How Building a MicroSaaS in R Might Make You a Better Data Scientist

Get In Touch

[TEST SCREEN]

David John Baker, PhD

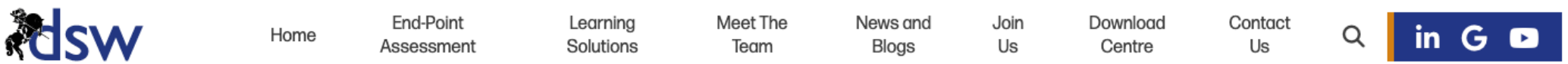
Redgranite Group

Tableau User Group Meeting, March 2025

Upskilling as a Data Scientist

Solving real world problems, not finishing assigned projects is what makes you a better data scientist

Why on mind?



Research Scientist – Level 7

Standard Number
ST0759 - Version 1.0

Maximum Funding
£18,000

Priced From £2,520
Discounts available based on volume

Standard Overview | Facilitation of Assessment | Assessment Methods | Grading | Resources | Benefits and Support

Standard Overview

This occupation is found in a wide range of industries including Pharmaceutical, Clinical Trials, Personal Care, Analytical, Manufacturing, Water/Environmental, Energy, Agricultural, Food Science, FMCG, Petro-Chemical, Nuclear, Aerospace, Oil, Gas, Materials, Renewable, Bio medical, NHS, Diagnostics and MOD/Defense.

The broad purpose of the occupation is someone who is primarily involved in planning, leading and conducting experiments and analysing results, either with a definite end use, for example to develop new products, processes or commercial applications, or to broaden scientific understanding in general.

Why upskill?

- Technology landscape is always changing
- Double Edged Sword
  - Intellectually stimulating
  - Anxiety inducing, especially if you’re more junior
- Juniors: need to differentiate yourself
  - LLMs
  - Mass Layoffs
- **Make your own life easier, more hireable, more value to colleagues and organization**

Not enough just to tell, must show

- Showing with **Projects** proof
  - Externally motivated (a.k.a. boss tells you what to do)
  - Data already exists (a.k.a pre-packaged data)
  - Tools > Everything Else
  - “Solution in search of a problem”
- Nothing wrong with this, siloed for a reason
- *Alternative approach if you want to really improve your skillset*

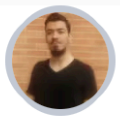
Problems, not projects

- Showing proof by solving **Problems**
- Intrinsically motivated: Unbounded problem space (a.k.a **you** need to decide the bounds)
- Data does not already live in a CSV or database
- Problem drives tech choices
- “Problem in search of solution”
- Basically… find a small real world problem and solve it

What is a MicroSaaS?

What is a MicroSaaS?

What is a MicroSaaS?



> *"A small SaaS company from a solo entrepreneur or a small team that's typically bootstrapped and going after a niche market. It's also prioritizing profits and business sustainability as opposed to growth."*

As you can see, it's very similar to a regular SaaS business except that the creator of a micro SaaS isn't trying to apply for Y Combinator or build "the next big thing".

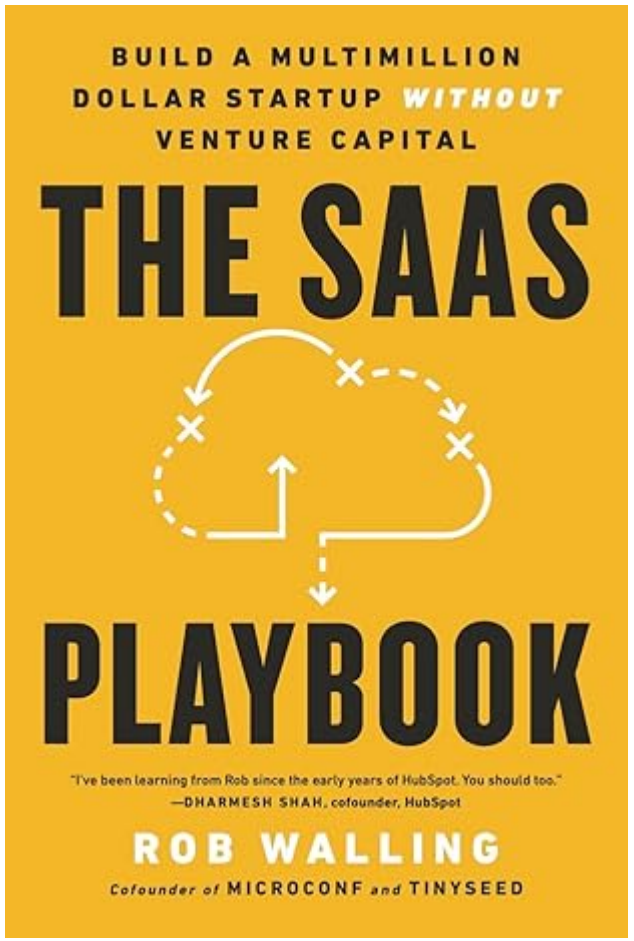
This is a business model that is more in tune with the way indie builders work.

