TALL

SOM FORTELLER

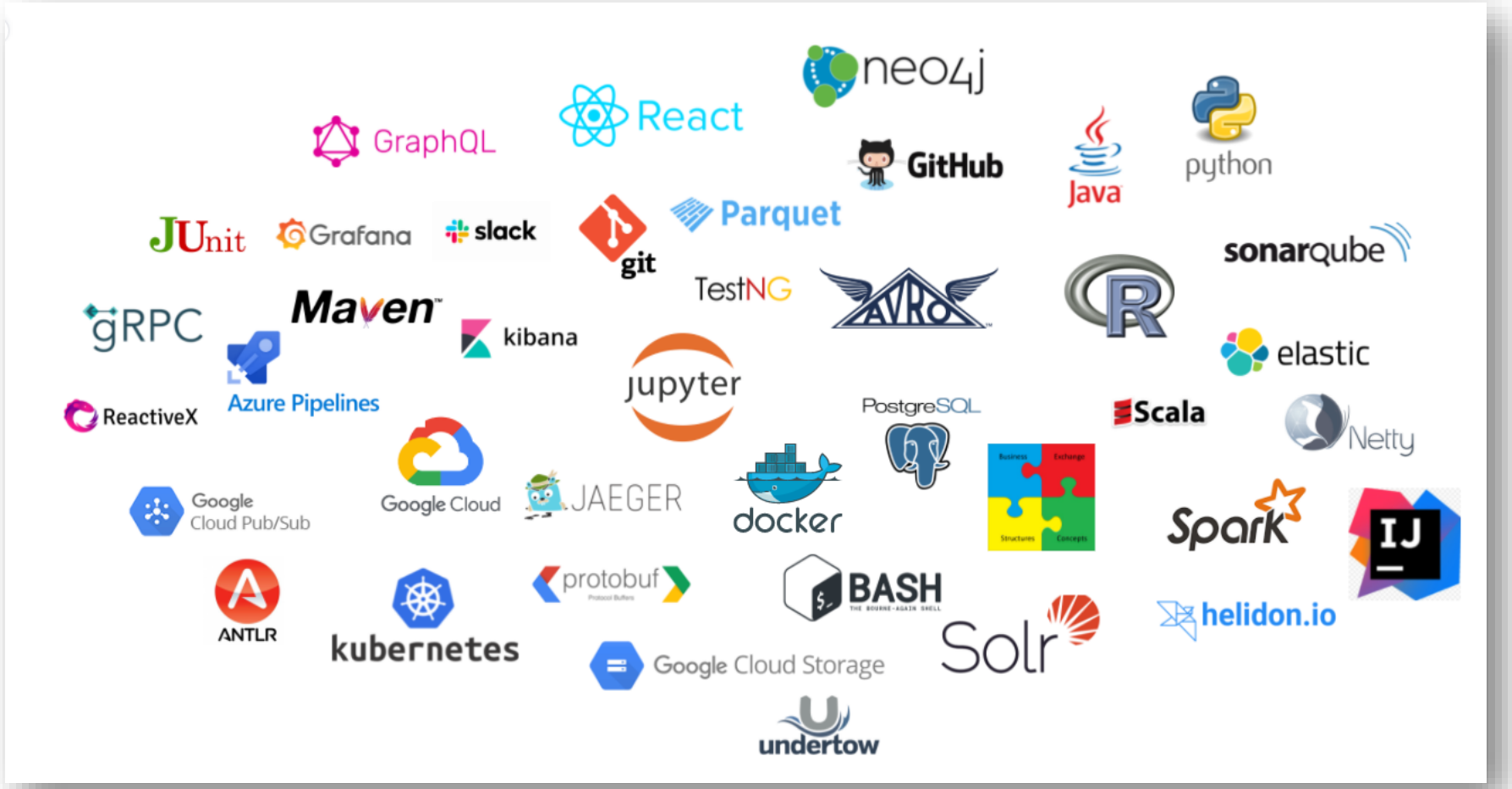
Using R in Jupyter for statistical production

SUSIE.JENTOFT@SSB.NO

Statistisk sentralbyrå
Statistics Norway

Why?

