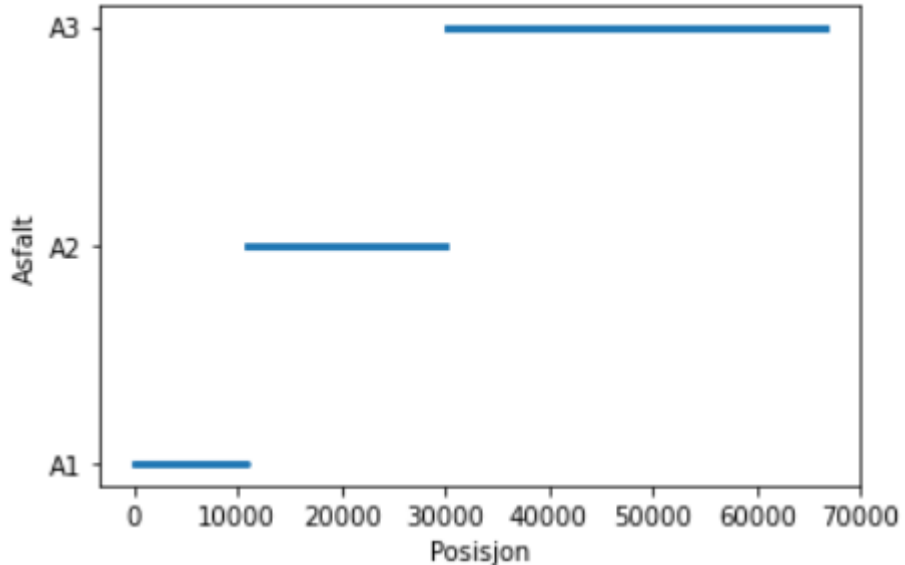


SVARARK for: Tellende prosjekt i ISTx1003 - høst 2022

Oppgave 1 (50%)	Oppgavetekst	SVAR
Q1a.1)	Hvor mange observasjoner har vi i datasettet? Matcher det med lengden på hver del av veien og den totale veilengden?	Vi har totalt 3355 observasjoner fra dataene som var samlet inn. Ut fra dataene som ble analysert, er det ikke en konkret match med observasjonene og den totale veilengden ettersom vi kan se at den siste posisjonen stoppet på omtrent 66693.6 meter som er ekvivalent med ca. 66.7 km. Dette matcher ikke helt med 67.8 km som ble oppgitt i rådata seksjonen.
Q1a.2)	Hva tror du en negativ verdi av `sporing` betyr i praksis? Hvordan tolker vi betydningen av `sporing_trafikk`?	I praksis betyr dette at veien har blitt utbedret (lagt ny vei). Dette kan også oppstå av eksterne faktorer eller ved mindre sporing målefeil, men om en kun ser på det idealistisk, er det kun ved ny vei. Betydningen av sporing_trafikk kan tolkes som hvor mye sporing oppstår ved et konstant antall av biler (i dette tilfellet 10000), og avhenger i stor grad av asfalt typen som er brukt. Kan anses som et mål på hvor lett asfalten får sporing.
Q1a.3)	Vi manglet originalt 17 av 3355 målinger av veibredden. Hvordan løste vi det?	De som har gitt oppgaven spurte kilden til dataen om hvorfor det manglet og fikk riktig data tilbake. Ofte er man ikke så heldig og må arbeide seg rundt manglende informasjon, og man kan bli nødt til å tilpasse modellen til manglende data.
Q1a.4)	Hva er det fulle navnet på asfalttypen vi har mest av? (skjelettasfalt, asfaltbetong eller asfaltgrusbetong)	Asfalttypen som det er mest av er Asfaltgrusbetong (forkortet til A3 i resten av tabellene). Vi kan se dette fra funksjon av asfalt mot posisjon. A3 er mest brukt på våres veideler.

		 <p>The chart displays three horizontal bars representing different road types (Asfalt) across a range of positions (Posisjon). The y-axis is labeled 'Asfalt' and has three categories: A1, A2, and A3. The x-axis is labeled 'Posisjon' and ranges from 0 to 70000. Bar A1 is located at the bottom, spanning from position 0 to approximately 10000. Bar A2 is in the middle, spanning from position 10000 to 30000. Bar A3 is at the top, spanning from position 30000 to approximately 65000.</p>
Q1a.5)	Hvorfor tror du veibredden er mye større i starten av datasettet?	<p>Veibredden i starten av datasettet er større ettersom starten inneholder veier som består av flere felt (vanligvis 4), og veiene nærmere den svenske grensen består vanligvis av 2 felt eller enda mindre. Grunnen til dette kan være fordi det er forventet å være høyere/mer trafikk ved Stjørdal enn ved den svenske grensen.</p> <p>Kontekst: Datasettet inneholder koordinater som begynner i stjørdal (breddegrad: 63.46724, lengdegrad: 10.933685)</p>
Q1a.6)	SVV vil at nye veier skal holde i 20-30 år. Basert på dataene fra 2020, er det realistisk? Hvorfor/hvorfor ikke? Veien skal repareres når spordybden blir for stor. Blir veien reparert ofte nok? Gi kort begrunnelse.	<p>Basert på dataene så ser vi at om veiene opprettholder en sporing på over 1.25 i snitt, så vil ikke veien holde i 20 år.</p> <p>Vi ser at veiene som også har lav sporing_trafikk, som antyder at veien er nyere gitt tilsvarende asfalttype, også har høyere sporing enn grensen for å holde i over 20 år. Det er dermed fornuftig å konkludere at de fleste veistrekningene ikke vil være i stand til å holde ut i 20-30 år som SVV</p>

