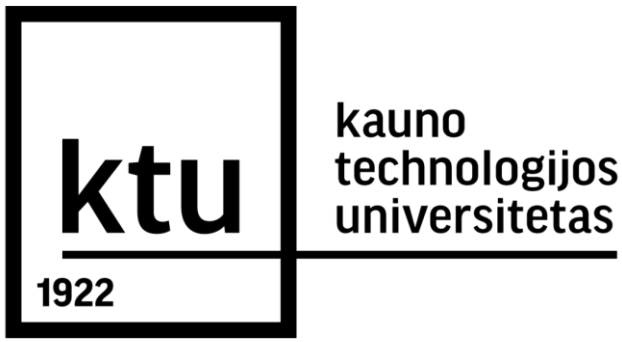
KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS



Intelektikos pagrindai (P176B101)

Laboratorinis darbas Nr.1

Atliko:

IFF-8/3 gr. studentas

Dovydas Zamas

2021 m. kovo 4 d.

Priėmė:

Lekt. PaulauskaitėTarasevičienė Agnė

KAUNAS 2021

Turinys

1. Duomenų rinkinys

Laboratoriniui darbui pasirinktas automobilių specifikacijų rinkinys. Iš automobilio duomenų rinkinio buvo pašalinti unikalūs atributai, pvz., „car_ID“, „CarName“. Modifikuotą automobilio duomenų rinkinį sudaro šie atributai: „symboling“[1], „drivewheel“[2], „enginelocation“[3], „cylindernumber“[4], „enginesize“[5], „compressionratio“[6], „horsepower“[7], „citympg“[8], „highwaympg“[9].

2. Duomenų rinkinio kokybės analizė

Atributo pavadinimas	Kiekis	Trūkstamos reikšmės, %	Kardinalumas	Min. reikšmė	Max. Reikšmė	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Stand. Nuokrypis
enginesize	841	0.23	42	-1	326	98	146	128.64	121	30.12
compressionratio	839	0.47	32	7	23	8.7	9.5	10.23	9	3.97
horsepower	839	0.47	59	48	288	70	116	107.25	95	42.09
citympg	843	0	29	13	49	19.5	30	25.01	24	6
highwaympg	839	0.47	30	16	54	25	34	30.68	30	6.27

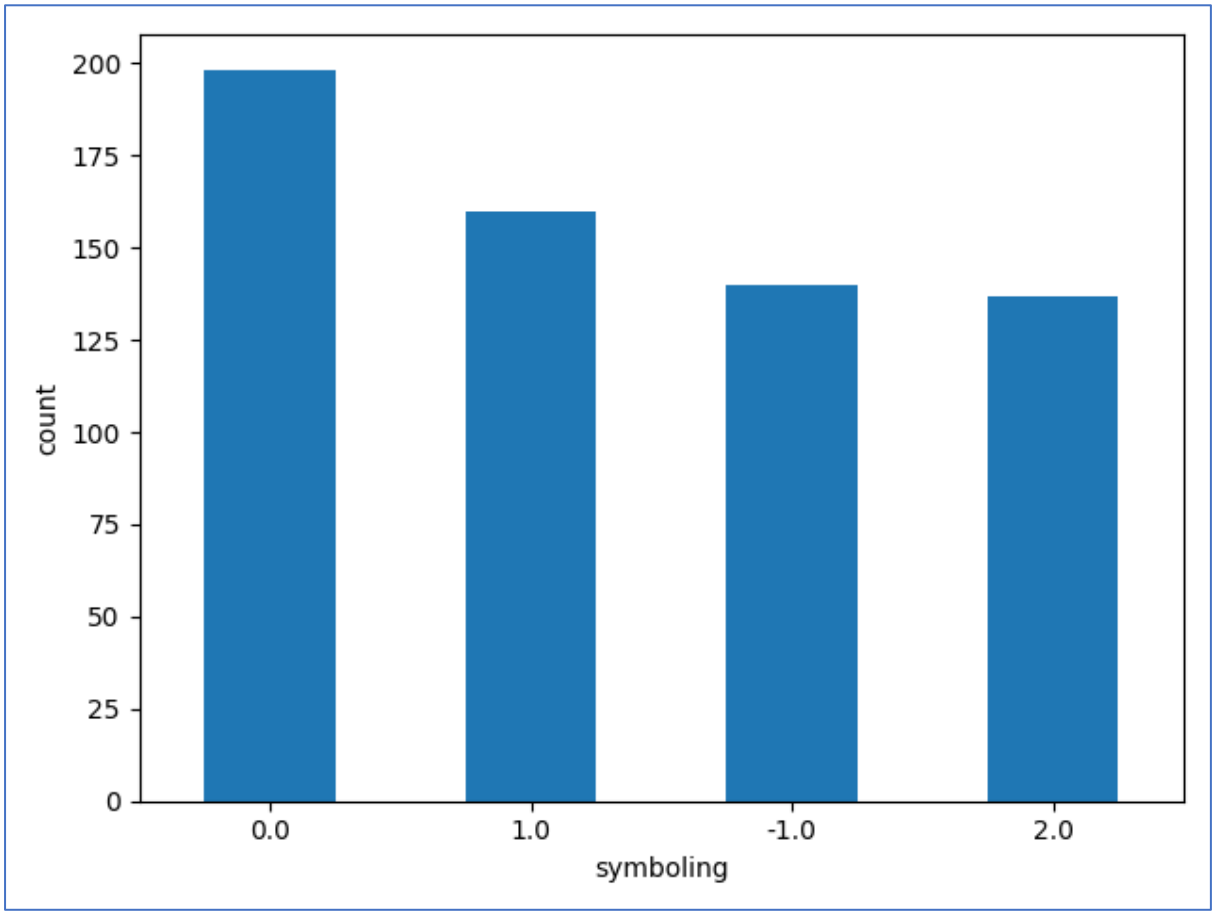
pav. 1 Tolydinio tipo atributų kokybės analizės lentelė

Atributo pavadinimas	Kiekis	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji moda	2-osios modo dažnumas	2-oji moda, %
Symboling	835	0.95	6	0	225	26.95	1	168	20.12
DriveWheel	843	0	3	fwd	461	54.69	rwd	344	40.81
Enginelocation	843	0	2	front	735	87.19	rear	108	12.81
Cylindernumber	839	0.47	7	four	665	79.26	six	112	13.35

pav. 2 Kategorinio tipo atributų kokybės analizės lentelė

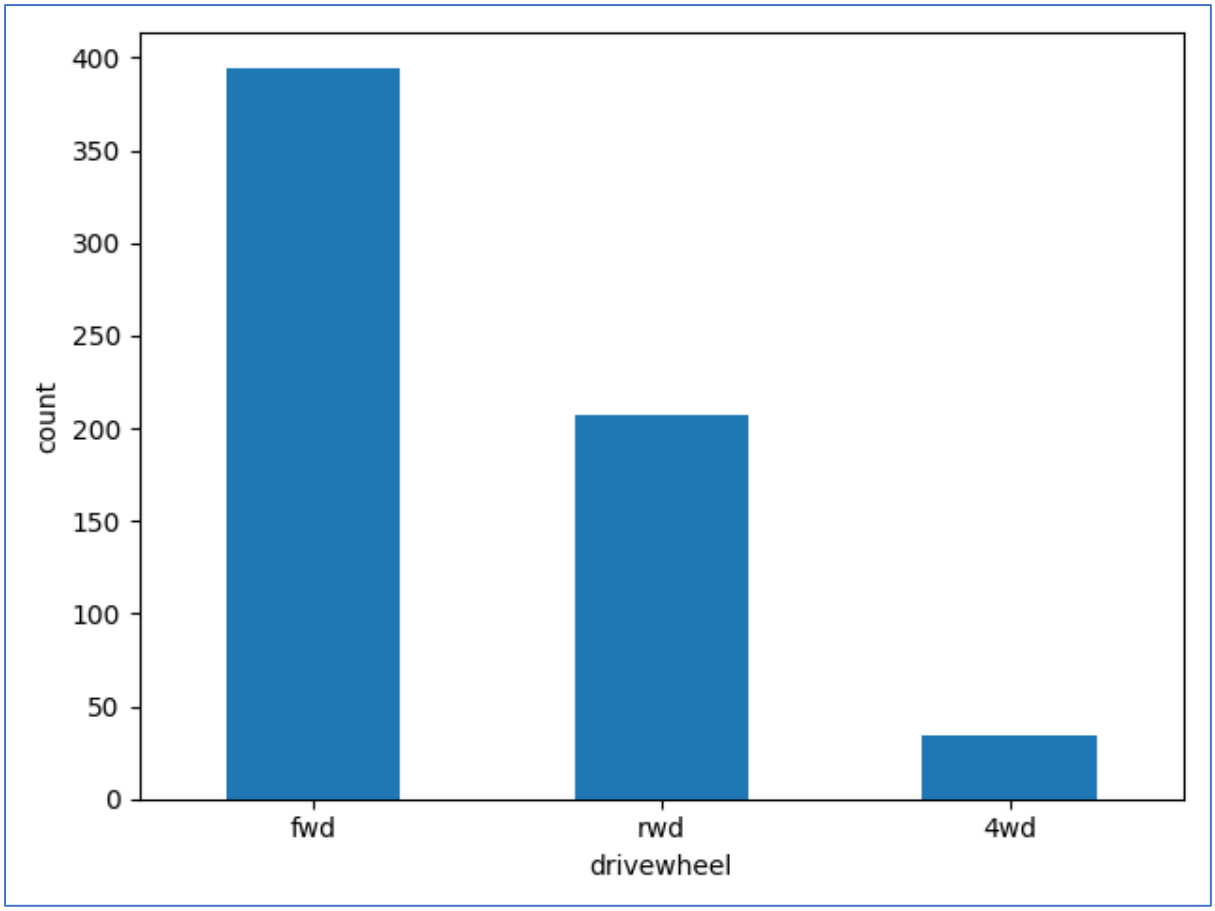
[1] Simboliai nusako mašinos rizikingumo laipsnį
[2] Varomieji ratai
[3] Variklio pozicija
[4] Cilindrų skaičius esantis mašinoje
[5] Variklio dydis
[6] Suspaudimo laipsnis
[7] Arklio galios
[8] Automobilio sąnaudos mieste
[9] Automobilio sąnaudos užmiestyje