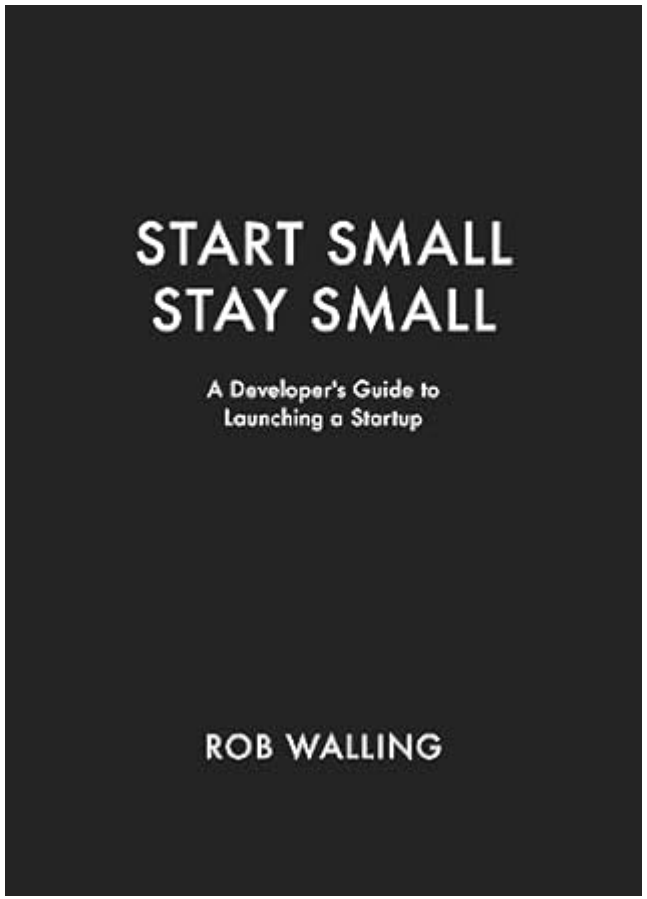
A micro SaaS can start as an experiment or a way to capitalize on a specific need from a niche. The idea here is to solve the need in the best possible way while making the most out of the available resources.

Where traditional startups are all about growth and scaling, micro SaaS companies focus on consistent profits, a loyal group of users, and a lean business operation.

It's a much more "beginner-friendly" approach to building a SaaS product without taking on too many of the risks that a traditional startup would have.