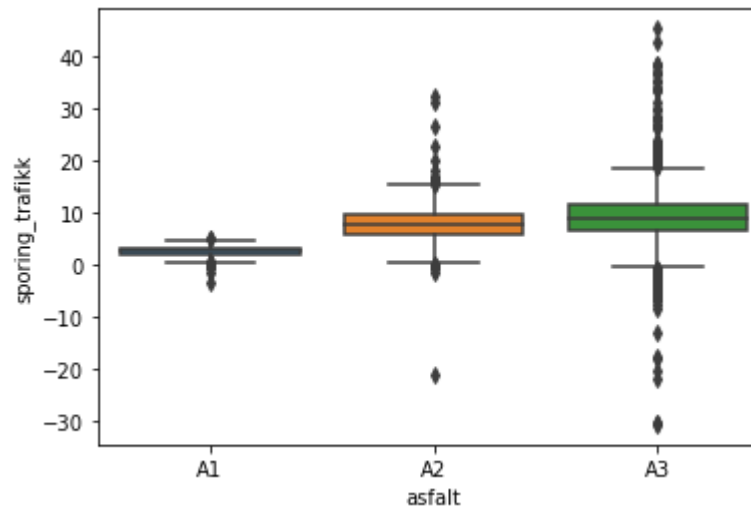
		<p>ønsker. Vi kan også se at veiene ikke blir reparert ofte nok ettersom frekvensplottet til strekninger med spordybde over 25 mm viser at de fleste strekningene har fortsatt betydelig positiv endring i sporing som tilsier at strekningen har fortsatt ikke blitt reparert. Dette betyr at flere og mer frekvente reparasjoner er nødvendige for å opprettholde SVVs grenser.</p>
Q1b.1)	Skriv ned ligningen for den estimerte regresjonsmodellen og forklar de ulike elementene.	<p>Vi kan ta ligningen av regresjonsmodellen som $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$, hvor $\hat{\beta}_0$ er konstant og $\hat{\beta}_1$ er koeffisienten av forklaringsvariabelen "spordybde". Vi finner disse talene som interception mellom rader "intercept" og "spordybde" og søylen "coef" i summary og</p> <p>Likning for den estimerte regresjonsmodellen: $\hat{y} = 4.6113 + 0.2665x$</p>
Q1b.2)	Hva er den estimerte verdien til skjæringspunktet (intercept) β_0^{\wedge} , og hvordan vil du tolke den?	Verdien til skjæringspunktet er 4.6113 og vi kan tolke det som sporing_trafikk på en vei med sporing lik 0.
Q1b.3)	Vi ser at for "spordybde" er "coef" lik 0.2665. Hva er formelen som er brukt for å regne ut denne verdien? Hvordan vil du forklare dette tallet til noen som ikke har hørt om lineær regresjon?	<p>Formel for å regne ut spordybde:</p> $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ <p>Vi kan forklare dette tallet som en økning i sporing_trafikk for 1 mm.</p>
Q1b.4)	For "spordybde" har vi også tallene 0.239 og 0.294 i kolonnene "[0.025 0.975]". Hva er disse to tallene og hvordan tolker du dem?	De to tallene er begrensninger av et konfidensintervall på 95% i regresjonsmodellen (dvs. 0.239 er nedre grenseverdi, og 0.294 er høyere grenseverdi). For oss viser tallene en intervall der vi forventer at de fleste (95/100) observasjoner skal falle inni dette intervallet.

Q1b.5)	Videre står det for `spordybde` at `P> t ` er 0.000. Hvilken hypotese er testet her? Hva er konklusjonen fra hypotesetesten hvis vi bruker signifikansnivå 0.05? Hvordan henger dette sammen med tallene 0.239 og 0.294 fra forrige punkt?	I denne delen tester vi en hypotese som sier at spordybde er 0. Ved å bruke signifikansnivå 0.05 kan vi se at $P > t $ er 0.000 som betyr at P er mindre enn 0.0005, som hjelper oss til å konkludere med at nullhypotesen må forkastes ettersom P-verdien er statistisk signifikant. Ut fra konfidensintervallene kan vi også se at nullhypotesen bør forkastes, ettersom det er forventet at en gjennomsnittlig observasjon vil vanligvis falle innenfor konfidensintervallet som ikke dekker 0.
Q1b.6)	Hvilke modellantagelser gjør vi i en enkel lineær regresjon?	Antagelsene for en enkel lineær regresjon er at vi har observert et par av verdier (x, y) og vi antar at hvert observert y-verdi er en realisasjon av den tilhørende stokastiske variabelen Y. Samtidig skal sammenhengen mellom x og Y være gitt av denne ligningen: $Y = B_0 + B_1 x + \epsilon$ Vi må også være nøye på å sjekke at målingene vi gjør ligger spredt rundt en omtrentlig linje for å anta at enkel lineær regresjon er en passende modell for dataene.
Q1b.7)	Hva er en predikert verdi og hva er et residual? Skriv også ned relevante formler.	Predikert verdi er verdien som vi forventer fra å bruke ligningen til enkel lineær regresjon dersom vi setter en bestemt verdi av x. Residualet er differansen mellom den egentlige observasjonen og den predikerte verdien som vi fikk ved å bruke ligningen. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$ $\hat{e}_i = y_i - \hat{y}_i$
Q1b.8)	Hvordan kan vi bruke predikert verdi og residual til å sjekke modellantagelsene?	Vi kan bruke predikert verdi og residual for å sjekke størrelsen på avviket mellom predikerte observasjoner og reelle observasjoner i regresjonsmodellen. Dette gir oss et innsikt i hvordan sammenhengen mellom x og y egentlig er, og vi kan bruke det for å oppnå mer nøyaktige predikerte verdier ved denne ligningen: $Y_i = B_0 + B_1 x_i + \epsilon_i$ der ϵ er residualet.

