Sveikatos draudimo kainos ir jai darančių įtaką kintamųjų analizė

Out[1]: Tam, kad analizę skaityti būtų paprasčiau, kodas yra paslėptas

Ši analizė skirta nustatyti kriterijus, veikiančius sveikatos draudimo kainą. Naudojami duomenys yra paimti iš autoriaus Brett Lantz knygos "Machine Learning with R". Duomenys yra paremti JAV cenzu ir galima teigti, kad atitinka realią situaciją šalyje. TUrimi duomenys yra korespondentų amžius, lytis, kūno masės indeksas (KMI, angliškai - BMI), turimų vaikų skaičius, rūkymo istorija, regionas ir bendra kaina mokama už draudimą.

Analizės pradžioje, peržiūrėjau bendrą informaciją apie duomenis bei koreliacijas tarp kintamųjų.

Lentelė 1

Skaitinių reikšmių vidurkiai, standartiniai nuokrypiai, didžiausios ir mažiausios reikšmės, kvartiliai ir medianos.

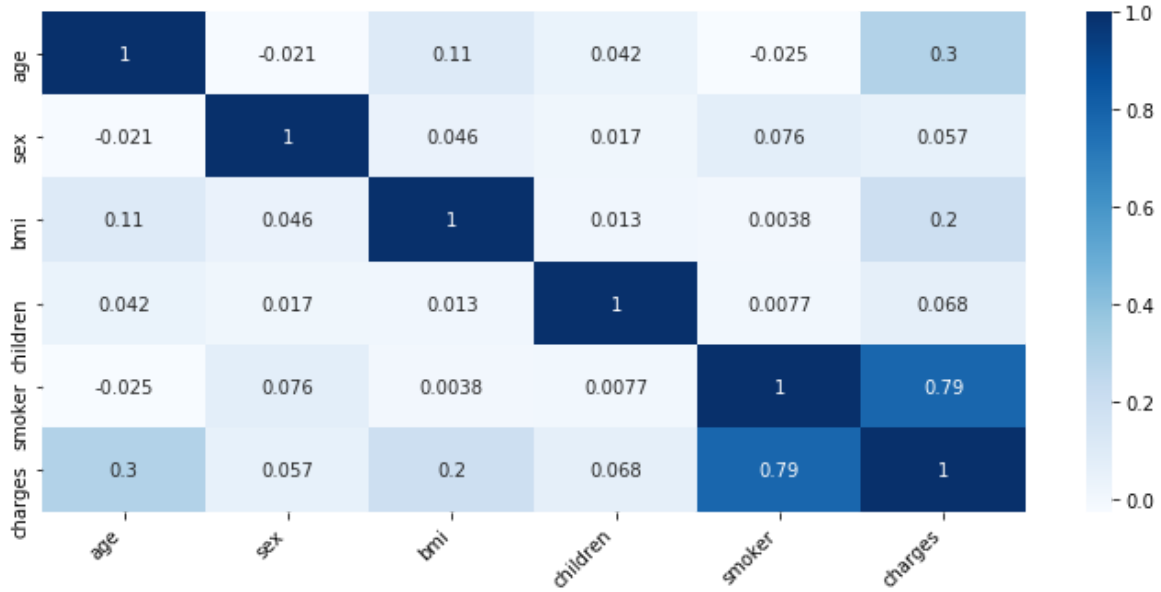
Out[5]:

	age	sex	bmi	children	smoker	charges
mean	39.207025	0.505232	30.663397	1.094918	0.204783	13270.422265
std	14.049960	0.500160	6.098187	1.205493	0.403694	12110.011237
min	18.000000	0.000000	15.960000	0.000000	0.000000	1121.873900
25%	27.000000	0.000000	26.296250	0.000000	0.000000	4740.287150
50%	39.000000	1.000000	30.400000	1.000000	0.000000	9382.033000
75%	51.000000	1.000000	34.693750	2.000000	0.000000	16639.912515
max	64.000000	1.000000	53.130000	5.000000	1.000000	63770.428010

Lentelė 2

Koreliacijos tarp kintamųjų. Kuo skaičius arčiau 1 tuo artimesnė koreliacija. Neigiami skaičiai reiškia atvirkštinę koreliaciją (kai vienas kintamasis didėja, kitas mažėja)

Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x5aab850>



Kadangi analizės tikslas yra nustatyti koreliacijas tarp draudimo kainos ir kitų kriterijų, kitos koreliacijos nebus analizuojamos. Matome, jog koreliacijos tarp draudimo kainos ir vaikų skaičiaus ar lyties yra statistiškai nereikšmingos. Todėl galima teigti, kad lytis ir vaikų kiekis neturi beveik jokios įtakos draudimo kainai.

