

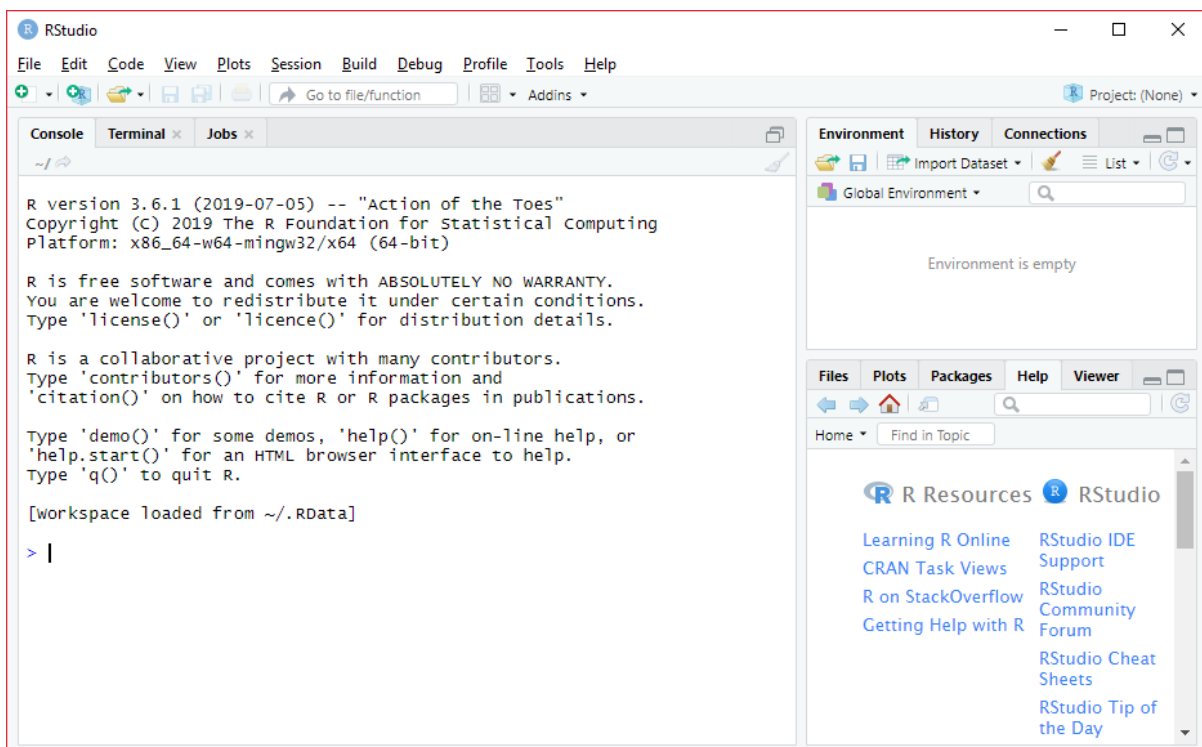
Johdatus yhteiskuntatilastotieteeseen, 2021 (#tilastoMOOC 2021-2022)

R, RStudio ja R Markdown : ensiaskeleita yhteiskunta *datatieteen* polulla

Kurssin MOOC-alustalla¹ on linkit ja ohjeet tehtävissä tarvittavien R- ja RStudio-ohjelmien lataamiseen ja asentamiseen. Molemmat ohjelmat ovat vapaasti saatavilla verkosta:

1. R: <https://www.r-project.org/>
2. RStudio: <https://www.rstudio.com/>

Asennettuasi ohjelmat koneellesi käynnistä RStudio. Sen aloitusnäky on alla olevan kaltainen:



Pienenhköjä eroja esiintyy riippuen tietokoneestasi, sen käyttöjärjestelmästä yms. Käsillä oleva ohje on tehty Windows 10 –ympäristössä 2019 ja päivitetty olennaisilta osin 2021. Olennaisia eroja ei pitäisi ilmetä, sillä ohjelmat ovat käytettävissä varsin samalla tavoin kaikissa tyypillisissä ympäristöissä (Mac/Windows/Linux).

