Første obligatoriske oppgavesett

Oppgavene løses individuelt. Dere kan diskutere oppgavene og mulige løsninger… men dere skal løse oppgavene selv. Oppgaven besvares som et Github-repository som dere kan sette public, eller private med invitasjon til meg (magbak@gmail.com). Innleveringsfristen er tirsdag 10. oktober.

Oppgave 1

Denne oppgaven baserer seg på data om trafikkulykker tilgjengelig fra SSB i Tabell: “08329: Drepte eller skadde i trafikkulykker, etter alder, kjønn, skadegrad, trafikantgruppe og ulykkestype 1999M01 - 2023M08”

En jobb som henter ut informasjon fra SSB finnes i git-repoet. Dere kan velge om dere vil løse oppgavene i Python eller i R - eller en blanding! Hvis du vil bruke R må du først skrive datasettet til en fil - for eksempel csv eller parquet. Det finnes også et ggplot-basert visualiseringsbibliotek til Python, kan være verdt å prøve: https://realpython.com/ggplot-python/

A) Formater datokolonnen som datetime, bruk f.eks. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.to_datetime.html
Sett dagen til 15. i hver måned.
Fjern også “\xa0” fra ulykkestype-kolonnen

B) Hva har skjedd med totalt antall drepte over tid? Lag et plott som viser utviklingen av antall drepte for hvert år i datasettet fra første år med data i datasettet til nå.

C) Er det sesongmessig variasjon i skadetallene? Plott pr. måned og se om du ser et mønster.

D) Lag et plott som viser andel av de drepte eller skadde med de ulike skadegradene (inkl. drepte) for hvert år. Ikke ta med “Skadde i alt”.

E) Lag et plott som viser utviklingen av de ulike typene møteulykker for hvert år (“Møting ved forbikjøring” og “Andre møteulykker”). Det har blitt bygget mange midtdelere, så disse bør det bli færre av.

F) Vegtrafikkindeksen sier noe om hvordan trafikkmengden har utviklet seg over tid, og er relevant informasjon for å forstå hvor mye tryggere det har blitt å være i trafikken. https://www.vegvesen.no/globalassets/fag/trafikk/trafikkdata/vegtrafikkindeksen_2023-05.pdf
Tabell 8 viser kumulativ trafikkutvikling siden 2005. Bruk tabellen til å framskrive antall i hver skadegrad fra 2005 år i datasettet til 2022. Inkluder denne informasjonen i grafen fra deloppgave B.

Oppgave 2

I denne oppgaven bruker vi NIST-data:
https://pvdata.nist.gov/
Last ned Ground-data (Bulk Download) med 1 minutt oppløsning for 2015. Her kan dere gjenbruke løsningen på oppgavene fra tredje forelesning i github-repoet for å lage en stor Parquet-fil. Les deretter denne inn i R og gjennomfør selve oppgaven i R.

A) Et pyranometer måler innstrålingen av sol som et gitt antall Watt pr. kvadratmeter. Velg en tilfeldig uke på sommeren 2015, og plott tidsstempel på X-aksen, og verdien til "Pyra1_Wm2_Avg" på Y-aksen.

B) "InvPDC_kW_Avg" er gjennomsnittlig effekt produsert av solcelleanlegget pr. minutt. Legg til denne verdien på Y-aksen.

C) Er "Pyra1_Wm2_Avg" og "InvPDC_kW_Avg" avhengige eller uavhengige variabler? Beregn Pearson Correlation mellom de to variablene. Begrunn svaret ditt. Hva tenker du om årsak og virkning?

D) Plott sammenhengen mellom de to variablene (scatterplot), med "InvPDC_kW_Avg" på X-aksen, "Pyra1_Wm2_Avg" på Y-aksen.

E) Estimer parametrene til en lineær regresjonsmodell med kun data fra 2015 hvor:
InvPDC_kW_Avg = α + $\boldsymbol{\beta}$*Pyra1_Wm2_Avg + $\boldsymbol{\epsilon}$
Vi antar at $\boldsymbol{\epsilon}$ er i.i.d. normalt distribuert. Hvor stor andel av variansen i "InvPDC_kW_Avg" gjør vi rede for med "Pyra1_Wm2_Avg"? Er dette en god modell? Finn et 95% konfidensintervall for $\boldsymbol{\beta}$. Forklar hva det betyr at $\boldsymbol{\beta}$ har dette konfidensintervallet.

F) Finn 95% konfidensintervallet til InvPDC_kW_Avg når Pyra1_Wm2_Avg = 400.

G) Lag et plott for verdier mellom 0 og Pyra1_Wm2_Avg på X-aksen, og på Y-aksen:
- De faktiske verdiene for InvPDC_kW_Avg (scatterplot)
- Estimatet av E[InvPDC_kW_Avg | Pyra1_Wm2_Avg] som en linje
- Nedre og øvre grenseverdi for 95%-konfidensintervallet til estimatet av InvPDC_kW_Avg som linjer.

H) Finn residualen ($\boldsymbol{\epsilon}$) for alle tidsstemplene og lag et histogram som viser distribusjonen til $\boldsymbol{\epsilon}$. Ser $\boldsymbol{\epsilon}$ normaldistribuert ut?

I) Bruk regresjonsmodellen til å predikere InvPDC_kW_Avg for de neste tre årene (du må laste ned litt mer data). Plott $\boldsymbol{\epsilon}$ på y-aksen og tidsstempel på x-aksen. Hva skjer med $\boldsymbol{\epsilon}$ over tid? Hvorfor tenker du at dette skjer?

Gjør en hypotesetest med α=0.05 for om gjennomsnittsverdien til $\boldsymbol{\epsilon}$ for 2016 kommer fra en normaldistribusjon med μ=0. Hva konkluderer du? Hvor ofte gir denne hypotesetesten falsk positiv?