

R Notebooks og reproduserbarhet

Assignment 1 i MSB105 Data Science - innleveringsfrist 24.09.20

Katrine Hope

Karl-Gunnar Severinsen

Innhold

| | |
|-----------------------------------------------|----|
| 1. Innledning | 2 |
| 2. Teori | 2 |
| 2.1 Replikerbarhet/reproduserbarhet | 2 |
| 2.2 Problemets omfang | 3 |
| 3. Analyse | 5 |
| 3.1 Titanic | 5 |
| 3.2 Eksepel fra forelesning | 7 |
| 3.3 Sessioninfo og koder | 9 |
| 4. Konklusjon | 11 |
| Referanse | 12 |
| Appendiks | 13 |
| Koder | 14 |

1. Innledning

I denne oppgaven ønsker vi å se på reproduserbarhet og viktigheten av det. Vi vil i denne sammenhengen ta oss innom temaet om bruk av “R Notebook” i Rstudio.

Vi vil se på teori rundt reproduserbarhet og R Notebook, som vi også vil knytte opp til en analyse.

2. Teori

2.1 Replikerbarhet/reproducerbarhet

Det er ønskelig at vitenskapelige oppdagelser og fremskritt skal være robuste og pålitelige fordi vi ønsker å ha tillit til at resultatene er riktig og at undersøkelsene er gjort på en tilfredsstillende måte (Bollen et al., 2015). I følge McNutt (2014) er det viktig for at vitenskapen skal utvikle seg at funnene baserer seg på troverdige funn. Mange forskere mener at reproduserbarhet er en viktig fremgangsmåte for å kunne validere funnene sine, men dette har vist seg å være veldig vanskelig å gjennomføre (McNutt, 2014). Noe av det som kan være med å gjøre det vanskelig å reproducere, men også replikere, er at tidligere studier kan mangle data eller koder, det kan være feil i programvaren som er sendt med, det kan mangle dokumentasjon, samt er det noen forskere som ikke ønsker å gi fra seg sine data og koder på studiene sine. National Science Foundation i USA viser til at for at vitenskap skal være robust og pålitelig må funnene være reproduserbar, replikerbart og generaliserbart (Bollen et al., 2015). I henhold til National Science Foundation blir de 3 påstandene definert som (Bollen et al., 2015):

1. Reproduserbarhet:

- Å reproducere vil si å gjøre studiet på nytt, med de samme dataene og med den samme metodikken, men gjort av en uavhengig part.

2. Replikerbarhet:

- Å replikere vil si at et studie gjøres på nytt av en uavhengig part, gjerne med nye data og at resultater og konklusjon er lik som tidligere.

3. Generaliserbarhet:

- Vil si at et studie kan gjøres på nytt med ny data og ny sammensetning, men komme frem til lignende konklusjoner.

Reproduksjon har potensiale til å være en minimumsstandard i følge Peng (2011), når full uavhengig replikasjon ikke er tilgjengelig. Dersom den nye studien kan bekrefte resultatene og konklusjonen fra et tidligere studie via replikasjon blir dette sett på som en vitenskapelig gullstandard (Jasny et al., 2011). Peng (2011) anser replikasjon som den ultimate standarden i likhet med Jasny et al. (2011).

2.2 Problemets omfang

Som forklart tidligere så kan tidligere studier mangle data, koder, fremgangsmåte, feil i programvare osv, samt er det flere forskere som ikke ønsker å gi fra seg all informasjon om sine studier. Dette gjør at det er nærmest umulig for forskere i senere tid å reprodusere en tidligere studie. Siden det er mer enn et problem vil det også være flere løsninger som må implementeres for at full reproduksjon skal være mulig.

