**AI/ML Business and Technical Introduction curriculum**

**Do-It-Yourself Demo of RStudio**

**Description:**

The module shows how to get started with R for machine learning related development work. It shows an example case where one loads a geographical dataset[[1]](#footnote-1), visualizes it, and applies unsupervised learning (hierarchical clustering) to it.

**Learning Objectives:**

* Install and run RStudio Desktop.
* Become familiar with the main components of RStudio Integrated Development Environment (IDE).
* Install new libraries in R.
* Load a dataset into R Studio, investigate it, and apply unsupervised learning to it.

**Prerequisites:**

* 64-bit Windows 10 or Windows 7, with the permissions to install new software. *The Anaconda distribution package, which contains R Studio, can also be installed on Linux and Mac OS.*
* Access to the Internet, i.e., without going through the Nokia Virtual Private Network (VPN) or via proxy. *As noted in the demo instructions (below), the Nokia Firewall can cause problems with installing the components (Anaconda, RStudio) or using them (leaflet). These problems can be solved by changing your Internet connection, e.g. to a mobile hotspot or other trusted WiFi access point, or modifying the proxy settings of the application in question.*

**Duration:**

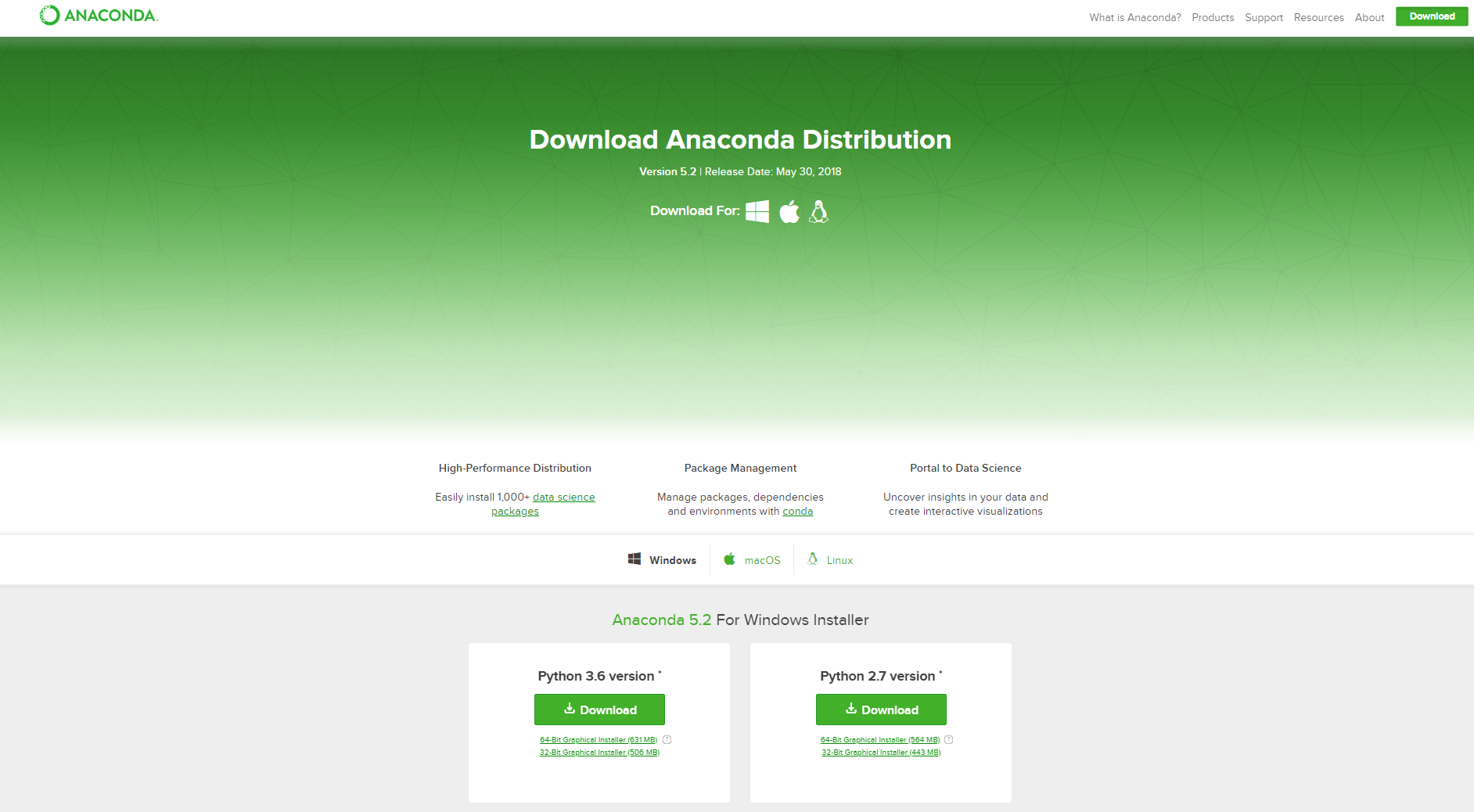
* Video: 15 min
* Installing Anaconda + RStudio + Packages: 30-45min (depends on the internet connection)
* Running the demo source code: 15-30 min

