FYS103 - Måleteknikk og dataanalyse, Labøvelse 2

Dataanalyse og Principal Component Analyse

Lars Berggren og Jorid Holmen

16. januar 2022

Sammendrag

I denne labøvelsen har vi sammenlignet data med informasjon fra 153 land i verden. Målet med øvelsen er å finne ut om det er mulig å gruppere land etter regimetype eller kontinent, i tillegg til andre interessante korrelasjoner blant variablene i datasettet. Vi gjennomfører dette ved hjelp av en principal component analyse. Vi fant ut at det er høy korrelasjon mellom regimetype og politikk i landene. Land med for eksempel fullt demokrati og ufullstendig demokrati får bedre resultater på politisk kultur, sivile friheter og fungerende regjering. Vi fant også en korrelasjon med GDP per innbygger og første prinsipalkomponent.

1 Innledning

Verden i dag består av mange forskjellige samfunn, hvor samtlige blir stadig utsatt for mer og mer globalisering. Dette fører til likheter mellom land som tidligere har vært svært forskjellige.

I denne rapporten foreligger det et datasett som tar for seg ulike variabler som omhandler politikk, kriminalitet og alkoholvaner i 153 ulike land. Vi ønsker å se om det er mulig å bruke disse variablene til å gruppere nasjoner etter kontinent eller regimetype. Det vil si, har geografiske forhold eller politiske forhold en innvirkning på hvordan hvert land scorer på de ulike variablene? Videre vil vi se på om det er korrelasjon mellom de ulike variablene fra dataen som er blitt innhentet. Dette gjør vi ved hjelp av principal component analyse.

2 Teori og metode

2.1 Principal Component Analyse

Principal Component analyse (PCA) er en måte å analysere store mengder med data. PCA reduserer dimensjonene av et stort datasett, ved å transformere de til mindre datasett som fortsatt inneholder mesteparten av informasjonen.

Grunnen til at vi bruker PCA, er for å observere trender, hull, grupper og liknende som er utenfor normen.

2.2 Datasettet

Datasettet som brukes i denne PCA-analysen inneholder informasjon om 153 land i verden, som er fra alle de 6 levedyktige verdensdelene på jorden. Landene har også fått klassifisert hvilken regimetype de har, delt inn i 4 ulike grupperinger: fullt demokrati, mangelfullt demokrati, hybrid regime og autoritært regime. Kontinentet de tilhører og regimetypene er metaene i denne analysen. Datasettet inneholder en score mellom 0 og 10, innenfor kategoriene:

- Rapportert alkoholforbruk
- Urapportert alkoholforbruk
- Prosentvis salg av; øl, vin, brennevin og andre alkoholtyper
- Demokrati score
- Valgprosessen
- Hvor velfungerende regjeringen er
- Politisk deltakelse

- Politisk kultur
- Borgerrettigheter i landet
- Kriminalrate
- GDP per innbygger (Inneholder ikke en verdi mellom 0 og 10)

Dataene er hentet fra ulike Wikipedia-artiklene [2], [3], [4] og [5]. GDP-verdiene er basert på tall fra World Economic Outlook databasen, som ble hentet i oktober 2021. Dataene som omhandler alkohol er blitt hentet fra en WHO publikasjon fra 2018. Dataene ble innarbeidet i 2016. Disse dataene tar for seg alkoholvaner for de som er 15 år og eldre.

Dataene for de politiske variablene, som for eksempel demokrati score eller politisk kultur, er blitt innhentet av Economist Intelligence Unit (EIU). Dette er en forskningsavdeling for britiske Economist Group, selskapet som publiserer bladet The Economist". Disse tallene oppdateres med jevne mellomrom, siste oppdatering var i 2020. Framgangsmåten består av et vektet gjennomsnitt av svarene på 60 spørsmål, som i stor grad er blitt besvart av eksperter, men også noe blir besvart av folket i den aktuelle nasjonen som undersøkes. EIU har også definert hvilken regimetype hver enkelt nasjon har, basert på bruk av eksperter innenfor fagfeltet.