Lentelė 3

Kainos koreliacijos su kitais kintamaisiais.

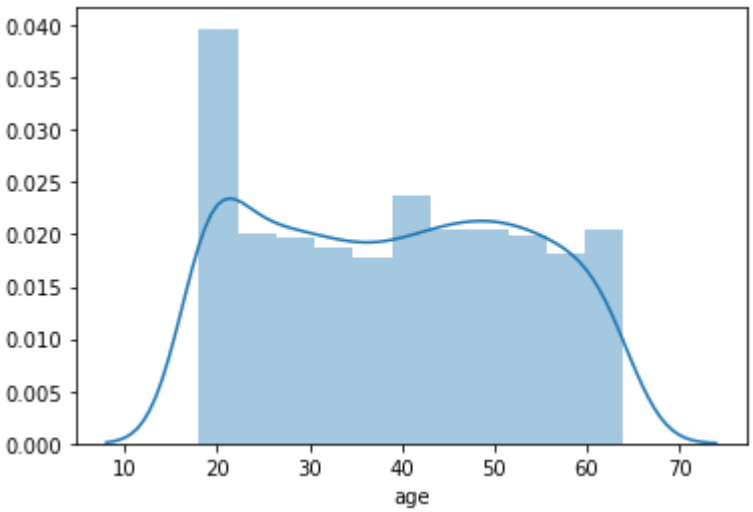
Out[7]:
```
smoker      0.787251
age         0.299008
bmi         0.198341
children    0.067998
sex         0.057292
Name: charges, dtype: float64
```

Peržiūrint duomenų pasiskirstymus, galima padaryti kelis įdomius pastebėjimus. Pirma, mūsų imtyje yra neproporcingai daug ~20 metų amžiaus žmonių (Lentelė 4). Taip pat galima pastebėti, kad didžiosios dalies žmonių kūno masės indeksas (BMI) viršija 30, kas žymi nutukimą (Lentelė 5).

Lentelė 4

Korespondentų pasiskirstymas pagal amžių.

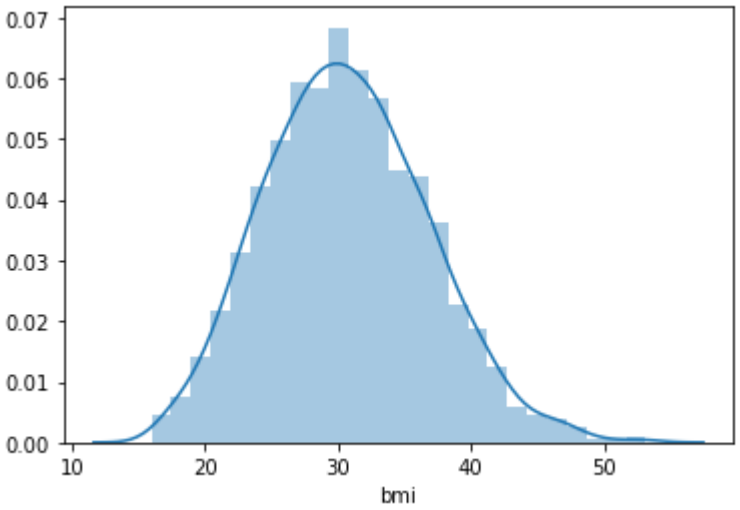
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x13631e38>



Lentelė 5

Korespondentų pasiskirstymas pagal kūno masės indeksą.

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x134dad18>

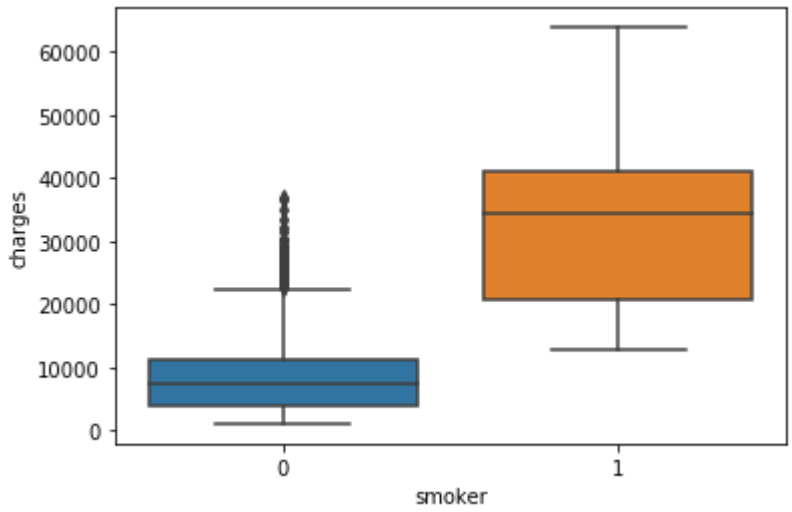


Kaip jau prieš tai matėme lentelėje 3, draudimo kaina turi didžiausią koreliaciją su rūkymo istorija. Sugrupavus korespondentus į rūkančius ir nerūkančius, galima matyti, jog rūkantieji už draudimą vidutiniškai moka 20000 - 30000 daugiau.

