**Sprawozdanie na przedmiot wykład specjalizujący – regresja liniowa**

Informatyka ogólna grupa 3

Wykonał: Murawski Sebastian

Index: 155082

Sprawozdanie wersja 1.0

Uniwersytet Warmińsko-Mazurski

Wydział Matematyki i Informatyki (WMiI)

Olsztyn, 17.11.2022r.

Spis treści

[Wstęp 3](#_Toc119606245)

[Opis atrybutów w bazie danych: 4](#_Toc119606246)

[Wczytać zestaw danych, który odpowiada ostatniej cyfrze z numeru albumu, do systemu R. 5](#_Toc119606247)

[Wykonać badania eksploracyjne (EDA, ang. exploratory data analysis) zestawu danych. 6](#_Toc119606248)

[Struktura danych 6](#_Toc119606249)

[Udział poszczególnych marek w samochodach. 7](#_Toc119606250)

[Udział poszczególnych paliw w samochodach. 8](#_Toc119606251)

[Udział poszczególnych typów nadwozia w samochodach. 9](#_Toc119606252)

[Udział ilości drzwi w samochodach. 10](#_Toc119606253)

[Udział poszczególnych systemów paliwa w samochodach. 11](#_Toc119606254)

[Udział poszczególnych miejsca silnika w samochodach. 12](#_Toc119606255)

[Udział poszczególnych typów podawania powietrza w samochodach. 13](#_Toc119606256)

[Udział poszczególnych rodzajów napędu w samochodach. 14](#_Toc119606257)

[Udział poszczególnej ilości cylindrów w samochodach. 15](#_Toc119606258)

[Udział poszczególnych typów silnika w samochodach. 16](#_Toc119606259)

[Spalanie samochodów w różnych środowiskach. 17](#_Toc119606260)

[Korelacja spalania samochodów od pojemności silnika. 18](#_Toc119606261)

[Korelacja spalania samochodów od wymiarów samochodu podczas jazdy na autostradzie. 20](#_Toc119606262)

[Korelacja spalania samochodów od wymiarów samochodu podczas jazdy w mieście. 22](#_Toc119606263)

[Korelacja spalania samochodów od ilości cylindrów. 24](#_Toc119606264)

[Korelacja stopnia sprężania do użytego paliwa. 25](#_Toc119606265)

[Korelacja spalania samochodów do sposobu doładowania podczas jazdy na autostradzie oraz mieście. 26](#_Toc119606266)

[Korelacja spalania samochodów do masy własnej podczas jazdy na autostradzie oraz mieście. 27](#_Toc119606267)

[Korelacja spalania samochodów do rodzaju napędu jazdy na autostradzie oraz mieście. 28](#_Toc119606268)

[Wyznaczyć cechę decyzyjną i cechy objaśniające. 29](#_Toc119606269)

[Oszacować regresję liniową cechy decyzyjnej na cechy objaśniające. 29](#_Toc119606270)

[Wyznaczyć efektywność zbudowanego modelu na podstawie współczynnika determinacji i średniej kwadratowej błędów RMSE (Root Mean Squared Error). 31](#_Toc119606271)

# Wstęp

Tematem sprawozdania jest praca z różnymi bazami danych w celu wykonania badania eksploracyjnego w języku R.

Podczas wykonywania zadania, zostały użyte następujące narzędzia:

* IDE – Rstudio – narzędzie ułatwiające pracę z językiem R. RStudio to zintegrowane środowisko programistyczne dla R , języka programowania do obliczeń statystycznych i grafiki.
* R – interpretowany język programowania oraz środowisko do obliczeń statystycznych i wizualizacji wyników. R rozprowadzany jest w postaci kodu źródłowego oraz w postaci binarnej wraz z wieloma dystrybucjami GNU/Linuksa. Dostępna jest także wersja dla Microsoft Windows i Mac OS.
* Baza danych „Automobile Data Set”
* Dodatkowe biblioteki takie jak:
  + Tidyverse — pakiety R do analizy danych, w tym ggplot2 , dplyr , tidyr i purrr
  + Caret -(skrót od Classification And RE gression Training ) to zestaw funkcji , które mają na celu usprawnienie procesu tworzenia modeli predykcyjnych .
  + Moments -Funkcje do obliczenia: momenty, kurtoza Pearsona, kurtoza i skośność Geary'ego; związane z nimi testy (Anscombe-Glynn, D'Agostino, Bonett-Seier).
  + ggplot2 - to pakiet wizualizacji danych typu open source dla statystycznego języka programowania R .Służy jako zamiennik podstawowej grafiki w R i zawiera szereg ustawień domyślnych do wyświetlania.

Zadania które trzeba było wykonać w ramach zajęć:

1. Wczytać zestaw danych, który odpowiada ostatniej cyfrze z numeru albumu, do systemu R.
2. Wykonać badania eksploracyjne (EDA, ang. exploratory data analysis) zestawu danych.
3. Wyznaczyć cechę decyzyjną i cechy objaśniające.
4. Oszacować regresję liniową cechy decyzyjnej na cechy objaśniające.
5. Wyznaczyć efektywność zbudowanego modelu na podstawie współczynnika determinacji i średniej kwadratowej błędów RMSE (Root Mean Squared Error).
6. Polepszyć efektywność modelu za pomocą walidacji krzyżowej k-krotnej (K-Fold Cross-Validation).
7. Napisać sprawozdanie i wysłać pod adres [matematyka@gmx.com](mailto:matematyka@gmx.com)

# Opis atrybutów w bazie danych:

* symbolizacja: -3, -2, -1, 0, 1, 2, 3.
* znormalizowane straty: ciągłe od 65 do 256.
* Modele: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury,mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
* rodzaj paliwa: olej napędowy, gaz.
* Styl doładowania: std, turbo.
* liczba-drzwi: cztery, dwa.
* styl nadwozia: hardtop, kombi, sedan, hatchback, kabriolet.
* Rodzaj napędu: 4wd, fwd, rwd.
* lokalizacja silnika: przód, tył.
* rozstaw osi: od 86,6 do 120,9.
* długość: od 141,1 do 208,1.
* szerokość: od 60,3 do 72,3.
* wzrost: od 47,8 do 59,8.
* masa własna: od 1488 do 4066.
* typ silnika: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
* liczba cylindrów: osiem, pięć, cztery, sześć, trzy, dwanaście, dwa.
* Wielkość silnika: od 61 do 326.
* układ paliwowy: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
* Szerokość tłoka: od 2,54 do 3,94.
* Skok tłoka: od 2.07 do 4.17.
* stopień kompresji: od 7 do 23.
* moc: ciągła od 48 do 288.
* Obroty maksymalne: od 4150 do 6600.
* city-mpg: od 13 do 49.
* autostrada-mpg: od 16 do 54.
* cena: od 5118 do 45400.

# Wczytać zestaw danych, który odpowiada ostatniej cyfrze z numeru albumu, do systemu R.

Na samym początku trzeba wczytać dane do programu oraz załadować potrzebne biblioteki.

#1.Wczytać zestaw danych, który odpowiada ostatniej cyfrze z numeru albumu, do systemu R.

