KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai (P176B101)

Laboratorinių darbų ataskaita

Atliko:
IFF-1/4 gr. studentas
Mildaras Karvelis
2024 m. kovo 21 d.
Priėmė:
lekt. Andrius Nečiūnas
lekt. Aušra Gadeikytė

KAUNAS 2023

TURINYS

1. Įvadas

Laboratorinio darbo tikslas yra surasti, apdoroti ir išanalizuoti tinkamą duomenų rinkinį. Darbo tikslo atlikimo eiga yra:

1. Pasirinkti tinkamą duomenų rinkinį;
2. Atlikti duomenų rinkinio kokybės analizę;
3. Nupaišyti ir aprašyti duomenų rinkinio atributų histogramas;
4. Nustatyti sąryšius tarp atributų;
5. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą;
6. Atlikti duomenų normalizaciją;
7. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.

2. Duomenų rinkinys

Duomenų rinkinį sudaro:

- „City“ – miesto pavadinimas;
- „Vehicle Type“ – transporto priemonės tipas;
- „Weather“ – oro sąlygos;
- „Economic Condition“ – ekonominė padėtis;
- „Day of Week“ – savaitės diena;
- „Hour of Day“ – valanda (1-24 h.);
- „Speed“ – greitis, km/h;
- „Is Peak Hour “ – ar tai piko valanda? (True arba false, 0 arba 1);
- „Random Event Occured“ – ar kažkas įvyko? (True arba false, 0 arba 1);
- „Energy Consumption“ – energijos suvartojimas;
- „Traffic Density“ – eismo tankumas.

3. Duomenų rinkinio kokybės analizė

3.1. Tolydinis tipas

Tolydiniam tipui kokybės analizei reikia apskaičiuoti:

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

Rezultatai matomi 1 lentelėje. Matome, jog nei vienas atributas neturi trūkstamų reikšmių, todėl šioje dalyje nieko taisyti nereikės. Taip pat, matome, kad kardinalumas ties „Speed“ ir „Energy Consumption“ yra vidutinis, tačiau prie „Traffic Density“ jis yra gan žemas.

Atributo pavadinimas	Kiekis	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
Speed	1219567	0.00%	670544	6.6934	163.0886	37.5331	80.5345	59.944	58.4711	26.632
Energy Consumption	1219567	0.00%	665672	4.9296	189.9489	29.27395	65.9055	49.464	45.7826	25.280
Traffic Density	1219567	0.00%	14209	0.0059	3.3776	0.1059	0.396	0.277	0.2186	0.219

1 lentelė. Tolydinio tipo atributų kokybės analizės lentelė

3.2. Kategorinis tipas

Kategoriniam tipui kokybės analizei 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

Rezultatai matomi 2 lentelėje. Galima pastebėti, kad nei vienas atributas neturi trūkstamų reikšmių, todėl modos puikiai susideda į 100%.

Atributo pavadinimas	Kiekis	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-osios Modos dažnumas, %
City	1219567	0.00%	6	Ecoopolis	204179	16.74%	AquaCity	203405	16.68%
Vehicle Type	1219567	0.00%	4	Autonomous Vehicle	757454	62.11%	Drone	304951	25.00%
Weather	1219567	0.00%	5	Solar Flare	244237	20.03%	Snowy	244195	20.02%
Economic Condition	1219567	0.00%	3	Booming	406684	33.35%	Recession	406571	33.34%
Day of Week	1219567	0.00%	7	Tuesday	174783	14.33%	Wednesday	174778	14.33%
Hour of Day	1219567	0.00%	24	11	51206	4.20%	15	51182	4.20%

Is Peak Hour	1219567	0.00%	2	0	1030901	84.53%	1	188666	15.47%
Random Event Occured	1219567	0.00%	2	0	1158726	95.01%	1	60841	4.99%

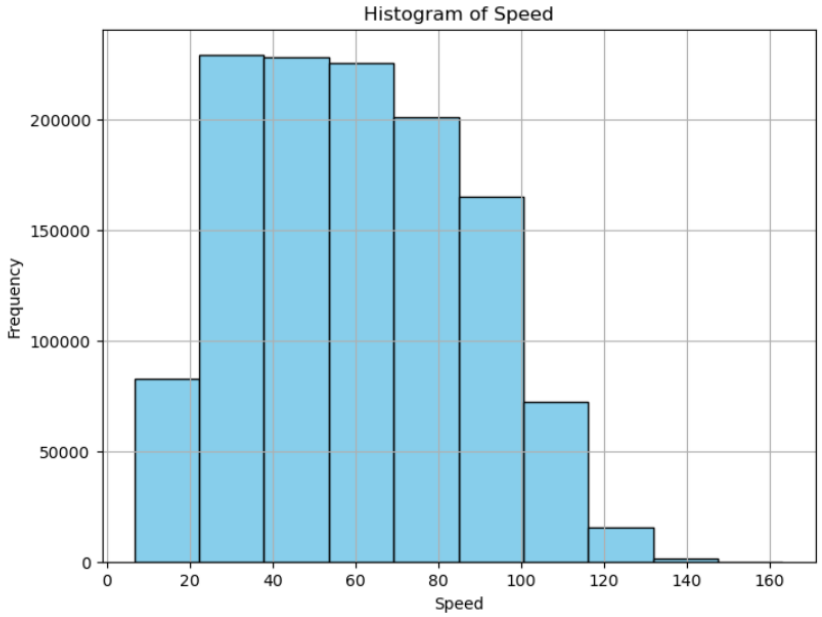
2 lentelė. Kategorinio tipo atributų kokybės analizės lentelė

4. Atributų grafikai

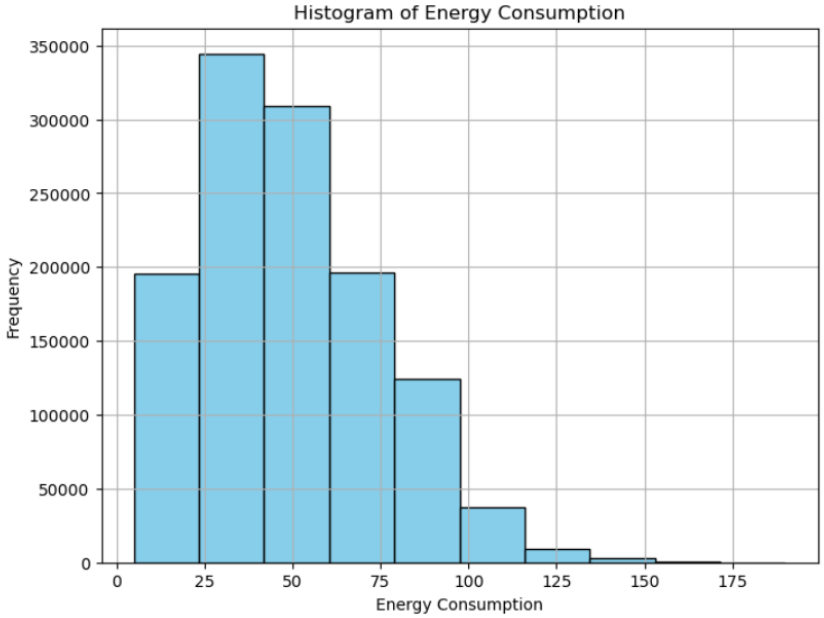
Grafikai buvo sukurti su programavimo kalba „Python“ naudojant biblioteką „Python Pandas“.

