KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai (P176B101)

Pirmo laboratorinio darbo ataskaita

Atliko:

IFF-1/1 gr. Studentas

Vytenis Kriščiūnas

Priėmė:

lekt. Nečiūnas Audrius

lekt. Budnikas Germanas

KAUNAS 2024

TURINYS

PAVEIKSLĖLIAI

1. Duomenų rinkinys

Darbui atlikti reikia pasirinkti teisingą duomenų rinkinį, kuris tūrėtų nemažiau nei 500 eilučių ir nemažiau nei 8 stulpelius. Pasirinkau – keliavimo rūšies pasirinkimo duomenų analizę. Šaltinis: https://vincentarelbundock.github.io/Rdatasets/doc/AER/TravelMode.html.

Šį rinkinį sudaro 9 stulpeliai ir 840 eilučių.

Šie duomenys perteikia keliavimo rūšies pasirinkimą asmenų, kurie trokšta keliauti tarp Sidnėjaus, Melburno ir Australijos.

Apie duomenis:

- „Individual“ – faktorius nurodantis individą nuo 1 iki 210 lygio;
- „Mode“ – faktorius indikuojantis kelionės rūšį: mašina, oru, traukiniu ar autobusu;
- „Choise“ – faktorius nurodantis pasirinkimą taip ar ne;
- „Wait“ – laukimo laikas terminale, 0 keliaujant mašina;
- „Vcost“ – transporto priemonės kaina;
- „Travel“ – kelionės trukmė transporto priemonėje;
- „Gcost“ – bendra kelionės kaina;
- „Income“ – uždarbis;
- „Size“ – žmonių kiekis.

2. Tolydinių tipų analizė

Reikia apskaičiuoti:

- Bendrą reikšmių skaičių;
- Trūkstamų reikšmių procentą;
- Kardinalumą;
- Minimalią ir maksimalią reikšmes;
- 1-ąjį ir 3-ąjį kvartilius;
- Vidurkį;
- Medianą;
- Standartinį nuokrypį.

Kadangi „individual“ reikšmės yra unikalios, jos nebuvo įtrauktos į analizę.

Atributo pavadinimas	Kiekis (Eiluciu sk.)	Trukstamos reiksmes, %	Kardinalumas	Minimali reiksme	Maksimali reiksme	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
wait	840	0	26	0	99	1	53	34.58928571	35	24.94860757
vcost	840	0	135	2	180	23	67	47.76071429	39	32.37100381
travel	840	0	405	63	1440	235	797	486.1654762	397	301.4391069
gcost	840	0	184	30	269	71	144	110.8797619	102	47.97835298
income	840	0	24	2	72	20	50	34.54761905	35	19.67604423
size	840	0	6	1	6	1	2	1.742857143	1	1.010349981

Galima pastebėti, kad trūkstamų reikšmių nėra. Reikšmių kardinalumas nėra labai didelis, „size“ duomenų unikalumas yra lygus 6, todėl šiuos duomenis būtų galima pakeisti į kategorinius duomenis.

Išmetus „size“ atributą yra gaunama ši lentelė:

Atributo pavadinimas	Kiekis (Eiluciu sk.)	Trukstamos reiksmes, %	Kardinalumas	Minimali reiksme	Maksimali reiksme	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
wait	840	0	26	0	99	1	53	34.58929	35	24.94861
vcost	840	0	135	2	180	23	67	47.76071	39	32.371
travel	840	0	405	63	1440	235	797	486.1655	397	301.4391
gcost	840	0	184	30	269	71	144	110.8798	102	47.97835
income	840	0	24	2	72	20	50	34.54762	35	19.67604

3. Kategorinių tipų analizė

Reikia apskaičiuoti:

- Bendrą reikšmių skaičių;
- Trūkstamų reikšmių procentą;
- Kardinalumą;
- Modą;
- Modos dažnumo reikšmę;
- Modos procentinę reikšmę;
- 2-ąją modą;
- 2-osios modos dažnumo reikšmę;
- 2-osios modos procentinę reikšmę.

Atributo pavadinimas	Kiekis (Eiluciu sk.)	Trukstamos reiksmes, %	Kardinalumas	Moda	Modos daznumas	Moda, %	2-oji moda	2-osios modos daznumas	2-oji moda, %
mode	840	0	4	car	210	25	bus	210	25
choise	840	0	2	no	630	75	yes	210	25

Galima pastebėti, kad trūkstamų reikšmių nėra. Kadangi visų reikšmių „mode“ pasiskirstymas buvo vienodas, po 25%, todėl modų dažnumas sutampa.

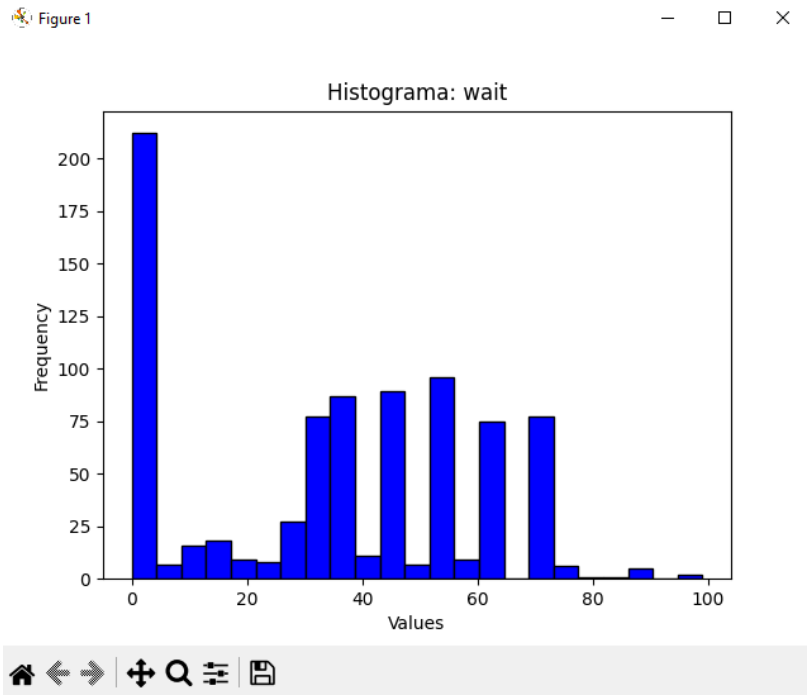
Papildžius lentelę „size“ atributu ji atrodo šitaip:

Atributo pavadinimas	Kiekis (Eiluciu sk.)	Truksta mos reiksmes, %	Kardinalumas	Moda	Modos daznumas	Moda, %	2-oji moda	2-osios modos daznumas	2-oji moda, %
mode	840	0	4	car	210	25	bus	210	25
choise	840	0	2	no	630	75	yes	210	25
size	840	0	6	Labai mazai	456	54.28571	Mazai	232	27.61905

„Size“ reikšmės buvo nuo 1 iki 6 ir jos pakeistos atitinkamais žodžiais: labai mažai, mažai, vidutiniskai, virš vidurkio, daug, labai daug. Galima pastebėti, kad dominuoja 1 arba 2 keleivių grupės.

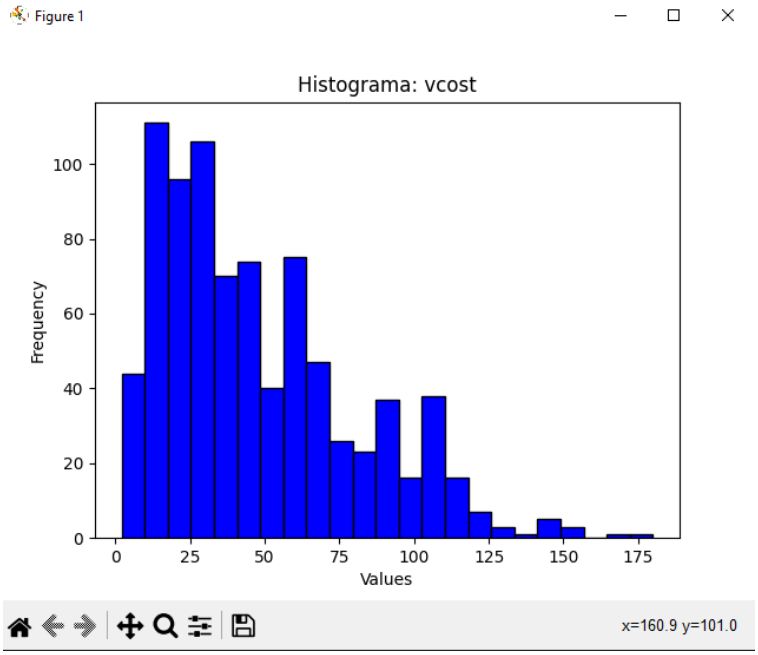
4. Atributų histogramos

Stulpelių skaičius randamas pagal formulę: 1 + 3.22 * log_e^n. Imties dydis n: 840. Gautas stulpelių skaičius: 23.



