

**Kauno technologijos universitetas**  
Informatikos fakultetas

**Komandinis darbas**

P176B101 Intelektikos pagrindai

**Arnas Bradauskas IFF-1/1**

**Ignas Survila IFF-1/5**

Studentai

**Arnas Nakrošis**

Dėstytojas

Kaunas, 2024

# Turinys

<b>1</b>	<b>Duomenų rinkinys</b>	<b>2</b>
<b>2</b>	<b>Metodų aprašymas</b>	<b>3</b>
2.1	K-means . . . . .	3
2.2	SOM . . . . .	3
<b>3</b>	<b>Atstumo metrika</b>	<b>4</b>
<b>4</b>	<b>K-means</b>	<b>5</b>
<b>5</b>	<b>SOM</b>	<b>13</b>
<b>6</b>	<b>K-means ir SOM palyginimas</b>	<b>14</b>

# 1 Duomenų rinkinys

Duomenų rinkinyje egzistuoja tokie atributai:

- work year - metai.
- experience level - darbuotojo patirties lygmuo.
- job title - darbo pozicijos pavadinimas.
- salary in usd - darbo atlyginimas JAV doleriais.
- company location - kompanijos lokacija.
- company size - kompanijos dydis.

## 2 Metodų aprašymas

Naudoti metodai: K-means ir SOM

### 2.1 K-means

Algoritmas suskirsto duomenis į K klasterių, kur kiekvienas taškas priklauso klasteriui su artimiausiu vidurkiu. Algoritmas susidaro iš šių žingsnių:

1. Inicializacija - atsitiktinai parenkame klasterių centrus.
2. Priskytimas - Priskiriame kiekvieną duomenų tašką artimiausiam centrui priklausomai nuo Euklido atstumo.
3. Atnaujinimas - apskaičiuojamas naujas kiekvieno klasterio centras pagal visų taškų vidurkį.
4. Iteravimas - kartojame šiuos žingsnius kol pasiekiamo konvergavimą (centrai nebesikeičia ar keičiasi labai nežymiai).

### 2.2 SOM

Dirbtinis neuroninis tinklas, naudojamas neprižiūrimam mokymui. Žingsniai:

1. Inicializacija - atsitiktinai parenkame neuronų svorius.
2. Apmokymas - Kiekvienam įvesties vektoriui randame BMU - neuroną, kuris turi artimiausią svorio vektorį. Atnaujiname BMU ir jo kaimyninius neuronus.
3. Iteravimas - kartojame apmokymą daug iteracijų, po truputį mažindami mokymosi spartą ir kaiminystės dydį.

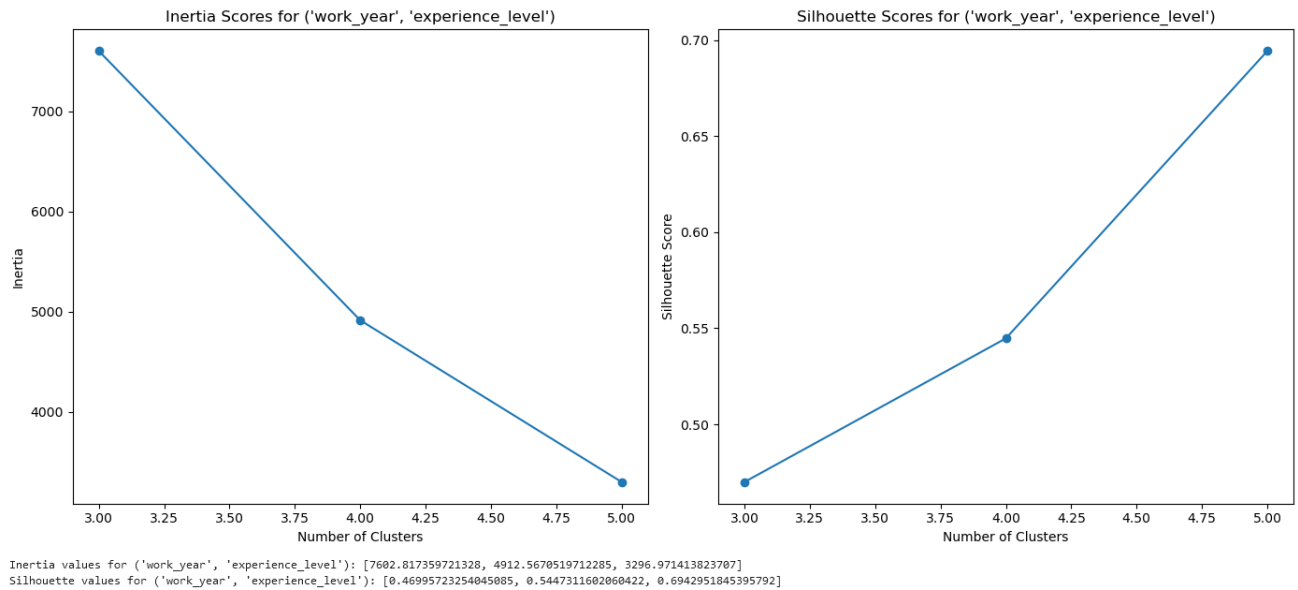
### 3 Atstumo metrika

Algoritmų palyginimui naudojome Euklido atstumo metriką. Turint du taškus  $\mathbf{p} = (p_1, p_2, \dots, p_n)$  ir  $\mathbf{q} = (q_1, q_2, \dots, q_n)$   $n$ -dimencijų erdvėje, Euklido atstumas  $d(\mathbf{p}, \mathbf{q})$  tarp jų yra skaičiuojamas taip:

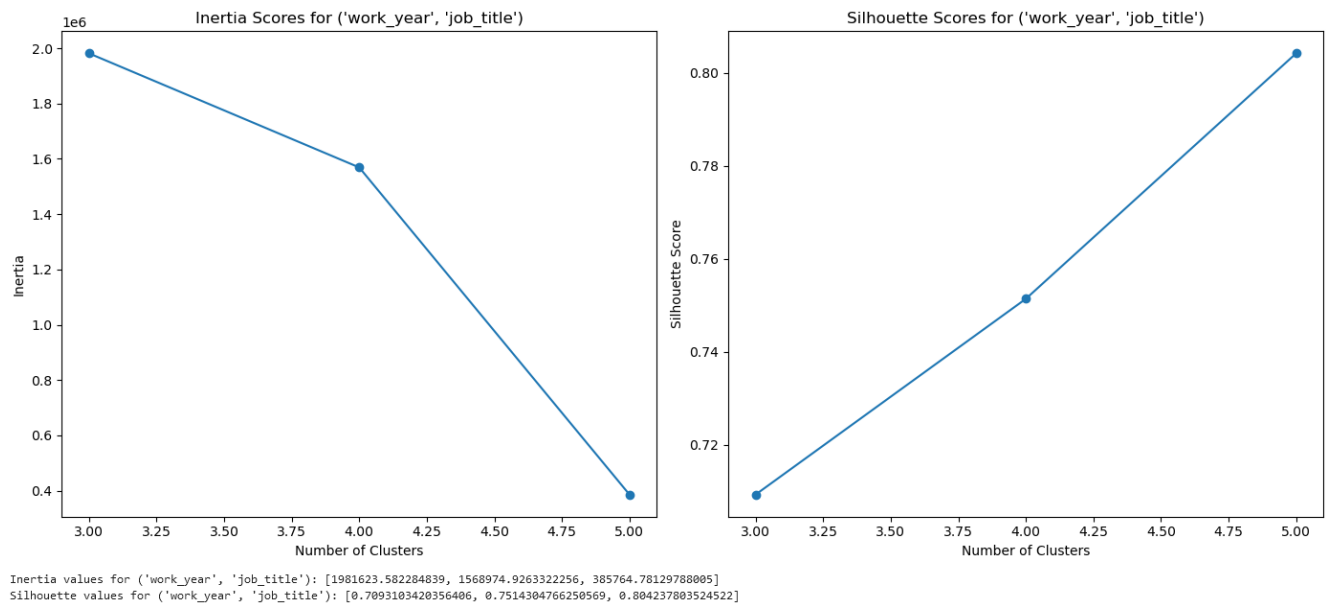
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

## 4 K-means

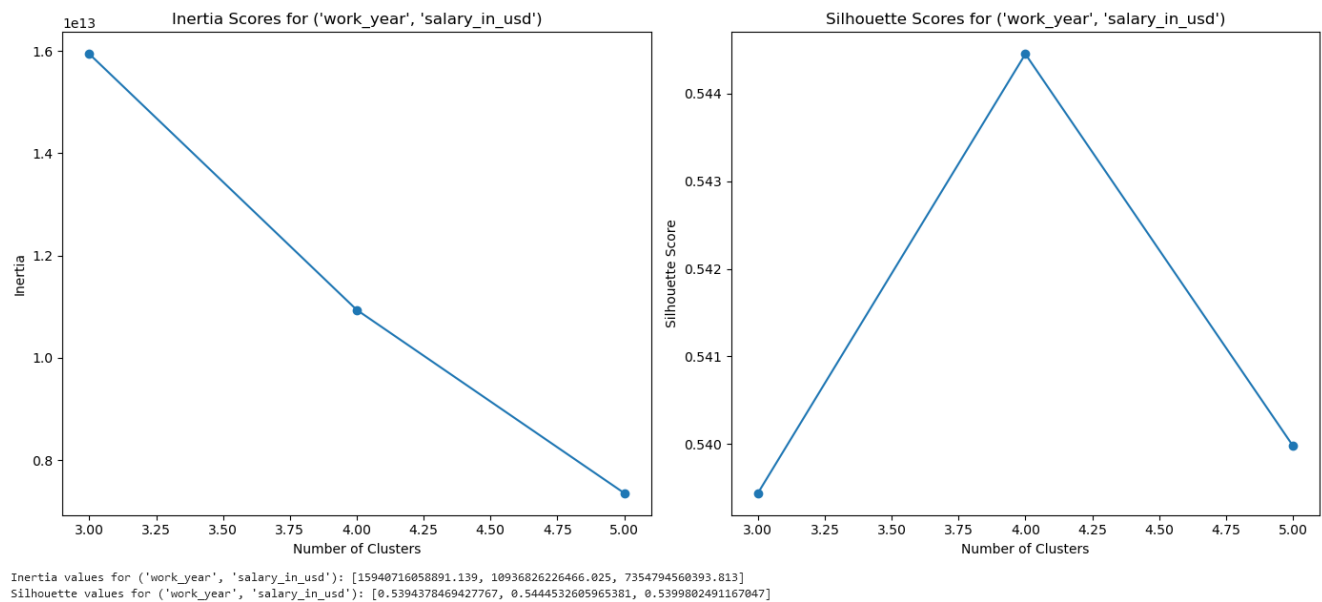
Skaiciuojame silueto koeficientus ir inercijos reikšmes visoms įmanomoms atributų kombinacijoms, su klasterių kiekiu = [3,4,5].



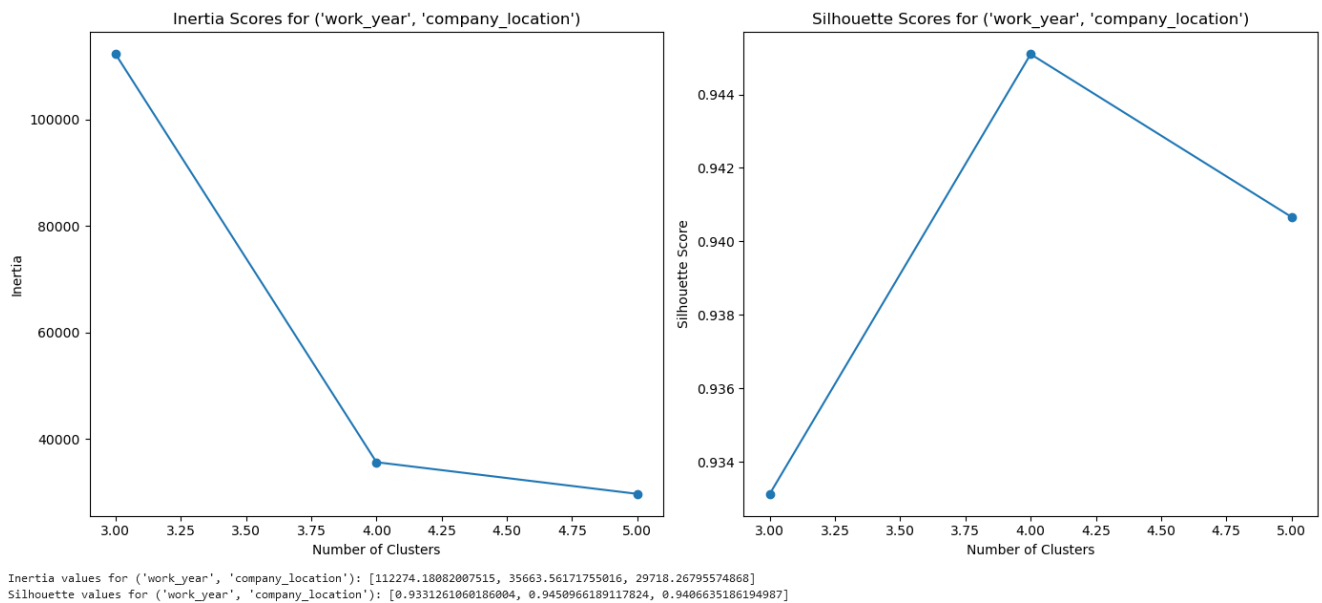
1 pav. Grafikai "work year" x "experience level".



