Writing Data Science Reports

Yelp Personalized Reviews

1 Introduction

Using data from Yelp Dataset Challenge, we wanted to provide personalized restaurant recommendations to users based on the reviews they previously left on Yelp. To do so, we used one user's past reviews and ratings, as well as other users' reviews and ratings of the establishment we are predicting the rating of. The training is therefore made for a specific user and the predictions based only on his or her specific preferences.

One of the main applications of this project is to be able to recommend restaurants that a user is likely to appreciate. This could lead to a new business opportunity for Yelp: they could partner with restaurants willing to offer discounts to users likely to enjoy the restaurant and make a commission for every positive recommendations. This business model is currently used by TheFork (LaFourchette), a european company launched in 2007 linking users and restaurants offering discounts through their app with TheFork taking a commission on every transaction. Offering this service without personalized recommendations, TheFork managed to expand to 4 countries in 7 years and was acquired by TripAdvisor in 2014 for $140M. Leveraging Yelp's data, such a service could easily be implemented and become a solid source of income for Yelp before other competitors enter this niche.

2 Dataset

Yelp has made a subset of its data publicly available in the context of the Yelp Dataset Challenge [1]. Data is provided for users, tips, reviews check-ins and businesses. There are 1.1 million reviews and 250,000 users in the dataset provided, from which we removed all the reviews that were not about restaurants.Then we only kept users that had given at least 20 reviews.

1

Nike Says Its $250 Running Shoes Will Make You Run Much Faster. What if That's Actually True?

By KEVIN QUEALY and JOSH KATZ JULY 18, 2018

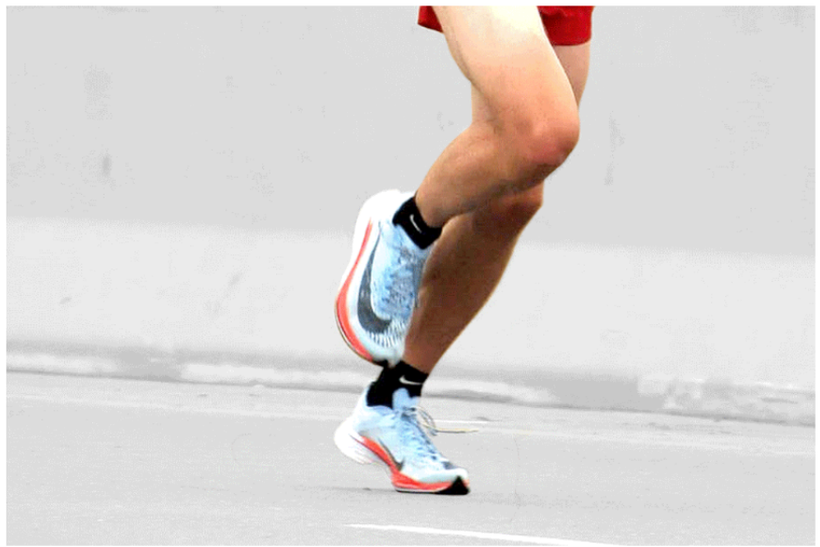


Illustration by Agnes Lee

If a running shoe made you 25 percent faster, would it be fair to wear it in a race? What about 10 percent? Or 2 percent? The Nike Zoom Vaporfly 4% — a bouncy, expensive shoe released to the public one year ago — raises these questions like no shoe in recent distance running history.

Nike says the shoes are about 4 percent better than some of its best racing shoes, as measured by how much energy runners spend when running in them. That is an astonishing claim, an efficiency

IN-CLASS ACTIVITY

Work in groups of 2 to 3 students.

Read and/or skim at the reports, and note all that you find be **good** in the reports, and all that you would suggest **should be improved** and why.

DATA SCIENCE REPORT: CHECKLIST

- Clear narrative.
- Problem solving.
- Interpretation and interpretability.
- Reproducible research.
- Discuss limitations.

DATA SCIENCE REPORT: CHECKLIST

- **Clear narrative.**
- Problem solving.
- Interpretation and interpretability.
- Reproducible research.
- Discuss limitations.