4.1. Tolydinio tipo histogramos

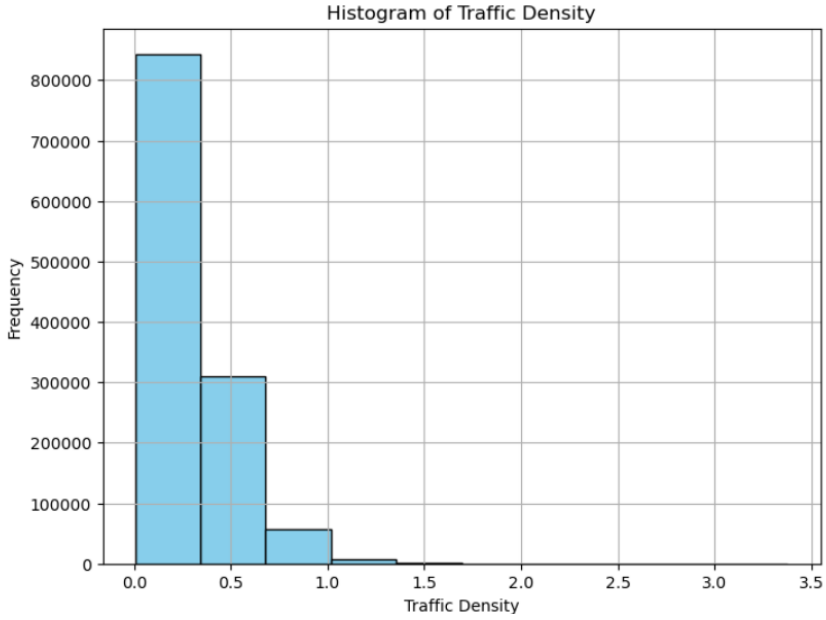
Matome, kad pirmoji (1 pav.) histograma yra nomaliąjame pasiskirstyme, o 2 ir 3 pav. yra ekponentiškame pasiskirstyme.



1 pav.



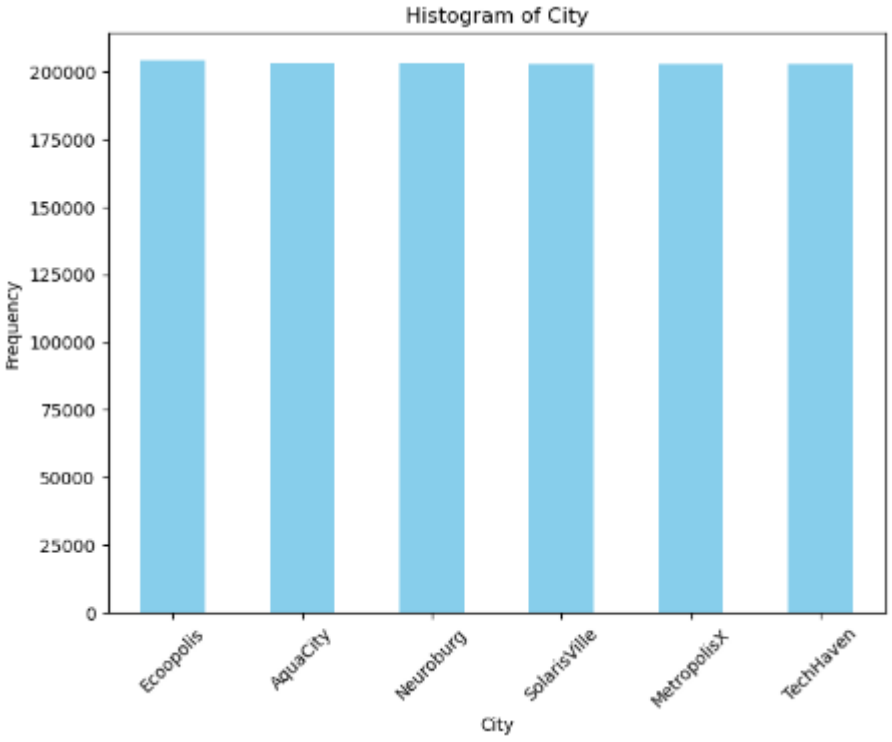
2 pav.



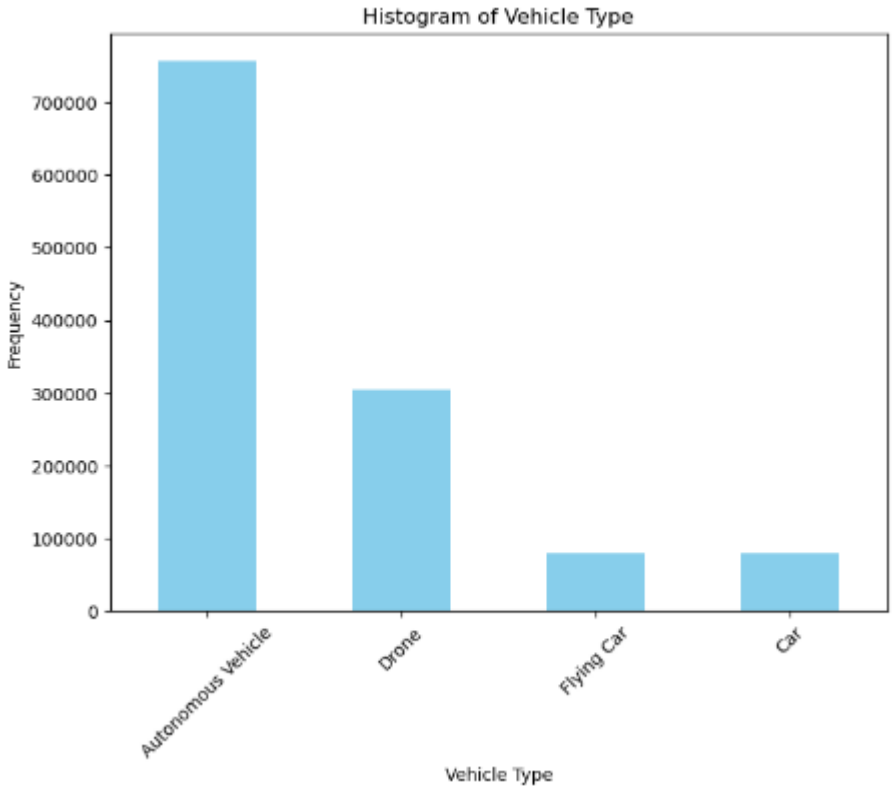
3 pav.

