



TARTU ÜLIKOOL

Eestikeelse subjektiivsuse andmestiku arendamine

Karl Gustav Gailit

Juhendajad Kadri Muischnek ja Kairit Sirts

Rakenduslingvistika kevadkonverents

24. aprill 2025

Subjektiivsus

- Subjektiivne tekst väljendab autori arvamusit või hoiakuid
- Objektiivne tekst väljendab fakte, reaalsust
- Levinud teema lingvistikas
 - Pragmatika
 - Subjektiivsus = öeldu/kirjutatu sisusse suhtumine
 - Intersubjektiivsus = kaassuhtlejasse suhtumine (Hennoste jt 2022)
 - Subjektiivsus kui subjekti esile toomine (Sanders 1999)
 - Seotud modaalsusega, aga iseseisev
 - Formaalne semantika
 - Väide on subjektiivne siis, kui see ja selle eitus võivad samaaegselt tõesed olla

Subjektiivsus keeletehnoloogias

- Sageli alamülesanne
 - Eesmärk ei ole subjektiivsus ise
 - Seega väga lihtsustatud lähenemine
 - Meelestatuse analüüs ja arvamuskaeve
 - Et vaadata arvamusi, tuleb need esmalt leida
 - Vahel meelestatust ja subjektiivsust käsitletud samana
 - Pigem meelestatus kui subjektiivsuse alamidee (Liu 2015)
- Võib olla ka iseseisev ülesanne
 - Tekstile silt objektiivne või subjektiivne
 - Lisasildid mitesobivale (nt küsimused)

Subjektiivsus keeletehnoloogias

- Lausete klassifitseerimine
 - Andmestik siltidega, kas lause on objektiivne või subjektiivne
 - Přibáň ja Steinberger (2022) tšehhi keele subjektiivsuse klassifitseerija
 - Laused pärinesid filmide arvustustest ja sisukirjeldustest
 - Võrreldi inimeste ja allika alusel märgendatud lauseid.
- Terviktekstide klassifitseerimine subjektiivseks ja objektiivseks
 - Tänapäeval harvem ülesanne, reeglina vaid allika põhjal
 - DEFT'09 ajaleheartiklite klassifitseerimise ülesanne
 - Määratud kategooria alusel, nt uudis objektiivne, arvamuslugu subjektiivne
 - Prantsuse ja inglise keel (Toprak ja Gurevych 2009)

Varasema liigitamise probleemid

- Pigem on tegu tekstide liigitamisega muudesse klassidesse
 - Arvustus või kokkuvõte; uudis või arvamuslugu
 - Domeenispetsiifilisus
- Tekstisisene varieerumine
 - Arvamuslood sisaldavad arvamuse tõestamiseks fakte
 - Uudistes sageli edastatud teiste arvamusi
 - Sealhulgas intervjuud ja poliitika
- Mitte-arvamuslugudes on 44% lauseid subjektiivsed (Wiebe jt 2001)

Mis teeb teksti subjektiivseks?

- Eesmärk esitada arvamust (mitte fakte või teavet)?
 - Inimene saab aru, aga masinal läheb see keerulisemaks
 - Autori eesmärk võib ebaõnnestuda
 - Uudis sisaldab liiga palju arvamusi, ning on pigem arvamusloona loetav
 - Muud eesmärgid
 - Lugejas emotsiooni tekitamine, ilukirjandus
- Tekstis esinevate lausete subjektiivsus?
 - Kui üle poolte lausete tekstis on subjektiivsed, on tekst ise subjektiivne
 - Aga kõik laused tekstis ei ole võrdse tähtsusega
 - Kui tekst sisaldab arvamuse tõestamiseks palju fakte?
- Kombinatsioon? Midagi kolmandat?

Alternatiivne andmestik

- Katsetada, kas tekstide subjektiivsust võrdlevat andmestikku on keeletehnoloogilistes ülesannetes parem kasutada
 - Skalaarne lähenemine
 - Sildi asemel anda tekstile arv skaalal
 - 0 kindlalt objektiivne, 100 kindlalt subjektiivne, kõik muu vahepealne
 - Saab teisendada siltideks
- Kasutada ära inimese subjektiivsuse intuitsiooni
 - Teksti eesmärk, kuidas see edastab arvamusi ja informatsiooni
- Võimaldab vaadata, kuidas suhestub subjektiivsuse hinnang teiste teksti aspektidega
 - Suhted tekstiliigi ja kirjutamisviisiga

Alternatiivne andmestik

- Pole uus lähenemine
 - Wiebe (2000) subjektiivsete üksuste otsimine lauses
 - Igale üksusele hinnang, kui tugevalt subjektiivne
 - Skaala üks kuni kolm
 - Belgia ajakirjandustekstidest subjektiivsete lõikude otsimine (Escouflaire, Descampe ja Fairon 2024)
 - Likerti skaala üks kuni viis
 - Mitu märgendajat, lõigu hinnang märgendajate keskmine
 - Seega kasutab kogu skaalat
- Eesmärk on arv, mitte aste skaalal
 - Küsida otse arvu

Pilootprojektist

- 60 teksti Ühendkorpusest (2021)
 - 40 suvaliselt valitud veebiteksti
 - 5 uudist ja 5 arvamuslugu ERR-ist
 - 5 uudist ja 5 arvamuslugu Uutest Uudistest
- 2 märgendajat
- Liuguriga märkida teksti subjektiivsuse taset
 - Vahemik 0 kuni 100
- Lisaks märkida kindlus 3-palli skaalal
 - Võimalus lisada kommentaare
- Limesurvey

Pilootprojektist

- Hinnang sõltub allikast
 - Uudised objektiivsed, arvamuslood subjektiivsed
 - Mõne allika uudised aga hinnatud subjektiivsel
- Mida rohkem skaala keskel, seda rohkem ebakindlust
- Korrelatsioon hindajate vahel kõrge
 - Spearman = 0.8738
 - Pearson = 0.8599
- Suurim erinevus nende vahel 50
 - Üks märgendas 26, teine 77
- Pilootprojekt õnnestus, suurem andmestik võimalik

Märgendamisest

- Märgendamine liuguriga skaalal 0 kuni 100
- 3 märgendajat
 - (2 naist ja 1 mees, 19-25, Tartu Ülikooli üliõpilased)
- 3 näidisteksti
 - Objektiivne, subjektiivne, ja üks mis ei kuulu kumbagi otsa
- Ei ole palutud välja tuua, mis tingis otsuse
 - Hoiduda sellele keskendumisest
- Saadetud märgendamiseks 4 portsus

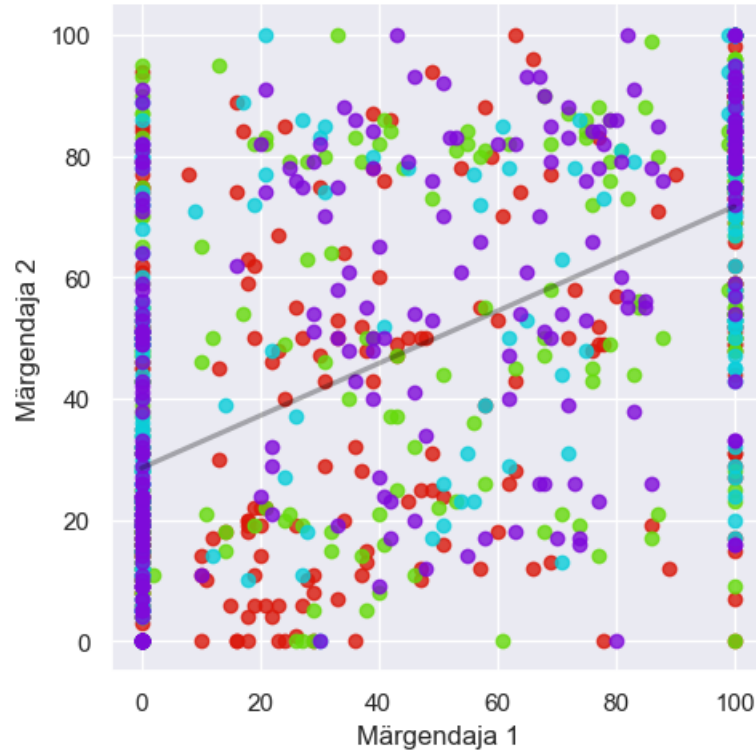
Andmestikust

- 1000 teksti
 - 300 ajakirjandus, 700 veebitekstid
 - 150 uudised, 150 arvamuslood
 - Portsus proportsioonid samad
 - 37 uudist, 38 arvamuslugu, 175 veebiteksti
- Pikkus 100 kuni 6000 tähemärki
 - (keskmine 1278, mediaan 970)
- Sõnu 19 kuni 1096
 - (keskmine 197, mediaan 145)
- Lauseid 1 kuni 85
 - (keskmine 12, mediaan 9)