#Ladowanie bibliotek

library**(**tidyverse**)**

library**(**mlbench**)**

library**(**ggplot2**)**

library**(**moments**)**

library**(**caret**)**

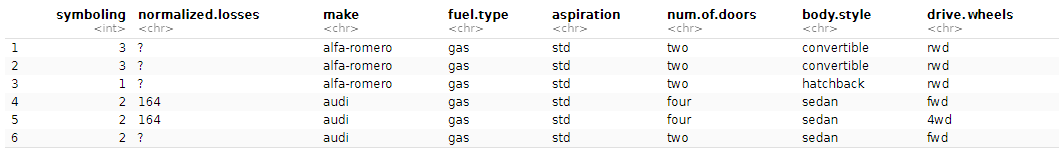
dataset **<-** read.csv**(**"daneS/imports-85.data",sep**=**",", header**=FALSE)**

Dla wygody dodałem nazwy odpowiadającej każdej z kolumn.

colnames**(**dataset**)** **<-** c**(**"symboling","normalized.losses","make","fuel.type","aspiration","num.of.doors","body.style","drive.wheels","engine.location","wheel.base","length","width","height","curb.weight","engine.type","num.of.cylinders","engine.size","fuel.system","bore","stroke","compression.ratio","horsepower","peak.rpm","city.mpg","highway.mpg","price"**)**

A następnie wyświetliłem w celu sprawdzenia

head**(**dataset**)**



Wykonać badania eksploracyjne (EDA, ang. exploratory data analysis) zestawu danych.

## Struktura danych

W celu sprawdzenia struktury danych oraz typów danych użyłem komendy str (dataset).

Wynikiem tej operacji jest wydruk w konsoli przydatnych informacji takie jak:

* Nazwa kolumn, wymiar danych, typ danych oraz przykładowe wartości które się w nich znajdują.

Wydruk z konsoli:

'data.frame'**:** 205 obs. of 26 variables**:**

**$** symboling **:** int 3 3 1 2 2 2 1 1 1 0 ...

**$** 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 67.9 ...

**$** height **:** num 48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...

**$** curb.weight **:** int 2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...

**$** engine.type **:** chr "dohc" "dohc" "ohcv" "ohc" ...

**$** num.of.cylinders **:** chr "four" "four" "six" "four" ...

**$** engine.size **:** int 130 130 152 109 136 136 136 136 131 131 ...

**$** fuel.system **:** chr "mpfi" "mpfi" "mpfi" "mpfi" ...

**$** bore **:** chr "3.47" "3.47" "2.68" "3.19" ...

**$** stroke **:** chr "2.68" "2.68" "3.47" "3.40" ...

**$** compression.ratio**:** num 9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...

**$** horsepower **:** chr "111" "111" "154" "102" ...

**$** peak.rpm **:** chr "5000" "5000" "5000" "5500" ...

**$** city.mpg **:** int 21 21 19 24 18 19 19 19 17 16 ...

**$** highway.mpg **:** int 27 27 26 30 22 25 25 25 20 22 ...

**$** price **:** chr "13495" "16500" "16500" "13950" ...

## Udział poszczególnych marek w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne marki samochodów w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nUdział producentów w ilości samochodów\n\n"**)**

make **<-** dataset**$**make

cbind**(**freq**=**table**(**make**)**, percentage**=**prop.table**(**table**(**make**))\***100**)**

Wydruk z konosli:

Udział producentów w ilości samochodów

freq percentage

alfa**-**romero 3 1.4634146

audi 7 3.4146341

bmw 8 3.9024390

chevrolet 3 1.4634146

dodge 9 4.3902439

honda 13 6.3414634

isuzu 4 1.9512195

jaguar 3 1.4634146

mazda 17 8.2926829

mercedes**-**benz 8 3.9024390

mercury 1 0.4878049

mitsubishi 13 6.3414634

nissan 18 8.7804878

peugot 11 5.3658537

plymouth 7 3.4146341

porsche 5 2.4390244

renault 2 0.9756098

saab 6 2.9268293

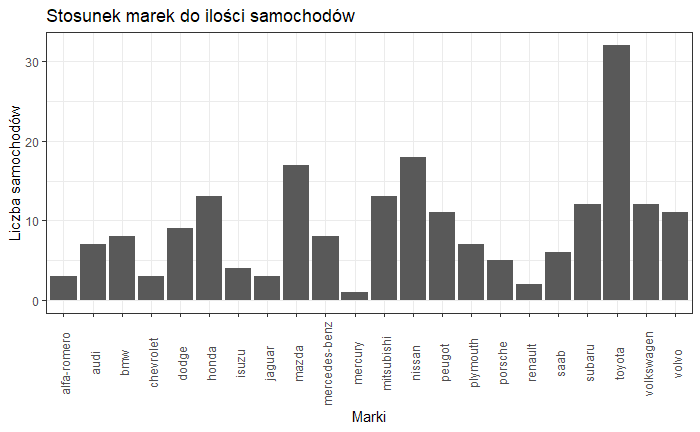
subaru 12 5.8536585

toyota 32 15.6097561

volkswagen 12 5.8536585

volvo 11 5.3658537

Wykres:



## Udział poszczególnych paliw w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne typy paliw w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania rodzajów paliw\n\n"**)**

fuel.type **<-** dataset**$**fuel.type

cbind**(**freq**=**table**(**fuel.type**)**, percentage**=**prop.table**(**table**(**fuel.type**))\***100**)**

**)**

Wydruk z konosli:

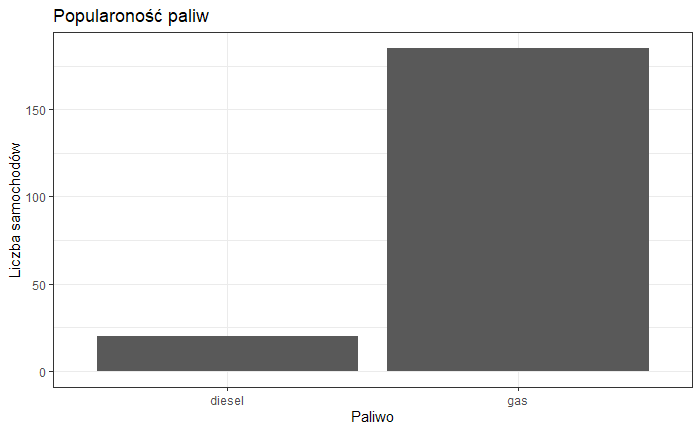
Częstość występowania rodzajów paliw

freq percentage

diesel 20 9.756098

gas 185 90.243902

Wykres:



## Udział poszczególnych typów nadwozia w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne typy nadwozia w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania rodzajów nadwozia\n\n"**)**

body.style **<-** dataset**$**body.style

cbind**(**freq**=**table**(**body.style**)**, percentage**=**prop.table**(**table**(**body.style**))\***100**)**

Wydruk z konosli:

Częstość występowania rodzajów nadwozia

freq percentage

convertible 6 2.926829

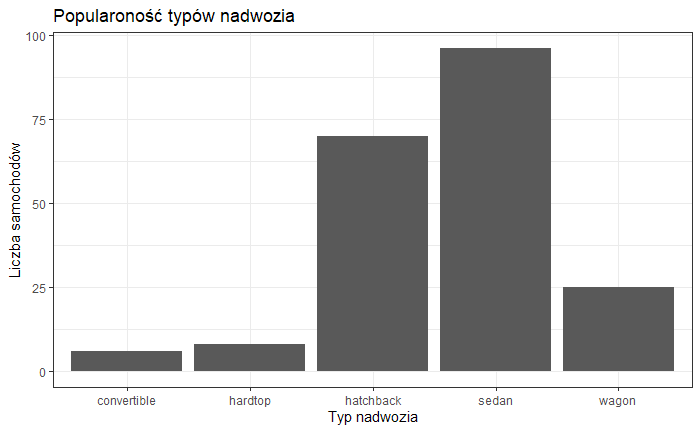
hardtop 8 3.902439

hatchback 70 34.146341

sedan 96 46.829268

wagon 25 12.195122

Wykres:



## Udział ilości drzwi w samochodach.

Następnie sprawdziłem jaki udział ma ilość drzwi w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania ilości drzwi\n\n"**)**

num.of.doors **<-** dataset**$**num.of.doors

cbind**(**freq**=**table**(**num.of.doors**)**, percentage**=**prop.table**(**table**(**num.of.doors**))\***100**)**

Wydruk z konosli:

Częstość występowania ilości drzwi

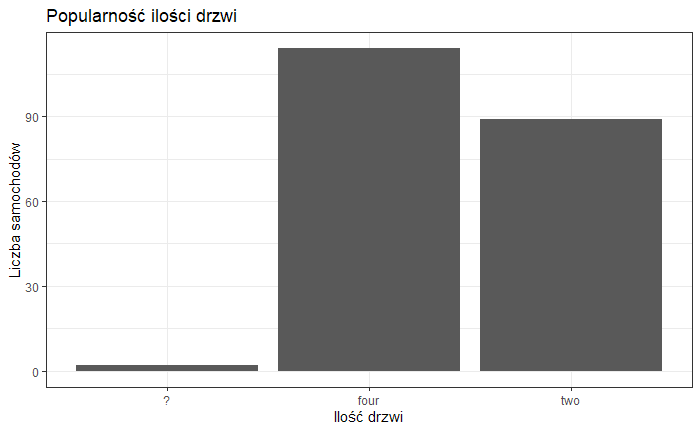
freq percentage

**?** 2 0.9756098

four 114 55.6097561

two 89 43.4146341

Wykres:



## Udział poszczególnych systemów paliwa w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne systemy podawania paliwa w całym zbiorze, za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania systemów paliwa\n\n"**)**

fuel.system **<-** dataset**$**fuel.system

cbind**(**freq**=**table**(**fuel.system**)**, percentage**=**prop.table**(**table**(**fuel.system**))\***100**)**

Wydruk z konosli:

Częstość występowania systemów paliwa

freq percentage

1bbl 11 5.3658537

2bbl 66 32.1951220

4bbl 3 1.4634146

idi 20 9.7560976

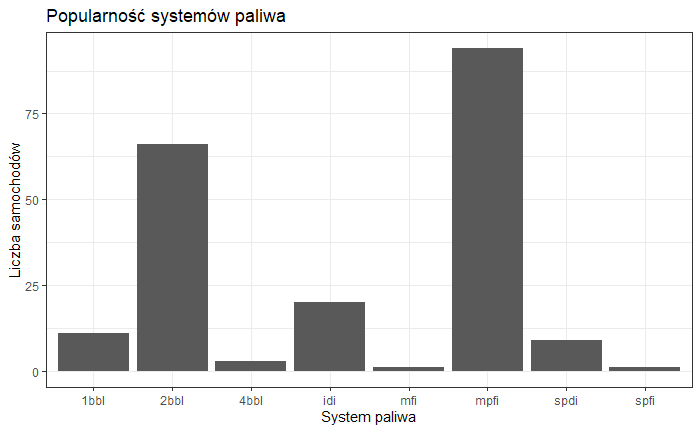
mfi 1 0.4878049

mpfi 94 45.8536585

spdi 9 4.3902439

spfi 1 0.4878049

Wykres:



## Udział poszczególnych miejsca silnika w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne umiejscowienia silnika w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania miejsca na silnik\n\n"**)**

engine.location **<-** dataset**$**engine.location

cbind**(**freq**=**table**(**engine.location**)**, percentage**=**prop.table**(**table**(**engine.location**))\***100**)**

Wydruk z konosli:

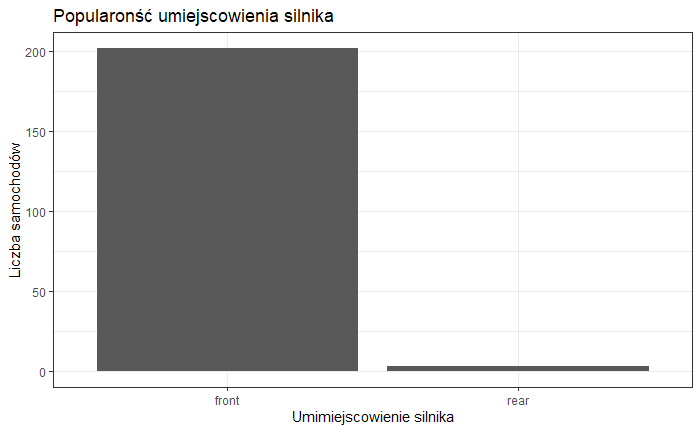
Częstość występowania miejsca na silnik

freq percentage

front 202 98.536585

rear 3 1.463415

Wykres:



## Udział poszczególnych typów podawania powietrza w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne typy podawania powietrza w całym zbiorze, za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania rodzaju podawania powietrza\n\n"**)**

aspiration **<-** dataset**$**aspiration

cbind**(**freq**=**table**(**aspiration**)**, percentage**=**prop.table**(**table**(**aspiration**))\***100**)**

Wydruk z konosli:

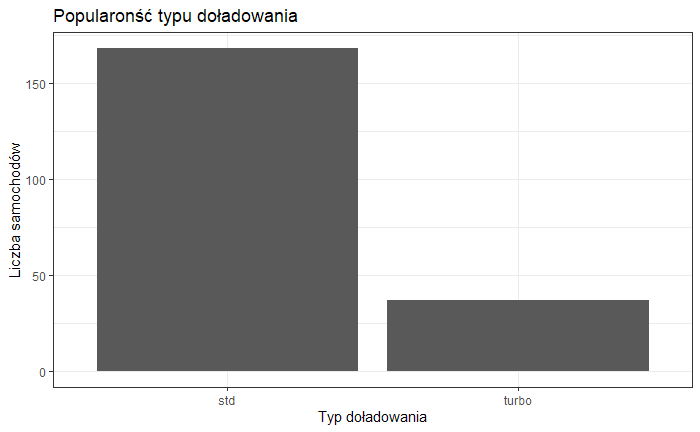
Częstość występowania rodzaju podawania powietrza

freq percentage

std 168 81.95122

turbo 37 18.04878

Wykres:



## Udział poszczególnych rodzajów napędu w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne rodzaje napędu w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania stylu napędu \n\n"**)**

drive.wheels **<-** dataset**$**drive.wheels

cbind**(**freq**=**table**(**drive.wheels**)**, percentage**=**prop.table**(**table**(**drive.wheels**))\***100**)**

Wydruk z konosli:

Częstość występowania stylu napędu

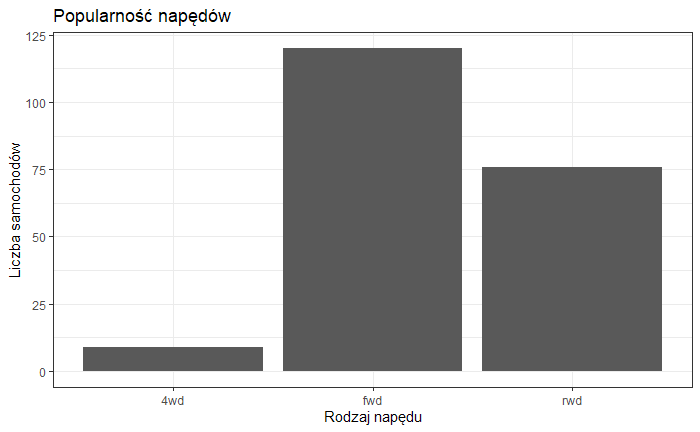
freq percentage

4wd 9 4.390244

fwd 120 58.536585

rwd 76 37.073171

Wykres:



## Udział poszczególnej ilości cylindrów w samochodach.

Następnie sprawdziłem jaki udział ma ilość cylindrów w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania ilości cylindrów \n\n"**)**

num.of.cylinders **<-** dataset**$**num.of.cylinders

cbind**(**freq**=**table**(**num.of.cylinders**)**, percentage**=**prop.table**(**table**(**num.of.cylinders**))\***100**)**

Wydruk z konosli:

Częstość występowania ilości cylindrów

freq percentage

eight 5 2.4390244

five 11 5.3658537

four 159 77.5609756

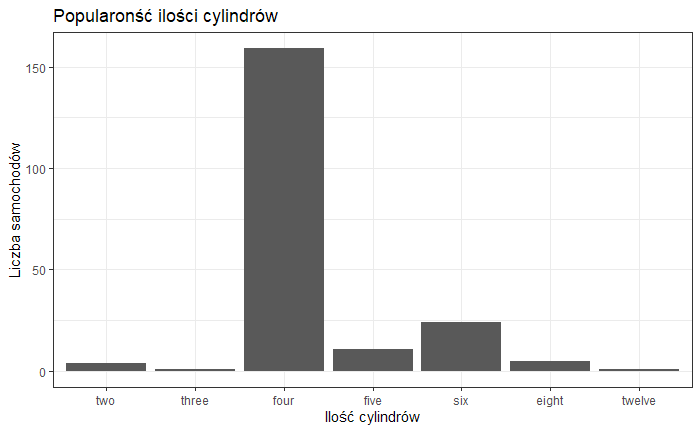
six 24 11.7073171

three 1 0.4878049

twelve 1 0.4878049

two 4 1.9512195

Wykres:



## Udział poszczególnych typów silnika w samochodach.

Następnie sprawdziłem jaki udział mają poszczególne typy silników w całym zbiorze,

za pomocą poniższej formuły:

cat**(**"\n\nCzęstość występowania rodzaju slinków \n\n"**)**

engine.type **<-** dataset**$**engine.type

cbind**(**freq**=**table**(**engine.type**)**, percentage**=**prop.table**(**table**(**engine.type**))\***100**)**

Wydruk z konosli:

Częstość występowania rodzaju slinków

freq percentage

dohc 12 5.8536585

dohcv 1 0.4878049

l 12 5.8536585

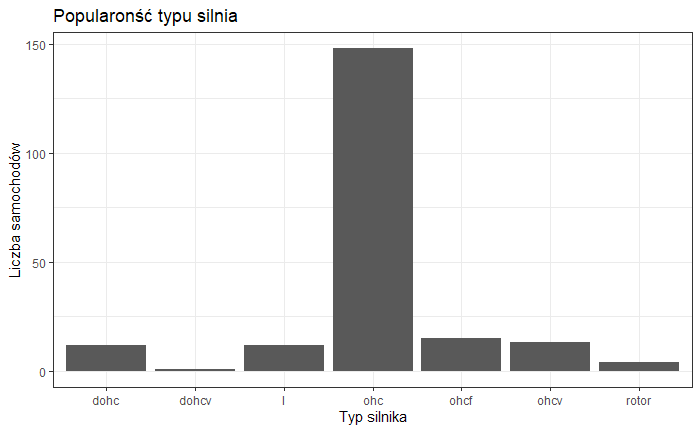
ohc 148 72.1951220

ohcf 15 7.3170732

ohcv 13 6.3414634

rotor 4 1.9512195

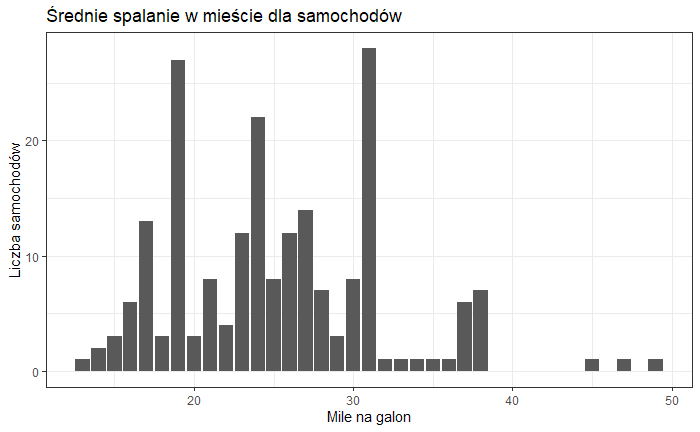
Wykres:

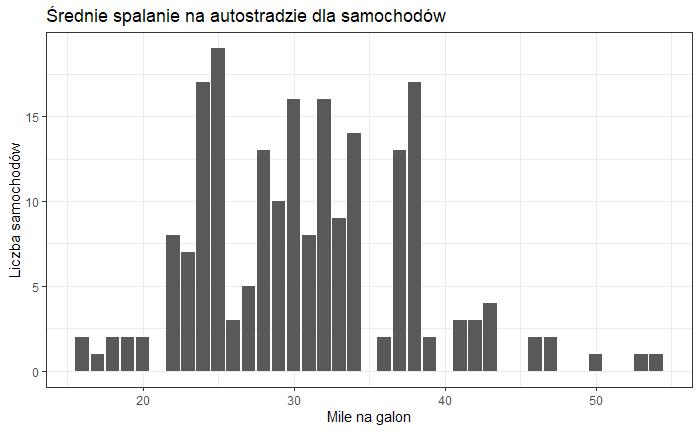


## Spalanie samochodów w różnych środowiskach.

Teraz zajmiemy się porównywaniem danych oraz wizualizacją różnic. Poniższe wykresy wizualizują średnie spalanie dla całej grupy samochodów dla 2 różnych sytuacji. Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa.

W odległości od 5-10 km od ścisłego centrum średnie prędkości oscylują w granicach 48-49 km/h. Miasto jest stosunkowo wolne na tle całej Polski. W godzinach porannego szczytu (6:00 - 9:00) w ścisłym centrum pojedziemy z średnią prędkością 21 km/h. Natomiast średnia prędkość na trasie wynosi ok. 75km/h.