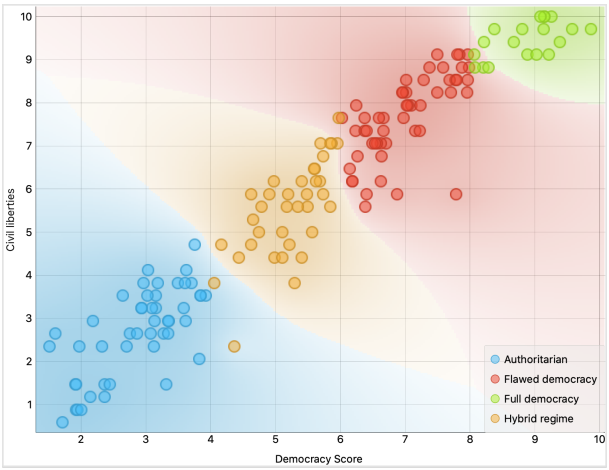
2.3 Gjennomføring av PCA

For å gjennomføre PCA ble datasettet lastet opp i programmet Quasar. Der ble det gjennomført 4 forskjellige PCA basert på alkohol, politikk, kriminalitet og alt samlet. For å gjennomføre dette på bare spesifikke temaer, må det velges spesifikke kolonner fra datasettet. Før PCA blir gjennomført må dataen deles på standardavviket, slik at de forskjellige variablene får riktige størrelser i forhold til hverandre.

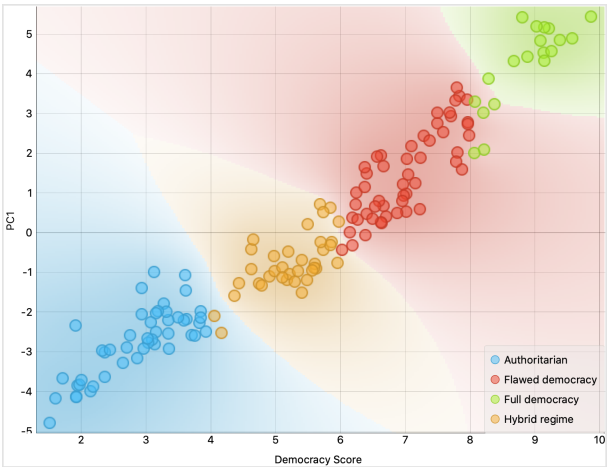
Resultatene av PCA kan enten plottes i et score-bilde, eller så kan det overføres til Excel for å lage et loading image. Begge deler blir gjort i denne oppgaven, for å få et mer helhetlig bilde.

3 Resultater

Ved å sortere all data etter regimetype oppdages det mest interessante resultater ved å sammenligne demokrati score med borgerrettighetene i landet (figur 1), og PC1 (figur 2). Ladningsplottet til PC1 kan man se i figur 3. Det var mye korrelasjon med alle variablene som handlet om politikk, slik som valgprosessen, hvor velfungerende regjeringen er, politisk deltakelse og politisk kultur.

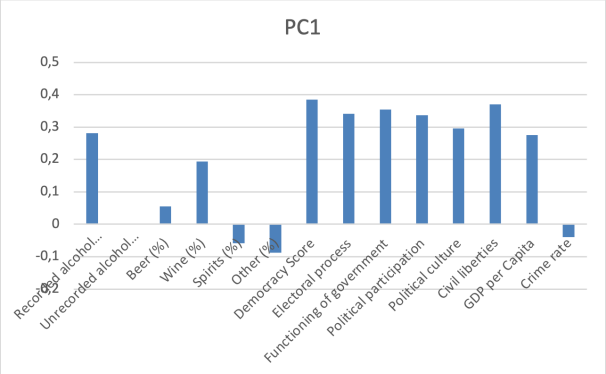


Figur 1: PCA av "Democracy score" sammenlignet med "Civil liberties". Fargene er forskjellige regimetyper.



Figur 2: PCA av "Democracy score" sammenlignet med PC1. Fargene er forskjellige regimetyper.