Q1b.9)	Vi får oppgitt tallet `R-Squared` til å være 0.098 (skrives ofte som 9.8%). R^2 har i enkel lineær regresjon en sammenheng med korrelasjonskoeffisienten, men det er også en annen definisjon som er relatert til sum av kvadrerte residualer. Hvilken formel er det? Forklar alle symboler. Hvordan vil du forklare R^2 til noen som ikke har hørt om enkel lineær regresjon?	$R^2 = \frac{SST - SSE}{SST}$ <p>$SSE = \sum \epsilon_i^2$ ϵ -residualet. Sum av kvadrerte residualer.</p> <p>$SST = \sum (y_i - \bar{y})^2$, der \bar{y} - gjennomsnittsverdien av observasjoner.</p> <p>Vi kan forklare R^2 som et mål på hvor godt en lineær regresjon som bestemmer observasjonene våre passer til datasettet vi har. En annen måte å forklare R^2 er hvor godt observert data passer til modellen vi lager.</p>
Q1b.10)	Studer plottet av predikert verdi mot residual. Hvordan skal et slikt plott se ut hvis modellantagelsene er oppfylt? Hvordan vil du evaluere plottet?	Et diagram av predikert verdi mot residual skal være ustrukturert med verdier sentrert rundt null dersom modellantagelsen er oppfylt. Dette plottet har verdier konsentrert rundt null, og det er ingen tydelig struktur.
Q1b.11)	Studer QQ-plottet av residualene. Hvordan vil du evaluere plottet?	Observasjonene i QQ-plottet følger ikke linja fullstendig og er kun optimal i midten. Ut fra QQ-plottet kan vi si at modellen er ikke særlig passende til data som vi har i modellen og er dermed ikke et ideelt utgangspunkt for oss.
Q1b.12)	Vil du konkludere med at modellen passer godt?	Vi konkluderer med at modellen passer dårlig fordi i vårt tilfelle ettersom R^2 er svært lav på 0.098. Dette indikerer at modellen er ikke passende til dataene som vi har og at veldig lite av variasjonen er forklart med den uavhengige variabelen i regresjonsmodellen.
Q1c.1)	Oppsummer kort hva du ser i plottene. Fokus skal være om du tror at det er noen sammenheng mellom `sporing_trafikk` (som respons) og de tre forklaringsvariablene (`spordybde`, `veibredde` og `asfalttype`). Hvilken asfalttype har høyest verdi for `sporing_trafikk`?	I plottet med spordybde og sporing_trafikk ser det ut til at det er en vifteform der de fleste verdiene er konsentrert litt over 0 sporing_trafikk. Ved lave verdier av spordybde er alle verdier av sporing_trafikk også rundt 0, og spredningen i sporing_trafikk øker med spordybden. Ved verdier av spordybde over 20/30mm ser det ut som at verdiene til sporing_trafikk går mot 0 igjen. Dette kan skyldes at det er færre målinger i dette området.

Når veibredden er rundt 3 meter er verdiene til sporing_trafikk veldig spreddt. Ved alle andre verdier av veibredden er sporing_trafikk rundt null. Det ser ut som at veibredden er knyttet til hvilken type asfalt som er brukt, og at en kan forvente lave verdier av sporing_trafikk ved alle veibredder utenom rundt 3 meter.

Asfalttypen A1 har de fleste av sine målinger med verdier av sporing_trafikk omtrentlig mellom 0 og 5. Asfalttype A2 sine målinger har generelt litt høyere verdier av sporing_trafikk og holder seg mellom 0 og 20. Asfalttype A3 har de høyeste verdiene av sporing_trafikk med verdier både litt høyere og litt lavere enn asfalttype A2.