Lentelė 6

Draudimo kainos pasiskirstymas pagal rūkymo istoriją. 0 žymi nerūkančiuosius, 1 žymi rūkančiuosius.

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x13549958>

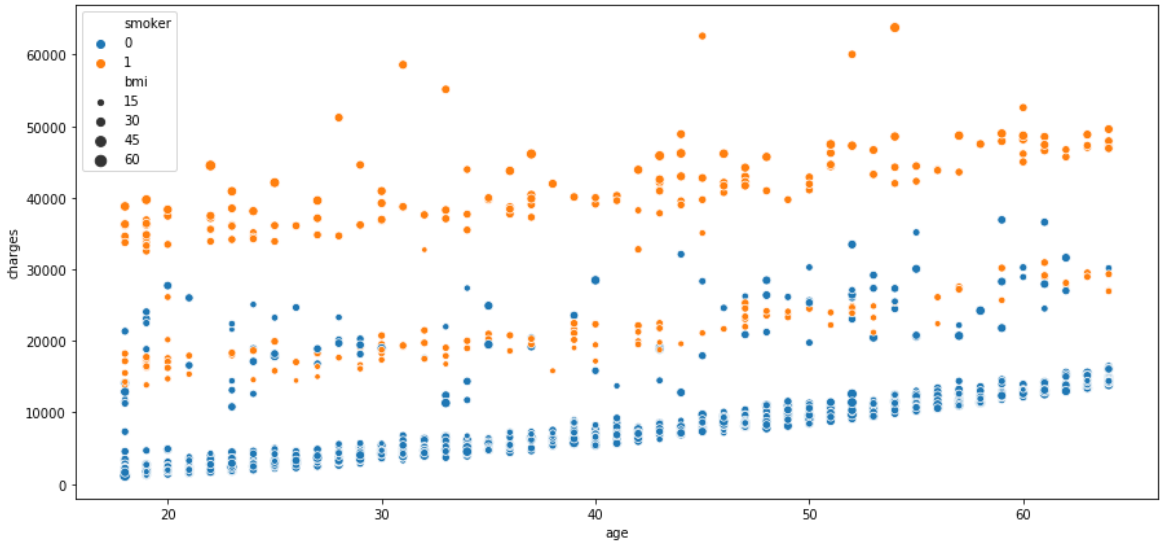


Sekančioje lentelėje atvaizduoti visi keturi pagrindiniai rodikliai: amžius (horizontali ašis), rūkymo statusas (apskritimo spalva), KMI (apskritimo dydis) ir draudimo kaina (vertikali ašis). Galima pastebėti, kad amžius ir kaina sudaro ryškią linijinę progresiją. Taip pat, kaip ir lentelėje 6, matomas ryškus kainos pasiskirstymas pagal rūkymo istoriją. Tačiau lentelėje 7 rūkantieji pasiskirstė į dar dvi grupes.

Lentelė 7

Rodiklių pasiskirstymas.

