KodeKlubben 2.0

Øvelsesgang 2

Kristian Urup Olesen Larsen, Jakob Jul Elben

November 23, 2018

Økonomisk Institut, KU

Velkommen (igen)!

Hvem er vi?

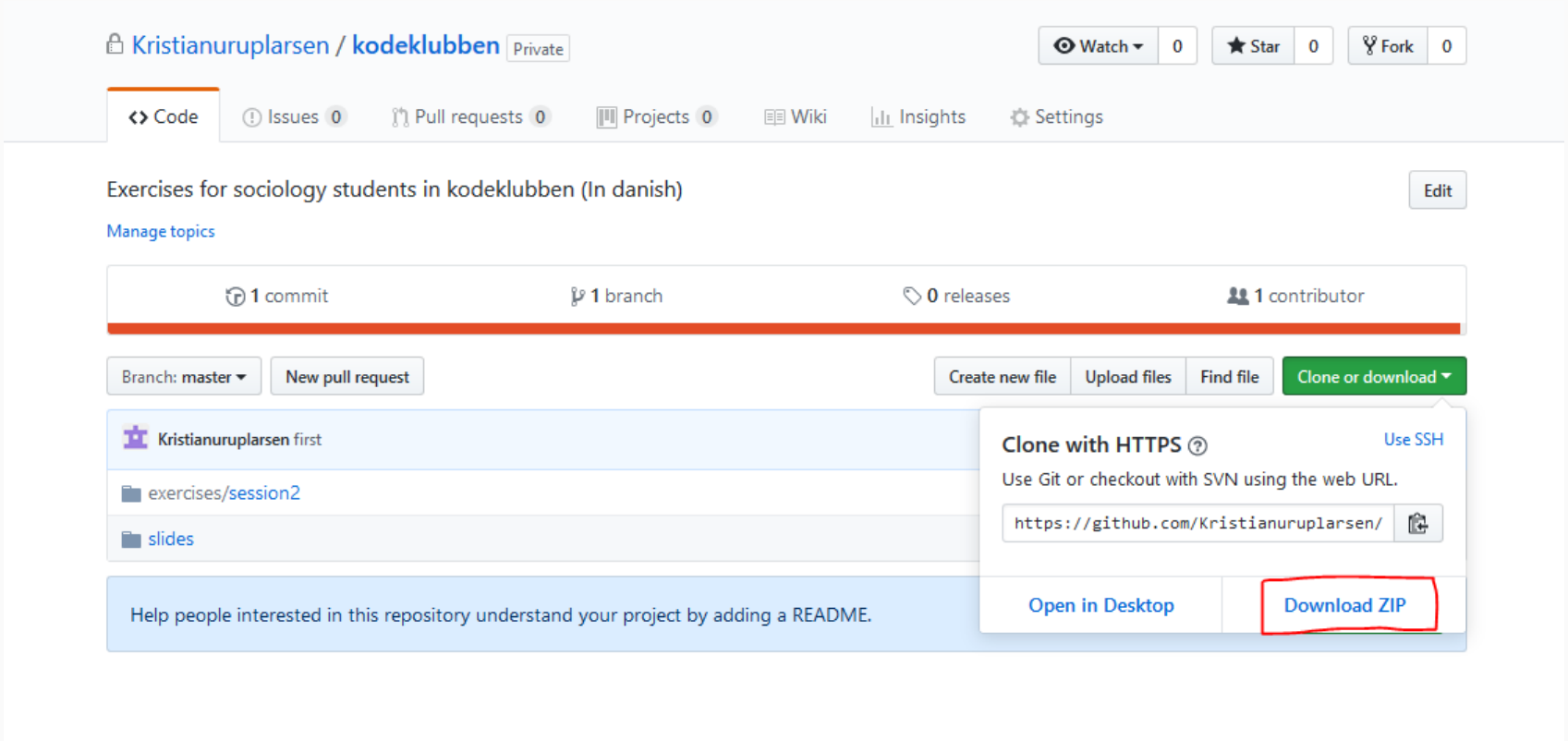
- Økonomistuderende
- RA's på Økonomisk Institut
- Arbejder typisk i Python, R, STATA eller SAS

Hvem er i?

