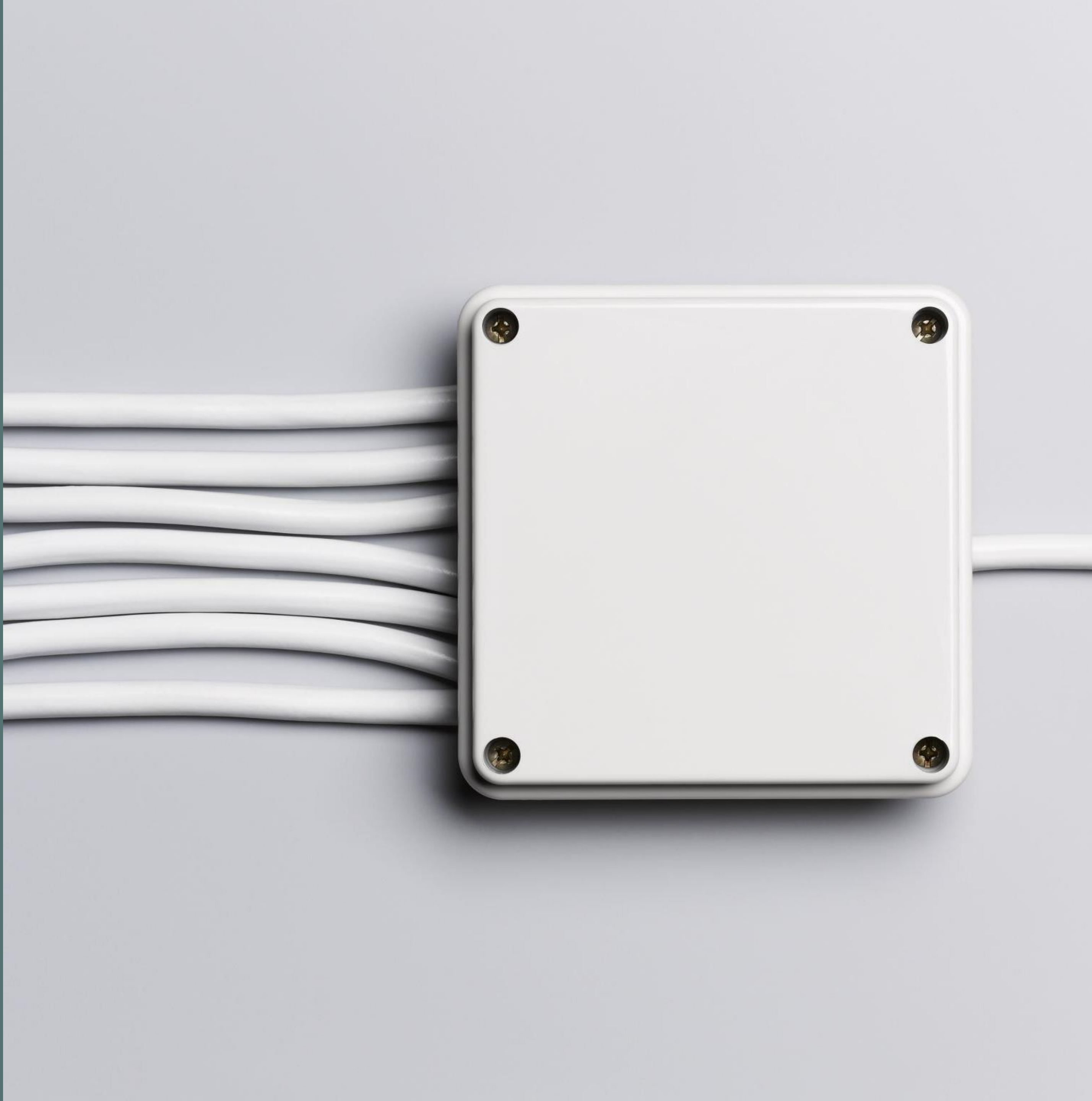
R + DATA SCIENCE FUNDAMENTALS

Justin Waterfield, MMCi

RTI International



WHAT EVEN IS R?

- Created by John Chambers in 1976 at Bell Labs.
    - Professor of Statistics from Stanford who was part of the monolith that was once the Bell Telephone Laboratories.
- Main purpose has originally been as a statistical language and data analysis.
    - Think modeling, classic statistics, and graphing.
- However, it has become popular and powerful overtime with the ability to import custom packages and especially the data science powerhouse “Tidyverse”.
- Open-source + Free!