DAPLA



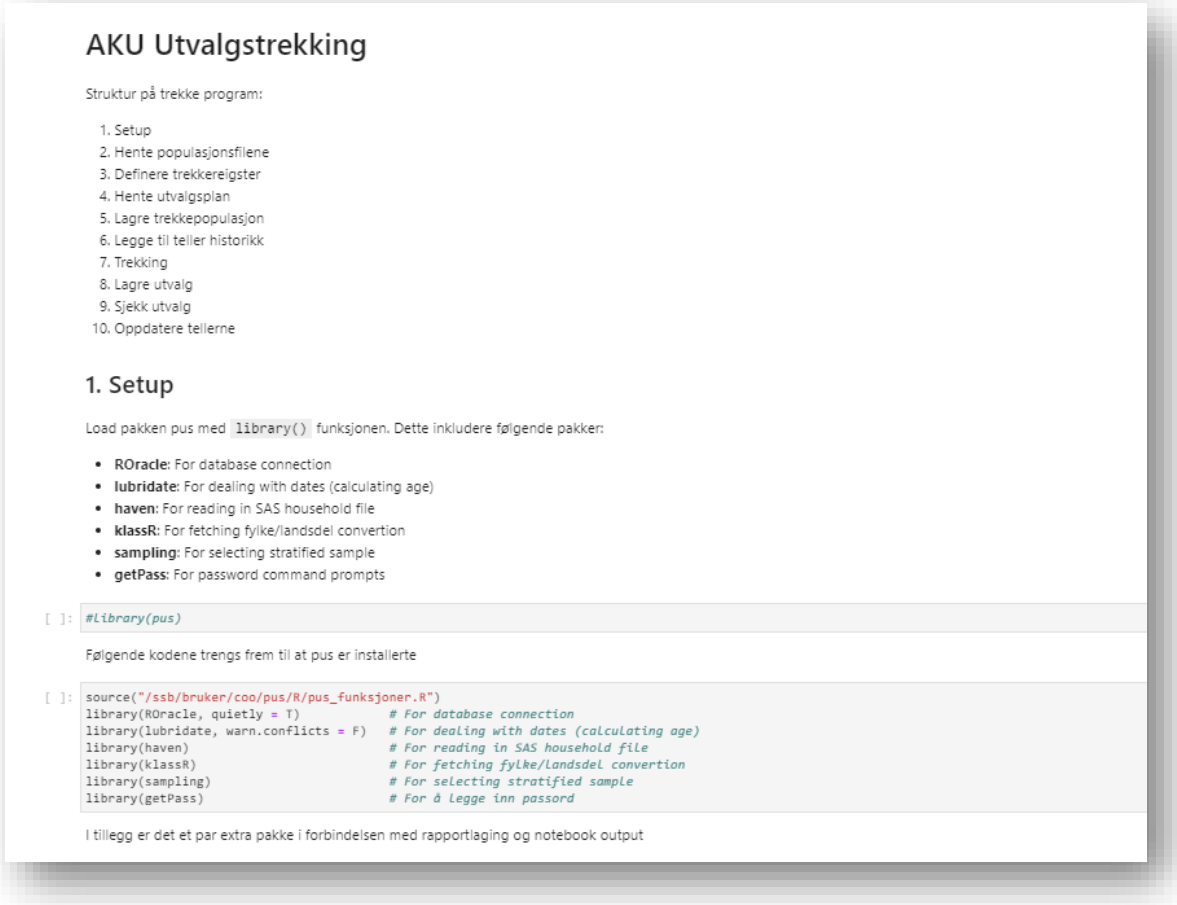
Statistisk sentralbyrå
Statistics Norway

First experiences

- General:
  - Learning R can take a while
  - R beginner and intermediate courses
- Jupyter:
  - Common environment for R & python
  - People mostly liked notebook environment for developing code
  - IT happy(ish)

Example: Sampling for the Norweigan labour force survey

- R–package with functions
- Run code in Jupyter
- Documentation of processes within the notebook.
- Easy to run for someone without coding background



AKU Utvalgstrekking

Struktur på trekke program:

1. Setup
2. Hente populasjonsfilene
3. Definere trekkereigster
4. Hente utvalgsplan
5. Lagre trekkepopulasjon
6. Legge til teller historikk
7. Trekking
8. Lagre utvalg
9. Sjekk utvalg
10. Oppdatere tellerne

1. Setup

Load pakken pus med `library()` funksjonen. Dette inkludere følgende pakker:

- **ROracle:** For database connection
- **lubridate:** For dealing with dates (calculating age)
- **haven:** For reading in SAS household file
- **klassR:** For fetching fylke/landsdel convertion
- **sampling:** For selecting stratified sample
- **getPass:** For password command prompts

```
[ ]: #library(pus)
```

