   : 48.0   Min.   :4150   Min.   :13.00
 1st Qu.:3.150   1st Qu.:3.11   1st Qu.: 8.50     1st Qu.: 70.0   1st Qu.:4800   1st Qu.:19.50
 Median :3.310   Median :3.29   Median : 9.00     Median : 95.0   Median :5100   Median :25.00
 Mean   :3.329   Mean   :3.25   Mean   :10.19     Mean   :103.3   Mean   :5099   Mean   :25.37
 3rd Qu.:3.590   3rd Qu.:3.41   3rd Qu.: 9.40     3rd Qu.:116.0   3rd Qu.:5500   3rd Qu.:30.00
 Max.   :3.940   Max.   :4.17   Max.   :23.00     Max.   :262.0   Max.   :6600   Max.   :49.00
  highway.mpg        price
 Min.   :16.00   Min.   : 5118
 1st Qu.:25.00   1st Qu.: 7756
 Median :30.00   Median :10245
 Mean   :30.84   Mean   :13248
 3rd Qu.:35.00   3rd Qu.:16509
 Max.   :54.00   Max.   :45400
>
```

Zastosowanie funkcji *summary()* na zbiorze.

Sprawdzam również odchylenie standardowe danych.

[10] Programowanie w języku R, Marek Gągolewski, str. 51, 107

Tabela 4 Odchylenie standardowe dla poszczególnych danych

```
Odchylenie standardowe danych
>   lapply(auto_summary, sd)
$wheel.base
[1] 6.132038

$length
[1] 12.47644

$width
[1] 2.132484

$height
[1] 2.396778

$curb.weight
[1] 524.7158

$engine.size
[1] 41.43392

$bore
[1] 0.2718657

$stroke
[1] 0.3141145

$compression.ratio
[1] 4.062109

$horsepower
[1] 37.86973

$peak.rpm
[1] 468.2714

$city.mpg
[1] 6.401382

$highway.mpg
[1] 6.829315

$price
[1] 8056.33
```

Z wyniku wstępnie można zauważyć, że rozkłady są skośne, o czym może sugerować różna wartość mediany i średniej. Jeśli mediana jest większa od średniej możemy domniemywać, że dany rozkład jest lewostronny – natomiast – jeśli mediana jest mniejsza od średniej przyjmujemy, iż rozkład jest prawostronny. W przypadku, gdy wartości są sobie równe mamy do czynienia z rozkładem symetrycznym. Jest to najprostszy sposób oceny skośności, jednak warto dokonać również dodatkowej analizy w tym kierunku.

Tabela 5 Współczynnik skośności dla danych

```
>   library (moments)
>   skewness(auto_summary)
      wheel.base           length            width           height       curb.weight
       0.97769934       0.13918738       0.86070637       0.03326970       0.67353728
      engine.size             bore           stroke compression.ratio       horsepower
       2.01225011      -0.01876304      -0.75585204       2.51251765       1.13462778
        peak.rpm         city.mpg      highway.mpg            price
       0.09591257       0.65965016       0.52025317       1.76318816
>
```

Za pomocą funkcji *skewness* możemy sprawdzić, że intuicja nas nie zawiodła.

Wyniki interpretujemy w następujący sposób:

-dane symetryczne: wartości od -0,5 do 0,5

-dane z umiarkowaną skośnością: wartości od -1 do -0,5 lub od 0,5 do 1

-dane mocno skośne: wartości mniejsze niż -1 lub większe niż 1

4.4 Czy większość aut ma silniki wolnossące (std)? Czy wiekszość samochodów napędzana jest dieslem? Jakich marek samochodowych jest najwięcej? Które marki samochodów są najdroższe?

By uzyskać informację na temat ilości aut danych z silnikami wolnossącymi (std) oraz z turbosprężarkami (turbo) z podziałem na marki oraz typ paliwa tworzę tabelę częstotliwości.

Tabela 6 Zmienne "fuel.type" z podziałem na "aspiration"

	diesel	gas
alfa-romero	0	3
audi	0	5
bmw	0	8
chevrolet	0	3
dodge	0	6
honda	0	13
isuzu	0	2
jaguar	0	3
mazda	2	11
mercedes-benz	0	4
mercury	0	0
mitsubishi	0	7
nissan	1	16
peugot	0	5
plymouth	0	5
porsche	0	4
saab	0	4
subaru	0	10
toyota	2	29
volkswagen	2	8
volvo	0	6

	diesel	gas
alfa-romero	0	0
audi	0	1
bmw	0	0
chevrolet	0	0
dodge	0	3
honda	0	0
isuzu	0	0
jaguar	0	0
mazda	0	0
mercedes-benz	4	0
mercury	0	1
mitsubishi	0	6
nissan	0	1
peugot	5	1
plymouth	0	2
porsche	0	0
saab	0	2
subaru	0	2
toyota	1	0
volkswagen	2	0
volvo	1	4

Tabela 7 Ilość aut z silnikami std i turbo

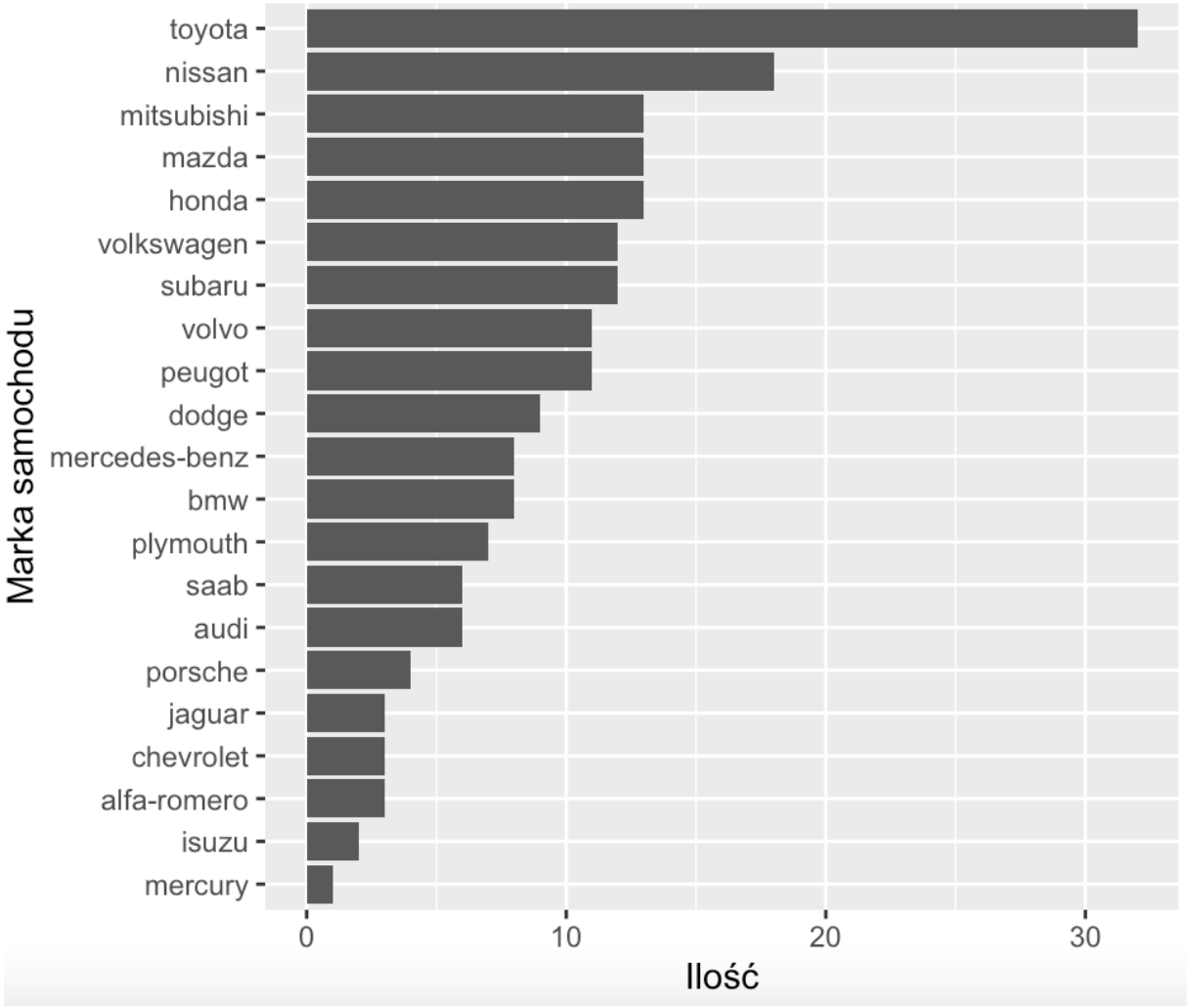
```
> table(automobil_clean$aspiration)

 std turbo
 159    36
```

Tabela 8 Ilość aut napędzanych na diesel i gaz

```
> table(automobil_clean$fuel.type)