3. Atributų histogramos



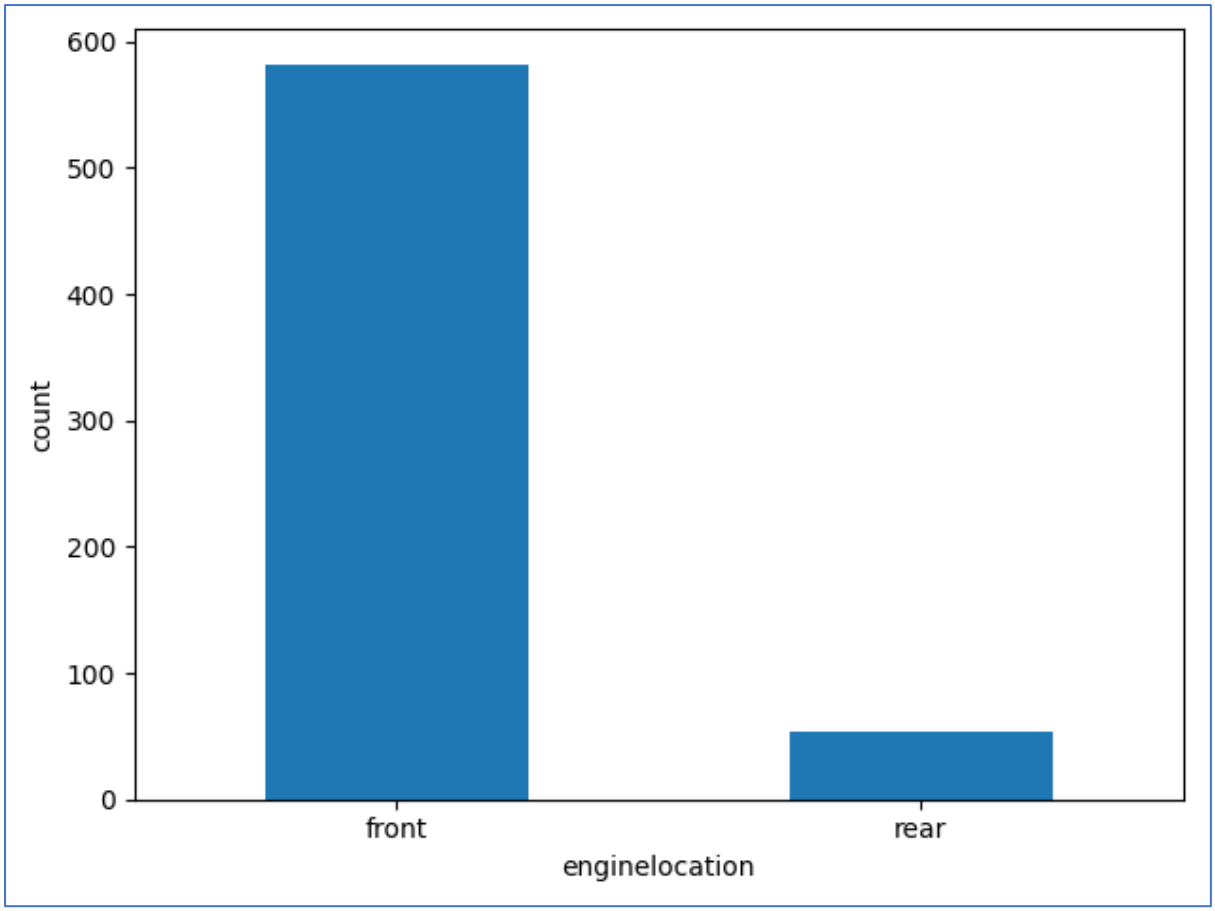
pav. 3 "Symboling" atributo histograma

Histograma (pav.3) nurodo normalųjį atributo „symboling“ reikšmių pasiskirstymą.



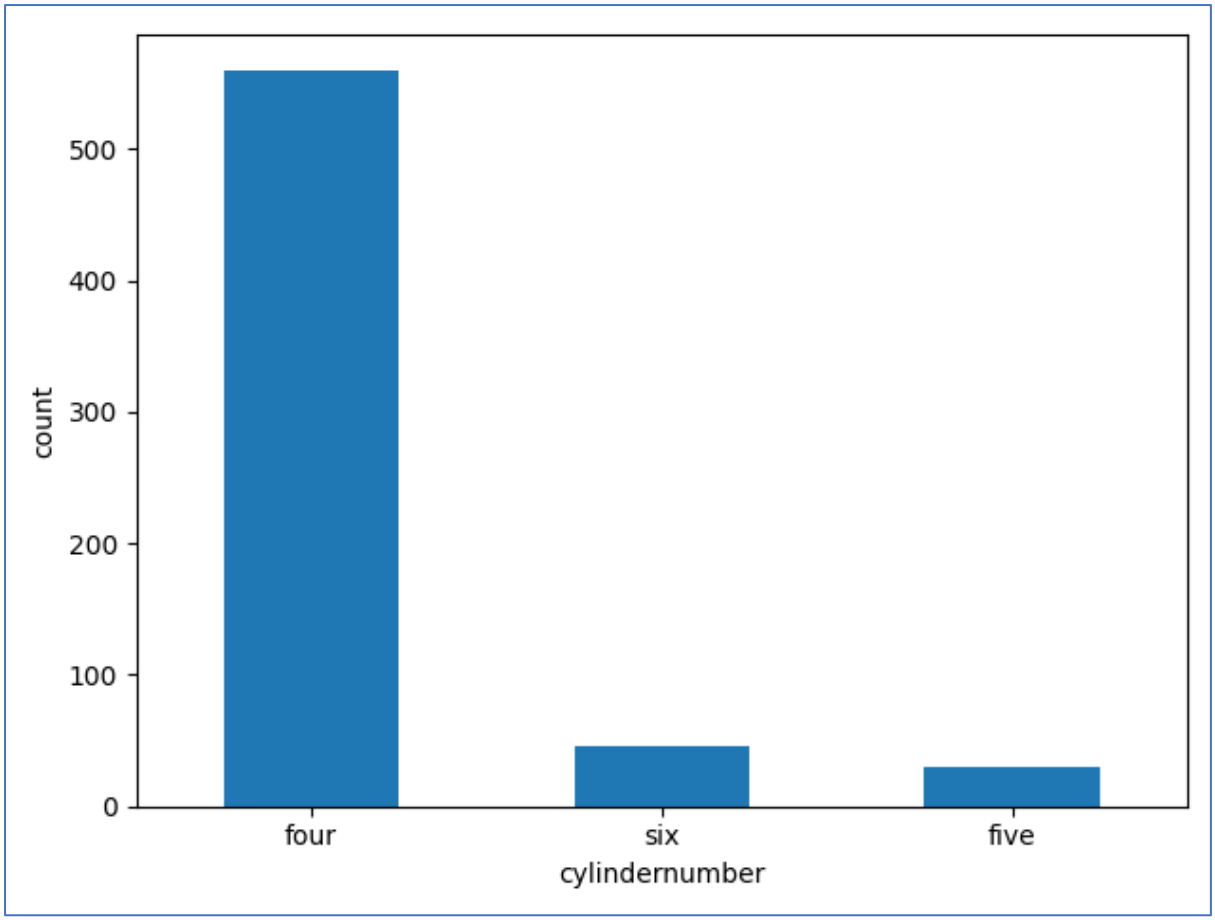
pav. 4 "drivewheel" atributo histograma

Histograma (pav. 4) nurodo normalųjį atributo “drivewheel” reikšmių pasiskirstymą.



pav. 5 "enginelocation" atributo histograma

Histograma (pav. 5) nurodo, kad duomenų rinkinyje yra didžioji dalis “front” reikšmę turintys duomenys - automobiliai su priekiniais varančiaisiais ratais.



pav. 6 "cylindernumber" atributo histograma

Histogramoje (pav. 6) matome, kad didžiausią duomenų rinkinio dalį sudaro automobiliai turintys 4 cilindrus.

4. Duomenų kokybės problemos ir sprendimai

Duomenų rinkinio atributai turėjo trūkstamų reikšmių bei išskirčių. Įrašai kurie turėjo tuščių reikšmių, bei išskirčių buvo ištrinti. Išskirčių radimui buvo pasinaudota „python“ biblioteka „pandas“, randami kvantiliai ir pagal juos atrenkami duomenys.

