**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**

1. Duomenų rinkinys 3

2. Tolydinių tipų analizė 4

3. Kategorinių tipų analizė 5

4. Atributų histogramos 6

5. Duomenų kokybės problemos 9

6. Ryšiai tarp atributų 9

6.1. Scatter plot diagramos tolydiniams duomenims atvarizduoti 9

6.2. SPLOM diagrama 16

6.3. Bar plot diagrama kategoriniams duomenims atvaizduoti 17

6.4. Bar plot ir box plot diagramos atvaizduojančios kategorinio ir tolydinio tipo kintamųjų sąryšius 20

6.4.1. Bar plot diagramos 20

6.4.2. Box plot diagramos 22

7. Kovariacijos ir koreliacijos reikšmės 23

7.1. Kovariacija 23

7.2. Koreliacija 24

8. Duomenų normalizacija 25

9. Vertimas tolydiniais duomenimis 26

10. Išvados 27

**PAVEIKSLĖLIAI**

1 pav. „Wait" histograma 8

2 Pav. „Vcost" histograma 8

3 Pav. „Travel" histograma 9

4 Pav. „Gcost" histograma 10

5 Pav. „Income" histograma 10

6 Pav. „Gcost" ir „travel" scatter plot grafikas 11

7 Pav. „Vcost" ir „gcost" scatter plot grafikas 12

8 Pav. „Travel" ir „vcost" scatter plot grafikas visoms transporto priemonės atvaizduoti 13

9 Pav. „Travel" ir „vcost" scatter plot grafikas lėktuvo duomenims perteikti 13

10 Pav. „Vcost" ir „wait" scatter plot grafikas 14

11 Pav. „Income" ir „vcost" scatter plot grafikas 15

12 Pav. „Travel" ir „income" scatter plot grafikas 16

13 Pav. „Wait" ir „income" scatter plot grafikas 17

14 Pav. SPLOM matrica 17

15 Pav. Žmonių sutikusių rinktis atitinkamas transporto priemones diagrama 18

16 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra labai mažas 19

17 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra mažas 19

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

19 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra virš vidurkio 21

20 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra didelis 21

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

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

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

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

25 Pav. Tolydinių duomenų koreliacijos matrica 26

26 Pav. „Mode" kategorinių duomenų vertimas tolydiniais 28

27 Pav. „Choice" kategorinių duomenų vertimas tolydiniais 28

# 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.

# 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 | Standartitnis 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 | Standartitnis 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 |

# 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.

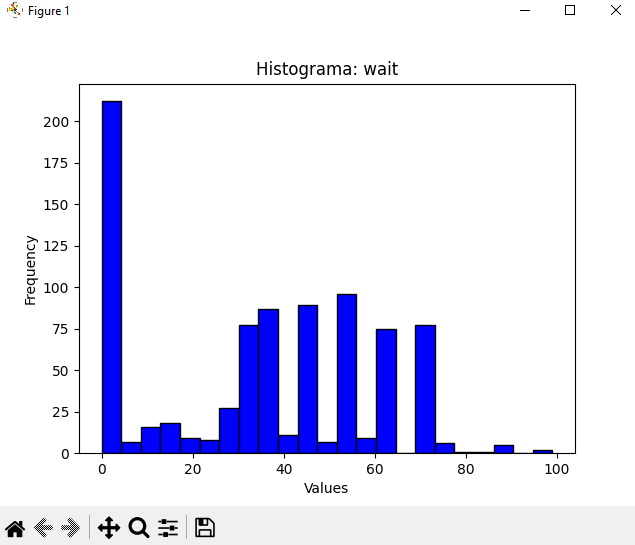
Papildžius lentelę „size“ atributu ji atrodo šitaip:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atributo pavadinimas | Kiekis (Eiluciu sk.) | 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 |
| 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.