DATA SCIENCE WORKFLOW

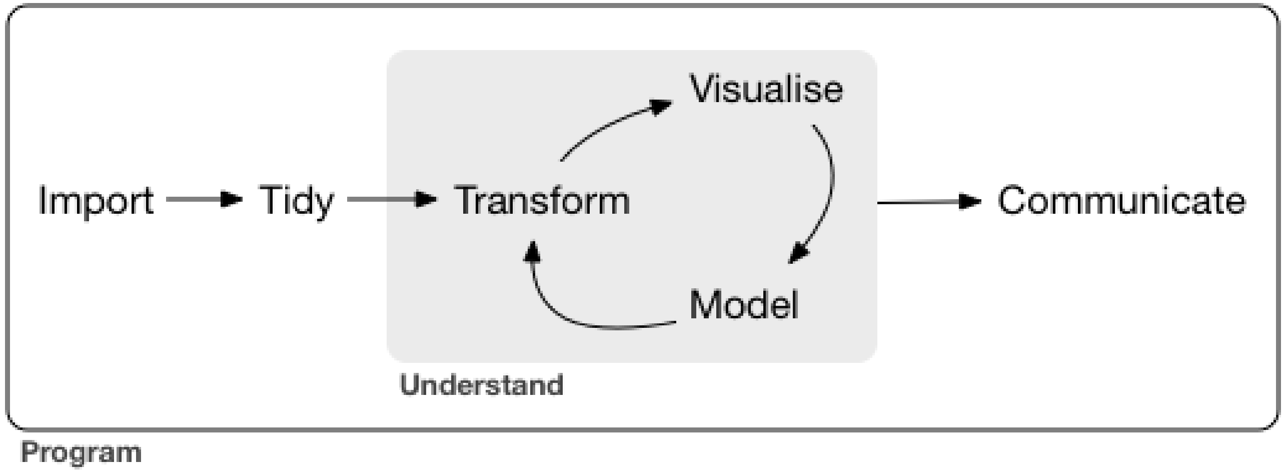
Data Collection + Storage
[Data Engineer]

Data Preparation
[Data Scientist]

Exploration & Visualization
[Data Analyst]

Experimentation &
Prediction
[Machine Learning Scientist]