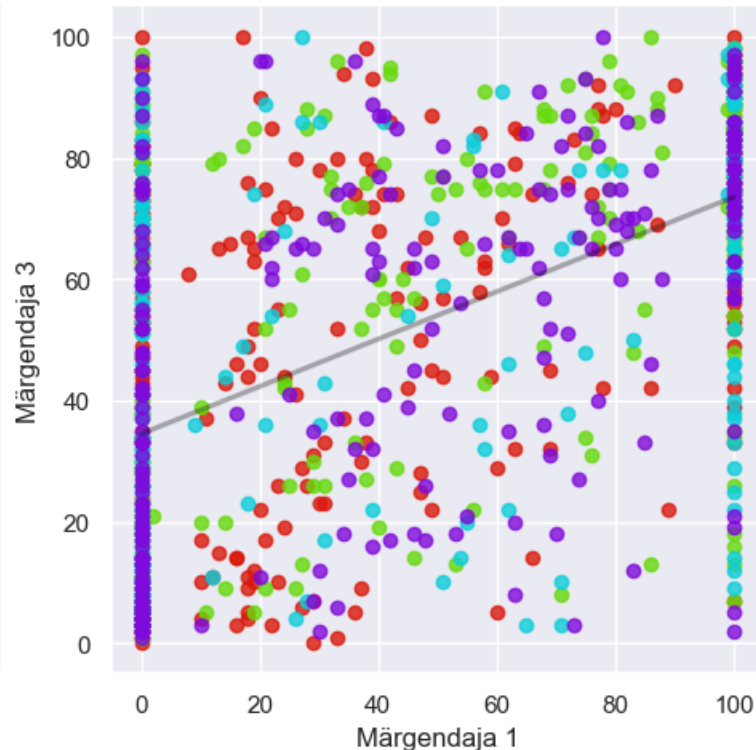
Andmestikust

- Ajakirjandustekstid Ühendkorpuse (2021) Feeds alamkorpusest
- Veebitekstid kõikidest alamkorpustest
- Valitud juhuslikult
 - Esmalt juhuslik allikas
 - Seejärel juhuslik tekst sellest allikast
- Tekstide seest asendatud isiklikku infot
 - Telefoninumbrid, täisnimed, meiliaadressid
 - Avalikud andmed alles jäetud
 - Tuntud inimeste nimed, ettevõtete kontaktandmed

