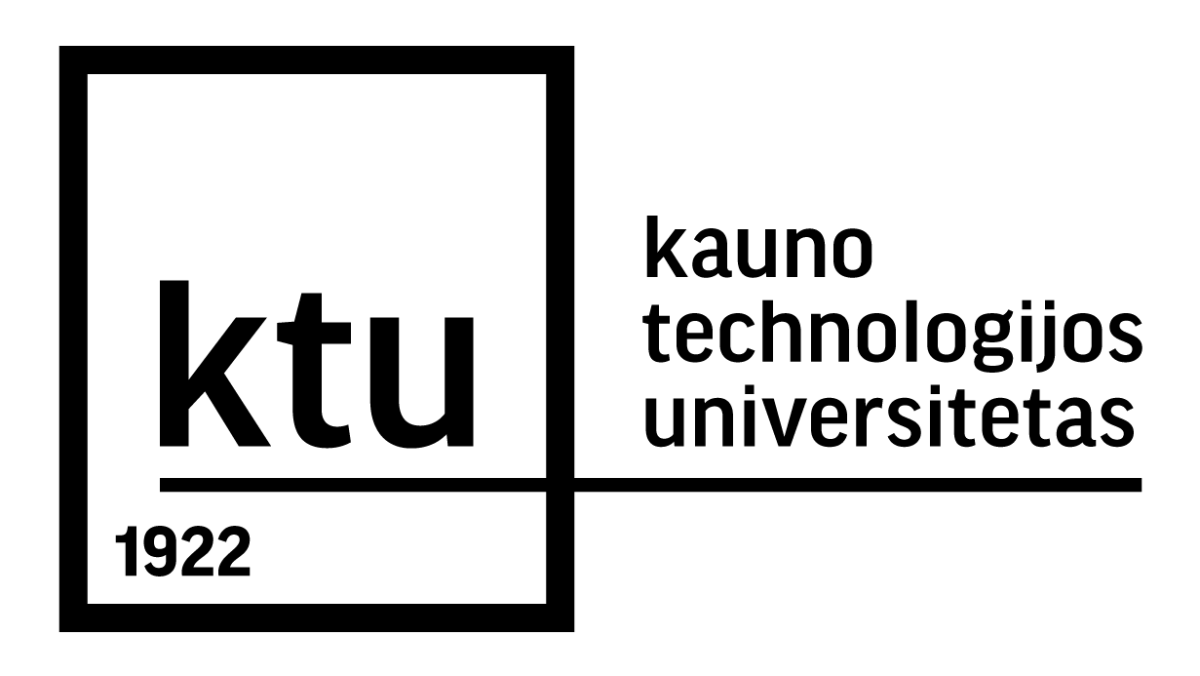
KAUNO TECHNOLOGIJOS UNIVERSITETAS

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 3](#_Toc65785004)

[2. Duomenų rinkinio kokybės analizė 3](#_Toc65785005)

[3. Atributų histogramos 4](#_Toc65785006)

[4. Duomenų kokybės problemos ir sprendimai 7](#_Toc65785007)

[5. Tolydinio tipo atributų vizualizacija 8](#_Toc65785008)

[6. Scatter plot Matrix diagrama 13](#_Toc65785009)

[7. Kategorinio tipo atributų vizualizacija 14](#_Toc65785010)

[8. Koreliacijos matrices diagrama 21](#_Toc65785011)

[9. Duomenų normalizacija 22](#_Toc65785012)

[10. Išvados 23](#_Toc65785013)

[Papildymas 24](#_Toc65785014)

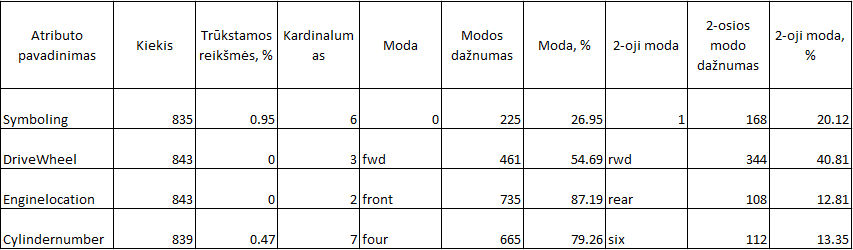
[Nuorodos 25](#_Toc65785015)

