BRITISH COLUMBIA | BCStats

Data analytics using R: The BC Stats experience (so far)

Statistics Canada: Economic Statistics Forum
2018-05-15

BRITISH COLUMBIA

README.md https://github.com/bcgov/CANSIM-dataviewer

CANSIM-dataviewer

This repo contains the code for the extraction, tabulation, and visualisation of data from Statistics Canada's CANSIM database.

The data presented here are tailored for users in British Columbia.

The repo contains the code, written in R, including for applications using the Shiny package.

Data series

These pages contain links to data sources and related reference materials at Statistics Canada, and to the R scripts developed for extracting, tabulating, and viewing those data.

- Consumer Price Index
- Labour Force Survey
- New Housing Price Index

more about CANSIM

CANSIM is Statistics Canada's repository of socio-economic data.

- CANSIM table directory, by subject

CANSIM is the comprehensive database of socio-economic statistics, maintained by Statistics Canada. The database contains a wide variety of data series, from the familiar (unemployment rates from the Labour Force Survey and the Consumer Price Index) to more obscure topics (total sales by vending machine operators; vector v101256214 in Table 080-0028).

In January, Manolo introduced himself to me, and said he's spotted this page with my name attached … and was curious about how BC Stats is using R.

Why R

Why R – the tool

- Statistical programming environment
- Open source
- Continuously developing
- Being taught widely
  - University programs – from economics to physics
  - Online (e.g. Coursera data science program)

3

The program R was—and continues to be—developed by people doing statistics (or "data science"), from a statistics background. So it's people like us developing a tool for people like us.

There are packages already developed that do a huge range of statistical and data science functions, from reading messy Excel files to X13-ARIMA seasonal adjustment to interactive web display of tables and charts.

Open source means it's got packages that people want and use. And it's free!

Lots of university programs are teaching R as the statistical / quantitative tool, as well as legitimate online programs such as the Coursera "Data Science Specialization" https://www.coursera.org/specializations/jhu-data-science

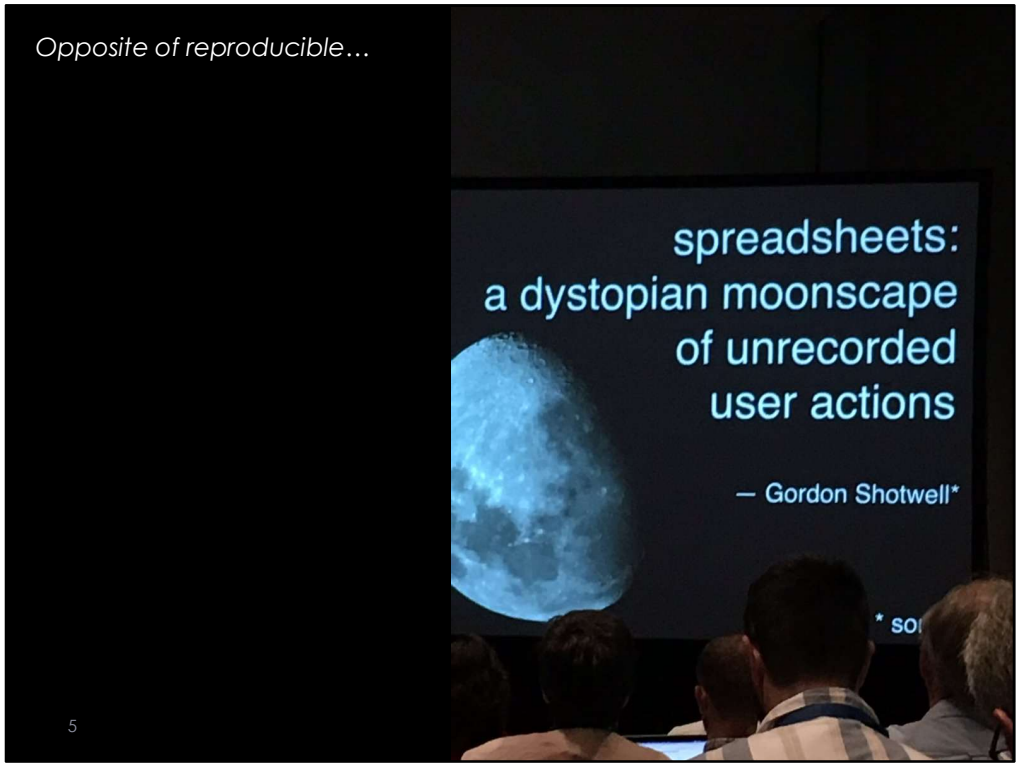
And because it's open source it's continuously developing; there's no need to wait for some corporate suits (e.g. Microsoft) to add a new function.

