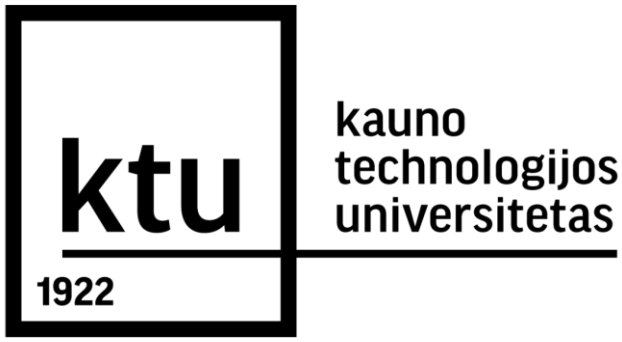
KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS



Intelektikos pagrindai (P176B101)

Laboratorinis darbas Nr.1

Atliko:
IFF-9/8 gr. studentas
Lukas Navašinskas
2022 m. kovo 8 d.
Priėmė:
lekt. Nečiūnas Audrius
doc. Paulauskaitė-Tarasevičienė Agnė

KAUNAS 2022

Contents

1. Duomenų rinkinys

Laboratoriniui darbui pasirinktas automobilių specifikacijų rinkinys. Rinkinį sudaro šie atributai:

Tolydiniai atributai: Year (Metai), Engine Cylinders (Cilindrų kiekis), Engine Displacement (Variklio darbinis tūris), City MPG (Kuro sanaudos mieste Mylios per Galoną), Highway MPG (Kuro sanaudos užmiestyje Mylios per Galoną), Annual Fuel Cost (Kasmetinė kuro kaina), Tailpipe CO2 in Grams/Mile (CO2 gramų išmetimas per mylią)

Kategoriniai atributai: Drive (Varomieji ratai), Transmission (Transmisija), Turbocharger (Turbina), Supercharger (kompresorius), Fuel Type (kuro tipas)

2. Duomenų rinkinio kokybės analizė

Tolydinių atributų analizė:

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-iasis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
Year	38113	0	34	1984	2017	1991	2009	2000.195	2001	10.465
Engine Cylinders	37977	0.357	9	2	16	4	6	5.737	6	1.752
Engine Displacement	37979	0.352	66	0	8.4	2.2	4.3	3.318	3	1.362
City MPG	38113	0	93	6	150	15	20	17.981	17	6.850
Highway MPG	38113	0	83	9	122	20	27	24.081	24	7.027
Annual Fuel Cost	38113	0	60	500	6050	1600	2350	1970.675	1950	532.555
Tailpipe CO2 in Grams/Mile	38113	0	592	0	1269.571	388	555.438	472.761	467.737	122.200

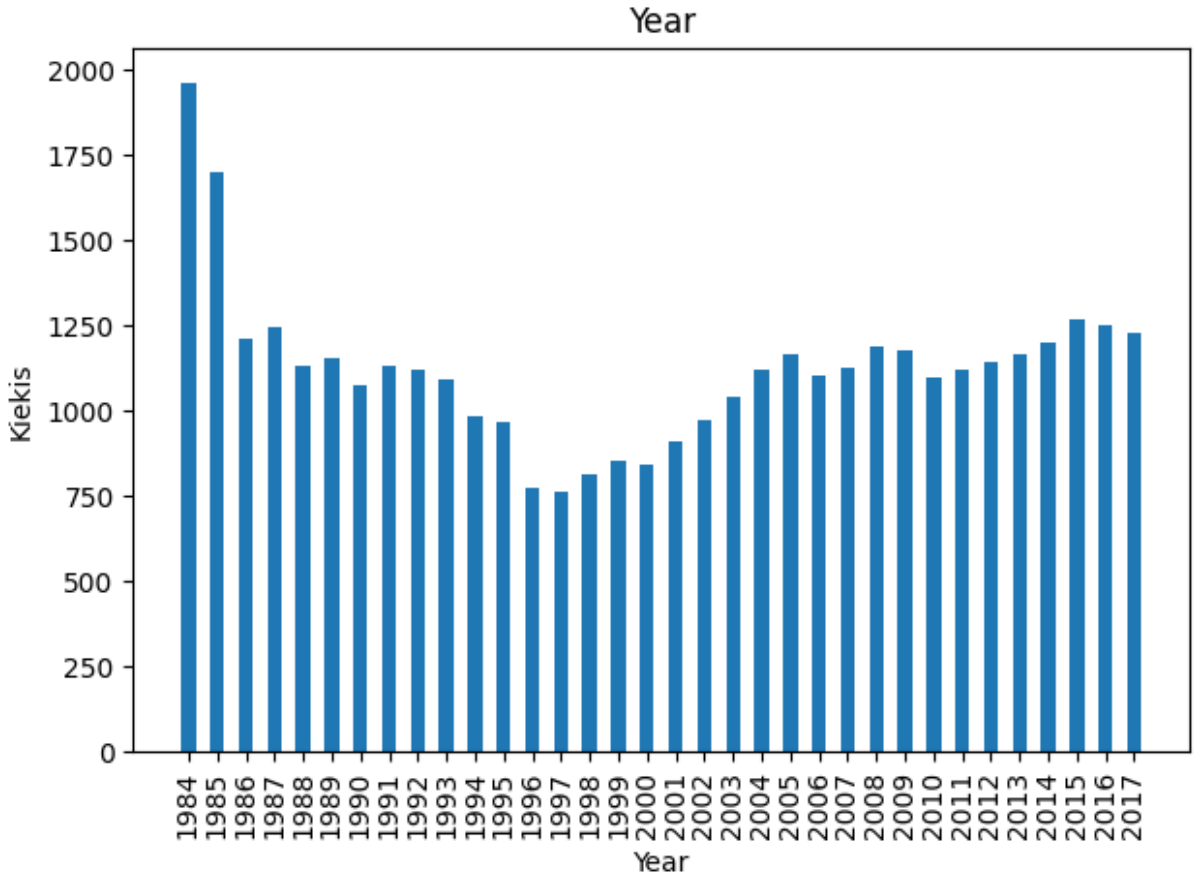
pav. 1 Tolydinio tipo atributų kokybės analizės lentelė

Kategorinių atributų analizė:

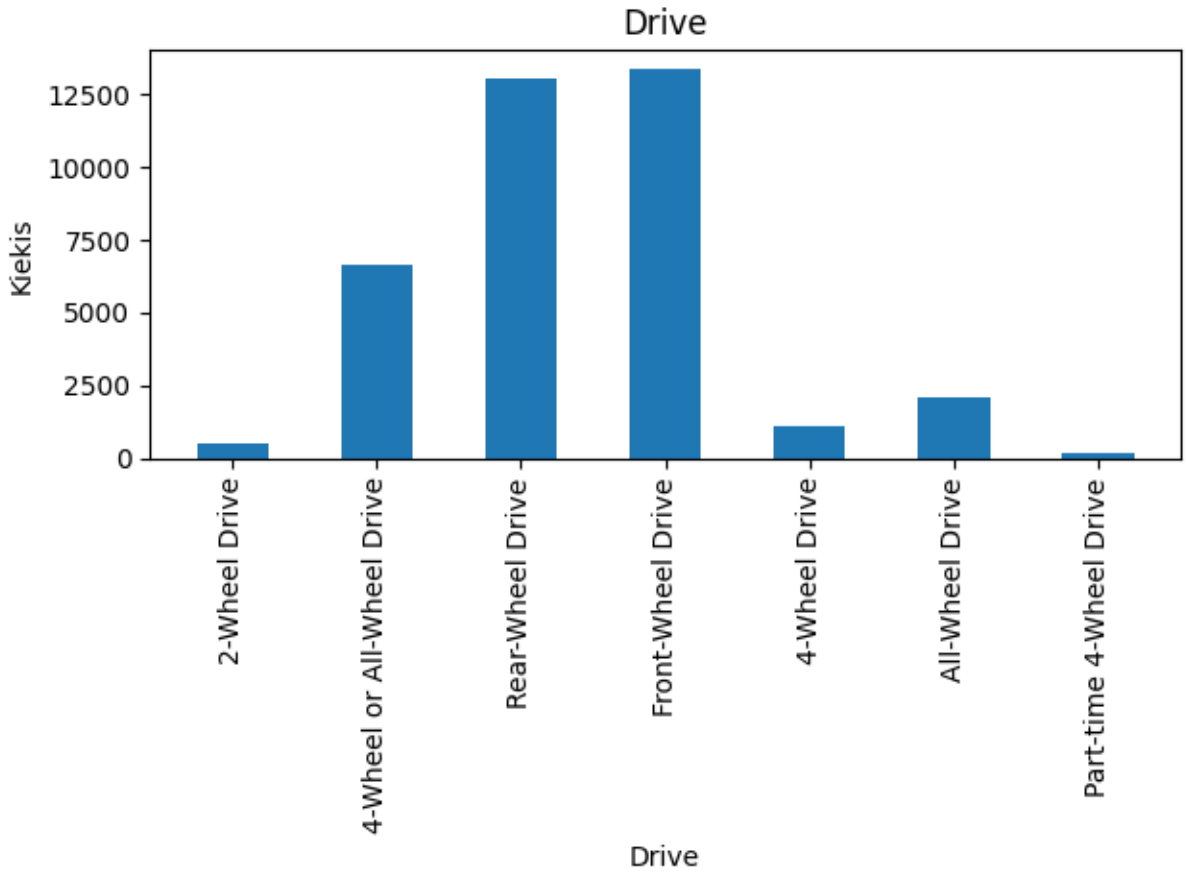
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %
Drive	36924	3.120	7	Front-Wheel	13351	35.030	Rear-Wheel Drive	13018	34.156
Transmission	38102	0.0289	46	Automatic 4-Speed	11042	28.972	Manual 5-Speed	8323	21.838
Turbocharger	5239	86.254	2	None	32874	13.746	Yes	5239	13.756
Supercharger	693	98.182	2	None	37420	1.818	Yes	693	1.818
Fuel Type	38113	0	14	Regular	25258	66.271	Premium	10133	26.587