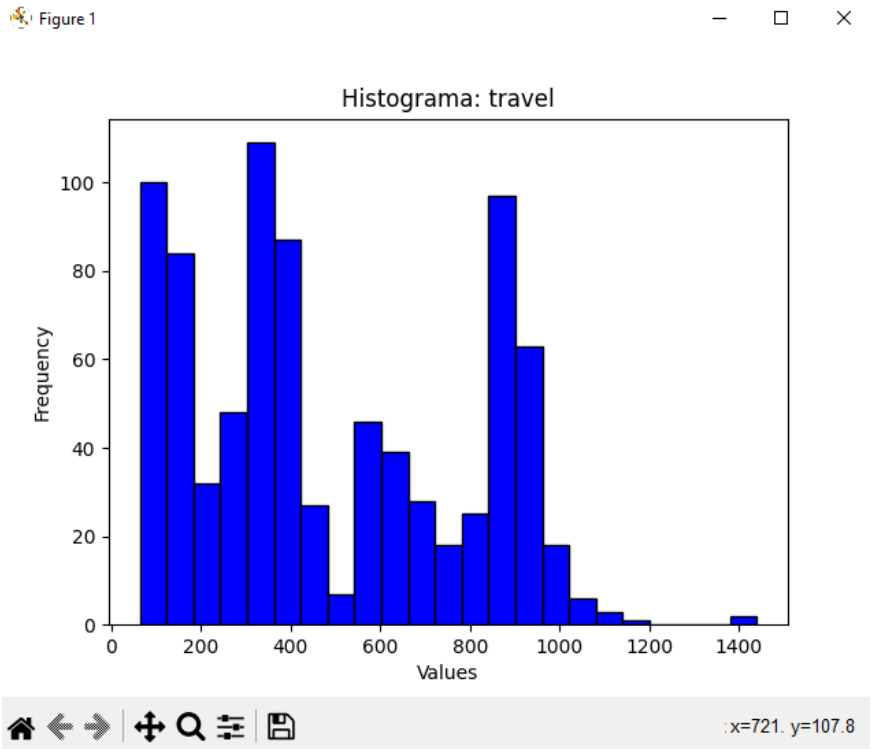
1 pav. „Wait" histograma

Galima teigti, kad labai daug žmonių pasirinko keliones automobiliu, nes yra labai daug reikšmių artimų 0.



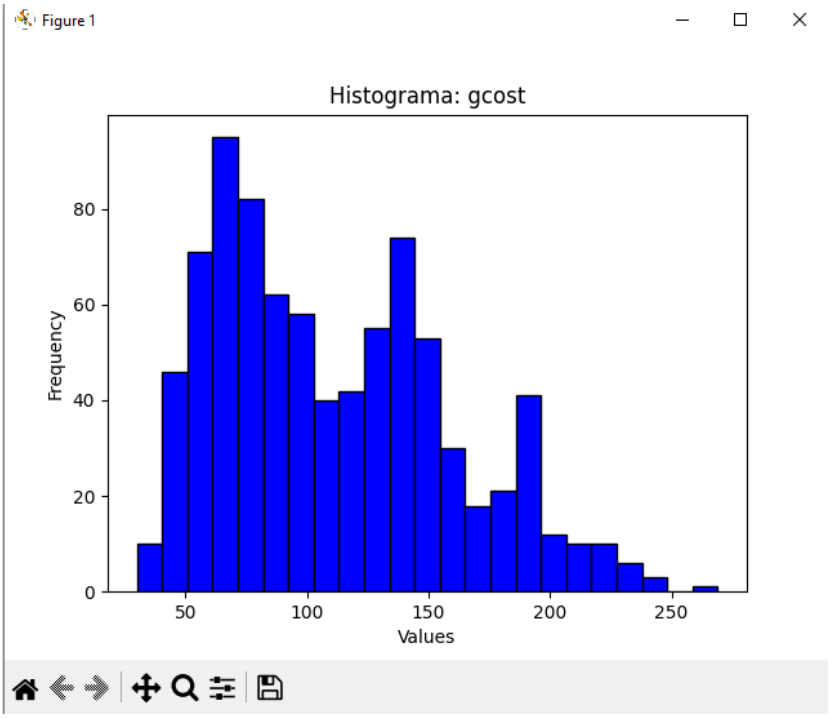
2 Pav. „Vcost" histograma

Žmonės yra linkę rinktis keliones su nedidelėmis transporto priemonių kainomis.



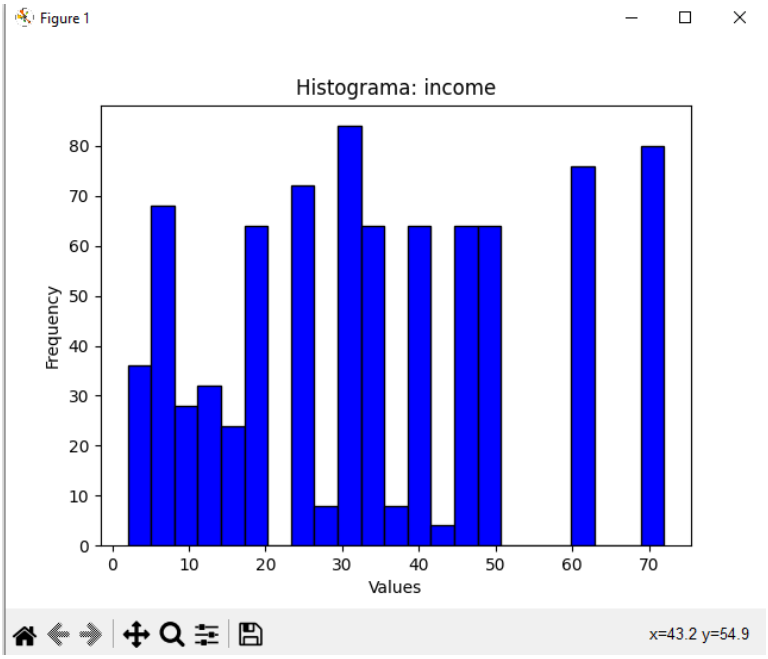
3 Pav. „Travel" histograma

Kelionių laikas labai svyruoja.



4 Pav. „Gcost" histograma

Bendros kelionės kainos duomenų išsidėstymas yra panašus į transporto priemonės kainos duomenis, žmonės yra linkę mokėti mažiau.



5 Pav. „Income" histograma

Iš žmonių uždarbio duomenų nieko tikslaus negalima pasakyti.

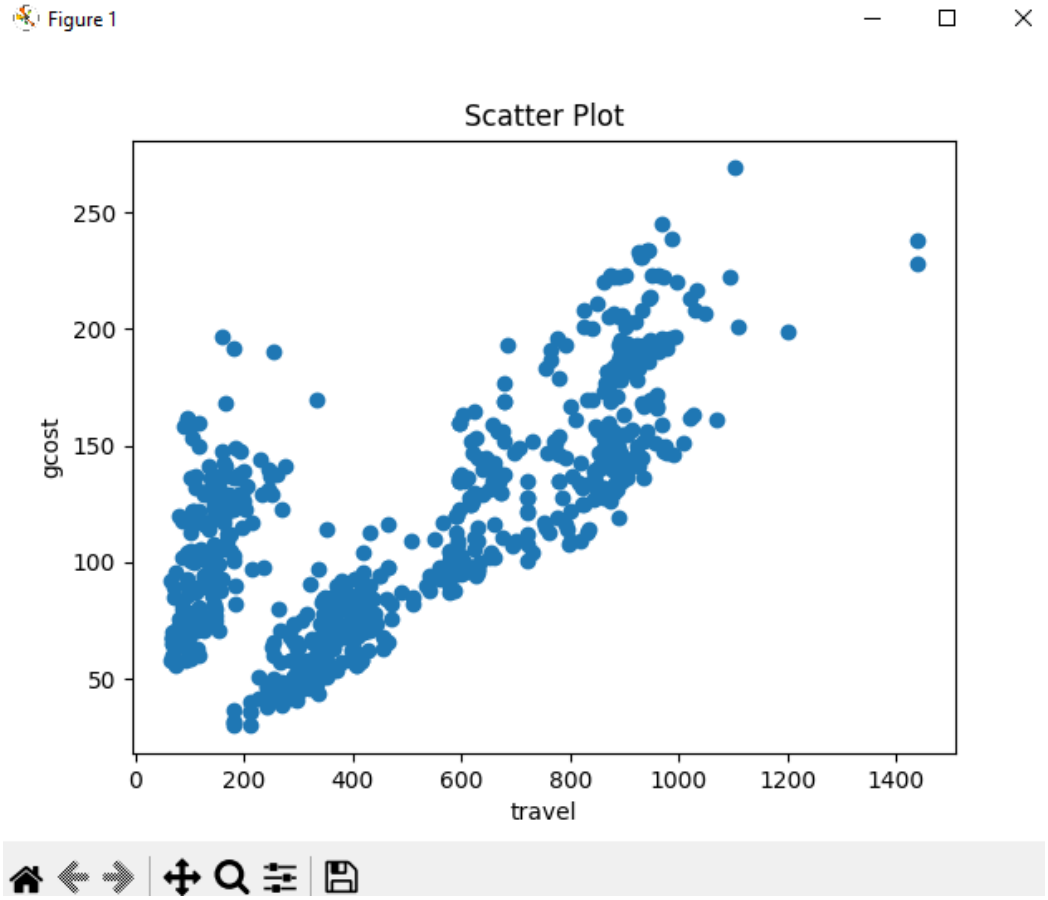
5. Duomenų kokybės problemos

Galima teigti, kad tolydinių duomenų kardinalumas nėra labai didelis ir jų pasiskirstymas histogramose nenusako tam tikros aiškios tendencijos tarp žmonių kiekio ir reikšmės. Taip pat labai mažo kardinalumo duomenis galima pakeisti kategoriniais duomenimis.