Setup

Alt materiale ligger på GitHub

- Åben https://github.com/Kristianuruplarsen/kodeklubben
- Klik på knappen *Clone or download* og på *Download ZIP*



Sidste gang

Sidste gang downloadede og rensede vi et datasæt over danske kommunalpolitikeres holdninger:

- Åben det datasæt i selv gemte sidste gang, eller
- Hent datasættet online:

```
import pandas as pd

url = 'https://raw.githubusercontent.com/'\
      'Kristianuruplarsen/kodeklubben/master/'\
      'data/candidates.csv'
df = pd.read_csv(url)

```

Projektet

Sidste gang:

- Download data fra DR
  - internet-hacks
  - Interagere med nettet gennem python
- Rens datasættet og gør klar til alt det sjove
  - jonglere med data og formatter

Denne gang:

- Alt det sjove.
  - *Dimensionality reduction*
  - Interaktive plot

Øvelser

Structure data

Load data, enten dem i gemte sidste gang, eller fra github (brug koden fra slides) *Hint:* denne gang skal i bruge pakken `pandas` (installer den med pip hvis i ikke allerede har den) som er pythons dataset-workhorse. Som regel kan man google sig til en løsning med søgeord a-la "pandas load dataset from github", "pandas convert to string" osv.

Structure data

Dette gøres ved

```
import pandas as pd

url = 'https://raw.githubusercontent.com/'\
      'Kristianuruplarsen/kodeklubben/master'\
      '/data/candidates.csv'

df_raw = pd.read_csv(url)
```

Structure data

Lige nu er kandidaternes svar registreret i en enkelt streng af tal mellem 1 og 5. Et 1-tal svarer til at kandidaten har svaret det mest venstrestille ("Meget uenig"), 5 svarer ligeledes til det meste højrestillede ("Meget enig").

Konverter kolonnen 'answers' til en streng-variable og drop de kandidater der ikke har svaret på alle spørgsmålene fra datasættet. Hvor mange observationer mister i?

Skriv dernæst en funktion der spreder answerstring ud over 15 variable (hint: stackoverflow).

Structure data

Dette gøres ved

```
df = df_raw

df['answers'] = df['answers'].astype(str)   # convert to string
df['Ncomplete'] = df.answers.apply(len)     # find no.
                                            # answered questions
df = df.query("Ncomplete == 15").reset_index(drop = True)
            # remove if less than 15 answers are answered

df = pd.concat([df, df['answers'].
    apply(lambda x: pd.Series(list(x)))], axis = 1)

df
```

PCA

Nu har i et pænt, færdigt datasæt. Næste step er at lave den interaktive figur i så til at starte med sidste fredag. For at lave den skal vi igennem et par trin:

- Vi skal have kollapset svarende på de 15 spørgsmål til tre akser - til det bruger vi PCA
- Vi skal have skrevet noget kode der producerer interaktive figurer, i får noget af koden udleveret her, fordi det kræver en hel del ligegyldig kode at sætte op.

PCA

Opgave 2.1 - PCA (machine learning)

Ligesom at pandas er pythons workhorse for datasæt, er scikit-learn goto pakken for machine learning i python. Start med at installere scikit learn. Læs derefter lidt af denne guide til PCA modulet. Forsøg at få et overblik over

- Hvordan man importerer PCA modulet, og hvordan man bruger det
- Hvordan PCA virker - hvad er idéen bag algoritmen?

Estimer en PCA model med 3 komponenter på de 15 kolonner med politikernes svar, jeres output skal have tre kolonner og lige så mange rækker som i har observationer i datasættet.