Vi kan dele løsningene inn i tekniske løsninger og menneskelige løsninger. Den menneskelige komponenten i problemstillingen er at mange forskere ikke ønsker full åpenhet til data, koder, programvare osv i sine studier. Dette har ikke vært standard retningslinjer tidligere og det viser seg at dette kan fremdeles i dag være en utfordring, det trengs derfor klare retningslinjer om hva som bør anses som god forskning i henhold til kravene om reproduksjon og replikasjon. I den tekniske delen kan problemet ofte være at det mangler data, koder, fremgangsmåte eller at det er feil i programvare. Ved å integrere koder i selve artikkelen, selv om de ikke trenger nødvendigvis å være synlig, skal det være mulig for andre forskere å reprodusere og replikere studien.

Her er en oversikt over hva som skal sendes til tidsskriftene:

- Dokument med tekst.
- Kode til å lese inn dataen med.
- Kode til å kalkulere de ulike modellene.
- Kode for å teste modellene.
- Kode for å generere rapport av selve resultatene.

Hovedpunktet er at all data og koder vil bli sendt sammen med et fullstendig reproduserbart dokument.

2.2.1 Mulig løsning (teoretisk plan) I følge Gentleman og Lang (2007) er det viktig, kanskje også essensielt, å integrere beregninger og koder som brukes i dataanalyser, metodebeskrivelser og simuleringer. Dette kan gjøres via et kompendium i henhold til Gentleman og Lang (2007). Kompendium vil si en samling av de ulike elementene, som tekst, kode, data, metodikk og lignende, og dette skal settes sammen som en enhet for å kunne distribueres, håndteres og oppdateres. Dersom kompendiet er laget riktig skal forfattere enkelt kunne reproducere resultatene.

Dynamiske dokumenter er en ordnet sammensetning av stykker (“chunks”) med kode eller tekst. “Code chunk” kan brukes som et middel som gjengir utdata i dokumenter, eller bare for å vise kode for illustrasjon, siden “code chunks” utfører beregninger som trengs for å produsere riktig utdata i dokumentet, men også for å produsere mellomresultater brukt på tvers av forskjellige code chunks (Gentleman og Lang, 2007). “Text chunks” er ment å være formatert for lesing og beskriver som oftest problemet, koden, resultatene og tolkningen (Gentleman og Lang, 2007). Et optimalt kompendium vil derfor være et dynamisk dokument siden alle komponenter er til stedet for reproduksjon. I analysedelen av denne oppgaven vil vi vise i R Notebook i programmet Rstudio hvordan vi legger inn text- og code chunks.

2.2.2 Mulig løsning (R Notebooks): Tidligere var det R Markdown som ble brukt, men der fikk man ikke all tekst, koding og utdata (resultat) inn i samme dokument, som oftest ble dette delt opp i ulike vindu. R Notebook er den nyeste utgivelsen fra Rstudio på programutvikling. Rstudio er en IDE (Integrated Developer Enviroment) for alle R relaterte