R ja RStudio ovat huippuluokan työkaluja, joiden kehittämiseen osallistuu aktiivisesti laaja kansainvälinen käyttäjäyhteisö. Kannattaa ladata ohjelmien uusimmat versiot ja päivittää niitä joko automaattisesti tai tarpeen mukaan. Esimerkiksi R:stä ilmestyy yleensä uusi versio pari kertaa vuodessa². Pitkältä voi ajatella, että ”*versiosta viis*”, mutta oudot virhetilanteet yms. korjaantuvat usein kaikkein helpoimmin päivittämällä ohjelmat uusimpiin versioihin. Mitä suurempi ero versionumeroissa on, sitä merkittävämpiä muutoksia siinä on tehty (erittäin yksityiskohtaiset tiedot muutoksista eri versiossa ovat myös avoimesti saatavilla).

¹ <https://mooc.helsinki.fi> -> Johdatus yhteiskuntatilastotieteeseen 2021-2022

² Aloitettuani tämän ohjeen kirjoittamisen (syyskuussa 2019) huomasin koneeni R:n olevan maaliskuulta, versio 3.5.3 (”Great Truth”), ja päivitin sen nyt uusimpaan, heinäkuussa julkaistuun versioon 3.6.1 (”Action of the Toes”). Jos nuo versioiden nimet alkavat askarruttaa ☺☺... ks. <https://livefreeordichotomize.com/2018/04/23/r-release-names/> tai <https://stackoverflow.com/questions/13478375/is-there-any-authoritative-documentation-on-r-release-nicknames>.

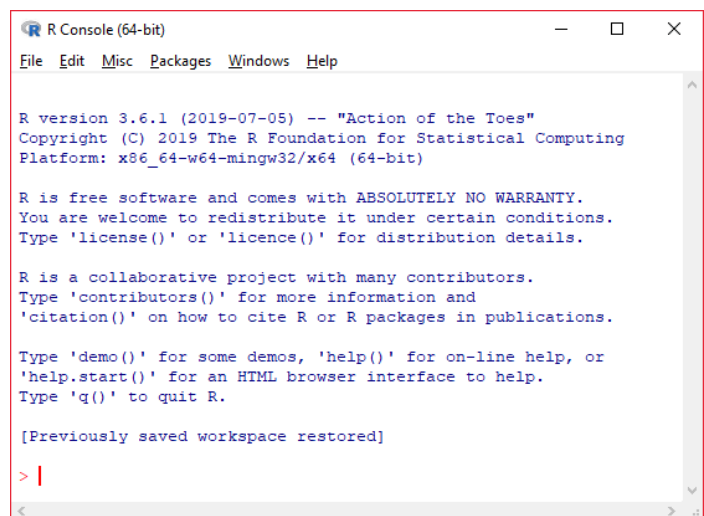
R on arvokas aikaansaannos, mutta raakana se kypsyyttää kokeneenkin

R sai alkunsa Aucklandin yliopiston tilastotieteen laitoksella 1990-luvulla, kun kaksi siellä työskennellyttä kaveria, **Robert Gentleman** ja **Ross Ihaka**, toteuttivat unelmansa vapaasti jaettavasta, avoimen lähdekoodin tilastollisesta ohjelmistosta (vrt. **Linux**, jonka **Linus Torvalds** päätti laittaa vapaasti jakeluun samoihin aikoihin HY:n tietojenkäsittelytieteen laitoksella). **R** perustuu 1970-luvulla Bellin Laboratoriossa kehitettyyn tilastolliseen **S**-ohjelmointikielen. **S**:n varaan 1990-luvulla rakennettu kaupallinen ohjelmisto **S-PLUS** oli käytössä 2000-luvun alkupuolelle, kunnes ilmainen ja vapaa **R** yleistyi, jolloin **S-PLUS** kävi käytännössä tarpeettomaksi³. Myös moni muu tilastollinen ohjelmisto on **R**:n myötä käynyt, jos ei tarpeettomaksi, niin ainakin aiempaa vähemmän tarpeelliseksi.