Why R – the workflow

- "Opinionated Analysis Development"
  - Hilary Parker
  - https://peerj.com/preprints/3210/
- Maximize the probability that your analysis is:
  - Reproducible
  - Accurate
  - Collaborative

4

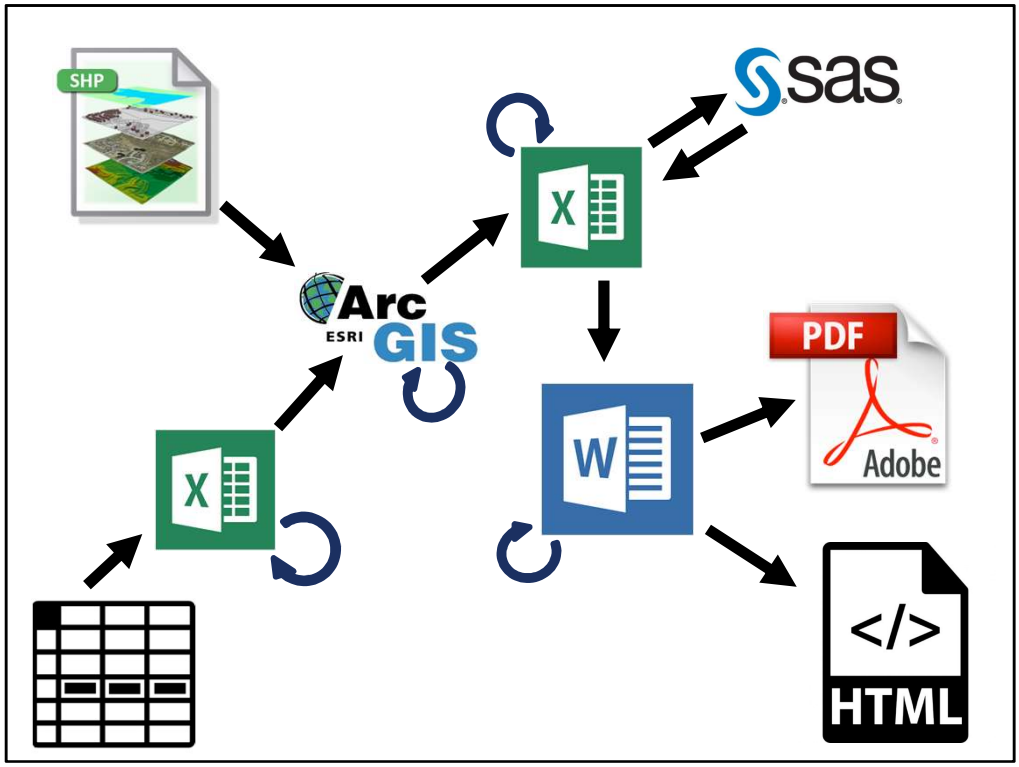
Hilary Parker (2017), "Opinionated analysis development", PeerJ Preprints, https://peerj.com/preprints/3210/