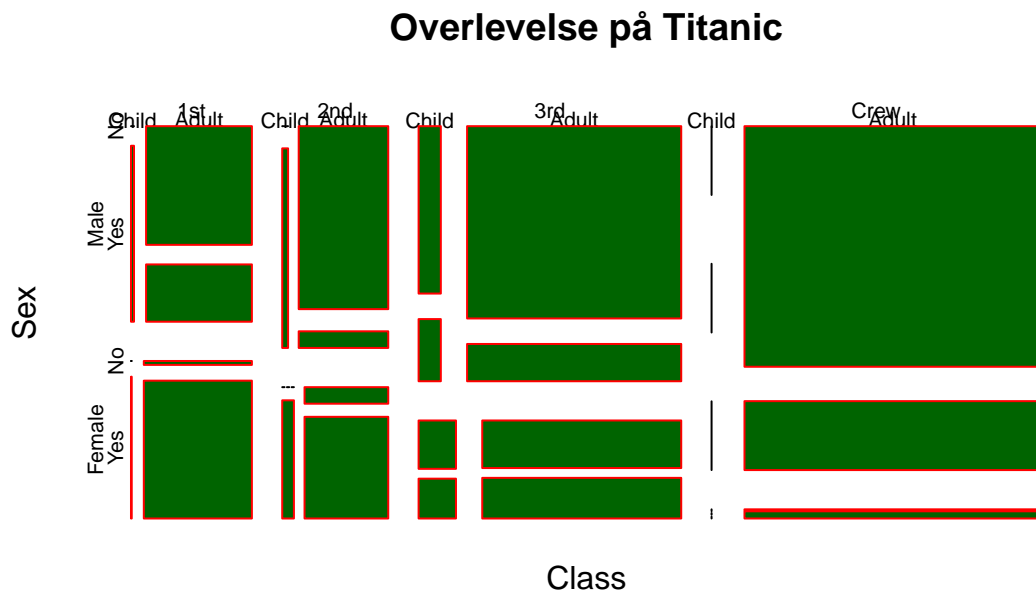
ting. Rstudio er et gratis programvare som du både kan laste ned lokalt på datamaskinen eller du kan jobbe over internett. Alle de vanlige platformene skal ha mulighet til å bruke R Studio som Mac, Window og Linux. R studio må knyttes sammen med andre programvarer og pakker for å kunne oppnå reproduserbarhet og replikerbarhet som er målet, se punkt 3.3 for mer info. RStudio kan knyttes opp mot github og det gjør at man har arbeidet sitt skylagret. En R Notebook er et R Markdown dokument med “code- og text chunks” som kan hente inn data og koder, utføre beregning ut i fra formlene som legges inn i “code chunks” og R Notebook vil vise oss resultatene direkte i samme dokument som vi arbeider i [xie2018].

3. Analyse

I dette kapitlet vil vi gå gjennom noen eksempler hvordan vi kan lage og vise frem data ved hjelp av *RStudio*. Vi vil også se på hva som kan gjøres for å vurdere viktigheten av å kunne presentere reproduserbare data, og om dette vil la seg gjøre på en enkel måte.

3.1 Titanic

Vi begynner først med å lage en oversikt over hvor mange som overlevde den skjebnesvangre jomfruturen til Titanic i 1912. Dette kan vi gjøre via datasettet “*Titanic*”, utviklet av R Core Team (2020).



Denne oversikten er gjerne litt rotete og forteller oss lite om nøyaktig hvor mange passasjerer som tilhører de forskjellige boksene. Den gir oss altså ikke stort mer enn en indikasjon på forholdet mellom død og overlevelse. Men vi kan tydelig se at det var generelt svært mange fra de høyere klassene, og da særlig blant barn og kvinner som overlevde.

Vi kan vise dette på følgende måte, som viser det totale antallet overlevende blant barn og voksne. Dette vil også fungere som et eksempel på en “chunk” som inneholder både kode og tekst.

```
## Vi summerer variablene hentet fra "Help"-funksjonen i RStudio.
apply(Titanic, c(3, 4), sum)
```

```
##      Survived
## Age      No Yes
## Child   52  57
## Adult 1438 654
```

```
apply(Titanic, c(2, 3, 4), sum)
```