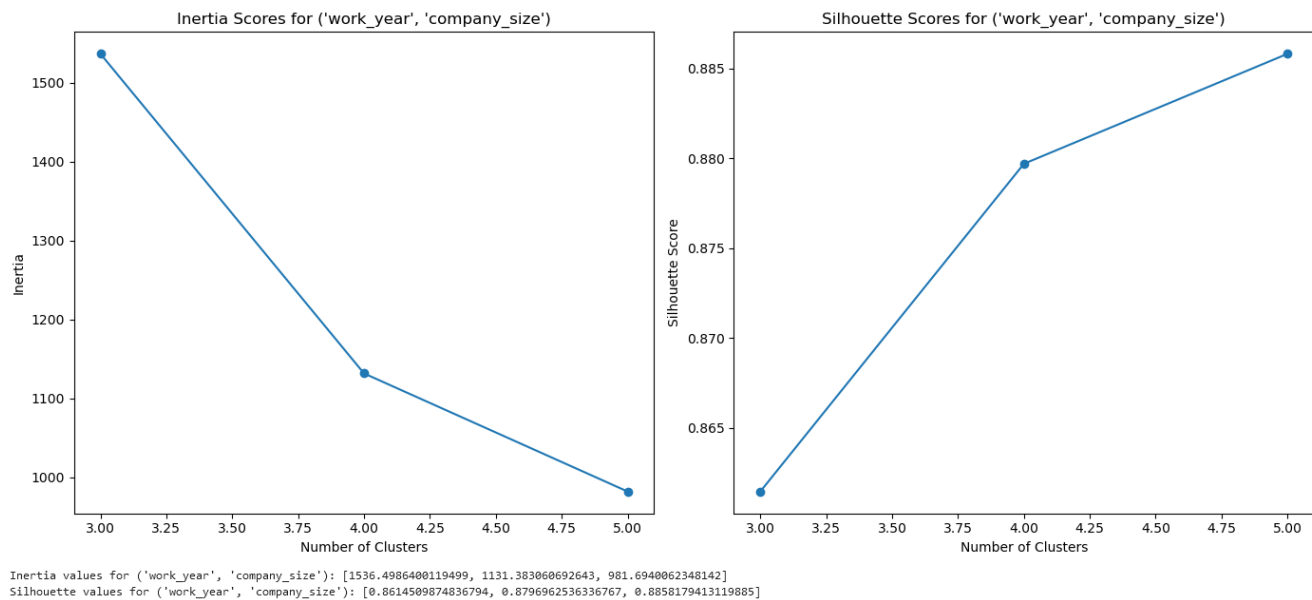
2 pav. Grafikai "work year" x "job title".



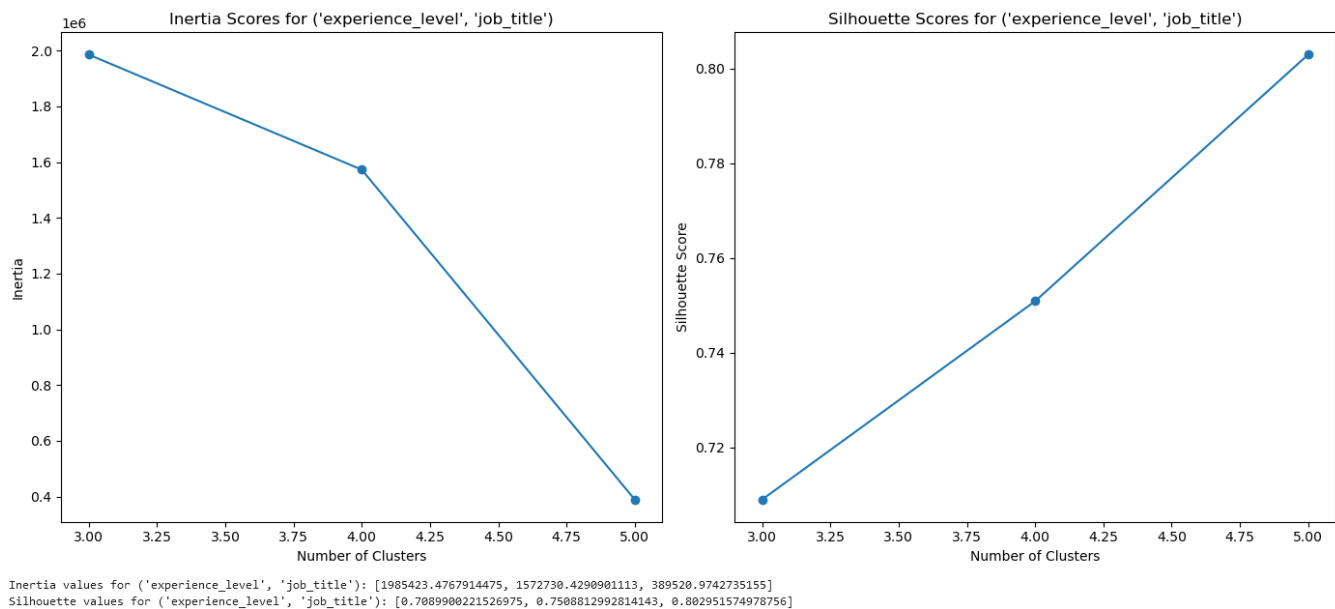
3 pav. Grafikai "work year" x "salary in usd".