**High-level instructions:**

1. Install RStudio – Anaconda Distribution or standalone version
2. Follow the instructions in the Nokia EDU course description to download the files used in this module:
   1. source code: ml-module.R
   2. dataset: api\_tweet.json
3. Launch RStudio, change default text encoding to UTF-8, and open the downloaded source code file.
4. Run the source code file in parts:
   1. Load libraries
   2. Load downloaded dataset into data table. *Remember to change the path in the R code to match the location of your dataset.*
   3. Modify the data table.
   4. Visualize the data on the map.
   5. Calculate and visualize the distance matrix.
   6. Apply hierarchical clustering.
   7. Visualize calculated clusters on the map.

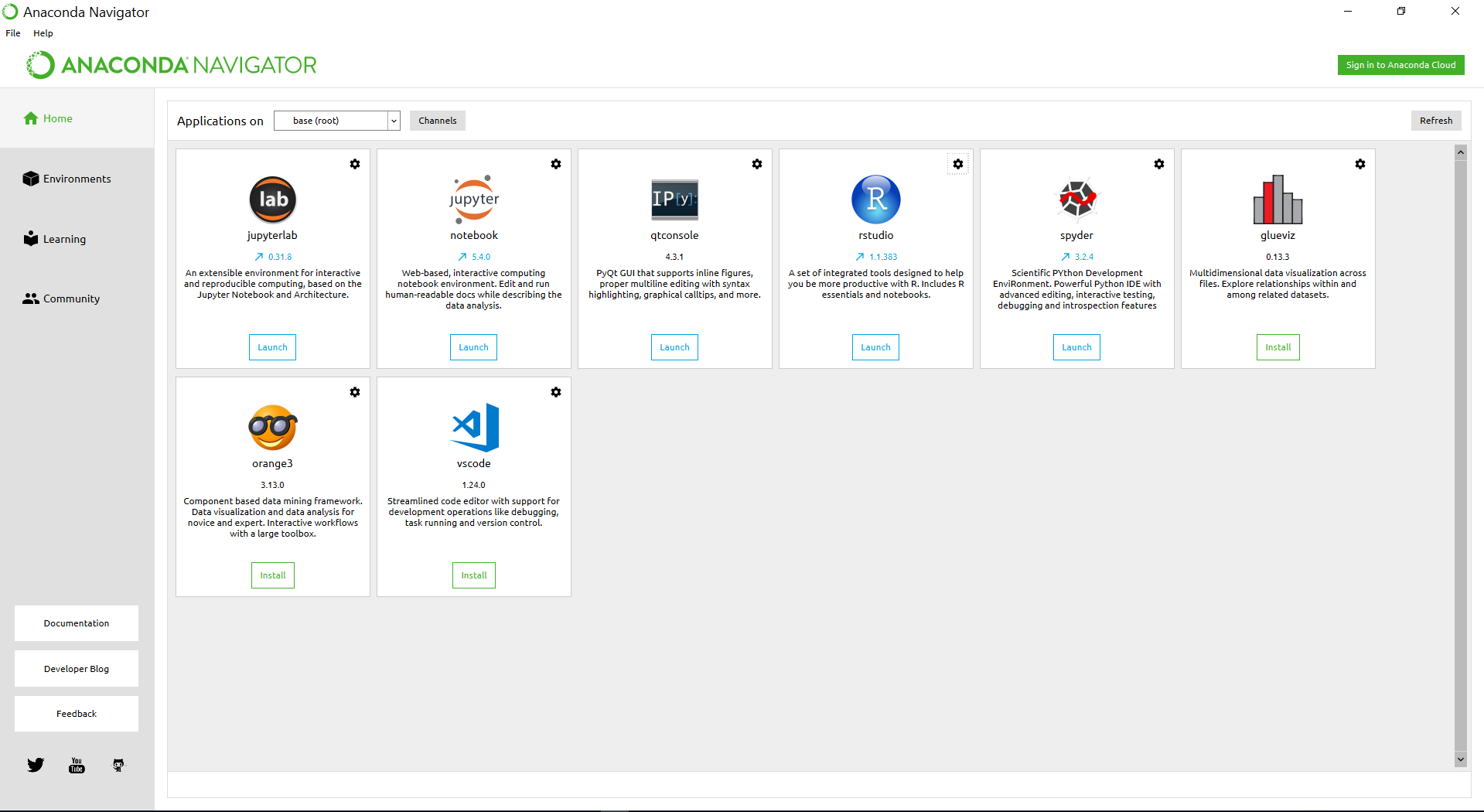
**Step-by-step instructions:**

1. **View introduction [Video #1]**
2. **Install Anaconda if you’ve not already installed it [Video #2]**
   * Go to <https://www.anaconda.com/download/> and follow the instructions.
   * Be sure to select the download package for your Operating System; usually it is the 64-bit version.
   * The installation is straightforward, but it takes about 20 minutes.