Rob Walling's Books

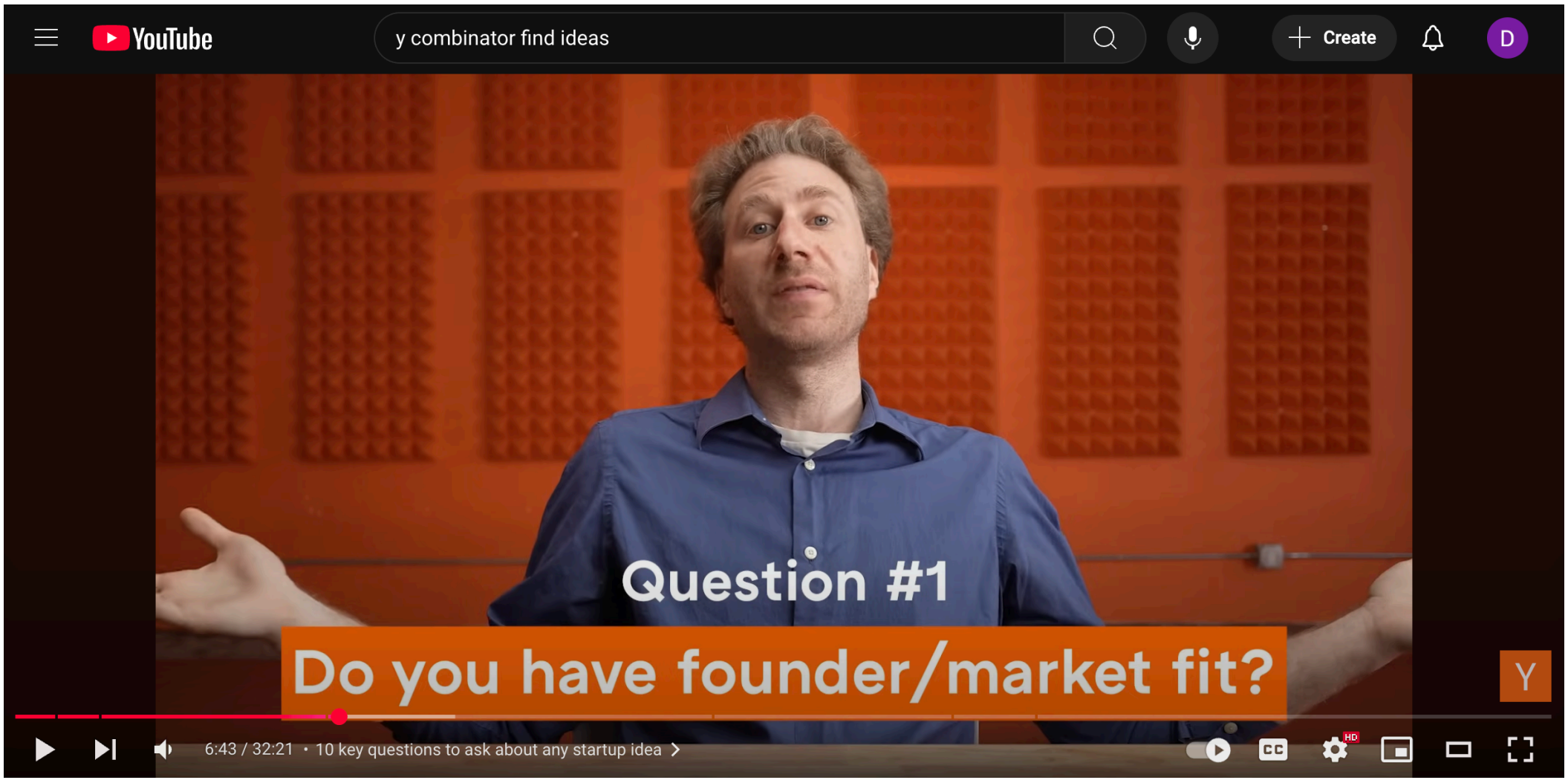




Thesis in a Nutshell

1. Try to build a MicroSaaS
2. Get forced to wear more than your data scientist hat
3. See your data science work through a novel lens
4. (Maybe use R to do it?)

What problem can I solve?



Keyword: *I*

- Recent freelance fractional CFO work
- Knowledge and network in performing arts
- Recently read book on how churn affects revenue in arts
- Asked network what kinds of problems existed for real people in this space

Analyze ticketing data so performing arts organizations can understand their

Truly, what are the specs of an MVP?

Requirements:

1. Upload historical ticket data
2. Faster and more streamline than Excel
3. Output easily understood by non-technical user
4. Maximum two months (Jan, Feb of 2025)

What do I want out of this?

- No idea if this would work or not
- Establish principles at start so however it goes, it’s investment in self
  - (Goal: become a better data scientist)

Wish List

- Improve my ability to make interactive dashboards
  - (mostly Quarto before, never serious shiny)
- See what all the AI LLM hype is about to assist with coding
- Productionize what I built
- Write code in a way that future-me (and future employers) would approve of

Developing the MVP

Why R?

- Open Source Programming Language
- Suite of tools that centralize data centric development
  - *tidyverse*, Posit
- I can work very quickly in it

Tidyverse Crash Course

Import and Assignment

```
1 library(readr)
2 library(janitor)
3 library(tidyverse)
4 library(knitr)
5 library(kableExtra)
6
7
8 superstore_orders <- read_csv("data/superstore_orders.csv")
```

Superstore Orders

Category	City	Country/Region	Customer ID	Customer Name	Order Date	Order ID
Office Supplies	Houston	United States	DP-13000	Darren Powers	1/3/2021	US-2021-103800
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Philadelphia	United States	MB-18085	Mick Brown	1/5/2021	US-2021-141817
Furniture	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Athens	United States	JO-15145	Jack O'Briant	1/6/2021	US-2021-106054
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Los Angeles	United States	LS-17230	Lycoris Saunders	1/6/2021	US-2021-130813
Technology	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Technology	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Furniture	Huntsville	United States	VS-21820	Vivek Sundaresam	1/7/2021	US-2021-105417

The Pipe: |>

- Pipe read as “and then”
- Start with the table called `superstore_orders`
- And then
- Use the function `clean_names()` on it
- `f(x)` becomes `x |> f()`

```
1 superstore_orders <- read_csv("data/superstore_orders.csv")
2
3 clean_names(superstore_orders)
4
5 superstore_orders |>
6   clean_names()
```

