Arbeidskrav 4 – nettskraping, datavask og lineær regresjonsmodell

Dette arbeidskravet er en variant av et tidligere arbeidskrav tilknyttet R-delen av kurset, men nå skal det gjøres i Python og det etterspørres noen andre aspekter. Utgangspunktet er igjen denne nettsiden og tabellen som illustrerer forskjellen mellom leverandørers påståtte kjørelengde (WLTP) og Motor/NAFs faktiske kjørelengde (STOP) på vinteren.

Innleveringen skal være en python notebook (ipynb-fil) fra din Github-konto, og lenken skal legges inn i Canvas sitt innleveringssytem. Det er fullstendig lov å bruke KI/LLM-modeller etter UiT sitt eget regelvert for eksamensrelatert KI-bruk, men det bemerkes det at dere må kunne forklare all kode som er levert, da det blir viktig i neste arbeidskrav.

Arbeidsfilen (.ipynb) du leverer skal bestå av to deler (D1 og D2) og oppsettet er angitt under.

D1 – markdown. Her skal du presentere problemet og resultatene med beskrivende tekst og figurer. Tenk på denne delen som en rapport som kunne vært levert til en sjef som ga deg i oppdrag å undersøke problemet oppgaven angir.
PS: om du trenger en kjapp introduksjon til markdown, er det meste man trenger angitt her.

D2 – python. Her skal all koden du bruker for å løse problemet ligge. Koden skal være veldokumentert og besvare alle deler av de programmeringstekniske aspektene tilknyttet oppgaven.

Oppgaven er i korte trekk angitt under.

1) Skrap den angitte nettsiden og lagre resultatet i en dataramme.
2) Vask dataen til en tabell med bare relevante tall og gode overskrifter ved det følgende:
    a. Del kolonnen med WLTP-tall i to: en for kjørte km og en for energibruk.
    b. Hvis en kolonne har benevninger: legg alle benevningene til overskriften for kolonnen slik at det bare er tall i kolonnene.
    c. Fjern rader med meningsløse eller manglende verdier fra datasettet.
3) Lag en ny kolonne der forholdet mellom leverte antall kilometer og påståtte antall kilometer er angitt.
4) Lag en regresjonsmodell av typen $f(x) = ax + b$ der WLTP-km er uavhengig variabel.
5) Plott modellen fra (4) mot datapunktene og leverandørens påståtte kjørelengde og sørg for at alle elementene i plottet er angitt tydelig for leseren. Plottet skal angis i to varianter: en variant som tar utgangspunkt i de x og y verdier som framkommer av dataen i tabellen, en annen som inkluderer origo.
6) Sammenfatt en rapport som forklarer utgangspunktet for undersøkelsen og funnene dine. I tillegg til funnene skal du kommentere det følgende:
    - sammenhengen mellom (3) og kolonnen «avvik».
    - om konklusjonen av plottene fra (5) er et tilfelle av tukling med y-akser.
    - kan vi med tilstrekkelig grad av sannsynlighet påstå at leverandørene holder det de lover?
    - gjennomsnittlig avvik for alle bilene i testen.