Rapport

Statistikkprosjekt

Del A

Av:

Gruppe 3, 2. klasse, bachelor dataingeniør ved NTNU Kalvskinnet

Frode Halvorsen

Markus Koteng

Oddbjørn Olsen

Kim Røstgård

Innholdsfortegnelse

Del A

Teori

I prosessen med med bearbeiding av de innsamlede dataene har vi brukt programmeringsspråket; “Python 3” sitt innebygde funksjonsbibliotek for prosessering av dataene, der vi har valgt å avrunde til ett desimals nøyaktighet ettersom datagrunnlaget bare hadde ett desimals nøyaktighet. Funksjonene i bibliotekene baserer seg på følgende formler fra læreboka:

Sentralmål

Gjennomsnitt

Gjennomsnittet er summen av alle tallverdiene delt på antall tall som er summert.

$$\frac{(a_1 + a_2 + ... + a_n)}{n}$$

Median

For et utvalg der antall observasjoner er et oddetall, er medianen den midterste verdien der utvalget er sortert i rekkefølge.
For et utvalg der antall observasjoner er et partall er medianen gjennomsnittet av de to midterste verdiene.

n er antall elementer/data

Hvis *n* er et oddetall er median:

$$a_{(n+1)/2}$$

Hvis *n* er partall er median:

$$\frac{a_{n/2} + a_{(n/2)+1}}{2}$$

Modus

Tallverdien som har størst antall observasjoner. Modus kalles også typetall. Dersom det er likt antall observasjoner av 2 eller flere observasjoner med størst antall i et datasett, blir modusen en liste med flere tallverdier.

Spredningsmål

Variasjonsbredde

Variasjonsbredde er forskjellen mellom største observasjon og minste observasjon i datasettet. Variasjonsbredden beskriver bredden på utvalgets histogram.

Variasjonsbredde regnes ut slik

$$variasjonsbredde = største\_observasjon - minste\_observasjon$$

Varians

Varians beskriver hvor stort det gjennomsnittlige kvadratet av (måleverdi - gjennomsnittlig måleverdi) er. Dette sier noe om i hvor stor grad en gjennomsnittlig måling avviker fra gjennomsnittet av målingene.

Varians regnes ut slik:

$$Var = \sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standardavvik

Standardavvik er et mål for spredningen av verdiene i et datasett eller av verdien av en stokastisk variabel. Standardavviket gir verdienes gjennomsnittlige avstand fra gjennomsnittet.

Standardavvik regnes ut slik:

$$\sigma = \sqrt{Var}$$

Beskrivelse av datasett

Datasett 1 - Tidsserie fra pålitelig kilde fra internett

Værobservasjoner ble hentet fra nettsiden; www.sharki.oslo.dnmi.no. Dataene viser gjennomsnittlig nedbør i januar måned, per år. Rapporten vi valgte å hente fra nettsida var: månedlig, månedlige verdier, tidsperiode 1890 - 1994, nedbør (mm) - rapportmal, for Hakloa i Nordmarka (Oslo).

Siden det er snakk om mange målepunkter over tid, har vi valgt å bruke linjediagram for dette datasettet.

Datasett 2 - Parkerte biler

Data ble samlet inn 31.01.18 på vei hjem fra NTNU Kalvskinnet.
Telling av biler ble gjort ved observasjon av bilskilt, fra Ilen Kirke og nedover mot Ilsvika. Tellingen startet kl. 1510 - og varte til 1540.

Vi valgte å bruke kakediagram her for å vise hvor stor andel av bilene som var elbiler.

Datasett 3 - Passerte sykler på 30 min

Data ble samlet inn fredag 02.02.18 i trafikkrysset der Kjøpmannsgata krysser Olav Tryggvasons gate. Innsamlingen ble gjort i tidsrommet fra kl. 09.30 til kl. 10.00. Dette ble gjort ved telling.

Siden datainnsamlingen kun resulterte i ett tall (slik vi fikk beskjed om), fant vi det svært vanskelig å lage diagram på denne. Vi valgte å bruke stolpediagram, da det er enkelt å lese av måleverdien på enkeltpunkter på denne.