6. Ryšiai tarp atributų

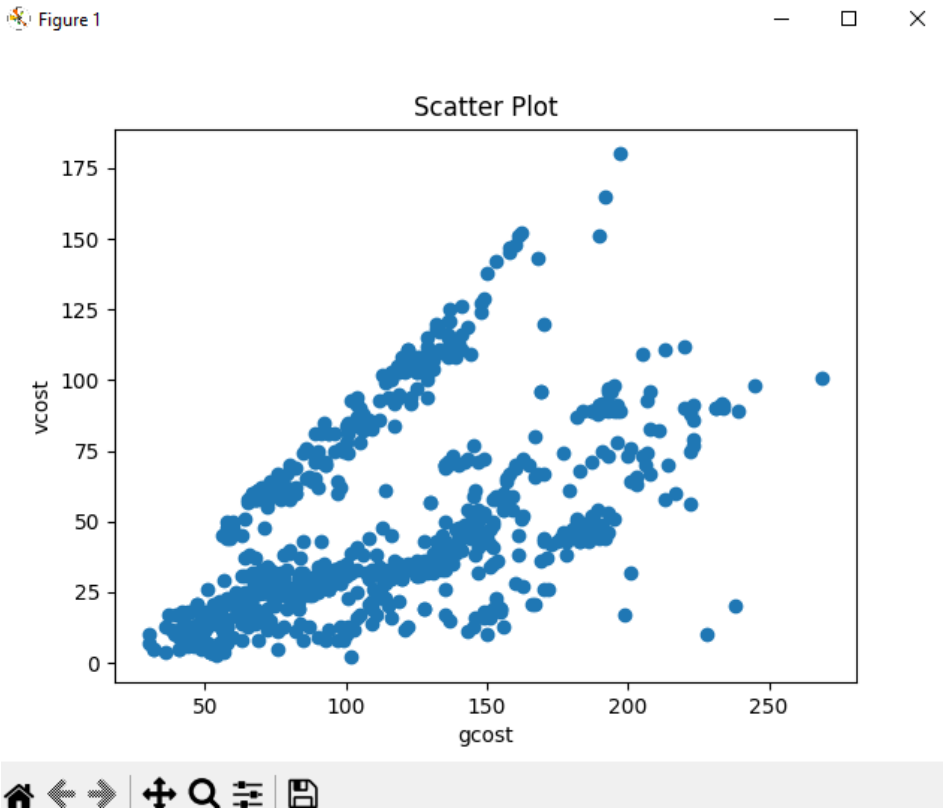
6.1. Scatter plot diagramos tolydiniams duomenims atvarizduoti

Gana stiprias tiesiogines atributų priklausomybes galima pastebėte šiuose grafikuose:



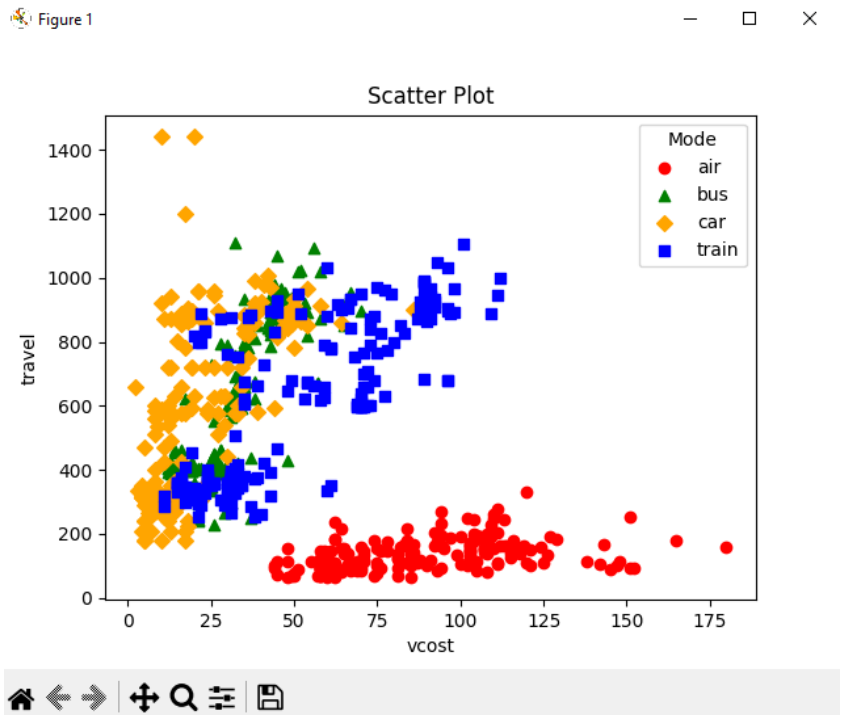
6 Pav. „Gcost" ir „travel" scatter plot grafikas

Matosi, kad didžioji duomenų dalis rodo tiesioginę priklausomybę tarp visos kelionės kainos ir kelionės laiko. Kuo ilgiau trunka kelionė tuo didesnė visos kelionės kaina.



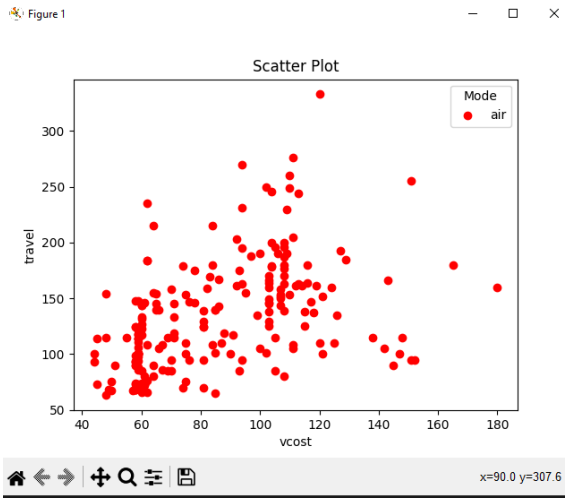
7 Pav. „Vcost" ir „gcost" scatter plot grafikas

Akyvaizdu, kad transporto kainos ir visos kelionės kainos duomenys tiesiogiai priklauso vienas nuo kito, nes transporto kaina įeina į visos kelionės kainos bendrą sumą. Kuo didesnės išlaidos yra skiriamos transportui tuo didesnė gaunasi visos kelionės kaina.



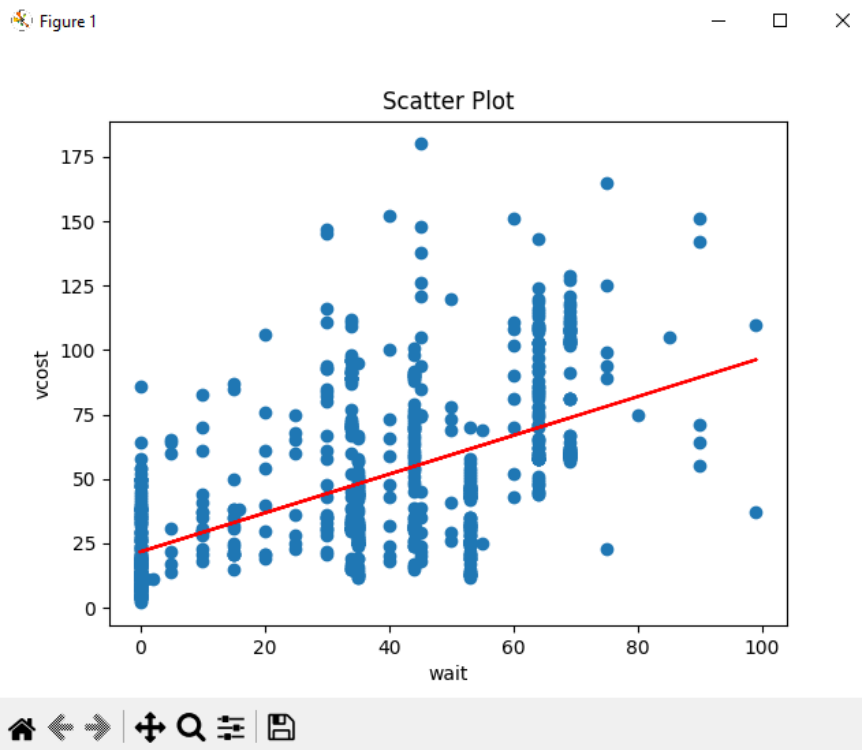
8 Pav. „Travel" ir „vcost" scatter plot grafikas visoms transporto priemonės atvaizduoti

Iš gautų rezultatų butų galima išskirti dvi grupes reikšmių: vieną perteikenčią tiesišką priklausomybę tarp kelionės laiko ir transporto priemonės kainos – kuo ilgiau trunka kelionė tuo transporto priemonė yra brangesnė ir kita grupę, kuriai kelionė truko daug trumpiau, todėl atrodo, kad kelionės laikas nedarė įtakos jos kainai (tai netiesa). Akyvaizdu, kad kelionės lėktuvu trukdavo daug trumpiau nei kitomis transporto priemonėmis.