pav. 2 Kategorinio tipo atributų kokybės analizės lentelė

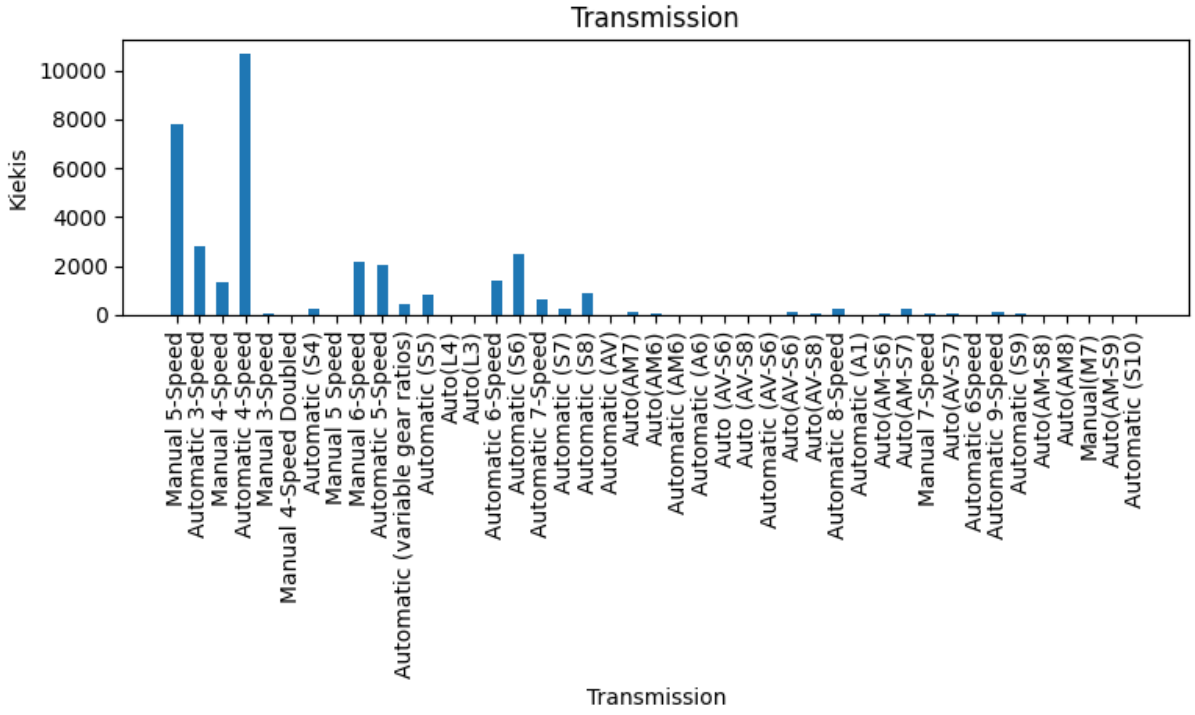
3. Atributų histogramos



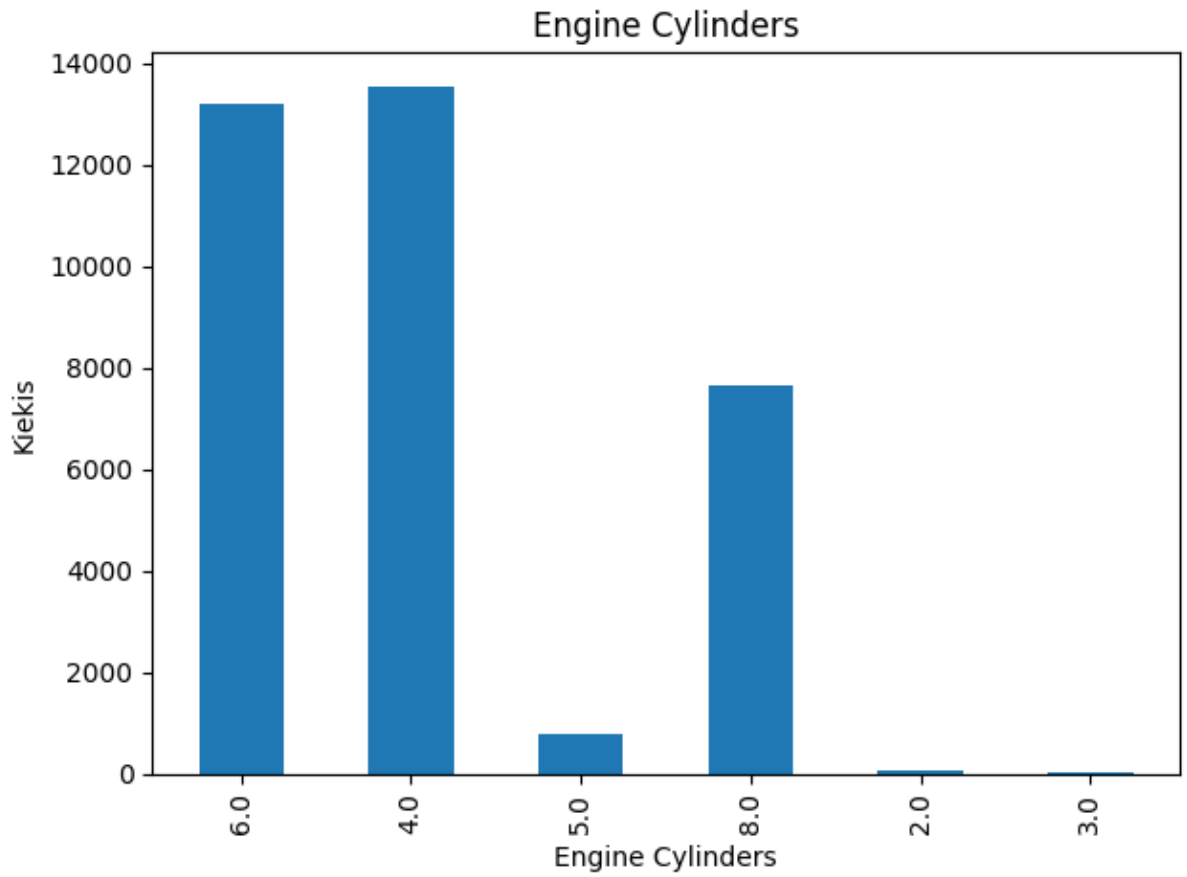
Tolydinio atributo „Year“ reikšmės pasiskirsčiusios netolygiai, nuo 1984 iki 1997 reikšmių kiekiai mažėja eksponentiškai, o toliau auga tolygiai. Matome, kad įrašų apie mašinas 1984 ir 1985 metais buvo ženkliai daugiau.