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x13595850>

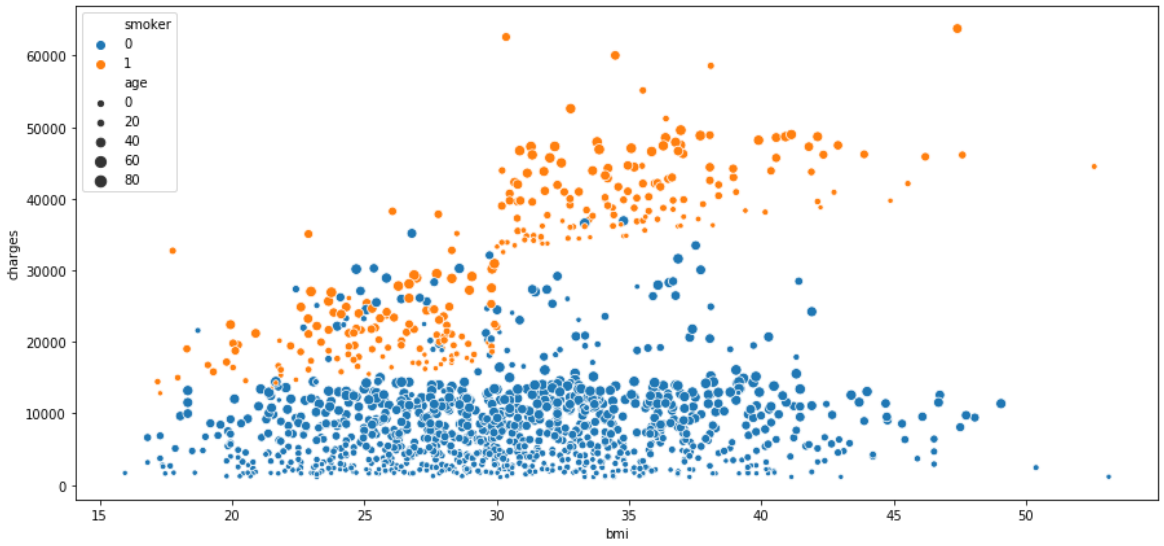


Rūkančiųjų pasiskirstymą galima paaiškinti KMI perkėlus į horizontalią ašį, o amžių atvaizdavus apskritimo dydžiu. Kainos lūžis ivyksta, kai KMI perkopia 30. Kaip minėta anksčiau, KMI virš 30 žymi nutukimą. Verta pastebėti, kad KMI neturi ryškios įtakos nerūkančiųjų draudimo kainai.

Lentelė 8

Visų korespondentų sveikatos draudimų kainos pagal KMI:

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x135e6f58>



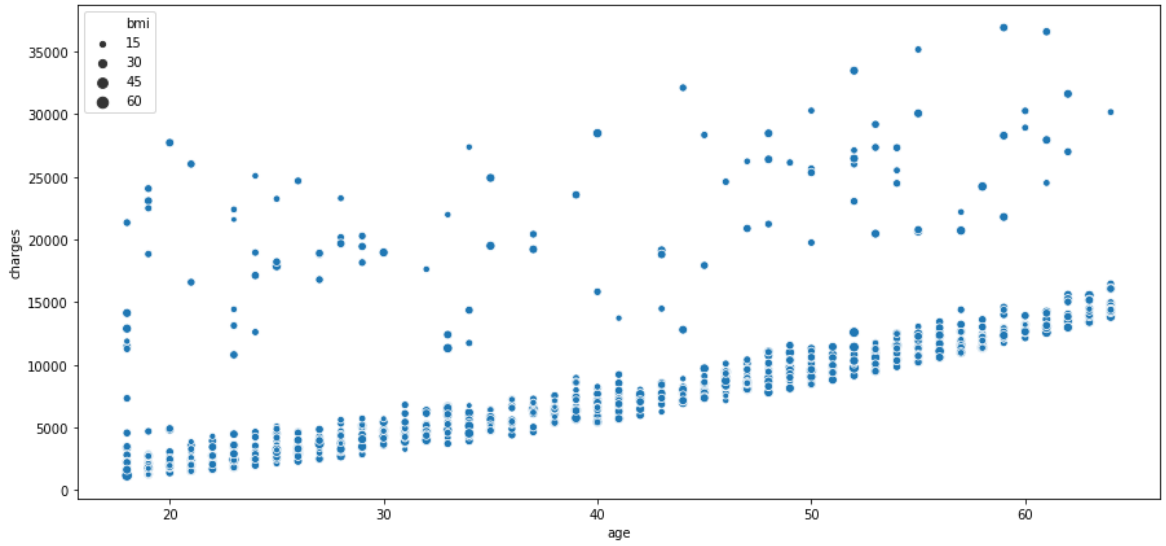
Toliau galima pastebėti, kad išskirsčius koresppondentus į nerūkančius (lentelė 9), rūkančius be nutukimo (lentelė 10) ir rūkančius su nutukimu (lentelė 11), mes vėl matome tuos pačius pasiskirstymus tarp amžiaus ir draudimo kainos - linijinę progresiją.

Taip pat matome, kad yra gana nemažas kiekis korespondentų, kurių draudimo kaina yra ryškiai aukštesnė, nei jų kategorija turėtų diktuoti. Kadangi turimuose duomenyse nepavyko rasti daugiau jokių korealiacijų, galima kelti hipotezę, jog šie kainos neatitikimai susiję su korespondentų ligų istorijomis. Tolimesnes įžvalgas būtų galima daryti gavus prieeigą prie šios informacijos.