diesel    gas
    20    175
```

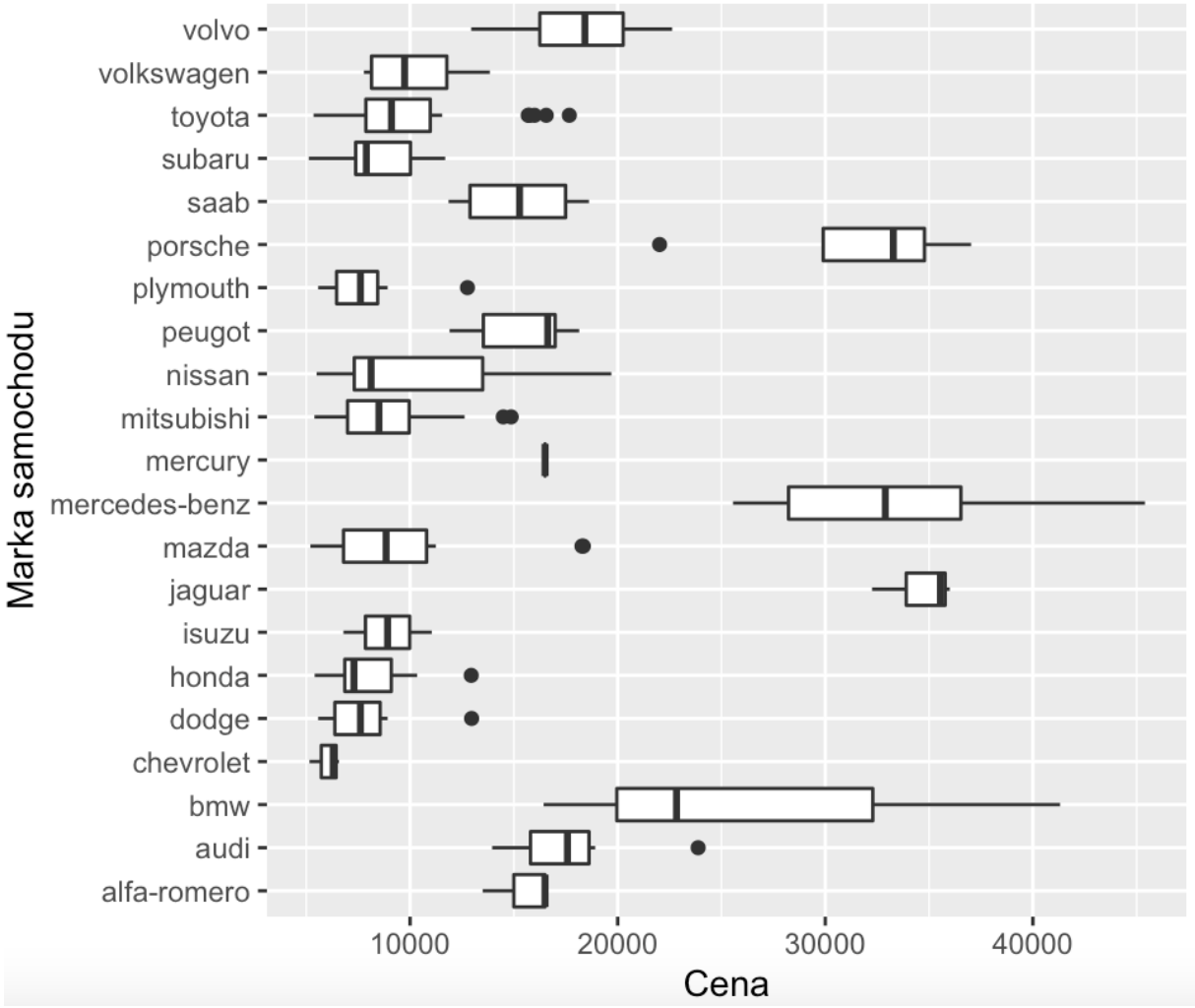
Dowiadujemy się z niej, że przeważająca ilość aut wyposażona jest w silniki std oraz napędzana jest gazem.

Poniżej przedstawiam zaś wykres obrazujący ilość samochodów z podziałem na ich markę z segregacją na od najbardziej do najmniej popularnych. Jak możemy zauważyć w datasecie znajduje się najwięcej samochodów marki toyota.



Rysunek 1 Ilość poszczególnych marek samochodowych

Z kolei na poniższym wykresie możemy zauważyć, że najdroższymi markami są: mercedes-benz, jaguar oraz porsche.



Rysunek 2 Wykresy pudełkowe dla cen poszczególnych marek samochodowych

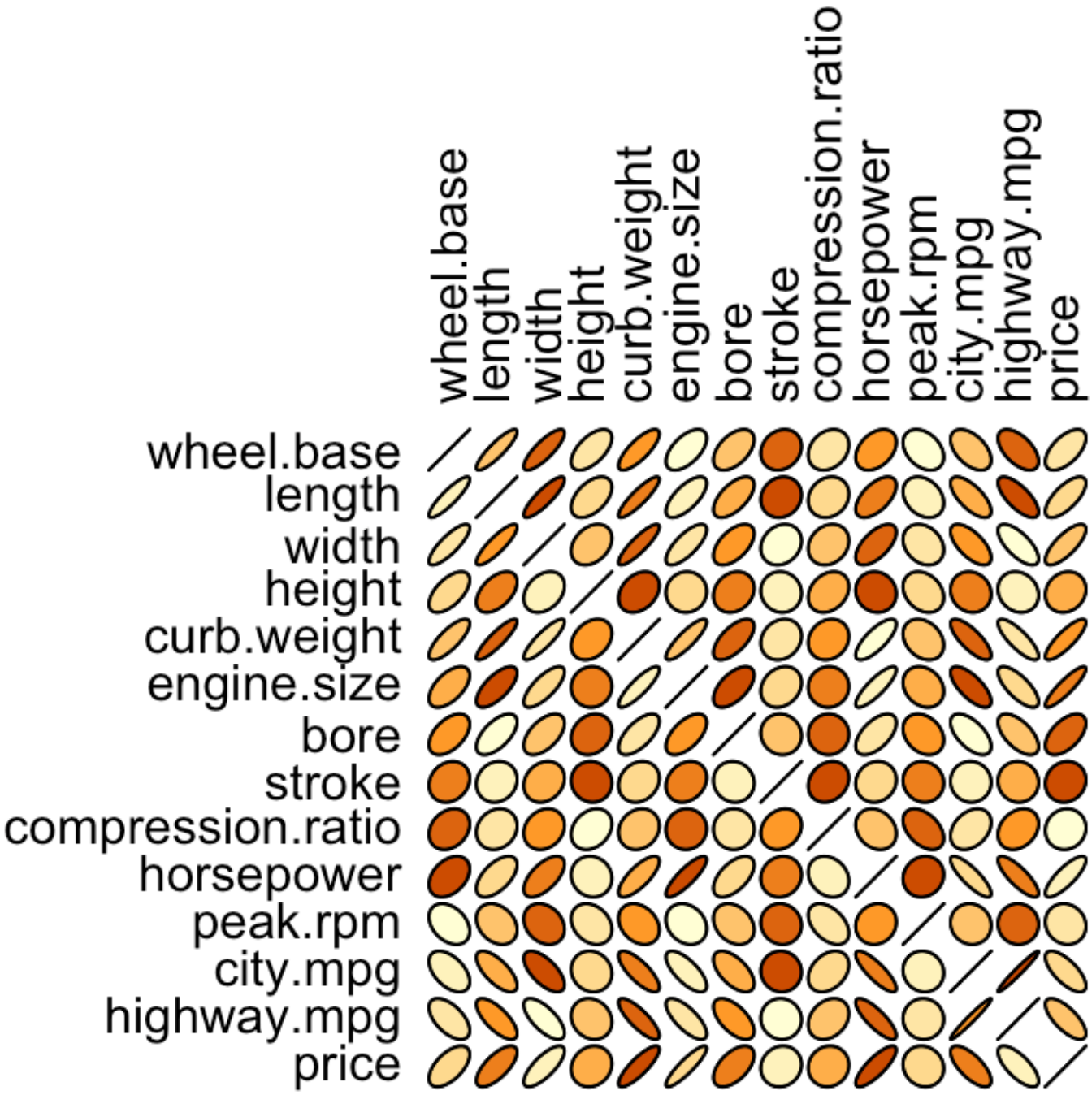
4.5 Jak skorelowane są dane? Czy wielkość silnika jest powiązana z wielkością auta?

Tabela 9 Tabela korelacji danych

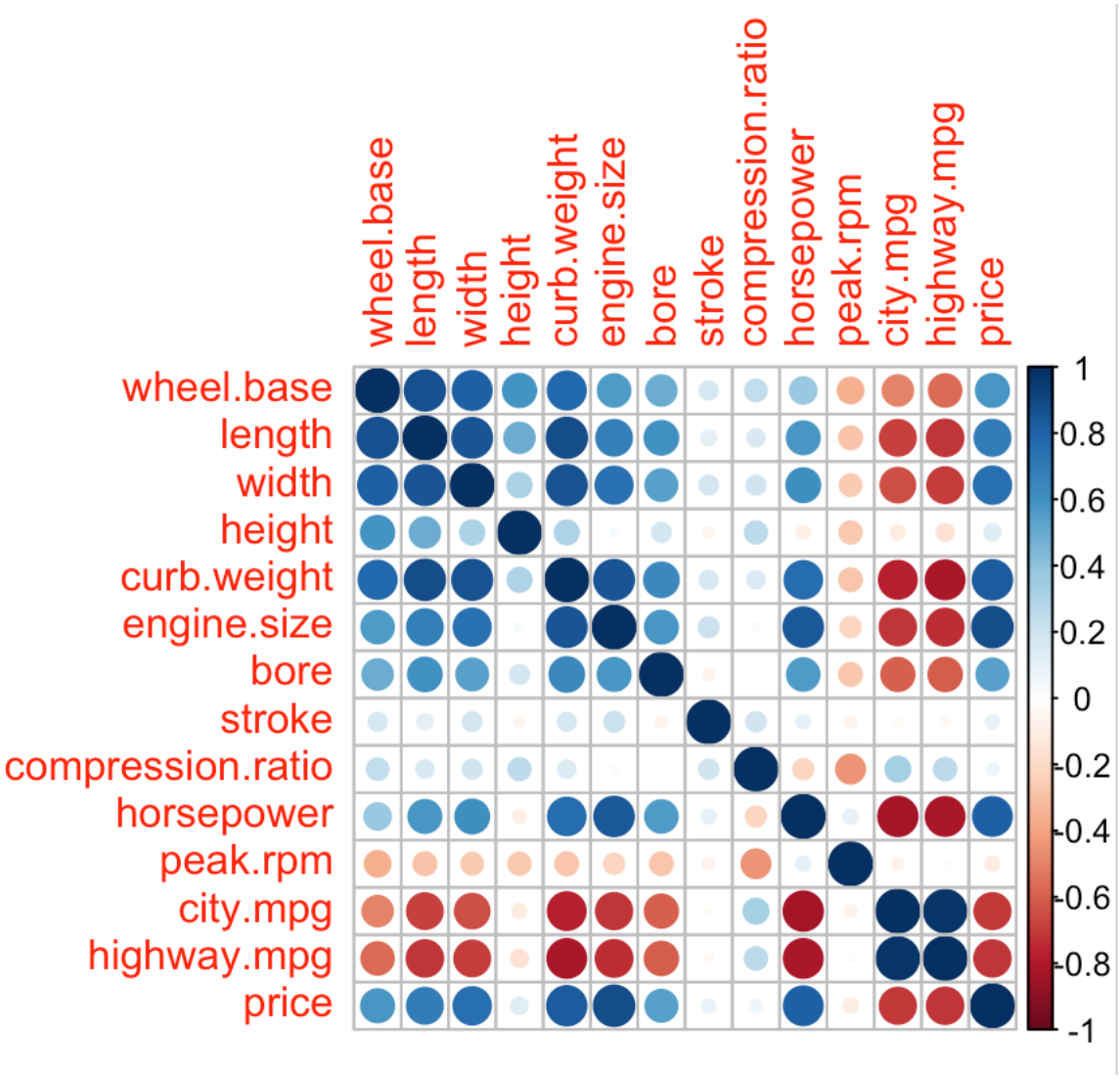
```
>   korelacja
                  wheel.base     length      width      height curb.weight engine.size         bore
wheel.base         1.0000000  0.8792218  0.8190086  0.59250037   0.7827202  0.56970408  0.498227517
length             0.8792218  1.0000000  0.8580838  0.49621812   0.8816650  0.68747862  0.609436758
width              0.8190086  0.8580838  1.0000000  0.31583448   0.8673146  0.74031969  0.544310504
height             0.5925004  0.4962181  0.3158345  1.00000000   0.3077318  0.03128552  0.189282865
curb.weight        0.7827202  0.8816650  0.8673146  0.30773178   1.0000000  0.85757324  0.645806034
engine.size        0.5697041  0.6874786  0.7403197  0.03128552   0.8575732  1.00000000  0.583091333
bore               0.4982275  0.6094368  0.5443105  0.18928286   0.6458060  0.58309133  1.000000000
stroke             0.1717218  0.1186643  0.1864323 -0.05552478   0.1727852  0.21198929 -0.066793158
compression.ratio  0.2477298  0.1601722  0.1909973  0.26115993   0.1553818  0.02461689  0.003056787
horsepower         0.3755413  0.5838128  0.6167786 -0.08441172   0.7602852  0.84269102  0.568527205
peak.rpm          -0.3523307 -0.2809857 -0.2516270 -0.26407787  -0.2789441 -0.21900779 -0.277661680
city.mpg          -0.4991263 -0.6896598 -0.6470992 -0.10236659  -0.7721709 -0.71062448 -0.591950375
highway.mpg       -0.5663546 -0.7193238 -0.6922199 -0.15118826  -0.8127097 -0.73213800 -0.600039574
price              0.5857928  0.6953308  0.7542734  0.13829069   0.8357293  0.88894226  0.546872917
                      stroke compression.ratio  horsepower    peak.rpm    city.mpg highway.mpg
wheel.base         0.17172176       0.247729806  0.37554127 -0.35233072 -0.49912630 -0.56635460
length             0.11866435       0.160172167  0.58381279 -0.28098572 -0.68965978 -0.71932378
width              0.18643226       0.190997343  0.61677860 -0.25162703 -0.64709916 -0.69221989
height            -0.05552478       0.261159933 -0.08441172 -0.26407787 -0.10236659 -0.15118826
curb.weight        0.17278521       0.155381807  0.76028522 -0.27894406 -0.77217086 -0.81270968
engine.size        0.21198929       0.024616889  0.84269102 -0.21900779 -0.71062448 -0.73213800
bore              -0.06679316       0.003056787  0.56852720 -0.27766168 -0.59195038 -0.60003957
stroke             1.00000000       0.199881893  0.10003999 -0.06829951 -0.02764104 -0.03645288
compression.ratio  0.19988189       1.000000000 -0.21440100 -0.44458194  0.33141295  0.26794095
horsepower         0.10003999      -0.214401004  1.00000000  0.10565391 -0.83411654 -0.81291687
peak.rpm          -0.06829951      -0.444581936  0.10565391  1.00000000 -0.06949336 -0.01695001
city.mpg          -0.02764104       0.331412952 -0.83411654 -0.06949336  1.00000000  0.97234992
highway.mpg       -0.03645288       0.267940954 -0.81291687 -0.01695001  0.97234992  1.00000000
price              0.09374644       0.069500205  0.81102684 -0.10433340 -0.70268485 -0.71558976
                      price
wheel.base         0.58579283
length             0.69533083
width              0.75427339
height             0.13829069
curb.weight        0.83572934
engine.size        0.88894226
bore               0.54687292
stroke             0.09374644
compression.ratio  0.06950020
horsepower         0.81102684
peak.rpm          -0.10433340
city.mpg          -0.70268485
highway.mpg       -0.71558976
price              1.00000000
```

Z powyższej tabeli możemy odczytać, jak skorelowane są ze sobą zmienne. Przedstawia ona wiele informacji, jednak w dość nieczytelnej formie. Poniżej przedstawiam wykres, który w bardziej widoczny sposób przedstawia powyższe badanie.

Na poniższym wykresie korelacja = 1 jest przedstawiona jako linia prosta. Korelacja zerowa daje w wyniku koło. Intensywność koloru użytego na wykresie wskazuje na wielkość korelacji. Dodatkowo orientacja elipsy służy do podkreślenia dodatniej lub ujemnej wartości korelacji. W przypadku korelacji dodatniej elipsa przechyla się w prawo, a dla korelacji ujemnej jest odwrotnie.



Rysunek 3Wykres korelacji - pierwsza metoda wizualizacji



Rysunek 4 Wykres korelacji - druga metoda wizualizacji

Powyższy wykres jest innym sposobem przedstawienia wyników korelacji.

Na oby dwóch wykresach z łatwością możemy odczytać, że niektóre pary zmiennych są słabo skorelowane – na przykład rozstaw osi (*wheel.base*) i moc (*horsepower*) - podczas gdy niektóre są skorelowane dużo mocniej - na przykład wielkość silnika (*engine.size*) i masa własna pojazdu (*curb.weight*).

Co więcej, na poniższym teście korelacji dla zmiennych *engine.size* oraz *curb.weight* możemy zauważyć, że p wartość wynosi 2.2e-16, czyli mniej niż poziom istotności alfa = 0,05. Odrzucamy zatem hipotezę zerową mówiącą o tym, że zmienne nie są skorelowane. Możemy zatem przyjąć hipotezę alternatywną i wywnioskować, że zmienne *curb.weight* oraz *engine.size* są istotnie skorelowane ze współczynnikiem korelacji równym -0.8575732 i wartością p równą

2.2e-16. Jest to zależność dodatnia, co oznacza, że im większy silnik, tym większa jest masa własna pojazdu.[11]

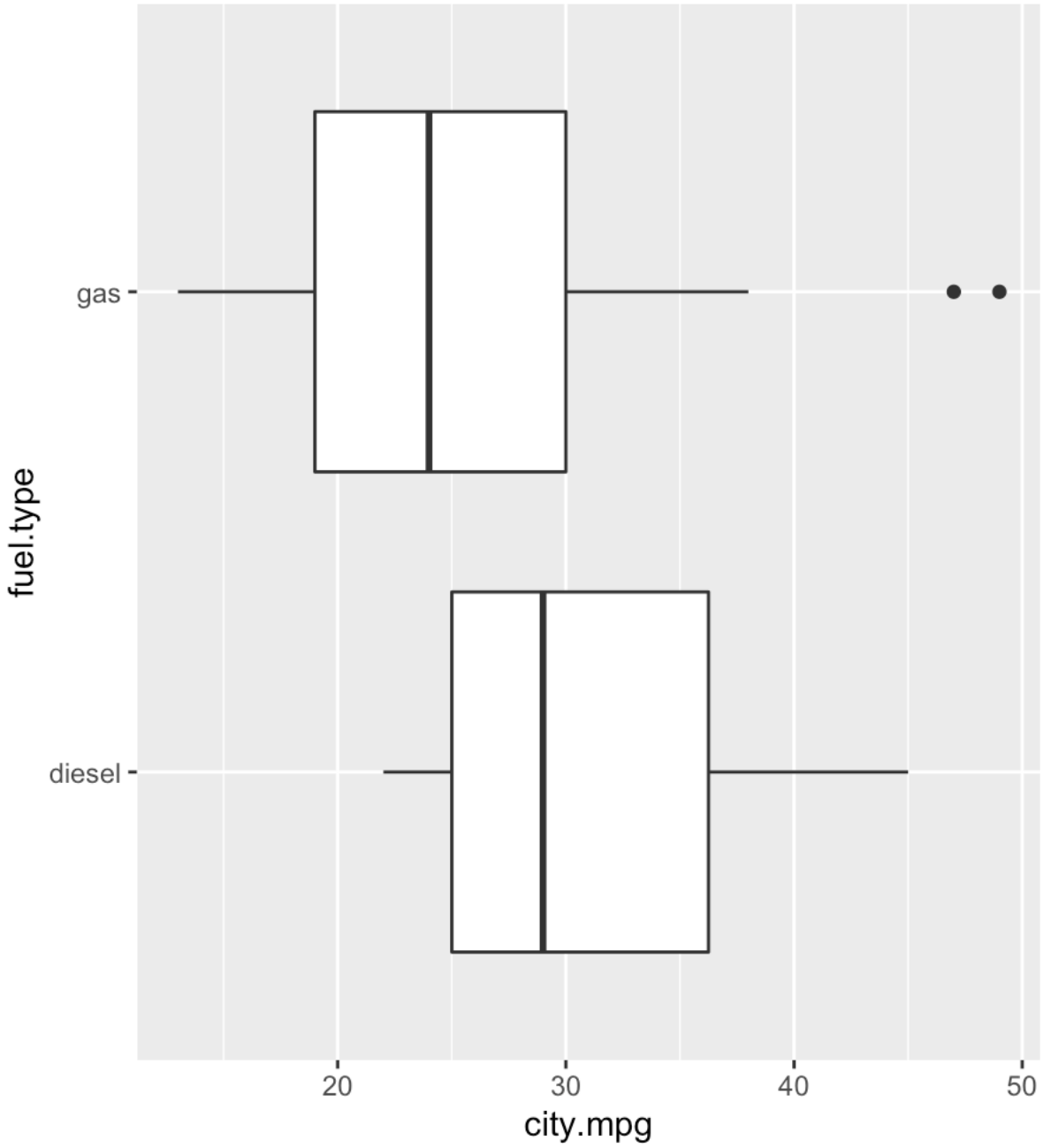
```
        Pearson's product-moment correlation