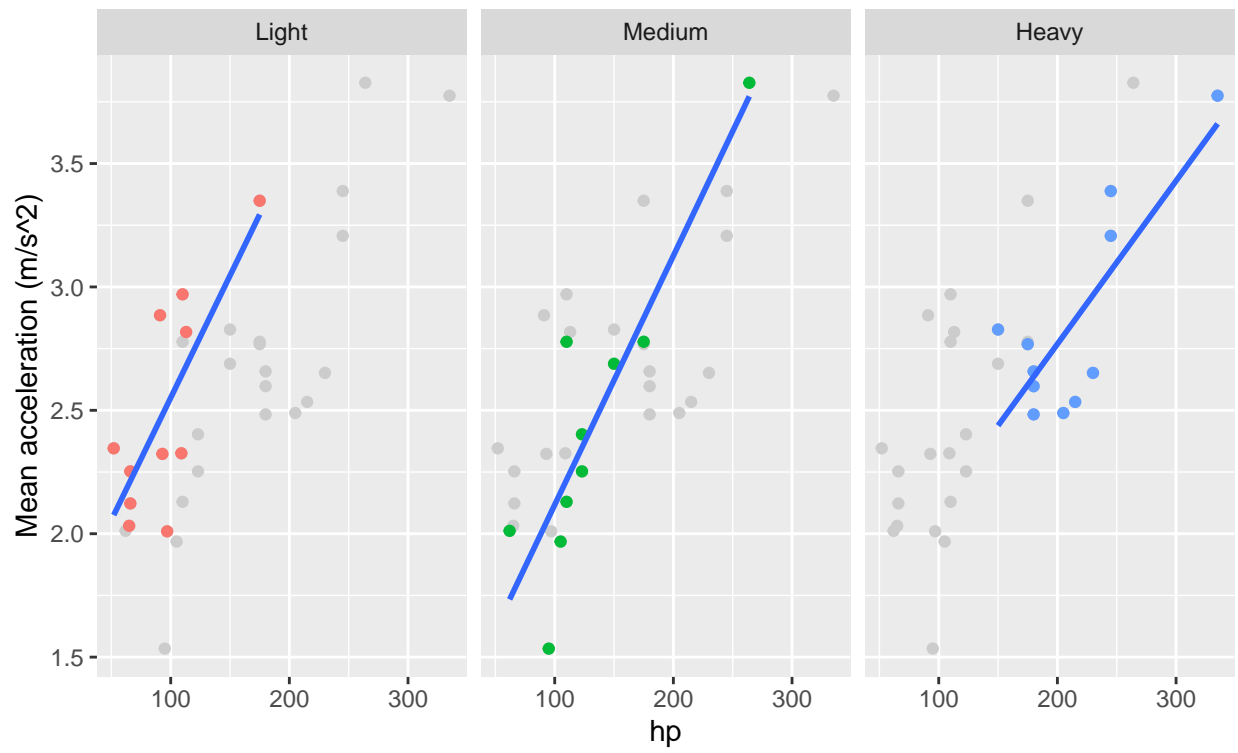
```
## , , Survived = No
##
##      Age
## Sex      Child Adult
##  Male      35  1329
##  Female    17   109
##
## , , Survived = Yes
##
##      Age
## Sex      Child Adult
##  Male      29   338
##  Female    28   316
```