Lentelė 9

Nerūkančių žmonių sveikatos draudimų kainos pagal amžių:

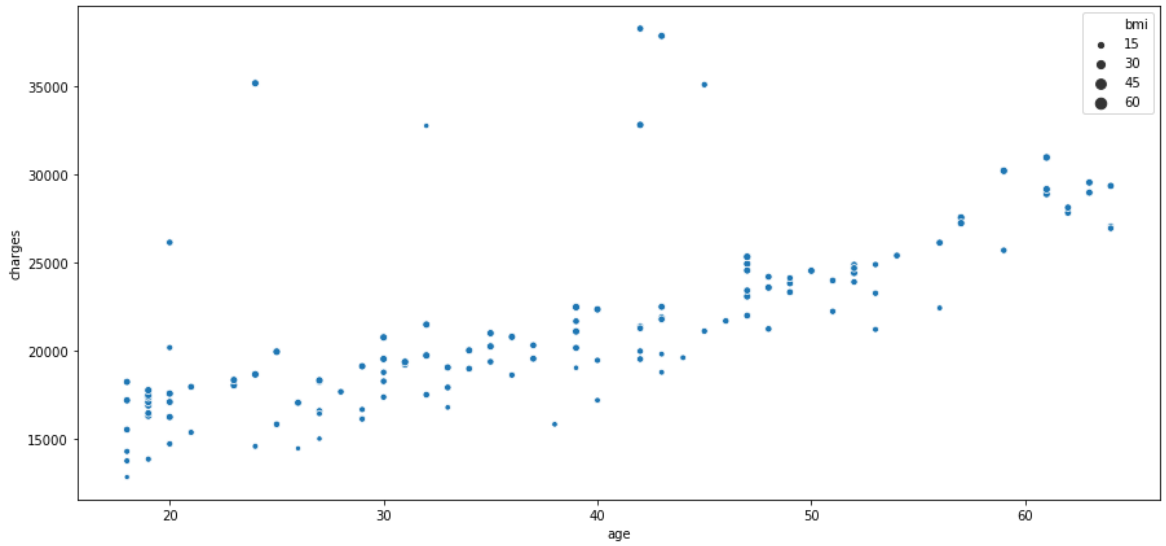
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x135957a8>



Lentelė 10

Rūkančių žmonių be nutukimo draudimo kainos pagal amžių:

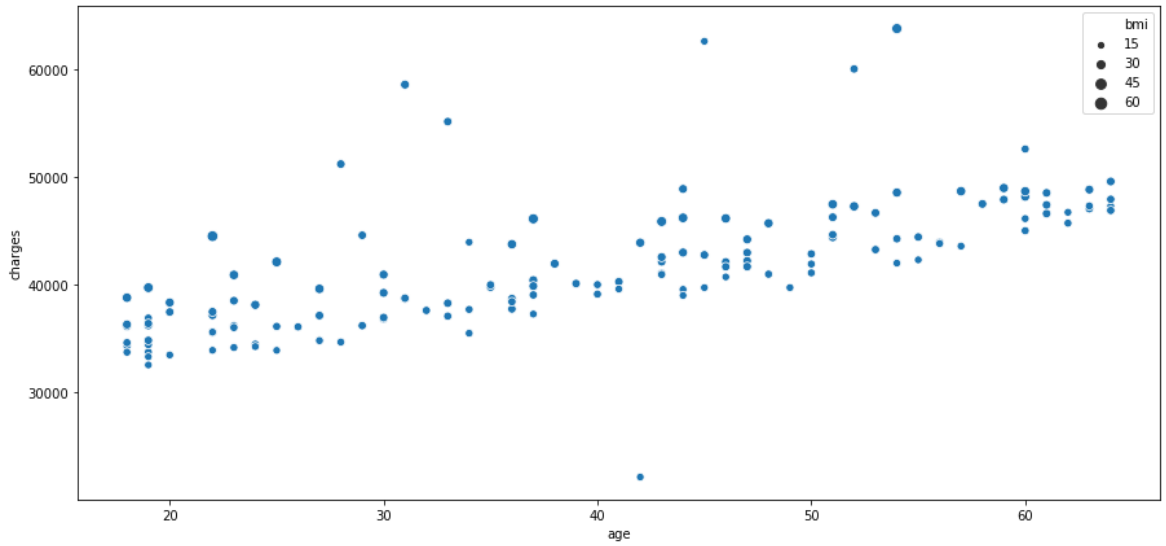
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x13fdf478>



Lentelė 11

Rūkančių žmonių su nutukimu draudimo kainos pagal amžių:

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x13fd8808>



Išvados

1. Didžiausią įtaką draudimo kainai daro korespondento rūkymo istorija.
2. Rūkančiųjų korespondentų draudimo kainai didelę įtaką daro jų kūno masės indeksas.
3. Nerūkančiųjų korespondentų draudimo kainai jų KMI įtakos neturi.
4. Atmetus kitus rodiklius, draudimo kaina ir korespondento amžius sudaro linijinę progresiją.
5. Korespondento lytis, turimų vaikų skaičius ir regionas draudimo kainai įtakos neturi.
6. Išimtinius duomenis galima bandyti paaiškinti per korespondentų ligų istoriją tolimesniuose tyrimuose