Cleaner Names

category	city	country_region	customer_id	customer_name	order_date	order_id
Office Supplies	Houston	United States	DP-13000	Darren Powers	1/3/2021	US-2021-103800
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Naperville	United States	PO-19195	Phillina Ober	1/4/2021	US-2021-112326
Office Supplies	Philadelphia	United States	MB-18085	Mick Brown	1/5/2021	US-2021-141817
Furniture	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Athens	United States	JO-15145	Jack O'Briant	1/6/2021	US-2021-106054
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Office Supplies	Los Angeles	United States	LS-17230	Lycoris Saunders	1/6/2021	US-2021-130813
Technology	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Technology	Henderson	United States	ME-17320	Maria Etezadi	1/6/2021	US-2021-167199
Furniture	Huntsville	United States	VS-21820	Vivek Sundaresam	1/7/2021	US-2021-105417

Select and Filter

```r
superstore_orders <- read_csv("data/superstore_orders.csv")

cleaned_data <- superstore_orders |>
  clean_names() |>
  select(category, region, quantity) |>
  filter(region != "South")
```

Column and Rows Logic

category	region	quantity
Office Supplies	Central	2
Office Supplies	Central	2
Office Supplies	Central	3
Office Supplies	Central	3
Office Supplies	East	3
Office Supplies	West	3
Furniture	Central	3
Office Supplies	Central	7
Office Supplies	Central	2
Technology	Central	3
Furniture	East	2
Furniture	West	3
Office Supplies	West	5
Office Supplies	East	2
Office Supplies	East	7

Group By and Summarise

```r
cleaned_data <- superstore_orders |>
  clean_names() |>
  select(category, region, quantity) |>
  filter(region != "South") |>
  group_by(category, region) |>
  summarise(total_quantity = sum(quantity))
```