Source: Jenny Bryan, https://speakerdeck.com/jennybc/spreadsheets

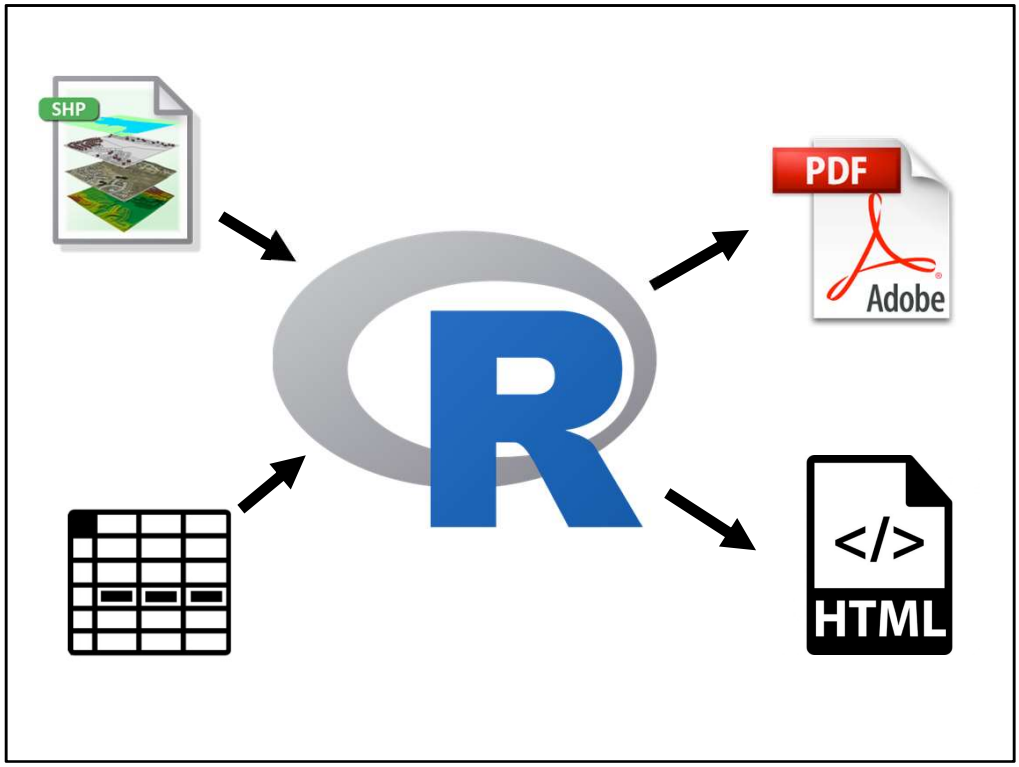
We have all seen spreadsheets with formulas that link all over the place, including to other sheets in the file.

And columns that all have a formula, except that one cell where the creator decided to hard-code a number



This is what a typical workflow (including spatial layers) might look like, with four different tools (Excel, ArcGIS, SAS, and Word) employed to create a PDF and HTML outputs.

Source: Andy Teucher, Ministry of Environment, Province of B.C.



R can work with spatial data, provide all of the data manipulation and visualization functionality of Excel, the statistical functions of SAS, and the text / writing functions of Word.

This leads to a much more streamlined and efficient workflow

Source: Andy Teucher, Ministry of Environment, Province of B.C.

More than just R

More than just R

- RStudio
  - IDE
  - incorporate Python & SQL code
  - Notebook and report output options
- Developer's Exchange
  - https://bcdevexchange.org/
- GitHub
  - https://github.com/bcgov
  - also, bcgov-c – behind the curtain
- Open data
- Open science

9

GitHub note:
https://twitter.com/JennyBryan/status/966903491259121666

Valid reasons not to participate in open science practices

Casper J. Albers*

Abstract

The past years have seen a sharp increase in the attention for open science practices. Such practices include pre-registration and registered reports, sharing of materials, open access publishing and attention to reproducibility of research. Despite the overwhelming amount of evidence highlighting the benefits of open science, some researchers remain reluctant. In this paper, I will outline valid reasons for researchers not to participate in open science practices.