Kodo fragmentas išskirčių radimui ir šalinimui:

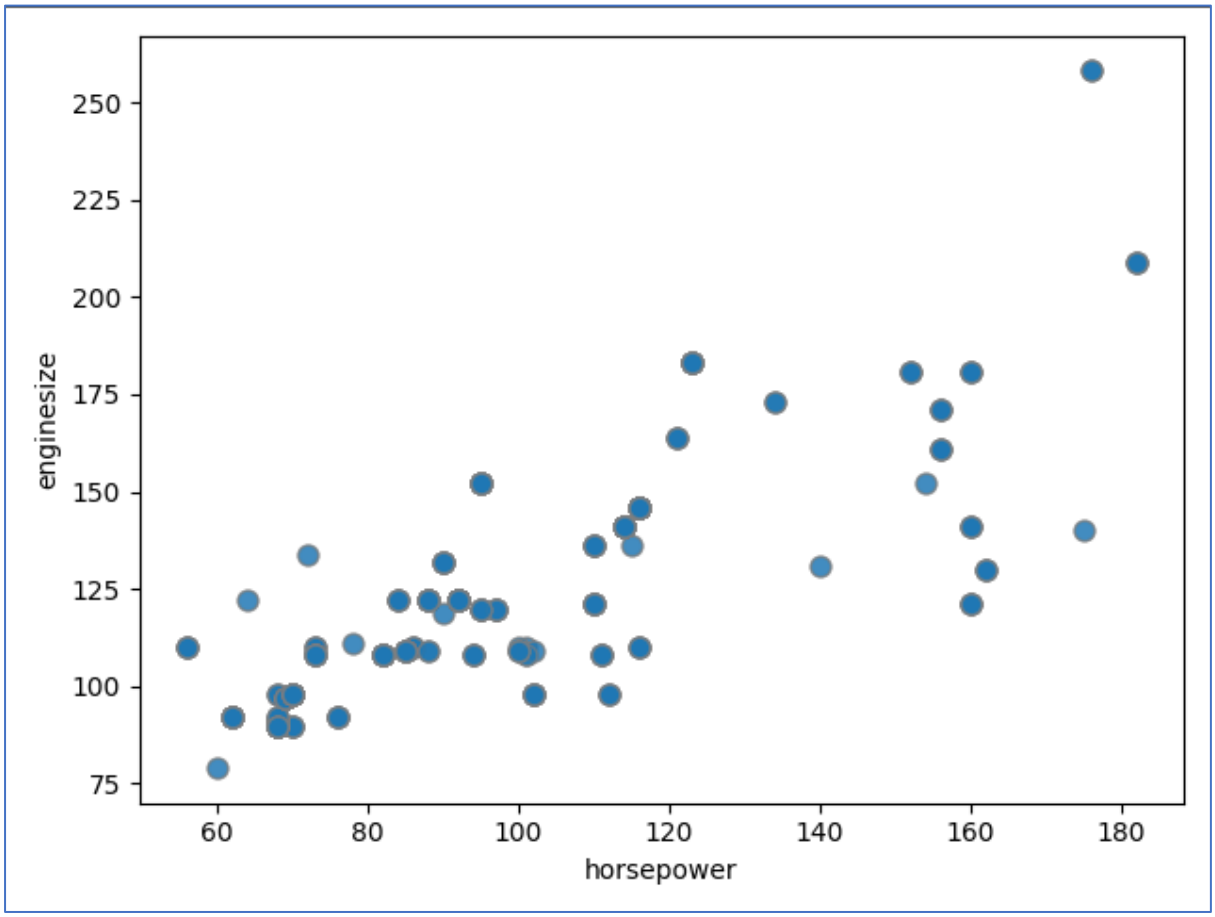
```
def deleteOutliers(data):
    q_low = data["enginesize"].quantile(0.0005)
    q_hi = data["enginesize"].quantile(0.9995)
    data = data[(data["enginesize"] < q_hi) & (data["enginesize"] > q_low)]

    q_low = data["symboling"].quantile(0.0005)
    q_hi = data["symboling"].quantile(0.9995)
    data = data[(data["symboling"] < q_hi) & (data["symboling"] > q_low)]

    q_low = data["compressionratio"].quantile(0.005)
    q_hi = data["compressionratio"].quantile(0.9995)
    data = data[(data["compressionratio"] < q_hi) &
(data["compressionratio"] > q_low)]

    q_low = data["horsepower"].quantile(0.0005)
    q_hi = data["horsepower"].quantile(0.9995)
```

```
    data = data[(data["horsepower"] < q_hi) & (data["horsepower"] > q_low)]

    q_low = data["citympg"].quantile(0.0005)
    q_hi = data["citympg"].quantile(0.9995)
    data = data[(data["citympg"] < q_hi) & (data["citympg"] > q_low)]

    q_low = data["highwaympg"].quantile(0.0005)
    q_hi = data["highwaympg"].quantile(0.9995)
    data = data[(data["highwaympg"] < q_hi) & (data["highwaympg"] > q_low)]
    return data
```

Kodo fragmentas ištrinti eilutes su trūkstamomis reikšmėmis:

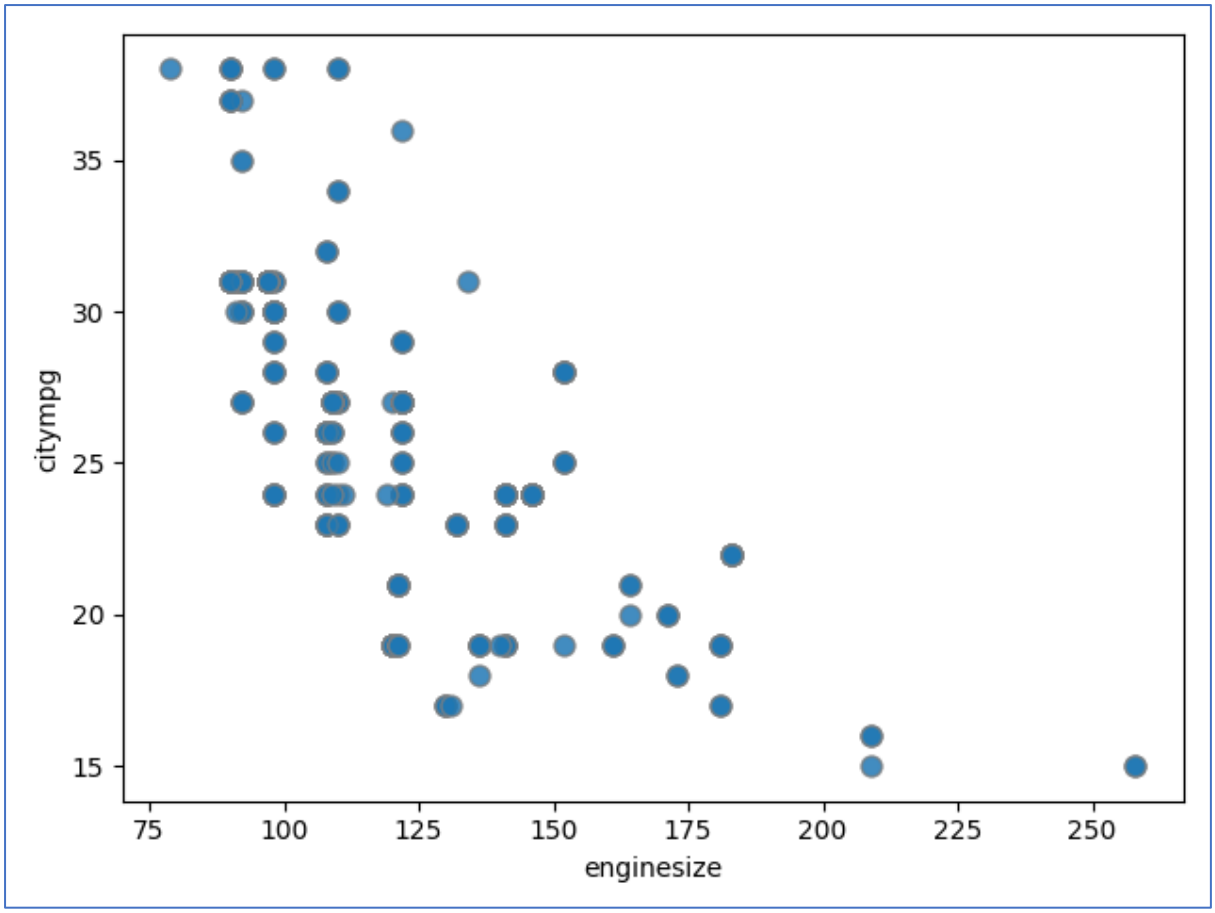
```
def deleteNaN(data):
    data = data.dropna()
```