Q1c.2)	Tilpass modellen ved å bruke koden for steg 3 og 4 over, med den nye formelen. Skriv ned ligningen for den estimerte regresjonsmodellen.	$\widehat{y}_i = 4.5999 + 0.1248 * x_1 - 0.8895 * x_2 + 4.4683 * x_3 + 5.6040 * x_4$
Q1c.3)	Hvor mange regresjonsparametere er estimert? Hva er betydningen av de ulike regresjonsparameterene?	<p>5 regresjonsparametere er estimert: ett konstant(intercept) og 4 forklaringsvariablene: spordybde, veibredde og asfalttype (består av 2 kategorier).</p> <p>$\beta_{\text{intercept}}=4.5999$: verdien når vi har asfalttype 1, spordybde 0, og veibredde 0.</p> <p>$\beta_{\text{spordybde}}=0.1248$. Vei 1 og vei 2 helt like unntatt av spordybde, spordybde av vei 1 mindre enn vei 2 i 1 mm, sportrafikk av vei 2 er 0.1248 større enn sportrafikk vei 1.</p> <p>$\beta_{\text{veibredde}}=-0.8895$. Vei 1 og vei 2 helt like unntatt av veibredde, veibredde av vei 1 mindre enn vei 2 i 1 m, sportrafikk av vei 2 er 0.8895 mindre enn sportrafikk vei 1.</p> <p>$\beta_{\text{asfalt[T.A2]}}=4.4683$. Hvis vi har to helt samme veier, men en har asfalttype A1 og andre har asfalt type 2. Estimering av sportrafikk av vei med asfalttype A2 skal bli 4.4683 større en vei med asfalttype 1.</p> <p>$\beta_{\text{asfalt[T.A3]}}=5.6040$. Hvis vi har to helt samme veier, men en har asfalttype A1 og andre har asfalt type 3. Estimering av sportrafikk av vei med asfalttype A3 skal bli 5.6040 større en vei med asfalttype 1.</p>
Q1c.4)	Sammenlign den estimerte regresjonskoeffisienten for `spordybde` i denne modellen med samme koeffisient i den enkle lineære regresjonsmodellen. Har disse to samme tolkning?	De har ikke samme tolkning siden vi i enkel lineær regresjon ikke har med ulike asfalttyper. I enkel regresjon model har vi tolkning for økning av sporing trafikk i 1 mm for hver asfalttype, men i multipl regressjon bare for asfalttype A1.
Q1c.5)	Hva er predikert sporing per ti tusen biler for en veistrekning med asfaltbetong, 10 mm spordybde	Utrekning: $4.5999+4.4683*1+5.6040*0+0.1248*10-0.8895*4 = 6.7582$

	og 4 meter veibredde? (Regn ut for hånd ved å bruke relevante tall fra `resultat.summary()`.)	For å finne predikert sporing per ti tusen biler trenger vi å sette variabler fra spørsmålet til modellen som vi fikk i Q1c.2 som vist i utregningen ovenfor.
Q1c.6)	Forklaringsvariabelen `asfalt` er kategorisk og vi har brukt en såkalt dummy-variabelkoding, der `A1` (skjelettasfalt) er referansekategorien. Er effekten av de andre asfalttypene på sporing per tusen biler signifikant forskjellig fra effekten for asfaltbetong på nivå 0.05? Hvis vi sammenligner tre deler av veien med lik spordybde og veibredde, men ulik asfalttype, hva er gjennomsnittlig forskjell i sporing per tusen biler? Hvilken type asfalt ser ut til å gi den mest solide veien?	<p>a) På grunn av at parametrene til alle andre asfalttyper ikke er lik 0, vil andre asfalt typer ha en effekt på sporing.</p> <p>b) Vi måler forskjell per ti tusen biler, men oppgaven spør om gjennomsnittlig forskjell per tusen biler, så vi må multiplisere svaret med 10. Gjennomsnittlig forskjell per tusen biler er $(4.4683+5.6040)*10/3 = 33.574$</p> <p>c) Mest solide asfalttype er A1. Fordi hvis vi har de samme parametrene av spordybde og veibredde vil prediksjonen til sporing_trafikk med veitype A3 bli 5.6040mm større enn vei med A1, og sporing_trafikk til A2 blir 4.4683mm større enn A1.</p>
Q1c.7)	Hva er andel forklart variasjon? Ville du forventet at andelen forklart variasjon gikk opp da vi la til flere forklaringsvariabler enn bare `spordybde`? Hvis vi nå la til en forklaringsvariabel som forteller noe om dyrelivet langs veien, ville da R^2 økt?	<p>justert $R^2=0.239$.</p> <p>Flere forklaringsvariabler vil gjøre at R^2 øker.</p> <p>Dersom vi legger til en forklaringsvariabel om dyrelivet vil R^2 øke, men Adj. R-squared vil antakelig ikke øke.</p>
Q1c.8)	Basert på utskrifter og plott, vil du konkludere med at modelltilpasningen er god?	Den modell passer ikke godt for oss, men den er bedre enn modell enn det vi hadde før etter enkel lineær regresjon. Vi kan konkludere det pga. R^2 er ganske lav og plottene ikke er langt bedre enn ved enkel lineær regresjon.
Q1c.9)	Utfør en ny multippel lineær regresjon (steg 2-5) med `sporing_trafikk` som respons og `veibredde` og `spordybde` som forklaringsvariabler. Du må nå modifisere modellformelen ved å ta bort `asfalt`, og så kopiere inn kode for steg 2-5. Hvor mange regresjonsparametere er nå estimert? Hva er de signifikante forklaringsvariablene? Ikke	<p>Det er 3 regresjonsparametre : konstant(intercept) og 2 forklaringsvariabler : spordybde,veibredde</p> <p>Alle forklaring variabler er signifikante siden hver har p-verdien som er mindre enn 0.0005</p>