REPORT CONTENT

What

- what is the topic of the report?
- what are the big questions that the report aims to answer?

Why

- why is it important to answer these questions?

How

- how did you answer the questions (data, methods)?

What (cont'd)

- what are the conclusions?
- what are the limitations of your data/analyses?

INTRODUCTION

- Summary of the study, data and methods, as well as any relevant substantive context and background.

- The "big questions" answered by your data analyses, and summaries of your conclusions about these questions.

- The limitations of the data and of your analyses and how these may affect your conclusions.

Guideline #1

Your introduction / report should allow to answer the questions: *what is the problem? why it is an interesting problem? how did you solve the problem?*

TELL A STORY

Clear, concise story of how you approached the problem:

- Every analysis should be **motivated**:
  why did you decide to plot this graph and/or model this relationship to start with?
- Every result should be **discussed**:
  how can we interpret this result? how does it help us advance our understanding of the problem at hand?
- Every insight should be followed by a relevant **next step**:
  usually, the motivation for the next set of graphs and analyses, or future work

Guideline #2

Your report should flow as a story that allows the reader to understand your reasoning process about the problem.

MEANINGFUL TITLES

MEANINGFUL TITLES

Where these estimates come from

Measuring shoe effects using statistical models

Pros of this approach: Tries to control for race conditions, weather, gender, age, pre-race training and a runner's previous race times.

Cons of this approach: Still not a randomized controlled trial.

Comparing groups of runners who completed the same two races

Pros of this approach: Follows athletes of similar ability who ran in identical conditions.

Cons of this approach: Runners could save their special shoes for when they expect to have a fast race.

Guideline #3

Favour meaningful titles over generic ones.
This will allow different levels of reading of your report, and help the reader navigate your document.