5. Tolydinio tipo atributų vizualizacija



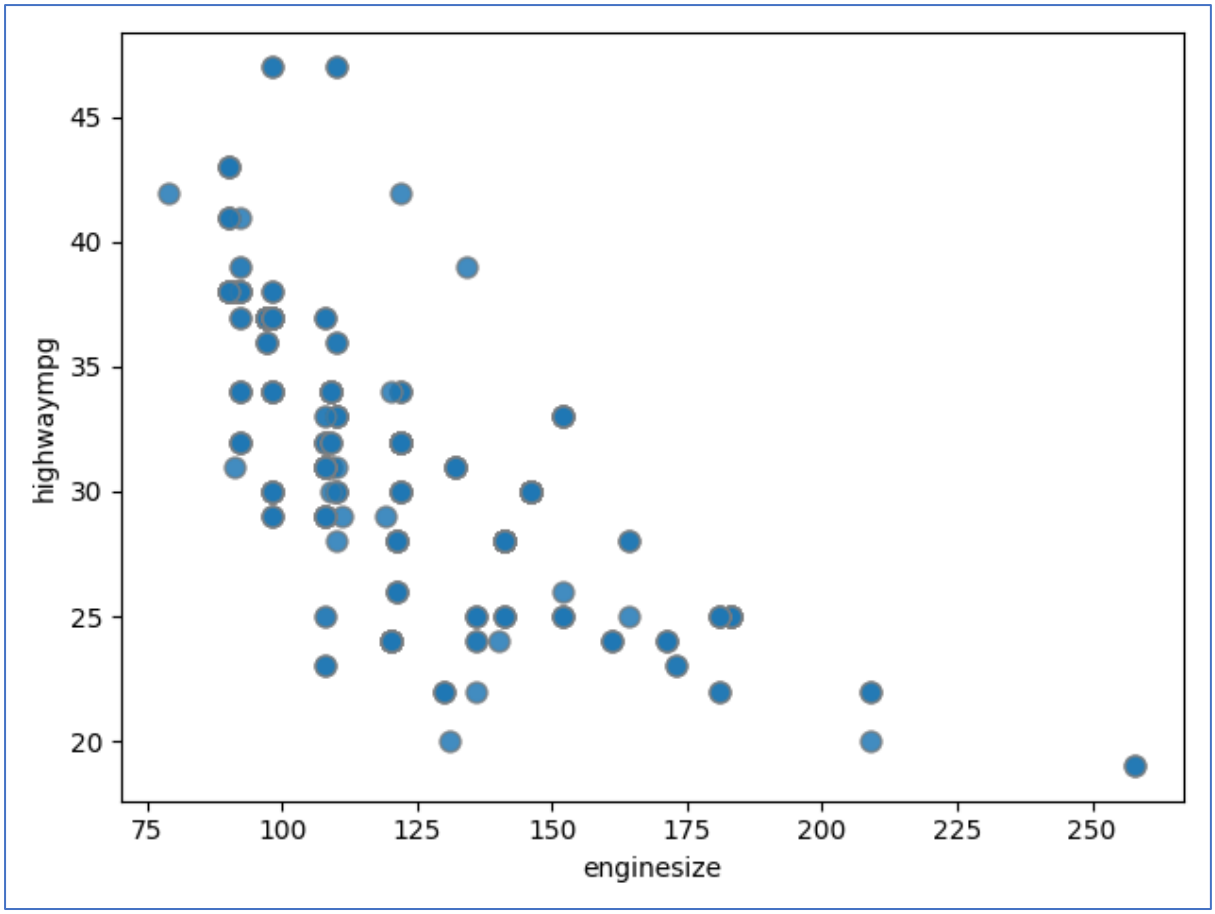
pav. 7 "enginesize" ir "horsepower" atributų 'scatter plot' diagrama

Diagrama vaizduoja teigiamą koreliaciją tarp variklio dydžio ir arklio galių – kuo variklis yra didesnis tuo daugiau turi arklio galių.



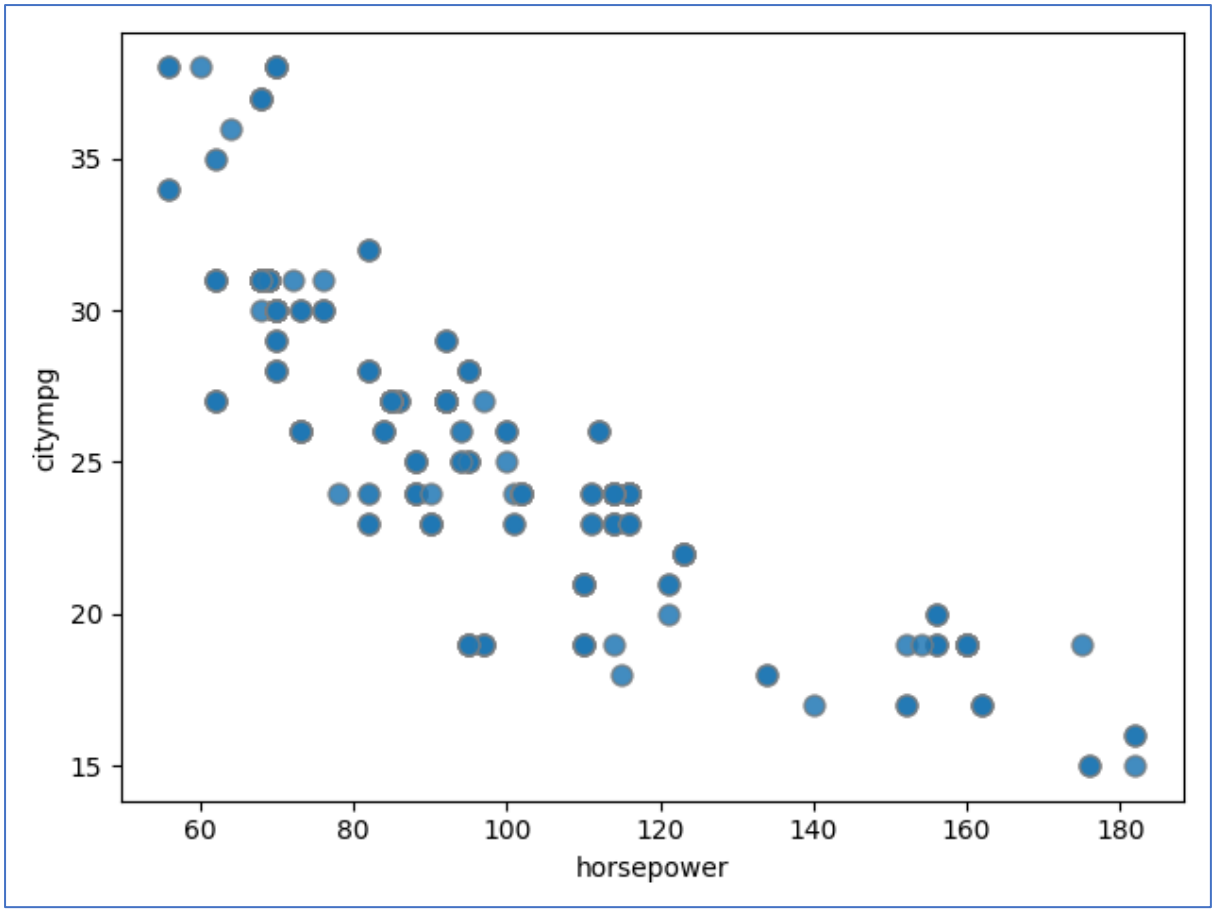
pav. 8 "citymmpg" ir "enginesize" atributų 'scatter plot' diagrama

Diagrama vaizduoja neigiamą koreliaciją tarp variklio ir kuro sąnaudų mieste – kuo variklio dydis yra didesnis tuo jis yra ekonomiškesnis, taip pat yra kriterijų, kurie yra neįvertinti ir daro įtaką variklio ekonomiškumui, pvz., suspaudimo laipsnis, ar automobilis važiuoja pastoviu greičiu.



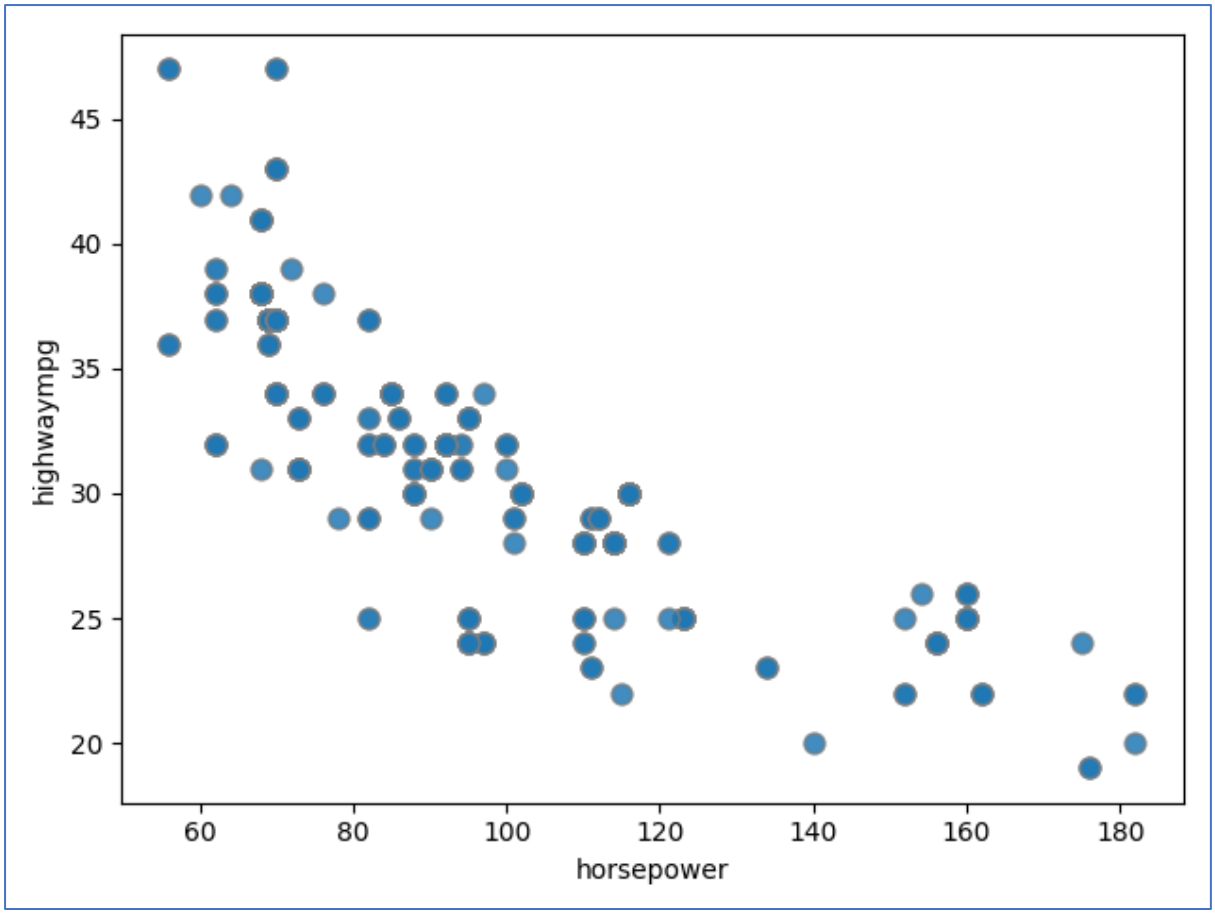
pav. 9 "highwaympg" ir "enginesize" atributų 'scatter plot' diagrama