R:ää voi käyttää sellaisenaankin, ilman **RStudio**a, mutta pelkän **R**:n käyttöliittymä on aika (k)ankea, etenkin aloittelijalle. **R** on ennen kaikkea **tilastollinen ohjelmointikieli**. Se sisältää tietorakenteita ja valmiita funktioita, joilla dataa analysoidaan **ohjelmoimalla** (siis ei vain valitsemalla valmiita toimintoja valikoista, kuten on tapana eräissä muissa tilastollisissa ym. ohjelmistoissa).

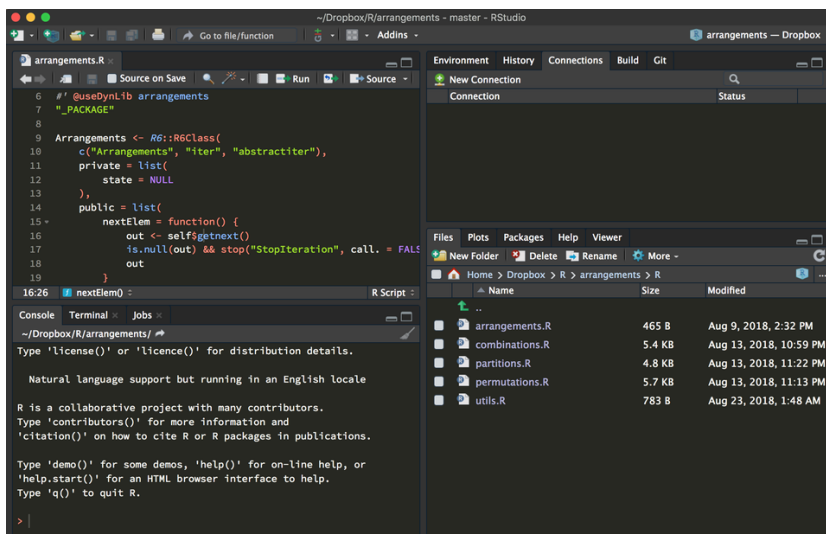
Erikseen ladattavat **R-paketit**, joita **R**-kielen edistyneet osaajat voivat laatia ja jakaa muiden käyttöön, laajentavat **R**:n sovellusalueita yhä useammille aloille ja uudennaisiin tarkoituksiin.

Jos miettii, voiko jonkin jännän jutun tehdä **R**:llä, kannattaa ensin googlata, sillä todennäköisesti siihen(kin) tehtävään on jo olemassa **R**-paketti, jolla tuo jännä juttu saattaa hoitua, jopa ihan helposti.



RStudio on monipuolinen käyttöliittymä **R**:lle (ja moniin muihin kieliopintoihin)

Onneksi **R**-pakettien lisäksi **R**:ään on myös saatavilla parempia käyttöliittymiä kuin yllä näkyvä "raaka **R**". **RStudio**, jota tällä kurssilla käytämme, on ylivoimaisesti suosituin. Jätetään raaka **R** ja palataan **RStudio**lle:

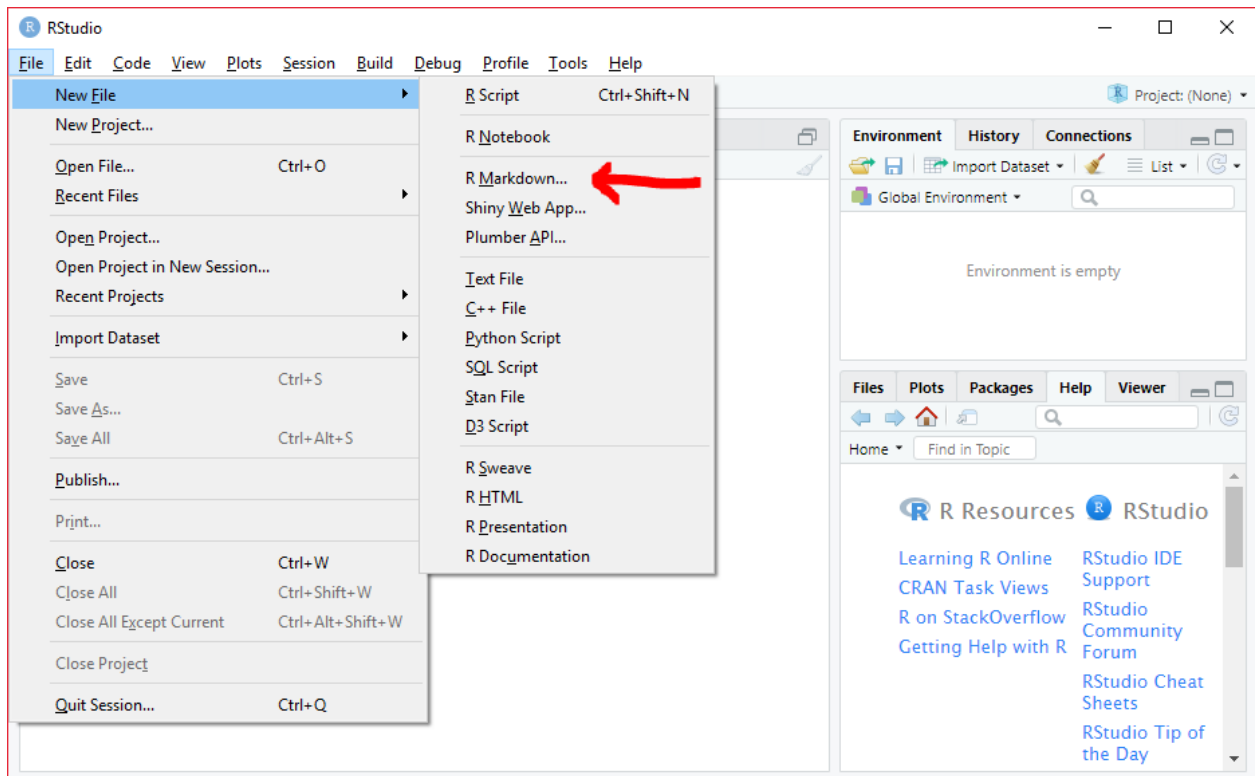


Kuvan lähde: https://www.reddit.com/r/rstats/comments/99v8qp/rstudio_with_a_dark_theme/

³ Jonkin aikaa oppikirjojen nimissä esiintyi sekä **S-PLUS** että **R**, mutta kun **R** vei voiton, **S-PLUS** katosi kirjoistakin.

Dynaamista dokumentointia R Markdown –merkkaukielellä

RStudio on varsin monipuolinen ohjelmisto, jolla voi työskennellä vaikka mitä: eri ohjelmointikielisiä skriptejä, dynaamisia ja staattisia raportteja ja dokumentteja kuten kirjoja, datapohjaisia esitysmateriaaleja jne. Leegio vaihtoehtoja ilmestyy näkyviin, jos avaat File-valikon kohdan New File:



Vaihtoehtojen runsaudesta ja teknisistä termeistä (C++, SQL, D3, Python jne.) ei ole mitään syytä ahdistua.

Kurssilla perehdytään yllä **punaisella nuolella** osoitettuun vaihtoehtoon nimeltään **R Markdown**, joka on 2000-luvulla kehitetty, yksinkertainen ja helppo **merkkaukielellä** rakenteisen dokumentin (eli sellaisen, joka pelkän tekstin ohella sisältää hierarkkisia rakenteita kuten eri tasoisia otsikoita, listoja, kaavioita, taulukoita, kuvia, lähde- ja muita luetteloita jne.) kirjoittamiseen ja hallintaan. Sen tällä kurssilla opeteltava versio **R Markdown** (kuten nimestä voi arvata) sujuvoittaa **R-ohjelmoinnin** ja rakenteisen dokumentin yhteiselo.

Opit kurssin aikana työskentelemään **RStudio**lla rakenteisia, **dynaamisia dokumentteja**. Dynaamisuus tarkoittaa **interaktiivista data-analyysia**, jossa koodataan **R**-kielellä ja yhdistetään sillä aikaan saadut tulokset (kuvat, taulukot ja analyysien tulokset ym.) omiin tulkintoihin ja muuhun tekstiin – siis kaikki samaan raporttiin. Opit itse tekemällä soveltamaan tieteellisen työskentelyn kannalta todella keskeisiä ja tärkeitä periaatteita: **dokumentointia** ja **toistettavuutta** (*reproducibility*) – samalla, kun opit data-analyysia ja **R**-ohjelmointia.