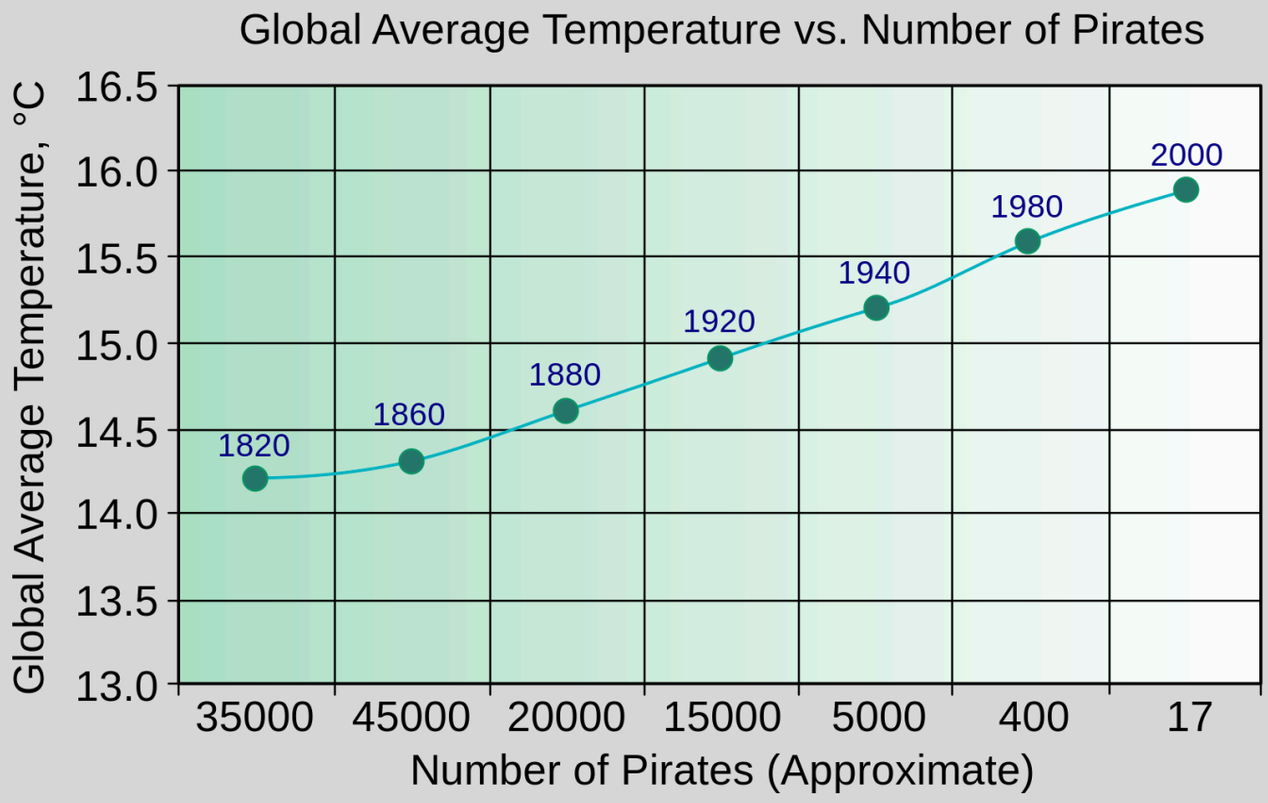
DATA SCIENCE REPORT: CHECKLIST

- Clear narrative.
- **Problem solving.**
- Interpretation and interpretability.
- Reproducible research.
- Discuss limitations.

PROBLEM SOLVING

Does the report build on a true reasoning about the problem at hand?

Our explorations revealed a correlation between the number of pirates and overall temperature over time.

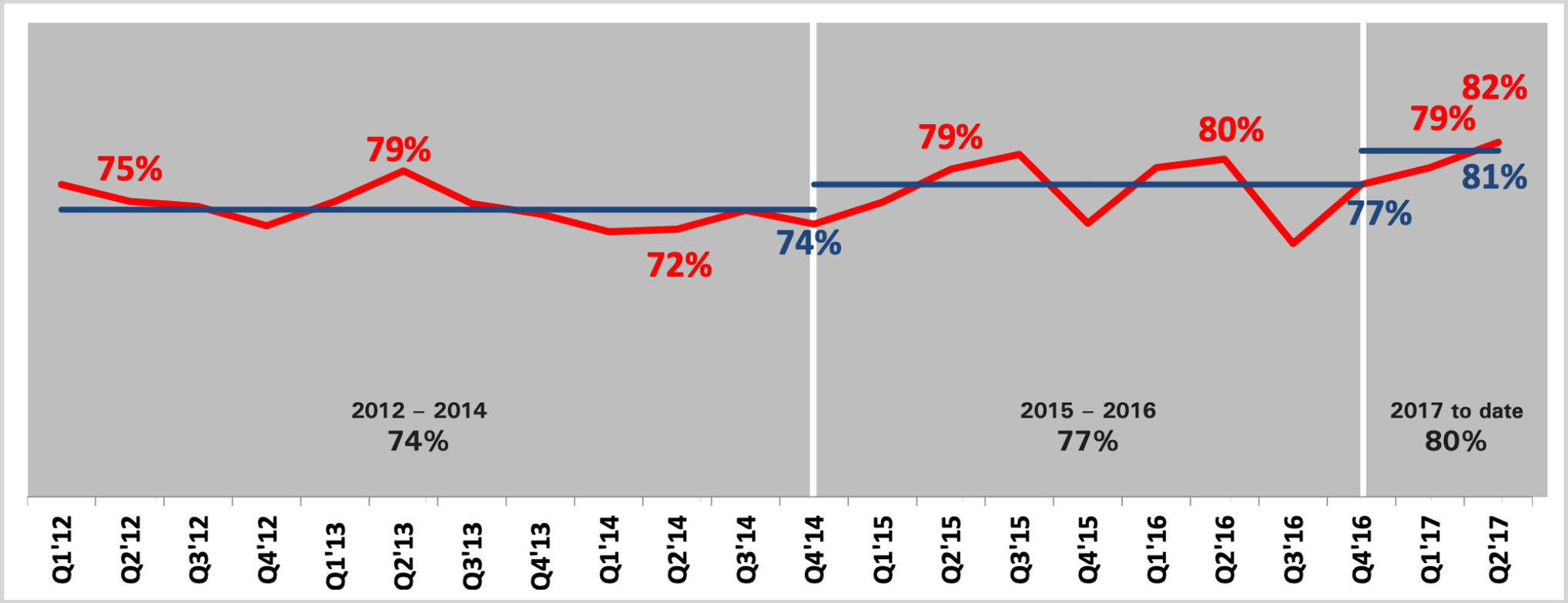


PROBLEM SOLVING

Does the report build on a true reasoning about the problem at hand?