Diagrama vaizduoja neigiamą koreliaciją tarp variklio ir kuro sąnaudų užmiestyje – kuo variklio dydis yra didesnis tuo jis yra ekonomiškesnis.



pav. 10 "citymmpg" ir "horsepower" atributų 'scatter plot' diagrama

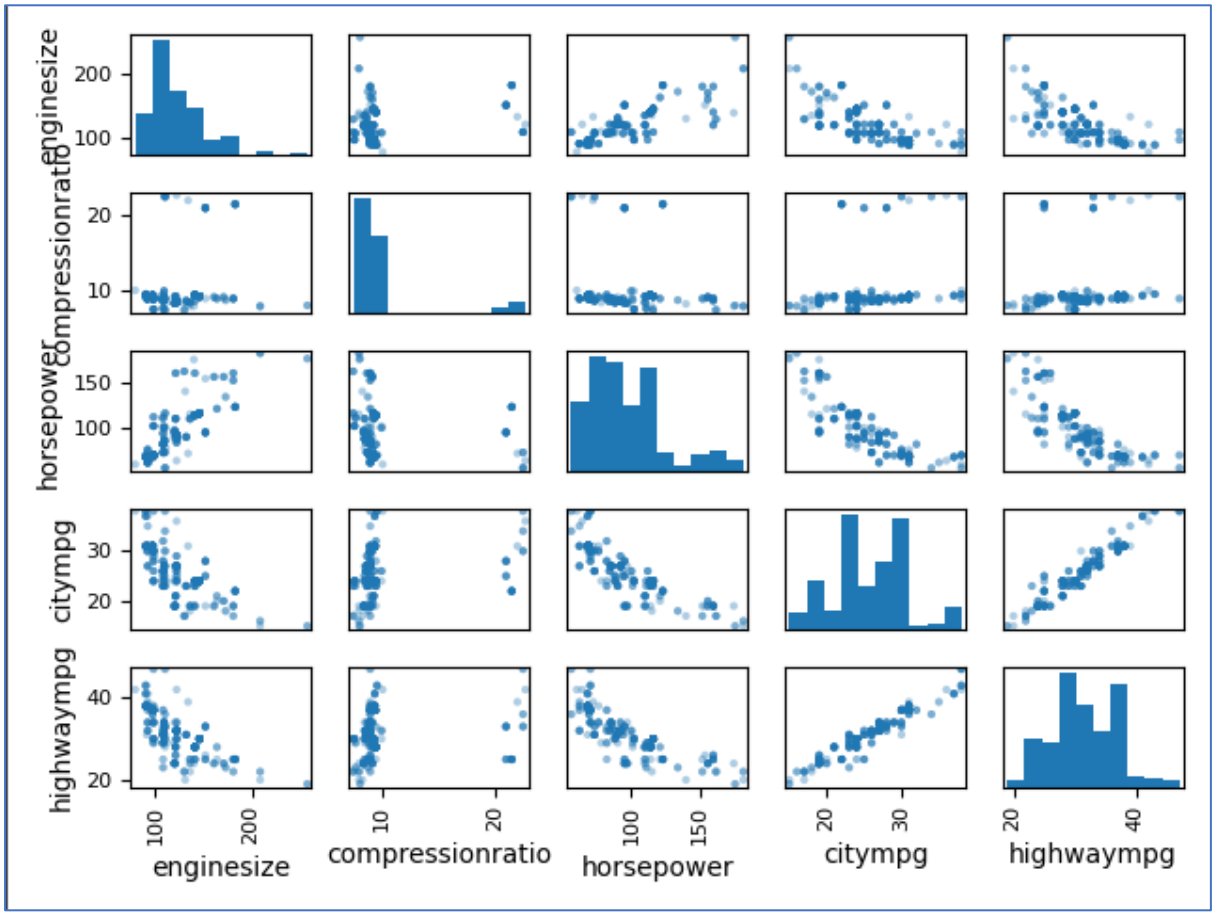
Diagrama vaizduoja neigiamą koreliaciją tarp arklio galių ir kuro sąnaudų mieste – kuo variklis turi daugiau arklio galių tuo jis yra ekonomiškesnis.



pav. 11"highwaympg" ir "horsepower" atributų 'scatter plot' diagrama

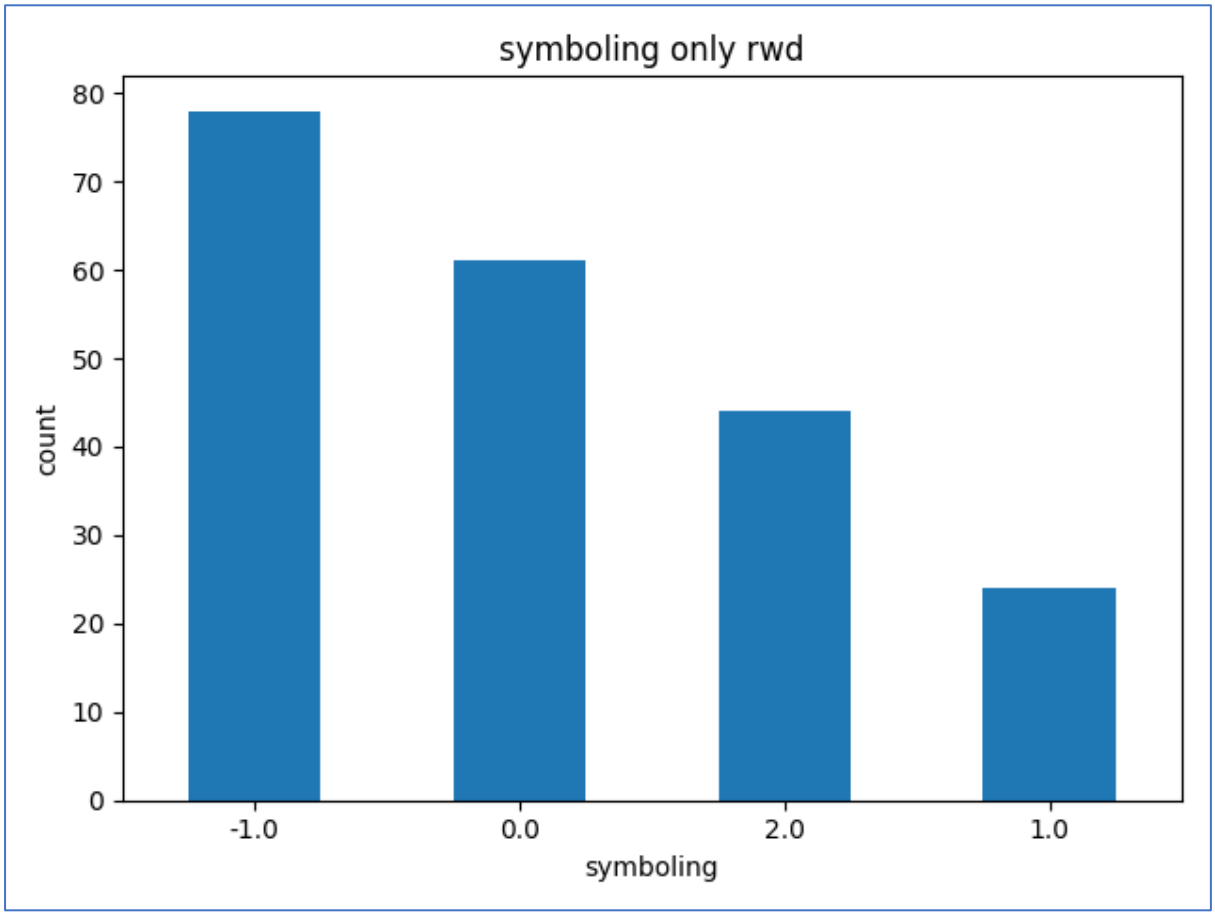
Diagrama vaizduoja neigiamą koreliaciją tarp arklio galių ir kuro sąnaudų užmiestyje – kuo variklis turi daugiau arklio galių tuo jis yra ekonomiškesnis.