	sikkert at passer for svar, men håper at våres forklaring variablene er signifikante.	
Q1c.10)	Er modelltilpasningen god?	Nei, modellen passer ikke godt fordi R^2 er 0.123, og ettersom denne verdien er lav indikerer dette at modellen er ikke passende til dataene som vi har og at veldig lite av variasjonen er forklart med de uavhengige variablene i modellen.
Q1c.11)	Sammenlign `Adj. R-squared` (også kalt justert R^2) for modellen med og uten `asfalt`. Hvis vi skal avgjøre om `asfalt` skal være med som forklaringsvariabel ved å bruke justert R^2 , hva blir konklusjonen? Begrunn svaret ditt.	Adj. R-squared (modell med asfalttype) = 0.239 Adj. R-squared (modell uten asfalttype) = 0.123 Adj. R-squared (modell med asfalttype) > Adj. R-squared (modell uten asfalttype) Adj. R-squared verdien i modell med asfalttype er betydelig større enn i modell med uten asfalt. Forklaringsvariablene for asfalttype gir oss derfor en bedre modell.

Oppgave 2 (30%)	Oppgavetekst	SVAR
Q2a.1)	Hvorfor ønsker vi å dele dataene inn i trening-, validering- og test-sett?	<p>Dette gjør vi for å kunne bruke data på en bedre måte.</p> <p>Gjennom å splitte opp dataen på denne måten kan vi sørge for at vi ikke får en overspesialisert modell ved å benytte noen av dataene til å sjekke hvor godt modellen fungerer. Vi må ha separate tester for modellen. Om man ikke bruker forskjellige data, ville du i det ekstreme tilfellet kunne lage en funksjon som passer perfekt til alle datapunktene og få en "perfekt modell" som ikke klarer å gi en prediction for ny data og er derfor ubrukelig.</p>
Q2a.2)	Hva brukes hver av disse delene til i våre analyser?	<p>Trening: Kan benytte dette settet til å lage eller trene opp modellen.</p> <p>Validering:</p>


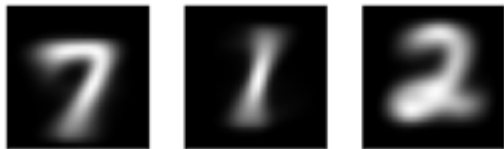
		<p>valideringsettet kan benyttes til å validere at modellen fra treningssettet fungerer som den skal, når den skal estimere nye data. Om feilraten på treningsdataen er liten, men feilraten på valideringsdataen er stor så er modellen overspesialisert.</p> <p>Test: benyttes helt på slutten for å sjekke hvordan modellen faktisk håndterer upartiske data.</p>
Q2a.3)	Hvor stor andel av dataene er nå i hver av de tre settene? Ser de tre datasettene ut til å ha lik fordeling for de tre forklaringsvariablene og responsen?	<p>60% trening, 20% validering og 20% testsett. <i>(utregnet fra 80% trening/validering og 20%test, der trening/validering var 75%/25%)</i></p> <p>Fordelingen av forklaringsvariablene på de forskjellige settene ser ut til å være relativt god. min/max mellom settene har ikke en større forskjell enn på 20%. Snittet på dataene mellom kampene ser ut til å variere en del.</p>
Q2a.4)	Kommenter hva du ser i plottene og utskriften.	<p>i plottet ser vi en sammenheng mellom hjemmeseier og noen av de forklaringsvariablene vi brukte.</p> <p>For eksempel (grafene med skudd_paa_maal_diff og responsvariabel forseelse_diff) $\text{skudd_paa_maal_diff} = \text{HST} - \text{AST}$ og $\text{forseelse_diff} = \text{HF} - \text{AF}$, vi ser at økning i differansen har sammenheng med økning i antall oransje punkter som tilsvarer mer hjemmeseiere. For akkurat den grafen er det mulig å se en tilnærmet skillelinje når $\text{skudd_pa_maal_diff}$ er ca. 0 (skiller hjemmeseier på høyre og ikke hjemmeseier på venstre siden av den tilnærmet vertikalt skillelinjen). På den samme grafen kan man se at det er større verdier for forseelser_diff (større positive eller negative verdier) når differansen på mål er ca. 0 (logisk, siden da er det mye spenning mellom de to lagene og laget som ligger bak eller føler seg presset mest sannsynlig</p>