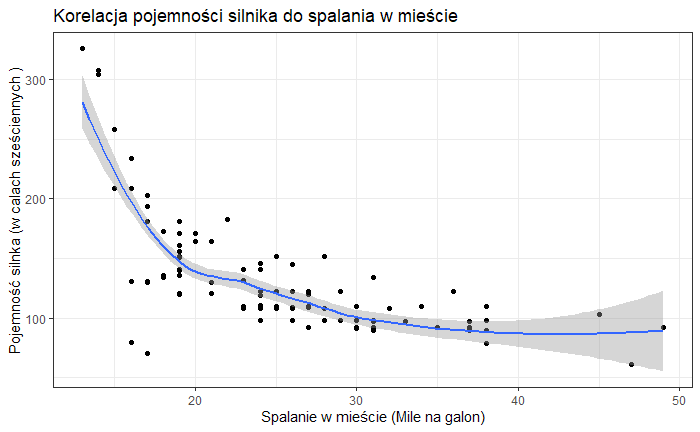
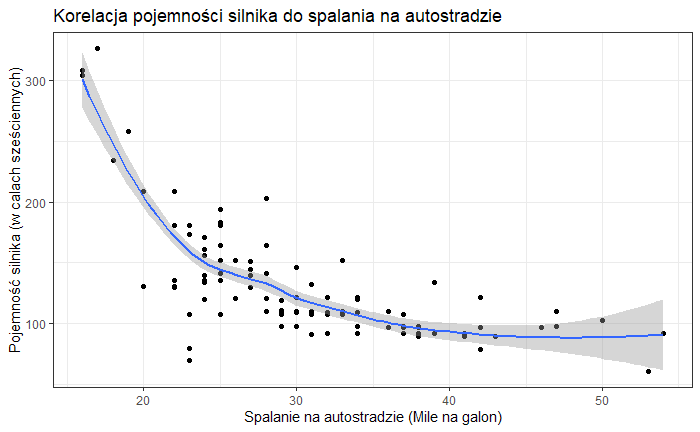


Wnioski:

Jak możemy zauważyć średnie spalanie samochodów znacząco się różni. Samochody są bardziej ekonomiczne podczas jazdy po autostradzie.

## Korelacja spalania samochodów od pojemności silnika.

Poniższe wykresy wizualizują średnie spalanie dla całej grupy samochodów w mieście oraz na autostradzie uzależnionej od wielkości silnika . Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa, natomiast pojemność silnika jest wyrażona w calach sześciennych [in^3]. Jeden cal sześcienny [in^3] odpowiada 16,4 centymetrowi sześciennemu [cm^3].

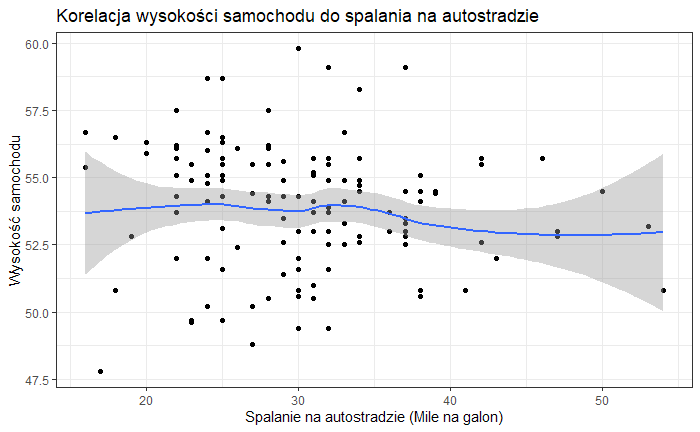
 

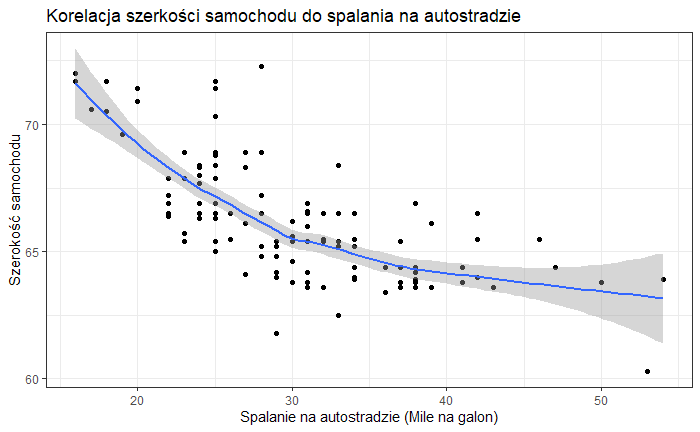
Wnioski:

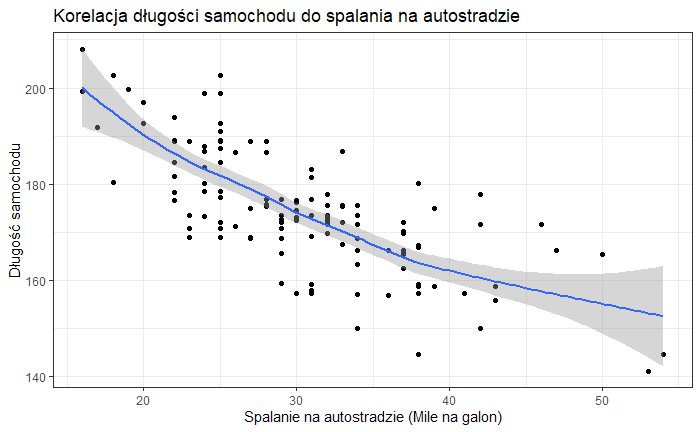
Jak możemy zauważyć średnie spalanie samochodów znacząco się różni. Samochody są bardziej ekonomiczne gdy mają mniejsza pojemność silnika. Różnica między jazdą po mieście oraz autostradzie jest marginalna.

## Korelacja spalania samochodów od wymiarów samochodu podczas jazdy na autostradzie.

W tym przypdaku duże znaczenie ma prędkość więć będziemy rozważać tylko jazdę po autostradzie. Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa. Wysokość, szerokość oraz długość jest wyrażona w calach.





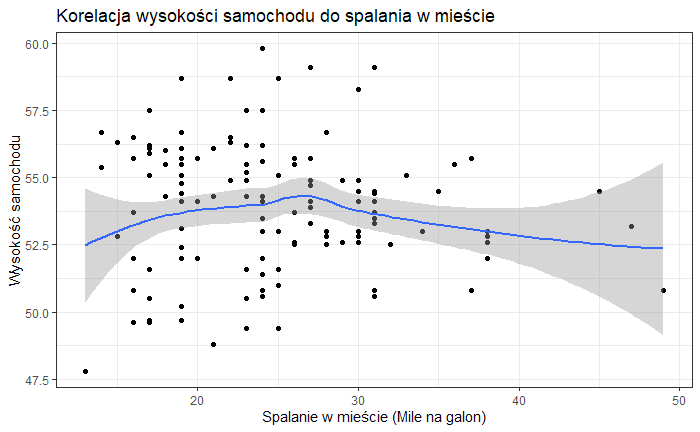


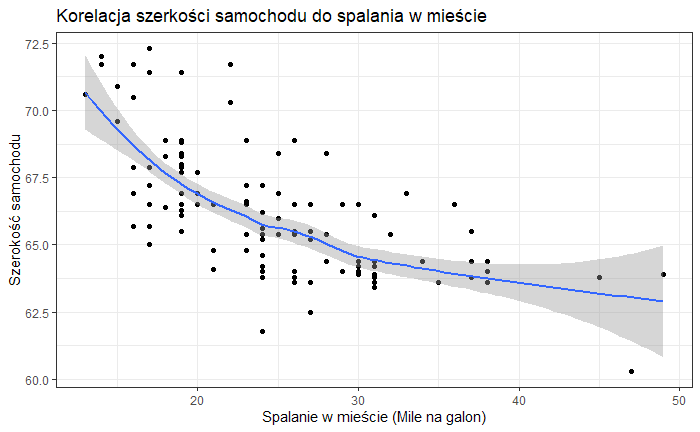
Wnioski:

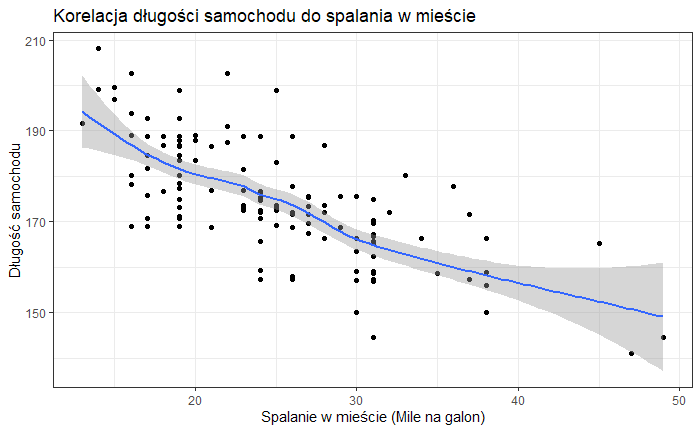
Jak możemy zauważyć średnie spalanie samochodów podczas jazdy po autostradzie znacząco się różni. Samochody są bardziej ekonomiczne gdy mają mniejsza szerokość oraz długość nadwozia. Wysokość nie wpływa w znaczący sposób na spalanie samochodu.

## Korelacja spalania samochodów od wymiarów samochodu podczas jazdy w mieście.

W tym przypdaku duże znaczenie ma prędkość więć będziemy rozważać tylko jazdę po autostradzie. Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa. Wysokość, szerokość oraz długość jest wyrażona w calach.







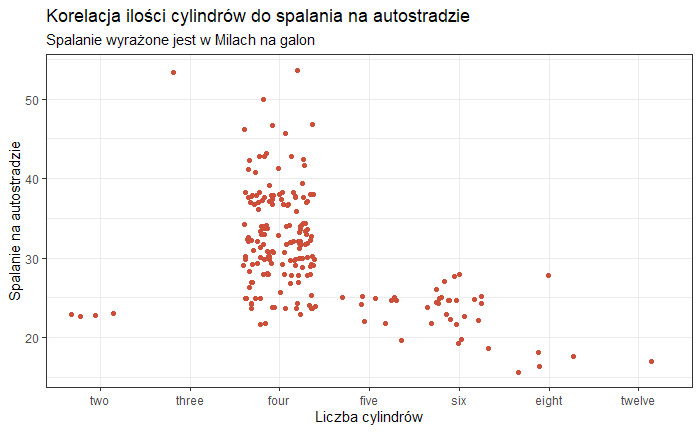
Wnioski:

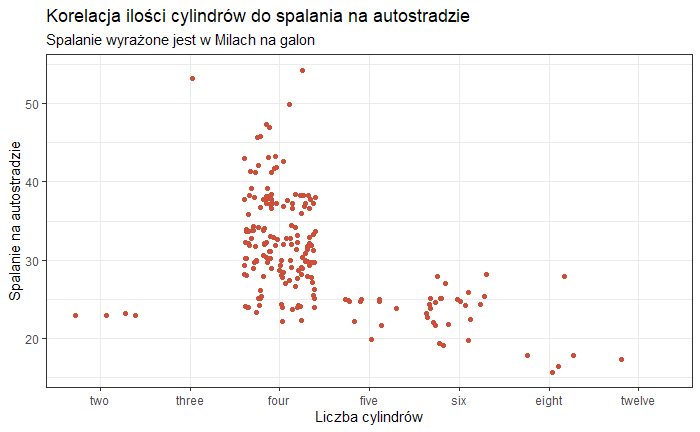
Jak możemy zauważyć średnie spalanie samochodów podczas jazdy w mieście znacząco się różni. Samochody są bardziej ekonomiczne gdy mają mniejsza szerokość oraz długość nadwozia. Wysokość nie wpływa w znaczący sposób na spalanie samochodu.

## Korelacja spalania samochodów od ilości cylindrów.

Będziemy rozważać koreleacje ilości spalonego paliwa podczas jazdy w mieście oraz po autostradzie do ilości cylindrów.

Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa. Ilość cylindrów jest wyrażona w angielskich nazwach.



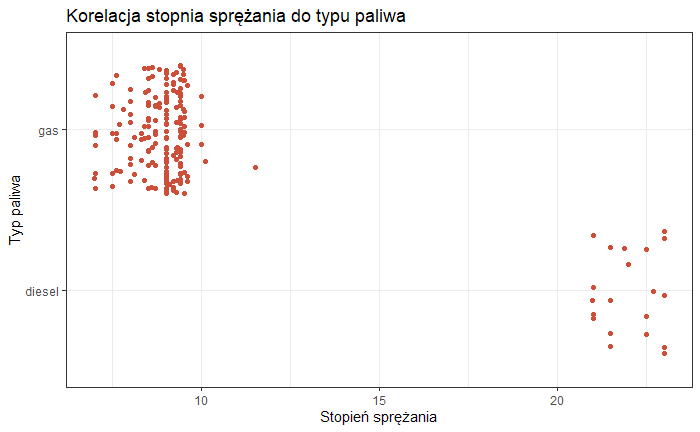


Wnioski:

Jak możemy zauważyć najczęściej występującą ilością cylindrów jest 4(four). Silniki o mniejszej oraz większej ilości cylindrów nie są aż tak popularne. Im więcej cylindrów tym silnik jest mniej ekonomiczny.

## Korelacja stopnia sprężania do użytego paliwa.

Stopień sprężania występuje bez jednostki (bez miana). W aktualnie produkowanych silnikach o zapłonie iskrowym stopień sprężania wynosi: ε = 8 ÷ 12 a w silnikach z zapłonem samoczynnym: ε = 14 ÷ 23.



Wnioski:

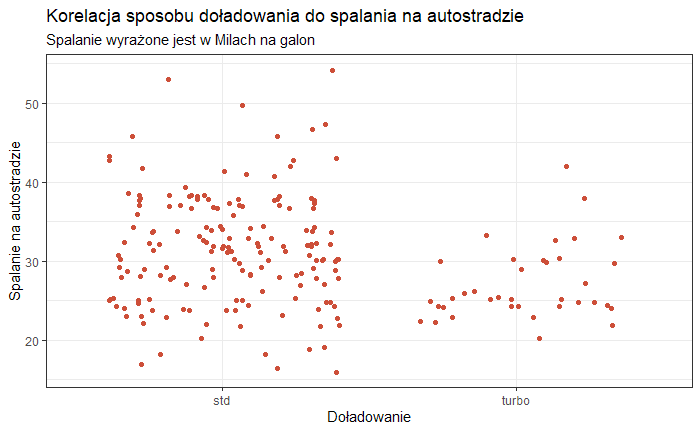
Silniki diesla maja dużo wyższy współczynnik sprężania od silników benzynowych.