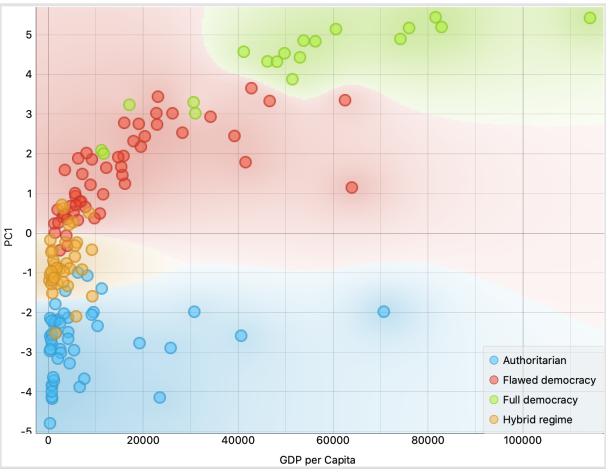
I tillegg til politikk ser man en korrelasjon mellom GDP per innbygger og PC1, slik som i figur 4. Sammenligner vi de to ser man at de med høyere GDP per innbygger har veldig høye tall på alt som omhandler politikk, og lavere tall på



Figur 3: Ladningsplottet til PC1

kriminalitet og urapportert alkoholforbruk, med unntak av noen få land. Landene i denne gruppen er alle sammen fullt demokrati.

Det er ingen åpenbar korrelasjon mellom kontinentene.



Figur 4: PCA av "GDP per capita" og PC1.

4 Diskusjon

Resultatene viser at det er korrelasjon når en sammenligner to politiske variabler etter regimetype, mens det ikke blir like tydelig når en sorterer dataene etter kontinent. Som nevnt i avsnitt 2.2, var det EIU som innhentet disse dataene basert på ekspertuttalelser og meninger fra folket. En antakelse kan være at ekspertene som svarte på spørsmålene, var de samme ekspertene som kategoriserte de ulike landene etter regimetype. Derfor vil det heller ikke være overraskende at det er korrelasjon mellom de politiske variablene, sortert etter regimetype, hvis regimetypen ble bestemt eller fikk innflytelse fra svarene på spørsmålene.

Det eksisterer nok en enighet blant folk flest, både i Norge og verden, at et fullt demokrati er designet for å score høyt på de politiske variablene i denne undersøkelsen, sammenlignet med et mer autoritært regime. Derav vil det ikke være overraskende at det er korrelasjon mellom de politiske variablene når en sorterer de etter regimetype. Derfor er det forståelig at det eksisterer en korrelasjon, men samtidig kan måten dataene har blitt innhentet og definert på, muligens gjort at korrelasjon er høyere enn den egentlig er.

Samtidig kan en stille spørsmålstegn ved at det ikke eksisterer noen særlig korrelasjon når en sorterer dataene etter kontinent. Dette kan ligne på et resultat av et mer globalisert samfunn.

Korrelasjonen mellom GDP per innbygger og PC1 viser at formuen til innbyggerne har veldig mye å si på hvordan politikken i landet er. Samtidig kan man også tenke at land med bra politikk får høyere formue. Basert på resultatene, kan begge teorier være sanne.

4.1 Feilkilder

Det er noen av dataene hvor det er ukjent når de ble registrert. Mye av dataene er også fra forskjellige år. Dette fører til mye usikkerhet. Om vi sammenligner data fra forskjellige år risikierer vi at landet har utviklet seg etter at dataen ble registrert, og at de nå for eksempel har en annerledes politisk situasjon.

Noen land har korrupt regime, og der kan det være typisk at informasjonen blir registrert feil [1]. Dette kan også føre til feil i sammenligningen.

5 Konklusjon

Målet med labøvelsen var å se om vi kan bruke variablene til å sortere land etter regimetype eller kontinent, samt om vi kunne finne andre interessante korrelasjoner. Variablene kan definitivt brukes til å sortere landene etter regimetype, men ikke etter kontinent. Dette tilsier at politikk er viktigere enn geografi, når det kommer til klassifisering av land. I tillegg kan politikk har

en innflytelse på GDP per innbygger, eller GDP har en innflytelse på politikken. Videre forskning kreves for å fastsette hvilken av teoriene som gjelder.

Referanser

[1] Luis Martínez. How much should we trust the dictator's gdp growth estimates?, 2021. [Online; accessed 15. 01, 2022].

[2] Wikipedia contributors. List of countries by past and projected gdp (nominal) per capita, 2021. [Online; accessed 15. 01, 2022].

[3] Wikipedia contributors. Democracy index, 2022. [Online; accessed 15. 01, 2022].

[4] Wikipedia contributors. List of countries by alcohol consumption per capita, 2022. [Online; accessed 15. 01, 2022].

[5] Wikipedia contributors. List of countries by intentional homicide rate, 2022. [Online; accessed 15. 01, 2022].