Datasett 4 - Joggetur

Datasettet er samlet inn fra joggetur som ble utført søndag 04.02.18 ved start- og målpunkt ved trappa på Ilen kirke, og ruten; Voldgata, Erling Skakkes gate, Kalveskinnsgata, Elvegata, og Voldgata. Faktorer som spilte inn i datainnsamlingen: Temperaturen var -4° C, Glatte sko, isdekt vei, snøfall underveis, biler og mennesker i veibanen førte til mindre variasjoner i ruten.

Kl. start: 1321, kl. slutt: 1416. Både tidtaking og observasjon ble gjort av samme person med iPhone 6 - stoppeklokkeapp.

Her valgte vi å bruke linjediagram, siden det er snakk om måleverdi som endrer seg litt over tid.

Datasett 5 - 100 lapper, 50 med kryss

Datasettet ble samlet inn ved at det ble utført totalt 1000 simulerte trekninger ved hjelp av et python-program. I hver trekning ble det generert 100 lapper der 50 av dem hadde kryss på. I hver trekning ble det trekt 10 lapper, og det ble notert ned hvor mange av dem som hadde kryss. Antall kryss per trekning ble samlet opp i en tabell og lagret i csv-format.

Fordi det er en gitt range (0 til og med 10) med mulige utfallsverdier, valgte vi å presentere disse dataene ved hjelp av et histogram. Ved hjelp av histogrammet kan man kjapt se hvor mange trekninger som ga 5 kryss, 6 kryss, osv.

Datasett 6a - Nordmenns kontakt med innvandrere i dagliglivet

Datasettet viser prosentandel nordmenn som har rapportert regelmessig kontakt med innvandrere i nabolaget sitt. Dataene er per år. Dataene er hentet fra SSB sin statistikkbank:

http://www.ssb.no/statbank/sq/10002084/

Siden det er snakk om gradvis endring av måleverdi over lang tid, og vi har mange målepunkter, har vi også her valgt å bruke linjediagram.

Datasett 6b - Nordmenns holdninger til innvandrere

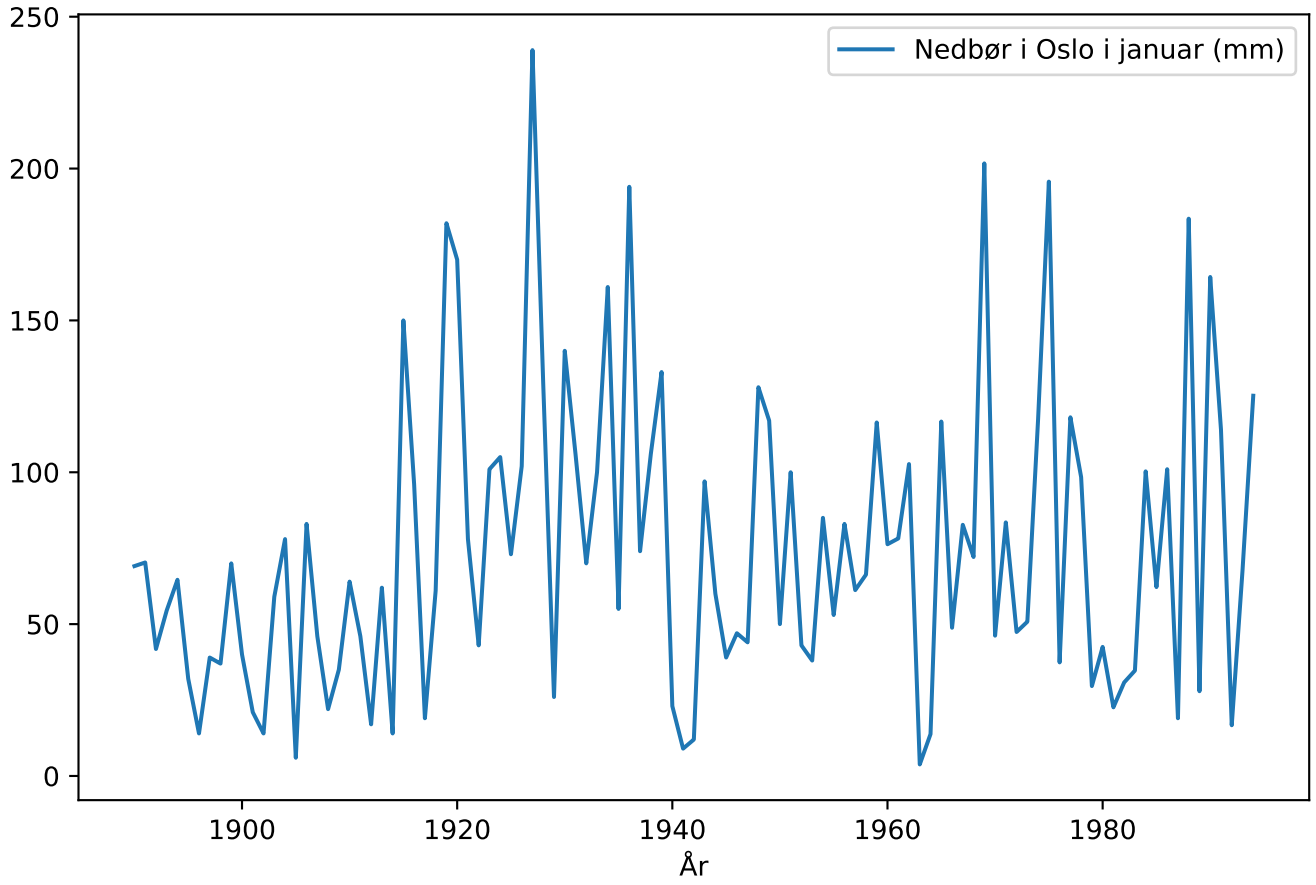
Datasettet viser prosentandel nordmenn som synes det er greit at deres sønn/datter er sammen med en innvandrer. Dataene er per år. Dataene er hentet fra SSB sin statistikkbank:

http://www.ssb.no/statbank/sq/10002083/

Siden det er snakk om gradvis endring av måleverdi over lang tid, og vi har mange målepunkter, har vi også her valgt å bruke linjediagram.

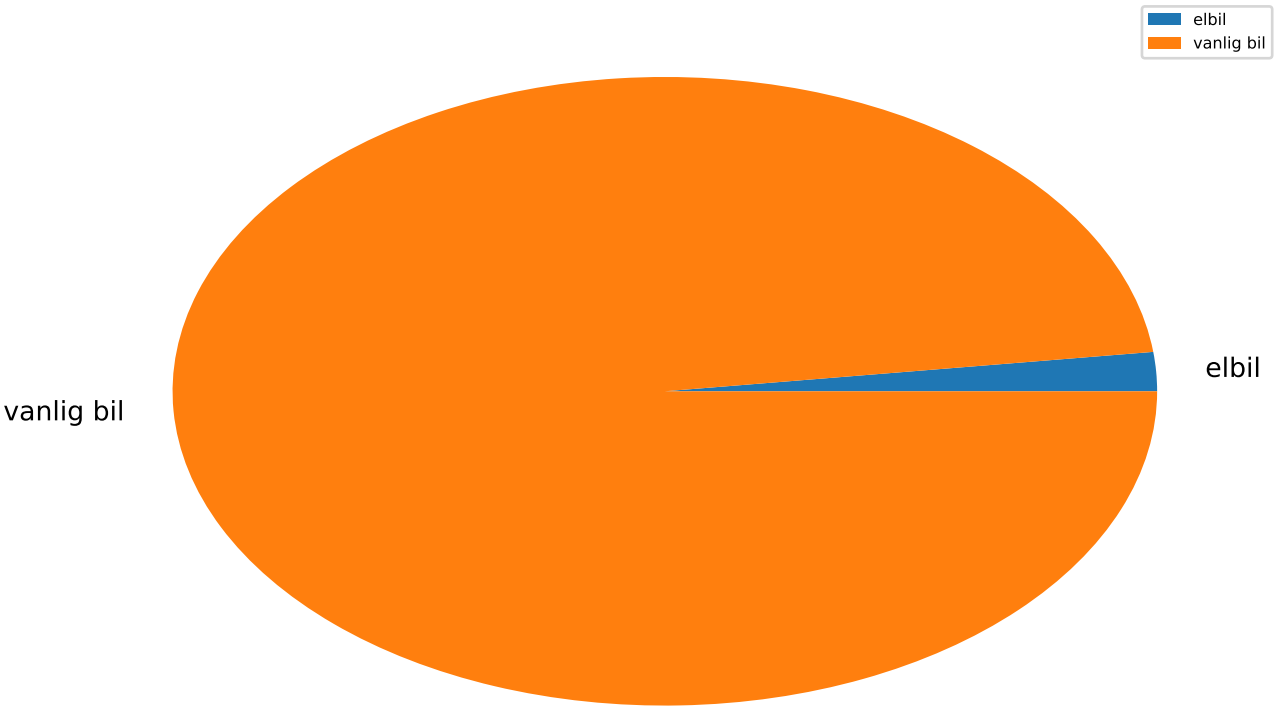