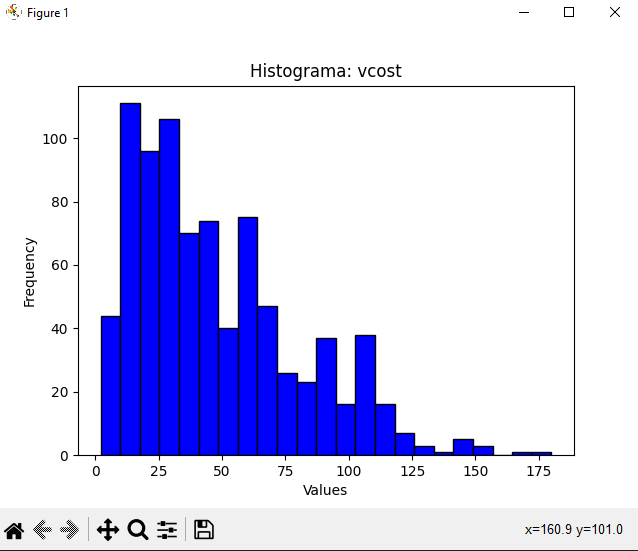
# Atributų histogramos

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



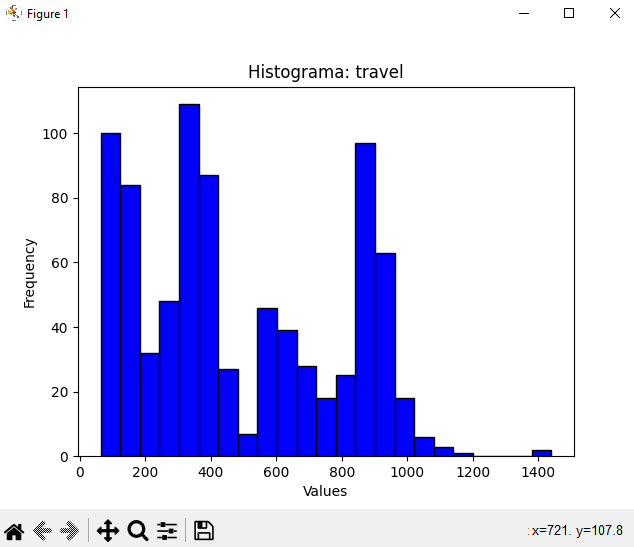
1 pav. „Wait" histograma

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



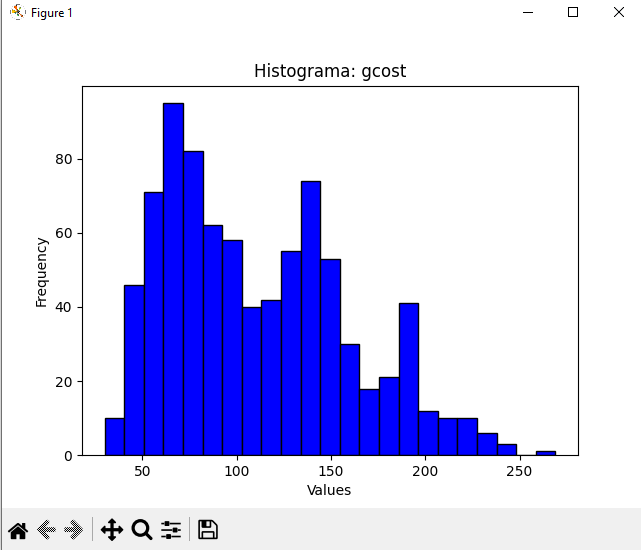
2 Pav. „Vcost" histograma

Žmonės yra linkę rinktis keliones su nedidelėmis transporto priemonių kainomis.



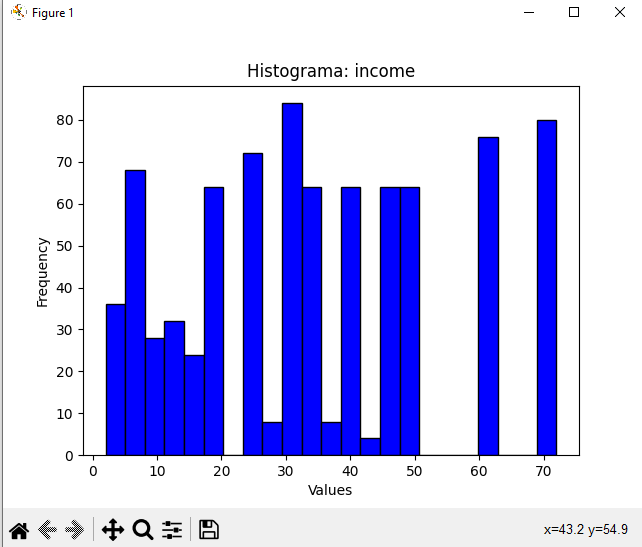
3 Pav. „Travel" histograma

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.

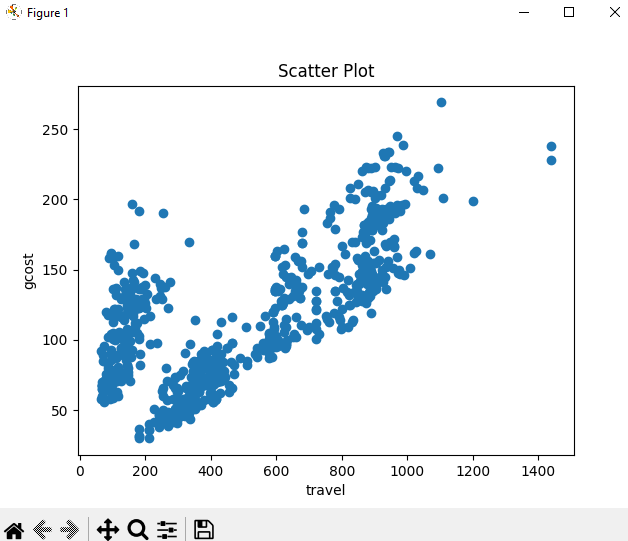
# Duomenų kokybės problemos

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.