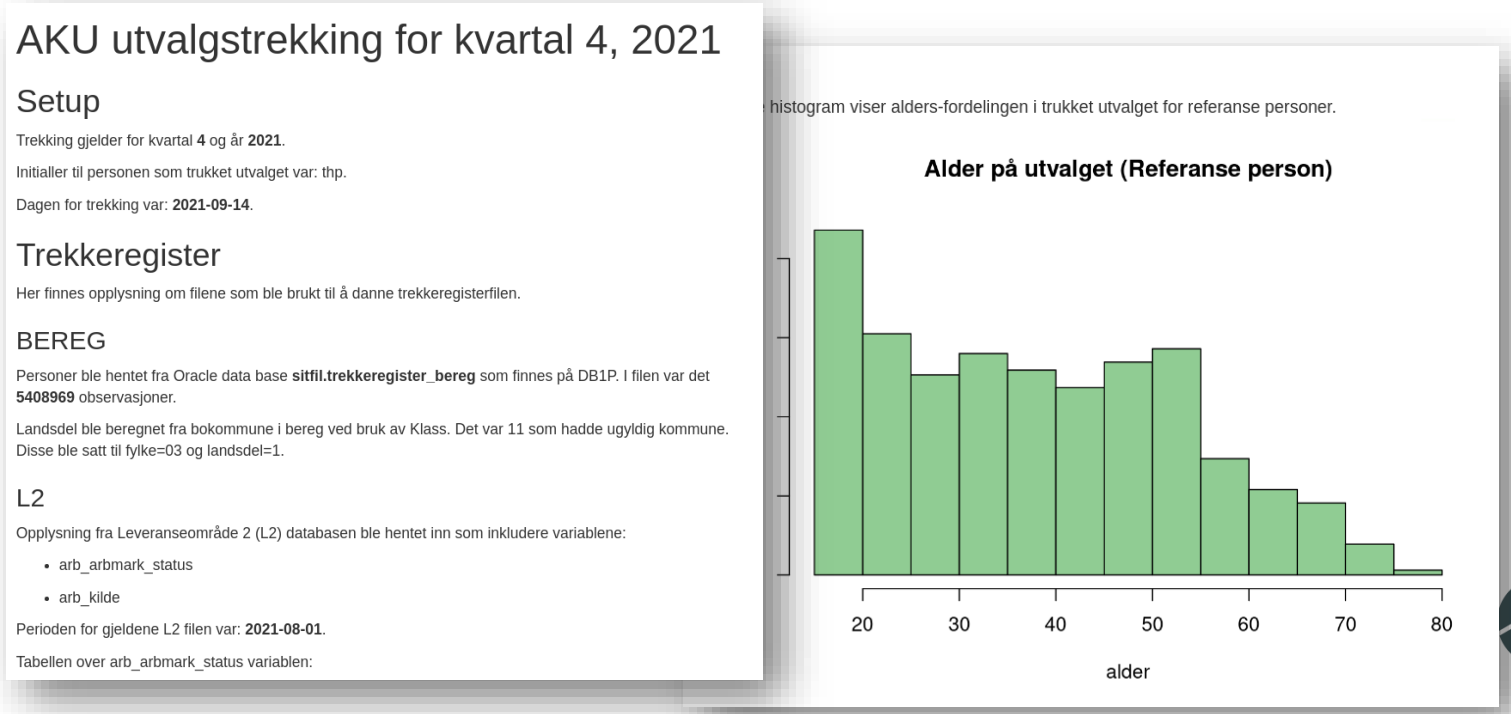
Følgende kodene trengs frem til at pus er installerte

```
[ ]: source("/ssb/bruker/coo/pus/R/pus_funksjoner.R")
library(ROracle, quietly = T)           # For database connection
library(lubridate, warn.conflicts = F)  # For dealing with dates (calculating age)
library(haven)                          # For reading in SAS household file
library(klassR)                         # For fetching fylke/landsdel convertion
library(sampling)                       # For selecting stratified sample
library(getPass)                        # For å legge inn passord
```

I tillegg er det et par extra pakke i forbindelsen med rapportlaging og notebook output

Render documents

- Jupyter documents don’t render well
- Instead we rendered a .rmd fil for the notebook



AKU utvalgstrekking for kvartal 4, 2021

Setup

Trekking gjelder for kvartal **4** og år **2021**.

Initialler til personen som trukket utvalget var: thp.

Dagen for trekking var: **2021-09-14**.

Trekkeregister

Her finnes opplysning om filene som ble brukt til å danne trekkeregisterfilen.

BEREG

Personer ble hentet fra Oracle data base **sitfil.trekkeregister_bereg** som finnes på DB1P. I filen var det **5408969** observasjoner.

Landsdel ble beregnet fra bokommune i bereg ved bruk av Klass. Det var 11 som hadde ugyldig kommune. Disse ble satt til fylke=03 og landsdel=1.

L2

Opplysning fra Leveranseområde 2 (L2) databasen ble hentet inn som inkludere variablene:

- arb_arbmark_status
- arb_kilde

Perioden for gjeldene L2 filen var: **2021-08-01**.

Tabellen over arb_arbmark_status variablen:

histogram viser alders-fordelingen i trukket utvalget for referanse personer.

Statistisk sentralbyrå
Statistics Norway

Production process

- Limited to R or python within the notebook
- Not so easy to run streamlined in a production larger production process (papermill + yaml files)
- No classic data view/ no view of enironment objects
- No dash
- Package creating tools

Conclusion