Vedlegg A: Data og utregninger

Datasett 1: Månedlig nedbør i Oslo januar måned



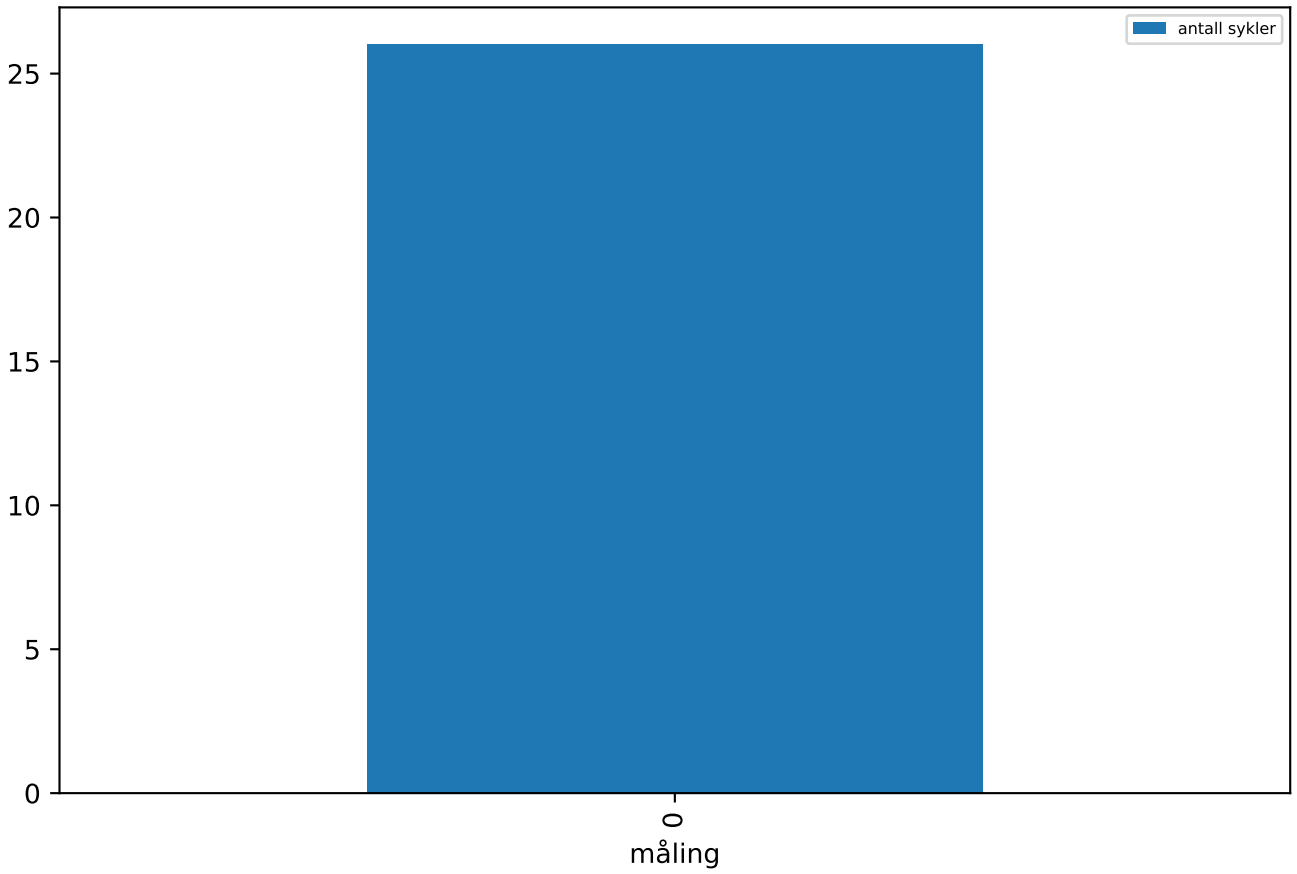
gjennomsnitt	74.5
median	64.6
modus	[14.0]
variasjonsbredde	235.2
varians	2439.2
standardavvik	49.4