# Ryšiai tarp atributų

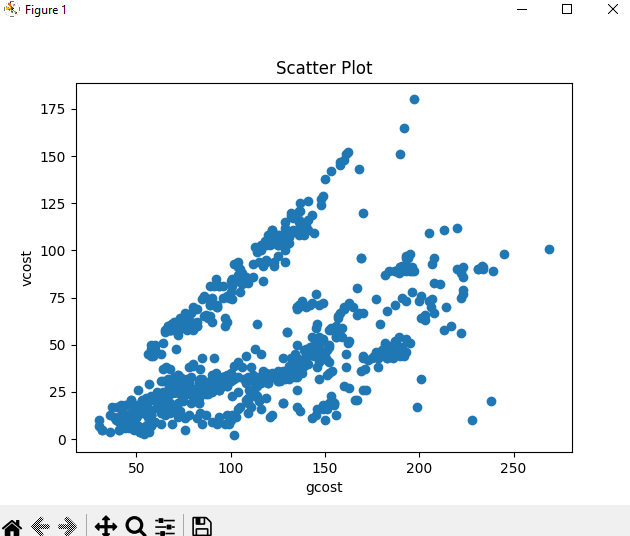
## Scatter plot diagramos tolydiniams duomenims atvarizduoti

Gana stiprias tiesiogines atributų priklausomybes galima pastebėte šiuose grafikuose:



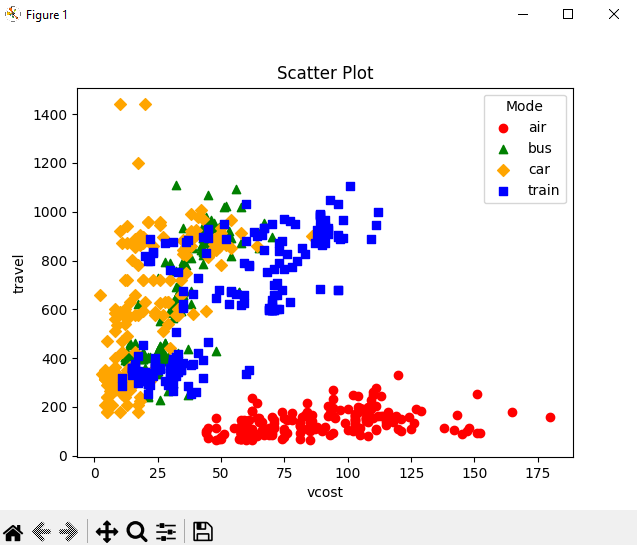
6 Pav. „Gcost" ir „travel" scatter plot grafikas

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



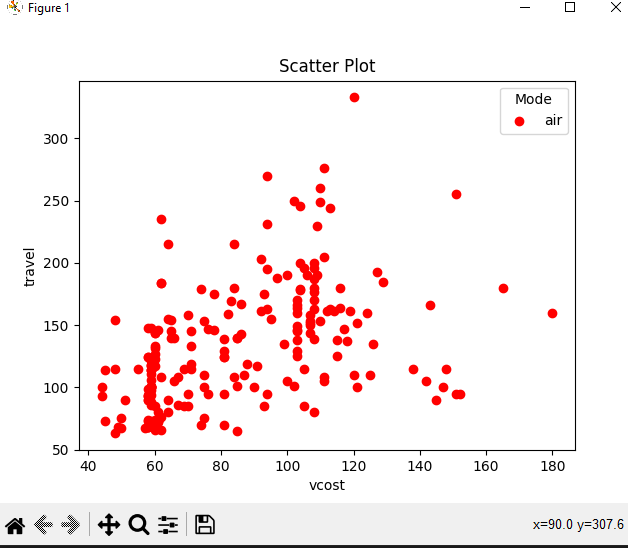
7 Pav. „Vcost" ir „gcost" scatter plot grafikas

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.



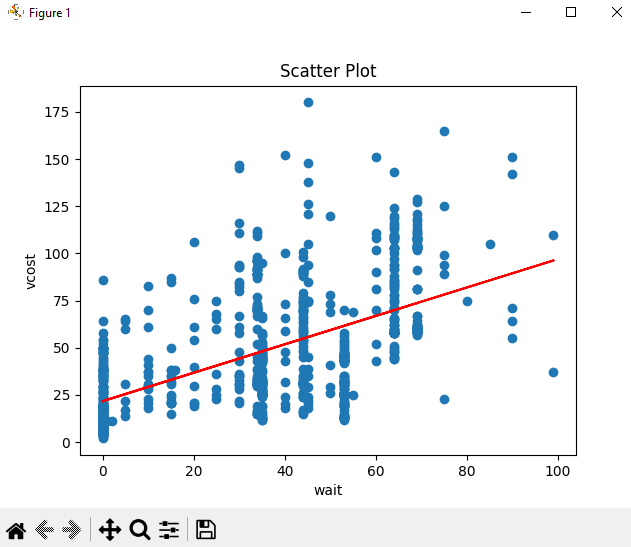
8 Pav. „Travel" ir „vcost" scatter plot grafikas visoms transporto priemonės atvaizduoti