Machine Learning

```
1 Sekantis tikslas - pasitelkti mašininio mokymosi modelius draudimo kainos
  apskaičiavimui. Atsakymų tikslumui patikrinti naudoju 5 sluoksnių
  kryžminio patikrinimo (angl. cross-validation) vidurkius. Modeliams
  pateikti duomenys buvo išvalyti nuo reikšmių, kurių kainų paaiškinti
  nepavyko. Išvadas apie kiekvieną modelį galima matyti po jo rezultatais.
```

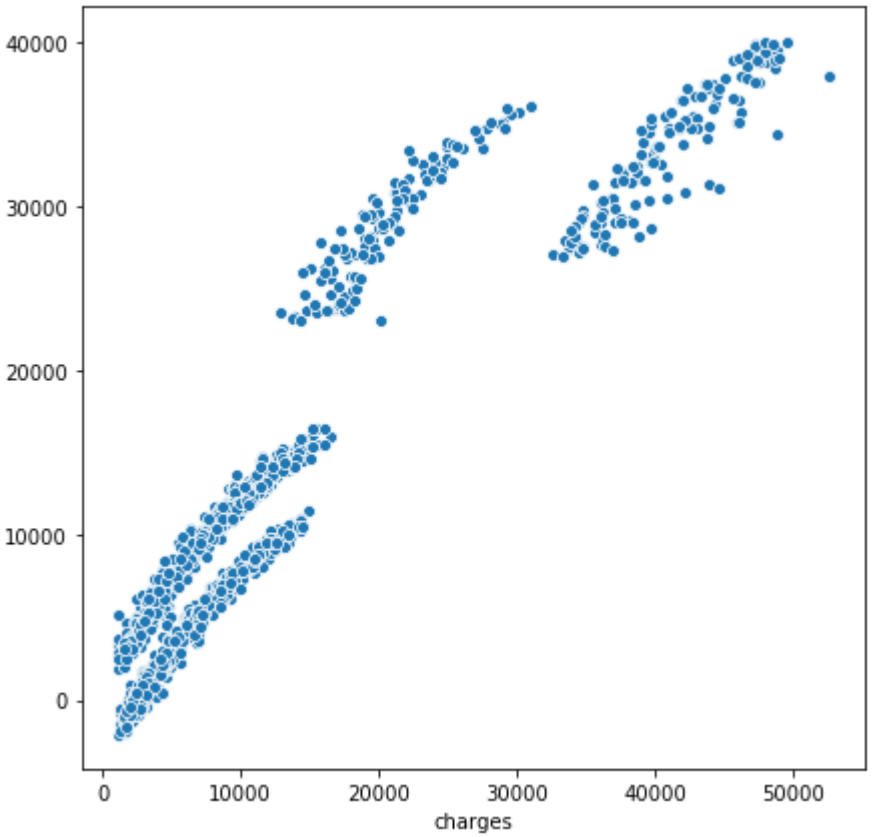
Linear Regression modelis

Modelio skaičiavimo rezultatai:

Out[24]:
```
fit_time                        9.089565e-03
score_time                      5.972576e-03
test_neg_mean_absolute_error   -3.341645e+03
test_neg_median_absolute_error -2.353803e+03
test_r2                         8.684466e-01
test_neg_mean_squared_error    -1.790452e+07
dtype: float64
```

Originalių reikšmių ir gautų rezultatų palyginimas. Horizontali ašis - Originalios reikšmės Vertikali ašis - Modelio skaičiavimai

Out[25]: `<matplotlib.axes._subplots.AxesSubplot at 0x13fd8aa8>`




```
1 Linear regression modelis darbą atliko, bet akivaizdžiai matomi keli
  trūkiai atsakymuose. Analizuodamas duomenis, modelis neįvertino KMI
  reikšmės svorio pasikeitimo tarp rūkančių ir nerūkančių žmonių ir paėmė jų
  vidurkį.
```