Datasett 2: Andel elbiler av parkerte biler (totalt 100)



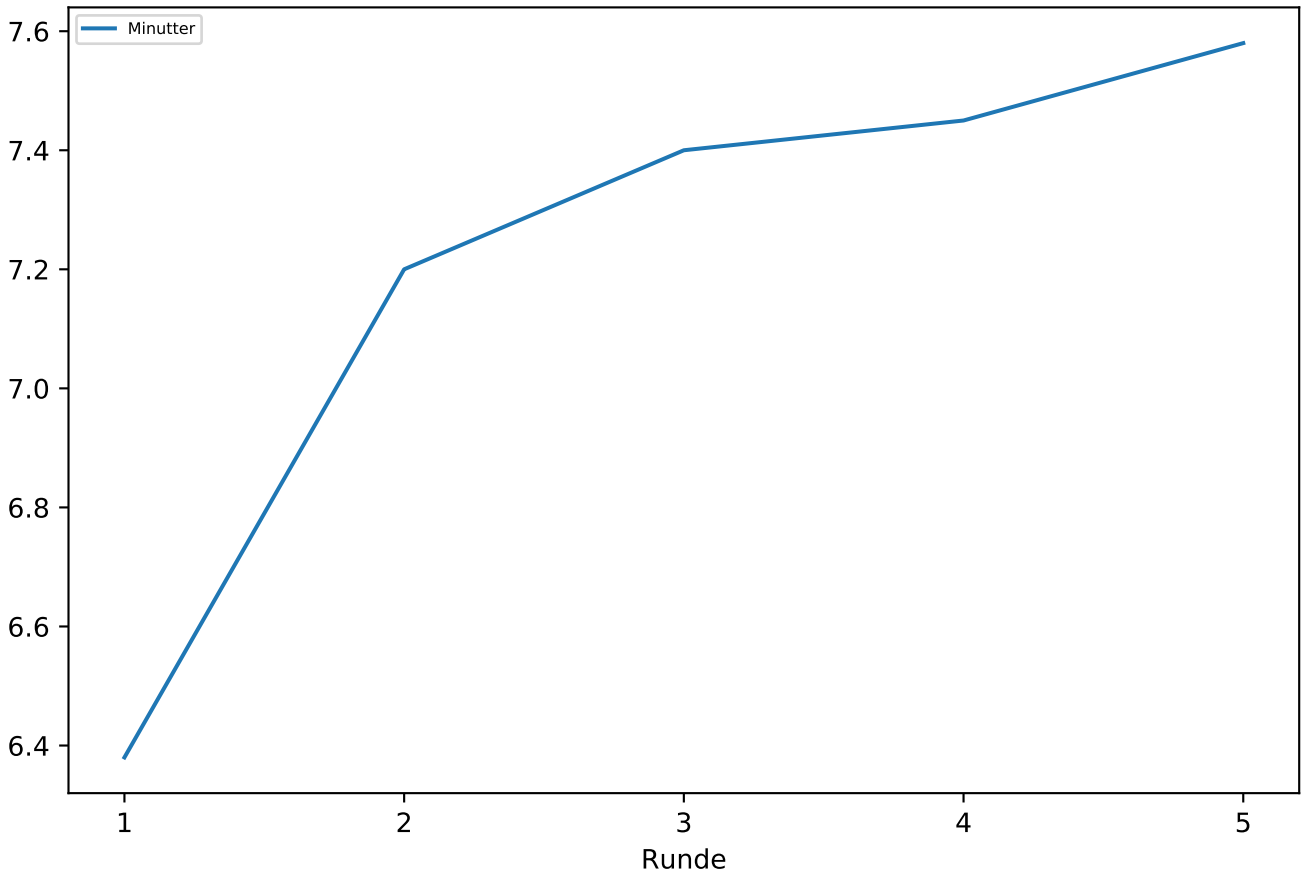
gjennomsnitt	2.0
median	2.0
modus	[2]
variasjonsbredde	0
varians	nan
standardavvik	nan