# 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]](#footnote-1)], „drivewheel“[[[2]](#footnote-2)], „enginelocation“[[[3]](#footnote-3)], „cylindernumber“[[[4]](#footnote-4)], „enginesize“[[[5]](#footnote-5)], „compressionratio“[[[6]](#footnote-6)], „horsepower“[[[7]](#footnote-7)], „citympg“[[[8]](#footnote-8)], „highwaympg“[[[9]](#footnote-9)].

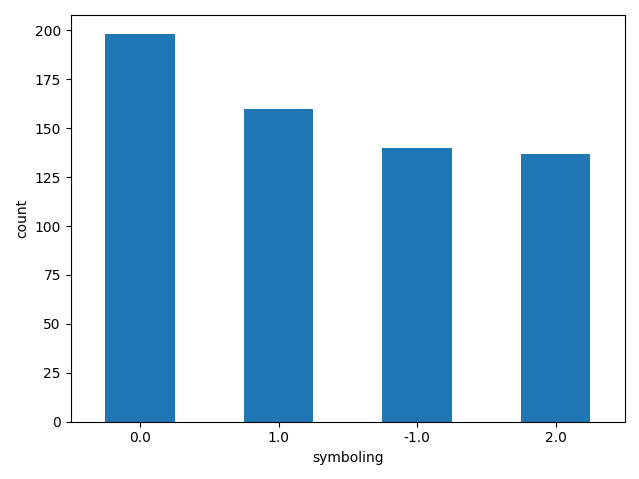
# Duomenų rinkinio kokybės analizė

|  |
| --- |
| pav. 1 Tolydinio tipo atributų kokybės analizės lentelė |
|  |



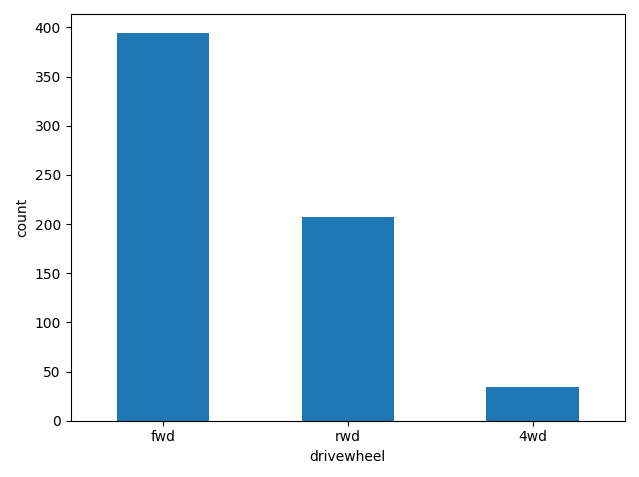
pav. 2 Kategorinio tipo atributų kokybės analizės lentelė