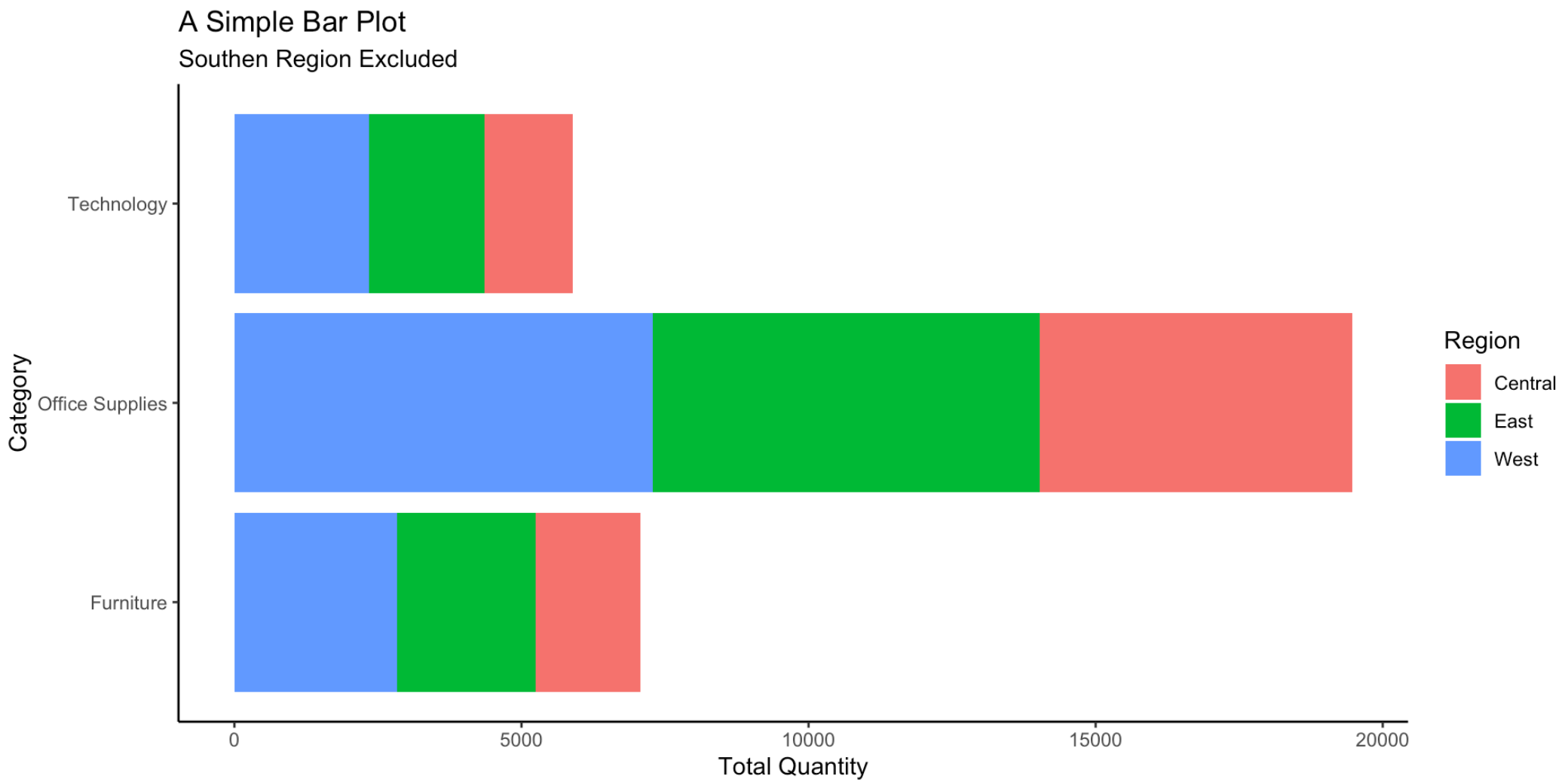
Summary Table

category	region	total_quantity
Furniture	Central	1838
Furniture	East	2414
Furniture	West	2826
Office Supplies	Central	5438
Office Supplies	East	6740
Office Supplies	West	7290
Technology	Central	1544
Technology	East	2005
Technology	West	2350

Plot It

```r
my_plot <- cleaned_data |>
  ggplot(aes(x = category, y = total_quantity, fill = region)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "A Simple Bar Plot",
       subtitle = "Southen Region Excluded",
       x = "Category",
       y = "Total Quantity",
       fill = "Region") +
  theme_classic()
```

Final Plot


A Simple Bar Plot
Southen Region Excluded
Category
Technology
Office Supplies
Furniture
0
5000
10000
15000
20000
Total Quantity
Region
Central
East
West

Posit Tools

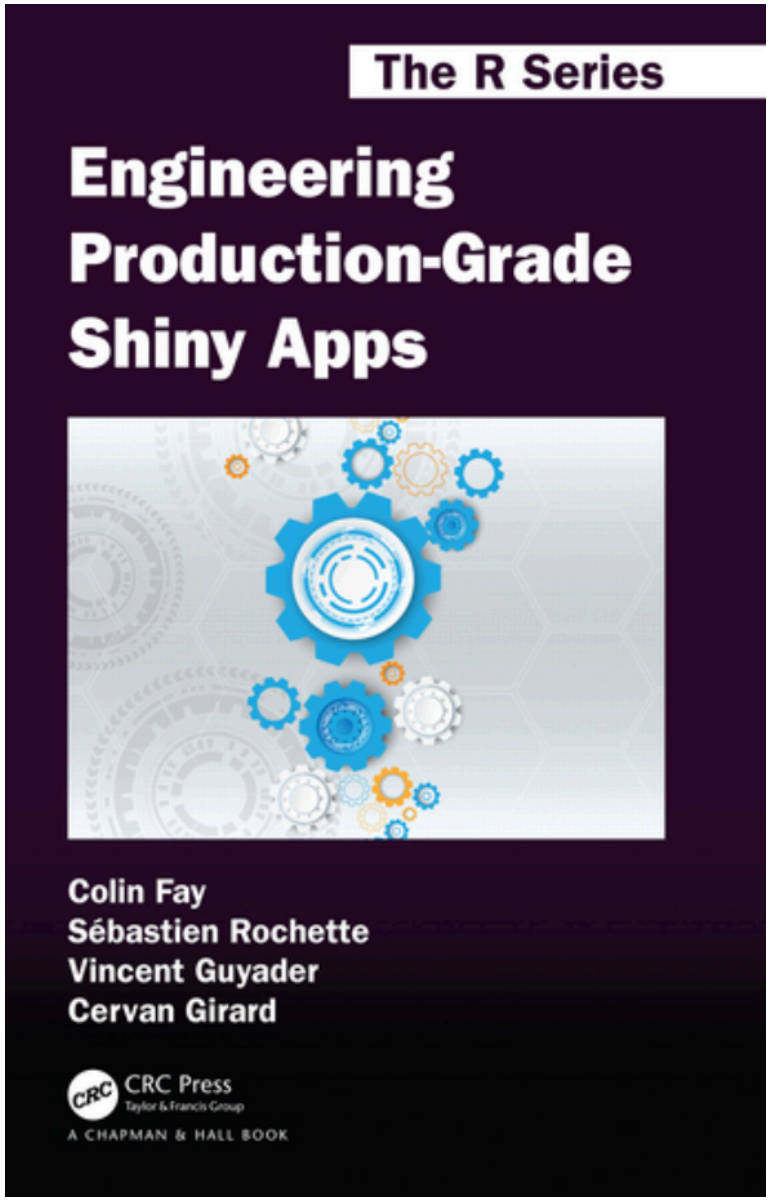
- Quarto
  - Static and Parameterized Reports
- Shiny
  - Interactive Dashboards
- Shiny Assistant
  - LLM for Shiny

