Milepæl 2 - Big data

Ola Wethal Petersen

September 2020

1 Teoridel

Hva er en Dataframe?

En dataframe er et dataset organisert i kolonner. Et dataset kan "puttes" inn i en dataframe og vi kan så definere de forskjellige kolonnene som tilhører dataen. Basert på språket som implementerer dataframen og datasettet kan det også hende at innholdet kan inferes, altså at dataframen klarer å finne ut hvilke kolonner som har hvilket navn selv.

Hva er forskjellen på et dataset og en dataframe?

Et dataset er da dataen selv, som er umodifisert. Dette kan bety at dataen kan være vanskelig å jobbe med, uten at vi setter opp en schema for dataen. Om dette er et problem eller ikke spørs på hvordan dataen er satt opp og om den er strukturert eller ikke.

Å legge dataen inn i en frame gjør det rett og slett enklere for oss å utføre arbeid med dataen. Når vi har en frame kan vi enkelt definere kolonner og datatyper. Ved å legge det inn i en frame kan vi også kanskje utføre jobber med dataen raskere enn vi kanskje vill gjort originalt. Dette spørs selvfølgelig helt på hvor stor datamendgen faktisk er og hva slags arbeid som skal utføres.

Hva er en partisjon?

Har partisjonering noe å si hvis vi bare har en maskin?

2 Programmering

Jeg bruker bitcoin_cash_price. csv som er tatt fra Crypto currency price history datasettet. Datasettet kan finnes på linken `https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory?select=bitcoin_cash_price.csv` men blir også vedlagt i innleveringen.

Mine 5 spørringer:

- Hva er den høyeste registrerte market capen?
- Hva er differansen mellom den høyeste highen og den laveste highen?
- Hvilken dato hadde den høyeste åpningskursen? Hvordan så resten av dataen ut for den dagen?
- Hvilken dato hadde den laveste åpningskursen? Hvordan så resten av dataen ut denne dagen?
- Hvilken dag hadde den største differansen mellom høyeste high og laveste low?

2.1 Hva er den høyeste registrerte market capen?

- **Hvorfor er dette interessant?**
  Dette er interessant å vite siden dette forteller oss hvilken dag Bitcoin Cash hadde sin høyeste samlede markedsverdi. Ved å se på annen data fra denne dagen kan det hjelpe oss med å forutse eller forstå fremtidige trender.
- **Antagelser for at denne spørringen skal være gyldig**
  For at denne spørringen skal være gyldig må vi anta at datasettet har korrekte verdier og at datasettet inneholder data som er samlet over lang tid. Hvis vi fant høyeste market cap over den siste uka kan det gi oss et bilde om hvordan valutaen ligger ann nå, men det gir oss hele bildet om markedsverdien sett fra et historisk perspektiv.
- Skjermdump av plans
- Annen implementasjon

2.2 Hva er differansen mellom den høyeste highen og den laveste highen?

- **Hvorfor er dette interessant?**
  Jeg syntes dette er interessant å vite for da kan jeg se hvor stor varianse det har vært i bitcoin cash sin verdi. Dette gir et innblikk i hvor mye som kan skje med noe så "sensitivt" som en kryptovaluta over tid.
- **Antagelser for at denne spørringen skal være gyldig**
  Jeg har ingen spesifikk antagelser for denne dataen annet enn at dataen må stemme for at spørringe skal gi oss noen verdifull info.
- Skjermdump av plans
- Annen implementasjon

2.3 Hvilken dato hadde den høyeste åpningskursen? Hvordan så resten av dataen ut for den dagen?

- **Hvorfor er dette interessant?**
  Ved å finne datoen med den høyeste åpningskursen kan vi se på perioden som bygget opp til denne datoen. Med denne infoen kan vi bruke tidligere trender til å hjelpe oss med å forutse lignende peaks i fremtiden. Ved å finne datoen og da all tilhørende data for den dagen kan vi bruke dette for eventuelle fremtidige kjøp.
- **Antagelser for at denne spørringen skal være gyldig**
  Ingen spesielle antagelser bortsett fra de nevnt i forrige spørring.
- Skjermdump av plans
- Annen implementasjon

2.4 Hvilken dato hadde den laveste åpningskursen? Hvordan så resten av dataen ut denne dagen?

- **Hvorfor er dette interessant?**
  Det omvendte av det som ble beskrivet i forrige spørring. Ved å se på dato fra dårlige dager og tidsperioden rundt denne dataoen kan vi finne trender som fører til dårlige perioder for valutaen.
- **Antagelser for at denne spørringen skal være gyldig**
  Ingen spesielle antagelser bortsett fra de som er nevnt tidligere.
- Skjermdump av plans
- Annen implementasjon

2.5 Hvilken dag hadde den største differansen mellom høyeste high og laveste low?

- **Hvorfor er dette interessant?**
  Det er interresant å se variasjonsbredden mellom høyeste og laveste verdi.
- **Antagelser for at denne spørringen skal være gyldig**
  Antar at all data er korrekt.
- Skjermdump av plans
- Annen implementasjon