9 Pav. „Travel" ir „vcost" scatter plot grafikas lėktuvo duomenims perteikti

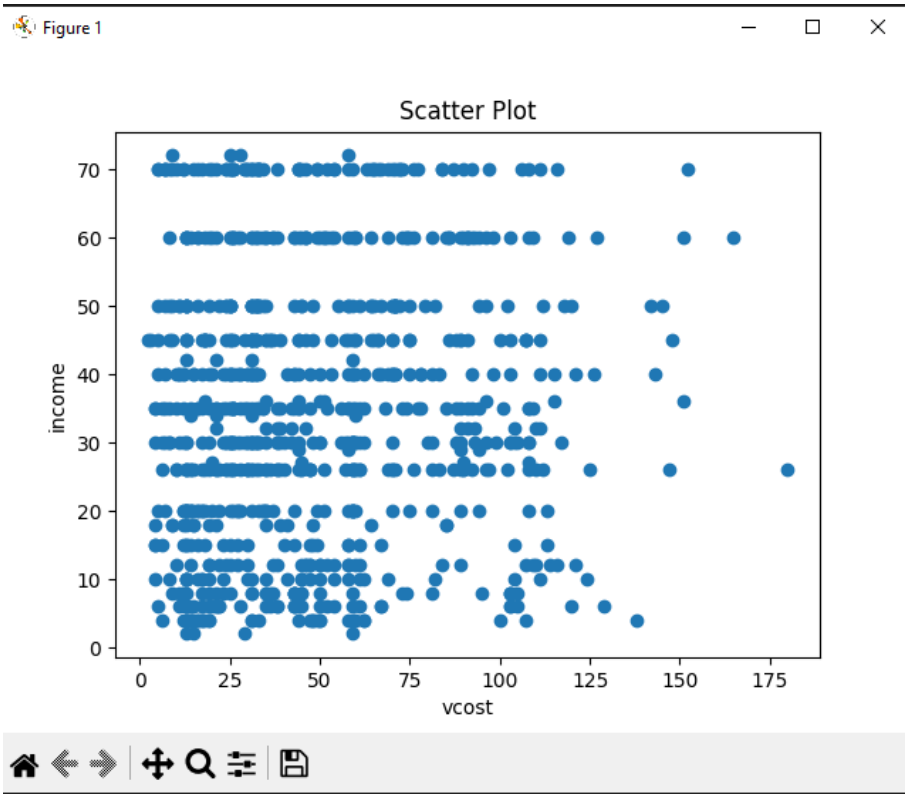
Jei atsižvelgtume tik į kelionę oru, būtų galima taip pat pastebėti tiesišką duomenų priklausomubę.



10 Pav. „Vcost" ir „wait" scatter plot grafikas

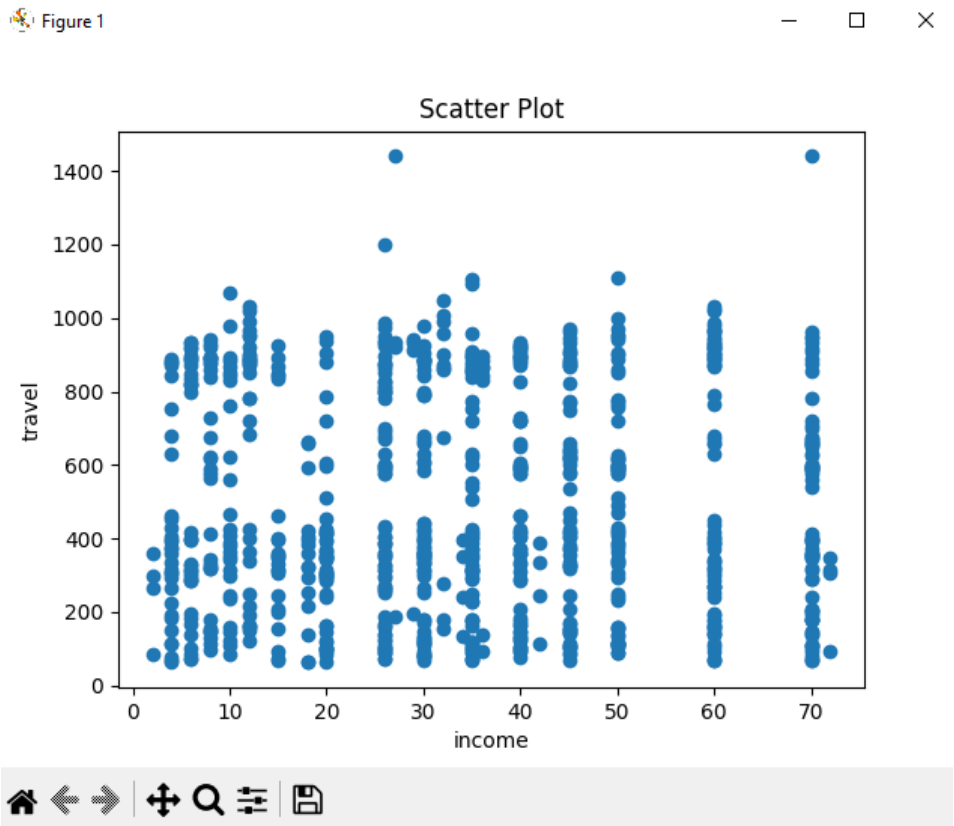
Nubrėžus linijinės progresijos tiesę per duomenis galima pastebėti gana stiprę priklausomybę tarp “vcost-wait” atributų – tai patvirtins vėliau apskaičiuoti kovariacijos ir koreliacijos koficientai. Nors duomenys nėra linkę stipriai grupuotis tarpusavyje, vis vien galima pastebęti jų sąryšį.

Tarpusavyje nekoreliuojantys atributai:



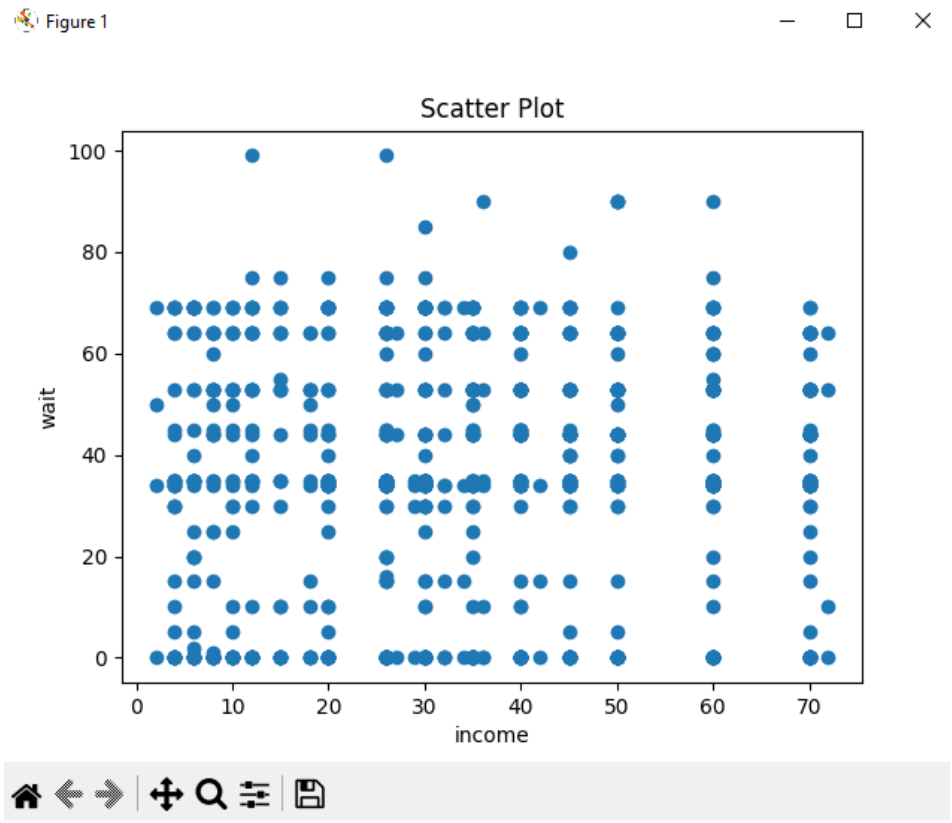
11 Pav. „Income" ir „vcost" scatter plot grafikas

Akyvaizdu, kad duomenys labai silpnai tarpusavyje koreliuoja, negalima pastebėti jokios tendencijos tarp transporto priemonės kainos ir žmonių uždarbio.