4 pav. Grafikai "work year" x "company location".

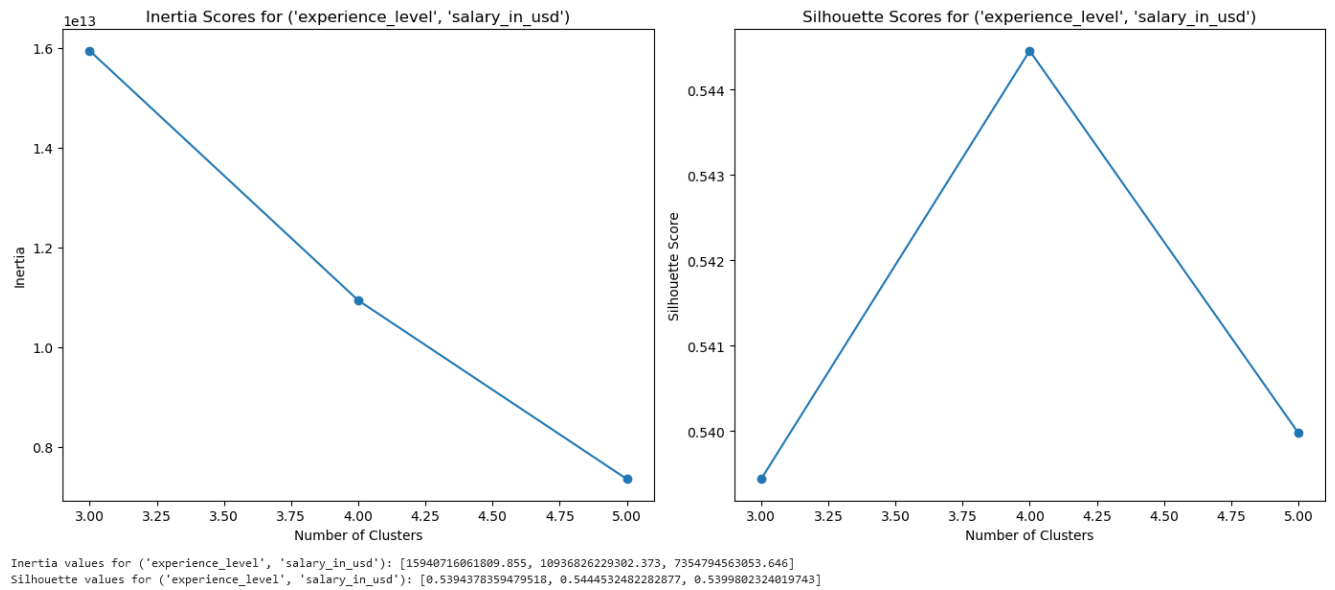


5 pav. Grafikai "work year" x "company size".

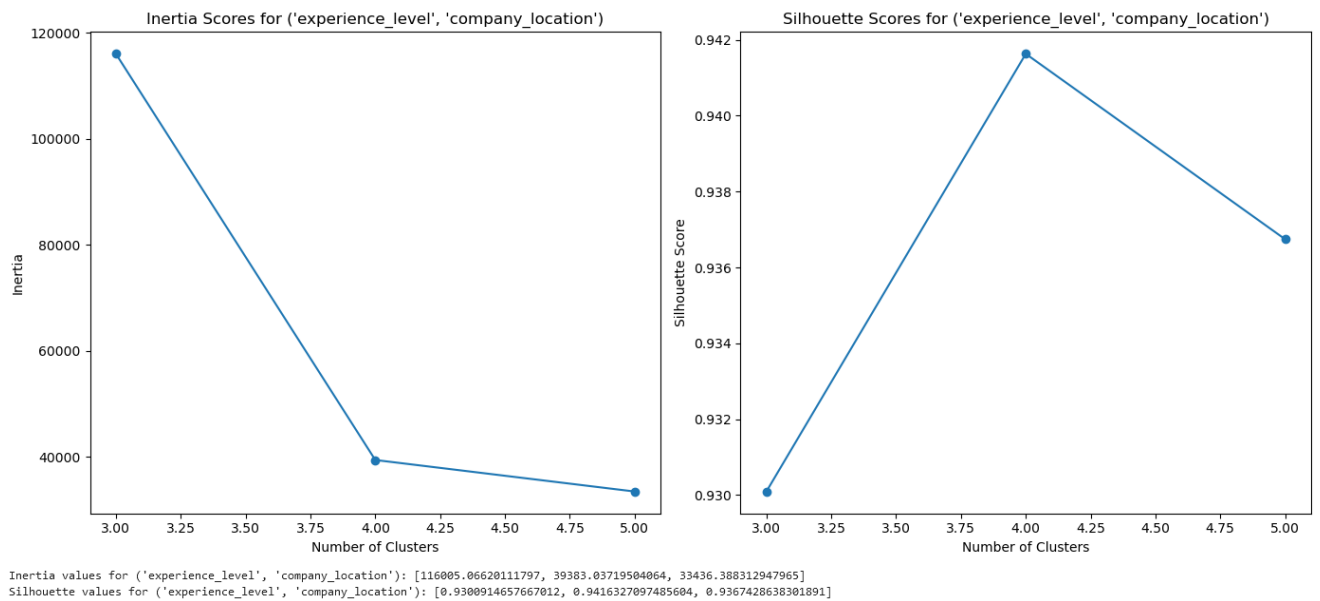


6 pav. Grafikai "experience level" x "job title".