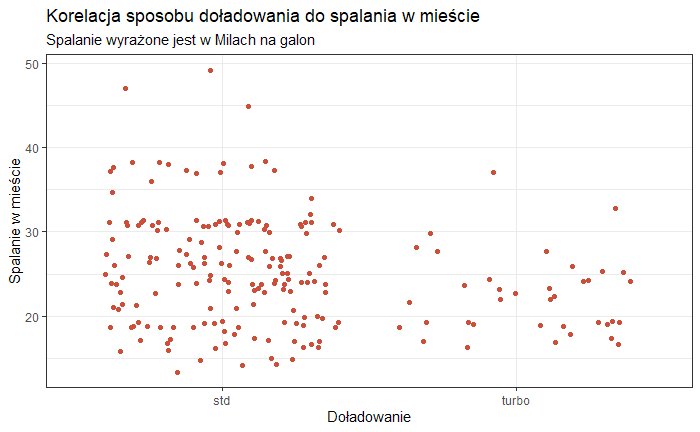
## Korelacja spalania samochodów do sposobu doładowania podczas jazdy na autostradzie oraz mieście.

Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa.

STD- jest to silnik atmosferyczny pracujący z ciśnieniem powietrza ok. 1000hPa.

Turbo- jest to silnik turbo doładowany pracujący z ciśnieniem wyższym od standardowego >1000hPa.





Wnioski:

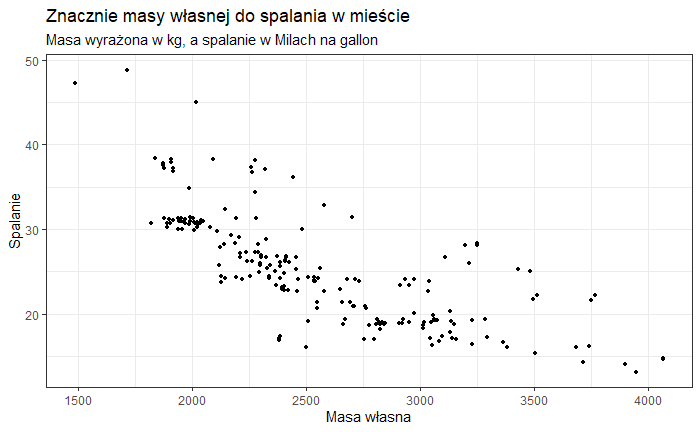
Mniejsze spalanie osiągają silniki standardowe.

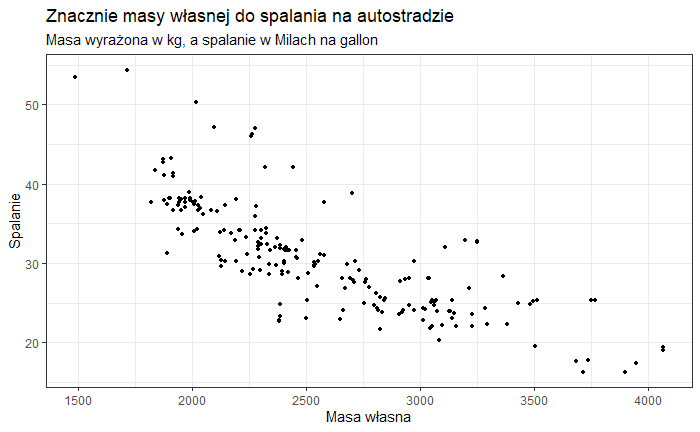
## Korelacja spalania samochodów do masy własnej podczas jazdy na autostradzie oraz mieście.

Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa.

Masa własna jest wyrażona w funtach

1 funt =0.40951718 kilogramów.





Wnioski:

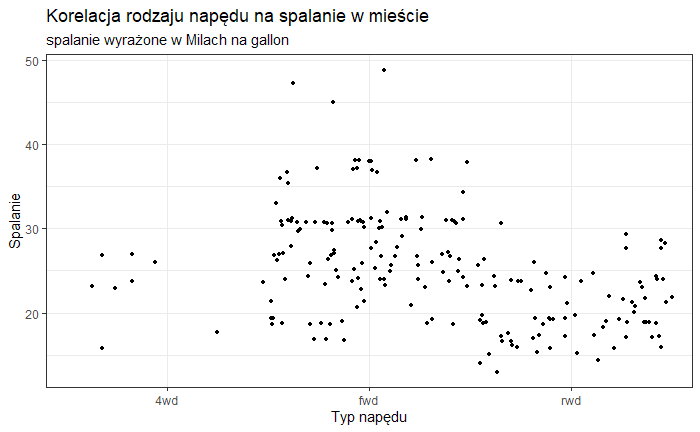
Masa własna samochodu ma bezpośrednie znaczenie na spalanie samochodu zarówno w mieście jak i na autostradzie. Im wyższa masa samochodu tym wyższe jest jego spalanie.

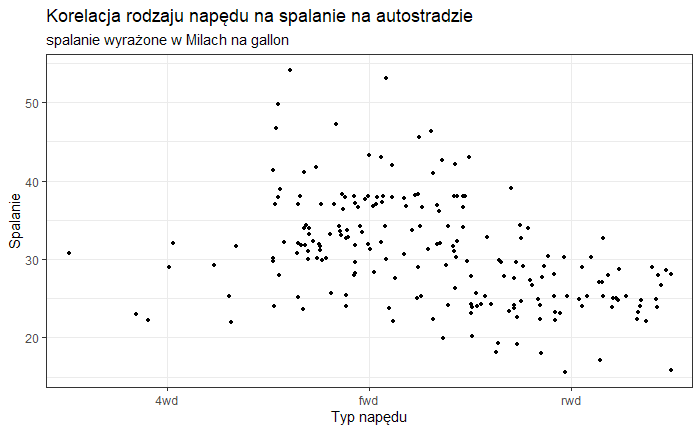
## Korelacja spalania samochodów do rodzaju napędu jazdy na autostradzie oraz mieście.

Spalanie jest wyrażone w milach na galon paliwa (MPG)- im większa liczba MPG tym samochód spala więcej paliwa.

Typ napędu:

* 4wd- napęd na 4 koła
* Fwd- napęd na 2 przednie koła
* Rwd – napęd na 2 tylnie koła





Wnioski:

Rodzaj napędu ma znacznie na spalanie samochodu. Najbardziej ekonomicznym napędem jest napęd „Fwd”. Napęd „4wd” jest najbardziej nieekonomicznym. Powoduje znaczące zwiększenie spalania. Pośrednim napędem jest „Rwd”.

# Wyznaczyć cechę decyzyjną i cechy objaśniające.

Cechą decyzyjną będzie spalanie w mieście lub na autostradzie, gdyż ma korelacje z cechami typu:

pojemność silnika, szerokość samochodu, długość samochodu, liczba cylindrów, waga własna oraz typ napędu.

# Oszacować regresję liniową cechy decyzyjnej na cechy objaśniające.

Za pomocą języka R oszacujemy regresje liniową dla spalania w mieście oraz na autostradzie.