# Atributų histogramos



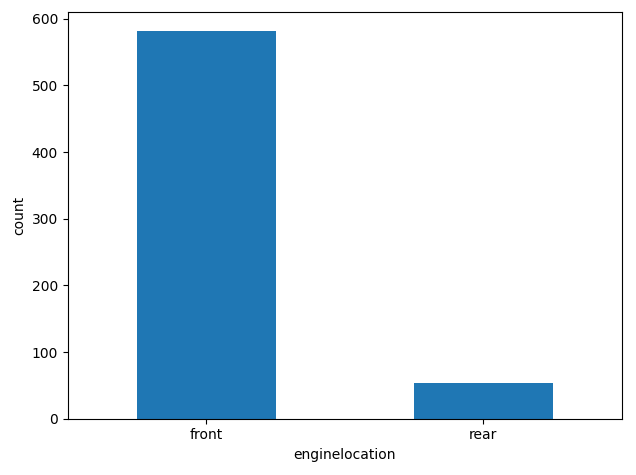
pav. 3 "Symboling" atributo histograma

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



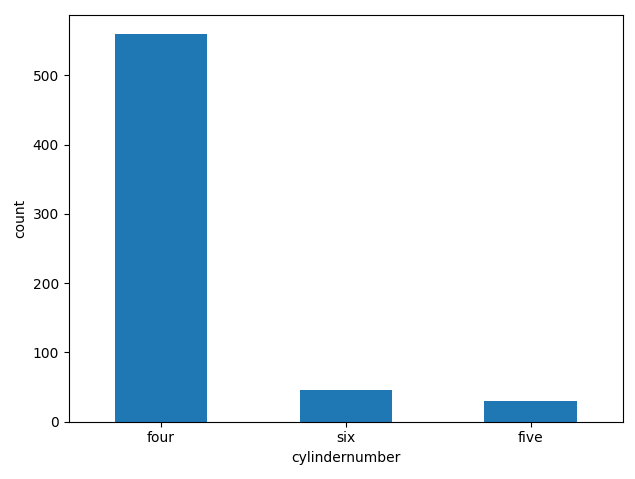
pav. 4 "drivewheel" atributo histograma

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.

# 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.

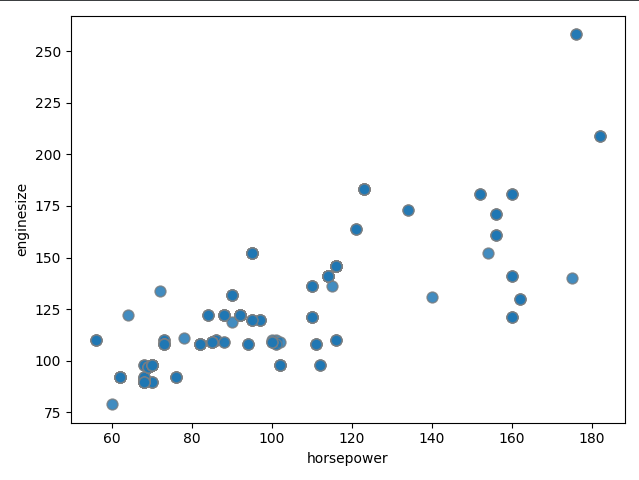
Kodo fragmentas išskirčių radimui ir šalinimui:

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:

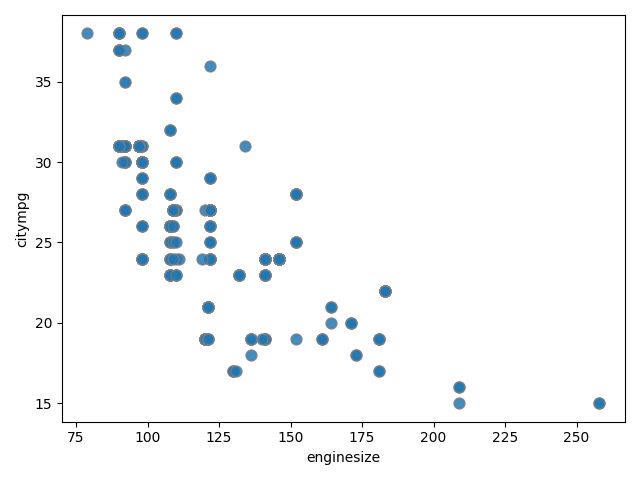
def deleteNaN(data):  
 data = data.dropna()