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

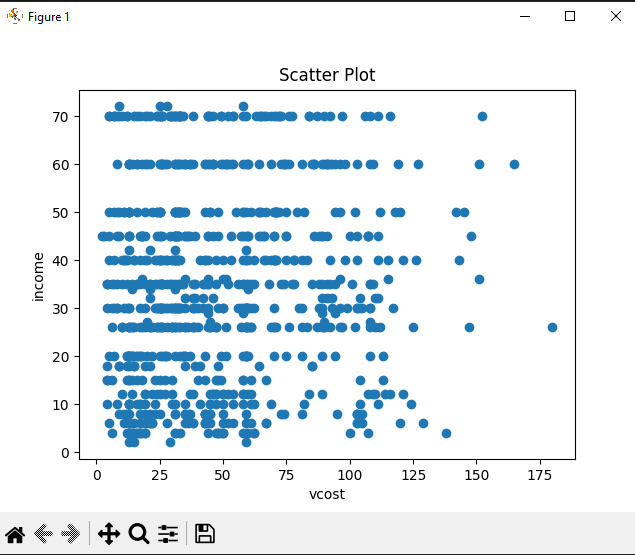
Jei atsižvelgtume tik į kelionę oru, būtų galima taip pat pastebėti tiesišką duomenų priklausomubę.



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

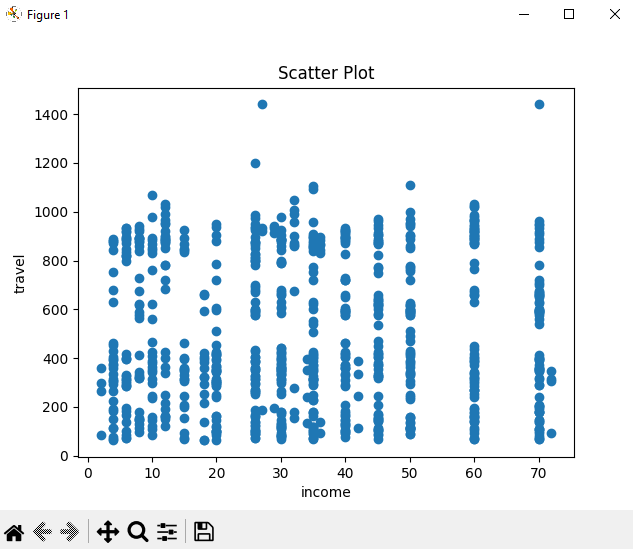
Nubrėžus linijinės progresijos tiesę per duomenis galima pastebėti gana stiprę priklausomybę tarp “vcost-wait” atributų – tai patvirtins vėliau apskaičiuoti kovariacijos ir koreliacijos koficientai. Nors duomenys nėra linkę stipriai grupuotis tarpusavyje, vis vien galima pastebęti jų sąryšį.

Tarpusavyje nekoreliuojantys atributai:



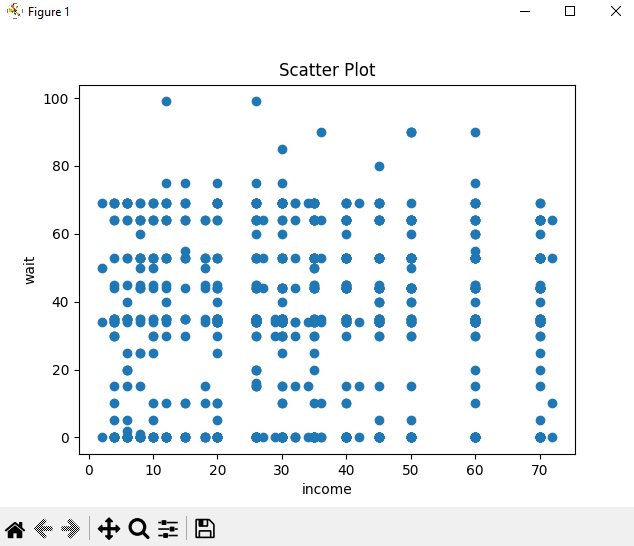
11 Pav. „Income" ir „vcost" scatter plot grafikas

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.