THE "R" DATA SCIENCE WORKFLOW

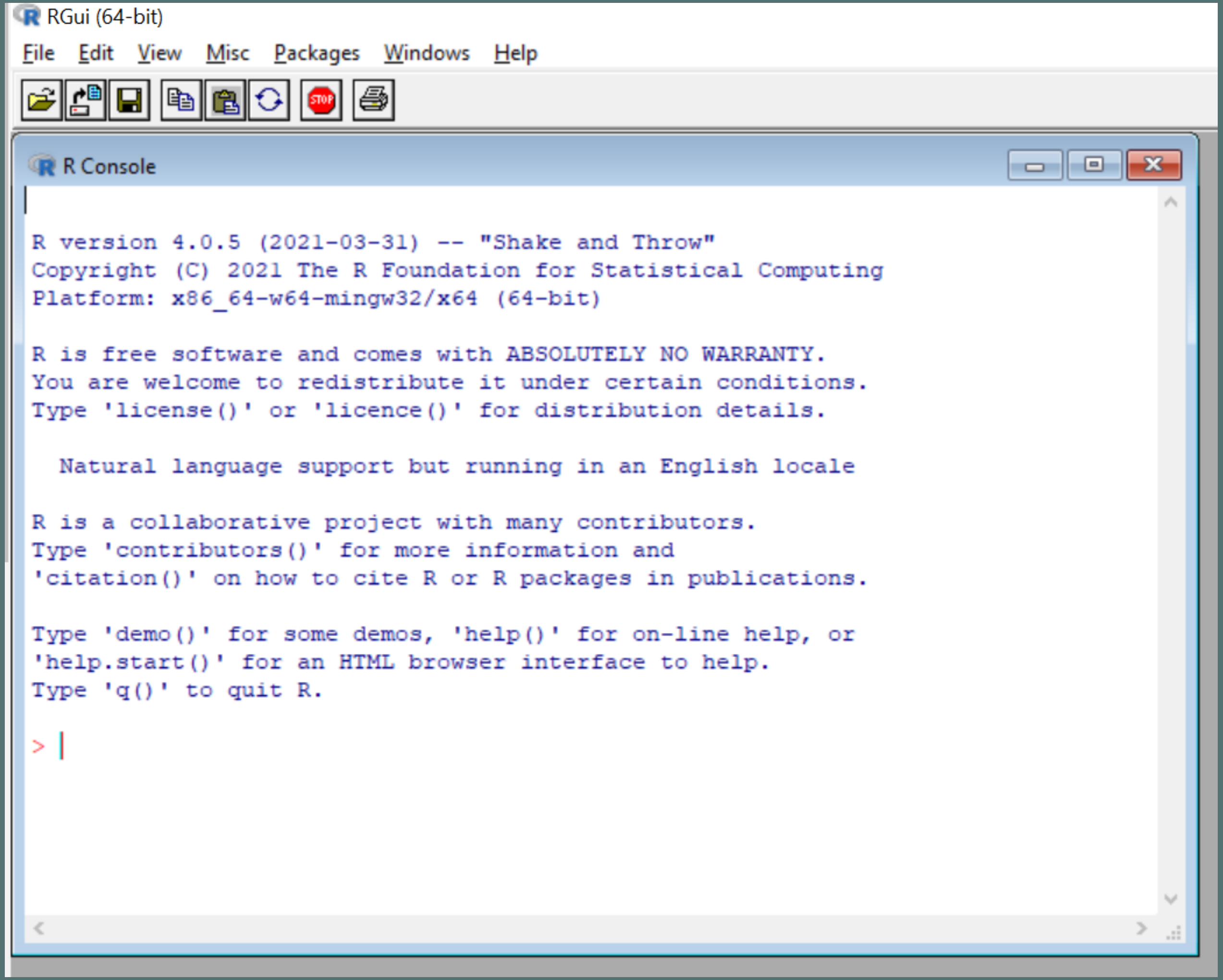


- Import: Take data from a dataset/database and load it into R.
- Tidy: Storing data in a consistent form. Think of this as labeling your axis or your cells in excel.
- Transform: Narrowing in on what you want to know, or what is of interest. How many people per city, married, kids, etc?
- Visualize: This is for human. Make it pretty!
- Models: This is for machines. Turn it into a tool.
- Communicate: Can you explain your results?

LET'S GET STARTED!

- Step 1:
  - Download R
  - Go to: The Comprehensive R Archive Network (r-project.org)
  - Pick your OS (Likely Windows or Mac)
  - Click "install R for the first time" **(Windows)**
    - Current version 4.1
  - Click R-4.0.5.pkg **(Mac)**
  - Run and install

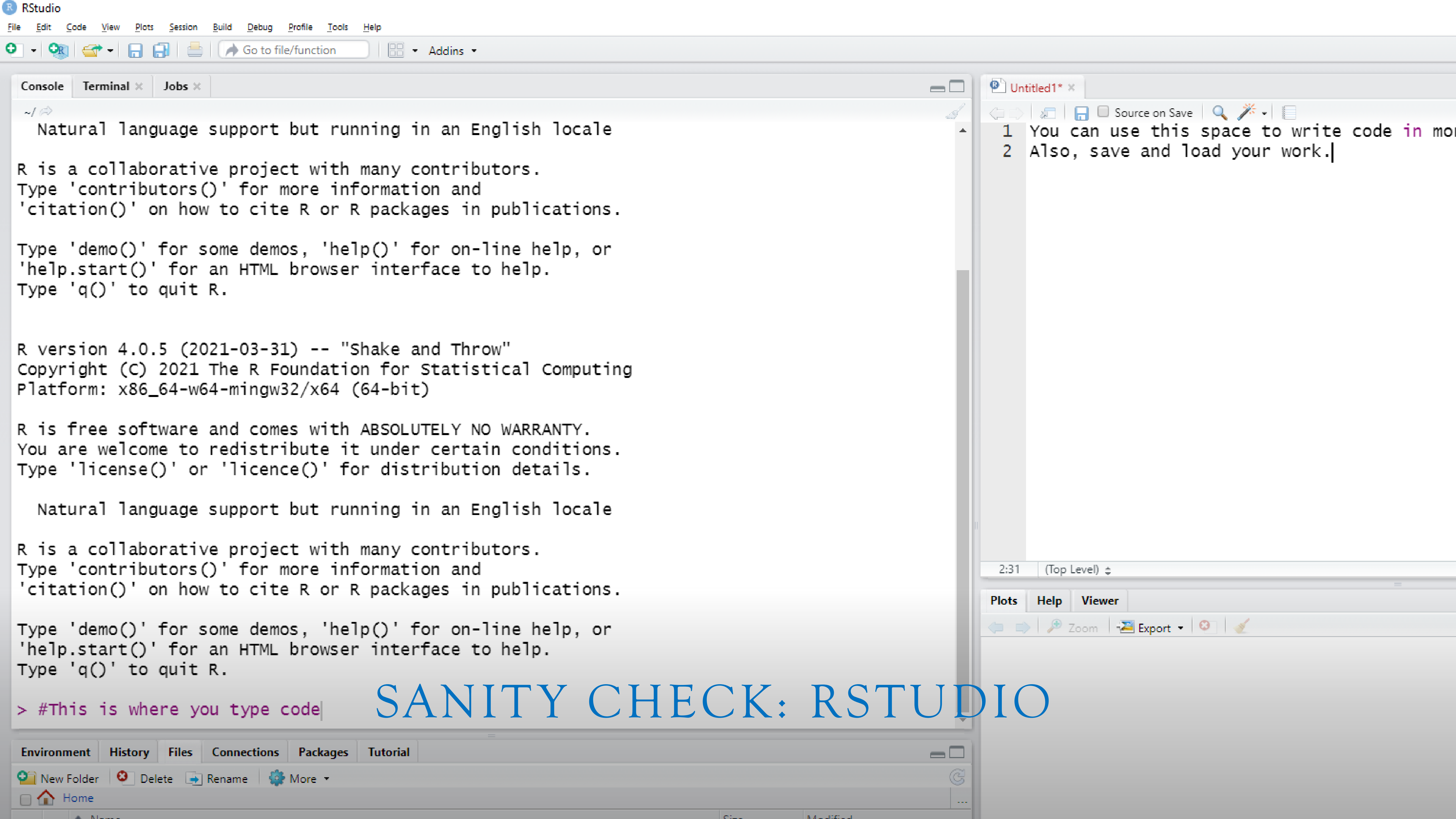
SANITY CHECK: BASE R



```
R version 4.0.5 (2021-03-31) -- "Shake and Throw"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

STEP 2: LET'S GET FANCY! (RSTUDIO)

- While optional, Rstudio is the best thing since sliced bread
- It is an integrated development environment or IDE
- In short, it gives you a lot of quality of life features you'll learn to appreciate over time. (Just trust me)
- Go to: Download the RStudio IDE – Rstudio
  - Should auto select Windows or Mac
  - You want "Free" version
  - Download and install