# Tolydinio tipo atributų vizualizacija



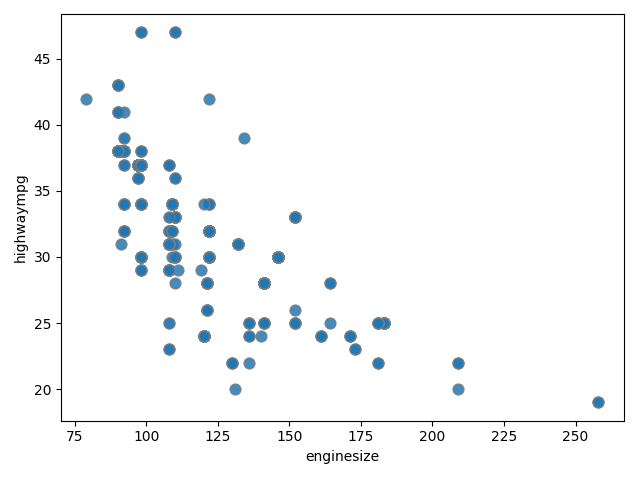
pav. 7 "enginesize" ir "horsepower" atributų 'scatter plot' diagrama

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



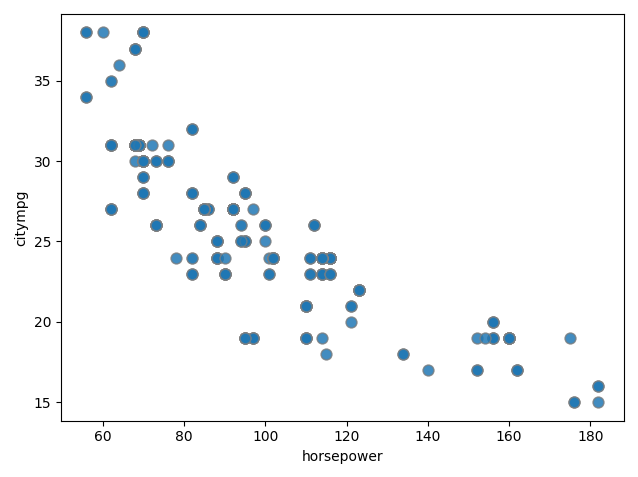
pav. 8 "citymmpg" ir "enginesize" atributų 'scatter plot' diagrama

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.



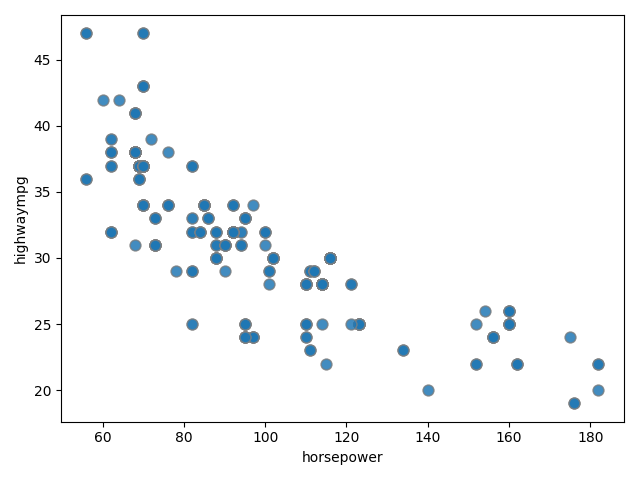
pav. 9 "highwaympg" ir "enginesize" atributų 'scatter plot' diagrama

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

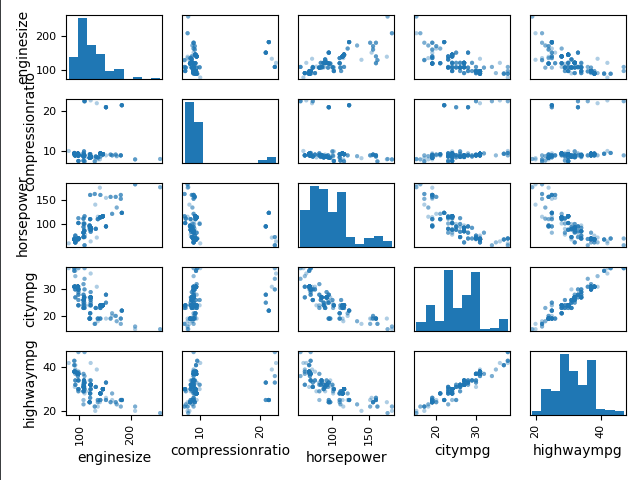
Diagrama vaizduoja neigiamą koreliaciją tarp arklio galių ir kuro sąnaudų mieste – kuo variklis turi daugiau arklio galių tuo jis yra ekonomiškesnis.