6. Scatter plot Matrix diagrama

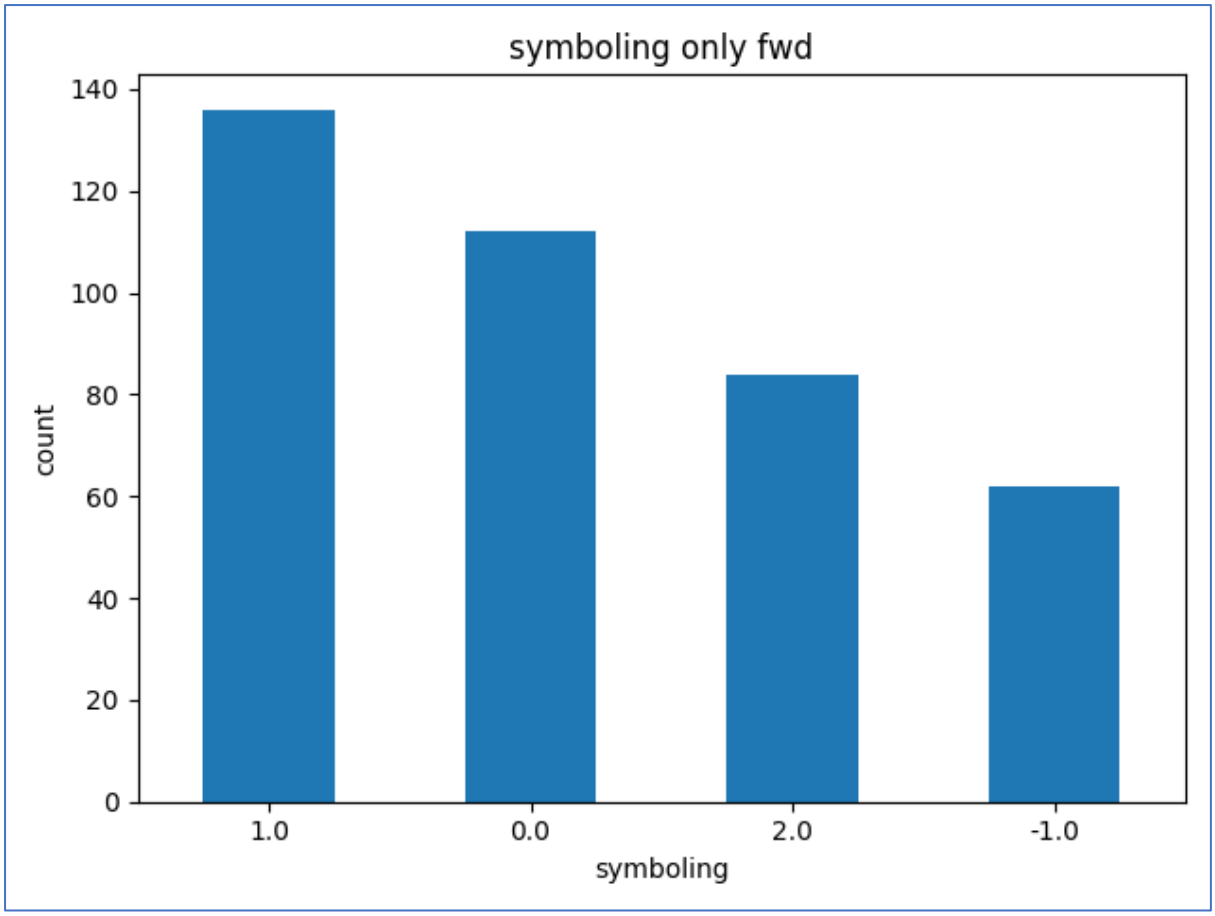


pav. 12 SPLOM diagrama

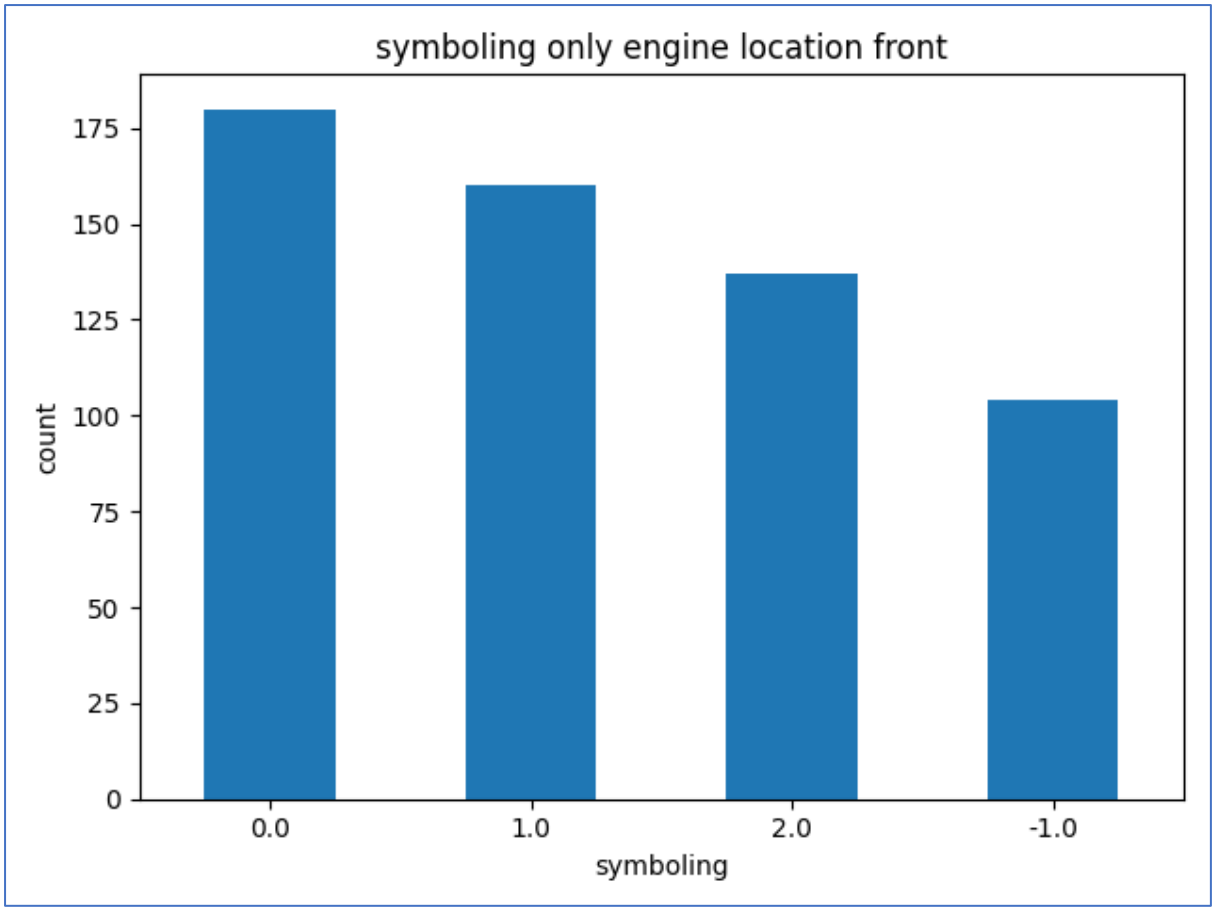
7. Kategorinio tipo atributų vizualizacija



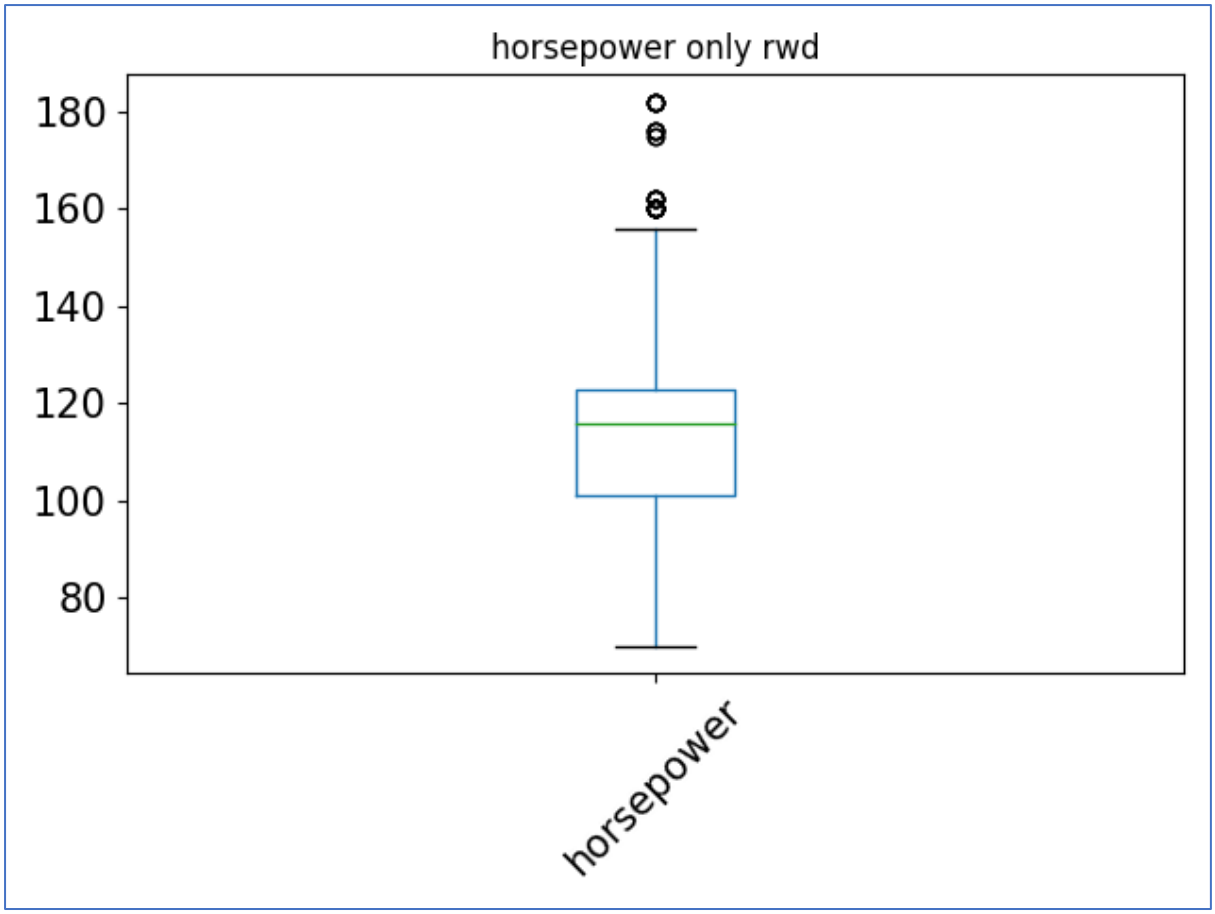
pav. 13 "symboling" ir "drivewheel" atributų 'bar plot' diagrama