4.2. Kategorinio tipo stulpelinės diagramos

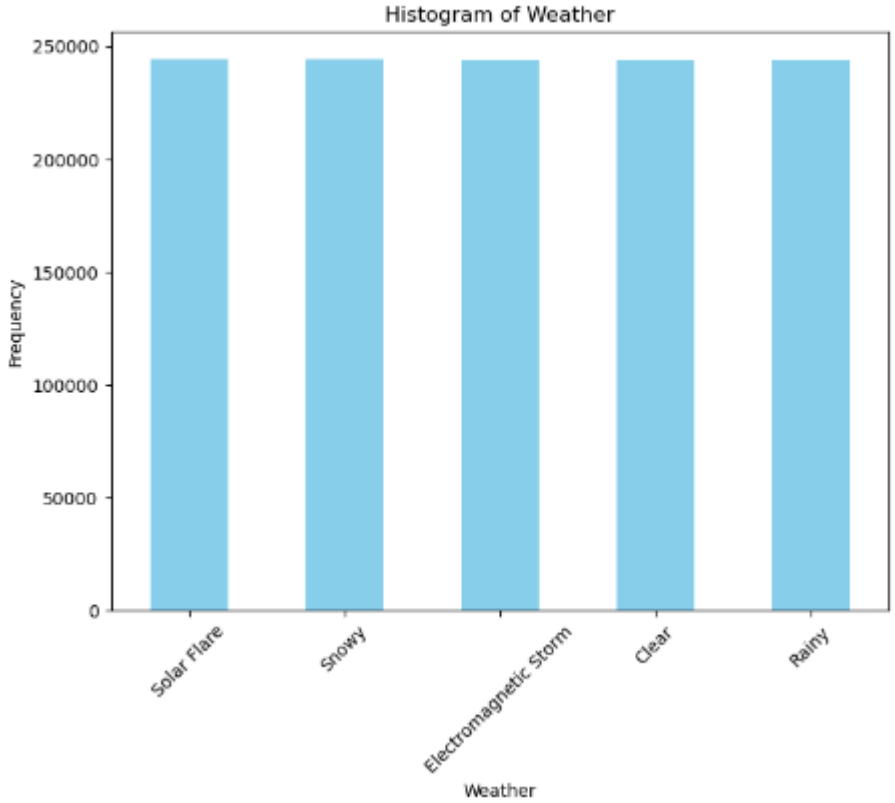
Matome, kad 4, 6, 7, 8 ir 9 pav. histogramos yra nomaliąjame pasiskirstyme, o 5, 10 ir 3 pav. yra ekponentiškame pasiskirstyme.



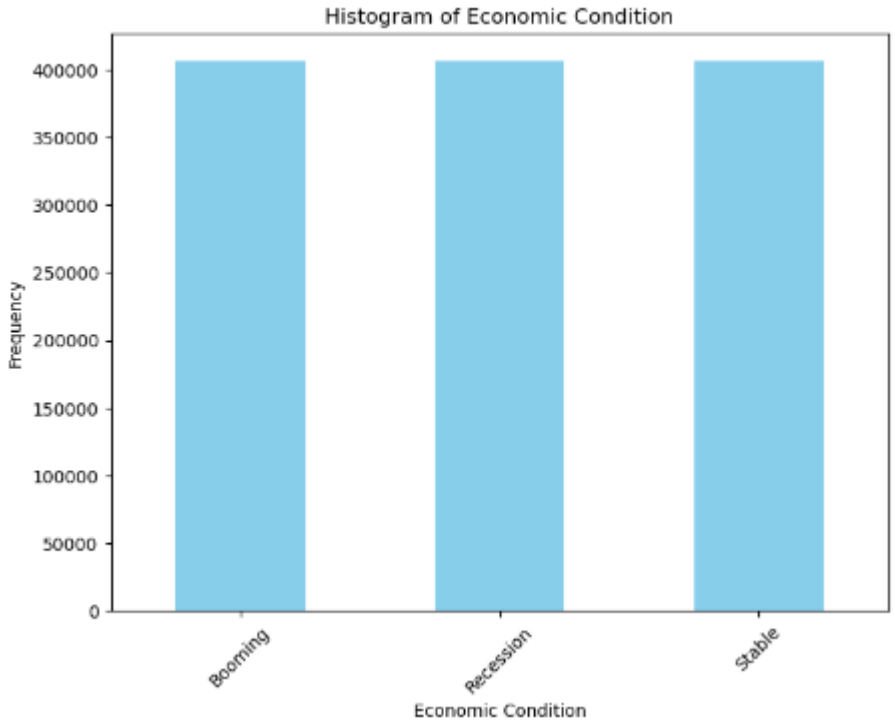
4 pav.


Histogram of Vehicle Type
Frequency
700000
600000
500000
400000
300000
200000
100000
0
Autonomous Vehicle
Drone
Flying Car
Car
Vehicle Type

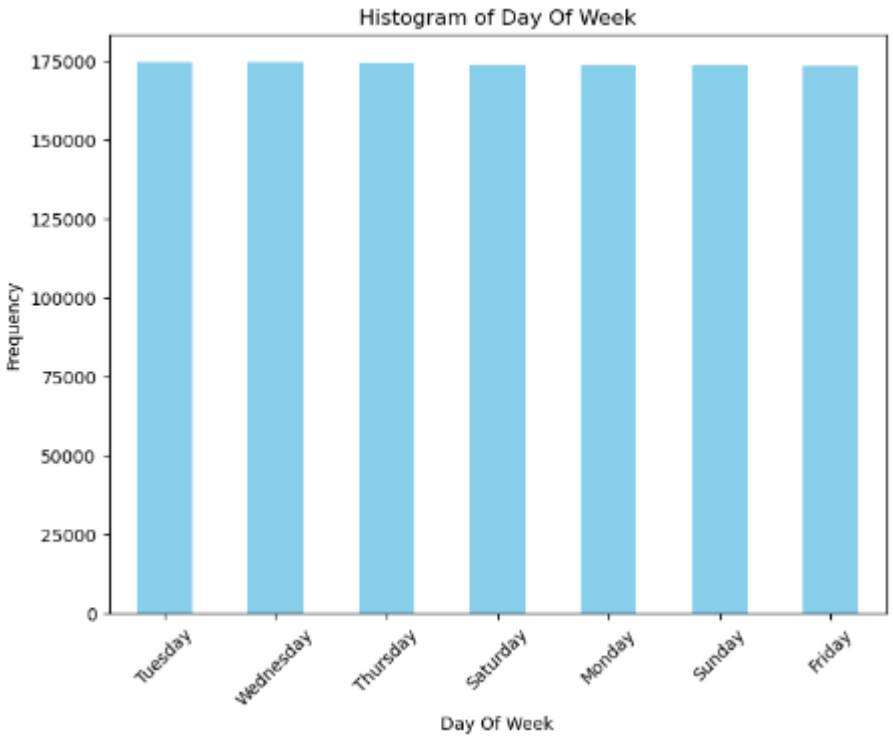
5 pav.


Histogram of Weather
Frequency
250000
200000
150000
100000
50000
0
Solar Flare
Snowy
Electromagnetic Storm
Clear
Rainy
Weather

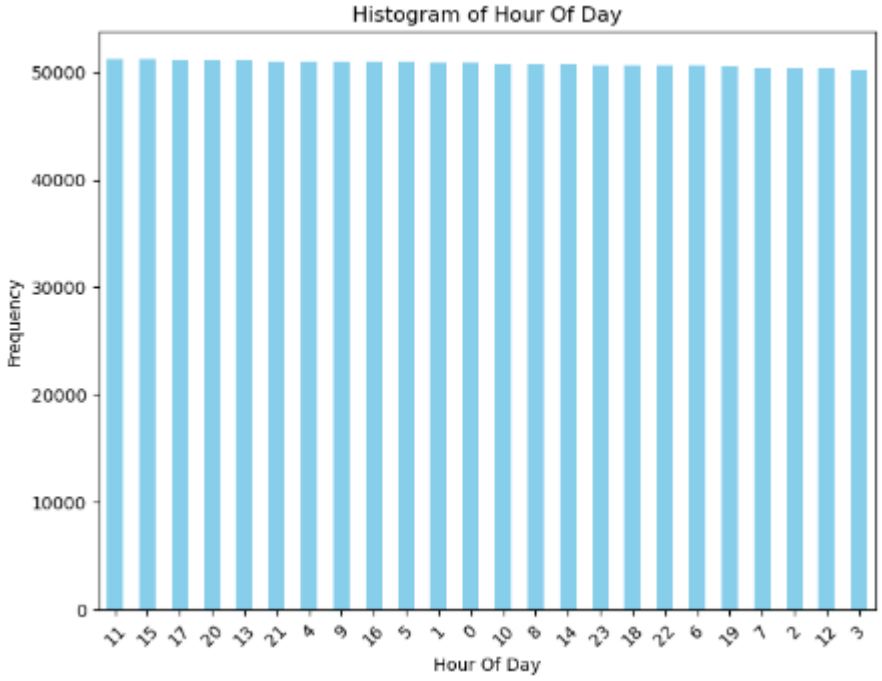
6 pav.