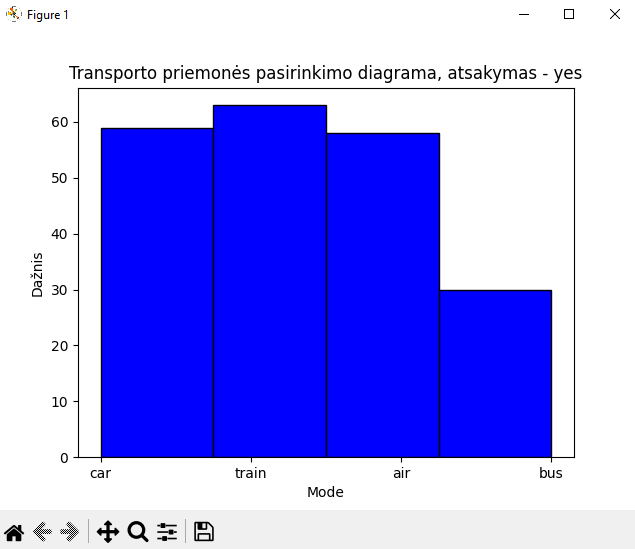
## SPLOM diagrama



14 Pav. SPLOM matrica

## Bar plot diagrama kategoriniams duomenims atvaizduoti

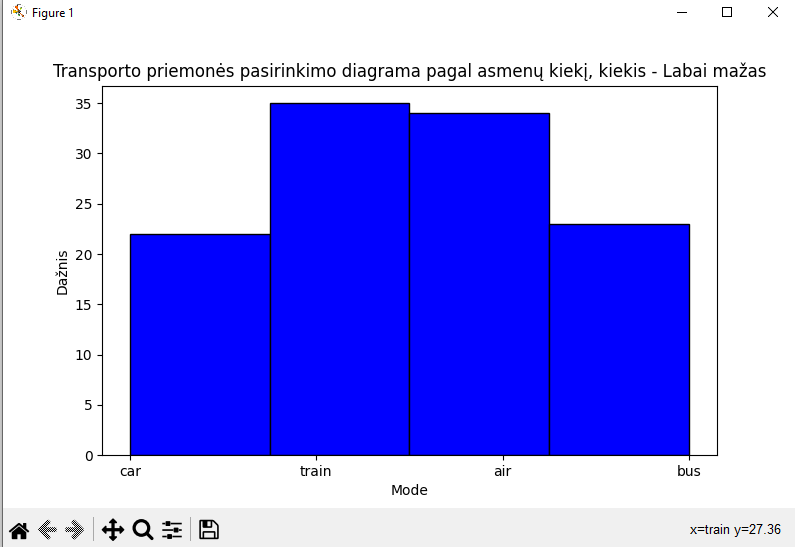
Iš pradžių nagrinėsiu kategorinių duomenų priklausomybę tarp transporto priemonės ir kiek žmonių ją rinkosi.



15 Pav. Žmonių sutikusių rinktis atitinkamas transporto priemones diagrama

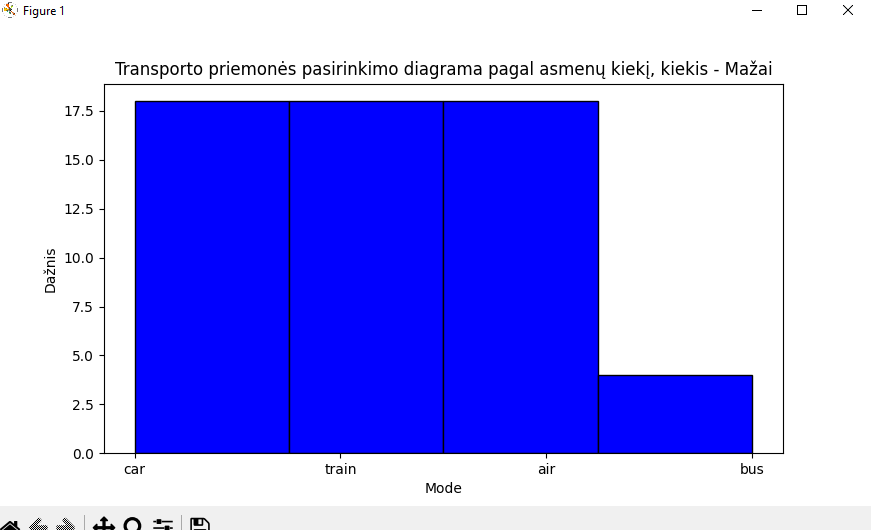
Iš gautų rezultatų diagramoje matosi, kad daugiausiai žmonių rinkosi keliauti traukiniu ir mažiausiai autobusu.

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



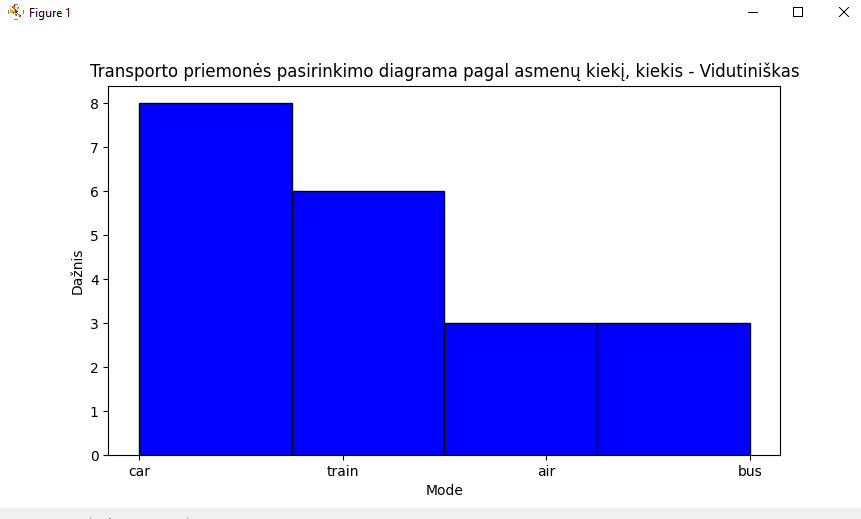
16 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra labai mažas

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.



