

Generative AI for Data Science 101: Coding Without Learning to Code

Jacob Bien

Data Sciences and Operations
University of Southern California
May 2024

*Joint project with
Gourab Mukherjee*

Retreat for Stats/Biostats Curriculum Reform

A perennial dilemma

Teach coding in intro stats?

To code or not to code?

Pro

- * Working with data is **fun**; can inspire students to want to learn more
- * Better prepares students for working with non-textbook datasets

Con

- Teaching coding takes a lot of time
- Distracts from statistical material, which is already difficult
- If too hard, will students become disillusioned?

Full-Time MBA

FTMBA

A full-time two-year MBA program for early career professionals to sharpen business fundamentals, learn to leverage technology for business and social impact, and empower leadership of diverse global teams. Opportunities for networking and specialization enable students to use the MBA to transform their career.

Marshall is located in one of the most vibrant global economies and is part of a large, diverse university ready to engage the world's demanding challenges. The Trojan family, known for its strong network, includes more than 450,000 alumni, where 90,000+ are from the Marshall School. Our campus community includes a world-class faculty, high performing students and a dedicated professional staff.

- ~200 students
- Wide-ranging technical backgrounds
- Work experience: nonprofits, media, retail, finance, consulting, ...



GSBA 524: Data Science for Business

- Required, first-semester core course
- 3 sections, ~65 students/section
- Co-taught with Gourab Mukherjee



To code or not to code?

Approaches

No software


























Point-and-click

Modify code chunks

Excel

Teach coding

Approaches

	Work with data?	Easy to teach?	Extend to new operations?	Used by data scientists IRL	Language agnostic?
No software					
Point-and-click					
Modify code chunks					
Excel					
Teach coding					

Approaches

	Work with data?	Easy to teach?	Extend to new operations?	Used by data scientists IRL	Language agnostic?
No software					
Point-and-click					
Modify code chunks					
Excel					
Teach coding					
Copilot					

Past years: Point-and-click

The screenshot shows the Radiant web interface. The top navigation bar includes 'Radiant', 'Data', 'Design', 'Basics', 'Model', 'Multivariate', 'Report', and 'Base dir: Home'. A dropdown menu is open under 'Model', listing various estimation methods: Estimate (Linear regression (OLS), Logistic regression (GLM), Multinomial logistic regression (MNL), Naive Bayes, Neural Network), Trees (Classification and regression trees, Random Forest, Gradient Boosted Trees), Evaluate (Evaluate regression, Evaluate classification), Recommend, and Collaborative Filtering. The 'Logistic regression (GLM)' option is highlighted. On the right, a data table is visible with columns for 'depth', 'table', 'x', 'y', 'z', and 'date'. The table contains several rows of data, including values like 61.00, 56.00, 4.43, 4.45, 2.71, and 2012-02.

Datasets:

diamonds

Add/edit data description

Rename data

Display:

preview str summary

Load data of type:

rds | rda | rdata

Load

Save data to type:

rds

Save

Estimate

- Linear regression (OLS)
- Logistic regression (GLM)**
- Multinomial logistic regression (MNL)
- Naive Bayes
- Neural Network

Trees

- Classification and regression trees
- Random Forest
- Gradient Boosted Trees

Evaluate

- Evaluate regression
- Evaluate classification

Recommend

- Collaborative Filtering

depth	table	x	y	z	date
61.00	56.00	4.43	4.45	2.71	2012-02
63.40	57.00	4.45	4.42	2.81	2012-02
63.10	58.00	4.27	4.23	2.68	2012-02
59.20	56.00	4.60	4.65	2.74	2012-02
62.60	58.00	4.72	4.68	2.94	2012-02
62.50	53.70	5.35	5.43	3.38	2012-02
61.70	56.00	6.14	6.18	2.80	2012-02

Radiant (Vincent Nijs)

Fall 2023: Try something new

- **Idea:** Use AI to get MBAs coding... without teaching them how to code??

Github Copilot

- **What:** Extremely capable auto-complete for code
- **How:** LLM for code by OpenAI, trained on Github
- **Marketed as:** “AI pair programmer”
- **Our off-label use:** English-to-code translator

redfin.R*

Source on Save

```

1 # Load the tidyverse package
2 library(tidyverse)
3
4

```

Console Background Jobs

```

R 4.3.0 · ~/Dropbox/talks/talk_teaching-with-copilot/
>
>
>
>
>
>
>
>
>

```

Files Plots Packages Help Viewer Pre

ilot > michigan-stat-curriculum-reform > demos

	Name	Size
	..	
<input type="checkbox"/>	demos.Rproj	205 B
<input type="checkbox"/>	redfin.csv	94 KB
<input type="checkbox"/>	redfin.R	76 B

3:1 (Top Level) Copilot: No completions available.