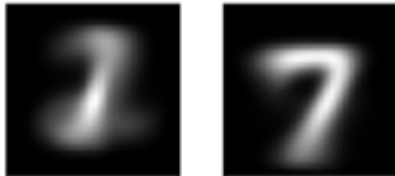

		<p>gjør mer forseelser), for eksempel hvis et lag dominerer det andre laget kan den andre laget gjøre opp for det ved flere forseelser. Det går an å se sammenhengen fordi de blå punktene holder seg til venstre siden og oransje punktene til høyre.</p> <p>På samme måte kan man tolke de to andre grafene (skudd_på_mål_diff med responsvariabelen corner_diff og corner_diff med responsvariabelen forseelse_diff). Grafen som bruker corner_diff med responsvariabelen forseelse_diff er vanskeligere å tolke. Siden corner spark og forseelse-diff har ikke en sterk relasjon i en fotballkamp.</p> <p>På utskriften ser vi en korrelasjonstabell av ulike kombinasjoner av forklaringsvariabler. Høyest korrelasjon ser vi i tabellen der forklaringsvariabelen er den samme som responsvariabelen, og da er korrelasjonen = 1. Vi ser at det er ganske lav korrelasjon på resten, med unntak av skudd_på_mål_diff sammen med y (hjemmeseier eller ikke) som har korrelasjon = 0.462209 (dette viser at jo større differansen på mål er jo større sannsynlighet er det for at det er hjemmelaget som vinner kampen).</p>
Q2a.5)	Hvilke av de tre variablene tror du vil være gode til å bruke til å predikere om det blir hjemmeseier? Begrunn svaret kort.	<p>Skudd_på_mål_diff fordi den har størst korrelasjon med y (hjemmeseier eller ikke) med korrelasjonsverdi = 0.462209.</p> <p>Corner_diff kan ved første øyekast se ut som en god estimator men i grunn er den ganske dårlig, selv om den korrelerer en del med Skudd_på_mål 0.330959, denne korrelasjonen er ikke relevant for dette spørsmålet så vi benytter ikke denne estimatoren da den korrelerer heller dårlig med y (seier) med 0.047838.</p>

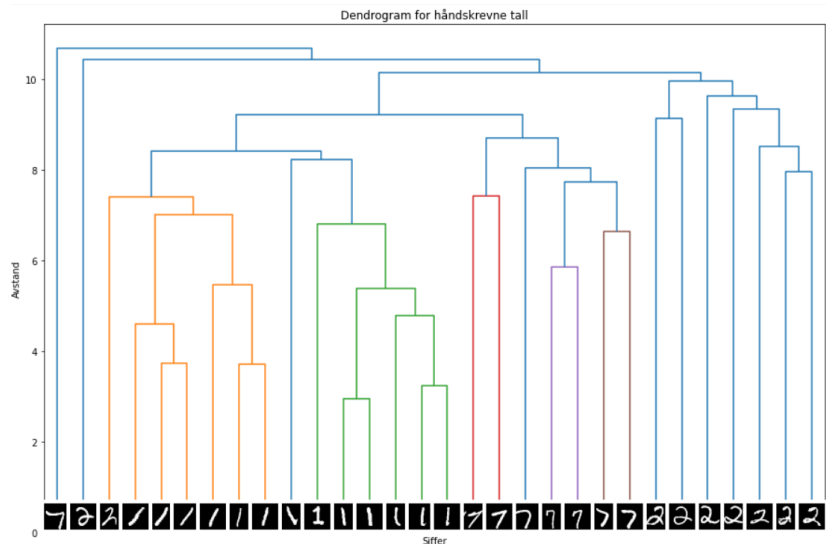
		Resten har tilnærmet ubetydelig korrelasjon utenom med seg selv (selvfølgelig).																																		
Q2b.1)	Hvilke forklaringsvariabler er signifikante i modellen på signifikansnivå 0.05?	<table><tr><td></td><td>coef</td><td>std err</td><td>z</td><td>P> z </td><td>0.025</td><td>0.975</td></tr><tr><td>skudd_paa_maal_diff</td><td>0.3909</td><td>0.027</td><td>14.710</td><td>0.000</td><td>0.339</td><td>0.443</td></tr><tr><td>corner_diff</td><td>-0.0727</td><td>0.016</td><td>-4.429</td><td>0.000</td><td>-0.105</td><td>-0.041</td></tr><tr><td>forseelse_diff</td><td>-0.0051</td><td>0.013</td><td>-0.377</td><td>0.706</td><td>-0.032</td><td>0.021</td></tr></table> <p>leser i tabellen fra P> z kolonnen, for å hente ut pverdiene. (=0 betyr bare at det rundes av til 0 om en har 3 desimaler som presisjon)</p> <p>ser at skudd_paa_maal_diff og corner_diff har p verdi under 0.05 mens forseelse_diff har gått over med verdi 0.706. Det vil si at forseelse_diff er ikke signifikant mens de to andre forklaringsvariabler er signifikante (corner_diff og skudd_paa_maal_diff).</p>								coef	std err	z	P> z	0.025	0.975	skudd_paa_maal_diff	0.3909	0.027	14.710	0.000	0.339	0.443	corner_diff	-0.0727	0.016	-4.429	0.000	-0.105	-0.041	forseelse_diff	-0.0051	0.013	-0.377	0.706	-0.032	0.021
	coef	std err	z	P> z	0.025	0.975																														
skudd_paa_maal_diff	0.3909	0.027	14.710	0.000	0.339	0.443																														
corner_diff	-0.0727	0.016	-4.429	0.000	-0.105	-0.041																														
forseelse_diff	-0.0051	0.013	-0.377	0.706	-0.032	0.021																														
Q2b.2)	Hvordan kan du tolke verdien av $\exp(\text{skudd_paa_maal_diff})$?	Odds er definert som forholdet mellom sannsynlighet for suksess og for fiasko. Hvis vi skal øke verdien til forklaringsvariabel skudd_paa_maal_diff fra $x_{1,i}$ til $x_{1,i} + 1$ vil odds(forholdet mellom sannsynlighet for suksess og fiasko) multipliseres på $\exp(\text{skudd_paa_maal_diff}) = 1.478327$. Større skudd_paa_maal_diff øker sannsynligheten for seiren.																																		