12 Pav. „Travel" ir „income" scatter plot grafikas

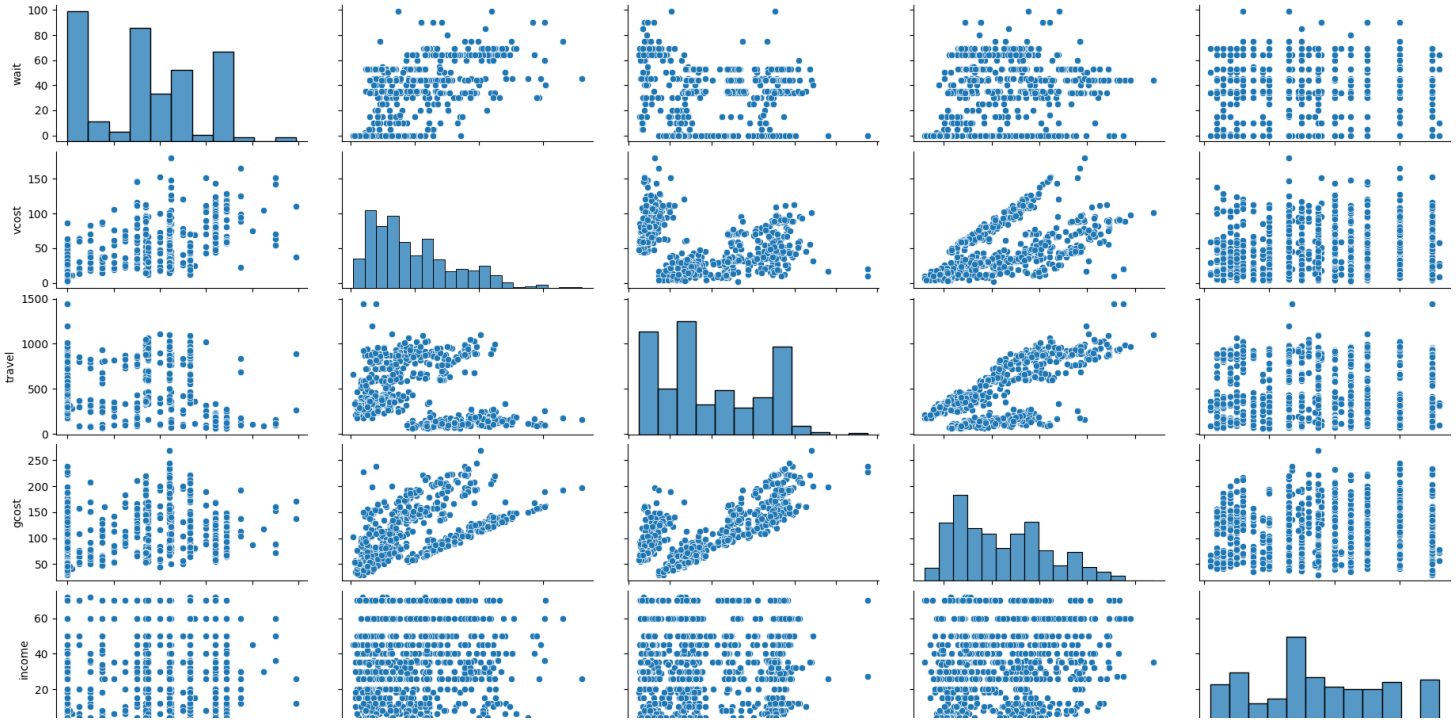
Kelionės laikas ir uždarbis tarpusavyje nekoreliuoja.



13 Pav. „Wait" ir „income" scatter plot grafikas

Laukimo laikas ir uždarbis, taip pat yra tarpusavyje nekorialiuojantys atributai.

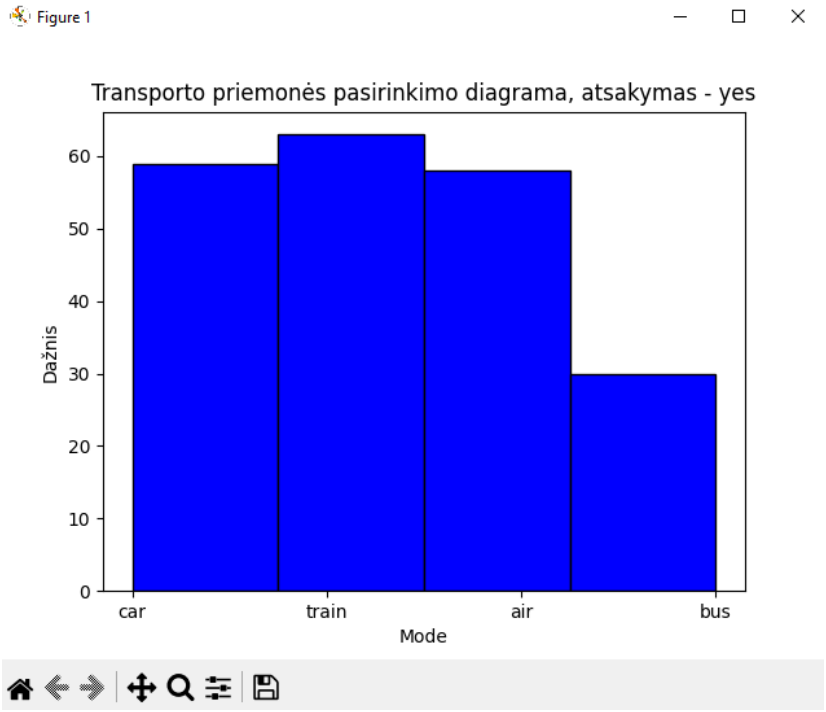
6.2. SPLOM diagrama



14 Pav. SPLOM matrica

6.3. Bar plot diagrama kategoriniams duomenims atvaizduoti

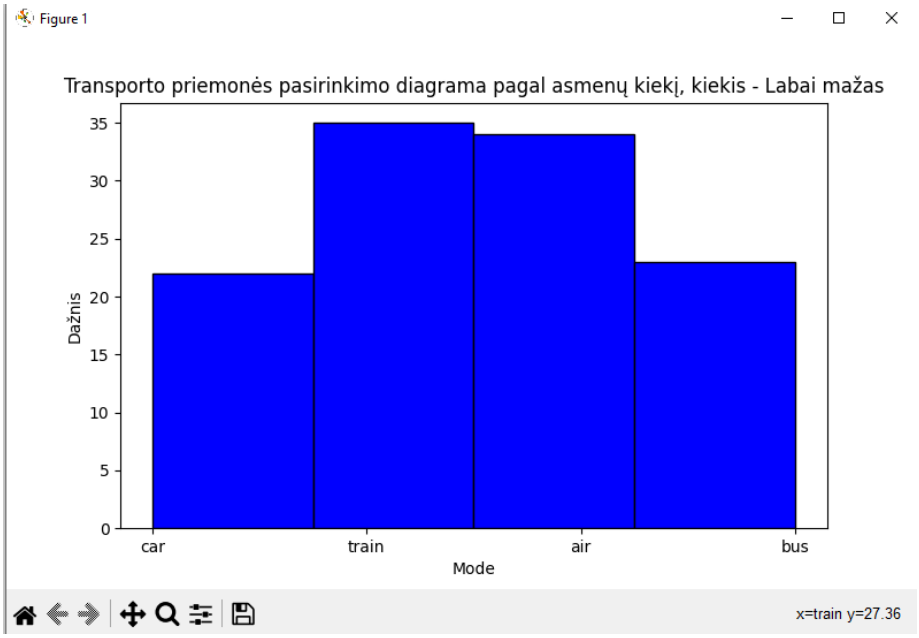
Iš pradžių nagrinėsiu kategorinių duomenų priklausomybę tarp transporto priemonės ir kiek žmonių ją rinkosi.



15 Pav. Žmonių sutikusių rinktis atitinkamas transporto priemones diagrama

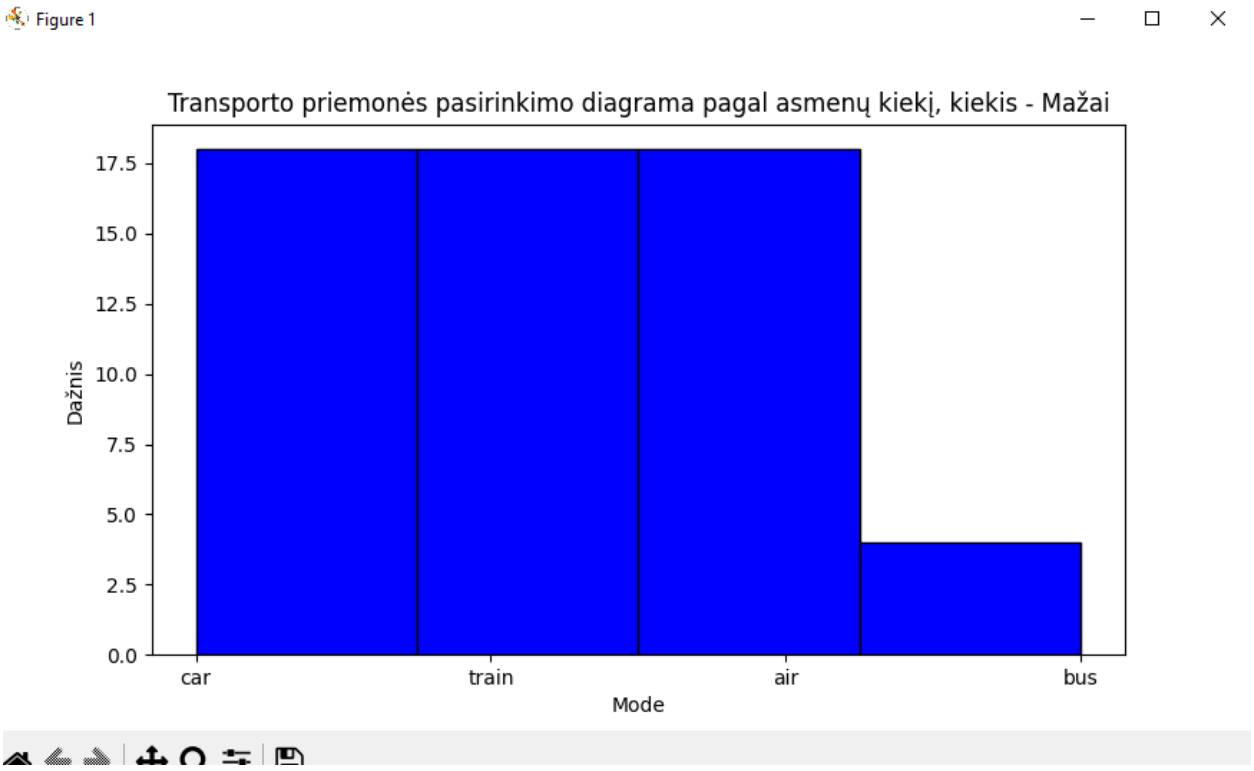
Iš gautų rezultatų diagramoje matosi, kad daugiausiai žmonių rinkosi keliauti traukiniu ir mažiausiai autobusu.