How can TTC improve their service?

We observe that the satisfactory rate of TTC riders has consistently increased, overall, over the past few years.



https://www.ttc.ca/About_the_TTC/Commission_reports_and_information/Commission_meetings/2017/September_5/Reports/3_Q1-Q2_2017_Customer_Satisfaction_Survey_Results.pdf

PROBLEM SOLVING

Does the report build on a true reasoning about the problem at hand?

Variations in the questionnaires and survey methods used to collect riders' feedback can have an impact on the responses. First, we check whether there is a significant change in the survey methods.

[Investigation finds no change]

One must be wary of comparing data from different samples. We verify that the difference is not due to sampling error, that is, we compare the number, demographics and characteristics of the surveyed populations in the different samples.

[Investigation finds comparable samples]

A variety of factors could explain the increased satisfaction in the service. In the following, we investigate whether the following are true:

- **faster vehicles / less travel duration**
- **less waiting time / increased frequencies of the trains**
- **less crowded vehicles**
- **better customer service**

Guideline #4

Your questions and analyses, and therefore your report narrative, should focus on a meaningful approach to further understanding and solving the problem at hand.

DATA SCIENCE REPORT: CHECKLIST

- Clear narrative.
- Problem solving.
- **Interpretation and interpretability.**
- Reproducible research.
- Discuss limitations.

INTERPRETATION AND INTERPRETABILITY

Does the report provide sufficient relevant information for the reader to interpret the results?

The levels of carbon in blood of [the sampled people] ranged from 3% to 5.8%.

INTERPRETATION AND INTERPRETABILITY

Does the report provide sufficient relevant information for the reader to interpret the results?

The levels of carbon in blood of [the sampled people] ranged from 3% to 5.8%.

On average, a healthy adult has a level less than 2.3%; an adult smoker has levels between 2.1% to 4.2%.

Guideline #5

Compare effects to other known effects to facilitate interpretation and interpretability of the **sign** and **magnitude** of the effect.

INTERPRETATION AND INTERPRETABILITY

Does the report allow for human-friendly explanations?

We want to understand what are the ingredients that make students successful.

We use data from the University, that comprises of students demographics, their background, their grades, their extra-curricular activities, etc…

We build a neural network to predict student success.

Read more about human-friendly explanations in ML:
https://christophm.github.io/interpretable-ml-book/explanation.html

INTERPRETATION AND INTERPRETABILITY

Does the report allow for human-friendly explanations?

We want to understand what are the ingredients that make students successful.

We use data from the University, that comprises of students demographics, their background, their grades, their extra-curricular activities, etc…

We use a regression model to predict student success.

Read more about human-friendly explanations in ML:
https://christophm.github.io/interpretable-ml-book/explanation.html