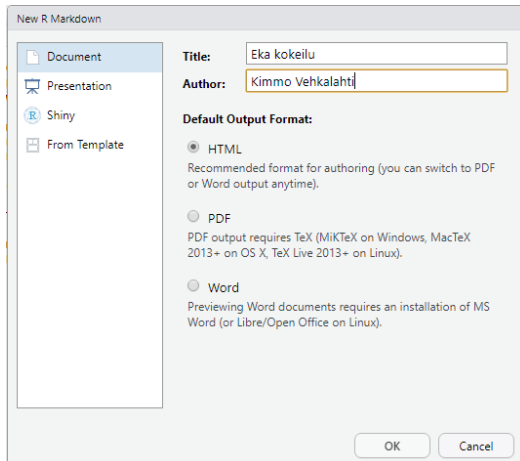
Mahtava paketti, eikö totta? 😊

”Dynaaminen dokumentti” kuulostaa hienolta, mutta on oikeastaan vain tavallinen tekstitiedosto, jonka lataat MOOC-alustalta ja muokattuasi **RStudio**lla talletat sen (File – Save) tiedostopäätteellä `.Rmd`. Siitä (*automaagisesti, Knit-toiminnolla*) syntyvän raportin tallettaa **RStudio** puolestasi HTML-tiedostona samalla nimellä mutta päätteellä `.html`, jolloin lopputulosta voi ihaila nettiselaimella. Oletuksena **RStudio** näyttää sitä omalla selailuohjelmallaan (viewer), kun dokumentti on ”kudottu kasaan” (knit). **Jännää termejä...!** 😊

Ryhdytään nyt katsomaan, miltä tämä kaikki näyttää käytännössä! ☺ Kokeillaan tässä vain **RStudio** omaa R Markdown –mallidokumenttia, joka tarjoaa yksinkertaistetun esimerkin työskentelytavasta. Sen jälkeen oletkin valmis lataamaan MOOCilta **Koodaustehtävän 1** ja aloittamaan varsinaiset kurssitehtävät! **YES!! ☺**