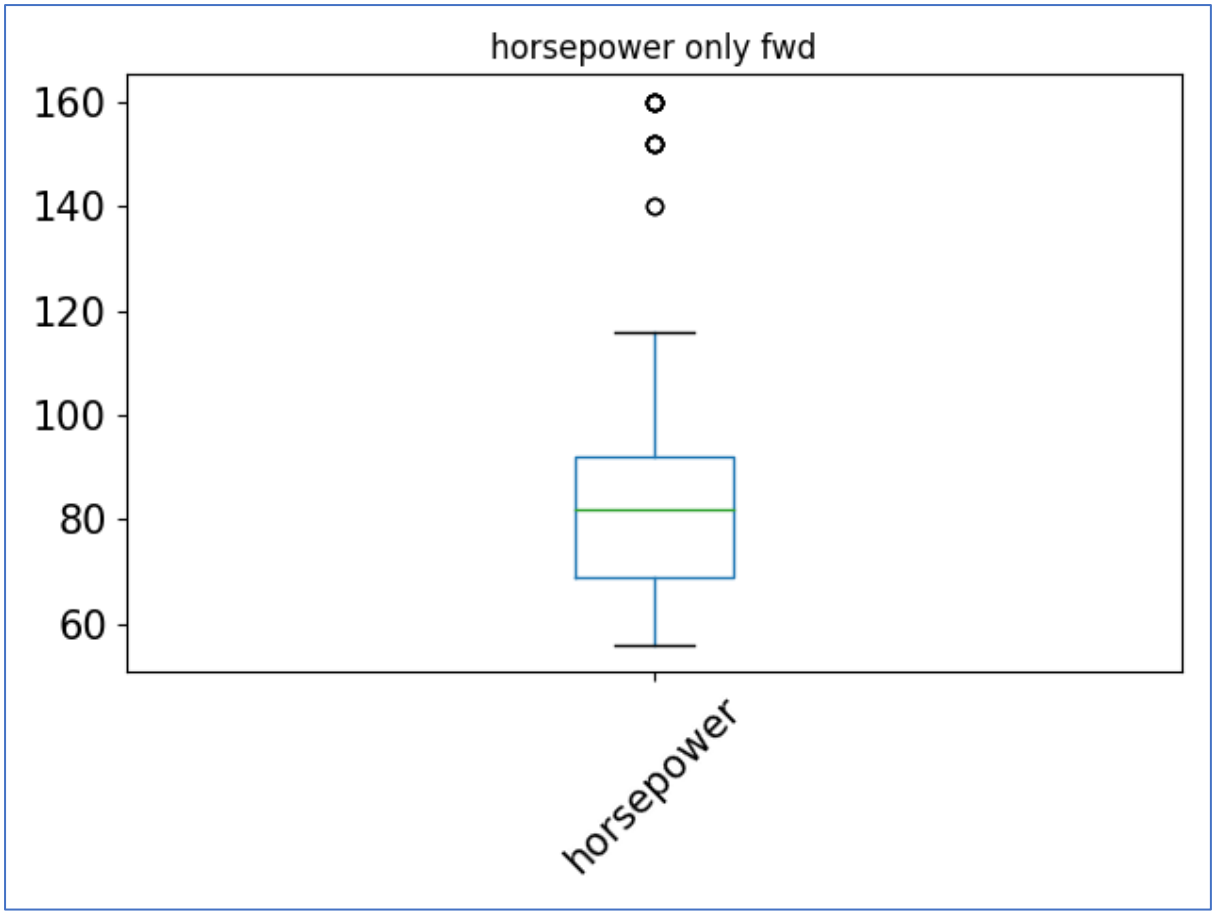
pav. 14 "symboling" ir "drivewheel" atributų 'bar plot' diagrama



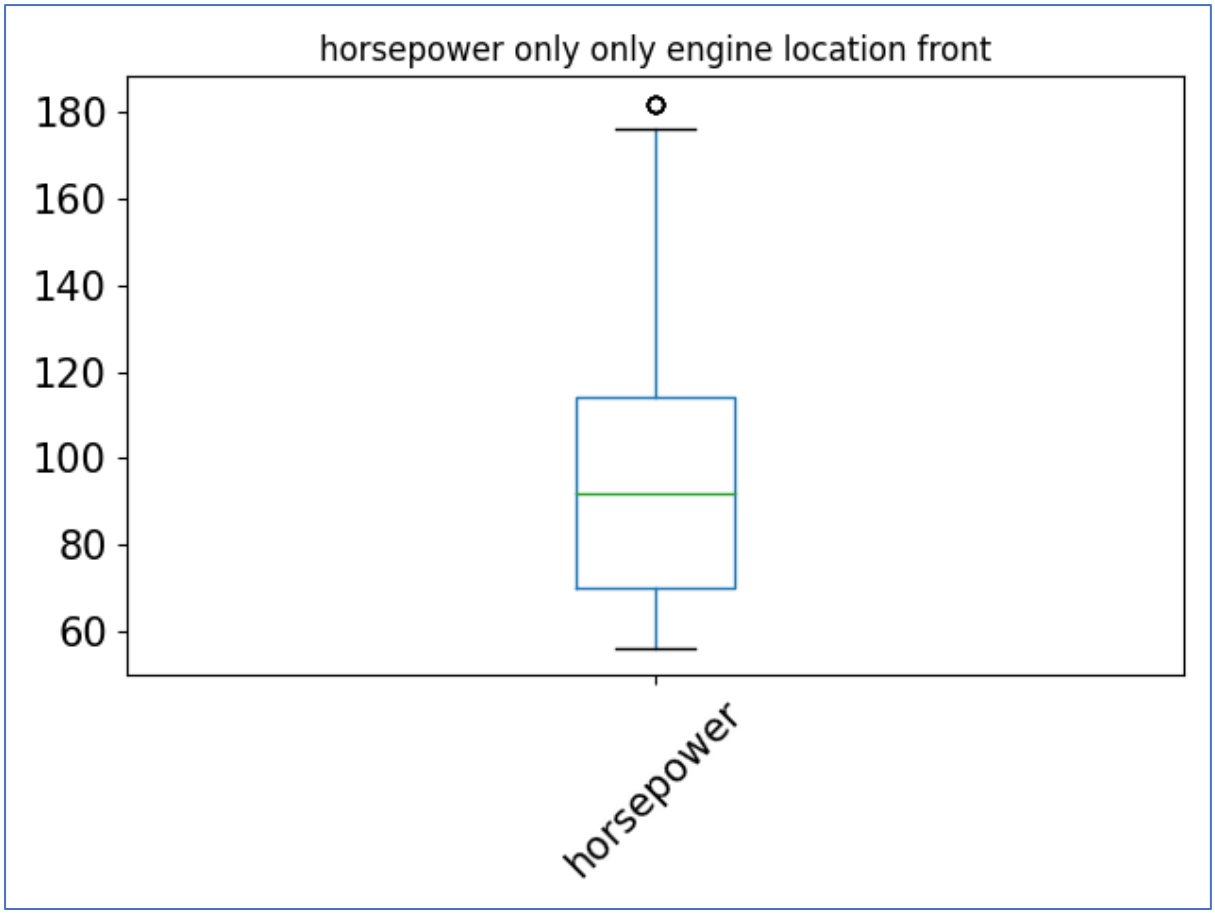
pav. 15 "symboling" ir "enginelocation" atributų 'bar plot' diagrama