Datasett 3: Antall registrerte sykler i løpet av 30 min



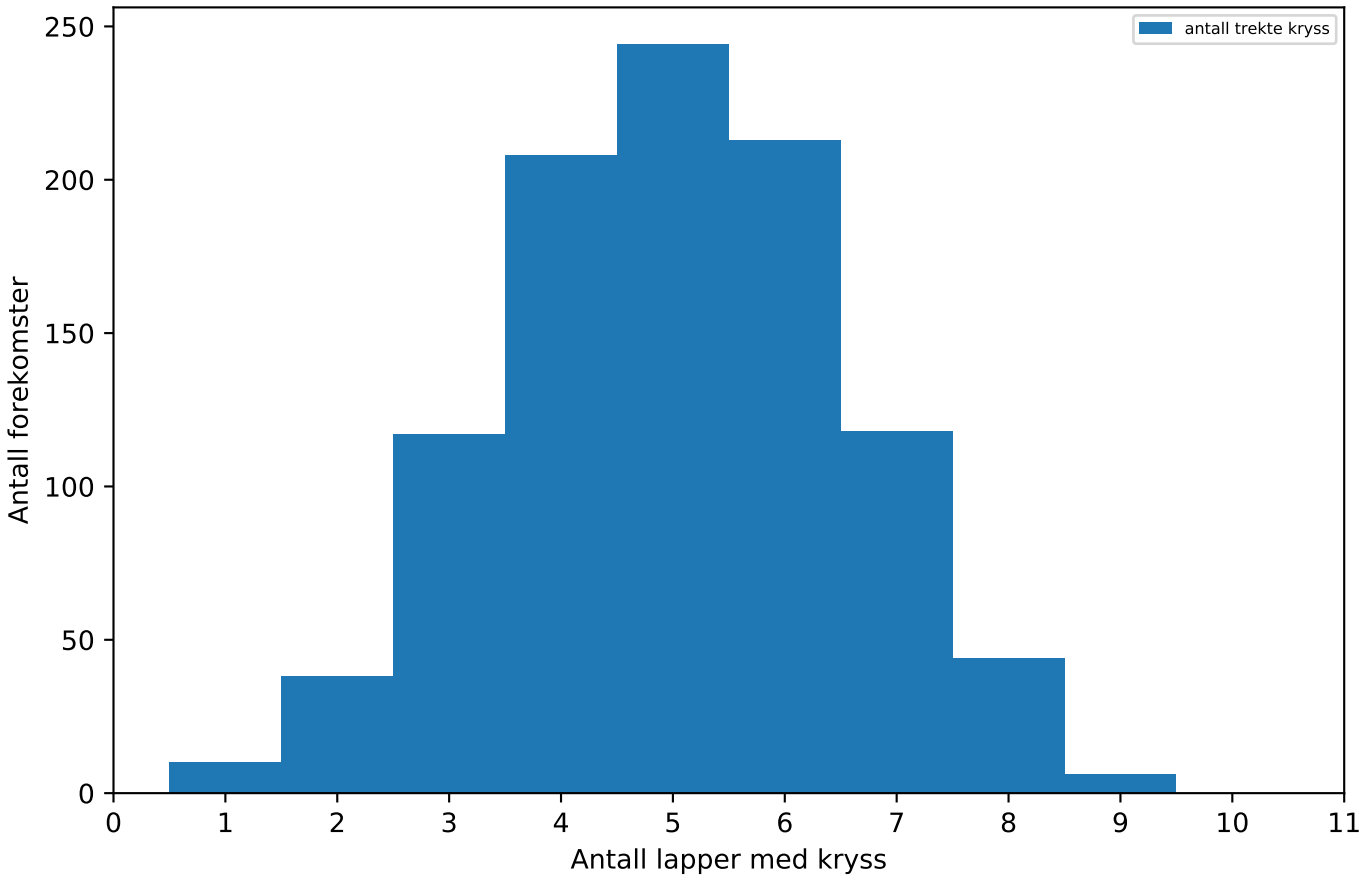
gjennomsnitt	26.0
median	26.0
modus	[26]
variasjonsbredde	0
varians	nan
standardavvik	nan