R Markdown –mallidokumentin avaaminen RStudio editoriin

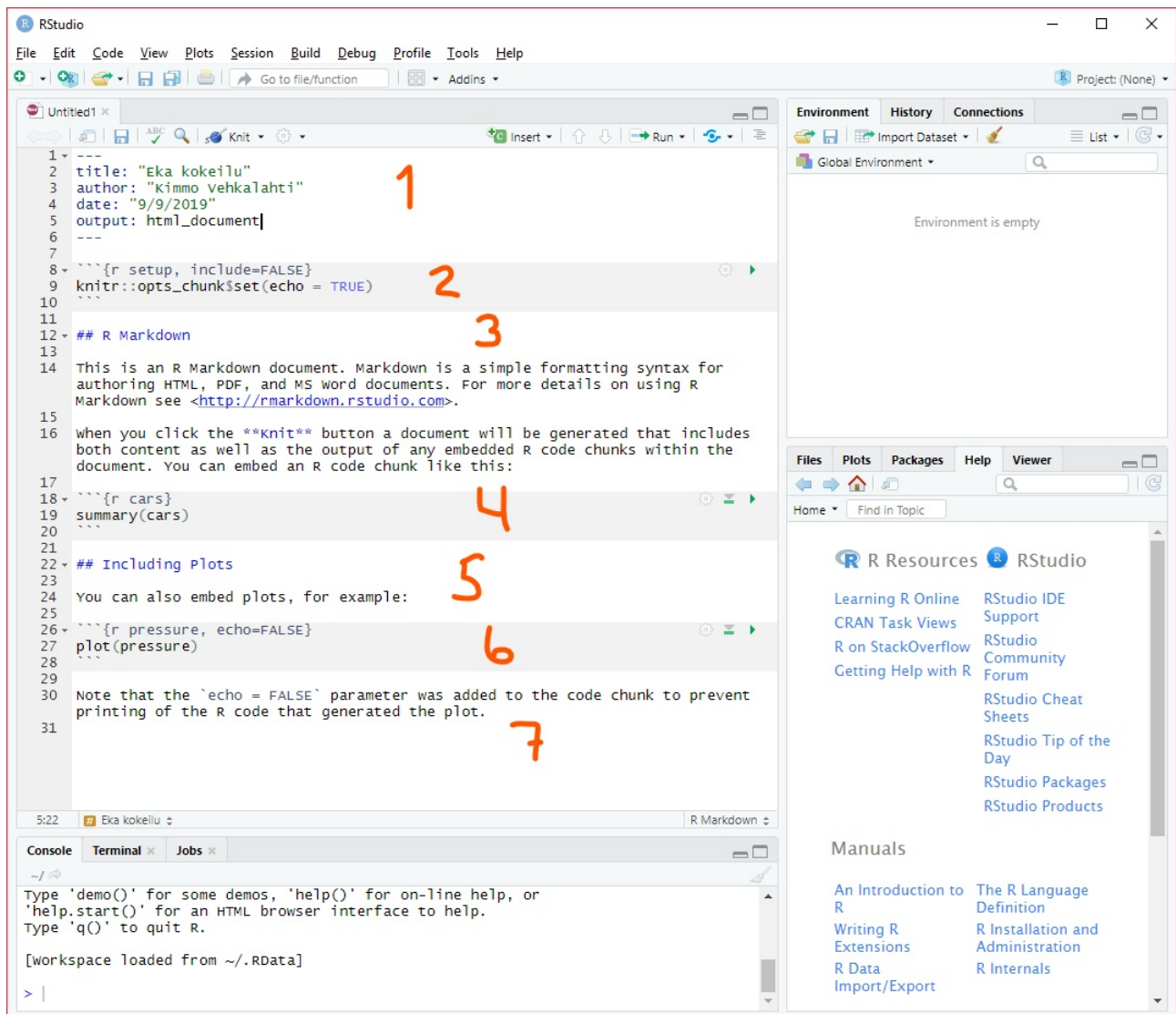
Alkuun pääset valitsemalla File – New File – R Markdown ja antamalla nimen (dokumentin ja omasi):



Samalla voit panna merkille, että raportteja voisi (pienin lisäasennuksin) tuottaa myös suoraan PDF:ksi tai Wordiksi.

Hyvien nimien keksiminen on usein yllättävän haastavaa – esimerkiksi *”Eka kokeilu”* ei nyt ehkä ole kovin kuvaava nimi, mutta kelvatkoon nyt tässä. Oman nimen kirjoittaminen on onneksi helpompaa ☺ Jatkossa näitä vaiheita ei tarvita ollenkaan, mikäli onnistut lataamaan koodaustehtävät suoraan MOOCilta koneellesi ja **RStudioon**. Siinä voi ilmetä pieniä hankaluuksia riippuen koneesta, selaimesta yms., mutta niistäkin selvittää heittämällä, katso tarvittaessa neuvoja MOOCin keskustelualueelta (Q&A).

OK-napin painallus synnyttää R Markdown –mallidokumentin ja avaa sen **RStudio** editoriin:



Katsotaan seuraavaksi, mitä nuo **punaisella** numeroidut kohdat **1–7** tarkoittavat.

Dynaaminen R Markdown -dokumentti ja siitä kutomalla (knit) muodostettava HTML-raportti
Dynaaminen dokumentti R Markdown –muodossa koostuu kolmen tyyppisestä osista:

1. **Metadata** otsikoi dokumentin ja ohjaa **RStudio**ta raportin tuottamisessa ns. YAML-muodossa, josta sinun ei tarvitse tietää muuta kuin että tämä osa on aina dokumentin alussa (edellä kohta **1**).
2. **Koodilohkot** ovat **R**-koodia varten. Lohkon aloittaa ````{r}` ja lopettaa ````` (edellä **2, 4** ja **6**).
3. **Tekstilohkot** sisältävät rakenteista tekstiä merkattuna Markdown-syntaksilla (edellä **3, 5** ja **7**).