data:  automobil_clean$engine.size and automobil_clean$curb.weight
t = 23.162, df = 193, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8153048 0.8907505
sample estimates:
      cor
0.8575732
```

Rysunek 5 Test korelacji Pearsona

4.6 Czy auta napędzane LPG odznaczają się mniejszym spalaniem w mieście niż pojazdy napędzane dieslem?

Aby wstępnie zapoznać się ze wynikami reprezentowanymi przez obie zmienne, przedtawię je poniżej na dwóch wykresach pudełkowych zestawionych obok siebie.

[11] Statistics in Medicine (Third Edition), R.H. Riffenburgh, 2012



Rysunek 6 Wykresy pudełkowe zmiennej city.mpg w zależności od zmiennej fuel.type

Na podstawie tej wizualizacji możemy sformułować jasną hipotezę: wydaje się, że samochody napędzane gazem odznaczają się mniejszą liczbą mil na galon, a tym mniej efektywnie zużywają paliwo (zużywają go więcej na tym samym dystansie), niż samochody napędzane dieslem. Jest jednak możliwe, że ta pozorna zależność zdarzyła się przypadkowo - to znaczy, że po prostu przypadkowo wybraliśmy grupę samochodów napędzanych LPG o niskiej wydajności i grupę samochodów dieslowych o wyższej wydajności. Aby sprawdzić, czy

tak jest, należy użyć testu statystycznego – na potrzeby tej hipotezy posłużę się testem t-Studenta.

Na początku sprawdzam zatem założenia testu, które powinny zostać spełnione:

1. Rozkład w grupach zbliżony jest do rozkładu normalnego.

Do sprawdzenia tego założenia możemy się posłużyć testem Shapiro – Wilka. Co ważne - w przypadku, gdy liczebność próby przekracza 50, test jest odporny na niespełnienie tego założenia.

```
>   shapiro.test(automobil_clean$city.mpg)

        Shapiro-Wilk normality test

data:  automobil_clean$city.mpg
W = 0.95854, p-value = 1.762e-05
```

Rysunek 7 Test Shapiro-Wilka dla zmiennej city.mpg

Wynik testu dla zmiennej *city.mpg*. Jeśli test Shapiro-Wilka osiąga istotność statystyczną ($p < 0,05$), świadczy to o rozkładzie oddalonym od krzywej Gaussa.

W grupie fuel.type reprezentuje zmienną *„gas”* jako 1 zaś zmienną *„diesel”* jako 0. Zmiany zapisuje w tabeli *gaz_diesel_0_1*. Następnie jak powyżej wykonuję test Shapiro-Wilka.

```
>   shapiro.test(gaz_diesel_0_1$fuel.type)

        Shapiro-Wilk normality test

data:  gaz_diesel_0_1$fuel.type
W = 0.34703, p-value < 2.2e-16
```

Rysunek 8 Test Shapiro-Wilka dla zmiennej fuel.type

Jak możemy zauważyć założenie o normalności rozkładu jest niespełnione w obydwóch grupach, ponieważ p value < 0.5, co wskazuje na odrzucenie hipotezy zerowej.

Jak wspomniałam, warto jednak zwrócić uwagę na fakt, iż test ten jest dość odporny na złamanie tego założenia i w praktyce, niektórzy analitycy nie biorą złamanie tego założenia pod uwagę.

2. Porównywane grupy mają podobną liczebność – jak możemy zauważyć z tabeli częstotliwości założenie to nie jest spełnione.
3. Wariancje w porównywanych grupach są do siebie podobne – homogeniczność wariancji - aby sprawdzić te założenie stosuje się dodatkowy test, przykładowo test Levene'a. Jest to jednak założenie klasycznego testu Studenta. Należy zauważyć, że używany przeze mnie test *t* Welcha (test statystyczny równości wartości oczekiwanych w dwóch populacjach) jest uogólnieniem testu *t* Studenta na populacje o różnych wariancjach. Stanowi przybliżone rozwiązanie problemu Behrensa-Fishera. Mimo wszystko dokonam testu wariancji pod kątem jej homogeniczności.