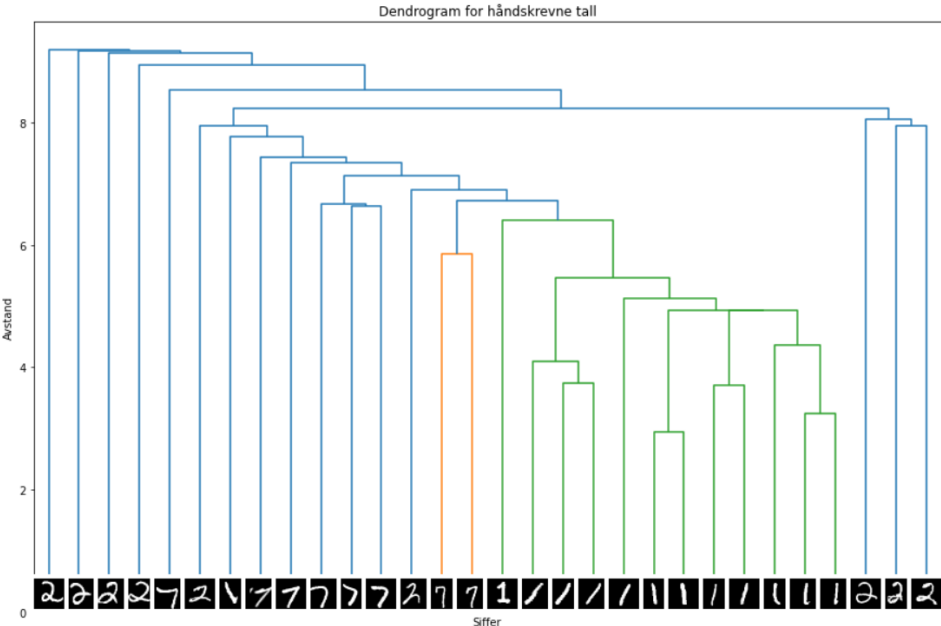
		$\frac{p_i}{1-p_i} \cdot \exp(\beta_1)$
Q2b.3)	Hva angir feilraten til modellen? Hvilket datasett er feilraten regnet ut fra? Er du fornøyd med verdien til feilraten?	Feilraten angir andelen av ukorrekte klassifikasjoner. Feilraten er regnet ut fra y (hjemmeseier eller ikke) og de tre forklaringsvariablene skudd_paa_maal_diff, corner_diff, forseelse_diff. Feilraten er ca 0.3, og vi er ikke helt fornøyd med modellen siden vi får feil ca 30% av gangen.
Q2b.4)	Diskuter hva du ser (modell uten forseelse_diff).	Forseelse_diff som forklaringsvariabel er nå fjernet. Feilraten og odds har endret seg.
Q2b.5)	Som din beste modell for logistisk regresjon, vil du velge modellen med eller uten `forseelse_diff` som kovariat? Begrunn svaret.	Uten forseelse diff, fordi feilraten er litt lavere. (0.2948 mot 0.2973)
Q2c.1)	Forklar kort hva som er gjort i koden over, og hvilken verdi av k du vil velge.	Koden over viser oss feilrater hvis vi tar hyperparameter (k) som alle oddetall til 49 ikke inkludert. Koden bygger også plott som viser feilraten i følge av antall naboer k for valideringssett. De 2 siste linjene av koden finner oss hvilke naboer som har den minste feilraten. Vi må velge en k som minimerer feilraten av valideringssett. I vår situasjon er det k = 37 og feilraten = 0.28992629
Q2d.1)	Vil du foretrekke å bruke logistisk regresjon eller k-nærmeste-nabo-klassifikasjon på fotballkampdataene?	kNN gir feilraten 0.26225490196078427 som er mindre enn feilraten av logistisk regresjon som er 0.2671568627450981. Ifølge feilraten er det mest optimalt å bruke k-nærmeste-nabo-klassifikasjon på fotballkampene, så det er den som er foretrukket å bruke.
Q2d.2)	Oppsummer hva du har lært at kan være en god metode for å predikere om hjemmelaget vinner en fotballkamp.	En metode som er muligens god på å predikere om hjemmelaget vinner eller ikke er å måle hvor mange ganger hjemmelaget skyter skudd på mål og antall cornere som hjemmelaget får. En betydelig stor positiv differanse av skudd på mål og cornere (dvs. at hjemmelaget skyter flere ganger og

		har flere cornere enn det andre laget) gjør det mer sannsynlig for at hjemmelaget vinner.
--	--	---

Oppgave 3 (20%)	Oppgavetekst	SVAR
Q3a.1)	Hvilke 3 siffer har vi i datasettet? (Alle 3 sifrene er representert blant de første 10)	Sifrene som ligger i datasettet er 1, 2 og 7.
Q3a.2)	Hvor mange bilder har vi i datasettet?	Vi har 6000 bilder i datasettet.
Q3a.3)	Hvilket siffer ligner det 500. bildet i datasettet vårt på? Lag et bilde som viser dette sifferet. (Husk at Python begynner nummereringen med 0, og derfor refereres det 500. bildet til `[499]`)	Det 500. bildet i datasettet ligner på sifferet 2. 
Q3b.1)	Tegn sentroidene av de 3 klyngene fra \$K\$-gjennomsnitt modellen. Tilpass koden over for å plote. Her kan du ta skjermbilde av sentroidene og lime inn i svararket.	
Q3b.2)	Synes du at grupperingen i klynger er relevant og nyttig? Forklar. Maks 3 setninger.	Grupperingen vi ville ha er 7, 1 og 2. Her har vi alle 3 og de har kommet tydelig frem. Grupperingen er relevant og nyttig
Q3b.3)	Vi har valgt $K = 3$ for dette eksempelet fordi vi vil finne klynger som representerer de 3 sifferene. Men generelt er K vilkårlig. Kom opp med et forslag for hvordan man (generelt, ikke nødvendigvis her) best kan velge K . (Se her, for eksempel: https://medium.com/analytics-vidhya/how-to-de	Albue-metoden går ut på å velge K til der det er en knekk på kurva mellom K og optimeringskriteriene. Silhuett-metoden kan gi oss en verdi mellom 1 og -1 som viser oss hvor likt et punkt er til sin egen klynge sammenlignet med andre klynger. Hvis man kombinerer disse to metodene nevnt over har vi en god estimat for hvilken verdi for k vi kan velge.