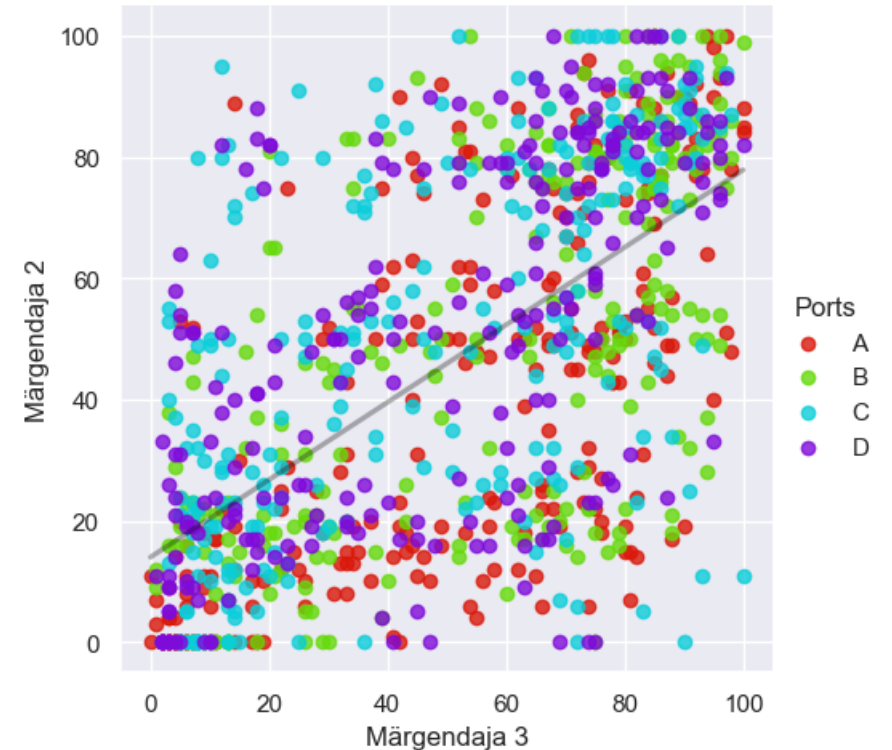
Kooskõlad



Spearman: 0.5632
Pearson: 0.5697

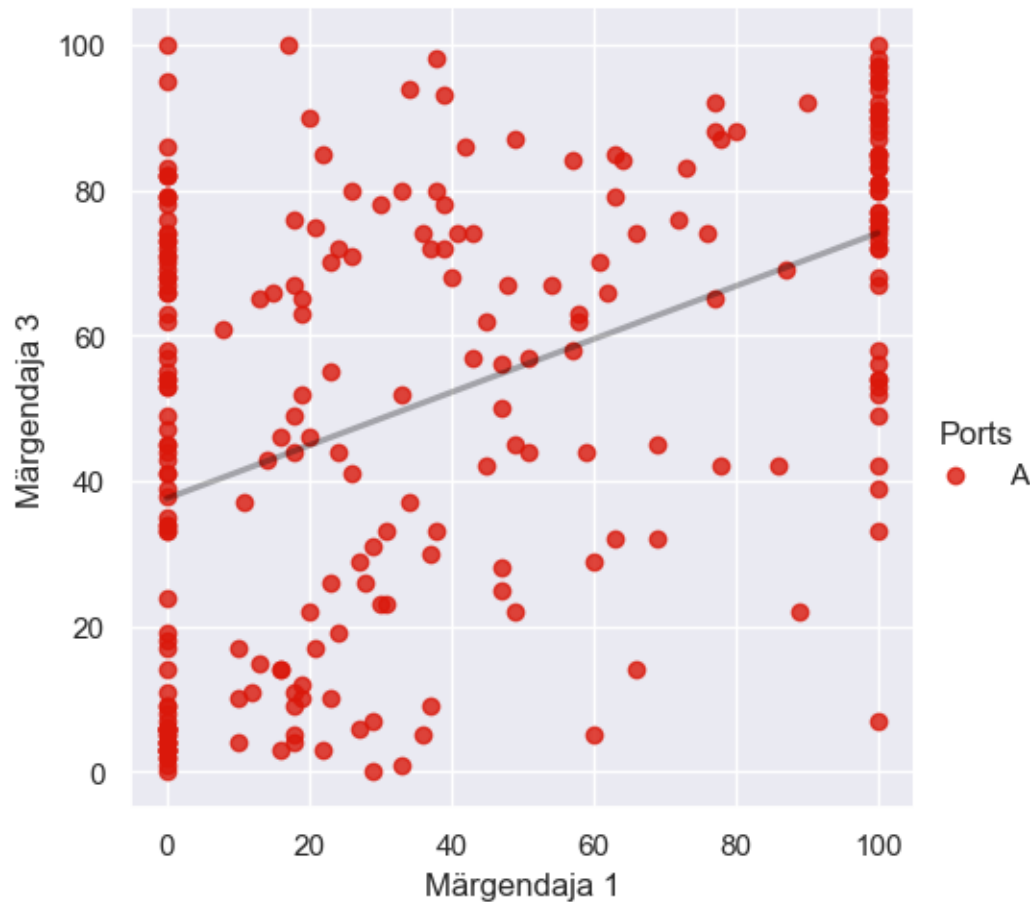


Spearman: 0.5212
Pearson: 0.5251



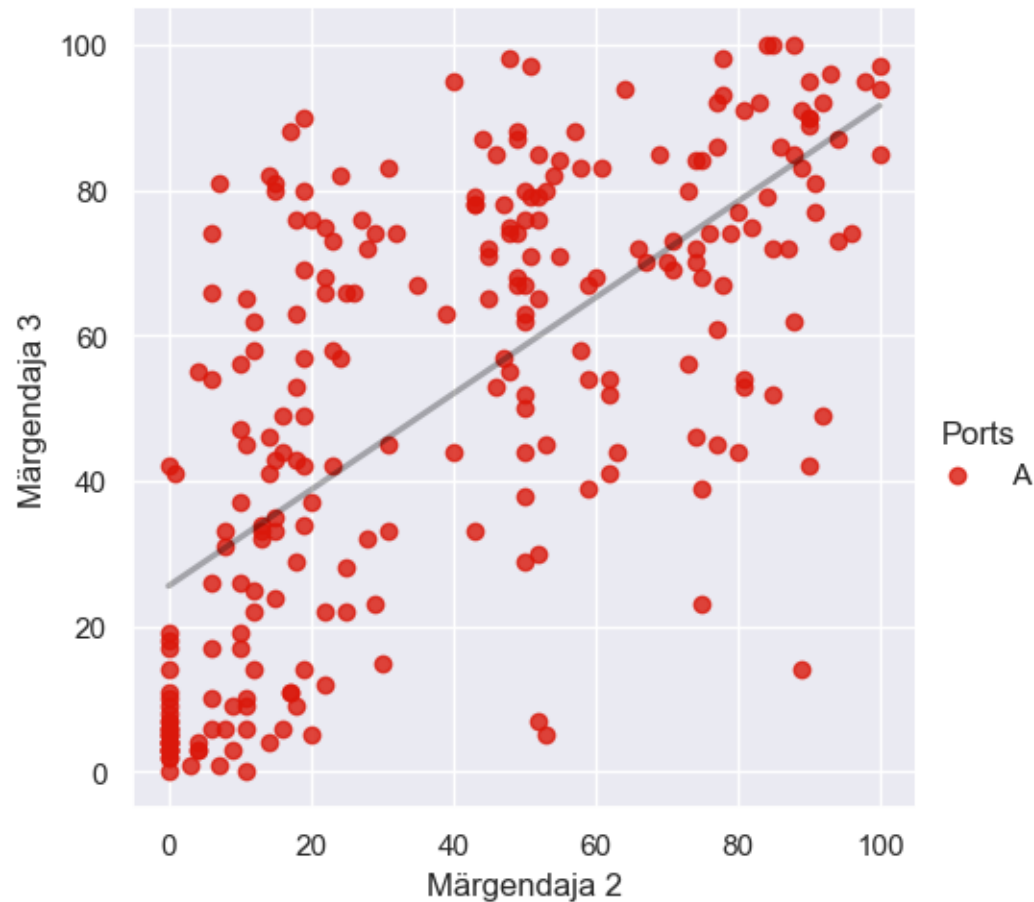
Spearman: 0.6293
Pearson: 0.6270

Suurim ja väikseim kooskõla



Spearman: 0.4463

Pearson: 0.4614



Spearman: 0.6904

Pearson: 0.6719

Kategooriad

Arvamuslugu – 150 tk

Uudis – 150 tk

Arvustus – 14 tk

Reklaam – 331 tk

Veebist ajakirjandus – 88 tk

Informeeriv – 46 tk

Juhis – 43 tk

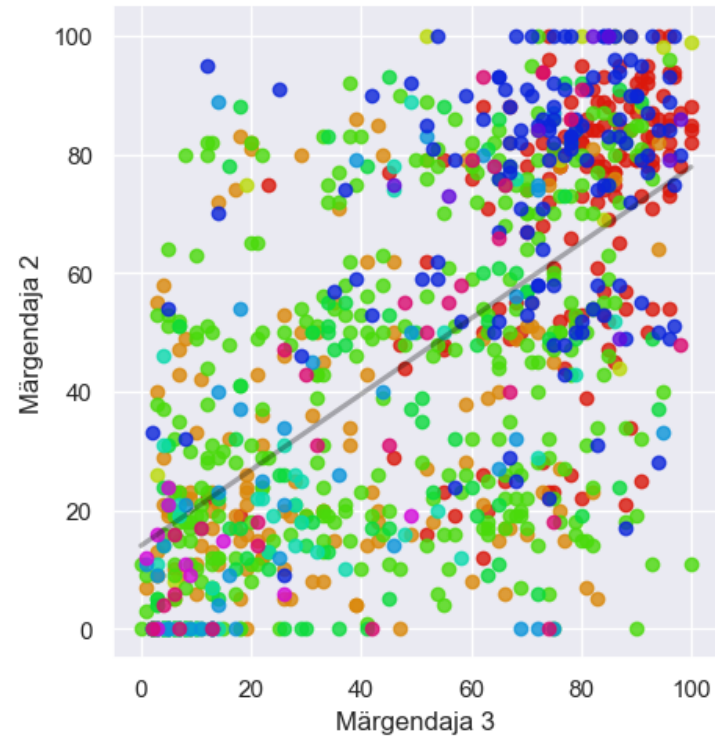
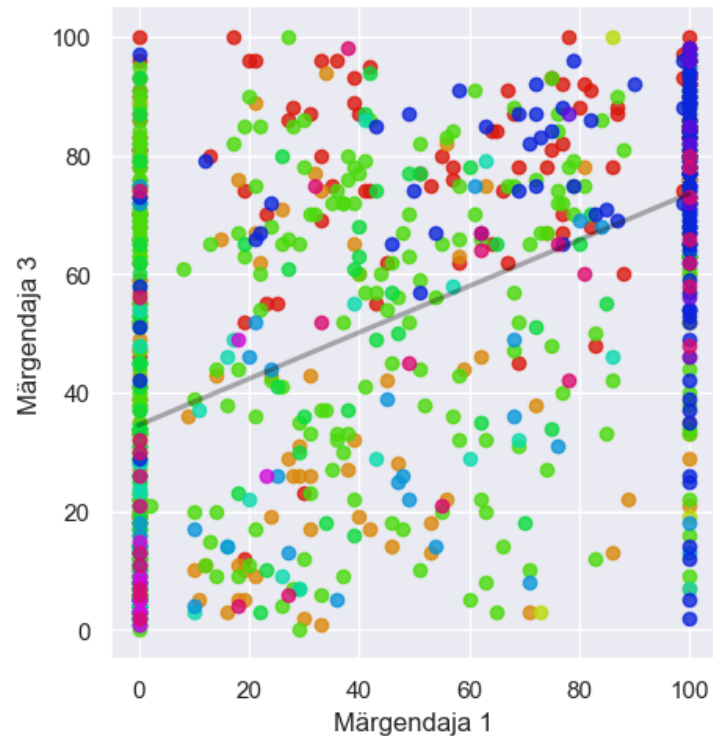
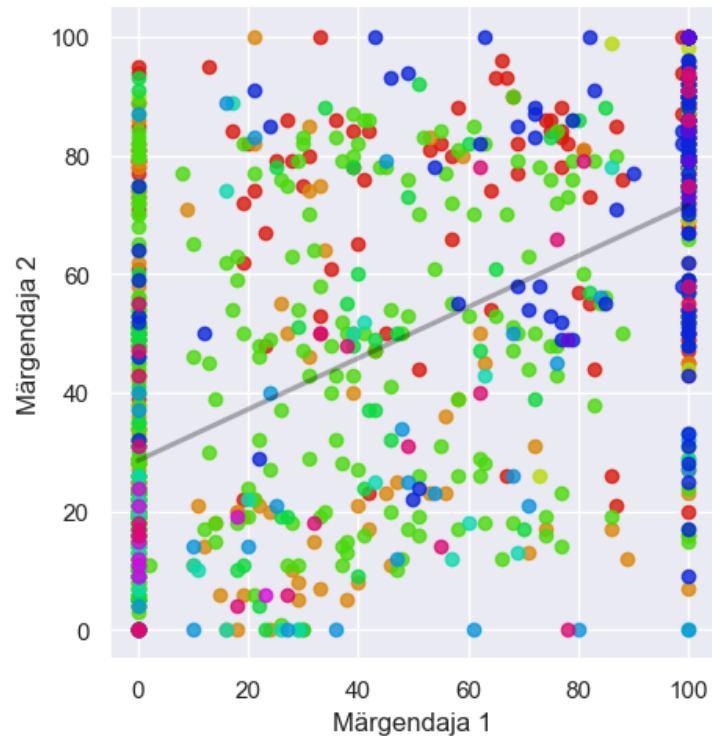
Sotsiaalmeedia – 130 tk