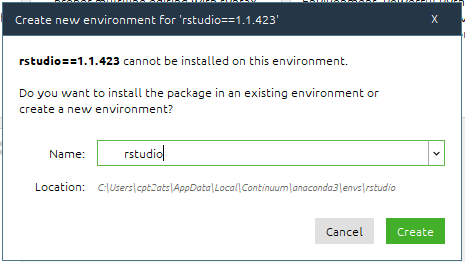


Download the newest Python version (3.x), unless there is special reason to use the older one.

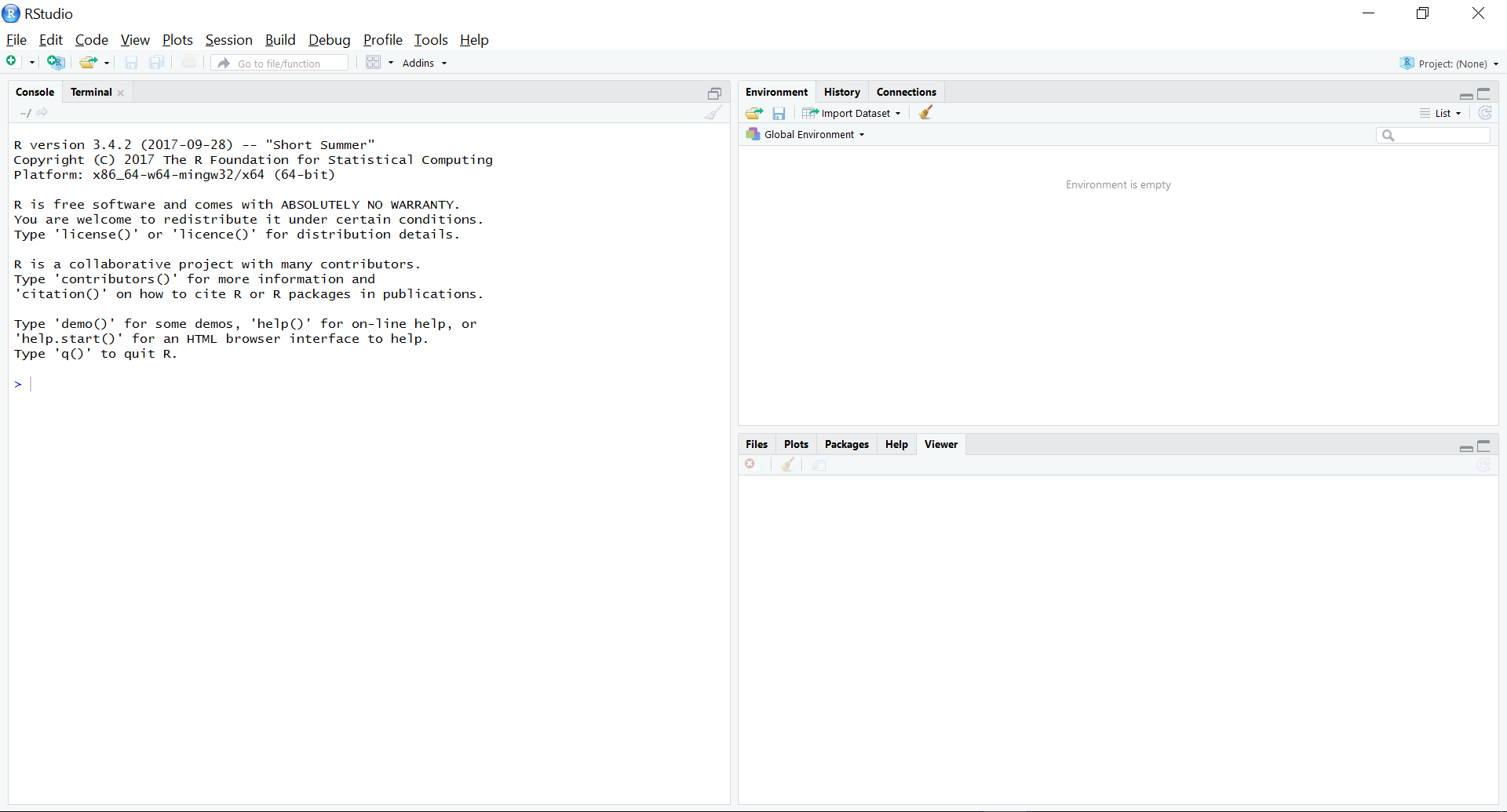
1. **Follow the instructions in the Nokia EDU course description to download the files used in this module.** They are also available from the AI/ML Competence Development Team Site:
   * Download the sample source code file: https://nokia.sharepoint.com/sites/CoDe-ML-DataScience/Shared%20Documents/General/Business%20and%20Technical%20Introduction%20Curriculum/Content%20for%20EDU/DIY%20Demo%20with%20R%20Studio/ml-module.R
   * Download the sample data file: <https://nokia.sharepoint.com/sites/CoDe-ML-DataScience/Shared%20Documents/General/Business%20and%20Technical%20Introduction%20Curriculum/Content%20for%20EDU/DIY%20Demo%20with%20R%20Studio/api_tweet.json>
2. **RStudio basics**
   * Launch RStudio via Anaconda Navigator. If RStudio is installed, you will see “Launch” button attached to it. **[Video #2]**