```
>   leveneTest(city.mpg ~ fuel.type, data = automobil_clean)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.5362 0.4649
      193
```

Rysunek 9 Test Levene

F value >0.05, co wskazuje na brak różnicy między wariancjami grup.

Gdyby kryteria 1. oraz 2. zostały spełnione, wyniki testu prezentowałyby się następująco:

```
>   t.test(city.mpg ~ fuel.type, data=automobil_clean)

        Welch Two Sample t-test

data:  city.mpg by fuel.type
t = 3.5423, df = 22.919, p-value = 0.001746
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.282669 8.694474
sample estimates:
mean in group diesel    mean in group gas
            30.30000             24.81143
```

Rysunek 10 Test Studenta dla zmiennej city.mpg oraz fuel.type

(~) oznacza w języku R „wyjaśnione przez" – możemy zatem interpretować ten zapis jako: „czy wydajność spalania zależy od zmiennej *fuel.type* w zbiorze *automobil_clean*?".

Wartość p wskazuje prawdopodobieństwo, czy ta pozorna różnica między dwiema grupami może pojawić się przypadkowo. W powyższym teście obserwujemy niską wartość p, więc możemy być dość pewni, że istnieje rzeczywista różnica między grupami (przyjmujemy *„alternative hipotesis: true difference in means is not equal to 0"*).

W wyniku widzimy również 95% przedział ufności. Ten przedział opisuje, o ile więcej metrów na galon jest w stanie przejechać samochód zasilany dieslem. Możemy się spodziewać, że z 95% prawdopodobieństwem wartość ta zmieści się w przedziale między ok. 2.28 m a 8.69 m.

Ze względu na wątpliwości co do kwestii spełnienia pierwszego i drugiego założenia wykonuję test nieparametryczny U-Manna-Whitney'a-Wilcoxona

```
>   wilcox.test(city.mpg ~ fuel.type, data=gaz_diesel_0_1)

        Wilcoxon rank sum test with continuity correction

data:  city.mpg by fuel.type
W = 2535, p-value = 0.0009855
alternative hypothesis: true location shift is not equal to 0
```

Rysunek 11 Test U-Manna-Whitney'a-Wilcoxona

Możemy ponownie zauważyć, że wartość p jest mniejsza niż 0,05. Na podstawie tego wyniku możemy stwierdzić, że dystrybuanty tych dwóch rozkładów różnią się. Hipoteza alternatywna jest dosłownie określona jako „ przesunięcie lokalizacji nie jest równe 0". Inaczej mówiąc - „rozmieszczenie jednej populacji jest przesunięte na lewo lub na prawo od drugiej" - co oznacza różne mediany. Spalanie w oby dwóch grupach nie jest takie samo.

Podsumowanie

Hipoteza zerowa zatem nie potwierdziła się. Analizując wykres pudełkowy oraz wyniki testu możemy przepuszczać, że auta napędzane LPG nie odznaczają się lepszą wydajnością spalania w mieście od aut napędzanych dieslem.

4.7 Czy ceny pojazdów podlegają rozkładowi Gaussa? Jaki ma to wpływ na dalsze badania? Czy logarytm zmiennej cena lepiej odwzorowuje rozkład normalny?

Ze statystyki opisowej przeprowadzonej na początku pracy wynika, iż dane w naszym zbiorze charakteryzują się mniejszą lub większą skośnością. Należy jednak zauważyć, że współczynnik skośności jest dość silnie podatny na szumy i nie świadczy całkowicie o prawdziwym rozkładzie zmiennych.

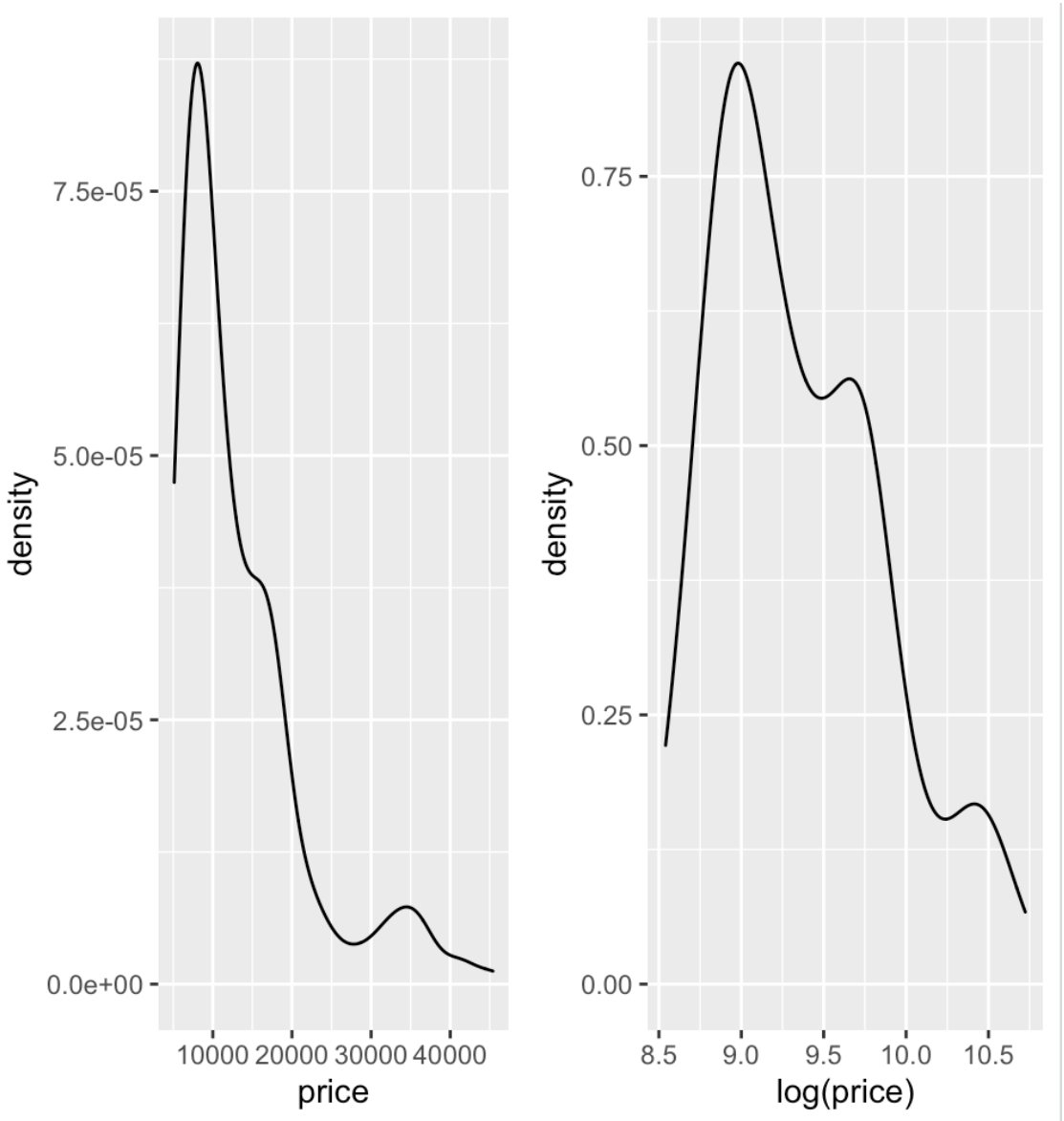
Chciałabym zatem sprawdzić, jak rzeczywiście kształtuje się rozkład zmiennej *price* i czy jest on zbliżony do rozkładu normalnego. Bazując na wstępnej statystyce formułuję następującą hipotezę zerową: Rozkład cen pojazdów nie podlega rozkładowi Gaussa.

Jak sprawdzić, czy hipoteza ta jest prawdziwa? Możemy to zrobić na wiele sposobów. Jednym z nich są analizy graficznie - do których używyjemy wykresów Q-Q - oraz testy numeryczne – takie jak test Kołmogorowa-Smirnowa dla rozkładów.

Zaczynam od ustalenia średniej, wariancji oraz odchylenia standardowego cen samochodów, a także dla logarytmu tych cen. Wyniki przedstawione są poniżej

```
Srednia cen =  13248.02 , wariancja cen =  64904454.57 , odchylenie standardowe cen =  8056.33>