Discussion

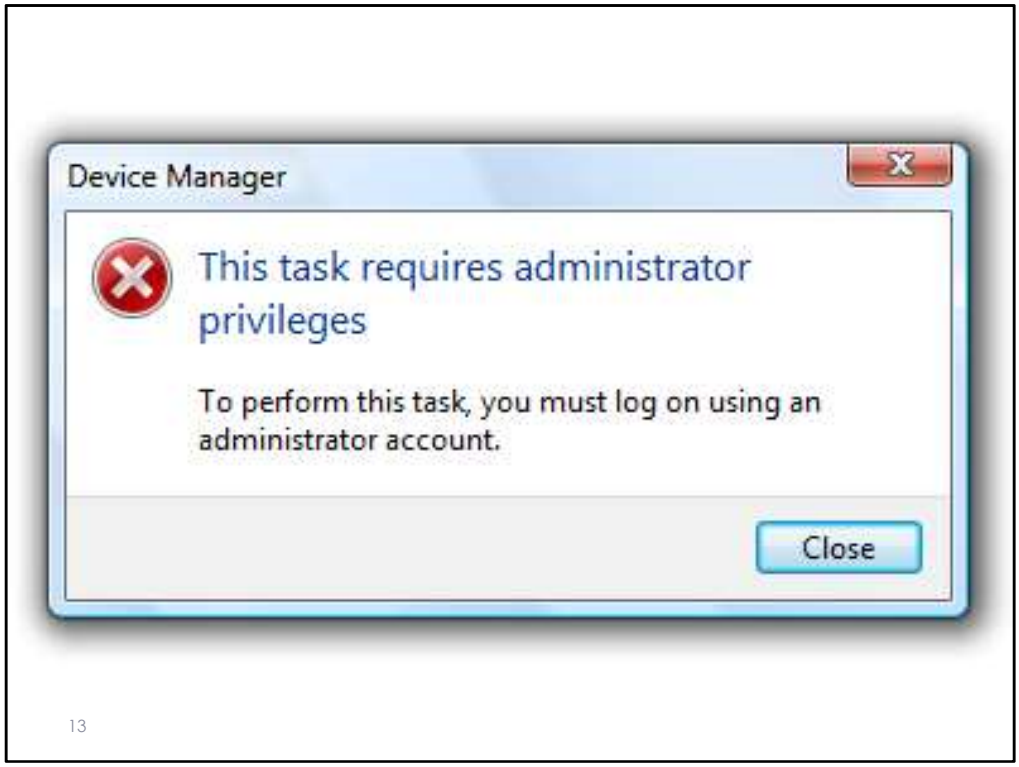
There are no valid reasons.

*Heymans Institute for Psychological Research, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. c.j.albers@rug.nl

Enablers

How do you start?

- Administrative privileges
- Open data & licenses
  - B.C. government
  - Statistics Canada
- Open development & licenses
  - Apache 2.0
  - Creative Commons
- A community of users
  - across BC Stats and the B.C. Public Service … and beyond



Four or five years ago, this is what I would have seen if I'd try to install R on my work computer, including all of the packages. It was a labour-intensive, time-consuming hassle to get R installed, and tough to keep maintained—remember what I said about packages being constantly updated.

Now I have admin privileges, and can do my own R maintenance.

README.md

bcgovr

BCDevExchange Delivery build passing

Overview

An R package to support development of R-based projects and packages following bcgov open source guidelines and policies.

Features

Currently there are two main functions for auto-populating a new R-based data analysis or package project directory with folders & files that encourage best practice in scientific computing and ensure the project has all the required bcgov items:

- `analysis_skeleton()` # starting a new data analysis project
- `package_skeleton()` # starting a new R package

These functions are most easily used by using the *bcgovr Project Template* as described below.

https://github.com/bcgov/bcgovr

14

Some of our colleagues in the Ministry of Environment have created an R package that auto-populates a new github repo with all of the licenses and other standard documentation. They've shared it with their colleagues across the BC Public Service, first through the bcgov github, and now it's on CRAN

cancensus

build passing | CRAN 0.1.7 | downloads 244/month

Access, retrieve, and work with Canadian Census data and geography.