Toliau nagrinėsime transporto pasirinkimo tendencijas atsižvelgiant į keliaujančių asmenų grupės dydį.



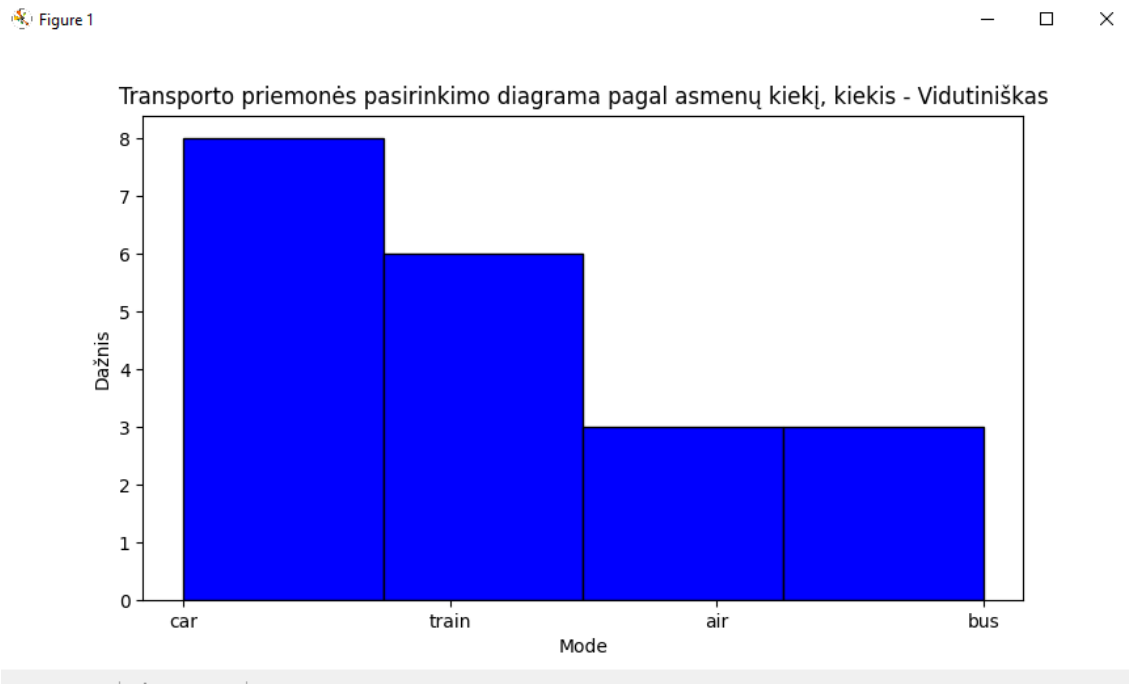
16 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra labai mažas

Pagal gautus rezultatus akivaizdu, kad gautas duomenų pasiskirstymo rezultatas yra labai panašus į transporto priemonės pasirinkimo rezultatą, kai į žmonių grupės kiekį nebuvo atsižvelgta. Daugiausiai žmonių vis dar renkasi traukinį ir mažiausiai – autobusą arba mašiną.



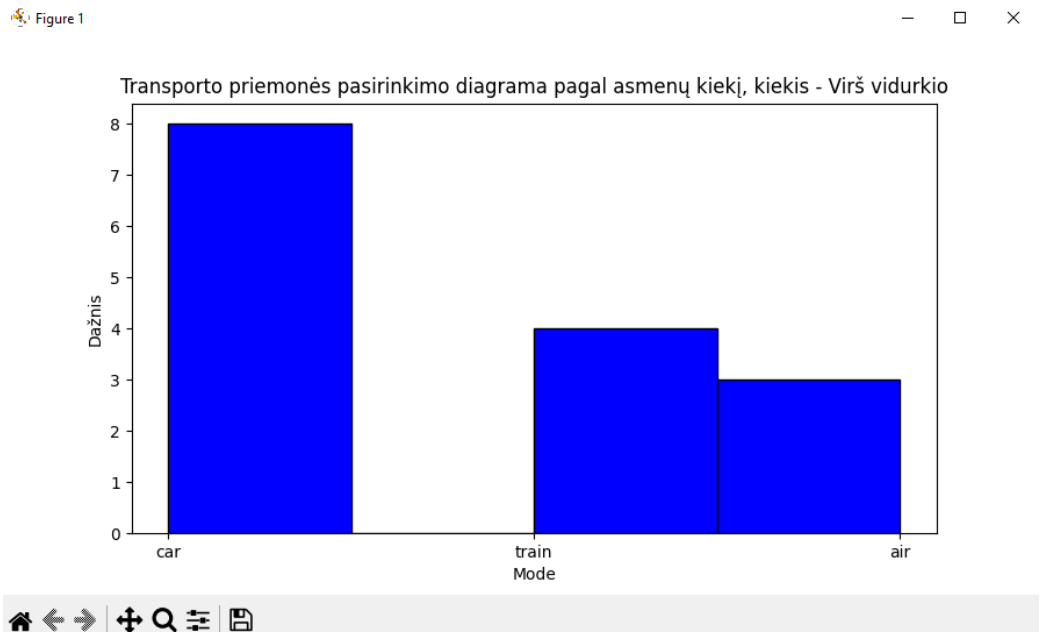
17 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra mažas

Asmenų kiekiui grupėje pasikeitus – išaugus iki dviejų asmenų mašinos, traukinio ir kelionės lėktuvu pasirinkimai susivienodina.

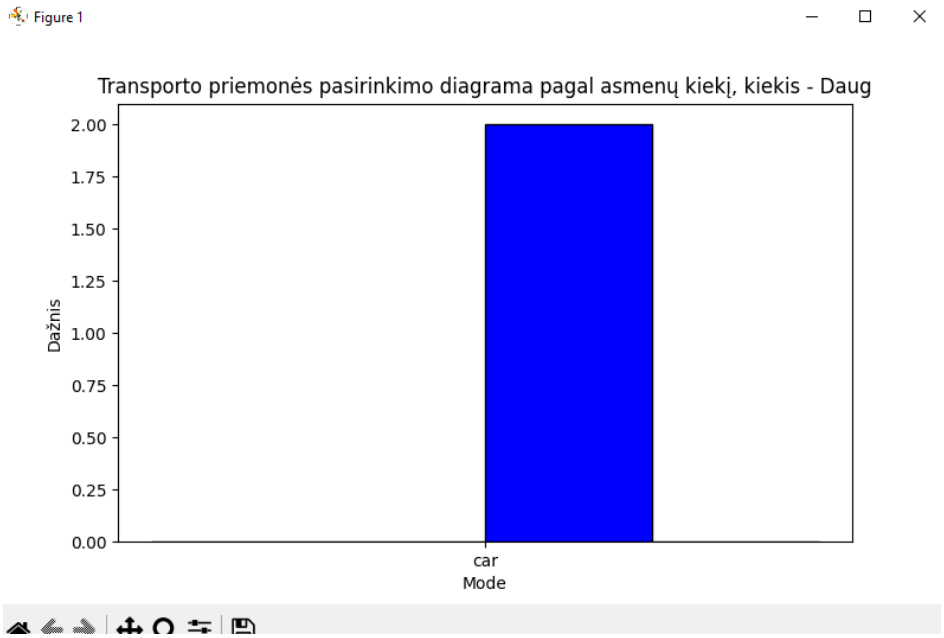


18 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra vidutiniškas

Šioje diagramoje galima pastebėti, kad išaugus asmenų grupėje dydžiui žmonės linkę rinktis mašiną. Vis dar labai mažai žmonių renkasi autobusą.



19 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra virš vidurkio



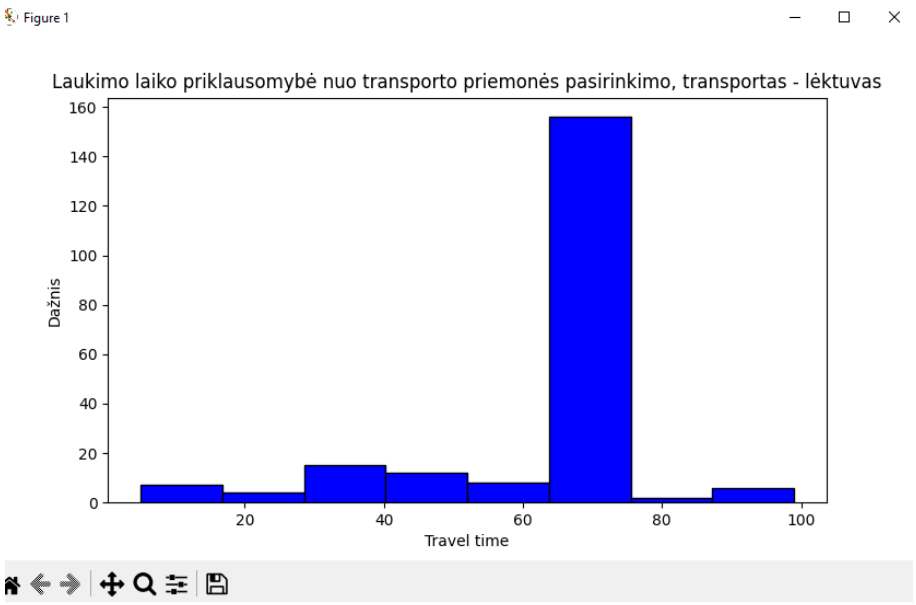
20 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra didelis

Akyvaizdu, kad žmonių kiekiui augant transporto priemonės pasirinkimas išlieka mašina. Svarbu pastebėti, kad duomenų kiekis vis mažėja, panašu, kad daugiau žmonių yra linkę keliauti labai mažose grupėse (vieni). Mažiausiai populiarumo susilaukė kelionės autobusu.