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

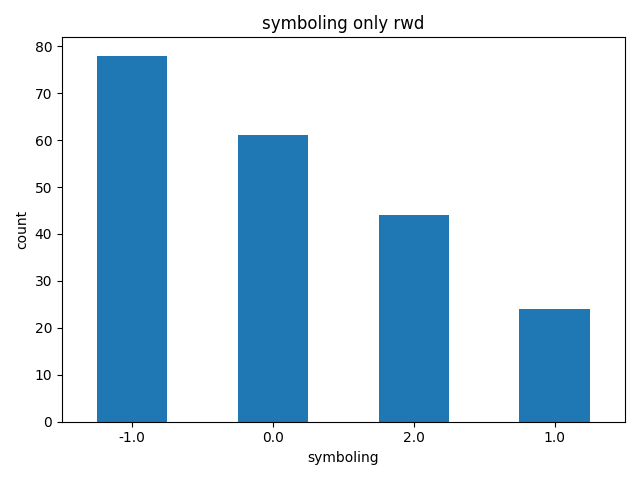
Diagrama vaizduoja neigiamą koreliaciją tarp arklio galių ir kuro sąnaudų užmiestyje – kuo variklis turi daugiau arklio galių tuo jis yra ekonomiškesnis.

# Scatter plot Matrix diagrama

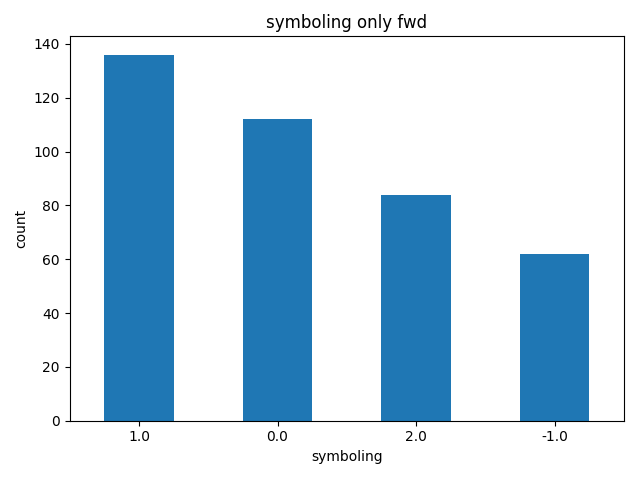


pav. 12 SPLOM diagrama

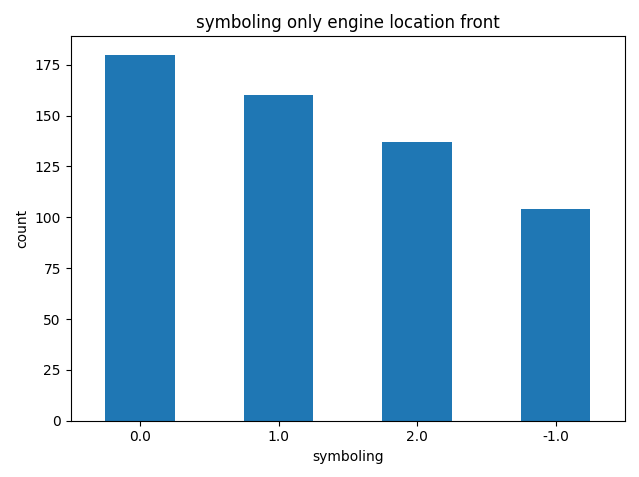
# Kategorinio tipo atributų vizualizacija