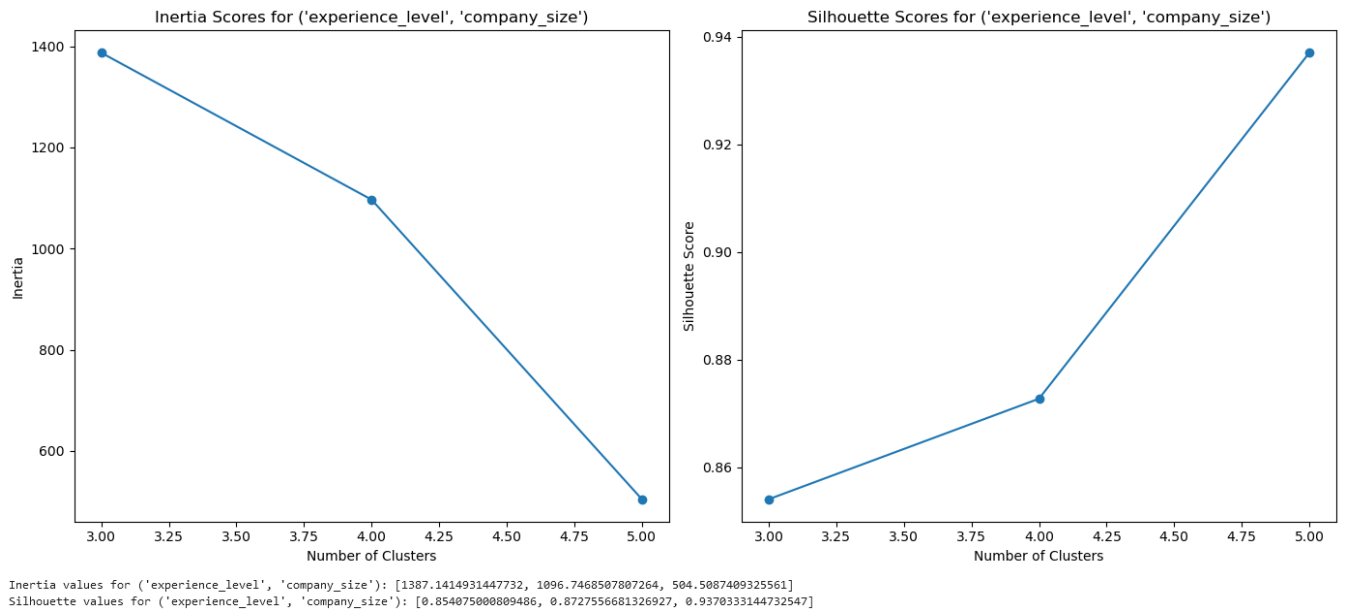




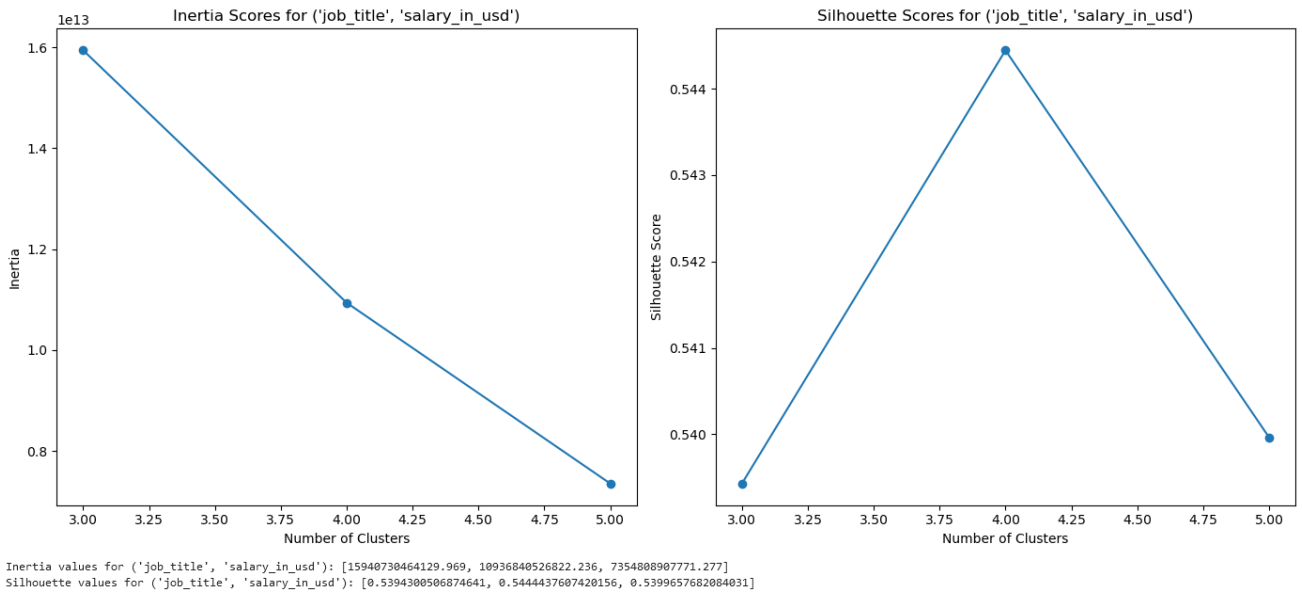
7 pav. Grafikai "experience level" x "salary in usd".



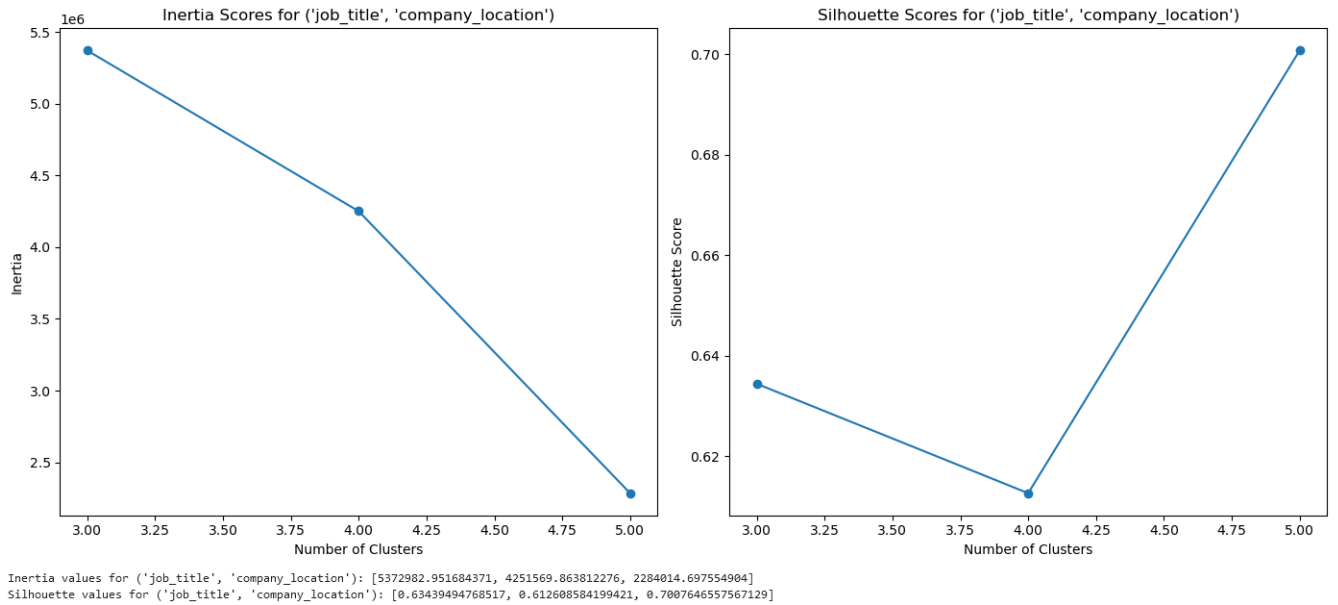
8 pav. Grafikai "experience level" x "company location".