	termine-the-optimal-k-for-k-means-708505d204e b). Beskriv i egne ord med maks 3 setninger.	
Q3b.4)	Kjør analysen igjen med $K = 2$ og $K = 4$. Synes du de nye grupperingene er relevante?	<p>Nei. Enten blir ikke alle tallene tatt med eller så får vi flere like tall. Det kan kanskje være hensiktsmessig å dele inn i flere 3 siden noen av tallene kan ha forskjellige skrivemåter som ikke likner hverandre?</p> <div data-bbox="1081 403 1473 579">  </div> <p style="text-align: right;">$K=2$</p> <div data-bbox="1081 676 1760 820">  </div> <p style="text-align: right;">$K=4$</p>
Q3c.1)	Vurder dendrogrammet nedenfor. Synes du at den hierarkiske grupperingsalgoritmen har laget gode/meningfulle grupper av bildene?	<p>Ja, for det meste ser grupperingen ganske bra ut, noen få symboler skiller seg ut litt med hvilken gruppe de hører til på venstresiden av dendogramet. Spesielt det tredje 2-tallet fra venstre har havnet midt inn i klyngen med 1-tall, og første 7-tall fra venstre vil også gjøre inndelingen av klynger mer upresis. Det at alle 2-tallene har veldig stor avstand mellom</p>

		<p>seg gjør også det vanskeligere å dele inn i klynger.</p> 
Q3c.2)	I koden under har vi brukt gjeonnomsnittskobling ('method = 'average'). Hvordan fungerer gjeonnomsnittskobling? Maks 2 setninger.	Gjennomsnittskobling beregner alle avstandene mellom de ulike punktene i klynge 1 og klynge 2, og regner ut gjennomsnittet av avstandene.

Q3c.3)	<p>Velg en annen måte enn `method = 'average'` til å koble klyngene sammen (vi har lært om dette i undervisningen) og lag et nytt dendrogram ved å tilpasse koden nedenfor. Kommenter resultatene. Ser det bedre/verre ut?</p>	<p>Vi brukte “singel” til å lage dendrogrammet. Dendrogrammet viser at ett-tallene havne som en underkategori av syv-tallene som igjen ligger under to-tallene. Dette gjør det vanskelig å dele inn i gode klynger. I tillegg er det flere bilder som har tilknytning til feil tallgruppe enn når vi brukte “average”.</p>  <p>The dendrogram, titled "Dendrogram for håndskrevne tall", shows the hierarchical clustering of handwritten digits. The y-axis is labeled "Avstand" (Distance) with values 0, 2, 4, 6, 8. The x-axis is labeled "Siffer" (Digits) and shows 27 handwritten digit samples. The tree structure indicates that the '1' digits (samples 15-18) cluster together at a distance of approximately 4, then merge with the '7' digits (samples 19-22) at a distance of approximately 6. This combined group then merges with the '2' digits (samples 23-26) at a distance of approximately 8. The '2' and '7' groups then merge with the remaining '1' digits at a distance of approximately 9. The '1' and '7' groups then merge with the remaining '2' digits at a distance of approximately 10. Finally, the '1' and '7' groups merge with the remaining '2' digits at a distance of approximately 11. The '2' and '7' groups then merge with the remaining '1' digits at a distance of approximately 12. The '1' and '7' groups then merge with the remaining '2' digits at a distance of approximately 13. Finally, the '1' and '7' groups merge with the remaining '2' digits at a distance of approximately 14.</p>
Q3d.1)	<p>Hvis vi skulle brukt en metode for å predikere hvilket siffer fra et håndskrevet tall er, og ikke bare samle dem i klynge, hva ville du brukt?</p>	<p>Maskinlæring med en CNN modell er en ganske godt testet måte å kunne gjenkjenne sifre og symboler fra håndskrevne tall ved bruk av nevrale nettverk. Men da dette ikke faller innenfor omfanget av denne oppgaven, og ikke er en direkte statistisk modell og krever korrekte datasett, vil vi gå litt anderledes frem.</p>

		<p>Det vi har gjort til nå er ikke-veiledet klyngeanalyse men for best mulig predikerings resultat er veiledet klyngeanalyse veien å gå. Med veiledet maskinlæring brukes merket data (labeled data). Det vil si at dataen som skal behandles må være behandlet (f.eks. av en person) og merket slik at algoritmen kan bruke informasjonen for trenings delen av modellen.</p>
--	--	--