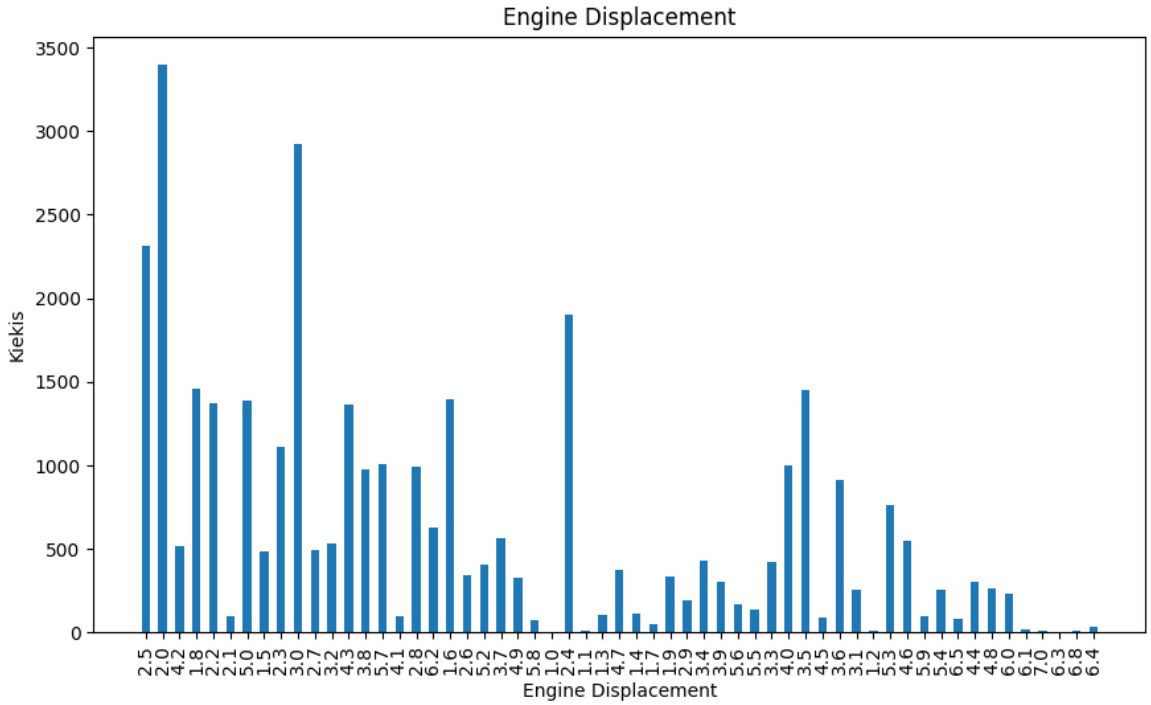
Kategorinio atributo „Drive“ reikšmės pasiskirčiusios netolygiai. Matome, kad duomenų rinkinyje populiariausios dokumentuotos mašinos buvo varomos galu, o kitos priekiu.



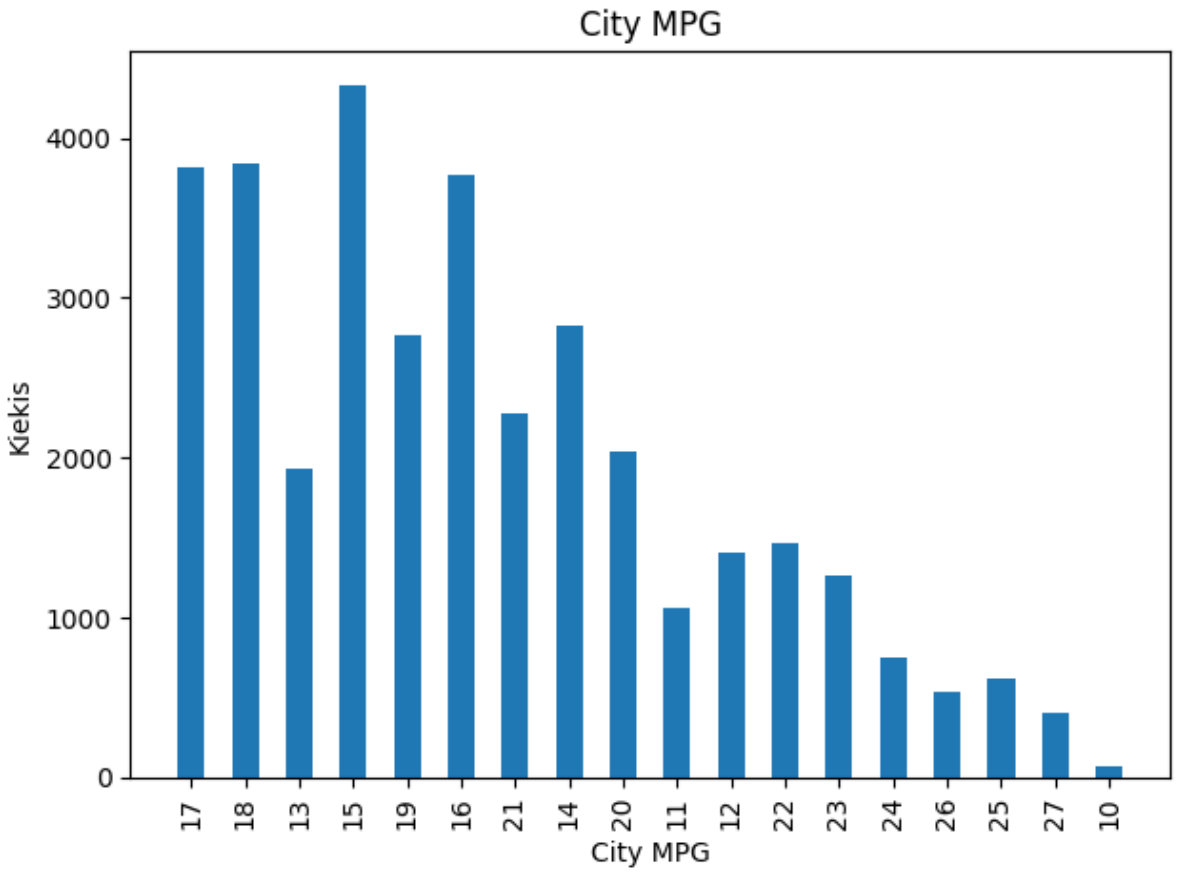
Iš šio kategorinio atributo „Transimssion“ histogramos galime spręsti, kad reikšmės yra pasiskirsčiusios netolygiai. Taip pat matome, kad yra nereikšmingų atributų, kurių vertė palyginus su populiariausiais atributais yra beveik nulinė. Reikėtų šiuos atributus šalinti.



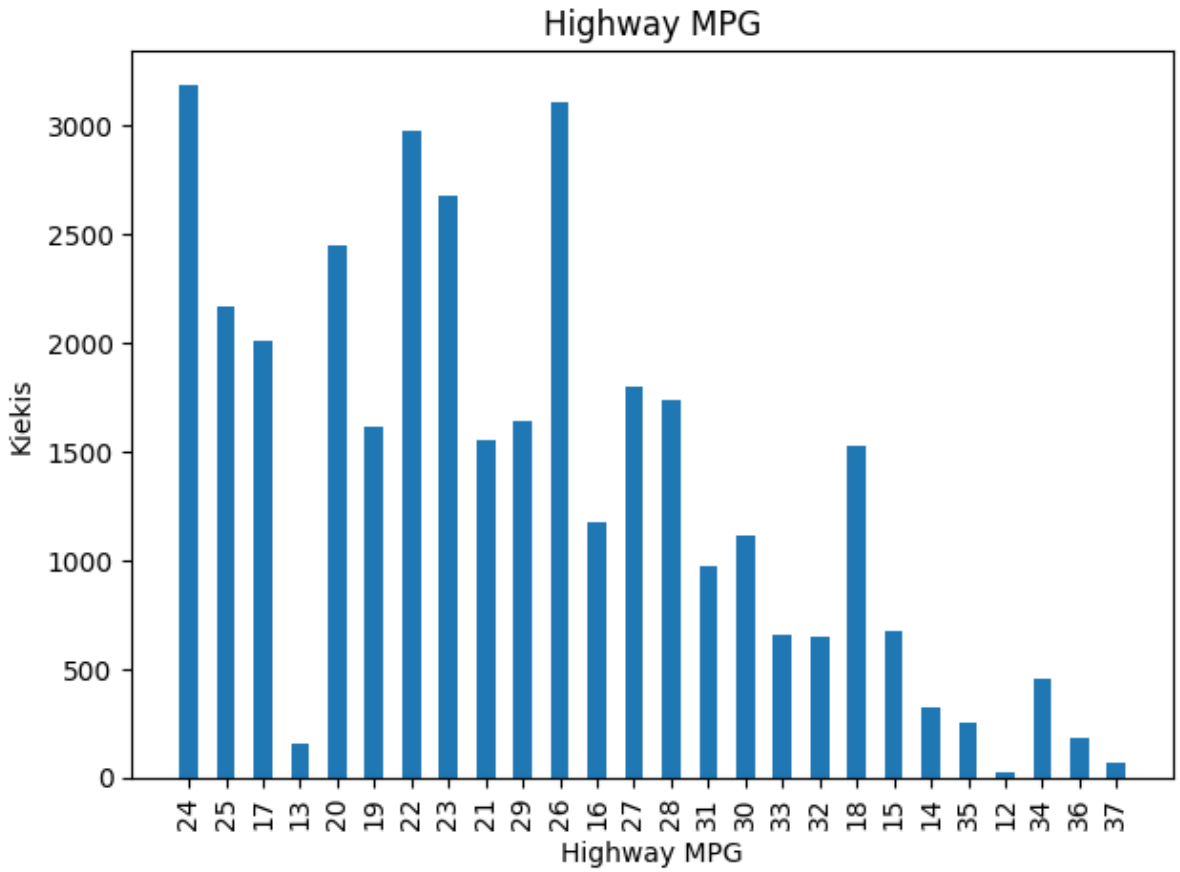
Šio tolydinio atributo “Engine Cylinders” galime matyti, kad duomenų rinkinyje mašinos turėjo daugiausiai 4 ir 6 cilindrus. Atributų vertės histogramoje pasiskirsčiusios netolygiai