3.2 Eksepel fra forelesning

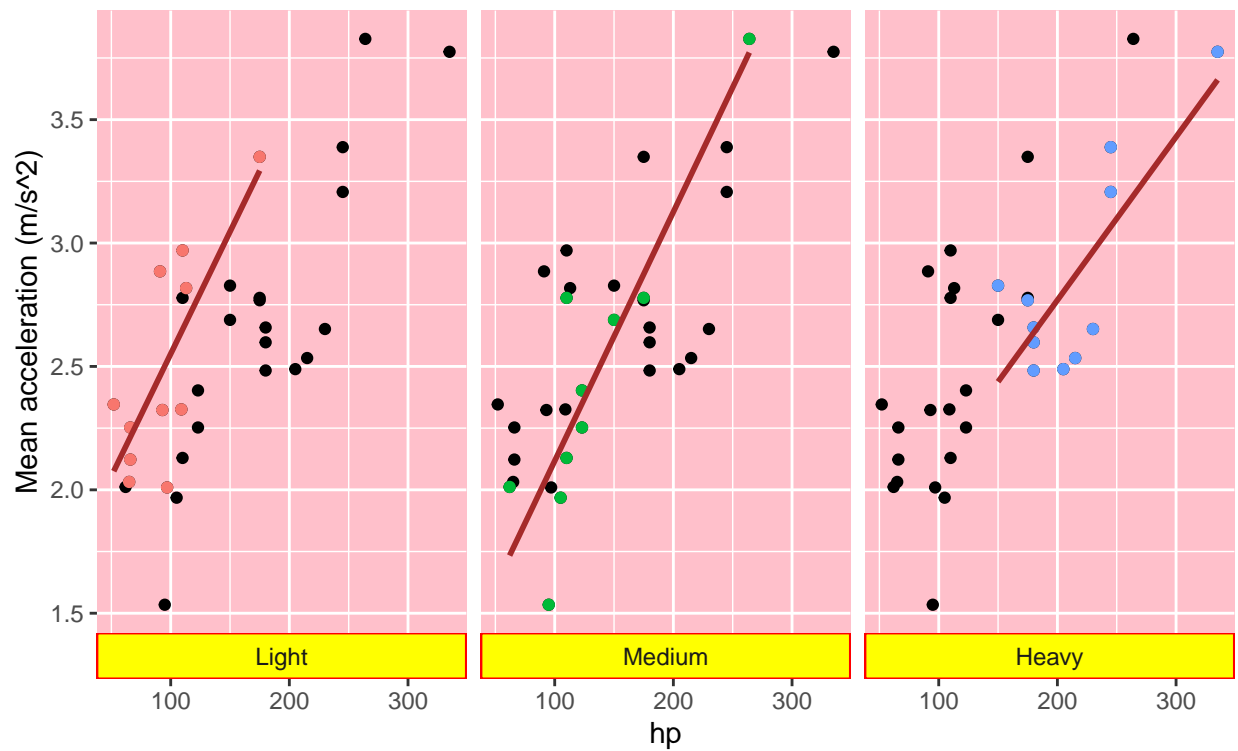
Vi kan også bruke pakken ggplot (utviklet av: Wickham, 2016) til å lage spennende og forklarende grafer dersom vi ønsker å forklare variabler i datasett.

De følgende grafene er da hentet fra forelesningsnotatene (Gjestland, 2020) og tar utgangspunkt i datasettet *mtcars*. Den første grafen er original, mens den andre lettere modifisert for å vise hvor lett det er å endre utseende og farger på grafen etter eget ønske.

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
## 'geom_smooth()' using formula 'y ~ x'
```



3.3 Sessioninfo og koder

I dette kapitlet har vi vist eksempler på “forskningsresultater”, der vi både viser og unnlater å vise hvordan vi kom frem til “sluttproduktet”. Det skal være forholdsvis enkelt å reproducere våre resultater. **Men** det forutsetter at andre brukere har installert akkurat nøyaktig de samme pakkene i *RStudio*, som vi har benyttet oss av. En løsning på denne utfordringen kan være å legge ved følgende “chunk”. Den vil produsere en oversikt over nødvendig programvare, for å kunne reproducere nøyaktig samme resultater som her.

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Norwegian Bokmål_Norway.1252
## [2] LC_CTYPE=Norwegian Bokmål_Norway.1252
## [3] LC_MONETARY=Norwegian Bokmål_Norway.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=Norwegian Bokmål_Norway.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
##  [1] ggpubr_0.4.0    forcats_0.5.0  stringr_1.4.0  dplyr_1.0.1
##  [5] purrr_0.3.4     readr_1.3.1    tidyr_1.1.1    tibble_3.0.3
##  [9] tidyverse_1.3.0 tinytex_0.25    ggplot2_3.3.2
```

```
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.5          lattice_0.20-41      lubridate_1.7.9      assertthat_0.2.1
## [5] digest_0.6.25       R6_2.4.1            cellranger_1.1.0     backports_1.1.8
## [9] reprex_0.3.0        evaluate_0.14       http_1.4.2          pillar_1.4.6
## [13] rlang_0.4.7         curl_4.3            readxl_1.3.1         rstudioapi_0.11
## [17] data.table_1.13.0   car_3.0-9           blob_1.2.1           Matrix_1.2-18
## [21] rmarkdown_2.3       labeling_0.3         splines_4.0.2        foreign_0.8-80
## [25] munsell_0.5.0       broom_0.7.0         compiler_4.0.2       modelr_0.1.8
## [29] xfun_0.16           pkgconfig_2.0.3     mgcv_1.8-31          htmltools_0.5.0
## [33] tidyselect_1.1.0    rio_0.5.16          fansi_0.4.1          crayon_1.3.4
## [37] dbplyr_1.4.4        withr_2.2.0         grid_4.0.2           nlme_3.1-148
## [41] jsonlite_1.7.0      gtable_0.3.0        lifecycle_0.2.0      DBI_1.1.0
## [45] magrittr_1.5        scales_1.1.1        zip_2.1.1            carData_3.0-4
## [49] cli_2.0.2           stringi_1.4.6       farver_2.0.3         ggsignif_0.6.0
## [53] fs_1.5.0            xml2_1.3.2          ellipsis_0.3.1       generics_0.0.2
## [57] vctrs_0.3.2         openxlsx_4.1.5      tools_4.0.2          glue_1.4.1
## [61] hms_0.5.3           abind_1.4-5         yaml_2.2.1           colorspace_1.4-1
## [65] rstatix_0.6.0       rvest_0.3.6         knitr_1.29           haven_2.3.1
```

