Øving 4 i Big Data om Pandas og dataanalyse

1. Vi har et datasett med salgsdata. Dette kommer som tolv forskjellige filer, 1 fil for hver måned. Legg innholdet av alle filene i én dataramme (det anbefales å bruke parameteren `index_col=0` på `read_csv()`, for å unngå å få flere kolonner enn man trenger). Legg merke til her at du sannsynligvis har en kolonne som brukes som indeks, men dersom du begynner å se nærmere på denne kolonnen, vil du finne at den settes tilbake til 1 hver gang du leser inn en ny fil. Bruk metoden `reset_index()` for å fikse opp i dette, og sjekk etterpå at resultatet ble korrekt.

2. Vi har to kolonner, hhv. `x_t` og `perf`. Vi vet ikke hva disse er, og de regnes derfor bare som støy for vårt formål. Fjern disse kolonnene.

3. Vi er interessert i å finne den beste måneden når det gjelder omsetning, og samtidig finne hvor mye det ble omsatt for hver måned. For å gjøre det litt enklere for oss selv, er vi interessert i å legge til en kolonne som vi kaller `Month` til ramma vår.

   (a) Månedsnummeret finner vi kolonna `Order date`, men vi kan ikke bruke denne direkte, siden dette inneholder hele datoen og tidspunktet. Bruk de to første tegnene fra `Order Date` og legg disse i en ny kolonne `Month`.

   (b) Det foregående punktet vil gi oss ei kolonne med månedsnummeret som strenger, som ikke er en fornuftig representasjon. Prøv å skrive kode for å konvertere dette til tall i stedet (`pd.to_numeric()` er en fornuftig metode å se på her). Du vil møte på en error; å fikse denne er neste deloppgave.

   (c) Du vil finne at en kan ikke konvertere hele kolonna i b) til tall, fordi det finnes rader med verdier som ikke lar seg konvertere. Finn i disse radene og fjern dem. Merk at det kan være lurt å resette indeksen igjen når en fjerner rader. Inkluder svaret du hadde i b), slik at du endelig får endret datatypen til tall.

   (d) For å videre gjøre det mer oversiktlig for oss selv, vil vi ha en kolonne vi kaller `Sum`, som inneholder summen for for hver ordrelinje (antall, multiplisert med pris). Legg til denne kolonnen.

   (e) For å få ut en oversikt over hvilke måneder som hadde størst inntekt, må vi guppere på hver måned (`groupby()`) og summere. Gjør dette, og sorter deretter på kolonnen `Sum`, med de høyeste tallene først.

   (f) Skriv dataramma til ei ny fil.