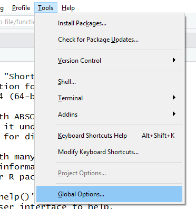
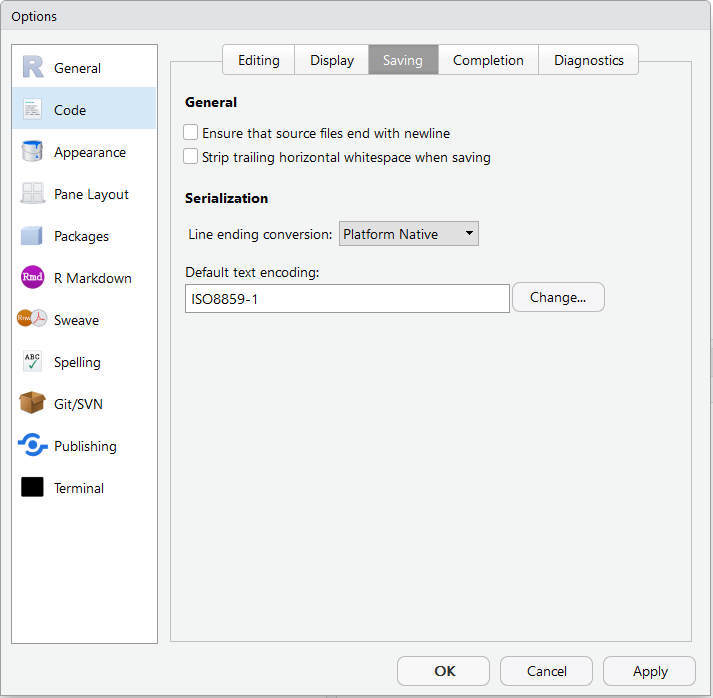
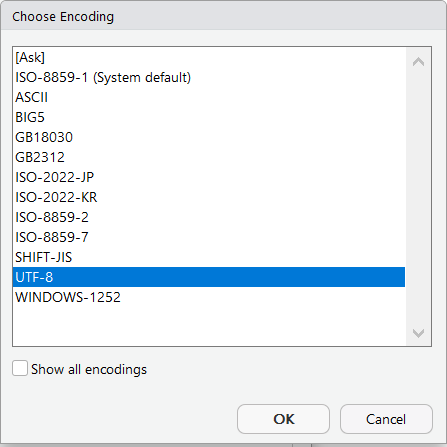
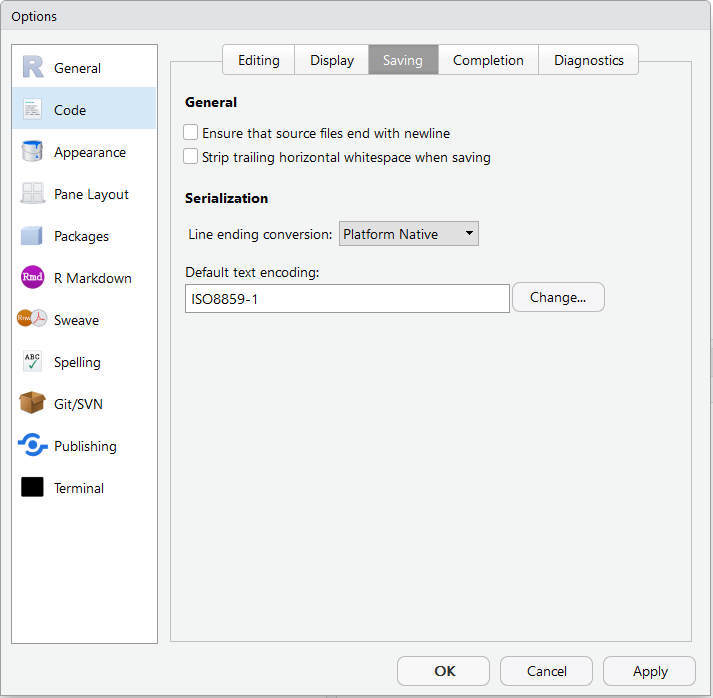
Note: If RStudio is not installed by default, you have to first click “Install” instead of “Launch”. If installation does not work, check your internet connection (see Prerequisites). During the installation process, you may be asked to create a new environment – you can safely create it:



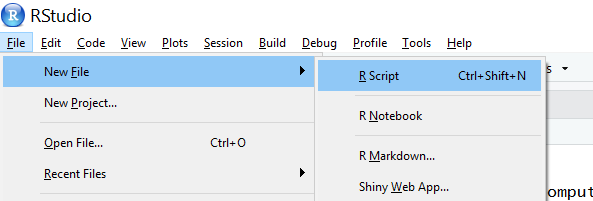
* + After RStudio is launched, you should see the main screen: **[Video #2]**



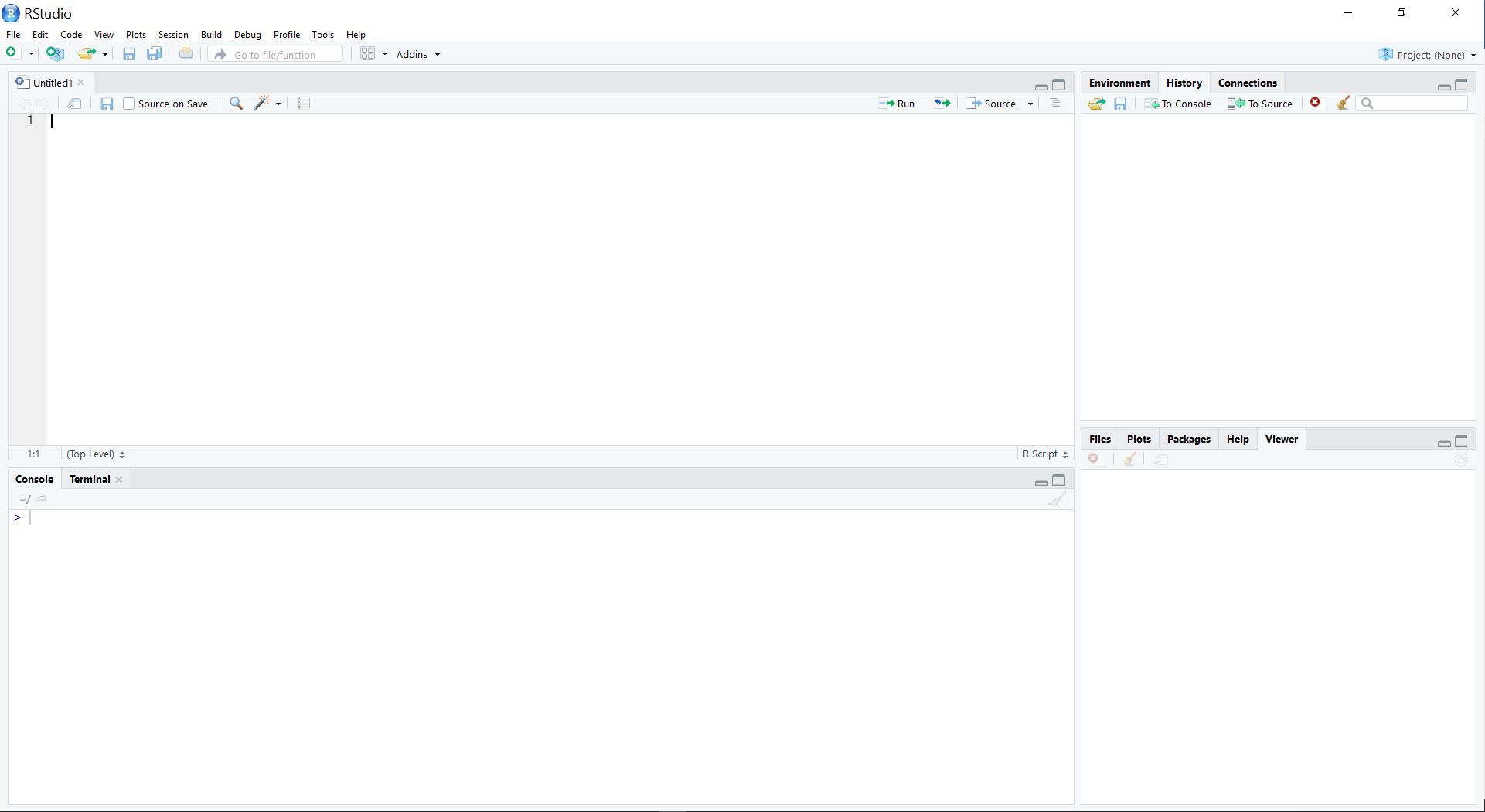
* + Set up default encoding for R files: **[Video #3]**
    - Tools → Global Options → Code → Saving → Default text encoding → Change to UTF-8 → OK → Apply

→ →→

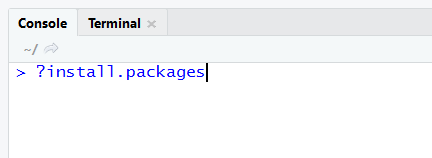
* + Click File → New File → R Script: **[Video #4]**



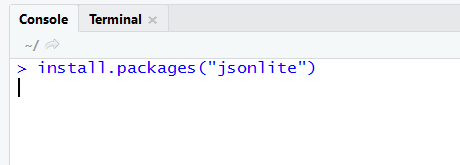
* + Investigate RStudio main screen: **[Video #4]**