Environment History Connections Tutorial

Import 114 MiB List

R Global Environment

Environment is empty

Teaching to code without teaching syntax

- **Principle #1:** Be specific.
- **Principle #2:** Think about context window.
- **Principle #3:** Break down complex operations into simple steps.

(Formulated as semester progressed)

redfin.R*

```
1 # Load the tidyverse package
2 library(tidyverse)
3
4 # Load redfin.csv
5 redfin <- read_csv("redfin.csv")
6
7 # Show the names of the variables
8 names(redfin)
9
10 # Show the first 6 rows of the data
11
```

10:2 (Top Level) Copilot: Completion response received.

Environment History Connections Tutorial

R Global Environment

Data

redfin 350 obs. of 27 variables

Console Background Jobs

R 4.3.0 · ~/Dropbox/talks/talk_teaching-with-copilot/

```
[23] "MLS#"
[24] "FAVORITE"
[25] "INTERESTED"
[26] "LATITUDE"
[27] "LONGITUDE"
>
> |
```

Files Plots Packages Help Viewer Pre:

Observations

- **Principle #1:** Be specific.

Student must specify exact variable names.

- **Principle #2:** Think about context window

Copilot's context window did not include variable names.

- **Principle #3:** Break down complex operations into simple steps

However, once "PRICE" was mentioned in file, one could assume that knowledge.

Observations

- **Principle #1:** Be specific.
- **Principle #2:** Think about context window
- **Principle #3:** Break down complex operations into simple steps



Next example

redfin.R*

Source on Save

```

10 # Make a scatter plot of PRICE vs.
    "SQURE FEET"
11 ggplot(redfin, aes(x = `SQURE FEET`, y =
    PRICE)) +
12   geom_point()
13
14 # Show the condo with price above 3.5
    million
15 redfin %>%
16   filter(PRICE > 3500000)
17
18 # What is its address?
19 redfin %>%
20   filter(PRICE > 3500000) %>%
21   select(ADDRESS)
22

```

22:1

(Top Level) Copilot: No completions available.

R S

Environment

History

Connections

Tutorial

Import

222 MiB

List

R Global Environment

Data

redfin

350 obs. of 27 variables

Console

Background Jobs

R 4.3.0 · ~/Dropbox/talks/talk_teaching-with-copilot/

```

+ filter(PRICE > 3500000) %>%
+ select(ADDRESS)
# A tibble: 1 × 1
  ADDRESS
  <chr>
1 2025 Avenue Of The Stars #1202
>