Histogram of Economic Condition
Frequency
400000
350000
300000
250000
200000
150000
100000
50000
0
Booming
Recession
Stable
Economic Condition

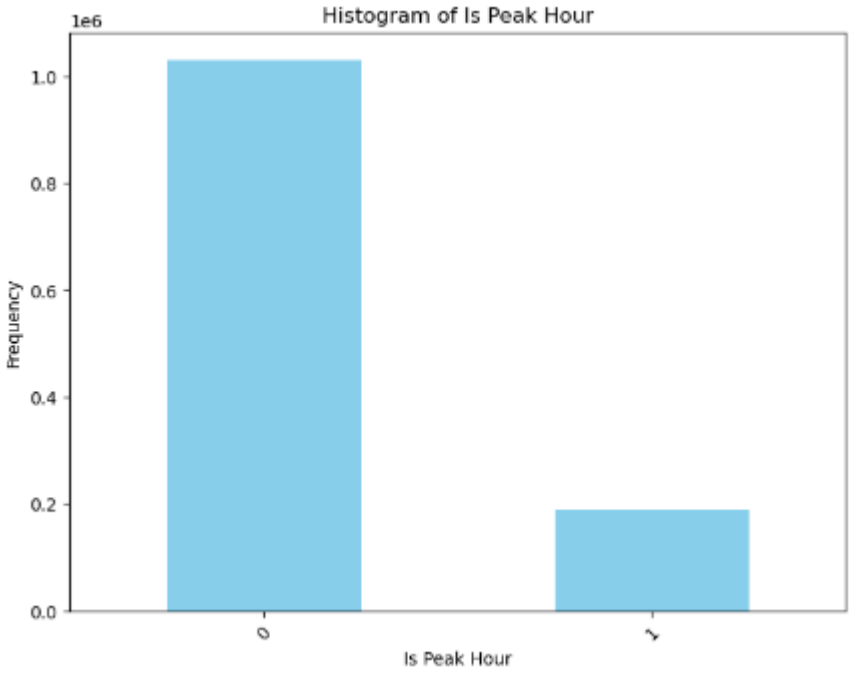
7 pav.


Histogram of Day Of Week
Frequency
175000
150000
125000
100000
75000
50000
25000
0
Tuesday
Wednesday
Thursday
Saturday
Monday
Sunday
Friday
Day Of Week

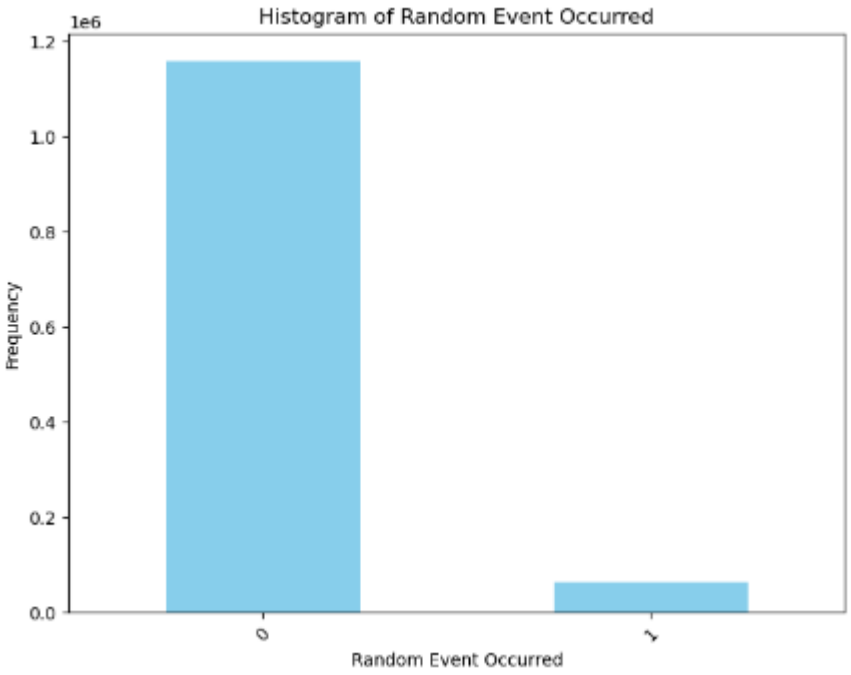
8 pav.


Histogram of Hour Of Day
Frequency
50000
40000
30000
20000
10000
0
11 15 17 20 13 21 4 9 16 5 1 0 10 8 14 23 18 22 6 19 7 2 12 3
Hour Of Day

9 pav.