Det passer nok best å kjøre en slik kode helt til slutt i dokumentet, da det gjerne tar en del plass og kan virke forstyrrende inne i en artikkel (som her).

Som vi var inne på i forrige kapittel, kan en god løsning være å samle alle kodene i et eget *kode-appendiks* i slutten av dokumentet. Dette er noe vi har valgt å gjøre for å vise frem muligheten. Vi benytter oss av følgende kode i en egen “chunk” i appendikset “Koder”:

```
{r ref.label=knitr::all_labels(), echo = T, eval = F}, som vi har hentet fra Xie (2018).
```

4. Konklusjon

Vi ser at ved å bruke R Notebook og lage et dynamisk dokument med både data, koder, fremgangsmåte, resultat og referanser, skal det være mulig å reprodusere, replikere og generalisere et studie. Det som kan gjøre det litt vanskelig er kompleksiteten på programmet, fordi selve kode- og data delen i RStudio er ikke det som er vanskeligst siden man kan skrive både formel og kode og få utdata i samme dokument, men det er alle programmer og pakker som skal snakke i lag som gjør arbeidet litt mer komplekst. Den viktigste komponenten vil uansett alltid være at forskeren ønsker å dele sin fulle utredelse slik at forskningen faktisk kan brukes av andre uavhengige parter, det er derfor viktig at det blir flere standard retningslinjer for hva som er minimumskrav for forskere ved utgivelse av undersøkelser.

Referanse

Bollen, K., Cacioppo, J. T., Krosnick, J. A., Olds, J. L., og Kaplan, R. M. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences). NSF.

Gentleman, R., og Lang, D. T. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23. <https://doi.org/10.1198/106186007X178663>

Gjestland, A. (2020). *R Notebook Showcasing Ggplot*. <https://elastic-turing-41462a.netlify.app/presentation>

Jasny, B. R., Chin, G., Chong, L., og Vignieri, S. (2011). Again, and Again, and Again. *Science*, 334(6060), 1225–1225. <https://doi.org/10.1126/science.334.6060.1225>

McNutt, M. (2014). Reproducibility. *Science*, 343(6168), 229–229. <https://doi.org/10.1126/science.1250475>

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* [Manual]. R Foundation for Statistical Computing.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Xie, Y. (2018). *How to Put All Your Code in the Appendix in R Markdown*. <https://yihui.org/en/2018/09/code-in-the-appendix/>.

Appendiks

Changes History Innledning

1 Your branch is ahead of 'origin/master' by 1 commit.

Staged Status Path

MSB105-Ass

MSB105-Kat

MSB105-Kat

Show Staged Unstaged

@@@ -27,12 -27,13

27 27

28 28 ## Innledning

29 29

30 30 Reproduserbarhet, R notebooks

31 31

32 32 <<<<<< HEAD

33 33 I denne oppgaven ønsker vi å se på reproduserbarhet og viktigheten av det.