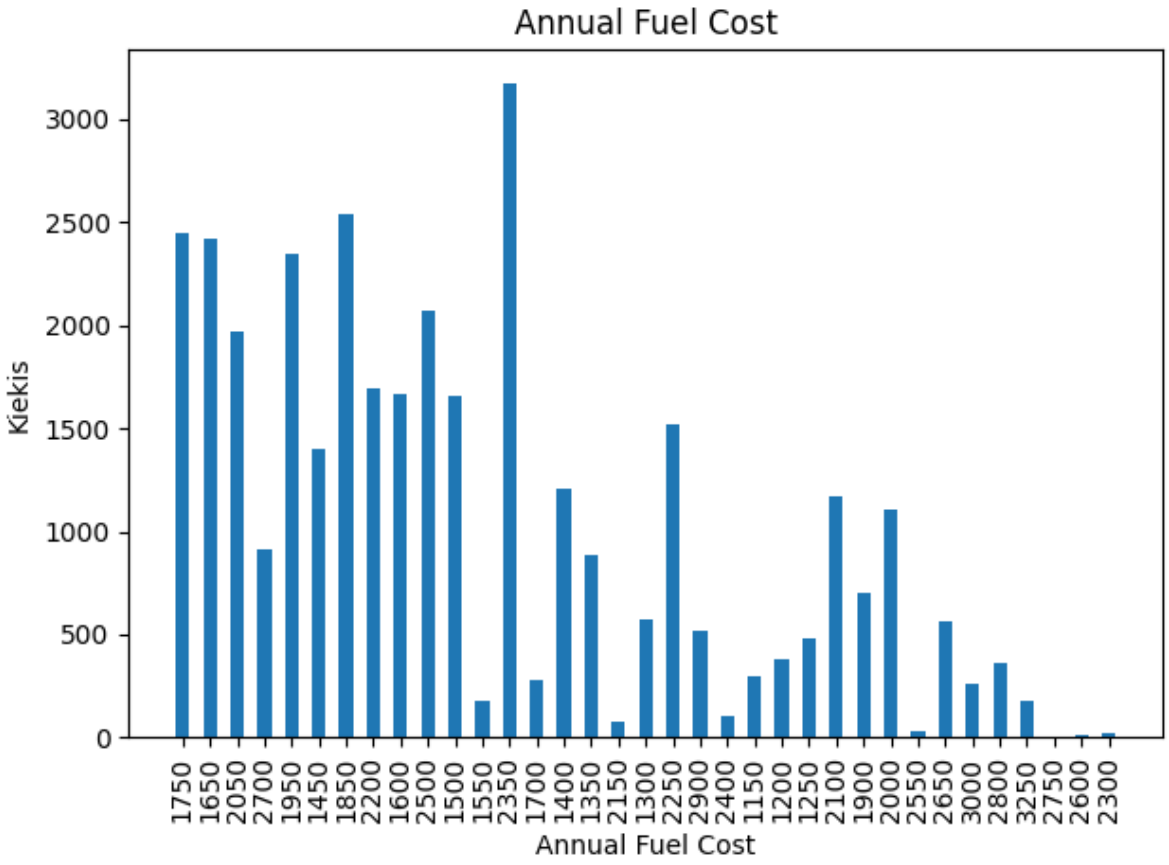
Iš šio Tolydinio atributo histogramos “Engine Displacement” galime matyti, kad reikšmės yra pasiskirsčiusios netolygiai. Šiame duomenų rinkinyje daugiausia mašinų turėjo 2, 3 ir 2.5 litrų variklio darbinius tūrius.



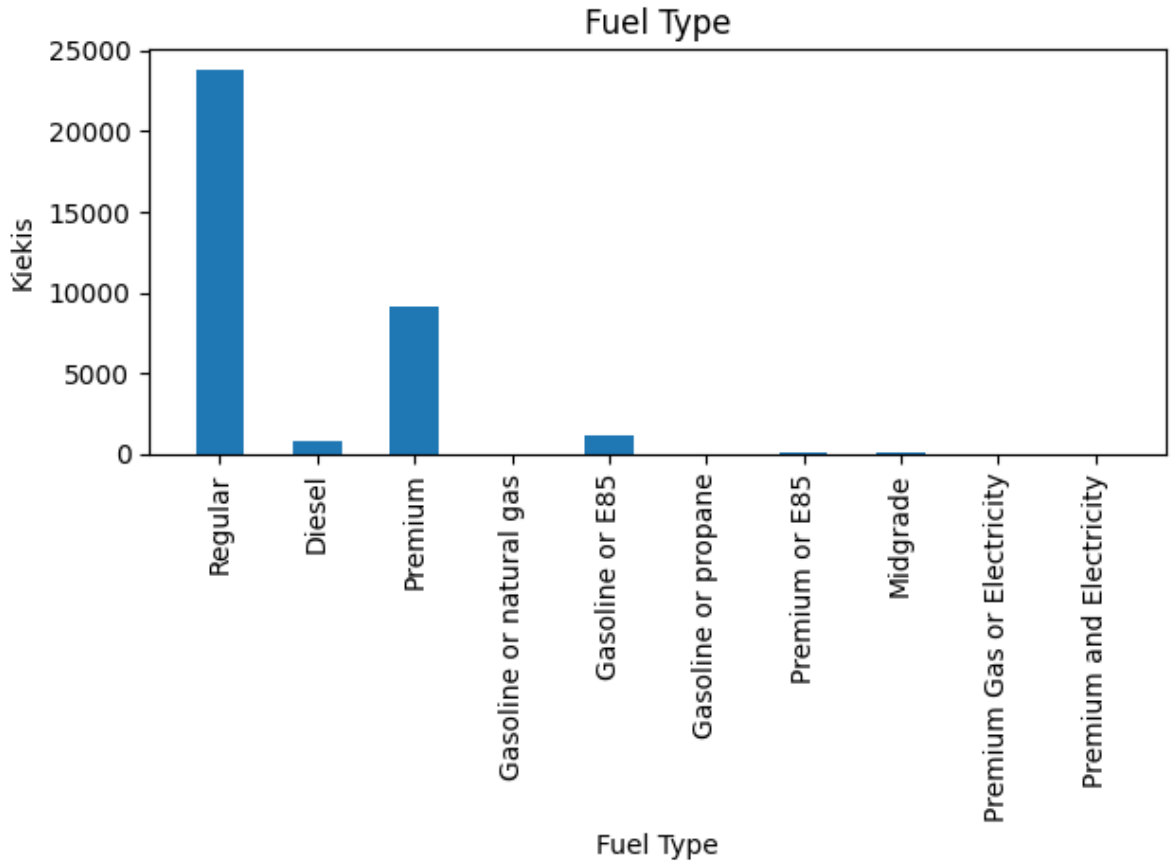
Iš šio tolydinio atributo „City MPG“ histogramos galime matyti, kad šiame duomenų rinkinyje mašinos dažniausiai galėdavo nuvažiuoti 15 mylių per vieną galoną kuro mieste. Histogramoje vertės pasiskirsčiosios netolygiai



Iš šio tolydinio atributo „Highway MPG“ histogramos galime matyti, kad šiame duomenų rinkinyje mašinos dažniausiai galėdavo nuvažiuoti 24 mylias per vieną galoną kuro užmiestyje. Histogramoje vertės pasiskirsčiosios netolygiai



Iš šio atributo „Annual Fuel Cost“ histogramos galime matyti, kad žmonės daugiausia sumokėdavo 2350 dolerių per metus už kurą. Histogramoje atributų vertės pasiskirsčiuios netolygiai



Iš šio kategorinio atributo „Fuel Type“ galime matyti, kad populiariausias kuras mašinom tarp 1984 ir 2017 buvo „Regular“, tai benzinas kurio kuro oktaninis skaičius yra ~87

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.