Ilukirjandus – 8 tk

Legaal – 10 tk

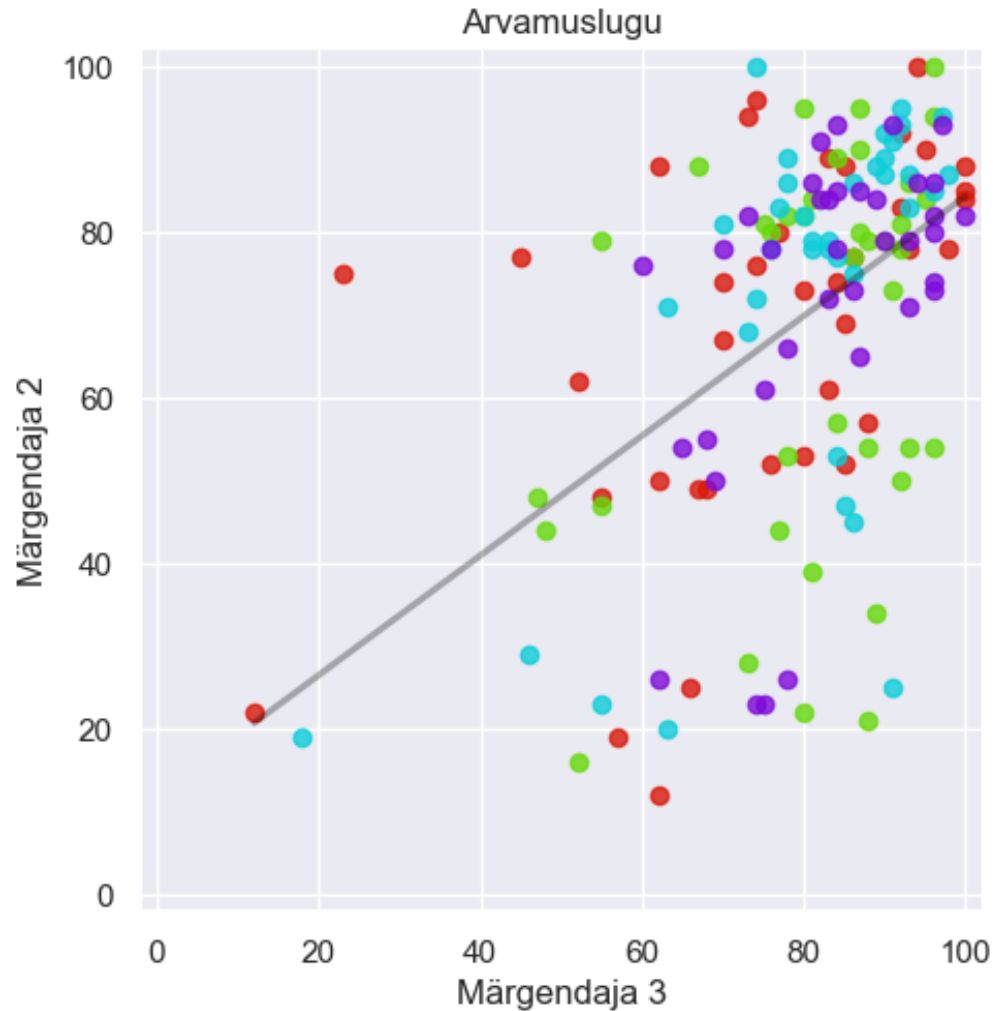
Varia – 30 tk

Kategooriad



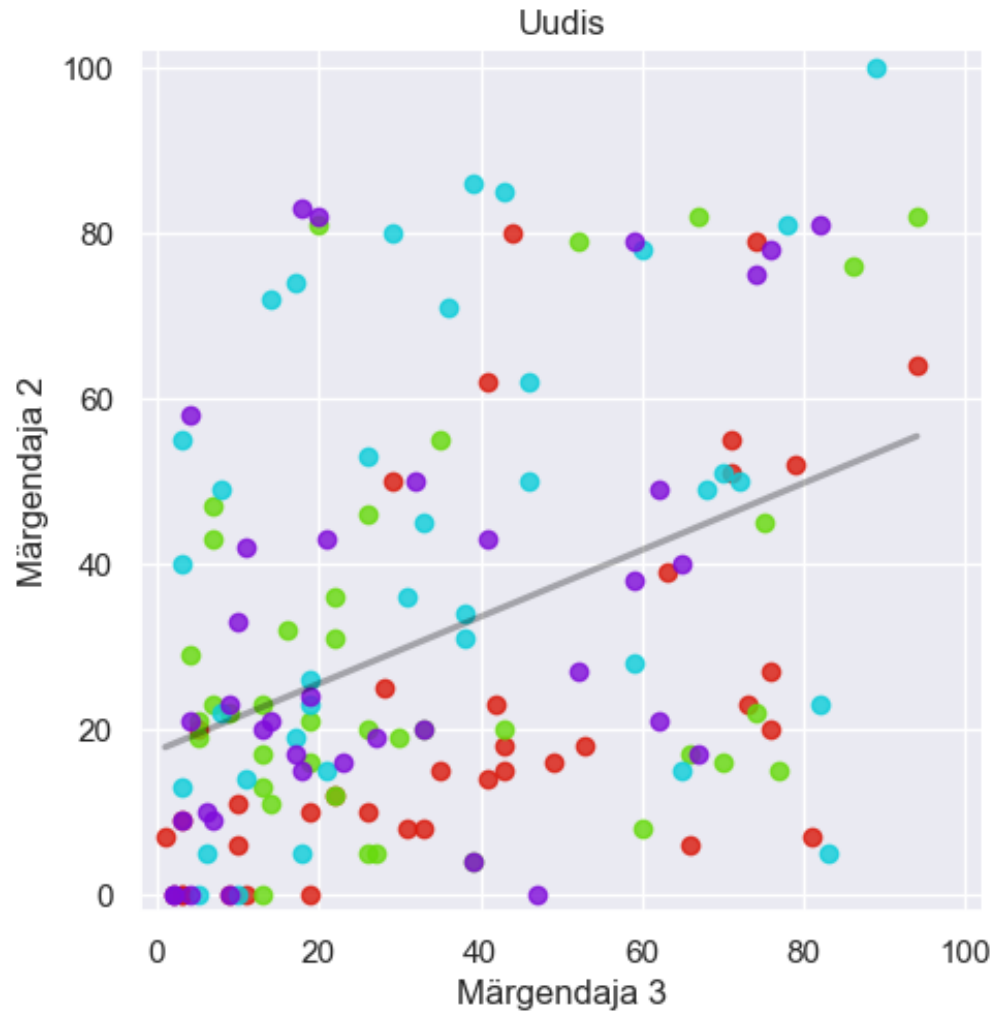
- Kategooria
- Arvamuslugu
 - Uudis
 - Arvustus
 - Reklaam
 - Veebist ajakirjandus
 - Informeeriv
 - Juhis
 - Sotsiaalne media
 - Ilukirjandus
 - Legaal
 - Varia

Arvamuslood



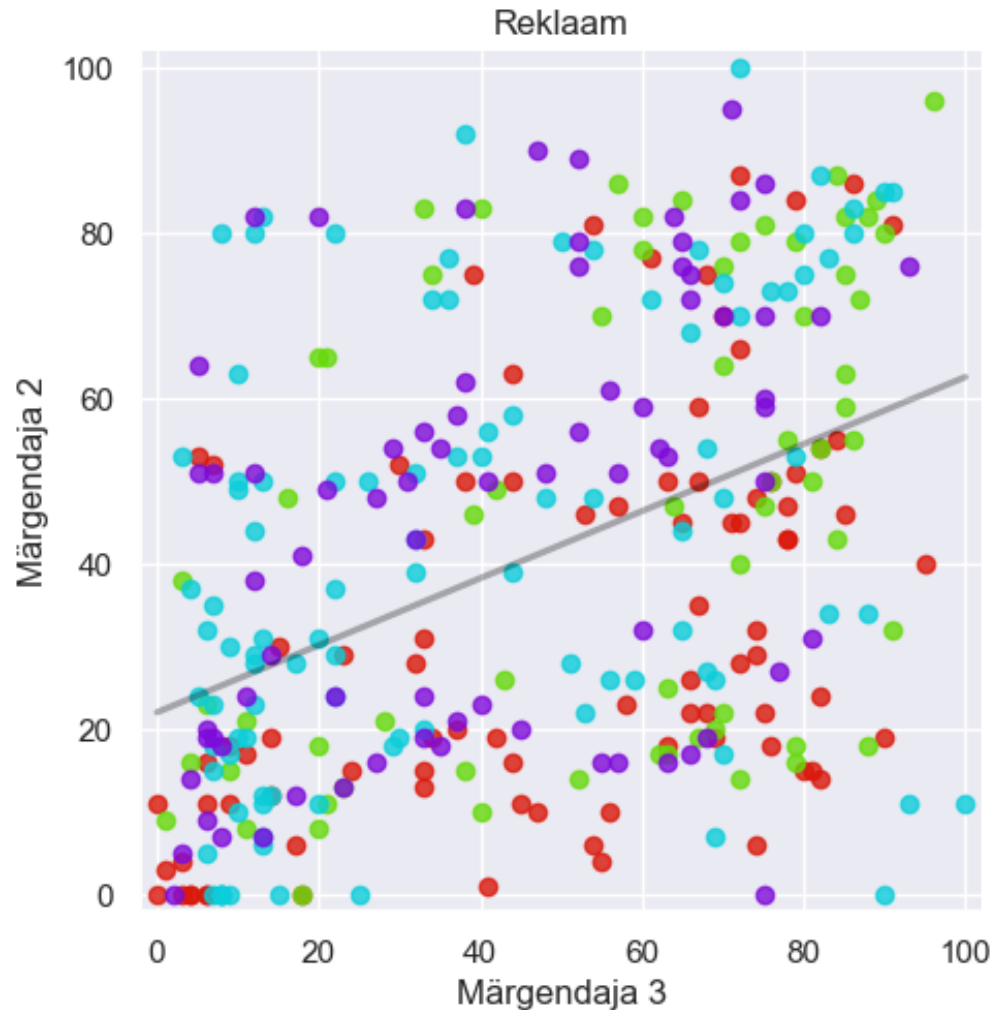
- 1 ja 2 – 0.2709 (Spearman)
0.3453 (Pearson)
- 1 ja 3 – 0.3016 (Spearman)
0.3297 (Pearson)
- 2 ja 3 – 0.4762 (Spearman)
0.5022 (Pearson)