34 34 Vi vil i denne sammenhengen ta oss innom temaet om bruk av "R Notebook" i Rstudio.

35 35 =====

36 36

37 37 Vi vil se på teori rundt reproduserbarhet og R Notebook, som vi også vil knytte opp til en analyse.

38 38 >>>>>> b80bc303adb015f1d8efd8bba23dccc35ecac6d76

39 39

40 40 ## Teori

41 41

42 42 ## Analyse

43 43

Git Pull

>>> /usr/bin/git pull

Auto-merging MSB105-Katrine,Karl-Gunnar.Rmd

CONFLICT (content): Merge conflict in MSB105-Katrine,Karl-Gunnar.Rmd

Automatic merge failed; fix conflicts and then commit the result.

Close

Commit message

Amend previous commit

MSB105-Katrine,Karl-Gunnar.Rmd

Documentation for package 'reproducibility'

1 title: "R Notebooks og reproduserbarhet"

2 subtitle: "Assignment 1 i MSB105 Data Science"

3 author: "Karl-Gunnar Severinsen"

4

5 "Katrine Hope"

6 "Karl-Gunnar Severinsen"

7 classoption: twoside

8 fontsize: 12pt

9 linesstretch: 1.5

10 output:

11 pdf_document:

12 df_print: paged

13 toc: true

14 tocdepth: 3

15 html_document:

16 df_print: paged

17 toc: true

18 tocdepth: 3

19 tocfloat: true

20 word_document:

21 df_print: paged

22 toc: true

23 tocdepth: 3

24 bibliography: reproducibility.bib

25 lang: nb-NO

26 csl: apa-no-ampersand.csl

27

28

29 ## (r setup, message=FALSE, quiet=F)

30 library(ggplot2)

31 library(tidyverse)

32 library(tidyverse)

33 library(ggpubr)

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

RStudio: Review Changes

Changes History master (all commits)

Subject Author Date SHA

Lagt inn i-line kode i 3.3 Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-23 b3ed2f1d

Merge branch 'master' of https://github.com/dunna99/MSB105-Assignment-1 Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-23 4a56c6fa

origin/master: merge Lagt inn i-line kode i 3.3 Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-23 f71d362a

Sleiv mer på punkt 2.1 og 2.2.2. Katrine Hope <katrine_hope@hotmail.com> 2020-09-22 8aef0d6f

Regn ut på analyse. Katrine Hope <katrine_hope@hotmail.com> 2020-09-22 46fa3ed9

Sleiv litt på 2.1 og 2.2. Katrine Hope <katrine_hope@hotmail.com> 2020-09-21 8a625890

Oppdatert VAML og lagt til sidefelt + plassering referanseliste. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-20 a995a5a3

Sleiv mer på punkt 2.1. Reproduserbarhet, replikerbarhet og generaliserbarhet. Katrine Hope <katrine_hope@hotmail.com> 2020-09-20 e807ae03

Oppdatert overskrifter. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 5427a003

La inn litt teori om reproduksjon og replikasjon. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 98cc01c4

Merge branch 'Innledning' into master. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 4986a693

bilde. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 a098a5dd

origin/master: merge Slettet merge conflict bildet under innledningsbrant. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 61463f63

Merge branch 'master' of https://github.com/dunna99/MSB105-Assignment-1. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 c9a04c4a

Lagt inn appendiks + bilde av merge conflict. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 a3f144a8

La ikke inn noe tekst. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 f6cd5336

La inn stjernebilde på merge conflict. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 de883e1f

Fikset merge conflict. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 34d2e573

La inn min tekst på innledning. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 26660b67

forlag til tekst i innledning. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 b80bc303

Fjernet tekst som er lagt inn på teori-branchen. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 30d8a25d

Lagt til disposisjonsforslag teori. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 ea19a3e9