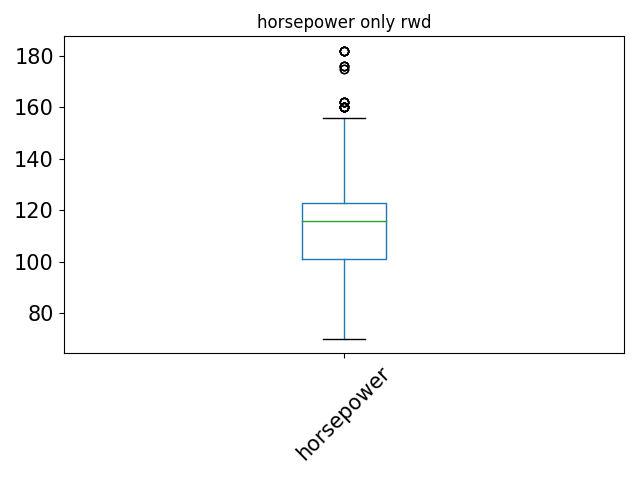
pav. 13 "symboling" ir "drivewheel" atributų 'bar plot' diagrama



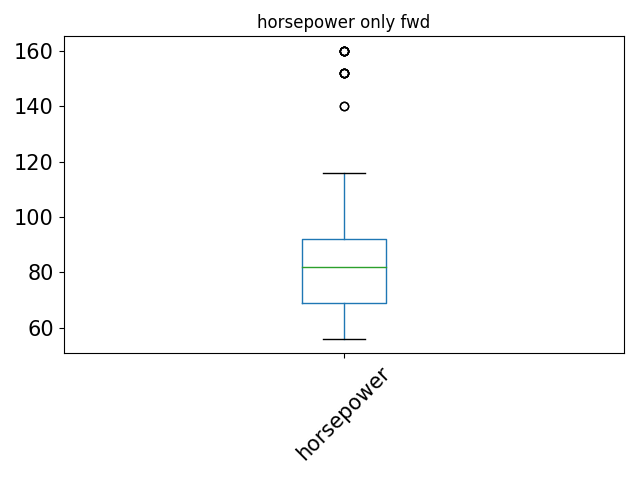
pav. 14 "symboling" ir "drivewheel" atributų 'bar plot' diagrama



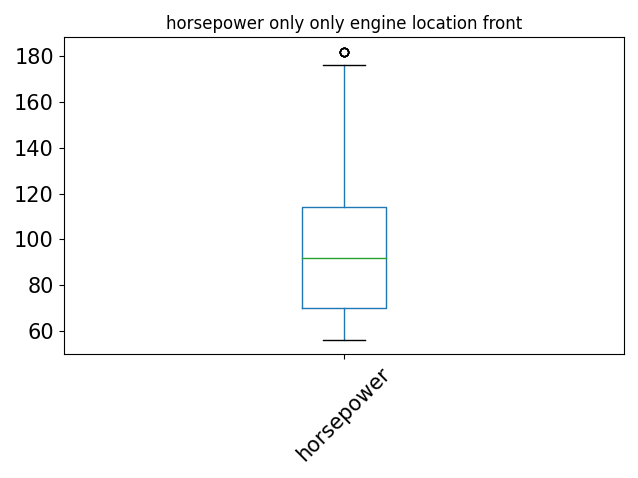
pav. 15 "symboling" ir "enginelocation" atributų 'bar plot' diagrama



