INF161 Prosjekt

Trafikkdata

Etter at jeg har lest inn **'Trafikkdata.csv'** velger jeg å fjerne kolonnene 'Trafikkregistreringspunkt', 'Navn' og 'Vegreferanse' fordi alle verdiene i hver kolonne er like (bevist i ipynb filen).

Deretter velger jeg å fjerne alle radene som ikke inneholder verdien «Total» i kolonnen 'Felt', fordi det disse verdiene forteller meg er bare hvilken retning folk sykler i, og det er ikke informasjon jeg føler er nødvendig å vite videre. Det var også i nesten alle de andre kolonnene som ble repeter unødvendig mange ganger. Da velger jeg så å fjerne hele 'Felt' kolonnen og det er fordi den nå bare inneholder verdien «Total», så er ikke nødvendig å ha den med lenger. Jeg velger også å fjerne kolonnene 'Fra' og 'Til' fordi de forteller oss det samme som kolonnene Dato, Fra tidspunkt og Til tidspunkt.

Vær data

Først leser jeg inn all informasjonen om været fra 2015-2023 og legger dette inn i en tabell. Jeg velger å ikke lese inn data fra tidligere enn 2015 fordi det ikke finnes trafikkdata før 2015.

Merger trafikk dataen og vær dataen sammen

Nå som jeg har fikset begge datasettene slik jeg vil ha de, merger jeg disse to sammen til et datasett. Dette gjør jeg ved å legge til en ny kolonne i begge datasettene som heter 'Datetime'. Denne kolonnen består av datoen og tidspunktet en rad skal gjelde fra. Siden hver rad i vær dataen bare har ti minutter forskjell, resampler jeg vær dataen til å være per time istedenfor, slik at det matcher trafikkdataen. Når jeg gjør dette velger jeg å summere verdiene i 'Solskinstid' fordi solskinstiden viser hvor mange minutter det er sol på de ti minuttene. Da gir det mer mening å legge disse verdiene sammen enn å ta gjennomsnittet. Så nå kan man se hvor mange minutter det er sol i løpet av en time.

Resten av kolonnene tar jeg gjennomsnittet, fordi jeg har sett på yr at de gjør det med blant annet Vindstyrke og Vindkast.

Siden trafikkdata ikke inneholder noe informasjon før 2015-07-16 15:00, filtrerer jeg datasettet slik at det ikke inneholder noe informasjon før dette tidspunktet. Jeg velger også og filtrerte det slik at datasettet ikke inneholder noe fra 2023, fordi det er det året vi skal predikere.

Nå som jeg har laget kolonnen 'Datetime' er det ikke vits å ha kolonnene 'Dato', 'Fra tidspunkt' eller 'Til tidspunkt' så jeg fjerner disse. Siden vi har fått vite at vær dataen bruker koden 9999.99 for manglende data, velger jeg å erstatte denne koden med **NaN**. Jeg printer deretter summen av hvor mange rader som inneholder NaN verdier i hver kolonne. Da ser jeg at kolonnen 'Relativ luftfuktighet' inneholder veldig mange NaN verdier og velger derfor å fjerne denne kolonnen. Finner også ut av at det er ca 3% av radene i datasettet som inneholder NaN, og velger å fjerne alle disse siden det ikke er en så stor del av datasettet.

I alle vær relaterte kolonner er det noen ekstremt høye verdier (printet ut max verdiene for å vise). Dett kan skyldes en feil i målingen av dataen. Derfor lager jeg grenser på hva som er en gyldig verdi for hver kolonne. Hva den grensen skal være har jeg funnet ut av å google meg frem til hva som f.eks. var høyeste temperatur i tillegg til at jeg har laget test plots og sett på de. Men med kolonnen 'Solskinstid' og 'Vindretning' sier grensa seg selv. Solskinstiden kan ikke være mer enn 60 minutter siden vi måler etter hver time, og vindretnings tallene symboliserer grader i en sirkel og en sirkel er 360 grader. Alle verdier som er over grensa, får verdien **NaN**

Visualisering

Først visualiserer jeg kolonnene 'Globalstraling', 'Vindretning' og 'Lufttrykk' i forhold til 'Trafikkmengde for å sjekke om jeg vil ha med disse kolonnene videre.

Jeg finner ut av at grafen for 'Globalstraling' viser at det ikke har så mye å si om det er mye stråling eller ikke. Man skulle trodd at det hadde syklet flere når det var mer stråling, men grafen viser omvendt. Jeg velger derfor å fjerne denne kolonnen. Selv om grafen for vindretningen også er ganske jevn kan man se at det er veldig mange færre som velger å sykle når det kommer vin fra nord, så det er noe man kan ta med seg videre. Velger også å beholde Lufttrykk siden man kan se at det er relativt flere som velger å sykle når det er høyt lufttrykk enn når det er lavt. Høytrykk pleier som regel å bli assosiert med klart vær og rolige forhold, mens lavtrykk blir assosiert med ustabilt vær og økt sannsynlighet for nedbør.