Kun dokumentti talletetaan (File – Save) ja painetaan nappia **Knit**, **RStudio** ”kutoo” kaikki osat kasaan⁴ ja avaa näin muodostuvan HTML-raportin selainäkymässä (vertaa alla olevia kohtia **1–7** edellä oleviin!):

⁴ Kun **RStudio** on juuri asennettu (tai päivitetty), tässä kohtaa tapahtuu tarvittavien **R**-pakettien asennuksia. **Odota kärsivällisesti**, ja jos näet kysymyksiä lähdekoodista kääntämisestä (source code, compiling), vastaa kieltävästi (No), *paitsi jos toimit Linuxissa* (jossa asennukset tehdään yleensä lähdekoodista kääntämällä eikä valmiina paketteina).

D:/O/JYT19/RStudio-ohjeet/Eka.html

Eka.htmlOpen in BrowserFindPublish

Eka kokeilu

Kimmo Vehkalahti
9/9/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

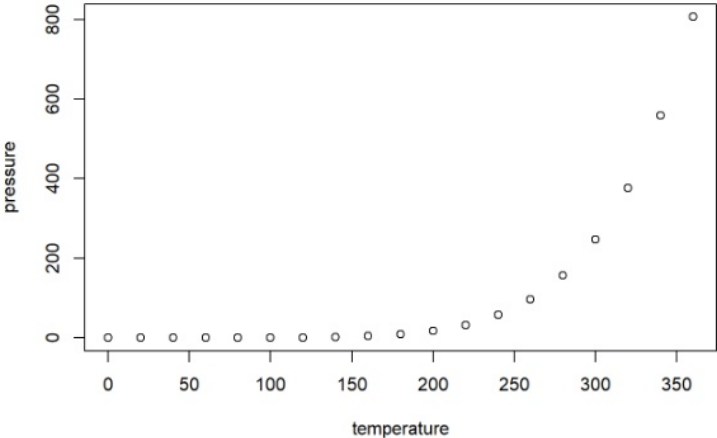
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

##	speed	dist
## Min.	: 4.0	Min. : 2.00
## 1st Qu.	:12.0	1st Qu.: 26.00
## Median	:15.0	Median : 36.00
## Mean	:15.4	Mean : 42.98
## 3rd Qu.	:19.0	3rd Qu.: 56.00
## Max.	:25.0	Max. :120.00

Including Plots

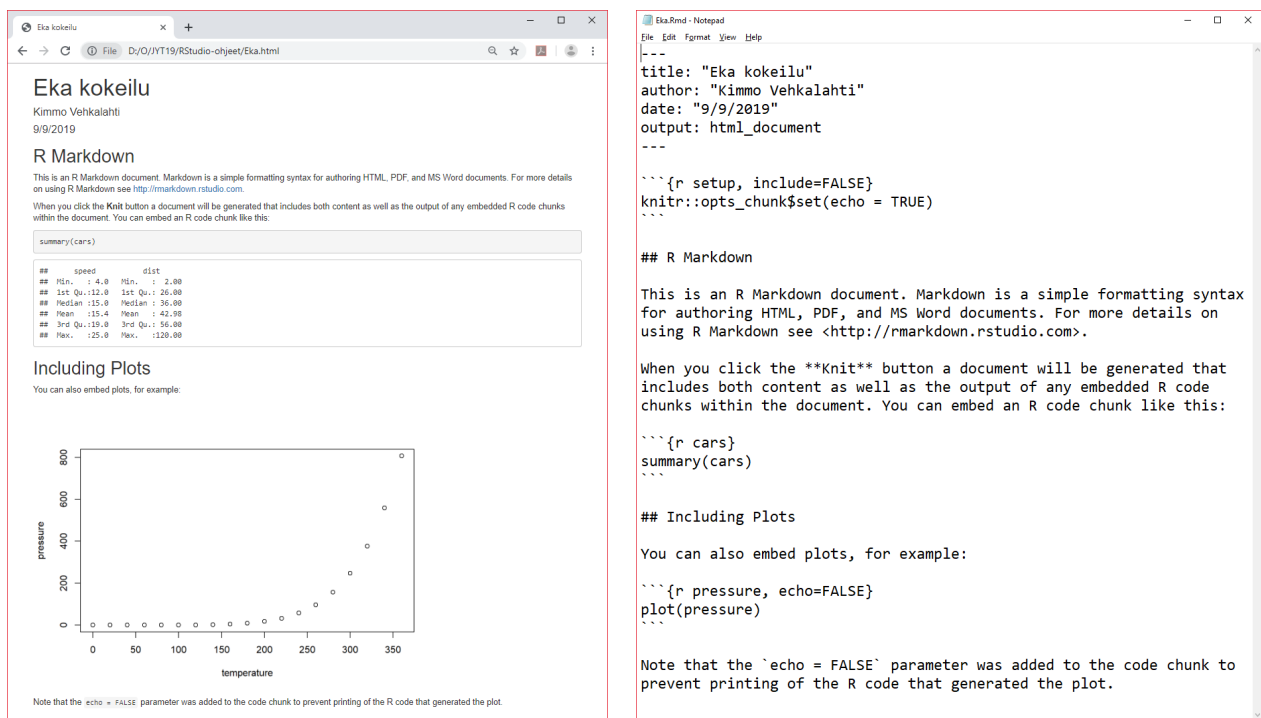
You can also embed plots, for example:



A scatter plot showing the relationship between temperature (x-axis) and pressure (y-axis). The x-axis ranges from 0 to 350 with major ticks every 50 units. The y-axis ranges from 0 to 800 with major ticks every 200 units. The data points show a clear upward trend, starting near (0, 0) and reaching approximately (350, 800).

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Open in Browser –napilla (ed. sivulla) raportti avautuu nettiselaimessa (tässä Google Chrome, vasemmalla):



The left screenshot shows a web browser displaying an R Markdown document titled "Eka kokeilu" by Kimmo Vehkalahti, dated 9/9/2019. The document includes a summary of the 'cars' dataset and a scatter plot of pressure vs. temperature. The right screenshot shows the source R Markdown file 'Eka.Rmd' in a Notepad window, showing the YAML header, R code chunks for summary(cars) and plot(pressure), and the corresponding text content.

Oikealla näkyy raportin perustana oleva dynaaminen dokumentti (tiedosto `Eka.Rmd`) pelkkänä tavallisena tekstitiedostona Windowsin Notepadilla. *Mikään ei estä sinua kirjoittamasta R Markdown –dokumenttia R-koodeineen suoraan Notepadilla tms.*, mutta siinä ei ole paljoa pointtia. **RStudio** idea on nimenomaan **interaktiivinen työskentely** (jota on aika vaikea kuvata paperille litistettynä!). Lisäksi tulevat monet kätevät toiminnot kuten koodien automaattiset syntaksivärjäykset, lukuisat työskentelyä nopeuttavat valikot ja pikanäppäimet, työvaiheiden välituloksia havainnollistavat ikkunat, suorat linkit netissä oleviin ohjeisiin, R-pakettien hallinta ja päivitys tarpeen mukaan jne. (tai vaikka jakaminen ym. [GitHubissa](#), jota käytämme [Introduction to Open Data Science](#) –kurssillani ja jonne olet tervetullut JYT:n jälkeen – ehkä jo 2021?).

Hands on data (science)!

Nyt sinun on aika ryhtyä omin käsin kokeilemaan **RStudio** käyttöä. 😊 Jos et heti keksi, mistä aloittaisit, tässä pari vinkkiä. Lisää edellä **punaisella 4:lla** merkittyyn osioon seuraavia rivejä ja **aktivoi niitä yksitellen** napilla **Ctrl + Enter** (Mac: **Cmd + Enter**). Kokeile kutoa kasaan (Knit)! *Tarkkaile, mitä RStudio:ssa tapahtuu!*

```
cars
str(cars)
head(cars)
mean(cars$speed)

boxplot(cars$dist, ylab = "Stop distance")

plot(cars)

library(stats)
plot(cars); lines(lowess(cars))
```