Production during the process

- Spend minimal time on MVP
- Faster in someone’s hands, the better
- Write in environment where it’s easy to deploy
- Keep **production** in mind

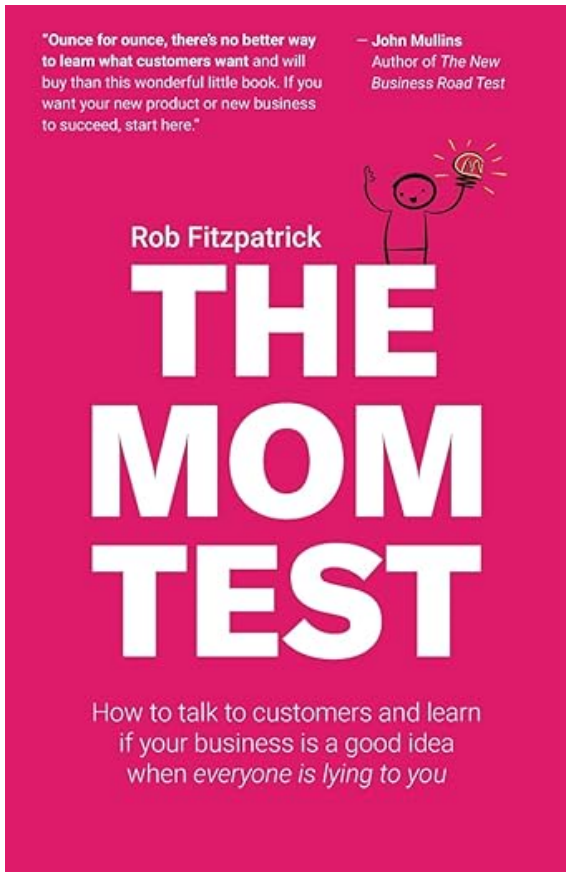
Production for Users and Developers

- It is used, even if only by one person.
- It is relied upon by its user(s).
- It has real life impact if something goes wrong.