Srednia logarytmu cen =  9.35 , wariancja logarytmu cen =  0.26 , odchylenie standardowe logarytmu cen =  0.51
```

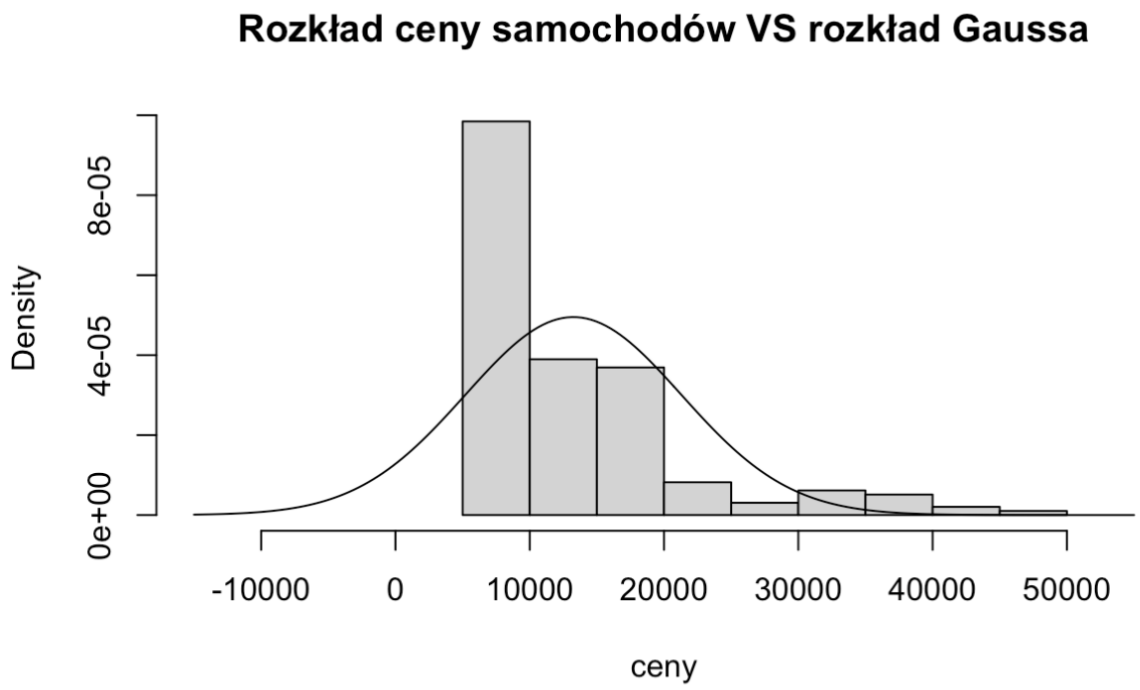
Na ich podstawie możliwa jest wizualizacja krzywej gęstości dla zmiennej *price* oraz *log(price).*

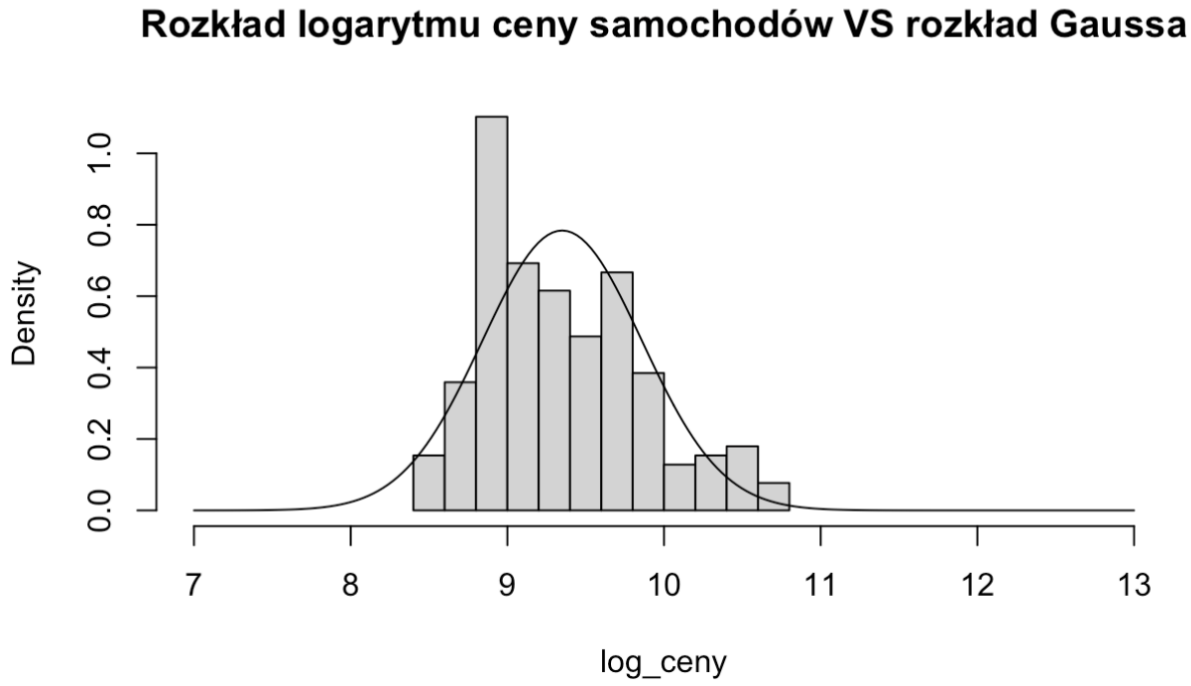


Rysunek 12 Krzywe gęstości dla zmiennych price oraz log(price)

Analizując te wykresy zwróćmy uwagę, że rozkład jest mocno pochylony w lewo. Potwierdza to dotychczasowe wnioski o skośności danych ze statystyki opisowej, która została przedstawiona powyżej. Na pierwszy rzut oka wydaje się, że ceny nie mają rozkładu normalnego. W dalszej części przyjrzymy się temu bliżej, jednak wcześniej chciałabym wytłumaczyć, czemu posłużyłam się w wykresie także logarytmem zmiennej *price.*

Rozkład logarytmicznie normalny jest często lepszym od rozkładu normalnego przybliżeniem rozkładów cech, w których istotne są stosunki pomiędzy wartościami, a nie różnice pomiędzy nimi. Na przykład przybliżony rozkład logarytmicznie normalny mają kursy akcji giełdowych, gdzie ważniejsze jest o ile procent zmniejszyła się lub zwiększyła wartość akcji, a nie o ile złotych. Jak możemy zauważyć w dokonanej powyżej statystyce *summary* minimalna wartość zmiennej *price* to 5118, zaś maksymalna 45400 - skala logarytmiczna umożliwia właśnie analizowanie danych z rozległego przedziału





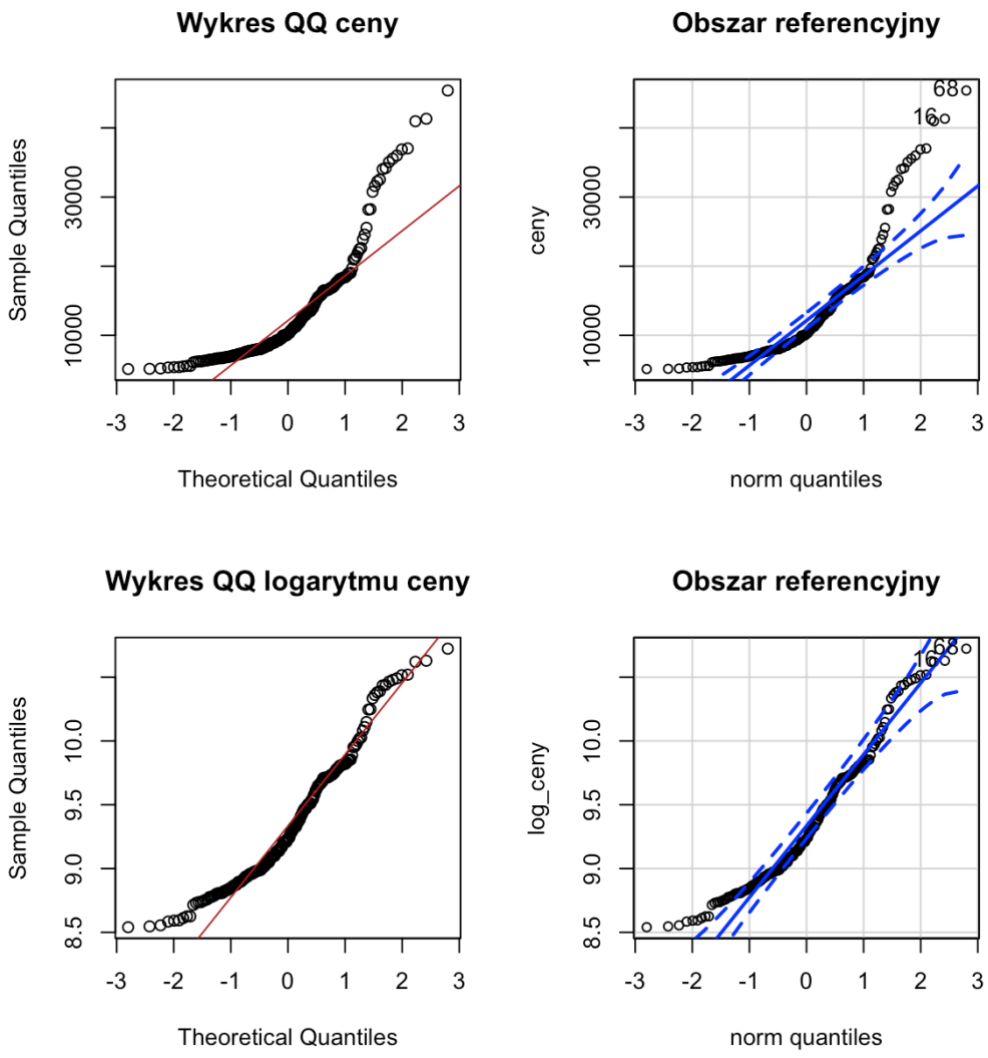
Rysunek 13 Rozkład cen samochodów oraz rozkład Gaussa dla zmiennej price oraz dla logarytmu zmiennej price

Powyżej znajduje się histogram zmiennej *price* i *log (price)* z odpowiadającymi im rozkładami normalnymi. Możemy zauważyć, że *log (price)* bardziej przypomina swoim kształtem rozkład normalny rozkład niż wykres pierwszy.

Wykresy QQ

Oczywiście istnieją również inne metody sprawdzenia, czy dane mogą pochodzić z rozkładu normalnego. Jedną z najczęściej używanych technik jest wykres QQ (qqplot), czyli wykres kwantyl kwantyl. Technika ta jest bardzo ściśle związana z testem Shapiro-Wilka. Interpretacja takiego wykresu jest następująca: jeżeli punkty wykresu leżą blisko prostej i są równomiernie rozłożone po jej jednej i drugiej stronie (np. naprzemiennie), to dane pochodzą z rozkładu normalnego.

Wizualny test przy pomocy wykresu QQ jest często podstawą do przyjęcia normalności jakiegoś zbioru. Stać się tak może na przykład gdy dane są mocno zdyskretyzowane (np. zaokrąglone).



Rysunek 14 Wykresy QQ

Ponieważ wykres Q-Q w górnej części powyższej grafiki wyraźnie nie jest linią prostą, zmienna *price* najprawdopodobniej nie jest zgodna z rozkładem normalnym. Chociaż wykres w dolnej częsci wydaje się mniej zakrzywiony niż wykres Q-Q na górze, nadal trudno powiedzieć, czy zmienna będąca logarytmem ceny charakteryzuje się rozkładem normalnym czy też nie. W dalszych krokach przeprowadzę zatem test Kolmogorova-Smirnova.

Test Kołmogorowa-Smirnowa (Kolmogorov-Smirnov test)

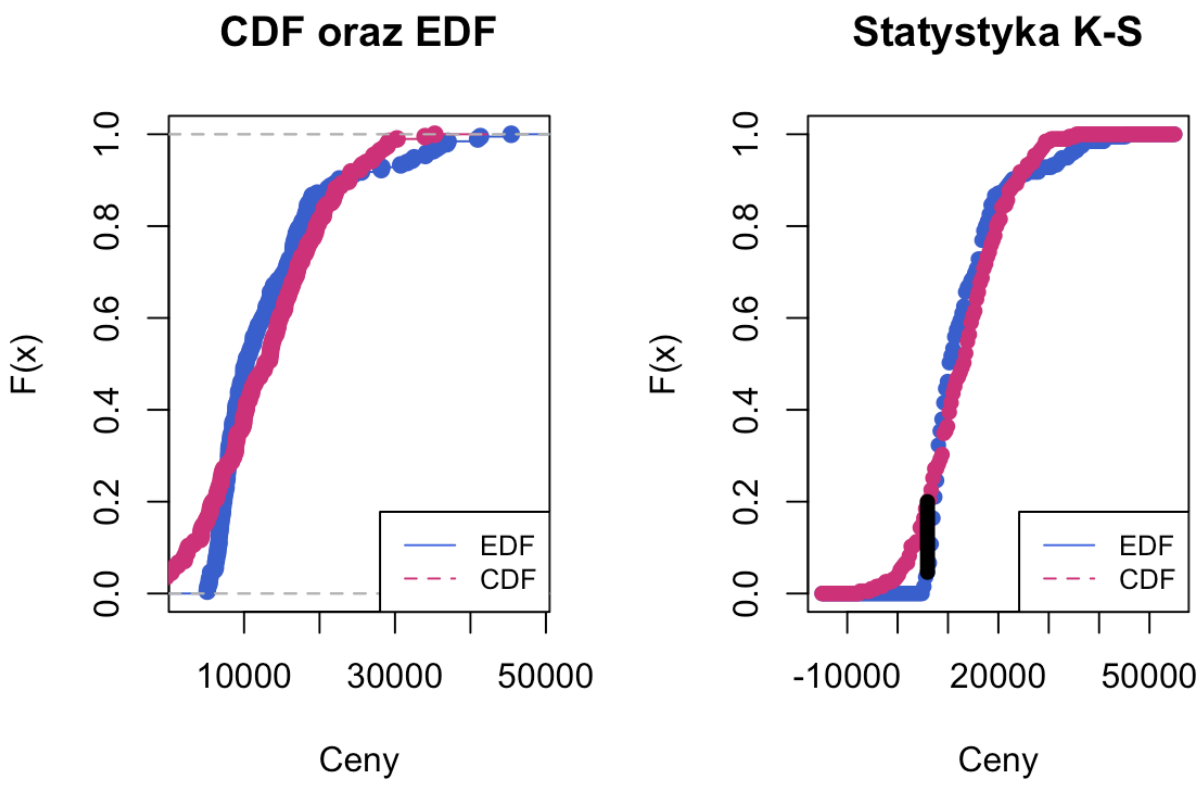
W 1930s Kołmogorow i Smirnow opracowali test zgodności dla ciągłych danych służący do testowania, czy próbka pochodzi z danego rozkładu prawdopodobieństwa opisanego hipotezą zerową. Dziś nadal jest on jednym z najbardziej znanych i najczęściej używanych testów zgodności. Wynika to z jego prostoty i dlatego, że opiera się na dystrybuancie

empirycznej (empirical distribution function EDF), która jest zbieżna do skumulowanego rozkładu populacji (cumulative distribution function CDF) (twierdzenie Glivenko-Cantelli). Chociaż mnóstwo różnych testów zgodności zostało opracowane w ostatnich dziesięcioleciach (patrz na przykład, D'Agostino i Stephens 1986), wiele z większa moc statystyczna niż test Kołmogorowa - Smirnowa (KS), to KS pozostaje popularny, ponieważ jest prosty i intuicyjny. Celem testu Kołmogorowa - Smirnowa jest sprawdzenie czy próba zbudowana ze zmiennych losowych pochodzi z określonego rozkładu. Dlatego też hipoteza zerowa musi określać zarówno typ rozkładu jak i jego parametry. Hipoteza alternatywna zakłada, że funkcja rozkładu prawdopodobieństwa nie pasuje do tej opisanej w hipotezie zerowej. Idea testu Kołmogorowa-Smirnowa jest dość prosta: maksymalna różnica między przyjętym CDF i EDF próby losowej jest używana do decydowania, czy losowa należy do zadanego rozkładu czy nie. Dla pojedynczej próbki danych test Kołmogorowa-Smirnowa stosuje się w celu sprawdzenia, czy próbka danych jest czy nie jest zgodna z określonym rozkładem. Jeśli istnieją dwie próbki danych, to można sprawdzać, czy te dwie próbki pochodzą z tego samego rozkładu prawdopodobieństwa.[12]

Sprawdzam zatem, ile wynosi maksymalna odległość pomiędzy wspomnianymi dystrybuantami, a wynik obliczeń zaprezentowany jest poniżej:

```
Maksymalna różnica między dystrybuantą empiryczną a teoretyczną 0.17948717948718
```

[12] https://algolytics.com/



Rysunek 15 Dystrybuanty empiryczne oraz teoretyczne zmiennej price

Graficznie odległość ta przedstawiona jest powyżej za pomocą pogrubionej linii na wykresie z prawej strony. Wykres z lewej strony został zestawiony dla porównania.

Ostatecznym zwieńczeniem moich dociekań będzie wykonanie *ks.test()* przy 95% ufności, zakładając następujące hipotezy:

H0 – próbka pochodzi z rozkładu normalnego

H1 – próbka nie pochodzi z rozkładu normalnego

```
        Two-sample Kolmogorov-Smirnov test

data:  ceny and rozkl_norm_ceny
D = 0.17949, p-value = 0.003739
alternative hypothesis: two-sided
```

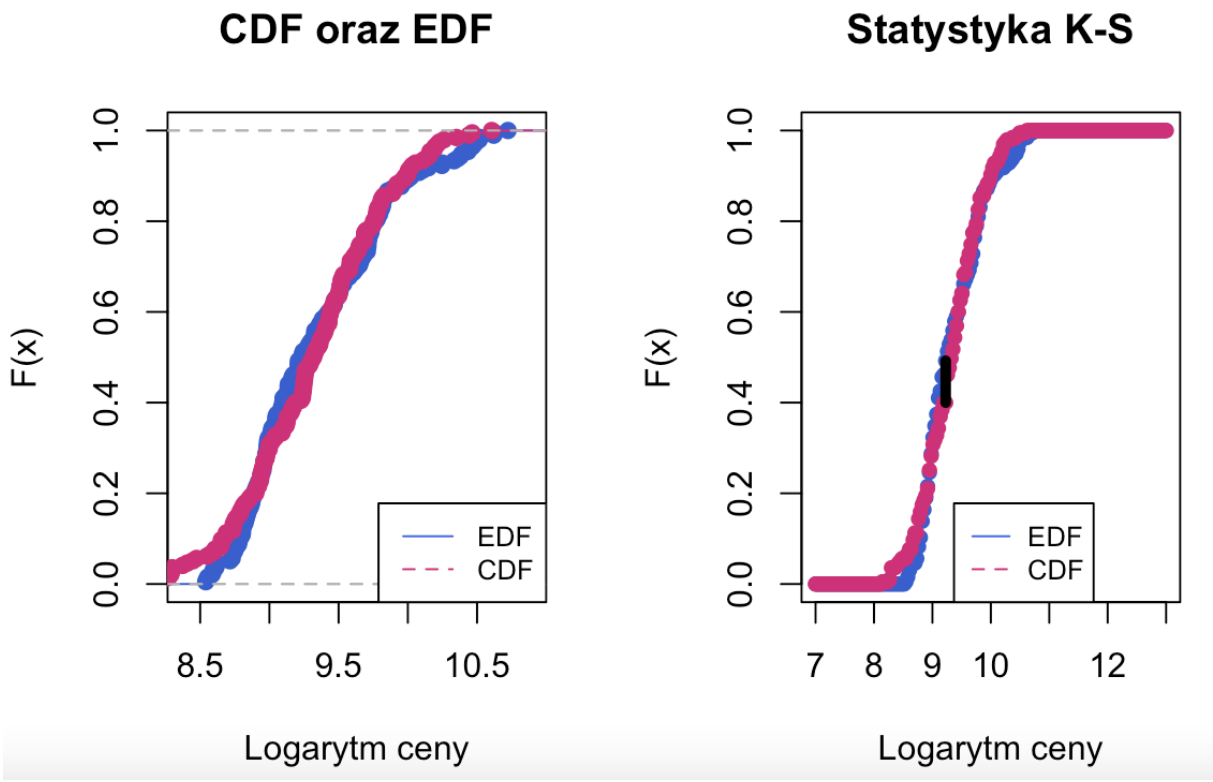
Rysunek 16 Test K-S dla zmiennej price

Jak możemy zauważyć p-value <0.05, więc odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej – analizowana próbka (*price)* nie pochodzi z rozkładu normalnego.

Warto zadać także pytanie, jak wypadnie test dla logarytmu cen. Dotychczasowa analiza wskazywała, że jego rozkład był nieco bardziej zbliżony do normalnego. Zacznę od dystrybuanty i statystyki KS:

```
Maksymalna różnica między dystrybuantą empiryczną a teoretyczną logarytmu cen 0.0923076923076923
```

Odległość ta oznaczona jest na poniższym wykresie czarną linią. Z lewej strony dla porównania wykres bez naniesionej statystyki.



Rysunek 17 Dystrybuanty empiryczne oraz teoretyczne dla logarytmu zmiennej price

Poniżej zaprezentuję również wyniki testu K-S o 95% ufności,, przyjmując hipotezy jak w poprzednim badaniu:

H0 – próbka pochodzi z rozkładu normalnego

H1 – próbka nie pochodzi z rozkładu normalnego

```
        Two-sample Kolmogorov-Smirnov test

data:  log_ceny and log_rozkl_norm_ceny
D = 0.097436, p-value = 0.3129
alternative hypothesis: two-sided
```

Rysunek 18 Test K-S dla logarytmu zmiennej price

Jak możemy zauważyć p-value >0.05, więc przyjmujemy hipotezę zerową –istnieje prawdopodobieństwo, że analizowana próbka będąca logarytmem zmiennej *price* pochodzi z rozkładu normalnego.

Należy jednak podkreślić, że test K-S jest raczej ogólny, ponieważ może być stosowany do testowania dowolnego rozkładu. Oznacza to, że moc tego testu jest ograniczona. Ostatnimi czasy statystycy podkreślają fakt, że test Shapiro-Wilka ma większą moc niż test Kołmogorowa-Smirnowa. Jednakże zarówno jeden i drugi może być wykorzystywany przy określaniu podobieństwa rozkładu zmiennej do rozkładu normalnego.

Test Shapiro-Wilka dla zmiennej *ceny*:

```
        Shapiro-Wilk normality test

data:  ceny
W = 0.79939, p-value = 4.265e-15
```

Rysunek 19 Test Shapiro-Wilka dla zmiennej price

Test Shapiro-Wilka dla zmiennej *log_ceny*:

```
        Shapiro-Wilk normality test

data:  log_ceny
W = 0.9464, p-value = 1.141e-06
```

Rysunek 20 Test Shapiro-Wilka dla logarytmu zmiennej price

W obu przypadkach – *ceny* jak i *log_ceny* –wartość p jest istotnie <0,05, zatem test Shapiro-Wilka wskazuje na odrzucenie hipotezy zerowej dotyczącej normalności rozkładów

Podsumowując – w dotychczasowej analizie rozkładu zmiennej *price* nie udało mi się przedstawić dowodów na to, że pochodzi ona z rozkładu Gaussa. Można jednak domniemywać, na bazie wykresu QQ oraz testu K-S, iż logarytm zmiennej *price* w pewnym stopniu odwzorowuje rozkład normalny. Celem transformacji cen na postać logarytmiczną było właśnie otrzymanie danych przypominających rozkład normalny oraz redukcja skośności. Jako, że założenie to zostało w pewnym stopniu zrealizowane (a przynajmniej pewne testy to potwierdzają takie prawdopodobieństwo), w dalszej części badań będę posługiwać się właśnie zlogarytmizowaną postacią zmiennej *price*.

4.8 Czy ceny aut z silnikami turbo są większe od cen aut z silnikami wolnossącymi?

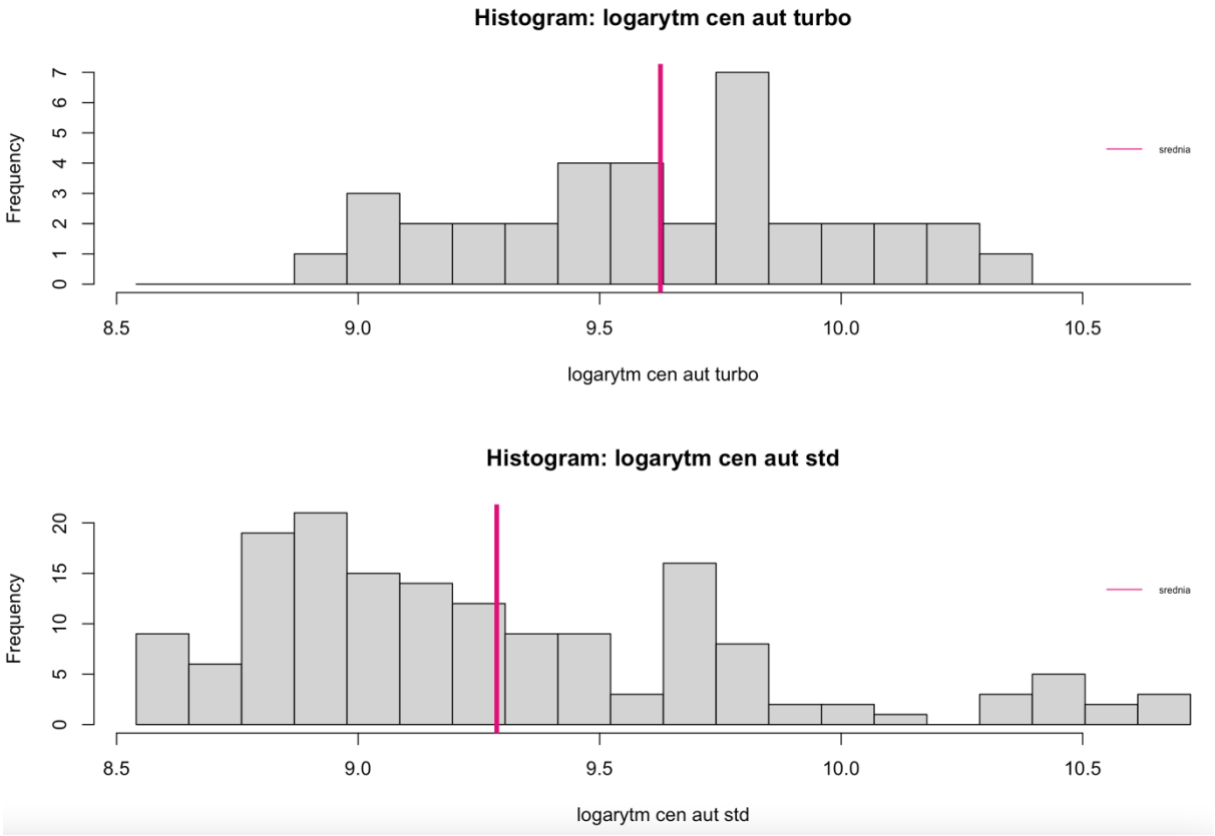
Różnica średnich między dwoma rozkładami normalnymi z nieznaną wariancją jest zgodna z rozkładem t-Sudenta. Jest oczywiste, że dla hipotez dotyczących dwóch wartości oczekiwanych można budować również jednostronne obszary krytyczne (w zależności od postaci hipotezy alternatywnej). Użyję zatem tego testu, by dowiedzieć się, czy ceny aut z silnikami turbo są większe od cen aut wyposażonych w silniki wolnossące. Posłużę się logarytmem cen, gdyż tak jak przedstawiłam we wcześniejszym badaniu, jest one bliższy rozkładowi normalnemu niż tylko ceny.

Poniżej przedstawiona została ilość silników z aspiracją standardową i turbodoładowaniem.

Tabela 10 Ilość silników std oraz turbo

std	turbo
159	36

Zanim wykonam test sprawdzam jak wyglądają histogramy oby dwóch zmiennych.



Rysunek 21 Histogramy dla cen aut z silnikami turbo oraz std

Z powyższych wykresów możemy zauważyć różnicę w średniej dwóch zmiennych. Wykonując test Studenta sprawdzę, czy róznica jest na tyle istotna, by odrzucić H_0.

Na podstawie poprzedniego badania możemy założyć, że rozkład prób (logarytm cen) jest zbliżony do rozkładu normalnego. Zakładam również, że analizowany przeze mnie problem ilustruje test dwóch niezależnych próbek.

Hipoteza zerowa testu brzmi następująco: różnica w średniej cenie samochodów z turbodoładowaniem oraz samochodów *std* wynosi 0, czyli $m_1 = m_2$. Hipoteza alternatywna to: $m_1 > m_2$, gdzie m_1 oznacza średnią cenę w populacji aut z silnikiem turbo, a m_2 średnią cenę w populacji aut z silnikiem turbo. W teście wykorzystamy zatem jednostronny obszar krytyczny.