Guideline #6

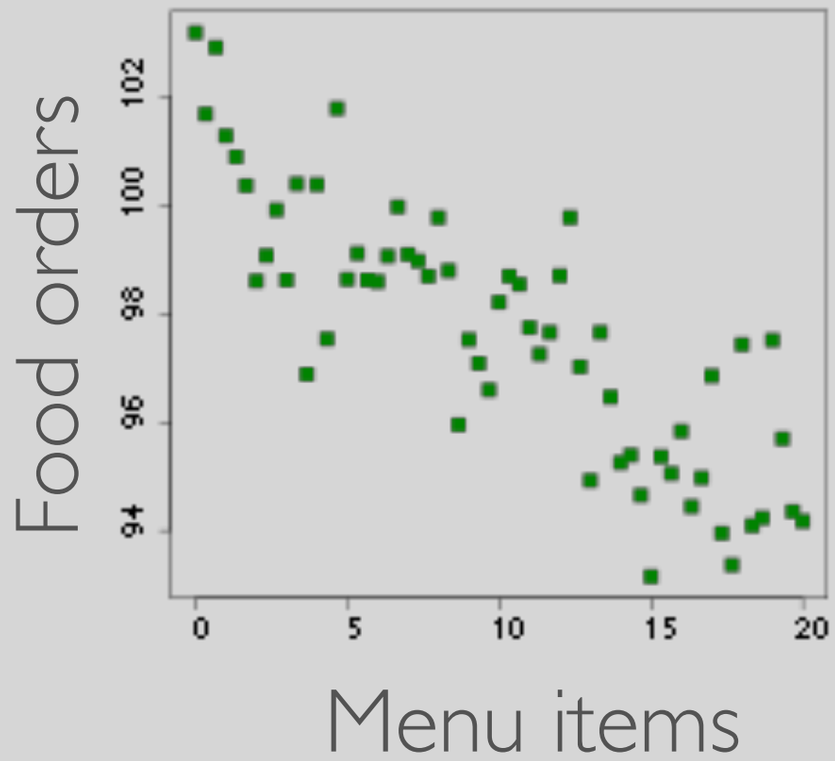
Favour human-interpretable methods and models whenever possible. This will facilitate the interpretability of your approach and results.

INCONCLUSIVE RESULTS
RESULTS CONTRADICTING EXPECTATIONS

All results are potentially insightful!

Because different customers have different taste, a restaurant that offers a greater choice of menu items may attract more customers.

The following graph shows food orders against number of menu items.



Contrary to what we expected, the data suggest an inverse correlation between orders and choices.

Several factors may explain these results.

- **restaurants offering fewer items are faster at delivering, due to easier logistic at their hand.**
- **ordering a full meal is easier and faster when there are less options.**
- **restaurants offering more items tend to be upper scale restaurants with “à la carte” expensive dishes.**

Guideline #7

Inconclusive results, or results contradicting what you expected should not be discarded too easily, especially when a strong effect is expected.

Rather, these results should be the entry point to a re-framing of the problem and further analyses.

DATA SCIENCE REPORT: CHECKLIST

- Clear narrative.
- Problem solving.
- Interpretation and interpretability.
- **Reproducible research**.
- Discuss limitations.

More guidelines and tips on reproducible research:
https://github.com/rdpeng/courses/blob/master/05_ReproducibleResearch/Checklist/Reproducible%20Research%20Checklist.pdf

REPRODUCIBLE RESEARCH

Does the report contain all of the details to reproduce your analyses?

To determine what people like to eat, we look at restaurant reviews.

We analysed correlations in our datasets.

We find a greater correlation between the ambiance of the restaurant and the review ratings, suggesting that the type of cuisine plays a less important of a role than the conditions in which people eat in their preference of a location.

REPRODUCIBLE RESEARCH

Does the report contain all of the details to reproduce your analyses?

We are interested in what type of cuisine people like the most when they eat out.

To answer this question, we look at reviews on a popular reviewing website: Yelp. Yelp releases a subset of its data in the Yelp Challenge. For this study, we use the Challenge #3, that comprises of the reviews for the period of 2013-2015, for 12,345 restaurants in North America. The dataset can be accessed at : [URL]

...

REPRODUCIBLE RESEARCH

Does the report contain all of the details to reproduce your analyses?

Method

There is no direct way of measuring what type of food people prefer to eat when they go out, as many other factors, such as the price of the meal, the distance of the restaurant, the wait time, etc… may impact their decision.

We suggest to reframe the problem as: what is the most popular type of cuisine that people happen to eat, when they eat out?

We look at the number of reviews against the type of cuisine served in restaurants.

REPRODUCIBLE RESEARCH

Does the report contain all of the details to reproduce your analyses?

Data processing

The dataset comprises meta-data specifying the type of cuisine served in a restaurant. The different values are [...].

Several restaurants happen to offer different types of cuisine. Below are two versions of a chart showing the number of reviews against type of cuisine. In the first chart, we retain only the restaurants who have only one type of cuisine.

In the second chart, we retain all establishments, and add the associated number of reviews for each different type of cuisine that the restaurant serve, meaning that the same review may be counted multiple times.