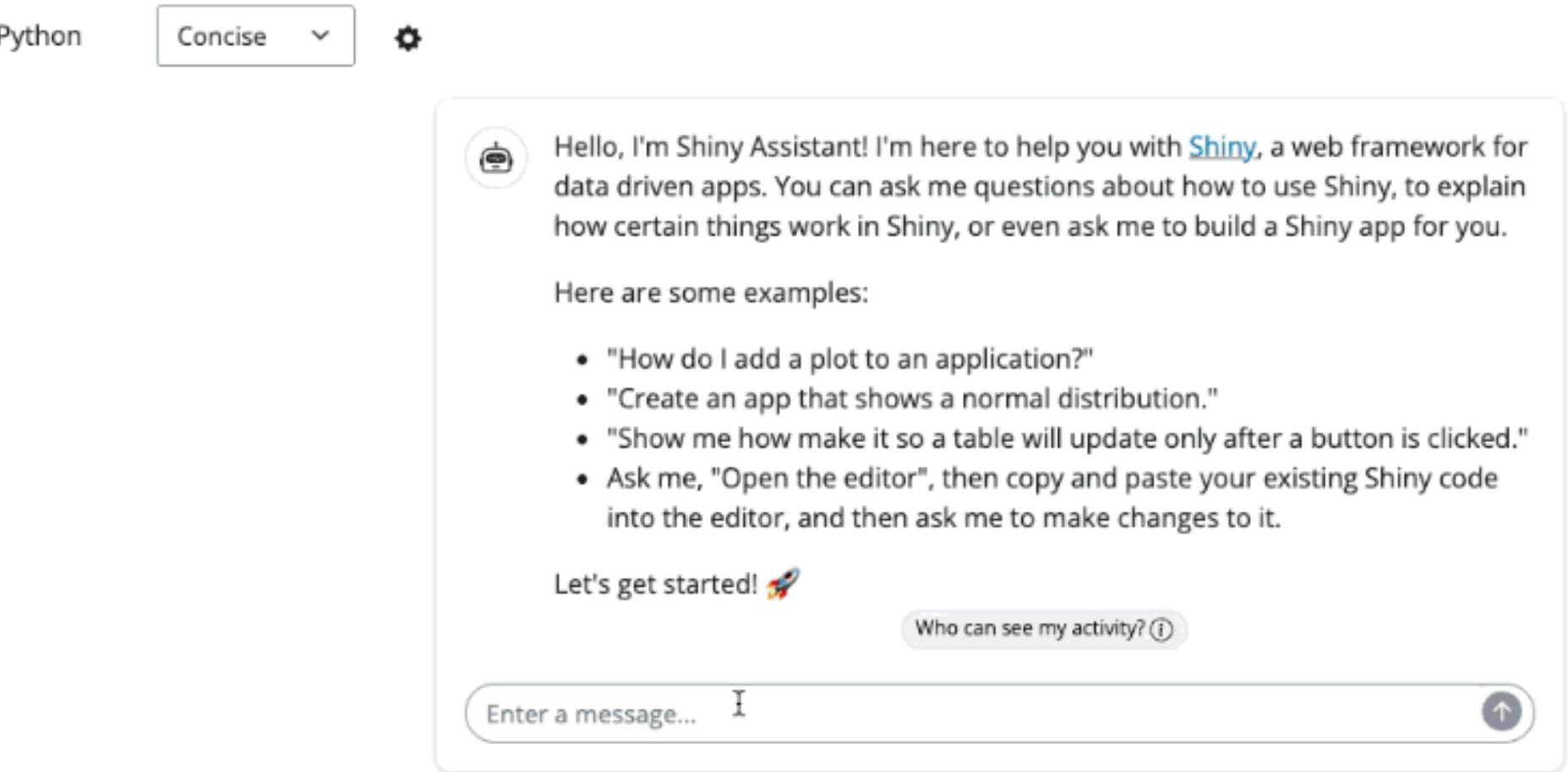


Getting Buy-In, AKA Living in Reality

- Goal during development: get buy in
- Do this fast
- Age of LLMs, use this to your advantage