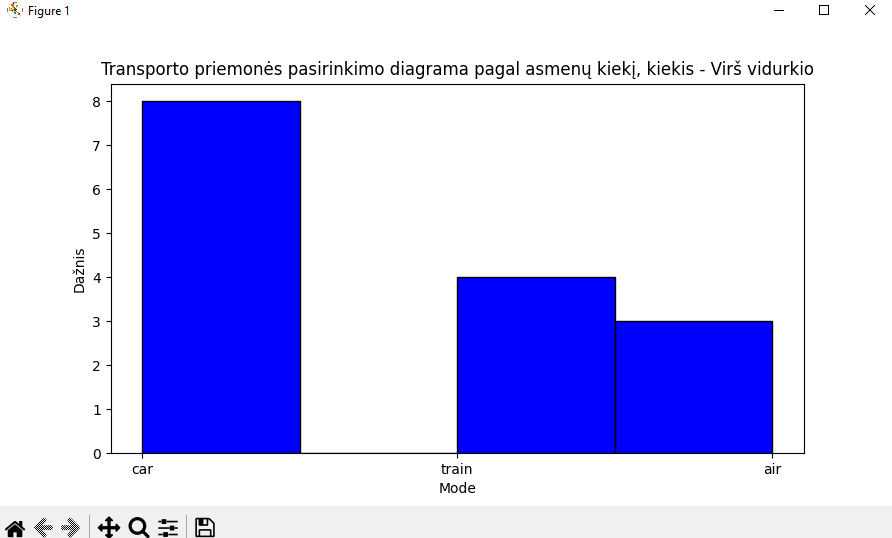
17 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra mažas

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

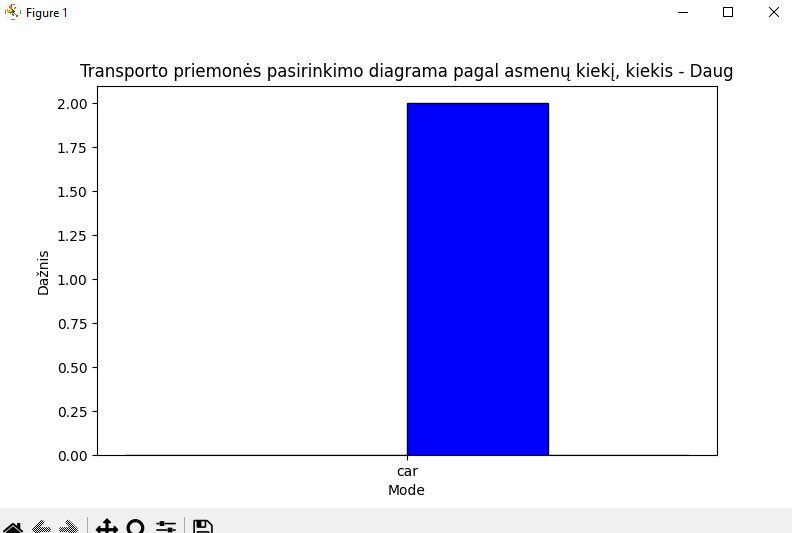


18 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra vidutiniškas

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



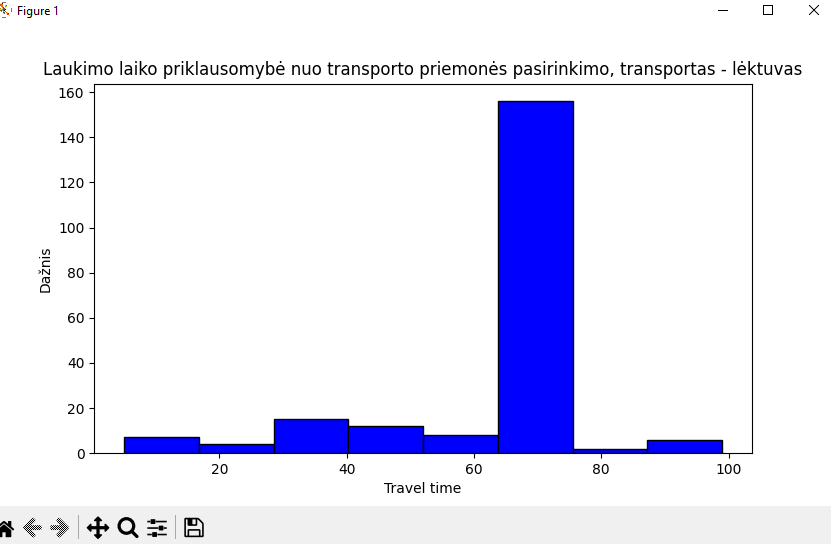
20 Pav. Transporto priemonės pasirinkimo diagama, kai asmenų kiekis yra didelis

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.