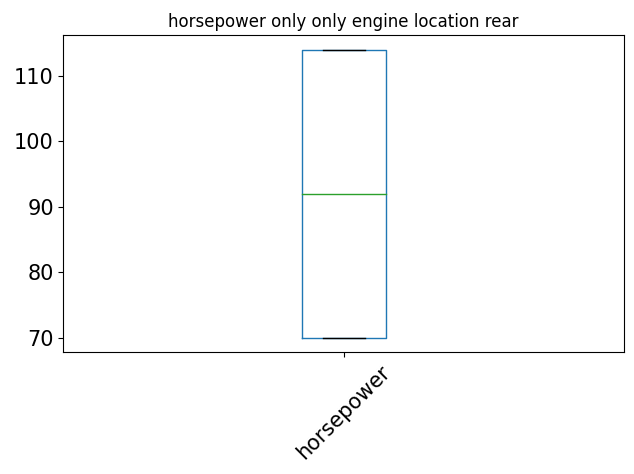
pav. 16 "horsepower" ir "wheeldrive” ‘box plot’ diagrama



pav. 17 "horsepower" ir "wheeldrive” ‘box plot’ diagrama

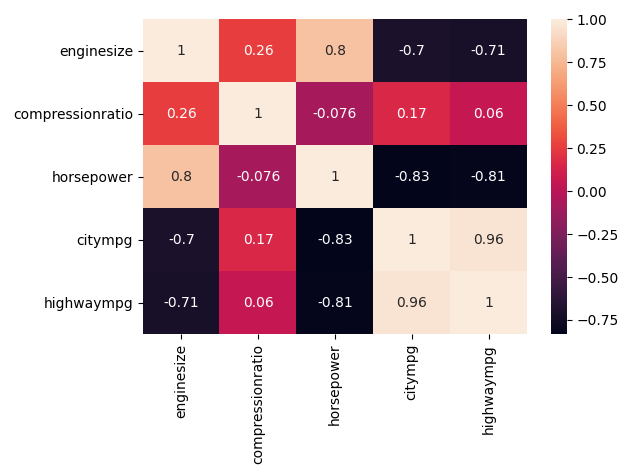


pav. 18 "horsepower" ir "enginelocation” ‘box plot’ diagrama



pav. 19 "horsepower" ir "wheeldrive” ‘box plot’ diagrama

# Koreliacijos matrices diagrama



pav. 20 Koreliacijos matricos diagrama

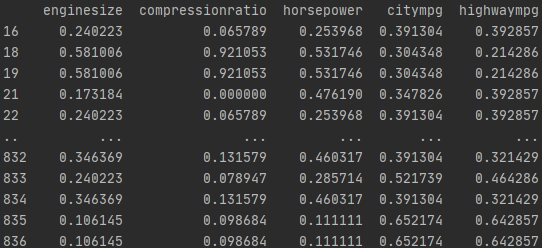
Koreliacijos matrices diagram (pav. 20) vaizduoja koreliacijas tarp tolydinių atributų. Dvi atributų („enginesize“ su „highwaympg“ ir „citympg“) poros turi neigiamą koreliaciją, tačiau „-0.7“ nėra stiprus ryšys, dvi poros („enginesize“ su „horsepower“ ir „citympg“ su „highwaympg“) turi teigiamą koreliaciją,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

# 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:



pav. 21 duomenų normalizacijos rezultatai

# 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).

# Nuorodos

<https://stackoverflow.com/>

[„Intelektikos pagrindai“](https://moodle.ktu.edu/pluginfile.php/401043/mod_resource/content/1/KNYGA%20ma%C5%A1inio%20mokymosi%20algoritmai%202021_01_26.pdf) (Paulauskaitė-Tarasevičienė Agnė),

[“IFF 7-7 Nojus Rimeisis Lab 1.pdf”](https://moodle.ktu.edu/pluginfile.php/393756/mod_resource/content/2/IFF%207-7%20Nojus%20Rimeisis%20Lab%201.pdf) (Nojus Rimeisis, 2020)

[„Indrė Pabijonavičiūtė IFF7-6.pdf“](https://moodle.ktu.edu/pluginfile.php/393753/mod_resource/content/1/Indr%C4%97%20Pabijonavi%C4%8Di%C5%ABt%C4%97%20IFF7-6.pdf) (Indrė Pabijonavičiūtė, 2020)

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