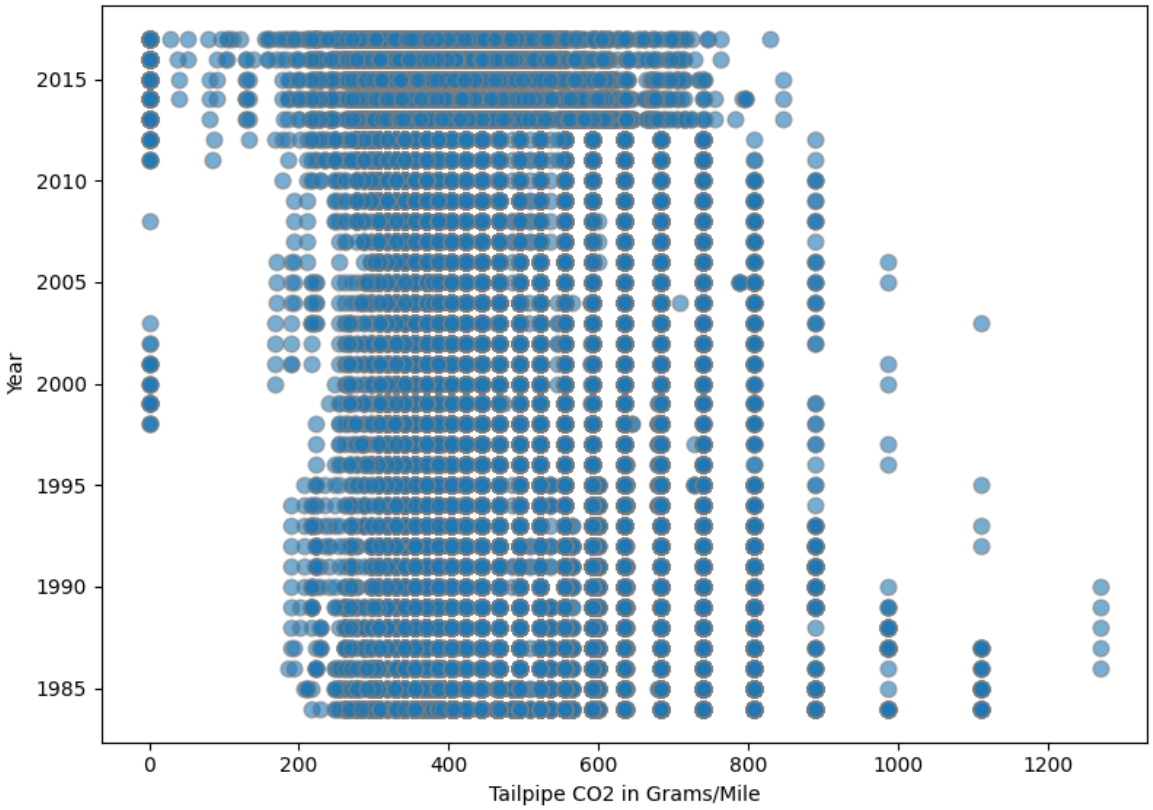
Kodo fragmentas išskirčių radimui ir šalinimui:

```
def salintiOutliers(df):
    Q1 = df.quantile(0.25)
    Q3 = df.quantile(0.75)
    IQR = Q3 - Q1
    df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
    return df
```

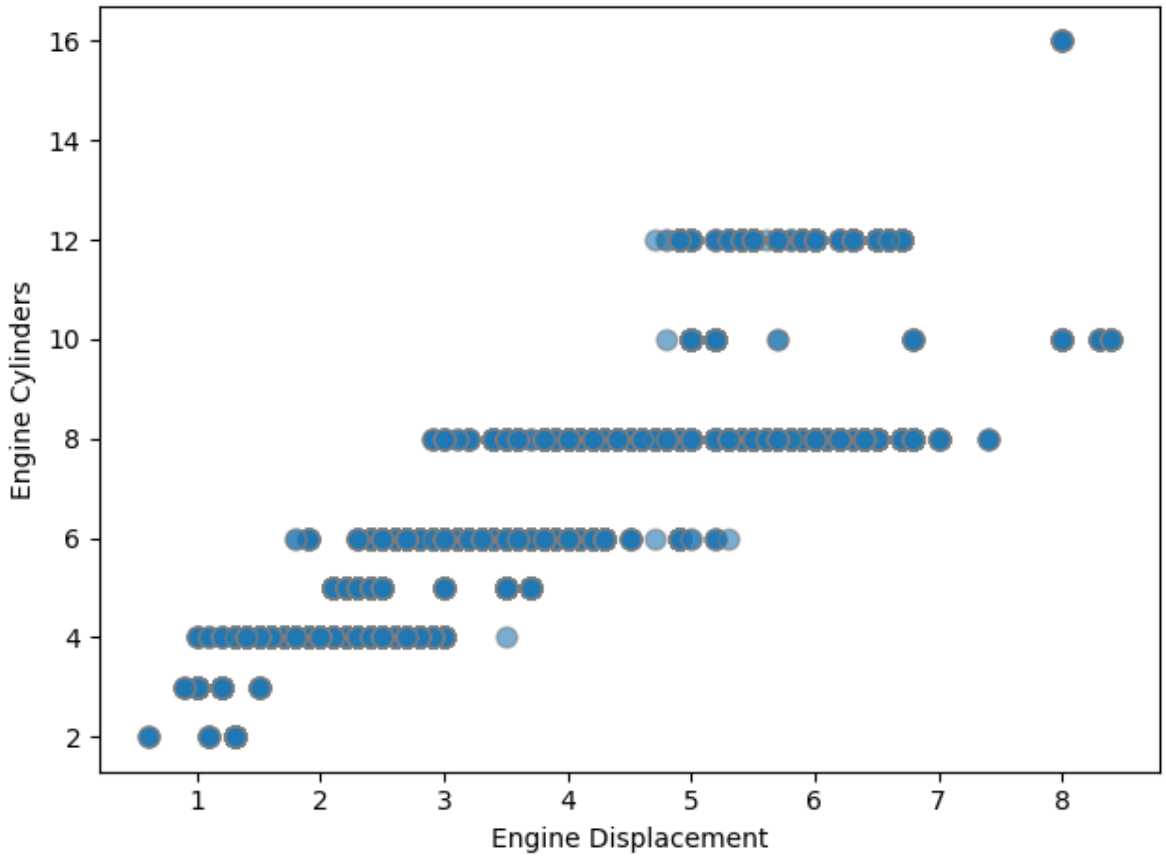
Kodo fragmentas ištrinti eilutes su trūkstamomis reikšmėmis:

```
data.dropna()
```