## Bar plot ir box plot diagramos atvaizduojančios kategorinio ir tolydinio tipo kintamųjų sąryšius

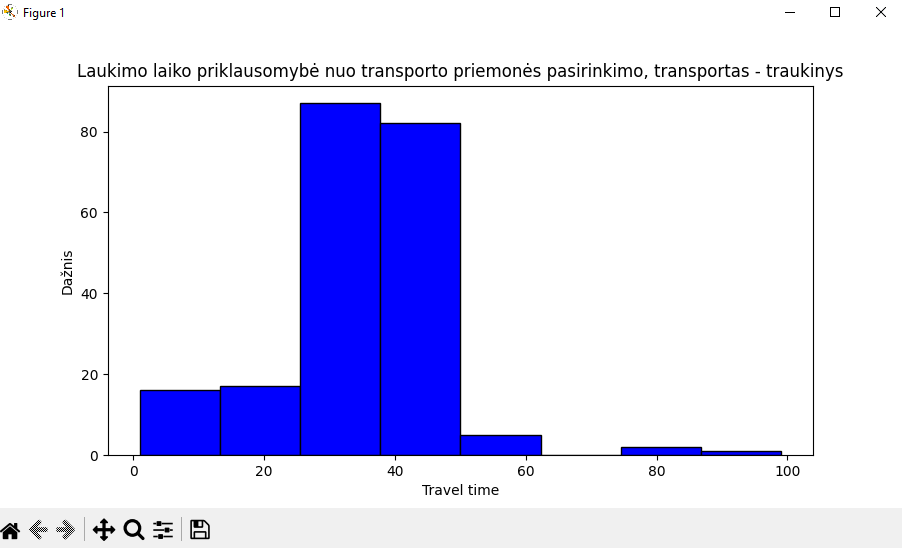
### Bar plot diagramos

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



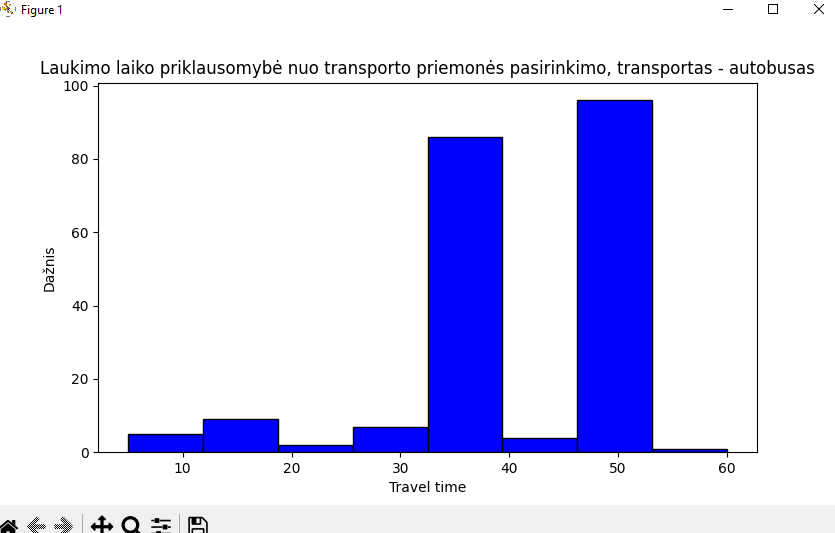
21 Pav. Laukimo laiko priklausomybės nuo transporto priemonės pasirinkimo diagrama, kai ji yra lėtuvas

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.

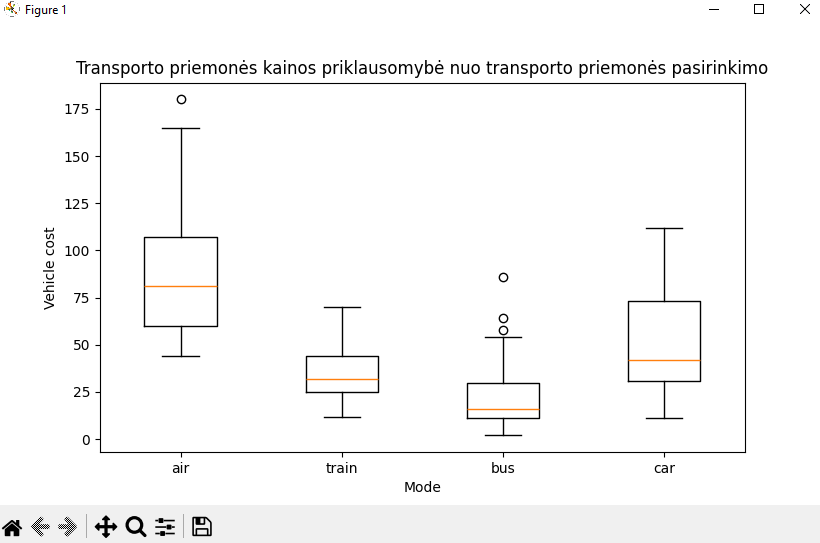


23 Pav. Laukimo laiko priklausomybės nuo transporto priemonės pasirinkimo diagrama, kai ji yra autobusas

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

### Box plot diagramos

Bus nagrinėjami transporto priemonės kainos duomenys.