```
  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.


R version 4.0.5 (2021-03-31) -- "Shake and Throw"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #This is where you type code
```

```
You can use this space to write code in mo
Also, save and load your work.
```

SANITY CHECK: RSTUDIO

STEP 3: LET'S INSTALL SOME PACKAGES!

- A package is a collection of functions
  - Think plugins, add-ons, or apps
- They enhance what R can do
- One of the coolest and most powerful things about R. Feel free to explore and experiment…but also maybe play around with just Tidyverse at first. It is a deep rabbit hole!

1. install.packages("tidyverse")
2. Press enter (or run). Should see a bunch of stuff start to happen. That is normal.
3. **You only have to install packages once, but you always have to "load" them whenever you start a new session.**

STEP 4: LOADING PACKAGES

- **library(tidyverse)**
- #> ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
- #> ✔ ggplot2 3.3.2   ✔ purrr   0.3.4
- #> ✔ tibble 3.0.3   ✔ dplyr   1.0.2
- #> ✔ tidyr   1.1.2   ✔ stringr 1.4.0
- #> ✔ readr   1.4.0   ✔ forcats 0.5.0
- #> ── Conflicts ── tidyverse_conflicts() ──
- #> ✖ dplyr::filter() masks stats::filter()
- #> ✖ dplyr::lag()   masks stats::lag()

WRAP-UP: A QUICK EXERCISE

- Tidyverse comes with a built-in dataset called “mpg”
  - “mpg” is a data frame that has variables collected by the EPA on 38 models of car
- Type: mpg
- You see a tibble (tidyverse’s way of showing data, aka a dataframe)
- Two key variables:
  - displ, a car’s engine size, in litres.
  - hwy, a car’s fuel efficiency on the highway, in MPG.

THE LEGENDARY: GGPLOT

- ggplot(data = mpg) +

  geom_point(mapping = aes(x = displ, y = hwy))
- You always need three things with a ggplot
  - Data (data = mpg)
  - Aesthetics aes(x = variable1, y = variable2)
  - Geometry (geom_point()) → scatter plot
- What happens?
- What is the big take-away?

RESOURCES:

- The R bible: Welcome | R for Data Science (had.co.nz)
  - Hadley Wickham is the truth and this book will show you the way.

- The link of all links: 2 New to R? Start here | Big Book of R

- Learn R, Python & Data Science Online | DataCamp
  - My recommended online resource for hands-on coaching/training.