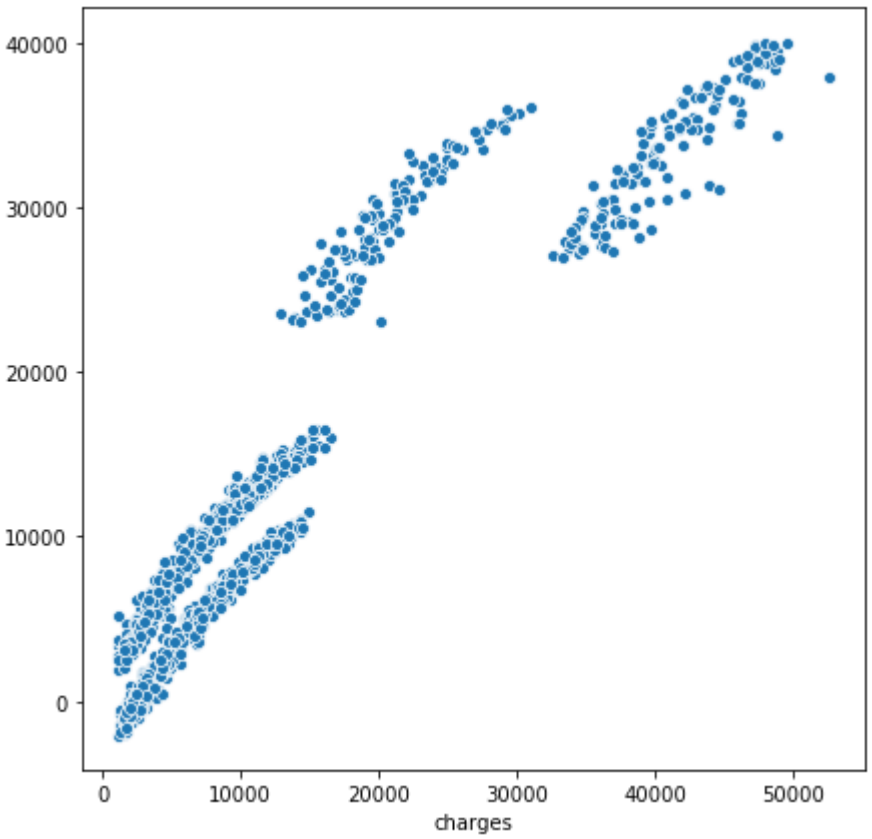
Lasso modelis

Modelio skaičiavimo rezultatai:

Out[27]:
```
fit_time                          9.302425e-03
score_time                        3.880978e-03
test_neg_mean_absolute_error     -3.340536e+03
test_neg_median_absolute_error   -2.353458e+03
test_r2                           8.684549e-01
test_neg_mean_squared_error      -1.790339e+07
dtype: float64
```

Originalių reikšmių ir gautų rezultatų palyginimas. Horizontali ašis - Originalios reikšmės Vertikali ašis - Modelio skaičiavimai

Out[28]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1457d820>
```



Nors lasso modelis davė kiek tikslesnius rezultatus, esminė problema yra tokia pati, kaip ir linear regression modelyje.

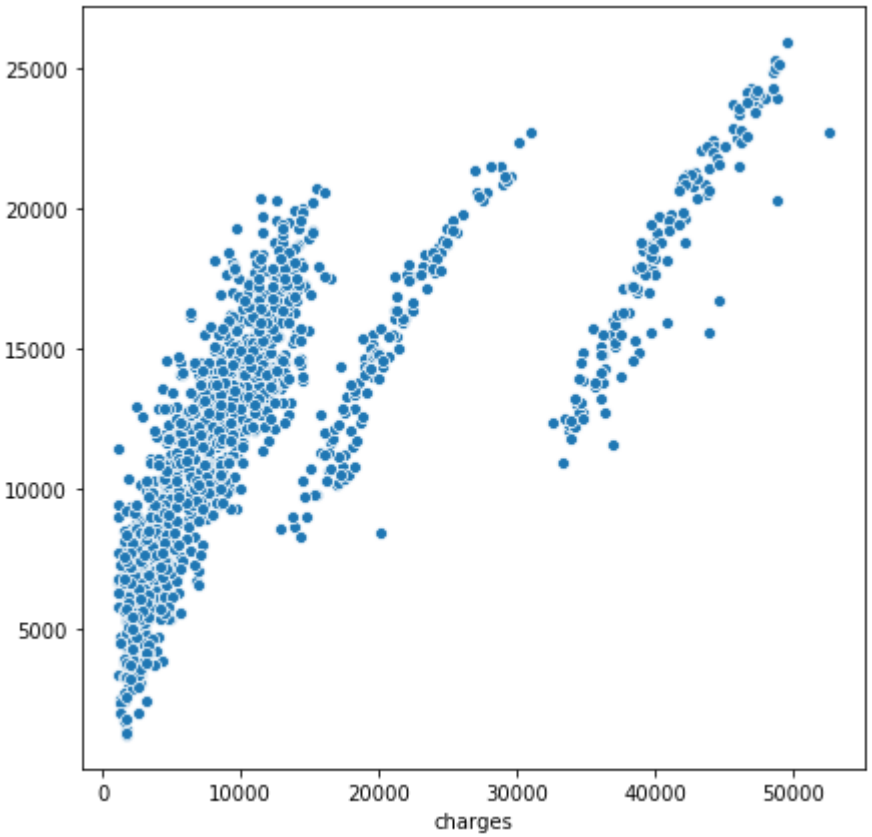
ElasticNet modelis

Modelio skaičiavimo rezultatai:

Out[30]:
```
fit_time                          9.499788e-03
score_time                        2.433920e-03
test_neg_mean_absolute_error     -6.146894e+03
test_neg_median_absolute_error   -4.497415e+03
test_r2                           4.501517e-01
test_neg_mean_squared_error      -7.485921e+07
dtype: float64
```

Originalių reikšmių ir gautų rezultatų palyginimas. Horizontali ašis - Originalios reikšmės Vertikali ašis - Modelio skaičiavimai

Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x145aa1c0>
```