```

Files

Plots

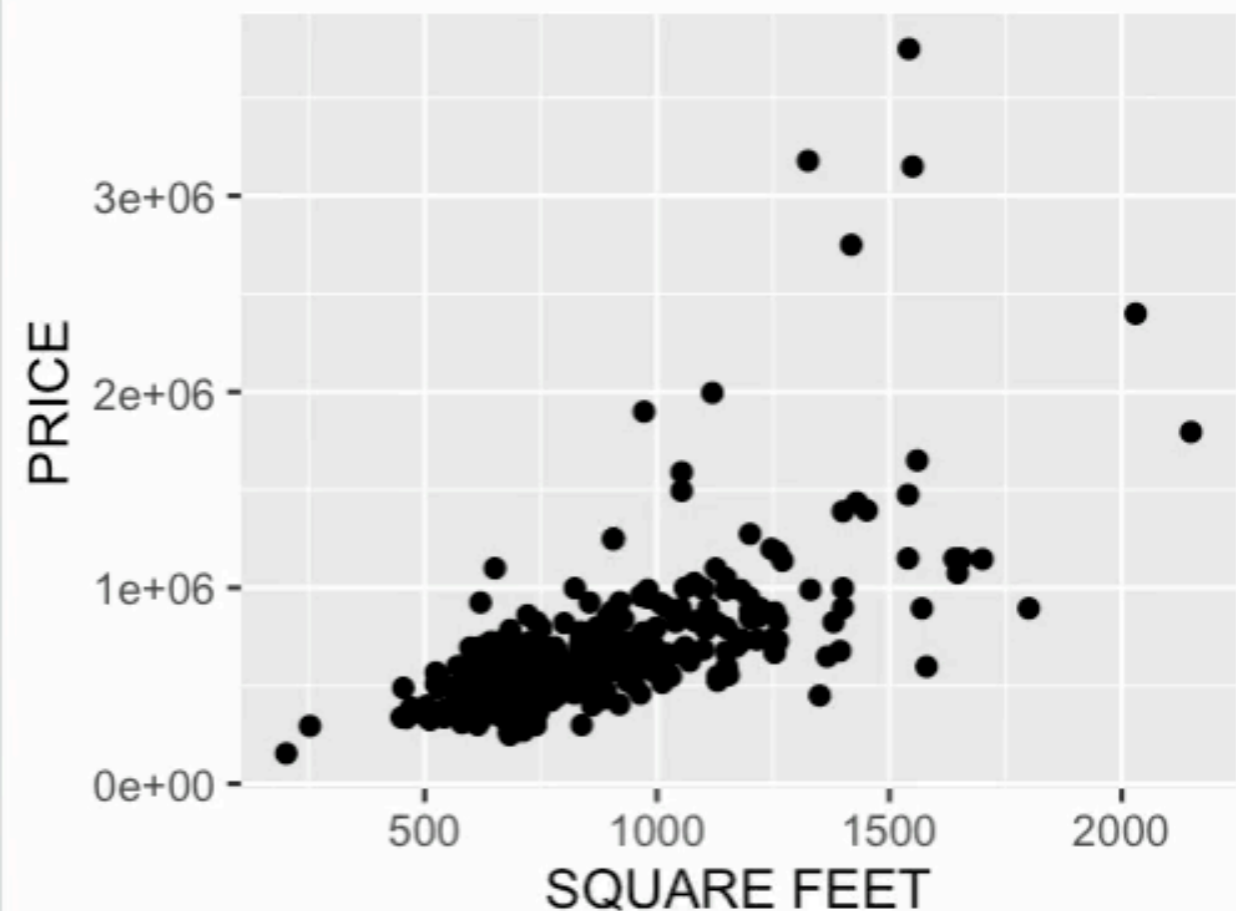
Packages

Help

Viewer

Pre

Zoom Export



Challenges

- Copilot's outputs are random!
- For famous data sets, Copilot seems to read your mind!
- Lack of transparency and an evolving product

Benefits

- Enables students to work flexibly with data
- Allows curiosity to drive data exploration, unconstrained by some limited list of operations
- A student could take what we taught in R and do it in Python with no change — **syntax is irrelevant!**

Thanks!

arXiv > stat > arXiv:2401.17647

Statistics > Other Statistics

[Submitted on 31 Jan 2024]

Generative AI for Data Science 101: Coding Without Learning To Code

Jacob Bien, Gourab Mukherjee

<https://arxiv.org/abs/2401.17647>

Questions?

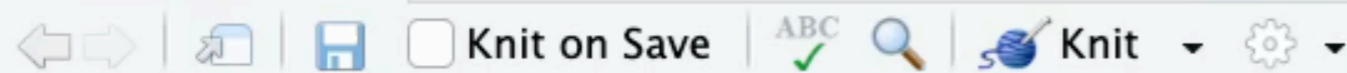
jbien@usc.edu



Works in R Markdown



redfin.Rmd*



Source

Visual

Outline

```

1 ---
2 title: "Redfin Data Analysis"
3 author: "Jacob Bien"
4 date: "2024-05-15"
5 output: html_document
6 ---
7
8 |
9

```

8:1

(Top Level)

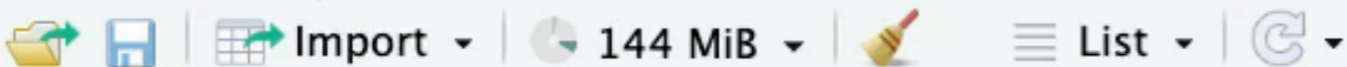
Copilot: Completion response received.

Environment

History

Connections

Tutorial



R Global Environment

Environment is empty

Console

Background Jobs

R 4.3.0 · ~/Dropbox/talks/talk_teaching-with-copilot/

>

>

>

>

>

>

>

Files

Plots

Packages

Help

Viewer

Pre:



ilot > michigan-stat-curriculum-reform > demos

Name

Size



..



demos.Rproj

205 B



redfin.csv

94 KB



redfin.R

1.1 KB



redfin.Rmd

100 B

Redfin Data Analysis

Jacob Bien

2024-05-15

In this notebook, we explore the redfin data. Let's start by loading the tidyverse package.

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
## Warning: package 'tidyr' was built under R version 4.3.1
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
## Warning: package 'stringr' was built under R version 4.3.1
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ dplyr      1.1.4      ✓ readr      2.1.4
## ✓ forcats   1.0.0      ✓ stringr    1.5.1
## ✓ ggplot2   3.5.0      ✓ tibble     3.2.1
## ✓ lubridate 1.9.2      ✓ tidyr      1.3.1
## ✓ purrr     1.0.2
## — Conflicts — tidyverse_conflicts() —
## * dplyr::filter() masks stats::filter()
## * dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Let's load the file, redfin.csv.

```
redfin <- read_csv("redfin.csv")
```

```
## Rows: 350 Columns: 27
## — Column specification —
## Delimiter: ","
## chr (15): SALE TYPE, SOLD DATE, PROPERTY TYPE, ADDRESS, CITY, STATE OR PROVI...
## dbl (12): ZIP OR POSTAL CODE, PRICE, BEDS, BATHS, SQUARE FEET, LOT SIZE, YEA...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```