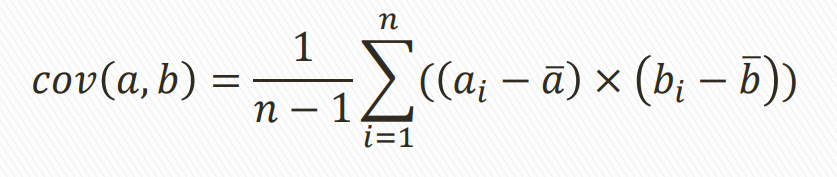
24 Pav. Box plot diagrama perteikianti transporto priemonės pasirinkimo ir kainos sąryšius

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.

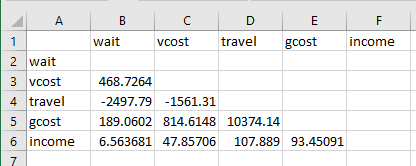
# Kovariacijos ir koreliacijos reikšmės

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

## Kovariacija



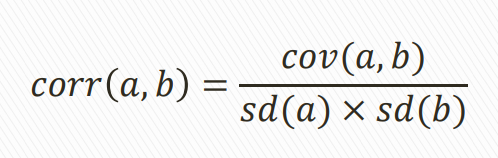
Naudojant šią formulę galima paskaičiuoti kovariacijas tarp atributų. Kuo didesnis gaunamas koeficientas tuo geresnis ryšys tarp atributų.

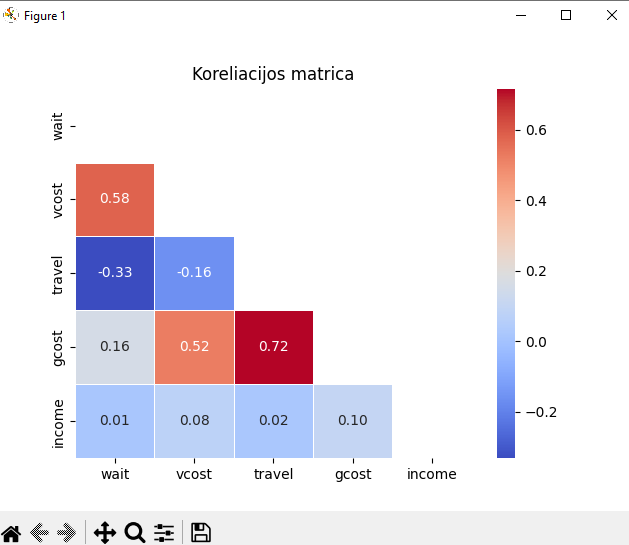


Pagal gautus duomenis kai kurie atributai tarpusavyje yra stipriai susiję, pvz: „travel-gcost“, „vcost-gcost“. Tuo tarpu egzistuoja reikšmių, kurios yra neigiamai susijusios, kai viena reikšmės auga, o kita krenta: „wait-travel“.

## Koreliacija

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



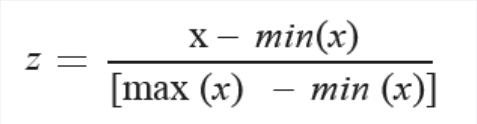


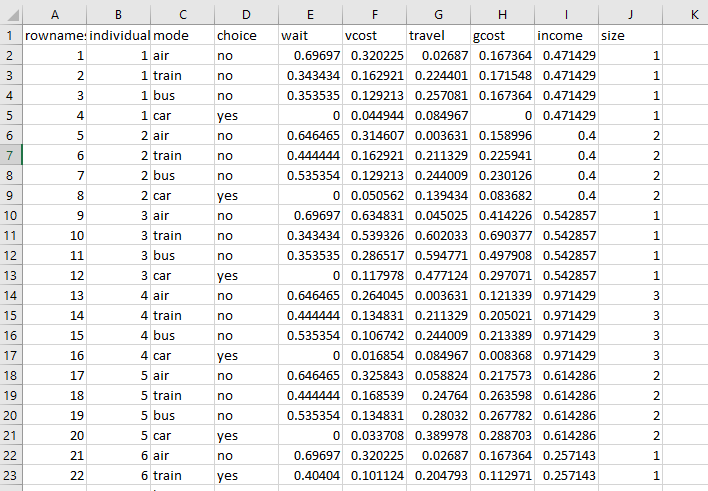
25 Pav. Tolydinių duomenų koreliacijos matrica

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.

# 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:





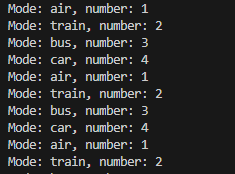
Normalizuoti 5 stulpeliai: wait, vcost, travel, gcost ir income.

# Vertimas tolydiniais duomenimis

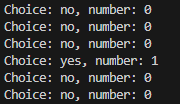
Bus kovertuojami mode ir choice stulpeliai į kategorinius duomenis.

Šių stulpelių eilutės keičiamos skaičiais:

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



26 Pav. „Mode" kategorinių duomenų vertimas tolydiniais



27 Pav. „Choice" kategorinių duomenų vertimas tolydiniais

# 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.