Kod:

regresja.liniowa.miasto **=** lm**(**city.mpg**~**engine.size **+** num.of.cylinders **+** curb.weight **+** width**+** length **+** drive.wheels, data **=** dataset**)**

summary**(**regresja.liniowa.miasto**)**

cat**(**"\n\n\n"**)**

regresja.liniowa.autostrada **=** lm**(**highway.mpg**~**engine.size **+** curb.weight **+** width **+** length**+** drive.wheels **+** num.of.cylinders, data **=** dataset**)**

summary**(**regresja.liniowa.autostrada**)**

Wydruk z konsoli:

Residuals**:**

Min 1Q Median 3Q Max

**-**6.1075 **-**2.6441 **-**0.5083 0.7940 15.2561

Coefficients**:**

Estimate Std. Error t value Pr**(>|**t**|)**

**(**Intercept**)** 24.175449 18.133487 1.333 0.184048

engine.size **-**0.044013 0.021613 **-**2.036 0.043078 **\***

num.of.cylindersthree 23.896529 4.558885 5.242 4.17e**-**07 **\*\*\***

num.of.cylindersfour 12.398136 2.280411 5.437 1.64e**-**07 **\*\*\***

num.of.cylindersfive 11.033592 2.800777 3.939 0.000114 **\*\*\***

num.of.cylinderssix 11.159982 2.854726 3.909 0.000128 **\*\*\***

num.of.cylinderseight 13.863359 4.068536 3.407 0.000799 **\*\*\***

num.of.cylinderstwelve 16.183365 6.132495 2.639 0.008999 **\*\***

curb.weight **-**0.005449 0.002074 **-**2.627 0.009304 **\*\***

width 0.313968 0.337602 0.930 0.353541

length **-**0.084997 0.054470 **-**1.560 0.120305

drive.wheelsfwd 2.782094 1.516847 1.834 0.068183 .

drive.wheelsrwd 2.641781 1.551366 1.703 0.090210 .

**---**

Signif. codes**:** 0 ‘**\*\*\***’ 0.001 ‘**\*\***’ 0.01 ‘**\***’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error**:** 3.925 on 192 degrees of freedom

Multiple R**-**squared**:** 0.6613, Adjusted R**-**squared**:** 0.6401

F**-**statistic**:** 31.24 on 12 and 192 DF, p**-**value**:** **<** 2.2e**-**16

Residuals**:**

Min 1Q Median 3Q Max

**-**6.9740 **-**2.0826 **-**0.3861 1.1207 14.0419

Coefficients**:**

Estimate Std. Error t value Pr**(>|**t**|)**

**(**Intercept**)** 27.746546 17.661645 1.571 0.117828

engine.size **-**0.036822 0.021050 **-**1.749 0.081847 .

curb.weight **-**0.006861 0.002020 **-**3.396 0.000831 **\*\*\***

width 0.392516 0.328817 1.194 0.234060

length **-**0.088021 0.053053 **-**1.659 0.098726 .

drive.wheelsfwd 3.988117 1.477378 2.699 0.007565 **\*\***

drive.wheelsrwd 3.552920 1.510998 2.351 0.019718 **\***

num.of.cylindersthree 22.471089 4.440261 5.061 9.74e**-**07 **\*\*\***

num.of.cylindersfour 11.361296 2.221073 5.115 7.56e**-**07 **\*\*\***

num.of.cylindersfive 9.382827 2.727899 3.440 0.000715 **\*\*\***

num.of.cylinderssix 10.552933 2.780445 3.795 0.000198 **\*\*\***

num.of.cylinderseight 11.488387 3.962670 2.899 0.004178 **\*\***

num.of.cylinderstwelve 13.965921 5.972924 2.338 0.020406 **\***

**---**

Signif. codes**:** 0 ‘**\*\*\***’ 0.001 ‘**\*\***’ 0.01 ‘**\***’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error**:** 3.823 on 192 degrees of freedom

Multiple R**-**squared**:** 0.71, Adjusted R**-**squared**:** 0.6919

F**-**statistic**:** 39.17 on 12 and 192 DF, p**-**value**:** **<** 2.2e**-**16

Dla miasta:

Wielokrotność R^2 wynosi 0,6613, podczas gdy dostosowana wartość to 0,6401

Oznacza to, że > 64% zmienności można wyjaśnić przez pojemność silnika, szerokość samochodu, długość samochodu, liczba cylindrów, waga własna oraz typ napędu

Dla autostrady:

Wielokrotność R^2 wynosi 0,71, a skorygowana wartość to 0,6919

Oznacza to, że > 69% zmienności można wyjaśnić przez pojemność silnika, szerokość samochodu, długość samochodu, liczba cylindrów, waga własna oraz typ napędu

# Wyznaczyć efektywność zbudowanego modelu na podstawie współczynnika determinacji i średniej kwadratowej błędów RMSE (Root Mean Squared Error).

Kod:

#Dla miasta

predicted\_data **<-** predict**(**regresja.liniowa.miasto, dataset, interval **=** 'confidence'**)**

df\_predicted\_data **<-** as.data.frame.matrix**(**predicted\_data**)**

df\_mpg\_data **<-** as.data.frame.matrix**(**dataset**)**

sqrt**(**mean**((**df\_mpg\_data**$**city.mpg **-** df\_predicted\_data**$**fit**)^**2**))**

#Dla autostrady

predicted\_data2 **<-** predict**(**regresja.liniowa.autostrada, dataset, interval **=** 'confidence'**)**

df\_predicted\_data2 **<-** as.data.frame.matrix**(**predicted\_data2**)**

df\_mpg\_data2 **<-** as.data.frame.matrix**(**dataset**)**

sqrt**(**mean**((**df\_mpg\_data2**$**highway.mpg **-** df\_predicted\_data**$**fit**)^**2**))**

Wydruk z konsoli:

3.798161 – dla miasta

6.688898- dla autostrady

Polepszyć efektywność modelu za pomocą walidacji krzyżowej k-krotnej (K-Fold Cross-Validation).

Została użyta 10 krotna walidacja krzyżowa.

library**(**caret**)**

ctrl **<-** trainControl**(**method **=**"cv", number **=** 10**)**

model **<-** train**(**city.mpg**~**engine.size **+** num.of.cylinders **+** curb.weight **+** width **+** engine.size **+** height **+** drive.wheels, data **=** df\_mpg\_data,

method **=** "lm", trControl **=** ctrl**)**

print**(**model**)**

#Dla autostrady

ctrl2 **<-** trainControl**(**method **=**"cv", number **=** 10**)**

model2 **<-** train**(**highway.mpg**~**engine.size **+** num.of.cylinders **+** curb.weight **+** width **+** engine.size **+** height **+** drive.wheels, data **=** df\_mpg\_data2,

method **=** "lm", trControl **=** ctrl**)**

print**(**model2**)**

Wydruk z konsoli:

205 samples

6 predictor

No pre**-**processing

Resampling**:** Cross**-**Validated **(**10 fold**)**

Summary of sample sizes**:** 184, 185, 185, 185, 183, 186, ...

Resampling results**:**

RMSE Rsquared MAE

3.84689 0.6735265 2.863141

205 samples

6 predictor

No pre**-**processing

Resampling**:** Cross**-**Validated **(**10 fold**)**

Summary of sample sizes**:** 184, 184, 185, 184, 185, 185, ...

Resampling results**:**

RMSE Rsquared MAE

3.844853 0.7031896 2.77134