C11
Eva Lotta Lepp, Karin Rosenberg, Annaliisa Vask
Analyzing divorces in Estonia

Task 2. Business understanding (1 point)

Identifying your business goals

Abielud ja lahutused on suur osa meie kultuurist. Traditsiooniliselt on see seotud kooselamise ja laste saamise ning kasvatamisega. Eesti Statistikaamet on selleteemalisi andmeid kogunud pea 100 aastat. Uurides, kuidas erinevad näitajad (sugu, elukoht, sotsiaal-majanduslik seisund, ühiste alaealiste laste arv, mehe elukoht, vanuserühm) mõjutavad lahutuste arvu, võime leida huvitavaid seoseid. See informatsioon võib kasulik olla mitmetes valdkondades. See aitab riigil perepoliitikat paremini kujundada ning rakendada meetmeid, mis toetavad peret. Neid andmeid kasutavad ka näiteks pangad ja kindlustusseltsid.

Lisaks on aastate jooksul ühiskond palju muutunud ning abielu tähendus inimeste jaoks kohati ümber kujunenud. Väga põnev on uurida, kuidas see on mõjutanud abielu ja lahutuste trende.

Projekti lõppeesmärgiks on visualiseerida, milliseid seoseid me uuritavate andmete vahel leidsime. Saame analüüsida, kas need on eeldatavad või leidub selliseid seoseid, mida ei ootaks. Plaanis on luua Eesti kaart lahutuste statistikaga, lahutaja nais- ja meesprofiil ning veel eraldi tuua välja huvitavamad joonised, kui neid tekib.

Saadud tulemust on võimalik subjektiivselt hinnata. Kui kõik grupi liikmed on leitud tulemuste ning loodud kaardi ning joonistega rahul ning ei soovi midagi lisada, on eesmärk täidetud.

Assessing your situation

Töö jaoks kasutatavad vahendid:

- andmetega tegelevad projekti kolm autorit
- statistikaameti kodulehel asuvad abielude ja abielulahutuste andmed
- Python, Jupyter Notebook + laiendused, GitHub

Tööd tuleb esitleda 15. detsembril kell 14:00 TÜ Delta keskuses. Selleks on vaja disainida plakat, mille juures kõik töö autorid seda esitleda saavad.

Kõik kasutatavad andmed on statistikaameti kodulehel avalikult tasuta kättesaadavad. Toorandmete saamiseks on vaja esitada taotlus Eesti Statistikaametile, kuna tegu on konfidentsiaalsete isikuandmetega. Nendele ligipääsemiseks tegime ka päringu. Andmete töötlemine ja menetlemine on tasuline. Võimalik, et sellega kaasnev rahaline panus projekti ei tasuks ennast selle koolitöö raames ära ning seetõttu kasutame arvatavasti avalikult kättesaadavaid andmeid, kuna projekt on võimalik lõpuni viia ka ilma toorandmete olemasoluta.

Töös ei kasutata selgitust vajavat terminoloogiat.

Defining your data-mining goals

Enne andmete põhjal kokkuvõtvate jooniste tegemist on vaja neid veidi ümber teha. Avalikud andmed ei ole toorandmed ning on statistikaameti poolt juba tabeliteks grupeeritud. Kõigepealt leidsime, milliseid saadaval olevaid tabeleid oleks üldse mõistlik kasutada. Suur osa tööle kuluvast ajast läheb nende andmete üksteisega vastavusse viimisele ja puhastamisele. Hilisemaks eesmärgiks on saadud andmete järgi jooniste ja graafikute loomine.

Andmeanalüüs on olnud edukas, kui oleme algsed tabelid endale sobivaks modifitseerinud ja nende põhjal piisavalt seoseid leidnud, et koostada vajalikud joonised. Kui kõik grupi liikmed on tulemusega rahul, võib andmeanalüüsi edukaks lugeda.

Task 3. Data understanding (2 points)

Gathering data:
Meie projekti eesmärkide täitmiseks on meil vaja abielu lahutanud inimeste andmeid mõlema abielu osapoole kohta. Selleks kasutame andmestikke, kust saame valitud parameetrite info eraldi nii naise kui mehe jaoks ning projektis analüüsitakse läbivalt kahte sugu eraldi. Parameetrid, mida projektis vaatame:

- sugu (eraldav tunnus)
- elukoht
- sotsiaalmajanduslik seisund
- ühiste alaealiste laste arv
- mehe elukoht
- vanuserühm

Enamus andmetest on tekstilisel kujul ning kategooriatesse jaotuvad. Samamoodi on ka näiteks vanused kättesaadavad kategooriatena, mitte eraldi täisarvudena.