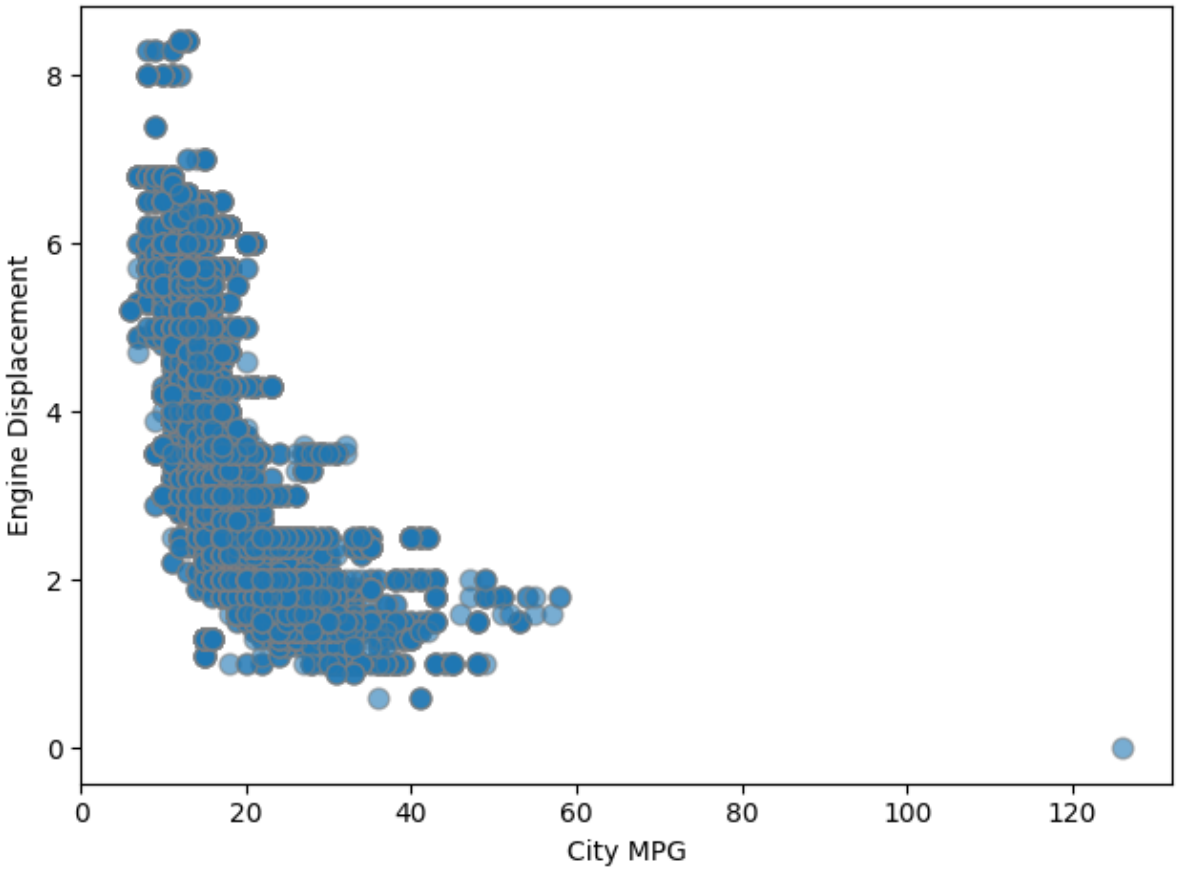
5. Sąryšiai tarp atributų



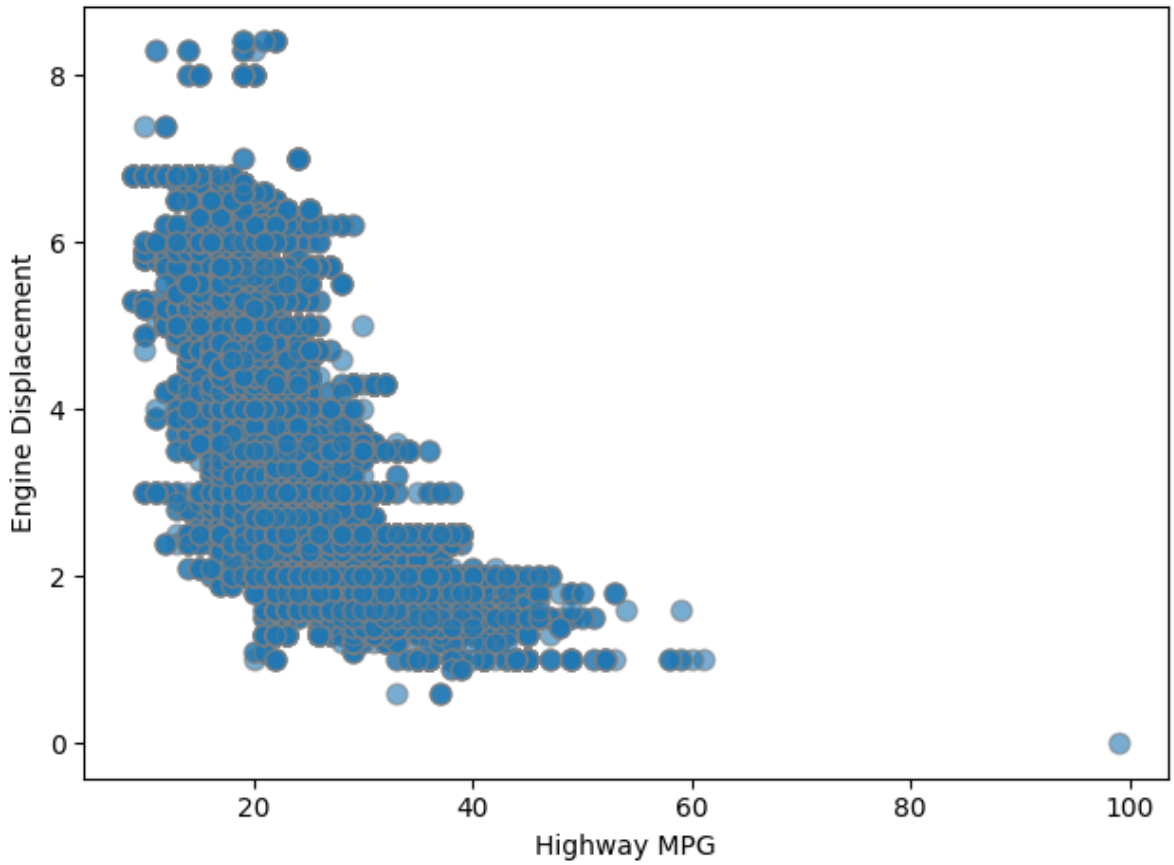
Šiame atributų sarišyje tarp “Year” ir “Tailpipe CO2 in Grams/Mile” galime matyti, kad daugiausią teršalų išmetančios mašinos (>900 CO2 g/myl) egzistavo tarp 1984 ir 2005 metų.



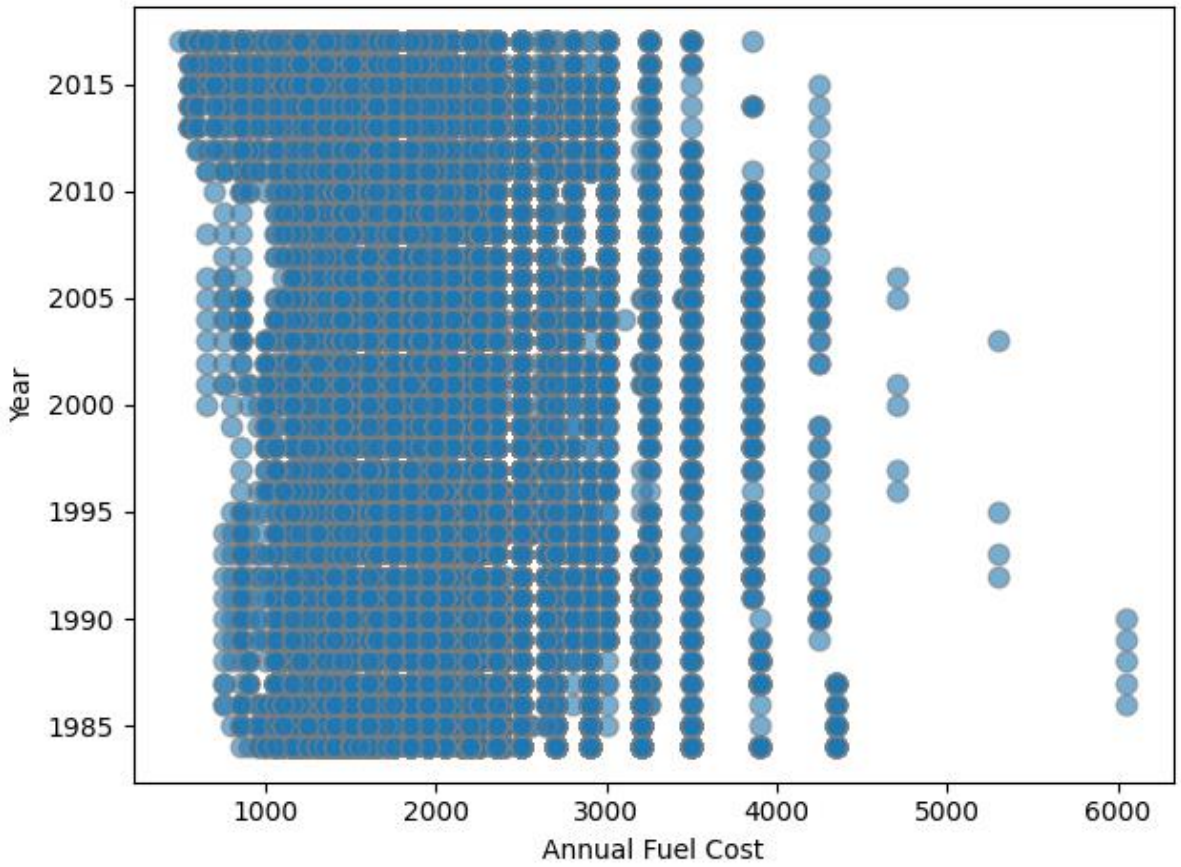
Iš šito sąryšio tarp “Engine Cylinders” ir “Engine Displacement” galime matyti, kad variklio darbinis tūris stipriai priklauso nuo cilindrų kiekio.