Datasett 4: Tiden det tar å jogge en runde



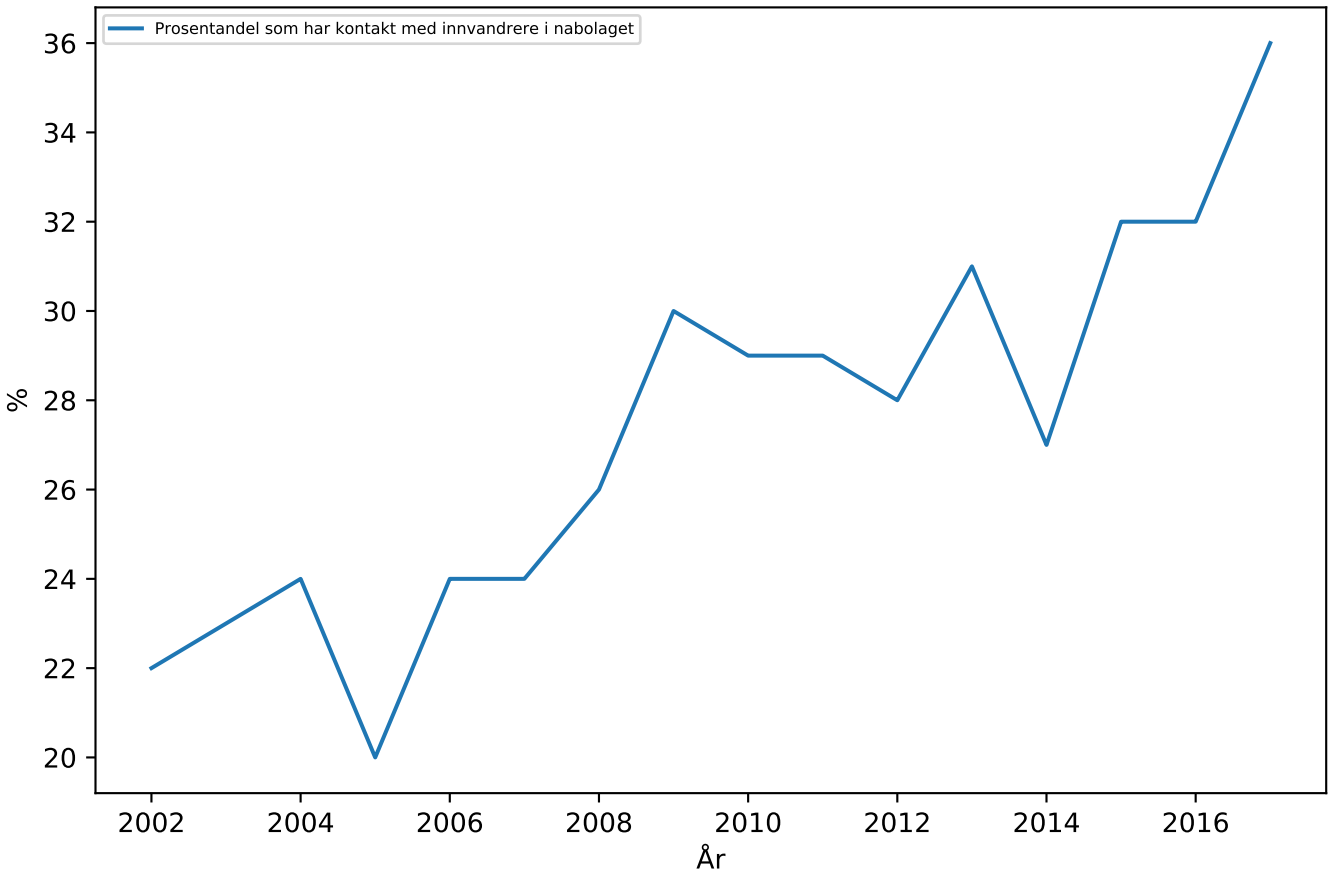
gjennomsnitt	7.2
median	7.4
modus	[6.38, 7.2, 7.4, 7.45, 7.58]
variasjonsbredde	1.2
varians	0.2
standardavvik	0.5

Datasett 5: Antall trekte lapper med kryss
Det ble trekt 10 lapper, av 100 der halvparten hadde kryss