Verify data availability:
Kõik meie kasutatavad andmed pärinevad Eesti Statistika leheküljelt, kust erinevad andmestikud on ka vabalt allalaetavad. Oleme ka juba kirjutanud Statistikaametisse, et saada ligipääs toorandmetele, kuna meie jaoks oleks kõige ideaalsem viis saada andmed lahutanud paaride kaupa. Kuna need on aga konfidentsiaalsed andmed, on sellega seotud erinevad menetlustoimingud ning ülejäänud raport on koostatud eeldusel, et töötleme ise andmeid meile sobivale kujule. Andmed on allalaetavad csv vormis, mille saame kohe Jupyter Notebook`i tabelitesse sisse lugeda.

Define selection criteria:
Oma töös kasutame Eesti Statistika leheküljelt pärinevaid tabeleid:
RV37U: LAHUTANUD SOTSIAAL-MAJANDUSLIKU SEISUNDI, ELUKOHA, SOO JA VANUSERÜHMA JÄRGI, HALDUSJAOTUS SEISUGA 01.01.2018;
RV35U: LAHUTUSED ÜHISTE ALAEALISTE LASTE ARVU, MEHE ELUKOHA JA ASUSTUSÜKSUSE LIIGI JÄRGI, HALDUSJAOTUS SEISUGA 01.01.2018;
R33U: LAHUTANUD ABIELU KESTUSE, ELUKOHA, SOO JA VANUSERÜHMA JÄRGI, HALDUSJAOTUS SEISUGA 01.01.2018;

Plaanime andmeid ühendada haldusjaotuse järgi nii palju, kui võimalik. Kasutame tabelitest kõiki parameetreid.

Describing data:
Statistikaameti andmepäringu juures on meie töö jaoks sobiv märkida “tabel - sorditud”, kuna selles valikus koondatakse kõik samade väärtustega isendid. See annab meile võimaluse read kerge vaevaga eraldada ning seega tekitada olukord, kus iga rida vastab ühele kindlale lahutanud isikule.

Nagu eelnevalt mainitud on omandatud andmestik enamasti kategooriate kaupa jagatud, mis teeb meie jaoks analüüsi mingis mõttes mugavaks, kuna ei pea numbrilisi andmeid isendite rohkuse tõttu ise grupeerima. Samas piirab see meie võimalusi andmeid analüüsida ning ei saa kasutada pidevaid graafikuid.
Leiame, et erinevate parameetrite hulk on piisav, et nende põhjal kaardistada n-ö Eesti tavalise lahutaja profiil, sealhulgas leida, millised parameetrid teevad abielulahutuse tõenäolisemaks või vastupidi, milliste parameetrite millised väärtused teevad lahutuse ebatõenäoliseks.

Exploring data:
Järgnevalt loetleme üles tunnused, mille põhjal hakkame Eesti lahutusi kaardistama:
Kategoorilised tunnused:

- Vanuserühm - erinevad vanuserühmad, noorim 15-19 ning vanim 75 ja vanemad
- Asustusüksus - Eesti, väljaspool Eestit, linnaline ja väikelinnaline, maaline
- Mehe elukoht - Eesti 15 maakonda, kus abielu sõlmimise hetkel mees oli sisse kirjutatud
- Sotsiaalmajanduslik seisund - 10 kategooriat inimese positsiooni kohta ühiskonnas, näiteks töötav, töötu, üliõpilane jne
- Sugu - naine või mees, mille põhjal analüüsime ka lahutajate profiile eraldi, kuna paaride kaupa andmete põhjal analüüsi teha ei saa.

Arvulised tunnused:

- Abielu kestus - mitu aastat oli paar enne lahutust abielus
- Ühiste alaealiste laste arv - mitu last on abielus oleval paaril

Kuna andmed on statistikaameti poolt juba eeltöödeldud, siis on näiteks kõik puuduvad väärtused lahterdatud eraldi kategooriasse, näiteks “Elukoht teadmata” jne.
Internetist saadavatel csv failidel on ka iga kategooria järgselt n-ö summa rida, mida peab andmetöötlusel kindlasti jälgima, et see andmestikust välja võtta, vastasel juhul saame dubleeritud andmed.
Andmekvaliteedi probleeme andmestike peal me ei täheldanud, peaasjalikult seetõttu, et andmed on juba statistikaameti poolt töödeldud visuaalselt (tabeli vormis) esitatavale kujule ning see eeldab korralikku andmetöötlust.