La inn disposisjonsforslag på innledning. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 069317c5

La til overskrifter. Katrine Hope <katrine_hope@hotmail.com> 2020-09-16 0ca57df7

Commit bil+csl. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 21d1af66

Create VAML. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 98b4a554

Project creation. Karl-Gunnar Severinsen <dunna99@gmail.com> 2020-09-16 b55961dc

SHA: ea8d07477761a0d791a6b6ddce615a09b0d0

Author: Karl-Gunnar Severinsen <dunna99@gmail.com>

Date: 2020-09-23 07:02

Subject: Lagt inn i-line kode i 3.3

Parent: b3ed2f1d6e5898d37497acdd1c4ee97e813181

MSB105-Katrine,Karl-Gunnar.Rmd

Environment History Connections Git Tutorial

Files Plots Packages Help Viewer

1. Innledning

2. Teori

3. Analyse

4. Konklusjon

Referanse

Appendiks

Koder

R Notebooks og reproduserbarhet

Assignment 1 i MSB105 Data Science - innleveringsfrist 24.09.20

Katrine Hope

Karl-Gunnar Severinsen

1. Innledning

I denne oppgaven ønsker vi å se på reproduserbarhet og viktigheten av det. Vi vil i denne sammenhengen ta oss innom temaet om bruk av "R Notebook" i Rstudio.

Vi vil se på teori rundt reproduserbarhet og R Notebook, som vi også vil knytte opp til en analyse.

2. Teori

2.1 Replikerbarhet/reproduserbarhet

Det er ønskelig at vitenskapelige oppdagelser og fremskritt skal være robuste og pålitelige fordi vi ønsker å ha tillit til at resultatene er riktige og at undersøkelsene er gjort på en tilfredsstillende måte (Bollen et al., 2015). I følge McNutt (2014) er det viktig for at vitenskapen skal utvikle seg at funnene baserer seg på troverdige funn. Mange forskere mener at reproduserbarhet er en viktig fremgangsmåte for å kunne

Koder

```
library(ggplot2)
library(tinytex)
library(tidyverse)
library(ggpubr)

mosaicplot(Titanic, main = "Overlevelse på Titanic", color = "darkgreen", border = "red")
## Vi summerer variablene hentet fra "Help"-funksjonen i RStudio.
apply(Titanic, c(3, 4), sum)
apply(Titanic, c(2, 3, 4), sum)

hp_acc <- data.frame(hp = mtcars$hp, acc = 1609.347/(2*mtcars$qsec^2))

ggplot(data = mtcars, mapping = aes(x = hp, y = 1609.347/(2*qsec^2))) +
  geom_point(data=hp_acc, mapping = aes(x = hp, y = acc), colour= "grey80") +
  facet_wrap(~cut_number(wt, n = 3 , labels = c("Light", "Medium", "Heavy"))) +
  geom_point(aes(colour = cut_number(wt, n = 3, labels = c("Light", "Medium", "Heavy"))),
    show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(y="Mean acceleration (m/s^2)", colour = "Weight")

hp_acc <- data.frame(hp = mtcars$hp, acc = 1609.347/(2*mtcars$qsec^2))

ggplot(data = mtcars, mapping = aes(x = hp, y = 1609.347/(2*qsec^2))) +
  geom_point(data=hp_acc, mapping = aes(x = hp, y = acc), colour= "black") +
  facet_wrap(~cut_number(wt, n = 3 , labels = c("Light", "Medium", "Heavy")), strip.position = "top") +
  theme(panel.background = element_rect(fill = "pink"), strip.background = element_rect(fill = "pink")) +
  geom_point(aes(colour = cut_number(wt, n = 3, labels = c("Light", "Medium", "Heavy"))),
    show.legend = F) +
  geom_smooth(method = "lm", color = "brown", se = FALSE) +
  labs(y="Mean acceleration (m/s^2)", colour = "Weight")

sessionInfo()
```