DAT 2000: Obligatorisk oppgave V24

Oversikt

I denne oppgaven skal vi sette sammen noen av de ulike bitene vi har jobbet med til en løsning. Oppgaven skal gjennomføres og leveres som et **privat** GitHub-repository.

Løses alene eller i grupper på maks 4 studenter. Frist kommer senere.

Denne versjonen inneholder kun første oppgave i oblig, dokumentet blir oppdatert litt senere med resten.

1. Vi skal lese inn et datasett med Polars (øvelsene 29. Januar), og så skal vi bearbeide det litt.
2. Så skal vi sette opp PostgreSQL i Docker (øvelsene 8. Januar) og laste opp datasettet vi nettopp bearbeidet (øvelsene 22. Januar).
3. Så skal vi kjøre noen spørringer mot PostgreSQL (øvelsene 8. Januar) for å få svar på noen spørsmål.
4. Så skal vi gjøre dette til en RESTful HTTP / JSON tjeneste med REST (øvelsene 5 februar). Vi skal bruke verktøyet cURL til å sjekke om tjenesten oppfører seg riktig.
5. Vi skal bruke requests-biblioteket i Python til å skrive noen tester (øvelsene 8. januar og øvelsene 5 februar).
6. Til slutt skal vi sette opp en Github-action som kjører testene (øvelsene 8. januar)

En siste valgfri oppgave blir å Dockerisere tjenesten, og å sette opp både databasen og tjenesten med Docker Compose (senere øvelser).

Oppsett

Vi skal bruke datasettet “Teknisk Kjøretøyinformasjon” fra Statens Vegvesen. Dette datasettet inneholder informasjon om kjøretøyene i Norge. Datasettet finnes her: https://data.norge.no/datasets/a8533876-cca7-4417-90be-b368f7d9542c
Datasettet kjoretoyinfo.csv er imidlertid litt stort (3,7GB!) og det var vrient å parse, så jeg har lagt ut noen forhåndsprosesserte varianter:

- Begrenset variant for årene fra og med år 2020. Bruk dette hvis du er usikker på hva maskinen din klarer: https://drive.google.com/file/d/17dcdXNZnk90uXXDJXxcvA29Dlv9ueShJ/view?usp=sharing
- Begrenset variant for årene fra og med år 2000. Litt mer data her. https://drive.google.com/file/d/1w6OXjP-1JKUsZNf7wh7iTczPhAroSur3/view?usp=sharing

Det ligger imidlertid noen støtte-data (oppslag av koder og lignende) som det vil være nyttig å laste ned fra Felles Datakatalog.

Oppgave 1: Polars

Innlesning og lett bearbeiding av datasettet (**10 poeng**):
Begynn med å gjøre dette i en Jupyter Notebokk (**prepp.ipynb**), dette skal vi senere konvertere til et Python-skript: **prepp.py**

1. Les inn **kjoretoyinfo_fra_[putt_inn_år].parquet**. Bruk pl.scan_parquet.
2. "tekn_reg_f_g_n" er dato for førstegangsregistrering i Norge. Konverter denne kolonnen til en datetime. Gjør tilsvarende for "neste_pkk" og "tekn_reg_eier_dato".
3. Join inn navnet på fargen, det vil si, join "tekn_farge" med "kode" i fargekode.csv. Pass på at du ikke mister noen biler som vi ikke vet fargen på.
4. Lag en kolonne "elbil" som er sann hvis drivstofftypen er kun elektrisk, false ellers.
5. Hent ut bare kolonnene:
    - tekn_reg_f_g_n: Dato for førstegangsregistrering i Norge
    - tekn_reg_eier_dato: Dato bilen ble registrert på nåværende eier.
    - tekn_aksler_drift: To aksler er firehjulsdrift
    - tekn_merke
    - tekn_modell
    - tekn_drivstoff
    - tekn_neste_pkk: Dato for neste EU-kontroll
    - Farge på bilen fra fargekode.csv
    - tekn_drivstoff: Drivstofftypen
    - Kolonne som indikerer om bilen er en elbil.
6. Skriv hele datasettet til en fil: **kjoretoyinfo_preppet.parquet**

Ti kjappe analyser for å trene litt (**10 poeng**):
Gjør disse i sin egen notebook: **analyse.inpynb** Les gjerne inn **kjoretoyinfo_preppet.parquet** fra forrige oppgave, og bruk denne som utgangspunkt.

7. Hvor mange elbiler (drivstofftype 5) ble førstegangsregistrert i 2022?
8. Hvor mange prosent av personbilene som ble solgt i 2022 var elbiler?
9. Hvilken bilmodell var den mest populære i 2022?
10. Hvor mange gule kjøretøy ble det solgt i Mai 2022?
11. Hvor stor andel av personbilene bilene som selges i Norge har firehjulstrekk?
12. Hvilken måned i året førstegangsregistreres det flest biler i Norge?
13. Hva var den mest populære fargen for biler som var førstegangsregistrert i Norge i hvert år?
14. Hvilken farge var den mest populære for traktorer i 2022? (se teknisk-kode.csv)
15. Hvilke bilmerker var de fem mest populære i 2022?
16. Hva var den mest populære fargen for de fem mest populære bilmerkene i 2022 (førstegangsregistrert 2022)?

En valgfri utfordring (belønning i himmelen):

17. Klarer du å gjøre det samme for drivstofftype.csv ("tekn_drivstoff") som du gjorde for fargen på bilen? her må du bruke str.split og explode, og joine, og deretter gjøre group by hvor du grupperer på de andre kolonnene.