Uudised



- 1 ja 2 – 0.2103 (Spearman)
0.2697 (Pearson)
- 1 ja 3 – 0.2408 (Spearman)
0.2600 (Pearson)
- 2 ja 3 – 0.4068 (Spearman)
0.4100 (Pearson)

Reklaam



- 1 ja 2 – 0.4065 (Spearman)
0.4063 (Pearson)
- 1 ja 3 – 0.3494 (Spearman)
0.3438 (Pearson)
- 2 ja 3 – 0.4329 (Spearman)
0.4305 (Pearson)

Objektiivse teksti näide

Parameetrid

Igast Ankarsrum Originals komplekti kuulub 7-liitrine roostevabast terasest kauss, taignarull, taignakraapija, taignakonks, vispel ja plastikust 3,5-liitrine segamiskauss. 5-aastase garantiiga, reguleeritava kiiruse ja taimeriga 800-vatine mootor on robustne ja vastupidav. Korraga on võimalik segada kuni 5 kg taigent. Enamikke tarvikuid on võimalik kasutada ühtemoodi nii vanematel kui ka uuematel mudelitel.

Sarnased tooted

(Ühendkorpus, Web 2017, id: 1639650)

Märgendajate hinnangud: 0, 0, 7

Subjektiivse teksti näide

Killuke helgust ja tänu igasse päeva

Aitäh! on justkui pisike päike, mis soovib tervitada, hinge soojendada ja tänada kõiki möödujaid. Me väärtustame värskust, käsitööd, tervislikku toitumist, perekonda ja koos kasvamist. Hoole ja armastusega valmistatud toit on meie kirg ning usume, et inimese toitumine on suuresti tema õnne nurgakiviks.

(Ühendkorpus, Web 2021, id: 5530781)

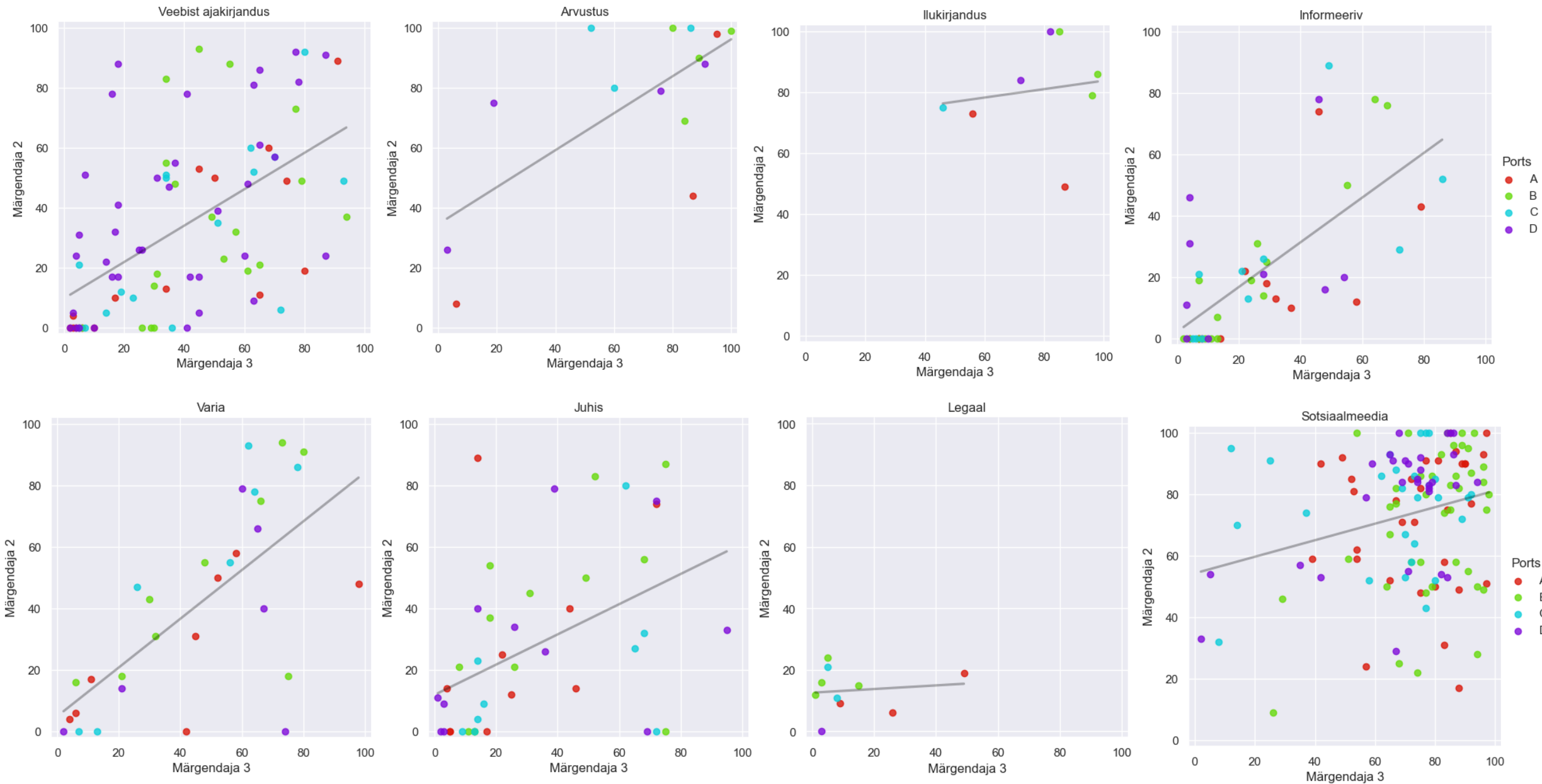
Märgendajate hinnangud: 100, 96, 96

Keskmise teksti näide

OÜ Almar Reisid pakub transporditeenuseid bussidega reisijateveoks uutes mugavates 8- 22-kohalistes mikrobussides. Meie firma pakub Teile laia teenuste ringi – vedude korraldamine, ekskursioonide, spordiürituste, pulmade ja ärikohtumiste teenindamine, aga ka lihtsalt väljasõidud loodusesse. Teenindame kliente nii Eesti territooriumil kui ka väljaspool (Euroopa, Skandinaavia, Venemaa). Austusega suhtumine igasse klienti ja individuaalne lähenemine igale reisile lubab meil olla kindel, et Te jääte meie teenindusega rahule ning meil on alati hea meel, et saame Teile kasulikud olla.

