Liisa Jullinen

Projekti raport

Eesmärgid

Antud projekti eesmärgiks oli leida iseseisvalt sobiv andmeallikas, andmeid vajadusel töödelda ning tranformeerida (antud juhul *parquet*-failiks), laadida andmestik *flat*-andmebaasi ja visualiseerida andmeid Superset'i abil. Sisuliseks eesmärgiks oli luua juhtpaneel, mis visualiseeriks Eestis toimunud inimkannatanutega liiklusõnnetuste statistikat ja aitaks luua sellesse huvipakkuvaid sissevaateid.

Andmeallikad

Andmeallikaks oli valitud Eesti Avaandmetest pärinev andmestik „Inimkannatanutega liiklusõnnetuste andmed". Tegu on Politsei- ja Piirivalveameti poolt kogutud andmetega, mille haldusega tegeleb Transpordiamet ning mis koondab andmeid inimkannatanutega liiklusõnnestustest Eestis ajavahemikul 2011-2025.

Andmestik on saadaval mitmetes erinevates formaatides, seal hulgas siin töös kasutatud .csv formaadis ning metaandmed formaadis .json. Kokku oli andmestikus 17370 rida ja 54 veergu ning andmestik hõlmas erinevaid andmetüüpe (numbriline, kategooriline, tekstiline, kuupäev ja kellaaeg).

- Täpsem andmestiku kirjeldus leidub esitatud kodutöös nr 1.
- Link andmetele: https://avaandmed.eesti.ee/datasets/inimkannatanutega-liiklusonnetuste-andmed

ETL protsess

ETL-protsessis kasutatav Superseti rakendus käivitati avalikult kättesaadava Docker'i konteineri kaudu.

Kuigi (näiteks Pythoni pakettide) versioonikonfliktide vältimiseks oleks mõistlik ka selle tarbeks ehitada eraldi Dockeri konteiner, ei näinud ma antud projekti puhul selleks vajadust (kuna andmeanalüüsi osa ei olnud just kuigi keerukas ega hõlmanud kuigipalju erinevaid pakette) ja kogu *transform* osa teostati ühe Pythoni skriptiga.

Kuna andmestiku allalaadimine otselingilt ei olnud võimalik (allalaadimiseks oli vaja manuaalselt nõustuda kasutustingimustega), laeti andmed esmalt alla käsitsi .csv-failina (seega on modifitseerimise koodis kasutatud kõvakettale salvestatud csv-faili, mitte otselinki).

Seejärel loeti allalaetud andmefail sisse, töödeldi (nt. toimumisaeg muudeti kuupäevaks, mõningaid muutujaid grupeeriti ümber, transformeeriti X ja Y koordinaadid õigesse koordinaatsüsteemi) ja transformeeriti andmefaili (parquet-failiks).

- Superseti käivitamise ja kohandamise info on leitav projekti GitHubi repositooriumist: https://github.com/liisa-j/DE_project/tree/main/superset_build

- Andmete sisselugemise, töötlemise ja transformeerimise kood on leitav siit: https://github.com/liisa-j/DE_project/blob/main/downloaddata.py

Visualisatsioonid

Visualisatsioonide loomiseks kasutati aine nõuetest lähtuvalt rakendust Apache Superset.

Eelmises punktis kirjeldatud *parquet*-fail laaditi *flat*-andmebaasi (Duck.db) ja andmestiku visualiseerimiseks kasutati Superseti rakendust.

- Selle juhtpaneeli seaduse täpsem kood on leitav siit (laetud alla Supersetist): https://github.com/liisa-j/DE_project/blob/main/dashboard_export.zip
- Supersetis loodud visualisatsioonidest koosneva juhtpaneeli jpg-formaati salvestatuna leiab siit: https://github.com/liisa-j/DE_project/blob/main/dashboard_liiklus.jpg

Raskused ja lahendused

Suurim raskus antud projekti puhul kaasnes Superseti kasutamisega, täpsemalt sellega seonduvate piirangutega. Kuigi tegu on väga intuitiivse ja lihtsa keskkonnaga, ei luba Superset andmestikku lihtsasti modifitseerida (nt. vaheldada laia ja pikka andmeformaati).

Lisaks tekkis probleem andmestiku visualiseerimisega kaardil. Superset kasutab lahendust Mapbox, mille abil saab visualiseerida ruumilise elemendiga andmeid (nt. saab kasutada GPSi süsteemi X ja Y koordinaate, ilmselt on võimalik lisaks punktidele visualiseerida ka jooni ja polügone), kuid aluskaardi kasutamiseks peab tegema tasulise (või vähemalt krediitkaardi andmeid nõudva) kasutajakonto. Seetõttu jäid huvitavamad ruumiandmete visualisatsioonid ka antud tööst välja ja jätsin töösse ainult illustratiivsena mõjuva kaardi valgel taustal. See on iseenesest kahetsusväärne, sest teiste vahenditega (nt. Python'i Geopandas või R'i Leaflet) võimalik väga lihtsasti luua väga huvitavaid ruumiandmete (või ruumilise elemendiga andmestike visualisatsioone), eriti arvestades, et tegelikult on tasuta kaardiandmeid väga palju (nt. Eesti kohta käivad andmed Maa- ja Ruumimetist ja kogu maailma kohta käivad OpenStreetMaps'ist).

Kuna on ka muid võimalusi juhtpaneelide loomiseks (nt. Dash Pythonile ja Shiny R'ile) kasutaksin mainitud probleemi lahendusena edaspidi lihtsalt mõnda muud platvormi. Mainin ka ära, et sellisel juhul oleksin ka andmete töötlemise ja visualisatsioonide jaoks loonud eraldi Pythoni konteineri, et vältida pakettide versioonikonflikte, mis tuleksid suurema hulga (ja veidi spetsiifilisemate) andmeanalüüsi pakettide (nt. Pythoni implementatsioon node2vec algoritmist) kasutamisel kindlasti ette.