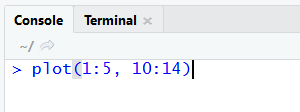
1. Code editor panel: used for writing/editing R code
2. Panel that shows session related information
3. Source code execution: write code here, execute it, and see results immediately.
4. Panel used mainly for the visualizations and help screen
   * Try out some basic R functionalities: **[Video #4]**
     + Get help for a command: write following code in Panel 3 and press enter:



* + - * + The help screen is displayed in Panel 4.
    - Install new packages: write the following code in Panel 3, and press enter:

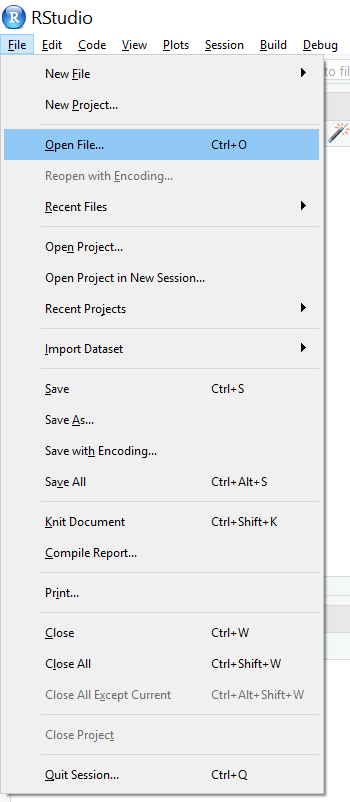


* + - * + The package gets installed, and you see the output in the Console screen. If you get an error message, it indicates an internet problem. Solve the connectivity problem, and try again – you need to be able to install new packages to complete the demo.
    - Plot a data set: write the following code in Panel 3, and press enter:

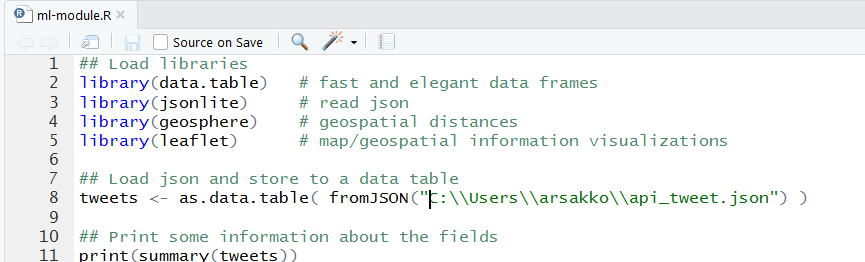


* + - * + The plot is shown in Panel 4.

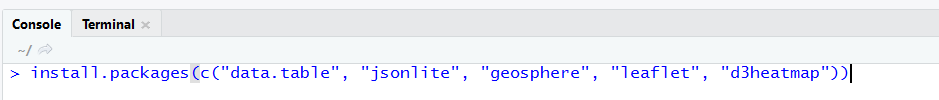
1. **Load the data and visualize it [Video #5]**
   * Open the sample source code file (ml-module.R) that was downloaded:

****

* + In the code editor (Panel 1), modify line 8 so that it points to the correct location (where you downloaded the sample data file).

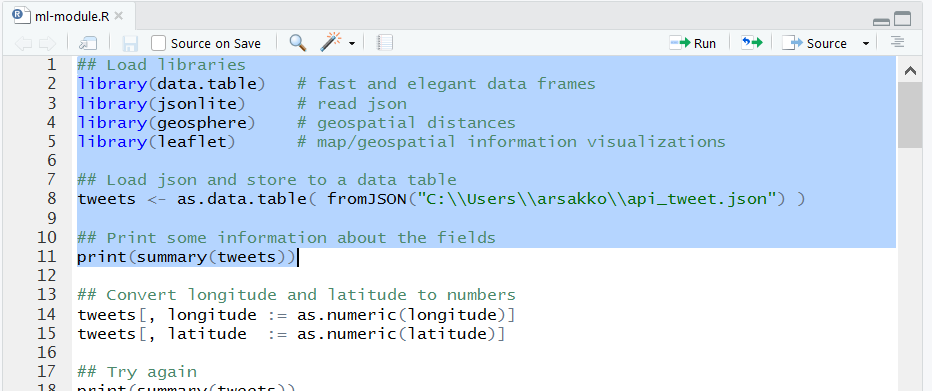


* + **Hint: To avoid possible typos when specifying the path you can:**
    - Navigate to the folder containing the file api\_tweet.json.
    - Click on the file path, which is the top line above the directory window.
    - When you click on the top line, it will change to the DOS path name. (e.g., C:\users\...)
    - Press CTRL-C to copy the path name.
    - In line 8 of the R code, replace the existing path, as shown above, with the path to your api\_tweet.json file.
    - Be sure to change each single backslash (\) to a double backslash (\\). This is necessary for R to correctly read your path name.
    - Be sure to include the name of the input file (api\_tweet.json).
  + Before executing the code, install the packages that are needed to run the ml-module.R example; write the following code in Panel 3 (lower-left panel), and execute it:

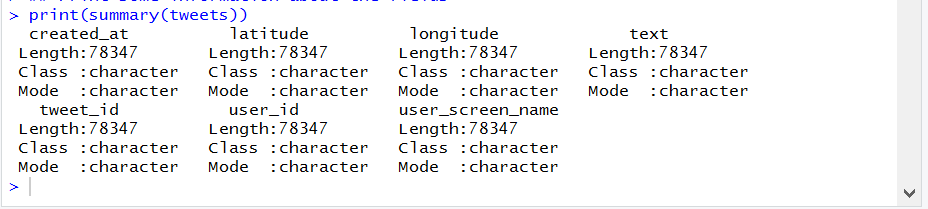


install.packages(c("data.table", "jsonlite", "geosphere", "leaflet", "d3heatmap"))

* + - * + Packages are installed (this takes some time)
  + Select lines 1-11 in Panel 1 and execute it by pressing the button “Run” in the same panel:

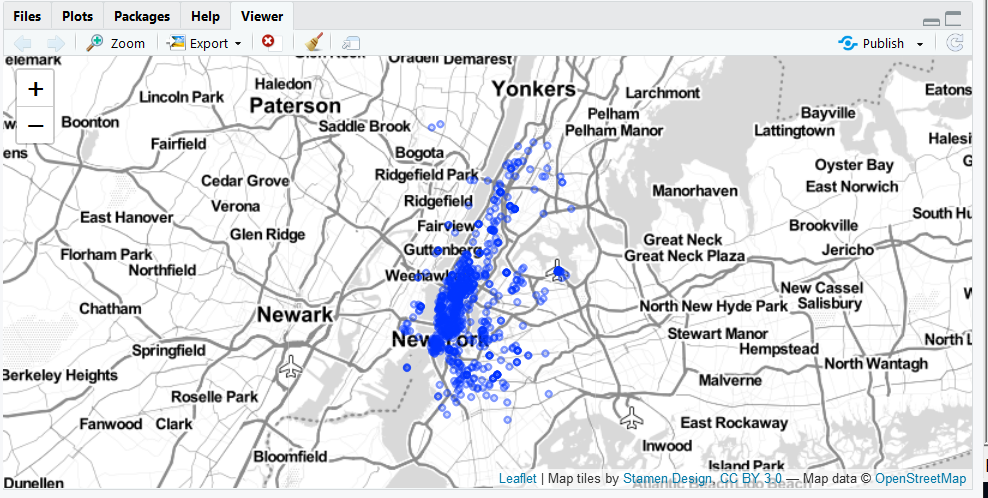


* + - * + If the data is correctly loaded, you will see a summary of the fields printed in Panel 3:



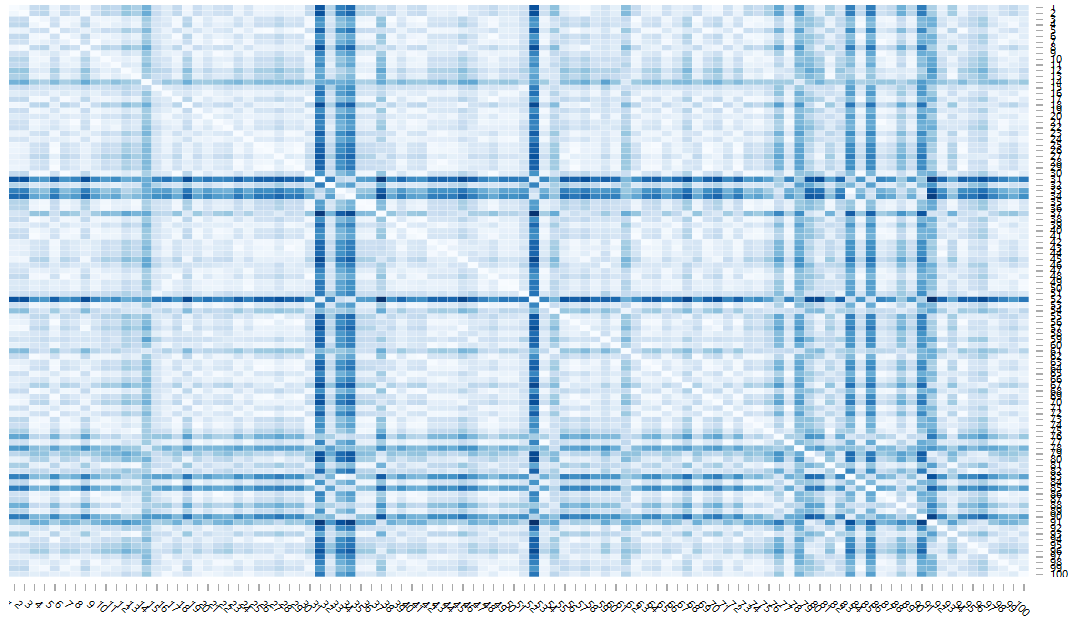
If you get an error, check that the file location is correct.