Verifying data quality
Esmase pilgu järgi andmetele tundub, et on mõned parameetrid, mida ei ole mõistlik töös kasutada. Üks neist on näiteks statistikaameti tabeli RV37U tunnust “sotsiaalmajanduslik seisund”, mis pakuks küll põnevat eristamise võimalust, kuid lähemal uurimisel selgub, et

ainult 16.6% kannetest on vastav väli täidetud. Seega kuigi see võiks pakkuda väga huvitavat infot lahutajate kohta, on andmeid ilmselt liiga vähe ning nende põhjal üldistusi teha oleks vale. Samuti on nimetatud andmestik ainult aasta 2017 kohta. Seega tundub, et antud parameetri peame enda kasutatavate andmete loetelust välja jätma.
Samas elukoha andmed on peaaegu sajaprotsendiliselt kajastatud, mis toetab ühte meie töö eesmärki - teha valmis visuaalne Eesti kaart lahutustega ning lahutuste muutusega ajas.

Teisi suuri andmekvaliteedi probleeme me ei tuvastanud.

Task 4. Planning your project (0.5 points)

[ajakulu tundides, ülesande lahendaja]
Projektis kasutame juba eelnevalt mainitud Jupyter Notebook`i, keeleks on Python oma teekidega ja lahutuste kaardistamisel võib vaja minna ka pilditöötlusprogrammi nt Photoshop. Grupikaaslastega suhtlemiseks kasutame Facebooki, Zoomi ja saame kokku päriselus.

Task 1: Andmete taotlemine- eeldatav meiepoolne ajakulu kokku 0.5h

- ➢ Esitame Statistikaametile taotluse [0.25, Karin]
- ➢ Olenevalt toimingu maksumusest ja menetluse ajast, otsustame, kas jääme Statistikaameti lehel olevate andmete juurde või tegeleme toorandmetega [0.25, koos]

Task 2: Andmete ettevalmistamine ja puhastamine- eeldatav ajakulu 20h

- ➢ Sobivate tabelite allalaadimine csv-failina ja nende sisselugemine Jupyter Notebook`i [1x3, kõik eraldi]
- ➢ Andmete kokkusobitamine [10, Annaliisa]
- ➢ Andmete puhastamine [7, Eva Lotta]

Task 3: Meeslahutaja profiili kujutamine - eeldatav ajakulu kokku 15h

- ➢ Ettevalmistatud andmetest arvutuste ja joonise tegemiseks vajaliku info koondamine [5, Karin]
- ➢ Vajalike arvutuste tegemine, kasutades arvutusliku statistika erinevaid meetodeid [5, Karin]
- ➢ Tulemuste visualiseerimine [5, Karin]

Task 4: Naislahutaja profiili kujutamine - eeldatav ajakulu kokku 15h

- ➢ Ettevalmistatud andmetest arvutuste ja joonise tegemiseks vajaliku info koondamine [5, Eva Lotta]

- ➢ Vajalike arvutuste tegemine, kasutades arvutusliku statistika erinevaid meetodeid [5, Eva Lotta]
- ➢ Tulemuste visualiseerimine [5, Eva Lotta]

Task 5: Lahutuste kaardistamine Eesti kaardil - eeldatav ajakulu kokku 20h

- ➢ Andmete ettevalmistamine asukoha järgi [1, Eva Lotta]
- ➢ Arvutuste tegemine [7, Karin]
- ➢ Andmete visualiseerimine [7, Annaliisa]
- ➢ Kaardi kujundamine ja kokkupanemine saadud visuaalidest ja arvutustest [5, Annaliisa]

Task 6: Ruum põnevate avastuste jaoks- eeldatav ajakulu kokku 15h

- ➢ Arutame mõtteid, mis on potentsiaalselt tekkinud andmetega tegelemise käigus ja otsustame, kas need on piisavalt huvitavad ja informatiivsed, et neid meie projekti lisada [0.5, koos]
- ➢ …

Task 7: Plakati kujundamine- eeldatav ajakulu kokku 7.5-8.5h

- ➢ Arutelu, mida kujutada esitletaval plakatil [0.5, koos]
- ➢ Plakatile selgituste/teksti lisamine [2, Annaliisa]
- ➢ Plakati kujundamine ja vajadusel tehtud graafikute ilustamine [3, Eva Lotta]
- ➢ Plakati kokkusobitamine ja vajadustel paranduste/muudatuste tegemine, kui mõni grupiliige pole rahul [1-2, koos]
- ➢ Esitluseks ettevalmistamine [1, koos]