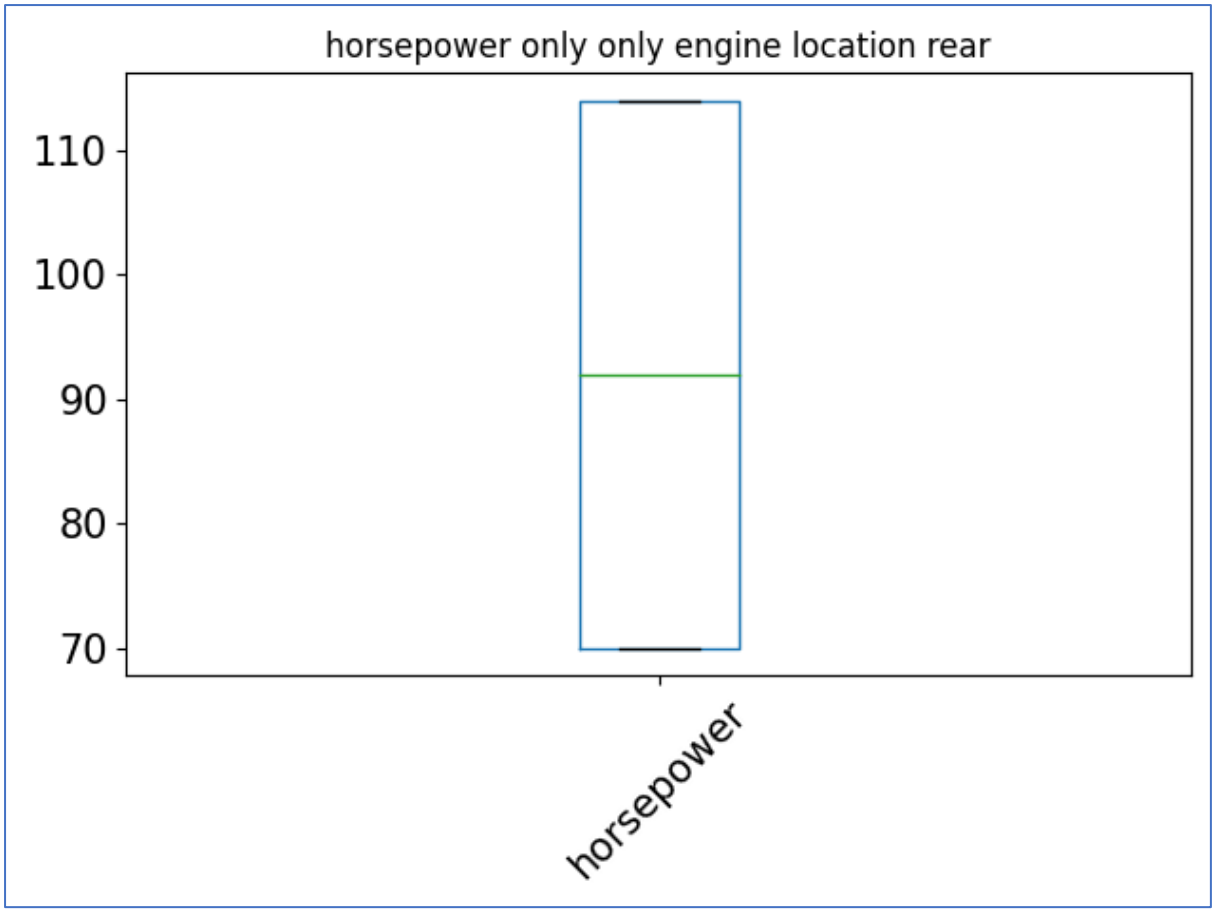
pav. 16 "horsepower" ir "wheeldrive" 'box plot' diagrama



pav. 17 "horsepower" ir "wheeldrive" 'box plot' diagrama

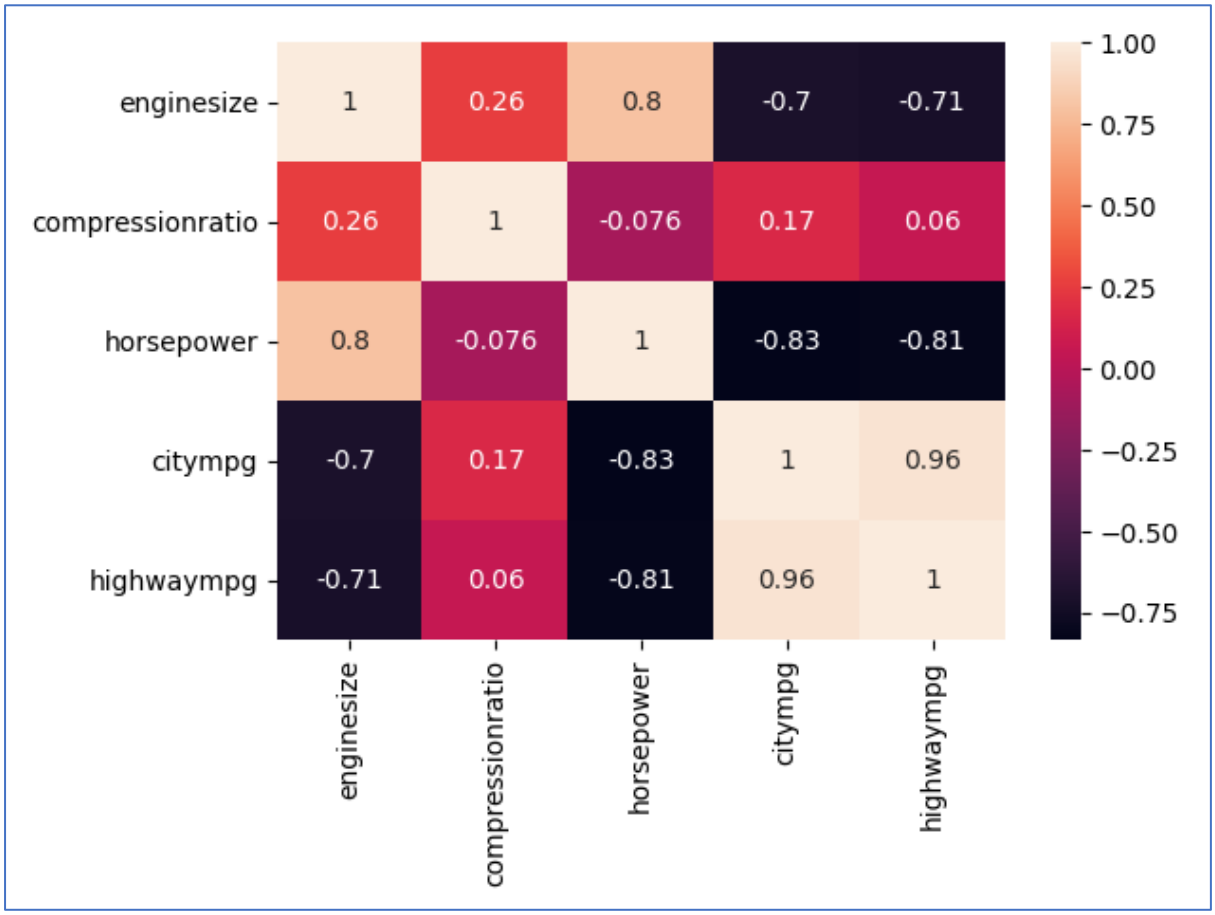


pav. 18 "horsepower" ir "enginelocation" 'box plot' diagrama



pav. 19 "horsepower" ir "wheeldrive" 'box plot' diagrama

8. Koreliacijos matrices diagrama



pav. 20 Koreliacijos matricos diagrama

Koreliacijos matrices diagram (pav. 20) vaizduoja koreliacijas tarp tolydinių atributų. Dvi atributų („enginesize“ su „highwaympg“ ir „citympg“) poros turi neigiamą koreliaciją, tačiau „-0.7“ nėra stiprus ryšys, dvi poros („enginesize“ su „horsepower“ ir „citympg“ su „highwaympg“) turi teigiamą koreliaciją, atributai “citympg” ir “highwaympg” turi stiprų ryšį.

9. Duomenų normalizacija

Programos kodas atlikti duomenų normalizacijai:

```
def normalize(data):
    result = data.copy()
    for feature_name in data.columns:
        max_value = data[feature_name].max()
        min_value = data[feature_name].min()
        result[feature_name] = (data[feature_name] - min_value) /
(max_value - min_value)
    return result
```

Programos rezultatai:

	enginesize	compressionratio	horsepower	citympg	highwaympg
16	0.240223	0.065789	0.253968	0.391304	0.392857
18	0.581006	0.921053	0.531746	0.304348	0.214286
19	0.581006	0.921053	0.531746	0.304348	0.214286
21	0.173184	0.000000	0.476190	0.347826	0.392857
22	0.240223	0.065789	0.253968	0.391304	0.392857
..	...	...	...	...	...
832	0.346369	0.131579	0.460317	0.391304	0.321429
833	0.240223	0.078947	0.285714	0.521739	0.464286
834	0.346369	0.131579	0.460317	0.391304	0.321429
835	0.106145	0.098684	0.111111	0.652174	0.642857
836	0.106145	0.098684	0.111111	0.652174	0.642857

pav. 21 duomenų normalizacijos rezultatai

10. Išvados

Analizės rezultatai parodo, kad dalis atributų neturi ryšio, pvz., „horsepower“ ir „compressionratio“(pav. 20), ryšys tarp šių atributų yra -0,076, tuo tarpu, kai atributai „citympg“ ir „highwaympg“ turi stiprų ryšį, t.y., 0.96.

Papildymas

Laboratorinis darbas yra sėkmingai atliktas – gauti ir išanalizuoti duomenys. Laboratoriniui darbui buvo naudojamos šios „python“ bibliotekos: „pandas“ – duomenų analizei, „Counter“ – antrosios modos radimui, „matplotlib“ – grafikų atvaizdavimui, „seaborn“ – koreliacijos radimui ir šios nuorodos.

Nuorodos

https://stackoverflow.com/

„Intelektikos pagrindai“ (Paulauskaitė-Tarasevičienė Agnė),

“IFF 7-7 Nojus Rimeisis Lab 1.pdf“ (Nojus Rimeisis, 2020)

„Indrė Pabijonavičiūtė IFF7-6.pdf“ (Indrė Pabijonavičiūtė, 2020)