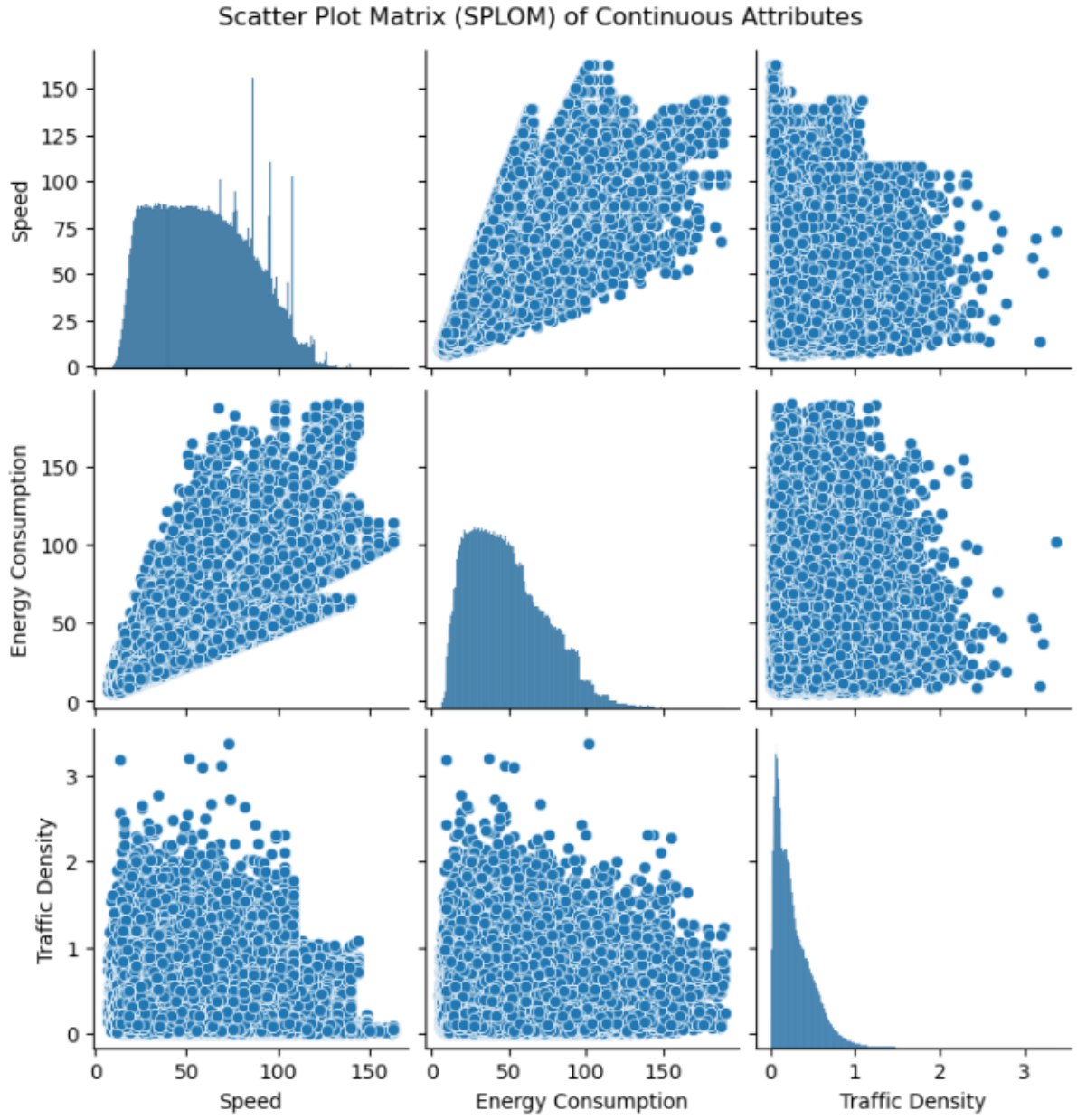

Histogram of Is Peak Hour
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
1
Is Peak Hour

10 pav.


Histogram of Random Event Occurred
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
1
Random Event Occurred

11 pav.

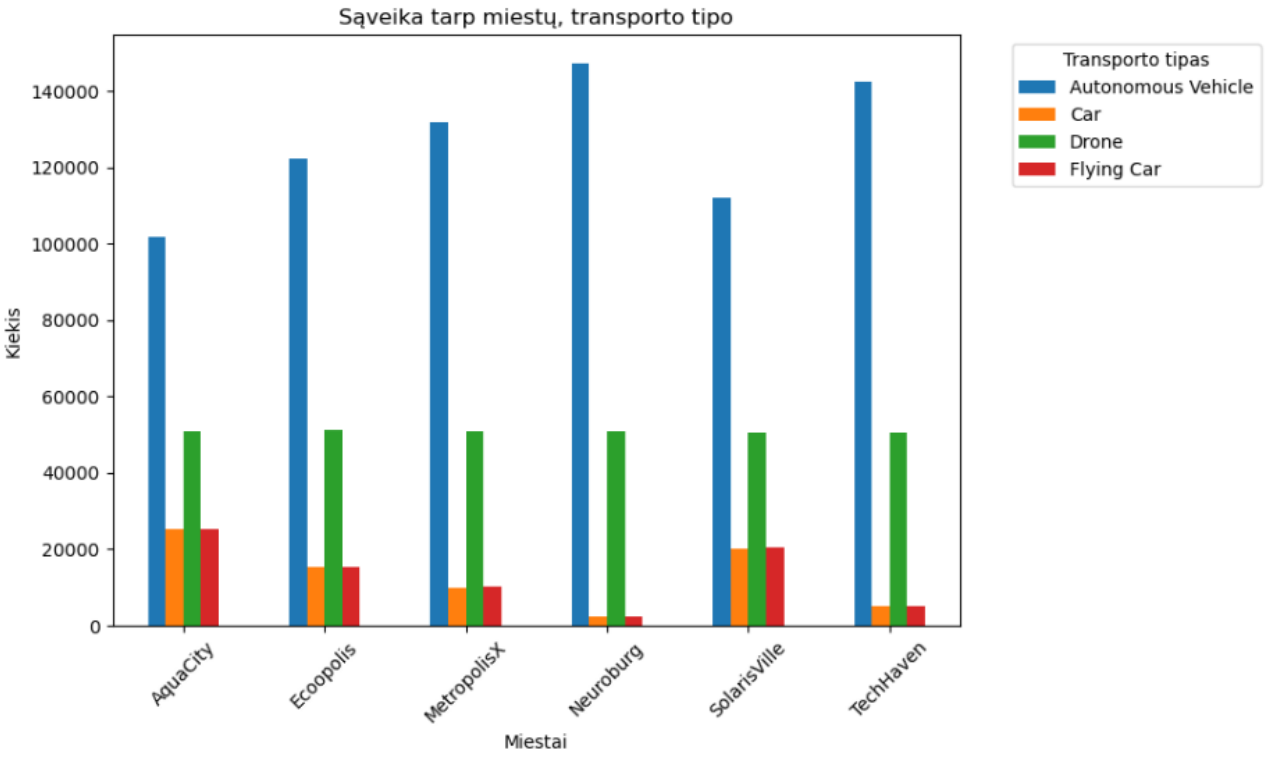
4.3. Scatter plot ir SPLOM diagrama tolydiems atributams