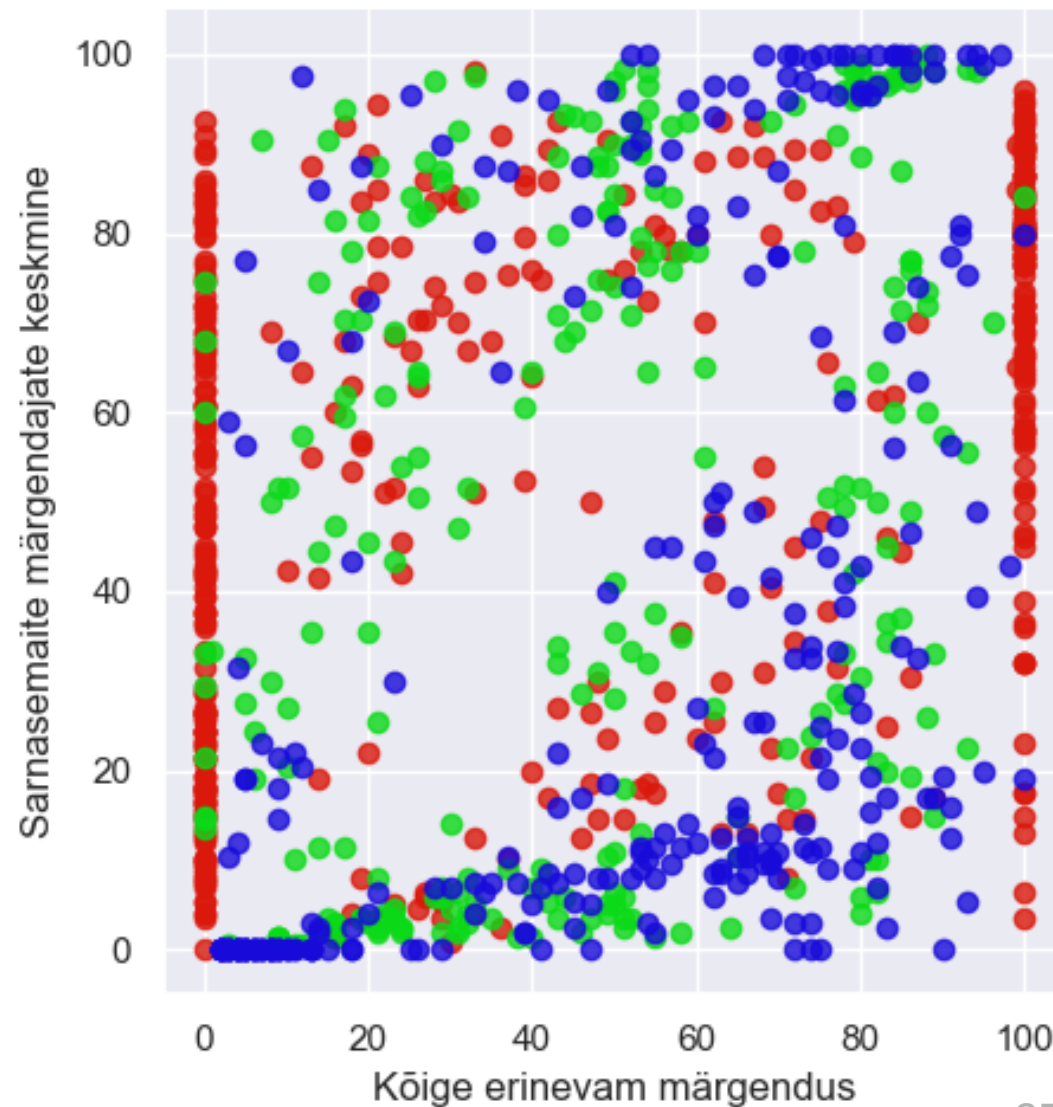
(Ühendkorpus, Web 2019, id: 4431419)

Märgendajate hinnangud: 54, 61, 56



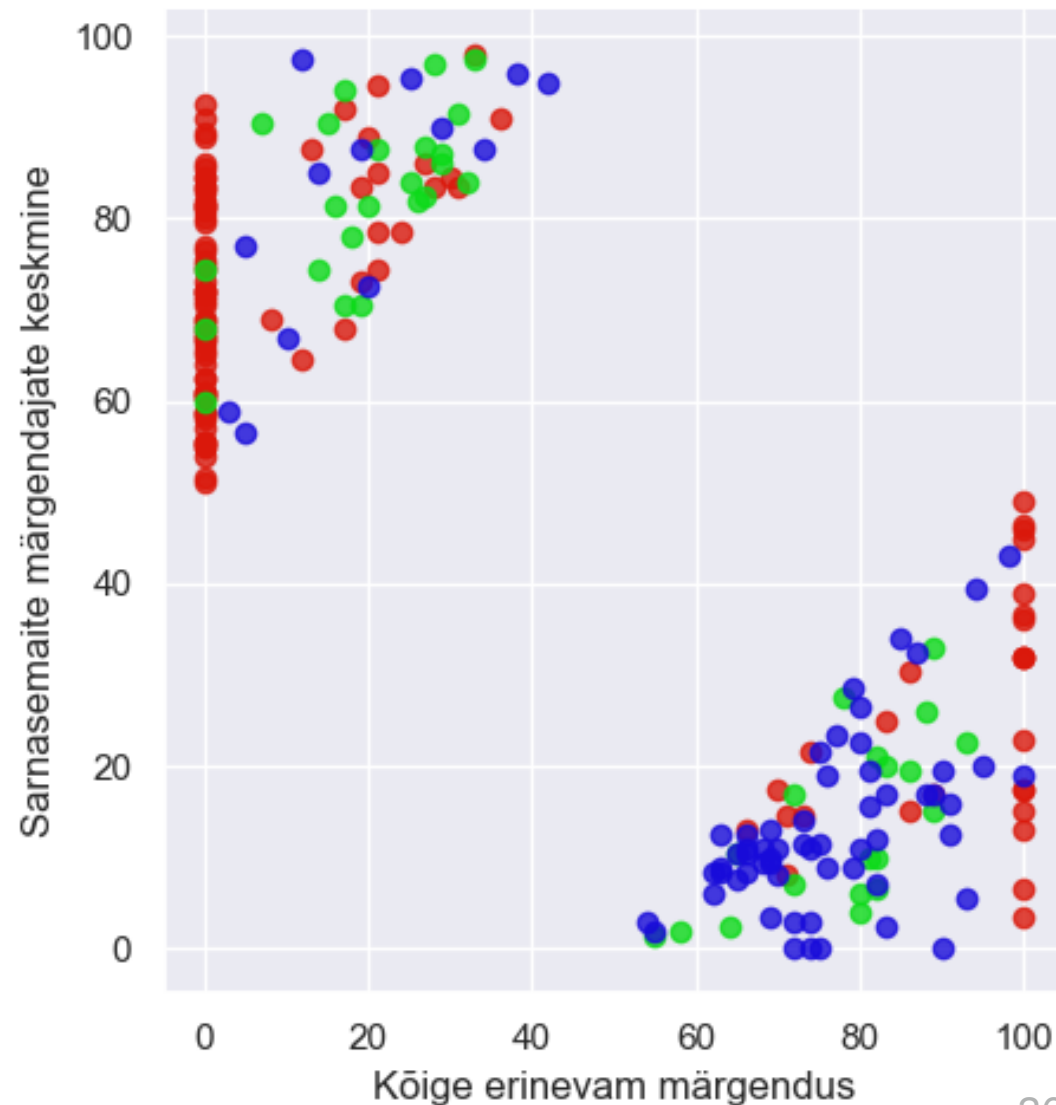
Märgendajatevahelised erinevused

- Enamus tekstidel erines ühe märgendaja hinnang teiste omadest
 - 1 – 446
 - 2 – 263
 - 3 – 291
- Erinevused eeldatavad
 - Subjektiivne ülesanne
- Suured erinevused huvitavad
 - Mida ja miks on märgendatud teistmoodi