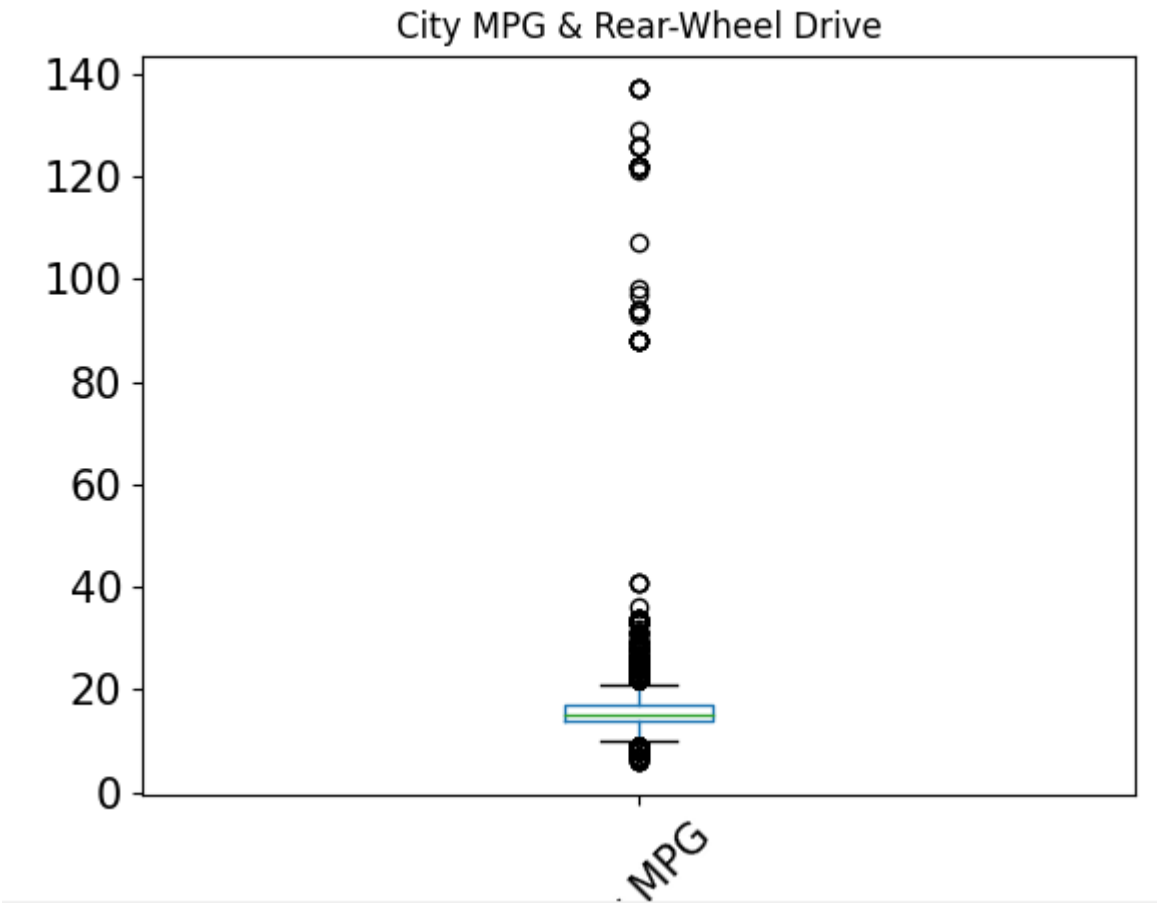
Iš šio sąryšio tarp “City MPG” ir “Engine Displacement” galime matyti, kad mašinos su didesniu variklio tūriu mieste sunaudoja mažiau kuro. Tai kelia klausimų ar šis duomenų rinkinys yra tikslus, nes tai nėra logiška.

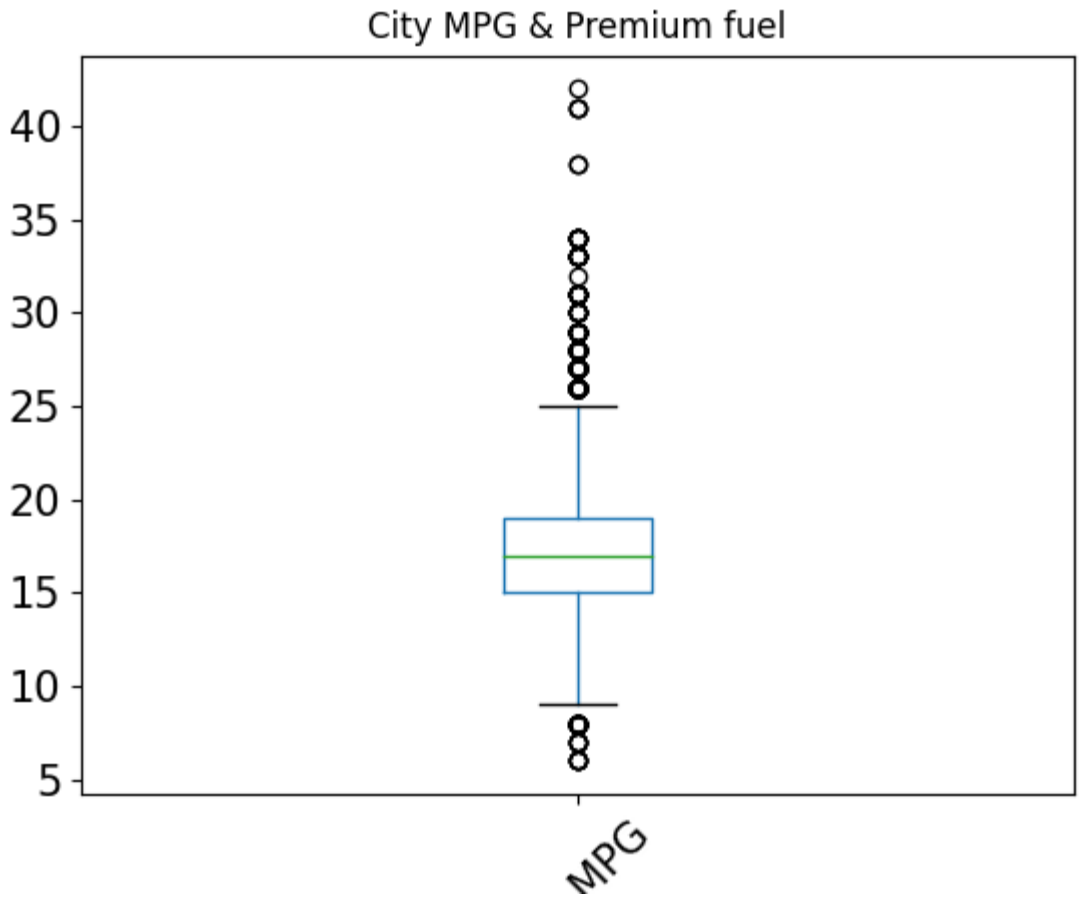


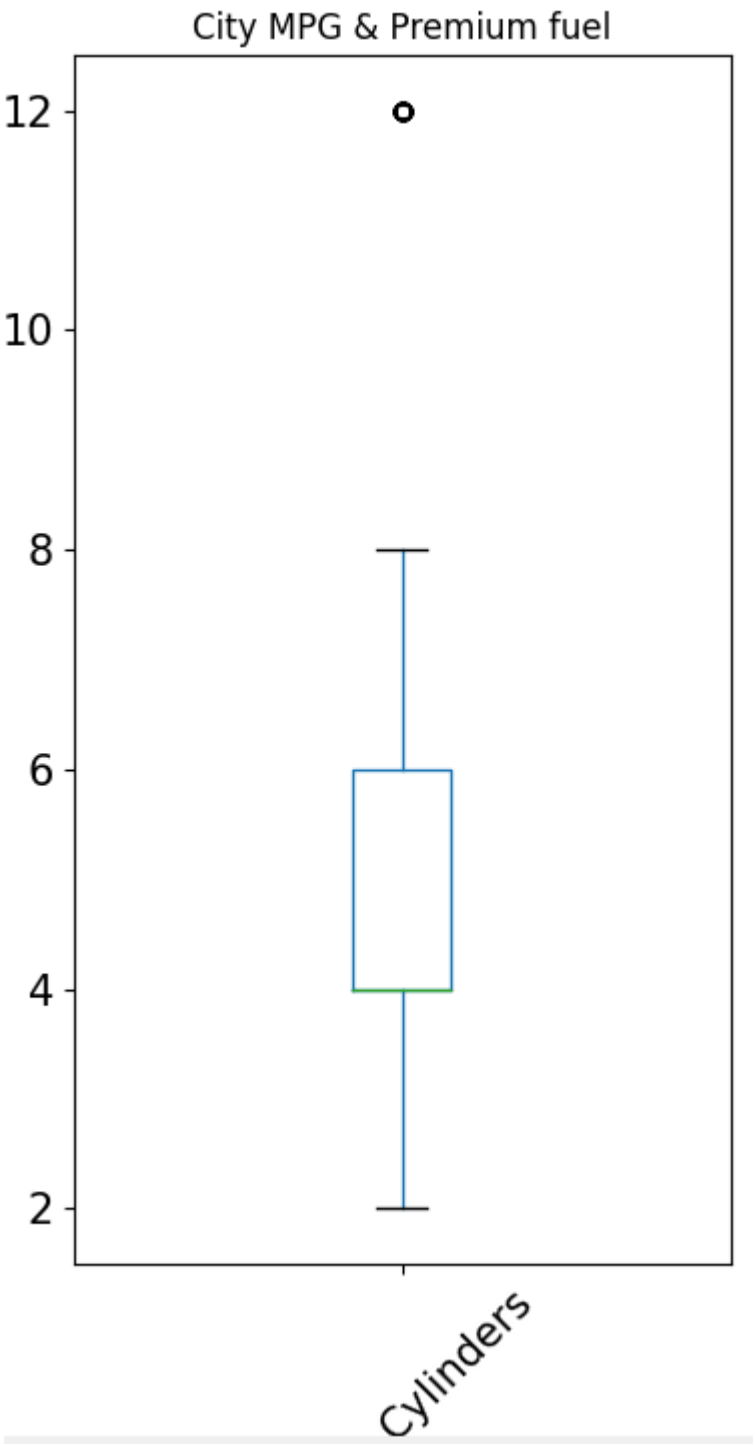
Iš šio sąryšio tarp “Highway MPG” ir “Engine Displacement” galime matyti, kad mašinos su didesniu variklio tūriu mieste sunaudoja mažiau kuro. Tai kelia klausimų ar šis duomenų rinkinys yra tikslus, nes tai nėra logiška