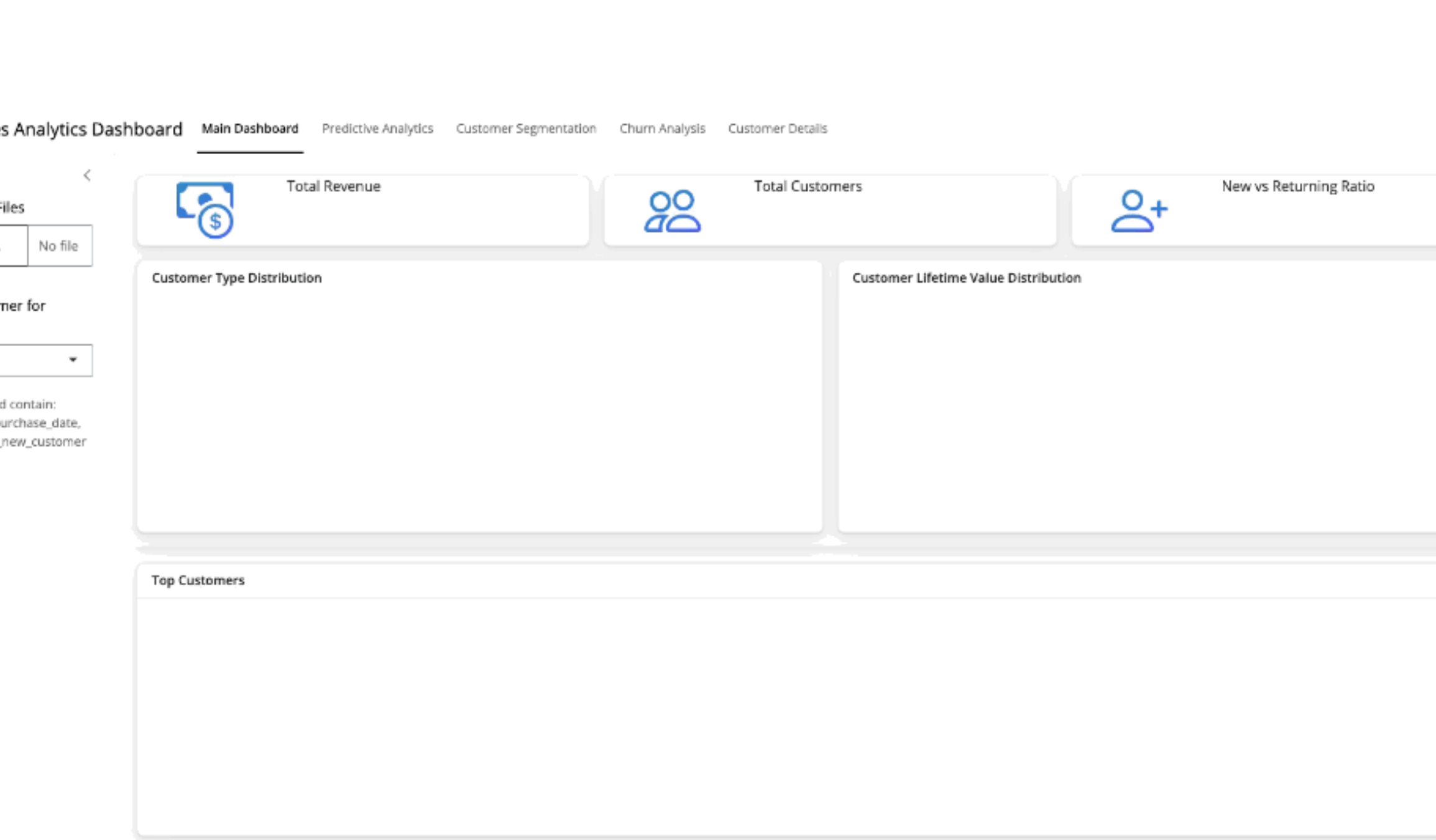


Move fast using LLMs


Python
Concise
Hello, I'm Shiny Assistant! I'm here to help you with Shiny, a web framework for data driven apps. You can ask me questions about how to use Shiny, to explain how certain things work in Shiny, or even ask me to build a Shiny app for you.
Here are some examples:
"How do I add a plot to an application?"
"Create an app that shows a normal distribution."
"Show me how make it so a table will update only after a button is clicked."
Ask me, "Open the editor", then copy and paste your existing Shiny code into the editor, and then ask me to make changes to it.
Let's get started!
Who can see my activity?
Enter a message...

Getting buy-in from small orgs, on right track?

- Posted video of 45 minutes playing with Shiny Assitant
- Group of Arts Admin people
- Looking for data trade, insights for data
- Got 10 years of data from Orlando Family Stage


s Analytics Dashboard
Main Dashboard
Predictive Analytics
Customer Segmentation
Churn Analysis
Customer Details
Files
No file
ner for
d contain:
urchase_date,
new_customer
Total Revenue
Total Customers
New vs Returning Ratio
Customer Type Distribution
Customer Lifetime Value Distribution
Top Customers

Iteration Alpha

- Two weeks of development
  - (refreshing Shiny knowledge)
- Feedback
  - Made worse version of their CRM
  - Want analysis
  - Want things they couldn’t do with basic tools

Bolero

Overview of data used to generate this report, current season's revenue progress, and essential KPIs.

Season-on-Season Revenue Progress

Bolero Score
73

Health
Amber

YoY Revenue Progress
53 %

Days Left in Season
88

Years of Data
14

Historical Transactions
114,587

Shows
4,413

Patrons
47,283

Take Aways

MicroSaaS allows you to own the entire process

- Pro: Forces you to consider the “so what” from Day 1
- Con: Context switching is difficult
- Pro: Wearing your “product” hat changes how you see your development
- Con: Generate a lot of technical debt if you commit to moving fast
- Pro: Live in an time with freak-level tooling and access to learning

Tether yourself to reality

- Long road from problem to payment
- Set self up for success by:
  - How can I work in way that even if it fails, I win?
  - Getting your ideas in front fast
- Real people’s real problems
- Skin-in-Game milestones for success
  - Introductions
  - Trading of resources
  - Money

Practice kindness to future you

- Develop with “what if this does work?” mindset
- Finding balance between moving fast and being kind to future you
- Code is meant to be read by people
  - (e.g. you, future colleagues)

Next Steps

Launch

- WeInc. Marketing hosts launch in Spring
- Potential for small revenue
- Continued learning
  - Data engineering
  - Sales
  - Delegating

Open to Work!

- Spent last year freelancing
- Currently looking for **full time** or **long contract work**
- Would love to talk to anyone hiring **data scientists**

davidjohnbaker1@gmail.com

https://www.linkedin.com/in/david-john-baker-phd/

@davidjohnbaker.bsky.social

Abstract

Building a MicroSaaS forces data scientists to think beyond analysis and requires developing skills in software engineering, writing code that can be used in production, and encourages higher level business thinking. In this talk, I'll demonstrate how tools from the R (mostly Posit!) ecosystem– like R Shiny– can be used to develop a minimum viable product using a ticket sales analysis application. I highlight how this process not only improves technical skills, but also enables rapid development of business objectives.