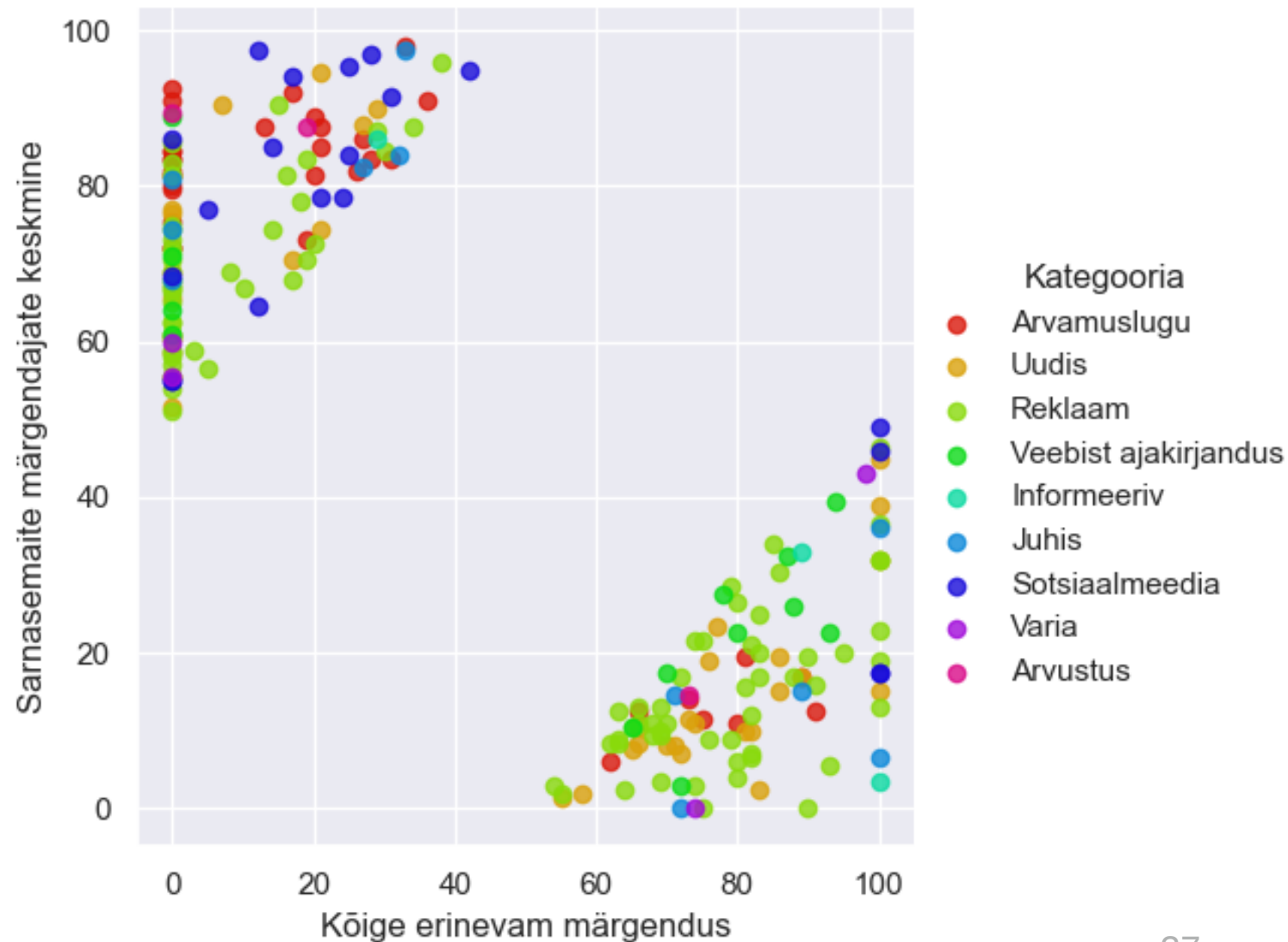
Märgendajatevahelised erinevused

- Erinevused, kus erinev tekst erines vähemalt 50 võrra
 - 1 – 107
 - 2 – 42
 - 3 – 71
- Kokku 220
- Tekste vaadates ei leidunud kindlaid ühisjooni
 - Ehk vead?
 - Samade märgendajatega kontrollmärgendus



Märgendajatevahelised erinevused

- Reklaam – 94 (28%)
- Uudis – 38 (25%)
- Arvamuslugu – 34 (23%)
- Sotsiaalmeedia – 20 (15%)
- Veebist ajakirjandus – 13 (15%)
- Juhis – 11 (26%)
- Varia – 4 (13%)
- Informeeriv – 3 (7%)
- Arvustus – 3 (21%)



Vastuolulise teksti näide

Huulte korrigeerimiseks kasutame Juvéderm täitesüste, mis jätavad kauni ja loomuliku tulemuse. Juvéderm® joonistab märkamatuult uuesti välja huulte kontuurid, muudab need täidlasemaks ning piiritleb need taas.

Kasutame JUVÉDERM® ULTRA® ja JUVÉDERM® SMILE®, mille mõju kestab 9-12 kuud

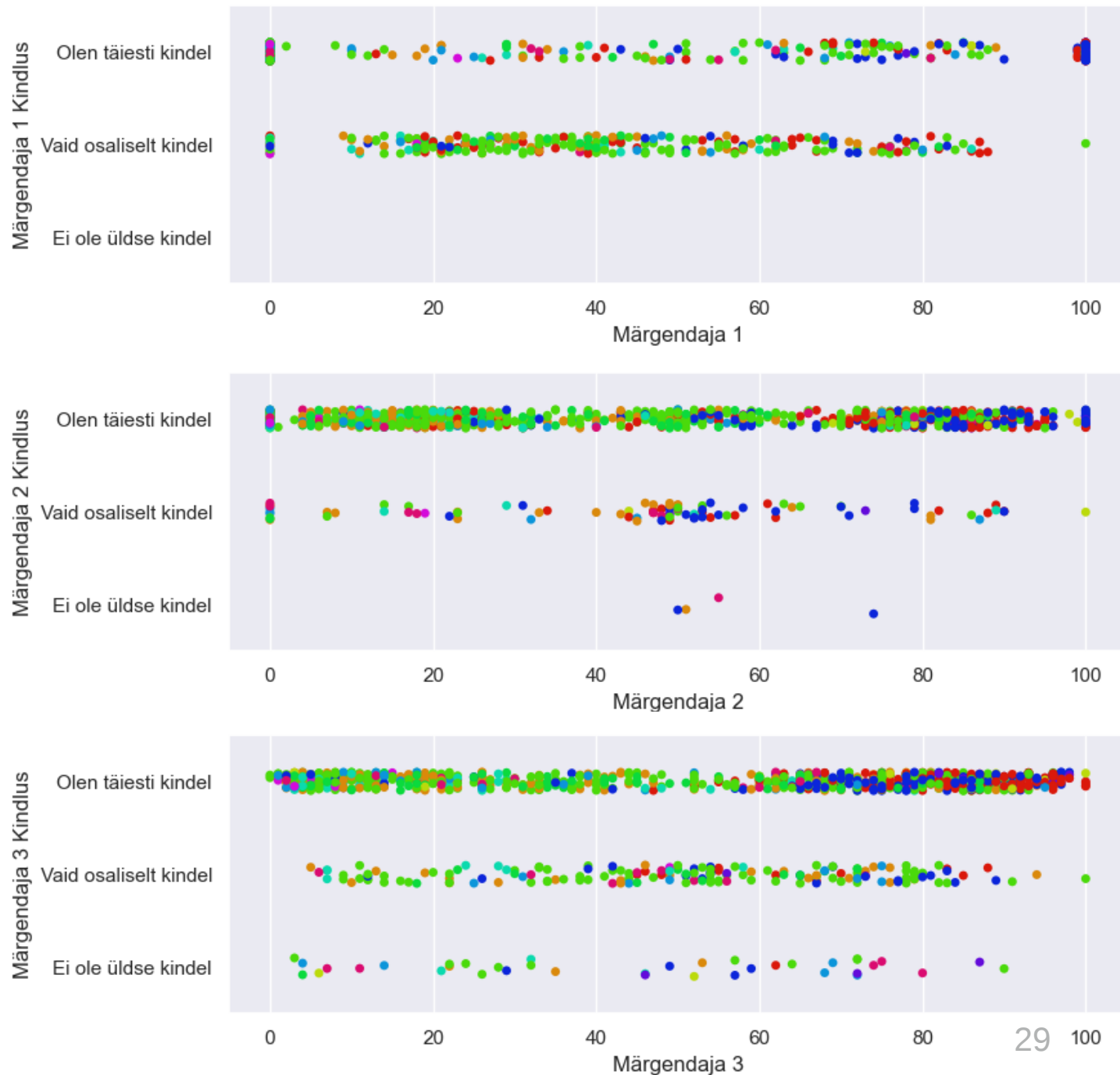
Uudisena pakume uue põlvkonna Juvéderm® VOLIFT® Vycross tehnoloogiaga valmistatud täitesüste, mis on vähem valulikud ning mille toime kestab 18-24 kuud.

(Ühendkorpus, Web 2021, id: 1135902)