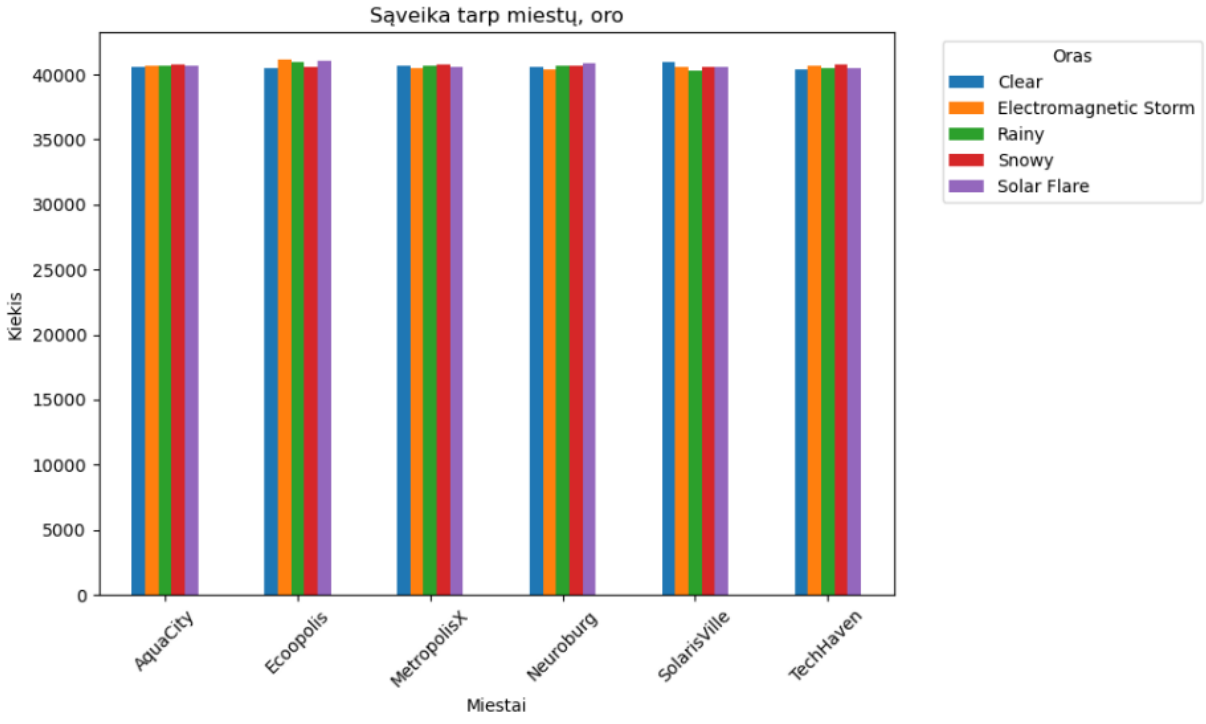
12 pav. Scatter plot ir SPLOM diagrama

4.4. Kategorinio tipo atributų priklausomybės bar-plot diagramos

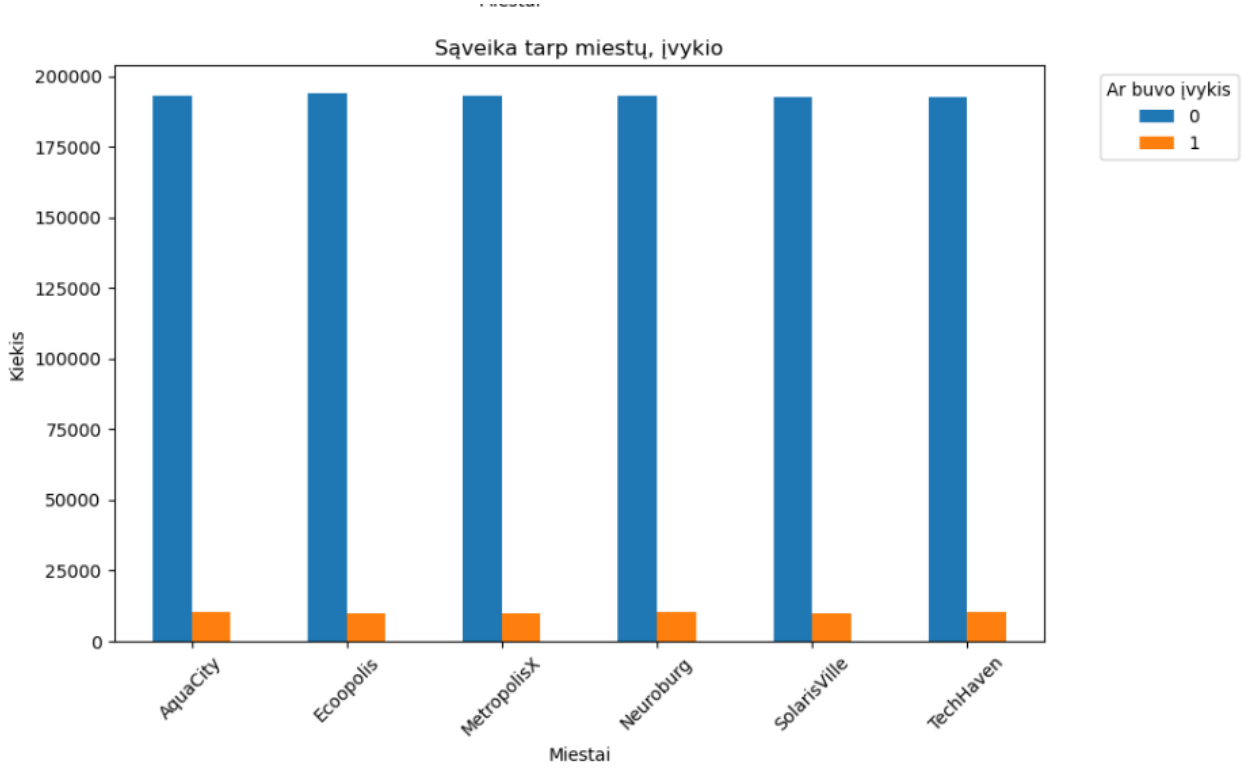
13 pav. matome, kad visuose miestuose dominuoja Autonomous Vehicle tipas, tačiau oro sąlygos ir įvykiai yra vienodai pasiskirstę kiekviename mieste.


Sąveika tarp miestų, transporto tipo
Kiekis
0
20000
40000
60000
80000
100000
120000
140000
AquaCity
Ecoopolis
MetropolisX
Neuroburg
SolarisVille
TechHaven
Miestai
Transporto tipas
Autonomous Vehicle
Car
Drone
Flying Car

13 pav.


Sąveika tarp miestų, oro
Kiekis
0
5000
10000
15000
20000
25000
30000
35000
40000
AquaCity
Ecoopolis
MetropolisX
Neuroburg
SolarisVille
TechHaven
Miestai
Oras
Clear
Electromagnetic Storm
Rainy
Snowy
Solar Flare

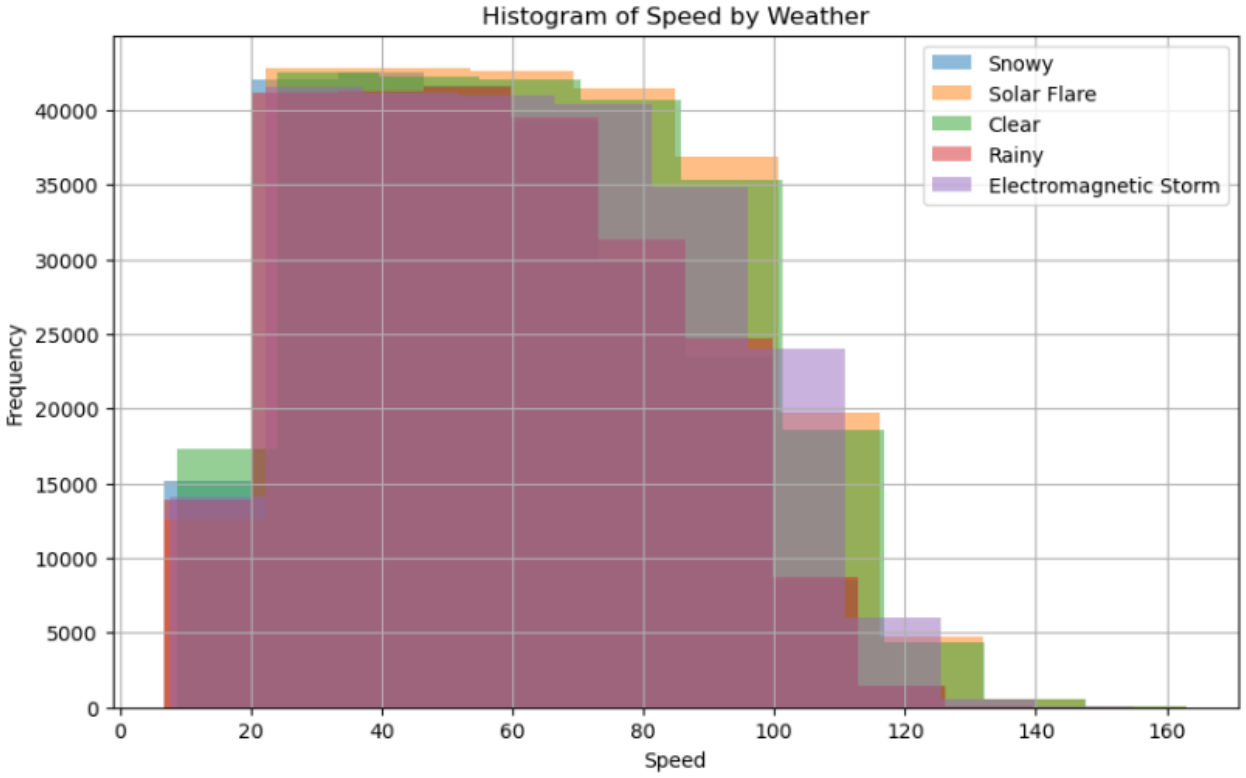
14 pav.