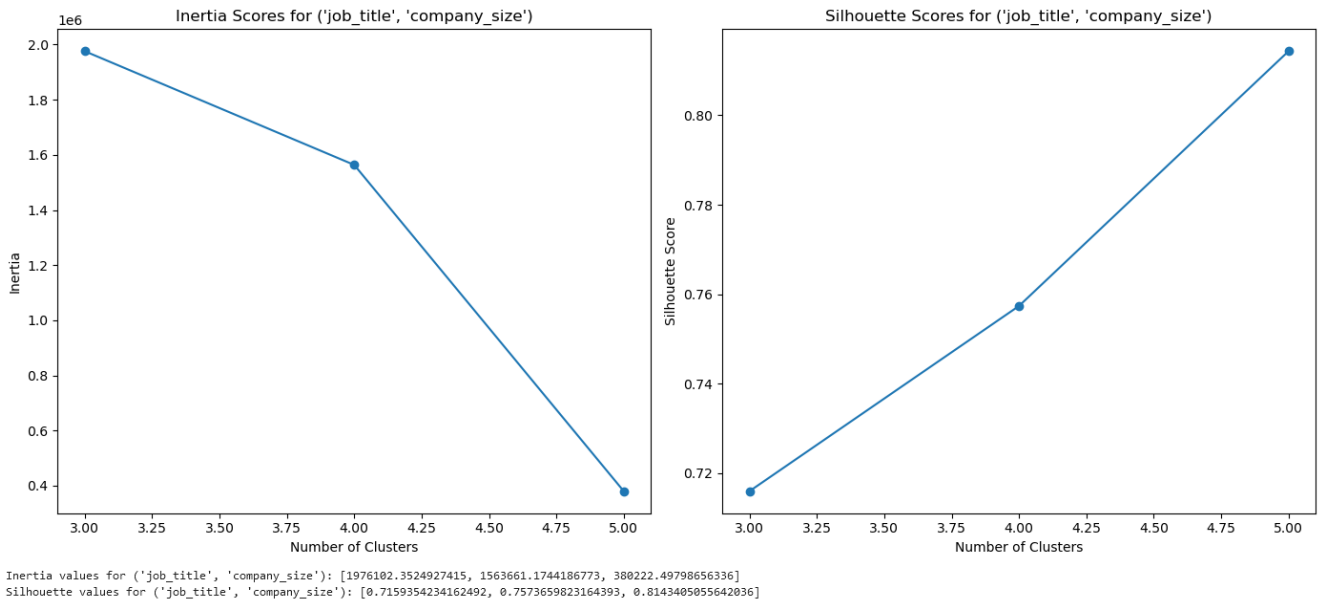
9 pav. Grafikai "experience level" x "company size".



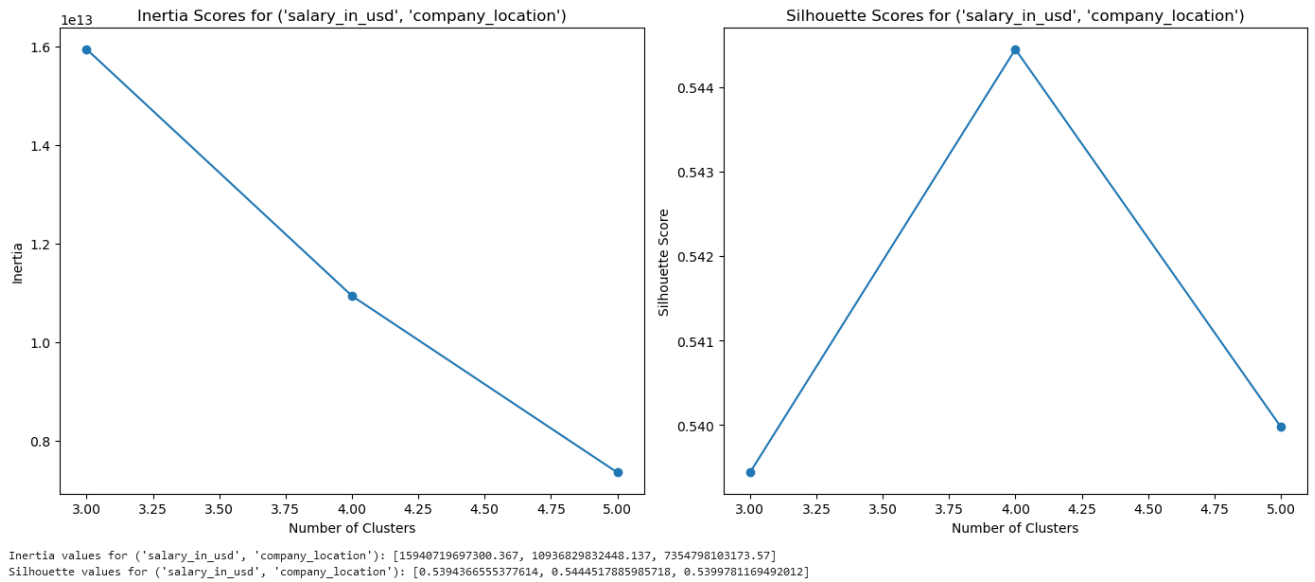
10 pav. Grafikai "job title" x "salary in usd".