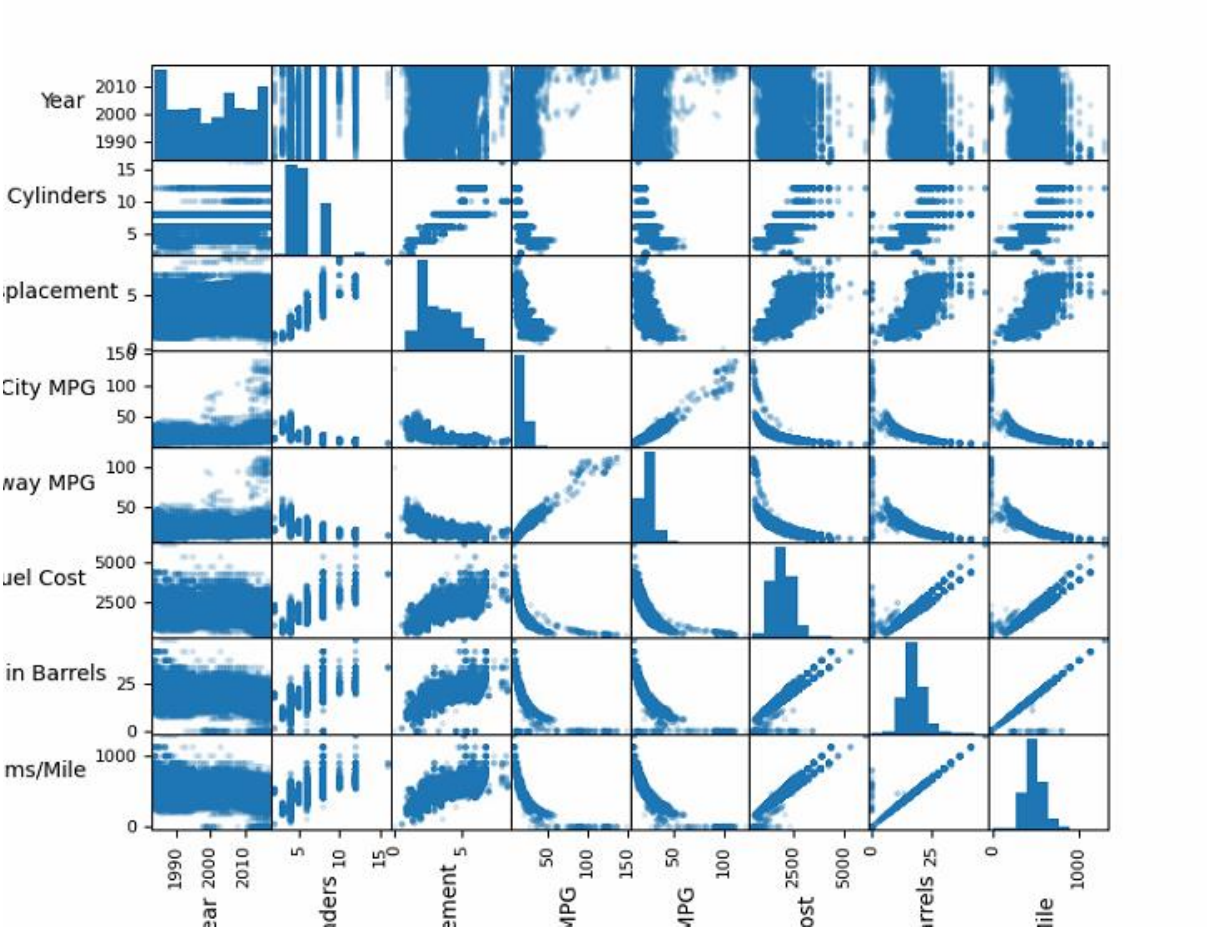
Iš šio sąryšio tarp “Annual Fuel Cost” ir “Year” galime matyti, kasmetinės kuro kainos buvo didžiausios tarp 1985 ir 1995


City MPG & Rear-Wheel Drive
140
120
100
80
60
40
20
0
MPG


City MPG & Premium fuel
40
35
30
25
20
15
10
5
MPG

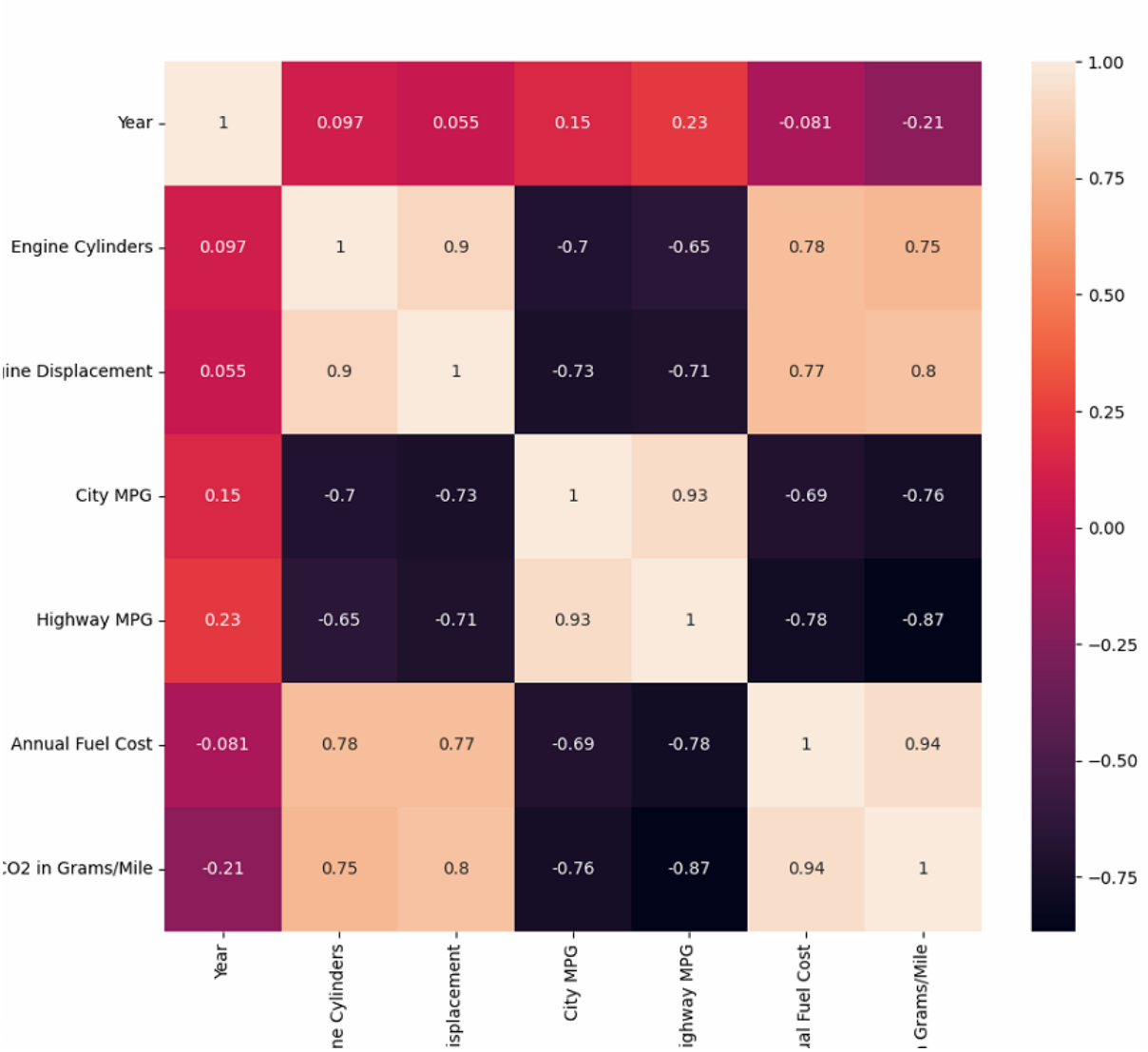

City MPG & Premium fuel
12
10
8
6
4
2
Cylinders

6. Scatter plot matrix diagrama



Kaip matome iš Scatter Plot Matrix, duomenų rinkinyje turime stipriai susijusių tolydinių atributų.

7. Koreliacijos matrices diagrama



Koreliacijos matricoje matome, kad stipriai susiję atributai yra “Engine Cylinders” ir “Engine Displacement” koreliacijos koeficientu 0.9, taip pat “City MPG” ir “Highway MPG” koeficientu 0.93 . Taip pat matome, kad “Annual Fuel Cost” ir “Engine cylinders” bei “Engine Displacement” atributai koreliuoja su koeficientu ~0.78

8. Duomenų normalizacija

Programos kodas atlikti duomenų normalizacijai:

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