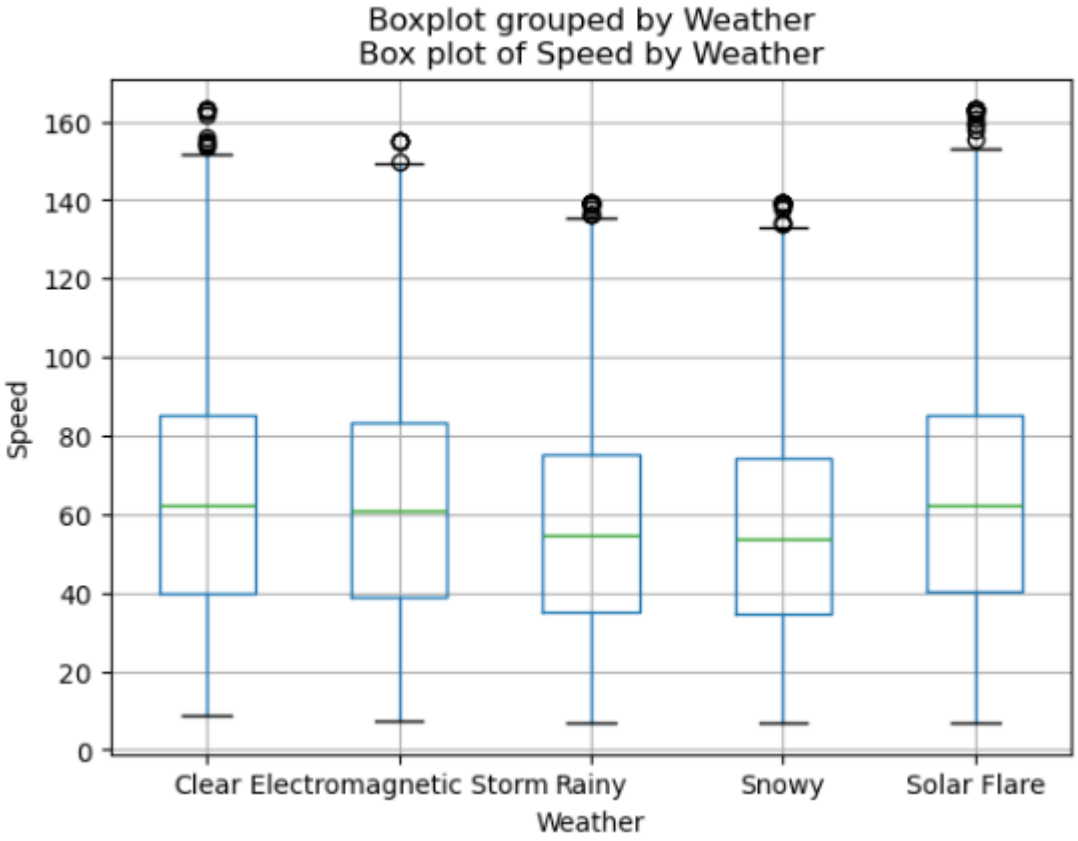
15 pav.

4.5. Kategorinių ir tolydžiųjų atributų priklausomybės histogramos ir box-plot diagramos

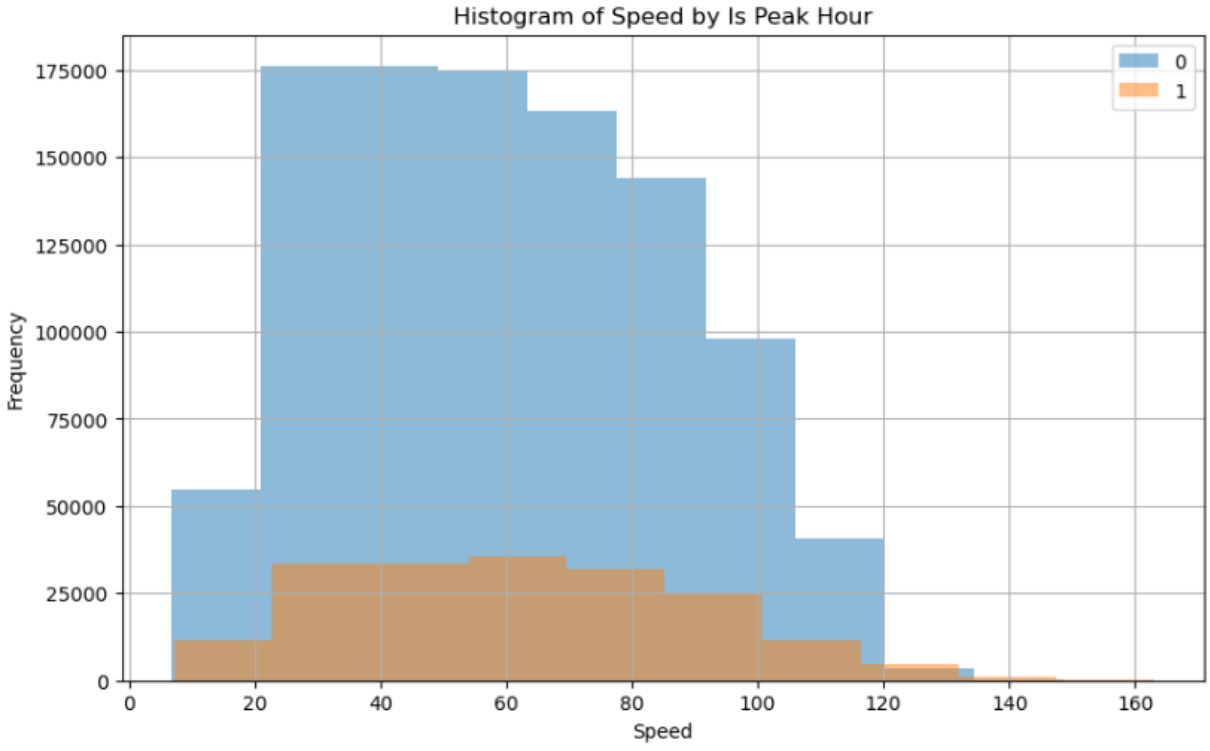
16 pav. matome, kad esant betkokioms oro sąlygoms, greitis labai mažai kinta, Clear ir Solar Flare oro sąlygomis greitis yra didžiausias. 18 pav. Matome panašius rezultatus.