6.4. Bar plot ir box plot diagramos atvaizduojančios kategorinio ir tolydinio tipo kintamųjų sąryšius

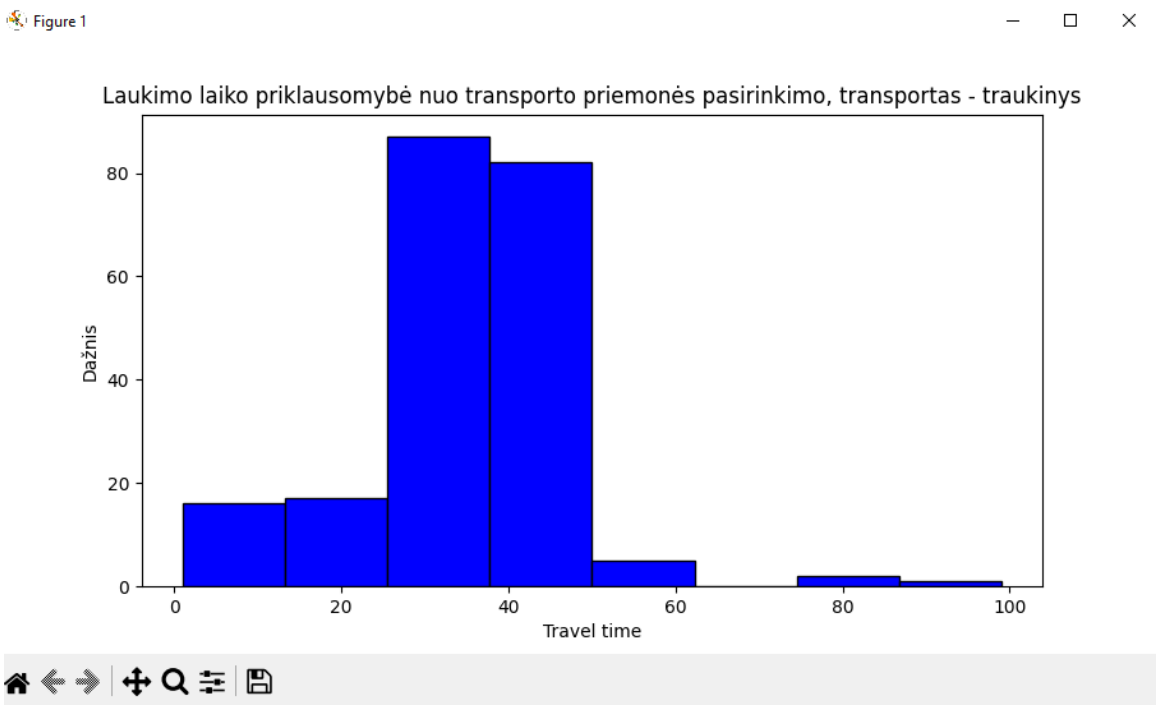
6.4.1. Bar plot diagramos

Nagrinėsiu laukimo laiką terminale, kai yra pasirenkama atitinkama transporto priemonė. Laukimo laikas yra 0, kai keliaujama mašina.



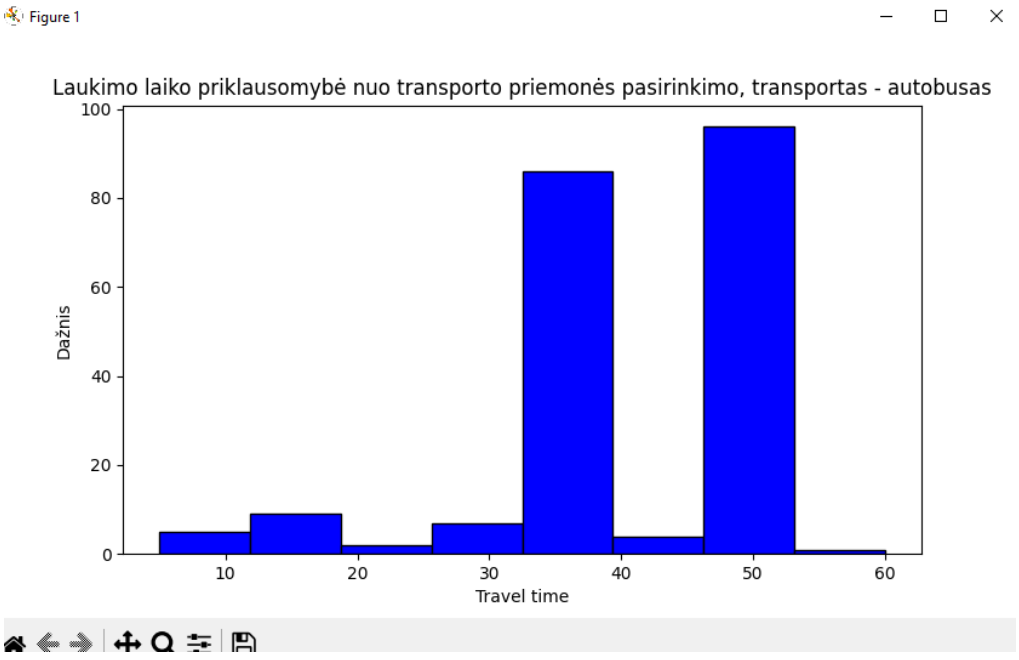
21 Pav. Laukimo laiko priklausomybės nuo transporto priemonės pasirinkimo diagrama, kai ji yra lėtuvas

Dažniausiai kelionės lėktuvu laukimas terminale užtrunka gana ilgai, būtent tai atspindi diagramos rezultatai.



22 Pav. Laukimo laiko priklausomybės nuo transporto priemonės pasirinkimo diagrama, kai ji yra traukinys

Išmetus triukšmus – ekstremalias reikšmes galima pastebėti, kad traukinio laukimas užtrunka mažiau nei lėktuvo ir kad duomenys yra gana vienareikšmiški.

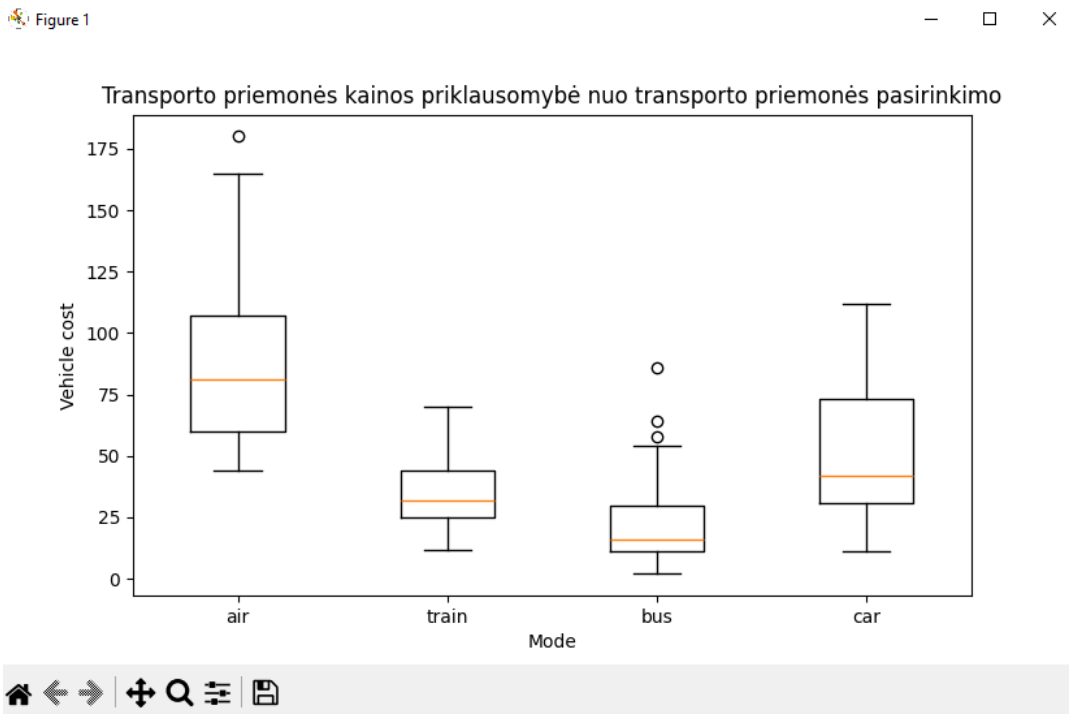


23 Pav. Laukimo laiko priklausomybės nuo transporto priemonės pasirinkimo diagrama, kai ji yra autobusas

Autobuso laukimas užtrunka ilgiau negu traukinio, bet mažiau negu lėktuvo.