Guideline #8

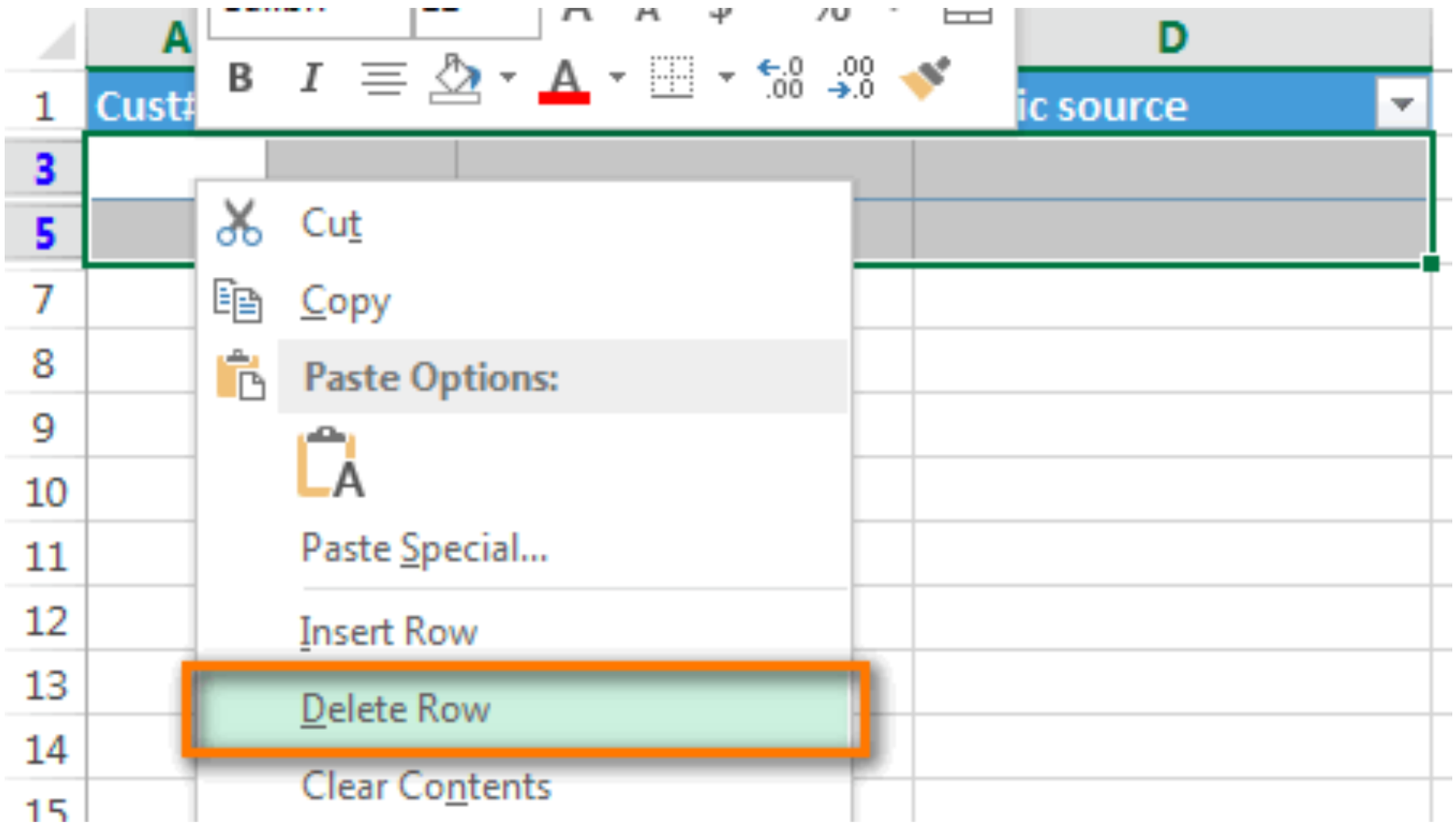
Make sure to include all of the details allowing for reproducibility of your analyses.

This includes:

- what are the data, and where to find them?
- what are the data processing steps you went through? (data cleaning, aggregation, derived measures, etc…)
- what are the methods you used? (specify parameters and external libraries where relevant)

REPRODUCIBLE RESEARCH

Did I do any manipulation that is not captured in my report?