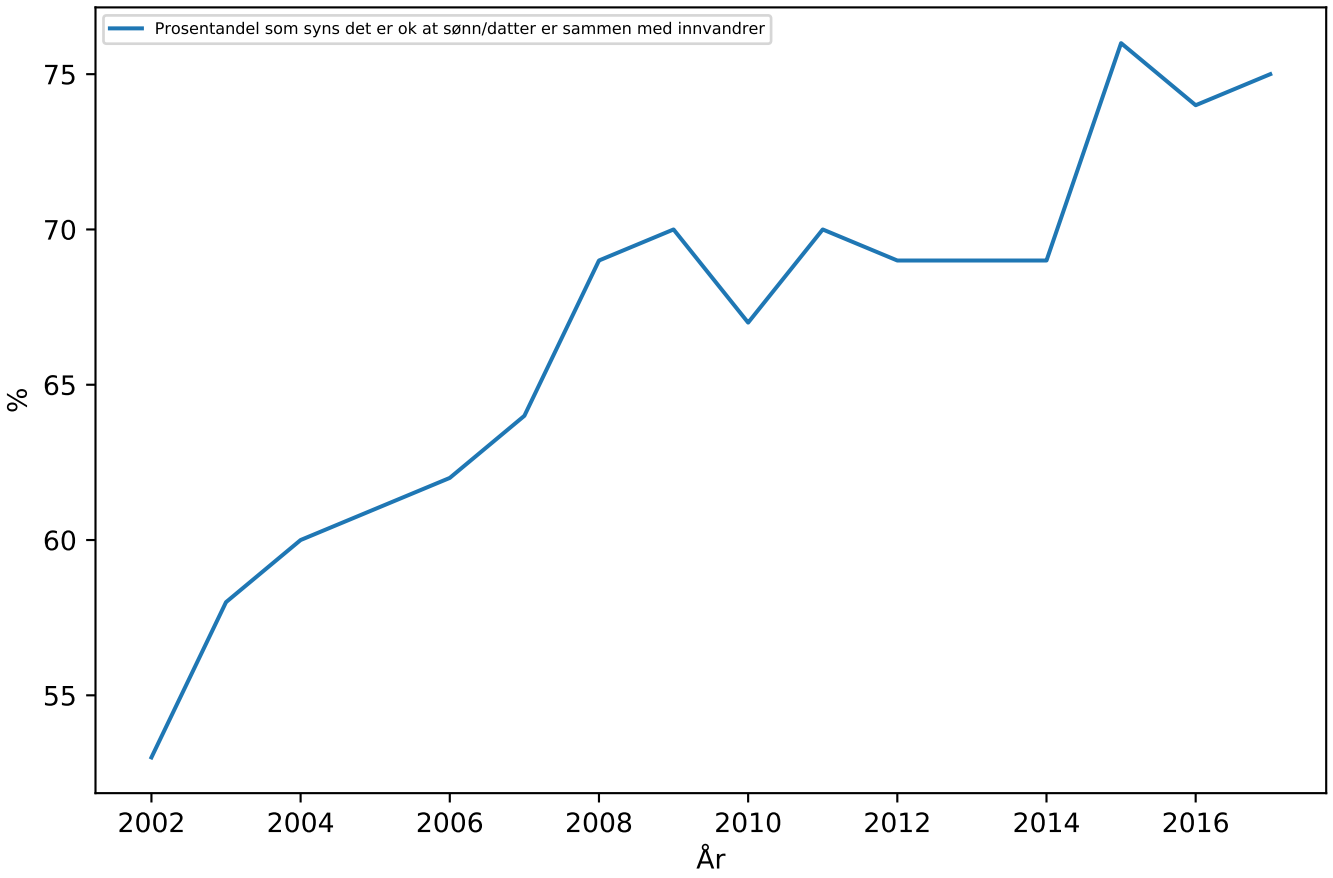
gjennomsnitt	5.0
median	5.0
modus	[5]
variasjonsbredde	9
varians	2.4
standardavvik	1.6

Datasett 6a: Prosentandel nordmenn
som hadde kontakt med innvandrere i nabolaget



gjennomsnitt	27.3
median	27.5
modus	[24]
variasjonsbredde	16
varians	18.8
standardavvik	4.3

Datasett 6b: Prosentandel nordmenn som hadde positiv holdning til at deres sønn/datter var sammen med innvandrer.



gjennomsnitt	66.6
median	69.0
modus	[69]
variasjonsbredde	23
varians	41.4
standardavvik	6.4