```
        Welch Two Sample t-test

data:  log_ssanie_ceny_turbo and log_ssanie_ceny_std
t = 4.4723, df = 66.863, p-value = 1.538e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2124856       Inf
sample estimates:
mean of x mean of y
 9.625932  9.287064
```

Rysunek 22 Test Studenta dla log(price) wśród aut z silnikami turbo oraz std

Jak możemy zauważyć, ponieważ wartość p jest znacznie mniejsza niż 0,05 odrzucamy hipotezę zerową, mówiącą o tym, że że ceny samochodów z turbodoładowaniem nie różnią się znacząco od cen samochodów standardowych. Zamiast tego przyjmujemy, że ceny aut z silnikami turbo są większe od cen aut z silnikami wolnossącymi na poziomie istotności alfa = 0.05.

4.9 Czy rodzaj nadwozia auta wpływa na jego cenę? Z jakim rodzajem nadwozia ceny aut najbardziej różnią się od pozostałych?

Poniżej przedstawiam, jak wygląda podział ze względu na nadwozia oraz ile obserwacji wyróżniamy w poszczególnych jego typach.

Tabela 11 Poszczególne typy nadwozi

convertible	hardtop	hatchback	sedan	wagon
6	8	63	94	24

Jak możemy zauważyć w datasecie znajdziemy 5 typów nadwozia, jednak w moim badaniu będę rozważać jedynie trzy z nich: *hatchback, sedan* oraz *wagon* (czyli kombi), ponieważ pozostałe dwie grupy odznaczają się zbyt małą ilością obserwacji.

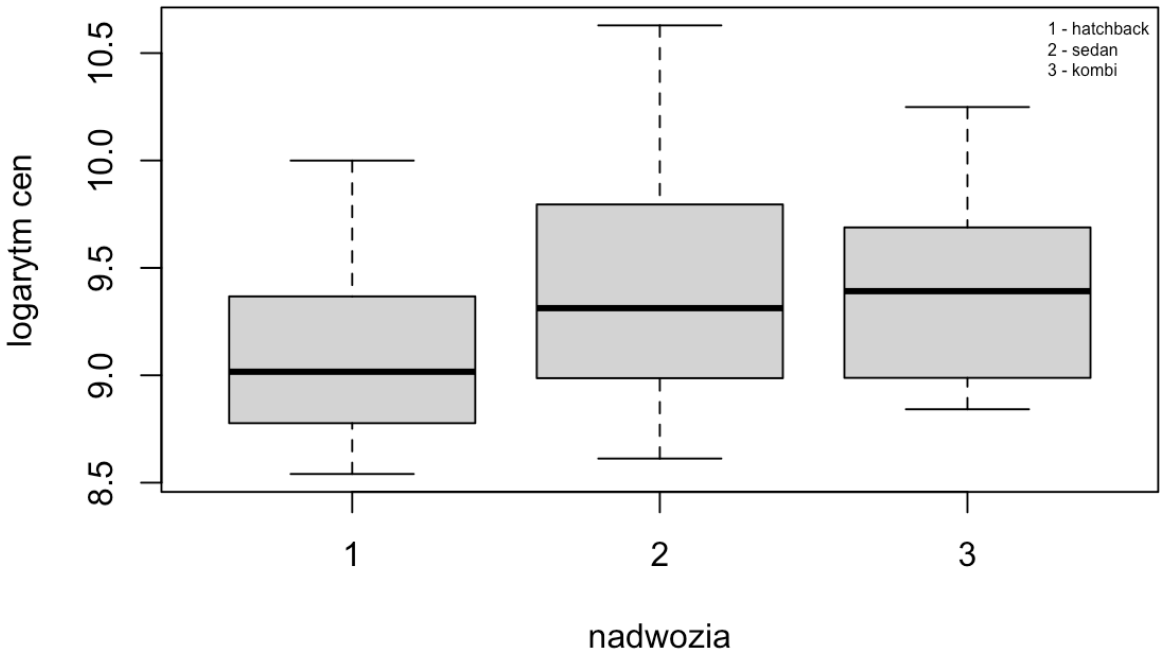
Co więcej, ponownie posłużę się logarytmem cen, gdyż tak jak udowodniłam wcześniej, jego rozkład jest w większym stopniu zbliżony do rozkładu normalnego.

Dane przekształcam do postaci ramki, a wspomniane trzy czynniki kategoryzuje za pomocą funkcji factors(). W tym momencie warto wspomnieć, że faktory to obiekty danych, które służą do kategoryzowania danych i przechowywania ich jako poziomów. Mogą przechowywać zarówno ciągi znaków, jak i liczby całkowite. Są przydatne w kolumnach, które mają ograniczoną liczbę unikalnych wartości. Takie jak „mężczyzna", „kobieta" i „prawda, fałsz", czy tak jak w moim przypadku *„hatchback", „sedan", „wagon".*

Struktura ramki danych – faktor 1 oznacza *hatchback,* 2 - *sedan, 3 – wagon* (kombi)..

Tabela 12 Struktura ramki danych

```
'data.frame':   181 obs. of  2 variables:
 $ nadwozia: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ log_ceny: num  9.71 8.55 8.75 8.63 8.76 ...
```



Rysunek 23 Wykresy pudełkowe dla logarytmu cen w zależności od typu nadwozia

Z wykresów pudełkowych odczytujemy, że średnia dla *hatchback* może różnić się od pozostałych. By potwierdzić to przypuszczenie wykonam test ANOVA, przyjmując następujące hipotezy:

H_0: wartości oczekiwane w podgrupach są sobie równe,

H_1: przynajmniej jedna wartość oczekiwana w podgrupie jest różna

```
> summary(anova)
             Df Sum Sq Mean Sq F value   Pr(>F)
nadwozia      2   4.26  2.1280   10.07 7.17e-05 ***
Residuals   178  37.60  0.2113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>   print(anova)
Call:
   aov(formula = log_ceny ~ nadwozia, data = ramka_danych)

Terms:
                nadwozia Residuals
Sum of Squares   4.25600  37.60412
Deg. of Freedom        2       178

Residual standard error: 0.4596293
Estimated effects may be unbalanced
```

Rysunek 24 Wyniki testu ANOVA

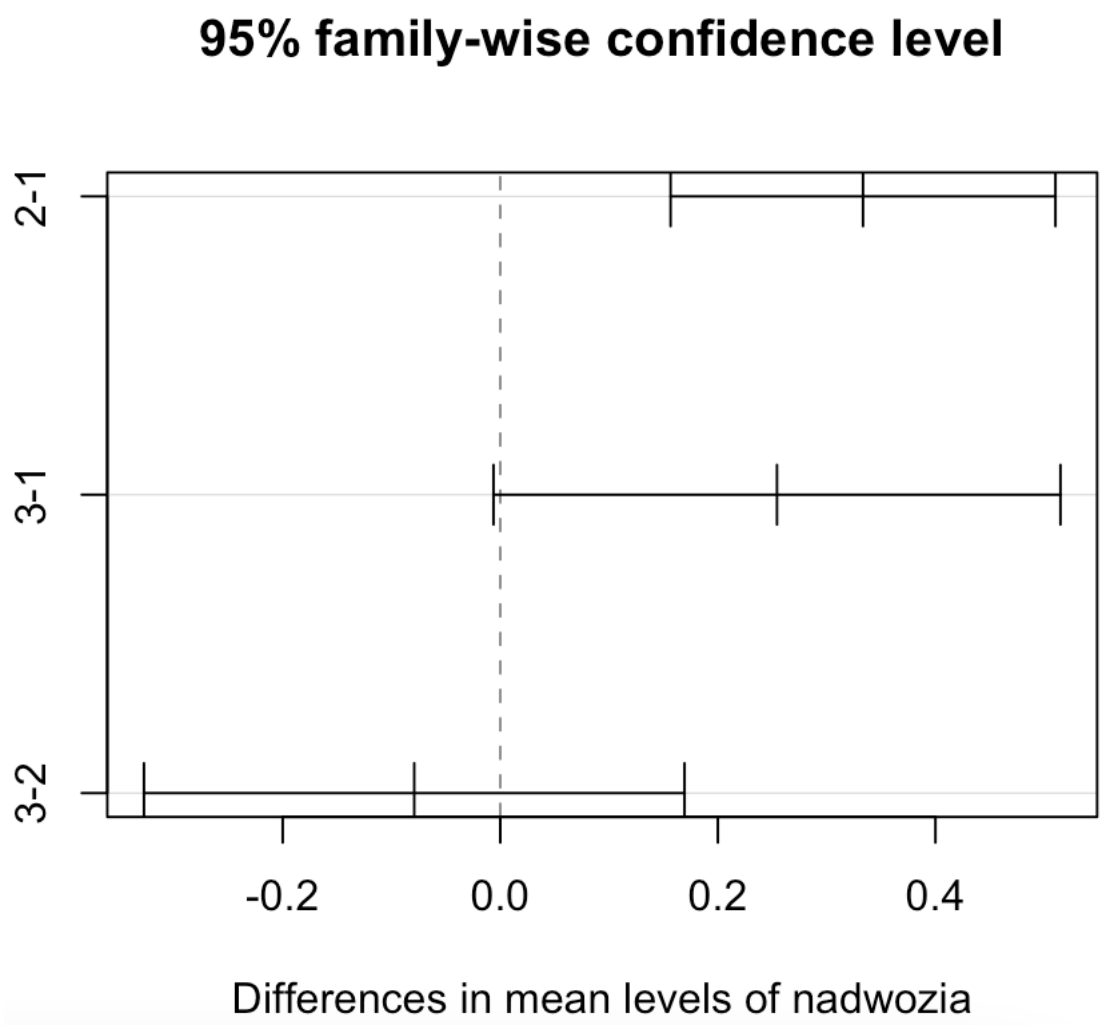
Z przedstawionych powyżej wyników testu ANOVA widać, że istnieje znacząca różnica między co najmniej kilkoma grupami, ponieważ p value jest znacznie mniejsze niż 0,05. Można więc podejrzewać, że rodzaju nadwozia ma wpływ na cenę auta. Na podstawie wykresu pudełkowego możemy przypuszczać, że to właśnie pierwsza grupa różni się od pozostałych, jednak ANOVA nie jest testem, wskazującym na to, który zestaw grup może być inny. Do tego posłuży nam test post-hoc dla ANOVY.

Test Tukeya (lub procedura Tukeya), zwany także testem Tukey's Honest Signiant Difference, jest testem post-hoc opartym rozkładzie t-Studenta. Tak jak wspomniałam, test ANOVA może powiedzieć, czy wyniki są ogólnie znaczące, ale nie określa dokładnie, gdzie leżą te różnice. Po przeprowadzeniu ANOVY i znalezieniu znaczących wyników warto przeprowadzić test HSD Tukeya, aby dowiedzieć się, które średnie grupy (w porównaniu ze sobą) są różne. Test porównuje wszystkie możliwe pary średnich.