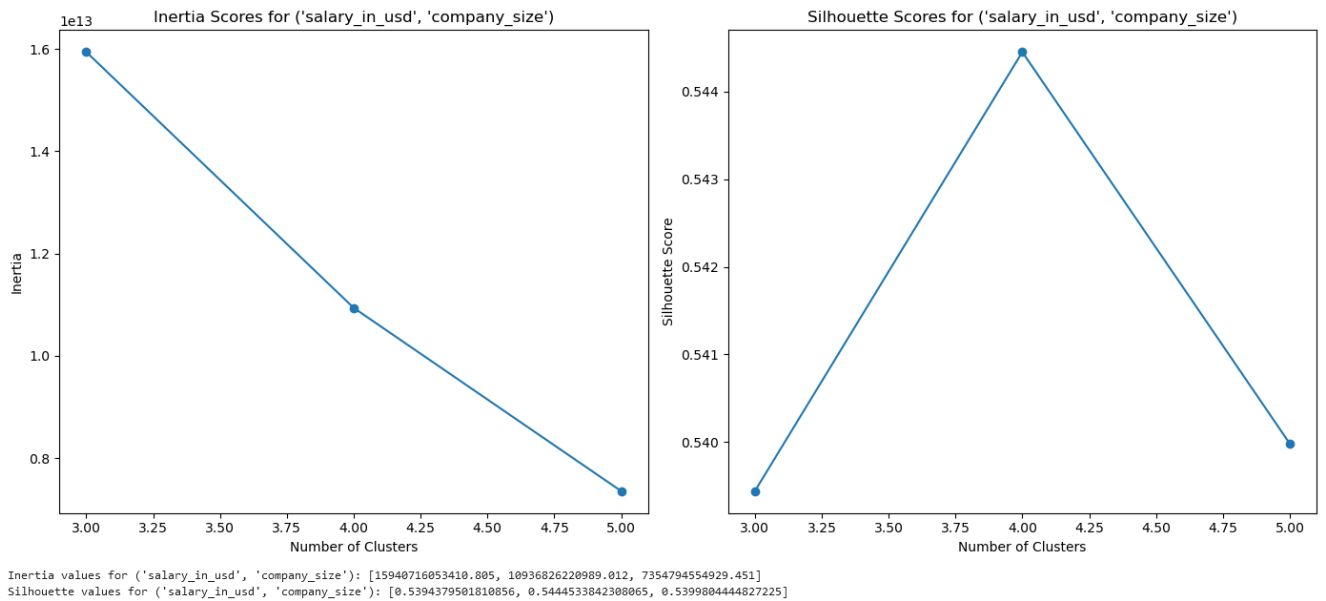
11 pav. Grafikai "job title" x "company location".



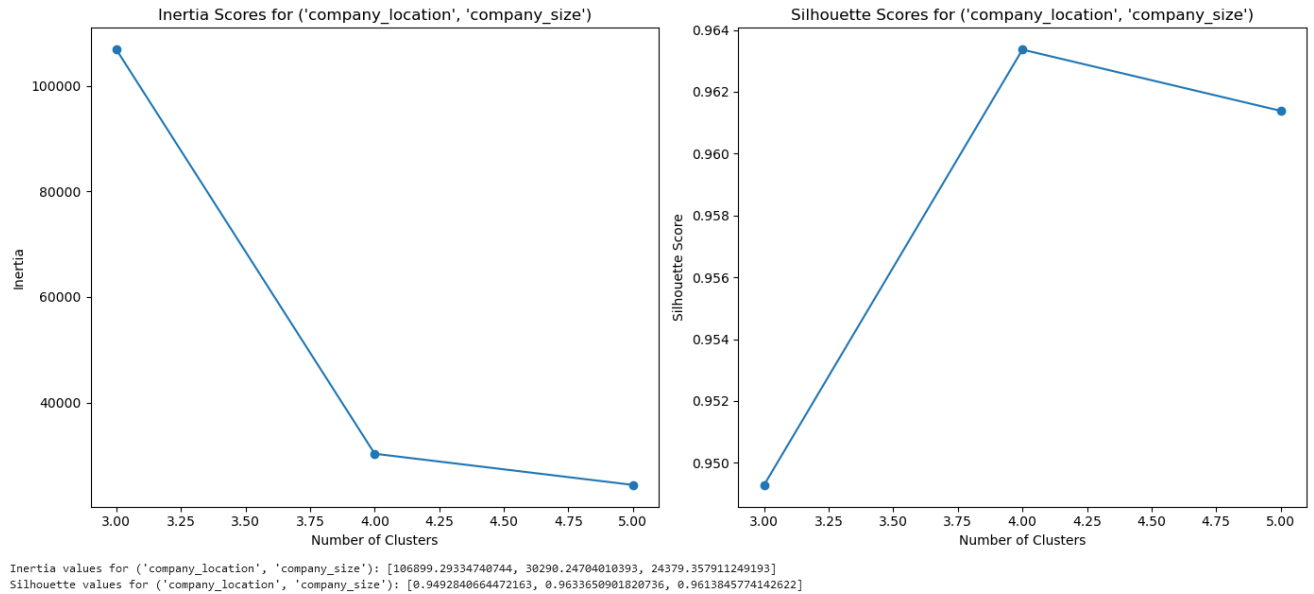
12 pav. Grafikai "job title" x "company size".



13 pav. Grafikai "salary in usd" x "company location".



14 pav. Grafikai "salary in usd" x "company size".