Guideline #9

Do NOT edit data or results by hand!

Or if you do, things done by hand should be thoroughly documented.

REPRODUCIBLE RESEARCH

Is there anything that I have done by hand, that I can do programmatically instead?

CENTRAL INTELLIGENCE AGENCY

THE WORK OF A NATION.
THE CENTER OF INTELLIGENCE.

Report Information | Contact

Search CIA.gov... SEARCH

عربي 中文 Français Русский Español More ▸

HOME | ABOUT CIA | CAREERS & INTERNSHIPS | OFFICES OF CIA | NEWS & INFORMATION | LIBRARY | KIDS' ZONE

Library

Library
Publications
Center for the Study of Intelligence
Freedom of Information Act Electronic Reading Room
Kent Center Occasional Papers
Intelligence Literature
Reports
Related Links
Video Center

Home » Library » Publications » **Download**

Download the World Factbook

What: Posted below is the Central Intelligence Agency's most popular publication -- *The World Factbook* -- in compressed, .Zip file format, for the convenience of visitors who wish to download a personal copy.

Updates: We occasionally update these publications, so if you rely on the data for more than simple informational purposes, we recommend that you check back periodically for updates.

Format: The files for both publications are in standard HTML format. They have been compressed in .Zip format for ease of distribution.

Created with: The files were created with WinZip®, but any program that supports the .Zip format may be used to uncompress them.

To unzip (uncompress) the files:

1. Create a file folder or directory on your computer's hard drive or on a peripheral storage device in which to store each publication downloaded.
2. Download the desired file or files listed below to the location of your choice.
3. Unzip the file or files you downloaded into the folder or directory you created for storing the publication.

WORLD FACTBOOK ARCHIVES

- The World Factbook 2018
- The World Factbook 2017
- The World Factbook 2016
- The World Factbook 2015
- The World Factbook 2014
- The World Factbook 2013
- The World Factbook 2012
- The World Factbook 2011
- The World Factbook 2010
- The World Factbook 2009
- The World Factbook 2008
- The World Factbook 2007
- The World Factbook 2006
- The World Factbook 2005
- The World Factbook 2004
- The World Factbook 2003
- The World Factbook 2002

REPRODUCIBLE RESEARCH

Is there anything that I have done by hand, that I can do programmatically instead?

```
# import Country Comparison :: Area
import pandas as pd

url = 'https://www.cia.gov/library/publications/the-world-factbook/fields/279rank.html'
areadat = pd.read_html(url); type(areadat) # returns a list of dataframes

# select the first element in the list to access dataframe and
# print out first few observations
areadat[0].head()
```

Guideline #10

Have your computer do as much of the process as possible (i.e. programmatic data collection, programmatic cleaning of data, …).

DATA SCIENCE REPORT: CHECKLIST

- Clear narrative.
- Problem solving.
- Interpretation and interpretability.
- Reproducible research.
- **Discuss limitations.**

DISCUSS LIMITATIONS

Does the report clearly state the limitations of the data being used, and the impact on conclusions?

Round 13

Our dataset has been updated for this iteration of the challenge - we're sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 10 metropolitan areas across 2 countries. Round 13 has kicked off starting January 15, 2019 and will run through December 31, 2019.

Our results are derived from information about local businesses in 10 metropolitan areas across 2 countries (North America). As such, they may not generalise to businesses in smaller towns or geographical areas outside of North America.

DISCUSS LIMITATIONS

Does the report clearly state the limitations of the methods being used, and the impact on conclusions?

Measuring shoe effects using statistical models

Pros of this approach: Tries to control for race conditions, weather, gender, age, pre-race training and a runner's previous race times.

Cons of this approach: Still not a randomized controlled trial.

Comparing groups of runners who completed the same two races

Pros of this approach: Follows athletes of similar ability who ran in identical conditions.

Cons of this approach: Runners could save their special shoes for when they expect to have a fast race.

Guideline #11

Make sure you explicitly discuss the limitations of both the **data** and the **methods**, and the implications of these limitations on the conclusions.