```
> HSD <- TukeyHSD(anova)
>   HSD
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_ceny ~ nadwozia, data = ramka_danych)

$nadwozia
          diff          lwr       upr     p adj
2-1  0.3334366  0.156557631 0.5103156 0.0000438
3-1  0.2543519 -0.006230054 0.5149339 0.0573985
3-2 -0.0790847 -0.327530761 0.1693614 0.7326102
```

Rysunek 25 Wyniki testu TukeyHSD

Jak możemy zauważyć p value w grupie 2-1 (sedan-hatchback) jest mniejsze od 0.05. Wskazuje to na odrzucenie naszej hipotezy zerowej: $m_1=m_2$. W pozostałych grupach p value > 0.05, więc nie ma powodów do odrzucenia hipotezy zerowej.



Rysunek 26 Wykres dla wyników testu TukeyHSD

Analizując powyższy wykres, możemy znaleźć pary grup, które mają istotną różnicę na poziomie ufności 95%. Przedział ufności (pokazany przez poziomy słupek) nie będzie nakładał się na 0 (pozioma przerywana linia), gdy różnica jest znacząca. Widzimy, że grupy 2 i 1 (cena samochodów typu hatchback i cena sedanów) znacznie się różnią.

Korzystając z testu Tukey'a HSD oraz ANOVY mogliśmy zauważyć, że cena samochodów typu hatchback różni się znacznie od ceny sedanów, cena hatchbacków i kombi może się różnić lub nie, a ceny sedanów i kombi nie różnią się znacząco. Zarówno w teście t, jak i ANOVA / Tukeya, założenie normalności jest ważne. Innymi słowy, ważne jest, aby obserwacje, na których przeprowadza się testy, pochodziły z populacji o rozkładzie normalnym. Dlatego sprawdziłam wcześniej czy cena lub log (cena) jest zgodna z rozkładem normalnym i ostatecznie użyłam log(cena) do wszystkich moich obliczeń, ponieważ okazało się, że jest ona bardziej zbliżona do rozkładu normalnego.

W teście ANOVA/TUKEY należy jednak pamiętać również o innych ważnych założeniach, które powinny zostać spełnione:

1. normalność rozkładu reszt modelu
2. równość wariancji w podgrupach

```
>  #TESTOWANIE ZAŁOŻENIA O NORMALNOŚCI ROZKŁADU RESZT ANOVY
>   # wydobywamy reszty
>   anova_reszty <- residuals(object = anova )
>   # test Shapiro- Wilka (H0: normalność rozkładu)
>   shapiro.test(x = anova_reszty )

        Shapiro-Wilk normality test

data:  anova_reszty
W = 0.95152, p-value = 7.418e-06
```

Rysunek 27 Test Shapiro-Wilka dla reszt modelu ANOVA

1. Z powyższego testu możemy odczytać, że wartość p value < 0.05, co skłania do odrzucenia hipotezy zerowej – co za tym idzie warunek normalności rozkładu reszt modelu jest niespełniony.

```
>  #TESTY WERYFIKUJĄCE JEDNORODNOŚĆ WARIANCJI W WIĘCEJ NIŻ 2 GRUPACH (H0: wariancje
 są równe, H1: przynajmniej jedna jest różna)
>   #test Levene'a
>   leveneTest(log_ceny ~ nadwozia, data = ramka_danych)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   2  3.3333 0.03791 *
      178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>   #test Bartletta
>   bartlett.test(log_ceny ~ nadwozia, data = ramka_danych)

        Bartlett test of homogeneity of variances

data:  log_ceny by nadwozia
Bartlett's K-squared = 6.054, df = 2, p-value = 0.04846
```

Rysunek 28 Testy równości wariancji w podgrupach

2. Również z tych dwóch powyższych testów odczytujemy wartość p value < 0.05, co oznacza, że odrzucamy hipotezę zerową świadczącą o jednorodności wariancji w podgrupach. Drugie założenie testów także nie zostało spełnione.

Na szczęście nie oznacza to, że moje pytanie badawcze pozostanie bez odpowiedzi. Test Kruskala-Wallisa jest jedną z najpopularniejszych alternatyw dla jednoczynnikowej analizy wariancji ANOVA. Test Kruskala-Wallisa przeprowadzamy w przypadku, gdy zostały złamane założenia analizy wariancji ANOVA bądź gdy charakter naszych zmiennych nie pozwala na wykorzystanie analizy wariancji ANOVA.

Tak jak jednoczynnikowa analiza wariancji jest rozszerzeniem testów t-Studenta dla prób niezależnych, tak test Kruskala-Wallisa jest rozszerzeniem testu U Manna-Whitneya. W przypadku testu Kruskala-Wallisa mamy do czynienia z większą niż 2 liczbą porównywanych grup – czyli np. sedan, hatchback, kombi. Jeżeli mielibyśmy dwie grupy niezależne przeprowadzilibyśmy test U Manna-Whitneya. Gdy mamy więcej niż 2 porównywane grupy korzystamy z testu Kruskala-Wallisa.

Test Kruskala-Wallsa nie wymaga tak jak analiza wariancji ANOVA spełnienia szeregu założeń. Rozkłady zmiennych nie muszą być zbliżone do rozkładu normalnego. Nie ma wymogu równoliczności grup pod względem liczby osób jak również nie wymagana jest równość wariancji w grupach - homogeniczoność wariancji. Jedynymi wymogami do przeprowadzenia testu Kruskala-Walisa są:

- zmienna zależna powinna być mierzona na skali co najmniej porządkowej (może być również mierzona na skali ilościowej)
- obserwacje w analizowanych grupach powinny być niezależne wobec siebie, co oznacza, że osoba będąca w jednej grupie nie powinna być również w innej porównywanej grupie[13].

Ideą testu jest badanie miar położenia (rozkładu) cechy we wszystkich porównywanych grupach. Hipotezą zerową jest testu równość dystrybuant rozkładów w porównywanych populacjach, hipotezą alternatywną zaś nierówność dystrybuant.

[13] https://www.naukowiec.org/

```
        Kruskal-Wallis rank sum test

data:  log_ceny by nadwozia
Kruskal-Wallis chi-squared = 18.261, df = 2, p-value = 0.0001083
```

Rysunek 29 Test Kruskala-Wallisa

Jak możemy zauważyć powyżej, również test Kruskala-Wallisa wskazuje nam na odrzucenie hipotezy zerowej – istnieje znacząca róznica między grupami, a więc rodzaj nadwozia auta rzeczywiście wpływa na jego cenę. Warto jednak zwrócić uwagę, iż istotny wynik testu Kruskala-Wallisa wymaga jeszcze przeprowadzenia porównań wielokrotnych (tzw. testów post-hoc). Z racji, że istotny statystycznie wynik testu Kruskala-Wallisa informuje, że są różnice pomiędzy grupami to w celu sprawdzenia pomiędzy którymi grupami należy przeprowadzić porównania wielokrotne z odpowiednimi poprawkami. Prawdopodobnie najpopularniejszym testem do tego celu jest test Dunn.

```
        Pairwise comparisons using Dunn's-test for multiple
                        comparisons of independent samples

data:  log_ceny by nadwozia

  1     2
2 9e-05 -
3 0.023 0.753

P value adjustment method: holm
```

Rysunek 30 Test Dunna

Ponownie możemy zauważyć, że cena samochodów typu hatchback (1) różni się znacznie od ceny sedanów (2) – p value < 0.05, zaś ceny sedanów i kombi (3) nie różnią się znacząco - p value > 0.05.

5. Wnioski

1. **Czy większość aut ma silniki wolnossące (std)? Czy wiekszość samochodów napędzana jest dieslem? Jakich marek samochodowych jest najwięcej? Które marki samochodów są najdroższe?**

Przeważająca ilość aut wyposażona jest w silniki std oraz napędzana jest gazem. Ponadto w datasecie znajduje się najwięcej samochodów marki toyota. Najdroższymi markami są: mercedes-benz, jaguar oraz porsche.

2. **Jak skorelowane są dane? Czy wielkość silnika jest powiązana z wielkością auta?**

Niektóre pary zmiennych są słabo skorelowane – na przykład rozstaw osi (*wheel.base*) i moc (*horsepowe*r) - podczas gdy niektóre są skorelowane dużo mocniej - na przykład wielkość silnika (*engine.size*) i masa własna pojazdu (*curb.weight*), co potwierdza również test korelacji dla dwóch ostatnich zmiennych. Jest to zależność dodatnia, co oznacza, że im większy silnik, tym większa jest masa własna pojazdu.

3. **Czy auta napędzane LPG odznaczają się mniejszym spalaniem w mieście niż pojazdy napędzane dieslem?**

Hipoteza zerowa nie potwierdziła się. Po analizuie wykresu pudełkowego oraz wyniku testu U-Manna-Whitney'a-Wilcoxona (wartość p jest mniejsza niż 0,05, a więc dystrybuanty tych dwóch rozkładów istotnie różnią się między sobą) możemy przepuszczać, że auta napędzane LPG nie odznaczają się lepszą wydajnością spalania w mieście od aut napędzanych dieslem.

4. **Czy ceny pojazdów podlegają rozkładowi Gaussa? Jaki ma to wpływ na dalsze badania? Czy logarytm zmiennej cena lepiej odwzorowuje rozkład normalny?**

W analizie rozkładu zmiennej *price* nie udało się przedstawić dowodów na to, że pochodzi ona z rozkładu Gaussa. Można jednak domniemywać, na bazie wykresu QQ oraz testu K-S, iż logarytm zmiennej *price* w pewnym stopniu odwzorowuje rozkład normalny. Celem transformacji cen na postać logarytmiczną było właśnie otrzymanie danych przypominających rozkład normalny oraz redukcja skośności. Jako, że założenie to zostało w pewnym stopniu zrealizowane (a przynajmniej pewne testy to potwierdzają takie prawdopodobieństwo), w dalszej części badań posługiwałam się właśnie zlogarytmizowaną postacią zmiennej *price*.

5. **Czy ceny aut z silnikami turbo są większe od cen aut z silnikami wolnossącymi?**

Wartość p otrzymana z testu jest znacznie mniejsza niż 0,05, a więc odrzuciliśmy hipotezę zerową, mówiącą o tym, że że ceny samochodów z turbodoładowaniem nie różnią się znacząco od cen samochodów standardowych. Zamiast tego przyjeliśmy, że ceny aut z silnikami turbo są większe od cen aut z silnikami wolnossącymi na poziomie istotności alfa = 0.05.

6. **Czy rodzaj nadwozia auta wpływa na jego cenę? Z jakim rodzajem nadwozia ceny aut najbardziej różnią się od pozostałych?**

Test Kruskala-Wallisa wskazał na odrzucenie hipotezy zerowej – istnieje znacząca różnica między grupami, a więc rodzaj nadwozia auta rzeczywiście wpływa na jego cenę. Warto przypomnieć, iż istotny wynik testu Kruskala-Wallisa wymagał jeszcze przeprowadzenia porównań wielokrotnych z odpowiednimi poprawkami (tzw. testów post-hoc). Z racji, że istotny statystycznie wynik testu Kruskala-Wallisa informuje, że są różnice pomiędzy grupami to w celu sprawdzenia pomiędzy którymi grupami zachodzą te różnice przeprowadziłam test Dunna. Dzięki niemu mogliśmy zauważyć, że cena samochodów typu hatchback różni się znacznie od ceny sedanów (p value < 0.05), zaś ceny sedanów i kombi (3) nie różnią się znacząco (p value > 0.05).

6. Bibliografia

- Rocznik Statystyczny Rzeczypospolitej Polskiej 2018, Główny Urząd Statystyczny, Warszawa 2018
- Statistics in Medicine (Third Edition), R.H. Riffenburgh, 2012
- Introductory Statistics (Third Edition), Sheldon M. Ross, 2010
- Programowanie w języku R, Marek Gągolewski
- Statystyka od podstaw, Janina Jóźwiak, Jarosław Podgórski
- J. Baudrillard, Społeczeństwo konsumpcyjne. Jego mity i struktury, Wydawnictwo Sic!,
- https://www.naukowiec.org/
- https://algolytics.com/

7. Spis tabel

8. Spis rysunków