ElasticNet modelio rezultatai yra dar prastesni - šis modelis yra dar toliau nuo norimo atsakymo. Modelis neteisingai įvertino rūkymo reikšmės svorį, vietoj to visą svarbą atiduodamas amžiaus kriterijui.

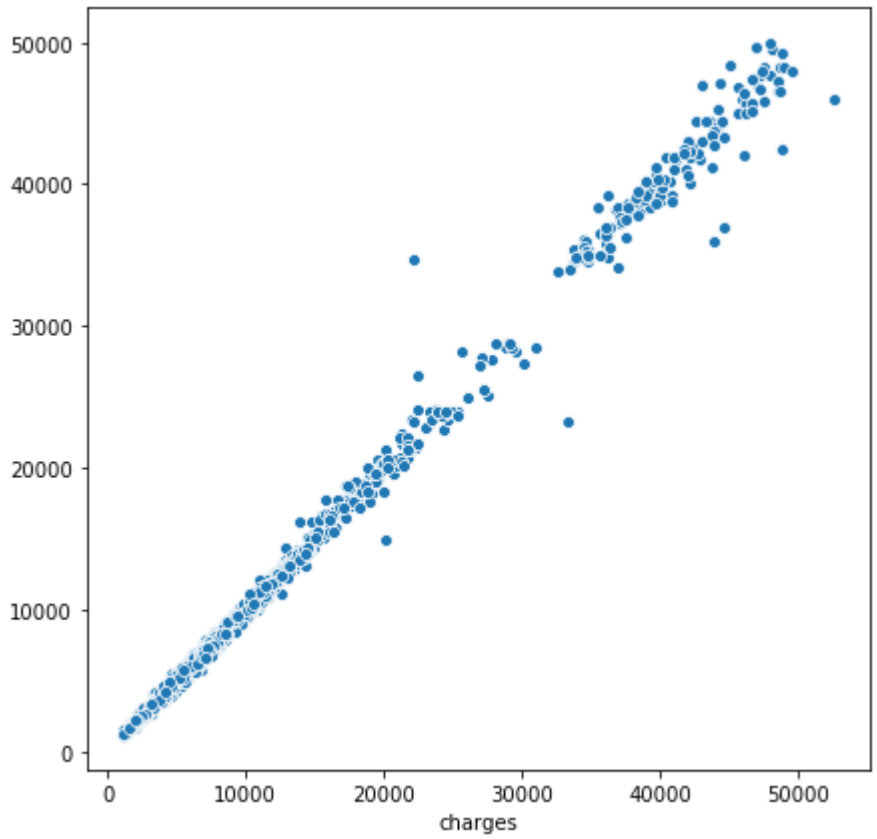
Random forest modelis

Modelio skaičiavimo rezultatai:

Out[33]:
```
fit_time                          0.725613
score_time                        0.020573
test_neg_mean_absolute_error      -390.184908
test_neg_median_absolute_error    -214.304911
test_r2                           0.994592
test_neg_mean_squared_error       -729067.710968
dtype: float64
```

Originalių reikšmių ir gautų rezultatų palyginimas. Horizontali ašis - Originalios reikšmės Vertikali ašis - Modelio skaičiavimai

Out[34]: `<matplotlib.axes._subplots.AxesSubplot at 0x1478e040>`



Random forest modelio rezultatai tikslesni. Apart kelių neatitikimų, visi kriterijai buvo įvertinti teisingai. Pabandžius tam pačiam algoritmui pateikti neišvalytus duomenis, modelio atsakymų tikslumas sumažėjo, todėl nusprendžiau visus modelius vertinti tik pagal jų sugebėjimą apdoroti išvalytus duomenis. Šis modelis duoda patenkinamus rezultatus šios analizės rėžiuose. Tolimesni skaičiavimo tobulinimai turėtų stengtis paaiškinti iki šiol nepaaiškintus duomenis.