Märgendajate hinnangud: 0, 89, 52

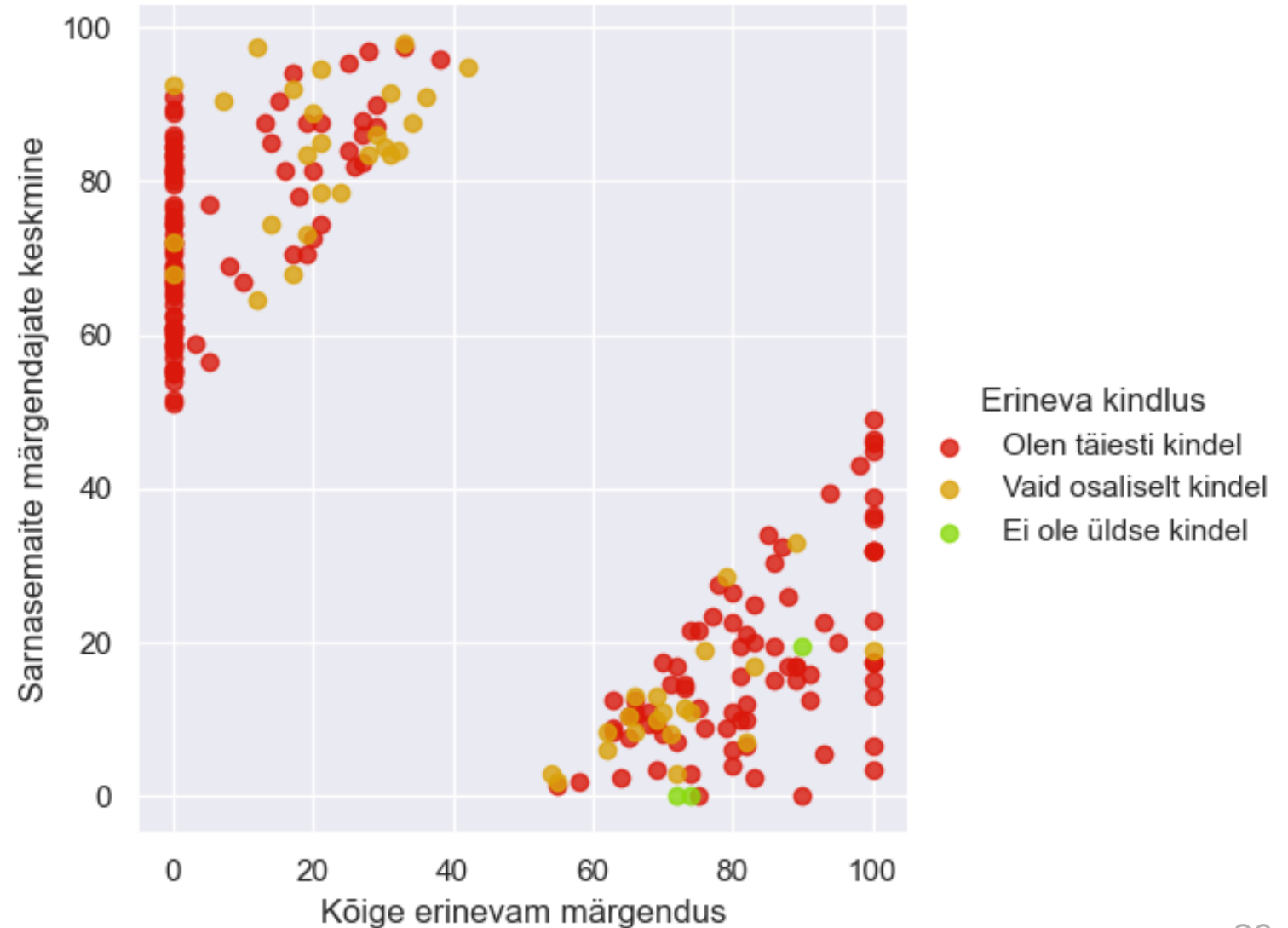
Kindlused

- Kolm taset
- Kohustuslik märkida
- Otstes kindlam, kui keskel
- Täiesti kindlad:
 - 1 – 738
 - 2 – 898
 - 3 – 798



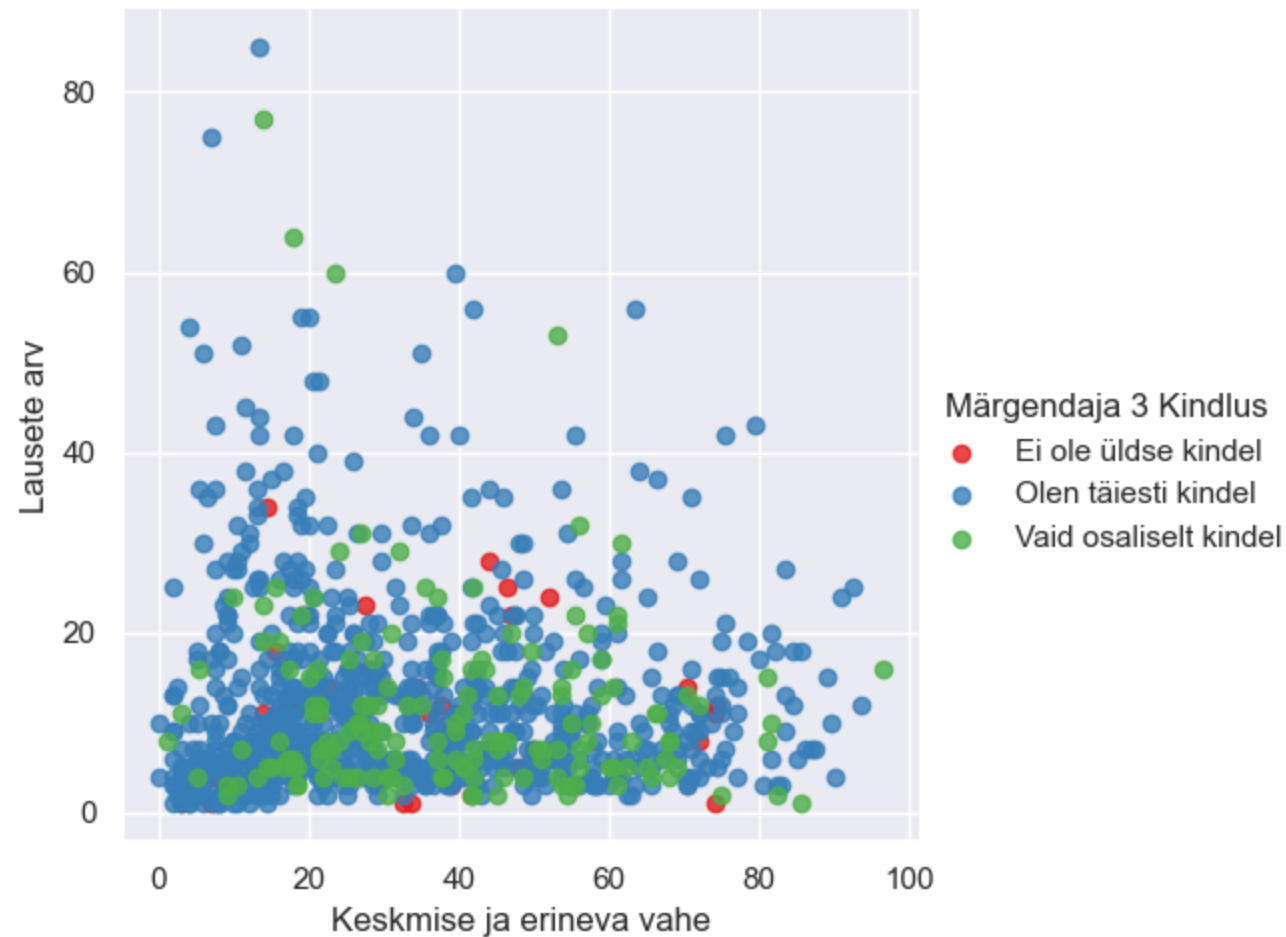
Kas ebakindlus tekitab erinevust?

- Täiesti kindel – 171
- Osaliselt – 46
- Ei ole – 3



Teksti pikkuse mõju

- Teksti pikkus ei mõjutanud märgendaja kindlust
- Leidus korrelatsioon lausete arvu ja erinevuse vahel
 - Spearman: 0.071
 - ($p = 0.0238$)
- Tähemärkide ja sõnade arv ei mõjutanud



Allikad

- Hennoste, Tiit; Gerson Klumpp; Helle Metslang 2022. Diskursusemarkerid ja pragmaatika. *Sissejuhatuseks. Keel ja Kirjandus*, 65 (1–2), 3–18.
- Liu, Bing 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 70–76.
- Přibáň, Pavel; Josef Steinberger 2022. Czech Dataset for Cross-lingual Subjectivity Classification. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1381–1391.
- Toprak, Cigdem; Iryna Gurevych 2009. Document Level Subjectivity Classification Experiments in DEFT'09 Challenge. *Proceedings of the DEFT 2009 Text Mining Challenge*, 89–97.
- Wiebe, Janyce; Theresa Wilson; Matthew Bell 2001. Identifying Collocations for Recognizing Opinions. *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*.
- Jiang, Zhengping, Anqi Liu, Benjamnin Van Durme 2024. Addressing the Binning Problem in Calibration Assessment through Scalar Annotations. *Transactions of the Association for Computational Linguistics* 2024; 12 120–136.
- Wiebe, Janyce. 2000. Learning Subjective Adjectives from Corpora. In *AAAI/IAAI*.
- Escouflaire, Louis, Antonin Descampe, Cédrick Fairon. 2024. Unveiling Subjectivity in Press Discourse: A Statistical and Qualitative Study of Manually Annotated Articles. *Discours* 34.
- Sanders, José. 1999. Degrees of subjectivity in epistemic modals and perspective representation. In Leon De Stadler & Christoph Eyrych (eds.), *Issues in Cognitive Linguistics*. Berlin, New York: DE GRUYTER MOUTON.



TARTU ÜLIKOOL

Täna kuulamast!