6.4.2. Box plot diagramos

Bus nagrinėjami transporto priemonės kainos duomenys.



24 Pav. Box plot diagrama perteikianti transporto priemonės pasirinkimo ir kainos sąryšius

Galima pastebėti, kad nestandartinių duomenų nėra daug. Ryšys tarp atributų nėra labai geras, nes duomenys tarpusavyje persidengia, reiškia transporto priemonių kainos yra įvairios ir nėra glaudžiai susijusios. Jeigu butų imamos 1 ir 3 kvartilų reikšmės, persidengimas gautųsi geresnis.

7. Kovariacijos ir koreliacijos reikšmės

Galima įrodyti ryšį tarp atributų paskaičiuojant kovariacijos ir koreliacijos koeficientus, jiems rasti yra naudojamos skirtingos formulės

7.1. Kovariacija

$$cov(a,b) = \frac{1}{n-1}\sum_{i=1}^{n}((a_i - \bar{a}) \times (b_i - \bar{b}))$$

Naudojant šią formulę galima paskaičiuoti kovariacijas tarp atributų. Kuo didesnis gaunamas koeficientas tuo geresnis ryšys tarp atributų.

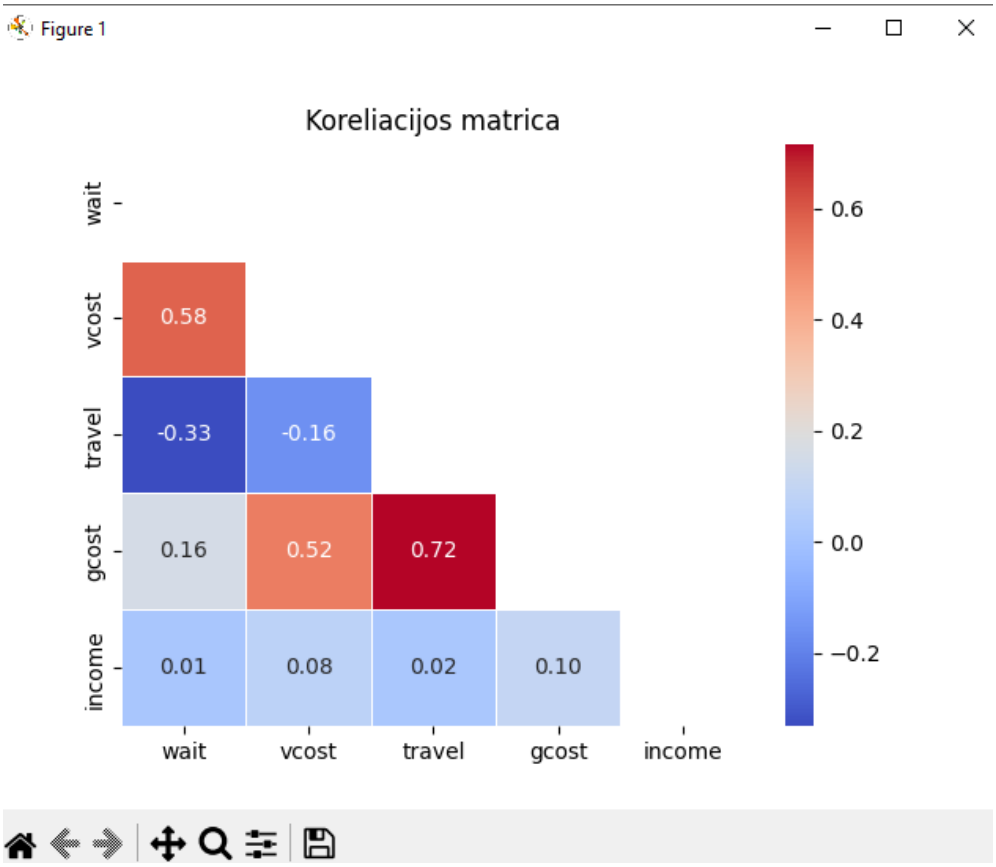
	A	B	C	D	E	F
1		wait	vcost	travel	gcost	income
2	wait					
3	vcost	468.7264				
4	travel	-2497.79	-1561.31			
5	gcost	189.0602	814.6148	10374.14		
6	income	6.563681	47.85706	107.889	93.45091	
7						

Pagal gautus duomenis kai kurie atributai tarpusavyje yra stipriai susiję, pvz: „travel-gcost“, „vcost-gcost“. Tuo tarpu egzistuoja reikšmių, kurios yra neigiamai susijusios, kai viena reikšmės auga, o kita krenta: „wait-travel“.

7.2.Koreliacija

Koreliacija, taip pat skirta parodyti ryšį tarp atributų, tačiau ji yra normalizuota. Gaunami rezultatai intervale [-1;1].

$$corr(a,b) = \frac{cov(a,b)}{sd(a) \times sd(b)}$$



25 Pav. Tolydinių duomenų koreliacijos matrica

Stipriausia koreliacija pastebima tarp „vcost-wait“, „gcost-travel“ ir „gcost-vcost“. Yra kelios reikšmės kurios gan artimos 0, jos reikštų, kad atributai vienas su kitu yra nesusiję.

8. Duomenų normalizacija

Normalizavimas reikšmių leidžia pakeisti jų diapazonus taip, kad būtų išlaikyti santykiniai jų skirtumai ir nebūtų labai išsiskiriančių reikšmių. Bus naudojamas „range normalization“ būdas tai atlikti. Reikšmės yra intervale [0;1]. Formulė:

$$z = \frac{x - min(x)}{[\max(x) - min(x)]}$$

	A	B	C	D	E	F	G	H	I	J	K
1	rowname	individual	mode	choice	wait	vcost	travel	gcost	income	size	
2	1	1	air	no	0.69697	0.320225	0.02687	0.167364	0.471429	1	
3	2	1	train	no	0.343434	0.162921	0.224401	0.171548	0.471429	1	
4	3	1	bus	no	0.353535	0.129213	0.257081	0.167364	0.471429	1	
5	4	1	car	yes	0	0.044944	0.084967	0	0.471429	1	
6	5	2	air	no	0.646465	0.314607	0.003631	0.158996	0.4	2	
7	6	2	train	no	0.444444	0.162921	0.211329	0.225941	0.4	2	
8	7	2	bus	no	0.535354	0.129213	0.244009	0.230126	0.4	2	
9	8	2	car	yes	0	0.050562	0.139434	0.083682	0.4	2	
10	9	3	air	no	0.69697	0.634831	0.045025	0.414226	0.542857	1	
11	10	3	train	no	0.343434	0.539326	0.602033	0.690377	0.542857	1	
12	11	3	bus	no	0.353535	0.286517	0.594771	0.497908	0.542857	1	
13	12	3	car	yes	0	0.117978	0.477124	0.297071	0.542857	1	
14	13	4	air	no	0.646465	0.264045	0.003631	0.121339	0.971429	3	
15	14	4	train	no	0.444444	0.134831	0.211329	0.205021	0.971429	3	
16	15	4	bus	no	0.535354	0.106742	0.244009	0.213389	0.971429	3	
17	16	4	car	yes	0	0.016854	0.084967	0.008368	0.971429	3	
18	17	5	air	no	0.646465	0.325843	0.058824	0.217573	0.614286	2	
19	18	5	train	no	0.444444	0.168539	0.24764	0.263598	0.614286	2	
20	19	5	bus	no	0.535354	0.134831	0.28032	0.267782	0.614286	2	
21	20	5	car	yes	0	0.033708	0.389978	0.288703	0.614286	2	
22	21	6	air	no	0.69697	0.320225	0.02687	0.167364	0.257143	1	
23	22	6	train	yes	0.40404	0.101124	0.204793	0.112971	0.257143	1	

Normalizuoti 5 stulpeliai: wait, vcost, travel, gcost ir income.

9. Vertimas tolydiniais duomenimis

Bus kovertuojami mode ir choice stulpeliai į kategorinius duomenis.

Šių stulpelių eilutės keičiamos skaičiais:

- Mode: air – 1, train – 2, bus – 3, car – 4.
- Choice: yes – 1, no – 0.

```
Mode: air, number: 1
Mode: train, number: 2
Mode: bus, number: 3
Mode: car, number: 4
Mode: air, number: 1
Mode: train, number: 2
Mode: bus, number: 3
Mode: car, number: 4
Mode: air, number: 1
Mode: train, number: 2
```

26 Pav. „Mode" kategorinių duomenų vertimas tolydiniais

```
Choice: no, number: 0
Choice: no, number: 0
Choice: no, number: 0
Choice: yes, number: 1
Choice: no, number: 0
Choice: no, number: 0
```

27 Pav. „Choice" kategorinių duomenų vertimas tolydiniais

10. Išvados

Pasirinktame duomenų rinkinyje nebuvo tuščių reikšmių, todėl nei pildyti eilutes duomenimis ar jas trinti nereikėjo. Reikšmių kardinalumas tolydiniams duomenims nebuvo labai geras, vis dėl to išsiskiriančių duomenų pakako, kad būtų galima pastebėti tam tikras jų tendencijas.

Pagal gautus kovariacijos ir koreliacijos koeficientus buvo rasta tarpusavyje susijusių nagrinėjamų reikšmių. Šios reikšmės leido daryti išvadas apie duomenų išsidėstymą.

Buvo tolydinių duomenų, kurių kardinalumas labai mažas, todėl jie buvo pakeisti į kardinalias reikšmes.