- Download data and Census geography in tidy and analysis-ready format
- Convenience tools for searching for and working with Census regions and variable hierarchies
- Provides Census geography in multiple R spatial formats
- Provides data and geography at multiple Census geographic levels including province, Census Metropolitan Area, Census Division, Census Subdividision, Census Tract, and Dissemination Areas
- Provides up-to-date data for the 2016, 2011, and 2006 Censuses

Reference

Cancensus home page and reference guide

https://github.com/mountainMath/cancensus

15

The `cancensus` package was developed by a private individual in Vancouver, who decided to make his wrangled and reformatted Census data tables accessible to other users.

Examples

README.md https://github.com/bcgov/CANSIM-dataviewer

CANSIM-dataviewer

This repo contains the code for the extraction, tabulation, and visualisation of data from Statistics Canada's CANSIM database.

The data presented here are tailored for users in British Columbia.

The repo contains the code, written in R, including for applications using the Shiny package.

Data series

These pages contain links to data sources and related reference materials at Statistics Canada, and to the R scripts developed for extracting, tabulating, and viewing those data.

- Consumer Price Index
- Labour Force Survey
- New Housing Price Index

more about CANSIM

CANSIM is Statistics Canada's repository of socio-economic data.

- CANSIM table directory, by subject

CANSIM is the comprehensive database of socio-economic statistics, maintained by Statistics Canada. The database contains a wide variety of data series, from the familiar (unemployment rates from the Labour Force Survey and the Consumer Price Index) to more obscure topics (total sales by vending machine operators; vector v101256214 in Table 080-0028).

For this project, we're experimenting with using R to access CANSIM tables, and run summary analysis and create charts

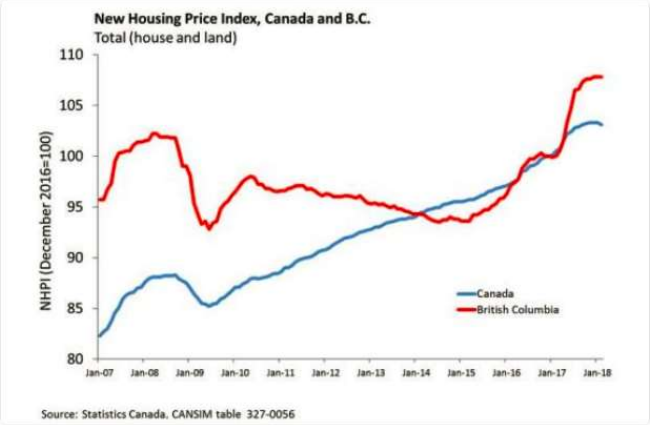
Example: CANSIM data manipulation

- Downloading a table from CANSIM
- Wrangling to select relevant series
- Summary tables
  - Including month-over-month percentage change
- Creating chart for publication

BC Stats
@bcstats

Following

The cost of new housing in B.C. was unchanged in February from January, but rose 7.8% over the same month in 2017. Find out more at www2.gov.bc.ca/gov/content/da ...



19

11:59 AM - 12 Apr 2018

Example: Voter List Quality measurement

- Objective: assess the accuracy of records in the B.C. list of voters at three points in time
- Wrangling the voter list
  - ~3.2 million records at each point in time
- Survey sampling (from CSV)
- Modeling and estimation
  - logistic regression
- Summarization of lists and models
- Visualization (tables and charts)
- Text / reporting
  - using Rmarkdown and bookdown
  - the full report is rendered using R

What's next?

Lots!

- Getting more people started with R
- Expand expertise
- Further training
- More analysis and modeling
- Packaging data
  - Extending `cancensus`
- More reporting
  - Including web-based reporting with Shiny

Martin Monkman
Provincial Statistician & Director
BC Stats
martin.monkman@gov.bc.ca
250.216.5848

www.bcstats.gov.bc.ca
@BCStats