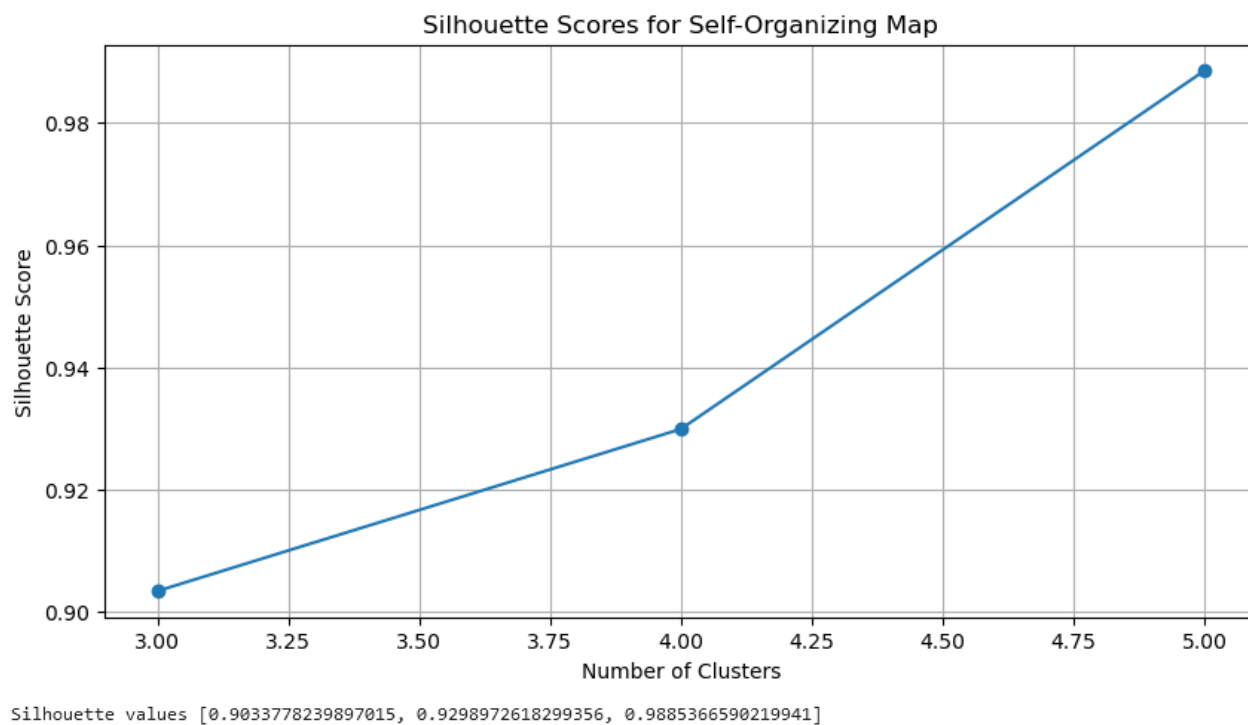


15 pav. Grafikai "company location" x "company size".

Matome, jog geriausi rezultatai (didžiausias silueto koeficientas bei mažiausia inercija) gavosi su "experience level" x "company size" palyginimu, todėl toliau jį naudosime kitiems skaičiavimams ir palyginimams.

## 5 SOM

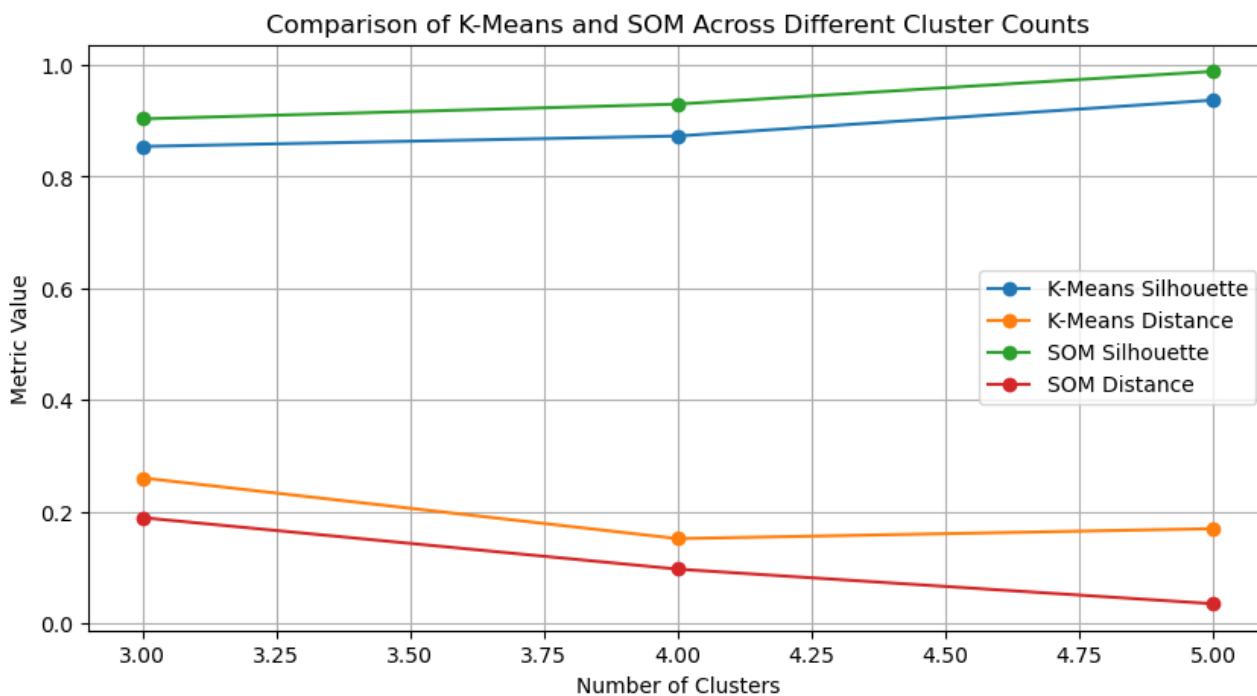
Skaičiuojame silueto koeficientus "experience level" x "company size" su klasterių kiekiu = [3,4,5].



16 pav. SOM silueto grafikas "experience level" x "company size".

## 6 K-means ir SOM palyginimas

Galiausiai naudojame apskaičiuojame Euklido atstumą K-means ir SOM rezultatams, naudojame šiuos rezultatus kartu su siluetų koeficientais, kad palyginti abu algoritmus.



17 pav. K-means ir SOM silueto koeficiento bei Euklido reikšmės palyginimas.

Matome, jog SOM algoritmas šiek tiek lenkia K-means lyginant ir siluetų koeficientus, ir Euklido atstumus. Tačiau abu algoritmai tiekia sąlyginai gerus rezultatus.