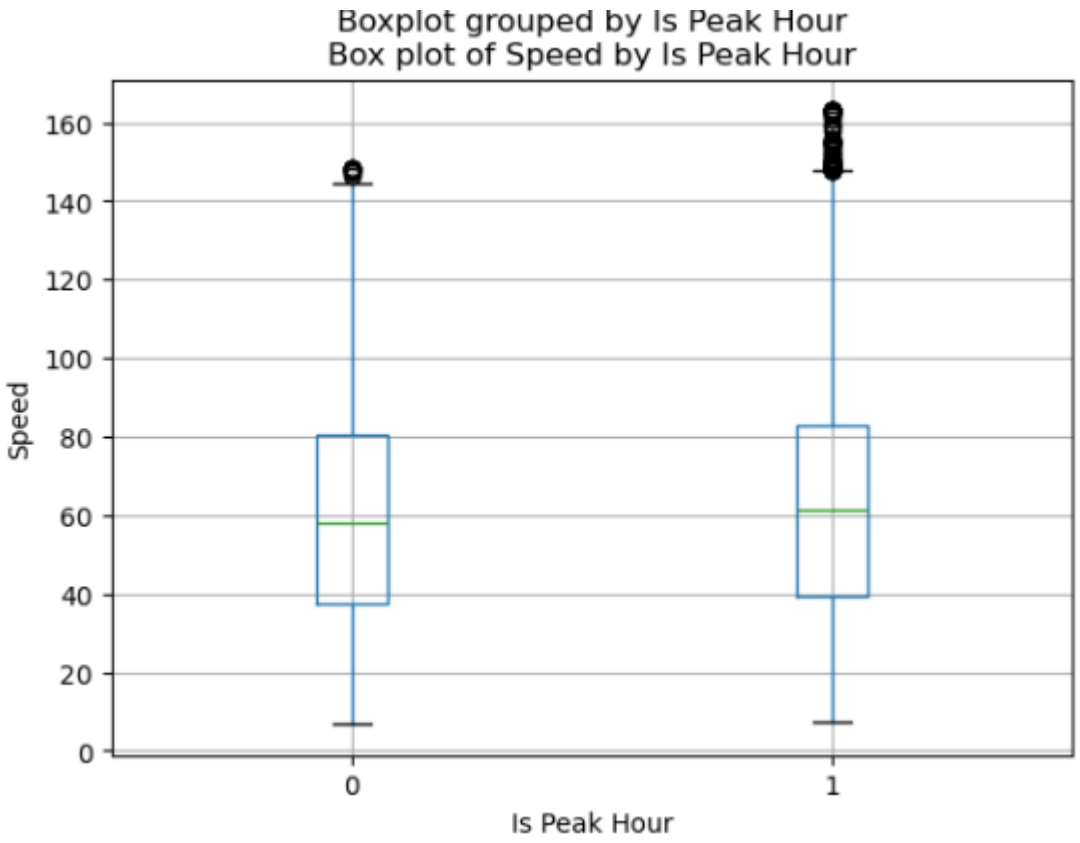
16 pav.


Boxplot grouped by Weather
Box plot of Speed by Weather
Speed
0
20
40
60
80
100
120
140
160
Clear
Electromagnetic Storm
Rainy
Snowy
Solar Flare
Weather

17 pav.


Histogram of Speed by Is Peak Hour
0
1
Frequency
0
25000
50000
75000
100000
125000
150000
175000
0
20
40
60
80
100
120
140
160
Speed

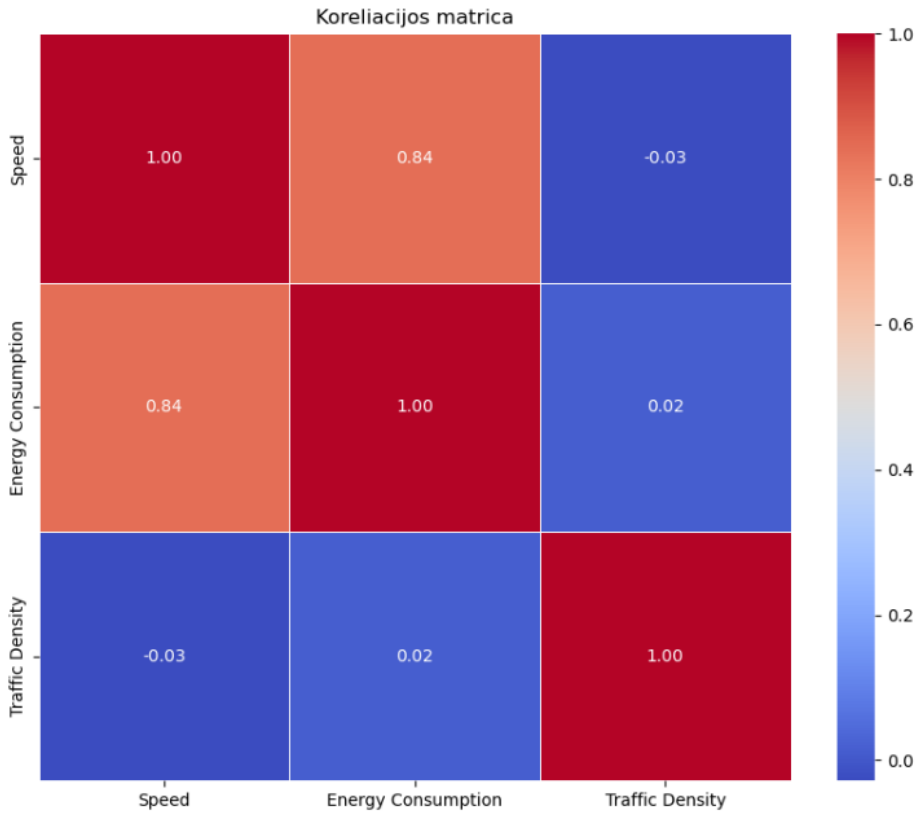
18 pav.



19 pav.

4.6. Korealiacija ir kovariacija

Korealiacijos intervalas yra (-1, 1)
Su koreliacijos reikšmėmis įmanoma lengvai pastebėti, kurie atributai turi sąryšį tarpusavyje. Minimalios koreliacijos reikšmės yra ~-0,10 ir ~0,90. Galima pastebėti, kad nėra koreliacijų, kurios atitiktų minimumą.



20 pav.

Pagal gautas kovariacijos reikšmes įmanoma atspėti, kurios atributų poros ryšys yra stipresnis už kitus. Pavyzdžiui, įmanoma teigti, kad „Energy Consumpion–Speed“ ryšys yra silpnesnis už „Energy Consumption – Traffic Density“ ryšį, kadangi pirminio kovariacija yra mažesnė. Taip pat galima teigti, kad ryšys tarp „Traffic Density“ su kitais atribtais yra labai prastas, kadangi dauguma kovariacijos reikšmių yra arti 0.

Kovariacijos matrica:

	Speed	Energy Consumption	Traffic Density
Speed	709.264923	565.621944	-0.164616
Energy Consumption	565.621944	639.085475	0.087184
Traffic Density	-0.164616	0.087184	0.048006

21 pav.

5. Duomenų normalizacija

Dažnai pasitaiko didelių reikšmių duomenys, kurių analizę ir supratimą gali palengvinti duomenų normalizacija. Savo duomenų rinkinio normalizavimui buvo naudojama formulė:
z = x – min(X)/max(X) – min(X)

Čia X – duomenų aibė, x – iš duomenų aibės X išrinkta reikšmė, z – normalizuota x reikšmė. Panaudojus formulę su kiekviena duomenų reikšme buvo sukurtas duomenų rinkinys, kurios reikšmės yra intervale [0;1]. 22 pav. matoma dalis normalizuoto duomenų rinkinio.

	Hour Of Day	Speed	Is Peak Hour	Random Event Occurred	Energy Consumption	Traffic Density
0	20	0.145359	0	0	0.052880	0.153691
1	2	0.716816	0	0	0.749320	0.093395
2	16	0.599104	0	0	0.466624	0.010558
3	8	0.448266	1	0	0.222386	0.051962
4	16	0.246326	0	0	0.190595	0.133019

22 pav.

6. Išvados

- Analizavus duomenų rinkinio kokybę buvo pastebėta, kad nei vienas atributas neturėjo trūkstamų duomenų, todėl jų keisti nereikėjo.
- Išanalizavę tolydžių ir kategorinių atributų priklausomybės histogramas galime teigti, kad atributai šiek tiek priklauso vienas nuo kito.