* + Select lines 1-33 and run the code the same way. This shows the tweets on the map:



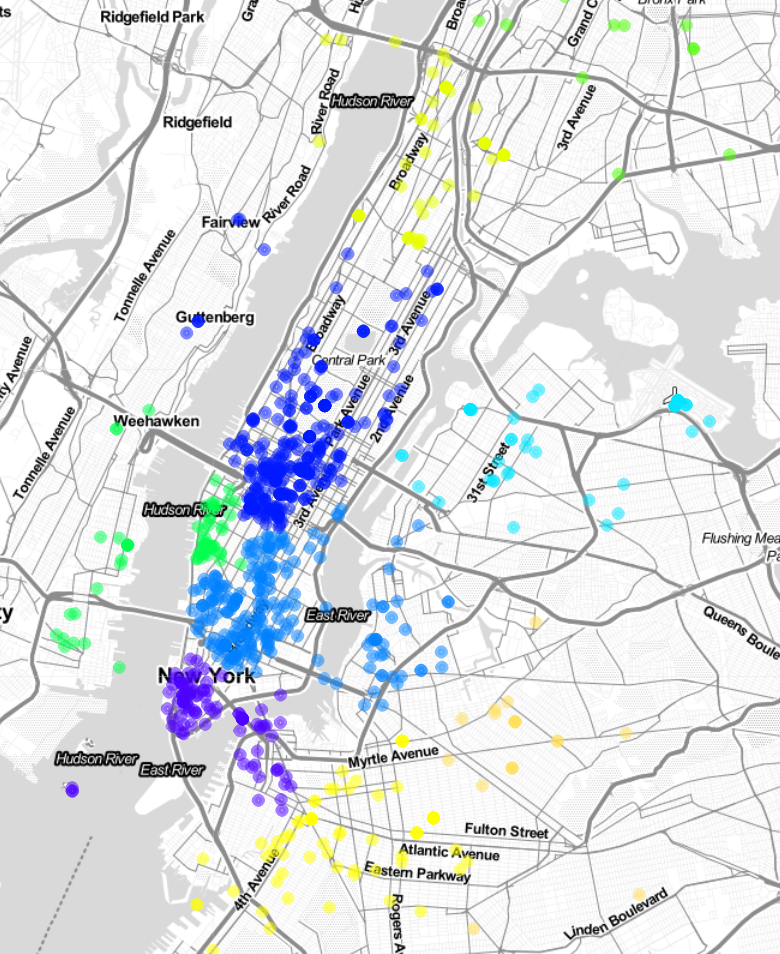
* + The background map (of New York City and surrounding areas) is loaded from the Internet during the visualization. If you can’t see it, you should check the internet connection. If you are working through the Nokia VPN, then another possible reason for not being able to see the background map, is that proxy settings have not been specified properly. Nevertheless, **you can continue running the remaining part of this demo even without the background map.**
  + You can zoom in the map and click the circles. To see the map in a separate window, click the “Zoom” button above the map.
  + Run the source code in lines 35-46; it plots the same map, but the color of the circles is different depending on the timestamp of the tweet.

1. **Unsupervised learning example: [Video #6]**
   * Run the source code in lines 67-75. This calculates the distances between pairs of tweets, and shows the distance matrix as a heatmap. The result looks like this:



These are 100 random tweets from the sample, so your result will not look exactly the same as the above graphic (or in the video for this course).

* + The source code in line 78 prints the same matrix but in such a way that the rows and columns are ordered using hierarchical clustering.
  + In lines 80-84 you execute the hierarchical clustering yourself.
  + In line 86 the result of clustering (a cluster identifier) is attached to the tweets.
  + Lines 88-99 visualize the tweets on the map, but this time the color indicates the newly calculated cluster:



**Revision History:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author(s)** | **Notes** |
| 0.8 | 6/15/2018 | Arto Sakko | Near-final version, based on the June 1, demo for AI Espoo Newscast. Original video has been split into seven shorter videos according to the sections of this demo. |
| 0.82 | 6/15/2018 | Tom Tirpak, Henna Falk, Marika Rytkonen | Small updates, per beta-testing. |
| 0.9 | 6/19/2018 | Arto Sakko, Tom Tirpak | Added Revision History. Updated reference for input data used in this demo. Clarified prerequisites. Included link for download files from AI/ML CoDe Team Site. Small edits for clarity. |

1. The dataset contains a set of twitter messages with latitude/longitude coordinates. It was collected by A. Himanen (Nokia Software, Digital Intelligence) for his Master’s thesis “Human-AI interaction in intelligent decision support systems”, Aalto University, School of Science, Department of Computer Science, Espoo, Finland, 2018. [↑](#footnote-ref-1)